STATISTICAL INFERENCE: CONCEPTS AND ISSUES †

Harrison B. Prosper

Department of Physics, Florida State University, Tallahassee, Florida 32306, USA

(Dated: February 10, 2004)

Abstract

I review the key concepts of frequentist and Bayesian statistics. Although the focus is on concepts, I have given enough detail in the hope that these notes might serve as a useful, succinct, introduction.

†Adapted from lectures given at the 1996 NATO ASI on Techniques and Concepts of High Energy Physics, St. Croix, USVI
# CONTENTS

## I  INTRODUCTION

A  Basic Notions ..............................................  7
   1  Loss and Risk .......................................... 10
   2  Risk Minimization ..................................... 11

## II  PROBABILITY THEORY  

A  Deductive Reasoning ........................................ 13
B  Inductive Reasoning ....................................... 15
C  The Basic Rules .......................................... 16
D  Relative Frequency Interpretation ...................... 18
E  Plausibility Interpretation ............................. 19
F  Bayes’ Theorem ........................................... 20
   1  Proof .................................................. 21
G  Moments, Correlation and Independence ............... 22
   1  Moments .............................................. 22
   2  Correlation .......................................... 23
   3  Independence .......................................... 23

## III  FREQUENTIST STATISTICS

A  Sampling Distribution, Bias and Variance ............ 25
   1  Discussion ........................................... 26
B  Efficiency, Likelihood and the Minimum Variance Bound 27
C  Maximum Likelihood and Consistency .................... 28
D  The Sampling Distribution of an Estimator ............ 30
E  Example: Estimating the Mean Lifetime ................. 32
F  Combining Results ....................................... 34
G  Confidence Intervals: Exact ............................. 37
1 Introduction .................................................. 37
2 Computing Intervals ........................................ 40
3 Example: Exact intervals for the Poisson distribution. .......... 41
4 Example: Exact intervals for the exponential distribution. ....... 43
5 Example: Exact intervals for the Gaussian distribution .......... 48
H Confidence Intervals: The Unified Approach .................. 49
  1 Example: KARMEN results .................................. 50
  2 Discussion .................................................. 51
I Confidence Intervals: Approximate ............................. 51
  1 Discussion ................................................. 53
J Hypothesis Testing ............................................ 54
  1 According to Neyman and Pearson ........................... 54
  2 According to Fisher ...................................... 57

IV BAYESIAN STATISTICS ........................................... 59
A The Bayesian Method ......................................... 60
B Continuous Sets of Propositions ................................ 60
C The Likelihood Principle ..................................... 61
D Parameter Estimation ......................................... 63
  1 Quadratic Loss ............................................ 63
  2 Absolute Loss ............................................ 64
  3 Bounded Loss ............................................. 65
  4 Uncertainty ............................................... 65
E Combining Results ............................................ 66
F Model Selection ............................................... 67

V BAYESIAN ANALYSIS ............................................ 68
A Optimal Event Selection ...................................... 69
ACKNOWLEDGEMENTS

References
I. INTRODUCTION

“We don’t know all about the world to start with; our knowledge by experience consists simply of a rather scattered lot of sensations, and we cannot get any further without some a priori postulates. My problem is to get these stated as clearly as possible.”

— Sir Harold Jeffreys, in a letter to Fisher dated 1 March 1934

The problem of statistical inference has been addressed in two radically different ways. The first dates back to the time of the English cleric Thomas Bayes (1702–1761) and the great mathematical physicists of the 18th and 19th centuries, which include Daniel and Jacob Bernoulli, Laplace and Gauss. During the latter half of the 20th century their approach, called Bayesian probability theory, has undergone a renaissance that was spearheaded by Sir Harold Jeffreys [1] (1891–1989) in the 1930s. Jeffreys’ ideas lay dormant until they were vigorously revived by mathematicians and physicists, notably Cox, de Finetti, Lindley, Savage and Jaynes [2]. In Bayesian theory, probability is regarded as a measure of the plausibility of an assertion. For example, when a meteorologist says “There is a 40% chance of rain tomorrow” she is asserting that 40% is her measure of the plausibility that the statement about rain tomorrow is true.

Towards the middle of the 19th century, this notion of probability and its associated methods fell into disfavor[9] because of dissatisfaction with what some writers saw as the arbitrary nature with which certain probabilities were assigned and the inherent subjectivity of the Bayesian mode of reasoning. To correct these alleged defects a different approach to inference was developed at the start of the 20th century, referred to variously as classical, orthodox or frequentist statistics. The Bayesian view was all but abandoned. The newer approach comprises the body of statistical ideas with which most physicists are familiar. Many of the ideas were developed by Sir Ronald Aylmer Fisher [4] (1890–1962). Significant contributions were made by Neyman, Pearson, Cramer, Rao, von Mises and Kolmogorov, to name but a few [4, 5].
It is common practice to view these thinkers as forming a single frequentist school. However, they did not all share the same point of view. There were, for example, severe disagreements between Jerzy Neyman and Ronald Fisher, the two principal architects of the frequentist theory. Neyman, for example, thought it necessary to specify alternative hypotheses when testing a given hypothesis. Fisher argued that an hypothesis can be tested without reference to alternatives and that moreover this was a useful thing to do. Neyman was a strict frequentist, whereas Fisher was rather more flexible and championed the use of the likelihood function as a tool of inference.

The frequentists did, however, agree (more or less) on one thing: probability is to interpreted not as a measure of plausibility or degree of belief but rather as the relative frequency with which something happens, or will happen [36]. From this point of view, statements such as “There is a 40% chance of rain tomorrow” made no sense because it is not possible to repeat that particular day and count how often it rained. However, the statement “There is a 40% chance of rain on days named July 14th” is meaningful because such days happen often and we can, therefore, assess the relative frequency with which it rains on days so named.

Fisher was a great geneticist and statistician who made numerous fundamental contributions to the subject of statistical inference [4]. Nonetheless, it is ironic that physicists have embraced his views, along with those of Neyman, rather than those of the mathematical physicists Laplace and Jeffreys. One goal of these notes is to provide a succinct introduction to the Bayes-Laplace-Jeffreys theory and to show how it can be used to conduct effective analyses in physics. However, since the frequentist view predominates it is important to have a clear understanding of that view, if only to appreciate more clearly how it differs from the Bayes-Laplace-Jeffreys theory, that is, the Bayesian theory [2, 6]. The notes therefore are divided roughly equally between frequentist and Bayesian ideas. After describing some general concepts I give a brief introduction to probability theory from a perspective that encompasses, indeed, is biased towards, the Bayesian viewpoint. I then cover frequentist statistics. The last section provides an introduction to Bayesian theory and its application.
The chief purpose of these notes is to give clear explanations of key statistical concepts along with what, I hope, is illuminating discussion. At the end of these notes there is a bibliography of books and articles that you might find helpful.

A Basic Notions

“Thou shalt not sit with statisticians nor commit a social science.”

— W. H. Auden

One goal of science is to infer facts about the world through a combination of observation, theory and experiment. The facts about the world are usually expressed as the parameters of a mathematical model. A model may have one parameter, or many parameters. But, even when a model has many parameters, our interest is generally focused on only one or two of them, for example the mass of the top quark. The mean rate of background events or the jet energy scale in our experiment may be a crucial part of the model but they are not usually thought to be of deep scientific interest. Sometimes, of course, the model is largely empirical and the parameters may have no physical significance. This is sometimes the case in the fitting of a function to data. In some circumstances, we may know the value of the parameters and it is the model we seek. Statistical inference includes these cases also.

A scientific investigation typically begins with models and direct knowledge of data $\mathbf{x} = (x_1, \ldots, x_N)$, which we have collected to help us answer two questions: 1) which model best describes the data and 2) what are the values of its parameters. The first task is called hypothesis testing or model selection; the second is called parameter estimation. To execute these tasks sensibly we need some self-consistent rules, which should then be applied consistently. Unfortunately, the choice of rules has occasioned much controversy since the time of Thomas Bayes. Everyone, it seems, has his or her idea about what constitutes a reasonable set of rules upon which to build a theory of inference. What is considered reasonable by one person may be considered contrived or absurd by another. The difficulty
is that we cannot appeal to Nature to ascertain who is right. It is precisely this aspect of statistical inference that causes controversy.

One of the rules, which happily everyone accepts, is that the connection between the parameters \( \theta \) of the model under consideration and data \( x \) is a probability, \( P(x|\theta) \). That symbol is to be read as “the probability of \( x \) given \( \theta \)”. We shall have more to say about probabilities in the next section. In general, the probability can be written as

\[
P(x|\theta) = \sum_{z \in \Omega} f(z|\theta),
\]

where the sum is over a neighborhood \( \Omega \) of the point \( x \). If the point \( x \) is a member of a continuous set the sum becomes an integral. If we choose sufficiently small neighborhoods we can write

\[
P(x|\theta) = f(x|\theta) dx,
\]

if \( x \) forms a continuous set, or

\[
P(x|\theta) = f(x|\theta),
\]

if \( x \) is discrete. If \( x \) is continuous the function \( f(x|\theta) \) is called a probability density function (pdf); if it is discrete it is called a probability mass function. The number \( P(x|\theta) \) is the probability that we assign to our data, assuming a particular model. In going from the parameters of a model to a prediction, about the set of all possible data we could in principle observe, we are performing deductive reasoning. Statistical inference is an attempt to solve the inverse problem: we have data and we wish to infer something about the correctness or otherwise of the model and the values of its parameters. When we do statistics we are engaged in inductive reasoning.

Before proceeding, we need to settle a few notational conventions. Statisticians make a distinction between a random variable and its value. A random variable is a map \( X \),

\[
X : \Omega \rightarrow \mathbb{R},
\]

between a set of possible events or outcomes \( \Omega = (\omega_1, \ldots, \omega_N) \) and the set of reals \( \mathbb{R} \). The map \( X \) assigns a real number \( x = X(\omega) \), called the value of the random variable, to
every outcome $\omega \in \Omega$. The height of persons who pass you in the street is an example of a
random variable. Its possible events are the people who can pass you and its value is the
height of a person. Since the outcome is random so too is the value of the random variable.
Note, however, that in spite of the name the map $X$ itself is generally not random! Rather
it is the set $\Omega$ of possible outcomes that possesses the (rather mysterious) quality called
randomness. One can think of that property as a manifestation of a randomizer whose
job it is to pick an outcome from the set of possibilities, according to a rule that is not readily
discernable. The randomizer, however, need not be governed by chance. Consider the set
of possible outcomes $\Omega = \{0, \ldots, 9\}$ and the function $X$ that maps this set to the subset
$\{0, \ldots, 9\} \in \mathbb{R}$. Their exists a random variable $X$ whose value is the next decimal digit of
$\pi$, starting, say, from the first. The digits of $\pi$ do not occur by chance even though they
form an excellent random sequence. The same is true of, so-called, pseudo-random number
generators, which provide sufficiently random sequences of real numbers—indispensable in
Monte Carlo-based calculations, even though, again, the randomizer is not governed by
chance. Usually, a random variable is denoted by an upper case symbol, while one of its
values is denoted by the corresponding lower case symbol. If $X$ is a random variable then
$x$ denotes one of its values. I shall try to adhere to this convention unless doing so becomes
burdensome.

Given a sample of data $\mathbf{x} = (x_1, \ldots, x_N)$, or to use the term we often use in our field—
data-set, a mathematical model $M$ and the associated probability $P(\mathbf{x}|\theta)$ we use statistical
analysis to decide the best values to assign to the parameters $\theta$. If we have several models
$M_1, M_2 \ldots$ then we will, in addition, want to decide which one is best. This, of course,
presupposes that we know what we mean by best. The mapping $(x_1, \ldots, x_N) \rightarrow (\theta_1, \ldots, \theta_M)$
from our data-set to the parameters, or to the set of models, is called a decision function,
which will be denoted by the symbol $d$. For now, let us limit our discussion to models that
depend upon a single parameter $\theta$. Denote by $\hat{\theta}$ any estimate of $\theta$. If the decision function
is such that $\hat{\theta} = d(\mathbf{x})$ then the function $d$ is called an estimator for $\theta$. One can think of
the estimator as a program, which when data are entered into it outputs estimates. The
estimator could be as simple as an averaging operation or as complex as several full-scale Monte Carlo simulations followed by heated discussions about what to publish.

1 Loss and Risk

Some decisions, of course, are better than others. For example, suppose you are outside on a sunny day and someone asks you for the time. Unfortunately, you left your GPS-enabled watch at home, but observe that the sun is high in the sky. That is your 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 surely safe, but not too useful perhaps. On the other hand the second one is likely to be true only approximately. So which is the better estimate?

To choose a decision function we need a way to quantify the quality of the associated decisions. Bad decisions entail loss. We therefore introduce a loss function, $L(\theta, d)$, to measure the loss arising from a decision. The loss function depends, in general, on both the decision function and the parameter.

The idea of a loss function is useful in both frequentist and Bayesian statistics. However, the two approaches use the loss function differently because of the following profound conceptual difference:

- **Frequentist:** In making inferences, data we could have observed are as relevant as data observed.

- **Bayesian:** In making inferences, only data observed are relevant.

Accordingly, in the frequentist approach we consider the loss pertaining to every data-set which could have been observed, as well as the loss pertaining to the data actually obtained. In the Bayesian theory, it is asserted that only the observed data are relevant in making an
inference. But because we not know which hypothesis about the parameter is correct all possible hypotheses must be considered.

In either case, the desire to average the loss function in some way motivates the definition of a new function

\[ R = \mathbb{E}[\mathcal{L}(\theta, d)]_*, \]  

(I.5)
called the **risk function**, where the symbol \( \mathbb{E}[\cdot]_* \) denotes the averaging or **expectation operator**. This operator is linear, that is, it obeys the rule

\[ \mathbb{E}[aX + bX] = a\mathbb{E}[X] + b\mathbb{E}[X]. \]  

(I.6)

In the frequentist theory, the risk function is an ordinary function of the parameter \( \theta \) and a functional of the decision function \( d \); that is, it depends on the set of all possible values of \( d \). In the Bayesian theory, the risk function is a functional of \( \theta \). However, it is generally not regarded as a function of \( x \) because the data are considered to be constant. In one case, the averaging is done with respect to all possible data-sets \( x \) for fixed \( \theta \) (frequentist), in the other with respect to all possible \( \theta \) for fixed \( x \) (Bayesian).

It should not be construed from the above that Bayesians do not care about data-sets that could have been observed. On the contrary, it is absolutely necessary during the *design* of an experiment, or an analysis, to consider what could be observed in order to design the best possible experiment or the most effective analysis. Bayesians merely assert that when the time comes to make an inference only the data actually acquired are relevant for making the inference.

### 2 Risk Minimization

A statistical analysis can be viewed as a procedure that minimizes *some* risk function in order to arrive at an optimal decision, usually an optimal decision about the value of a parameter or a model. If the risk function is known, then, in principle, an optimal decision can be had with respect to the underlying loss function. However, in many circumstances
although the loss function is known, since we choose it, the risk function is not. In these cases, we must make do with an estimate of the risk function, the most common of which is given by

$$R_{\text{emp}} = \frac{1}{n} \sum_{i=1}^{n} L(\theta, f(x_i, \omega)),$$

(1.7)

where $f(x_i, \omega)$ is a suitably parameterized function, with parameters $\omega$ and data $x_i$, that one hopes is flexible enough to include a good approximation to the optimal decision function $d$, say at the point $\omega = \omega_0$. The function $R_{\text{emp}}$ is called the empirical risk function. Its minimization, to obtain an approximation to the optimal decision function $d$, is a widely used strategy in data analysis, encompassing everything from curve-fitting to the training of sophisticated learning machines [7]. The strategy is referred to as empirical risk minimization.

The most important mathematical property of empirical risk, and the property that makes it useful in practice, is that the function $f(x_i, \omega_0)$, found by minimizing the empirical risk, is expected to converge to the optimal decision function $d$ as the sample size $n$ goes to infinity, provided that the function $f(x_i, \omega)$ is sufficiently flexible and the minimization algorithm is effective at finding the minimum. Much of the research on algorithms is devoted to finding efficient ways to minimize broad classes of empirical risk functions. We give an example later.
II. PROBABILITY THEORY

“Probability theory is nothing but common sense reduced to calculation.”

