4 Classical Inference I: Estimators

Statistical inference is a mathematical activity, or process, intent on extracting meaningful information from measurements in order to characterize systems, establish best model parameters, or determine the plausibility (truthfulness) of hypotheses and models. Indeed, once experimental data have been acquired, one wishes either to characterize the data by some ad hoc functional form, or compare the measured data distribution to some model predictions. In either case, the functional form may involve "free" parameters not constrained by theory and to be determined by the data. There are also cases where no specific functional form is known a priori and as simple a characterization of the data as possible is sufficient or required. It may be sufficient, for instance, to evaluate the mean, variance, and selected higher moments of the data. Inference also includes the notion of statistical test, that is, the use of statistical techniques to assess the adequacy of scientific models and hypotheses.

Traditionally, statistical inference is a subfield of probability and statistics concerned with the determination of the properties of the underlying distribution of datasets based on statistical analysis techniques largely based on the frequentist interpretation of probability. However, the field of inference has greatly expanded in recent decades, and inference may now be approached from a plurality of paradigms,[1] or schools of thought, corresponding to different interpretations of the notion of probability and approaches to the inference process. We will restrict our discussion to the frequentist and Bayesian paradigms, the former still most commonly used in the physical sciences, and the latter receiving a fast growing level of interest in physical and life sciences, engineering, as well as economics and financial fields. Other commonly less adopted paradigms, including the likelihood paradigm, the total survey error paradigm, and the survey sampling paradigm, are discussed in the statistics literature [15, 96, 167].

Similarities and distinctions between the goals and methods of the frequentist and Bayesian inference paradigms are briefly outlined in §4.1. The remainder of this chapter as well as Chapters 5 and 6 present detailed discussions of classical inference concepts and techniques while Bayesian inference methods are explored in Chapter 7.

[1] The term paradigm is here used in the sense defined by Kuhn [132]. Restricted to statistical inference, this amounts to classifying statisticians and scientists according to the inference framework they use in the conduct of their statistical analyses.

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

4.1 Goals of Frequentist Inference

Inference problems are traditionally classified according to the type of statements or conclusions they produce. These include:

1. Point estimates, which involve the determination of a particular observable, properties of a distribution, or model parameters designed to characterize a particular population
2. Interval estimates, most commonly known as confidence intervals within which a particular observable, property of a distribution, or model parameter would be found with a definite and specific probability if the experiment or sampling could be repeated indefinitely
3. Rejection of a hypothesis based on a statistical test predicated on a probabilistic model of the data
4. Clustering and classification of data points

By contrast, in the Bayesian paradigm the inference process essentially involves only two categories of tasks known as **model selection** and **parameter estimation** from which a plurality of subsidiary statements can be derived or summarized.

Data analyses in astronomy, particle physics, and nuclear physics rely increasingly on the Bayesian paradigm but nonetheless still make extensive use of the traditional frequentist methods. We thus endeavor to discuss a wide range of inference problems under both paradigms, and present, in this and the following chapters, discussions of inference techniques developed in both paradigms.

In traditional statistical inference, the characterization and modeling of datasets relies on mathematical functions or procedures known as **estimators**, which we formally define in §4.3. Essential properties of estimators are discussed in §4.4, while examples of implementation for basic notions of mean, variance, and covariance are presented in §4.5. Data modeling (commonly called **fitting**) based on the method of **Maximum Likelihood** ($\mathbb{ML}$) and the relatively simpler method of **Least Squares** ($\mathbb{LS}$) are introduced in §5.1 and §5.2, as are a number of associated technical problems, including interpolation/extrapolation, and weighted means of results of distinct experiments. A brief discussion of the notion of goodness-of-fit follows in §5.3. However, for a more in-depth discussion of the notion of statistical test applied to functional fits, see §6.6.5.

4.2 Population, Sample, and Sampling

Before proceeding with the introduction of the concepts of statistics and estimators, it is important to discuss the underlying notions of sample and population. Definitions of these concepts are presented in §4.2.1 while their characterization and representation are briefly discussed in §§4.2.2 and 4.2.3, respectively.

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

4.2.1 Population and Statistical Sample

In statistical analysis, a **population** is defined as a complete set of items or entities that share one or several properties. For instance, the population of Americans living in the continental United States shares a geographical location, a language, a political system, and so on. The concept of population is not restricted to humans; one may of course consider populations of animals, plants, or any other types of entities of interest. For instance, one may wish to study and characterize the fleet (population) of cars used in America, and contrast it to the fleets of cars used in Europe or Asia; or one might be interested in studying the population of stars composing the Milky Way. The notion of population, applied in statistics, is thus as broad and generic as human interests in studying the world around us.

By contrast, a **statistical sample** is a subset drawn from the population and used for the purpose of a statistical analysis. If the sample is sufficiently large and obtained properly, it is assumed that traits and characteristics of the entire population can be adequately inferred from corresponding characteristics of the sample.

It is useful to introduce the notion of **subpopulation** consisting of a specific subset of a population and defined based on one or more additional properties shared by the elements or items of this subset. For instance, the United States car fleet may be subdivided into several subpopulations according to the nationality (origin) of the manufacturers: Americans, Europeans, Asians, and so on. Items of a sample are in general heterogeneous and may not all share additional properties of interest. A sample must then be distinguished from a subpopulation.

Analyses in terms of subpopulations are useful, for instance, in pharmacology because different subpopulations (e.g., based on ethnic or hereditary history) may have different responses to medicines or treatments. In nuclear physics, one may distinguish the properties (e.g., momentum spectrum) of distinct particle subpopulations (e.g., pions, kaons, or protons) produced by a given type of elementary collision at a specific incident energy. In some cases, a great difference (variance) may exist between subpopulations. Modeling and understanding the population dynamics is then best accomplished by examining subpopulations individually. Distinctions between populations and subpopulations are somewhat arbitrary. A specific analysis may indeed focus on a particular subpopulation, which effectively becomes the "population of interest." The larger population, of which the given population is a subset, may then be referred to as the **parent population**. It is also common to refer the population from which a sample is drawn as the **parent population of the sample**.

4.2.2 Sample Characterization

It is typically not possible to acquire information about all elements of a population. The population may be too large or it may not be possible, ab initio, to access or enumerate all its elements. Population **sampling** is thus necessary. Sampling consists in selecting individuals or acquiring information about finitely many individuals or items of the population. The acquired subsets, called **samples**, may vary in size, but it is a general rule that the larger they are, the better the inference about properties of the whole population will be.

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

The sample space of a random variable X, defined in §2.2, corresponds to the entire space spanned by X. A population and measured samples of a population thus typically correspond to subsets of the sample space.

A **complete sample** is a set of items that includes all elements of a population.[2] As stated earlier, various practical or technical issues may limit the size of a sample. Statistical analyses involving complete samples are thus rare, in general. Since finite samples are typically the rule, statisticians must ensure that acquired samples are **unbiased**. Biases may occur, for instance, if subpopulations of a population feature distinct traits or behaviors in regard to an observable of interest. In this context, a sample consisting of elements drawn mostly from one particular subpopulation cannot be representative of the other subpopulations and thus the population as a whole.

An **unbiased sample** is a set of objects chosen from a complete sample using a drawing or sampling process that is not influenced by the properties of the elements of the population. Particular subpopulations are thus not favored or neglected and the ensuing statistical analysis and inference is thus representative of the whole population.

Truth be told, **sampling bias** occurs in all sciences and is often unavoidable. For instance, early studies of stars of the Milky Way greatly suffered from a brightness bias. Stars appear dimmer to an observer the farther away they are. Studies of stellar properties were thus initially biased, that is, influenced predominantly by classes of brighter stars. Similarly, studies of human populations may be biased based on geographical location, gender, religion, ethnicity, and so forth.

The problem arises, obviously, that it is generally not possible to draw complete samples, and one must devise techniques to avoid biasing collected samples. At the outset, it might appear that the collection of a **random sample**, also known as a **probability sample**, might solve this problem. In practice, observational biases are often introduced by the techniques or instruments used to acquire the data sample.

In astronomy, for instance, studies of galactic properties are somewhat biased by the observational wavelength window of the telescope: distant galaxies that recede from us at high velocity have large red shifts and may be difficult or impossible to measure with an optical (visible light) telescope.

In particle and nuclear physics, measurements of unbiased samples of events and produced particles are de facto impossible. Indeed, the triggering methods and devices used in the detection of collisions (events) and measurements of particle properties invariably feature technical limitations. Particles are measured within finite kinematic ranges, known as **detection acceptance**, and events may be missed if they fail to activate the devices used as trigger detectors. Collisional events acquired with detectors and procedures that minimally influence or bias the properties of measured events and the particles they contain are said to feature a **minimal bias**. By extension, triggering devices and procedures enabling the collection of such data are said to be **minimum bias triggers**. Interestingly, the volume of data produced in high-energy physics experiments is often so large that it is not possible to detect, digitize, and record all collisions. Special triggering detectors and procedures are then devised to bias and select collisions and particles of particular interest. It is then

[2] However, a complete sample is not required to cover the entire sample space of a random variable.

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

possible to selectively increase the relative abundance of specific types of particles or events (e.g., events susceptible of containing a top quark or a Higgs boson).

More generally, it is important to acknowledge that measurements can be influenced by various instrumental effects (discussed in Chapter 12) that cause loss or neglect of individual items of a population, or that smear the properties of these items. The loss or neglect of items is typically characterized by **sampling efficiency** and/or **detection efficiency** (commonly just called efficiency). Efficiency losses, if not uniform across all values of a particular observable, may severely bias the measured values and the conclusions reached about the population. In experimental physics, most particularly in nuclear and particle physics, particle and collision (event) losses are unavoidable and typically depend on observables of interest such as the momentum of particles, or the multiplicity produced in nuclear collisions. Fortunately, techniques exist to assess detection efficiencies and their dependencies on observables of interest, and it is then possible to correct sampled data for efficiency losses. Similarly, smearing or measurement resolution can also introduce measurement biases, but techniques to evaluate such effects are also available and may be used to correct the data. These topics are discussed at length in §12.3.

From the preceding discussion, it should be clear that a statistical analysis of a data sample is but a part of the measurement process. Obviously, measurement methods and the observables of interest vary greatly across scientific disciplines, but it remains that the sampling of a population and the statistical analyses based on collected samples are only a small part of the scientific process. Accordingly, techniques used to gather data and the types of data collected thus vary considerably across disciplines. However, statistical techniques to extract information about the parent population are essentially universal, that is, more or less independent of the specificities of the data or their origin.

4.2.3 Sample Representation and Random Variables

Statistical samples may consist of discrete data, continuous data, or mixtures of the two. They may be as simple as a collection of values $\{o_i\}$, with $i = 1, \ldots, n$, of a particular observable O (e.g., the momentum of particles produced by a certain type of nuclear collisions) representing a particular trait or characteristic of individuals of a population. They may also be arbitrarily complex and gather jointly (simultaneously) multiple observables representing traits or properties of individuals of a population. Computationally, such collections of observables may be represented as vectors $\vec{x}$ or n-tuples. An n-tuple is an ordered list of n elements that may be heterogeneous in nature or type. For instance, in nuclear collision studies, one may represent the information reconstructed about measured particles in terms of tuples $(q, p_T, \eta, \phi, \text{PID})$, where q is the charge of the particle, p_T its transverse momentum, η its pseudorapidity, ϕ the azimuthal angle of production, and PID a variable identifying the species of the particle.

In statistical analyses, one may be interested in characterizing a population based on one or more components of the n-tuples. More specifically, one may be interested in determining the moments, covariances (cumulants), and (probability) distributions of these components, either jointly or individually. Computational techniques required to obtain such quantities are the domain of statistical inference. We begin our discussion of statistical

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

inference techniques with the introduction of basic notions of statistics and estimators in §4.3. Properties of estimators are discussed in §4.4, and basic estimators presented in §4.5. Histograms, which provide graphical representations of the distribution of numerical data, are introduced and defined in §4.6. Classical inference techniques involving composite hypotheses and most particularly the optimization of model parameters toward the description of measured data are addressed in §§5.1 and 5.2.

4.3 Statistics and Estimators

4.3.1 Problem Statement

Let us consider an experiment consisting of a series of n distinct measurements of some phenomenon and yielding a dataset D. We will first assume that though distinct, the measurements are carried in the same conditions and from an unchanging system.[3] Let us consider that the dataset D consists of n measurements yielding random continuous vectors $\vec{x_i}$ of the same dimensionality. In the simplest case, each measurement would yield a single value x_i. In a slightly more sophisticated situation, the vectors could consists of pairs of numbers (x_i, y_i) considered as independent and dependent variables, respectively. More generally still, the vector $\vec{x}$ could represent an arbitrarily long list of values (i.e., an n-tuple) obtained from each measurement (i.e., each individual item of the sample). The sample space of each measurement would thus consist of all possible values of the multidimensional variable, $\vec{X}$, in other words, all values $\vec{x_i}$ a measurement of $\vec{X}$ can produce.

