Publishers' page

Publishers' page

Publishers' page

Publishers' page

MULTIVARIATE ANALYSIS IN HIGH ENERGY PHYSICS

By

Pushpalatha C. Bhat and Harrison B. Prosper

Preface

Contents

Preface vii

Chapter 1 Basic Concepts of Data Analysis 1

1.1 Elements of Probability Theory 1
1.1.1 Deductive Reasoning 1
1.1.2 Inductive Reasoning 2
1.1.3 The Axioms of Probability Theory 3
1.1.4 Bayes' Theorem 5
1.1.4.1 Proof of Bayes' Theorem 5
1.2 Elements of Statistical Inference 6
1.2.1 Loss and Risk Functions 9
1.2.2 Frequentist Inference 11
1.2.2.1 Sampling Distribution, Bias and Variance . . 11
1.2.2.2 Efficiency, Likelihood and the Minimum Variance Bound 12
1.2.2.3 Maximum Likelihood and Consistency 13
1.2.2.4 Sampling Distributions of Estimators 15
1.2.3 Bayesian Inference 15
1.2.3.1 Combining Results 17
1.2.3.2 Model Selection 17
1.3 Functional Approximation 18
1.4 Classification 18
1.5 Parameter Estimation 18
1.5.1 Frequentist 18
1.5.2 Bayesian 18

Chapter 1

Basic Concepts of Data Analysis

1.1 Elements of Probability Theory

> Probability theory is nothing but common sense reduced to calculation.
> Laplace, 1819

Probability theory is a vast mathematical continent. Nonetheless, it is possible to glean its essential features by judiciously selecting the parts to be explored. In this section we explore those aspects of probability theory of direct relevance to data analysis in the physical sciences.

1.1.1 *Deductive Reasoning*

In about 350 BC, Aristotle noted that when we reason correctly we do so according to definite rules, which can reduced to a few fundamental ones called *syllogisms*. In addition to the syllogisms listed in Table 1.1.1, we note the rule that if A is *true* then its negation, written as $\overline{A}$, is of necessity *false*.

Table 1.1 Syllogisms.

	modus ponens (ponere=affirm)	*modus tollens* (tollere=deny)
Major premise	If A is TRUE, then B is TRUE	If A is TRUE, then B is TRUE
Minor premise	A is TRUE	B is FALSE
Conclusion	Therefore, B is TRUE	Therefore, A is FALSE

Table 1.2 Syllogism Mnemonics.

	modus ponens	*modus tollens*
Major premise	$AB = A$	$AB = A$
Minor premise	$A = 1$	$B = 0$
Conclusion	$B = 1$	$A = 0$

We say that A *contradicts* $\overline{A}$. A mnemonic for the syllogisms are the set of symbolic expressions given in Table 1.1.1. The symbols A, B, 1, 0, and their *negations*—$\overline{A}$, $\overline{B}$, $\overline{1}$ and $\overline{0}$, are called *propositions*. The proposition 1 has a *truth value* which is always *true*; 0 is always *false*. Here is an example.

Example 1.1 Let $A =$ (You got more votes) and let $B =$ (You become president). Our major premise is *if* (You got more votes), *then* (You become president). Our minor premise is (You got more votes) is *true*. Therefore, our conclusion is (You become president) is *true*. Note, however, that if B is true then it does *not* follow that A is true. In our example, if (You become president) is false then we conclude that (You got more votes) is false. However, if (You become president) is true it does *not* follow that (You got more votes) is true. Conversely, if A is false we cannot conclude that B is false.

The simplest way to see this is to use the symbolic expressions and note that if B is set to 1 (that is, to the proposition that is always *true*) in $AB = A$ we get $A = A$ and we cannot determine the truth value of A; likewise, if $A = 0$ then the truth value of B is undefined.

It should be noted that syllogisms reflect merely the *logical* relationship between two propositions A and B. They are silent as to whether or not a causal relationship exists between them. Two propositions could be logically related yet fail to be causally related.

1.1.2 *Inductive Reasoning*

Deductive reasoning is extremely important to the practice of science. However, in scientific work we are frequently not sufficiently well informed to reason deductively. We may have performed an experiment and acquired data that are insufficient to allow for a precise deduction of the value of a physical parameter. In the example, given above, no one doubts that if the

Table 1.3 Weak Syllogisms.

Major premise	If A is TRUE, then B is TRUE	If A is TRUE, then B is TRUE
Minor premise	B is TRUE	A is FALSE
Conclusion	Therefore, A is plausibly TRUE	Therefore, B is plausibly FALSE

proposition B is false then so too is proposition A. While we risk being proved wrong, most of us would consider it reasonable that if B is, in fact, true then A is likely to be true. The issue is can such a vague statement be made precise?

Interestingly, an affirmative answer was provided by the physicist R.T. Cox in a 1946 article in the *American Journal of Physics* *. Cox showed that a caculus can be established that assigns truth values that lie between *false* and *true*. These truth values can be represented by real numbers in the interval $[0, 1]$. The rules of the calculus established by Cox are isomorphic to those of probability theory. Consequently, these truth values can be construed as providing one interpretation of **probabilities**, which strictly speaking are mathematical abstractions. These truth values are a measure of the *plausibility* of propositions. The Cox theory, which solidified the work of Thomas Bayes, Laplace, Sir Harold Jeffreys and others, is really a theory of inductive reasoning. It provides quantitative meaning to the weaker syllogisms shown in Table 1.1.2.