— Laplace, 1819

Probability theory, according to Laplace, is human reasoning abstracted. Not everyone agrees with him. In Sect. III we shall learn about a theory of statistical inference that was invented in the early part of the 20th century, by Fisher, Neyman, Pearson and others. They took issue with Laplace arguing that the demands of science required a statistical theory that was wholly objective; one that let the data “speak for themselves”. However, it is far from clear that such a theory exists. Unlike a physical theory, such as the Standard Model — whose ultimate arbiter is Nature, the arbiters of theories of inference are scientists who have been unable to agree on foundations, even after arguing for more than two hundred years.

For the reader who may not have discerned my preference let me say that it is for the Bayesian theory. I prefer that theory because it is elegant, well-founded and does what one would wish a theory of inference to do. Therefore, my cursory look at probability theory is slanted towards the Bayesian viewpoint. For those unmoved by mere beauty, however, here are two other good reasons for the Bayesian bias: 1) the Bayesian viewpoint includes the frequentist point of view as a special case and 2) it yields a theory of probability that is much more broadly applicable than one based on probability interpreted as a relative frequency.

Probability theory is a vast and complex subject. It is impossible to do justice to the theory in these brief notes. My aim is simply to give you a feel for what it is all about.

A Deductive Reasoning

Around 350 BC, Aristotle — described by his teacher Plato as the Nous (intelligence personified) of his Academy — noticed that when we reasoned well we did so according to definite rules all of which can be reduced to the 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

In addition, if $A$ is TRUE then its negation, written as $\overline{A}$, is FALSE. We say that $A$ contradicts $\overline{A}$. A useful mnemonic for the syllogisms are the set of symbolic expressions:

<table>
<thead>
<tr>
<th>modus ponens</th>
<th>modus tollens</th>
</tr>
</thead>
</table>
| Major premise | $AB = A$ | $AB = A$
| Minor premise | $A = 1$ | $B = 0$
| Conclusion | $B = 1$ | $A = 0$

The symbols $A$, $B$, 1, 0, and their negations, $\overline{A}$, $\overline{B}$, 1 and 0, are called propositions. The proposition 1 has a truth value of TRUE; 0 has a truth value of FALSE.

Here is an example. Let $A = \text{You hunt squirrels}$ and let $B = \text{You need an assault rifle}$.

Our major premise is: If $\text{You hunt squirrels}$ then $\text{You need an assault rifle}$. Let our minor premise be $\text{You hunt squirrels}$ is TRUE. Our logical conclusion is: therefore, $\text{You need an assault rifle}$ is TRUE. Note, however, that if $B$ is TRUE it does not follow that $A$ is TRUE: If $\text{You need an assault rifle}$ is TRUE it does not follow that $\text{You hunt squirrels}$ is TRUE. You may hunt squirrels, but it may be that $\text{You need an assault rifle}$ because you are a member of the armed forces on a combat mission. Conversely, if $A$ is FALSE we cannot conclude that $B$ is FALSE. But, if $\text{You need an assault rifle}$ is FALSE then we conclude that $\text{You hunt squirrels}$ is FALSE.

The simplest way to see how the syllogisms work 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 are non the wiser about the truth value of $A$; likewise, if $A = 0$ then the truth value of $B$ is left unknown. This is as far as we can get with purely deductive
reasoning. We say a bit more about these rules below.

\section*{B Inductive Reasoning}

Deductive reasoning can only get us so far. Most of the time we are not well enough informed to reason deductively. Alas, most of the time we do not even reason; witness some of the deliberations in the halls of Congress! In the example given above few would doubt that if $B$ is false then $A$ is false. But what if $B$ is true? What if it is true that \textit{You need an assault rifle}? There is always the possibility that one could be wrong, but a good many people would harbor the suspicion that if the proposition $B$, \textit{You need an assault rifle}, is true this renders the proposition $A$, \textit{You hunt squirrels}, more likely. Now suppose that I were given some new data, such as: you are in an active combat zone; that would surely color my conclusion differently than if I had been told that you are a stockbroker walking her dog in Batavia, Illinois. Bayes and Laplace saw this mode of reasoning as the essence of common sense.

Many compelling arguments lead to the conclusion that common sense reasoning, called \textbf{plausible reasoning} by Jaynes, follows definite rules. One can view the Bayes-Laplace theory and its subsequent developments by Sir Harold Jeffreys, R. T. Cox, E. T. Jaynes and others as an extension of logic to include truth values that lie between \textit{false} and \textit{true}. One makes the idealization that truth values can be represented by real numbers in the interval $[0,1]$. It can then be shown that these numbers satisfy the axioms of probability and, as such, are a quantitative measure of the plausibility of propositions. This theory assigns a quantitative meaning to the weaker 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 more likely | Therefore, $B$ is less likely.
C The Basic Rules

Probability theory, viewed as extended logic, is based upon two sets of rules. The first is the algebra of propositions, called **Boolean algebra** after George Boole (1854). If \( A, B, C, 1, 0 \) and their negations are propositions, and \(+\) and \(\cdot\) are binary operations then:

\[
\begin{align*}
A + 0 &= A \\
A + \overline{A} &= 1 \\
A \cdot 1 &= A \\
A \cdot \overline{A} &= 0
\end{align*}
\]

**Commutativity law**

\[
A \cdot B = B \cdot A \\
A + B = B + A
\]

**Distributivity law**

\[
\begin{align*}
A \cdot (B + C) &= A \cdot B + A \cdot C \\
A + B \cdot C &= (A + B) \cdot (A + C)
\end{align*}
\]

The above are called the **Huntington axioms**. Notice the duality between the operations “\(\cdot\)” and “\(+\)”. To simplify the notation we usually do not write the “\(\cdot\)” operation explicitly.

From these rules it is possible to derive thousands of theorems; here are a few:

\[
\begin{align*}
A + 1 &= 1 \\
\overline{0} &= 1 \\
A + AB &= A \\
A(A + B) &= A \\
A A &= A \\
A + A &= A \\
A(BC) &= (AB)C \\
A + (B + C) &= (A + B) + C \\
\overline{AB} &= \overline{A} + \overline{B} \\
\overline{A + B} &= \overline{A} \overline{B}
\end{align*}
\]

See if you can prove them.

The second set of rules are the **product rule**

\[
P(AB|C) = P(B|AC)P(A|C),
\]

\[
= P(A|BC)P(B|C), \tag{II.2}
\]

the **sum rule**

\[
P(A|B) + P(\overline{A}|B) = 1, \tag{II.3}
\]
and the conventional numerical assignments for certainty and its converse

\[ P(1|B) = 1, \]
\[ P(0|B) = 0. \] (II.5)

From just the two sets of rules, plus the rules of arithmetic, an impressive mathematical theory has been built.

To illustrate the use of the rules we shall prove a theorem. We want to relate \( P(A + B|C) \) to \( P(A|C) \) and \( P(B|C) \). All we need do is apply the above rules repeatedly in more or less the right order. As you “walk through” the proof, check off the rule that is being used:

\[
P(A + B|C) = 1 - P(A + B|C)
\]
\[
= 1 - P(A|C)
\]
\[
= 1 - P(B|AC)P(A|C)
\]
\[
= 1 - [1 - P(B|AC)] P(A|C)
\]
\[
= 1 - P(A|C) + P(B|AC)P(A|C)
\]
\[
= P(A|C) + P(B|AC)P(A|C)
\]
\[
= P(A|C) + P(AB|C)
\]
\[
= P(A|C) + [1 - P(A|BC)] P(B|C)
\]
\[
= P(A|C) + P(B|C) - P(A|BC)P(B|C)
\]
\[
P(A + B|C) = P(A|C) + P(B|C) - P(AB|C). \] (II.7)

Most of us would probably agree that the Huntington axioms are reasonably intuitive. What of the product and sum rules, Eqs. (II.2) and (II.3)? These seem less intuitively obvious. After the geneticist onslaught of the 1930s, it was time for physicists to go on the offensive. An important advance occurred in 1946 when, fittingly, a physicist, R. T. Cox, published a beautiful article [8] in which he derived the product and sum rules from more primitive axioms:
• **Axiom 1)** Plausibilities can be represented by real numbers.

• **Axiom 2)** The plausibility of a proposition and the plausibility of another proposition given the first determine the plausibility of both propositions taken together.

• **Axiom 3)** The plausibility of a proposition determines the plausibility of its converse.

Cox showed that plausibilities or degrees of belief follow rules that are isomorphic to those of probability and thus provide a **subjective interpretation** of the latter. Well before the advent of Cox’s theorem, Jacob Bernoulli, who, along with his contemporaries, regarded the subjective interpretation of probability as self-evidently sensible [9], proved a theorem that provides the link between relative frequency and probability.

### D  Relative Frequency Interpretation

In the **objective interpretation**, probability is interpreted as the relative frequency with which something happens, or could happen. Let $n$ be the number of experiments or trials; for example, this could be the number of proton-antiproton collisions at the Tevatron. Let $k$ be the number of successes; for example, the events of interest, such as $t\bar{t}$ events. The relative frequency of successes is

$$\frac{k}{n}.$$  \hspace{1cm} (II.8)

Our expectation, borne of experience, is that as $n \to \infty$ the relative frequency $k/n$ converges to a number, call it $p$, which we interpret as the probability of a success. Several attempts have been made to build a theory of probability in which the latter is defined as the limit of the relative frequency $k/n$. None has succeeded.

Any such theory of probability must deal with the following annoying possibility. It is possible, though exceedingly unlikely, that on every trial we get a success, or a failure, or we alternate between the two *ad infinitum*. We, therefore, have to be careful about what we mean by the limit of the (rational) number $k/n$. The correct statement, first noted by
Jacob Bernoulli (1703), is the **law of large numbers**, a version of which states that

\[
\lim_{n \to \infty} P\left[ \left| \frac{k}{n} - p \right| > \epsilon \right] = 0,
\]

for any real number \( \epsilon > 0 \). That is, the *probability* that the fraction \( k/n \) differs from the *probability* \( p \) by more than \( \epsilon \) goes to zero, as the number of trials goes to infinity. The previous sentence contains a potentially recursive definition of probability. While there is nothing wrong, in principle, with recursive definitions, in this case, the implied recursion is conceptually problematic. If probability is a relative frequency and nothing more, then both probabilities that occur in the statement of Bernoulli’s theorem perforce are relative frequencies. The second probability in that sentence is loosely the “limit” of the relative frequency \( k/n \); loose because the limit exists only in a probabilistic sense. What of the first probability? If we take the implied recursion seriously then the first probability is also loosely the limit of a relative frequency. But, alas, we shall then be ensnared in an infinite hierarchy of infinite sets of trials.

### E Plausibility Interpretation

Perhaps, one day, a mathematician with a Cantorian sensibility will be able to make sense of the infinite hierarchy of infinite trials. But until that day arrives we must find a way to cut the Gordian Knot. We can avoid the infinite regression by interpreting the first probability differently from the second: if we interpret the first as a measure of *plausibility* then the theorem is a statement about the plausibility of the proposition \( \lim_{n \to \infty} k/n = p \). Bernoulli’s theorem, as he himself interpreted it, declares that it is plausible to the point of certainty that \( k/n \to p \) as the number of trials grows without limit. The import of this theorem, and Bernoulli’s interpretation of it, is that probability as relative frequency is a *derived* notion pertaining to a special class of circumstances, namely, those in which one can entertain, in principle, performing *identically* repeated trials in which the relative frequency converges, probabilistically, to some number \( p \), which, because it satisfies the axioms of
probability, we are at liberty to call a probability. The Standard Model is an example of a physical theory that can predict the limiting numbers $p$ for the kind of identically repeated trials performed in high energy physics experiments.

The position advocated here is that probability is an abstraction that can usefully be interpreted in at least two different ways: relative frequency and degree of belief. Moreover, the first is best understood in terms of the second.

## F Bayes’ Theorem

In 1763, Thomas Bayes published a paper [37] in which he gave a special case of a theorem that bears his name. Bayes’ theorem

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

(II.10)

where $A$, $B_k$ and $C$ are propositions is a direct consequence of the product rule, Eq. (II.2), of probability theory. Consider two propositions $A$ and $B$. They 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 the theorem we proved earlier, we get

$$P(A + B|C) = P(A|C) + P(B|C),$$

(II.11)

which is easily 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:

$$\sum_k P(B_k|C) = 1.$$  

(II.12)

If $B_kD_j$ are a set of mutually exclusive and exhaustive **joint propositions**, then we can write Bayes’ theorem as

$$P(B_kD_j|AC) = \frac{P(A|B_kD_jC)P(B_kD_j|C)}{\sum_{i,l} P(A|B_iD_lC)P(B_iD_l|C)}.$$  

(II.13)
1 Proof

Start by replacing $B$, in the product rule Eq. (II.2), by the joint proposition $B_kD_j$

$$P(A|B_kD_jC)P(B_kD_j|C) = P(B_kD_j|AC)P(A|C),$$  \hspace{1cm} (II.14)

and then sum Eq. (II.14) over the set of joint propositions:

$$\sum_{k,j} P(A|B_kD_jC)P(B_kD_j|C) = P(A|C)\sum_{k,j} P(B_kD_j|AC)$$

$$= P(A|C).$$  \hspace{1cm} (II.16)

In the last step we have assumed that the joint propositions $B_kD_j$ form an exhaustive set. Now go back to the product rule Eq. (II.14), substitute in the formula for $P(A|C)$ and rearrange to get Bayes’ theorem, Eq. (II.13). At this point let us note that, to reduce Bayes’ theorem to a statement about the proposition $B_k$ only we need merely sum Eq. (II.13) over the propositions $D_j$. Bayes’ theorem can be extended easily to encompass larger and larger sets of joint propositions.

Here is a simple example of the use of Bayes’ theorem. You are being hounded by a pack of dogs of which 25% are rabid. You have a good nose for rabid dogs, so if a dog is in fact rabid then 80% of the time you would recognize it as such. This, if you will, is your rabid-dog identification efficiency. However, you have a 10% chance to mis-identify a dog as rabid when in fact it is not. But, in our politically correct world, you must refrain from visiting violence upon a dog if there is more than a 30% chance of harming one that is healthy, otherwise you risk a grisly encounter with the Animal Liberation Front. A dog pounces at you. Do you harm the dog? Let us do a small calculation. We summarize the probabilities as follows

$$P(Ok|C) = 0.75,$$

$$P(Bad|C) = 0.25,$$

$$P(rabid|Bad \cdot C) = 0.80,$$

$$P(rabid|Ok \cdot C) = 0.10,$$  \hspace{1cm} (II.18)
where $C$ could be some proposition like $C = \text{The dog is a pit-bull terrier}$. So we read $\mathbb{P}(\text{Bad}|C)$ as the probability that the dog is bad, that is, rabid, given that it is a pit-bull terrier; $\mathbb{P}(\text{rabid}|\text{Ok} \cdot C)$ is the probability to classify as rabid a dog that is, in fact, healthy. Applying Bayes’ theorem we get

$$
\mathbb{P}(\text{Ok}|\text{rabid}) = \frac{\mathbb{P}(\text{rabid}|\text{Ok} \cdot C)\mathbb{P}(\text{Ok}|C)}{\mathbb{P}(\text{rabid}|\text{Ok} \cdot C)\mathbb{P}(\text{Ok}|C) + \mathbb{P}(\text{rabid}|\text{Bad} \cdot C)\mathbb{P}(\text{Bad}|C)},
$$

$$
= \frac{0.1 \times 0.75}{0.1 \times 0.75 + 0.80 \times 0.25},
$$

$$
= 0.27. \quad (\text{II.20})
$$

We conclude that the chance that a dog, having been classified as rabid, is actually healthy is 27%. Being less than the cut of 30%, you confront the dog secure in the knowledge that the issue has been approached with due rigor!

I’m sure the reader can think of a less silly example from particle physics.

