1 Basic statistical concepts

1.1 INTRODUCTION

This chapter and the next chapter are essential for understanding the content of this book. By design, the chapters present statistics from a quite different perspective as usually the statistics is introduced and taught. The theory of statistics is presented from the pure Bayesian perspective where we attempt to make sure that concepts of the classical statistics are not mixed in the exposition of the theory. In our experience, the Bayesian statistics is frequently introduced in image and signal analysis texts as an extension of the classical treatment of probability. The classical treatment of probability is based on the interpretation of probability as the frequency of occurring of some phenomena based on repeated identical trials. The classical approach is often referred to as the *frequentist* statistics. From the Bayesian point of view, the probability describes the strength of beliefs in some propositions. One of the most frequently used terms, the probability distribution, in frequentist statistics means the "histogram" of outcomes of the infinite number of repetitions of some experiment. In Bayesian statistics, the probability distribution quantifies beliefs or in other words measure of uncertainty. Unfortunately, these two concepts of probability, Bayesian and frequentist, are not compatible and cannot be used together in a logically coherent way. What creates confusion is that both approaches are described mathematically by the probability calculus and because of that they can be intermingled and used together which, to us at least, is incomprehensible.

In this book we decided not to introduce classical concepts at all. To help the reader who is accustomed to thinking about the probability as a frequency, we intentionally do not use the term *random variable*. This is because the random variable is strongly associated with the concept of frequency. To avoid any unwanted associations, the term random variable is replaced in this book by the term *quantity*. The classical term *parameter* is not used in this book either. In the classical statistics, parameters describe unknown values and inference about those parameters is obtained in classical statistical procedures. Instead of the term "parameter" the term *quantity* is used as well. Both the "random variable" and the "parameter" are put on the same conceptual level and are referred to as *quantities*. Finally, in the classical statistics the term *data* is used to describe the outcome of experiments. Based on the data, inferences about parameters are made in frequentist statistics. In the Bayesian view utilized here, the term data is another *quantity* which is conceptually the same as quantities corresponding to random variables or quantities corresponding to parameters. For this quantity we relax our naming rule and use interchangeably the data and the quantity to describe outcomes of the experiments.

It may appear that such convention creates confusion because there is a single term "quantity" to describe so many phenomena. There is more to gain than lose as we believe that this naming convention helps considerably with understanding of Bayesian concepts. In order to help differentiating different quantities, we will use adjectives *observable* and *unobservable* added to the term quantity that identify which quantities are revealed in the experiment (correspond to "data" in classical treatment) and which are never revealed (correspond to parameters in classical statistics).

1.2 BEFORE- AND AFTER-THE-EXPERIMENT CONCEPTS

In this chapter, a specific view on processes that involve uncertainty will be considered. The author hopes that the approach will allow to smoothly introduce concepts that are frequently poorly explained or misunderstood. The content of this book is concerned about knowledge of *quantities* that can, or cannot, be observed directly in an experiment. Such quantities will be referred to as observable and unobservable quantities, respectively. Interchangeably, we will refer to knowledge about quantities as beliefs. We will also use uncertainty about the quantity which is the opposite term to knowledge. For *unobservable quantities* (UQs) the true value of the quantity is unknown (uncertain). For example, suppose we are interested in a true weight of some object. This quantity cannot be observed (determined) directly and the true weight is unknown. By unobservable directly we mean that there is no experiment that can reveal the true value of that quantity. The *observable quantities* (OQs) will be those where the true values are revealed by the experiment. For example when weighing an object the reading from the scale is an observable quantity. Obviously, the true weight of the object (unobservable quantity) and the reading from the scale (observable quantity) are two different quantities and are not necessarily equal.

> **Important:** Here an important distinction has to be made. The weight of the object and the result of the measurement are two different quantities. The weight is uncertain before and after the experiment; however, the measurement is uncertain before the experiment (we do not know what the reading on the scale will be), but it is known exactly after the experiment. Therefore the quantity which is the measurement is revealed and known exactly. The true weight remains uncertain.

The quantities that we will be interested in are going to be referred to in this book as the *quantities of interest* (QoIs) which include UQs and OQs. Sometimes quantities that are known will be required to fully describe a problem at hand (when considering the radioactive decay such quantities can be the half-life or decay constant for given radiotracer). These quantities will be referred to as *known quantities* (KQs). The values of all QoIs constitute the objective truth that will be referred to as the "state of nature" (SoN). Obvi-

ously the KQs also describe the SoN but since they are known at all stages of the experimentation they are not considered as a part of QoIs. We require that the SoN is defined by at least one QoI. We assume that knowledge of the true SoN implies the knowledge of all true values of QoIs that define it and vice versa. All true values of QoIs, observable and unobservable, define the SoN.

Although some QoIs are not observable, we will be able to make a guess about the true value based on our general knowledge and experience and maybe some experiments that shed some light on the true values of the QoIs that were done in the past. There are two extremes in the amount of information that we can have about a QoI. A perfect knowledge is when we know the true value of the quantity and the least knowledge is when we have no indication which of the possible values of the quantity is the true value.

One way to think about asking how accurate is the information regarding some QoI is to think about a range of possible true values of this quantity. If the number of such values is small, we say that our information is more precise, or better, than information in the case where the number of possible values is larger. In the extreme, for a single possible value, the knowledge is "perfect" and no uncertainty is involved. The knowledge is perfect from the definition for all QoIs that are observable after the experiment performed. If all QoIs are OQs, after the experiment all values are certain, the SoN defined by those quantities is therefore known and statistical description is not necessary.

The goal of any experiment is to improve knowledge about QoIs and the SoN defined by those QoIs. For OQs this improvement is obvious as the true values of those QoIs are simply revealed and the knowledge about them becomes perfect (we know the true values of the QoIs) once the experiment is performed. Sometimes we will refer to those true values of OQ as *experimental data*, *data*, or *observations*. We often will say that the OQ is revealed or realized in the experiment as opposed to hidden, uncertain, or unknown. Based on observable QoIs that are revealed, some additional information about unobservable QoIs will be obtained. This process will be referred to as the *statistical inference*. We deliberately do not use the term *random variable* to describe the QoI, because the word "random" is misleading and makes it difficult to understand the line of reasoning employed in this work. The quantities we refer to as UQ and OQ are deterministic and constant and using the term "random" when referring to them would be confusing. Another deviation from the other authors is that the term *parameter* typically used in the literature to describe some unknown property of the state of nature is not used. The closest correspondence to the classical term "parameter" used in this book is the UQ. We do however place OQs and UQs on the same conceptual level and consider them as quantities that define the SoN.

We consider two stages at which the information about the SoN is evaluated: before-experiment (BE) and after-experiment (AE). When considering the SoN after the experiment (AE), the uncertainty about SoN is described

only by the UQs. In the AE stage, the OQs are no longer uncertain and are known; therefore, no uncertainty about them can contribute to uncertainty about the SoN. Just to be sure that there is no misunderstanding, the OQ is the measurement (e.g., reading on the scale) and not the value of the quantity that it attempts to estimate. If the goal of the investigation is to obtain insights about the SoN, it is obvious to consider only the AE stage. However, BE state is also interesting when we do not want to tie conclusions about the SoN to actual observations of OQs, but rather consider all possible observations that can occur. This can be important when our task is to optimize imaging systems in which case we need to consider all possible values of OQs that can be obtained with that imaging system.

Let's consider the following example of the before-experiment (BE) and after-experiment (AE) concepts and a single OQ.

Example 1.1: Single-die roll (1)

The experiment is defined as a simple roll of a six-sided die. The result of the roll is the observable QoI. The SoN is defined by the number obtained in the roll. In the BE state, there are six possible true values of the QoI. The experiment is performed and the number six is obtained. Therefore, the AE state (after the roll) OQs are revealed (realized) so the true value of the QoI (six) is known. In the AE state, there is no uncertainty for this example as the SoN is defined by a single QoI that is known in the AE state.

The same concepts can be illustrated using a more sophisticated example in which the SoN is defined by two QoIs in which one is observable and one is unobservable:

Example 1.2: Radioactive sample (1)

Let's consider a sample of some radioactive material that is expected to create f radioactive decays (total activity) per second. The value of f is unknown and considered as the *unobservable quantity* since it cannot be observed directly. Therefore, per our definition, the value f is a UQ. The experiment is performed that involves observation of g decays from the radioactive sample during some period of time using a detector that registers photons emitted from the sample. The *sensitivity* of the detection is assumed known. The sensitivity is the deterministic constant (KQ) indicating the average of the ratio of the number of detected photons to the number of emitted photons. For simplicity we assume that we have a perfect efficiency and therefore 100% of emissions are registered by the detector. The number of detected counts is a OQ. The SoN (defined by f and g) is uncertain BE and AE; however, it seems that AE we have more information about the SoN as one of the QoIs that defines the SoN is known. Looking

slightly ahead, the main idea of statistical inference is that the observation of g counts registered by the detector not only reveals QoI g but also improves the knowledge about the activity f (the UQ); therefore, observations not only reduce the number of unknown QoIs but can also improve the knowledge about the UQs.

The concept of the before and after the experiment conditions introduced in this section is illustrated in Figure 1.1. Before experiment (left column of Figure 1.1) the possible true values of the QoIs f and g are from 0 to ∞. After the experiment is performed (right column of Figure 1.1) the OQ is observed and at this stage it is known and equal to g. The UQ f is still unknown (the true value is uncertain) and for the example presented here the initial range of possible true values remains unchanged (0 to ∞).

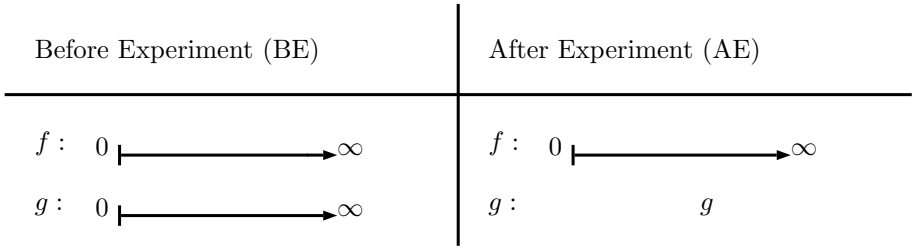


FIGURE 1.1 Before-experiment and after-experiment concepts.

To summarize, an important concept was introduced that will be used throughout this book. We refer to this concept as the BEAE concept where the BEAE stands for the before-experiment-after-experiment concept. The term experiment indicates a process in which the true values of some quantities are observed (referred to as observable quantities or OQs) which before experiment are unknown. At least one quantity of interest (QoI) should not be observed in the experiment, otherwise no uncertainty would be present in AE (see Example 1.1), all values of QoIs would be known and there would be no uncertainty in the description of the SoN as well since QoIs define the SoN.