1.1.3 *The Axioms of Probability Theory*

Probability theory is based upon two sets of rules. The first set is George Boole's algebra of propositions, called **Boolean algebra**. Let A, B, C, 1, 0 and their negations be propositions, and $+$ and $\cdot$ binary operations defined on them. The axioms of Boolean algebra are

$$
\begin{array}{llllll}
(1) & A+0 & = & A & A\cdot 1 & = & A \\
(2) & A+\overline{A} & = & 1 & A\cdot \bar{A} & = & 0 \\
(3) & A\cdot B & = & B\cdot A & A+B & = & B+A \\
(4) & A\cdot(B+C) & = & A\cdot B+A\cdot C & A+B\cdot C & = & (A+B)\cdot(A+C).
\end{array}
$$

*R. T. Cox, Probability, Frequency, and Reasonable Expectation, Am. J. Phys. **14** (1946) 1.

Axioms (3) and (4) are known as the commutative and distributive laws, respectively. Several equivalent sets of axioms can be written down. This particular set is referred to as the *Huntington axioms*. Note the duality between the operations $+$ and $\cdot$. For simplicity, the $\cdot$ operation is often omitted.

The second set of rules are the *product rule*

$$\begin{aligned} P(AB|C) &= P(B|AC)P(A|C), \\ &= P(A|BC)P(B|C), \end{aligned} \tag{1.1}$$

and the *sum rule*

$$P(A|B) + P(\overline{A}|B) = 1, \tag{1.2}$$

and the conventional numerical assignments for true and false

$$P(1|B) = 1, \tag{1.3}$$

$$P(0|B) = 0. \tag{1.4}$$

The symbol $P(*|*)$ denotes a probability. The Cox theorem derives rules Eqs. (1.2) and (1.2) as a consequence of a more basic set of axioms. Other axiom sets have been proposed, most notably by Kolmogorov who defined probability in terms of measures and sets. However, for our purposes the rules stated above are adequate.

All theorems of probability theory, when combined with the rules of arithmetic, can be deduced from these axioms. We list a few of the more important ones below.

(1)	$A+1$	$=$	1	$A0$	$=$	0
(2)	$\overline{0}$	$=$	1	$\overline{1}$	$=$	0
(3)	$A+AB$	$=$	A	$A(A+B)$	$=$	A
(4)	AA	$=$	A	$A+A$	$=$	A
(5)	$A(BC)$	$=$	$(AB)C$	$A+(B+C)$	$=$	$(A+B)+C$
(6)	$\overline{AB}$	$=$	$\overline{A}+\bar{B}$	$\overline{A+B}$	$=$	$\overline{A}\,\bar{B}$.

Theorems (4) and (5) are known as the *idempotency*, and *associative* laws, respectively, while (6) is known as *de Morgan's laws*. To illustrate the use of these rules we shall prove a theorem that relates $P(A+B|C)$ to $P(A|C)$ and $P(B|C)$. To do so we simply apply the rules repeatedly in more or less

the right order

$$
\begin{aligned}
P(A+B|C) &= 1 - P(\overline{A+B}|C) \\
&= 1 - P(\overline{A}\,\overline{B}|C) \\
&= 1 - P(\overline{B}|\overline{A}C)P(\overline{A}|C) \\
&= 1 - \left[1 - P(B|\overline{A}C)\right] P(\overline{A}|C) \\
&= 1 - P(\overline{A}|C) + P(B|\overline{A}C)P(\overline{A}|C) \\
&= P(A|C) + P(B|\overline{A}C)P(\overline{A}|C) \\
&= P(A|C) + P(\overline{A}B|C) \\
&= P(A|C) + P(\overline{A}|BC)P(B|C) \\
&= P(A|C) + [1 - P(A|BC)]\, P(B|C) \\
&= P(A|C) + P(B|C) - P(A|BC)P(B|C) \\
P(A+B|C) &= P(A|C) + P(B|C) - P(AB|C). \qquad (1.5)
\end{aligned}
$$

This is something one could have anticipated from a consideration of Venn diagrams. Indeed, many tentative theorems of probability theory can be guessed in this way. The hard part, of course, is establishing that the tentative theorem is either true or false using only the rules given above.

1.1.4 *Bayes' Theorem*

In 1763 Thomas Bayes published a paper † in which he gave a formula, the general version of which now bears his name. Bayes' theorem

$$P(B_k|AC) = \frac{P(A|B_kC)P(B_k|C)}{\sum_i P(A|B_iC)P(B_i|C)}, \qquad (1.6)$$

where B_k are considered to be a set of mutually exclusive–and exhaustive–propositions, is a direct consequence of the product rule. The theorem has become an extremely important, and controversial, one for data analysis. Here is a simple proof of it.

