Entropy **2012**, *14*, 2254-2264; doi:10.3390/e14112254

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Statistical Information: A Bayesian Perspective

Rafael B. Stern [1,]* and Carlos A. de B. Pereira [2]

[1] Department of Statistics, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA

[2] Instituto de Matemática e Estatística, Universidade de São Paulo, Rua do Matão 1010, 05508-900, São Paulo, Brazil; E-Mail: cpereira@ime.usp.br (C.A.B.P.)

* Author to whom correspondence should be addressed; E-Mail: rbstern@gmail.com.

Received: 15 August 2012; in revised form: 28 September 2012 / Accepted: 1 November 2012 / Published: 7 November 2012

Abstract: We explore the meaning of information about quantities of interest. Our approach is divided in two scenarios: the analysis of observations and the planning of an experiment. First, we review the Sufficiency, Conditionality and Likelihood principles and how they relate to trivial experiments. Next, we review Blackwell Sufficiency and show that sampling without replacement is Blackwell Sufficient for sampling with replacement. Finally, we unify the two scenarios presenting an extension of the relationship between Blackwell Equivalence and the Likelihood Principle.

Keywords: statistical information; Blackwell sufficiency; likelihood principle

---

1. Introduction

One of the goals of statistics is to extract information about unknown quantities of interest from observations or from an experiment to be performed. The intuitive definition of information that we adopt from Basu [1] is:

"Information is what it does for you, it changes your opinion".

One might further question:

- Information about what?

We are interested in information about a quantity of interest, $\theta \in \Theta$. A quantity of interest represents a state of nature that we are uncertain of. For example, one might be interested in the number of rainy days next year. For instance, θ can be this number and Θ all natural numbers smaller or equal to 366.

- Where is the information?

Stating Θ already uses previous knowledge about θ. In the example in the last paragraph, we have informed that any year has at most 366 days and, therefore, θ must be smaller than this number. Besides stating Θ, one might also think that some values are more probable than others. This kind of knowledge is used to elicit the prior distribution for θ. The prior distribution represents a description of our present state of uncertainty about θ. Usually, the scientists' goal is to decrease his uncertainty about θ. Thus, he collects data he believes to be related to the quantity of interest. That is, he expects that there is information about θ in the data he collects.

- How is information extracted?

We focus on the case in which one uses Bayes' theorem, to compute the posterior distribution for θ given the observation. The posterior distribution describes the uncertainty about the quantity of interest after calibrating the prior by the observation. (In practice, the posterior distribution can rarely be computed. In these cases, it usually is sufficient to compute a quantity proportional to the posterior or to sample from the posterior.) Information also depends on the statistical framework.

- How much information is extracted?

In Section 3 we question: How much information is extracted from a given observation? Section 3.1 reviews common principles in Statistics and their relationship with the Likelihood principle. Section 3.2 presents a simple example and discusses information functions compatible with the Likelihood principle.

In Section 4 we consider questions related to experimental design: "How much information do we expect to obtain by performing an experiment?" or "What is the best choice among possible experiments?". Blackwell Sufficiency is a strong criterion for the comparison of experiments. The definition of Blackwell Sufficiency, with a new example, is presented in Section 4.1. If Θ is finite, then two experiments are equally informative in Blackwell's sense iff the distribution of their likelihoods is the same (Torgersen [2]). In Section 4.2 we extend this result to a setting with no restrictions on Θ. Finally, since not all experiments are comparable in Blackwell's sense, Section 4.3 explores the metrics discussed in Section 3.2 within the framework of decision theory to compare experiments.

In the following Section, we formalize the definitions here introduced.

2. Definitions

A probability space is a triple $(\Omega, \Im, P)$ in which Ω is a set, $\Im$ is a σ-algebra on Ω and $P : \Im \mapsto [0,1]$ is a probability function. A quantity R corresponds to a function from Ω to a set $\Re$. We define the probability space induced by R, $(\Re, \Im_R, P_R)$, where $\Im_R = \{M \subset \Re : R^{-1}[M] \in \Im\}$ and $P_R(M) = P\left(R^{-1}[M]\right)$. Finally, the σ-algebra induced on Ω by a quantity R is called $\Im_{|R}$ and corresponds to $\{R^{-1}[M] : M \in \Im_R\}$.