In a general scientific context, various states of prior knowledge may exist about the phenomenon being measured. If the phenomenon is newly observed, it is possible that very little prior information is available about it and one may then wish to characterize the sample (and the populations it is meant to represent) by its moments or other similar quantities. Alternatively, a great deal of information may already exist about the phenomenon, and one may wish to characterize the sample according to a specific mathematical model, or hypothesis, M, stated in terms of a function, $f(\vec{X}|\vec{\theta}, M)$, which expresses some relation between the elements of the vector $\vec{X}$. In this context, $\vec{\theta}$ could represent a finite set of parameters of the model M believed to properly describe the phenomenon, or being tested. For instance, in an analysis of the speed of receding galaxies, the model M could state the hypothesis that the relation between the recession speed of galaxies is strictly proportional to their distance. The model would then involve a single parameter consisting of the slope of the relation between the speed and the distance. It might then be the purpose of the analysis to infer which value or range of values of the slope, H_0, known as the Hubble constant, best fits the data. So, whether one wishes to characterize the population of interest according to moments or some more sophisticated model of the phenomenon, one is faced with a similar problem, which involves the determination of parameters that represent the whole

[3] We will see later in this chapter how these conditions may be lifted.

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

population based on a specific data sample D. This is known as a problem of **parameter estimation** and alternatively referred to as a **single point estimate**.

Parameter estimation may be discussed in both the frequentist and Bayesian paradigms. We first introduce the more traditional frequentist treatment in §4.3.2 and then show, in §4.3.3, that the Bayesian approach yields a more general and comprehensive solution to parameter estimation problems.

4.3.2 Parameter Estimation within the Frequentist Paradigm

For simplicity, let us first restrict our discussion by assuming that n successive measurements of a random variable X yield values, x_i, with $i = 1, \ldots, n$. The dataset, D, may then be represented as a vector $\vec{x} = (x_1, x_2, \ldots, x_n)$. Within the frequentist inference paradigm, one assumes there exists a probability density, $f(x)$, describing the likelihood (frequency) of observing specific values x_i, with $i = 1, 2, \ldots, n$, of the random variable X in n distinct measurements. The sample space of the experiment thus consists of a space of n-dimensions, with each dimension corresponding to the sample space of X. Successive deployments of this experiment are expected to yield fluctuating values of the vector $\vec{x}$. In order to evaluate the probability density function (PDF), $f_n(\vec{x})$, of the vector $\vec{x}$, one must first determine the probability density of a given set of n values measured in one experiment. We will assume the measurements are carried out independently but with identical conditions. Remember from Chapter 2 that the probability of a subset A of the sample space **S** consisting of the union of disjoint subsets A_i (with $A_i \cup A_j$ for $i \neq j$) is simply the product of the probabilities of the subsets. The PDF of $\vec{x}$ thus consists of the product of the probability densities of each of the values x_i. Since the measurement conditions are deemed identical, each measurement x_i is governed by the same PDF $f(x)$. One thus gets:

$$f_n(\vec{x}) = f(x_1) f(x_2) \cdots f(x_n) = \prod_{i=1}^{n} f(x_i). \tag{4.1}$$

The problem, of course, is that the function $f(x)$, or some of its parameters, are not known a priori. It thus becomes our goal to infer what this function might be or, at the very least, determine some of its properties (e.g., its moments) based on the measurements x_i.

Let us assume that the PDF $f_n(\vec{x})$ might depend on m a priori unknown parameters $\vec{\theta} \equiv \{\theta_1, \theta_2, \ldots, \theta_m\}$. We wish to construct a function of the n measured values x_i, denoted $\vec{x} = (x_1, x_2, \ldots, x_n)$, to estimate the parameters $\vec{\theta}$. In the literature, a function of the measurements $\vec{x}$ with no unknown parameters is called a **statistic**, while a statistic used to estimate one or several properties of a PDF is known as an **estimator**.

An experiment yields random values $\vec{x}$. A function of these $\vec{x}$, that is, an estimator consequently yields values $\vec{\theta}$, known as **estimates**, that are randomly distributed. Let us denote these estimates $\hat{\theta}$ to distinguish them from the actual value $\vec{\theta}$ (i.e., the true value of the parameters). $\hat{\theta}$ is commonly called **estimator** of $\vec{\theta}$. The term **estimator** usually refers to the mathematical function of the sample, whereas the term **estimate** is used for a particular value of the estimator obtained within a specific experiment, that is, with a specific sample $\{x_i\}, i = 1, \ldots, n$.

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

By its very nature and construction, an estimator $\hat{\theta}$ is expected to yield estimates that fluctuate dataset by dataset and randomly deviate from the true value $\vec{\theta}$ when the experiment is repeated. One hopes, however, that given sufficiently many values x_n, the estimator might **converge** to the actual value $\vec{\theta}$. In this context, it is interesting to consider the convergence properties (vs. the size of a sample) and errors associated with particular estimator definitions: how close an estimator $\hat{\theta}$ comes to the true value $\vec{\theta}$ depends on the quality of the measurements, the size of the sample, as well as the actual formulation or definition of the estimator. We, however, delay a discussion of these considerations until §4.4 after a discussion of Bayesian parameter estimation in the next section.

4.3.3 Parameter Estimation within the Bayesian Paradigm

Parameter estimation within the Bayesian paradigm has the same basic goal as in classical inference: one wishes to find parameter values $\vec{\theta}$ that best fit or characterize the observed data, D. However, the Bayesian approach also accounts for prior information that might be available about the system, the model, M, and the parameters themselves, $\vec{\theta}$, before the data are measured. Prior information about the parameters of the models, in particular, is encoded as a **prior probability density** $p(\vec{\theta}|I)$, where I represents relevant facts known about the system prior to the measurement(s). The data D thus provides a complement of information, which through Bayes' theorem provides a **posterior probability distribution** (density) of the parameters $\vec{\theta}$:

$$p(\vec{\theta}|D, I) = \frac{p(\vec{\theta}|I)p(D|\vec{\theta}, I)}{p(D|I)} \tag{4.2}$$

The probability $p(D|I)$, known as global likelihood of the data, can be calculated using the law of total probability:

$$p(D|I) = \int p(\vec{\theta}|I)p(D|\vec{\theta}, I) \prod_i d\theta_i \tag{4.3}$$

In effect, the probability density function $p(\vec{\theta}|D, I)$, known as posterior PDF, plays the same role as the PDF $f_n(\vec{x})$ in the frequentist paradigm and provides the probability of measuring specific values of the model parameters $\vec{\theta}$ based on the measured data D. The posterior $p(\vec{\theta}|D, I)$ is then said to be **conditioned by** the data D.

For a model with a single continuous parameter, θ, this corresponds to the density of probability of finding the parameter in the range $[\theta, \theta + d\theta]$, while for a model with multiple parameters, it amounts to the joint density of probability for finding the parameters θ_i simultaneously in the ranges $[\theta_i, \theta_i + d\theta_i]$. An estimator of the model parameters, denoted $\hat{\theta}$, may then be determined on the basis of the mode of the distribution, that is, the value of $\vec{\theta}$ yielding the maximum value of the PDF $p(\vec{\theta}|D, I)$. While it might be tempting to use the expectation value of the parameters $\vec{\theta}$

$$\langle\vec{\theta}\rangle \equiv \int \vec{\theta} p(\vec{\theta}|D, I) \prod d\theta_i. \tag{4.4}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

in lieu of the mode, one should note that, in general, the distribution $p(\vec{\theta}|D, I)$ may not be symmetric about its mode. Indeed, if $p(\vec{\theta}|D, I)$ is asymmetric, skewed, and features long tails, the expectation value $\langle\vec{\theta}\rangle$ shall not have the highest probability and thus would not yield the most probable value of $\vec{\theta}$. It is thus essential to identify the true maximum position of $p(\vec{\theta}|D, I)$. In effect, whether using a frequentist or Bayesian approach, finding the parameters of a model that best match the data becomes a problem of optimization. Optimization of $f_n(\vec{x})$ and $p(\vec{\theta}|D, I)$ in the presence of a single free model parameter may be relatively simple, particularly for linear models, but it can be quite arduous for models involving several (free) parameters. Fortunately, many techniques exist to tackle this optimization task. Several of these techniques, relevant for linear and nonlinear data models, are discussed in Chapters 5 and 7.

We shall consider examples of estimators evaluated in the frequentist and Bayesian paradigms in §§4.5 and 7.4, respectively. But first, let us examine the properties of estimators as mathematical functions.

4.4 Properties of Estimators

Given estimators are arbitrary functions of the measured data values x_i, one should expect they provide estimates of observables of interest that may depend on their actual mathematical formulation, the size of the data sample used to compute them, as well as, perhaps, the probability distributions that govern the production and measurement of the observable X. It is thus sensible to wonder how precise or accurate estimates of an observable may be given a specific dataset, its size, and the actual mathematical formulation of the statistics used to estimate the observable. This implies statistics and estimators are not all born equal. It is thus of interest to briefly consider, in the following paragraphs, their properties or behavior under changes in the sample size or the probability distribution that governs an observable X of interest. The discussion presented here is at a similar level to that found in Cowan [67], but a more in-depth treatment of these concepts may be found in James [116].

4.4.1 Estimator Consistency

An estimator $\hat{\theta}$ is said to be **consistent** when estimates converge toward the true value of the parameter(s) as the number of observations x_i increases. A number of convergence criteria can be applied but it is common to require **consistency in probability** defined based on the convergence of the probability $p\big(|\hat{\theta}_n - \theta| > \varepsilon\big)$, where $\hat{\theta}_n$ is an estimator of θ based on n observations. An estimator is then said to be consistent if a value N exists such that

$$p(|\hat{\theta}_n - \theta| > \varepsilon) < \eta \tag{4.5}$$

for all $n > N$ and arbitrarily small (positive definite) numbers ε and η. This expression indicates $\hat{\theta}_n$ converges toward θ for increasing values of n, as schematically illustrated in Figure 4.1.

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

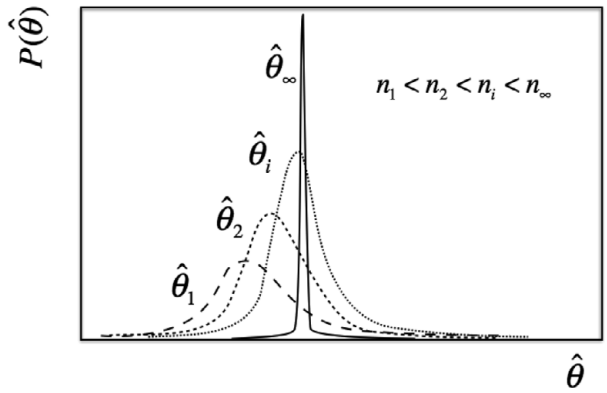


Fig. 4.1 An estimator $\hat{\theta}$ is said to be consistent if its probability density narrows and progressively converges to the true value (of an observable) as the number of observations n_i increases.

Consistency implies asymptotic convergence but does not guarantee that the precision of an estimator improves monotonically with the number of observations n. It is thus a necessary but insufficient requirement for an estimator to be useful. In other words, adding observations does not necessarily guarantee an increase in precision and smaller errors. Additional criteria are thus warranted.

4.4.2 Expectation Value of an Estimator

Statistics and estimators are multivariate functions of random variables. We can then apply the rules and techniques introduced in §§2.8–2.8.6 for such functions to determine their probability density functions and associated properties.

Let $\hat{\xi} \equiv \hat{\xi}(\vec{x})$ represent an estimator of interest calculated based on a sample of n measured values x_i that we shall represent, for convenience, in vector notation as $\vec{x} = (x_1, \ldots, x_n)$. Let us denote the PDF of the estimator $\hat{\xi}$ as $g(\hat{\xi}|\xi)$, where it is assumed that the estimator has some "true" value ξ. Calculation of this function, commonly known as **sampling distribution** of ξ, requires one makes assumptions concerning the probability distributions of each of the measured data x_i. In general, one requires a PDF $p(x_i|\vec{\theta}, H_i, I)$ based on a **probability model** or hypothesis H_i and prior information I for each of the observables x_i obtained in the data sample, where $\vec{\theta}$ represents one or several parameters of the models H_i which may be unknown a priori. In practice, if the x_i are independently sampled from the same population, it is reasonable to assume they may be described by the same model H and PDF $p(x|\vec{\theta}, H, I)$, that is, with the same model parameter(s) $\vec{\theta}$.

As we saw in §2.8.6, the probability distribution of a particular set of such values $\vec{x} = (x_1, \ldots, x_n)$, called the likelihood (of the data D) for parameter $\vec{\theta}$, and denoted $L(D|\vec{\theta}, H, I)$, is readily determined by multiplying the probability densities of each of the x_i. We thus write

$$L(D|\vec{\theta}, H, I) \equiv L(\vec{x}|\vec{\theta}, H, I) = \prod_{i=1}^{n} p(x_i|\vec{\theta}, H, I). \tag{4.6}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

Indeed, the **likelihood function** $L(\vec{x}|\vec{\theta}, H, I)$ represents the probability of the data $\vec{x}$ given a hypothesis or model H with parameters $\vec{\theta}$. Consequently, regarding the estimator $\hat{\xi}$ as a function of the continuous random variable $\vec{X}$, its expectation value may be calculated according to Eq. (2.144), and one gets

$$\mathrm{E}[\hat{\xi}(\vec{x})] = \int \hat{\xi} g(\hat{\xi}|\xi)\, d\hat{\xi}, \tag{4.7}$$

$$= \int \cdots \int \hat{\xi}(\vec{x}) L(\vec{x}|\vec{\theta}, H, I)\, dx_1 \cdots dx_n, \tag{4.8}$$

$$= \int \cdots \int \hat{\xi}(\vec{x}) p(x_1|\vec{\theta}, H, I) \cdots p(x_n|\vec{\theta}, H, I)\, dx_1 \cdots dx_n. \tag{4.9}$$

Within the frequentist paradigm, this expectation value represents the mean value of $\hat{\xi}$ produced by an infinitely large (or very large) number of similar experiments yielding samples $\{x_i\}, i = 1, \ldots, n$, while in the Bayesian paradigm, it is simply the expectation value of $\hat{\xi}$ implied by the likelihood function L, the model H, and the parameter value $\vec{\theta}$ without regard as to whether the measurement can be repeated.