1.1.4.1 *Proof of Bayes' Theorem*

Two propositions A and B are said to be *mutually exclusive* if the truth of one denies the truth of the other, that is: $P(AB|C) = 0$. In that case, from

†Posthumously; he died in 1761.

the theorem we proved above, we get

$$P(A+B|C) = P(A|C) + P(B|C), \tag{1.7}$$

which is readily generalized to any number of mutually exclusive propositions. A set of mutually exclusive propositions B_k is said to be *exhaustive* if their probabilities sum to unity

$$\sum_k P(B_k|C) = 1. \tag{1.8}$$

If $B_k D_j$ are a set of mutually exclusive and exhaustive **joint propositions** then we can write Bayes' theorem as

$$P(B_k D_j|AC) = \frac{P(A|B_k D_j C)P(B_k D_j|C)}{\sum_{i,l} P(A|B_i D_l C)P(B_i D_l|C)}. \tag{1.9}$$

To prove the theorem we replace B, in the product rule Eq. (1.2), by the joint proposition $B_k D_j$

$$P(A|B_k D_j C)P(B_k D_j|C) = P(B_k D_j|AC)P(A|C). \tag{1.10}$$

and then sum over the set of joint propositions

$$\begin{aligned} \sum_{k,j} P(A|B_k D_j C)P(B_k D_j|C) &= P(A|C)\sum_{k,j} P(B_k D_j|AC) \\ &= P(A|C). \end{aligned} \tag{1.11}$$

In the last step we have assumed that the joint propositions $B_k D_j$ form an exhaustive set. Using Eq. (1.10), we substitute into Eq. (1.11) the formula for $P(A|C)$ and re-arrange to arrive Bayes' theorem, Eq. (1.31). It is should be noted that to reduce Bayes' theorem to a statement about B_k only, we need merely sum Eq. (1.31) over the propositions D_j. Bayes' theorem is easily extended to encompass larger sets of joint propositions.

1.2 Elements of Statistical Inference

The aim of statistical inference is to extract useful information from data as efficiently as possible, even in the presence of uncertain knowledge about relevant aspects of the data collection system. In science the data collection system is usually a well defined experiment. This task is intrinsically one of induction, for which probability theory is the natural language.

Our statement of the aim of statistical inference, while true, is too vague to be immediately useful. A more useful way of stating the goal, from the point of view of a scientist, is that it is to establish "facts" about the world. A seismologist, for example, might be interested in establishing the total energy released in an earthquake. A high energy physicist may wish to know the mass of the top quark. More generally, a scientist may be interested in establishing to what extent one or more models, currently under consideration, accord with observation. To the degree that a model accords with observation it can be construed as a (tentative) fact about the world. In this book we are primarily concerned with facts about the world that can be described by the parameters of a mathematical model. These parameters need not necessarily have deep physical significance. Sometimes our interest may be to develop an empirical model of some phenomenon, in which case it may be sufficient to use a model having a convenient set of adjustable parameters none of which may have any physical significance.

The object of interest, therefore, is a vector of K **parameters** $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_K)$ of some mathematical model, $\mathrm{M}(\boldsymbol{\Theta})$. Together with one or more models, we have a vector $\mathbf{x} = (x_1, \ldots, x_N)$ of data, which has been acquired in the hope of shedding light on two things: 1) the degree to which the model accords with observation and 2) the values of its parameters. The first task is called **hypothesis testing** or, more generally, **model selection**; the second is called **parameter estimation**.

The fundamental idea, which happily every scientist accepts, is that the connection between the model $\mathrm{M}(\boldsymbol{\Theta})$ and our data $\mathbf{x}$ is a **probability**, $P(\mathbf{x}|\boldsymbol{\Theta})$. That symbol is to be read as "the probability of $\mathbf{x}$ given $\boldsymbol{\Theta}$". In general, the probability can be written as

$$P(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{z \in \Omega} f(\mathbf{z}|\boldsymbol{\Theta}), \tag{1.12}$$

where the sum is over a neighborhood Ω of the point $\mathbf{x}$. If the point $\mathbf{x}$ is a member of a continuous set the sum is to be replaced by an integral. In practice, the neighborhoods are chosen to be very small, in which case we can write

$$P(\mathbf{x}|\boldsymbol{\Theta}) = f(\mathbf{x}|\boldsymbol{\Theta})\, d\mathbf{x}, \tag{1.13}$$

if $\mathbf{x}$ forms a continuous set, or

$$P(\mathbf{x}|\boldsymbol{\Theta}) = f(\mathbf{x}|\boldsymbol{\Theta}), \tag{1.14}$$

if $\mathbf{x}$ is discrete. The function $f(\mathbf{x}|\mathbf{\Theta})$ is called the **probability density function** (pdf), if $\mathbf{x}$ is continuous.

The number $P(\mathbf{x}|\mathbf{\Theta})$ is the probability that we would *assign* to our data if we knew the values of the parameters, assuming the model to be correct. We say "assigned" because in most situations that is exactly what we do. The probabilities are generally derived by contemplating the nature of the experiment being performed. When we choose to use a Poisson distribution to describe the counts in an experiment it is generally because some mathematical reasoning led to that choice. It is quite rare that we are able to say with certainty that a particular probability describes our experiment. Ultimately, the probabilities used are based on what we believe we *know* about our experiment.