**G Moments, Correlation and Independence**

1 Moments

Given a random variable $X$, the $r$-moment, $m_r(a)$, about the point $a$ is defined by

$$
m_r(a) \equiv \mathbb{E}[(X - a)^r], \quad (\text{II.22})
$$

$$
= \sum_{k=0}^{r} \binom{n}{k} m_k(0)(-a)^{r-k}, \quad (\text{II.23})
$$

where $m_k(0) = \mathbb{E}[X^k]$ is the $k$-moment about zero and where the averaging is done with respect to $X$ [38]. If we have a sample containing $n$ values $x_1, \ldots, x_n$ of $X$ we can calculate an estimate $\hat{m}_r(a)$ of $m_r(a)$ as follows

$$
\hat{m}_r(a) = \frac{1}{n} \sum_{i=1}^{n} (x_i - a)^r. \quad (\text{II.24})
$$
2 Correlation

Consider two random variables $X$ and $Y$, each with $N$ elements. One can think of these variables as column matrices. The covariance matrix of the variables is defined by

$$
C[X, Y] = E[(X - E[X])(Y - E[Y])^T], \\
= E[XY^T] - E[X]E[Y^T],
$$

(II.26)

where $^T$ is the transpose operator. The averaging is done over all possible data-sets with $N$ elements each. If we set $Y = X$ this yields the variance matrix $\mathbb{V}[X] = C[X, X]$. Note that the variance matrix is square and diagonal and the covariance matrix is square and symmetric. The two are related as follows

$$\mathbb{V}[aX + bY] = a^2\mathbb{V}[X] + b^2\mathbb{V}[Y] + 2ab C[X, Y],
$$

(II.27)

which shows (try to show it) that, unlike the expectation operator, the variance operator is non-linear. The covariance matrix has the disadvantage that it has dimensions, so we generally prefer to use a dimensionless version of it called the correlation matrix $\rho$, whose matrix elements are given by

$$
\rho_{ij} = \frac{C[X, Y]_{ij}}{\sigma_i \sigma_j},
$$

(II.28)

where $\sigma_i = \sqrt{\mathbb{V}[X_i]}$ are the standard deviations. Note that $-1 \leq \rho_{ij} \leq 1$. If a matrix element is close to 1 then we say that the quantities $X_i$ and $Y_i$ are strongly linearly correlated. If it is close to $-1$ then the quantities are said to be strongly linearly anti-correlated.

3 Independence

Consider two random variables $X$ and $Y$, associated with the probability $\mathbb{P}(X, Y|\theta)$. We can write the probability as

$$
\mathbb{P}(X, Y|\theta) = f(X|Y, \theta) f(Y|\theta) \, dX \, dY,
$$

(II.29)
where \( f(X|Y, \theta) \) is the probability density of \( X \) given \( Y \) and \( \theta \) and \( f(Y|\theta) \) is the probability density of \( Y \) given \( \theta \). If \( f(X|Y, \theta) = f(X|\theta) \) that is, the probability of \( X \) does not depend on \( Y \), then the variables \( X \) and \( Y \) are said to be statistically independent. In that case, we can write

\[
P(X, Y|\theta) = f(X|\theta) dX \ f(Y|\theta) dY, \\
= P(X|\theta) \ P(Y|\theta).
\] (II.31)

The following points should be noted.

- If variables are statistically independent then they are uncorrelated.
- If variables are uncorrelated this does not imply that they are statistically independent.

For example, suppose we define a uniform probability on the circle \( X^2 + Y^2 = a^2 \). Obviously, \( X \) and \( Y \) are not independent. Their linear correlation, however, is zero!
III. FREQUENTIST STATISTICS

A Sampling Distribution, Bias and Variance

Consider the **quadratic loss function**

\[
\mathcal{L}(\theta, d) = (\theta - d)^2,
\]

where \(\theta\) is a parameter whose value is to be estimated with the estimator \(d\). The associated risk function

\[
\mathcal{R}(\theta) = \mathbb{E}[(\theta - d)^2]\big|_x,
\]

is called the **mean squared error**. In Eq. (III.2), we are averaging with respect to all possible data-sets \(x\), of which our particular data-set \(x_0\), say, is but a singular sample. The probability to obtain a data-set \(x\) is assumed to be \(\mathbb{P}(x|\theta)\). When data are assumed to be \textit{sampled} from a \textbf{population} or \textbf{ensemble} of possible data-sets the probability \(\mathbb{P}(X|\theta)\) is called the **sampling distribution** of the random variable \(X\). If \(\text{SomeDistribution}(\theta)\) is a sampling distribution, one uses the notation

\[
X \sim \text{SomeDistribution}(\theta),
\]

to indicate that the random variable \(X\) is distributed according to \(\text{SomeDistribution}\). The sampling distribution is an important quantity in frequentist statistics.

To write Eq. (III.2) in a more useful form let us apply the expectation operator to the estimator \(d\). The expectation value of \(d(x)\), with respect to all possible data-sets \(x\), can be written as

\[
\mathbb{E}[d(x)] \equiv \theta + b(\theta),
\]

where the quantity \(b(\theta)\) is called the \textit{bias}, which, in general, is a function of the parameter \(\theta\). We can re-write Eq. (III.2) as

\[
\mathcal{R}(\theta) = \mathbb{V}[d] + b^2(\theta),
\]
where

\[ \mathbb{V}[d] = \mathbb{E}[d^2] - \mathbb{E}[d]^2. \]  

(III.6)

If the estimates \( \hat{\theta} = d(x) \) are such that \( b(\theta) = 0 \) for every possible value of \( \theta \) then the estimates \( \hat{\theta} \) and the estimator \( d \) are said to be unbiased. In that happy circumstance the mean squared error is equal to the variance. From a frequentist viewpoint, this suggests a tentative answer to the question: what is the best estimator?

- The best estimator is unbiased and has the smallest variance.

But is this really the best answer? Well, not necessarily, if by best estimator one means that estimator which minimizes the risk function, Eq. (III.5). Equation (III.5) shows that the minimum risk — assuming a quadratic loss — obtains when one has achieved the best compromise between variance and bias, which need not be, always, when the bias is zero.

1 Discussion.

Fisher [4] was a staunch advocate of frequentist methods. But even he noted, pointedly, that the mere specification of a data-set does not uniquely determine to which population, or ensemble, it belongs. Many arguments arise simply because of a failure to acknowledge the obvious: that the choice of ensemble in statistical data analysis is often just that, a choice. The reason we say that the outcome of a counting experiment, such as one to count the number of \( t\bar{t} \) events, follows a particular sampling distribution is because of what we believe we know about the characteristics of the experiment. But, unless we actually repeated the experiment, we cannot know whether or not the distribution assumed is in fact correct. Let us pursue this a little further.

Consider an experiment that has run for two years and has found 10 Higgs boson events. To which ensemble does this data-set belong? If we assert that the experiment ran for a fixed amount of time, one might argue that the data-set belongs to an ensemble characterized by experiments each of which ran for two years. If, however, some colleagues asserted that
the experiment ran until we found 10 Higgs events, then they are at liberty to choose an ensemble consisting of experiments each of which contains exactly 10 events. Or perhaps the experiment ran until funding stopped, just after the 10th Higgs event was recorded. These three ensembles are different and each will give a different result because each corresponds to a different sampling distribution. By merely changing our mind about which ensemble we think is reasonable we can change our answers. Yet the data have not changed: we have 10 Higgs events that were collected in two years with a given amount of funding.

The point being made is this: Some ensembles of course objectively exist, such as the set of Fermilab scientists, however, the ones we use in statistical analysis, typically, are not. Which ensemble we use in an analysis is determined by tradition, collective agreement or probabilistic reasoning, not by some objectively verifiable procedure.

B Efficiency, Likelihood and the Minimum Variance Bound

Suppose we are lucky enough to be measuring a parameter for which a best estimator exists. How would we recognize this happy state? First some definitions. The **efficiency** of an estimator \( d \) is defined by

\[
e(\theta) \equiv \frac{\text{Var}[d^*]}{\text{Var}[d]}, \tag{III.7}
\]

where \( d^* \) is the best estimator, that is, it is unbiased and has the smallest variance. The **likelihood function** \( L(\theta) \), as defined by Fisher, is the probability density function \( f(x|\theta) \) evaluated at the observed value \( x = x_0 \), that is, \( L(\theta) \equiv f(x_0|\theta) \). Since \( x_0 \) represents the data actually observed, which are constant, the likelihood function is a function of \( \theta \) only. However, the reader is warned that other authors use the term likelihood as a synonym for probability density function, in which case the likelihood is a function of both \( \theta \) and \( x \).

An important theorem was proved by Aitken, Silverstone, Fréchet, Cramér and Rao. Sadly for the others it is usually called the **Cramér-Rao inequality**; a non-partisan name
is the **minimum variance bound**. The theorem states that

\[ \mathcal{V}[d] \geq \frac{(1 + D_\theta b(\theta))^2}{\mathbb{E}[(D_\theta \ln f(X|\theta))^2]}, \]  

(III.8)

with the corollary that

\[ \mathcal{V}[d^*] = \frac{1}{\mathbb{E}[(D_\theta \ln f(X|\theta))^2]}, \]  

(III.9)

is the variance of the best estimator for \( \theta \), if it exists. The symbol \( D_\theta \) is the derivative operator with respect to \( \theta \). The main 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, be independent of the parameter \( \theta \). Given an estimator \( d \) for \( \theta \), one can determine its efficiency and thereby discover if it is the best frequentist estimator for that parameter.

### C Maximum Likelihood and Consistency

Amongst the many ideas introduced by Fisher his method of **maximum likelihood** is justly famous [39]. It provides a general and systematic way to construct frequentist estimators. The principle is:

- Choose as an estimate \( \hat{\theta} \) the value of \( \theta \), which will be equal to some function \( d(x) \), that maximizes the likelihood function \( L(\theta) \). That is, solve \( D_\theta L(\theta) = 0 \), for \( \hat{\theta} = d(x) \).

Like all good mathematical principles this one boasts many important theorems of which one of the most important is

- If a best, that is, **maximally efficient unbiased**, estimator exists for a parameter \( \theta \) then the maximum likelihood principle will find it.

Bias, or a lack thereof, is an example of a statistical property about which opinions differ. Some consider bias to be “bad”, while others look on bemused. Again, we cannot appeal to Nature to decide which opinion is “correct”, if any. But here are some points to ponder. In high energy physics, we usually do not perform repeated sampling with a fixed sample size,
thereby generating an actual ensemble of experimental results. Suppose we have obtained a sample of 10 Higgs boson events. In high energy physics, it would be very unusual to repeat the experiment to obtain many samples of size 10 in order to assess the bias in a sample of that size. What is more common is to append a new data-set to an older one, or to combine a new result with an older. If subsequent to our 10 Higgs boson events we obtain 15 more, the new sample would be combined with the previous one to obtain what one hopes is a more precise estimate, say, of the Higgs boson mass.

We can, of course, perform virtual experiments on a computer. But the bias we would calculate pertains to the ensemble of virtual experiments, not to the single experiment we have actually performed. This is because when we speak of a biased, or unbiased, estimate we are making a statement about the ensemble to which the estimate is presumed to belong; in particular, it is a statement about the mean of the ensemble of estimates. The difference \( \epsilon = \hat{\theta} - \theta \) between an estimate \( \hat{\theta} \) and the parameter \( \theta \) is not the bias, but the error in the estimate, which is something, alas, we can never know — except, of course, when the experiments are virtual. The bias is, by definition, the mean error, \( b(\theta) \equiv E[\epsilon] \).

Reasonable people disagree about whether or not an estimate should be corrected so that the bias in the ensemble of (virtual) experiments to which the corrected estimate is presumed to belong is zero. For example, bias it is something that the Particle Data Group [10] does consider when they combine the results from different experiments, a point that we shall return to.

Although opinion differs about the relevance of bias and bias correction, there is one rule that everyone agrees is sensible: an estimator should be consistent. I find it ironic, therefore, that while we often fret about bias, and spend much time correcting for it, we seldom check whether or not our estimators are consistent! An estimator is said to be consistent if, as the size of the data-set grows, the estimates derived from the same converge to the true value of the parameter being estimated. This implies that the bias \( b(\theta) \) of a consistent estimator vanishes in the limit of an infinitely large data-set. Maximum likelihood estimators have the virtue that they are usually consistent. For large samples, the distribution of the
estimator \( d \) has a mean that gets ever closer to \( \theta \) and a variance that attains that prescribed by the minimum variance bound. Consistency is a very desirable property and one would not knowingly use an inconsistent estimator. After all, it is surely desirable that as more and more data are acquired we are able to report better and better results.

Another useful property of estimates obtained from the maximum likelihood principle is their invariance with respect to re-parametrization. What this means is this: If we had found \( \hat{\theta} \) by solving
\[
D_\theta L(\theta) = 0, \tag{III.10}
\]
or what is usually much more convenient
\[
D_\theta \ln L(\theta) = 0, \tag{III.11}
\]
and then solved the above equation using instead the new parameter \( \alpha = g(\theta) \) then we would find that \( \hat{\alpha} = g(\hat{\theta}) \). (Why this is true?) Therefore, one might as well solve the likelihood equation, Eq. (III.11), using the most convenient parameter and, at the end, transform to the parameter of interest.

All is not roses, however. For a given likelihood function, if the estimates for one parameter are unbiased those of any function thereof will almost certainly be biased! The easiest way to see this is to do a Taylor series expansion of \( \hat{\alpha} = g(\hat{\theta}) \) about the parameter \( \theta \):
\[
\hat{\alpha} \approx g(\theta) + (\hat{\theta} - \theta) D_\theta g(\theta) + \frac{1}{2} (\hat{\theta} - \theta)^2 D^2_\theta g(\theta), \tag{III.12}
\]
and then apply the expectation operator to both sides, where the averaging is done with respect to the values \( x \), of the random variable \( X \). If \( \hat{\theta} \) is unbiased, we get
\[
E[\hat{\alpha}] \approx \alpha + \frac{1}{2} \nu[\hat{\theta}] D^2_\theta g(\theta), \tag{III.13}
\]
which shows that the estimate \( \hat{\alpha} \) is, in general, biased even if \( \hat{\theta} \) is not.

### D The Sampling Distribution of an Estimator

As noted above the probability \( P(X|\theta) = f(X|\theta) \, dX \) is interpreted as the sampling distribution of the random variable \( X \). By maximizing the probability density function \( f(X|\theta) \),
that is, by finding its mode, we obtain an estimator $d$ for the parameter $\theta$, which yields an estimate $\hat{\theta} = d(x)$.

Let me address, again, Fisher’s definition of likelihoods. Fisher defined the likelihood to be a function, of the parameters of the probability density, that is proportional to the probability density function evaluated at the observed data. Since the observed data are constant the likelihood is a function of the parameters only. Unfortunately, if likelihood is defined in this way then one cannot speak of an estimator derived from a likelihood function, since an estimator is usually regarded as a function of the data. If we wish to speak of an estimator, yet adhere to Fisher’s notion of likelihood, then we must derive the estimator $d$ not from the likelihood function $L(\theta)$ but from the probability density function $f(x|\theta)$ from which it is derived. That is, estimates are derived from likelihoods, while estimators are derived from probability densities, via (somewhat confusingly) the principle of maximum likelihood.

A critic might suggest that the preceding discussion is so much useless pedantry. I admit to being sympathetic to that point of view. I tend to find discussions less tortuous if $L(\theta)$ is just another name for $f(x|\theta)$, in which estimators and estimates can be derived from likelihoods.

In the frequentist theory, the sampling distribution of an estimator is the basis for quantifying uncertainty. Uncertainty, from a frequentist viewpoint, is a statement about the typical size of the (unknowable) error $\epsilon = \hat{\theta} - \theta$, over some ensemble. In high energy physics we sometimes refer to this as the ensemble error. One consequence of this view of uncertainty is that we forego any inferential benefit that might be had from a data-set that happens to be tightly clustered. On the other hand, we would not suffer from a data-set that, by chance, happened to be more spread out than average.