Unobservable quantities will be studied in AE state. The uncertainty about UQs will be studied in light of the observed values of OQs. Since OQs are revealed in the experiment, the beliefs about their true values BE are irrelevant.

Although the terms before-experiment and after-experiment suggest that a time series is analyzed, it is not how these two terms should be interpreted. Simply, the term before-experiment indicates that the results of the experiments are unknown but the experiment will be performed and they will become known with certainty. Therefore, in BE state the OQs are uncertain, but their true values are assumed constant. This may sound a bit paradoxical

because if we think in terms of time, the experiment has not been performed yet. It is therefore easier to associate the BE state with a stage at which the actual experiment is already performed but simply the results are not yet revealed. This view of the BEAE concept removes the logical paradox. The BEAE concept does not address prediction of future experiments, but rather indicates two stages of knowledge about QoIs considered.

Figure 1.2 presents a summary of all quantities introduced in this section. In this book it is assumed that quantities are numbers and in general can be vectors. All three types of quantities define the state of nature of the system that is being investigated. The true value of observable quantities are revealed in the experiment and the true values of KQ are assumed known in all stages of analysis. The KQs are not part of quantities of interest (QoIs) because the full knowledge about them is available in every state of experimentation. The true values of unobservable quantities are uncertain and the main goal of the statistical analysis is to use available information about the true values of other quantities to reduce uncertainly about the UQs.

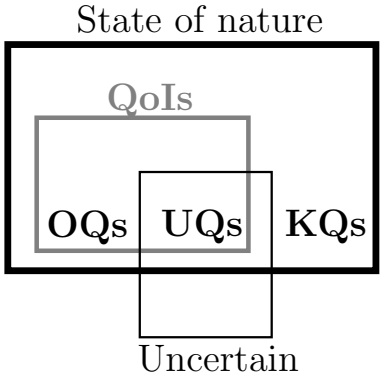


FIGURE 1.2 Observable, unobservable, and known quantities (OQs, UQs, KQs) are identified. UQs are uncertain at the AE stage. Both OQs and UQs are QoIs.

1.3 DEFINITION OF PROBABILITY

At first the probability will be introduced for a single QoI. The term "probability" is used to indicate the strength of our beliefs that some particular value of QoI is true. This definition is very close to the standard use of this term in everyday life. Therefore if we think that some value of QoI is true with a high probability, it indicates a high confidence in this proposition and conversely if the probability that some value of QoI is true is low, our confidence is low.

Suppose we are considering a single QoI in BE state and that the number of possible true values of the QoI is very small. For this case the SoN is defined

by a single QoI and therefore probability of the QoI being true is equivalent to the probability of the SoN. The belief that QoI is true is equivalent to the belief that SoN defined by this QoI is true. For example, in a coin toss only two true values are possible: heads and tails. Before the experiment (coin toss) the measure of belief is assigned to each of those values that describes our belief that the result of the toss (true value of QoI) is heads or tails. We will use the term *probability* to refer to those beliefs.

At this point we are ready to define a measure of our beliefs that a value of some QoI is true with more mathematical formality. We assume that for a particular QoI we know all possible true values that the QoI can have. In this book we will use small letters to denote the values of QoIs such as g, f, or y, and corresponding capital letters to denote the set of all possible true values of that QoIs such G, F or Y, respectively. We assume for a moment that the QoIs are scalars and extend the theory to vector quantities in Section 1.8.

As mentioned above, a measure of our beliefs about a given value of a QoI being true is the *probability*. For a given G, any possible QoI $g \in G$ has an assigned measure $p(g)$ (the probability) which is a scalar value from the range 0 to 1. The value of probability 0 about the value of g indicates that the proposition that the g is the true value is impossible and the value of 1 indicates that the true value of QoI is known with certainty and equal to g (this will always be the case for OQs in the AE state).

The following properties of the probability measure are postulated and defined below. Non-trivial extension of the properties of the probability measures to more QoIs will be given in Section 1.4.

1. We postulate that for QoI g where $g \in G$, the probability that the value g is true is described by a number $p(g)$ where $p(g) \geq 0$.
2. We require that the sum over all possible true values of QoI of the probability measure is equal to 1. Therefore $\int_{g \in G} p(g) = 1$.
3. We require that the probability of QoI is either g_1 or g_2 is the sum of probability measures for g_1 and g_2. Mathematically this is denoted by $p(g_1 \cup g_2) = p(g_1) + p(g_2)$.

Example 1.3: Single-die roll (2)

The example of the roll of a die is re-used. Before the experiment (which is the actual roll of a die), assuming the die is fair, we believe that each number that will be revealed in the experiment is equality probable. For this example the true state of nature corresponds to the number that occurs in a die roll. Therefore, denoting by g the QoI indicating the number obtained, the probability a given number is rolled $p(g)$ is $1/6$.

The a priori knowledge or belief is the information about QoI that is available before the experiment is performed. This knowledge is summarized by

assigning for each possible value of QoI a measure of belief that this value is the true value. In the example above, the a priori knowledge that a given number would be rolled was 1/6.

All QoIs can have a prior knowledge assigned to them, but it is meaningful to specify prior knowledge only for QoIs that are unobservable. The reason for this is that a priori knowledge becomes irrelevant for OQs that are revealed in the experiment. Since during the course of planning of the experiment which QoIs will be observed is known, the a priori knowledge for OQ will not be considered.

Note how irrelevant is the fact that we assigned probability 1/6 to every possible number that can be obtained in a roll once we actually know the number that occurred. Since we know this number in AE condition, all consideration about probability of this number in BE are irrelevant[1].

As introduced in this chapter, two stages of the experiment are identified, BE and AE, and the knowledge (probability, beliefs) about the true value of QoIs may change when considering QoIs in BE and AE states. By convention we will refer to these probabilities as *prior* and *posterior* probabilities at before- and after-experiment stage. We will adhere to this convention throughout this book.

1.3.1 COUNTABLE AND UNCOUNTABLE QUANTITIES

By definition, a *countable QoI* is such that all possible true values of this quantity can be enumerated by integers. Conversely, if the possible true values of a QoI cannot be enumerated by integers, they are labeled as *uncountable QoI*.

To illustrate these countable and uncountable QoIs, consider the two following examples:

Example 1.4: Roll of die and random number generator

The simplest example of a countable QoI is a result of six-sided die roll. In the BE condition the possible values of the QoI are 1, 2, 3, 4, 5, or 6. Therefore the six possible ways can be trivially enumerated by integers 1 through 6 and therefore it is a countable QoI. Another example is an experiment in which a number from range of $[0, 1]$ is selected. For this case in the BE conditions there are an infinite number of values (real number from $[0, 1]$) that cannot be enumerated by integers and therefore the quantity is uncountable. The generation of random number is a common task in computing when using Monte Carlo methods. However, one needs to take into account that real numbers are represented using binary system with limited precision. If double precision is used for

[1] In fact the actual prior of OQs can be used to test some assumptions made about the model of the experiment, but this application of the prior of OQ is not discussed in this book. For more on this topic see Berger [8] and Robert [84].

example (64 bits per number) only 2^{64} numbers can be represented. Therefore, when using computers and double precision statistics we actually use countable quantities. For most applications (including medical imaging), this limited precision in representing real number can be ignored, but one needs to be mindful of the limitation of digital machines in representing the uncountable values.

To simplify the notation the symbol $\int_{g\in G}$ is used for both (1) the summation for countable QoI and (2) the integral for uncountable QoI. Which of these two (whether g is countable or uncountable) applies will hopefully be clear from the context. When not obvious it will be specified explicitly if the symbol $\int_{g\in G}$ is a summation or an integral.

We will use the symbol p to indicate the probability or probability density for countable and uncountable QoIs. However, we will use sometimes the term probability for both countable and uncountable QoIs and based on the type of QoI it will be clear probability or probability density is referred to. If the term probability density is used, it will always imply the uncountable QoI.

By $p(g)$ the distribution is indicated, where g can be any of the possible values from the set G. In AE condition some QoIs are observed and at this point their probability distribution is trivial. Only a single value QoI that was observed has a posterior distribution that is non-zero and for all other gs the posterior is zero. The posterior distribution has to obey the normalization condition; therefore, for OQ in AE state, the non-zero posterior for countable and uncountable QoIs is either 1 or the Dirac delta function[2]. Without losing generality, the G will be used to indicate QoIs that are observable (their true value is revealed in the experiment). The following example is used to illustrate the definitions introduced in this section:

Example 1.5: Radioactive sample (2)

Let's revisit the counting experiment (Example 1.2) in which the number of radioactive decays is measured by a perfect sensitivity detector (all radioactive decays are registered). In BE state we have two QoIs f and g where f is a uncountable QoI indicating the amount of radioactivity in the sample. We unrealistically assume that this amount of activity does not change over time and therefore it is assumed constant in time and reflects the average number of emissions per unit time. G is countable QoI and represents the number of decays that will be measured during the experiment. All possible true values of both QoIs are known. The values of f are from a range $[0, \infty]$ and the number of detected radioactive decays g can take integer values $0, 1, \ldots, \infty$. In BE state, we express our beliefs about the true values of the QoIs by the specification of $p(f)$ and $p(g)$ for every possible $f \in F$ and $g \in G$. After the experiment is

[2]The Dirac delta function $\delta_d(x)$ is defined such that $\int_{f\in F} \delta(f) = 1$, respectively.

performed and g is measured (observed, realized), the posterior of g is trivial as it is zero for all other than observed number of counts and 1 for the observed number of counts. The prior probability of UQ f, $p(f)$, after the experiment is "updated" to the posterior probability. Interestingly, if we consider another experiment that follows, the posterior from the previous experiment becomes the prior for the new experiment, and is updated again by the data. This type of analysis is called the *sequential analysis* and plays important role in many applications. For more details on sequential analysis refer to Berger [8].

1.4 JOINT AND CONDITIONAL PROBABILITIES

In the preceding sections we considered SoNs that were defined by a single QoI. There, $p(g)$ was the probability that the SoN defined by g was true and similarly $p(f)$ was the probability that SoN defined by f was true. These two different SoNs were considered independently.

Here, we assume that there is only a single SoN defined jointly by f and g. By virtue of this assumption we generalize the probability of such SoN as $p(f, g)$. The comma signifies that we consider a SoN which is defined by particular values f and g. If the SoN is defined by more than one QoI, the probability will be referred to as *joint probability distribution*. For each pair of $\{f, g\}$ that define a possible SoN the probability is assigned. We note the symmetry in the definition. The identical SoN is described by a pair $\{f, g\}$ and by $\{g, f\}$ as there is no significance in the order that we specify the QoIs. This symmetry implies that $p(f, g) = p(g, f)$, so the order of the symbols in notation of joint probabilities is irrelevant.

