entropy

MDPI

Article

An Information-Theoretic Measure for Balance Assessment in Comparative Clinical Studies

Jarrod E. Dalton [1,*], William A. Benish [2] and Nikolas I. Krieger [1]

[1] Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland Clinic Lerner College of Medicine at Case Western Reserve University, 9500 Euclid Avenue, Cleveland, OH 44126, USA; kriegen@ccf.org

[2] Department of Internal Medicine, Case Western Reserve University, Cleveland, OH 44106, USA; wab4@cwru.edu

* Correspondence: daltonj@ccf.org

Received: 4 December 2019; Accepted: 12 February 2020; Published: 15 February 2020

Abstract: Limitations of statistics currently used to assess balance in observation samples include their insensitivity to shape discrepancies and their dependence upon sample size. The Jensen–Shannon divergence (JSD) is an alternative approach to quantifying the lack of balance among treatment groups that does not have these limitations. The JSD is an information-theoretic statistic derived from relative entropy, with three specific advantages relative to using standardized difference scores. First, it is applicable to cases in which the covariate is categorical or continuous. Second, it generalizes to studies in which there are more than two exposure or treatment groups. Third, it is decomposable, allowing for the identification of specific covariate values, treatment groups or combinations thereof that are responsible for any observed imbalance.

Keywords: balance; Jensen–Shannon divergence; observational study; relative entropy; selection bias

1. Introduction

The goal of comparative studies is to measure the effect of two or more treatment (or exposure) groups on an outcome. A potential source of bias in these studies is the association between the treatment groups and one or more confounding variables. Randomized clinical trials mitigate this risk through randomization of treatments, resulting in balanced groups with respect to the confounding variables. We say that the relationship between treatment T and outcome O is confounded by a covariate C if C is associated with O and T but is not a consequence of T (i.e., not a mediator of the effect of T on O) [1].

A common strategy for evaluating the potential for confounding in such a study is to identify all covariates that may meet these criteria and evaluate their association with T. When treatment groups T are balanced on a variable C, that is, when T and C are probabilistically independent, then C cannot confound the estimation of the relationship between T and O.

A variety of techniques are typically employed to assess balance in observational samples, including estimation of simple univariate descriptive statistics, univariate tests of association, and estimation of standardized difference scores (defined as the difference in means between groups divided by a combined estimate of standard deviation). Depending on the situation, however, each of these three approaches may lead to erroneous conclusions. Univariate descriptive statistics may not adequately capture complex distributions (e.g., those with multiple modes) [2]. Tests of association are heavily dependent on sample size, and thus can be as indicative of sample size as they are of imbalance. And standardized difference scores—despite their popularity—are not sensitive to discrepancies in higher order moments (e.g., skewness, kurtosis) and/or multimodalities among continuous distributions.

In this article, we propose the use of an information-theoretic measure known as the Jensen–Shannon divergence (JSD) [3] to assess treatment group balance. The JSD offers several advantages over the aforementioned approaches. First, it is universally defined for binary, multilevel, and continuous distributions (although, in practice, computation for continuous distributions is facilitated by binning the variables into a number of discrete levels), for any number of treatment groups, and for multivariate distributions (i.e., vectorized covariate values $\vec{C}$) across treatment groups. Second, it allows for the identification of specific levels of C or T—and, moreover, specific combinations of C and T—that contribute most to imbalances across groups or treatments in relation to others. And third, it is sensitive to high order imbalances (e.g., differences in variability, skewness, bimodality, etc.) in addition to location shifts.

A brief introduction to information theory and the JSD is presented in the next section of this report (Section 2). Properties of the JSD as a measure of covariate imbalance are discussed in Section 3. Examples are presented in Section 4. We conclude with a brief summary (Section 5).

2. Information Theory and the JSD

The JSD is an information-theoretic measure of dissimilarity among two or more probability distributions [3]. It is derived from relative entropy (or Kullback–Leibler divergence) [4] and is therefore related to mutual information [5] (pp. 18–21). These measures are fundamentally tied to Shannon's entropy [6]. The goal of this section is to describe the JSD in intuitive terms, beginning with the definition of entropy.