A case of particular interest arises when $\hat{\xi} \equiv \hat{\theta}_i$ is an estimator of one of the model parameters θ_i, that is, a function of the variables x_i chosen to yield an estimate of a particular model parameter θ_i of the distribution $p(x|\vec{\theta}, H, I)$. Two situations are typically encountered in this context. In one, the estimator consists of a simple statistic, that is, a formula involving no free parameters and whose calculation yields an estimate of the model parameter directly. In others, the statistics formula depends on the value of the parameter itself which is then considered unknown. An optimization procedure is then required to determine which parameter value is best compatible with the measured data.

Whether an estimator ξ is designed to determine some generic property of a distribution (e.g., one of its moments or cumulants), or a particular parameter θ_i of the model H that defines the distribution, it is legitimate to wonder how precise and accurate an estimate obtained from a finite dataset may be. This requires a discussion of the notions of **estimator bias** and **estimator precision**.

4.4.3 Estimator Bias

The **bias**, b_n, of an estimator is defined as the difference between its expectation value $\mathrm{E}[\hat{\xi}(\vec{x})]$ and the true value ξ for an "experiment" consisting of n observations. For an elementary (scalar) estimator, this is written

$$b_n[\hat{\xi}] \equiv \mathrm{E}[\hat{\xi}] - \xi = \mathrm{E}[\hat{\xi} - \xi]. \tag{4.10}$$

The bias b_n is an intrinsic property of the estimator $\hat{\xi}$, which depends on its formulation (i.e., its functional form), the sample size n, and the properties of the PDF being evaluated, but not on the sampled values. An estimator is said to be **unbiased** for all values of n and ξ if

$$b_n[\hat{\xi}] = 0, \tag{4.11}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

or

$$\mathrm{E}[\hat{\xi}] = \xi. \tag{4.12}$$

If the bias is finite for finite values of n but vanishes in the limit $n \to \infty$, then the estimator is considered to be **asymptotically unbiased**. We present examples of unbiased and asymptotically unbiased estimators in §4.5.

It is worth noticing that the notion of bias could also be based on the median or the mode of the distribution $g(\hat{\xi}|\xi)$. The choice of $\mathrm{E}[\hat{\xi}]$ is indeed somewhat arbitrary, but its linear property as an operator greatly simplifies calculations of bias and thus makes for a convenient and practical choice. However, this choice implies that an estimator $\hat{\xi}$ is not necessarily the median or the mode of the PDF $g(\hat{\xi}|\vec{\xi})$. It might indeed be more or less probable to obtain estimates with values below (or above) the expectation value $\mathrm{E}[\hat{\xi}]$. It is also important to realize that an estimator may be biased even if it is consistent or unbiased even if inconsistent. For a more in depth discussion of this notion, see, for instance refs. [67, 116].

The linear nature of the expectation value operator also implies that while an estimator $\hat{\xi}$ might be unbiased, functions of this estimator are not necessarily unbiased. For instance, if η is an unbiased estimator of ξ, the square η^2 or the inverse $1/\eta$ are not unbiased because linear operators are not invariant under nonlinear variable transformations. A condition based on the median would guarantee such invariance but is often considered far less practical. Be that as it may, if η^2 is a consistent estimator of ξ^2, one can show that it will tend to be asymptotically unbiased. Consistency is thus usually considered, by statisticians, a more important property than the lack of bias. In practice, however, physicists typically do not have to worry too much about such subtleties.

4.4.4 Estimator Mean Square Error

The bias of an estimator determines its **accuracy**, that is, whether it tends toward the correct value (i.e., the true value) on average. One must also consider the estimator's **precision** and characterize the dispersion of values that might be obtained with successive measurements and the estimates computed from them. This is accomplished via a calculation of the **mean square error** (MSE), defined as

$$\mathrm{MSE} = E[(\hat{\theta} - \theta)^2]. \tag{4.13}$$

It is convenient to rewrite this expression by inserting $-E[\hat{\theta}] + E[\hat{\theta}]$ in the argument of the expectation value. One gets

$$\mathrm{MSE} = E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2], \tag{4.14}$$

$$= E[(\hat{\theta} - E[\hat{\theta}])^2] + (E[\hat{\theta} - \theta])^2. \tag{4.15}$$

The first term of the second line corresponds to the variance $\mathrm{Var}[\hat{\theta}]$ of the estimator, whereas the second term is the square of its bias b. One thus finds that the MSE of an estimator equals the sum of its variance and the square of its bias:

$$\mathrm{MSE} = \mathrm{Var}[\hat{\theta}] + b^2. \tag{4.16}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

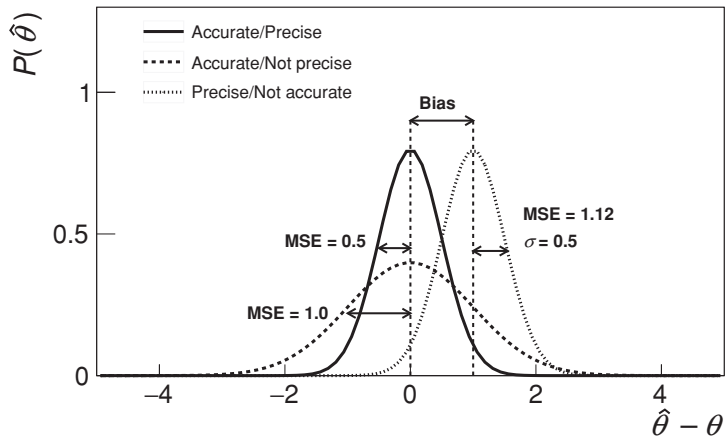


Fig. 4.2 Illustration of the concepts of **bias** and **Mean Square Error (**MSE**)**.

The notions of bias and estimator mean square error were here introduced in terms of simple estimators or functions of the measured values, $\xi(\vec{x})$. These concepts are, however, readily extended to cases where model parameters (alternatively viewed as continuous model hypotheses) are obtained by maximization of a likelihood function or minimization of a χ^2 function. They in fact also apply to any type of measurement that extracts estimates of model parameters, whether based on a simple or composite hypotheses [116].

4.4.5 Epilog: Accuracy and Precision of Estimators

There is no unique method to construct estimators. Indeed, we will show later in this chapter that there exist several distinct criteria to define and build estimators, for a given model parameter θ, which produce distinct functional forms and associated PDF $g(\hat{\theta}|\theta)$. But, given a particular estimator, one can usually establish whether it has the desired properties, that is, whether it is consistent, whether it features a bias, has a small MSE, and so on. The notions of bias, estimator variance, and MSE are illustrated in Figure 4.2, which displays the PDFs of three hypothetical estimators $\hat{\theta}$, plotted as functions of the deviation $\hat{\theta} - \theta$.

An estimator and, more generally, a measurement method, are said to be unbiased if repeated measurements yield estimates that cluster around the actual value of the model parameter (null deviation on average). However, an estimator is said to be **biased** if the estimates it produces tend to systematically deviate from the actual value of the parameter. An experimental procedure and an estimator are said to be **accurate** if they yield results (e.g., estimates of model parameters or physical quantities) with no bias.

The variance of an estimator determines its precision. Obviously, smaller values of variance imply a more **precise** estimator and experimental technique. An estimator may, however, be very precise but biased and consequently not accurate. An experimental method and the estimator used to obtain parameter estimates should thus ideally be designed to have both high precision and accuracy. An estimator is deemed **optimal** if it has zero bias

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

and minimum variance. Likewise, a measurement procedure is considered **valid** if it is both accurate and precise.

It is in the very nature of stochastic phenomena that by construction, all estimators have a limited precision, that is, a finite MSE. We show in §4.7.6 that the minimal variance of an estimator $\hat{\theta}$, called the **minimal variance bound** (MVB), is given by the following expression:

$$\mathrm{MVB}(\hat{\theta}) = \left\langle \left(\frac{d}{d\theta} \ln L \right)^2 \right\rangle^{-1}, \tag{4.17}$$

where L is the likelihood function defined in Eq. (4.6). We will make use of this bound on several occasions later in this book.

We consider the formulation of basic estimators and examine some of their properties in the next section. Maximum likelihood and least square estimators are introduced in §5.1 and §5.2, respectively.

4.5 Basic Estimators

Consider a random sample $D = \{x_i\}$ consisting of n values x_i distributed according to a common but unknown PDF $p(x)$. Although it may not be possible to fully determine $p(x)$ based on the sample $\{x_i\}$, it is possible and may suffice, in many practical applications, to determine its lowest moments.

4.5.1 First Moment: Sample Mean

We first seek to obtain a function of the n values x_i that provides an estimator $\hat{\mu}$ of the expectation value $\mu = \mathrm{E}[x]$. An obvious candidate for such an estimator is the **arithmetic mean**, $\bar{x}$, defined as

$$\hat{\mu} \equiv \bar{x} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{4.18}$$

It is commonly also called **sample mean** or just **mean**, and is often also noted $\langle x \rangle$.

By construction as a sum of the random variables x_i, $\bar{x}$ is a statistic and we can thus apply the techniques introduced in the previous paragraphs to calculate its value and properties. One expects $\bar{x}$ to deviate randomly from the expectation value $\mu = \mathrm{E}[x]$ of the PDF $p(x)$. Indeed, one expects that repeated measurements of n values $\{x_i\}$ drawn from the same population, characterized by the PDF $p(x)$, should yield values of $\bar{x}$ that fluctuate, but nonetheless cluster about μ. For finite samples, deviations from μ are by construction finite and the variance of $\bar{x}$ is consequently also finite. One can show that the weak law of large numbers guarantees that $\bar{x}$ is a consistent estimator of the population mean μ [116].

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

It is also relatively straightforward to show that the estimator $\bar{x}$ converges to μ with zero bias in the limit $n \to \infty$.

The proof of unbiasedness is indeed relatively simple and also instructional. It proceeds on the basis of a calculation of the expectation value of $\bar{x}$ as follows:

$$\mathrm{E}[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]. \tag{4.19}$$

The evaluation of the expectation values requires an integral. However, the order of evaluation of this integral and the sum is inconsequential. One can thus write

$$\mathrm{E}[\bar{x}] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[x_i]. \tag{4.20}$$

The expectation value $\mathrm{E}[x_i]$ is by definition equal to

$$E[x_i] = \int \cdots \int x_i p(x_1) p(x_2) \cdots p(x_n)\, dx_1 dx_2 \cdots dx_n, \tag{4.21}$$

$$= \left(\int p(x_1)\, dx_1\right) \times \cdots \times \left(\int x_i p(x_i)\, dx_i\right) \times \cdots \times \left(\int p(x_n)\, dx_n\right), \tag{4.22}$$

$$= \int x_i p(x_i)\, dx_i = \mu, \tag{4.23}$$

in which we used the normalization of the PDF, $\int p(x_i)\, dx_i = 1$, and the definition of the first moment μ. Inserting this result in Eq. (4.20), we find

$$\hat{\mu} = \mathrm{E}[\bar{x}] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu. \tag{4.24}$$

We then conclude that the sample mean, $\bar{x}$, is indeed an unbiased estimator of the mean μ of the PDF $p(x)$.

4.5.2 Mean Square Error of the Sample Mean

It is useful to consider the variance $\mathrm{Var}[\bar{x}]$, or **Mean Square Error** (MSE), as an indicator of typical deviations of the sample mean, $\bar{x}$, from the actual value μ. $\mathrm{Var}[\bar{x}]$ gauges the variation of $\bar{x}$ for repeated experiments, that is, different sets $\{x_i\}$. As such, it can be used to estimate the error, or uncertainty, of $\bar{x}$ relative to the true value μ.

We evaluate the variance $\mathrm{Var}[\bar{x}]$ by inserting the expression of the sample mean into the definition of the variance:

$$\mathrm{Var}[\bar{x}] = \mathrm{E}[\bar{x}^2] - (\mathrm{E}[\bar{x}])^2, \tag{4.25}$$

$$= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] - \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]^2. \tag{4.26}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

The calculation proceeds similarly as for the expectation value $\mathrm{E}[\bar{x}]$ (see Problem 4.5). One gets

$$\mathrm{Var}[\bar{x}] = \frac{1}{n^2}\sum_{i,j=1}^{n}\mathrm{E}[x_i x_j] - \frac{1}{n^2}\left(\sum_{i,=1}^{n}\mathrm{E}[x_i]\right)^2, \tag{4.27}$$

$$= \frac{1}{n^2}\left(n\mu_2' + n(n-1)\mu^2\right) - \mu^2, \tag{4.28}$$

$$= \frac{1}{n}\left(\mu_2' - \mu^2\right), \tag{4.29}$$

$$= \frac{\sigma^2}{n}, \tag{4.30}$$

where μ_2' is the second moment of the PDF $p(x)$, and σ its (unknown) variance. We find that the variance of the sample mean $\bar{x}$ equals the variance of the PDF $p(x)$ representing the fluctuations of the random variable X divided by the number of measurements n. In other words, the standard error on the sample mean $\bar{x}$ is equal to the standard deviation of the PDF divided by the square root of the number of measurements carried out, that is, the size of the sample $\{x_i\}$:

$$\sigma_{\bar{x}} = (\mathrm{MSE}\,[\bar{x}])^{1/2} = \frac{\sigma}{\sqrt{n}}. \tag{4.31}$$

The error on the sample mean thus depends on the PDF itself through σ, as well as the size of the sample used to estimate the mean. It can be arbitrarily reduced by increasing the sample size n. However, improvements are slow: to reduce the error by a factor of 2, one needs to quadruple the size of the data sample.

