Some New Results on Statistical Information and Evidence

By
Maryam Moghimi[1]

Presented to the Faculty of the Graduate School of The University of Texas at Arlington in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF TEXAS AT ARLINGTON
May 2020

Supervising Committee:

Dr. Bill Corley
Dr. Victoria Chen
Dr. Jay Rosenberger
Dr. Shan Sun-Mitchell

[1] Department of Industrial and Manufacturing Systems Engineering, The University of Texas at Arlington, TX 76019 USA
Maryam.moghimi@mavs.uta.edu

Acknowledgements

At the early steps of my Ph.D. research, it became obvious to me that a researcher cannot and does not work without the help and support of their peers. While the list of individuals I wish to thank extends beyond the limits of this format, I would like to acknowledge the dedication and support of the following people:

First, I express my deep sense of gratitude to my supervising professor Dr. Bill Corley, who has the attitude and the substance of true professional researchers. Without his guidance and continuous support, I would not be where I am.

I also thank the members of my Ph.D. committee, Dr. Victoria Chen, Dr. Jay Rosenberger and Dr. Shan Sun-Mitchell, for their valuable suggestions.

Finally, I warmly thank my current and former colleagues at the center on Stochastic Modeling, Optimization, & Statistics (COSMOS) who helped me with my research and shared memorable times: Dr. Nilabh Ohol, Ms. Nahal Sakhavand, Dr. Alireza Fallahi, Dr. Ashkan Aliabadi Farahani, and Mr. Shiris Roa.

Maryam Moghimi
May 2020

Dedication

To my parents, Fariba and Ali, who made me leave my country and get a Ph.D. Hopefully, they will let me visit after this.

To my little sister, Mahya, who believed in me much as a 4-year-old believes in Wonder Woman.

To my professor, Dr. Corley, who taught me not only that I should think outside the box for solutions but that I should sometimes check under the box.

And finally to the novel coronavirus, which made me stay home and finish my dissertation.

Maryam Moghimi
May 2020

Some New Results on Statistical Information and Evidence

Maryam Moghimi

The University of Texas at Arlington, 2020

Supervising Professor: Dr. Bill Corley

Table of Contents

Chapter 1 ..... **7**

Introduction **7**

Chapter 2 ..... **9**

Information Loss due to the Compression of Sample Data from Discrete Distributions **9**

Chapter 3 ..... **26**

New Concepts in Information Theory with Applications in Data Analysis **26**

Chapter 4 ..... **41**

Comparison and Extension of Measures of Evidence in Hypothesis Testing **41**

Chapter 5 ..... **57**

General Conclusions **57**

Chapter 1

Introduction

"I just invent. Then I wait until man comes around to needing what I've invented."

— Buckminster Fuller

This dissertation represents an attempt to relate some fundamental statistical problems using the notions of information, entropy, and evidence. The dissertation is presented in the format of article-based dissertation including 3 papers.

The first paper is entitled "Information Loss due to the Compression of Sample Data from Discrete Distributions". It is a study about the information lost when a real-valued statistic $T(X_1, \ldots, X_n)$ is used to summarize the sample data $\mathbf{x} = (x_1, \ldots, x_n)$ of a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from a discrete random variable X with a one-dimensional parameter θ. The process where the data sample X is compressed to the summary statistic T(X) is irreversible and always involves some information loss. For instance, if $T(\mathbf{X}) = \overline{X}$, the original measurements x cannot be reconstructed from $\bar{\mathbf{x}}$, and some information about $\mathbf{x}$ is lost. Nonetheless, such data compression is frequently used to make inferences about, for example, the true mean μ of X.

In general, the first paper presents a decomposition on the total information available about $\mathbf{X}$ in $\mathbf{x}$ and give various expressions for the Shannon information lost by compressing $\mathbf{x}$ to $T(\mathbf{x})$. The focus is on sufficient statistics for the parameter θ, which are used to develop a general formula independent of θ for this lost information as well as for an associated entropy that depends only on T. This approach would also work for non-sufficient statistics, but the lost information and associated entropy would involve θ**.**

The second paper is entitled "New Concepts in Information Theory with Applications in Data Analysis". The previous paper involved an examination of Shannon information and entropy for discrete random variables. The second paper then continues the study and develops new concepts of information and entropy. In particular, a new type of information called gambler's information is introduced, which views events prospectively, as compared with Shannon information, which views them retrospectively. The Shannon information of an event is defined as the negative log to the base 2 of the probability p for the event. Based on this definition, an observer obtains more information if an unlikely event occurs than if a likely one does.

Shannon information may not be the appropriate information for modeling some decisions. The Shannon information of an event is not obtained until the event actually occurs and causes a level of surprise appropriate to the likelihood of it occurring. On the other hand, gambler's information stems from the probability of the event and not its occurrence. The notion of outer entropy is also defined in which the log in Shannon (or inner) entropy is placed outside the expectation to provide simpler calculations and often intuitive results. As applications of these new concepts, they are used instead of Shannon information and inner entropy in the framework of paper 1 and then also used to provide intuitive measures of evidence for the parameter values of a discrete random variable.

The third paper is entitled "Comparison and Extension of Measures of Evidence in Hypothesis Testing". A principal goal of statistics is to obtain evidence from data for comparing alternative decisions. For

example, one may need to decide whether a population mean μ satisfies $\mu \leq \mu_0$ as opposed to $\mu > \mu_0$ for some specified μ_0. There are numerous attempts to define evidence in statistics. This paper considers the likelihood ratio, confidence distributions, the Bayesian posterior odds ratio, and P-value as measures of evidence. Moreover, a new definition of P-value utilizing the maximum likelihood estimator is proposed that in the limit obtains the frequentist probability that a null hypothesis is true without reference to error or test statistics.

The dissertation is organized as follows. Chapter 2 includes the study on information loss due to the compression of sample data from discrete distributions. In Chapter 3, new concepts in information theory are presented. In Chapter 4, various measures of evidence, including a new P-value, are presented and compared. General conclusions are offered in Chapter 5.

Chapter 2

Information Loss due to the Compression of Sample Data from Discrete Distributions

Maryam Moghimi*[1,2], H.W. Corley[1,2]

[1] Center on Stochastic Modeling, Optimization, and Statistics (COSMOS), The University of Texas at Arlington, Arlington, TX, USA
[2] The authors contributed equally to this paper.
* Correspondence: maryam.moghimi@uta.edu; +1-214-971-0904 (M.M.), corley@uta.edu; Tel.: +1-817-272-3092 (H.C.)

Abstract: In this paper we study the information lost when a real-valued statistic $T(X_1, \ldots, X_n)$ is used to summarize the sample data $\mathbf{x} = (x_1, \ldots, x_n)$ of a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from a discrete random variable X with a one-dimensional parameter θ. We compare the probability that the random sample $\mathbf{X}$ yields $\mathbf{x}$ to the probability that the compressed sample $T(\mathbf{X})$ yields $T(\mathbf{x})$. The former probability measures the total information about $\mathbf{x}$, while the latter measures the compressed information about $\mathbf{x}$, both of which are expressed here as Shannon information. The difference is the information lost about $\mathbf{X}$ by its compression to $T(\mathbf{X})$. We focus on sufficient statistics for the parameter θ and develop a general formula independent of θ for this lost information as well as for an associated entropy that depends only on T. Our approach would also work for non-sufficient statistics, but the lost information and associated entropy would involve θ. Examples are presented for some standard discrete distributions.

Keywords: discrete distributions, Shannon information, lost information, sampling, data reduction, data compression, entropy, sufficient statistics, likelihood

1. Introduction

We consider the data sample $\mathbf{x} = (x_1, \ldots, x_n)$ from a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ for a discrete random variable X with sample space S and one-dimensional parameter θ. Here a statistic $T(\mathbf{X})$ is a real-valued function of the random sample but not a function of any parameter θ associated with X, though θ may fixed at an arbitrary value. The data sample $\mathbf{X}$ is compressed to the summary statistic $T(\mathbf{X})$, which could be used to characterize $\mathbf{X}$ or to estimate θ. Such data compression is an irreversible process [1] and always involves some information loss. For instance, if $T(\mathbf{X}) = \bar{X}$, the original measurements $\mathbf{x}$ cannot be reconstructed from $\bar{x}$, and some information about $\mathbf{x}$ is lost. Nonetheless, such data compression is frequently used to make inferences about, for example, the true mean μ of X. Our information-theoretic approach to data compression generalizes the observation in [2] that a binomial random variable loses all the information about the order of successes in the associated sequence of Bernoulli trials.

For any real-valued statistic T and the given sample data $\mathbf{x}$, we decompose the total information about $\mathbf{X}$ available in $\mathbf{x}$ into the sum of (a) the information available in the compressed data $T(\mathbf{x}) = \bar{x}$ and (b) the information lost in the compression. When T is a sufficient statistic for θ this lost information is independent of θ. Moreover, by taking the expected value of this lost information over all possible data sets, we define an associated entropy measure that depends on T but neither $\mathbf{x}$ nor θ. Our approach also works for non-sufficient statistics, but the lost information and associated entropy would then involve θ, and so θ must be estimated to computing these quantities.

The paper is organized as follows. In Section 2, we present the necessary definitions, notation, and preliminary results. In Section 3, we decompose the total information available about $\mathbf{X}$ in $\mathbf{x}$ and give various expressions for the Shannon information lost by compressing $\mathbf{x}$ to $T(\mathbf{x})$. In Section 4, we develop an entropy measure associated with this lost information**.** In Section 5, we present examples of our results for some standard discrete distributions and several statistics sufficient for θ. Conclusions are offered in Section 6.

2. Preliminaries

The following definitions, notation, and results are used here. Further details can be found in [3,4] and elsewhere. An important class of statistics is first defined.

Definition 2.1 (Sufficient Statistic). A statistic $T(\mathbf{X})$ is a sufficient statistic (SS) for the parameter θ if the probability

$$P[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})] \tag{1}$$

is independent of θ.

Note that P instead of P_θ is used in (1) since this probability is independent of θ. Also observe that (1) is not a joint conditional distribution for $\mathbf{X}$ since its n condition changes with $\mathbf{x}$. This observation becomes significant in Section 4. The fact that (1) does not involve θ is used to prove the Fisher Factorization Theorem (FFT), which is the usual method for determining if a statistic is an SS for θ. We use the notation $f(\mathbf{x}|\theta)$ to denote the joint pmf of $\mathbf{X}$ evaluated at the variable $\mathbf{x}$ for a fixed value of θ.

Result 2.2 (Fisher Factorization Theorem). The real-valued statistic $T(\mathbf{X})$ is sufficient for θ if and only if there exist functions $g: R^1 \longrightarrow R^1$ and $h: S^n \longrightarrow R^1$ such that for any sample data $\mathbf{x}$ and for all values of θ the joint pmf $f(\mathbf{x}|\theta)$ of $\mathbf{X}$ can be factored as

$$f(\mathbf{x}|\theta) = g[T(\mathbf{x})|\theta] \times h(\mathbf{x}) \tag{2}$$

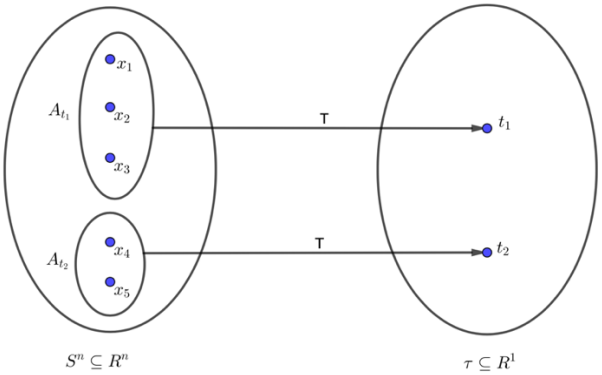
for real-valued, nonnegative functions g on R^1 and h on S^n. The function h does not depend on θ, while g does depend on $\mathbf{x}$ but only through $T(\mathbf{x})$.

We focus on a sufficient statistic T for θ in Section 3, where we need the notion of a partition [5] as defined next.

Definition 2.3 (Partition). Let S be the denumerable sample space of the discrete random variable X, and thus let S^n be the denumerable sample space of the random sample $\mathbf{X}$. For any statistic $T: S^n \longrightarrow R^1$, let τ_T be the denumerable set $\tau_T = \{t|\exists \mathbf{x} \in S^n \text{ for which } t = T(\mathbf{x})\}$, which is the range of T. Then T partitions the sample space S^n into the mutually exclusive and collectively exhaustive partition sets $A_t = \{\mathbf{x} \in S^n | T(\mathbf{x}) = t\}, \forall t \in \tau_T$.

Figure 2.1 below illustrates the situation.

Figure 2.1



We also need the well-known likelihood function.

Definition 2.4 (Likelihood Function). Let $\mathbf{x}$ be sample data from a random sample $\mathbf{X}$ from a discrete random variable X with sample space S and real-valued parameter θ, and let $f(\mathbf{x}|\theta)$ denote the joint pmf of the random sample $\mathbf{X}$. For any sample data $\mathbf{x}$, the likelihood function of θ is defined as

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta). \tag{3}$$

The likelihood function in (3) is a function of the variable θ for given data $\mathbf{x}$. However, the joint pmf $f(\mathbf{x}|\theta)$ as a function of $\mathbf{x}$ for fixed θ is frequently called the likelihood function as well. In this case we also write the joint pmf as $L(\mathbf{x}|\theta)$. We distinguish the two cases since $L(\theta|\mathbf{x})$ is not a statistic but $L(\mathbf{x}|\theta)$ is one that incorporates all available information about $\mathbf{X}$. Moreover, $L(\mathbf{x}|\theta)$ is an SS for θ [4] and uniquely determines an associated SS called the likelihood kernel to be used in subsequent examples.

Definition 2.5 (Likelihood kernel). Let S be the sample space of $\mathbf{X}$. For fixed θ, suppose that $L(\mathbf{x}|\theta)$ can be factored as

$$L(\mathbf{x}|\theta) = K(\mathbf{x}|\theta) \times R(\mathbf{x}),\ \forall \mathbf{x} \in S^n, \tag{4}$$

where $K: S^n \longrightarrow R^1$ and $R: S^n \longrightarrow R^1$ have the following properties:

(a) every nonnumerical factor of $K(\mathbf{x}|\theta)$ contains θ;

(b) $R(\mathbf{x})$ does not contain θ;

(c) for $\forall \mathbf{x} \in S^n$, both $K(\mathbf{x}|\theta) \geq 0$ and $R(\mathbf{x}) \geq 0$; and

(d) $K(\mathbf{x}|\theta)$ is not divisible by any positive number except 1.

Then $K(\mathbf{x}|\theta)$ is defined as the likelihood kernel of $L(\mathbf{x}|\theta)$ and $R(\mathbf{x})$ as the residue of $L(\mathbf{x}|\theta)$.

Theorem 2.6. The likelihood kernel $K(\mathbf{x}|\theta)$ has the following properties.

(i) $K(\mathbf{x}|\theta)$ uniquely exists.

(ii) $K(\mathbf{x}|\theta)$ is an SS for θ.

(iii) For any θ_1 and θ_2, the likelihood ratio $\frac{L(\mathbf{x}|\theta_1)}{L(\mathbf{x}|\theta_2)}$ equals $\frac{K(\mathbf{x}|\theta_1)}{K(\mathbf{x}|\theta_2)}$.

Proof. To prove (i), for fixed θ we first show that the likelihood kernel $K(\mathbf{x}|\theta)$ of **Definition 2.5** exists by construction. Since the formula for $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$ must explicitly contain θ, the parameter θ cannot appear only in the range of $\mathbf{x}$. Hence $L(\mathbf{x}|\theta)$ as a function of $\mathbf{x}$ can be factored into $K(\mathbf{x}|\theta) \times R(\mathbf{x})$ satisfying (a) and (b) of **Definition 2.5,** where $K(\mathbf{x}|\theta) \geq 0, \forall \mathbf{x} \in S^n$, and the numerical factor of $K(\mathbf{x}|\theta)$ is either $+1$ or -1. Then $R(\mathbf{x}) \geq 0,\ \forall \mathbf{x} \in S^n$, since $K(\mathbf{x}|\theta) \geq 0,\ \forall \mathbf{x} \in S^n$, and $K(\mathbf{x}|\theta) \times R(\mathbf{x}) = f(\mathbf{x}|\theta) \geq 0$. Thus (c) is satisfied. Finally, the only positive integer that evenly divides $+1$ or -1 is 1, so (d) holds. It follows that the likelihood kernel $K(\mathbf{x}|\theta)$ and its associated $R(\mathbf{x})$ in **Definition 2.5** are well defined and exist.

We next show that $K(\mathbf{x}|\theta)$ as constructed above is unique. Let $K_1(\mathbf{x}|\theta)$ with residue $R_1(\mathbf{x})$ and $K_2(\mathbf{x}|\theta)$ with $R_2(\mathbf{x})$ both satisfy **Definition 2.5.** Thus for $j = 1,2,\ R_j(\mathbf{x})$ does not contain θ while every nonnumerical factor of $K_j(\mathbf{x}|\theta)$ does contain θ. It follows that $K_1(\mathbf{x}|\theta) \geq 0$ and $K_2(\mathbf{x}|\theta) \geq 0$ must be identical or else be a positive multiple of one another. Assume that $K_2(\mathbf{x}|\theta) = \lambda K_1(\mathbf{x}|\theta)$ for some $\lambda > 0$. If $\lambda \neq 1, K_2(\mathbf{x}|\theta)$ is divisible by a positive number other than 1 to avoid (d). Thus, $K(\mathbf{x}|\theta)$ is unique.

To prove (ii) we show that this unique $K(\mathbf{x}|\theta)$ is an SS for θ. For $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$, let $g[z] = z$ and $h(\mathbf{x}) = R(\mathbf{x})$ in (2). Then $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = g[K(\mathbf{x}|\theta)] \times h(\mathbf{x}) = K(\mathbf{x}|\theta) \times R(\mathbf{x})$. Thus $K(\mathbf{x}|\theta)$ is an SS by the FFT of **Result 2.2**.

Finally, (iii) follows immediately from **Definition 2.5** and the fact that $L(\mathbf{x}|\theta_2) \neq 0$ for $\mathbf{x} \in S^n$. ■

We next discuss the notion of information to be used here. Actually, probability itself is a measure of information in the sense that it captures the surprise level of an event. An observer obtains more information, i.e., surprise, if an unlikely event occurs than if a likely one does. Instead of probability, however, we use the additive measure known as Shannon information [6, 7] defined as follows.

Definition 2.7 (Shannon Information). Let $\mathbf{x}$ be sample data for the random sample $\mathbf{X}$ from the discrete random variable X with a one-dimensional parameter θ, and let $f(\mathbf{x}|\theta)$ be the joint pmf of $\mathbf{X}$ at $\mathbf{x}$. The Shannon information obtained from the sample data $\mathbf{x}$ is defined as

$$I(\mathbf{x}|\theta) = -\log f(\mathbf{x}|\theta), \tag{5}$$

where the units of $I(\mathbf{x}|\theta)$ is bits if the base of the logarithm is 2, which is to be used here.

The expected information over $\forall \mathbf{x} \in S^n$ will also be used.

Definition 2.8 (Entropy). Under the conditions of **Definition 2.7**, the entropy $H(\mathbf{X}|\theta)$ is defined as the expected value of $I(\mathbf{X}|\theta)$; i.e,

$$H(\mathbf{X}|\theta) = \sum_{\mathbf{x}} f(\mathbf{x}|\theta) I(\mathbf{x}|\theta). \tag{6}$$

Since entropy is the expected information over all possible random samples, it measures the available information about $\mathbf{X}$ better than would a single data set $\mathbf{x}$, which might not be typical [8]. We next give a method to obtain the information loss about $\mathbf{X}$ that occurs when a data set $\mathbf{x}$ is compressed to $T(\mathbf{x})$. In our approach, we focus on a sufficient statistic T so there will be no θ in (5) for the lost information below. However, our approach is applicable to a non-sufficient statistic as well if θ is estimated from the data.

3. Information Decomposition under Data Compression by a Real-Valued Statistic

We now develop a procedure to determine how much information about $\mathbf{X}$ contained in a data set $\mathbf{x}$ is lost when the data is compressed to $T(\mathbf{x})$ by the sufficient statistic T. Consider the joint conditional probability

$$P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})], \tag{7}$$

which is identified with the probabilistic information lost about the event $\mathbf{X} = \mathbf{x}$ by the data compression of $\mathbf{x}$ to $T(\mathbf{x})$. The notation P_θ refers to the fact that the discrete probability (7) in general involves the parameter θ. We next express (7) using the definition of conditional probability to obtain the basis of our development. **Result 3.1** is given in [3, p. 273] and proved below to illustrate the reasoning.

Result 3.1. Let $\mathbf{x}$ be sample data for a random sample $\mathbf{X}$ from a discrete random variable X with sample space S and real-valued parameter θ, and let $T(\mathbf{X})$ be any real-valued statistic. Then

$$P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})] = \frac{P_\theta[\mathbf{X} = \mathbf{x}]}{P_\theta[T(\mathbf{X}) = T(\mathbf{x})]}. \tag{8}$$

Proof. Using the definition of conditional probability, rewrite (7) as

$$\frac{P_\theta[\mathbf{X} = \mathbf{x}; T(\mathbf{X}) = T(\mathbf{x})]}{P_\theta[T(\mathbf{X}) = T(\mathbf{x})]}. \tag{9}$$

But $T(\mathbf{X}) = T(\mathbf{x})$ whenever $\mathbf{X} = \mathbf{x}$, so (8) follows. ■

Observe that if T is an SS for θ, the right side of (8) is independent of θ and hence so is the left. Now taking the negative logarithm of (8) and rearranging terms gives

$$-\log P_\theta[\mathbf{X} = \mathbf{x}] = -\log P_\theta[T(\mathbf{X}) = T(\mathbf{x})] - \log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]. \tag{10}$$

From (8) note that $P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})] \geq P_\theta[\mathbf{X} = \mathbf{x}]$ since $P_\theta[T(\mathbf{X}) = T(\mathbf{x})] \leq 1$, so $-\log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})] \leq -\log P_\theta[\mathbf{X} = \mathbf{x}]$. Similarly, $-\log P_\theta[T(\mathbf{X}) = T(\mathbf{x})] \leq -\log P_\theta[\mathbf{X} = \mathbf{x}]$. These facts suggest that the left side of (10) is the total Shannon information in bits about $\mathbf{X}$ contained in the sample data $\mathbf{x}$. On the right side of (10), the term $-\log P_\theta[T(\mathbf{X}) = T(\mathbf{x})]$ is considered the information about $\mathbf{X}$ contained in the compressed data summary $T(\mathbf{x})$, and the term $-\log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]$ is identified as the information about $\mathbf{X}$ that has been lost as the result of the data compression by $T(\mathbf{x})$.

