SANDIA REPORT
SAND2022-15094
Printed September 2022

Sandia National Laboratories

Entropy and its Relationship with Statistics

R. B. Lehoucq, C. D. Mayer, J. D. Tucker

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185
Livermore, California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@osti.gov
Online ordering: http://www.osti.gov/scitech

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.gov
Online order: https://classic.ntis.gov/help/order-methods

ABSTRACT

The purpose of our report is to discuss the notion of entropy and its relationship with statistics. Our goal is to provide a manner in which you can think about entropy, its central role within information theory and relationship with statistics. We review various relationships between information theory and statistics—nearly all are well-known but unfortunately are often not recognized.

Entropy quantifies the "average amount of surprise" in a random variable and lies at the heart of information theory, which studies the transmission, processing, extraction, and utilization of information. For us, data is information. What is the distinction between information theory and statistics? Information theorists work with probability distributions. Instead, statisticians work with samples. In so many words, information theory using samples is the practice of statistics.

Acknowledgements

We thank Danny Dunlavy, Carlos Llosa, Oscar Lopez, Arvind Prasadan, Gary Saavedra, Jeremy Wendt for helpful discussions along the way.

Our report was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Adminstration under contract DE-NA0003525.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The purpose of our report is to discuss the notion of entropy and its relationship with statistics. Our goal is to provide a manner in which you can think about entropy, its central role within information theory and overlap with statistics. We review various relationships between information theory and statistics—nearly all are well-known but unfortunately are often not recognized. A good example is that the log likelihood ratio converges to the relative entropy as the number of samples increases.

Entropy quantifies the "average amount of surprise" in a random variable. Consider the three distributions shown in the leftmost graphic of Figure 1-1. If you randomly sample from each of the distributions and create a histogram as depicted in the rightmost graphic of Figure 1-1, which distribution (or histogram) surprises you the most? The least? And why? Our question assumes that the notion of surprise is well-defined—the purpose of an entropy definition is to quantify average surprise. Because the three distributions range from flat to more concentrated, the samples, will tend to cluster as the spread decreases. Hence there is less surprise among the samples associated with less spread. Equivalently, as the distribution concentrates about some point, then sampling the distribution results in less spread, hence less surprise.

Entropy lies at the heart of information theory, which studies the transmission, processing, extraction, and utilization of information. For us, data is information. What is the distinction between information theory and statistics? The two graphics in Figure 1-1 provide the answer. Information theorists work with probability distributions. Instead, statisticians work with samples. Indeed, statistics is the practice of describing the population (i.e., the probability distribution) in terms of a random sample (e.g., the histogram). In so many words, information theory using samples is the practice of statistics.

The origins of information theory lie in statistical mechanics, which attempts to connect the movement of molecules to observable physical phenomena. Entropy, as introduced by Boltzmann, is a measure of molecular disorder. Shannon observed that entropy is a general principle with application outside of physics and could be applied to the transmission of digital bits, or limits on the transmission of bits. Both for Boltzmann and Shannon, entropy quantifies the amount of information in a collection of objects (e.g., molecules and bits). And so information theory provides a framework to understand what is available in the data. For Boltzmann and Shannon, the populations are the molecular trajectories and the (uncompressed) signal, respectively. Hence information theory and statistics are intimately related.

Chapter 2 introduces entropy via definitions and examples. Chapter 3 reviews "p-value" based hypothesis testing and the relationship with a max-entropy distribution. Chapter 4 reviews the relationship between Neyman-Pearson hypothesis testing and relative entropy. Chapter 5 reviews Fisher Information and its uses within mathematical statistics. Chapter 6 reviews the relationship

