Statistical Information: A Bayesian Perspective

R. B. STERN and C. A. de B. PEREIRA*

Department of Statistics, IME-USP,
São Paulo, SP, Brazil
**E-mail: cpereira@ime.usp.br*

We explore the concept of information in statistics: information about unknown quantities of interest, the parameters. We discuss intuitive ideas of what should be information in statistics. Our approach on information is divided in two scenarios: observed data and planning of an experiment. On the first scenario, we discuss the Sufficiency Principle, the Conditionality Principle, the Likelihood Principle and their relationship with trivial experiments. We also provide applications of some measures of information to an intuitive example. On the second scenario, the definition and new applications of Blackwell Sufficiency are presented. We discuss a new relationship between Blackwell Equivalence and the Likelihood Principle. Finally, the expected values of some measures of information are calculated.

Keywords: Blackwell Equivalence, Blackwell Sufficiency, Conditionality Principle, Likelihood Principle, Sufficiency Principle.

1. Introduction

The concept of information is perhaps one of the most controversial in Statistics. One can find innumerous different measures of the amount of information an experiment or a given data set brings about unknown quantities of interest — parameters. It is important to explore these measures because, after all, extracting information seems to be the ultimate goal of Statistics. More precisely, the goal is to extract — from an observed data or from an experiment to be performed — information about unknown quantities of interest. The intuitive definition of information which will guide this paper is given by Ref. 1 and is as follows:

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

This conceptual definition leads one to the following four questions:

- Information about what?

- Where is the information?
- How is information extracted?
- How much information is used?

We are interested in defining information about a parameter, θ, which assumes possible outcomes in Θ, the parameter space. Hence, the answer to the first question is straightforward.

A parameter represents a state of nature which we are uncertain of. For example, one can be interested in the number of days in which it will rain next year. For instance, θ can be this number and Θ all natural numbers smaller or equal to 366.

Next, we try to answer the second question. It is important to note that when defining Θ we are already using some previous knowledge about θ. In the example, 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 elicitate the prior distribution for θ. Here the prior distribution represents our description of our present state of uncertainty about θ. Mainly, the statisticians' goal is to decrease her/his uncertainty about this unknown quantity of interest. In order to reach such objective, (s)he observes data that, in his/her opinion, is related to the parameter. Consequently, one expects that there is information about θ in the data to be observed. That is, answering the second question, statistical (expected) information is contained in the collected data set (experiment to be performed).

It seems natural at this point to ask: How to extract information contained in the observed data? In order to answer this question properly, the "scientist" considers a global probability space involving a prior distribution on θ and the experimental distribution for every possible θ. Next, the "scientist" uses the Bayes operation to obtain the posterior distribution. The posterior distribution describes the uncertainty about the parameter after calibrating the prior by the observed data. Thus, we could say that the new information also depends on the statistical framework. The act of observing the data corresponds to a mechanism of transforming unknown quantities in known ones. We also call such a mechanism an *experiment*.

The last question which is important in design of experiments is: "How much information is extracted after the experiment is performed (after the data is observed)?". After obtaining an adequate answer, it is possible to modify the question to a *pre-posterior analysis*: "How much information do we expect to obtain in a specific experiment to be performed?". The heart

of this paper is to explore possible answers for both questions.

Section 3 analyzes how informative a particular data set is. Section 3.1 introduces common principles in Statistics and their relationship with the Likelihood Principle. Section 3.2 presents a simple example and three information functions compatible with the principles of Section 3.1.

Section 4 is related to experimental design and tries to answer the following question: Among the possible alternative experiments, which is our best choice? Blackwell Sufficiency is a possible criterion to compare experiments. The definition of Blackwell Sufficiency, with examples, is presented in Section 4.1. The Likelihood Principle and its relationship with Blackwell's Equivalence are discussed in Section 4.2. We argue that Blackwell Sufficiency is the best criterion whenever the experiments are comparable in that sense. Finally, since not all experiments are comparable in Blackwell's sense, Section 4.3 explores the metrics exposed in Section 3.2 using the framework of decision theory to compare experiments.