In particular, this lost information represents a combinatorial loss in the sense that multiple $\mathbf{x}$'s may give the same value $T(\mathbf{x}) = t$ as depicted in **Figure 2.1** above. In other words, the lost information $-\log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]$ is a measure of the knowledge unavailable about the data sample $\mathbf{x}$ when only

the compressed data summary $T(\mathbf{x})$ is known and not $\mathbf{x}$ itself. For a sufficient statistic $T(\mathbf{X})$ for θ, this lost information is independent of θ. It is a characteristic of $T(\mathbf{X})$ for the given data sample $\mathbf{x}$.

In terms of **Figure 2.1** above, the situation may be described as follows. On the left is the sample space $S^n \subseteq R^n$ over which probabilities on $\mathbf{X}$ are computed. On the right is the range $\tau_T \subseteq R^1$ of T over which the probability of $T(\mathbf{X})$ are computed. T compresses the data sample $\mathbf{x}$ into $T(\mathbf{x})$, where multiple $\mathbf{x}$'s may give the same $T(\mathbf{x}) = t$. In **Figure 2.1** the distinct data samples $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ are all compressed into the same value t_1. But knowing that $T(\mathbf{x}) = t_1$ for some data sample $\mathbf{x}$ does not provide sufficient information to know unequivocally, for example, that $\mathbf{x} = \mathbf{x}_1$. Information is lost in the compression. One can also say that the total information $-\log P_\theta[\mathbf{X} = \mathbf{x}]$ deriving from the left side of **Figure 2.1** is compressed to $-\log P_\theta[T(\mathbf{X}) = T(\mathbf{x})]$ deriving from the right. The reduction of information from the left to the right side is precisely the lost information $-\log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]$. For fixed t, it is lost due to the ambiguity as to which data sample on the left actually gave t when only t is known. There is no ambiguity when T is one-to-one.

The general decomposition of information in (10) is next summarized in **Definition 3.2**, where T does not need to be sufficient for θ.

Definition 3.2 ($\mathbf{I_{total}}$, $\mathbf{I_{comp}}$, $\mathbf{I_{lost}}$). Let $\mathbf{x}$ be sample data for a random sample $\mathbf{X}$ from a discrete random variable X with sample space S and real-valued parameter θ. For any real-valued statistic $T(\mathbf{X})$, the Shannon information about $\mathbf{X}$ obtained from the sample data $\mathbf{x}$ can be decomposed as

$$I_{total}(\mathbf{x}|\theta) = I_{comp}(\mathbf{x}|\theta, T) + I_{lost}(\mathbf{x}|\theta, T), \tag{11}$$

where

$$I_{total}(\mathbf{x}|\theta) = -\log P_\theta[\mathbf{X} = \mathbf{x}], \tag{12}$$

$$I_{comp}(\mathbf{x}|\theta, T) = -\log P_\theta[T(\mathbf{X}) = T(\mathbf{x})], \tag{13}$$

and

$$I_{lost}(\mathbf{x}|\theta, T) = -\log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]. \tag{14}$$

Both **Result 3.1** and **Definition 3.2** are valid for any real-valued statistic for $\mathbf{X}$. The notation $I_{total}(\mathbf{x}|\theta)$ indicates that I_{total} is a function of the sample data $\mathbf{x}$ for a fixed but arbitrary parameter value θ. Similarly, both $I_{comp}(\mathbf{x}|\theta, T)$ and $I_{lost}(\mathbf{x}|\theta, T)$ are functions of $\mathbf{x}$ for fixed θ and T. However, in this paper we focus on sufficient statistics, which provide a simpler expression for $I_{lost}(\mathbf{x}|\theta, T)$ that does not involve θ. For a sufficient statistic T for θ, we use the notation $I_{lost}(\mathbf{x}|T)$ for the lost information, though $I_{total}(\mathbf{x}|\theta)$ and $I_{comp}(\mathbf{x}|\theta, T)$ still require θ. The next result is an application of the FFT of **Result 2.2**.

Theorem 3.3 (Lost Information for an SS). Let $\mathbf{x}$ be sample data for a random sample $\mathbf{X}$ from a discrete random variable X with sample space S and real-valued parameter θ. Let T be an SS for θ, $f(\mathbf{x}|\theta)$ be the joint pmf of $\mathbf{X}$, and $f(\mathbf{x}|\theta) = g[T(\mathbf{x})|\theta] \times h(\mathbf{x})$ as in **Result 2.2**. Then for all $\mathbf{x} \in S^n$

$$I_{lost}(\mathbf{x}|T) = -\log \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \in A_{T(\mathbf{x})}} h(\mathbf{y})}, \tag{15}$$

where $A_{T(\mathbf{x})}$ is defined in **Definition 2.3** for $t = T(\mathbf{x})$.

Proof. Let $\mathbf{x} \in S^n$. Then $f(\mathbf{x}|\theta) > 0$ since $\mathbf{x}$ is a realization of $\mathbf{X}$. Since T is an SS, we write (7) without θ. Then it suffices to establish that

$$P[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})] = \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \in A_{T(\mathbf{x})}} h(\mathbf{y})}, \tag{16}$$

from which (15) immediately follows. Rewrite (8) as

$$P[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})] = \frac{P_\theta[\mathbf{X} = \mathbf{x}]}{P_\theta[T(\mathbf{X}) = T(\mathbf{x})]} = \frac{f(\mathbf{x}|\theta)}{\sum_{\mathbf{y} \in A_{T(\mathbf{x})}} f(\mathbf{y}|\theta)}, \tag{17}$$

so from (17) and (2), then

$$P[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x})]=\frac{g[T(\mathbf{x})|\theta]\times h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} g[T(\mathbf{y})|\theta]\times h(\mathbf{y})}. \tag{18}$$

But $T(\mathbf{y})=T(\mathbf{x}), \forall \mathbf{y}\in A_{T(\mathbf{x})}$ in (18), so

$$P[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x})]=\frac{g[T(\mathbf{x})|\theta]\times h(\mathbf{x})}{g[T(\mathbf{x})|\theta]\times \sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})}, \forall \mathbf{x}\in S^n. \tag{19}$$

Since $f(\mathbf{x}|\theta)>0$ and hence $g[T(\mathbf{x})|\theta]\neq 0$, this term can be canceled on the right side of (19) to yield (16). Taking $-\log$ of (16) completes the proof. ■

Now consider **Theorem 3.3** when each A_t is a singleton in (16), i.e., when T is a one-to-one function. In this extreme case, $P[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x})]=1$ since $\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})=h(\mathbf{x})$ in the denominator of the right side of (16). Thus $I_{lost}(\mathbf{x}|T)=0$ from which $I_{comp}(\mathbf{x}|\theta,T)=I_{total}(\mathbf{x}|\theta)$ for all $\mathbf{x}$ in S^n. Thus the special case of a one-to-one T justifies the identification of the lost information as $I_{lost}(\mathbf{x}|\theta,T)=-\log P_\theta[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x})]$. In other words, for all data samples $\mathbf{x},\mathbf{y}\in S^n$, if $\mathbf{x}\neq\mathbf{y}$ whenever $T(\mathbf{x})\neq T(\mathbf{y})$, then $P[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x})]$ is not diminished by the compression of the singleton $A_{T(\mathbf{x})}$ to the number $T(\mathbf{x})$.

More generally, it is also true that $I_{lost}(\mathbf{x}|\theta,T)=0$ when T is one-to-one but not sufficient for θ. In this case, write $P_\theta[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x})]=\frac{P_\theta[\mathbf{X}=\mathbf{x}]}{P_\theta[T(\mathbf{X})=T(\mathbf{x})]}=\frac{f(\mathbf{x}|\theta)}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} f(\mathbf{y}|\theta)}$. But since T is one-to-one, $\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} f(\mathbf{y}|\theta)=f(\mathbf{x}|\theta)$, $P_\theta[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x})]=1$, and again $I_{lost}(\mathbf{x}|\theta,T)=0$.

Now consider the other extreme case where $T(\mathbf{x})=c$ is constant on S^n. Then $P_\theta[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=c]=\frac{P_\theta[\mathbf{X}=\mathbf{x}]}{P_\theta[T(\mathbf{X})=c]}$. But $P_\theta[T(\mathbf{X})=c]=1$, so $P_\theta[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=c]=P_\theta[\mathbf{X}=\mathbf{x}]$ and $I_{lost}(\mathbf{x}|\theta,T)=I_{total}(\mathbf{x}|\theta,T)$ on S^n. In this case, $I_{comp}(\mathbf{x}|\theta,T)=0$ because the event $T(\mathbf{x})=c$ gives no information about $\mathbf{x}$.

Next, in the following corollary we show that (16) can be simplified when T is the likelihood function.

Corollary 3.4 (Information Loss for Likelihood Function). Under the assumptions of **Theorem 3.3**, if $T(\mathbf{x})=L(\mathbf{x}|\theta)$, then

$$I_{lost}(\mathbf{x}|L)=-\log\frac{1}{\left|A_{L(\mathbf{x}|\theta)}\right|}, \tag{20}$$

where $\left|A_{L(\mathbf{x}|\theta)}\right|$ is the cardinality of the partition set A_t for $t=L(\mathbf{x}|\theta)$.

Proof. For $T(\mathbf{x})=L(\mathbf{x}|\theta)=f(\mathbf{x}|\theta)$ in (2), let g be the identity function and $h(\mathbf{x})=1$. Then substituting $h(\mathbf{x})=1$ into (16) gives the denominator $\sum_{\mathbf{y}\in A_{L(\mathbf{x}|\theta)}} 1=\left|A_{L(\mathbf{x}|\theta)}\right|$ to yield (20). ■

We next state a reproductive property of a statistic T′ that is a one-to-one function of a sufficient statistic T for θ.

Theorem 3.5. If there is a one-to-one function between a sufficient statistic T for θ and an arbitrary real-valued statistic T′ on S^n, the following hold.

(i) T′ is also an SS.
(ii) T and T′ partition the sample space S into the same partition sets.
(iii) $I_{lost}(\mathbf{x}|T)=I_{lost}(\mathbf{x}|T'), \forall \mathbf{x}\in S^n$.

Proof. To prove (i), let u be a real-valued one-to-one function of T′ such that

$$T(\mathbf{x})=u[T'(\mathbf{x})]. \tag{21}$$

Since T is an SS, by equation (2) there are real-valued functions g on R^1 and h on S^n for which

$$f(\mathbf{x}|\theta)=g[T(\mathbf{x})|\theta]\times h(\mathbf{x}). \tag{22}$$

By substituting $T(\mathbf{x})$ from (21) in (22), we get

$$f(\mathbf{x}|\theta) = g(u[T'(\mathbf{x})]|\theta) \times h(\mathbf{x}), \tag{23}$$

which can be rewritten as

$$f(\mathbf{x}|\theta) = (g \circ u)[T'(\mathbf{x})|\theta] \times h(\mathbf{x}). \tag{24}$$

Since T' in (24) satisfies the condition of **Result 2.2** for $g' = g \circ u$, T' is an SS.

To prove (ii), we use **Definition 2.3**. Let T partition the sample space S^n into the mutually exclusive and collectively exhaustive sets $A_t = \{\mathbf{x}|T(\mathbf{x}) = t\}, \forall t \in \tau_T$. By equation (21) we can also write A_t as

$$A_t = \{\mathbf{x}|u[T'(\mathbf{x})] = t\}, \forall t \in \tau_T. \tag{25}$$

Since u is a one-to-one function, it has an inverse u^{-1}. Letting $u^{-1}(t) = t'$, we apply u^{-1} to the right side of (25) and get

$$A_t = \{\mathbf{x}|T'(\mathbf{x}) = t'\}, \forall t' \in u(\tau_T). \tag{26}$$

But $u(\tau_T) = \tau_{T'}$ and the cardinalities $|\tau_T| = |\tau_{T'}|$, so the right side of (26) is $A_{t'}$ and

$$A_t = A_{t'}. \tag{27}$$

Finally, to get (iii) we use **Theorem 3.3** to calculate information lost over two statistics T and T'. Since $h(\mathbf{x})$ is the same in (22) and (24) and since equation (27) holds, we sum $h(\mathbf{x})$ over the same sets in the denominator of equation (16) for both T and T' to give

$$I_{lost}(\mathbf{x}|T) = I_{lost}(\mathbf{x}|T') \tag{28}$$

and complete the proof. ■

We next compare the information loss of the sufficient statistic $L(\mathbf{x}|\theta)$ to other sufficient statistics. For the sufficient statistic $K(\mathbf{x}|\theta)$, a lemma is needed.

Lemma 3.6. Let $\mathbf{x}$ be any data sample for a random sample $\mathbf{X}$ from the discrete random variable X with real-valued parameter θ. Then $K(\mathbf{x}|\theta)$ is a function of $L(\mathbf{x}|\theta)$ and $\tau_L \geq \tau_K$.

Proof. From [3, p. 280], $K(\mathbf{x}|\theta)$ is a function of $L(\mathbf{x}|\theta)$ if and only if $K(\mathbf{x}|\theta) = K(\mathbf{y}|\theta)$ whenever $L(\mathbf{x}|\theta) = L(\mathbf{y}|\theta)$. For all data samples $\mathbf{x}$ and $\mathbf{y}$, we thus prove that if $L(\mathbf{x}|\theta) = L(\mathbf{y}|\theta)$, then $K(\mathbf{x}|\theta) = K(\mathbf{y}|\theta)$. Thus suppose that $L(\mathbf{x}|\theta) = L(\mathbf{y}|\theta)$. By **Definition 2.5** we can decompose $L(\mathbf{x}|\theta)$ and $L(\mathbf{y}|\theta)$ into $K(\mathbf{x}|\theta)R(\mathbf{x})$ and $K(\mathbf{y}|\theta)R(\mathbf{y})$, respectively. Note that $K(\mathbf{y}|\theta) \neq 0$. Otherwise $L(\mathbf{y}|\theta) = 0$ in contradiction to $\mathbf{y}$ being sample data with a nonzero probability of occurring. Write

$$\frac{K(\mathbf{x}|\theta)}{K(\mathbf{y}|\theta)} = \frac{R(\mathbf{y})}{R(\mathbf{x})}. \tag{29}$$

Suppose that $K(\mathbf{x}|\theta) \neq K(\mathbf{y}|\theta)$ so that $\frac{K(\mathbf{x}|\theta)}{K(\mathbf{y}|\theta)} = \frac{R(\mathbf{y})}{R(\mathbf{x})} \neq 1$ in (29). From **Definition 2.5**, every nonnumerical factor of $K(\mathbf{x}|\theta)$ and $K(\mathbf{y}|\theta)$ contains θ, and neither $K(\mathbf{x}|\theta)$ nor $K(\mathbf{y}|\theta)$ is divisible by any positive number except the number 1. Hence, since $\frac{R(\mathbf{y})}{R(\mathbf{x})}$ does not contain θ, the nonnumerical factors of $K(\mathbf{x}|\theta)$ and $K(\mathbf{y}|\theta)$ must cancel in (29) and the remaining numerical factors could not be identical. Thus at least one of these factors would be divisible by a positive number other than 1 in contradiction to **Definition 2.5**. It now follows that $K(\mathbf{x}|\theta) = K(\mathbf{y}|\theta)$, so $K(\mathbf{x}|\theta)$ is some function u of $L(\mathbf{x}|\theta)$. Finally, $\tau_L \geq \tau_K$ since this function u is surjective from S^n onto its image $u(S^n)$. ■

Lemma 3.7. Under the conditions of **Lemma 3.6**, the sufficient statistics L and K satisfy

$$I_{comp}(\mathbf{x}|\theta, L) \geq I_{comp}(\mathbf{x}|\theta, K), \forall \mathbf{x} \in S^n. \tag{30}$$

Proof. Let $\mathbf{x} \in S^n$ and suppose that $\mathbf{y} \in A_{L(\mathbf{x})}$. Then $L(\mathbf{y}|\theta) = L(\mathbf{x}|\theta)$, so it follows from **Lemma 3.6** that $K(\mathbf{y}|\theta) = K(\mathbf{x}|\theta)$ and thus $\mathbf{y} \in A_{K(\mathbf{x})}$. Hence $A_{L(\mathbf{x})} \subseteq A_{K(\mathbf{x})}$, and so

$$P_\theta[L(\mathbf{X}|\theta) = L(\mathbf{x}|\theta)] = \sum_{\mathbf{y}\epsilon A_{L(\mathbf{x})}} f(\mathbf{x}|\theta) \le \sum_{\mathbf{y}\epsilon A_{K(\mathbf{x})}} f(\mathbf{x}|\theta) = P_\theta[K(\mathbf{X}|\theta) = K(\mathbf{x}|\theta)], \forall \mathbf{x} \in S^n \quad (31)$$

Taking the Shannon information of both sides of the inequality in (31) and using (13) gives (30). ■

Theorem 3.8. Let $\mathbf{x}$ be sample data for a random sample $\mathbf{X}$ from a discrete random variable X with the real-valued parameter θ. Then for all $\mathbf{x} \in S^n$,

$$I_{lost}(\mathbf{x}|L) \le I_{lost}(\mathbf{x}|K). \quad (32)$$

Proof. Let $\mathbf{x} \in S^n$. Note that $I_{total}(\mathbf{x}|\theta)$ in (12) does not depend on the arbitrary sufficient statistic T of (11). Hence

$$I_{total}(\mathbf{x}|\theta) = I_{comp}(\mathbf{x}|\theta, L) + I_{lost}(\mathbf{x}|L) = I_{comp}(\mathbf{x}|\theta, K) + I_{lost}(\mathbf{x}|K). \quad (33)$$

Then (32) follows immediately from (30) and (33). ■

As a consequence of **Theorem 3.5**, **Theorem 3.8** has an immediate corollary.

Corollary 3.9. Under the conditions of **Theorem 3.8**, let T be a sufficient statistic for θ for which there is a one-to-one function between T and K. Then for all $\mathbf{x} \in S^n$,

$$I_{lost}(\mathbf{x}|L) \le I_{lost}(\mathbf{x}|T). \quad (34)$$

Corollary 3.9 raises the question whether (34) holds for all sufficient statistics $\mathbf{T}$ for $\boldsymbol{\theta}$ or even for all real-valued statistics $\mathbf{T}$. It is conjectured that the first conclusion is false and hence is the second, but the question remains open. It is conceivable that notion of a minimal sufficient statistic [3] is relevant. Regardless, the proofs of **Lemma 3.7** and **Theorem 3.8** illustrate the fact that the relation between the lost information for two statistics $\mathbf{T}$ and $\mathbf{T}'$ is determined by the relation between their partition sets $\mathbf{A_t} = \{\mathbf{x}|\mathbf{T}(\mathbf{x}) = \mathbf{t}\}$ and $\mathbf{B_{t'}} = \{\mathbf{x}|\mathbf{T}'(\mathbf{x}) = \mathbf{t}'\}$. For example, if for every $\mathbf{A_t}$ there exists a $\mathbf{B_{t'}}$ for which $\mathbf{A_t} \subset \mathbf{B_{t'}}$, then the partition of $\mathbf{S^n}$ by the $\mathbf{B_{t'}}$ of $\mathbf{T}'$ is said to be coarser than the partition by the $\mathbf{A_t}$ of $\mathbf{T}$. In that case, $\mathbf{I_{lost}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{T}) \le \mathbf{I_{lost}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{T}')$ because each $\mathbf{x} \in \mathbf{S^n}$ has more $\mathbf{y} \in \mathbf{S^n}$ with $\mathbf{T}'(\mathbf{y}) = \mathbf{T}'(\mathbf{x})$ than there are with $\mathbf{T}(\mathbf{y}) = \mathbf{T}(\mathbf{x})$. In words, $\mathbf{T}'(\mathbf{y}) = \mathbf{t}'$ is at least as ambiguous as $\mathbf{T}(\mathbf{y}) = \mathbf{t}$ in determining the data sample giving the value of the respective statistics.

4. Entropic Loss for an SS

For a sufficient statistic T for θ we now propose an entropy measure to characterize T by the expected lost information incurred by compressing the random sample $\mathbf{X}$ into $T(\mathbf{X})$. This expectation is taken over all possible data sets $\mathbf{x}$. This nonstandard entropy measure is called entropic loss, and it depends on neither a particular data set $\mathbf{x}$ nor the value of θ. Before defining this measure, we need to determine the appropriate pmf to use in taking an expectation. The following results are used.

Result 4.1. Under the assumptions of **Theorem 3.3**, for any data sample let $t = T(\mathbf{x})$ and consider the partition set A_t. Then

$$\sum_{\mathbf{x}\in A_t} P[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t] = 1. \quad (35)$$

Proof. Summing (16) over $\mathbf{x} \in A_t$ yields

$$\sum_{\mathbf{x}\in A_t} P[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t] = \frac{\sum_{\mathbf{x}\epsilon A_{T(\mathbf{x})}} h(\mathbf{x})}{\sum_{\mathbf{y}\epsilon A_{T(\mathbf{x})}} h(\mathbf{y})} = 1. \quad (36)$$

to give (35). ■

Result 4.2. Under the assumptions of **Theorem 3.3**,

$$\sum_{\mathbf{x}\in S^n} P[\mathbf{X}=\mathbf{x} \mid T(\mathbf{X})=T(\mathbf{x})] = |\tau_T|. \tag{37}$$

Proof. We perform the sum on the left of (37) by first summing over $\mathbf{x} \in A_t$ for fixed t and then summing over each $t \in \tau_T$ to give

$$\sum_{\mathbf{x}\in S^n} P[\mathbf{X}=\mathbf{x} \mid T(\mathbf{X})=T(\mathbf{x})] = \sum_{t\in\tau_T}\sum_{\mathbf{x}\in A_t} P[\mathbf{X}=\mathbf{x} \mid T(\mathbf{X})=t], \tag{38}$$

The inner series on the right side of (38) sums to one by **Result 4.1**. Hence the outer sum yields $|\tau_T|$ for $\tau_T = \{t|\exists \mathbf{x} \in S^n$ for which $t = T(\mathbf{x})\}$. ■

From (37) it follows that the left side of (37) is not a probability distribution on S^n unless $|\tau_T| = 1$. Moreover, $P[\mathbf{X}=\mathbf{x} \mid T(\mathbf{X})=T(\mathbf{x})]$ is not a conditional probability distribution even if $|\tau_T| = 1$ since the condition $T(\mathbf{X}) = T(\mathbf{x})$ varies with $\mathbf{x}$. However, we use **Result 4.2** to normalize $P[\mathbf{X}=\mathbf{x} \mid T(\mathbf{X})=T(\mathbf{x})]$ and obtain the appropriate pmf for calculating the expectation of $I_{lost}(\mathbf{X}|T)$.