An experiment corresponds to a mechanism that allows observing a given quantity. The performance of an experiment corresponds to the observation of this quantity. In order to be concise, from this point forward, we use the word experiment for both the experiment itself and the quantity that is observed when the experiment is performed.

Many quantities of interest are not observable. Therefore, it is only possible to learn about them in an indirect manner. Here, we restrict ourselves to performing experiments that are related to the quantity of interest and applying Bayes' Theorem to update our knowledge about the latter. For example, in Section 1, the quantity of interest θ corresponds to the number of rainy days next year. A possible experiment to learn about θ would be to collect pluviometric data from recent years. Let X be the quantity representing this yet unobserved data. For brevity, we also call X the experiment. Our uncertainty about θ after performing X and observing $X = x$ is given by $P(\theta|X = x)$.

Let X be an experiment in $\mathcal{X}$. A function $T : \mathcal{X} \mapsto \tau$ is called a statistic of X. Therefore, $T(X)$ is also an experiment. Whenever there is no confusion, we use the letter T both to indicate the statistic T and the experiment $T(X)$.

From now on, we restrict ourselves to quantities in $\mathcal{R}$ with probability distributions that are discrete or are absolutely continuous with respect to the Lebesgue measure. $p_X(x|\theta)$ is the conditional probability (density) function of the experiment X given the quantity of interest θ. After the experiment is performed we write $L(\theta|X = x)$ for the likelihood function of X at point x. Whenever clear in the context, we write $p(x|\theta)$ and $L(\theta|x)$ for the former functions. The prior distribution of θ is denoted by $p(\theta)$ and the posterior by $p(\theta|X = x)$. Bayes's Theorem provides $p(\theta|X = x) \propto p(\theta)L(\theta|x)$.

Finally, we say that an experiment $X : \Omega \mapsto \mathcal{X}$ is trivial for a quantity of interest $\theta : \Omega \mapsto \Theta$ if $\Im_{|X}$ is independent of $\Im_{|\theta}$. This condition is equivalent to the assertion that, $\forall\theta' \in \Theta, \forall x \in \mathcal{X}$, $p(x|\theta') = p(x)$. We use the word trivial to emphasize that X and θ are not associated. Consequently, performing X alone does not bring "information" about θ.

3. Information after an Experiment is Performed

3.1. Statistical Principles and Information

Let $Inf(X, x, \theta)$ denote the information gained about the quantity of interest θ after observing outcome x in experiment X. We follow Basu [1] and Birnbaum [3] in restricting the possible forms for $Inf(X, x, \theta)$ by assuming common statistical principles.

A statistic $T : \mathcal{X} \mapsto \tau$ is sufficient if X and θ are conditionally independent given T, that is, X is a trivial experiment for θ given T. Thus, since X is a trivial experiment for θ given T, all the information about θ in X is gained by observing T alone. The Sufficiency Principle states that for any sufficient statistic T, for any x and y in $\mathcal{X}$, if $T(x) = T(x')$ then $Inf(X, x, \theta) = Inf(X, x', \theta)$. This principle is usually followed by all scientists, although not always explicitly mentioned: for inference about θ the scientist only needs to consider a sufficient statistic.

The Conditionality Principle is another important statistical principle: it can be seen as the reciprocal of Sufficiency. The latter states that a trivial experiment performed after T does not bring extra information about θ. The former states that a trivial experiment performed before another experiment does not bring extra information about θ. Let X_1 and X_2 be two arbitrary experiments. Let $Y \in \{1, 2\}$ be an experiment jointly independent from θ, X_1 and X_2. Let X_Y be the mixture of X_1 and X_2. X_Y is performed in the following way: Perform Y. If the result of Y is 1 then perform X_1, else perform

X_2. The Conditionality Principle states that $Inf((Y, X_Y), (i, x), \theta) = Inf(X_i, x, \theta), \forall i \in \{1, 2\}$. This principle is more controversial than that of Sufficiency.