The axioms that were specified for probabilistic description of SoN described by a single QoI (Section 1.3) apply the same for the SoN described by two (or more as it will be shown in Section 1.8) and therefore:

1. We postulate the probability that the SoN defined by g and f is true is described by a number $p(f, g)$ where $p(f, g) \geq 0$.
2. We require that the sum over the probability of true SoNs (the probability $p(f, g)$) is equal to 1. Therefore $\int_{g \in G} \int_{f \in F} p(f, g) = 1$.
3. We require that the probability of the SoN defined by $\{g_1, f_1\}$ or $\{g_2, f_2\}$ is the sum of probabilities for those two SoNs. Mathematically this is denoted by $p(\{g_1, f_1\} \cup \{g_2, f_2\}) = p(g_1, f_1) + p(g_2, f_2)$.

As defined in the beginning of this section, the true SoN is defined by f and g. If one of those quantities becomes known by obtaining the experimental data g, the uncertainty about the true SoN is manifested only through uncertainty in f. In other words, the probability distribution reflecting our beliefs about the true SoN is the function of only f as the other QoI is known. We indicate this "partial" knowledge of the SoN through the *conditional distribution* $p(f|g)$. We define this distribution in the BE state and therefore only "pretend" that g is known. The conditional distribution can be obtained

from the joint distribution simply by extracting values of the joint distribution corresponding to known QoIs and normalizing them by $\int_{f\in F} p(f,g)$. This process is illustrated with Example 1.6.

Example 1.6: Conditional distribution from joint distribution

The concept of the joint probability distribution is illustrated in Figure 1.3. For clarity, we assume that f and g are one-dimensional QoIs and for each pair $\{g,f\}$ the probability is assigned. We first define all possible true values of f and g which is the region $[0,1]$. An analytical function $p(f,g) = 144(f-0.5)^2 \times (g-0.5)^2$ is chosen to represent the joint distribution and plotted in Fig. 1.3(A). We consider a line on 2D plot corresponding to a value $g = 0.3$ (we "simulate" that g is known) and plot values of joint distribution in Figure 1.3(B). Therefore, we "simulate" an experiment in which value 0.3 of QoI g was observed. The analytical form of this distribution is $p(f,g=0.3) = 144/25(f-0.5)^2$. Normalization of values of $p(f,g=0.3)$ by the normalization constant $\int_{f\in[0:1]} p(f,g=0.3)$ leads to the *conditional probability* which is denoted as $p(f|g=0.3)$. This notation indicates a conditional probability distribution of QoI f if hypothetically the true value of g is 0.3. The actual shape of the conditional distribution is identical to the joint distribution evaluated at $g = 0.3$ and they differ only by a scaling factor. Although the latter finding is demonstrated on a simple example, it is true in general.

It is easy to demonstrate that all three axioms are obeyed for $p(g|f)$ if they are obeyed for $p(f,g)$. The normalization factor ($\int_{f\in F} p(f,g)$) that was used to obtain the conditional probability we denote as $p(g)$. In fact the normalization factor can be interpreted as a function of g which can be shown to also obey the axioms. The $p(g)$ is the *marginalized* probability distribution obtained from the joint $p(f,g)$ by *marginalization* (integrating out) the other QoI that the joint probability is dependent on (in our example it is f). Sometimes, if the joint distribution has a closed form, the marginalization can be performed analytically (see Examples 1.6 and 1.7). The notation is a little unambiguous as the same symbols were used to describe distribution of g when SoN was described by a single QoI and here where $p(g)$ indicates the marginalized distribution. However, based on the context of whether the SoN is defined by a single QoI or by multiple QoIs would unambiguously indicate the type of distribution that $p(g)$ represents. It follows that if the SoN is defined just by a single QoI the $p(g)$ is the distribution of QoI g in BE condition. If more than one QoIs describe the SoN, the notation $p(g)$ always indicates marginalized distribution.

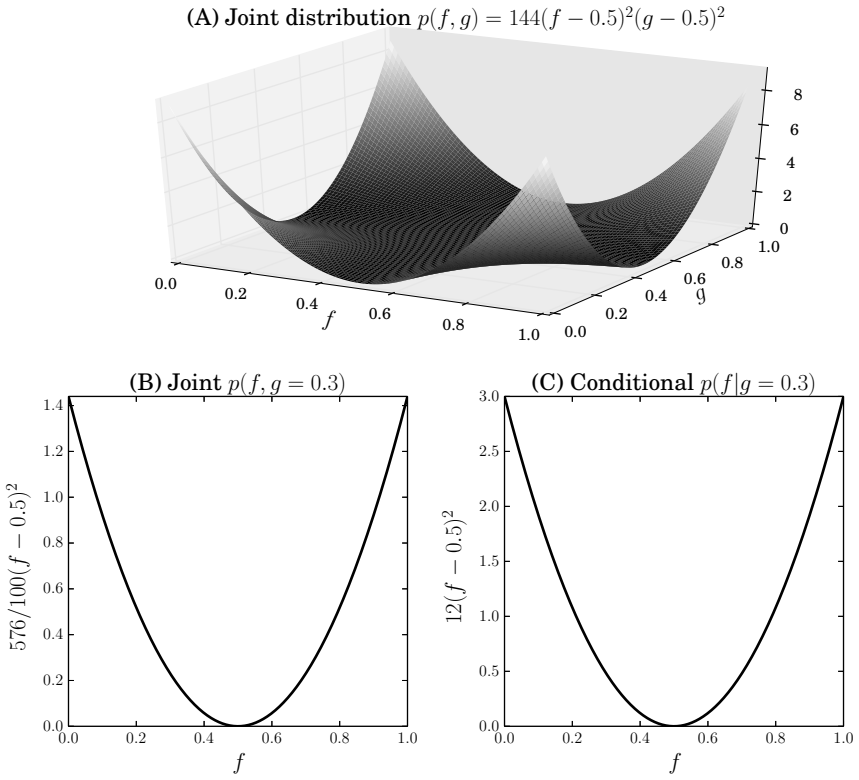


FIGURE 1.3 (A) 2D joint distribution defined for $\{f, g\}$. (B) Values of the joint distribution for QoI $g = 0.3$. (C) Conditional distribution of f conditioned on $g = 0.3$. The shapes of both distributions in (B) and (C) are identical and they differ only by a scaling factor.

Example 1.7: Analytic marginalization

Suppose a joint probability distribution $p(f, g)$ is considered that expresses our beliefs that f and g define the true SoN:

$$p(f,g) = \frac{6}{\pi^2}\frac{1}{g^2}\frac{f^g e^{-f}}{g!} \tag{1.1}$$

The range of possible values of uncountable QoI f and countable QoI g is $F : f \in [0, \infty]$ and $G : g \in [1, 2, ..., \infty]$. We have that:

$$p(g) = \int_{f\in F} p(f,g) = \frac{6}{\pi^2 g^2 g!}\int_0^\infty df\, f^g e^{-f} = \frac{6}{\pi^2 g^2} \tag{1.2}$$

Unfortunately, situations where the analytic marginalization is available is extremely rare in practice and numerical methods are used in order to obtain the numerical approximation of marginal distributions. For example for the case considered here, the marginalization over g yields a sum that cannot be evaluated in a closed-form expression as the sum has no simple mathematical form:

$$p(f) = \int_{g \in G} p(f, g) = \frac{6e^{-f}}{\pi^2} \sum_{g=1}^{\infty} \frac{f^g}{g^2 g!} \tag{1.3}$$

It is left for the reader to check that $p(f, g)$ and $p(g)$ are proper probability distributions and they integrate to 1 over the range of all possible values of f and g.

1.5 STATISTICAL MODEL

There are two motivations that lead to the introduction of conditional distributions. First, once the experiment is performed and OQ are known, there is no point of considering the joint distribution, but rather we "extract" the conditional distribution from the joint, which corresponds to observations, and make inferences based on that. The other reason for conditional distribution is that based on the knowledge of the experiment, we can propose a statistical *model* of the experiment $\mathcal{M}$ by means of conditional distributions.

Before we can define the model, the statistical independence of QoIs needs to be introduced. We define the f *statistically independent* of g when beliefs about f are insensitive to knowledge of true value of g. It follows that statistical independence of f and g implies statistical independence of g and f. Before the experiment, we pretend that g is known and therefore the independence applies to any possible value of f and g.

$$p(f|g) = p(f). \tag{1.4}$$

In other words, our beliefs about f are independent on knowledge of true value of g. The QoIs f and g are *statistically dependent* if $p(f|g) \neq p(f)$. We leave it for the reader to show that if f and g are statistically dependent/independent, then g and f are also statistically dependent/independent. We consider the dependence/independence of f and g in the BE state and this implies that the property is true for any value of QoIs g and f.

We use Example 1.8 to illustrate the concept.

Example 1.8: Conditional distribution from joint distribution—independence

Let's consider an analytical function describing the joint distribution $p(f, g) = 144(f - 0.5)^2(g - 0.5)^2$. This distribution is in fact an example of a joint distribution of two statistically independent QoIs. This can be demonstrated using

simple algebra and by showing that the particular $p(f, g)$ considered in this example implies that Equation (1.4) holds (this is left to the reader). An alternative approach to showing the independence is a graphical demonstration in Figure 1.4 showing that all conditional distributions extracted from the joint distribution are equal.

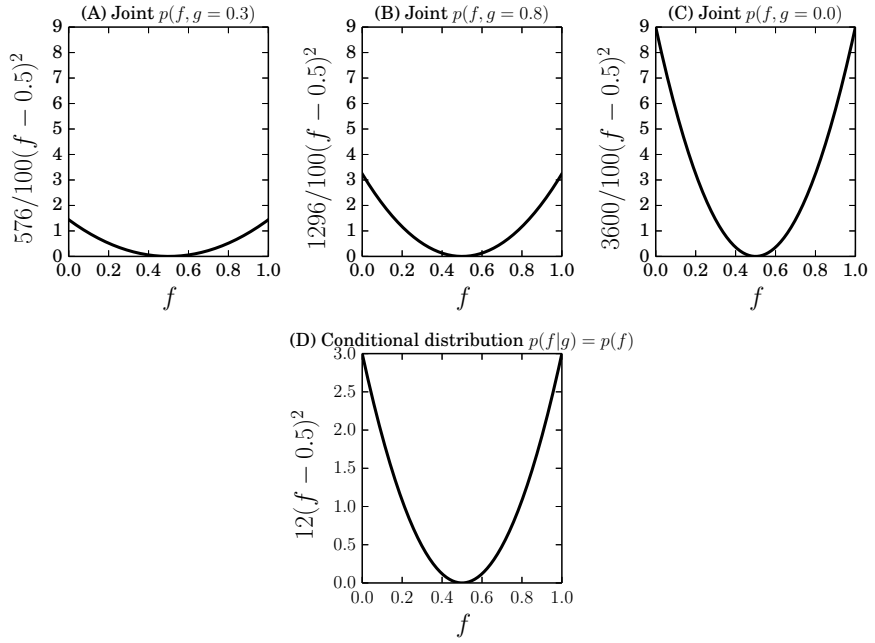


FIGURE 1.4 Three distributions extracted from the joint distribution $p(f, g) = 144(f - 0.5)^2(g - 0.5)^2$ for different values of g equal to (A) 0.3, (B) 0.8, and (C) 0.0. After normalization the same conditional distribution is obtained (D).