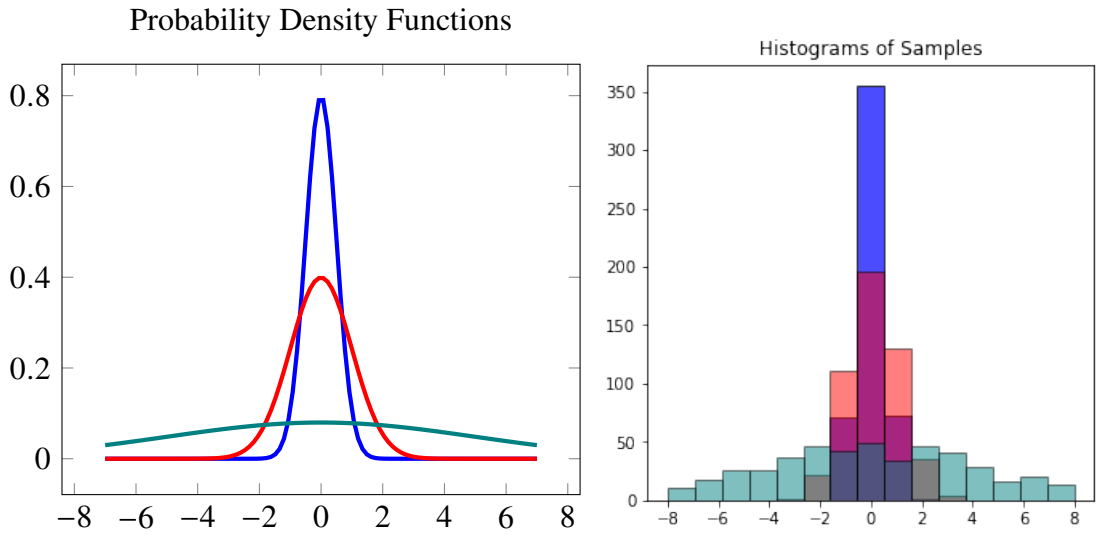


Figure 1-1. Left: Probability density functions of three probability distributions with different spread. Right: Histograms of 500 samples from the three distributions sorted into 15 bins.

between random variables including the concept of a statistical copula. Our report concludes with a discussion of the Wasserstein metric in §7 including a comparison with relative entropy.

1.1. Notes and References

Our report makes no claims that the material presented is original. Instead, our report documents well-known relationships in a conversational tone as we struggled to clarify them and apply to our work. We have few references, a limitation we hope to address in revisions to our report. One obvious reference is the excellent book [Cover and Thomas, 2012], in particular Chapter 7 reviewing relationships between information theory and statistics. Our report lists several references helpful to us along the way. If you the reader are compelled to search the internet to better understand the topic at hand, then at some level our report has met one of its goals. We welcome pointers to other useful references, suggestions for additional topics and improving our discussion, in particular our inevitable errors.

2. ENTROPY

The (differential) entropy of a continuous random variable X distributed with respect to the probability density function (pdf) p is

$$h(X) = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx \tag{2.1a}$$

while the (discrete) entropy of a discrete random variable X distributed with respect to the probability mass function (pmf) p is

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)\,. \tag{2.1b}$$

where n denotes the number of elements in the support of p where $\log p$ denotes the logarithm of the function p. The case of the letter "h" denotes whether the entropy under discussion is differential or discrete, i.e., whether the random variable X is continuous or discrete. In a slight abuse of notation, p denotes the distribution for a continuous or discrete random variable given the context.

Entropy quantifies the "average amount of surprise" in a random variable. Let's parse the words "average" and "surprise". Both entropies can be rewritten as

$$-\mathbb{E} \log p(X) \tag{2.2}$$

where the context specifies whether the expectation is with respect to (pdf or pmf) p. And so the entropy of X is an average, namely the expectation of a function of X with respect to the probability density or mass function p.

Figure 2-1 shows a comparison of differential entropy for normally distributed random variables with different standard deviations first depicted in Figure 1-1. The entropy increases with the standard deviation σ. This formalizes our intuition that if we were to observe outcomes of the random variables in the figure, we would, on average, be surprised by observations of Z and less so by observations of X.

Why the function logarithm? To quantify the amount of surprise $\mathcal{I}$ of an event E, it is reasonable to require the following properties:

1. $\mathcal{I}(E_1) > \mathcal{I}(E_2)$ when $p(E_1) < p(E_2)$, i.e., $\mathcal{I}$ is monotonically decreasing; a larger probability of an event is associated with a decrease in the surprise.
2. $\mathcal{I}(E) = 0$ when the event E always occurs, for instance $p(X = x_i) = \delta_{i,j}$ where $\delta_{i,j}$ is the Kronecker delta function, which is zero except at the one event x_i.