2. Definitions

In the context of experimental information we will always be concerned with a probability space, that 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 random quantity R corresponds to a function from Ω to some set $\Re$. We define the probability space induced by R, $(\Re, \Im_R, P_R)$, where $\Im_R = \left\{M \subset \Re : R^{-1}[M] \in \Im\right\}$ and $P_R(M) = P\left(R^{-1}[M]\right)$. Finally, the σ-algebra induced on Ω by a random quantity R is called $\Im_{|R}$ and corresponds to $\left\{R^{-1}[M] : M \in \Im_R\right\}$.

We call an experiment any random quantity which can be observed, that is, which can be known. The realization of an experiment corresponds to the observation of this random quantity. It is important to observe that classifying a random variable as an experiment has nothing to do with the probability space, but with the limitations which exist in the world.

On the other hand, a parameter is a random quantity of interest. If the parameter were an experiment, its value could be known and the work of the statistician would be easy. Nevertheless, in many cases the parameter is not an experiment. Therefore, it is necessary to learn about it in an indirect manner, that is, observing those random quantities which are experiments and applying the Bayes Theorem. Again, classifying a random variable as a parameter has nothing to do with the probability space, but with an aspect of the world, our interest.

Therefore, in our representation, experiments and parameters are

treated essentially in the same way: both are random quantities. The given names only reflect some aspects of the world not contemplated in the probability space, that is, the observability of the random quantity and our interest in it.

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

From now on, we restrict ourselves to random quantities whose probability distributions are absolutely continuous or discrete. Following this restriction, $p_X(x|\theta)$ is the conditional probability (density) function of the random quantity X given 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.

Finally, we say that an experiment $X : \Omega \mapsto \mathcal{X}$ is trivial for $\theta : \Omega \mapsto \Theta$ 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, X by itself does not carry "information" about θ.

3. Data Information

3.1. *Statistical Principles and Information*

At this point, we hope to have convinced the reader that it is an important task to define how informative is the observed data of an experiment. Therefore it seems reasonable to assume the existence of an information function which has as arguments the experiment, its observed data, and the parameter. Although this function is still not defined, to make reference to it we will use the notation $Inf(X, x, \theta)$, where X is an experiment, x the observed data and θ the parameter.

Before establishing an exact form for such a function, we look for properties, as in Ref. 1 and Ref. 2, which follow from common statistical principles.

A statistic $T : \mathcal{X} \mapsto \tau$ is called sufficient if X and θ are conditionally independent given T, that is, X is a trivial experiment for θ given T. Previously, it was discussed that there are reasonable reasons to believe that a trivial experiment for θ does not bring information about the parameter. Therefore, since X is a trivial experiment for θ given T, it seems that all the information in X is gained by observing only T. This is the heart of the Sufficiency Principle. 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 statisticians, although not always explicitly mentioned: for inference about θ the statistician only needs to consider a sufficient statistic.

The Conditionality Principle is another important statistical principle: it might be seen as the reciprocal of the Sufficiency one. In the latter we say that the trivial experiment realized after T does not bring information about θ. Considering X_1 and X_2 as possible experiments, the Conditionality Principle states that a trivial experiment for θ, X_1 and X_2 does not bring information about θ. Let Y be some trivial experiment for (θ, X_1, X_2) assuming values in $\{1, 2\}$ and X_Y a mixture of X_1 and X_2. X_Y is observed in the following way: If the result of Y is 1 then we observe the result of X_1, if it is 2 then we observe the result of X_2. The Conditionality Principle states that $Inf((Y, X_Y), (i, x), \theta) = Inf(X_i, x, \theta)$, $\forall i \in \{1, 2\}$. In the sequel we show that this principle is more controversial than that of Sufficiency.

The Likelihood Principle is the object of the last part of this section. The Likelihood Principle states that any two possible outcomes having proportional likelihood functions must provide the same information about the parameter. 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 $Inf(X_1, x_1, \theta) = Inf(X_2, x_2, \theta)$. It is possible to prove that this principle is stronger than the Sufficiency Principle and than the Conditionality Principle. Below, we prove that both the Sufficiency and the Conditionality Principles imply the Likelihood Principle. Ref. 2 and Ref. 3 also presented proofs of this important result.