So far we used a slightly artificial example of an analytic function representing the joint distribution of two QoIs that are represented by scalar values. To make the concept of statistical dependence/independence more intuitive, a simple experimental model in which two dice are used is considered next. This will also help with the introduction of the concept of the statistical dependence and how it can be used to obtain more information about UQs based on OQs.

Example 1.9: Statistically independent QoIs—dice

Consider rolls of two "fair" dice. The result on only one of the die (say die number two) is revealed to the observer in AE stage. The true values of the numbers obtained in the experiment (the rolls of two dice) are denoted by f and g, respectively, for dice one and two. The possible true values of the QoIs for each roll is 1, 2, 3, 4, 5, or 6. With no other information about the experiment (the model) the result of roll two g (revealed to the observer) does not affect beliefs of the result of the roll one f and vice versa. Therefore, it can be stated that experiment in which the true value g is revealed does change our beliefs about f. This is another way of saying that f and g are statistically independent ($p(f|g) = p(f)$).

It should be quite clear from Example 1.9 that if the QoIs are statistically independent, then in AE the OQ (g) is known and knowledge about UQ (f) is unchanged. In fact, for any statistical problem in which there are some OQs and UQs, the independence would preclude gaining any additional knowledge (reduce uncertainty) about any of the UQs when some statistically independent OQs are observed.

The statistical dependence between observable and unobservable quantities is one of the most important concepts in statistical inference used in this book and it is introduced through the definition of the statistical *model* of the experiment. The model is simply the specification of conditional dependence of OQs and UQs. The model is defined as the conditional distribution where the OQs are conditioned on UQs. Adhering to our convention that an OQ is denoted as g and an UQ denoted as f the model is denoted by $p(g|f)$. If the model is defined, it will imply that f and g are statistically dependent ($p(g|f) \neq p(g)$) otherwise the model would be quite useless for statistical inference (see Example 1.9). The knowledge of the model will be derived from the known physical description of the experiment. When formulating a model, various considerations have to be taken into account. Typically the reality is much more complicated than what can actually be modeled with experiments and the assumed models will simplify the reality in order to be tractable. A trade-off between model complexity and tractability has to be considered in almost every problem.

To better understand the concept of the model, let's consider the extension of the problem with two dice:

Example 1.10: Conditional two-dice roll (1)

Let's consider an experiment in which we have two people. Person A rolls two dice, one after the other. The f describes the result of the first roll and g describes the result of the second roll. Person B ("The observer") observes

the result of only the second roll g (since g by convention is used to indicate observable quantity). The value of roll g is the only observable quantity. The result of roll f is unknown to person B. If no other conditions of the experiment are specified (see Example 1.9), the numbers obtained in each of the two rolls are independent, and we will not be able to infer any additional information about f based on g.

However, the experiment is modified such that the person A repeats the second roll g until the number is equal or less than the number obtained in the first roll f and only the result of the last roll is revealed to the observer. By this modification, the statistical dependence is introduced. Based on the description of the experiment, the model of the experiment can be defined. Intuitively, the statistical dependence is obvious since the number on the second roll g will be dependent of the number obtained in the first roll f.

In real systems person A embodies the laws of physics or laws of nature. In this example, the unknown number obtained in the first roll we interpreted as the UQ and the rules governing the process of rolling the second die until the number is equal or less than the number in the first roll are interpreted as the "law of nature."

Just to signal types of problems that will be discussed in this book, the typical question that will be asked is as follows: Having observed the number in the second roll (which is OQ), what can be said about an unobservable quantity of the number obtained in the first roll?

In this book the laws of nature are always described by conditional probabilities $p(g|f)$ (the model) based on the description of the experiment, knowledge of physical principles governing the experimental processes, and logic. For this particular example based on provided description of rules of the experiment, the model $p(g|f)$ is defined in Table 1.1.

TABLE 1.1
The model: values of conditional probability of OQ g if UQ has value f

		$p(g\|f)$					
	$f \rightarrow$	**1**	**2**	**3**	**4**	**5**	**6**
$g \downarrow$	**1**	1	1/2	1/3	1/4	1/5	1/6
	2	0	1/2	1/3	1/4	1/5	1/6
	3	0	0	1/3	1/4	1/5	1/6
	4	0	0	0	1/4	1/5	1/6
	5	0	0	0	0	1/5	1/6
	6	0	0	0	0	0	1/6

1.6 LIKELIHOOD

So far several quantities such as joint and conditional distributions were discussed and considered mostly in BE conditions. The model of the experiment summarized by conditional probability $p(g|f)$ where g is the OQ and f is UQ was also introduced. We now move to the AE regime, and in this section we introduce the most important distribution that will be used throughout this book[3], namely, the *likelihood function*. There are two conditions needed to determine the likelihood function: (1) The model has to be known, (2) the data (some OQs used in model definition) need to be observed. Paraphrasing those two conditions, the likelihood function is the "model" ($p(g|f)$) in the AE condition.

The likelihood function (LF) exists only in the AE conditions and exists only when at least one of the QoIs is observed and there is a statistical dependence between the OQs and UQs.

We denote the LF as $p(G = g|f)$ and interchangeably refer to this quantity as the *likelihood function* (LF), the *data likelihood*, or simply as the *likelihood*. The fact that g is observed (AE) is indicated by using the notation $G = g$. The LF for an observed value of g assigns a value of likelihood to every possible $f \in F$ that indicates how likely it is to observe the result $G = g$ if the value of the UQ is f. The LF is therefore a function of the UQ f. The difference between the model $p(g|f)$ and $p(G = g|f)$ is that the first is the distribution defined in the BE and is a function of both g and f, where the latter is a function of f for observed data $G = g$ and defined in AE the condition.

Example 1.11: Conditional two-dice roll (2)—likelihood function

Using the example of the model summarized in Table 1.1 the example of two likelihood functions $p(g = 1|f)$ and $p(g = 5|f)$, for different values of OQ g are shown in Figure 1.5.

The likelihood function is not a probability distribution and therefore we cannot interpret the value of the likelihood function as a measure of probability (belief) of f because in general the likelihood function is not normalized i.e., $\sum_{f \in Y} p(G = g|f)$ is not guaranteed to be 1.

Example 1.12: Conditional two-dice roll (3)

Suppose we now consider the model summarized in Table 1.1 in AE condition

[3]The likelihood function is the most important distribution not only here in this book but also in classical statistics [26, 66, 67, 106].

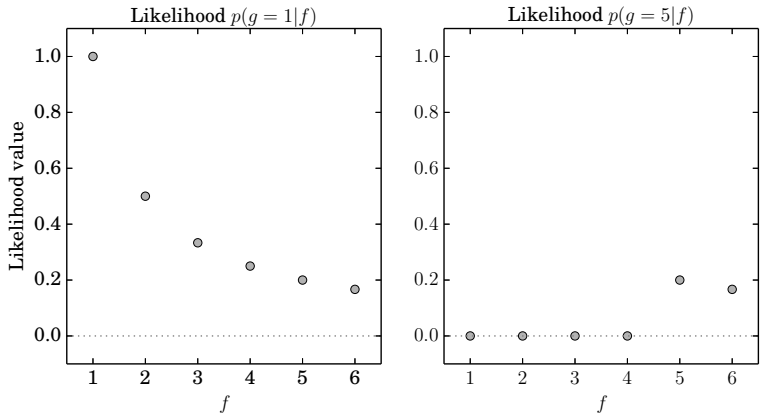


FIGURE 1.5 The likelihood function for the model described in Example 1.10 corresponding to $g = 1$ (left) and $g = 5$ (right). The values in the figure correspond to rows 1 and 5 in Table 1.1.

when $g = 3$ is observed. From the Table 1.1 we extract a row that corresponds to $g = 3$ since these are the only relevant values in AE state. The row has the following values 0, 0, $1/3$, $1/4$, $1/5$, and $1/6$. Because these are the values in the AE condition, from the definition they define the likelihood function $p(G = g|f)$. We immediately see that the sum over all possible f does not amount to 1 as the likelihood function is not a probability distribution. Based of the values of the likelihood a likelihood ratio of the form $\Lambda(g; f_1, f_2)$ can be constructed. For example the likelihood ratio $\Lambda(3; 4, 6)/ = 1.5$ indicates that if $g = 3$ is observed it indicates that in the first roll f number 4 is 1.5 times more likely than number 6. We are careful not to use word ″more probable″ as the likelihood is not a measure of probability. One way to interpret the likelihood is to think about it as a "measure of probability" that comes only from the data. In the next chapter we discuss the likelihood ratio in more details.

We assume that for all problems considered in this book the functional form of the likelihood is known and derived from the model $p(g|f)$. It should be clear that if the model is known then the likelihood is known because from all the distributions $p(g|f)$ defined in the BE state, in the AE state we simply select the distribution corresponding to the observed data g. Interestingly, this function alone can be used to make inferences in classical statistics (e.g., maximum likelihood (ML) methods) but these inferences are quite limited and cannot be easily extended to decision theory. For this reason, the methods that are based solely on likelihood function are not utilized in this book. In general, methods based on data likelihood (e.g., maximum likelihood estimation) do not perform well for systems in which the data contain large amounts of noise and systems with a substantial null space which frequently is the case in

nuclear imaging and therefore the ML solution to imaging problems is never used unless some *regularization methods* are employed.

1.7 PRE-POSTERIOR AND POSTERIOR

Throughout this book all decisions about UQ f will be based on *posterior* distributions $p(f|G = g)$. This distribution is defined in AE condition as indicated by the fact that the OQ g is known ($G = g$). Earlier for the case of single QoI, we noted that our general approach to reasoning is that we start with some initial beliefs about a true value of f which we described by the prior distribution $p(f)$. After the experiment is performed and some OQ (g) is observed, the original beliefs are modified to reflect the additional information that is obtained by revealing the true value of the OQ g. Here we describe it mathematically by defining the posterior which is the original beliefs $p(f)$ updated by experiment outcome g. From the definition, the posterior is the probability distribution on f when g is true (observed).