4.5.3 Estimator of the Variance

We next seek an estimator $\widehat{\sigma^2}$ of the variance σ^2 of the unknown distribution $p(x)$. Based on the expression of the sample mean, Eq. (4.18), it would seem reasonable to write

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\,. \tag{4.32}$$

However, it turns out that this expression does not provide an unbiased estimator of σ^2. Indeed, let us evaluate its expectation value:

$$E\left[S^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right], \tag{4.33}$$

$$= \frac{1}{n}\sum_{i=1}^{n}E\left[\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^2\right]. \tag{4.34}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

We expand the square within the expectation value to get

$$\frac{1}{n}\sum_{i=1}^{n} E\left[{x_i}^2\right] - \frac{2}{n^2}\sum_{i=1}^{n} E\left[x_i \sum_{j=1}^{n} x_j\right] + \frac{1}{n^3}\sum_{i=1}^{n} E\left[\sum_{j=1}^{n} x_j \sum_{k=1}^{n} x_k\right]. \tag{4.35}$$

Since sum and expectation value operations commute, one can write

$$\frac{1}{n}\sum_{i=1}^{n} E\left[{x_i}^2\right] - \frac{2}{n^2}\sum_{i,j=1}^{n} E\left[x_i x_j\right] + \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j,k=1}^{n} E\left[x_j x_k\right]. \tag{4.36}$$

One gets $E\left[x_i x_j\right] = E\left[{x_i}^2\right] = {\mu_2}'$ for $i = j$ and $E\left[x_i x_j\right] = E\left[x_i\right] E\left[x_j\right] = \mu^2$ for $i \neq j$ since the measurements are assumed to be uncorrelated. Equation (4.36) becomes

$$\tfrac{1}{n}\sum_{i=1}^{n} {\mu_2}' - \tfrac{2}{n^2}\sum_{i=1}^{n} {\mu_2}' - \tfrac{2}{n^2}\sum_{i\neq j=1}^{n} \mu^2 + \tfrac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n} {\mu_2}' + \tfrac{1}{n^3}\sum_{i=1}^{n}\sum_{i\neq j=1}^{n} \mu^2. \tag{4.37}$$

The sums with $i \neq j$ involve $n(n-1)$ identical terms. Equation (4.37) thus simplifies to

$$\mathrm{E}[S^2] = E\left[\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2\right] = \frac{n-1}{n}\sigma^2, \tag{4.38}$$

from which we conclude that S^2 is indeed a biased estimator of the variance σ^2. Obviously, given that $\lim_{n\to\infty} (n-1)/n = 1$, we conclude S^2 is asymptotically unbiased. Note, however, that if the mean μ is known, then the estimator $\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$ is unbiased (see Problem 4.6). In general, since the mean μ is a priori unknown, it remains desirable to find a nonbiased estimator of the variance. It is then possible to show (see Problem 4.7) that the estimator s^2, defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{4.39}$$

is in fact unbiased.

4.5.4 Mean Square Error of the Variance Estimator s^2

Following the procedure used in §4.5.2 to calculate the MSE of the sample mean, it is relatively simple, although somewhat tedious (see Problem 4.14), to show that the MSE of the estimator s^2 is given by

$$\mathrm{MSE}[s^2] = \mathrm{Var}[s^2] = \frac{1}{n}\left(\hat{\mu}_4 - \frac{n-3}{n-1}\hat{\mu}_2\right), \tag{4.40}$$

where the $\hat{\mu}_k$ are estimates of central moments (see Problem 4.9) of order k given by

$$\hat{\mu}_k = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^k. \tag{4.41}$$

Equation (4.40) reduces to the following expression for a Gaussian PDF:

$$\mathrm{MSE}[s^2] = \frac{2\sigma^4}{n^2}(n-1) \approx \frac{2\sigma^4}{n-1} \quad \text{Gaussian PDF.} \tag{4.42}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

One thus concludes that the uncertainty on the estimate of the standard deviation, σ_s, is proportional to the actual variance of the PDF and scales inversely with the square root of the sample size n. This implies that for broadly distributed PDFs, a rather "large" sample must be acquired in order to obtain a reasonably precise estimate of the standard deviation.

4.5.5 Covariance and Pearson Coefficients

The result obtained in the previous section for the variance also applies to the covariance of joint measurements of two variables X and Y. Let us indeed consider a sample of n pairs of measurements $\{x_i, y_i\}$. One can show (see Problem 4.8) that the expression

$$\hat{V}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{4.43}$$

provides an unbiased estimator of the covariance V_{xy} of the two variables X and Y.

It is also useful to consider the ratio, r_{xy}, of the covariance $\hat{V}_{xy}$ by the square root of the product of the variances of X and Y:

$$\hat{r}_{xy} = \frac{\hat{V}_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{\left(\overline{x^2} - \bar{x}^2\right)\left(\overline{y^2} - \bar{y}^2\right)}}. \tag{4.44}$$

This ratio is known as Pearson correlation coefficient.[4] It measures the strength of the correlation between the variables X and Y relative to their respective variances. When calculated for an entire population, the coefficient is typically designated by the Greek letter ρ (see §2.9.5) whereas the Latin letter r is used for a calculation based on a sample. It can be shown that $\hat{r}_{xy}$ is biased but converges to the Pearson coefficient ρ_{xy} in the large n limit (see Problem 4.15). One can also show (see Problem 4.16) that the coefficient lies in the range $[-1, 1]$. A value of 1 indicates that the variables Y and X are maximally correlated, and that a strict linear relation exists between the two variables:

$$y = |a|x. \tag{4.45}$$

Similarly, a value of -1 indicates that all data points lie on a single line $y = -|a|x$. A null value suggests there is no correlation between the variables.

4.5.6 Robust Moments

We saw in previous sections that the estimators of the moments of a distribution based on a sample of n values boil down to expressions of the form

$$\hat{\mu}'_k \equiv \langle x^k \rangle_n \equiv \frac{1}{n} \sum_{i=1}^{n} x_i^k \tag{4.46}$$

[4] First introduced by Francis Galton in the 1880s, and named after English mathematician Karl Pearson (1857–1936) generally credited with establishing the discipline of mathematical statistics.

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

for regular moments and

$$\hat{\mu}_k \equiv \langle (x - \hat{\mu}_1)^k \rangle_n \equiv \frac{1}{n} \sum_{i=1}^{n} (x_i - \langle x \rangle)^k \,. \tag{4.47}$$

for centered moments. While straightforward to implement in a computer program, these expressions feature important drawbacks. First, given the sums may proceed over a very large number of values, numerical errors are likely to occur, particularly for higher moments, due to rounding errors imposed by the finite number of bits used in floating point operations. Second, in the case of centered moments, the preceding formula requires one runs the analysis twice: once to obtain the mean $\langle x \rangle$, and once again to obtain the centered moments. Additionally, the two preceding formulas also present the disadvantage that the calculation must be run over all items x_i of the data sample before one obtains an estimate of the moments. Fortunately, these technical difficulties are easily circumvented by a simple reformulation of the preceding expressions.

Let us first consider the mean of a sample of n values and write

$$\begin{aligned} \langle x \rangle_n &= \frac{1}{n} \sum_{i=1}^{n} x_i, \\ &= \frac{1}{n} \left(x_n + \sum_{i=1}^{n-1} x_i \right), \qquad (4.48) \\ &= \frac{1}{n} x_n + \frac{(n-1)}{n} \langle x \rangle_{n-1}, \qquad (4.49) \end{aligned}$$

where in the third line, we have inserted $\langle x \rangle_{n-1}$ corresponding to the mean of the first $n-1$ values of the sample. Rearranging, we get

$$\langle x \rangle_n = \langle x \rangle_{n-1} + \frac{x_n - \langle x \rangle_{n-1}}{n}, \tag{4.50}$$

which says that the mean $\langle x \rangle_n$ obtained after n values is equal to the mean $\langle x \rangle_{n-1}$ obtained with $n-1$ values **plus** a correction Δ_n given by

$$\Delta_n = \frac{x_n - \langle x \rangle_{n-1}}{n}. \tag{4.51}$$

The same derivation technique can be applied to centered moments of order $k \geq 2$ (see Problem 4.17), and one obtains

$$\begin{aligned} \mu_k &= \langle (x - \langle x \rangle)^k \rangle \\ &= \tfrac{n-1}{n} \left\{ \left[1 - (1-n)^{k-1} \right] \Delta_n^k + \textstyle\sum_{l=2}^{k} \binom{l}{k} \langle (x - \langle x \rangle_n)^l \rangle_n \Delta_n^{k-l} \right\}, \end{aligned} \tag{4.52}$$

which for $k = 2, 3$ may be written

$$\langle (x - \langle x \rangle)^2 \rangle_n = \frac{n-1}{n} \left\{ n\Delta_n^2 + \langle (x - \langle x \rangle)^2 \rangle_{n-1} \right\} \tag{4.53}$$

$$\langle (x - \langle x \rangle)^3 \rangle_n = \frac{n-1}{n} \left\{ (2n - n^2)\Delta_n^3 + 3 \langle (x - \langle x \rangle)^2 \rangle_n \Delta_n + \langle (x - \langle x \rangle_n)^3 \rangle_n \right\}. \tag{4.54}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

The preceding expressions for centered moments are said to be **robust** because they involve numerically stable operations. Calculations proceed iteratively and do not involve addition or subtraction of huge numbers. Indeed, the correction Δ_n vanishes asymptotically for growing values of n, the calculated moments thus naturally tend to their "true" value as n iteratively increases. Values involved in the calculation all remain finite and rounding errors are thus minimized. The preceding expressions also involve the practical advantage that it is not necessary to proceed to an analysis of the full dataset in order to obtain the moments. The calculation of moments of all orders may be carried out iteratively and in parallel: one gets an estimate of all the required moments at each step n of the calculation. This enables "on the fly" analyses in which estimates of the moments may be obtained during the data acquisition process. It may then be possible to monitor the data collection, in real time, for shifts and varying trends that could indicate changes in the performance of detector components.

4.5.7 Other Estimators Commonly Used

Geometric Mean

Whenever the magnitude of a random variable is proportional to a product of n elementary random variables, x_i, it is convenient to replace the concept of arithmetic mean, which involves a sum of variables, with the **geometric mean**, defined as follows:

$$\bar{x}_{\text{geom}} = \sqrt[n]{x_1 x_2 \ldots x_n}. \tag{4.55}$$

The geometric mean finds applications in physics and engineering. A specific example involves the notion of average attenuation. Suppose a beam of light is directed through a series of n attenuators, each contributing a random attenuation, q_i, with $i = 1, \ldots, n$. The overall attenuation, Q, of the "signal" is thus equal to the product of all n attenuation factors, $Q = q_1 \times q_2 \times \ldots \times q_n$. The average attenuation, in other words, the attenuation per element, is thus $\bar{q} = \sqrt[n]{q_1 \times q_2 \times \ldots \times q_n}$, which is the geometric mean of the attenuations q_i. This means one could replace all n attenuators, q_i, by n identical attenuators, each yielding an attenuation of $\bar{q}$, and get the exact same result. A similar reasoning can be done with gain devices, such as amplifiers or photomultipliers.

Harmonic Mean

Another distribution characterization seldom encountered is the **harmonic mean**, $\bar{x}_{\text{harm}}$, of n values x_i defined as

$$\bar{x}_{\text{harm}} = \frac{n}{1/x_1 + 1/x_2 + \ldots + 1/x_n}. \tag{4.56}$$

Although less frequently used, this type of mean also has various applications. For instance, notes produced by the different strings of a guitar or violin sound pleasing (or harmonic) to the human ear[5] if their effective lengths are in the ratio 1: $\frac{1}{2}$: $\frac{1}{3}$:…The

[5] A fact originally noted by the ancient Greek Pythagoras.

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

harmonic mean of two notes is thus conveniently defined as twice the inverse of the sum of the reciprocal of their lengths. This stems from the fact that the human ear is sensitive to the frequency of the sound (the pitch), while the wavelength is determined by the length of the strings. The arithmetic mean of several frequencies is consequently equivalent to the harmonic mean of several string lengths.

4.6 Histograms

4.6.1 Basic Concept and Definition

Experimentally, various constraints limit the number of times a measurement of a particular observable X might be carried out. It is nonetheless desirable and of interest to generate simple visual representations of the acquired data. Indeed, whether one has acquired a sample of 10, 1,000, 1,000,000, or several billions data points, it is clearly useful to visualize these data in a graph. Different scenarios arise. It could be that the observable X of interest has a definite value, that is, is expected to take a single value. Measurement uncertainties may however lead to deviations from the actual (i.e., true) value of the observable. One may then be interested in visualizing how finitely many measured values $\{x_i\}$, $i = 1, \ldots, n$, are distributed. Alternatively, values of a continuous observable X could instead be expected to follow some model or probability distribution. It is then of interest to compare the density of measured values vs. x with the model or theoretical distribution. Either way, it is clear that plotting a graph of the number of instances or "events" as a function of the value of x is of interest. Indeed, one would like to compare the measured density distribution to the model. The problem, of course, is that we are dealing with a continuous variable and only dispose of a finite sample. One must then invoke a statistic, or estimator, that enables a sensible representation of the distribution of the continuous variable X. This is readily achieved by representing a set of measured values $\{x_i\}$, $i = 1, \ldots, n$, according to a construction consisting of the numbers of observed values, n_j, with $j = 1, \ldots, m$, in m finite width elements, called bins, of a partition of the range of allowed values of X. This construction is called a **histogram**.