As we shall see later, in the Bayesian theory the sampling distribution of estimators does not play an inferential role. It is used mainly to investigate how Bayesian results are likely to behave on the average, in what we physicists call ensemble tests. To compute sampling distributions often demands a considerable amount of mathematical ingenuity. It certainly
In frequentist theory, it helps to be a Fisher! The basic method, however, is simple to state. Here it is:

$$P(\hat{\theta}|\theta) = d\hat{\theta} \int_x \delta(\hat{\theta} - d(x)) P(x|\theta).$$  \hspace{1cm} (III.14)

The integral is over the values $x$ of the random variable $X$ subject to the constraint that the estimate $\hat{\theta}$ be equal to $d(x)$, which is implemented with a $\delta$-function [11]. Equation (III.14) is an elegant formula, but rarely can we actually use it. Most of the time we must resort to Monte Carlo methods to compute the sampling distribution, or as we high energy physicists would say the ensemble distribution.

Given the sampling distribution, we can compute some measure of its width such as the standard deviation, $\sqrt{\nu[\hat{\theta}]}$. If the calculations can be done analytically it is often useful first to calculate the function

$$F(\omega) = \mathbb{E}[e^{i\omega\hat{\theta}}],$$

$$= \int e^{i\omega\hat{\theta}} f(\hat{\theta}|\theta) d\hat{\theta},$$  \hspace{1cm} (III.16)

which we recognize as the Fourier transform of the probability density function $f(\hat{\theta}|\theta)$. The quantity $F(\omega)$ is called the characteristic function. It provides an elegant way to compute the moments $m_r(0)$, about zero, of a distribution:

$$m_r \equiv \int_{\hat{\theta}} \hat{\theta}^r P(\hat{\theta}|\theta),$$

$$= i^{-r} D^r \omega F(\omega)_{\omega=0}. $$  \hspace{1cm} (III.18)

The standard deviation is, by definition, $\sigma = \sqrt{m_2 - m_1^2}$.

### E  Example: Estimating the Mean Lifetime

Let us apply some of the frequentist theory described above to a simple example. Suppose we have $N$ lifetime measurements, $x = (x_1, \ldots, x_N)$, with which we wish to estimate the mean lifetime of an ensemble of systems that we assume decay according to the exponential law

$$X \sim \text{Exp}(\theta) \equiv \exp(-X/\theta) dX/\theta.$$  \hspace{1cm} (III.19)
We shall do the calculation in several steps.

- **Step 1**
  
  First, we write down the probability density function for the random variables $X_i$:
  
  $$f(X|\theta) = \prod_{i=1}^{N} \exp(-X_i/\theta)/\theta$$
  
  $$= \theta^{-N} \exp \left(-\frac{1}{\theta} \sum_{i=1}^{N} X_i \right). \quad (\text{III.21})$$
  
  Be sure to include *all* terms that depend upon the parameter.

- **Step 2**
  
  We invoke the maximum likelihood principle
  
  $$D_\theta f(X|\theta) = 0, \quad (\text{III.22})$$
  
  to derive the estimator
  
  $$d(X) = \frac{1}{N} \sum_{i=1}^{N} X_i, \quad (\text{III.23})$$
  
  which is just what our intuition told us it had to be!

- **Step 3**
  
  Next we calculate the sampling distribution of $\hat{\theta} = d(X)$ using Eq. (III.14):
  
  $$\mathbb{P}(\hat{\theta}|\theta) = \hat{\theta} \int_{x} \delta \left(\hat{\theta} - \frac{1}{N} \sum_{i=1}^{N} x_i \right) \theta^{-N} \prod_{i=1}^{N} e^{-x_i/\theta},$$
  
  $$= \hat{\theta} \theta^{-N} \int_{0}^{\infty} dx_1 e^{-x_1/\theta} \cdots \int_{0}^{\infty} dx_N e^{-x_N/\theta} \delta(\cdots). \quad (\text{III.25})$$
  
  The most convenient way to handle the $\delta$-function is to use its integral representation
  
  $$\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \ e^{i\omega x}, \quad (\text{III.26})$$
  
  whereupon we can write
  
  $$\mathbb{P}(\hat{\theta}|\theta) = \hat{\theta} \theta^{-N} \frac{1}{2\pi i} \int_{-\infty}^{\infty} d\omega \ e^{i\omega \hat{\theta}} G(\omega), \quad (\text{III.27})$$
where the function $G(\omega)$ is given by

$$
G(\omega) = i\theta^{-N} \prod_{j=1}^{N} \int_{0}^{\infty} dx_j e^{-(i\omega/N+1/\theta)x_j},
$$

$$
= i \left( \frac{N}{i\theta} \right)^N \left( \frac{1}{\omega - N/i\theta} \right)^N.
$$

(III.29)

The integral is readily done by continuing $\omega$ into the complex plane and then applying the residue theorem, noting that the function $G(\omega)$ has a simple pole of order $N$ at $\omega = N/i\theta$. This yields

$$
P(\hat{\theta} | \theta) = \left( \frac{N\hat{\theta}/\theta}{N-1} \right)^{N-1} e^{-N\hat{\theta}/\theta} d(N\hat{\theta}/\theta),
$$

(III.30)

which is a special case of the **gamma distribution**

$$
Gamma(p) = x^{p-1}e^{-x}/\Gamma(p).
$$

(III.31)

As far as the frequentist approach goes we are almost done; the rest is just a matter of computing the moments of the sampling distribution.

In practice, lifetimes are measured with some error; so the sampling distribution we have calculated is generally a poor model. A more realistic model is to approximate the distribution of measurement errors by a Gaussian density function with known variance. It turns out that even for this more complicated case the sampling distribution of $\hat{\theta} = d(x)$ can be worked out exactly [12].

One question we have not answered as yet is this: is the lifetime $\theta$ a parameter for which a best estimator exists; that is, one that is unbiased and has the smallest variance? If it is then we are guaranteed to have found it because we have derived our estimator by maximum likelihood. The answer is yes. Try to prove it.

### F Combining Results

The frequentist procedure to solve an inference problem is:
• Given a model $M$, characterized by the parameter $\theta$, derive the sampling distribution $P(X|\theta)$ for the random variable $X$.

• Construct the likelihood function for the observed data $x$ and solve the likelihood equation $D_\theta \ln L(\theta) = 0$ to obtain the estimate $\hat{\theta} = d(x)$.

• Compute the sampling distribution $P(\hat{\theta}|\theta)$ of the estimates $\hat{\theta}$ from the underlying probability distribution $P(x|\theta)$.

• Compute the moments of $P(\hat{\theta}|\theta)$.

There are at least two things that we usually report, after following this procedure: 1) the estimate $\hat{\theta}$ itself and 2) the standard deviation of the distribution of estimates. The latter quantifies the uncertainty of the estimate. Within a frequentist perspective, there is another important reason why it should be reported: it is needed to combine the results from different experiments.

Suppose there are several measurements of the mass of the top quark. The Particle Data Group’s job is to combine these results to produce a world average, $\hat{\theta}_w$, of the top quark mass, which they do by computing the weighted average

$$\hat{\theta}_w = \sum_i w_i \hat{\theta}_i,$$

(III.32)

whose uncertainty must be assessed. Usually, the Particle Data Group assumes that the results from the different experiments are linearly uncorrelated, in which case they can use

$$\mathbb{V}[\hat{\theta}_w] = \sum_i w_i^2 \mathbb{V}[\hat{\theta}_i],$$

(III.33)

as the variance of the probability density function $f(\hat{\theta}_w|\theta)$ of the world average. But how are they to choose the weights?

Their aim is to combine the results so that the probability density function $f(\hat{\theta}_w|\theta)$ of the world average has the smallest possible variance. This requires then to solve $D_{w_i} \mathbb{V}[\hat{\theta}_w] = 0$.
subject to the constraint $\sum_i w_i = 1$, a problem solved easily (try it) using the method of Lagrange multipliers. The answer is

$$w_i = \frac{1/\sqrt{\hat{\theta}_i}}{\sum_j 1/\sqrt{\hat{\theta}_j}}.$$  

(III.34)

The optimal weights are determined by the variances. Note that we have placed no restrictions on the form of the distributions $\mathbb{P}(\hat{\theta}_i|\theta)$, other than that their variances be finite; in particular, they need not be Gaussian distributions. Therefore, Eq. (III.34) is valid for all sensible distributions.

If we apply the expectation operator to $\hat{\theta}_W$, which performs an average over an ensemble of possible world averages, we get

$$\mathbb{E}[\hat{\theta}_W] = \theta + \sum_i w_i b_i(\theta),$$  

(III.35)

for the mean of the distribution $f(\hat{\theta}_W|\theta)$. Note the conceptual difference between Eq. (III.32) and Eq. (III.35): the first pertains to a set of experimental results that constitute an objectively existing ensemble, while the second pertains to an abstraction, namely, an infinite ensemble of possible world averages.

The Particle Data Group could be lucky, or very well informed, and choose just the right combination of experimental results so that the weighted sum of biases $b_i(\theta)$ is zero for the ensemble of possible world averages. However, it is much easier if each experiment reports an unbiased estimate. Then the bias in the ensemble of world averages would be zero.

A weighted average is not the only way to combine results. We shall discover another way when we discuss the Bayesian theory. Moreover, as I alluded to earlier, the critical requirement is not so much a lack of bias but the convergence of the world “average” to the true value of the parameter of interest, as more and more results are combined.
G  Confidence Intervals: Exact

1  Introduction

Variances are needed in order to form an optimally weighted average. Unfortunately, there is a problem. The variance usually depends upon the parameter whose value we are trying to estimate. Since we have only an estimate of this parameter we can compute only an estimate of the variance. If the variances are insufficiently accurate the weights in the weighted average, Eq. (III.32), will not be quite correct.

There is another difficulty with the weighted average, Eq. (III.32), a conceptual one. The averaging of Eq. (III.32) to obtain the variance, Eq. (III.33), and mean, Eq. (III.35), assumes that the weights remain constant during the averaging. That is, each repetition of an experiment leads to the same reported variance and hence the same weight \( \omega_i \). But is this reasonable? Well, that is a matter of opinion! It could be argued that since each experiment provides only an estimate of the variance, that estimate would vary from one repetition of the experiment to another. Therefore, the weights would not be constant over the ensembles of experiments and that possibility ought to be taken into account when performing the ensemble average. The weights \( \omega_i \) would be correlated with the estimates \( \hat{\theta}_i \), with the possible consequence that the weighted average could be biased even if each estimate \( \hat{\theta}_i \) were not. This is yet another example of arguments that can arise about virtual ensemble is appropriate in studying the distributional properties of estimates.

In a classic paper of 1937, Neyman [13] sidestepped these difficulties with his invention of the confidence interval. These intervals are another measure of uncertainty. The confidence interval \([l(\hat{\theta}), u(\hat{\theta})]\) is two numbers, with \(l(\hat{\theta}) < u(\hat{\theta})\), that depend on the estimate \(\hat{\theta}\) so that

\[
\mathbb{P}[l(\hat{\theta}) < \theta < u(\hat{\theta})] \geq \beta,
\]  

(III.36)

whatever the true value of \(\theta\) and whatever the true values of the other parameters of the problem. We shall refer to this as the Neyman criterion. The quantity \(\beta\) is a probability.
called the confidence level. The confidence interval is very much a frequentist notion which, curiously — in spite of its frequentist pedigree, Fisher disliked. Their is an analogous concept in the Bayesian theory called a credible interval, however, its interpretation is quite different.

Unfortunately, the meaning of Eq. (III.36) is a little complicated. What Eq. (III.36) means is this. The parameter \( \theta \) is an unknown, but fixed, number; for example, the assumed pole mass of the Standard Model Higgs boson. An experiment yields an estimate \( \hat{\theta} \) from which an interval \([l(\hat{\theta}), u(\hat{\theta})]\) can be calculated, which, in a sense to be made clear, has a high probability of containing the true value of the Higgs boson mass. Imagine the experiment repeated infinitely many times. In general, each repetition would yield a different estimate and, therefore, a different interval. Sometimes the interval \([l(\hat{\theta}), u(\hat{\theta})]\) will contain the true value of \( \theta \) and sometimes it will not. The relative frequency with which intervals contain \( \theta \) is called the coverage probability. In other words, it is the fraction of intervals, over the ensemble, that contain the true value of the parameter to be estimated. But we must broaden our gedanken experiment beyond a single ensemble of experiments. Imagine not just one ensemble but a possibly infinite set of ensembles each characterized by a coverage probability and a different fixed point within the parameter space, of which \( \theta \) is one coordinate [40]. If the minimum coverage probability over the set of ensembles is \( \beta \) we say that \([l(\hat{\theta}), u(\hat{\theta})]\) is a confidence interval with a confidence level of 100 \( \times \) \( \beta \)%.

In what sense are we to understand the word “confidence”? Again, consider the measurement of the mass of the Higgs boson. We construct a set of ensembles for which \( \beta = 0.9 \) and in which each ensemble is indexed by, that is, associated with, a different Higgs boson mass, as well as fixed values of any other parameters of the problem, such as the jet energy scale and mean background count.

By construction, for every ensemble within the set at least 90% of intervals contain the value of the Higgs boson mass associated with that ensemble. This holds, in particular, for the ensemble to which our actual interval belongs, presumably the one indexed by the Higgs boson mass of the world we inhabit. Given that 90% is close to 100%, we are entitled to be
reasonably confident that, by chance, the interval actually obtained contains the true value of the true Higgs boson mass. However, the 90% is not a direct measure of that confidence — frequentists do not admit such a concept, rather it is a property of the set of ensembles we have chosen to perform our analysis.

Another aspect of confidence intervals that Neyman himself noted is the following. Consider an ensemble of 68.3% confidence intervals from actual experiments, each, perhaps, having measured a different physical quantity. Then, goes the frequentist argument, since 68.3% of the intervals over this actual ensemble contain the true value, 68.3% of statements of the form \( \theta \in [l(\hat{\theta}), u(\hat{\theta})] \) are true. It is certainly true that every ensemble, actual or otherwise, has a coverage probability. What is less clear, is that for an actual ensemble of putative 68.3% intervals the coverage probability is necessarily at least 68.3%. More to the point, since we do not know the true values of the physical quantities, whose measurement was the object of the various experiments, the claim that the coverage is at least 68.3% over the ensemble of published intervals cannot be verified, except perhaps by some supreme being privy to the inner workings of the Universe.

Fisher was critical of the concept of confidence intervals and their associated levels for what I consider to be sound reasons. The confidence level is a property of the ensemble to which a confidence interval it is presumed to belong. As Fisher [4] noted, in connection with tests of significance,

“.. if we possess a unique sample on which significance tests are to be performed, there is always ... a multiplicity of populations to each of which we can legitimately regard our sample as belonging; so the phrase ‘repeated sampling’ from the same population does not enable us to determine which population is to be used to define the probability level, for no one of them has objective reality, all being products of the statistician’s imagination.”

The same can be said of confidence levels.
2 Computing Intervals

How can we construct a set of ensembles with a given minimum coverage probability $\beta$, that is, confidence level? It is a remarkable fact that ensembles with such properties can be constructed, and rather easily it turns out, for any sampling distribution that depends upon a single parameter. If, however, we have a sampling distribution that depends on many parameters of which only one is of interest the task becomes much more difficult.