Definition 4.3 (Entropic Loss). Under the assumptions of **Theorem 3.3**, the entropic loss resulting from the data compression by T is defined as

$$H_{lost}(\mathbf{X},T) = \frac{-1}{|\tau_T|}\sum_{\mathbf{x}\in S^n} P[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x})]\log P[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x})], \tag{39}$$

which from (15) and (16) can be rewritten as

$$H_{lost}(\mathbf{X},T) = \frac{-1}{|\tau_T|}\sum_{\mathbf{x}\in S^n} \frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})}\log\frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})}. \tag{40}$$

Note that (39) and (40) are independent of both $\mathbf{x}$ and θ. Also, as noted in Section 3 for $I_{lost}(\mathbf{x}|T)$, if each $A_{T(\mathbf{x})}$ is a singleton in (40), then $H_{lost}(\mathbf{X},T) = 0$. We now compute $H_{lost}(\mathbf{X},T)$ for the sufficient statistic $T(\mathbf{X}) = L(\mathbf{X}|\theta)$.

Theorem 4.4 (Entropic Loss for Likelihood Function). Under the assumptions of **Theorem 3.3**, the entropic loss resulting from data compression by $T(\mathbf{x}) = L(\mathbf{x}|\theta)$ is

$$H_{lost}(\mathbf{X},L) = \frac{-1}{|\tau_L|}\sum_{t\in\tau_L}\log\frac{1}{|A_t|}. \tag{41}$$

Proof. From (20) write

$$H_{lost}(\mathbf{X},L) = \frac{-1}{|\tau_L|}\sum_{\mathbf{x}\in S^n}\frac{1}{|A_{L(\mathbf{x})}|}\log\frac{1}{|A_{L(\mathbf{x})}|}. \tag{42}$$

We decompose the sum over $\mathbf{x} \in S^n$ in (42) to consecutive sums over $\mathbf{x} \in A_t$ and then $t \in \tau_T$ to get

$$H_{lost}(\mathbf{X},L) = \frac{-1}{|\tau_L|}\sum_{t\in\tau_L}\sum_{\mathbf{x}\in A_t}\frac{1}{|A_t|}\log\frac{1}{|A_t|} = \frac{-1}{|\tau_L|}\sum_{t\in\tau_L}\frac{|A_t|}{|A_t|}\log\frac{1}{|A_t|}. \tag{43}$$

Equation (41) now follows from (43). ■

Since $H_{lost}(\mathbf{X},T)$ has been defined only for a sufficient statistic T for θ and is independent of θ, as well as the data sample $\mathbf{x}$. $H_{lost}(\mathbf{X},T)$ could thus be used to compare sufficient statistics. In particular, if the sufficient statistics $T_1(\mathbf{X})$ and $T_2(\mathbf{X})$ are considered as estimators for θ, then entropic loss could serve as a metric for regarding, say, T_1 as a better estimator for θ than T_2 if $H_{lost}(\mathbf{X},T_1) < H_{lost}(\mathbf{X},T_2)$.

Result 4.5. If there is a one-to-one function between two sufficient statistics T and T′ for θ, then they have the same entropic loss for a random sample $\mathbf{X}$; i.e.,

$$H_{lost}(\mathbf{X},T) = H_{lost}(\mathbf{X},T'). \tag{44}$$

Proof. For all $\mathbf{x} \in S^n$, $I_{lost}(\mathbf{x}|T) = I_{lost}(\mathbf{x}|T')$ from **Theorem 3.5**, so

$$-\log\frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})} = -\log\frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T'(\mathbf{x})}} h(\mathbf{y})}, \tag{45}$$

from which

$$\frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})} = \frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T'(\mathbf{x})}} h(\mathbf{y})}. \tag{46}$$

Thus from (45) and (46)

$$\frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})}\log\frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})} = \frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T'(\mathbf{x})}} h(\mathbf{y})}\log\frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T'(\mathbf{x})}} h(\mathbf{y})}. \tag{47}$$

Now summing (47) over $\mathbf{x} \in S^n$ yields

$$\sum_{\mathbf{x}\in S^n}\frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})}\log\frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})} = \sum_{\mathbf{x}\in S^n}\frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in T'(\mathbf{x})} h(\mathbf{y})}\log\frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T'(\mathbf{x})}} h(\mathbf{y})}. \tag{48}$$

But from **Theorem 3.5**, $|\tau_T| = |\tau_{T'}|$. Thus dividing the left side of (48) by $-|\tau_T|$ and the right side by $-|\tau_{T'}|$ yields (44). ■

Given (32), it might be anticipated that

$$H_{lost}(\mathbf{X}, L) \le H_{lost}(\mathbf{X}, K). \tag{49}$$

However, we conjecture that (49) is not always true, but we have no counterexample. If this conjecture is true, then $L(\mathbf{x}|\theta)$ would not in general have the minimum entropic loss among sufficient statistics for θ.

5. Examples and Computational Issues

In this section we present examples involving the discrete Poisson, binomial, and geometric distributions [9]. For each distribution, three sufficient statistics for some parameter θ are analyzed. Thus the right side of (8) is independent of θ, as well as the information $I_{lost}(\mathbf{x}|T)$ conveyed by the data sample $\mathbf{x}$ about $\mathbf{X}$. Even for sufficient statistics, calculating the information quantities of this paper may present computational issues, some of which are discussed in this section. Our examples are therefore simple in order to focus on the definitions and results of Sections 3 and 4.

Example 5.1 (Poisson Distribution). Consider the random sample $\mathbf{X} = (X_1, \ldots, X_n)$ with the data sample $\mathbf{x} = (x_1, \ldots, x_n)$ from a Poisson random variable X. We consider three sufficient statistics for the parameter $\theta > 0$. These sufficient statistics are $T_1(\mathbf{X}) = \sum_{i=1}^n X_i$, the likelihood kernel $T_2(\mathbf{X}) = K(\mathbf{X}|\theta)$ for fixed but arbitrary θ and the likelihood function $T_3(\mathbf{X}) = L(\mathbf{X}|\theta)$ for fixed but arbitrary θ. In particular, we use $T_1(\mathbf{X})$ as a surrogate for $T_1'(\mathbf{X}) = \frac{\sum_{i=1}^n X_i}{n}$. Neither $T_1(\mathbf{X})$ or $T_1'(\mathbf{X})$ involves θ and can thus be used either to characterize $\mathbf{X}$ or to estimate θ. Moreover, since there is an obvious one-to-one function relating $\frac{\sum_{i=1}^n X_i}{n}$ and $\sum_{i=1}^n X_i$, **Theorems 3.5** and **4.5** establish that $I_{lost}(\mathbf{x}|T_1') = I_{lost}(\mathbf{x}|T_1)$ and $H_{lost}(\mathbf{X}, T_1') = H_{lost}(\mathbf{X}, T_1)$, respectively. We consider $T_1(\mathbf{X})$ because it is also Poisson, whereas $T_1'(\mathbf{X})$ is not since $\frac{\sum_{i=1}^n X_i}{n}$ is not necessarily a nonnegative integer. In contrast to $T_1(\mathbf{X})$, both $T_2(\mathbf{X})$ and $T_3(\mathbf{X})$ contain θ and can only be used to characterize $\mathbf{X}$. For each of these three sufficient statistics we develop an expression for $I_{lost}(\mathbf{x}|T)$ and describe how to obtain a numerical value. We then illustrate previous results with simple data and present computational results in **Table 5.1**.

Case 1: Let $T_1(\mathbf{X}) = \sum_{i=1}^{n} X_i$. Observe that $T_1(\mathbf{X})$ is a sufficient statistic for θ from **Result 2.2** since $f(\mathbf{x}|\theta) = P_\theta[\mathbf{X} = \mathbf{x}] = \frac{\theta^{\sum_{i=1}^{n} x_i}\, e^{-n\theta}}{\prod_{i=1}^{n} x_i!}$ can be factored in (2) into the functions $g[T_1(\mathbf{x})|\theta] = \theta^{\sum_{i=1}^{n} x_i}\, e^{-n\theta}$ and $h(\mathbf{x}) = \frac{1}{\prod_{i=1}^{n} x_i!}$. Next recall that the statistic $\sum_{i=1}^{n} X_i$ has a Poisson distribution with parameter $n\theta$ [9]. Thus $P_\theta[\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i] = \frac{(n\theta)^{\sum_{i=1}^{n} x_i}\, e^{-n\theta}}{(\sum_{i=1}^{n} x_i)!}$, and so (8) becomes

$$P[\mathbf{X} = \mathbf{x}|\, \textstyle\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i] = \frac{1}{n^{\sum_{i=1}^{n} x_i}} \binom{\sum_{i=1}^{n} x_i}{x_1, \dots, x_n}, \tag{50}$$

where the multinomial coefficient $\binom{\sum_{i=1}^{n} x_i}{x_1,\dots,x_n} = \frac{(\sum_{i=1}^{n} x_i)!}{\prod_{i=1}^{n} x_i!}$. It follows from (50) and (10) that

$$I_{lost}(\mathbf{x}|T_1) = -\log \binom{\sum_{i=1}^{n} x_i}{x_1, \dots, x_n} + (\log n) \sum_{i=1}^{n} x_i, \tag{51}$$

which is also $I_{lost}(\mathbf{x}|T_1')$.

For a data sample $(x_1, \dots, x_n)$, the evaluation of $I_{lost}(\mathbf{x}|T_1)$ in (51) involves computing factorials. For realistic data, the principal limitation in calculating them by direct multiplication is their magnitude. See [11] for a discussion. However, (51) can be approximated using either the well-known Stirling formula or the more accurate Ramanujan approximation [12]. The online multinomial coefficient calculator [13] can evaluate multinomial coefficients for both x_i and n less than approximately 50 if any $x_i = 0$ is removed from $\binom{\sum_{i=1}^{n} x_i}{x_1,\dots,x_n}$. Such deletions do not affect the calculation since $0! = 1$.

As a numerical example, consider a data sample $\mathbf{x}$ of size $n = 34$ from a Poisson random variable X with $\theta = 3$. On the average, $T_1(\mathbf{X}) = \sum_{i=1}^{n} X_i = n\theta = 102$, so we take $\sum_{i=1}^{n} x_i = 102$ for the data sample $\mathbf{x} = (4, 7, 1, 3, 4, 2, 5, 0, 1, 2, 3, 6, 8, 0, 1, 2, 4, 9, 0, 2, 3, 1, 4, 2, 0, 1, 5, 6, 2, 7, 0, 1, 4, 2)$. Then the calculator at [13] gives that $\binom{\sum_{i=1}^{n} x_i}{x_1,\dots,x_n} \approx 1.574 \times 10^{123}$ in (50). Moreover, $(\log n) \sum_{i=1}^{n} x_i = 518.915$. Hence from (51), $I_{lost}(\mathbf{x}|T_1) = I_{lost}(\mathbf{x}|T_1') \approx$ 109.667 bits of Shannon information. This value corresponds to 13.708 bytes at 8 bits per byte or to 0.013 kilobytes (KB) at 1024 bytes per kilobyte [14]. It thus follows from the discussion at the beginning of this example that

$$I_{lost}(\mathbf{x}|T_1) = I_{lost}(\mathbf{x}|T_1') \approx \text{ 0.013 KB.} \tag{52}$$

Case 2: Let $T_2(\mathbf{X}) = K(\mathbf{X}|\theta)$ for fixed but arbitrary $\theta > 0$. For a data sample $(x_1, \dots, x_n)$ write

$$L(\mathbf{x}|\theta) = f(\mathbf{x}|\theta) = \frac{\theta^{\sum_{i=1}^{n} x_i}\, e^{-n\theta}}{\prod_{i=1}^{n} x_i!}, \tag{53}$$

from which

$$K(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^{n} x_i}\, e^{-n\theta} \tag{54}$$

and $R(\mathbf{x}) = \frac{1}{\prod_{i=1}^{n} x_i!}$ in (4). Note that for all fixed $\theta > 0$ except $\theta = 1$, there is an obvious one-to-one function between $T_1(\mathbf{x}) = \sum_{i=1}^{n} x_i$ and (54). Hence in the numerical example of **Case 1**, $I_{lost}\big(\mathbf{x}\big|K(\mathbf{x}|\theta)\big) = I_{lost}(\mathbf{x}|T_1) \approx$ 0.013 KB from **Theorem 3.5** for all $\theta > 0$ except $\theta = 1$. For $\theta = 1$, $K(\mathbf{x}|\theta) = e^{-n}$ and is constant with respect to any data sample $\mathbf{x}$. Thus $I_{comp}(\mathbf{x}|1, K) = 0$ and $I_{lost}\big(\mathbf{x}\big|K(\mathbf{x}|1)\big) = I_{total}(\mathbf{x}|1, K)$. It follows that $K(\mathbf{x}|1)$ provides no information about $\mathbf{X}$.

Case 3: Let $T_3(\mathbf{X}) = L(\mathbf{X}|\theta)$ for fixed but arbitrary $\theta > 0$. We attempt to obtain $I_{lost}\big(\mathbf{x}\big|L(\mathbf{x}|\theta)\big)$ for a data sample $\mathbf{x} = (x_1, \dots, x_n)$ by determining $\big|A_{L(\mathbf{x}|\theta)}\big|$ and using (20). From (53), note that for all fixed $\theta > 0$ except $\theta = 1$, $\mathbf{y} \in A_{L(\mathbf{x}|\theta)}$ if and only if

$$\frac{\theta^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} = \frac{\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}. \tag{55}$$

Thus for any fixed θ satisfying $\theta > 0$ and $\theta \neq 1$, $\mathbf{y} \in A_{L(\mathbf{x}|\theta)}$ if both $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i$ and $\prod_{i=1}^{n} y_i! = \prod_{i=1}^{n} x_i!$. However, for some $\theta > 0$ and $\theta \neq 1$, it is possible that $\mathbf{y} \in A_{L(\mathbf{x}|\theta)}$ when neither $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i$ nor $\prod_{i=1}^{n} y_i! = \prod_{i=1}^{n} x_i!$. For example, let $\theta = 2$, $\mathbf{x} = (4,1,1,0)$, and $\mathbf{y} = (3,2,0,0)$. Then $\sum_{i=1}^{n} x_i = 6$, $\sum_{i=1}^{n} y_i = 5$, $\prod_{i=1}^{n} x_i! = 24$, and $\prod_{i=1}^{n} y_i! = 12$. However, (55) is satisfied.

Such complications suggest that an efficient implicit enumeration of the $\mathbf{y}$ satisfying (55) would be required to obtain $\left|A_{L(\mathbf{x}|\theta)}\right|$ for calculating $I_{lost}\left(\mathbf{x}\middle|L(\mathbf{x}|\theta)\right)$ from (20). Using such an algorithm, a conventional computer could probably compute $I_{lost}\left(\mathbf{x}\middle|L(\mathbf{x}|\theta)\right)$ for the numerical data and value of θ in **Case 1** since there is now a 250 petabyte, 200 petaflop conventional computer [15]. Substantially larger problems, if not already tractable, will likely be so in the foreseeable future on quantum computers. Recently the milestone of quantum supremacy was achieved where the various possible combinations of a certain randomly generated output were obtained in 110 seconds, whereas this task would have taken the above conventional supercomputer 10,000 years [16]. Regardless, for the data of **Case 1**, we have the upper bound $I_{lost}\left(\mathbf{x}\middle|L(\mathbf{x}|\theta)\right) \leq 0.013$ KB from (32)**.**

We present some simple further simple computational results for the Poisson example distribution to illustrate relationships among T_1, T_2, T_3. **Table 5.1** below summarizes the results for sample data (x_1, x_2, x_3) with $\sum_{i=1}^{3} x_i \leq 2$. In particular, a complete enumeration of $A_{L(\mathbf{x}|\theta)}$ gives $I_{lost}\left(\mathbf{x}\middle|L(\mathbf{x}|\theta)\right)$ from (20).

Table 5.1. Poisson Example

$\mathbf{x} = (x_1, x_2, x_3)$	$T_1(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_1)$	$T_2(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_2)$	$T_3(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_3)$
(0,0,0)	0	0	$e^{-3\theta}$	0	$e^{-3\theta}$	0
(0,0,1)	1	$\log 3$	$\theta e^{-3\theta}$	$\log 3$	$\theta e^{-3\theta}$	$\log 3$
(0,1,0)						
(1,0,0)						
(1,1,0)	2	$\log\frac{9}{2}$	$\theta^2 e^{-3\theta}$	$\log\frac{9}{2}$	$\theta^2 e^{-3\theta}$	$\log 3$
(1,0,1)						
(0,1,1)						
(2,0,0)	2	$\log 9$	$\theta^2 e^{-3\theta}$	$\log 9$	$\frac{\theta^2 e^{-3\theta}}{2}$	$\log 3$
(0,2,0)						
(0,0,2)						

Example 5.2 (Binomial Distribution). Consider a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from a binomial random variable X with parameters m and θ, where θ is the probability of success on any of the m Bernoulli trials associated with the X_i, $i = 1, \ldots, n$. Let m be fixed, so the only parameter is θ. Moreover, the sample space of the underlying random variable X is now finite.

Case 1: $T_1(\mathbf{X}) = \sum_{i=1}^{n} X_i$. Again $\sum_{i=1}^{n} X_i$ is an SS for θ. From [9], $\sum_{i=1}^{n} X_i$ has a binomial distribution with parameter θ for fixed nm. Hence

$$P_\theta\left[\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i\right] = \theta^{\sum_{i=1}^{n} x_i}\, \theta^{mn - \sum_{i=1}^{n} x_i} \binom{mn}{\sum_{i=1}^{n} x_i} \tag{56}$$

and

$$P_\theta[\mathbf{X} = \mathbf{x}] = \theta^{\sum_{i=1}^{n} x_i}\, \theta^{mn - \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \binom{m}{x_i}. \tag{57}$$

From (1), dividing (57) by (56) gives

$$P[\mathbf{X} = \mathbf{x} | \textstyle\sum_{i=1}^{n} X_i = t] = \dfrac{\prod_{i=1}^{n} \binom{m}{x_i}}{\binom{mn}{t}}. \tag{58}$$

By taking the $-\log$ of (58) gives the lost information as

$$I_{lost}(\mathbf{x}|T_1) = -\log \frac{\prod_{i=1}^{n} \binom{m}{x_i}}{\binom{mn}{t}} = -\sum_{i=1}^{n} \log \binom{m}{x_i} + \log \binom{mn}{t}. \tag{59}$$

Case 2: $T_2(\mathbf{X}) = K(\mathbf{X}|\theta)$. In this case we use (16) as in **Example 5.1**. Write

$$L(\mathbf{x}|\theta) = f(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{mn-\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \binom{m}{x_i}, \tag{60}$$

from which $K(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{mn-\sum_{i=1}^{n} x_i}$ and $R(\mathbf{x}) = \prod_{i=1}^{n} \binom{m}{x_i}$ in (4). To factor the right side of (60) as in (2), let g be the identity function and $h(\mathbf{x}) = \prod_{i=1}^{n} \binom{m}{x_i}$. Hence,

$$I_{lost}(\mathbf{x}|T_2) = -\log \frac{\prod_{i=1}^{n} \binom{m}{x_i}}{\sum_{\mathbf{y} \in A_{K(\mathbf{x}|\theta)}} \prod_{i=1}^{n} \binom{m}{y_i}}, \tag{61}$$

and (61) yields

$$I_{lost}(\mathbf{x}|T_2) = -\sum_{i=1}^{n} \log \binom{m}{x_i} + \log \sum_{\mathbf{y} \in A_{K(\mathbf{x}|\theta)}} \prod_{i=1}^{n} \binom{m}{y_i}, \tag{62}$$

where

$$A_{K(\mathbf{x}|\theta)} = \left\{ \mathbf{y} \in S^n \middle| \theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{mn-\sum_{i=1}^{n} y_i} = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{mn-\sum_{i=1}^{n} x_i} \right\}. \tag{63}$$

From (63), for any fixed θ satisfying $0 < \theta < 1$ and $\theta \neq 1/2$, it can easily be shown that $\mathbf{y} \in A_{K(\mathbf{x}|\theta)}$ if and only if $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i$. Thus in general, for a given $\mathbf{x}$ and fixed θ, determining $A_{K(\mathbf{x}|\theta)}$ in **Case 2** would require an enumeration of the $\mathbf{y}$ satisfying (63) to compute (62). We perform such an enumeration below for a simple example.

Case 3: $T_3(\mathbf{X}) = L(\mathbf{X}|\theta)$. For a data sample $\mathbf{x} = (x_1, \ldots, x_n)$ we now have

$$L(\mathbf{x}|\theta) = \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{n} x_i} (1-\theta)^{mn} \prod_{i=1}^{n} \binom{m}{x_i} \tag{64}$$

with g be the identity function and $h(\mathbf{x}) = 1$ in (2). For fixed θ satisfying $0 < \theta < 1$ and $\theta \neq 1/2$, we obtain that $\mathbf{y} \in A_{L(\mathbf{x}|\theta)}$ if and only if

$$\left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{n} y_i} \prod_{i=1}^{n} \binom{m}{y_i} = \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \binom{m}{x_i}. \tag{65}$$

As in **Case 3** of **Example 5.1**, developing an algorithm to use (65) and determine $\left|A_{L(\mathbf{x}|\theta)}\right|$ for calculating $I_{lost}\left(\mathbf{x}\middle|L(\mathbf{x}|\theta)\right)$ from (20) is beyond the scope of this paper.

As a simple example, consider the experiment of flipping a possibly biased coin twice ($m = 2$). The total number of heads follows a binomial distribution with the parameter θ, which is the probability of getting a head on any flip. By doing this experiment three times we generate the random variables X_1, X_2, X_3 with possible values 0, 1, 2. **Table 5.2** shows all the possibilities and the lost information for the statistics. The small size of this example allows the computation of I_{lost} in **Cases 2** and **3** via total enumeration.