In order to better understand the need for histograms, consider that many measurements can be essentially reduced to an experimental determination of the PDF of specific processes. In particle and nuclear physics, for instance, it is common to measure the momentum spectrum and angular distributions of particles produced in elementary interactions such as proton–proton, proton–nucleus, or nucleus–nucleus collisions at high energy. While one can model, at least in principle, the overall momentum and angular distributions, it is not possible to predict on a collision-by-collision, and particle-by-particle basis, what momenta and angle the particles will have. The momenta and angles are consequently random variables. However, a very large (infinite) number of measurements of momenta or production angles is expected to yield distributions of these variables that are representative of the nuclear interactions under consideration, that is, their respective PDFs. In practice, it is obviously impossible to carry out an infinite number of measurements. In

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

fact, in many cases, the number of occurrences of a specific measurement outcome may be severely limited by small cross section, finite beam intensity (luminosity), experimental acceptance, or detection efficiency (see Chapter 12 for definitions of these concepts). Additionally, even if many occurrences are measured for a continuous variable X, it is in fact impossible to measure the probability of a specific value. It is, however, conceptually and practically meaningful to determine the number of occurrences in a finite interval of width Δx. In principle, one would prefer to carry out measurements with as small a value of Δx as possible. In practice, the measurement is meaningful only if a finite number of occurrences is observed in a given interval. One can therefore study histograms, in other words, the number of occurrences in segments of finite width Δx.

Histograms may be defined in one, two, or any number of dimensions and are used for the study of a single random variable as well as the simultaneous (joint) study of two or more variables. We begin our discussion of histograms with the definition of one-dimensional histograms in §4.6.2 and extend to two- and multidimensional histograms in §4.6.3.

4.6.2 One-Dimensional Histograms

A one-dimensional (1D) histogram of a single continuous variable X is defined over a finite range $x_{\min} \leq x < x_{\max}$. This range is divided or **binned** in a specific (but arbitrary) number, m, of contiguous segments, called **bins**. The bins are numbered from $j = 1$ to m. They may all be of equal width, $\Delta x = (x_{\max} - x_{\min})/m$, or defined to cover subranges of unequal widths with lower and upper bin boundaries noted $x_{\min,j}$ and $x_{\max,j}$, respectively. In the latter case, bin boundaries are usually defined to leave no gaps between contiguous bins, that is, such that $x_{\max,j} = x_{\min,j+1}$ for $j = 1, \ldots, m-1$.

The analysis of a dataset $\{x_i\}$ and the filling of the histogram proceed as follows. The bin contents, H_j, $j = 1, \ldots, m$ are initially set to zero. The values $\{x_i\}$, $i = 1, \ldots, n$, are then sequentially considered. Values x_i satisfying $x_{\min} \leq x_i < x_{\max}$ are set to generate an entry in the bin j such that $x_{\min,j} \leq x_i < x_{\max,j}$. Depending on the specific circumstances of an analysis, the entry may involve incrementing the bin content H_j by one unit (i.e., +1) or by some weight w_i determined by the value x_i. Once all entries have been carried out, the shape of the histogram is expected to be representative of the PDF of variable X. The histogram may finally be divided by a suitable normalization constant consisting, for instance, of the number of events analyzed, the total number of entries in the histogram, or the total weight of the entries. For unit increments (+1) per entry, division by the total number of entries yields normalized bin contents

$$h_i = H_i/n, \tag{4.57}$$

corresponding to the relative frequency of each bin. In the limit of an infinite number of measurements, $n \to \infty$, h_i is then expected to represent the probability, p_i, of occurrences of the random variable X in the range $x_{\min,i} \leq x < x_{\max,i}$,

$$\lim_{n\to\infty} h_i = p_i \equiv \int_{x_{\min,i}}^{x_{\max,i}} f(x)\,dx, \tag{4.58}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

with the normalization[6]

$$\sum_{i=1}^{m} p_i = \int_{x_{\min}}^{x_{\max}} f(x)\,dx = 1. \tag{4.59}$$

The normalized bin content h_i is thus an estimator of the probability of obtaining a value of X in the range $x_{\min,i} \leq x < x_{\max,i}$. If resolution smearing and efficiency effects can be ignored (or at the very least neglected), the definition of h_i in terms of an arithmetic mean of the number of entries in the bin constitutes an unbiased estimator of the probability p_i of observing entries in that bin. Altogether, the bin contents h_i, $i = 1, \ldots, n$, thus provide an estimator of the probability distribution of the observable X.

When weights w_i are used to increment a histogram, H_i corresponds to the sum of all weights used to increment bin i. If the weights are themselves random variables, $h_i = H_i/n$ no longer represents a number of occurrences in bin i but is instead the mean weight in that bin:

$$\lim_{n\to\infty} h_i = \frac{1}{n}\langle \Sigma w_i \rangle = \int_{x_{\min,i}}^{x_{\max,i}} w f(x)\,dx. \tag{4.60}$$

The mean weight in bin i, denoted $\langle w_i \rangle$, is given by the ratio of $\langle \Sigma w_i \rangle$, and the number of entries in that same bin:

$$\langle w_i \rangle = \frac{\langle \Sigma w_i \rangle}{\langle \Sigma 1_i \rangle} = \frac{\int_{x_{\min,i}}^{x_{\max,i}} w f(x)\,dx}{\int_{x_{\min,i}}^{x_{\max,i}} f(x)\,dx}. \tag{4.61}$$

With weights equal to unity, and in the limit $n \to \infty$, we find that a histogram provides the probabilities p_i that the variable X occurs in the range $x_{\min,i} \leq x < x_{\max,i}$. Evidently, it is not possible to carry out an experiment with an infinite number of measurements. The values h_i one obtains experimentally thus provide estimates of the probabilities p_i, as defined in the foregoing. Globally, a normalized histogram h_i, $i = 1, \ldots, m$ thus provides an estimator of the relative probabilities p_i, $i = 1, \ldots, m$. Uncertainties associated with the estimates h_i are discussed later in this section.

Various plotting techniques are available to represent histograms graphically. Histograms are often represented as bar graphs such as those displayed in Figure 4.3. These histograms were created with a Monte Carlo (random) number generator (§13.3) according to a Gaussian distribution (§3.9) of mean 0.5, and standard deviation equal to 0.75. The histograms shown in panels (a), (b), (c), and (d) were obtained with sample sizes of 100, 1,000, 10,000, and 1,000,000 entries, respectively. While it is difficult to discern the exact shape of the distribution based on histogram (a), the shape becomes progressively more obvious with increased statistics, and with one million entries in panel (d), one can identify a Gaussian distribution. Panels (e) and (f) present histograms obtained with 100 distinct entries. The

[6] Obviously, this normalization applies only if there are no gaps between bins.

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

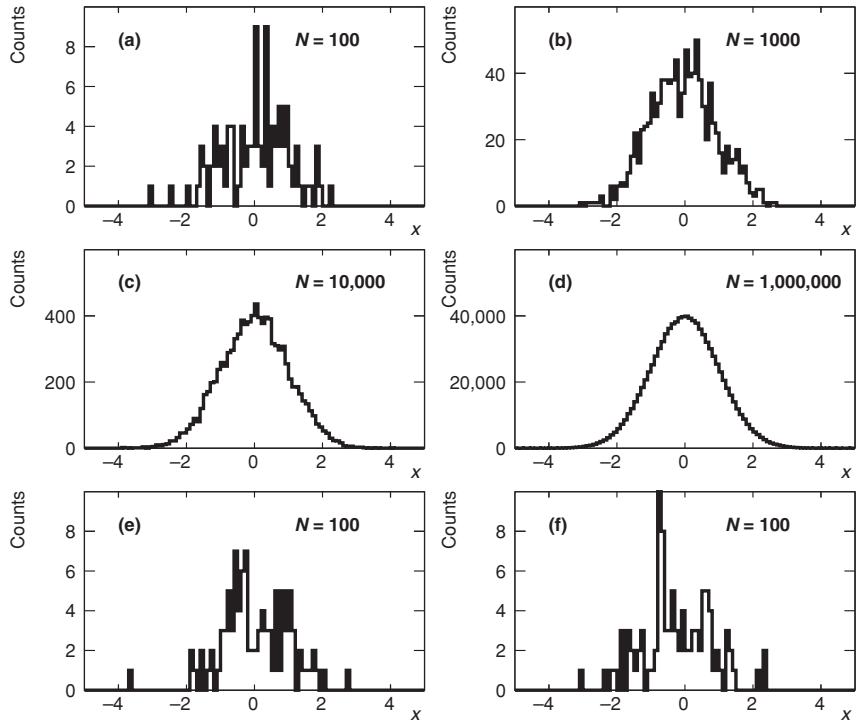


Fig. 4.3 Histograms obtained with a Monte Carlo generator designed to produce a standard normal distribution ($\mu = 0$, $\sigma = 1$). The size of the generated sample is shown in each panel. Histograms (a), (e), and (f) present distinct samples of size equal $n = 100$ which exemplify that distributions measured with fewer statistics (i.e., small sample) may considerably deviate from the actual PDF.

entries are mutually exclusive and different from those shown in panel (a). Comparing the three distributions (a), (e), and (f), one observes much variance in the shape, width, and height of the distributions. Obviously, with little statistics, it is difficult to precisely characterize the **parent** PDF, that is, the PDF that determines the measured distribution. But as the number of samples increases, one progressively becomes able to better discern and identify the shape of the distribution. This is typical of all measurements relying of finitely many samples or events.

While of evident interest, the quantities p_i defined earlier do not exactly correspond to the PDF, $f(x)$, one seeks to determine, but instead correspond to integrals of the PDF in bins $i = 1, n$. A theorem of elementary calculus, however, enables the determination of an estimate of the PDF. The theorem (used here without demonstration) stipulates that for each bin i, there exists a value $x_{i,0}$ such that

$$\int_{x_{\min,i}}^{x_{\max,i}} f(x)\,dx = f(x_{i,0})\,(x_{\max,i} - x_{\min,i}) = f(x_{i,0})\Delta x_i, \tag{4.62}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

where $\Delta x_i = x_{\max,i} - x_{\min,i}$. It is important to realize that in general, the values $x_{i,0}$ do not correspond to the center of bins i. In the limit $n \to \infty$, given the value p_i, the value $f(x_{i,0})$ of the PDF at $x_{i,0}$ is then

$$f(x_{i,0}) = \frac{p_i}{\Delta x_i}. \tag{4.63}$$

For finite n, the value p_i is replaced by h_i, and one gets estimates $\hat{f}_i$ of the PDF value in each bin:

$$\hat{f}_i \equiv f_{\text{estimate}}(x_{i,0}) = \frac{h_i}{\Delta x_i}. \tag{4.64}$$

In histogram plots, the estimate $\hat{f}_i$ is usually represented as a continuous horizontal line across the bin width, as shown in Figure 4.3. If instead one wishes to plot $f(x)$ as an actual function, one must determine the abscissa x_i corresponding to the ordinate $\hat{f}_i$ in order to plot the data as a set of points $(x_i, \hat{f}_i)$. This is further discussed in Problem 2.10.

Given the finite size of the sample used in a measurement, one expects the measured estimates $\hat{f}_i$ might randomly deviate from the actual value of the function $f(x_i)$. It is thus also important to estimate the errors on each of these values in order to evaluate the precision of the measurement, and enable meaningful comparisons with other measurements or with models (see §6.4). If there are no bin-to-bin correlations, for experiments where the successive values of x measured are uncorrelated, the number of entries H_i in each bin is a random number determined by a Poisson distribution (§3.3) of mean μ_i given by

$$\mu_i = n \int_{x_{\min,i}}^{x_{\max,i}} f(x)\,dx = np_i, \tag{4.65}$$

and variance $\text{Var}[H_i] = \mu_i \approx H_i$. One therefore concludes that the values H_i shall converge to np_i in the large n limit. It also implies one can estimate the error (standard deviation) on the bin content as $\sqrt{H_i}$. The standard deviation of h_i is thus of the order of

$$\delta h_i \approx \text{Var}[h_i]^{1/2} = \sqrt{H_i}/n, \tag{4.66}$$

and since H_i equals nh_i, this amounts to the commonly known result $\delta h_i = \sqrt{h_i/n}$. The relative error of the estimate $\hat{f}_i = h_i/\Delta x_i$ is then

$$\frac{\delta \hat{f}_i}{\hat{f}_i} = \frac{\sqrt{h_i/n}}{h_i} = \frac{1}{\sqrt{n}}\frac{1}{\sqrt{h_i}} \to \frac{1}{\sqrt{n}}\frac{1}{\sqrt{p_i}}. \tag{4.67}$$

One concludes, as expected, that the relative error, $\delta \hat{f}_i/\hat{f}_i$, on an estimate of the PDF improves as the square root of the number of measurements n carried to determine the PDF. This is costly. Indeed, to get a reduction of the errors by a factor of 2, one must quadruple the statistics, whereas to achieve an improvement by a factor of 10, one must increase the dataset by a factor of 100, and so on. Consequently, one finds that significantly improving known results requires considerable increase of the size of existing data samples. In nuclear and particle physics experiments, this translates into demands for substantial increases in

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

the beam luminosity (intensity), the duration of data acquisition periods, as well as detector acceptance and efficiency.