We consider Neyman’s method, applied to a sampling distribution that depends upon a single parameter $\theta$. It is convenient to define the probabilities

$$L(\hat{\theta}|\theta) = \int_{z \leq \hat{\theta}} P(z|\theta),$$  \hspace{1cm} (III.37)

and

$$R(\hat{\theta}|\theta) = \int_{z \geq \hat{\theta}} P(z|\theta),$$  \hspace{1cm} (III.38)

which could be called the left and right cumulative distribution functions, respectively. But because this is an awful mouthful, let us just refer to them as the left and right probabilities. The functions $u(\hat{\theta})$ and $l(\hat{\theta})$ are obtained by solving

$$L(\hat{\theta}|u) = \alpha_L, \quad \text{(Upper Limit)}$$  \hspace{1cm} (III.39)

and

$$R(\hat{\theta}|l) = \alpha_R, \quad \text{(Lower Limit)}$$  \hspace{1cm} (III.40)

where

$$\beta = 1 - \alpha_L - \alpha_R.$$  \hspace{1cm} (III.41)

Remember the rule: LEFT is UP and RIGHT is LOW! Notice there are infinitely many sets of intervals that could be computed, corresponding to our freedom to choose the value of $\alpha_L$, or $\alpha_R$. One choice is to set $\alpha_L = \alpha_R = (1 - \beta)/2$, which leads to intervals called central confidence intervals. Equations (III.37) and (III.38) are written in a form valid for both continuous and discrete distributions.
It may not be immediately obvious why this construction leads to intervals with the stated property. To understand that they do consider Fig. 1, which shows a plot of \( \theta \) versus \( \hat{\theta} \). That figure depicts a mapping between the parameter space on the vertical axis and the space of estimates, or observations, on the horizontal axis. The two curves \( \theta = u(\hat{\theta}) \) and \( \theta = l(\hat{\theta}) \), which are solutions to Eqs. (III.39) and (III.40), respectively, map an estimate \( \hat{\theta} \) to an interval in the parameter space. Consider the horizontal line \( \theta = \theta_0 \). That line crosses the upper limit curve \( \theta = u(\hat{\theta}) \) at the estimate \( \hat{\theta}_1 \) and the lower limit curve \( \theta = l(\hat{\theta}) \) at the estimate \( \hat{\theta}_2 \). The estimates \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) partition the space of estimates into three regions, labelled I, II and III. Also note, for each value of \( \theta \), the horizontal line segment between the curves \( u(\hat{\theta}) \) and \( l(\hat{\theta}) \) defines an interval in the space of observations with probability content \( \geq \beta \), by construction.

Suppose an estimate lands in region I. Its associated interval will fail to include the true value. Should an estimate land in region III it too will fail to bracket the true value \( \theta_0 \). On the other hand, any estimate landing in region II will contain \( \theta_0 \). The question is how often does that happen? Well, by construction, the probability to obtain an estimate in region I is given by the left probability \( L(\hat{\theta}_1 | u) \equiv P[\theta \leq \hat{\theta}_1] = \alpha_L \); likewise, the probability to obtain an estimate in region III is just \( R(\hat{\theta}_2 | l) \equiv P[\hat{\theta} \geq \hat{\theta}_2] = \alpha_R \). Consequently, the probability to obtain an estimate in region II — and, therefore, an interval that brackets the true value \( \theta_0 \) — is \( \geq \beta = 1 - \alpha_L - \alpha_R \). We have made no assumption about \( \theta_0 \) other than that it is a fixed number, therefore, this reasoning holds for all values of \( \theta_0 \) that are possible a priori.

3 Example: Exact intervals for the Poisson distribution.

Let us use these rules to compute the exact 68.3% central confidence intervals for the Poisson distribution

\[
N \sim \text{Poisson}(\theta) \equiv \exp(-\theta) \theta^N / N!,
\]

(III.42)
FIG. 1: The plot shows how an estimate is mapped into an interval in the parameter space. As the estimates vary, so do the intervals. Estimates that land in region II lead to intervals that bracket the true value $\theta_0$, while estimates that land in either region I or region II exclude $\theta_0$. 
where the random variable $N$ is the observed count. This is the high energy physicist’s favorite distribution. Applying the “left is up” and “right is low” rules we get

$$L(N|u) = \sum_{k=0}^{N} p(k|u) = (1 - \beta)/2,$$  \hspace{1cm} (III.43)

and

$$R(N|l) = \sum_{k=N}^{\infty} p(k|l) = (1 - \beta)/2.$$ \hspace{1cm} (III.44)

Note that the “right is low” equation can be re-written as

$$\sum_{k=0}^{N-1} p(k|l) = (1 + \beta)/2.$$ \hspace{1cm} (III.45)

In Table I we give the first 26 intervals obtained by solving these equations.

Everyone knows the rule-of-thumb $l(N) \approx N - \sqrt{N}$ and $u(N) \approx N + \sqrt{N}$ for computing approximate 68.3% confidence intervals for the Poisson distribution. But not everyone knows how bad an approximation this is even for values of N as large as 25! This is shown in Fig. 2 which is a plot of the ratio of the width of the exact central interval $u(N) - l(N)$ to the width, $2\sqrt{N}$, of the $\sqrt{N}$ interval. Another noteworthy observation is the peculiar behavior of the coverage probability as $\theta$ varies over the parameter space. This is shown in Fig. 3.

The complex structure is due to the discreteness of the Poisson distribution. As expected, at no point does the coverage probability fall below a probability of 0.683. This is in contrast to the behavior of the $\sqrt{N}$ intervals which, as shown in Fig. 4, have a confidence level that is much lower than 0.683. Remember the confidence level is the smallest coverage probability over the parameter space.

4 Example: Exact intervals for the exponential distribution.

We observe a single unstable system decay after a time interval $t$ and from that we wish to infer something about its mean lifetime $\theta$. We shall assume an exponential decay law (see Eq. (III.19))

$$X \sim \text{Exp}(\theta).$$ \hspace{1cm} (III.46)
TABLE I: Exact 68% Poisson Central Intervals.

<table>
<thead>
<tr>
<th>N</th>
<th>l(N)</th>
<th>u(N)</th>
<th>([u(N) - l(N)] / 2\sqrt{N})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0000</td>
<td>1.8420</td>
<td>——</td>
</tr>
<tr>
<td>1</td>
<td>0.1726</td>
<td>3.3008</td>
<td>1.5641</td>
</tr>
<tr>
<td>2</td>
<td>0.7077</td>
<td>4.6393</td>
<td>1.3900</td>
</tr>
<tr>
<td>3</td>
<td>1.3666</td>
<td>5.9198</td>
<td>1.3144</td>
</tr>
<tr>
<td>4</td>
<td>2.0848</td>
<td>7.1645</td>
<td>1.2699</td>
</tr>
<tr>
<td>5</td>
<td>2.8393</td>
<td>8.3844</td>
<td>1.2399</td>
</tr>
<tr>
<td>6</td>
<td>3.6189</td>
<td>9.5857</td>
<td>1.2179</td>
</tr>
<tr>
<td>7</td>
<td>4.4172</td>
<td>10.7725</td>
<td>1.2010</td>
</tr>
<tr>
<td>8</td>
<td>5.2305</td>
<td>11.9474</td>
<td>1.1874</td>
</tr>
<tr>
<td>9</td>
<td>6.0550</td>
<td>13.1126</td>
<td>1.1762</td>
</tr>
<tr>
<td>10</td>
<td>6.8897</td>
<td>14.2694</td>
<td>1.1668</td>
</tr>
<tr>
<td>11</td>
<td>7.7327</td>
<td>15.4191</td>
<td>1.1588</td>
</tr>
<tr>
<td>12</td>
<td>8.5829</td>
<td>16.5625</td>
<td>1.1517</td>
</tr>
<tr>
<td>13</td>
<td>9.4394</td>
<td>17.7003</td>
<td>1.1455</td>
</tr>
<tr>
<td>14</td>
<td>10.3015</td>
<td>18.8332</td>
<td>1.1401</td>
</tr>
<tr>
<td>15</td>
<td>11.1685</td>
<td>19.9617</td>
<td>1.1352</td>
</tr>
<tr>
<td>16</td>
<td>12.0400</td>
<td>21.0861</td>
<td>1.1308</td>
</tr>
<tr>
<td>17</td>
<td>12.9156</td>
<td>22.2065</td>
<td>1.1267</td>
</tr>
<tr>
<td>18</td>
<td>13.7948</td>
<td>23.3242</td>
<td>1.1230</td>
</tr>
<tr>
<td>19</td>
<td>14.6774</td>
<td>24.4384</td>
<td>1.1197</td>
</tr>
<tr>
<td>20</td>
<td>15.5631</td>
<td>25.5499</td>
<td>1.1166</td>
</tr>
<tr>
<td>21</td>
<td>16.4517</td>
<td>26.6586</td>
<td>1.1137</td>
</tr>
<tr>
<td>22</td>
<td>17.3429</td>
<td>27.7648</td>
<td>1.1109</td>
</tr>
<tr>
<td>23</td>
<td>18.2366</td>
<td>28.8688</td>
<td>1.1085</td>
</tr>
<tr>
<td>24</td>
<td>19.1327</td>
<td>29.9706</td>
<td>1.1061</td>
</tr>
<tr>
<td>25</td>
<td>20.0309</td>
<td>31.0702</td>
<td>1.1039</td>
</tr>
</tbody>
</table>
FIG. 2: We plot the ratio of the width of the exact 68% central interval, for the Poisson distribution, to the that of the $\sqrt{N}$ interval. We see that for small counts, $N$, the $\sqrt{N}$ rule underestimates the width.
FIG. 3: The coverage probability for the 68% central intervals is plotted as a function of the Poisson parameter $\theta$. The strange behavior is due to the discrete nature of the Poisson distribution. The confidence level, which by definition is the minimum coverage probability over the parameter space, is 0.683.
FIG. 4: The coverage probability for the $\sqrt{N}$ intervals is plotted as a function of the Poisson parameter $\theta$. The confidence level is much less than the nominal value of 0.683. Indeed, it is closer to 0.2.
The maximum likelihood estimate for $\theta$ is obviously $\hat{\theta} = x$. A $100 \times \beta\%$ confidence interval can be obtained by setting the left and right probabilities, Eq. (III.37) and Eq. (III.38), equal to $\alpha = (1 - \beta)/2$. The answer is

$$l(\hat{\theta}) = \frac{\hat{\theta}}{\ln \left[ \frac{2}{1 - \beta} \right]}, \quad (III.47)$$

and

$$u(\hat{\theta}) = \frac{\hat{\theta}}{\ln \left[ \frac{2}{1 + \beta} \right]}.$$  \hspace{1cm} (III.48)

(Try to show this.) The interval behaves as we would expect, namely, as we demand a higher confidence level, the wider the interval becomes.

5 Example: Exact intervals for the Gaussian distribution

Along with the Poisson distribution the Gaussian is the high energy physicist’s most important distribution. Statisticians seem positively besotted by it! Ironically, because of its beautiful symmetries it is the worst distribution to use to explain the pitfalls of statistical analysis. It is easy to muddle concepts that are distinct. For example, for a Gaussian there is no numerical distinction between a 68.3% confidence interval and the interval based upon the standard deviation, even though conceptually those intervals are very different. For a Gaussian distribution,

$$X \sim \text{Gaussian}(\theta, \sigma) \equiv \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{X - \theta}{\sigma} \right)^2 \right\}, \quad (III.49)$$

the boundaries of its 68.3% confidence interval are given by $l(x) = x - \sigma$ and $u(x) = x + \sigma$ where $\theta$ is the mean of the distribution and $\sigma$ is the standard deviation. The Gaussian distribution owes its importance to the fact that many likelihood functions assume a Gaussian form when the data-sets become sufficiently large. We may then avail ourselves of the Gaussian’s many useful mathematical properties to simplify analyses that might otherwise be extremely difficult, if not impossible.
H Confidence Intervals: The Unified Approach

In 1998 [14], Gary Feldman and Bob Cousins considered the bothersome problem of creating confidence intervals for a Poisson signal \( s \) masked by background, assuming that the mean background count \( b \) is precisely known. The likelihood function is given by

\[
L(s) \propto P(n|s, b) = \text{Poisson}(s + b) = \exp[-(s + b)](s + b)^n/n!,
\]

where \( n \) is the observed count. Since this is a one parameter problem one could use the solution given above. However, that solution will yield negative limits for the signal if the background count is large enough, which many physicists regard as unsatisfactory. These authors achieved two goals: 1) constructing non-negative intervals for the signal \( s \) and 2) constructing confidence intervals that switch smoothly between one-sided and two-sided intervals. Feldman and Cousins observed, as noted in Sect. III G 2, that for a given value of the parameter \( s \) there is considerable freedom in how an interval in the space of observations \( \{n\} \) can be chosen. Central intervals are just the simplest of infinitely many intervals containing the a given probability content \( \beta \). In the Feldman and Cousins method, for each value of \( s \) one builds an interval in the space of observations \( n \), incrementally, in decreasing order (called Feldman Cousins ordering) of a ratio of likelihoods until the interval contains (or just exceeds) the desired confidence level \( \beta \). Here is the algorithm described for the case of a precisely known mean background count \( b \):

**Algorithm**

```plaintext
for each \( s \) in \([0, s_{\text{max}}]\) 
{
    for each \( n \) in \([0, n_{\text{max}}]\) 
    {
        compute \( \lambda(n) = L(s)/L(\hat{s}) \),
        with
    }
}
```
\[ \hat{s} = \begin{cases} \ n - b \ & \text{if } n > b \\
0 & \text{otherwise} \end{cases} \]

\textbf{sort sequence} \( \{ \lambda(n_i) \} \) in decreasing order

\( \text{sum} = 0 \)

\( i = 1 \)

\textbf{while} \( \text{sum} < \beta \)

\{ 

\( \text{sum} = \text{sum} + P(n_i|s, b) \)

\textbf{with} 

\( n_i = \arg\max_n \lambda(n_i) \)

(i.e., the \( n \) associated with the given \( \lambda \))

\( i = i + 1 \)

\}

\}

Each \( s \) is now associated with a set \( \{ n \} \) and each \( n \) is associated with a set \( \{ s \} \).

For each \( n \) we find the lower and upper bounds of its associated set \( \{ s \} \).

\textbf{for each} \( n \) in \([0, n_{\text{max}}]\)

\{ 

\( l(n) = \min \{ s \} \)

\( u(n) = \max \{ s \} \)

\}

1 \textbf{ Example: KARMEN results} 

The KARMEN experiment [15] recorded \( n = 0 \) events with a mean background of \( b = 2.8 \).

If we assume the background uncertainty to be negligible, one obtains the Feldman Cousins interval \([l(0), u(0)] = [0.0, 1.71]\), which by construction is non-negative.
2 Discussion

In this example of the Feldman Cousins method the likelihood \( L(s) \) depends on the single parameter \( s \), since \( b \) is assumed to be perfectly known. Consequently, by construction, the intervals satisfy the Neyman criterion (see Sect. III G). However, if \( b \) is not precisely known, but is replaced by an estimate, exact coverage is not guaranteed [41]. Moreover—and this is true of all frequentist methods for constructing intervals, it is a matter of debate which ensemble is to be used to define the confidence level. One can think of at least two: 1) the ensemble in which \( n \) varies randomly, but the background estimate \( \hat{b} \) is fixed or 2) the ensemble in which both \( n \) and \( \hat{b} \) vary randomly from one hypothetical experiment to the next. There is the further complication that the distribution of \( \hat{b} \) may itself depend on parameters whose true values are unknown. The Neyman criterion requires exact coverage for all possible values of all parameters. Parameters other than the ones for which inferences are desired are referred to as nuisance parameters.

Why are the intervals non-negative? Because of the reasonable but ad hoc choice \( \hat{s} = 0 \) for the signal estimate whenever \( n < b \). However, one can imagine other ad hoc choices; here is one: set \( \hat{s} = \frac{n}{n'} \hat{s}' \), where \( \hat{s}' \) is the first non-negative estimate occurring at count \( n' \). Note, however, that any choice other than \( \hat{s} = 0 \) will yield wider intervals. One can imagine other ordering principles also.

I Confidence Intervals: Approximate

Most of the time the exact procedure for computing confidence intervals is simply impossible to put into practice either because of insurmountable mathematical difficulties or computational constraints. When the exact approach is computationally burdensome we must resort to approximations, the most widely used of which is based upon the following theorem. Let
\[ Y = -2 \ln \frac{f(x_1, \ldots, x_N|\theta)}{f(x_1, \ldots, x_N|\hat{\theta})}, \]  

(III.51)

then, under fairly mild assumptions, as \( N \to \infty \), \( Y \to \chi^2 \) variable with one degree of freedom. The theorem implies that in the large data-set limit \( Y \) becomes independent of the parameters of the probability density function. Consequently, a probabilistic statement about the quantity \( Y \) can be re-cast as one for the quantity \( \theta \), by inverting Eq. (III.51). The theorem states, in effect, that in the limit of large data-sets all reasonable probability densities functions approach a Gaussian form, as alluded to above.