The Likelihood Principle states that any two possible outcomes having proportional likelihood functions must provide the same information about the quantity of interest. Therefore, for any experiments X_1 and X_2 and any $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$, if $L(\theta|x_1) \propto L(\theta|x_2)$, then $Inf(X_1, x_1, \theta) = Inf(X_2, x_2, \theta)$. This principle is stronger than the Sufficiency Principle and the Conditionality Principle. Birnbaum [3], Basu [4] present the converse statement,

Theorem 1 *The Sufficiency and the Conditionality Principles hold iff the Likelihood Principle holds.*

A scientist who follows the Likelihood Principle can perform inference about the quantity of interest solely based on the likelihood function. Lindley and Philips [5] and Pereira and Lindley [6] provide examples in which some frequentist methods violate the Likelihood Principle. Indeed, Frequentist Statistics does not follow the Conditionality Principle. On the other hand, Wechsler *et al.* [7] shows that Bayesian Statistics follows the Likelihood Principle.

3.2. Information in the Observation

After performing an experiment, how much information about θ does one obtain? In the last section, we argued that the information obtained from points with proportional likelihoods should be the same. Nevertheless, this property only gives a vague idea about how the information function should be. In order to add precision to the definition of information we again rely on: "Information is what it does for you, it changes your opinion".

Before one performs an experiment, his opinion about θ is given by his prior distribution. On the other hand, his opinion after the experiment is performed is given by his posterior distribution. Hence, since the information should represent the change in opinion, it should be a function of prior and posterior distributions. If prior and posterior distributions are equal, there is no gain of information.

Next, we use an intuitive example to illustrate some information functions that satisfy this property. Consider that there are 4 balls, 2 of them are black and 2 are white. 3 of these 4 balls are put in an urn. You do not know which ball was left out. You are offered the possibility of performing one of the following three experiments; Experiment 1 consists of taking only one ball from the urn; Experiment 2 consists of taking two balls with replacement and; Experiment 3 consists of taking two balls without replacement. Your goal is to guess the number of white balls in the urn, 1 or 2. Assume that, a priori, you do not believe any combination of balls is more likely, a uniform prior. Also assume that all balls in the urn have equal probability of being selected. Let θ be the number of white balls in the urn and X_i be the number of white balls observed in the i-th experiment. The posterior probabilities $P(\theta = 1|X_i = j)$ are provided in Table 1.

Table 1. $P(\theta = 1|X_i = j)$.

i/j	0	1	2
1	0.33	0.67	-
2	0.20	0.50	0.80
3	0	0.50	1

Some information functions that can be applied to these experiments are:

1. The Euclidean distance: $Inf_E(X_i, x_i, \theta) = \sqrt{\sum_j [P(\theta = j) - P(\theta = j|X_i = x_i)]^2}$.
2. $Inf_V(X_i, x_i, \theta) = [E(\theta|X_i = x_i) - E(\theta)]^2$.
3. Kullback–Leibler divergence: $Inf_{KL}(X_i, x_i, \theta) = \sum_j P(\theta = j|X_i = x_i) \log\left(\frac{P(\theta=j|X_i=x_i)}{P(\theta=j)}\right)$.

Which of the experiments is the most informative? That is, which experiment do you expect to most change your opinion? Table 2 does not provide a straightforward answer. For example, Experiment 1, in a worst case scenario, brings more information than Experiments 2 and 3. Similarly, $P(X_2 = 1) < P(X_3 = 1)$ and, thus, obtaining no information in Experiment 2 is less likely than in Experiment 3. On the other hand, Experiment 3 provides the largest possible increments in information. In the next section, we discuss how to decide which experiment is the most informative.

Table 2. From left to right, tables for $Inf_E(X_i, j, \theta)$, $Inf_V(X_i, j, \theta)$ and $Inf_{KL}(X_i, j, \theta)$.

i/j	0	1	2
1	0.23	0.23	-
2	0.42	0	0.42
3	0.7	0	0.7

i/j	0	1	2
1	0.03	0.03	-
2	0.09	0	0.09
3	0.25	0	0.25

i/j	0	1	2
1	0.02	0.02	-
2	0.08	0	0.08
3	0.30	0	0.30

4. Information before an Experiment is Performed

4.1. Blackwell Sufficiency

Consider two experiments, X and Y, that depend on θ. One usually wants to choose between X and Y for inferences about θ based solely on the conditional distributions of X given θ and Y given θ. In this section we review the concept of Blackwell Sufficiency Blackwell [8] and show that it is a generalization of the Sufficiency Principle for comparison of experiments.