4.6.3 Multidimensional Histograms

Two-dimensional (2D) and multidimensional histograms are used to represent data with dependencies on two or more variables. They are defined and filled in much the same way as 1D histograms. The following discussion is limited to 2D histograms but can be readily extended to multidimensional histograms.

To construct a 2D histogram, one must first define the ranges $x_{\min} \leq x < x_{\max}$, $y_{\min} \leq y < y_{\max}$ and number of bins m_x and m_y used to plot the two random variables (here labeled x and y) of interest. Modern plotting software packages typically enable the creation of 2D histograms with different types of partitions: one can use bins of fixed or variable width on both axes, or mix fixed and variable widths on the two axes. It is customary to identify the bins using a pair of indices (i, j) where i and j range from 1 to m_x and m_y, respectively. A histogram is filled using a data sample consisting of n pairs $(x, y)_k$, with $k = 1, \ldots, n$. Given the variables are random observables, one expects the content of the bins, H_{ij}, to also be random variables. The content of each histogram is readily converted into an estimate of the joint probability, h_{ij}, of observing pairs (x, y) in bin (i, j) by dividing H_{ij} by the size, n, of the event sample. Given a joint PDF, $f(x, y)$, and in the large n limit, the probability, p_{ij}, of observing counts in bin (i, j) is given by

$$p_{ij} \equiv \lim_{n\to\infty} h_{ij} = \int_{x_{\min,i}}^{x_{\max,i}} \int_{y_{\min,j}}^{y_{\max,j}} f(x, y)\, dxdy. \tag{4.68}$$

As for 1D histograms, one notes that although the values p_{ij} enable a representation of $f(x, y)$, they are not equal to the PDF $f(x, y)$ but rather constitute, as per Equation (4.68), integrals of the probability density. However, values $(x_{o,i}, y_{o,j})$ exist for each bin such that

$$p_{ij} = f(x_{o,i}, y_{o,j})\Delta x_i \Delta y_j \tag{4.69}$$

where $\Delta x_i = x_{\max,i} - x_{\min,i}$ and $\Delta y_i = y_{\max,i} - y_{\min,i}$. Note that $(x_{o,i}, y_{o,j})$ does not usually coincide with the middle of the bin. Given a measured sample, h_{ij}, one can therefore obtain an approximation $\hat{f}$ of the joint PDF using

$$\hat{f}(x_{o,i}, y_{o,j}) = \frac{h_{ij}}{\Delta x_i \Delta y_j}. \tag{4.70}$$

As for 1D distributions, the relative errors $\delta\hat{f}_{ij}/\hat{f}_{ij}$ are given by

$$\frac{\delta\hat{f}_{ij}}{\hat{f}_{ij}} = \frac{\sqrt{h_{ij}/N}}{h_{ij}} = \frac{1}{\sqrt{n}}\frac{1}{\sqrt{h_{ij}}} \to \frac{1}{\sqrt{n}}\frac{1}{\sqrt{p_{ij}}}. \tag{4.71}$$

There are several techniques to graphically represent 2D histograms. Figure 4.4 illustrates various popular representations: (a) a scatterplot where the density of points is proportional to the bin content, (b) z-color where a range of colors (or shades of gray) is used

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

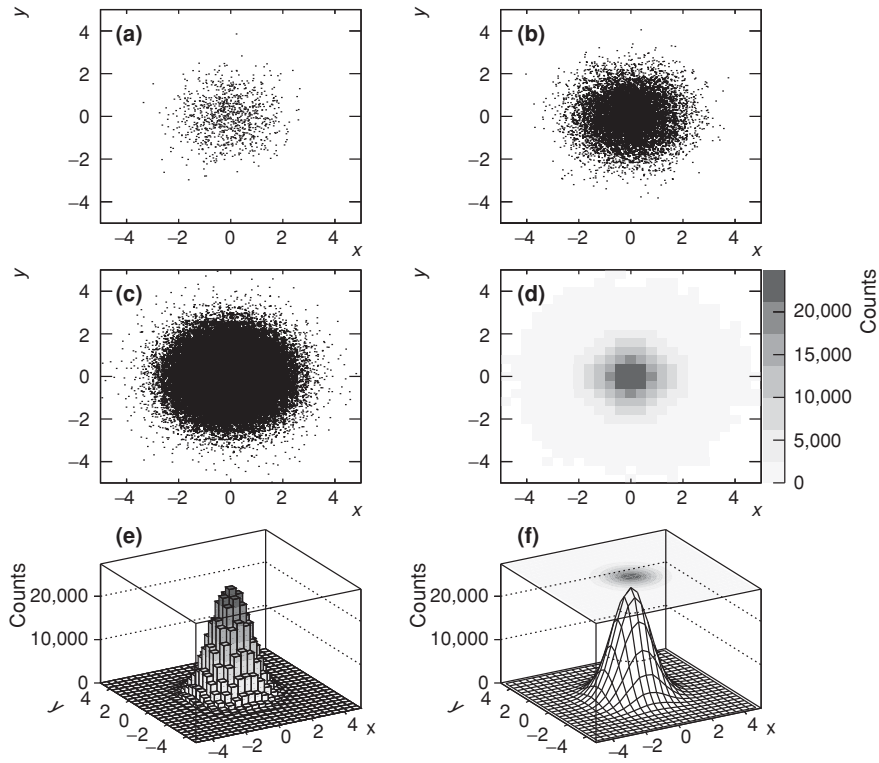


Fig. 4.4 Illustration of the notion of 2D histograms based on a 2D standard normal distribution ($\mu_x = \mu_y = 0$, $\sigma_x = \sigma_y = 1$): (a–c) scatterplots with 1,000, 10,000, and 1,000,000 entries. Panels (d, f) present alternative plotting methods (contour, lego, hybrid) for a histograms with 1,000,000 entries.

to represent the amplitude of the signal, (c) contour plot where lines show values of equal amplitude, as well as (d) lego plot, (e) surface plot, and (f) mixed z-color + surface where the amplitude is shown explicitly in a three-dimensional plot. Various other graphing techniques are available in plotting packages such as ROOT, MATLAB®, or Mathematica™. In principle, projections of the estimate of joint probability, h_{ij}, onto the x- and y-axes yield estimates of the marginal probabilities $\hat{f}_x(x)$ and $\hat{f}_y(y)$, respectively. Note, however, that since 2D histograms are typically filled using specific boundaries (both along the x- and y-axes), the projections are likely to be incomplete and not yield the full integrals given by Eqs. (2.106, 2.107) unless the probabilities p_{ij} are negligible outside the boundaries of the histogram. This means that the projections are estimates of the conditional probability densities $h_x(x|y_{\min} \leq y < y_{\max})$ and $h_y(y|x_{\min} \leq x < x_{\max})$. They provide PDFs of x and y given y and x, respectively, are in the ranges $y_{\min} \leq y < y_{\max}$ and $x_{\min} \leq x < x_{\max}$ defined by the boundaries of the histogram.

Slices of the histogram h_{ij} in y and x projected onto the x- and y-axis, respectively, may be used to obtain estimates of the conditional probabilities $h_x(x|y)$ and $h_y(y|x)$. Note, however, that the finite bin width used to create and fill histograms implies these estimates are valid

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

for ranges of y and x, respectively, corresponding to the widths of the bins along these two axes. Slices consisting of several rows or columns can of course be added together to improve the statistical significance of the data. But this brings about a loss in the resolution of the conditions on y and x used to define the slices.

4.6.4 Profile Histograms

In the study of multivariate systems, one often encounter situations in which a variable Y exhibits significant event-to-event fluctuations as well as an explicit dependence on some independent (control) variable X. A two-dimensional histogram can evidently be used to study the joint probability of the two variables, but in many cases one is interested primarily in the dependence of the average $E[Y]$ on the variable X. Creating a 2D histogram Y vs. X to study the dependence of Y on X may then be fastidious and somewhat of an overkill. Fortunately, **profile histograms** provide an easy and convenient alternative.

Profile histograms are based on the same notion of partition of the domain of one of several control variables X, and can be drawn in one, two, or multiple dimensions. However, rather than storing multiple content bins for the dependent variable (Y), for each bin of the variable X, one reduces the information to three quantities for each X bin: the number of entries in the bin i, hereafter noted n_i; an estimate of the mean of Y, $\langle y_i \rangle$; and the variance, $\langle \Delta y_i^2 \rangle$, of the signal Y in the bin. A profile histogram is "filled" similarly as a regular histogram by calling a "fill" function

$$\text{profileHisto.fill}(x, y) \tag{4.72}$$

for a pair of values (x, y), where x is an instance value of the independent variable X and y a jointly measured value of the dependent variable Y. However, rather than incrementing a bin content, a profile histogram implements Eqs. (4.50, 4.53) to obtain robust estimates of the mean, $\langle y_i \rangle$, and the variance, $\langle \Delta y_i^2 \rangle$, of the signal Y in each bin i based on successive entries y. The mean and variance of Y in a given bin i (of X) are thus calculated for successive entries in that bin according to

$$\Delta_i = \frac{y - \langle y_i \rangle}{n_i + 1}, \tag{4.73}$$

$$\langle y_i \rangle' = \langle y_i \rangle + \Delta_i, \tag{4.74}$$

$$\langle \Delta y_i^2 \rangle' = \frac{n_i}{n_i + 1} \left[(n_i + 1)\Delta_i^2 + \langle \Delta y_i^2 \rangle \right], \tag{4.75}$$

$$n_i' = n_i + 1, \tag{4.76}$$

where primed quantities stand for updated values obtained after the inclusion of the $(n_i + 1)$th entry. The profile histogram may then be examined at any stage of an analysis to obtain estimates of the mean, $\langle y_i \rangle$, and centered moment, $\langle \Delta y_i^2 \rangle$, in each bin i. Additionally, for sufficiently large number of entries, n_i, the variance $\langle \Delta y_i^2 \rangle$ may be used to estimate the error on the mean in each bin according to

$$\delta \langle y_i \rangle = \sqrt{\frac{\langle \Delta y_i^2 \rangle}{n_i}}. \tag{4.77}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

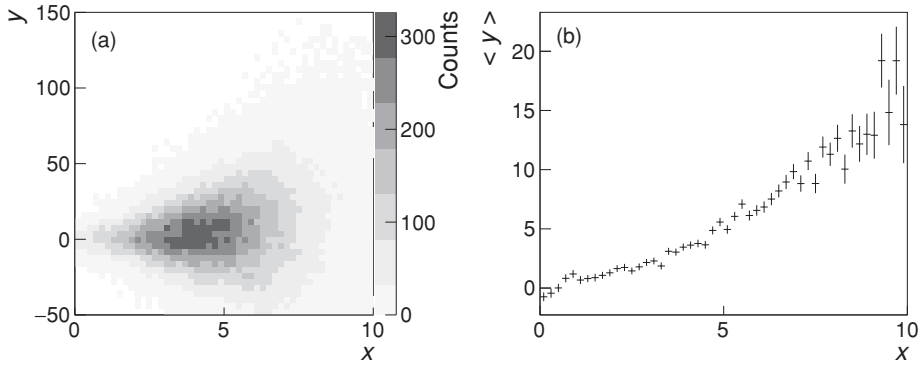


Fig. 4.5 Illustration of the concept of profile histogram. (a) 2D histogram showing the joint distribution of variables X and Y. (b) Profile histogram displaying the dependence of the average of Y on X. Error bars displayed indicate uncertainties on the mean of the signal Y.

This provides the additional advantage that it is readily possible to produce plots of the average $\langle Y \rangle(x)$ with errors bars corresponding to the error on the mean, or alternatively, the standard deviation of the signal.

Figure 4.5a presents an example of a joint distribution (2D histogram) of two variables X and Y, where the dependent variable Y exhibits large fluctuations at any given value of X. The dependence of Y on X, difficult to discern based on this 2D histogram, is readily identified in the profile histogram shown in panel (b) of the same figure.

4.7 Fisher Information

Given a probability model $p(x|\vec{\theta}, H, I)$ for an observable X and model H, it is legitimate to ask how measurements of X may constrain the values of the model parameters $\vec{\theta}$, that is, how much information is contained in the data and the functional form of $p(x|\vec{\theta}, H, I)$ that can constrain the value of the parameters $\vec{\theta}$. Two distinct quantities, the entropy S and the Fisher information $\mathbb{I}(\theta)$, are commonly used to characterize this information. The Fisher information and examples of calculation for specific PDFs are presented in this section, whereas the notion of entropy S and its connection to the Fisher information are discussed in §7.3.1. A technique to obtain a minimal bound on the variance of model parameters based on the Fisher information is presented in §4.7.6.

This section may be skipped in a first reading of this chapter.

4.7.1 A Basic Derivation of the Fisher Information

The function $p(x|\vec{\theta}, H, I)$ amounts to the probability (density) of observing values of x given model parameters $\vec{\theta}$. Changing the values of the model parameters modifies the probability and the likelihood of x being observed in any given range: the parameters determine

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

the distribution of observable values x. One can thus say that the PDF $p(x|\vec{\theta}, H, I)$ contains some kind of "information" that dictates what values x may be observed given specific parameters $\vec{\theta}$. Conversely, given a measured value of x, the functional form of a PDF should also provide information about the (unknown) model parameters $\vec{\theta}$. Let us then examine how one can quantify this information.