More generally, if there are \( K \) un-constrained parameters in the numerator and we divide by the absolute maximum of the probability density function (or likelihood function if you prefer) then the random variable \( Y \) will follow a \( \chi^2 \) distribution with \( K \) degrees of freedom, provided that the data-set is large enough.

The utility of this theorem is best illustrated by an example. During the presidency of Bill Clinton, let us suppose that we watched American TV for one minute and counted the number of times president Clinton uttered the phrase “bridge to the 21st century” and found that he did so 4 times. We wish to estimate the average number of times he uttered the phrase per minute and to quantify the accuracy of our estimate by computing a 68.3% confidence interval for the estimated mean count. Let the mean count, in one minute, be denoted by \( \theta \) and let the measured count be denoted by \( n \). We shall assume a Poisson model for this problem with distribution \( \mathcal{P}(n|\theta) = \text{Poisson}(\theta) \). The maximum likelihood principle gives \( \hat{\theta} = n \) as our estimate of the mean count; no surprise here! As shown in the previous section, the confidence intervals can be computed exactly because this is a single parameter problem. But let us pretend that the problem is one that requires an approximation. Even for president Clinton \( n \) is a tad short of infinite, but let us nonetheless apply the theorem stated above for \( n = 4 \) and write

\[ y = 2(n \ln n - n - n \ln \theta + \theta) \approx \chi^2. \]  

(III.52)

Note that \( y = 0 \) at \( \theta = \hat{\theta} = n \). For \( \theta \neq n \), \( y \to +\infty \). The function \( y \) has a roughly parabolic
form, becoming more so as $n \rightarrow +\infty$. Accordingly, in that limit, we should be able to write Eq. (III.52) as

$$y \approx \chi^2 = (\theta - \hat{\theta})^2 / \sigma^2, \quad \text{(III.53)}$$

and find that for every value of $y$ we get two values of $\theta$: $l(\hat{\theta}) = \hat{\theta} - \sigma$ and $u(\hat{\theta}) = \hat{\theta} + \sigma$. However, it is not necessary to re-write Eq. (III.52) in that way to obtain the lower and upper limits; we just solve Eq. (III.52) directly.

From the properties of the $\chi^2$ distribution we know that the 68.3% boundary occurs at $\chi^2 = 1$, and that

$$P[\chi^2 < 1] = 0.683. \quad \text{(III.54)}$$

Therefore, setting $y = 1$, corresponding to $\chi^2 = 1$, in Eq. (III.52) we find $l(4) = 2.3$ and $u(4) = 6.3$. We, therefore, report the results of our little counting experiment as

$$\hat{\theta} = 4^{+2.3}_{-1.7} \text{ annoying utterances per minute.} \quad \text{(III.55)}$$

The interval (2.3,6.3) is close to that provided by the $\sqrt{N}$ rule: $l(4) = 2$ and $u(2) = 6$. But, the $\sqrt{N}$ rule itself is only an approximation to the exact 68.3% central confidence interval. The exact 68.3% central confidence interval for $n = 4$ is (2.1,7.2), according to the results given in Table 1. The interval (2.3,6.3) obtained from the log-likelihood ratio is an approximate 68.3% confidence interval in the sense that it belongs to an ensemble in which the minimum coverage probability approximately 0.683.

1 Discussion.

The method just described is how uncertainties are estimated from a negative log-likelihood curve: find the minimum, go up by 0.5 and read off the “error” interval. From a frequentist viewpoint, this procedure is simply an efficient method for computing approximate confidence intervals. For single-parameter problems, the exact procedure requires first a calculation of the sampling distribution of estimates from which any number of exact intervals can be computed. But the “up-0.5" rule is so handy that it is the method of choice for
all sample sizes, not just asymptotically large ones. Physicists, being free spirits, feel free to use the “up-0.5” rule regardless of its provenance—and then argue interminably about what kind of function to fit to the log-likelihood curve in order to find its minimum and extract the confidence interval. But, if we insist on using this rule for small data-sets, and we claim to be frequentists, we have no right to claim that we have computed a “one-sigma”—that is, a 68.3% interval. It may, or may, not be a 68.3% interval.

One reason, I suspect, that we cling so stubbornly to the 0.5-rule is that it is easy to use. Another is our desire to have a definition of uncertainty that reflects the intrinsic accuracy of the data-set we actually have, rather than one that reflects the average accuracy of an ensemble of hypothetical data-sets. If, by chance, data are tightly clustered most physicists accept the proposition that such data are more informative than data that are more scattered; consequently, they are perfectly happy to quote the uncertainty derived from the negative log-likelihood function rather than that pertaining to the ensemble of estimates. This is the sensible policy adopted by both the CDF and DØ collaborations. For a justification of this commendable bit of intuition, we must await our discussion of the Bayesian theory.

**J Hypothesis Testing**

The last topic to be covered within the frequentist theory is hypothesis testing. Since Neyman and Fisher disagreed about how this should be done, we should, and shall, treat their ideas separately.

1 **According to Neyman and Pearson**

Consider two hypotheses $H_0$ and $H_1$. By convention, the conservative hypothesis — the one to be rejected only if one has compelling evidence to do so — is called the **null hypothesis** and is usually denoted by $H_0$. The other hypothesis, $H_1$, is called the **alternative**
hypothesis. For example, in searches for new phenomena associated with supersymmetry
the null hypothesis $H_0$ is the Standard Model, while the alternative hypothesis $H_1$ is some
SUSY model. If the evidence is insufficiently strong to suggest something new one may fail
to reject the null hypothesis. It is then a matter of taste, tradition, or convenience, whether
or not one chooses to accept the null hypothesis. Failing to reject the null hypothesis does
not, in itself, mandate that one accepts it. Often, however, we do because the null hypothesis
is usually worth keeping, at least until something better comes along. The Standard Model
is an exceedingly good theory. Therefore, if some SUSY search failed to reject the Standard
Model — the null hypothesis, it would be considered eccentric if one chose not to accept the
null. On the other hand, if the evidence is strongly against the null hypothesis then one may
well wish to reject it. But again, it is a matter of taste whether the alternative hypothesis
should be accepted. The point is that in the frequentist theory the goal of hypothesis testing
is a negative one: it is to ascertain whether or not the null hypothesis should be rejected.
That being said, since it is rather tortuous to keep saying “failed to reject” I shall use the
shorthand “accept”.

When testing a null hypothesis we can make one of two mistakes: 1) we could mistakenly
reject the null when it is true or 2) accept it when it is false. The first kind of error is called
a type I error, the probability of which we are usually anxious to keep rather small. The
second kind of error is called a type II error, which ideally should be small also. Can we
compute the probability of these errors? In general, no! It could be that neither hypothesis
is true and even if one were we do not know which one. What can be computed is the
probability of error assuming that one or the other hypothesis is true. If we assume $H_0$ to
be true then we can compute the probability $\alpha$ of rejecting it, that is, making a type I error.
The probability $\alpha$ is called the size, or significance of the test. The smaller the significance
the more demanding we are of the evidence that may lead us to reject $H_0$. If we assume $H_1$
to be true, and therefore $H_0$ to be false, then we can compute the probability $\beta$ of accepting
the false hypothesis $H_0$, that is, a type II error. The probability $1 - \beta$ is called the power
of the test; it is the probability to reject the null hypothesis when it is false. Since this is a
good thing to do, we would like the power of the test to be as large as possible.

The idea of the **Neyman Pearson hypothesis test** is to place an upper bound \( \alpha \) on the probability to reject a null hypothesis that is true — a false rejection, while minimizing the probability to accept a null hypothesis that is false — a false acceptance, or, equivalently, while maximizing the power. Hypotheses that are fully specified are called **simple hypotheses**, otherwise they are **compound**, or **composite**, hypotheses. A simple hypothesis is one whose specification is sufficient to allow an unambiguous prediction of all the possible data-sets that could be observed. Such an hypothesis has no free parameters. The Neyman Pearson test is optimal for simple hypotheses, which are the only ones we consider here. In order to construct such a test it is necessary to define a **critical** or **rejection region** such that if the data fall within it one would be led to reject the null.

If the null hypothesis \( H_0 \) is true we shall assume it yields the sampling distribution \( \mathbb{P}(X|H_0) \) for random variable \( X \), which could be multi-dimensional. Likewise, \( \mathbb{P}(X|H_1) \) is the distribution associated with the alternative hypothesis \( H_1 \). To define the test we need to minimize the false acceptance probability \( \beta \) subject to the constraint that the false rejection probability has an upper bound, \( \alpha \). This is a problem readily solved using Lagrange multipliers. Let the critical region be denoted by \( R \) and its complement by \( \overline{R} \). The probability to reject the null hypothesis if true is the probability for data to appear in the rejection region \( R \), that is,

\[
\alpha' = \int_R \mathbb{P}(X|H_0) \leq \alpha ,
\]  

(III.56)

which is the constraint to be adhered to. The probability to accept the null hypothesis if false is the probability to obtain data in the complementary region \( \overline{R} \), assuming that \( H_1 \) is true, that is,

\[
\beta = \int_{\overline{R}} \mathbb{P}(X|H_1) ,
\]  

(III.57)

which is the quantity to be minimized, subject to the constraint, Eq. (III.56). Introducing
the Lagrange multiplier $\kappa$ yields the function

$$
\int_{\mathbb{R}} p(X|H_1) + \kappa \left[ \int_{\mathbb{R}} p(X|H_0) - \alpha' \right],
$$

(III.59)

$$
= 1 - \int_{\mathbb{R}} f(x|H_1) dx + \kappa \int_{\mathbb{R}} f(x|H_0) dx - \kappa \alpha',
$$

$$
= \int_{\mathbb{R}} \left[ \kappa f(x|H_0) - f(x|H_1) \right] dx + [1 - \kappa \alpha'],
$$

which is minimized when $\kappa f(x|H_0) - f(x|H_1) < 0$, that is, for the rejection region $\mathbb{R}$ defined by

$$
L_R = \frac{f(x|H_1)}{f(x|H_0)} > \kappa.
$$

(III.60)

The quantity $L_R$ is called the **likelihood ratio**.

Even though the likelihood ratio test is optimal only for simple hypotheses, it is still used even when there are nuisance parameters. Typically, what is done, as in the Feldman Cousins method, is to replace the nuisance parameters with reasonable estimates.

2 **According to Fisher**

Fisher argued that it was satisfactory to perform hypothesis testing even in the absence of an explicit alternative, an opinion not shared by Neyman. The archetypical example in particle physics of the sort of problem Fisher addressed is that of deciding whether an observed excess of events over background is “significant”, while not committing oneself to any particular hypothesis about the putative signal. Another is deciding whether a fit to data is satisfactory, which is known as a **goodness of fit** test. I have stated these examples in the language that is commonly used. However, what we are really doing in the first example is deciding whether or not to reject the background, that is, null, hypothesis. In the case of a goodness of fit test, we are really deciding whether to reject the fit. Therefore, it would be more appropriate to speak of a **badness of fit** test! However, the idea of getting a “good fit” is so ingrained that it is unlikely that the phrase “badness of fit” will ever catch on.
So how did Fisher, and how do we, get away without specifying an alternative hypothesis? By having in mind an implicit set of alternatives in mind! Given a probability distribution \( P(X|H_0) \), one constructs a rejection region \( R \) based upon a judgement about what values of \( X \) are suggestive of “a direction away” from the null hypothesis such that more extreme values of \( X \) cast greater doubt on the veracity of the null hypothesis. In order to effect such a construction we need *some* idea of what direction away from the null means. That is, one must have in mind an implicit set of alternatives that are judged to be worse than the null in explaining the data. In many practical situations, \( X \) is 1-dimensional and the more extreme values tend to be in one or both tails of the distribution. The size of the test, or significance, is defined as

\[
\alpha = \int_{X > X_0} P(X|H_0),
\]

where the rejection region is \( R = \{ X : X > X_0 \} \). In principle, the test is to be used in the same way as the Neyman Pearson test: one commits ahead of time to a significance, say \( \alpha = 0.1\% \); if the observed value \( x > X_0 \) the null hypothesis is to be rejected, without flinching. In practice, one computes

\[
p = \int_{X > x} P(X|H_0),
\]

called the **p-value** and reports it. In high energy physics, if \( H_0 \) is a (fully specified) background hypothesis the p-value is usually described with words such as “the probability for the background to fluctuate to, and above, the observed count \( x \) is so-and-so”. Tradition then requires that this be converted into an equivalent number of standard deviations assuming a Gaussian model for the background. If the p-value is small enough, then we might be willing to say that the data are not described by background alone.
IV. BAYESIAN STATISTICS

Bayesian statistics is firmly grounded in the subjective interpretation of probability. Whereas the frequentist theory deals only with the distributional properties of data, that is, statements of the form

$$P(Data|Theory),$$

the Bayesian theory admits, in addition, statements of the form

$$P(Theory|Data),$$

that is, the probability that a given Theory is correct, in light of evidence provided by Data. This is precisely the kind of statement that most people would wish to get from a theory of inference. The connection between the two probabilities, Eqs. (IV.1) and (IV.2), can be sketched as follows

$$P(Theory|Data, I) \propto P(Data|Theory, I) P(Theory|I),$$

which is an application of Bayes' theorem, Eq. (II.13). We have introduced the symbol I to remind us of the fact that every probability is conditioned on some prior information and, or, assumptions. The probability $$P(Theory|I)$$ is called the prior probability. It encodes what we believe we know about the Theory independently of the Data. The probability $$P(Data|Theory, I)$$, without its measure, is the likelihood, while the probability $$P(Theory|Data, I)$$ is called the posterior probability. The power of the Bayesian theory is due in large measure to the fact that one can speak, meaningfully, of the probability of a theory, or an hypothesis. Moreover, since Theory can be anything whatsoever one anticipates that the domain of applicability of the Bayesian theory is considerable larger than that of a theory where the notion of a probability of an hypothesis is absent. However, this conceptual gain comes at a price. In order to arrive at a posterior probability the price to be paid is the specification of a prior probability for the Theory, independently of the Data. There is simply no way around this if one wishes to adhere to the rules of probability theory.
A The Bayesian Method

We begin with an abstract statement of the Bayesian inference procedure:

- Assign a prior probability \( P(B_k D_j | I) \) to each proposition \( B_k D_j \) in light of what we know about them. This \textbf{prior knowledge} is represented by the proposition \( I \).

- Acquire some pertinent evidence, represented by proposition \( A \), and assign to these data a probability \( P(A | B_k D_j I) \).

- Compute the posterior probability \( P(B_k D_j | AI) \) from Bayes’ theorem.

- If we are interested only in the set of propositions \( \{B_k\} \), we excise the uninteresting ones, \( \{D_j\} \), from the problem by summing over them, \( P(B_k | AI) = \sum_j P(B_k D_j | AI) \), in accordance with the rules of probability theory. This operation is called \textbf{marginalization}. Now that we have the posterior probability \( P(B_k | AI) \) we can answer all rational questions about the proposition \( B_k \).

The above procedure is a mathematical model of inductive reasoning. Being a mere model, it captures only some aspects of human reasoning, which surely is vastly more complex than the model implies. Yet even this simplified model of reasoning has proven to be extraordinarily powerful.