A statistic T is sufficient for an experiment X, if X and θ are conditionally independent given T. Consequently, T is sufficient iff $p(x|\theta) = p(t|\theta)p(x|t)$. The conditional distribution of X given θ can be generated by observing T and sampling from $p(x|t)$.

Let $X \in \mathcal{X}(X)$ and $Y \in \mathcal{X}(Y)$ be two statistical experiments. X is Blackwell Sufficient for Y if there exists a map $H : \mathcal{X}(X) \times \mathcal{X}(Y) \mapsto [0, 1]$, a transition function, satisfying the following properties:

- For any $y \in \mathcal{X}(Y)$, $H(\cdot, y)$ is measurable on the σ-algebra induced by X, $\Im_{|X}$.

- For any $x \in \mathcal{X}(X)$, $H(x, \cdot)$ is a probability (density) function defined on $(\mathcal{X}(Y), \Im_{|Y})$.
- For any $y \in \mathcal{X}(Y)$, $p(y|\theta) = E(H(X, y)|\theta)$, the conditional expectation of $H(X, y)$ given θ.

Let $\mathcal{X}(X)$ and $\mathcal{X}(Y)$ be countable sets and define for all $x \in \mathcal{X}(X)$, $Z_x \in \mathcal{X}(Y)$ as a trivial experiment such that $P(Z_x = y) = H(x, y)$. From the definition of Blackwell Sufficiency, the quantities (Z_X, θ) and (Y, θ) are equally distributed: X is Blackwell Sufficient for Y if and only if one can obtain an experiment with the same distribution as Y by observing $X = x$ and, after that, performing the "randomization", Z_x.

Next, we provide two examples of Blackwell Sufficiency that address the question in the end of Section 3.2. Example 1 is a version of that in Basu and Pereira [9]. Example 2 is new and shows that sampling without replacement is Blackwell sufficient for sampling with replacement. Other examples of Blackwell Sufficiency can be found, for example, in Goel and Ginebra [10] and Torgersen [2].

Example 1 *Let X and Y be two experiments, π a quantity of interest in $[0, 1]$ and q and p known constants in $[0, 1]$. Representing the Bernoulli distribution with parameter p by Ber(p), consider also that the conditional distributions of X and Y given π are, respectively:*

$$X \sim \text{Ber}(\pi) \text{ and } Y \sim \text{Ber}(q\pi + (1 - q)p)$$

X is Blackwell Sufficient for Y regarding π.

Proof. Let $A \sim \text{Ber}(q)$ and $B \sim \text{Ber}(p)$, both independent of all other variables, then defining $Y' = AX + (1 - A)B$, (Y', π) and (Y, π) are equally distributed. Therefore, X is Blackwell Sufficient for Y.

Example 2 *Next, we generalize the example of Section 3.2. Consider an urn with N balls. θ of these balls are black and $N - \theta$ are white. n $(\leq N)$ balls are drawn from the urn.*

By stating that $(X_1, \ldots, X_n)$ is a sample with replacement from the urn, we mean:

1. *Conditionally on θ, $X_1 \sim \text{Ber}\left(\frac{\theta}{N}\right)$;*
2. *Conditionally on θ, $X_1, \ldots, X_n$ are identically distributed;*
3. *X_{i+1} is conditionally independent of $(X_i, \ldots, X_1)$ given θ, $\forall i \in \{1, \ldots, n - 1\}$.*

Analogously, $(Y_1, \ldots, Y_n)$ corresponds to a sample without replacement, that is:

1. *Conditionally on θ, $Y_1 \sim \text{Ber}\left(\frac{\theta}{N}\right)$;*
2. *$Y_{i+1}|(y_i, \ldots, y_1, \theta) \sim \text{Ber}\left(\frac{\theta - \sum_{j=1}^{i} y_j}{N - i}\right)$,*
   $\forall i \in \{1, \ldots, n - 1\}$, $\forall (y_i, \ldots, y_1) \in \{0, 1\}^i$.

$(Y_1, \ldots, Y_n)$ is Blackwell Sufficient for $(X_1, \ldots, X_n)$ regarding θ.