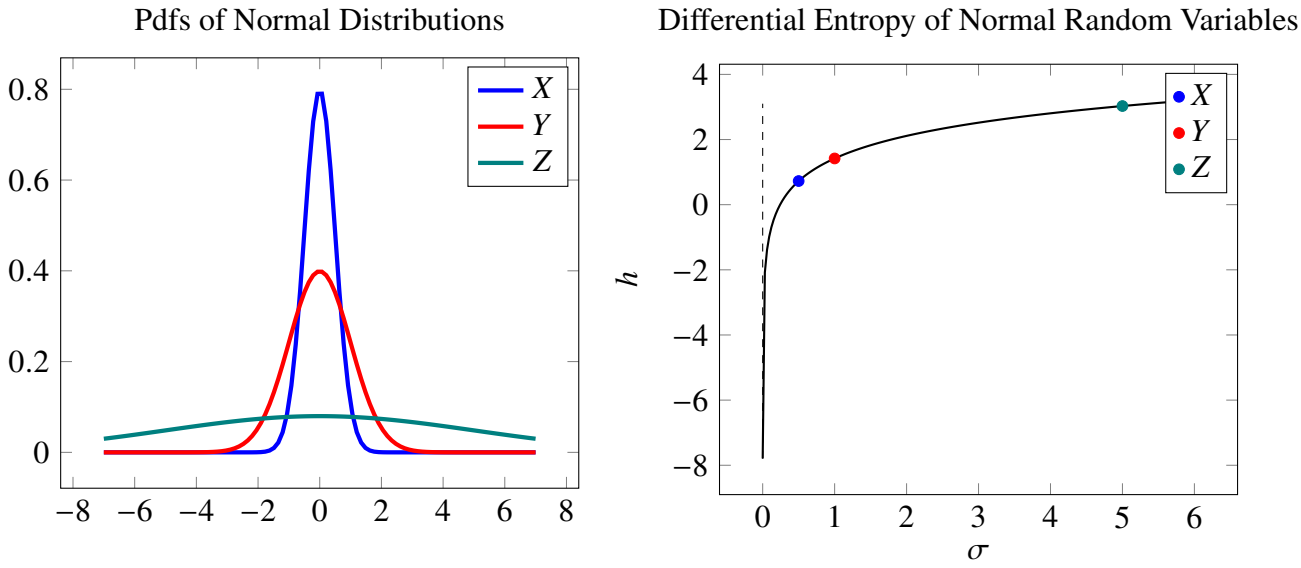


Figure 2-1. Left: The pdfs for three normally distributed random variables with different standard deviations first depicted in Figure 1-1. Right: The differential entropy of a normally distributed random variable as a function of σ. For X with $\sigma = \frac{1}{2}$, $h(X) \approx 0.725791$. For Y with $\sigma = 1$, $h(Y) \approx 1.41894$. For Z with $\sigma = 5$, $h(Z) \approx 3.02838$. As σ increases, so does the expected surprise.

3. $\mathcal{I}(E_1 \text{ and } E_2) = \mathcal{I}(E_1) + \mathcal{I}(E_2)$ when E_1 and E_2 are independent events.

The choice $\mathcal{I}(E) = -\log p(E)$ satisfies these properties (and in fact is the only function possible) and its average is the entropy H.

An important distinction between h and H is that the former takes on values in the interval $(-\infty, \infty)$ while the latter is non-negative and bounded by $\log n$. In the next section, we consider the relationship between h and H for an important class of probability measures.

2.1. Maximum entropy

The entropy of a random variable varies with choice of parameter. Consider, for example, the entropy of a Bernoulli random variable with probability of success p, as shown in Figure 2-2. A Bernoulli random variable models a coin flip, where p is the probability of the coin displaying a head. Note that if $p = 0$ or 1, there is no surprise in the outcome. The entropy is maximized when $p = \frac{1}{2}$.

A related concept is that of a maximum entropy distribution. Instead of maximizing the entropy of a random variable by varying its parameters, we seek the distribution of maximum entropy within a class of distributions. Such distribution is the basis for the principle of maximum entropy: if nothing is known about a distribution, then the distribution with the largest entropy is the least-informative.