B Continuous Sets of Propositions

In many applications in high energy physics the propositions we make are of the form \( A = \theta \in [a, a + da] \), that is, a parameter has a value within some continuous set. Let

\[
P(x | \theta, \lambda, I) = \int_{\Omega} f(z | \theta, \lambda, I) dz,
\]  

be the probability assigned to the data-set \( x \), contained in a neighborhood \( \Omega \) of \( x \), and \( \theta \) and \( \lambda \) the parameters of the model currently under consideration. Perhaps, \( \theta \) is the parameter of interest, say the top quark mass, while \( \lambda \) is not — that is, \( \lambda \) represents one
or more nuisance parameters, such as the mean background rate or the jet energy scale. If 
\( \mathbb{P}(\theta, \lambda|I) = f(\theta, \lambda|I)d\theta d\lambda \) is the prior probability assigned to the proposition that \( \theta \) and \( \lambda \) have certain values — where \( f(\theta, \lambda|I) \) is the prior density, we can write Bayes’ theorem as

\[
\mathbb{P}(\theta, \lambda|x, I) = \frac{\mathbb{P}(x|\theta, \lambda, I)\mathbb{P}(\theta, \lambda|I)}{\int_{\theta, \lambda} \mathbb{P}(x|\theta, \lambda, I)\mathbb{P}(\theta, \lambda|I)} = f(\theta, \lambda|x, I) d\theta.
\]

(IV.6)

In the frequentist theory, no one has found a general algorithm to rid a problem of nuisance parameters in a manner consistent with the precepts of the theory. Typically, what is done is to replace the nuisance parameters with estimates. In the Bayesian theory, the way to excise nuisance parameters is specified by the rules of probability theory: they are to be removed by marginalization, that is, by integrating them out of the problem

\[
\mathbb{P}(\theta|x, I) = \int_{\lambda} \mathbb{P}(\theta, \lambda|x, I).
\]

(IV.7)

This elegant formula summarizes what we know about the parameter \( \theta \), given the data we have acquired and our prior knowledge, that properly accounts for whatever knowledge we may have about the other parameters.

C The Likelihood Principle

The posterior probability, \( \mathbb{P}(\theta|x, I) = f(\theta|x, I) d\theta \) — the final result of our inference about \( \theta \), displays a very important philosophical, and practical, difference between frequentist and Bayesian statistics that we have already alluded to, namely,

- An inference depends only on the data observed,

a principle that is referred to as the **likelihood principle**. Clearly, to base an inference on an ensemble of outcomes is to be is sharply at odds with the likelihood principle. Consequently, this principle is at odds with Neyman and Pearson’s notion of confidence intervals and a host of standard frequentist practice. Since these methods are still firmly entrenched, one
is naturally led to ask: is the likelihood principle sensible? Certainly, Jeffreys thought so. Ironically, even his nemesis, Fisher — a forceful critic of all things Bayesian, was an advocate of the likelihood principle. Indeed, Fisher was extremely critical of what he regarded as Neyman’s “extreme frequentism” and sought to develop a more likelihood based theory of inference. Fisher’s development of the theory of **fiducial inference** was his enigmatic attempt to make good on the likelihood principle, without conceding an inch, so he thought, to Bayesians. A further irony is that, according to a theorem due to Birnbaum [16], the likelihood principle follows from ideas that many frequentist statisticians consider sound.

My own view is that the likelihood principle makes a lot of sense. If we are lucky enough to have obtained, by chance, data that are tightly clustered these data are surely telling us something useful, more so than if they had been widely scattered. It does not seem sensible to throw away this information and, instead, base an inference on the average behavior of the ensemble from which these data are assumed to have been sampled, which ensemble is accessible to us only via simulations. There is nothing amiss with the use of simulations; they are indispensable in any non-trivial analysis. What is amiss is to fail to recognize that, useful as they are, these ensembles remain hypothetical. In an experiment to search for new phenomena, for example, we can always compute the probability for background to fluctuate to a count equal to or greater than that observed. However, the p-value computed pertains to a hypothetical circumstance: an experiment that we could, in principle, repeat but, in practice, seldom do and in any case, can never do infinitely many times. Moreover, as long as we are unable to see the future we cannot possibly *know* what we would observe were we to actually repeat an experiment once, let alone infinitely many times. All we can do is *predict* what could be observed. It is logically possible that were we to actually repeat our experiment we might observe that which we failed to predict. Indeed, that is often how progress is made.
D Parameter Estimation

The posterior probability is a complete statement of the results of an inference. In practice, however, one usually provides a summary of it in terms of a few numbers, such as its mean and some measure of its width. In some circumstances, after marginalization over nuisance parameters, it may be useful to take the mean of the resulting posterior density as an estimate of the parameter of interest. However, the mean is not the only possibility. One way to formalize the construction of estimates is through loss functions, which we discussed in general terms in Sect. I A 1 and which we discuss in more detail below. In the Bayesian theory, it is natural to speak of our knowledge being uncertain, in particular, our knowledge of the value of a parameter. Moreover, this uncertainty is measured not by the expected scatter of estimates over an ensemble but rather by some measure of the width of the posterior density, which, in accordance with the likelihood principle, depends only on the observed data.

A loss function is a way to measure the quality of a decision. A typical decision is: given a data-set \( x \) decide that the estimate of \( \theta \) is \( \hat{\theta} = d(x) \), where \( d \) is a special kind of decision function called an estimator. We consider three different loss functions.

1 Quadratic Loss

The quadratic loss, introduced earlier, is

\[
L(\theta, d) = (\theta - d)^2. \tag{IV.8}
\]

Earlier, we also introduced the average loss, that is, the risk function. In the frequentist theory, the averaging is done with respect to all possible data-sets \( x \). In the Bayesian theory, we average over all possible propositions about the value of \( \theta \), constrained by the fact that
we have a definite data-set. Therefore, we are led to consider the risk function

\[ R(x) = \mathbb{E}[L(\theta, d)]_{\theta}, \]

\[ = \int_{\theta} L(\theta, d) \mathbb{P}(\theta|x), \]  

(IV.10)

that is,

\[ R(x) = \int_{\theta} (\theta - d)^2 \mathbb{P}(\theta|x), \]  

(IV.11)

for the quadratic loss, where \( \mathbb{P}(\theta|x) \) is the posterior probability.

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

\[ D_d R(x) = D_d \int_{\theta} L(\theta, d) \mathbb{P}(\theta|x), \]

\[ = \int_{\theta} D_d L(\theta, d) \mathbb{P}(\theta|x), \]

\[ = 0. \]  

(IV.13)

The symbol \( D_d \) is the derivative operator with respect to \( d \). (Being physicists, we naturally assume that the derivative and integral operators commute.) After minimization, we obtain the intuitively pleasing result

\[ \hat{\theta} = d(x) = \int_{\theta} \theta \mathbb{P}(\theta|x). \]  

(IV.14)

In words:

- For a quadratic loss, the optimal estimate is the mean of the posterior probability.

2. Absolute Loss

The absolute loss, defined by

\[ L(\theta, d) = |\theta - d|, \]  

(IV.15)

is more tolerant of deviations than is the quadratic loss, consequently, one would expect the resulting estimator \( d \) to be less sensitive to the tails of the posterior density and in that
sense more robust. As before, we obtain the estimator \( d \) by minimizing the risk

\[
\mathcal{R}(x) = \int_\theta |\theta - d| \mathbb{P}(\theta|x).
\]

Differentiating the risk function with respect to \( d \) yields

\[
D_d \mathcal{R}(x) = 0
\]

\[
= \int_\theta D_d |\theta - d| \mathbb{P}(\theta|x)
\]

\[
= - \int_\theta \frac{\theta - d}{|\theta - d|} \mathbb{P}(\theta|x),
\]

that is,

\[
\int_{\theta < d} f(\theta|x) \, d\theta = \int_{\theta > d} f(\theta|x) \, d\theta,
\]

which shows that the optimal estimator \( d \) for the absolute loss function is the **median** of the posterior density.

3 **Bounded Loss**

Finally, we consider the bounded loss

\[
\mathcal{L}(\theta, d) = \begin{cases} 
0, & |\theta - d| < a \\
1, & |\theta - d| \geq a
\end{cases}
\]

(IV.20)

4 **Uncertainty**

The uncertainty in our knowledge of a parameter is quantified by some measure of the width of the posterior density. One such measure is the variance

\[
\mathrm{var}[x] = \mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2.
\]

Another is the Bayesian analog of a confidence interval, \([l(\hat{\theta}), u(\hat{\theta})]\), called a **credible interval** or sometimes a Bayesian confidence interval, obtained from the formulae

\[
\int_{\theta \leq l(\hat{\theta})} \mathbb{P}(\theta|x) = \alpha_L
\]

(IV.22)
and

\[ \int_{\theta \geq u(\hat{\theta})} P(\theta | x) = \alpha_R, \]  

(IV.23)

where \( \alpha_L \) and \( \alpha_R \) as chosen so that \( \beta = 1 - \alpha_L - \alpha_R \). These formulae appear to be rather like those used to compute intervals in the frequentist theory. However, the similarity is superficial and misleading. Here the confidence level \( \beta \) is not defined by reference to an ensemble of intervals. The interpretation of a \( 100 \times \beta \% \) Bayesian interval is rather direct: \( \beta \) is the probability that the proposition \( A\theta \in \left[l(\hat{\theta}), u(\hat{\theta})\right] \) is true.

E Combining Results

In the frequentist theory the results from different experiments are combined using a weighted average. However, more generally, results can be combined using use Bayes’ theorem, Eq. (II.13). Let \( f(x_k | \theta, \lambda, \alpha_k) \) be the likelihood for experiment \( k \), where \( \theta \) is the parameter of interest, \( \lambda \) represents any nuisance parameters that are common to all experiments — this could be, for example, a measured cross section used by all experiments — and \( \alpha_k \) represents nuisance parameters specific to experiment \( k \). Ideally, for each experiment the marginal likelihood,

\[ f(x | \theta, \lambda) = \int f(x | \theta, \lambda, \alpha_k) f(\alpha_k) d\alpha_k, \]  

(IV.24)

would be reported, that is, the likelihood function marginalized with respect to one or more parameters, here the nuisance parameters \( \alpha_k \) specific to the experiment. We do not marginalize with respect to \( \lambda \) because it is common all experiments. The function \( f(\alpha_k) \) is the prior density for \( \alpha_k \), where for simplicity we have dropped the dependence on the prior information \( I \). In writing Eq. (IV.24) we have implicitly factorized the full prior density \( f(\theta, \lambda, \alpha_k) \) as follows

\[ f(\theta, \lambda, \alpha_k) = f(\theta, \lambda | \alpha_k) f(\alpha_k). \]  

(IV.25)
We shall assume that for every experiment, whose results are to be combined, the prior density \( f(\theta, \lambda | \alpha_k) \) is independent of \( \alpha_k \), in which case we may write
\[
f(\theta, \lambda, \alpha_k) = f(\theta, \lambda) f(\alpha_k).
\] (IV.26)

Given this assumption, each experiment can produce an inference about \( \theta \) and \( \lambda \) by supplying a prior density \( f(\theta, \lambda) \). This observation provides the clue about how to combine results. The prior density \( f(\theta, \lambda) \) for a given experiment is simply the posterior density from a previous experiment \( f(\theta, \lambda | x) \). Therefore, by recursively combining the results from \( K \) experiments we obtain the overall posterior density
\[
f(\theta, \lambda | x_1, \ldots, x_K) = \frac{f(x_1 | \theta, \lambda) \cdots f(x_K | \theta, \lambda) f(\theta, \lambda)}{\int_{\theta, \lambda} f(x_1 | \theta, \lambda) \cdots f(x_K | \theta, \lambda) f(\theta, \lambda)}.
\] (IV.27)

This is essentially just the joint likelihood function for the combined results times an overall prior density for \( \theta \) and \( \lambda \). Unlike the weighted average, this method will converge to the true value as more and more experiments are combined, even if each experiment is biased, provided that the result from each experiment is consistent.

### F Model Selection

Let’s say we have a set of competing models \( M \), which may depend upon different sets of parameters \( \theta_M \). Given some prior information \( I \) and some data-set \( x \), how do we decide between the various models? This is the problem of hypothesis testing or model selection.

The first thing we do is assign a probability, \( \mathbb{P}(x | \theta_M, M, I) \), to our data-set given a model \( M \) and hypotheses about the value of the corresponding parameters \( \theta_M \). We must also assign a prior probability \( \mathbb{P}(\theta_M, M | I) \). Then we write down Bayes’ theorem
\[
\mathbb{P}(\theta_M, M | x, I) = \frac{\mathbb{P}(x | \theta_M, M, I) \mathbb{P}(\theta_M, M | I)}{\sum_M \int_{\theta_M} \mathbb{P}(x | \theta_M, M, I) \mathbb{P}(\theta_M, M | I)}.
\] (IV.28)

The symbol \( \mathbb{P}(\theta_M, M | x, I) \) represents the probability of the proposition that \( M \) is the “correct” model with parameter values \( \theta_M \), given the evidence \( x \) and \( I \).
It is very important to understand that the probabilities $\mathbb{P}(\theta_M, M|x, I)$ are conditioned on the set of models considered. That’s why the word correct is in quotes. Correct, in this context, means the best of the current bunch. If someone came up with yet another model, then the probabilities would, in general, change were we to include the new model in the set under consideration. Therefore, $\mathbb{P}(\theta_M, M|x, I)$ cannot be construed as an absolute measure of the validity of a model. But it is a measure of the conditional validity of a model: it provides a way to compare models within a given set in light of what we know. If a rational thinker had to choose a single model she would opt for the model with the highest posterior probability. Of course, should she acquire further pertinent information that information, via Bayes’ theorem, could cause our rational thinker to change her mind. This is not a defect, but a virtue.

One last thing. We can marginalize $\mathbb{P}(\theta_M, M|x, I)$ with respect to $\theta_M$ to obtain $\mathbb{P}(M|x, I)$, the probability of model $M$. This is potentially very useful if each model within the set are identical, except for a single parameter $\alpha$. For example, the $M$ could label models that differ by the assumed top quark mass. We then have a way to estimate that parameter:

$$\hat{\alpha} = \sum_M \alpha_M \mathbb{P}(M|x, I), \quad (IV.29)$$

and its associated uncertainty

$$\sigma^2_{\alpha} = \sum_M (\alpha_M - \hat{\alpha})^2 \mathbb{P}(M|x, I). \quad (IV.30)$$

V. BAYESIAN ANALYSIS

We now have enough of the Bayesian theory in hand to see how it works in practice. To this end we shall conclude these notes with a look at a typical application in high energy physics: Analyzing a counting experiment.
A Optimal Event Selection

Before we can measure something, we must have events. A basic task of data analysis, therefore, is to separate signal from background. It is one of the more enjoyable tasks because we get a chance to show off our cleverness. The most important work, of course, is using one’s physical intuition to devise fiendishly clever variables to dig out the signal. But, having found our fantastic variables, the drudgery begins: finding cuts.

The traditional method to find cuts combines a judicious use of common sense with a dose of trial and error. We all have experienced, I’m sure, how extremely time consuming that can be; especially if, along the way, we have to refine the odd variable or two. Surely, there is a better way. Well, of course there is; and it should come as no surprise that it has something to do, again, with Mr. Bayes!

It helps to think about the problem geometrically. Suppose we have found $n$ variables that we consider useful for separating signal from background. The $n$ variables can be thought of as a point in an $n$-dimensional space, sometimes referred to as feature space. Presumably, by construction, the signal tends to cluster in one part of this space while the background occupies a different region. However, inevitably, there will be some overlap between the signal and background distributions. The problem to be solved is to find the boundary that separates optimally signal from background. Usually we do the simplest thing: we construct a boundary from planes that are perpendicular to the axes, where each plane corresponds to a cut on a specific variable. However, in general, the optimal boundary cannot be built from such intersecting planes; in general, it will be a curved surface.