2.1. Entropy

Let X be a discrete random variable which takes on values $x_i \in \{x_1, x_2, \ldots, x_M\}$. Let the probability distribution of X be denoted as $f(X)$. The entropy of X, denoted $H(X)$, is a measure of the uncertainty of the outcome of X and is defined as:

$$H(X) = E(-\log_2 f(X)) = -\sum_{i=1}^{M} f(x_i) \log_2 f(x_i). \quad (1)$$

The base of the logarithm is arbitrary. Log base two is often used, giving entropy units of bits (binary digits).

One approach to understanding the concept of entropy is to explore its relationship to the average number of bits (e.g., 0 s and 1 s) required to efficiently encode a sequence of outcomes of the random variable. Consider, for example, the case where the sample space is $\{A, B, C, D\}$ with corresponding probabilities $f(X) = \{0.25, 0.125, 0.5, 0.125\}$. With four possible outcomes, it may be tempting to encode a single outcome using two bits, e.g., $00 \rightarrow A$, $01 \rightarrow B$, $10 \rightarrow C$, and $11 \rightarrow D$. A more efficient mapping is $0 \rightarrow C$, $10 \rightarrow A$, $110 \rightarrow B$, and $111 \rightarrow D$. Since A, B, C, and D have probabilities of 0.25, 0.125, 0.5, and 0.125, respectively, and are encoded with 2, 3, 1, and 3 bits, respectively, the expected value of the number of bits required to transmit the outcome of X with this coding scheme is 0.25×2 bits + 0.125×3 bits + 0.5×1 bit + 0.125×3 bits = 1.75 bits.

Shannon demonstrated that $H(X)$ defines a limit beyond which codes cannot be made more efficient. Using either of the above coding schemes allows for the unambiguous encoding of a series of outcomes of X, but the second scheme is optimal in that the expected number of bits required to transmit the outcome of X is $H(X) = 1.75$ rather than two. To achieve (or to become arbitrarily close) to the efficiency specified by $H(X)$ may require a mapping that associates each code with a sequence of outcomes of X [5] (p. 104). For example, in the case of two possible outcomes A and B, with respective probabilities 2/3 and 1/3, a code that is more efficient than simply $0 \rightarrow A$ and $1 \rightarrow B$ is $0 \rightarrow AA$, $10 \rightarrow AB$, $110 \rightarrow BA$, and $111 \rightarrow BB$. The length of the inefficient code required to indicate the outcome is 1 bit, but the average length of the more efficient code, per outcome, is 0.9444 bits (compared to the ideal of $H(X) = 0.9183$ bits).

2.2. Joint and Conditional Entropy

The joint entropy $H(X, Y)$ of two random variables is a natural extension of the concept of entropy for a single random variable:

$$H(X,Y) = E(-\log_2 f(X,Y)) = -\sum_{i=1}^{M}\sum_{j=1}^{N} f\left(x_i, y_j\right)\log_2 f\left(x_i, y_j\right). \tag{2}$$

Similar to that described above for a single random varible, the joint entropy defines the lower limit of the average number of bits required to encode the observations from the joint distribution.

Conditional entropy, denoted $H(X|Y)$, is a measure of residual uncertainty in X, given the observation of some other random variable Y. It is defined as:

$$H(X|Y) = E(-\log_2 f(X|Y)) = -\sum_{i=1}^{M}\sum_{j=1}^{N} f\left(x_i, y_j\right)\ \log_2 f(x_i|y_j). \tag{3}$$

Conditional entropy is also equal to the difference between joint and marginal entropies, i.e., $H(X|Y) = H(X,Y) - H(Y)$. In this sense, conditional entropy represents the number of bits needed to encode X after the value of Y is observed. Joint and conditional entropy naturally extend to distributions that are defined across three or more random variables (we omit these equations for the purposes of this discussion).