Table 5.2. Binomial Example

$\mathbf{x} = (x_1, x_2, x_3)$	$T_1(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_1)$	$T_2(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_2)$	$T_3(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_3)$
(0,0,0)	0	0	$(1-\theta)^6$	0	$(1-\theta)^6$	0
(0,0,1)	1	$\log 3$	$(1-\theta)^5\theta^1$	$\log 3$	$2(1-\theta)^5\theta^1$	$\log 3$
(0,1,0)						
(1,0,0)						
(1,1,0)	2	$\log\frac{15}{4}$	$(1-\theta)^4\theta^2$	$\log\frac{15}{4}$	$4(1-\theta)^4\theta^2$	$\log 3$
(1,0,1)						
(0,1,1)						
(2,0,0)	2	$\log 15$	$(1-\theta)^4\theta^2$	$\log 15$	$(1-\theta)^4\theta^2$	$\log 3$
(0,2,0)						
(0,0,2)						
(1,1,1)	3	$\log\frac{5}{2}$	$(1-\theta)^3\theta^3$	$\log\frac{5}{2}$	$8(1-\theta)^3\theta^3$	0
(2,1,0)	3	$\log 10$	$(1-\theta)^3\theta^3$	$\log 10$	$2(1-\theta)^3\theta^3$	$\log 6$
(2,0,1)						
(1,0,2)						
(1,2,0)						
(0,1,2)						
(0,2,1)						
(2,1,1)	4	$\log\frac{15}{4}$	$(1-\theta)^2\theta^4$	$\log\frac{15}{4}$	$4(1-\theta)^2\theta^4$	$\log 3$
(1,2,1)						
(1,1,2)						
(2,2,0)	4	$\log 15$	$(1-\theta)^2\theta^4$	$\log 15$	$(1-\theta)^2\theta^4$	$\log 3$
(2,0,2)						
(0,2,2)						
(2,2,1)	5	$\log 3$	$(1-\theta)^1\theta^5$	$\log 3$	$2(1-\theta)^1\theta^5$	$\log 3$
(2,1,2)						
(1,2,2)						
(2,2,2)	6	0	θ^6	0	θ^6	0

Now using (40), we give in **Table 5.3** the entropic losses of **Example 5.2** for T_1, T_2, T_3. Note that $H_{lost}(\mathbf{X}, T)$ is the same for the sum T_1 and the likelihood kernel T_2, which are related by a one-to-one function. Hence **Result 4.5** is corroborated. Also observe that $H_{lost}(\mathbf{X}, T)$ is smallest for the likelihood function T_3.

Table 5.3. Entropic loss over different statistics for a binomial distribution

$H_{lost}(\mathbf{X}, T_1)$	$H_{lost}(\mathbf{X}, T_2)$	$H_{lost}(\mathbf{X}, T_3)$
1.4722	1.4722	1.2095

Example 5.3 (Geometric Distribution). Consider a random sample $\mathbf{X} = (X_1, \dots, X_n)$ with sample data $\mathbf{x} = (x_1, \dots, x_n)$ from a geometric random variable X, where the parameter θ is the probability of success on any of the series of independent Bernoulli trials for which X is the trial number on which the first success is obtained. It readily follows from [5] that

$$P[\mathbf{X}=\mathbf{x}] = \theta^n(1-\theta)^{\sum_{i=1}^n x_i - n}. \tag{66}$$

Case 1: $T_1(\mathbf{X}) = \sum_{i=1}^n X_i$. For fixed n, $\sum_{i=1}^n X_i$ has a negative binomial distribution with parameter θ. Hence,

$$P\left[\sum_{i=1}^n X_i = \sum_{i=1}^n x_i\right] = \binom{\sum_{i=1}^n x_i - 1}{n-1}\theta^n(1-\theta)^{\sum_{i=1}^n x_i - n}. \tag{67}$$

Thus $T_1(\mathbf{X}) = \sum_{i=1}^n X_i$ is an SS for θ since it satisfies (2) with $g[T_1(\mathbf{x})|\theta] = \theta^n(1-\theta)^{T_1(\mathbf{x})-n}$ and $h(x_1, \dots, x_n) = \binom{\sum_{i=1}^n x_i - 1}{n-1}$. Moreover, substitution of (66) and (67) into (8) gives

$$P\left[\mathbf{X}=\mathbf{x}\middle|\sum_{i=1}^n X_i = \sum_{i=1}^n x_i\right] = \frac{1}{\binom{\sum_{i=1}^n x_i - 1}{n-1}}. \tag{68}$$

Then from (14) and (68) we obtain that

$$I_{lost}(\mathbf{x}|T_1) = \log\binom{\sum_{i=1}^n x_i - 1}{n-1}. \tag{69}$$

Case 2: $T_2(\mathbf{X}) = K(\mathbf{X}|\theta)$. From (66), for all $\mathbf{x} \in S^n$, $R(\mathbf{x}) = 1$ and

$$K(\mathbf{x}|\theta) = L(\mathbf{x}|\theta) = \left(\frac{\theta}{1-\theta}\right)^n (1-\theta)^{\sum_{i=1}^n x_i}. \tag{70}$$

Thus for $0 < \theta < 1$, there is an obvious one-to-one function between $T_1(\mathbf{x}) = \sum_{i=1}^n x_i$ and $T_2(\mathbf{x}) = K(\mathbf{x}|\theta)$ in (70). Thus from **Theorem 3.5**, $I_{lost}(\mathbf{x}|T_2) = I_{lost}(\mathbf{x}|T_1)$ as given in (69).

Case 3: $T_3(\mathbf{X}) = L(\mathbf{X}|\theta)$. Since $K(\mathbf{X}|\theta) = L(\mathbf{X}|\theta)$ from (70), then

$$I_{lost}(\mathbf{x}|T_3) = \log\binom{\sum_{i=1}^n x_i - 1}{n-1} \tag{71}$$

from (69). However, there is an alternate derivation of (71). For $0 < \theta < 1$ it follows from (70) that then $\mathbf{y} \in A_{L(\mathbf{x}|\theta)}$ if and only if

$$\sum_{i=1}^n y_i = \sum_{i=1}^n x_i. \tag{72}$$

But for fixed positive integers $x_1, \dots, x_n$ we have from [17] that the number of solutions $\left|A_{L(\mathbf{x}|\theta)}\right|$ to (72) in positive integers $y_1, \dots, y_n$ is

$$\binom{\sum_{i=1}^n x_i - 1}{n-1}. \tag{73}$$

Thus (71) follows for $L(\mathbf{X}|\theta)$ from (73) and (20), so $I_{lost}(\mathbf{x}|T_1) = I_{lost}(\mathbf{x}|T_2) = I_{lost}(\mathbf{x}|T_3)$ from **Theorem 3.5**.

As a numerical illustration, let the random variable X denote the number of flips of a possibly biased coin until a head is obtained. Then X has a geometric distribution with the parameter θ as the probability of getting a head on any flip. Suppose this experiment is performed three times yielding the possible sample data $\mathbf{x} = (x_1, x_2, x_3)$ shown in **Table 5.4**. $I_{lost}(\mathbf{x}|T)$ is then calculated for each of the sufficient statistics for θ of **Example 5.3**. Observe that the individual statistics depend on θ while the lost information does not. Moreover, $I_{lost}(\mathbf{x}|T_1) = I_{lost}(\mathbf{x}|T_2) = I_{lost}(\mathbf{x}|T_3)$ for all the sample data as established analytically above.

Table 5.4. Geometric Example

$\mathbf{x} = (x_1, x_2, x_3)$	$T_1(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_1)$	$T_2(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_2)$	$T_3(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_3)$
(1,1,1)	3	0	θ^3	0	θ^3	0
(2,1,1)	4	$\log 3$	$\theta^3(1-\theta)$	$\log 3$	$\theta^3(1-\theta)$	$\log 3$
(1,2,1)						
(1,1,2)						
(2,2,1)	5	$\log 6$	$\theta^3(1-\theta)^2$	$\log 6$	$\theta^3(1-\theta)^2$	$\log 6$
(2,1,2)						
(1,2,2)						
(2,2,2)	6	$\log 10$	$\theta^3(1-\theta)^3$	$\log 10$	$\theta^3(1-\theta)^3$	$\log 10$

6. Conclusion

In this paper, the Shannon information obtained from a random sample $\mathbf{X}$ for a discrete random variable X with a single parameter θ was decomposed into two components: (i) the compressed information obtained by the value of a real-valued statistic $T(\mathbf{X})$ for the sample data $\mathbf{x}$ and (ii) the information lost by using this statistic to characterize $\mathbf{X}$. We focused on this lost information caused by multiple data sets having the same value of the statistic. This possibility is typical of data analysis, where the data uniquely determines the value of the statistic, but a value of the statistic does not uniquely determine the data yielding it. In other words, we answered the question: how much Shannon information is lost about a data sample when only the value of a sufficient statistic is known but not the original data. We also defined the entropic loss associated with a sufficient statistic T under consideration as the expected lost information over all possible samples to give a metric dependent only on T. Our approach is applicable to any T, but we focused on sufficient statistics for θ for simplicity. Applications of our results were computationally intensive.

References

1. Landauer, R. (1961). Irreversibility and heat generation in the computing process. IBM Journal of Research and Development.
2. Hodge, S.E.; Vieland, V.J. (2017). Information loss in binomial data due to data compression. Entropy.
3. Casella, G.; Berger, R.L. 2002). Statistical Inference, 2nd ed.; Cengage Learning, Delhi, India.
4. Pawitan, Y. (2013). In All Likelihood: Statistical Modeling and Inference Using Likelihood, 1st ed.; The Clarendon Press: Oxford, UK.
5. Rohatgi, V.K.; Ehsanes Saleh, A.K. (2001). An Introduction to Probability and Statistics, 2nd ed.; John Wiley & Sons, Inc., NY, USA.
6. Shannon, C.E.; Weaver, W. (1964). The Mathematical Theory of Communication, 1st ed.; The University of Illinois Press, Urbana, Illinois.
7. Shannon, C. (1948). A mathematical theory of communication. Bell. Syst. Tech. J.
8. Kapur, J.N.; Kesavan, H.K. (1992). Entropy Optimization Principles with Applications, 1st ed.; Academic Press, Inc., San Diego, CA, USA.
9. Johnson, J.L. (2003). Probability and Statistics for Computer Science, 1st ed.; John Wiley & Sons, Inc., NJ, USA.
10. Beeler, R.A. (2015). How to Count: An Introduction to Combinatorics and Its Applications, 1st ed.; Springer, Switzerland.
11. https://en.wikipedia.org/wiki/Factorial#Computation, accessed 7/1/2019.
12. Mortici, C. (2010). Ramanujan formula for the generalized Stirling approximation. Applied Mathematics and Computation.

13. https://mathcracker.com/multinomial-coefficient-calculator.php, accessed 7/27/2019.
14. https://en.wikipedia.org/wiki/Kilobyte, accessed 7/27/2019.
15. https://en.wikipedia.org/wiki/Summit_(supercomputer), accessed 7/1/2019.
16. Arute, F.; Arya, K.; Martinis, J.M. (2019). Quantum supremacy using a programmable superconducting processor. Nature.
17. Mahmoudvand, R.; Hassani, H.; Farzaneh, A.; Howell.G. (2010). The exact number of nonnegative integer solutions for a linear Diophantine inequality. IAENG International Journal of Applied Mathematics.

Chapter 3

New Concepts in Information Theory with Applications in Data Analysis

Maryam Moghimi*[1,2], H.W. Corley[1,2]

[1] Center on Stochastic Modeling, Optimization, and Statistics (COSMOS), The University of Texas at Arlington, Arlington, TX, USA
[2] The authors contributed equally to this paper.
* Correspondence: maryam.moghimi@uta.edu; +1-214-971-0904 (M.M.), corley@uta.edu; Tel.: +1-817-272-3092 (H.C.)

Abstract: In this paper we present new concepts of information and entropy for discrete distributions. In particular, we define a new type of information. The well-known Shannon information, also called surprisal, measures the surprise that would be gained retrospectively by an observer after the occurrence of an event. A low probability of occurrence gives much surprise and hence Shannon information. On the other hand, the new gambler's information uses the probability of an event prospectively as information to decide whether to bet on the event's occurrence. Higher probability gives more information. We also define the notion of outer entropy in which the log in Shannon (or inner) entropy is placed outside the expectation to yield to provide simpler calculations and some intuitive results. We then apply gambler's information and inner entropy the gambler's information lost when a statistic is used to characterize a random sample. Finally, we apply outer entropy and propose two new metrics that provide evidence whether one estimate of the parameter θ for a random variable X from a random sample $\mathbf{X}$ may be considered better than another. Examples are presented.

Keywords: Information, entropy, discrete distribution, evidence, data analysis

1. Introduction

Our previous study of the information loss due to data compression for a random sample [1] motivated an examination of the notions of Shannon information and entropy [2] for discrete random variables. The Shannon information of an event is defined as the negative log to the base 2 of the probability p for the event, while Shannon entropy is the expected value of Shannon information. The general properties of Shannon information and Shannon entropy can be found in [3, 4, 5, 6], for example. It should be noted that probability itself is a measure of information in the sense that an event with a small probability of occurrence is surprising. In other words, an observer obtains more information, sometimes called surprisal [7, p. 150], if an unlikely event occurs than if a likely one does. Since Shannon information involves a logarithm of the probability, it also measures surprise but is additive as opposed to probability itself. [2]

Shannon information may not be the appropriate information for modeling some decisions. To address this issue, we introduce here a new type of information called gambler's information, which views events prospectively, as compared with Shannon information, which views them retrospectively. The Shannon information of an event is not obtained until the event actually occurs and causes a level of surprise appropriate to the likelihood of it occurring. On the other hand, gambler's information derives from the probability of the event and not its occurrence. We also define a new entropy measure called outer entropy and designate Shannon entropy as inner entropy to avoid confusion. Outer entropy is easier to calculate but has similar characteristics to inner entropy. For instance, outer entropy, like inner entropy, is a measure

of diversity [6]. We then use our new definitions to provide intuitive measures of evidence for discrete random.

The paper is organized as follows. In Section 2, we present definitions, notation, and preliminary results. In Section 3, we find the gambler's information and outer entropy for some standard discrete distributions. In Section 4, we state some results for these new concepts and reformulate the results of [1] in terms of gambler's information and outer entropy as an application. We also introduce new evidence functions for the parameter of a discrete distribution, state some properties, and give examples. Conclusions are offered in Section 5.

2. Preliminaries

The following definitions, notation, and results are used here. Further details can be found in [2, 4, 6], for example. In this paper, we consider a discrete one-dimensional random variable X with a sample space S and one-dimensional parameter θ. The pmf for a discrete random variable X is denoted by $f(x|\theta)$ for a fixed but arbitrary value of θ.

Two types of information are first defined.

Definition 2.1 (Shannon Information). Let X be a discrete one-dimensional random variable with pmf f(x). Then the Shannon information associated with $x \in S$ is defined as

$$I(x|\theta) = -\log f(x|\theta), \tag{2.1}$$

where the units of $I(x)$ is bits if the base of the logarithm is 2.

Shannon information is an additive measure for an observer's level of surprise about the occurrence of x. Indeed, Shannon information is also called surprisal since a small value of $f(x|\theta)$ gives more information about x than a high one. In (2.1) Shannon information is a monotonically decreasing function of the probability $p = f(x|\theta)$. It measures the surprise that would be incurred after the occurrence of x. In other words, the associated information may be construed as retrospective. However, information may be desired prospectively. For example, a gambler often wants to bet on an event with a high probability of occurrence. Hence, Shannon information is not be appropriate as a criterion for modeling some decisions. The following definition addresses this issue. Unless otherwise stated, the log function will have base 2.

Definition 2.2 (Gambler's Information). Let X be a discrete one-dimensional random variable with pmf f(x). For $x \in S$ the associated gambler's is defined as

$$I^g(x|\theta)| = -\log[1 - f(x|\theta)]. \tag{2.2}$$

While Shannon is an additive information measure of the surprise level associated with f(x) for $x \in S$, gambler's information increases with the likelihood of an event as opposed to the likelihood of an event. We also call it certitude. Gambler's information, or certitude, is a monotonically increasing function of $f(x|\theta)$ and additive in $g(x|\theta) = 1 - f(x|\theta)$. Other definitions for information have been proposed. For example, Vigo [8, 9] has defined a measure of representational information. Further details on different types of information can be found in [10-15].

We develop some new types of entropy and compare them to Shannon entropy [1-6]. We use the notation $f(x|\theta)$ to denote that the parameter θ has given value for the variable x.

Definition 2.3 (Shannon Inner Entropy). Let X be a discrete one-dimensional random variable with pmf f(x) and sample space S. The inner entropy $H_I(\theta)$ of X with respect to Shannon information is defined as

$$H_I(\theta) = \sum_{x \in S} f(x|\theta) I(x|\theta), \tag{2.3}$$

or equivalently,

$$H_I(\theta) = E[I(X|\theta)] = E[-\log f(X|\theta)]. \tag{2.4}$$

Equations (2.3) and (2.4) simply give the expected Shannon information over S, i.e., the usual Shannon entropy [1-6]. However, there are difficulties using inner entropy. The principal one is that calculation of inner entropy is often difficult with $\log f(x_i)$ inside the summation of (2.3). To address this issue, we propose a new type of entropy with similar properties which can lead to some intuitive results.

Definition 2.4 (Shannon Outer Entropy). Let X be a discrete one-dimensional random variable with pmf $f(x|\theta)$ and sample space S. The Shannon outer entropy $H_O(\theta)$ with respect to Shannon information is defined as

$$H_O(\theta) = -\log \sum_{x \in S} [f(x|\theta)]^2. \tag{2.5}$$

Note that the Shannon outer entropy of X is the Shannon information of the expected value of the pmf $f(x|\theta)$ of X. It is only a function of a fixed but arbitrary value of θ. Thus

$$H_O(\theta) = I\big[E\big(f(X|\theta)\big)\big] = -\log E[f(X|\theta)]. \tag{2.6}$$

By comparing **Definitions 2.3** and **2.4**, we note that outer entropy in (2.6) results from the interchange of information and expected value for inner entropy in (2.4).

Other alternatives to Shannon entropy include the diversity index [15, 16], which is the probability that two data points in a sample have the same value. In contrast, $\sum_{x \in S}[f(x|\theta)]^2$ in (2.5) is interpreted as the expected value of the pmf of X leading to Shannon information. The diversity index is known as the Simpson index in ecology and the Herfindahl index in economics [17, 18]. DeDeo [19] discusses issues associated with using the diversity as a measure of uncertainty. In addition, Renyi entropy [20, 21] generalizes Shannon entropy, as does Tsallis entropy [22, 23]. Less popular alternatives of entropy can be found in [24]. A comparison different measures of entropy is given in [4, 22].

We now relate inner and outer entropy. First note from [6] that $H_I(\theta) \geq 0$.

Lemma 2.5. Let X be a discrete one-dimensional random variable with pmf f(x) and sample space S. Then

$$H_O(\theta) \geq 0. \tag{2.7}$$

Proof. For all $x \in S$, f(x) is between 0 and 1 and $\sum_{x \in S} f(x|\theta) = 1$. Hence, $0 \leq \sum_{x \in S}[f(x|\theta)]^2 \leq 1$, and $-\log \sum_{x \in S}[f(x|\theta)]^2 \geq 0$. ■

We next use a well-known inequality to compare the inner and outer entropy.

Theorem 2.6 (Jensen's Inequality) [6]. If f is a convex function and X is any random variable, then

$$E[f(X)] \geq f(E[X]), \tag{2.8}$$

where equality holds either if X has a single value or if f is linear.

Theorem 2.7. For any discrete random variable X with fixed θ,

$$H_I(\theta) \geq H_O(\theta). \tag{2.9}$$

Proof. Since $g(y) = -\log(y)$, then g(y) is a convex function of y on the convex set (0,1]. Letting $y = f(x|\theta)$ gives

$$E[-\log f(X|\theta)] \geq -\log E[f(X|\theta)] \tag{2.10}$$

from (2.9), so (2.10) follows. ■

Outer entropy can obviously be extended to the bivariate case with the multivariate case, with joint and conditional outer entropy analogs to those of inner entropy. Instead, we define further entropies using the gambler's information in (2.4) and (2.6).

Definition 2.9 (Gambler's Inner and Outer Entropy). Let X be a discrete one-dimensional random variable with pmf $f(x|\theta)$ and sample space S. Then gambler's inner entropy $H_I^g(\theta)$ with respect to gambler's information is defined as

$$H_I^g(\theta) = E[I^g(X|\theta)] = -\sum_{x \in S} f(x|\theta) \log\{1 - f(x|\theta)\}. \tag{2.11}$$

Similarly, the gambler's outer entropy $H_O^g(X)$ with respect to gambler's information is defined as

$$H_O^g(\theta) = I^g[E(X|\theta)] = -\log\left\{1 - \sum_{x\in S}[f(x|\theta)]^2\right\}. \tag{2.12}$$

In Section 4, we use (2.12) to define a new measure of evidence for the parameter θ of X.

3. Examples

We now present examples of the definitions of Section 2 for some standard discrete distributions. In each example, formulas for both the standard Shannon inner entropy of Shannon and the new Shannon outer entropy are given.

3.1 Uniform Distribution. Let X be uniformly distributed with $S = \{1, 2, \dots, N\}$. Then

$$f(x|N) = \begin{cases} \frac{1}{N} & x = 1, 2, \dots, N \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

Shannon Inner Entropy. The inner entropy is easily calculated from (2.3) to be

$$H_I(N) = \sum_{x=1}^{N} -\frac{1}{N}\log\left(\frac{1}{N}\right) = \log N, \tag{3.2}$$

Shannon Outer Entropy. Similarly, from (2.5)

$$H_O(N) = -\log\sum_{x=1}^{N}\left(\frac{1}{N}\right)^2 = \log N. \tag{3.3}$$

Equations (3.2) and (3.3) show that $H_I(X) = H_O(X)$ for a discrete uniform distribution, which gives the equality case in (2.9). Note that both the entropy of both (3.2) and (3.3) increases linearly with N. Despite the increase in average Shannon information, however, a rational gambler would be increasingly unlikely to bet on the occurrence of any particular value of X.

Gambler's Inner Entropy. From (2.11) we get

$$H_I^g(N) = -\sum_{i=1}^{N}\frac{1}{N}\log\{1 - \frac{1}{N}\} = -\log\{1 - \frac{1}{N}\}. \tag{3.4}$$

Gambler's Outer Entropy. Also, by equation (2.12) the gambler's outer entropy is

$$H_O^g(N) = -\log\left\{1 - \sum_{i=1}^{N}\left[\frac{1}{N}\right]^2\right\} = -\log\{1 - \frac{1}{N}\}. \tag{3.5}$$

From equations (3.4) and (3.5), for a uniform discrete random variable, the gambler's inner and outer entropies are identical as was the case for the Shannon inner and outer entropies. In this case, however, both the Shannon inner and outer entropies approach 0 as N gets large. In other words, as N increased a rational gambler could use either gambler's inner or outer entropy as increasing evidence against betting on the occurrence of any particular value of X as N increased.