The conditional distribution of UQ conditioned on OQ can also be defined in the BE state. For this case we simply assume that if g is observed it would result in the posterior $p(f|g)$. Because we only assume that g is observed we will refer to the $p(f|g)$ as the *pre-posterior*. There is only one posterior distribution $p(f|G = g)$ but there are many pre-posterior distributions and the number of pre-posteriors is the same as the number of possible true values of g. There are two ways to obtain the posterior distribution which we will discuss below.

1.7.1 REDUCTION OF PRE-POSTERIOR TO POSTERIOR

The most straightforward approach to obtain the posterior is when the pre-posteriors $p(f|g)$ are known. The experiment reveals the true value of g and from the family of pre-posteriors the one is selected that corresponds to g which was observed. The pre-posterior is "reduced" to posterior once the g is observed:

$$p(f|g) \xrightarrow{g \text{ is observed}} p(f|G = g). \tag{1.5}$$

1.7.2 POSTERIOR THROUGH BAYES THEOREM

In real-world scenarios the direct knowledge of the pre-posterior will not be available and the posterior will be obtained through the Bayes theorem. As explained in Section 1.4 if the joint distribution is known in BE, it is quite straightforward to determine the pre-posterior and once the pre-posterior is known finding posterior in AE state is trivial (see Section 1.7.1). Unfortunately the joint distribution is typically unknown. However, the joint distribution can be approximated using

$$p(f, g) = p(g|f)p(f) \tag{1.6}$$

where on the right-hand side we have a "model" $p(g|f)$ that we assumed is known and the $p(f)$ which is unknown in the large majority of cases. However, since $p(f)$ can be interpreted as the current knowledge about UQ f an educated guess about this distribution based on the current knowledge can be made and joint distribution can be approximated using Equation (1.6). We come back to the problem of the selection of $p(f)$ later in this section. Assuming that the joint probability is known (or approximated) obtaining the pre-posterior (or posterior) directly follows from Equation (1.6) and the fact that if the joint distribution is known, the posterior can be obtained by selecting values of the joint corresponding to the "hypothetically" observed g and then normalizing those values to 1. Using this approach the pre-posterior is readily calculated as follows:

$$p(f|g) = \frac{p(f,g)}{\int_{f' \in F} p(f',g)} = \frac{p(g|f)p(f)}{\int_{f' \in F} p(g|f')p(f')}. \tag{1.7}$$

Often in the literature the normalization term $\int_{f \in F} p(f,g)$ is abbreviated to $p(g)$. The posterior is derived from the pre-posterior replacing a value of g with the actual measurement $G = g$ and the following is obtained that we refer to as the *Bayes Theorem*.

$$p(f|G=g) = \frac{p(G=g|f)p(f)}{\int_{f' \in F} p(G=g|f')p(f')} \tag{1.8}$$

where $p(G = g|f)$ is the likelihood and $p(f)$ is the guess about the value UQ f. The $p(f)$ can also be interpreted as the current state of knowledge about the value of f. It is customary to refer to this distribution as the *prior* to indicate that it represents the knowledge about some QoI prior to the experiment.

The above equation is known as the *Bayes Theorem* named after Thomas Bayes who suggested using the above equation to update beliefs based on data. The theorem gained popularity after it was rediscovered by Pierre-Simon Laplace and published in 1812 [63].

1.7.3 PRIOR SELECTION

One of the most frequent critiques of the use of Bayes Theorem in science is that objectivity of the analysis using posterior requires the accurate knowledge of the prior which is almost never available. It follows that if the guess about $p(f)$ is inaccurate the resulting posterior may be inaccurate as well. We fully agree with these arguments; however, we would like the reader to consider the following three points:

Complete objectivity is an illusion. Drawing statistical inference without the use of Bayes Theorem relies on the model and the resulting likelihood after the experiment is performed. This, however, relies on the assumption that the model is correct. When complex biological processes are

involved this will never be the case and some level of approximation when constructing a model will have to be used. Therefore, there is some level of subjectivity used when a model is assumed. Therefore, so-called "objective" analysis frequently lacks objectivity as well because the model is never exact.

Subjectivity is unavoidable in decision making. We should not forget that the endpoint of any analysis in medical imaging or for that matter in medicine is not the posterior or likelihood but rather the decision. Assuming that statistical analysis is performed objectively the decision (disease present or not, optimal course of therapy, etc.) still has to be made and this decision will require a subjective judgment. The definition of prior should be interpreted as an attempt to quantify the use of subjective knowledge, and the use of subjective knowledge is unavoidable.

Bayesian analysis describes knowledge gain. The most importantly, the application of the Bayes Theorem should be interpreted as an update of the current beliefs (reflected by the prior $p(f)$) by the experimental data leading to the posterior $p(f|G = g)$. In this interpretation the correctness of the prior is quite irrelevant with respect to the correctness of the theory and logic of the approach. If the prior is inaccurate, it simply represents inaccurate beliefs that hopefully are improved by the results of the experiment. The correctness of the prior beliefs is not prerequisite for correctness of the analysis. While using wrong assumptions (the prior) may result in wrong conclusion (the posterior), there is no logical inconsistency in this process. The main idea of the analysis is that the beliefs are improved by the data and if the prior is "wrong" the posterior will be "less wrong."

There is a substantial research in the field investigating various approaches to prior selection. In this book we interpret the prior as the current (before experiment) beliefs about the UQs. The prior expresses the level of uncertainty about the true value of the UQ. However, we will also frequently use so-called *uninformative* prior which is attempt to express the belief that we know little about the true values of UQs. For example, one could assign the same prior probability to every possible value of the UQ (*flat prior*) naively believing that this assignment expresses the lack of knowledge. This approach is appealing from the point of view of ease of the implementation, but in fact it is incorrect and does not express our actual beliefs. The flat prior, in fact, implements prior beliefs that all possible true values of UQ have the same probability, which is a quite specific belief about the reality. By doing this we hope that information contained in the likelihood overwhelms incorrect beliefs in the prior. The benefits of the ease of use of flat prior have to be weighted against the possible inaccuracy in the posterior.

A interesting approach to the selection of objective prior (not requiring a subjective judgment) was introduced by Bernardo [10], Ye and Berger [113]. In this approach the selection of the prior is determined by the model in BE stage, and maximizes the expectation of the gain in information (information

defined in the Shannon sense [91]) that will be brought by the experiment. Therefore, the prior is used merely as a tool to obtain a posterior, and since the process is deterministic, the statistical analysis that leads to posteriors based on reference priors does not require any subjective judgments.

Another type of priors that we would like to mention is the *entropy priors* championed by Jaynes [47]. The idea is to maximize the entropy of the prior distribution in order to minimize the amount of information that the prior contains (we cover this topic in more detail in Section 6.1.2).

1.7.4 EXAMPLES

To help understand the distributions introduced in previous sections (joint distribution, likelihood, pre-posterior, and posterior), the following two examples briefly introduced in previous sections are used.

Example 1.13: Conditional two–dice roll (4)

The example of the conditional dice roll is used in which person A rolls the first die (f) and then rolls the second die g until number is equal or less than on the first. Only the final number obtained on the second die g is revealed for the observer. Therefore the true values of f and g (the true SoN) are known only to person A and the observer knows only the true value of g in the AE state.

First, we characterize the problem in BE condition. For both QoIs f (the value of first roll) and g (the value of second roll) there are six possible true values of f and g corresponding to the numbers 1 through 6. We assume that the dice are "fair" and therefore assign a priori probabilities for each value equal to $1/6$. This is to ensure that the distribution is normalized to 1. Interestingly, for this case the priors $p(f)$ and $p(g)$ that express the beliefs that the dice are fair are the same as the non-informative flat prior that assigns equal prior probability to every possible true value of the QoI. This is purely coincidental and only infrequently a flat prior will express our exact beliefs.

In order to determine the pre-posterior and posterior for this example, the joint distribution of $p(f, g)$ is determined first. This can be done using Equation (1.6). The statistical model is summarized in Table 1.1. Table 1.2 presents the calculated joint probability. The marginal values of the probability of obtaining g (Table 1.2) calculated as $\int_{f \in F} p(g|f)p(f)$ are presented. The joint distribution summed over all possible SoNs (pairs $\{f, g\}$) is equal to 1 which can be verified in Table 1.2:

Since the joint distribution is specified, the value of the pre-posteriors can be found by normalizing values in each row of Table 1.2 to 1. This is done by dividing values in each row by the corresponding value of $p(g)$ shown in the last column in Table 1.2. The result of this division is shown in Table 1.3.

Having determined the distribution of the pre-posterior, in AE (Table 1.3), the posterior can correspond to any row in the table. The selection of an appropriate row depends on the observed true number obtained in the second roll g. For

example, if $g = 2$ is observed, the probability of the value on the first roll f being 2 is three times larger than that of being 6.

TABLE 1.2
Value of the joint probability f and g

		$p(g,f)$						$p(g) = \sum_{f \in F} p(g,f)$
	$f \rightarrow$	**1**	**2**	**3**	**4**	**5**	**6**	
$g \downarrow$	**1**	1/6	1/12	1/18	1/24	1/30	1/36	147/360
	2	0	1/12	1/18	1/24	1/30	1/36	87/360
	3	0	0	1/18	1/24	1/30	1/36	57/360
	4	0	0	0	1/24	1/30	1/36	37/360
	5	0	0	0	0	1/30	1/36	22/360
	6	0	0	0	0	0	1/36	10/360

TABLE 1.3
Value of the pre-posterior of f conditioned on g where g is assumed to be observed in the experiment (dice rolls)

		$p(f\|g)$					
	$f \rightarrow$	**1**	**2**	**3**	**4**	**5**	**6**
$g \downarrow$	**1**	60/147	30/147	20/147	15/147	12/147	10/147
	2	0	30/87	20/87	15/87	12/87	10/87
	3	0	0	20/57	15/57	12/57	10/57
	4	0	0	0	15/37	12/37	10/37
	5	0	0	0	0	12/22	10/22
	6	0	0	0	0	0	1

1.7.5 DESIGNS OF EXPERIMENTS

Obtaining statistical inferences about UQ based on experimental data and prior beliefs will be done using the following. The experiments are considered as means of improvement of the information about the UQ. In BE state the knowledge about UQs is described by the prior and in AE state by the posterior. Often, the specification of the prior will be very difficult and a pragmatic

approach will be followed compromising between the accuracy in formulation of the prior with the ease of implementation. The goal will always be to obtain the pre-posterior or posterior and based on these distributions make decisions about a problem at hand. The following defines the experimental design that will be followed for every problem discussed in this book:

1. Define BE and AE conditions, identify QoIs that are unobservable (UQ) and QoI that we can measure (obtaining their true value) which are statistically dependent on UQs.
2. For each of QoIs specify the range of possible true values.
3. Based on the description of the experiment, formulate the model of the experiment defined by a conditional distribution.
4. For each of the UQs specify the initial beliefs (prior) in the form of the prior distribution.
5. For UQs determine the pre-posterior distribution based on the model and the prior using Bayes Theorem (Section 1.7). If decisions are based on the actual measurements of OQs determine the posterior.
6. Make a decision based on posterior or pre-posterior depending on a problem at hand.