When we go from the parameters of a model and predict all possible data that can be obtained, we are performing **deductive** reasoning. Statistical inference solves the inverse problem: we have data and we wish to infer something about the correctness or otherwise of the models and the values of their parameters. This entails **inductive** reasoning.

The problem of inference has been solved in two different ways. The first solution was presented by the English cleric Thomas Bayes (1763) and the great mathematical physicists of the 18th and 19th centuries Daniel Bernoulli, Laplace and Gauss. Their solution, now called *Bayesian statistics*, fell into disfavor towards the end of the 19th and the early part of the 20th century. However, since the middle of the 20th century Bayesian statistics has experienced a vigorous renaissance, which was initiated by Sir Harold Jeffreys (1891 - 1989) in the 1930s. The rivival was sustained by many scientists, notably the physicist Edwin T. Jaynes.

In the Laplace-Bayes theory, probability is interpreted as a measure of the plausibility, or "degree of belief", of a proposition, an interpretation that was given a sound underpinning by Cox in 1946. Consider the following everyday statement: "There is an 80% chance of rain tomorrow". This statement instructs us to assign a degree of belief of 0.8 to the statement "It will rain tomorrow".

The second solution, called variously *classical, orthodox* or *frequentist* statistics was developed in the early part of the 20th century. Its chief proponents were Sir Ronald Aylmer Fisher (1890 - 1962) and Jerzy Neyman. Important contributions were made by Pearson, Cramer, Rao, von Mises and Kolmogorov. In their approach probability is interpreted as the *relative*

frequency with which something occurs. Therefore the statement "There is an 80% chance of rain tomorrow" is considered devoid of meaning because of the impossibility of repeating that particular day. On the other hand, the statement "There is an 80% chance of rain on Mondays in Seattle" makes sense, since Mondays in Seattle occur rather frequently.

1.2.1 *Loss and Risk Functions*

Given a data-set $\mathbf{x} = (x_1, \ldots, x_N)$ and a mathematical model $\mathrm{M}(\mathbf{\Theta})$ and the associated probability $P(\mathbf{x}|\mathbf{\Theta})$ we use statistical analysis to decide the best values to assign to the parameters $\mathbf{\Theta}$. If we have several models $\mathrm{M}(\mathbf{\Theta}_1)_1, \mathrm{M}(\mathbf{\Theta}_2)_2, \ldots$ then we will, in addition, want to decide which one is best, which begs the question: What does one mean by best?

The mapping $(x_1, \ldots, x_N) \longrightarrow (\theta_1, \ldots, \theta_M)$ from our data-set to the parameters of a given model is called a **decision function**; we shall denote it by the symbol $d(x_1, \ldots, x_N)$. For now, we shall limit our discussion to models that depend upon a single parameter θ. We denote by $\hat{\theta}$ any **estimate** of θ. If the decision function is chosen so that $\hat{\theta} = d(x_1, \ldots, x_N)$ then d is called an **estimator** of θ. When we insert into it a given data-set, $\mathbf{x}_0$, the function d gives a specific **estimate**, $\hat{\theta}_0$, of θ.

Suppose someone asks you for the time one bright sunny day. In spite of the fact that you do not have a watch with you, you wish nonetheless to be helpful. You note that the sun is high in the sky, which you take as your datum. Given this datum, here are two possible decision functions

- The sun is high in the sky; therefore, say it is not midnight.
- The sun is high in the sky; therefore, say it is noon.

The first time estimate is rather safe, but not too useful, while the second one is likely to be true only approximately. The question is: Which is the better estimate?

To choose a decision function we need a way to quantify the quality of its estimates. Since bad decisions entail loss we construct a *loss function*, $\mathcal{L}(\theta, d)$, to measure the loss arising from a decision. The loss function depends both on our decision function, which in principle we know, and on the parameter, whose value we do not know. Moreover, the loss function, through its dependence on the decision function, is a function of the data.

The most commonly used loss function is the *quadratic loss*

$$\mathcal{L}(\theta, d) = (\theta - d)^2 . \tag{1.15}$$

Its main virtue is that it is both reasonable and mathematically convenient. In practice, one does not use the loss function directly, but rather some sort of average thereof, called the *risk function.* But, the averaging is done differently in Bayesian and frequentist statistics. The reason is because of an important conceptual difference between the two approaches, which is stated below.

- **Frequentist:** Inferences can be based on data that could have been observed as well as data observed.
- **Bayesian:** Inferences must be based only on data observed.

Accordingly, a frequentist asserts that the averaging should be done over all data-sets that could have been observed. This leads to the risk function

$$\mathcal{R}_d(\theta) = \mathrm{E}[\mathcal{L}(\theta, d)]_{\mathbf{x}}. \tag{1.16}$$

We note that the frequentist risk function is an ordinary function of the parameter θ, but a functional of the decision function d, that is, it depends on the set of all possible values of d. The symbol $\mathrm{E}[\cdot]_{\mathbf{x}}$ is the averaging or **expectation operator**.