Proof. Define $X_1^* = Y_1$, $T_i = \sum_{j=1}^{i} Y_j$ and $\forall i \in \{1, \ldots, n - 1\}$ two quantities A_{i+1} and B_{i+1}. These two quantities are such that:

1. $A_{i+1} \sim \text{Ber}\left(\frac{N-i}{N}\right)$, and is independent of all other variables;
2. $B_{i+1}|T_i = t_i \sim \text{Ber}\left(\frac{t_i}{i}\right)$;

3. $\forall i \in \{1, \ldots, n\}$, conditionally on $T_i = t_i$, B_i is jointly independent of $(A_1, \ldots, A_n), (B_1, \ldots, B_{i-1}), (Y_{i+1}, \ldots, Y_n)$ and θ.

Define:

$$X^*_{i+1} = A_{i+1}Y_{i+1} + (1 - A_{i+1})B_{i+1}$$

Conditionally on θ, $X^*_{i+1}|t_i \sim \text{Ber}(\theta/N)$, $\forall t_i \in \{0, \ldots, i\}$. Therefore, $X^*_{i+1} \sim \text{Ber}(\theta/N)$ and is conditionally independent of $(Y_i, \ldots, Y_1)$ given θ. Finally, since $(X^*_i, \ldots, X^*_1)$ is a function of $(Y_i, \ldots, Y_1)$, $(A_i, \ldots, A_2)$ and $(B_i, \ldots, B_2)$, conclude that X^*_{i+1} is independent of $(X^*_i, \ldots, X^*_1)$ given θ. By the previous conclusions, $(X^*_1, \ldots, X^*_n, \theta)$ is identically distributed to $(X_1, \ldots, X_n, \theta)$. Also, by construction, $(X^*_1, \ldots, X^*_n)|(Y_1 = y_1, \ldots, Y_n = y_n)$ is trivial, $\forall (y_1, \ldots, y_n) \in \{0, 1\}^n$. Hence, sampling without replacement is Blackwell Sufficient for sampling with replacement.

Hence, in Section 3.2, Experiment 3 is Blackwell Sufficient for Experiment 2. Similarly, Basu and Pereira [11] shows that Experiment 3 is Blackwell Sufficient for 1. One expects that the information gained about θ by performing Experiment 3 is at least as much as one would obtain by performing Experiments 1 or 2. Are experiments 1 or 2 also Blackwell Sufficient for 3? In this case, the experiments would be equally informative. In the next subsection we present a theorem that characterizes when two experiments are equally informative in Blackwell's sense and, thus, also answers the comparison of the experiments in Section 3.2.

4.2. Equivalence Relation in Experiment Information

In this section, the experiments can assume values in a countable set. For an experiment $X : \Omega \mapsto \mathcal{X}$, we assume that X is measurable on the power set of $\mathcal{X}$ and that $\forall \theta \in \Theta, \exists x \in \mathcal{X}, P(x|\theta) > 0$. No assumption is required of Θ.

Using Blackwell Sufficiency, it is possible to define an equivalence relation between experiments: X and Y are Blackwell Equivalent if any one is Blackwell Sufficient for the other, $X \approx Y$. This equivalence relates to the Likelihood Principle in Section 3.1 through:

Theorem 2 *Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be two experiments. $X \approx Y$ iff, for every likelihood function $L(\cdot)$,*

$$\forall \theta \in \Theta, P(\{x \in \mathcal{X} : L_X(\cdot|x) \propto L(\cdot)\}|\theta) = P(\{y \in \mathcal{Y} : L_Y(\cdot|y) \propto L(\cdot)\}|\theta)$$

The following notation reduces the algebra involved. Since all sets are countable, consider them to be ordered. Let, $\forall \theta \in \Theta$, $P(X = x|\theta)$ be a probability function, then we define that $p(.|\theta)$ is a vector such that in its i-th position the value assumed is $P(x_i|\theta)$; x_i is the i-th element of the ordering assumed in the set of values of X. Consider F to be an arbitrary map from $\mathcal{X} \times \mathcal{Y}$ into $[0, 1]$. We also use the symbol F for the countably infinite matrix that has in its j-th row and i-th column position the value of $F(x_i, y_j)$; x_i is the i-th element of the ordering in $\mathcal{X}$ and y_j is the j-th element of the ordering in $\mathcal{Y}$. Finally, a (transposed) transition matrix is such that all of its elements are greater or equal to 0 and for any column the sum of its elements is equal to 1.