2.3. Mutual Information

The mutual information between the random variables X and Y, denoted $I(X;Y)$, is the expected value of the amount of information that knowledge of the outcome of Y provides about the outcome of X. Mutual information is symmetric with respect to X and Y, and is a function of both the variables' marginal entropies and their joint entropy:

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= I(Y;X). \end{aligned} \quad , \tag{4}$$

2.4. Relative Entropy

Relative entropy is an information-theoretic measure expressing the divergence from a given probability distribution $f(X)$ to a reference (or target) distribution $g(X)$. It is defined as

$$D(g(X) \parallel f(X)) = E_{g(X)}\left[-\log_2 \frac{f(X)}{g(X)}\right] = \sum_{i=1}^{M} g(x_i)\log_2 \frac{g(x_i)}{f(x_i)}. \tag{5}$$

The relative entropy is interpreted as the number of bits required to "correct" the probabilities in the distribution f so that they match those of the reference distribution g (under an optimal coding scheme) [5] (p. 18).

Since the expectation in Equation (5) is taken with respect to the target distribution $g(X)$, the relative entropy function is asymmetric, i.e., it is not necessarily the case that $D(g(X) \parallel f(X)) = D(f(X) \parallel g(X))$. Given this asymmetry, it is not a suitable candidate for a measure of covariate balance among groups: the divergence between two groups would depend upon which group is taken to be the Reference group. Jeffrey's divergence (J) is a symmetric version of relative entropy, defined as $J(g(x); f(x)) = D(g(x)\big\|f(x)) + D(f(x)\big\|g(x))$ [7]. One reason why it is not a suitable candidate for the task of assessing covariate balance among groups is that there may be more than two groups.

2.5. Jensen–Shannon Divergence (JSD)

The JSD is a modified version of relative entropy, that addresses the asymmetry problem described above by expressing divergences with respect to a common distribution $\widetilde{f}(X)$. Assume that there are N distributions of X: $f_1(X)$, $f_2(X), \ldots, f_N(X)$. The common distribution is taken as the (unweighted) mean of the component densities:

$$\widetilde{f}(x) = \frac{1}{N}\sum_{k=1}^{N} f_k(x). \tag{6}$$

The JSD of the set of distributions $f_k(X)$ is defined as the average relative entropy from the common distribution $\widetilde{f}(X)$ to the specific distributions $f_k(X)$:

$$\text{JSD} = \frac{1}{N}\sum_{k=1}^{N} D\Big(f_k(X) \parallel \widetilde{f}(X)\Big). \tag{7}$$

2.6. The JSD of Covariate Distributions Across Treatment Groups

Equations (6) and (7) can be modified to calculate the JSD for a set of N treatment groups. We replace the continuous random variable X with the discrete covariate random variable C. Similary, we replace the probability density function f with the probability mass function p. Assuming that C can assume M values, we have, for $i = 1, \cdots, M$:

$$\widetilde{p}(c_i) = \frac{1}{N}\sum_{k=1}^{N} p_k(c_i), \tag{8}$$

and

$$\text{JSD} = \frac{1}{N}\sum_{k=1}^{N} D(p_k(C) \parallel \widetilde{p}(C)) = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{M} p_k(c_i)\log_2\left(\frac{p_k(c_i)}{\frac{1}{N}\sum_{k=1}^{N} p_k(c_i)}\right). \tag{9}$$

3. Properties of the JSD

The JSD is non-negative and is equal to zero when the covariate distributions are identical for all treatment groups. It is interpreted as the average relative entropy from the common covariate distribution, $\widetilde{f}(C)$, to the group-specific distributions. As noted in the Introduction, the JSD can be applied to binary random variables, categorical random variables, or continuous random variables.

Being defined additively in terms of units of information, the JSD is decomposable. One may calculate the JSD across all the treatment groups or determine the contribution of a subset of groups to the overall JSD. Similarly, specific levels of the covariate(s) of interest may be examined to identify regions of the covariate space exhibiting the greatest degree of imbalance across groups. Furthermore, contributions of individual treatment/covariate combinations to the overall JSD can be studied and compared. The decomposability of the JSD is illustrated in Section 4.