Theorem 3.1. *The Sufficiency and the Conditionality Principles imply the Likelihood Principle.*

Proof. Consider $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ outcomes of, respectively, experiments X_1 and X_2, such that $L(\theta|x_1) \propto L(\theta|x_2)$. The statistic $T : \{1, 2\} \times (\mathcal{X}_1 \cup \mathcal{X}_2) \mapsto (\{1, 2\} \times (\mathcal{X}_1 \cup \mathcal{X}_2)) \cup \{0\}$ is such that: 1) $T(i, z) = (i, z)$, if $(i, z) \neq (1, x)$ and $(i, z) \neq (2, y)$; 2) $T(i, z) = 0$, if $(i, z) = (1, x)$ or $(i, z) = (2, y)$. Define $Y \in \{1, 2\}$ in such a way that $P(Y = 1) = P(Y = 2)$ and Y is trivial for (θ, X_1, X_2). Y is used to define a mixture X_Y. Note that T is a sufficient statistic for (Y, X_Y). Therefore, using the Sufficiency Principle, $Inf((Y, X_Y), (1, x), \theta) = Inf((Y, X_Y), (2, y), \theta)$. Next, by Conditionality, $Inf(X_1, x, \theta) = Inf((Y, X_Y), (1, x), \theta)$ and $Inf(X_2, y, \theta) = Inf((Y, X_Y), (2, y), \theta)$. The Likelihood Principle follows straightforward. □

A "scientist" that follows the Likelihood Principle can perform inference about the parameter solely based on the likelihood function. However, this principle is not followed by many statisticians. Ref. 4 and Ref. 5 provide interesting examples in which some classical statistical methods violate the Likelihood Principle. In one example, someone observes 3 failures out of 10 trials on two experiments. On the first one, the person would realize 10 trials and therefore, the distribution of the total of failures is binomial. On the second one the person would realize trials until he observed 3 failures and, therefore, the distribution of the total of failures is the negative binomial. Even though the unbiased estimation produces different results, posterior distributions are the same. This way, since classical statisticians follow the Sufficiency Principle and do not follow the Likelihood Principle, it is possible to conclude that they do not follow the Conditionality Principle. On the other hand, as was also shown in Ref. 6, bayesian statisticians usually follow all three principles.

In our opinion, it is reasonable to follow both the Sufficiency and Conditionality Principles whenever performing statistical analysis. Therefore, when looking for a more rigorous definition of the information function in the remaining part of this paper, the Likelihood Principle will be assumed. That is, the information given by the observation of some data will not give different values for data with proportional likelihood functions.

3.2. *Information on observed data*

Our interest in this section is on a function of the information about θ contained in a particular observed data. In the last section it was seen that such a function must not give different values for points with proportional likelihoods. Nevertheless, this property only gives a vague idea about how such a function should be. Therefore, we will go back to the definition given in the introduction in search for intuition: "Information is what it does for you, it changes your opinion".

Recall that the opinion one has about the parameter before (s)he realizes the experiment is given by his/her prior distribution for it. On the other hand, the opinion one has after an experiment is realized becomes, by the use of Bayes Theorem, his/her posterior distribution. Hence, since the information should represent the changing in opinion, it seems reasonable for it to compare prior and posterior distributions. If prior and posterior distributions are equal, it seems reasonable to assume there is no gain of information. We will also define that the information given when the prior distribution is equal to that of the posterior is 0. Since for any posterior

different to the prior distribution there is some change in opinion, it seems reasonable to assume that information is always greater or equal to 0.

As discussed before, a trivial experiment should not bring any information about the parameter. Therefore, it should be less or equally informative than any other experiment. Note that prior and posterior are equal for trivial experiments, and the properties expected of these experiments are consistent with statement above.

Consider the following simple example: Someone chooses 3 balls among 4 balls, 2 black ones and 2 white. The 3 balls are put in an box. Another person is offered the possibility of observing one of three experiments; Experiment 1 consists of taking only one ball out of the urn; Experiment 2, two balls with replacement and; Experiment 3, two balls without replacement. The objective is to guess the number of white balls in the urn, 1 or 2. The person, a priori, does not believe any combination of balls is more likely than others, a uniform prior. Finally, the person assumes 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 the number of white balls observed in the i-th experiment, denote by $P(\theta|X_i = x_i) = (a, b)$, for $P(\theta = 1|X_i = x_i) = a$ and $P(\theta = 2|X_i = x_i) = b$. The posterior probabilities are as follows:

(1) $P(\theta|X_1 = 0) = \left(\frac{2}{3}; \frac{1}{3}\right)$, $P(\theta|X_1 = 1) = \left(\frac{1}{3}; \frac{2}{3}\right)$;
(2) $P(\theta|X_2 = 0) = \left(\frac{4}{5}; \frac{1}{5}\right)$, $P(\theta|X_2 = 1) = \left(\frac{1}{2}; \frac{1}{2}\right)$, $P(\theta|X_2 = 2) = \left(\frac{1}{5}; \frac{4}{5}\right)$;
(3) $P(\theta|X_3 = 0) = (1; 0)$, $P(\theta|X_3 = 1) = \left(\frac{1}{2}; \frac{1}{2}\right)$, $P(\theta|X_3 = 2) = (0; 1)$.

Some common information functions 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:
$Inf_{KL}(X_i, x_i, \theta) = \sum_j P(\theta = j|X_i = x_i) \log\left(\frac{P(\theta=j|X_i=x_i)}{P(\theta=j)}\right)$.

For experiment 1, the information given by these functions for each vector of possible outcomes is, respectively, $\left(18^{-1/2}; 18^{-1/2}\right)$, $\left(36^{-1}; 36^{-1}\right)$ and $(0.024; 0.024)$. For experiment 2, $(0.424; 0; 0.424)$, $(0.09; 0; 0.09)$ and $(0.084; 0; 0.084)$. Finally, for experiment 3, $\left(2^{-1/2}; 0; 2^{-1/2}\right)$, $(0.25; 0; 0.25)$ and $(0.301; 0; 0.301)$.

These measures have interesting properties. Nevertheless, they do not offer a straightforward way of comparing experiments. Next section, this

problem will be discussed in detail.

4. Information in an Experiment

4.1. *Blackwell Sufficiency*

A celebrated definition in Statistical Inference is that of Sufficiency. According to the Sufficiency Principle exposed in Section 3.1, any inference based on data or on a sufficient statistic should be the same. Nevertheless, this principle is only useful to compare statistics inside the same space. Consider two experiments, X and Y, that depend on the parameter θ. One usually wants to choose between X and Y for inferences about θ based solely on their marginal distributions — the conditional distributions of X given θ and Y given θ. In this case, clearly the Sufficiency Principle is useless. In this section we expose the concept of Blackwell Sufficiency[7] and show that it is a natural generalization of the Sufficiency Principle for comparison of experiments.

From Section 3.1, a statistic T is sufficient for an experiment X, if X and θ are conditionally independent given T. Consequently, T is sufficient if and only if $p(x|\theta) = p(t|\theta)p(x|t)$. After performing T, there exists a simple "randomization exercise" which produces an experiment like X.

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, one can see that the random 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 and, after that, performing a simple randomization exercise, Z_x.

The structure presented above leads us to conclude that Blackwell Sufficiency is the version of Sufficient Statistics for the comparison of experi-

ments. Hence, it seems reasonable to state that is better to realize X than Y. It is not difficult to see that every Sufficient Statistic is also Blackwell Sufficient.

Having presented the overall aspects of Blackwell Sufficiency, a few examples are needed. These are related to those found in Ref. 8 and seem to give support to well known beliefs.

Example 4.1. Let X and Y be two experiments, π a parameter in $[0,1]$ and q and p known constants in $[0,1]$. Representing the Bernoulli distribution with parameter p by $\mathrm{Ber}(p)$, consider also that the conditional distributions of X and Y given π are, respectively:

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

If $A \sim \mathrm{Ber}(q)$ and $B \sim \mathrm{Ber}(p)$, either one independent of all variables considered in this problem, then for $Y' = AX + (1-A)B$ we have that (Y', π) and (Y, π) are equally distributed. Therefore, X is Blackwell Sufficient for Y.

Example 4.2. Let X and Y be two experiments, π a parameter in $[0,1]$, q and p known constants in $[0,1]$ and $n \in N$. Consider also the following. Representing the Binomial distribution with parameters n and p by $\mathrm{Bin}(n,p)$, consider also that the conditional distributions of X and Y given π are, respectively:

$$X \sim \mathrm{Bin}(n, \pi) \text{ and } Y \sim \mathrm{Bin}(n, q\pi + (1-q)p).$$

We know that X has the same distribution as a sufficient statistic for a sequence $X_1, \ldots, X_n$ conditionally independent identically distributed given π with $X_1 \sim \mathrm{Ber}(\pi)$. Thus X is Blackwell Sufficient for $X_1, \ldots, X_n$.