Proof. ($\Leftarrow$) Let $S : \mathcal{X} \mapsto [0, 1]^\Theta$ and $T := \mathcal{Y} \mapsto [0, 1]^\Theta$, such that $S(x)$ and $T(y)$ are likelihood nuclei of x and y—a likelihood nucleus is a chosen likelihood between all of those that are proportional.

Recall from Basu [4] that S and T are, respectively, minimal sufficient statistics for X and Y. Therefore, $S \approx X$ and $T \approx Y$. By the hypothesis, (S, θ) and (T, θ) are identically distributed, therefore they are Blackwell Equivalent. By transitivity of Blackwell Equivalence $S \approx T$, since $S \approx X \approx Y \approx T$.

($\Rightarrow$) Consider the above statistics S and T. For simplicity, we call (For an arbitrary function f and set A, we define $f[A]$ as the image of A through f.) $S[X[\Omega]] = \xi_X$ and $T[Y[\Omega]] = \xi_Y$. We also call $P(S(X) = l_x|\theta) = p_X(l_x|\theta)$ and $P(T(Y) = l_y|\theta) = p_Y(l_y|\theta)$. Clearly, by construction, for every two points in ξ_X or in ξ_Y, if their likelihood functions are proportional, then they are the same point. Since S and T are minimal sufficient statistics, $S \approx X$, $T \approx Y$ and, therefore, $S \approx T$.

Since S is Blackwell Sufficient for T, there exists a map $A : \xi_X \times \xi_Y \mapsto [0, 1]$ such that A is a transition matrix and:

$$Ap_X(.|\theta) = p_Y(.|\theta), \forall\theta \in \Theta$$

On the other hand, T is also Blackwell Sufficient for S and, similarly, there exists a map $B : \xi_Y \times \xi_X \mapsto [0, 1]$ such that B is a transition matrix and:

$$Bp_Y(.|\theta) = p_X(.|\theta), \forall\theta \in \Theta$$

From these two equations, there exist two other transition matrices, $M = BA$ and $N = AB$, such that:

$$\begin{aligned} Mp_X(.|\theta) &= p_X(.|\theta), \forall\theta \in \Theta \\ Np_Y(.|\theta) &= p_Y(.|\theta), \forall\theta \in \Theta \end{aligned}$$

Since M and N are transition matrices, respectively, from ξ_X to ξ_X and from ξ_Y to ξ_Y, we consider the Markov Chains associated to them. All probability functions in the family $\{p_X(.|\theta) : \theta \in \Theta\}$ are invariant measures for M. Note that there are no transient states in M. If there were, let x be a transient state in M, consequently $P(x|\theta) = 0, \forall\theta \in \Theta$. This is a contradiction from the assumption that $\forall\theta \in \Theta, \exists x \in \mathcal{X}, P(x|\theta) > 0$; Conclude that there is no transient state in M.

Next, we use the following result found in Ferrari and Galves [12]:

Lemma 1 *Consider a Markov Chain on a countable space $\mathcal{X}$ with a transition matrix M and no transient states. Let M have irreducible components $C(1)$, ..., $C(n)$, ... Then, there exists an unique set of probability functions $\{p_j(\cdot) : j \in N\}$, with $p_j(x)$ defined in $\{1, \ldots, |C(j)|\}$, such that all invariant measures (μ) of M can be written as the following:*

If $c_{k,i}$ is the i-th element of $C(k)$, then $\mu(c_{k,i}) = p_k(i).q(k)$ and q is a probability function in N.

Recall that if a Markov Chain is irreducible, it admits a unique ergodic measure. This lemma states that any invariant measure of an arbitrary countable Markov Chain is a mixture of the unique ergodic measures in each of the irreducible components.

Using the lemma, since $C(1), \ldots, C(n), \ldots$ are irreducible components of M and $c(k, i)$ is the element of number i of $C(k)$, then $p_1(c(k, i)|\theta) = p_k(i)q_{k,\theta}$. Consequently,

$$p_1(c(k, i)|\theta) = p_1(c(k, j)|\theta) \left(\frac{p_k(i)}{p_k(j)}\right)$$

If two states are in the same irreducible component then their likelihood functions are proportional. The same proof holds to matrix N.