3.2 Bernoulli Distribution. Let X have a Bernoulli distribution with pmf

$$f(x|p) = \begin{cases} p & \text{for } x = 1 \\ q = 1 - p & \text{for } x = 0. \end{cases} \tag{3.6}$$

Shannon Inner Entropy. From (2.3)

$$H_I(p) = -p \log p - (1-p)\log(1-p). \tag{3.7}$$

Shannon Outer Entropy. Similarly, from (2.5)

$$H_O(p) = -\log[p^2 + (1-p)^2]. \tag{3.8}$$

As an example, consider the Bernoulli experiment of tossing a coin with a known probability p of obtaining heads. **Figure 3.1** below shows the Shannon inner and outer entropies have the same pattern as p changes.

Fig 3.1. Shannon Inner and Outer Entropy for Bernoulli Distribution

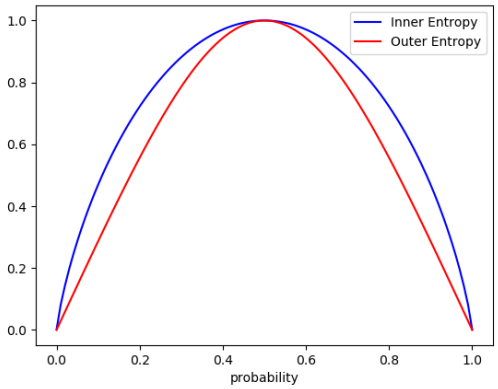


Figure 3.1 depicts the fact that the maximum of both the Shannon inner and outer entropy occurs at the $p = \frac{1}{2}$. From (3.7)

$$H_I\left(\frac{1}{2}\right) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = -\log\frac{1}{2} = 1 \text{ bit}, \tag{3.9}$$

and from (3.8)

$$H_O\left(\frac{1}{2}\right) = -\log\left[\frac{1^2}{2} + \left(1-\frac{1}{2}\right)^2\right] = 1 \text{ bit}. \tag{3.10}$$

Moreover, the inner and outer entropy achieve a maximum value when the coin is fair.

Gambler's Inner Entropy. From (2.12)

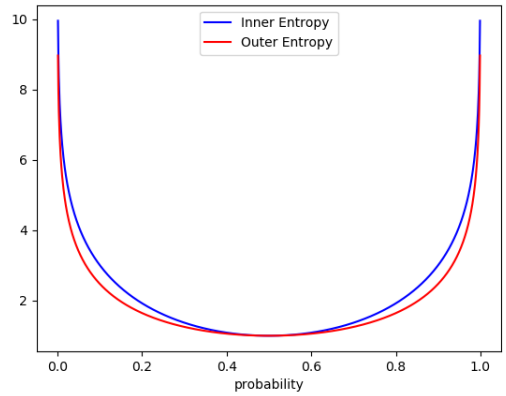
$$H_I^g(p) = -[p\log(1-p) + (1-p)\log p] \tag{3.11}$$

Gambler's Outer Entropy. Similarly, from (2.13)

$$H_O^g(p) = -\log\{2p - 2p^2\}. \tag{3.12}$$

In **Example 3.1**, it was noted that is an inverse relationship between Shannon and gambler's entropies for the uniform distribution. **Figure 3.2** plots gambler's entropies against p for a Bernoulli distribution. A similar inverse relationship is evident for the Bernoulli distribution by comparing **Figures 3.1** with **Figure 3.2**

Fig 3.2. Gambler's Inner and Outer Entropy for Bernoulli Distribution



3.3 Geometric Distribution. Let X have a geometric distribution with pmf

$$g(x|p) = p(1-p)^{x-1},\ x = 1, 2, \ldots., \tag{3.13}$$

where the parameter p is the probability of success for each of the associated Bernoulli trials.

Shannon Inner Entropy. In this case the Shannon inner entropy has the closed form [25]

$$H_I(p) = \frac{-p\log p - (1-p)\log(1-p)}{p}. \tag{3.14}$$

Shannon Outer Entropy. From (2.5) we have

$$H_O(p) = -\log\sum_{x=1}^{\infty}[p(1-p)^{x-1}]^2 = -\log\left[p^2\sum_{x=1}^{\infty}(1-p)^{2x-2}\right]. \tag{3.15}$$

Factoring $(1-p)^{-2}$ out of the series in (3.15) and summing the resulting geometric series gives

$$H_O(p) = -\log\frac{p}{2-p}. \tag{3.16}$$

Note that the outer entropy of (3.16) as a more concise closed form in comparison to (3.15).

Gambler's Inner Entropy. From equation (2.11), the gambler's inner entropy is

$$H_I^g(p) = -\sum_{x=1}^{\infty}p(1-p)^{x-1}\log\{1-p(1-p)^{x-1}\}, \tag{3.17}$$

which has no obvious closed form.

Gambler's Outer Entropy. By equation (2.12) the gambler's outer entropy is

$$H_O^g(p) = -\log\sum_{x=1}^{\infty}\{1-[p(1-p)^{x-1}]^2\}, \tag{3.18}$$

which can be simplified to

$$H_O^g(p) = -\log\frac{2-2p}{2-p}. \tag{3.19}$$

Again, the outer entropy of (3.19) has a closed form as opposed to (3.18).

3.4 Poisson Distribution. Now let X have a Poisson distribution with parameter $\lambda > 0$ and pmf

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \text{for } x = 0,1, \dots . \tag{3.20}$$

Shannon Inner Entropy. The Shannon inner entropy is given in [26] as

$$H_I(\lambda) = \lambda[1 - \log(\lambda)] + e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x \log(x!)}{x!}, \tag{3.21}$$

for which there is apparently no closed form.

Shannon Outer Entropy. From (2.5)

$$H_O(\lambda) = -\log \sum_{x=0}^{\infty} \left(\frac{e^{-\lambda}\lambda^x}{x!} \right)^2. \tag{3.22}$$

Thus

$$H_O(\lambda) = -\log \left[e^{-2\lambda} \sum_{x=0}^{\infty} (\frac{\lambda^x}{x!})^2 \right]. \tag{3.23}$$

The series in (3.23) is given [27], from which

$$H_O(\lambda) = 2\lambda \log e - \log I_0(2\lambda), \tag{3.24}$$

where I_0 is a modified Bessel function of the first kind. Again, the notion of outer entropy yields a closed form in (3.24) as opposed to the inner entropy of (3.21).

Gambler's Inner Entropy. From (2.11) the gambler's inner entropy is

$$H_I^g(\lambda) = -\sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \log \left\{ 1 - \frac{e^{-\lambda}\lambda^x}{x!} \right\}. \tag{3.25}$$

Gambler's Outer Entropy. From (2.12) the gambler's outer entropy is

$$H_O^g(\lambda) = -\log \left\{ 1 - \sum_{x=0}^{\infty} \left(\frac{e^{-\lambda}\lambda^x}{x!} \right)^2 \right\}, \tag{3.26}$$

which simplifies as in (3.24) to

$$H_O^g(\lambda) = -\log\{1 - e^{-2\lambda} I_0(2\lambda)\}. \tag{3.27}$$

Table 3.1 summarize the results of this section, also it provides a basic vision of calculation difficulties on inner and outer entropy.

Table 3.1. Summary of Different Entropies for Some Important Discrete Distributions

Distribution	pmf	H_I	H_O	H_I^g	H_O^g
Uniform	$\frac{1}{N}$ $x = 1, 2, \dots, N$	$\log N$	$\log N$	$-\log\{1 - \frac{1}{N}\}$	$-\log\{1 - \frac{1}{N}\}$
Bernoulli	$\begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases}$	$-p \log p - (1 - p) \log(1 - p)$	$-\log[1 - 2p + 2p^2]$	$-[p \log(1 - p) + (1 - p) \log p]$	$-\log\{2p - 2p^2\}$
Geometric	$p(1 - p)^{x-1}$ for $x = 1, 2, \dots$	$\frac{-p \log p - (1 - p) \log(1 - p)}{p}$	$-\log \frac{p}{2 - p}$	$-\sum_{x=1}^{\infty} p(1 - p)^{x-1} * \log\{1 - p(1 - p)^{x-1}\}$	$-\log \frac{2 - 2p}{2 - p}$
Poisson	$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$, for $x = 0,1, \dots$	$\lambda[1 - \log(\lambda)] + e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x \log(x!)}{x!}$	$2\lambda \log e - \log I_0(2\lambda)$	$-\sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \log\{1 - \frac{e^{-\lambda}\lambda^x}{x!}\}$	$-\log\{1 - e^{-2\lambda} I_0(2\lambda)\}$

4. Applications

In this section we develop some theorems and present some applications of the proposed new outer entropies and gambler's information. In particular, we use gambler's information to measure the amount of information loss due to data compression instead of using Shannon information as in [1]. We also use outer entropy to define a new measure of evidence about a single parameter of a discrete random variable.

4.1. Gambler's Information Approximation

We now use the well-known Maclaurin series to estimate gambler's information. The first order approximation of Maclaurin series [28] to be used is stated as **Result 4.1**.

Result 4.1. For sufficiently small $0 < p < 1$,

$$-\ln(1-p) = \sum_{n=1}^{\infty} \frac{p^n}{n} \approx p. \tag{4.1}$$

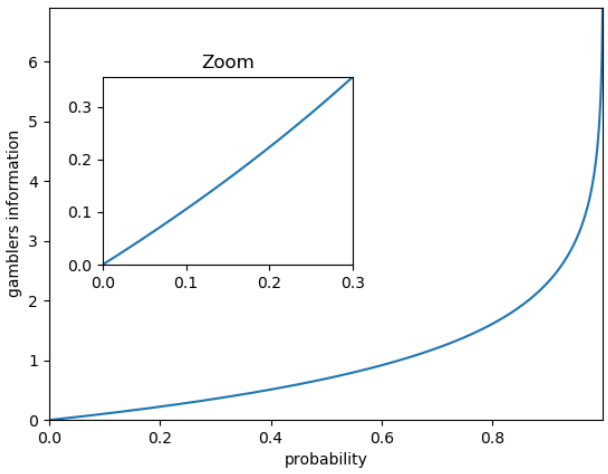
However, we change the left-hand side of (4.1) to bits, i.e., from the base $e \approx 2.71828$ to the base 2 used in the definition of gambler's information. Information defined for the natural logarithm has units of in nats (also called nits or nepits) instead of bits. To transform the units of (4.1) from nats to bits we can change the base of the natural log in (4.1) using the formula $\log_2 c = \frac{\ln c}{\ln 2}$ [29]. Then one nat is $\frac{1}{\ln 2} = 1.44$ bits to two decimal places.

We now apply (4.1) with $p = f(x|\theta)$, the pdf of the discrete random variable X with one-dimensional parameter of interest θ. Then from (2.2) the left-hand side of (4.1) is the gambler's information in nats associated with any $x \in S$. In other words,

$$I^g(x|\theta) \approx f(x|\theta) \text{ nats.} \tag{4.2}$$

Equation (4.2) is a good approximation for $x \leq 0.3$, after which the nonlinear terms become relevant. This fact is illustrated in the expanded part of **Figure 4.1.** For $f(x) = p = 0.3$ the gambler's information in nats is 0.357. In this case, the percentage change between the actual information at $p = 0.3$ and the estimate 0.3 is $\frac{-\ln(1-0.3)-0.3}{-\ln(1-0.3)} \times 100$, or 15.9%. Thus the difference in the actual gambler's information in nats and the approximation from (4.1) is less than 15.9% of the actual value for $0 < p < 0.3$. In comparison, for $0 < p < 0.2$ the percentage change is less than 9.1%. For $p = 0.4$ this change goes up to 21.7%, and for $p = 0.5$ the percentage change is 27.9%.

Fig 4.1. Gambler's Information Approximation



4.2. Information Loss

We now apply gambler's information instead of Shannon information, together with the approximation of Section 4.1, to the results of [1] concerning the information lost when a sufficient statistic is used to characterize a random sample. Consider the data sample $\mathbf{x} = (x_1, \dots, x_n)$ from a random sample $\mathbf{X} = (X_1, \dots, X_n)$ for a discrete random variable X with sample space S, pmf h(x), and one-dimensional parameter θ. For a data sample $\mathbf{x} = (x_1, \dots, x_n)$ let the joint pmf of $\mathbf{X}$ be $f(\mathbf{x}|\theta) = \prod_{i=1}^{n} h(x_i|\theta)$. This data sample is compressed to a real-valued summary statistic $T(\mathbf{X})$, which may be used to characterize $\mathbf{X}$ or to estimate θ. Such data compression is an irreversible process [30] and always involves some information loss. In [1], we developed a procedure to determine how much of the information about $\mathbf{X}$ contained in a data set $\mathbf{x}$ is lost when the data is compressed to a sufficient statistic $T(\mathbf{x})$. This lost information represents a combinatorial loss in the sense that multiple $\mathbf{x}$'s may give the same value $T(\mathbf{x}) = t$. In other words, the lost information $-\log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]$ is a measure of the knowledge unavailable about the data sample $\mathbf{x}$ when only the compressed data summary $T(\mathbf{x})$ is known and not $\mathbf{x}$ itself.

In contrast to using log to the base 2 in equation (2.2) above, in this section gambler's information will be in nats instead of bits. In other words, $I^g(x|\theta) = -\ln[1 - f(x|\theta)]$ now so that approximation (4.1) is not $-\log[1 - f(x|\theta)] \approx 1.44\, f(x|\theta)$ but rather $-\ln[1 - f(x|\theta)] \approx f(x|\theta)$. We therefore avoid the constant factor 1.44 in our equations.

The general decomposition of information of [1] is summarized in **Definition 4.2**, where T does not need to be sufficient for θ.

Definition 4.2 ($\mathbf{I_{total}}, \mathbf{I_{comp}}, \mathbf{I_{lost}}$). Let $\mathbf{x}$ be sample data for a random sample $\mathbf{X}$ from a discrete random variable X as described above, let P_θ be an appropriate probability function involving the parameter θ, and let $T(\mathbf{X})$ be any real-valued statistic. The Shannon information about $\mathbf{X}$ obtained from the sample data $\mathbf{x}$ can be decomposed as

$$I_{total}(\mathbf{x}|\theta) = I_{comp}(\mathbf{x}|\theta, T) + I_{lost}(\mathbf{x}|\theta, T), \tag{4.3}$$

where

$$I_{total}(\mathbf{x}|\theta) = -\log P_\theta[\mathbf{X} = \mathbf{x}], \tag{4.4}$$

$$I_{comp}(\mathbf{x}|\theta, T) = -\log P_\theta[T(\mathbf{X}) = T(\mathbf{x})], \tag{4.5}$$

and

$$I_{lost}(\mathbf{x}|\theta, T) = -\log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]. \tag{4.6}$$

In **Definition 4.2** the information loss and information decomposition is in bits of Shannon information since the logarithm base is 2. We now derive the gambler's information decomposition for a specific data set. We start with total information. Since the probability associated with total information is $P_\theta[\mathbf{X} = \mathbf{x}]$, we define the gambler's total information using the natural logarithm in (2.2) as

$$I^g_{total}(\mathbf{x}|\theta) = -\ln\{1 - P_\theta[\mathbf{X} = \mathbf{x}]\} = -\ln[\,1 - f(\mathbf{x}|\theta)]. \tag{4.7}$$

Note that $f(\mathbf{x}) = \prod_{i=1}^{n} h(x_i)$, where $0 < h(x_i) < 1$, for a sample space S with cardinality $\|S\| > 1$. For such a sample space, it follows that as the sample size increases, then $f(\mathbf{x})$ decreases until it is small enough for approximation (4.1) to apply. For example, $n = 25$ and $h(x_i) = 0.9, i = 1 \dots, 25$, give an unrealistically high joint probability of $f(\mathbf{x}) \approx 0.073$ on the right side of (4.1) for 0.074 on the left to illustrate the practical validity of the approximation. Thus for a sufficiently large sample size, we can write (4.7) in nats as

$$I^g_{total}(\mathbf{x}|\theta) \approx P_\theta[\mathbf{X} = \mathbf{x}] = f(\mathbf{x}|\theta)]. \tag{4.8}$$

This approximation can be used for total gambler's information but not for the compressed and lost information since it is shown in [1] that $P_\theta[T(\mathbf{X}) = T(\mathbf{x})] > P_\theta[\mathbf{X} = \mathbf{x}]$ and $P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})] > P_\theta[\mathbf{X} = \mathbf{x}]$. However, from **Figure 4.1** if $P_\theta[T(\mathbf{X}) = T(\mathbf{x})]$ and $P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]$ are less than 0.3 we can use (4.2) to write

$$I^g_{comp}(\mathbf{x}|\theta, T) = -\ln\{1 - P_\theta[T(\mathbf{X}) = T(\mathbf{x})]\} \approx P_\theta[T(\mathbf{X}) = T(\mathbf{x})] \tag{4.9}$$

and

$$I^g_{lost}(\mathbf{x}|\theta, T) = -\ln\{1 - P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]\} \approx P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})], \tag{4.10}$$

also in nats. It is shown in (1) that

$$P_\theta[\mathbf{X} = \mathbf{x}] = P_\theta[T(\mathbf{X}) = T(\mathbf{x})] \times P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]. \tag{4.11}$$

But now from (4.8) - (4.10), for sufficiently small enough probabilities, we can substitute for probability terms in (4.11) for the corresponded approximate gambler's information in nats to give the approximate decomposition

$$I^g_{total}(\mathbf{x}|\theta) \approx I^g_{comp}(\mathbf{x}|\theta, T) \times I^g_{lost}(\mathbf{x}|\theta, T). \tag{4.12}$$

By comparing (4.12) to (4.3), it is seen that in the probability range of approximation (4.2), Shannon is decomposed additively while gambler's information is decomposed multiplicatively.

Note further that taking the negative log to the base 2 of (4.8) gives

$$-\log I^g_{total}(\mathbf{x}|\theta) \approx -\log P_\theta[\mathbf{X} = \mathbf{x}]. \tag{4.13}$$

But the right-hand side of (4.13) is the Shannon information for $f(\mathbf{x})$. Hence, for a sufficiently large sample size, then

$$-\log I^g_{total}(\mathbf{x}|\theta) \approx I(\mathbf{x}|\theta), \tag{4.14}$$

where $I^g_{total}(\mathbf{x}|\theta)$ is in nats, and $I(\mathbf{x}|\theta)$ is in bits.

4.3. Entropic Evidence

Using the notation of Section 4.2 we propose in this section two new metrics involving outer entropy that provide evidence whether one estimate of the parameter θ for a random sample $\mathbf{X}$ may be considered better than another according to the metrics. In the process, we give two new numerical point estimates for θ independent of a data sample. However, we first comment on the relation of statistical data to evidence. The attempt to decrease the uncertainty may be considered within a hierarchical framework [31] in which data is transformed into information and this information is transformed into evidence. This evidence is then used to test hypotheses and check assertions, and the original data is thereby transformed into knowledge. In other words,

$$\text{data} \Rightarrow \text{information} \Rightarrow \text{evidence} \Rightarrow \text{knowledge}. \tag{4.15}$$

Within the context of (4.15) we now define the notion of entropic evidence about the parameter θ in terms of outer entropy.

Definition 4.3 (Entropic Evidence). Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample with joint pmf $f(\mathbf{x}|\theta)$ from a random variable X with sample space S. Then using (2.5) and (2.12), the Shannon entropic (SE) evidence about the parameter θ is defined as

$$SE(\theta) = H_O(\theta) = -\log\left\{\sum_{\mathbf{x}\in S^n} [f(\mathbf{x}|\theta)]^2\right\} \tag{4.16}$$

and the gambler's entropic (GE) evidence about the parameter θ

$$GE(\theta) = H_O^g(\theta) = -\log\left\{1 - \sum_{\mathbf{x}\in S^n} [f(\mathbf{x}|\theta)]^2\right\}. \tag{4.17}$$

Both $SE(\theta)$ and $GE(\theta)$ are obviously nonnegative. They can be used in a manner similar to the approach yielding a maximum likelihood estimator. A maximum likelihood estimator (MLE) is the statistic $\hat{\theta}(\mathbf{x})$ in terms of the random data sample $\mathbf{x} = (x_1, \ldots, x_n)$ that maximizes the joint probability $f(\mathbf{x}|\hat{\theta}(\mathbf{x}))$ - i.e., the likelihood function – that the random sample $\mathbf{X}$ takes the values of the observed data $\mathbf{x}$. Equivalently, the MLE is obtained by maximizing $\log f(\mathbf{x}|\theta)$ or minimizing the Shannon information given by $-\log f(\mathbf{x}|\theta)$.

Now write (4.16) as

$$SE(\theta) = H_O(\theta) = -\log \bar{f}(\theta) \tag{4.18}$$

and (4.17) as

$$GE(\theta) = H_O^g(\theta) = -\log\{1 - \bar{f}(\theta)\}, \tag{4.19}$$

where $\bar{f}(\theta)$ is the expected value $\sum_{\mathbf{x}\in S^n}[f(\mathbf{x}|\theta)]^2$ of the pmf $f(\mathbf{x}|\theta)$ over $\mathbf{x} = (x_1, \ldots, x_n)$, which is a function only of θ. Note that (4.18) and (4.19) are almost identical to (2.1) and (2.2), respectively, with the function f replaced by $\bar{f}$. Using inner entropies would not give this simple analog despite inner and outer entropy have similar graphical characteristics as indicated in **Figure 2.1**.

We propose that (4.18) and (4.19) are useful measures of evidence in estimating θ. In particular, from the discussion after **Definition 4.3**, the fact that an MLE $\hat{\theta}$ minimizes (2.1) suggests that minimizing (4.18) would give a useful numerical estimate θ^* for θ with no dependence on the data sample. Moreover, it follows that a parameter value θ_1 for X with a smaller value of (4.18) may be considered better than a value θ_2 with a larger value of (4.18). Equivalently, a minimum Shannon entropic (MSE) estimate would maximize $\bar{f}(\theta)$ over θ to give a numerical estimate θ^* independent of a data sample. On the other hand, minimizing (4.19) would over θ to give an analogous minimum gambler's entropic (MGE) estimate would be equivalent to minimizing $\bar{f}(\theta)$ over θ to give a non-MSE numerical estimate θ^* independent of a data sample. The analogy of the MSE to an MLE suggests that an MSE estimate would be a better estimator for θ than an MGE estimate. It should be noted that maximizing (4.19) is equivalent to minimizing (4.18).