In this chapter we cover steps 1 through 5, and in the next chapter we will describe approaches to decision making (point 6) based on the posteriors and pre-posteriors.

The steps 1 through 5 define the Bayesian analysis that will be used in all problems that are discussed in this book. The Bayesian analysis of any problem ends with providing the posterior or pre-posterior which contains the complete information about the UQs. This sometimes will not be sufficient in practical applications. For example, we doubt that providing physicians with a million-dimensional posterior distribution would be received enthusiastically. Therefore, the posterior will be summarized in one way or another to provide easily digestible information for the decision maker. For example, the image of the distribution of the radiotracer will be a much more reasonable result provided to physicians than the posterior distribution function. The formation of the image from the posterior is a decision problem, as a decision needs to be made about which possible true values of the UQs best represent the reality. The word "best" used in the previous sentence is undefined at this point and exact definition will be given in the next chapter discussing the decision making.

So far, we have introduced four key distributions pictured in Figure 1.6. Only two of those distributions will be used in this book for decision making, the pre-posterior and the posterior, covered in Chapter 2.

Before Experiment (BE)		After Experiment (AE)	
$p(g\|f)$	Model	$p(G=g\|f)$	Likelihood
$p(f\|g)$	Pre-posterior	$p(f\|G=g)$	Posterior

FIGURE 1.6 Distributions introduced in Chapter 1. It is assumed that OQ is g and UQ is f.

1.8 EXTENSION TO MULTI–DIMENSIONS

Up to this point, either a single QoI or a pair of single-dimensional QoIs f and g were considered. In real world applications many more QoIs will be used to characterize a problem. The case of multi-dimensional QoIs is a straightforward extension of the presented theory. To simplify the notation and make it clearer in most cases we will adhere to the convention that all QoIs are divided into two groups: UQs and OQs. We will use vector notation indicating UQs as $\mathbf{f}$ and OQs as $\mathbf{g}$ where I and K indicate the number of elements in those vectors, respectively. Bold non-italic small-letter font will always indicate vectors.

The probability distribution $p(\mathbf{f})$ is defined as the joint probability distribution of components of the vector $\mathbf{f}$:

$$p(\mathbf{f}) = p(f_1, f_2, \ldots, f_I) \tag{1.9}$$

Similarly, conditional probabilities $p(\mathbf{f}|\mathbf{g})$ are defined as the joint probability of elements of $\mathbf{f}$ conditioned on the joint probability of elements of vector $\mathbf{g}$ as:

$$p(\mathbf{f}|\mathbf{g}) = p(f_1, f_2, \ldots, f_I|g_1, g_2, \ldots, g_K) \tag{1.10}$$

Sometimes the notation will be used when more than two symbols will be used to indicate the distributions. For example $p(\mathbf{f}, \mathbf{g}, \mathbf{y})$ where $\mathbf{y}$ is a vector with J QoIs is a joint probability distribution of all elements of vectors $\mathbf{f}$, $\mathbf{g}$, $\mathbf{y}$:

$$p(\mathbf{f}, \mathbf{g}, \mathbf{y}) = p(f_1, f_2, \ldots, f_I, g_1, g_2, \ldots, g_K, y_1, y_2, \ldots, y_J) \tag{1.11}$$

All considerations from previous sections about two scalar QoIs are easily transferable to more than two vector QoIs.

In the following two sections we present rules that will allow transformations of the joint and conditional probabilities of multi-dimensional probability distributions of the QoIs.

1.8.1 CHAIN RULE AND MARGINALIZATION

The chain rule allows expressing the joint probability (e.g., probability distribution of vector QoI) and is the generalization of Equation (1.6),

$$p(\underbrace{f_1}, \underbrace{f_2, f_3, \ldots, f_I}) = p(f_1|f_2, f_3, \ldots, f_I)p(f_2, f_3, \ldots, f_I) \tag{1.12}$$

where underbraces indicate two probability distributions: probability distribution of f_1 and joint probability distribution $f_2, \ldots, f_I$. Applying the above $I-1$ additional times the original joint distribution $p(f_1, \ldots, f_I)$ can be expressed as a product:

$$p(f_1, f_2, f_3, \ldots, f_I) = p(f_1|f_2, f_3, \ldots, f_I)p(f_2|f_3, \ldots, f_I)\ldots p(f_{I-1}|f_I)p(f_I) \tag{1.13}$$

To marginalize multi-dimensional distributions we apply similar rules as in the two one-dimensional distributions a shown in Example 1.13 for two dimensions.

$$p(\mathbf{f}) = \int_{\mathbf{g}\in\mathbf{G}} p(\mathbf{f}, \mathbf{g}) \tag{1.14}$$

where by $\mathbf{G}$ we indicate all possible values of QoIs $\mathbf{g}$. If we are interested in the marginalized density of a single QoI which is an element of $\mathbf{f}$ the following applies

$$p(f_1) = \int_{\mathbf{g}\in\mathbf{G}} \int_{f_2,\ldots,f_I \in F_2 \ldots F_I} p(\mathbf{f}, \mathbf{g}). \tag{1.15}$$

Example 1.14: Application of chain rule, marginalization, and Bayes Theorem

Suppose we want to express the conditional probability distribution of a single element of vector $\mathbf{f}$ conditioned on a single element of $\mathbf{g}$ given conditional $p(\mathbf{g}|\mathbf{f})$. The chain rule, marginalization, and Bayes theorem are sufficient for this task:

It is obtained by

$$p(f_1|g_1) = \frac{1}{p(g_1)} \int_{f_2,\ldots,f_I \in F_2,\ldots,F_I} \int_{g_2,\ldots,g_K \in G_2,\ldots,G_K} p(\mathbf{g}|\mathbf{f})p(\mathbf{f}) \tag{1.16}$$

The distributions $p(\mathbf{f})$ and $p(g_1)$ were required in order to accomplish the task. Distribution $p(g_1)$ can be obtained from $p(\mathbf{g})$ through marginalization.

In order to express the Bayes Theorem and relations between the model and pre-posterior (or likelihood and posterior) using multi-dimensional QoIs,

the straightforward generalization of two one-dimensional QoIs is used, and the chain rule described in the previous section,

$$p(\mathbf{f}|\mathbf{g}) = \frac{p(\mathbf{g}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{g})} \tag{1.17}$$

1.8.2 NUISANCE QUANTITIES

Sometimes when considering many UQs, some of them will not be of direct interest. Therefore their value will influence the posterior and make it a higher dimensional than if the posterior is dependent only on UQs that are of direct interest. Those quantities that are not of direct interest will be referred to as *nuisance QoIs*. Since nuisance QoIs are always UQs they will be denoted as NUQs (N+UQs). In the formulation of the statistical analysis used in this book, nuisance quantities are handled with relative ease. It is done by first determining the pre-posterior or posterior distributions using methodology provided in this chapter and then by marginalizing the NUQs. This can be summarized in the following equations and illustrated by Example 1.15.

Suppose that the vector of NUQs is indicated by $\tilde{\mathbf{f}}$ and the posterior of all unobservable QoIs is indicated by $p(\mathbf{f}, \tilde{\mathbf{f}}|\mathbf{G} = \mathbf{g})$, then the posterior of UQ is obtained by marginalization:

$$p(\mathbf{f}|\mathbf{G} = \mathbf{g}) = \int_{\tilde{\mathbf{f}} \in \tilde{\mathbf{F}}} p(\mathbf{f}, \tilde{\mathbf{f}}|\mathbf{G} = \mathbf{g}) \tag{1.18}$$

Similar marginalization of NUQs can be used to obtain pre-posterior of the UQ.

Example 1.15: Three scalar QoIs—marginalization of nuisance QoIs

Suppose we consider a simple model of the experiment with three scalar QoI: the OQ g, the UQ f, and the NUQ $\tilde{f}$. Each of the QoIs has only two possible true values of either 0 or 1. The model and the prior are defined in Table 1.4

From Table 1.4 we can formulate the initial beliefs about f and $\tilde{f}$ by marginalizing the prior $p(f, \tilde{f})$ and obtain $p(f = 0) = 0.80$ and $p(f = 1) = 0.20$. Similarly the marginalized prior of NUQ $p(\tilde{f} = 0) = 0.40$ and $p(\tilde{f} = 1) = 0.60$. We also note that the joint prior $p(f, \tilde{f})$ as defined above indicates that f and $\tilde{f}$ are statistically dependent as the $p(f, \tilde{f}) \neq p(f)p(\tilde{f})$ which indicates statistical dependence based on definition in Equation (1.4) because

$$p(f|\tilde{f}) = \frac{p(f, \tilde{f})}{p(\tilde{f})} \neq \frac{p(f)p(\tilde{f})}{p(\tilde{f})} = p(f) \tag{1.19}$$

and therefore $p(f|\tilde{f}) \neq p(f)$.

TABLE 1.4

Definition of the model $p(g|f,\tilde{f})$, the prior $p(f,\tilde{f})$, the pre-posterior $p(f,\tilde{f}|g)$, and the pre-posterior with marginalized NUQ

$p(g|f,\tilde{f})$ **[model]**

	$\{f,\tilde{f}\}\rightarrow$	**0,0**	**0,1**	**1,0**	**1,1**
$g\downarrow$	**0**	0.00	0.10	0.30	0.50
	1	1.00	0.90	0.70	0.50

$p(f,\tilde{f})$ **[prior]**

$\{f,\tilde{f}\}\rightarrow$	**0,0**	**0,1**	**1,0**	**1,1**
	0.30	0.50	0.10	0.10

$p(f,\tilde{f}|g)$ **[pre-posterior]**

	$\{f,\tilde{f}\}\rightarrow$	**0,0**	**0,1**	**1,0**	**1,1**
$g\downarrow$	**0**	0.00	0.38	0.23	0.38
	1	0.34	0.52	0.08	0.06

$p(f|g)^a$ **[marginalized pre-posterior]**

	$f\rightarrow$	**0**	**1**
$g\downarrow$	**0**	0.38	0.62
	1	0.86	0.14

Note: The values are given with the precision of two decimal places.
[a] The values of $p(f|g)$ are obtained from $p(f,\tilde{f}|g)$ by adding the first two and the last two columns. For this simple example this addition corresponds to marginalization.