In accordance with the Bayesian viewpoint, however, the averaging is done not with respect to all possible data-sets, but instead over all possible propositions (that is, hypotheses) about the value of θ. The only data-set that enters the risk function, in this case, is that which was obtained. The probability of a given hypothesis about the value of θ, given the data-set $\mathbf{x}$, is denoted by $P(\theta|\mathbf{x})$. The risk function in Bayesian statistics is therefore given by

$$\begin{aligned} \mathcal{R}_d(\mathbf{x}) &= \mathrm{E}[\mathcal{L}(\theta, d)]_\theta, \\ &= \int_\theta \mathcal{L}(\theta, d) P(\theta|\mathbf{x}). \end{aligned} \tag{1.17}$$

Inserting into it the quadratic loss function, Eq. (1.15), we obtain

$$\mathcal{R}_d(\mathbf{x}) = \int_\theta (\theta - d)^2 P(\theta|\mathbf{x}). \tag{1.18}$$

The corresponding frequentist risk function, called the *mean squared error*, is

$$\mathcal{R}_d(\theta) = \mathrm{E}[(\theta - d)^2]_{\mathbf{x}}. \tag{1.19}$$

Owing to their different interpretations of probability, inference proceeds differently in frequentist and Bayesian statistics. While there are common concepts and features, we think it better to consider them separately from this point onwards. The following two sections consider inference from the frequentist and Bayesian perspectives, respectively.

1.2.2 *Frequentist Inference*

1.2.2.1 *Sampling Distribution, Bias and Variance*

The *mean squared error* can be written in a more useful form. To do so, we apply the expectation operator to the function d itself. Remember, in Eq. (1.19), we are averaging over all possible data-sets. The data-set we have obtained is regarded as having been drawn from an *ensemble* of all possible data-sets. The probability to draw a given data-set is assumed to be $P(\mathbf{x}|\theta)$. When this probability is interpreted in this way it is called the *sampling distribution* of $\mathbf{x}$. The expectation value of d can be written as

$$\mathrm{E}[d(x_1, \ldots, x_N)]_{\mathbf{x}} \equiv \theta + b(\theta). \tag{1.20}$$

The function $b(\theta)$ is called the *bias*. We can re-write Eq. (1.19) as

$$\mathcal{R}_d(\theta) = \mathrm{Var}[d]_{\mathbf{x}} + b(\theta)^2, \tag{1.21}$$

where

$$\mathrm{Var}[d]_{\mathbf{x}} \equiv \mathrm{E}[d^2]_{\mathbf{x}} - \mathrm{E}[d]^2_{\mathbf{x}}, \tag{1.22}$$

defines the *variance operator*, $\mathrm{Var}[\cdot]_{\mathbf{x}}$. The expectation value, or the *mean*, and the variance, or its square-root the *standard deviation*, are important characteristics of a sampling distribution.

If our estimates $\hat{\theta} = d(x_1, \ldots, x_N)$ are such that $b(\theta) = 0$ for every possible value of θ then our estimates $\hat{\theta}$ and our estimator d are said to be *unbiased*. In that happy circumstance the mean squared error is equal to the variance of the sampling distribution. If we have several estimators of the parameter of interest it seems prudent to use the best one. But to do

that we must answer the question: What is the best estimator? A sensible answer would be

Proposition 1.1 *The best estimator is that which minimizes the risk.*

However, in practice the answer accepted by most scientists is

Proposition 1.2 *The best estimator is unbiased and has the smallest variance.*

It is important to understand that these two answers are different, and would in general yield different choices for the "best" estimator. The first answer could well pick an estimator that is biased! But, by construction, this will be the estimator that minimizes the frequentist risk. Since the second answer insists on zero bias, it may not minimize the risk function.

1.2.2.2 *Efficiency, Likelihood and the Minimum Variance Bound*

If one had the best estimator for the inference problem under consideration, assuming that one exists for that problem, how would one recognize it? We begin with some definitions. The *efficiency* of an estimator d is defined by

$$e(\theta) \equiv \mathrm{Var}[d^*]/\mathrm{Var}[d], \tag{1.23}$$

where d^* is the best estimator, as defined in the previous section. The *likelihood function* $L(\theta)$, as defined by Fisher, is the probability density function $f(\mathbf{x}|\theta)$ regarded as a *function of the parameter*. The data-set $\mathbf{x}$ is to be regarded as a constant, equal to the observed data-set. Therefore, the likelihood function characterizes the specific data-set obtained. Interestingly, on this specific point, Fisher's point of view accords with that of the Bayesians, namely that the likelihood function provides a complete encoding of a given data-set. In this book, we shall try to make a clear distinction between a probability density function $f(*|*)$, which is a function both of data and parameters, and the likelihood function $L(*)$, which is a particular instance of the former obtained by fixing the data to the observed values.

The probability density function is an important quantity for which many theorems have been derived. One such is that proved by Aitken, Silverstone, Fréchet, Cramér and Rao, which is known as the *Cramér-Rao inequality* or the *minimum variance bound*. The theorem states that

Theorem 1.1

$$Var[d] \geq \frac{(1 + D_\theta b(\theta))^2}{E[(D_\theta \ln f(\mathbf{x}|\theta))^2]}, \tag{1.24}$$

with the corollary that

Corollary 1.1

$$Var[d^*] = \frac{1}{E[(D_\theta \ln f(\mathbf{x}|\theta))^2]} \tag{1.25}$$

is the variance of the best (unbiased) estimator of θ, if it exists. The symbol D_θ is the derivative operator with respect to θ. The main mathematical requirement on $f(\cdot|\cdot)$ for the theorem to hold true is that the boundary of the domain over which the probability density function is defined is independent of the parameter θ. (See the Appendix for a simple proof.)