From example 4.1, we know that the sequence previously considered is Blackwell Sufficient for a sequence $Y_1, \ldots, Y_n$ conditionally independent given π and such that, conditionally to π, $Y_1 \sim \mathrm{Ber}(q\pi + (1-q)p)$.

Finally, $Y_1, \ldots, Y_n$ is Blackwell Sufficient for Y, since it has the same distribution as a sufficient statistic for $Y_1, \ldots, Y_n$. Thus, since Blackwell Sufficiency is a transitive relation, X is Blackwell Sufficient for Y.

Example 4.3. 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 selected from the urn.

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

(1) Conditionally to θ, $X_1 \sim \text{Ber}\left(\frac{\theta}{N}\right)$;
(2) Conditionally to 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 to θ, $Y_1 \sim \text{Ber}\left(\frac{\theta}{N}\right)$;
(2) $Y_{i+1}|(y_i, \ldots, y_1, \theta) \sim \text{Ber}\left(\frac{\theta - \sum_{j=1}^{i} y_j}{N-i}\right)$,
$\forall i \in \{1, \ldots, n-1\}$, $\forall (y_i, \ldots, y_1) \in \{0,1\}^i$.

Next, we prove that $(Y_1, \ldots, Y_n)$ is Blackwell Sufficient for $(X_1, \ldots, X_n)$.

Define $X_1^* = Y_1$, $t_i = \sum_{j=1}^{i} y_j$ and $\forall i \in \{1, \ldots, n-1\}$ two random 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) $B_{i+1}|t_i \sim \text{Ber}\left(\frac{t_i}{i}\right)$;
(3) $\forall i \in \{1, \ldots, n\}$, B_i, conditionally to t_i, is independent of $((A_1, \ldots, A_n); (B_1, \ldots, B_{i-1}); (Y_{i+1}, \ldots, Y_n); \theta)$.

Define:

$$X_{i+1}^* = A_{i+1} Y_{i+1} + (1 - A_{i+1}) B_{i+1}.$$

We obtain that, conditionally to 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 $(B_i, \ldots, B_2)$, we conclude that X_{i+1}^* is independent of $(X_i^*, \ldots, X_1^*)$ given θ.

By the previous conclusions, we know that $(X_1^*, \ldots, X_n^*, \theta)$ is identically distributed to $(X_1, \ldots, X_n, \theta)$. By construction, we also know that $(X_1^*, \ldots, X_n^*)|(Y_1 = y_1, \ldots, Y_n = y_n)$ is trivial, $\forall (y_1, \ldots, y_n) \in \{0,1\}^n$. Therefore, it is proved that sampling without replacement is Blackwell Sufficient for sampling with replacement.

Example 4.4. A simple corollary of the above result is now presented. First we recall that T_n is a sufficient statistic for $(Y_1, \ldots, Y_n)$, the sampling without replacement presented in Example 3, and that, conditionally to θ, T_n is Hypergeometric with parameter (N, n, θ); Consequently, T_n is Blackwell Sufficient for $(Y_1, \ldots, Y_n)$. We know that $(Y_1, \ldots, Y_n)$ is Blackwell Sufficient for $(X_1, \ldots, X_n)$, , the sampling with replacement presented

in Example 3. Also $(X_1, \ldots, X_n)$ is Blackwell Sufficient for $\sum_{i=1}^{n} X_i = S_n$ and, conditionally to θ, $S_n \sim \text{Bin}(n, \theta/N)$. Using the transitive property of Blackwell Sufficiency we conclude that T_n is Blackwell Sufficient for S_n.

Example 4.5. In many inferential problems we are concerned with an experiment Y and a parameter μ such that $Y|\mu \sim N(\mu, s)$, a normal random quantity. We commonly think of this model as $Y = \mu + \epsilon$, $\epsilon \sim N(0, s)$. We interpret ϵ as some random noise on a measure of μ. The larger the value of s, the more intense the noise is. Therefore, it seems reasonable that if $X \sim N(\mu, s')$ and $s < s'$, then Y is more informative for μ than X. It is possible to prove that this intuitive idea is preserved by Blackwell Sufficiency. Let $Z \sim N(0, s'-s)$ independent of Y, then $(Y+Z, \theta)$ is equally distributed to (X, θ).