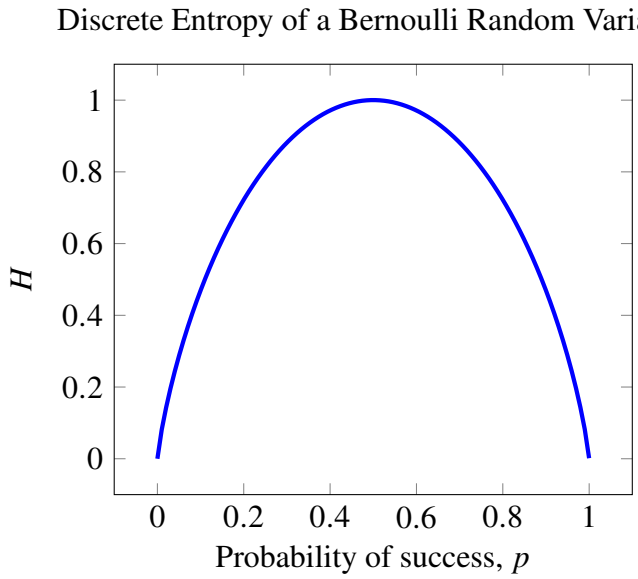


Figure 2-2. The entropy of a Bernoulli random variable.

For instance, consider the class of distributions continuous over the interval $(0, a)$ for positive a. Is there a maximum entropy distribution? Yes—the uniform distribution over $(0, a)$, i.e., $(1/a)\,\mathbb{1}_{(0,a)}$ where

$$\mathbb{1}_{(0,a)}(x) = \begin{cases} 1 & x \in (0, a) \\ 0 & x \notin (0, a) \end{cases} \tag{2.3}$$

is an indicator function. The differential entropy of $(1/a)\,\mathbb{1}_{(0,a)}$ is easily computed to be

$$h\big((1/a)\,\mathbb{1}_{(0,a)}\big) = (1/a)\,h\big(\mathbb{1}_{(0,a)}\big) = \log a\,. \tag{2.4}$$

By varying a, the differential entropy varies over the interval $(-\infty, \infty)$. Intuitively, as $a \to 0^+$, the differential entropy decreases to $-\infty$ and the resulting uniform distribution is nearly a point mass and so contains little surprise. As $a \to \infty$, the differential entropy increases without bound and the resulting uniform distribution contains substantial surprise. We now glean the idea behind the principle of maximum entropy because the uniform distribution favors all outcomes equally, i.e., is least-informative.

The choice of $a = 1$ results in the standard uniform distribution, i.e.,

$$\mathbb{1}_{(0,1)}(x) = \begin{cases} 1 & x \in (0, 1) \\ 0 & x \notin (0, 1) \end{cases} \tag{2.5}$$

is involved in the universality of the uniform, also referred to as the probability integral transform. Both phrases encapsulate the significant conclusion that when F is the continuous cumulative distribution function (cdf) for a real-valued random variable X, then the distribution for the random