The upshot of this, is that if the efficiency of an estimator turns out to be unity, then one is assured that this estimator is the best one for your parameter, according to the criterion given above. Although we have a test for "bestness" it would clearly be useful to have a systematic way to construct such estimators.

1.2.2.3 *Maximum Likelihood and Consistency*

The first person to use the idea of *maximum likelihood* was Gauss, the Prince of Mathematics. However, it was Fisher who elevated the principle to its justly famous status. The principle of maximum likelihood provides a frequentist with a general and systematic way to construct estimators. The principle is

Proposition 1.3 *Choose as an estimate the value of θ, call it $\hat{\theta}$—which will be equal to some function $d(\mathbf{x})$, that maximizes the likelihood function. That is, we solve $D_\theta L(\mathbf{x})\hat{\theta} = 0$, for $\hat{\theta}$.*

While this principle can be arrived at by choosing the appropriate loss function and minimizing the corresponding risk, its real justification is that it yields frequentist estimators with desirable properties. The most important consequence of the principle is the following theorem, stated without proof,

Theorem 1.2 *If an efficient unbiased estimator exists for a parameter θ then the likelihood principle will find it.*

An estimator is said to be *consistent* if, as the size of the data-set grows without limit, the bias $b(\theta)$ tends to zero. An important feature of maximum likelihood estimators is that they are, in general (although not always), consistent. For large samples, the distribution of the estimates $\hat{\theta}$ has a mean equal to the true value of θ and a variance that attains the minimum variance bound. This state of affairs is referred to as the *asymptotic limit.* It is reasonable to expect that as more and more data are accumulated inferences based on them will yield better and better results. Consistency is such a desirable property that one would be inclined to reject any method that failed to satisfy this criterion.

Another desirable property of estimators obtained from the maximum likelihood principle is invariance of the estimators under re-parameterization. Consider $\hat{\theta}$, found by solving

$$\mathrm{D}_\theta L(\mathbf{x})\hat{\theta} = 0, \tag{1.26}$$

or what is usually more convenient

$$\mathrm{D}_\theta \ln L(\mathbf{x})\hat{\theta} = 0. \tag{1.27}$$

Suppose, however, it proved to be more convenient to solve the above equation using instead the parameter $\alpha = g(\theta)$. It is readily shown that $\hat{\alpha} = g(\hat{\theta})$. Therefore, one might as well solve the problem using the most convenient parameterization and, at the end, transform to the parameter of interest.

There is, however, a subtlety. For a given likelihood function, if a best estimator exists it is unique; consequently, all other estimators will, in general, be biased! The easiest way to see this is to do a Taylor series expansion of $\hat{\alpha} = g(\hat{\theta})$ about the parameter θ:

$$\hat{\alpha} \approx g(\theta) + (\hat{\theta} - \theta)\,\mathrm{D}_\theta g(\theta) + \frac{1}{2}(\hat{\theta} - \theta)^2 \mathrm{D}_\theta^2 g(\theta), \tag{1.28}$$

and then apply the expectation operator to both sides, where in accordance with frequentist reasoning the averaging is done with respect to all possible data-sets $\mathbf{x}$. Assuming $\hat{\theta}$ to be unbiased we get

$$\mathrm{E}[\hat{\alpha}] \approx \alpha + \frac{1}{2}\mathrm{Var}[\hat{\theta}]\mathrm{D}_\theta^2 g(\theta), \tag{1.29}$$

which shows that the estimate $\hat{\alpha}$ is biased, in general, even if $\hat{\theta}$ is not. The above, however, suggests an approximate way to reduce the bias in $\hat{\alpha}$, which

we leave as a simple puzzle for the reader!

1.2.2.4 *Sampling Distributions of Estimators*

As noted above the probability $P(\mathbf{x}|\theta) = f(\mathbf{x}|\theta)\, d\mathbf{x}$ is interpreted as the sampling distribution of the data-set $\mathbf{x}$. From the maximum likelihood principle we can obtain an estimate $\hat{\theta} = d(x_1, \ldots, x_N)$. The next task is to compute the sampling distribution $P(\hat{\theta}|\theta)$ of the estimator $\hat{\theta}$. The basic method is simply stated. One computes the integral

$$P(\hat{\theta}|\theta) = d\hat{\theta} \int_{\mathbf{x}} \delta(\hat{\theta} - d(\mathbf{x}))P(\mathbf{x}|\theta). \tag{1.30}$$

The integral is over all possible data-sets $\mathbf{x} = (x_1, \ldots, x_N)$ subject to the constraint that the estimate $\hat{\theta}$ is equal to the value of the estimator $d(\mathbf{x})$. The δ-function is the simplest way to impose that constraint. In practice, it is seldom possible to calculate the integral in Eq. (1.30) analytically and one must resort to numerical methods, usually involving Monte Carlo techniques. The uncertainty in the estimate $\hat{\theta}$ is judged either from the width of the sampling distribution at a *fixed* parameter value θ, or more sensibly from the width of the likelihood function. Recall, that by definition the latter is the probability density function for a fixed data-set, here a particular estimate $\hat{\theta}$. Note, that this width will vary from one data-set to another. On the other hand, the width of the sampling distribution simply measures the spread of the possible estimates $\hat{\theta}$ for an assumed value of the parameter θ.