Example 4.6. Other two important distributions are the Poisson and Exponential. We show that if, conditionally to θ, $X_1 \sim \text{Poisson}(\theta)$ and $X_2 \sim \text{Poisson}(p\theta + k)$, $k \in R_+$, $p \in [0, 1]$, then X_1 is Blackwell Sufficient for X_2. Using the same technique we will also prove that, conditionally to θ, if $Y_1 \sim \text{Exp}(\theta)$ and $Y_2 \sim \text{Exp}(p\theta + k)$, $k \in R_+$, $p \in [0, 1]$, then Y_1 is Blackwell Sufficient for Y_2.

Conditionally to θ, let Z' and A be independent Poisson Processes in $[0, 1]$ with rates respectively θ and k. Define a new process, Z, in which we choose, independently and with probability $(1 - p)$, points of Z' to be discarded. Conditionally to θ, Z is still a Poisson Process in $[0, 1]$ with rate $p\theta$ and $Z \cup A$ is a Poisson Process in $[0, 1]$ with rate $p\theta + k$. If $n(Z'), n(Z), n(A)$ and $n(Z \cup A)$ are the number of occurrences in Z', Z, A and $Z \cup A$, then $n(Z) + n(A) = n(Z \cup A)$. Since $(n(Z'), \theta)$ has the same distribution of (X, θ) and $(n(Z \cup A), \theta)$ has the same distribution of (X_2, θ), then Blackwell Sufficiency is proven.

Here, conditionally to θ, let Z' and A be independent Poisson Processes in R_+ with rates respectively θ and k. Again, Z is the process which is obtained by disregarding points with probability $(1 - p)$: $Z \cup A$ is a Poisson Process in R with rate $p\theta + k$. If T'_Z, T_Z, T_A and $T_{Z \cup A}$ are the waiting times until the first occurrence, respectively, in Z', Z, A and $Z \cup A$, then $T'_Z \sim \text{Exp}(\theta)$, $T_Z \sim \text{Exp}(p\theta)$, $T_A \sim \text{Exp}(k)$ and $T_{Z \cup A} \sim \text{Exp}(p\theta + k)$. Since, $T_{Z \cup A} = \min(T_Z, T_A)$, it is proven that $\text{Exp}(\theta)$ is Blackwell Sufficient for $\text{Exp}(p\theta + k)$.

4.2. *Equivalence relation in experiment information*

In this section, Ω is a countable set and $F = \wp(\Omega)$, that is, the set of all subsets of Ω. Therefore, for any random quantity $X : \Omega \mapsto \mathcal{X}$, $F_X = \wp(\mathcal{X})$. It will also be assumed that $\forall\theta \in \Theta, \exists x \in \mathcal{X}, P(x|\theta) > 0$.

Blackwell Sufficiency was introduced in the last section. Using it, 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, $X \approx Y$. We show that this equivalence relates to the Likelihood Principle, Section 3.1. In the sequel, we prove the following result:

Theorem 4.1 (Blackwell likelihood (BL)). *Let $X \in \mathcal{X}_1$ and $Y \in \mathcal{X}_2$ be two experiments. $X \approx Y$ if and only if $P(\{x \in \mathcal{X}_1 : L_X(\cdot|x) \propto L(\cdot)\}|\theta') = P(\{y \in \mathcal{X}_2 : L_Y(\cdot|y) \propto L(\cdot)\}|\theta')$, $\forall\theta' \in \Theta$ and all likelihood function $L(\theta)$ derived from either X or Y.*

For simplicity, we use a special notation that reduces the algebra involved. Since all sets are countable, we consider an ordering inside them. 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. If F is a map from $\xi_1 \times \xi_2$ into $[0, 1]$, then we also use the symbol F as the matrix which 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 ξ_1 and y_j is the j-th element of that in ξ_2. Finally, we recall that a transition (transposed) matrix is a matrix in which all elements are greater or equal to 0 and for any column the sum of its elements is equal to 1.