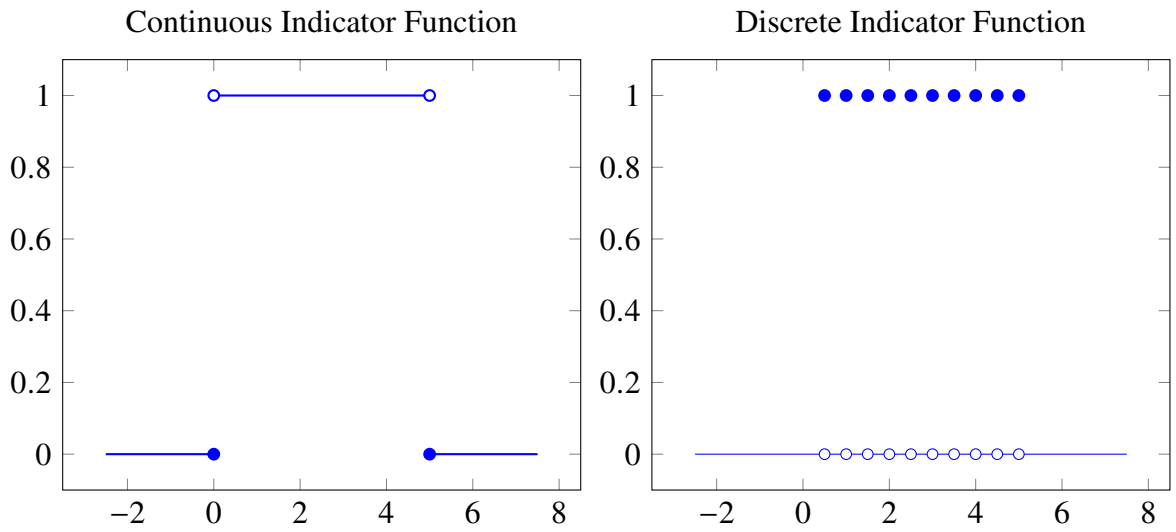


Figure 2-3. Left: The continuous indicator function $\mathbb{1}_{(0,5)}$. Right: The discrete indicator function $\mathbb{1}_{\{i(5/10)\}_{i=1}^{10}}$

variable $F(X)$ is (2.5). In other words, the distribution for the random variable $F(X)$ is the maximum entropy distribution among all continuous distributions supported over the unit interval and by (2.4) has value zero.

The maximum entropy pmf corresponding to n points distributed uniformly over $(0, a]$ (i.e., the uniform discrete distribution) $(1/n)\mathbb{1}_{\{i(a/n)\}_{i=1}^{n}}$ where

$$\mathbb{1}_{\{i(a/n)\}_{i=1}^{n}}(x) = \begin{cases} 1 & x \in \{i(a/n)\}_{i=1}^{n} \\ 0 & x \notin \{i(a/n)\}_{i=1}^{n} \end{cases} \tag{2.6}$$

with

$$H\big((1/n)\mathbb{1}_{\{i(a/n)\}_{i=1}^{n}}\big) = (1/n)\, H\big(\mathbb{1}_{\{i(a/n)\}_{i=1}^{n}}\big) = \log n\,.$$

Figure 2-3 shows an example of continuous and discrete indicator functions with $a = 5$ and $n = 10$.

A limiting relationship between the differential and discrete entropy is available. In terms of continuous and discrete uniform distributions, the limiting relationship holds for all n and so becomes an identity. When $a > 0$, a straight forward derivation shows that

$$H\big((1/n)\mathbb{1}_{\{i(a/n)\}_{i=1}^{n}}\big) + \log(a/n) = h\big((1/a)\mathbb{1}_{(0,a)}\big)\,. \tag{2.7}$$

The identity (2.7) explains that the entropies of the continuous and discrete uniform distributions over the interval $(0, a)$ and the n points $\{a/n, 2(a/n), \ldots, a\}$, respectively are related via the spacing a/n between points. In so many words, the differential entropy is the sum of the discrete entropy on the points and the spacing between the points. Note that the relationship is true for all n but we cannot take limit $n \to \infty$ since the lefthand side approaches $-\infty$ while the righthand side is independent of n.