An example is now presented to illustrate the procedure for calculating the SE and GE evidence associated with the parameter for a binomial distribution.

Example 4.4 (Binomial Distribution). Consider the experiment of flipping a possibly biased coin twice ($m = 2$). The total number of heads X in the experiment follows a binomial distribution with $S = \{0,1,2,\}$ and the parameter θ being the probability of getting a head on any flip. By doing this experiment three times ($n = 3$) we generate the random sample $\mathbf{X} = (X_1, X_2, X_3)$. **Table 4.1** shows all the possibilities and their average probabilities.

Table 4.1. Binomial Distribution Average PMF Probabilities

$\mathbf{x} = (x_1, x_2, x_3)$	$f(\mathbf{x})$	$f^2(\mathbf{x})$
(0,0,0)	$(1-\theta)^6$	$(1-\theta)^{12}$
(0,0,1)	$2(1-\theta)^5\theta^1$	$4(1-\theta)^{10}\theta^2$
(0,1,0)		
(1,0,0)		
(1,1,0)	$4(1-\theta)^4\theta^2$	$16(1-\theta)^8\theta^4$
(1,0,1)		
(0,1,1)		
(2,0,0)		
(0,2,0)		

(0,0,2)	$(1-\theta)^4\theta^2$	$(1-\theta)^8\theta^4$
(1,1,1)	$8(1-\theta)^3\theta^3$	$64(1-\theta)^6\theta^6$
(2,1,0)	$2(1-\theta)^3\theta^3$	$4(1-\theta)^6\theta^6$
(2,0,1)		
(1,0,2)		
(1,2,0)		
(0,1,2)		
(0,2,1)		
(2,1,1)	$4(1-\theta)^2\theta^4$	$16(1-\theta)^4\theta^8$
(1,2,1)		
(1,1,2)		
(2,2,0)	$(1-\theta)^2\theta^4$	$(1-\theta)^4\theta^8$
(2,0,2)		
(0,2,2)		
(2,2,1)	$2(1-\theta)^1\theta^5$	$4(1-\theta)^2\theta^{10}$
(2,1,2)		
(1,2,2)		
(2,2,2)	θ^6	θ^{12}

By summing over the last column of the **Table 4.1**, we calculate $\bar{f}(\theta) = \sum_{x\in S}[f(x)]^2$ as

$$\bar{f}(\theta) = (1-\theta)^{12} + 12(1-\theta)^{10}\theta^2 + 51(1-\theta)^8\theta^4 + 88(1-\theta)^6\theta^6 + 51(1-\theta)^4\theta^8 + 12(1-\theta)^2\theta^{10} + \theta^{12}. \quad (4.20)$$

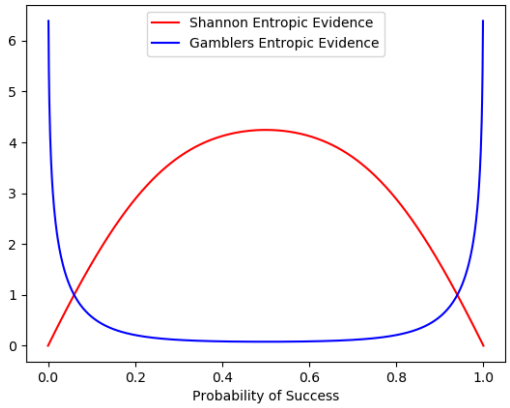
Substituting $\bar{f}(\theta)$ calculated in (4.20) into (4.18) and (4.19) gives the SE and GE evidence shown in **Table 4.2** for 5 different value of θ. **Table 4.2** and equation (4.20) indicate that $SE(\theta)$ and $GE(\theta)$ are symmetric functions for the binomial distribution. This property is not true for nonsymmetric distributions such as the Poisson distribution.

Table 4.2. Entropic Evidence for different θ

θ	SE(θ)	GE(θ)
0.2	2.884	0.210
0.25	3.352	0.149
0.5	4.245	0.078
0.75	3.352	0.149
0.8	2.884	0.210

Figure 4.2 is a plot of $SE(\theta)$ and $GE(\theta)$ vs θ in which $SE(\theta)$ is minimized for $\theta = 0$ or 1 with $SE(\theta) = 0$. This value confirms that a certain head or a sure tail produces no surprise. This result also gives insight into a general MLE $\hat{\theta}(\mathbf{x})$ of the parameter θ, which minimizes Shannon information as previously noted. By maximizing the joint pmf of a random sample **X** over θ, the MLE $\hat{\theta}(\mathbf{x})$ minimizes the surprise that the sample data would give. Similarly, $GE(\theta)$ is minimized by $\theta = 0.5$ in **Figure 4.2**. In this case, a fair coin minimizes the certitude of a flip of the coin. We conclude that the choice of an MSE or MGE estimator depends on whether the goal of a data analyst is to suppress surprise or certitude or, respectively equivalent, whether to emphasize certitude or surprise.

Fig 4.2. Entropic Evidence for Binomial Distribution



5. Conclusions

In this paper, we have considered only discrete random variables, but the results can be extended to continuous ones as well. We defined here a new measure of information called gambler's information (or certitude) in contrast to Shannon information (or surprisal) for discrete random variables. Gambler's information takes a prospective view of an event and measures the level of probabilistic certitude that it may occur. Gambler's information is obtained before the event occurs. In effect, this level of certitude is the probability itself and increases for an increasing probability. On the other hand, Shannon information takes a retrospective view of an event and measures the level of surprise incurred if the event did occur. The surprisal associated with an event increases for a decreasing probability of the event. The choice of information type to use in a particular model depends on the retrospective or prospective nature of the model.

We also defined a new type of entropy called outer entropy by moving the log function outside the expectation in contrast to Shannon entropy to facilitate both intuitive appeal and mathematical manipulation. Outer entropy can be defined using either Shannon or gambler's information. For Shannon information, it becomes the negative log of the mean pmf value of a random variable. For gambler's information, it becomes the negative log of 1 – (the mean pmf value). An intriguing question is whether gambler's information and outer entropy would be useful as thermodynamic tools since the use of inner entropy in physics predates Shannon and plays a major role in thermodynamics.

In addition, we provided examples of the concepts introduced here and gave two applications. The first application determines the gambler's information lost when a random data sample is characterized by a single statistic such as the mean of the underlying random variable. The second application uses outer entropy as a new metric for deciding between possible numerical parameter values for the underlying random variable.

References

1. Moghimi, M., Corley, H. W. (2020). Information loss due to the compression of sample data from discrete distributions. Entropy (Under second review).
2. Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal. https://doi.org/10.1002/j.1538-7305.1948.tb01338.
3. Jaynes, E. T. (1957). Information theory and statistical mechanics. Physical Review. https://doi.org/10.1103/PhysRev.106.620.

4. Csiszár, I. (2008). Axiomatic characterizations of information measures. Entropy. https://doi.org/10.3390/e10030261.
5. Kapur J.N., Kesavan H.K. (1992). Entropy Optimization Principles and Their Applications. Springer: Dordrecht. https://doi.org/10.1007/978-94-011-2430-0_1.
6. Cover, T. M., Thomas, J. A. (2005). Elements of Information Theory. John Wiley & Sons, Inc.: NJ, USA. https://doi.org/10.1002/047174882X.
7. Ibekwe-SanJuan, F., Dousa, T. (2014). Theories of Information, Communication and Knowledge: A Multidisciplinary Approach. Springer: London. https://doi.org/10.1007/978-94-007-6973-1.
8. Vigo, R. (2011). Representational information: A new general notion and measure of information. Information Sciences. https://doi.org/10.1016/j.ins.2011.05.020.
9. Vigo, R. (2013). Complexity over uncertainty in generalized representational information theory (GRIT): A structure-sensitive general theory of information. Information (Switzerland). https://doi.org/10.3390/info4010001.
10. Klir, G.J. (2006). Uncertainty and Information: Foundations of Generalized Information Theory. John Wiley & Sons, Inc.: NJ, USA. https://doi.org/10.1002/0471755575.
11. Devlin, K. (1991). Logic and Information. Cambridge University Press: Cambridge, UK.
12. Luce, R. D. (2003). Whatever happened to information theory in psychology? Review of General Psychology. https://doi.org/10.1037/1089-2680.7.2.183
13. Floridi, L. (2011). The Philosophy of Information. Oxford University Press: Oxford, UK. https://doi.org/10.1093/acprof:oso/9780199232383.001.0001.
14. Garner., W. R. (1974). The Processing of Information and Structure, Wiley: NY, USA. https://doi.org/10.2307/1421985.
15. Spellerberg, I. F., Fedor, P. J. (2003). A tribute to Claude-Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the "Shannon-Wiener" index. Global Ecology and Biogeography. https://doi.org/10.1046/j.1466-822X.2003.00015.x.
16. Tabner, I. T. (2007). A review of concentration, diversity or entropy metrics in economics, finance, ecology and communication science. The International Journal of Interdisciplinary Social Sciences: Annual Review 2. https://doi.org/10.18848/1833-1882/cgp/v02i04/52345.
17. Lad, F., Sanfilippo, G., Agrò, G. (2015). Extropy: Complementary dual of entropy. Statistical Science. https://doi.org/10.1214/14-STS430.
18. Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., Meiners, T., Müller, C., Obermaier, E., Prati, D., Socher, S. A., Sonnemann, I., Wäschke, N., Wubet, T., Wurst, S., Rillig, M. C. (2014). Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. Ecology and Evolution. https://doi.org/10.1002/ece3.1155.
19. DeDeo, S. Information theory for intelligent people. http://tuvalu.santafe.edu/~simon/it.pdf (accessed on April 22, 2020).
20. Renyi, A. (1961). On measures of entropy and information. Fourth Berkeley Symposium on Mathematical Statistics and Probability.
21. Bromiley, P. (2004). Shannon entropy, Renyi entropy, and information. Statistics and Information Series. https://doi.org/10.1016/j.patrec.2004.03.003.
22. Amigó, J. M., Balogh, S. G., Hernández, S. (2018). A brief review of generalized entropies. Entropy. https://doi.org/10.3390/e20110813.
23. Jäckle, S., Keller, K. (2017). Tsallis entropy and generalized Shannon additivity. Axioms. https://doi.org/10.3390/axioms6020014.
24. Schroeder, M. J. (2004). An alternative to entropy in the measurement of information. Entropy. https://doi.org/10.3390/e6050388.

25. Traylor, R. (2017). A generalized geometric distribution from vertically dependent Bernoulli random variables. Academic Advances of the CTO. https://www.themathcitadel.com/wp-content/uploads/2017/07/generalized-geometric.pdf (accessed on April 22, 2020).
26. Cheraghchi, M. (2018). Expressions for the entropy of binomial-type distributions. IEEE International Symposium on Information Theory. https://doi.org/10.1109/ISIT.2018.8437888.
27. https://www.wolframalpha.com/input/?i=sum+k+%3D+0+to+infinity+x%5Ek%2F%28k%21%29%5E2 (accessed on April 22, 2020).
28. Apostol, T.M. (1974). Mathematical Analysis. 2nd Edition, Addison-Wesley: Boston, USA.
29. https://en.wikipedia.org/wiki/Logarithm (accessed on April 22, 2020).
30. Landauer, R. (2010). Irreversibility and heat generation in the computing process. IBM Journal of Research and Development. https://doi.org/10.1147/rd.53.0183.
31. Dammann, O. (2019). Data, information, evidence, and knowledge: a proposal for health informatics and data science. Online Journal of Public Health Informatics. https://doi.org/10.5210/ojphi.v10i3.9631.

Chapter 4

Comparison and Extension of Measures of Evidence in Hypothesis Testing

Maryam Moghimi*[1,2], H.W. Corley[1,2]

[1] Center on Stochastic Modeling, Optimization, and Statistics (COSMOS), The University of Texas at Arlington, Arlington, TX, USA
[2] The authors contributed equally to this paper.
* Correspondence: maryam.moghimi@uta.edu; +1-214-971-0904 (M.M.), corley@uta.edu; Tel.: +1-817-272-3092 (H.C.)

Abstract. In this paper we present some measures for statistical evidence for testing hypotheses. These measures include both frequentist and Bayesian approaches. The likelihood ratio and confidence distribution first provide frequentist measures of evidence, where the confidence distribution is derived from the notion of confidence intervals. The Bayesian posterior distribution is then discussed. It is noted that a posterior distribution from a noninformative prior gives the standard frequentist P-value. We then define a novel P-value using the maximum likelihood estimator (MLE) and without using either Type I error or the assumption that null hypothesis is true. In fact, none of the approaches discussed here provide the probabilities of Type I or Type II error. However, an example for the confidence distribution illustrates how error rates can be approximated using simulation.

Keywords: hypothesis testing, evidence, likelihood ratio, confidence distribution, odds ratio, P-value, frequentist approach, Bayesian approach, maximum likelihood estimator

1. Introduction

A principal goal of statistics is to obtain evidence from data for comparing alternative decisions. For example, statistical evidence may allow one to decide that a population mean μ satisfies $\mu \leq \mu_0$ as opposed to $\mu > \mu_0$ for some specified μ_0. Unfortunately, evidence is an ambiguous concept in statistics, though [1-5], among others, have attempted to define it. However, in arguments about the likelihood principle, [6-12] have simply not defined evidence despite it being central to their arguments. Currently, the p-value [13,14,15], the likelihood ratio [4,16], the Bayes factor [17], and the posterior odds ratio [18] are the most frequently applied measures of evidence. Evans [2] suggested but did not pursue the idea that the evidence an event B gives about an event A is simply the difference between P(A|B) and P(A). Recently, Vieland [19, 20] has proposed an axiomatic approach by considering evidence to be analogous to temperature in thermodynamics. In addition, related to evidence is the notion of belief considered in [2, 21, 22, 23], which includes the well-known Dempster-Shafer theory. In this paper, we discuss the likelihood ratio, the confidence distributions, the Bayesian posterior distribution, and the P-value as measures for statistical evidence with regard to hypothesis testing. We then propose a new P-value that is not related to significance levels and not defined under the assumption that the null hypothesis is true.

The paper is organized as follows. In Section 2, we present some data science preliminaries to provide a metric used subsequently to compare the performance of the different measures of evidence. In Section 3, we review the likelihood ratio as a measure of comparative. In Section 4, we consider the confidence distribution as a measure of evidence. Confidence distributions are Bayesian-like yet frequentist probability

distributions derived from confidence intervals. Either the confidence that a null hypothesis is true can be used as evidence. Alternately, the ratio of the confidence distribution probabilities associated with the null and alternate hypotheses can be used as a measure of comparative evidence.

In Section 5, the Bayesian posterior odds ratio is summarized as a measure of evidence. The posterior odds ratio for noninformative priors leads to the same ratio obtained in Sections 4 for normal distributions resulting from the application of the Central Limit Theorem. In Section 6, the standard P-value of a hypothesis test is discussed. Then a new definition for P-value is proposed without reference to significance levels and without the assumption that the null hypothesis is true.

In Section 7, we present an example for some normally distributed samples by testing a hypothesis using confidence distributions. We simulate to determine values for the Type I and Type II errors of the test. Conclusions are offered in Section 8.

2. Data Analysis Preliminaries

In this section we summarize the metrics that will be subsequently used to measure the performance of our proposed evidence definitions in Section 7. We first define the confusion matrix, also known as the error matrix, which is a numerical table to visualize the performance of a method for choosing between two alternatives. [24, page 23]

Definition 2.1 (Confusion Matrix). In classical hypothesis testing [25] for the two alternatives H_0 vs H_1, the confusion matrix is the 2 by 2 matrix of **Table 2.1**.

Table 2.1. Confusion Matrix

Decision	H_0 True	H_0 False
Do not Reject H_0	True Negatives	False Negatives
Reject H_0	False Positives	True Positives

The entries of **Table 2.1** are described as follows in terms of a test for cancer to clarify the nonintuitive standard terminology. A positive test is a test result stating that the tested patient has cancer. A negative test states that the subject does not have cancer.

True Positives (TP) - the number of cases correctly predicted to be positive. For example, the number of valid predictions that a person with a cancer has cancer.

True Negatives (TN) – the number of cases correctly predicted to be negative. For example, the valid predictions that a person without cancer does not have cancer.

False Positives (FP) - the number of cases incorrectly predicted to be positive. For example, the invalid predictions that a person without cancer (H_0 true) has cancer. (Type I error rate)

False Negatives (FN) - the number of cases incorrectly predicted to be negative. For example, the invalid predictions that a person with cancer (H_0 false) does not have cancer. (Type II error rate)

The general confusion matrix involves n multiple alternatives for $n \geq 2$, but we only consider the case $n = 2$. Although such a small confusion matrix is an informative table, as its name suggests, it is not always easy to apply. Here we restrict our usage of confusion matrix to calculate the accuracy of our model.

Definition 2.2 (Accuracy). Under the assumption of **Definition 1**

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Population}}, \tag{2.1}$$

where total population = $\text{TP} + \text{TN} + \text{FP} + \text{FN}$.

Accuracy is a useful measure when the confusion matrix for a dataset is nearly symmetric in such a way that the number of false negatives and the number of false positives are roughly the same. If the numbers

of false positives and false negatives differ significantly, other measure such as precision may need to be used. [26]

3. Likelihood Ratio as a Measure of Evidence

Probability measures uncertainty, and frequentist probabilities provide a measure of evidence based on the past frequencies yielding them. On the other hand, likelihood ratios can be said to measure comparative evidence. A pmf represents the uncertainty about the value of a random variable. The likelihood function gives the joint probability or pmf for an arbitrary random sample. Since the likelihood incorporates all available information about the underlying random variable X before any data is observed, it must include all available evidence when evaluated at observed data. We review below the use of the likelihood function in comparing the evidence associated with testing simple hypotheses involving different parameter values θ_1 and θ_2. See [4, 10, 27, 28] for further details on the use of the likelihood function.

Definition 3.1 (Likelihood Function). Let $\mathbf{x} = (x_1, \dots, x_n)$ be sample data from a random sample $\mathbf{X} = (X_1, \dots, X_n)$ from a random variable X with sample space S and real-valued parameter θ, and let $f(\mathbf{x}|\theta)$ denote the joint pdf of the random sample **X**. For any sample data **x**, the likelihood function of θ is defined as

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta), \tag{3.1}$$

where $L(\theta|\mathbf{x})$ in (3.1) is a function of the variable θ for given data **x**.

Note that the joint pdf $f(\mathbf{x}|\theta)$ as a function $L(\mathbf{x}|\theta)$ of **x** for fixed θ may also be called the likelihood function as well. The two functions $L(\theta|\mathbf{x})$ and $L(\mathbf{x}|\theta)$ may be called dual likelihood functions. In this paper, we restrict the likelihood function to **Definition 3.1** and use it to compare two alternatives for the parameter θ of the discrete random variable X.

Definition 3.2 (Likelihood Ratio). Under the assumptions of **Definition 3.1**, consider a test of the hypotheses $H_0: \theta = \theta_1$ vs $H_1: \theta = \theta_2$. Then the likelihood ratio is defined as

$$\Lambda(\mathbf{x}) = \frac{L(\theta_1|\mathbf{x})}{L(\theta_2|\mathbf{x})} = \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_2)}. \tag{3.2}$$

The likelihood ratio metric of (3.2) compares the probability of making the observations **x** for the possible parameter value θ_1 with that for θ_2. The parameter value yielding the largest $L(\theta|\mathbf{x})$ is considered the more likely value much as in the notion of a maximum likelihood estimation [27, 28].

Example 3.3. Let $\mathbf{X} = (X_1, \dots, X_n)$ denote a random sample with the exponential pdf as

$$f(x|\theta) = \theta e^{-\theta x}, \quad 0 < x < \infty. \tag{3.3}$$

To test the hypothesis $H_0: \theta = \theta_1 = \frac{1}{2}$ vs $H_1: \theta = \theta_2 = 1$, we calculate

$$\frac{L(\theta_1|\mathbf{x})}{L(\theta_2|\mathbf{x})} = \left(\frac{1}{2}\right)^n \exp\left\{\left(\frac{1}{2}\right)\sum_{i=1}^{n} x_i\right\}. \tag{3.4}$$

For a data set as {1,3,5,3,2,2,0,1,3,0}, then $n = 10$ and $\sum_{i=1}^{n} x_i = 20$, and the likelihood ratio is 21.51. For a decision criterion that we reject H_0 only if the likelihood ratio $\Lambda(\mathbf{x}) < 0.5$ in (3.2), then we fail to reject the null hypothesis. ■

It should be noted that the likelihood ratio test for two simple hypotheses forms a central part of Neyman-Pearson statistical theory. However, Neyman-Pearson theory is aimed at finding good rules for choosing from a specified set of possible alternatives. It does not address the problem of interpreting statistical evidence. [4, p. 58].

4. The Confidence Distribution as a Measure of Evidence

The concept of confidence was introduced by Neyman in his papers on confidence intervals [29], but the notion of a confidence distribution of a parameter θ originated with Cox [30]. The word "confidence" is conventionally used to indicate that the concept does not involve a probability on θ. In particular, a 95% confidence interval for an unknown parameter θ means the true value of the parameter is contained in the confidence interval for 95% of all possible data sets $\mathbf{x}$. In other words, the notion of confidence has a frequentist interpretation with respect to the distribution of the random sample $\mathbf{X}$ and hence the underlying random variable X.

The confidence distribution, on the other hand, is a bridge between the Bayesian [31] and frequentist approaches in statistics. In particular, the confidence distribution provides a frequentist analog to a Bayesian posterior for θ. The confidence distribution function on the parameter space comprises all possible confidence levels for a confidence interval of the parameter. Thus Type I error is implicit in the definition. Moreover, a confidence distribution is a function of both the parameter and the random sample. A duality similar to that for the likelihood function plays an essential role in the following definition.

Definition 4.1 (Confidence Distribution). Let $\mathbf{x}$ be sample data for the random sample $\mathbf{X}$ from the discrete random variable X with cdf F(x) and a one-dimensional parameter θ. The function $C_F(\theta, \mathbf{x})$ onto the interval [0,1] is called confidence distribution (CD) for the parameter θ if

1. For each $\mathbf{x} \in S^n$, $C_F(\theta_0, \mathbf{x})$ is a cumulative distribution function for θ.
2. At the true parameter value $\theta = \theta_0$, $C_F(\theta_0, \mathbf{x})$ as a function of the random sample $\mathbf{X}$ follows the uniform distribution $U(0,1)$.