Proof. ($\Leftarrow$) Let $S : \mathcal{X}_1 \mapsto [0, 1]^\Theta$ and $T : \mathcal{X}_2 \mapsto [0, 1]^\Theta$, 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 which are proportional. Recall Ref. 3 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 the transitive property 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 $S(X(\Omega)) = \xi_X$ 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:

$$Mp_X(.|\theta) = p_X(.|\theta), \forall\theta \in \Theta,$$
$$Np_Y(.|\theta) = p_Y(.|\theta), \forall\theta \in \Theta.$$

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 $P(x|\theta) = 0$, $\forall\theta \in \Theta$. However, such a state cannot occur by the definition in the beginning of this section; there is no transient state in M.

To proceed with the proof we use the following result stated in Ref. 9:

Lemma 4.1. *Consider a Markov Chain on a countable space $\mathcal{X}$ with a transition matrix A and no transient states. Let A 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 A 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.

To interpret the above result, we consider the sub-matrix A_k associated to $C(k)$. Since the A_k is irreducible, it only has one invariant measure, p_k. Now suppose that at the initial position (X_0) of the Chain each component $C(k)$ has probability q_k of being chosen. As n increases the law of X_n converges to the one provided by the lemma.

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.

4.3. *Experiment Information Function*

In the last section, we did not define any information function but a representation of it, Inf. Now, we are interested possible functions capable of describing the information of a statistical experiment. A possible approach to this problem is considering that the information gained is an utility function[10] which the researcher wants to maximize. This way, by the results in decision theory $Inf(X, \theta) = E\left(Inf(X, x, \theta)\right)$. Since we consider the data

information function as non-negative, the utility function is concave, see Ref. 11 for instance.

Proceeding with this approach, we compare the different information functions presented in Section 3.2. Recall the example in Section 3.2 and remember that, for any of the information functions introduced, the maximum possible value 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.

For the three possible experiments described in Section 3.2, X_1, X_2, X_3, the information vectors using euclidean distance are: $\left(18^{-1/2}; 18^{-1/2}\right)$, $(0.424; 0; 0.424)$ and $\left(2^{-1/2}; 0; 2^{-1/2}\right)$. Since the maximum possible information is $\frac{1}{\sqrt{2}}$, 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.

Consider the second information function, $Inf(X, x, \theta) = [E(\theta) - E(\theta|X = x)]^2$. In the three possible experiments, the information vectors using this function are, $\left(\frac{1}{36}; \frac{1}{36}\right)$, $(0.09; 0; 0.09)$ and $(0.25; 0; 0.25)$. Dividing by the maximum, we obtain the following vectors: $\left(\frac{1}{9}; \frac{1}{9}\right)$, $(0.36; 0; 0.36)$ and $(1; 0; 1)$. Therefore, the expected information gain is, respectively, 11%, 20% and 33%. Thus, the third experiment is more informative than the second which is more informative than the first. It is interesting to note that the information of an experiment using this metric is: $Inf(X, \theta) = V(E(\theta|X))$.

Finally, considering the Kullback-Leibler divergence, the informations vectors are: $(0.024; 0.024)$, $(0.084; 0; 0.084)$ and $(0.301; 0; 0.301)$. Dividing by the maximum, .301, we obtain: $(0.081; 0.081)$, $(0.278; 0; 0.278)$ and $(1; 0; 1)$. The expected gain of information is respectively, 2.4%, 4.6% and 33%. Again, the informativeness order, X_1, X_2, X_3, is in complete agreement with the ordering induced by Blackwell Sufficiency.

5. Final Considerations

Our main objective in writing this note was to reflect on the concept of information. It seems that there is no universal concept of information. The reader might have noticed that we follow an approach different from the one provided by Information Theory. As a matter of fact, we were not concerned with the old and important concept of information of a (prior) distribution. Another possible approach is related to Fischer Information. In this case, the value of the information depends on the value of the parameter, which is unknown and the object of the inference. It must be recognized that Fischer Information is one of the most important tools for the construction of Modern Statistical Inference. What could be questioned is the title given for this measure: Information. In fact, we do not know how to answer the questions presented in Section 1 when using this measure.