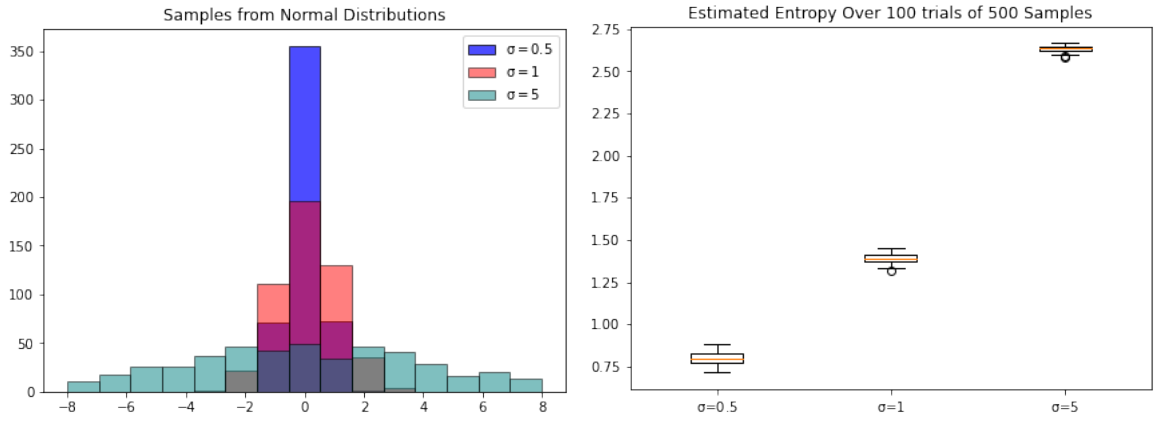


Figure 2-4. Approximating entropy using samples from the normal distributions in Figure 2-1. Left: Histograms of 500 samples from the distributions sorted into 15 bins. Right: Boxplots of the estimated entropy of the distributions over 100 trials of 500 samples each. The average entropy estimates for $\sigma = 0.5$, 1, and 5 were around 0.8002, 1.3914, and 2.6339, respectively.

2.2. Approximate entropy

Computing the entropy of a random variable requires knowledge of the probability density or mass function, quantities typically estimated from a sample so that the entropy is approximated. Because the entropy is an expectation and we assume the sample is random, Monte Carlo approximation is natural choice. The number of samples needed to achieve a prescribed level of approximation depends upon the problem at hand and the use of entropy. Our report will review several situations where the Monte Carlo approximation of entropy can by replaced by a standard statistical procedure.

Figure 2-4 shows histograms and estimated entropy based on samples taken from the distributions in Figure 2-1. The discrete entropy was estimated with 500 samples sorted into 15 bins with the probability of a bin given by normalizing the histogram counts by the sample size.

3. ENTROPY AND FISHER'S HYPOTHESIS TEST

Suppose we have a samples of some continuous distribution. Can we make a principled decision whether the samples are associated with a specific distribution f? Hypothesis testing represents such a principled approach and we'll review the relationship with entropy. In a sense we'll make precise, hypothesis testing is a decision procedure equivalent to determining whether the average surprise, i.e., the entropy, is maximized over the unit interval. Either hypothesis testing or checking whether entropy is maximized are decision procedures for associating the samples with the distribution f.

In the hypothesis testing approach advocated by Fisher [1925] a p-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis holds. A small p-value implies that the observed outcome is unlikely under the null hypothesis. Given a threshold, if the p-value is less than the threshold, then the null hypothesis is rejected.

Let's consider an example. Suppose we sample a distribution n times and denote each one by x_i. We decide that the null hypothesis is that the samples are drawn from a normal distribution with mean μ and standard deviation σ, in other words the null distribution is f. We'll use the test statistic

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)}{\frac{\sigma}{\sqrt{n}}} \tag{3.1}$$

i.e., the z-test to determine whether the n samples are distributed with mean μ and standard deviation σ. The law of large numbers implies that the z-test is distributed with respect to the standard normal distribution when the distribution for the samples has finite variance. Hence the p-value is computed by determining the probability that the null distribution, in this case the standard normal distribution, achieves a value at least as extreme of the test statistic. In other words, as the magnitude of the test statistic increases, the p-value decreases since it represents the area under the tails of the standard normal distribution.

This example illustrates that the p-value is a function of the z-test (the chosen test statistic), which summarizes the sample. Both the p-value and test statistic are random variables since they are functions of the random sample. The probability distribution for the test statistic is referred to as the sampling distribution and must be calculable under the null hypothesis. In the example, the law of large numbers implies that the z-test is distributed with respect to the standard normal distribution as the number of samples increases.