As a function of the densities themselves (and not their moments), the JSD allows for the evaluation of balance in a manner that does not assume that continuous densities belong to any particular family of distributions. It is sensitive to shape discrepancies among groups. In contrast, the standardized difference score converges to zero (with increasing group sample sizes) whenever the means of the two samples are equal (see Figure 1).

In practice, computation of the JSD using observational data can be difficult for continuous densities, especially mixture distributions [2]. Our approach relies on the binning of continuous variables (as is done with histograms). When small numbers of categories are used, this simplification can mask subtle features of group-specific probability densities. A further limitation of the JSD is that density estimates for categorical variables are increasingly variable among small samples.

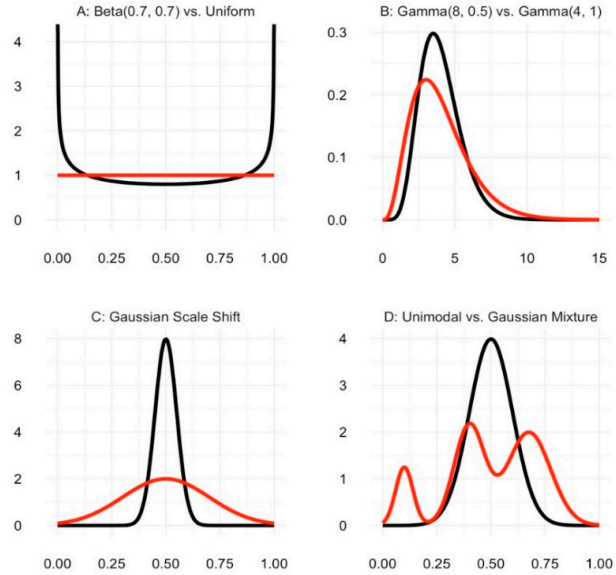

Figure 1. Four pairs of continuous distributions, each of which has a standardized difference score equal to zero.

4. Applications

Table 1 summarizes findings from 93,583 outpatients in the Cleveland Clinic Health System who had a lipid panel drawn between 2007 and 2010 (first visit meeting these criteria). The patients are partitioned into three treatment groups: Disadvantaged (age < 80 years and living in a census tract that is in the top 25% of all tracts in the United States with respect to the Area Deprivation Index [8]), Elderly (not living in a disadvantaged neighborhood per the above definition but aged 80 or older), and Reference (neither disadvantaged nor elderly). The covariate is baseline diabetes state defined by blood sugars < 109 mg/dL, 109–125 mg/dL, and > 125 mg/dL. A stand-alone R package for implementing the JSD computations illustrated in this section is provided at http://github.com/jarrod-dalton/jsd, and the code used for this section is given in the Appendix A.

Table 1. Number of individuals in three treatment groups (Disadvantaged, Elderly, Reference) and three covariate groups (defined by blood sugar ranges).

Glucose	Disadvantaged	Elderly	Reference
<109	7191	3637	64,265
109–125	1025	835	7298
>125	1715	685	6932

Table 2 presents the probability distributions of glucose levels within each treatment group. The average of these distributions, i.e., the common distribution, $\widetilde{f}(C)$, is shown in the final column.

Table 2. Probability distributions of glucose levels within each treatment group (Disadvantaged, Elderly, Reference). The common distribution, $\widetilde{f}(C)$, is shown in the final column.

Glucose	Disadvantaged	Elderly	Reference	$\widetilde{f}(C)$
<109	0.724	0.705	0.819	0.749
109–125	0.103	0.162	0.093	0.119
>125	0.173	0.133	0.088	0.131

Table 3 presents contributions of individual cells, the three treatment groups, and the three covariate groups to the overall JSD, which is 0.0144 bits. This is the average of the relative entropies

from the common distribution to the treatment group-specific distributions. Given three treatment groups, the maximum possible JSD is $\log_2(3) = 1.5850$ bits.