For simplicity's sake, let us first restrict the discussion to a single parameter PDF $p(x|\theta, H, I)$. Changing θ by an increment $\Delta\theta$ produces a change in probability Δp approximately equal to

$$\Delta p \approx \frac{\partial p(x|\theta, H, I)}{\partial \theta} \Delta\theta. \tag{4.78}$$

The derivative $\partial p/\partial\theta$ determines how a change in θ affects the probability distribution. It also influences how a change in the measured distribution (i.e., values of x being more or less probable) impacts the most likely values of the parameter θ. The slope function $\partial p/\partial\theta$ thus contains much of the information one needs to establish a connection between the data and the model parameter. The magnitude of the probability (density) also matters, however. A given increment Δp should be far more significant if it effects a comparatively large change on the probability density. It is thus more relevant to consider relative changes of the density, that is, how the ratio $\Delta p/p$ is affected by a change $\Delta\theta$ of the parameter θ. To this end, one introduces the notion of **score**, hereafter noted ξ, and defined as

$$\xi(x|\theta) = \frac{\partial}{\partial\theta} \ln p(x|\theta, H, I)\bigg|_\theta \tag{4.79}$$

$$= \frac{1}{p}\frac{\partial p}{\partial\theta}\bigg|_\theta. \tag{4.80}$$

A high score means a given increment $\Delta\theta$ has a large impact on the probability density whereas a small score produces only a modest change. Evidently, given the density is a function of x, the score is itself also a function of x. One should thus consider the average score of the distribution; that is, one should calculate its expectation value $E[\xi]$. It is straightforward to verify, however, that this expectation value vanishes for all values of θ. Indeed, for "well-behaved" distributions, one has

$$\mathrm{E}[\xi] = \int \frac{1}{p(x|\theta)} \frac{\partial p(x|\theta)}{\partial\theta} p(x|\theta)\, dx, \tag{4.81}$$

$$= \int \frac{\partial p(x|\theta)}{\partial\theta}\, dx = \frac{\partial}{\partial\theta} \int p(x|\theta)\, dx, \tag{4.82}$$

$$= \frac{\partial}{\partial\theta} 1 = 0, \tag{4.83}$$

where in the second line, we assumed the density $p(x|\theta)$ is a regular function such that the order of the integral vs. x and the derivative with respect to θ may be interchanged, while in the third line, we used the normalization of the function. Clearly, the foregoing null expectation value is of little interest because it does not provide any means to use the score to extract information about the probability density $p(x|\theta)$. The second moment of

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

the score, called Fisher information $\mathbb{I}(\theta)$, does, however.

$$\mathbb{I}(\theta) = \int \left[\frac{1}{p(x|\theta)} \frac{\partial p(x|\theta)}{\partial \theta} \right]^2 p(x|\theta)\, dx, \tag{4.84}$$

Since the integrant is by construction nonnegative, the Fisher information $\mathbb{I}(\theta)$ is positive definite for all values of θ, and its magnitude constitutes an indicator of the amount of information contained in the PDF. Probability densities with high scores in regions of high probability yield large integral values whereas densities with small scores yield small values of $\mathbb{I}(\theta)$. The Fisher information $\mathbb{I}(\theta)$ indeed constitutes an indicator of the information contained in a PDF $p(x|\theta)$.

It is interesting to note that given the expectation value of the score is null, the Fisher information amounts to the variance of the score. Densities with large Fisher information have a large score variance and provide better capabilities to determine and constrain model parameters.

The Fisher information may also be calculated in terms of the expectation value of the second derivative of the score. To demonstrate this statement, first compute the second derivative of the score relative to θ

$$\frac{\partial^2}{\partial \theta^2} \ln p(x|\theta) = \frac{\partial}{\partial \theta} \left(\frac{1}{p} \frac{\partial p}{\partial \theta} \right), \tag{4.85}$$

$$= \frac{1}{p} \frac{\partial^2 p}{\partial \theta^2} - \frac{1}{p^2} \left(\frac{\partial p}{\partial \theta} \right)^2, \tag{4.86}$$

and calculate its expectation value across the domain of x. Since

$$\mathrm{E}\left[\frac{1}{p} \frac{\partial p}{\partial \theta} \right] = \int \frac{\partial p}{\partial \theta}\, dx = \frac{\partial}{\partial \theta} \int p\, dx = 0, \tag{4.87}$$

$$\mathrm{E}\left[\frac{1}{p} \frac{\partial^2 p}{\partial \theta^2} \right] = \int \frac{\partial^2 p}{\partial \theta^2}\, dx = \frac{\partial^2}{\partial \theta^2} \int p\, dx = 0, \tag{4.88}$$

it follows, using Eq. (4.84), that

$$\mathbb{I}(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial \theta^2} \ln p(x|\theta) \right], \tag{4.89}$$

where the expectation value is evaluated at a fixed value of θ.

4.7.2 Fisher Information Matrix

For a probability model involving m parameters $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$, the Fisher information becomes a covariance matrix with elements

$$\mathbb{I}_{ij}(\vec{\theta}) = \mathrm{Cov}\left[\frac{\partial}{\partial \theta_i} \ln p(x|\vec{\theta}), \frac{\partial}{\partial \theta_j} \ln p(x|\vec{\theta}) \right], \tag{4.90}$$

which for regular functions may be written as (Problem 4.19)

$$\mathbb{I}_{ij}(\vec{\theta}) = -\mathrm{E}\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln p(x|\vec{\theta}), \right]. \tag{4.91}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

4.7.3 Fisher Information Properties

The Fisher information features several important properties.

Additivity

Given two observables X and Y, the information obtained from independent measurements of these observables may be written

$$\mathbb{I}_{x,y}(\theta) = \mathbb{I}_x(\theta) + \mathbb{I}_y(\theta), \tag{4.92}$$

where $\mathbb{I}_x(\theta)$ and $\mathbb{I}_y(\theta)$ are the information obtained from measurements of X and Y, respectively. The additivity results from the fact that the variance of a sum of two independent variables is equal to the the sum of their respective variances.

Scaling

The property of additivity is readily extended to an "experiment" consisting of n elementary measurements of an observable X sampled from the same distribution $p(x|\theta)$. The Fisher information of the likelihood $p(\vec{x}|\theta)$ of the measured data $\vec{x} = (x_1, x_2, \ldots, x_n)$ is equal to n times the Fisher information of an individual measurement.

$$\mathbb{I}_{\vec{x}}(\theta) = n\mathbb{I}_x(\theta). \tag{4.93}$$

In this context, the Fisher information obtained from a single measurement of an observable X is commonly called **unit Fisher information** to distinguish it from that obtained from a sample of n measurements.

Transformation

The Fisher information of a probability model parameter θ depends on the derivative of the distribution relative to this parameter. But the parameterization of a model can be changed without changing its meaning. For instance, with a Gaussian PDF, specification of the width of the distribution may be done in terms of the standard deviation σ, the variance σ^2, or the precision $\xi = \sigma^{-2}$. These three parameters have different meanings, physical units, and numerical values, but the Gaussian PDF they define remains the same. In some fashion, the information associated to these different parameters cannot be too different. Indeed, their information must be related to one another. More generally, it is interesting to consider how the Fisher information $\mathbb{I}(\vec{\theta})$ transforms under a **reparametrization** $\vec{\theta} \to \vec{\lambda}(\vec{\theta})$.

Let us first consider a probability model involving a single parameter θ and calculate the information in terms of λ based on $\theta \equiv \theta(\lambda)$

$$\mathbb{I}_\lambda(\lambda) = \int \left[\frac{1}{p(x|\theta(\lambda))} \frac{\partial p(x|\theta(\lambda))}{\partial \lambda} \right]^2 p(x|\theta(\lambda))\, dx, \tag{4.94}$$

where we included the dependency of the parameter θ on λ explicitly to emphasize that the information is considered a function of λ rather than θ. Using a chain rule for the partial

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

derivative, we get

$$\mathbb{I}_\lambda(\lambda) = \int \left[\frac{1}{p(x|\theta)}\frac{\partial p(x|\theta)}{\partial \theta}\frac{\partial \theta}{\partial \lambda}\right]^2 p(x|\theta)\,dx, \tag{4.95}$$

$$= \left(\frac{\partial \theta}{\partial \lambda}\right)^2 \int \left[\frac{1}{p(x|\theta)}\frac{\partial p(x|\theta)}{\partial \theta}\right]^2 p(x|\theta)\,dx, \tag{4.96}$$

where the probability is now considered a function of θ rather than λ and the square of the derivative $\partial\theta/\partial\lambda$ was factored out of the integral because it is independent of the observable x. We thus conclude that under a reparameterization $\theta \equiv \theta(\lambda)$, the Fisher information transforms according to

$$\mathbb{I}_\lambda(\lambda) = \mathbb{I}_\theta(\theta(\lambda))\left(\frac{\partial \theta}{\partial \lambda}\right)^2. \tag{4.97}$$

For a multiparametric model involving m parameters, $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$, we saw in §4.7.2 that the information takes the form of a matrix. We may then consider how the information changes under a transformation $\vec{\theta} \to \vec{\lambda}$ defined by the $m \times m$ Jacobian matrix

$$\mathbf{J}_{ij} \equiv \frac{\partial \theta_i}{\partial \lambda_j}, \tag{4.98}$$

where the indices i and j span the m model parameters. Repeating the chain rule derivation, Eq. (4.95), for each pair of parameters $i, i' \to j, j'$, one finds

$$\mathbb{I}_{\lambda_j,\lambda_{j'}}(\vec{\lambda}) = \mathbb{I}_{\theta_i,\theta_{i'}}(\vec{\theta})\left(\frac{\partial \theta_i}{\partial \lambda_j}\right)\left(\frac{\partial \theta_{i'}}{\partial \lambda_{j'}}\right). \tag{4.99}$$

The transformation may also be written in a convenient matrix form

$$\mathbb{I}_{\vec{\lambda}}(\lambda) = \mathbf{J}^T \mathbb{I}_{\vec{\theta}}(\vec{\theta}(\vec{\lambda}))\mathbf{J}. \tag{4.100}$$

We will use the transformation property of the Fisher information when discussing the notion of information invariance forming the basis of Jeffreys' priors introduced in §7.3.2.

4.7.4 Example 1: Fisher Information of a Binomial PDF

Consider a nuclear physics experiment involving a measurement of the rate of production of a certain particle species in a specific nuclear collision at a specific beam energy. The measurement is carried out with a given detector and within a specific kinematic range. Nominally, produced particles by the collisions are detected and counted within the apparatus. However, particle losses associated with various instrumental causes invariably occur. The number of particles detected, n, thus represents only a fraction of the number produced N and one defines the particle detection efficiency as the ratio of the number of observed particles by the number of produced particles:

$$\varepsilon \equiv \frac{n}{N}. \tag{4.101}$$

The efficiency is largely determined by the design of the apparatus but may be influenced by a host of factors internal and external to the apparatus. In practice, it is typically difficult

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

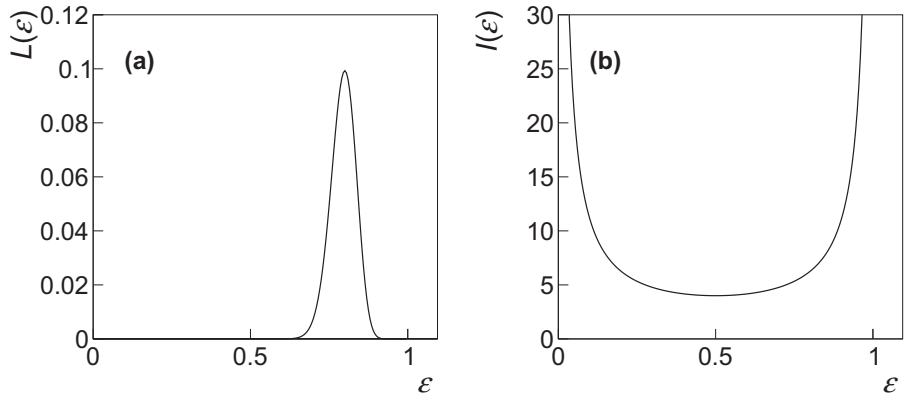


Fig. 4.6 (a) Plot of the likelihood function of an efficiency parameter ε given the successful detection of $n = 80$ particles and the prior knowledge that $N = 100$ particles actually passed through the detection apparatus. (b) Fisher information $\mathbb{I}(\varepsilon)$ contained in the binomial distribution plotted as a function of the efficiency (success rate) parameter ε.

to carry out, ab initio, a precise calculation of the efficiency of a specific device and one is often required to carry out a controlled measurement where both n and N are explicitly determined to calculate the efficiency. The difficulty arises that a measurement is intrinsically a stochastic process, akin to the toss of a coin. For a given particle passing through the detector, there is a probability ε that the particle will be detected and a probability $1 - \varepsilon$ that it will not. Assuming, as in §3.1.1, that the detection of any given particle is independent to that of others, the detection process can be modeled with a Bernoulli probability model

$$p(x|\varepsilon) = \varepsilon^x (1-\varepsilon)^{1-x}, \tag{4.102}$$

where a value $x = 1$ represents a success, that is, a particle being detected, and a value $x = 0$ corresponds to a "failure," that is, a particle remaining undetected. Clearly, just a single toss of a coin is insufficient to determine whether a coin is fair, one needs to attempt the detection of a large number of particles to obtain an estimate of the efficiency. Let us thus consider a succession of N independent particles passing through the detector. As for coin tosses, the total number of success n (i.e., particle detected) relative to the number of trials N shall be given a binomial distribution

$$p(n|N,\varepsilon) = \frac{N!}{n!(N-n)!}\varepsilon^n (1-\varepsilon)^{N-n}. \tag{4.103}$$

The function $p(n|N,\varepsilon)$ determines the likelihood of n particles being observed given N are produced when the detection efficiency is ε. But ε is unknown, and it is the purpose of the experiment to determine its value. One can then invoke the principle of maximum likelihood, discussed in §5.1, and obtain the $\mathbb{ML}$ estimate of ε as n/N, in which n is the number of actually detected particles, and N is the number of particles known to pass through the detector.