The problem of finding this surface, however, is indeterminate until we have specified what we mean by optimal. This is what most people mean by optimal: a boundary is optimal if it minimizes the probability to misclassify events. So all we have to do is to write down a function of this probability, minimize it and we’re done. For the moment, let’s suppose that we know the signal and background distributions: $p(x|S)$ and $p(x|B)$, respectively. Let us further assume that we know the signal and background prior probabilities $p(S)$ and $p(B)$. 
These prior probabilities are not controversial: $p(S)$ is just the chance to pick a signal event without regard to its feature vector $\mathbf{x}$, and likewise for $p(B)$. Since the event must be either signal or background we must have that $p(S) + p(B) = 1$. What’s the chance to misclassify a signal event, with feature vector $\mathbf{x}$? It is just the chance for the signal event to land on the background side of the boundary. It’s easiest to see what’s going on if we consider a one dimensional problem, with the boundary, say, at $x = x_0$. The probability $E_S$ to misclassify a signal event is

$$E_S(x_0) = p(S) \int_x p(x|S)h(x_0 - x). \quad (V.1)$$

Here $h(z)$ is the step function, which is defined by $h(z) = 1$, if $z > 0$, but zero otherwise. The probability to misclassify the background is likewise the probability for the background to land on the signal side:

$$E_B(x_0) = p(B) \int_x p(x|B)h(x - x_0). \quad (V.2)$$

Hence, the probability to misclassify events, regardless of whether they are signal or background, is just

$$E(x_0) = E_S(x_0) + r E_B(x_0), \quad (V.3)$$

where we have introduced the weight $r$ to allow for the possibility that we may wish to weight the background more (or less) than the signal. We now minimize $E(x_0)$ with respect to the choice of boundary, that is, we set $D_{x_0}E = 0$ and obtain

$$p(S) \int_x p(x|S)\delta(x_0 - x) + p(B) \int_x p(x|B)\delta(x - x_0) = 0. \quad (V.4)$$

The derivative operator has conveniently converted the step functions into delta functions, thereby rendering the integrals trivial. After almost no algebra we can savor our prize; the optimal boundary is obtained as the solution to the equation

$$r(x_0) = \frac{p(x_0|S)p(S)}{p(x_0|B)p(B)}. \quad (V.5)$$
The function \( r(\cdot) \) is called the **Bayes discriminant**, because of its intimate connection with Bayes’ theorem:

\[
p(S|x) = \frac{r}{1 + r} = \frac{p(x|S)p(S)}{p(x|S)p(S) + p(x|B)p(B)}.
\]  

(V.6)

The \( n \)-dimensional generalization of this has the same Bayesian form. The posterior probability \( p(S|x) \) is precisely that needed for event classification. It is the probability that an event characterized by the vector \( x \) is of the signal class. By using this probability we have succeeded in mapping the original \( n \)-dimensional problem into a more tractable one-dimensional one.

This is all very pretty, but there is a serious practical problem. Rarely do we have analytical expressions for the signal and background distributions \( p(x|S) \) and \( p(x|B) \). We seem, alas, to have achieved a pyrrhic victory! Well, this would be the sad conclusion were it not for the existence of some impressive mathematical results obtained in the last few years, regarding artificial neural networks. A neural network is a highly non-linear function that maps \( n \) inputs into one or more outputs. It has been shown\[^{19}\] that, under suitable circumstances, its output is a direct approximation to the probability \( p(S|x) \).

This is a significant conceptual breakthrough: we now know what the network output means. Before this understanding was gained we could enjoy the parody of neural networks as mysterious black boxes that perform miracles in unfathomable ways. The reality is more prosaic: a neural network is an example of a special kind of mathematical function\[^{20}\] that was first studied by Kolmogorov in the 1950s. He described a general way to build functions that could approximate, with arbitrary accuracy, any map of the unit \( n \)-cube into the unit interval. And that’s exactly what we need to do to perform an optimal demarcation between signal and background, which demarcation can be calculated straightforwardly with a neural network, often in a matter of minutes on today’s average desktop computer.
B Prior Probabilities

To solve an inference problem we must assign two probabilities: a prior and a likelihood. Many of the techniques of sampling theory can be used to determine the latter and there is little disagreement about the results. Everyone is happy to use the Poisson, binomial, gamma, $\chi^2$, Gaussian and other standard distributions as appropriate. However, even when people agree that prior probabilities are necessary, they often disagree about how to assign them. The problem is particularly acute when we have minimal prior information about the parameters to be estimated.

The critics of the Bayesian use of Bayes’ theorem charge that the choice of prior probabilities is subjective and arbitrary and, therefore, so too are the inferences derived from the same. It may be that the only prior information we have about a parameter $\theta$ is that it is positive. What prior probability should we assign to various hypotheses about its value? Laplace argued that if we know nothing about the value of a parameter then we should assign a flat prior probability: $P(\theta|I) \propto d\theta$. This seems reasonable, until we realize that any choice of prior probability for a given parameter implies that we have specified, implicitly, the prior probability for the infinity of parameters that are functions of $\theta$. Clearly, we have done a lot more than we bargained for!

For example, suppose we transform from $\theta$ to the parameter $\alpha = 1/\theta$ then mathematical consistency appears to demand that its prior probability distribution be $P(\alpha|I) \propto d\alpha/\alpha^2$; a form that looks, at best, non-intuitive.

This prior probability would be fine were it not for the following question: what reason do we have to suppose that the prior probability is uniform in $\theta$ rather than in the parameter $\alpha$ or some other parameter, like $\tau = \theta^{18}$? It seems that the assignment of prior probabilities for a parameter about which we are almost totally ignorant is, indeed, arbitrary. This is the core of the controversy about prior probabilities that has raged unabated for more than 200 years.

The problem of how to assign prior probabilities to represent a state of “almost complete
ignorance” about the value of a parameter was re-examined early this century by Sir Harold Jeffreys [1]. Later, Jaynes [21] suggested that the prior probability, representing a state of “almost complete ignorance”, should respect the symmetries of the problem. An example will make this clear.

Let $\theta$ be the rate at which accidents occur at a particular intersection. The probability that exactly $n$ accidents occur in a fixed time $t$ is

$$P(n|\theta) = e^{-\theta t} \frac{\theta^t}{n!}.$$  \hspace{0.5cm} (V.7)

We wish to infer something about the parameter $\theta$. All that we know beforehand is that the accident rate is a positive number. Actually, we know a bit more than that: we know that the probability to observe $n$ accidents would remain unchanged whether we measured the rate in accidents per day or in accidents per month, or for that matter whatever unit of time we used. That is, the probability to get $n$ accidents does not change if we apply the transformation $t \rightarrow qt'$ and $\theta \rightarrow \theta' / q$ to the Poisson distribution, where $q$ is an arbitrary positive scale change.

You could choose to measure things in accidents per day, while I might choose to use accidents per week. But, because we have the same state of prior knowledge about the accident rate, namely, we are completely ignorant of its value apart from knowing that it is positive—and because we think rationally, we must, Jaynes suggests, assign the same prior probabilities, just as we assign the same probability to observe $n$ accidents, irrespective of the unit used. Another way to state this is that if two rational people assign the same probability for a given data-set, then for equivalent hypotheses about the accident rate ($\theta'$ in one case and $\theta$ in the other), being rational, they will assign the same prior probabilities to them. The only reason why they might not do so is if they had different prior information; but, by assumption, their prior information is the same.

Okay, let’s for the moment buy this argument. If $P(\theta|I) = h(\theta)d\theta$ is your prior probability and $P(\theta'|I) = f(\theta')d\theta'$ is mine then, according to Jaynes, consistency demands that

$$P(\theta'|I) = P(\theta|I).$$  \hspace{0.5cm} (V.8)
Here the $I$ is absolutely essential, because it must be the same $I$ on both sides of the equation. We get

$$q f(q\theta) = h(\theta). \quad (V.9)$$

The above is true for all $q > 0$; it is true, in particular, for $q = 1$. Setting $q = 1$, in Eq. (V.9), we find that $f(\cdot)$ and $h(\cdot)$ must be the same function. We can, therefore, write

$$qh(q\theta) = h(\theta). \quad (V.10)$$

This is a simple example of a functional equation. Its general solution, to within an arbitrary constant, is $h(\theta) = 1/\theta$. We conclude, therefore, that

$$P(\theta|I) \propto d\theta/\theta, \quad (V.11)$$

is the prior probability that encodes the minimal information we have about the accident rate $\theta$. Such prior probabilities are called non-informative priors. This particular one is called the Jeffreys prior.

One might object that the Jeffreys prior and the uniform prior favored by Laplace do not make sense because they are non-normalizable. This objection is dismissed easily by noting that in the real world we always work within a finite parameter interval; we never go quite to zero nor do we ever go to infinity. Those limits are mathematical idealizations that we use to make our lives simpler. On any realistic interval the prior probabilities can always be normalized. Indeed, they have to be lest they fail to satisfy the laws of probability theory.

This is pretty slick reasoning; perhaps, a bit too slick! Why? Because other equally cogent arguments [32] lead to the prior probability

$$P(\theta|I) \propto d\theta/\sqrt{\theta}. \quad (V.12)$$

And there is something intuitively attractive about the flat prior. So, how is one to choose which prior probability to use? This is the Bayesian conundrum; it is the analog of the frequentist problem of choosing an ensemble. The problem of which prior to use is considered by many to be so severe as to negate any advantage that might be gained by the use of
Bayesian theory. However, I shall now argue that, in practice, the problem is less severe than it seems.

The basic thesis of the advocates of these various non-informative priors is that, for any given prior information, there must be a unique prior probability. But why should that be true? If it is true, then there is only one prior probability for any given problem and all others are wrong. Yet, the many arguments that have been advanced for particular prior probabilities are not all obviously crazy, but give nonetheless different answers. Perhaps, the real problem is the assumption that there must be a unique prior probability. Perhaps, when we have next to no information about something, it should not be expected that even rational beings would agree about how to encode such information, or rather a lack thereof, into a prior probability. It could be that there is an irreducible element of uncertainty when we are faced with reasoning based on almost no information. There is some fuzziness that no amount of mathematical wizardry can "de-fuzzify", as the fuzzy set enthusiasts would say.

What I’m suggesting is this: our search for uniqueness in our inferences may be misguided, because non-uniqueness of outcome may be an inherent property of inductive reasoning. There is uncertainty arising from the data-set as well as from the prior information. Even in the frequentist theory both kinds of uncertainty exist, except that the second source is more deeply hidden. It is hidden in the implicit choice of a flat prior probability when using the maximum likelihood principle. It is hidden in the infinite number of ways we can choose an ensemble.

What matters is that we make reasonable inferences. And, as a practical matter, experience shows that any one of the priors discussed above leads to approximately the same inferences. Furthermore, these inferences merge rapidly as we acquire more data and the likelihood overpowers the prior probability. My advice is this: use the prior probability that seems most reasonable to you or, better, the one that has been agreed upon by the community, for the given problem. Then check the robustness of your answers by trying different reasonable priors. If your answers are unduly sensitive to the choice of prior prob-
ability, then the honest conclusion should be that the data are inadequate and more should be acquired.

No one has succeeded in creating a theory of inference that is devoid of the influence of the inference-maker, in spite of some claims to the contrary. Perhaps such a thing is not possible, even in principle.

C Analyzing A Counting Experiment

Every Bayesian analysis contains at least four ingredients:

- A model
- A data-set
- A likelihood
- A prior probability

For a counting experiment, of which the Grenoble experiment is an example, the model is

\[ n = \theta + \mu, \tag{V.13} \]

where \( n \) is the mean number of events, \( \theta \) the mean signal count and \( \mu \) the mean background count. Let \( N \) be the total number of events observed. What probability should we assign to the data-set \( \{N\} \)?

In a typical high energy physics experiment the number of reactions produced is very large. For example, at the Fermilab Tevatron accelerator more than a trillion collisions were needed to find the top quark. Call that the number of trials \( T \). Suppose that out of the trillions of trials we succeed in recording \( N \) interesting events. What probability should we assign to the data-set? Straightforward combinatoric reasoning, combined with the assumption that at each trial the probability of success \( q \) is the same and that the order
of events is immaterial, leads uniquely to the binomial distribution

\[ P(N|q, T, I) = \binom{T}{N} q^N (1 - q)^{T-N}. \quad (V.14) \]

While rigorously correct, it would be absurd to use the probability in this form because of the fact that \( T \) is usually so much greater than \( N \). Noting that \( n = qT \), by definition, we can write

\[ P(N|q, T, I) = P(N|n, T, I) = \frac{T!}{(T - N)!N!} \frac{n^N}{T^N} (1 - \frac{n}{T})^{T-N}, \quad (V.15) \]

which, for \( T >> N \), becomes the Poisson distribution

\[ P(N|n, I) = \frac{n^N}{N!} \exp(-n)/N!, \quad (V.16) \]

which is the standard likelihood for a counting experiment.

Recall that the Grenoble experiment yielded a total of \( N = 3 \) events. However, that experiment was slightly more complicated in that it also yielded an independent background count of \( B = 7 \) events, a result that caused some grief. So what does the Reverend have to offer?

We begin, as always, with Bayes’ theorem:

\[ \mathbb{P}(\theta, \mu|\mathbf{x}, I) = \frac{\mathbb{P}(|\theta, \mu, I)P(\theta, \mu|I)}{\int_{\theta, \mu} \mathbb{P}(|\theta, \mu|I)P(\theta, \mu|I)}, \quad (V.17) \]

and eliminate the nuisance parameter \( \mu \) by marginalization:

\[ \mathbb{P}(\theta|N, B, I) = \int_{\mu} \mathbb{P}(\theta, \mu|N, B, I). \quad (V.18) \]

The likelihood for the data-set \((N, B)\) is

\[ \mathbb{P}(N, B|\theta, \mu, I) = (\theta + \mu)^N \frac{\exp(-\theta - \mu) \mu B e^{-\mu}}{N! B!}, \quad (V.19) \]

which is the product of two Poisson distributions, one for the count \( N \) and the other for the count \( B \). What about the prior probability \( P(\theta, \mu|I) \)?

Using Jaynes’ invariance argument one can show\[22\] that

\[ P(\theta, \mu|I) = \frac{d\theta d\mu}{(\theta + \mu)\mu}, \quad (V.20) \]
is the appropriate prior probability; appropriate, that is, for this argument. A striking feature of this strange looking prior probability is that it doesn’t factorize: it cannot be written as a product of a function of $\theta$ times a function of $\mu$. So here is another thing to think about: Is this prior reasonable and, if so, what does the non-factorizability mean? For this prior and likelihood the posterior probability can be worked out. The answer is

$$P(\theta|N, B, I) = e^{-\theta} \sum_{i=0}^{N-1} \omega_i \frac{\theta^{N-1-i}}{(N-1-i)!},$$  \hspace{1cm} (V.21)

where

$$\omega_i = \frac{(B-1+i)!/(2^i i!)}{\sum_{j=0}^{N-1} (B-1+j)!/(2^j j!)}. \hspace{1cm} (V.22)$$

Finally, by calculating the mean of the posterior probability we derive

$$\hat{\theta} = d(N, B) = N - \sum_{i=0}^{N-1} i \omega_i,$$  \hspace{1cm} (V.23)

as our best estimate (for a quadratic loss function) of the mean signal count $\theta$. To compute an upper limit $u$ on the mean signal count $\theta$ we need merely to solve

$$\beta = \int_{0}^{u} P(\theta|N, B, I).$$  \hspace{1cm} (V.24)

For $\beta = 0.9$ and $(N, B) = (3, 7)$ we obtain $u(N, B) = 3.3$ events, a perfectly plausible answer. It is possible to re-work the problem using a flat prior probability. Why don’t you try it, compute the 90% upper limit and see by how much it differs from the value given here.

**ACKNOWLEDGEMENTS**

Over the years I have benefited greatly from many discussions, and arguments, with colleagues and friends. I would like to thank in particular Roger Barlow, Maria Fidecaro, Daniel Drijard, Pushpa Bhat, Chip Stewart, Jim Linnemann, Tom Trippe, Rajendran Raja, Marc Paterno, Mark Strovink, Bob Cousins, Louis Lyons and Fred James. The late Edwin Jaynes, whom I never met, was particularly inspiring. His ideas are interesting and he wrote
about them beautifully, avoiding the disease that plagues contemporary technical writing, which strives to be as impersonal as possible and succeeds to a fault.

REFERENCES


[7] XXXX


[36] In fact, different frequentist authors had differing notions of “frequentist” probability. For a lucid discussion of this point see chapter VII of Ref. [1].

[37] Posthumously as it turned out—he died in 1761, even though there is some evidence that he had figured out his result as early as 1748.

[38] We have used the binomial expansion theorem \((a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}\).

[39] In fact, Fisher was not the first to use maximum likelihood. It was first used by Gauss. However, it was Fisher who developed it into a powerful statistical technique.

[40] This structure is what mathematicians call a *fibre bundle*: it is a *base space*, here the space of parameters — one of whose coordinates is \(\theta\), to each point of which is attached a *fibre*, here an ensemble of intervals.

[41] The authors have reported, however, that for the problems they have studied, coverage
is closely satisfied.