Table 3. Contributions of individual cells, treatment groups, and levels of the covariate to the overall JSD (in units of bits).

Glucose	Disadvantaged	Elderly	Reference	Total
<109	−0.0119	−0.0206	0.0349	0.0023
109–125	−0.0072	0.0237	−0.0112	0.0053
>125	0.0228	0.0008	−0.0168	0.0067
Total	0.0036	0.0039	0.0068	0.0144 *

* Note: row/column sums do not equal 0.0144 due to rounding error.

The Reference group is the largest treatment group contributor to the JSD, and the Glucose > 125 category is the largest covariate group contributor to the JSD. Moreover, by considering the absolute values of the individual cell components, we conclude that the largest contributor to the JSD is from individuals in the Reference group with serum glucose values less than 109 mg/dL.

A problem with using any method to quantify covariate imbalance among treatment groups is that there is no obvious point that defines an acceptable amount of imbalance [9]. For the current example, the JSD value of 0.0144 bits is small relative to its maximal possible value of 1.5850 bits, but it is clear from Table 2 that individuals in the Reference group tend to have lower blood sugars than individuals in the other two treatment groups. An important factor in deciding what constitutes acceptable balance is the potential of the covariate to affect the outcome [10].

In order to further examine differences between the JSD and standardized difference scores, we consider the case in which there are two treatment groups with normally distributed covariates. Figure 2 plots the JSD as a function of the standardized difference score, when the standard deviation of one of the two distributions is one and the standard deviation of the other distribution is either one (plotted in black), two (plotted in blue), or three (plotted in red). Since there are two treatment groups, the JSD curves asymptote at one bit (since $\log_2(2) = 1$). The standardized difference score curves, on the other hand, are unbounded in the positive direction. As expected, both the JSDs and the standard difference scores increase as the two distributions diverge. The plot also illustrates the point made in Section 3 that the JSD, but not the standardized difference score, is sensitive to differences between the standard deviations of the two distributions when the means of two distributions are identical.

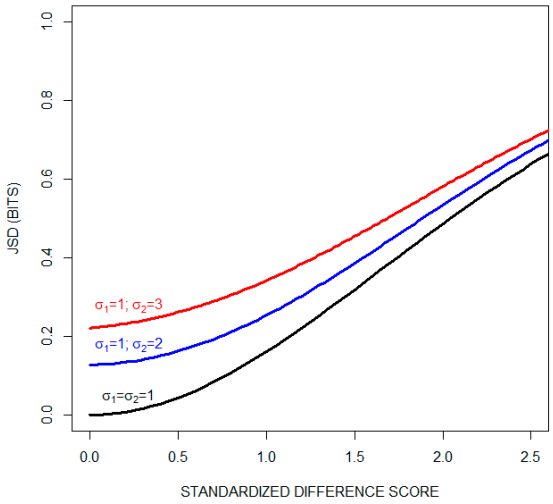

Figure 2. The JSD as a function of the standardized difference score when there are two treatment groups with normally distributed covariates. Three cases are shown: the standard deviation of one of the two distributions is set equal to one, while the standard deviation of the second distribution is set to equal either one (black curve), two (blue curve), or three (red curve).

5. Summary

We propose that the JSD be used to assess treatment group balance on known potential confounding variables in comparative clinical studies. This information-theoretic measure is equal to the average relative entropy between the covariate distributions for each treatment group and a common distribution, defined as the average of the individual distributions. Advantages of the JSD over alternative measures of treatment group balance include its sensitivity to the shape of distributions and its insensitivity to sample size. The JSD is applicable to both categorical and continuous random variables. Moreover, the JSD is decomposable, allowing for comparisons among specific levels of covariates of interest.

Author Contributions: Conceptualization, J.E.D. and W.A.B.; writing—original draft preparation, J.E.D. and W.A.B.; writing—review and editing, J.E.D. and W.A.B.; funding acquisition, J.E.D; software, J.E.D., N.I.K. and W.A.B. All authors have read and agreed to the published version of the manuscript.