Example 1.16: Wire-cube of joint probabilities

In this example the data from the previous Example 1.15 is re-used. The idea is to demonstrate using an illustrative model of wire-cube of joint probabilities (Figure 1.7(A)) that the joint-distribution contains all statistical information about the problem at hand and other distributions can be derived from the joint distribution.

In Figure 1.7, only a few examples are given which demonstrate how to use the wire-cube to obtain other distributions in BE and AE states. However, all distributions can be derived with ease. Although only three dimensions are used and only two possible true values for each QoI, the wire-cube model is correct in any number of dimensions and any number of possible true values for QoIs. Obviously it would be impossible to represent graphically higher dimensional wire-cubes.

In Figure 1.7(B), the wire-cube represents the model (conditional probabil-

ity) of future observation g conditioned on values of f and $\tilde{f}$. The values of the model are obtained from the joint probability (from the wire-cube Figure 1.7(A)) by selecting the corners in Figure 1.7(A) that corresponds to the same pair of values of $\{f, \tilde{f}\}$ and normalizing them to 1. This is indicated in Figure 1.7(B) by connecting those corners by a thick line. The representation of pre-posterior $p(f, \tilde{f}|g)$ in Figure 1.7(C) is obtained from the joint Figure 1.7(A) by identifying corners that correspond to the same values of the measurement g and normalizing them to 1. Those corners are connected by a thick line shown in Figure 1.7(C). The wire-cube in Figure 1.7(C) represents the pre-posteriors and if the data is measured in AE (either $g = 0$ or $g = 1$) one of the connected thick-line squares corresponding to observed g will become the posterior. Figure 1.7(D) illustrates the posterior $p(f|g)$ which is the result of marginalization of $p(f, \tilde{f}|g)$, which simply adds values of $p(f, \tilde{f}|g)$ in Figure 1.7(C) along $\tilde{f}$ axis. Figuratively speaking, the cube is squeezed and dimensionality of the distribution reduced.

Based on this example as an exercise, the reader may try to obtain various likelihoods or quantities as $p(f)$ etc.

1.9 UNCONDITIONAL AND CONDITIONAL INDEPENDENCE

The last idea introduced in this chapter could be one of the most important concepts for the design of efficient computational algorithms discussed in Chapter 3 and Chapter 6. At the same time, it is one of the hardest to properly understand and gain intuition about. The simple independence of some QoIs f and y was already defined in Section 1.5 and the usual mathematical property indicating this independence is

$$p(f, y) = p(f)p(y) \tag{1.20}$$

with equivalent formulations $p(y|f) = p(y)$ and due to symmetry $p(f|y) = p(f)$. It was explained that independence of y and f indicates that the knowledge of the true value of f does not change uncertainty about y.

If another QoI g is introduced, the joint of three QoIs is defined as $p(f, y, g)$. The joint can be marginalized over f to $p(g, y)$ and after the marginalization the unconditional independence (note that we use the term *unconditional independence* instead of simple independence) is considered again. If the marginalized joint of g and y is independent ($p(g, y) = p(g)p(y)$), then the g and y are *unconditionally independent*. Unconditional independence is defined in situations where at least three QoIs are considered and all QoIs other than g and y are marginalized. Once obtained, the independence for marginalized joint $p(g, y)$ is assessed using standard methods (e.g., Equation (1.20)). The difference between statements that g and y are independent or unconditionally independent is that in the first case the SoN is defined just by two QoIs, whereas unconditional independence is used when SoN is defined by more than two QoIs and other quantities are marginalized.

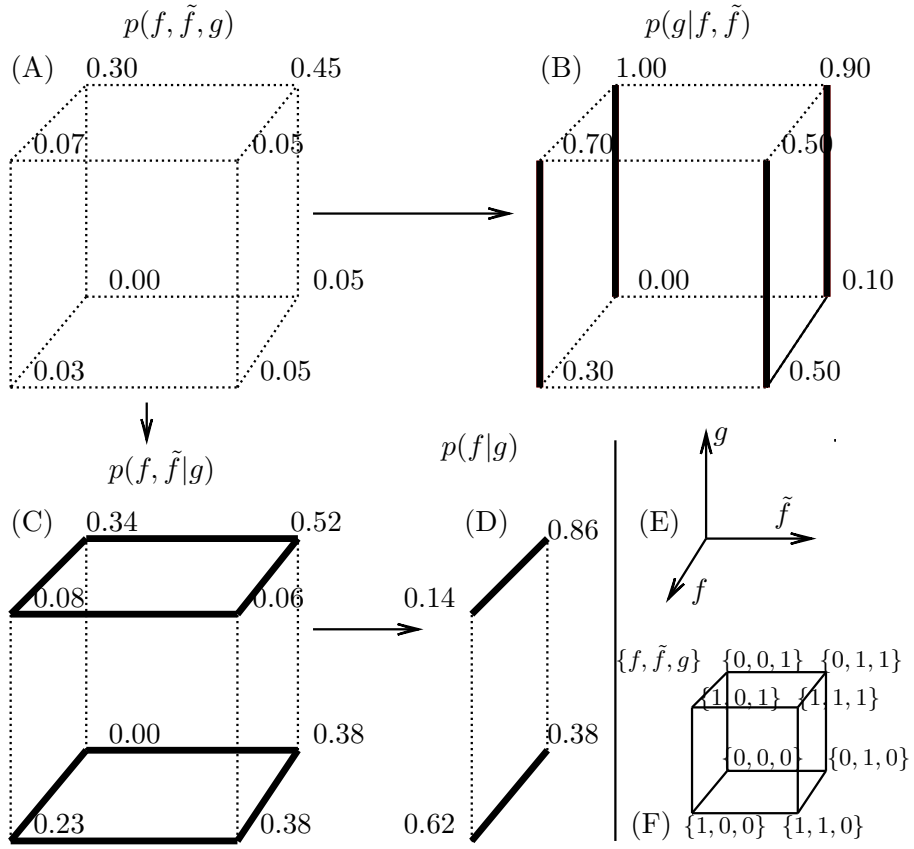


FIGURE 1.7 (A) The wire-cube represents the joint probability of three QoIs f, $\tilde{f}$, and g. There are only two possible true values for each QoI and each corner of the cube correspond to different values of f, $\tilde{f}$, and g, which are shown in (F). The values of each QoI are either 0 or 1. (B), (C), (D) show some conditionals and marginalizations described in detail in Example 1.16.

Consider a *conditional joint* $p(g, y|f)$ which is a joint distribution of g and y conditioned on f which is another way of saying the we consider joint distribution of g and y assuming that the true value of f is known.

Therefore, if the value of f is known we define conditional independence (conditioned on the knowledge of f). Similar to Equation (1.20) the *conditional independence* of g and y given f can be mathematically summarized by

$$p(g, y|f) = p(g|f)p(y|f). \tag{1.21}$$

The above equation has to be true for all $f \in F$ for g and y to be considered

conditionally independent given f. Unconditional independence of g and y ($p(g, y) = p(g)p(y)$) does not imply they are also conditionally independent (when the true value f is known) and vice versa.

The definitions are somewhat abstract and therefore we now put these ideas in some more intuitive context.

Notation First let's introduce a symbol $\perp$ which indicates the independence and therefore $g \perp y$ is read as g is independent of y. If g and y are dependent, that will be indicated simply by 'not $g \perp y$'. The conditional independence and dependence of g and y given f are indicated by $g \perp y|f$ and 'not $g \perp y|f$', respectively.

Using these conventions let's consider the following four scenarios:

1. (not $g \perp y$) and (not $g \perp y|f$)
The QoIs g and y are dependent without any information about f (not $g \perp y$) and with knowledge of true value f (not $g \perp y|f$). This indicates that if either g or y is known, that would change the knowledge about uncertain true value of y or g regardless of whether any knowledge about true value of f is available. However, this does not mean that the change in knowledge will be the same in cases when we have and we do not have information about the true value of f.

2. ($g \perp y$) and (not $g \perp y|f$)
In this situation without the knowledge of true value of f the g and y are independent and no gain in reducing the uncertainty in g or y can be obtained if the true value of y or g becomes known. Upon knowing the true value of f (more information available) the g and y become dependent. To illustrate this, let's consider the following:
Consider two people A and B. The probability of A getting a lung cancer and probability of B getting a lung cancer in the next 10 years without any other information seem to be independent. However, upon obtaining the information that they both smoke, the probabilities are no longer independent, as they both have a higher chance of getting the lung cancer. If the information that is obtained is that person A smokes and person B does not, it makes person A more likely to have a cancer than person B in which case dependence is introduced again, etc.

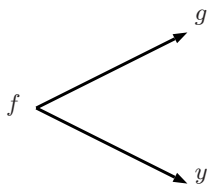
3. (not $g \perp y$) and ($g \perp y|f$)
This situation applies to imaging and is discussed in Section 3.3.3. For this case, the unconditional dependence is "removed" if the true value of f is known. Because this case is important for the applications discussed in this book, it is illustrated with Examples 1.17 and 1.18.

4. ($g \perp y$) and ($g \perp y|f$)
The final case is in a sense the least interesting out of 4 cases listed here. It states that no information can be gained about QoIs g or y

upon knowledge of true value of y or g without or with the knowledge of true value of f.

Example 1.17: Three-dice roll (1)—conditional independence

The conditional example with two-dice roll (Example 1.10) is somewhat modified and three dice are used. Here, a person A rolls die 1 and notes the result f. Then, the same person A rolls die 2 until the number is less than or equal to the number obtained in roll 1 and the result of the last roll becomes g. He repeats the last steps with die 3 obtaining y.



Only the result of g and y are made known to person B who analyzes the problem. Suppose that first we want to establish if there is unconditional independence between g and y. Intuitively, we suspect that there is a dependence because if, for example, $g = 1$ it makes it possible that the unknown f was also 1 in which case y must be 1 as well because of the description of the experiment. It seems that if g is 1 it makes it more probable that $y = 1$ than any other number 2 through 6 compared to a case when no dependence is considered.