The i-th element of ξ_X is said to connect to the j-th element of ξ_Y if $A(i,j) > 0$. Similarly, the i-th element of ξ_Y is said to connect to the j-th element of ξ_X if $B(i,j) > 0$. Note that every state in ξ_X connects to at least one state in ξ_Y and vice-versa. This is true because A and B are transition matrices.

For all $x_1 \in \xi_X$, if x_1 connects to $y \in \xi_Y$ then y only connects to x_1. If there were a state $x_2 \in \xi_X$ such that y connected to x_2, then x_1 and x_2 would be on the same irreducible component of M. Therefore x_1 and x_2 would yield proportional likelihood functions and, by the definition of S, $x_1 = x_2$. Similarly, if a state $y \in \xi_Y$ connects to a state $x \in \xi_X$ then x connects solely to y.

Finally, we conclude that every state in ξ_X only connects to one state in ξ_Y and vice versa. Also, if $x \in \xi_X$ connects to $y \in \xi_Y$, then y connects to x and vice-versa. This implies that if x connects to y, then $P(X = x|\theta) = P(Y = y|\theta), \forall\theta \in \Theta$. Since S and T are sufficient the Theorem is proved.

Applying the above Theorem and the Likelihood Principle, one obtains the following result: if X is Blackwell Equivalent to Y,

$$A_e = \{x : Inf(X, x, \theta) = e\} \subset \mathcal{X}_1; B_e = \{y : Inf(Y, y, \theta) = e\} \subset \mathcal{X}_2$$

then $P(A_e|\theta) = P(B_e|\theta), \forall\theta \in \Theta$, for all possible e—the value of information.

> For any information function, Inf, satisfying the Likelihood Principle — if x and y yield proportional likelihood functions, then $Inf(X, x, \theta) = Inf(Y, y, \theta)$ —, X is Blackwell Equivalent to Y, if and only if, the distribution of (Inf, θ) for X and Y are the same.

Also, since the likelihood nuclei are not equally distributed in the experiments in Section 3.2, conclude that no pair of them is Blackwell Equivalent. Hence, from the conclusions in 4.1, Experiment 3 is strictly more informative than Experiments 2 and 1.

4.3. Experiment Information Function

In the last section, we defined properties an information function should satisfy. We reviewed Blackwell Sufficiency as a general rule for comparing experiments. Nevertheless, not every two experiments are comparable through this criterion. Next, we explicitly consider functions capable of describing the information of an experiment. A possible approach to this problem is considering that the information gained is a utility function DeGroot [13] that the scientist wants to maximize. This way, it follows from DeGroot [13] that $Inf(X, \theta) = E\left(Inf(X, x, \theta)\right)$. Since we consider the data information function as non-negative, the utility function is concave, see DeGroot [14] for instance.

Proceeding with this approach, we compare the different information functions presented in Section 3.2. In this example, the maximum information is obtained when the posterior distribution is such that $P(\theta = 0|x) = 0$ or $P(\theta = 0|x) = 1$. Therefore, to compare those information functions, we divide all of them by these maxima.

First, we consider Euclidean distance as the information function. In the first experiment, with probability 1 the gain of information is 33%. That is, a small gain with a small risk. On the second experiment, with probability 56% the gain is 60% of the maximum and with probability 44% it is 0%

of the maximum, moderate gain with moderate risk. In the third experiment one can get 100% of the maximum possible information with probability 33% and can get 0% of the maximum possible information with probability 67%, maximum gain with great risk. In conclusion, if one uses the Euclidian's "utility", then he/she would have no preference among the three experimnents, since, for all of them, the expected information gain is of 33%. This is surprising as the third experiment is Blackwell Sufficient for both the others.

Next, consider $Inf(X, x, \theta) = [E(\theta) - E(\theta|X = x)]^2$. The information of an experiment using this metric is: $Inf(X, \theta) = V(E(\theta|X))$. The expected information gain for each of the three experiments is, respectively, 11%, 20% and 33%. Thus, the third experiment is more informative than the second, which in turn is more informative than the first.

Similarly, considering the Kullback–Leibler divergence, the expected gain of information for each of the three experiments is, respectively, 2.4%, 4.6% and 33%. Again, the ordering induced by information gain in X_1, X_2, X_3 agrees with the ordering induced by Blackwell Sufficiency. The difference of information between experiments 3 and 2 is much higher than that between 2 and 1 when using Kullblack–Leibler divergence than when using $V(E(\theta|X))$.