The likelihood function $p(n|N,\varepsilon)$ is plotted as a function of ε in Figure 4.6a for $N = 100$ and $n = 80$. One finds that the function has a maximum at $\varepsilon = 0.8$ corresponding indeed to the $\mathbb{ML}$ estimator value $80/100$. We evaluate the (unit) Fisher information of the binomial

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

PDF using Eq. (4.84):

$$\mathbb{I}(\varepsilon) = -\mathrm{E}\left[\frac{\partial^2}{\partial\varepsilon^2}\ln p(x|\varepsilon)\right] \tag{4.104}$$

$$= -\mathrm{E}\left[\frac{\partial^2}{\partial\varepsilon^2}\ln\left(\frac{N!}{n!(N-n)!}\varepsilon^n(1-\varepsilon)^{N-n}\right)\right] \tag{4.105}$$

$$= \mathrm{E}\left[\frac{n}{\varepsilon^2} + \frac{N-n}{(1-\varepsilon)^2}\right], \tag{4.106}$$

where ε is considered a constant for the purpose of the evaluation of the expectation value. Given the expectation value of n is equal to $N\varepsilon$, we find the Fisher information for a binomial distribution is

$$\mathbb{I}(\theta) = \frac{1}{\varepsilon^2}\mathrm{E}[n] + \frac{N-E[n]}{(1-\varepsilon)^2}, \tag{4.107}$$

$$= \frac{N\varepsilon}{\varepsilon^2} + \frac{N-N\varepsilon}{(1-\varepsilon)^2}, \tag{4.108}$$

$$= \frac{N}{\varepsilon(1-\varepsilon)}. \tag{4.109}$$

Equation (4.109) is plotted in Figure 4.6b. One observes that for values $\varepsilon = 1$ and $\varepsilon = 0$, the information diverges thereby reflecting the fact that the outcome of N trials is absolutely certain, that is, all or none of the particles are detected. The information reaches a minimum at $\varepsilon = 0.5$ corresponding to the value of ε for which outcomes are least certain. Indeed, for $\varepsilon = 0.5$, the odds of getting a success (particle detected) or failure (not detected) are equal and the outcome of a trial (measurement) is thus maximally uncertain.

4.7.5 Example 2: Fisher Information of a Gaussian Distribution

Let us calculate the Fisher information matrix of a Gaussian PDF:

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{4.110}$$

with unknown mean μ and variance σ^2. As per the definition, Eq. (4.91),

$$\mathbb{I}(\mu,\sigma^2) = -E\left[\begin{pmatrix} \frac{\partial^2}{\partial\mu^2}\ln p(x|\mu,\sigma^2) & \frac{\partial^2}{\partial\mu\partial\sigma^2}\ln p(x|\mu,\sigma^2) \\ \frac{\partial^2}{\partial\sigma^2\partial\mu}\ln p(x|\mu,\sigma^2) & \frac{\partial^2}{\partial\sigma^2\partial\sigma^2}\ln p(x|\mu,\sigma^2) \end{pmatrix}\right], \tag{4.111}$$

this requires calculation of derivatives of the log of the PDF with respect to μ and σ^2. First taking the log of $p(x|\mu,\sigma^2)$, we get

$$\ln p = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}. \tag{4.112}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

Second derivatives of $\ln p$ yield

$$\frac{\partial^2}{\partial \mu^2} \ln p = -\frac{1}{\sigma^2}, \tag{4.113}$$

$$\frac{\partial^2}{\partial \sigma^2 \partial \mu} \ln p = -\frac{x-\mu}{\sigma^4}, \tag{4.114}$$

$$\frac{\partial^2}{\partial \sigma^2 \partial \sigma^2} \ln p = \frac{1}{2\sigma^4} - \frac{(x-\mu)}{\sigma^6}. \tag{4.115}$$

Inserting the preceding expressions in Eq. (4.111), and remembering that the expectation value of x is the mean μ, and that of $(x-\mu)^2$ is σ^2, the Fisher information matrix of a Gaussian PDF may then be written

$$\mathbb{I}(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}. \tag{4.116}$$

We find that the information one may gain on the mean by measurements of x is inversely proportional to the variance σ^2. For a small or vanishing variance, the information diverges and the precision achievable in the determination of μ is thus very large. Conversely, for a large variance, the information is small, and the achievable precision rather limited. As for the variance, one finds that the information contained in $p(x|\mu, \sigma^2)$ is inversely proportional to the square of its value. Again in this case, one concludes that if the variance is small, it shall be relatively straightforward to obtain precise estimates of its value but estimates shall remain rather imprecise if σ^2 is large. It is important to note that the inverse square dependence implies that the variance may be more difficult to determine precisely than the mean. Additionally note that the diagonal terms are strictly null, implying that the information on μ and σ^2 are not correlated (although, technically speaking, the information on μ is entirely determined by σ^2).

In general, the off-diagonal elements of the Fisher information matrix of a PDF are nonvanishing. However, whenever $\mathbb{I}_{ij}(\vec{\theta}) = 0$ for $i \neq j$, the parameters θ_i and θ_j are said to be orthogonal. In the case of the Gaussian, we found the off-diagonal elements are null, thereby implying that μ and σ are orthogonal; that is, the determination of μ does not influence the determination of σ^2, and conversely. This largely stems from the symmetric nature of the Gaussian PDF. By contrast, PDFs that are not symmetric about their mean shall tend to have nonvanishing nondiagonal elements, and the determination of the corresponding parameters will be mutually correlated.

As we discuss next, the Fisher information of a given parameter determines a minimal variance bound achievable in an experimental measurement of its value. This means that the very shape and definition of a PDF determine the precision achievable in the experimental determination of its parameters.

4.7.6 Minimal Variance Bound

By definition, an unbiased estimator satisfies:

$$\langle \hat{\theta} \rangle = E[\hat{\theta}] = \int \hat{\theta} L \, dX = \theta, \tag{4.117}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

where L is the likelihood of the measured data $\vec{x} = (x_1, x_2, \ldots, x_n)$ and dX stands for $dx_1 dx_2 \cdots dx_n$. We differentiate this expression with respect to θ. Since the estimator depends on the x_i alone, and assuming the limits of integration are independent of θ, one gets

$$\int \hat{\theta} \frac{dL}{d\theta}\, dX = 1. \tag{4.118}$$

Next, consider the normalization integral of L:

$$\int L\, dX = 1. \tag{4.119}$$

Differentiating with respect to θ produces

$$\int \frac{dL}{d\theta}\, dX = 0. \tag{4.120}$$

We then rewrite Eqs. (4.118, 4.120) as follows:

$$\int \hat{\theta} \frac{d\ln L}{d\theta} L\, dX \equiv \left\langle \hat{\theta} \frac{d\ln L}{d\theta} \right\rangle = 1, \tag{4.121}$$

$$\int \frac{d\ln L}{d\theta} L\, dX \equiv \left\langle \frac{d\ln L}{d\theta} \right\rangle = 0. \tag{4.122}$$

Next, multiply the second line by θ and subtract from the first line to get

$$\int \left(\hat{\theta} - \theta\right) \frac{d\ln L}{d\theta} L\, dX = \left\langle \left(\hat{\theta} - \theta\right) \frac{d\ln L}{d\theta} \right\rangle = 1. \tag{4.123}$$

We can now use the Schwarz inequality

$$\int u^2\, dX \int v^2\, dX \geq \left(\int uv\, dX \right)^2 \tag{4.124}$$

with $u = (\hat{\theta} - \theta)\sqrt{L}$ and $v = \frac{d\ln L}{d\theta}\sqrt{L}$ to obtain

$$\left(\int (\hat{\theta} - \theta)^2 L\, dX \right) \left(\int \left(\frac{d\ln L}{d\theta} \right)^2 L\, dX \right) \geq 1. \tag{4.125}$$

The first integral corresponds to the variance $\langle(\hat{\theta} - \theta)^2\rangle$ of the estimator while the second integral is the average of the derivative of the logarithm of L with respect to θ. One then arrives at the lower bound:

$$V\left[\hat{\theta}\right] \geq \frac{1}{\left\langle (d\ln L/d\theta)^2 \right\rangle}. \tag{4.126}$$

The quantity $\left\langle (d\ln L/d\theta)^2 \right\rangle$ is the **Fisher information**, $\mathbb{I}(\theta)$, contained in the likelihood function L. Based on Eq. (4.89), we find that it may also be expressed (see also Problem 4.2)

$$I(\theta) = \left\langle d^2 \ln L/d\theta^2 \right\rangle. \tag{4.127}$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

Alternatively, one can also insert the expression for L in the preceding equation, and obtain (for a single parameter):

$$I(\theta) = -n \int \frac{d^2 \ln p(x|\theta)}{d\theta^2} p(x|\theta)\, dx. \tag{4.128}$$

This expression and Eq. (4.126) can be used to obtain the minimum sample size n required to achieve a given measurement precision of the parameter θ.

Exercises

4.1 Derive the expression given by Eq. (4.16) for the Mean Square Error (MSE).

4.2 Derive the expression Eq. (4.127) for $I(\theta)$.

4.3 Derive the expression Eq. (4.128) for $I(\theta)$.

4.4 Show that for a biased estimator, one must replace the numerator of Eq. (4.17) with $1 + db/d\theta$.

4.5 Verify that the standard error on the sample mean estimator is given by $\sigma/\sqrt{n}$ as in Eq. (4.30).

4.6 Show that S^2, defined by Eq. (4.32), is an unbiased estimator of the variance σ^2 when the population mean, μ, is known.

4.7 Show that s^2, defined by Eq. (4.39), is an unbiased estimator of the variance σ^2 when the population mean, μ, is not known.

4.8 Show that $\hat{V}_{xy}$, defined by Eq. (4.43), is an unbiased estimator of the covariance V_{xy}.

4.9 Show that the expression $\hat{m}_k = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^k$ provides an unbiased estimator of the k^{th} central moment, μ_k.

4.10 Find unbiased estimators for the moments μ'_k and central moments μ_k.

4.11 Show that the variance of estimators $\widehat{\mu}'_k$ of the moments μ'_k is given by

$$V[\widehat{\mu}'_k] = \frac{1}{n}\left(\mu'_{2k} - \left(\mu'_k\right)^2\right).$$

4.12 Show that the covariance of estimators of the moments μ'_r and μ'_q is given by

$$\mathrm{Cov}[\hat{\mu}'_r, \hat{\mu}'_q] = \frac{1}{n}\left(\mu'_{r+q} - \mu'_r \mu'_q\right).$$

4.13 Show that the variance of the estimator of the standard deviation $\hat{\sigma}$ is given by the following expression in the large n limit:

$$V[\hat{\sigma}] = \frac{\mu_4 - (\mu_2)^2}{4n\sigma^2}.$$

Additionally, show that for a Gaussian PDF, this expression reduces to

$$V[\hat{\sigma}] = \frac{\sigma^2}{2(N-1)}.$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press

4.14 Show that the standard error of the estimator $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is given by

$$V[s^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2\right),$$

where $\hat{\mu}_k$ are central moments of order k given by

$$\hat{\mu}_k = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^k .$$

4.15 Show that the expectation value and variance of the Pearson coefficient estimator $\hat{r}_{xy}$, defined by Eq. (4.44), are given by the following expressions:

$$E[\hat{r}_{xy}] = \rho - \frac{\rho(1-\rho^2)}{2n} + O(n^{-2})$$
$$V[\hat{r}_{xy}] = \frac{1}{n}(1-\rho^2)^2 + O(n^{-2}).$$

4.16 Prove the assertion that the estimator $\hat{r}_{xy}$ defined by Eq. (4.44) is bound in the range $[-1, 1]$.

4.17 Derive the expression Eq. (4.17) for robust centered moments of order k, and verify the expressions given by Eq. (4.53) and Eq. (4.54) for centered moments of second and third order, respectively.

4.18 Show the following expression yields a biased but asymptotically unbiased estimator $\hat{\lambda}$ of the decay constant $\lambda = 1/\tau$:

$$\hat{\lambda} = \frac{1}{\hat{\tau}} = n\left(\sum_{i=1}^{n} t_i\right)^{-1}. \tag{4.129}$$

4.19 Show that the Fisher information matrix defined by Eq. (4.90) may be written as

$$\mathbb{I}_{ij}(\vec{\theta}) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln p(x|\vec{\theta}),\right].$$

https://doi.org/10.1017/9781108241922.006 Published online by Cambridge University Press