1.2.3 *Bayesian Inference*

Since, Bayes' theorem

$$P(B_k D_j|AC) = \frac{P(A|B_k D_j C)P(B_k D_j|C)}{\sum_{i,l} P(A|B_i D_l C)P(B_i D_l|C)}, \tag{1.31}$$

is, after all, a theorem, it is true however one interprets probability. However, when probability is interpreted as a *degree of belief* the theorem assumes a profound role in Bayesian inference. But first a few more definitions. The probability $P(B_k D_j|C)$ is called the **prior** probability assigned to proposition $B_k D_j$; $P(A|B_k D_j C)$ is the **likelihood** assigned to proposition A given the joint proposition $B_k D_j C$ and $P(B_k D_j|AC)$ is the **posterior** probability of proposition $B_k D_j$. We shall suppose that the

proposition B_k is the object of interest and the various D_j are not. The Bayesian inference procedure is

- Assign a prior probability $P(B_k D_j|C)$ to each proposition $B_k D_j$ in light of prior knowledge, represented by proposition C.
- Acquire some pertinent data, represented by proposition A, and assign to these data a probability $P(A|B_k D_j C)$.
- Compute the posterior probability $P(B_k D_j|AC)$ from Bayes' theorem. However, since our interest is in the set of propositions $\{B_k\}$ we sum over the uninteresting ones D_j: $P(B_k|AC) = \sum_j P(B_k D_j|AC)$ to arrive at the posterior probability $P(B_k|AC)$ from which questions about B_k can be answered.

In many applications the propositions are just declarations that a set of parameters have a set of values. Moreover, these parameters are often assumed to form a continuous set. Suppose that

$$P(\mathbf{x}|\theta,\lambda) = \int_\Omega f(\mathbf{z}|\theta,\lambda)d\mathbf{z}, \tag{1.32}$$

is the probability assigned to the data-set $\mathbf{x}$, contained in a neighborhood Ω of $\mathbf{x}$, and that θ and λ are the parameters of the model being tested. Perhaps, θ is the parameter that is of interest while λ is not —that is, λ could represent one or more *nuisance parameters*. If $P(\theta,\lambda) = h(\theta,\lambda)d\theta d\lambda$ is the prior probability we have assigned to the proposition that θ and λ have certain values we can write Bayes' theorem as

$$P(\theta,\lambda|\mathbf{x}) = \frac{P(\mathbf{x}|\theta,\lambda)P(\theta,\lambda)}{\int_{\theta,\lambda} P(\mathbf{x}|\theta,\lambda)P(\theta,\lambda)}, \tag{1.33}$$

which is valid for a continuous set of propositions. Notice that we have dropped the dependence of the probabilities on the prior information C; this we do to simplify the expressions. The dependence on C is now implicit. We remove the nuisance parameters λ by integrating the posterior probability with respect to λ

$$P(\theta|\mathbf{x}) = \int_\lambda P(\theta,\lambda|\mathbf{x}). \tag{1.34}$$

This is the general method of performing inferences in Bayesian statistics and of dealing with nuisance parameters, the nature of which is unrestricted.

For example, the nuisance parameters could represent "theoretical uncertainties". The nuisance parameters represent everything about which we are uncertain. We might add, that no such general method exists in the Fisher-Neyman-von Mises theory.

1.2.3.1 *Combining Results*

Combining results is easy and systematic. The posterior probability obtained from one experiment serves as the prior probability for another. Actually, the order in which experiments are combined is completely immaterial. By repeatedly applying Bayes' theorem it is easy to show that for K independent data-sets $\mathbf{x}_1, \ldots, \mathbf{x}_K$ the overall posterior probability is

$$P(\theta, \lambda | \mathbf{x}_1, \ldots, \mathbf{x}_K) = \frac{P(\mathbf{x}_1|\theta, \lambda) \cdots P(\mathbf{x}_K|\theta, \lambda) P(\theta, \lambda)}{\int_{\theta,\lambda} P(\mathbf{x}_1|\theta, \lambda) \cdots P(\mathbf{x}_K|\theta, \lambda) P(\theta, \lambda)}. \tag{1.35}$$

1.2.3.2 *Model Selection*

Suppose we have two competing models $\mathrm{M}_1(\theta_1)$ and $\mathrm{M}_2(\theta_2)$, which may depend upon different sets of parameters θ_1 and θ_2. Given some prior information C and some data-set $\mathbf{x}$ how do we decide between the two models? The first thing we do is assign a probability to our data-set in light of each model, every hypothesis about the value of its parameter (or parameters) and our prior information

$$P(\mathbf{x}|\mathrm{M}_1 C) = \int_{\theta_1} P(\mathbf{x}|\theta_1 \mathrm{M}_1 C) P(\theta_1|C), \tag{1.36}$$

and

$$P(\mathbf{x}|\mathrm{M}_2 C) = \int_{\theta_2} P(\mathbf{x}|\theta_2 \mathrm{M}_2 C) P(\theta_2|C). \tag{1.37}$$