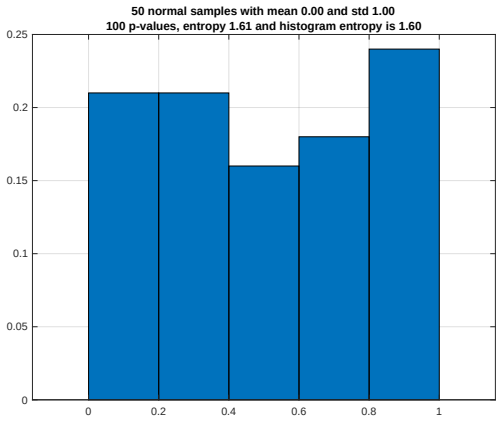
An important observation is that when the null hypothesis is true and the underlying random variable is continuous, then the probability distribution of the p-value is uniform on the unit interval because the z-test transforms a standard normal random variable to a uniform random variable;

see the discussion following (2.5). This is well understood but typically not recognized and demystifies the somewhat subtle notion of a p-value; see e.g., Murdoch et al. [2008] for an insightful discussion. The differential entropy of the p-value distribution is zero by (2.4) under the assumption that the null is true. Hence, if the null is false, then the entropy of the p-value distribution is negative.

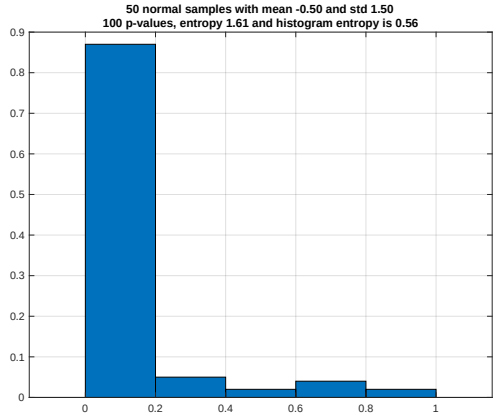
Recall that the null hypothesis is rejected if the p-value is smaller than a prescribed (positive) threshold. For instance, when the threshold is set at one percent, we'll incorrectly reject the null hypothesis (when it is true) with probability one-hundredth. This is an unavoidable error and occurs because there are numerous subsets of the unit interval with area less than one-hundredth—area under the uniform distribution. In contrast, the entropy requires the distribution for the random variable transformed by test statistic random variable—there is no error. The entropy is either negative or zero, however, a threshold to determine whether the entropy is "sufficiently" close to zero can be determined. If we modify the uniform distribution on subsets of the unit interval with area less than the threshold, the entropy of the modified distribution is a negative number within a threshold of zero.

In practice, however, we have samples of the random variable so that sampling error plays a role. Care must be taken to select a threshold that is on the same order or larger than sampling error. For example, we assumed that the sample size was sufficiently large so that the standard normal approximation is a good approximation for the z-test. We can check whether the approximation to the differential entropy approximated via the samples x_i is "sufficiently" negative or compute a p-value. The latter is typically more expedient than approximating the entropy.

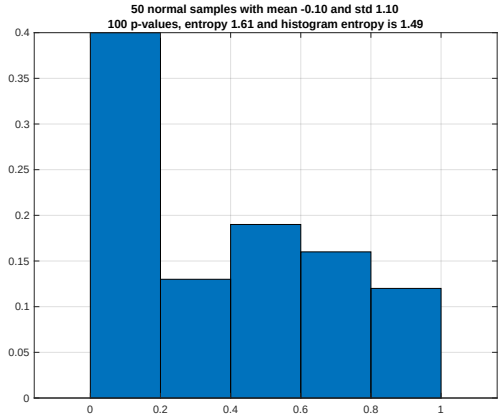
Figure 3-1 displays the results of four experiments with different sets of parameters used for the samples drawn from a normal distribution. The equality between the differential and discrete entropies (2.7) for a uniform distribution suggests that a histogram generated by sampling the test statistic is a simple approximation to the p-value distribution and its entropy is easily calculated using (2.1b). The four plots support the assertion that only when the samples are drawn from the standard normal distribution ($\mu = 0$ and standard deviation $\sigma = 1$) is the p-value distribution (approximated by the histogram) of maximum entropy.