5. Conclusions

We used Basu's concept of information as a starting point for reflection. To operationalize Basu's concept, we discussed some common Statistical Principles. While these principles are usually presented under a frequentist perspective, we chose a Bayesian one. For instance, the definition of the Conditionality Principle that we presented is slightly different from that in Birnbaum [3] and Basu [4]. Such principles are based on the idea that trivial experiments (or ancillary statistics) should not bring information about the parameter.

We also discussed comparison experiments. A known alternative to the classical sufficiency definition is that of Blackwell Sufficiency. Let X and Y be two experiments such that X is Blackwell Sufficient for Y, if you are restricted to choose only one, it should be X. We showed that sampling without replacement is preferable to with replacement in this sense. Blackwell Sufficiency is also useful for characterization of distributions, for instance Basu and Pereira [11].

Theorem 2 states that two experiments are Blackwell Equivalent if and only if their likelihood-function statistics are equally distributed conditionally on θ. Two applications of this Theorem are as follows. (i) If one believes in the Likelihood Principle and that two experiments are equally informative if the distribution of the information functions are equal, then the information equivalence between experiments induced by Blackwell Equivalence follows. (ii) To prove that an experiment is not Blackwell Sufficient for another is, in general, difficult: one must show that there is no transition function from one to the other. However, if X if Blackwell Sufficient for Y, using theorem 2, if the likelihood-function statistics, conditionally on θ, are not equally distributed, then Y is not Blackwell Sufficient for X. This is the case for both examples in section 4.1 and, thus, Blackwell Equivalence does not hold.

We end this paper by evoking the memory of D. Basu who, among other teachings, inspires the authors with the illuminating concept of information: "Information is what it does for you, it changes your opinion".

Acknowledgments

The authors are grateful to Basu for his legacy on Foundations of Statistics. We are grateful to Adriano Polpo, Estéfano Alves de Souza, Fernando V. Bonassi, Luis G. Esteves, Julio M. Stern, Paulo C. Marques and Sergio Wechsler for the insightful discussions and suggestions. We are grateful to the suggestions from the anonymous referees. The authors of this paper have benefited from the support of CNPq and FAPESP.

References

1. Basu, D. A Note on Likelihood. In *Statistical Information and Likelihood : A Collection of Critical Essays*; Ghosh, J.K., Ed.; Springer: Berlin, Germany, 1988.
2. Torgersen, E.N. *Comparison of Experiments*; Cambridge Press: Cambridge, UK, 1991.
3. Birnbaum, A. On the foundations of statistical inference. *J. Am. Stat. Assoc.* **1962**, *57*, 269–326.
4. Basu, D. Statistical information and likelihood: a collection of critical essays. In *Statistical Information & Likelihood*; Ghosh, J.K., Ed.; Springer: Berlin, Germany, 1988.
5. Lindley, D.V.; Philips, L.D. Inference for a bernoulli process (a bayesian view). *Am. Stat.* **1976**, *30*, 112–119.
6. Pereira, C.; Lindley, D.V. Examples questioning the use of partial likelihood. *Statistician* **1987**, *36*, 15–20.
7. Wechsler, S.; Pereira, C.; Marques, P. Birnbaum's theorem redux. *AIP Conf. Proc.* **2008**, *1073*, 96–100.
8. Blackwell, D. Comparison of experiments. In Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 31 July–12 August 1950.
9. Basu, D.; Pereira, C. Blackwell sufficiency and bernoulli experiments. *Braz. J. Prob. Stat.* **1990**, *4*, 137–145.
10. Goel, P.K.; Ginebra, J. When is one experiment "always better than" another? *Statistician* **2003**, *52*, 515–537.
11. Basu, D.; Pereira, C. A note on blackwell sufficiency and skibinsky characterization of distributions. *Sankhya A* **1983**, *45*, 99–104.
12. Ferrari, P.; Galves, A. *Coupling and Regeneration for Stochastic Processes*; Sociedad Venezoelana de Matematicas: Caracas, Venezuela, 2000.
13. DeGroot, M.H. *Optimal Statistical Decisions*; Wiley: New York, NY, USA, 1970.
14. DeGroot, M.H. Uncertainty, information, and sequential experiments. *Ann. Math. Stat.* **1971**, *33*, 404–419.