We then use Bayes' theorem to compute the posterior probability of each model; for example, the probability of model M_1 in light of the information we have is

$$P(\mathrm{M}_1|\mathbf{x}C) = \frac{P(\mathbf{x}|\mathrm{M}_1 C) P(\mathrm{M}_1|C)}{P(\mathbf{x}|\mathrm{M}_1 C) P(\mathrm{M}_1|C) + P(\mathbf{x}|\mathrm{M}_2 C) P(\mathrm{M}_2|C)}, \tag{1.38}$$

where $P(\mathrm{M}_j|C)$ are the prior probabilities we assign to the models. If each model is thought to be equally plausible *a priori*, we might choose to set $P(\mathrm{M}_1|C) = P(\mathrm{M}_2|C) = 0.5$.

It is very important to understand that the probabilities $P(\mathrm{M}_j|\mathbf{x}C)$ make sense only in the context of a given set of models. If someone came up with a third model then the probabilities would, in general, change were we to include the new model amongst the set under consideration. Therefore, it is wrong to regard $P(\mathrm{M}_j|\mathbf{x}C)$ as an absolute measure of the validity of a model. It serves merely as a way to *compare* the plausibility of a given set of models, in light of what we know. A rational thinker would opt for the model with the highest posterior probability and, of course, should we acquire further pertinent information that information, via Bayes' theorem, could cause our rational thinker to change her mind.

1.3 Functional Approximation

1.4 Classification

1.5 Parameter Estimation

1.5.1 *Frequentist*

1.5.2 *Bayesian*

Given the posterior probability of the parameter of interest, having integrated out all nuisance parameters, we can derive an estimator of the parameter from it. Furthermore, we can quantify the uncertainty of our estimate directly from the posterior probability. In fact, as noted again below, the variance of the posterior probability distribution is one such measure.

Earlier we introduced the idea of a loss function as a way to measure the quality of a decision. A typical decision is: Given a data-set $\mathbf{x} = (x_1, \ldots, x_N)$ decide that the estimate of θ is $\hat{\theta} = d(\mathbf{x})$, where $d(\mathbf{x})$ is a special kind of decision function called an estimator. We shall consider the quadratic loss, introduced earlier

$$\mathcal{L}(\theta, d) = (\theta - d)^2 \,. \tag{1.39}$$

As in the frequentist approach, the best estimator, for a specified loss function, is declared to be that which minimizes the risk:

$$\mathrm{D}_d \mathcal{R}_d(\mathbf{x}) \quad = \quad \mathrm{D}_d \int_\theta \mathcal{L}(\theta, d) P(\theta|\mathbf{x}),$$

$$
\begin{aligned}
&= \int_\theta \mathrm{D}_d \mathcal{L}(\theta, d) P(\theta|\mathbf{x}), \\
&= 0. \qquad (1.40)
\end{aligned}
$$

The symbol D_d is the derivative operator with respect to $d(\mathbf{x})$. We have made the usual assumption that the derivative and integral operators commute. When we perform the minimization using the quadratic loss function we obtain as the best Bayesian estimator for θ the intuitively pleasing result

$$\hat{\theta} = d(\mathbf{x}) = \int_\theta \theta P(\theta|\mathbf{x}). \qquad (1.41)$$

In words:

Proposition 1.4 *The best Bayesian estimator of a parameter, given a quadratic loss function, is the mean of the posterior probability.*

One measure of the accuracy of the estimator is the variance

$$\mathrm{Var}[d(\mathbf{x})] = \int_\theta \theta^2 P(\theta|\mathbf{x}) - d^2, \qquad (1.42)$$

(or its square-root, the standard deviation) of the posterior probability. Another possible measure is a $100 \times \beta\%$ Bayesian confidence interval, $\left(l(\hat{\theta}), u(\hat{\theta})\right)$, obtained from the formulae:

$$\int_{\theta \leq l(\hat{\theta})} P(\theta|\mathbf{x}) = \alpha_L \qquad (1.43)$$

and

$$\int_{\theta \geq u(\hat{\theta})} P(\theta|\mathbf{x}) = \alpha_R. \qquad (1.44)$$

The numbers α_L and α_R as chosen so that $\beta = 1 - \alpha_L - \alpha_R$, that is, so that

$$\int_{l(\hat{\theta})}^{u(\hat{\theta})} P(\theta|\mathbf{x}) = \beta. \qquad (1.45)$$

These formulae may appear to be similar to those used to compute intervals in the frequentist approach. However, the similarity is superficial and misleading because here *the intervals do not refer to an ensemble.* Of course, it is perfectly sensible, and indeed scientifically necessary, to study the average behavior of an inference method be it frequentist or Bayesian.

Such studies are sometimes referred to as *ensemble tests* in high energy physics. But, the ensemble itself has no bearing on the interpretation of the Bayesian results. The ensemble result merely informs us of the average behavior of the method with respect to that ensemble.