$C_F(\theta_0, \mathbf{x})$ is a cdf on X that gives the probability that $\theta \leq \theta_0$. It has an interpretation similar to that for a confidence interval on θ. The frequentist probability $C_F(\theta_0, \mathbf{x})$ is the percentage of all possible data sets $\mathbf{x}$ for which θ is in an appropriate confidence interval involving θ_0 and $\mathbf{x}$. Condition (2) of **Definition 4.1** is analogous to the assumption that the null hypothesis is assumed true in the standard definition of P-value. It should be pointed out that the confidence distribution has alternative definitions [32, 33] including one involving the pivot statistics [34] of standard confidence intervals. Given a confidence distribution, we can define an associated confidence density that intuitively is a frequentist version for X of a Bayesian posterior density for the parameter θ. It supports the notion that confidence distributions bridge frequentist and Bayesian statistics.

Definition 4.2 (Confidence Density). Under the condition of **Definition 4.1**, the derivative of $C(\theta, \mathbf{x})$ with respect to θ will be called the confidence density of θ. Hence,

$$c_F(\theta, \mathbf{x}) = \frac{\partial C_F(\theta, \mathbf{x})}{\partial \theta}. \tag{4.1}$$

We focus on the one-sided hypothesis test $H_0: \theta \leq \theta_0$ vs $H_1: \theta > \theta_0$ for the remainder of this paper. Most elementary expositions, e.g., [28] would consider the null hypothesis to be $H_0: \theta = \theta_0$. In applications, an equality null is usually not appropriate. Indeed, rejecting such a null provides only when the data provides sufficient evidence that $H_1: \theta > \theta_0$. Failing to reject thus is failing to reject that $H_0: \theta \leq \theta_0$. The equality null simply allows an exact determination of Type I error for the data rather than a lower bound on it.

Definition 4.3 (Confidence Ratio). For the one-sided hypothesis test $H_0: \theta \leq \theta_0$ vs $H_1: \theta > \theta_0$, we define the confidence ratio as

$$CR(\theta_0, \mathbf{x}) = \frac{C_F(\theta_0, \mathbf{x})}{1 - C_F(\theta_0, \mathbf{x})}, \tag{4.2}$$

where $C_F(\theta_0, \mathbf{x})$ is the frequentist probability.

There is no Type I and Type II error calculations associated with confidence distributions since the error probabilities associated with elementary confidence intervals become the probability distribution on θ. The decision criterion for a hypothesis test can be based on $C_F(\theta_0, \mathbf{x})$, which is the probability $\theta \leq \theta_0$, or on the

confidence ratio (4.2). A simulation as in Section 7 can provide estimate of the errors involved. The confidence ratio can also be generalized for more general hypotheses. For example, we can compare the evidence for θ being in a region Ω_1 vs in a disjoint region Ω_2 by considering the confidence ratio

$$\mathrm{CR}(\Omega_1, \Omega_2, \mathbf{x}) = \frac{\int_{\Omega_1} \mathrm{c_F}(\theta, \mathbf{x})\mathrm{d}\theta}{\int_{\Omega_2} \mathrm{c_F}(\theta, \mathbf{x})\mathrm{d}\theta}. \tag{4.3}$$

Example 4.4 (CD and CR for Normal Distribution Mean). Let $\mathbf{x} = (\mathrm{x_1}, \ldots, \mathrm{x_n})$ be sample data from a random sample $\mathbf{X} = (\mathrm{X_1}, \ldots, \mathrm{X_n})$ from a normal random variable X with parameters μ and σ^2. The confidence distribution for the mean μ of X is shown in [33] to be

$$\mathrm{C}_\varphi(\mu, \mathbf{x}) = \mathrm{F}_{\mathrm{T_{n-1}}}\left(\frac{\mu - \bar{\mathbf{x}}}{\mathrm{s_x}/\sqrt{\mathrm{n}}}\right) \tag{4.4}$$

where $\bar{\mathbf{x}}$ is the sample mean, $\mathrm{s_x}$ is the sample standard deviation, and $\mathrm{F_{T_{n-1}}}$ is the cdf of Student's t distribution with $\mathrm{n}-1$ degrees of freedom. For $\mathrm{n} > 30$, however, it follows by the CLT that a good approximation of (4.4) is

$$\mathrm{C}_\varphi(\mu, \mathbf{x}) = \varphi\left(\frac{\mu - \bar{\mathbf{x}}}{\mathrm{s_x}/\sqrt{\mathrm{n}}}\right). \tag{4.5}$$

For the one-sided hypothesis testing on μ, as $\mathrm{H_0}: \mu \leq \mu_0$ vs $\mathrm{H_1}: \mu > \mu_0$, (4.5) gives

$$\mathrm{C}_\varphi(\mu_0, \mathbf{x}) = \varphi\left(\frac{\mu_0 - \bar{\mathbf{x}}}{\mathrm{s_x}/\sqrt{\mathrm{n}}}\right). \tag{4.6}$$

Then from (4.2),

$$\mathrm{CR}(\mu_0, \mathbf{x}) = \frac{\mathrm{C}_\varphi(\mu_0, \mathbf{x})}{1 - \mathrm{C}_\varphi(\mu_0, \mathbf{x})}. \tag{4.7}$$

Hence

$$\mathrm{CR}(\mu_0, \mathbf{x}) = \frac{\varphi\left(\frac{\mu_0 - \bar{\mathbf{x}}}{\mathrm{s_x}/\sqrt{\mathrm{n}}}\right)}{1 - \varphi\left(\frac{\mu_0 - \bar{\mathbf{x}}}{\mathrm{s_x}/\sqrt{\mathrm{n}}}\right)}. \tag{4.8}$$

Example 4.5 (CD and CR for Normal Distribution Variance). Under the assumptions of **Example 4.4**, from [33] the confidence distribution for the variance of normal distribution is

$$\mathrm{C}_{\chi^2}(\sigma^2, \mathbf{x}) = 1 - \mathrm{F}_{\chi^2_{\mathrm{n-1}}}\left(\frac{(\mathrm{n}-1)\mathrm{S_x^2}}{\sigma^2}\right). \tag{4.9}$$

where $\mathrm{F}_{\chi^2_{\mathrm{n-1}}}$ is the cdf of $\chi^2_{\mathrm{n-1}}$ distribution. Thus $\mathrm{C}_{\chi^2}({\sigma_0}^2, \mathbf{x})$ is a cdf on X that $\sigma^2 \leq {\sigma_0}^2$ and

$$\mathrm{C}_{\chi^2}({\sigma_0}^2, \mathbf{x}) = 1 - \mathrm{F}_{\chi^2_{\mathrm{n-1}}}\left(\frac{(\mathrm{n}-1)\mathrm{S_x^2}}{{\sigma_0}^2}\right). \tag{4.10}$$

Hence, for one-sided hypothesis test $\mathrm{H_0}: \sigma^2 \leq {\sigma_0}^2$ vs $\mathrm{H_1}: \sigma^2 > {\sigma_0}^2$, from (4.2) and (4.10)

$$\mathrm{CR}({\sigma_0}^2, \mathbf{x}) = \frac{1 - \mathrm{F}_{\chi^2_{\mathrm{n-1}}}\left(\frac{(\mathrm{n}-1)\mathrm{S_x^2}}{{\sigma_0}^2}\right)}{\mathrm{F}_{\chi^2_{\mathrm{n-1}}}\left(\frac{(\mathrm{n}-1)\mathrm{S_x^2}}{{\sigma_0}^2}\right)}. \tag{4.11}$$

Numerical examples for Section 4, 5, and 6 will be presented in Section 7.

5. Bayesian Posterior Odds as a Measure of Evidence

In this section, we apply the Bayesian approach to hypotheses testing problems, by calculating the posterior ratio over two alternative hypotheses. In general, Bayesian inference requires a different interpretation of probability since it uses a probability to describe the degree of belief about an unknown parameter and treats parameters as random variables. For the parameter θ, the distribution f(θ) that summarizes the information about θ prior to get sample information is the prior distribution.

Consider a random data sample $\mathbf{x} = (x_1, \dots, x_n)$ from a random variable X with joint pdf $f(\mathbf{x}|\theta)$. Then the posterior pdf is

$$f(\theta|\mathbf{x}) = \frac{f(\theta)f(\mathbf{x}|\theta)}{f(\mathbf{x})}, \tag{5.1}$$

where

$$f(\mathbf{x}) = \int_{-\infty}^{\infty} f(\theta)f(\mathbf{x}|\theta)d\theta. \tag{5.2}$$

In other words, we update the prior distribution f(θ) with the sample data to get the posterior distribution $f(\theta|\mathbf{x})$. The corresponding posterior cdf for θ is denoted $F(\theta|\mathbf{x})$. To use the posterior distribution in the hypothesis testing, we can define the posterior odds ratio as follows.

Definition 5.1 (Posterior Odds Ratio). Consider a random data sample $\mathbf{x} = (x_1, \dots, x_n)$ from a random variable X with posterior pdf $f(\theta|\mathbf{x})$ and cdf $F(\theta|\mathbf{x})$. For the one-sided hypothesis test $H_0: \theta \leq \theta_0$ vs $H_1: \theta > \theta_0$, we define the posterior odds ratio as

$$\Omega(\theta_0, \mathbf{x}) = \frac{F(\theta_0|\mathbf{x})}{1 - F(\theta_0|\mathbf{x})}. \tag{5.3}$$

Equivalently, we can simply use $F(\theta_0|\mathbf{x})$ as the evidence that the null is true. The following example illustrates the use of (5.3) in the hypothesis testing problems.

Example 5.2. Let $\mathbf{x} = (x_1, \dots, x_n)$ be a random data sample from the normal density with mean μ and variance 1, where μ is unknown. Assume the prior pdf for μ is normal with mean 0 and variance 1; that is,

$$f(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\mu^2}, \qquad -\infty < \mu < \infty. \tag{5.4}$$

The joint pdf of the sample for fixed μ is

$$f(\mathbf{x}|\mu) = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum(x_i-\mu)^2}. \tag{5.5}$$

From (5.1), (5.2), (5.4) and (5.5) it is shown in [18] that

$$f(\mu|\mathbf{x}) = \frac{(n+1)^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{1}{2}(n+1)\left[\mu - \frac{n\bar{x}}{n+1}\right]^2\right\}. \tag{5.6}$$

Therefore, the posterior of μ is distributed as $N\left(\frac{n\bar{x}}{n+1}, \frac{1}{n+1}\right)$.

To test $H_0: \mu \leq 0$ vs $H_1: \mu > 0$, suppose that we take a sample $\mathbf{x} = (x_1, \dots, x_{31})$ with $n = 31$ and $\bar{\mathbf{x}} = 0.5$ so the posterior of μ becomes N(0.484,0.031). Then the posterior odds ratio (5.3) becomes

$$P(\mu \leq 0|\mathbf{X} = \mathbf{x}) = \emptyset\left(\frac{0 - 0.484}{\sqrt{0.031}}\right) = \emptyset(-2.748) = 0.003. \tag{5.7}$$

Hence, the odds ratio is

$$\Omega(0, \mathbf{x}) = \frac{\emptyset(-2.748)}{1 - \emptyset(-2.748)} = 0.003. \tag{5.8}$$

In other words, the odds are 333.33 to 1 that $H_1: \mu > 0$ is true as opposed to $H_0: \mu \leq 0$. It appears beyond reasonable doubt that H_0 should be rejected. ■

In general, a result from the Bayesian approach does not necessarily agrees with one from frequentist approach. One reason is that there are no significance levels in the Bayesian approach though errors can certainly occur. In one situation, however, the results are remarkably similar. A noninformative prior is one for which $f(\theta)$ is constant over the parameter space, perhaps in limit. For example, $f(\theta) = \frac{1}{a}$ for $0 < \theta < a$ is a noninformative prior over a parameter space $(0, \infty)$ as $a \to \infty$. A noninformative prior assumes that there is no information about the parameter before collecting data. When a noninformative prior is used, the posterior probability $F(\theta_0|\mathbf{x})$ is the P-value for the one-sided hypothesis test $H_0: \theta \leq \theta_0$ vs $H_1: \theta > \theta_0$. [31]

For **Example 5.2** with a noninformative prior, we obtain a $N\left(0.5, \frac{1}{31}\right)$ posterior for μ. In this case

$$F(0|\mathbf{x}) = \emptyset\left(\frac{0 - 0.5}{1/\sqrt{31}}\right) = 0.0027, \tag{5.9}$$

which is also the frequentist P-value. The posterior odds ratio is now $\Omega(0, \mathbf{x}) = 0.0027$.

6. A New Definition of P-value as a Measure of Evidence

The notion of P-value is a fundamental tool in statistical inference and has been widely used for reporting outcomes of hypothesis tests. For example, see Chavalarias et al. [35]. Yet in practice, P-value is often misinterpreted, misused, or miscommunicated. Moreover, there is no general definition that unequivocally reflects the available evidence for the null hypothesis since H_0 is assumed to hold in existing definitions. In this section we propose a new definition of P-value that gives different values in some cases from the existing definitions. It provides a simple intuitive interpretation of P-value. Our approach appears applicable to a wide range of hypothesis testing problems. However, we restrict ourselves here to the standard one-sided tests. Our definition yields an interpretation of P-value as both a cardinal and ordinal measure of the evidence.

There are two standard ways of defining P-value for the general hypothesis test $H_0: \theta \in \Theta_0$ vs $H_1: \theta \notin \Theta_0$ with a parameter space Θ_0 and test statistic $T(\mathbf{X})$ for a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from a random variable X with parameter θ. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be observed.

Definition 6.1 (Standard Definitions of P-value). Under the assumption $H_0: \theta \in \Theta_0$ is true:

(1) [e.g., 36]

$$P - \text{value1}(\mathbf{x}) = \sup_{\theta \in \Theta_0} P\{T(\mathbf{X}) \geq T(\mathbf{x}) | \theta\}. \tag{6.1}$$

(2) [e.g., 37]

$$P - \text{value2}(\mathbf{x}) = \inf \{\alpha : T(\mathbf{x}) \in R_\alpha\}, \tag{6.2}$$

where R_α is the rejection region for a level of significance α.

Equation (6.1) is usually interpreted as: under the assumption that H_0 is true, P-value is the probability that $T(\mathbf{X})$ is at least as extreme as its observed value $T(\mathbf{x})$. This interpretation can lead to the common misunderstanding that this definition of P-value is the probability that H_0 is true. On the other hand, under the assumption that H_0 is true, equation (6.2) is based on significant levels (Type I error probabilities) and can lead to the misunderstanding that P-value is only a measure of Type I error and not related to the likelihood that H_0 is true. Both definitions elicit the question: how can the assumption that H_0 is true produce evidence that H_0 is true? The answer is that P-value is actually the probability that H_0 is true if H_0 is assumed true. For a P-value of 0.05 there is thus a 0.95 probability, intuitively speaking, that H_0 is false

and yields a reductio ad absurdum. In other words, there is a 0.95 probability that the assumption is false and that H_0 should be rejected. For other issues with these definitions, see [38], for example.

To address such issues, we propose a new definition below that involves the well-known maximum likelihood estimator (MLE) [27, 28] $\hat{\theta}(\mathbf{X})$ of θ obtained as $\arg\max_{\theta} L(\theta|\mathbf{x})$ in terms of $\mathbf{x}$, with the likelihood function $L(\theta|\mathbf{x})$ as in (3.1). The MLE $\hat{\theta}(\mathbf{x})$ maximizes the probability and minimizes the surprise of obtaining the data sample $\mathbf{x}$.

Definition 6.2 (P-value). Let $\mathbf{x} = (x_1, \ldots, x_n)$ be observed random sample data for a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from a continuous random variable X with a real-valued parameter θ. In addition let $f(\mathbf{x}|\theta)$ be the joint pdf of $\mathbf{X}$, and $\hat{\theta}(\mathbf{X})$ denote the MLE for θ, and let $Y = \hat{\theta}(\mathbf{X})$ with pdf $f_Y(y|\theta)$. Then for the hypothesis test $H_0: \theta \in \Theta_0$ vs $H_1: \theta \notin \Theta_0$, the novel P-value (NPV) for the null hypothesis $H_0: \theta \in \Theta_0$ at $\mathbf{x}$ is defined as

$$\mathrm{NPV}(\mathbf{x}|\Theta_0) = \left[\int_{\Theta_0} f_Y(y|\theta)dy\right]|_{\theta = \hat{\theta}(\mathbf{x})}, \tag{6.3}$$

where the integration is over the possible values y of $Y = \hat{\theta}(\mathbf{X})$.

$\mathrm{NPV}(\mathbf{x}|\Theta_0)$ in (6.3) is not the frequentist probability that $H_0: \theta \in \Theta_0$ is true, but it is an approximation. The MLE is used as a substitute for θ, and the integration in (6.3) simply gives frequentist probability that $Y = \hat{\theta}(\mathbf{X}) \in \Theta_0$. Since the distributions of MLEs are typically well known, an analytical or numerical integration in (6.3) is feasible. However, the result of the integration involves θ itself as seen in **Result 6.3**, so $\hat{\theta}(\mathbf{x})$ is then used as a numerical approximation for θ after the integration. The reasoning is that the properties of the MLE mentioned above should make $\hat{\theta}(\mathbf{x})$ a useful surrogate for θ.

For the remainder of this paper we specialize (6.3) and consider only the hypothesis test $H_0: \theta \leq \theta_0$ vs $H_1: \theta > \theta_0$ for the parameters μ and σ^2 of X. Under the assumption that X is a normal random variable, we rewrite $\mathrm{NPV}(\mathbf{x}|\Theta_0)$ as $\mathrm{NPV}(\mathbf{x}|\theta_0)$, and equation (6.3) gives the usual P-value when $\theta = \mu$ but not when $\theta = \sigma^2$.

Result 6.3. Let $X_1, \ldots, X_n$ be a random sample from a random variable $X \sim N(\mu, \sigma^2)$ with unknown μ, and consider the one-sided hypothesis test $H_0: \mu \leq \mu_0$ vs $H_1: \mu > \mu_0$. Then

$$\mathrm{NPV}(\mathbf{x}|\mu_0) = \Phi\left(\frac{\mu_0 - \bar{\mathbf{x}}}{\sigma/\sqrt{n}}\right) \tag{6.4}$$

when σ^2 is known, and

$$\mathrm{NPV}(\mathbf{x}|\mu_0) = F_{t(n-1)}\left(\frac{\mu_0 - \bar{\mathbf{x}}}{s/\sqrt{n}}\right) \tag{6.5}$$

when σ^2 is unknown, where $F_{t(n-1)}$ is the cdf for Student's t-distribution.

Proof. To prove (6.4) recall that the MLE for μ is the sample mean $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. The integral of (6.3) hence becomes

$$P[\bar{X} \leq \mu_0] = \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right). \tag{6.6}$$

Substituting $\bar{\mathbf{x}}$ for μ in (6.6) gives (6.4). Equation (6.5) then follows from the usual distribution theory [28] of hypothesis testing ■

The right sides of (6.4) and (6.5) are the standard P-values for the hypothesis test $H_0: \mu \leq \mu_0$ vs $H_1: \mu > \mu_0$ from (6.1). We next consider hypothesis tests on variances. Recall [e.g., 28] that the MLE for the variance of a normal distribution is $\hat{\sigma}^2(\mathbf{X}) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$, and so $\frac{n\hat{\sigma}^2(\mathbf{X})}{\sigma^2} \sim \chi^2(n-1)$.

Result 6.4. Let $X_1, \ldots, X_n$ be a random sample from a random variable $X \sim N(\mu, \sigma^2)$ with unknown σ^2, and consider the one-sided hypothesis test $H_0: \sigma^2 \leq \sigma_0^2$ vs $H_1: \sigma^2 > \sigma_0^2$. Then

$$\mathrm{NPV}(\mathbf{x}|\sigma_0^2) = F_{\chi^2(n-1)}\left(\frac{n\sigma_0^2}{\hat{\sigma}^2(\mathbf{x})}\right). \tag{6.7}$$

Proof. Since the MLE $\hat{\sigma}^2(\mathbf{X})$ for σ^2 satisfies $\frac{n\hat{\sigma}^2(\mathbf{X})}{\sigma^2} \sim \chi^2(n-1)$, then the integral of (6.3) becomes

$$P[\hat{\sigma}^2(\mathbf{X}) \le \sigma_0^2] = P\left[\frac{n\hat{\sigma}^2(\mathbf{X})}{\sigma^2} \le \frac{n\sigma_0^2}{\sigma^2}\right] = F_{\chi^2(n-1)}\left(\frac{n\sigma_0^2}{\sigma^2}\right). \tag{6.8}$$

Substituting $\hat{\sigma}^2(\mathbf{x})$ for σ^2 in (6.8) gives (6.7). ■

The right side of (6.7) is not the standard P-value for the hypothesis test $H_0: \mu \le \mu_0$ vs $H_1: \mu > \mu_0$ from (6.1). The standard P-value from (6.1) is $1 - F_{\chi^2(n-1)}\left(\frac{(n-1)s^2}{\sigma_0^2}\right)$, which is obtained when the probability $P[S^2 \ge s^2]$ is computed with $\sigma^2 = {\sigma_0}^2$ in (6.1). The standard P-value differs from (6.7) as well as from $F_{\chi^2(n-1)}\left(\frac{(n-1)\sigma_0^2}{s^2}\right)$, which is obtained using the unbiased S^2 instead of the biased MLE $\hat{\sigma}^2(\mathbf{X})$ in (6.3).

To compare the novel P-value to the previous evidence approaches considered, we note that confidence distributions usually give the standard P-value for the hypothesis test $H_0: \theta \le \theta_0$ vs $H_1: \theta > \theta_0$ by computing $C_F(\theta_0, \mathbf{x})$, though [33] gives an example for which this is not true. Confidence distributions, though, are implicitly defined by the notion of significance and hence Type I error, which seems contrary to a direct measure of evidence. As for Bayesian posteriors, it was noted in Section 5 that the posterior probability that $\theta \le \theta_0$ equals the P-value if the prior distribution for θ is noninformative. The difficulty is that even a noninformative prior contains belief. Moreover, obtaining the usual P-values offers nothing new except reasoning that does not involve significance levels. The NPV approach has particular appeal since as the sample size increases, an MLE $\hat{\theta}(\mathbf{x})$ converges in probability to θ. In other words, (6.4), (6.5), and (6.7) converge in probability to the frequentist probability that the respective null hypotheses are true and approach 1 as n approaches infinity.