We used Basu's concept of information to develop our reflection. To operationalize Basu's concept, we had to strongly use the Statistical Principles. We alert the reader that we presented these principles with Bayesian "eyes" while, frequently, those principles are presented under a frequentist perspective. For instance, the definition of the Conditionality Principle that we presented is slightly different from that of Ref. 2 and Ref. 3. We are convinced that trivial experiments (or ancillary statistics) should not bring information about the parameter and, therefore, one should follow the Sufficiency and Conditionality Principles. Consequently, one must follow the Likelihood Principle and abandon approaches based on sample spaces: the frequentist way. One of the most celebrated hypothesis test is the asymptotic likelihood ratio test that does not depend on stopping rules. Consequently, it also does not depend on the sample spaces. The three Statistical Principles, although created by frequentist statisticians, are intrinsically considered whenever using the Bayesian operation.

On trying to understand the concept of information, we were leaded to the problem of comparison of experiments — a pre-posterior analysis. When comparing statistics in different sample spaces, a known alternative to the classical sufficiency definition is that of Blackwell Sufficiency. A principle can be naturally induced by Blackwell Sufficiency. Let X and Y be two experiments such that X is Blackwell Sufficient for Y. If you are restricted to chose only one, it should be X! This principle was used in some important cases, for example, to show that sampling without replacement is preferable than that with replacement. Blackwell Sufficiency is also useful for characterization of distributions, for instance Ref. 12.

The main achievements of this paper are the examples of the use of

Blackwell Sufficiency and the theorem of Section 4.2. Looking at the examples, some of the preferences were expected — the ones in the sampling and the normal distribution. Other examples, such as 1, 2 and 6, surprise the authors. An important question related to all of these examples is: Are the experiments Blackwell Equivalent? The answer is no: for Example 1, see Ref. 8. The general proof for the negative answer uses the BL Theorem, discussed below.

BL Theorem states that two experiments are Blackwell Equivalence, if and only if, the two likelihood-function statistics are equally distributed conditionally to 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 person believes in the information equivalence between experiments induced by Blackwell Equivalence. 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 BL Theorem, if the likelihood-function statistics, conditionally to θ, are not equally distributed, then Y is not Blackwell Sufficient for X. We leave to the reader to check that, in all the examples of Section 4.1, Blackwell Equivalence does not hold. Recall that we did not prove BL Theorem for the absolutely continuous cases.

In the first example, discussed in Section 3.2, we try three measures of information. The first one, based on the Euclidean distance, did not differentiate the experiments. This seems incoherent since there is a strict order of preference induced by Blackwell Sufficiency. The second and third measures of information both satisfy the strict order of preference induced by Blackwell Sufficiency. An important distinction between these two measures is that, while the Kullback-Leibler divergence does not take into account the possible values of the parameter space, the variance of the posterior mean does strongly depend on these values. Another difference is that the latter may not be defined, although the former always can be computed.

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 Professor 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. The authors of this paper have benefited from the support of CNPq and FAPESP.

References

1. D. Basu, A note on likelihood, in *Lecture Notes in Statistics* **45**, (Springer-Verlag, Berlin, 1988)
2. A. Birnbaum, *Journal of the American Statistical Association* **57**, 269 (1962).
3. D. Basu, Statistical information & likelihood, in *Lecture Notes in Statistics* **45**, (Springer-Verlag, Berlin, 1988)
4. D. V. Lindley and L. D. Philips, *The American Statistician* **30**, 112 (1976).
5. C. A. B. Pereira and D. V. Lindley, *The Stastician* **36**, 15 (1987).
6. S. Wechsler, C. A. B. Pereira and P. C. Marques, *AIP Conference Proceedings* **1073**, 96 (2008).
7. D. Blackwell, *Proceedings of the 2nd Berkeley Symposium* **57**, 93 (1971).
8. D. Basu and C. A. B. Pereira, *Brazilian Journal of Probability and Statistics* **4**, 137 (1990).
9. P. A. Ferrari and A. Galves, *Coupling and Regeneration for Stochastic Processes* (Sociedad Venezoelana de Matematicas, Caracas, 2000).
10. M. H. DeGroot, *Optimal Statistical Decisions* (Wiley, New York, 1970).
11. M. H. DeGroot, *Annals of Mathematical Statistics* **33**, 404 (1971).
12. D. Basu and C. A. B. Pereira, *Sankhya: The Indian Journal of Statistics, Series A* **45**, 99 (1983).