Funding: Research reported in this publication was supported by The National Institute on Aging of the National Institutes of Health under award number R01AG055480. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Acknowledgments: The authors are grateful to the reviewers for their helpful suggestions.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

The jsd R package can be found at https://github.com/jarrod-dalton/jsd. The package can be installed using the following command (run the first command if the remotes package has not already been installed):

```
#install.packages("remotes")
remotes::install_github("jarrod-dalton/jsd")
```

The library is then loaded as follows:

```
library(jsd)
```

The glucose dataset contains the data used for the example in Section 4. Note that the actual glucose values are simulated.

```
data(glucose)
head(glucose)
##        cohort    glucose
## 1 Reference   79.16898
## 2    Elderly  124.66537
## 3    Elderly   95.61820
## 4 Reference   90.11065
## 5 Reference  105.50799
## 6 Reference   89.35060
```

There is a helper function in the package, called chop, which will convert numeric variables into categorical variables. See help(chop) for details. Here, we convert the glucose variable into a categorical variable with 3 levels:

```
glucose$glucose_cat <- chop(glucose$glucose, cuts = c(0, 109, 125, Inf))
```

The jsd_balance function is then used to compute the JSD measures. The output of the jsd_balance function contains the cell contributions to the JSD, marginal contributions of each treatment group to the JSD, marginal contributions of each covariate level to the JSD and the overall JSD value (see Table 3 for details). The first argument to the jsd_balance function is a formula in which the group variable is on the left hand side of the tilde and the covariate(s) is/are on the right hand side of the tilde (separated by '+' – see help(jsd_balance) for details and examples):

```
jsd_balance(cohort ~ glucose_cat, data = glucose)
## $glucose_cat
## $freqs
##                cohort
## glucose_cat Disadvantaged Elderly Reference
##    [  0,109)         7191    3637     64265
##    [109,125)         1025     835      7298
##    [125,Inf]         1715     685      6932
##
## $cell_contribs
##                cohort
## glucose_cat  Disadvantaged       Elderly      Reference
##    [  0,109)  −0.0119399530 −0.0205710889  0.0348528681
##    [109,125)  −0.0072178122  0.0237387920 −0.0111726329
##    [125,Inf]   0.0227710815  0.0007506513 −0.0168364804
##
## $group_contribs
## Disadvantaged       Elderly      Reference
##    0.003613316   0.003918354   0.006843755
##
## $cov_contribs
##      [  0,109)     [109,125)       [125,Inf]
## 0.002341826  0.005348347  0.006685252
##
## $jsd
## [1] 0.01437543
##
## attr(,"class")
## [1] "jsd_balance"
```

References

1. Friedman, L.M.; Furberg, C.D.; DeMets, D.L.; Reboussin, D.M.; Granger, C.B. *Fundamentals of Clinical Trials*, 5th ed.; Springer: New York, NY, USA, 2010.
2. Contreras-Reyes, J.E.; Cortés, D.D. Bounds on Rényi and Shannon Entropies for Finite Mixtures of Multivariate Skew-Normal Distributions: Application to Swordfish (Xiphias gladius Linnaeus). *Entropy* **2016**, *18*, 382. [CrossRef]
3. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory* **1991**, *37*, 145–151. [CrossRef]
4. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *2*, 79–86. [CrossRef]
5. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2012.
6. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
7. Nielsen, F. On the Jensen-Shannon symmetrization of distances relying on abstract means. *Entropy* **2019**, *21*, 485. [CrossRef]
8. Kind, A.J.H.; Buckingham, W.R. Making Neighborhood-Disadvantage Metrics Accessible—The Neighborhood Atlas. *N. Engl. J. Med.* **2018**, *378*, 2456–2458. [CrossRef] [PubMed]
9. Austin, P.C. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Commun. Stat.-Simul. Comput.* **2009**, *38*, 1228–1234. [CrossRef]
10. Ho, D.E.; Imai, K.; King, G.; Stuart, E.A. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Anal.* **2007**, *15*, 199–236. [CrossRef]