In the next section, we present an example using the odds ratios considered here. That is, we compare to obtain insight into error rates.

7. Example

In this example, we use the confidence ratio to exemplify how simulation can provide Type I and Type II errors and the accuracy defined in (2.1). We first generate 1000 samples using Python. Let $i = 1, \dots, 1000$, and for each i, let $({X_1}^{(i)}, \dots, {X_{100}}^{(i)})$ be a random sample from the random variable X with the mean μ and variance σ^2, where μ and σ^2 is the same for all the 1000 samples. Then $\overline{X}^{(i)} \sim N(\mu, \frac{\sigma^2}{n})$ by the CLT. We then test the hypothesis as $H_0: \mu \le \mu_0$ vs $H_1: \mu > \mu_0$ and calculate the Type I and Type II errors. Finally, we use the confidence ratio in (4.8) and check if it is a good measure of evidence by calculating the accuracy. We could do this for all the measures of evidence presented here.

Note that a perfect test would have zero false positives and zero false negatives. However, in statistics, we deal with uncertainty and can never know whether statistical conclusions are correct. In this example, we use known parameters $\mu = 0$ and $\sigma^2 = 1$.

- **Testing over Population Mean μ**

By specifying μ_0, we can check if the null hypothesis is true or not. We consider two cases, each with two alternatives. In each case we used the $R = 1$ as a threshold CR - or equivalently for this example's hypothesis on μ, the novel odds ratio (NOR) for the new P-value in (6.4) - to fill the confusion matrix as it explained on **Definition 2.1**. Then based on the confusion matrix, we calculate the accuracy, Type I and Type II errors.

- **Case 1: H_0 is true and $\mu \leq \mu_0$.**

In the first case, we consider three different μ_0 as 0.01, 0.1 and 0.2. Since $\mu = 0$, we expect to fail to reject the null hypothesis. The general confusion matrix for this case is shown in **Table 7.1**. Since the assumption of this case is that H_0 is true, we do not have any case which H_0 is false.

Table 7.1. Case 1 General Confusion Matrix for the Mean of Normal Distribution

Decision	H_0 True	H_0 False
Do Not Reject H_0	# of cases where $CR(\mu_0) > 1$ out of 1000	0
Reject H_0	# of cases where $CR(\mu_0) \leq 1$ out of 1000	0

Next based on **Table 7.1** and the simulated data for the 1000 samples, we have the confusion matrix as **Table 7.2.**

Table 7.2. Case 1 Confusion Matrix for the Mean of Normal Distribution on 1000 samples

Decision	H_0 True			H_0 False
	$\mu_0 = 0.01$	$\mu_0 = 0.1$	$\mu_0 = 0.2$	
Do Not Reject H_0	508	825	982	0
Reject H_0	492	175	18	0

Data in **Table 7.2** has been classified by $R = 1$. However, $R = 1$ is not necessarily a good threshold. Jeffreys proposed the classification scheme shown in **Table 7.3** to summarize conclusions for the Bayes factor in terms of discrete categories of evidential strength. [39, p. 432]

Table 7.3. Evidence Categories for the Bayes Factor

Bayes Factor	Conclusion
> 100	Extreme evidence for H_1
30 – 100	Very strong evidence for H_1
10 – 30	Strong evidence for H_1
3 – 10	Moderate evidence for H_1
1 – 3	Anecdotal evidence for H_1
1	No evidence
1/3 – 1	Anecdotal evidence for H_0
1/3 – 1/10	Moderate evidence for H_0
1/10 – 1/30	Strong evidence for H_0
1/30 – 1/100	Very strong evidence for H_0
< 1/100	Extreme evidence for H_0

Here, we use the same classification and we calculate the accuracy and Type I and Type II errors in each of the classes. **Table 7.4** shows the accuracy plus Type I and Type II errors for the case where $\mu_0 = 0.1$ for each of the classes.

Table 7.4. Case 1 Performance Evaluation based on Different Ratio Thresholds

Decision Criteria	Accuracy	Type I Error (α)	Type II Error (β)
$R = \frac{1}{100}$	0.999	0.001	0.0
$R = \frac{1}{30}$	0.998	0.002	0.0
$\mathbf{R = \frac{1}{19}}$	**0.996**	**0.004**	**0.0**
$R = \frac{1}{10}$	0.99	0.01	0.0
$R = \frac{1}{3}$	0.953	0.047	0.0
$R = 1$	0.825	0.175	0.0
$R = 3$	0.596	0.404	0.0
$R = 10$	0.359	0.641	0.0
$R = 19$	0.256	0.744	0.0
$R = 30$	0.202	0.798	0.0
$R = 100$	0.1	0.9	0.0

In this case we check the evidence for H_0 since we already know that H_0 is true. Hence, we focus on the cases where $R \leq 1$. Based on the table, you can check how changing R effects the accuracy. In addition to the thresholds of ratio proposed by Jeffreys, we added another ratio as $R = \frac{1}{19}$. In this case the accuracy is the calculated for the ratio $\frac{0.05}{0.95} = \frac{1}{19}$.

- **Case 2: H_0 is False and $\mu > \mu_0$.**

In this case we consider three different μ_0 as -0.01, -0.1 and -0.2. Since $\mu = 0$, we expect to reject the null hypothesis. The general confusion matrix for this case has been shown on **Table 7.5**. Since in this case H_0 is false, we do not have any case for which H_0 is true.

Table 7.5. Case 2 General Confusion Matrix for the Mean of Normal Distribution

Decision	H_0 True	H_0 False
Do Not Reject H_0	0	# of cases where $CR(\mu_0) > 1$ out of 1000
Reject H_0	0	# of cases where $CR(\mu_0) \leq 1$ out of 1000

Based on **Table 7.5** and the simulated data for the 1000 samples, we have the confusion matrix as **Table 7.6** where $R = 1$. Note that the data and samples are fixed for both cases.

Table 7.6. Case 2 Confusion Matrix for the Mean of Normal Distribution on 1000 samples

Decision	H_0 True	H_0 False		
		$\mu_0 = -0.01$	$\mu_0 = -0.1$	$\mu_0 = -0.2$
Do Not Reject H_0	0	445	160	31
Reject H_0	0	555	840	969

We calculate the accuracy, Type I and Type II errors over the same classes used in the last case. **Table 7.7** below summarizes the results for $\mu_0 = -0.1$.

Table 7.7. Case 1 Performance Evaluation based on Different Ratio Thresholds

Decision Criteria	Accuracy	Type I Error (α)	Type II Error (β)
$R = \frac{1}{100}$	0.101	0.0	0.899
$R = \frac{1}{30}$	0.217	0.0	0.783
$\mathbf{R = \frac{1}{19}}$	**0.291**	**0.0**	**0.709**
$R = \frac{1}{10}$	0.401	0.0	0.599
$R = \frac{1}{3}$	0.643	0.0	0.357
$R = 1$	0.84	0.0	0.16
$R = 3$	0.952	0.0	0.048
$R = 10$	0.987	0.0	0.013
$\mathbf{R = 19}$	0.996	**0.0**	**0.004**
$R = 30$	0.999	0.0	0.001
$R = 100$	1	0.0	0.0

In this case we check the evidence for H_1 since H_0 is false. Hence we focus on the cases where $R \geq 1$. **Table 7.7** shows how changing R effects on the accuracy. In addition to the thresholds of ratio proposed by Jeffreys, again we added another ratio with $R = 19$.

- **Testing over Population Variance σ^2**

We use the same random variable and the same sample data to test the hypothesis testing $H_0: \sigma^2 \leq {\sigma_0}^2$ vs $H_1: \sigma^2 > {\sigma_0}^2$.

  - **Case 1: H_0 is true and $\sigma^2 \leq {\sigma_0}^2$.**

In the first case, we use three different values of ${\sigma_0}^2$: 1.01, 1.1 and 1.2. Since $\sigma^2 = 1$ is known, we expect to fail to reject the null hypothesis. **Table 7.8** below shows the general confusion matrix. Since the assumption is that H_0 is true, we do not have the occurrence of false H_0.

Table 7.8. Case 1 General Confusion Matrix for the Variance of Normal Distribution

Decision	H_0 True	H_0 False
Do Not Reject H_0	# of cases where $R > 1$ out of 1000	0
Reject H_0	# of cases where $R \leq 1$ out of 1000	0

From **Table 7.8** and the simulated data for the 1000 samples, for ${\sigma_0}^2 = 1.1$ and $R = 1$, we have the confusion matrix in **Table 7.9.**

Table 7.9. Case 1 Confusion Matrix for the Variance of Normal Distribution on 1000 samples

Decision	H_0 True			H_0 False
	${\sigma_0}^2 = 1.01$	${\sigma_0}^2 = 1.1$	${\sigma_0}^2 = 1.2$	
Do Not Reject H_0	535	777	917	0
Reject H_0	465	223	83	0

We calculate the accuracy, Type I and Type II errors for the classes used before. **Table 7.10** below summarizes the results for $\sigma_0{}^2 = 1.1$.

Table 7.10. Case 1 Performance Evaluation based on Different Ratio Thresholds

Decision Criteria	Accuracy	Type I Error (α)	Type II Error (β)
$R = \frac{1}{100}$	1.0	0.0	0.0
$R = \frac{1}{30}$	1.0	0.0	0.0
$\mathbf{R = \frac{1}{19}}$	**0.998**	0.002	**0.0**
$R = \frac{1}{10}$	0.99	0.01	0.0
$R = \frac{1}{3}$	0.929	0.071	0.0
$R = 1$	0.777	0.223	0.0
$R = 3$	0.495	0.505	0.0
$R = 10$	0.252	0.748	0.0
$\mathbf{R = 19}$	**0.157**	0.843	**0.0**
$R = 30$	0.113	0.887	0.0
$R = 100$	0.047	0.953	0.0

In this case, we check the evidence for H_0 since we know that H_0 is true. Hence, we focus on the cases where $R \leq 1$. The table shows how changing R effects on the accuracy. As before, we added ratio $R = \frac{1}{19}$.

- **Case 2: H_0 is False and $\sigma^2 > \sigma_0{}^2$.**

In this case we consider three different $\sigma_0{}^2$: 0.99, 0.90 and 0.8. Since $\sigma^2 = 1$, we expect to reject the null hypothesis. **Table 7.11** below shows the general confusion matrix for this case. Since the assumption is that H_0 is false, we do not have any case which H_0 is true.

Table 7.11. Case 2 General Confusion Matrix for the Variance of Normal Distribution

Decision	H_0 True	H_0 False
Do Not Reject H_0	0	# of cases where $R > 1$ out of 1000
Reject H_0	0	# of cases where $R \leq 1$ out of 1000

Based on **Table 7.11** and the simulated data for the 1000 samples, for $\sigma_0{}^2 = 0.90$ and $R = 1$, we have the confusion matrix as **Table 7.12.**

Table 7.12. Case 2 Confusion Matrix for the Variance of Normal Distribution on 1000 samples

Decision	H_0 True	H_0 False		
		$\sigma_0{}^2 = 0.99$	$\sigma_0{}^2 = 0.9$	$\sigma_0{}^2 = 0.8$
Do Not Reject H_0	0	473	243	69
Reject H_0	0	527	757	931

We calculate the accuracy, Type I and Type II errors over the previous classes before. **Table 7.13** below summarizes the results for $\sigma_0{}^2 = 0.9$.

Table 7.13. Case 2 Performance Evaluation based on Different Ratio Thresholds

Decision Criteria	Accuracy	Type I Error (α)	Type II Error (β)
$R = \frac{1}{100}$	0.05	0.0	0.95
$R = \frac{1}{30}$	0.126	0.0	0.874
$\mathbf{R = \frac{1}{19}}$	**0.17**	**0.0**	**0.83**
$R = \frac{1}{10}$	0.264	0.0	0.736
$R = \frac{1}{3}$	0.528	0.0	0.472
$R = 1$	0.757	0.0	0.243
$R = 3$	0.906	0.0	0.094
$R = 10$	0.974	0.0	0.026
$\mathbf{R = 19}$	**0.985**	**0.0**	**0.015**
$R = 30$	0.992	0.0	0.008
$R = 100$	1.0	0.0	0.0

8. Conclusion

In this paper, we presented some measures for statistical evidence with regard to hypothesis testing. We focused on the one-sided hypothesis test $H_0: \theta \leq \theta_0$ vs $H_1: \theta > \theta_0$. First, we reviewed the well-known likelihood ratio as a measure of evidence associated with testing simple hypotheses involving different parameter values θ_1 and θ_2.

Second, we considered the confidence distribution as a measure of evidence. Confidence distribution is a frequentist tool derived from confidence intervals. We used the confidence ratio to compare the relative evidence for null and alternative hypothesis.

Third, we used the posterior odds ratio from Bayesian approach to calculate evidence. In this case, the ratio of posterior probability for the parameter being in region of null vs alternative hypothesis gives the evidence. We also noted that the posterior odds ratio for noninformative priors leads to the same ratio obtained by confidence distribution.

Fourth, as the principal contribution of the paper, we proposed a novel P-value involving the MLE for the parameter of interest. We then discussed the benefits of the defined novel P-value as compared to the classic P-value. These included no assumption of H_0 being true in its calculation, having no dependence on Type I error, and approximating the probability that H_0 is true. As n approaches infinity, the probability approaches 1 that the new P-value is the probability that $H_0$is true. Interestingly, the classic and new P-values were the same for testing the mean of a normal distribution but differed in testing the variance.

Fifth, we showed that the errors and accuracy of an evidence-based test can be obtained by simulation.

References

1. Shafer, G. (1976). A Mathematical Theory of Evidence. Princeton University Press: Princeton, USA.
2. Evans, M. (2015). Measuring Statistical Evidence Using Relative Belief. Chapman and Hall/CRC: NY, USA. https://doi.org/10.1201/b18587.

3. Hacking, I. (2016). Logic of Statistical Inference. Cambridge University Press: Cambridge, USA. https://doi.org/10.1017/CBO9781316534960.
4. Royall, R. M. (2017). Statistical Evidence: A Likelihood Paradigm. Chapman and Hall: NY, USA. https://doi.org/10.1201/9780203738665.
5. Salicone, S., Prioli, M. (2018). Measuring Uncertainty within the Theory of Evidence. Springer, Cham: NY, USA. https://doi.org/10.1007/978-3-319-74139-0.
6. Birnbaum, A. (1962). On the foundations of statistical inference. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1962.10480660.
7. Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character. https://doi.org/10.1098/rsta.1922.0009.
8. Mayo, D. G. (2014). On the Birnbaum argument for the strong likelihood principle. Statistical Science. https://doi.org/10.1214/13-STS457.
9. Lindsey, J. K. (2014). Likelihood Principle. Wiley StatsRef: Statistics Reference Online. https://doi.org/10.1002/9781118445112.stat05881
10. Edwards, A.W.F. (1990). Likelihood. New Palgrave. Palgrave Macmillan: London, UK. https://doi.org/10.1007/978-1-349-20865-4_16.
11. Berger, J. O., Wolpert, R. L. (1988). The Likelihood Principle. 2nd Edition. Institute of Mathematical Statistics: Hayward, CA, USA.
12. Evans, M. (2014). Discussion of "On the Birnbaum argument for the strong likelihood principle.". Statistical Science. https://doi.org/10.1214/14-STS471.
13. Dollinger, M.B., Kulinskaya, E., Staudte, R.G. (1996). When is a P-value a Good Measure of Evidence?. Rieder H. (eds) Robust Statistics, Data Analysis, and Computer Intensive Methods. Lecture Notes in Statistics, vol 109. Springer: NY, USA. https://doi.org/10.1007/978-1-4612-2380-1_8
14. Hubbard, R., Lindsay, R. M. (2008). Why p-values are not a useful measure of evidence in statistical significance testing. Theory & Psychology. https://doi.org/10.1177/0959354307086923.
15. Liu, S., Liu, R., Xie, M. (2020). P-value as the strength of evidence measured by confidence distribution. Submitted to Statistical Science. https://arxiv.org/pdf/2001.11945.pdf (Updated January 31, 2020).
16. Blume, J. D. (2002). Likelihood methods for measuring statistical evidence. Statistics in Medicine. https://doi.org/10.1002/sim.1216.
17. Kass, R. E., Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1995.10476572.
18. Hines, W. W., Montgomery, D. C. (2000). Probability and Statistics in Engineering and Management Science. John Wiley & Sons.: NY, USA.
19. Vieland, V. J., Seok, S. C. (2016). Statistical evidence measured on a properly calibrated scale for multinomial hypothesis comparisons. Entropy. https://doi.org/10.3390/e18040114.
20. Vieland, V. J., Das, J., Hodge, S. E., Seok, S. C. (2013). Measurement of statistical evidence on an absolute scale following thermodynamic principles. Theory in Biosciences. https://doi.org/10.1007/s12064-013-0180-9.
21. Evans, M. (2020) The measurement of statistical evidence as the basis for statistical reasoning. Proceedings of the 5th International Electronic Conference on Entropy and Its Applications. 18--30 Nov 2019. https://arxiv.org/abs/1906.09484v1.
22. Evans, M. (2016). Measuring statistical evidence using relative belief. Computational and Structural Biotechnology Journal. https://doi.org/10.1016/j.csbj.2015.12.001.

23. Shafer, G. (2011). A betting interpretation for probabilities and Dempster-Shafer degrees of belief. International Journal of Approximate Reasoning. https://doi.org/10.1016/j.ijar.2009.05.012.
24. Marsland, S. (2014). Machine Learning: An Algorithmic Perspective. 2nd Edition. Chapman and Hall/CRC: NY, USA. https://doi.org/10.1201/b17476
25. Lehmann, E. L., Romano, Joseph P. (2005). Testing Statistical Hypotheses, 3rd Edition. Springer: NY, USA.
26. https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124
27. Pawitan, Y. (2013). In All Likelihood. Statistical Modeling and Inference Using Likelihood. 1st ed. The Clarendon Press: Oxford, UK.
28. Casella, G.; Berger, R.L. (2002). Statistical Inference. 2nd ed. Cengage Learning: Delhi, India.
29. Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. Biometrika. https://doi.org/10.2307/2332207.
30. Cox, D. R. (1958). Some problems connected with statistical inference. The Annals of Mathematical Statistics. https://doi.org/10.1214/aoms/1177706618.
31. Bolstad, W. M., Curran, J. M. (2016). Introduction to Bayesian Statistics. 3rd ed. John Wiley & Sons.: NY, USA. https://doi.org/10.1002/9781118593165.
32. Cox, D.R. (2006). Principles of Statistical Inference. Cambridge University Press: London, UK. https://doi.org/10.1017/CBO9780511813559.
33. Xie, M. G., Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. International Statistical Review. https://doi.org/10.1111/insr.12000.
34. Schweder, T., Hjort, N. (2016). Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press: Cambridge, UK. https://doi.org/10.1017/CBO9781139046671.
35. Chavalarias, D., Wallach, J., Li, A., Ioannidis, J. (2016). Evolution of reporting p values in the biomedical literature. 1990-2015, JAMA. http://dx.doi.org/10.1001/jama.2016.1952.
36. Abell, M. L., Braselton, J. P., Rafter, J. A. (1999). Statistics with Mathematica, Academic Press.
37. Lehmann, E. L., Romano, J. P. (2005). Testing Statistical Hypotheses. Springer-Verlag: NY, USA.
38. Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. Seminars in Hematology 45, 135.
39. Jeffreys, H. (1998) [1961]. The Theory of Probability, 3rd Edition. Oxford University Press: Oxford, England.

Chapter 5

General Conclusions

This dissertation consists of three papers. It is a comprehensive study on the hierarchical framework in which data is transformed into information and this information is transformed into evidence. This evidence is then used to test hypotheses and check assertions. In other words,

$$\text{data} \Rightarrow \text{information} \Rightarrow \text{evidence}.$$

In the first paper, we studied the data and the information loss over the data compression. We focused on this lost information caused by multiple data sets having the same value of the statistic. This possibility is typical of data analysis. The data uniquely determines the value of the statistic, but a value of the statistic does not uniquely determine the data yielding it. In other words, we answered the question: how much Shannon information is lost about a data sample when only the value of a sufficient statistic is known but not the original data. We also defined the entropic loss associated with a sufficient statistic T under consideration as the expected lost information over all possible samples to give a metric dependent only on T.

Next we pursued the transformation from data to information. We proposed a new measure of information called gambler's information (or certitude) in contrast to Shannon information (or surprisal) for discrete random variables. We also explained that the choice of information type to use in a particular model depends on the retrospective or prospective nature of the model. Then we defined a new type of entropy called outer entropy by moving the log function outside the expectation in contrast to Shannon entropy to facilitate both intuitive results and mathematical manipulation.

We provided two applications of these new concepts. The first application determines the gambler's information lost when a random data sample is characterized by a single statistic such as the mean of the underlying random variable as was done in paper 1 for Shannon information. The second application uses outer entropy as a new metric for deciding between possible numerical parameter values for the underlying random variable.

In the third paper, we considered the next concept in the hierarchical framework above: evidence. Here we presented some measures for statistical evidence with regard to hypothesis testing. We began by reviewing the well-known likelihood ratio as a measure of evidence associated with testing simple hypotheses involving different parameter values θ_1 and θ_2. Then we considered the confidence distribution as a measure of evidence. Confidence distribution is a frequentist tool generalized from confidence intervals. We used a confidence ratio to compare the evidence that the null hypothesis is true to the evidence that it is false. We next used the posterior odds ratio from Bayesian approach to calculate evidence. In this case, the ratio of posterior probability for the parameter being in region of null vs alternative hypothesis gives the evidence. We also noted the posterior odds ratio for noninformative priors leads to the same ratio obtained by confidence distribution. Finally, we used P-value as a measure of evidence, and we showed in the case of one-sided hypothesis testing for normal distribution, the P-value is equivalent to confidence. We then defined a novel P-value (NPV) independent of significance levels, error, and test statistics, as well as the assumption that H_0 is true. This NPV functions essentially as a frequentist posterior probability and, in effect, bridges the gap between Bayesian and frequentist statistics. Examples were presented.

Maryam Moghimi received her B.S. degree in Industrial Mathematics from Sharif University of Technology, Tehran, Iran, and her M.S. degree in Statistics from The University of Texas at Arlington (UTA), Texas, USA. She obtained her Ph.D. in Industrial Engineering at The University of Texas in Arlington (UTA) in 2020. She has conducted research at the Center on Stochastic Modeling Optimization and Statistics (COSMOS) and currently works as a Senior DevOps Engineer at CCC Information Services, Austin, Texas. Her research interests include Information Theory, Statistics and Data Science. She enjoys working on real-world problems.