In order to verify this let's construct the marginalized distributions $p(g, y)$, $p(g)$, and $p(y)$ and then confirm that $p(g, y) \neq p(g)p(y)$. In doing so we can also confirm our intuition that $p(f = 1|g = 1) > p(f = 1)p(g = 1)$. First, we determine the joint $p(g, y)$ by $p(g, y) = \sum_{f=1}^{6} p(g, y|f)p(f)$. The $p(f)$ is known and equal to $1/6$. The $p(g, y|f)$ can be easily deducted from the description of the experiment and for example for $f = 3$ it is:

	$y=1$	$y=2$	$y=3$	$y=4$	$y=5$	$y=6$
$g=1$	$1/9$	$1/9$	$1/9$	0	0	0
$g=2$	$1/9$	$1/9$	$1/9$	0	0	0
$g=3$	$1/9$	$1/9$	$1/9$	0	0	0
$g=4$	0	0	0	0	0	0
$g=5$	0	0	0	0	0	0
$g=6$	0	0	0	0	0	0

Similarly the $p(g, y|f)$ can be constructed for other values of f 1, 2, 4, 5, and 6. Note that we obtained the entries of this matrix by simply multiplying probabilities $p(g|f)$ and $p(y|f)$ because it is obvious that upon knowing f (hypothetically) the rolls 2 and 3 are independent. And therefore $p(g, y|f) = p(g|f)p(y|f)$ which is the definition of conditional independence. Although the conditional independence is obvious in this example, it will not always be the case with real-world problems considered in this book.

We multiply $p(g, y|f)$'s by $p(f) = 1/6$ and add results obtaining $p(y, g)$:

$$\begin{array}{c|cccccc} & y=1 & y=2 & y=3 & y=4 & y=5 & y=6 \\ g=1 & 5369 & 1769 & 869 & 469 & 244 & 100 \\ g=2 & 1769 & 1769 & 869 & 469 & 244 & 100 \\ g=3 & 869 & 869 & 869 & 469 & 244 & 100 \\ g=4 & 469 & 469 & 469 & 469 & 244 & 100 \\ g=5 & 244 & 244 & 244 & 244 & 244 & 100 \\ g=6 & 100 & 100 & 100 & 100 & 100 & 100 \end{array} \times \frac{1}{21600}$$

Because the above 2D discrete distribution is symmetric upon exchange of rows and columns, it follows that $p(g)$ must be equal to $p(f)$, and can be obtained by summing either the columns or the rows. Once this is done, the values of $p(g, y)$ vs. $p(g)p(f)$ are plotted in Figure 1.8. The plots show that g and y are not unconditionally independent and confirm our intuition about $p(f = 1|g = 1) > p(f = 1)p(g = 1)$.

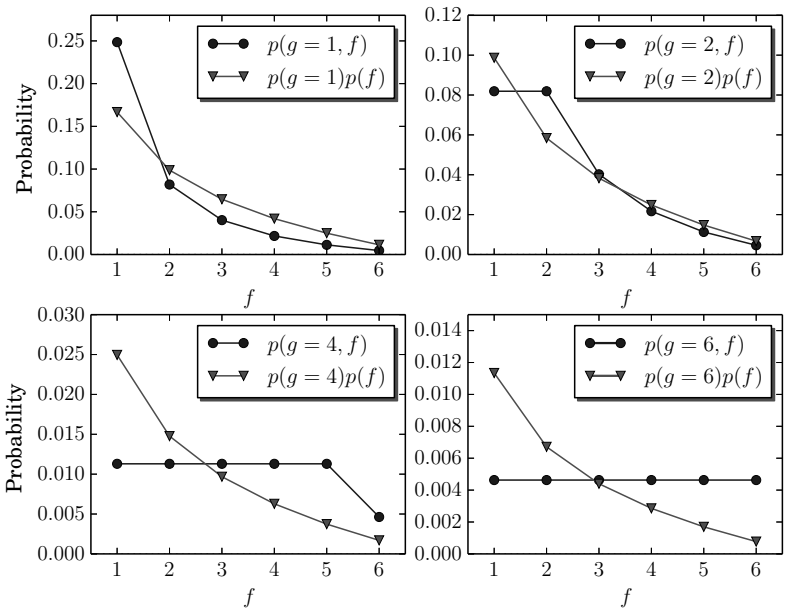


FIGURE 1.8 Comparison of $p(f, g)$ and $p(f)p(g)$ for Example 1.17. Lines connecting the points were added for clarity.

.

Few of the properties of the conditional independence are specified below. From the definition $g \perp y$ for $p(g, y|f) = p(g|f)p(y|f)$. The alternative definition is

$$p(g|f, y) = p(g|y) \tag{1.22}$$

The above is more telling than Equation (1.21) because it directly indicates that the knowledge of the true value of f is irrelevant for gaining any insights about the true value of g if the true value of y is known. The proof of equivalence of Equation (1.21) and Equation (1.22) is shown below:

$$p(g, y|f) = \frac{p(g, y, f)}{p(f)} = \frac{p(g|f, y)p(f, y)}{p(f)} = p(g|f, y)p(y|f) \tag{1.23}$$

Now since $p(g, y|f) = p(g|f)p(y|f)$ we obtain that $p(g|f, y) = p(g|y)$.

Using similar considerations it can be shown that if $f \perp g|y$ then $g \perp f|y$. An interesting property such that $f, y \perp g_1|g_2$ implies $f \perp g_1|g_2$ and $y \perp g_1|g_2$ can be shown as well (g_1 and g_2 are two different QoIs).

Example 1.18: Three-dice roll (2)–conditional independence

Another example relevant to some properties of imaging systems that will be discussed in Chapter 3 is considered. We have a person A rolling die 1 f and then, as before, rolling die 2 until the number (y) is smaller or equal to the number rolled on die 1. Then, he proceeds to rolling die 3; however, he rolls the die 3 until the number is smaller or the same as the number obtained in roll 2.

$$f \longrightarrow y \longrightarrow g$$

Note the difference with the previous Example 1.17 where roll 3 was done with respect to roll 1. In this example it can be shown that the result of roll 3 is conditionally independent of the result of roll 1 if the number obtained in roll 2 is known. In other words

$$p(g|f, y) = p(g|y) \text{ or } g \perp f|y. \tag{1.24}$$

If y is known the knowledge of f is superfluous and unnecessary and does not bring any additional information about g.

This and other properties of conditional independence listed above can be illustrated using a similar approach as used in Example 1.17. The exercise of demonstrating this is left for the reader.

1.10 SUMMARY

In this chapter the basics of the statistical approach that will be used in this book was introduced. The are two main points that have to be emphasized.

The first is that two stages of knowledge about the problem are identified. The *before experiment* and *after experiment* stages which are denoted as BE and AE. The second is that during the BE stage quantities of interests (QoI) are identified that will be considered and grouped into two categories: (1) quantities that will be observed (OQs) and their true values will be revealed in the experiment and (2) QoIs that will still be unknown in the AE stage and the knowledge about their true values uncertain (UQs). It is assumed that all quantities of interest have a deterministic true value and the knowledge (or uncertainty) about the true values is described by probability distributions.

Once the experiment is performed and probability distributions of known quantities are reduced to delta functions, the probability distributions of UQs (that BE are represented by priors) are modified if there is a statistical dependence between UQs and OQs. The formalism of modification of BE and AE distributions of UQs was introduced by the means of the Bayes Theorem. In order to be able to use the Bayes Theorem it was shown that the model of the experiment (conditional probability of OQs conditioned on UQs) is required as well as the prior probability of UQs. In all of the above, the probability distributions are interpreted as measures of plausibility that a given value of QoI is true [13, 22, 88]. This plausibility can be either a subjective belief which reflects the level of uncertainty in knowledge of the true value of the QoIs, or other distributions chosen for cases where subjective knowledge is poor or there are difficulties in summarizing the knowledge with a distribution.

This lack of objectivity in Bayesian formulation of the statistics is one of the major criticisms of the Bayesian approach. This issue was already discussed in Section 1.7.3 and here we reiterate our view on this subject. In applications of statistical analysis in medial imaging and medicine the ultimate goal of the imaging or medical procedures is to make a decision. In imaging one of the most frequently performed tasks is to summarize imaging data by a single image. This becomes a decision problem because usually there will be many images that could be obtained from data acquired by some physical apparatus that are plausible. For example, we can filter images and by adjusting parameters of the filter providing an infinite number of filtered images at which point we need to decide which of those images should be chosen to accomplish the task at hand. Other decisions are made such as whether disease is present or not, etc.

Any time a decision is made based on uncertain data, the subjectivity must be used to make this decision. In the Bayesian model of uncertain data, the subjectivity is introduced explicitly by the definition of the prior (probability distribution of UQ in BE condition) and by the "loss function" (see Section 2.2) when decisions are made. In classical statistics inferences from the experiment are objective[4]. For example, data may be summarized by the

[4]This assuming that the model is correct. In fact, a statistical analysis is always conditional and there are no objective analysis per se, because assumption about the model of the experiment has to be made.

P-value which is the probability of obtaining at least as extreme result as was observed assuming some hypothesis (or model of the experiment) is true. However, in order to make a decision (reject the hypothesis or not) the *significance* (the value of the threshold) has to be selected BE. This selection is a subjective choice that should be varied based on likelihood of the hypothesis.

The misunderstanding of the pseudo-objectivity of classical statistics leads to many incorrect conclusions found in medical literature [35, 46, 89, 101]. For example, the classical hypothesis testing used extensively in medical imaging and medicine, the experimental evidence that is summarized by the classical methods by the P-value should always be evaluated in the light of plausibility of the hypothesis (see the next example); however, the classical statistics provide only limited means for quantitative combination of the findings from the data and plausibility of the hypothesis, and plausibility of hypothesis is seldom discussed in the context of objective evidence summarized by the P-value.

Example 1.19: Summary of experimental evidence: P-values

Suppose we roll a six-sided die thee times and we obtain six all three times. Using measure of classical statistics the null hypothesis that obtaining any number is equally likely can be rejected with a high statistical significance (low P-value). Therefore the objective analysis of the data says that the die is rigged. Obviously any reasonable decision maker who uses prior knowledge would require much substantial evidence to be convinced that the die is "unfair" and based on this subjective judgment, the null hypothesis will not be rejected based on the mentioned experiment. In order to do so, the rejection region, has to be equal to much lower value than P=0.05 or P=0.01 used so extensively in the field. Perhaps for the example with die the significance level should be set at 10^{-6} level. Looking through scientific literature, it is rare that investigators ever discuss the reasoning behind choosing the classical statistical significance level used in their work.

We argue that the use of Bayesian methods (e.g., as the BEAE view) is justified and subjectivity is unavoidable when making decisions about uncertain events. The use of Bayesian approaches force investigators to express their beliefs in a form of the prior rather than combine in some unspecified way the experiential evidence (summarized for example with the P-value) with prior beliefs to make decisions. In medical imaging, decision making is the end-point of any imaging procedure (e.g., disease present or not, disease progresses or not, etc.) and therefore the issue of combining the experimental evidence with the prior beliefs in some coherent way is of utmost importance.