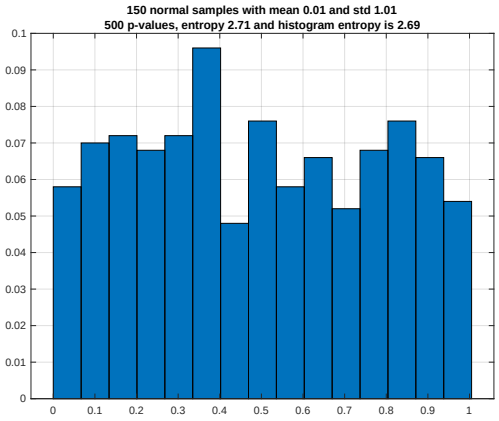
(a) $(\mu, \sigma) = (0, 1)$ **normal sampling and the p-value entropy is** 1.6**.**



(b) $(\mu, \sigma) = (-.5, 1.5)$ **normal sampling and the p-value entropy is** $.56$**.**



(c) $(\mu, \sigma) = (.1, 1.1)$ **normal sampling and the p-value entropy is** 1.49**.**



(d) $(\mu, \sigma) = (.01, 1.01)$ **normal sampling and the p-value entropy is** 2.69**.**

Figure 3-1. Investigating the relationship between entropy and the uniform distribution. Plots (a)-(c) depict a histogram of 100 **p-values each of which is a z-test using** 50 **samples drawn from a normal distribution. A histogram approximation to the p-value entropy can then be compared to the entropy of the uniform pmf or** $\log 5 \approx 1.61$**. Each histogram has 5 bins. Plot (d) depicts the histogram of** 500 **p-values each of which is a z-test using** 150 **samples drawn from a normal distribution. The p-value entropy can then be compared to the entropy of the uniform pmf or** $\log 15 \approx 2.71$**. The histogram has 15 bins.**

4. RELATIVE ENTROPY AND NEYMAN-PEARSON HYPOTHESIS TESTING

The KL divergence, or relative entropy [Kullback and Leibler, 1951] of the probability density of q from p is defined to be

$$\kappa(p,q) \coloneqq \int p(x) \log \frac{p(x)}{q(x)} dx \geqslant 0 \text{ with equality only when } p = q \tag{4.1a}$$

where $q(x) = 0$ implies $p(x) = 0$. If we recall that the log likelihood is given by the expression $\log(p/q)$, then

$$\kappa(p,q) = \mathbb{E}_p \log \frac{p}{q}\,. \tag{4.1b}$$

In words, the relative entropy is the average, or expectation, of the log likelihood with respect to the distribution p. Relative entropy is an example of an f-divergence that quantifies the difference between two distributions.

We also remark that $\kappa(p,q)$ is a pre-metric and not a metric since $\kappa(p,q)$ does not in general equal $\kappa(q,p)$. An important conclusion is that relative entropy is not a distance between p and q. However, relative entropy does distinguish between the distributions p and q if properly understood.

4.1. Relative entropy and hypothesis testing

Relative entropy is an idealization because knowledge of p and q is assumed. In practice, we have samples of p or q and want a principled procedure to select p or q. Can we decide in a statistically defensible manner whether the samples are drawn from p or q? One approach is to approximate the distributions p and q from the samples and then compute the relative entropy. Another approach is to use hypothesis testing, which allows us to work with the samples directly. This is an advantage since this does not require the distributions p and q or their approximation. An important consideration of Neyman-Pearson [Neyman and Pearson, 1933] hypothesis testing is the relationship between sample size and the probabilities of false negative and false positives. Quantifying the size of these two probabilities can be addressed by considering the relative entropy between the underlying distributions. The smaller the relative entropy, the larger the sample size needed. Relative entropy represents the best possible false positive and false negative rate achievable.

4.2. Hypothesis testing

Suppose we assume that $X_1, X_2, \ldots, X_n$ are distributed with respect to the distribution f. Consider the two simple hypotheses

$$\begin{aligned} H_p &: f = p \\ H_q &: f = q\,. \end{aligned} \tag{4.2}$$

We will denote H_p and H_q the null and alternate hypothesis, respectively. Hypothesis testing is a principled manner in which to accept or reject that the samples $X_1, X_2, \ldots, X_n$ are drawn from p. There are two primary errors

- Rejecting the null hypothesis H_p when it is true. This is deemed a type I error or a false positive. The probability of a type I error is the significance level of the test and is typically denoted by α.
- Accepting the alternate hypothesis H_q when it is false. This leads to a type II error or a false negative. The probability of a type II error is typically denoted by β.
- The power of the test is the probability of rejecting H_p when it is false and is equal to $1-\beta$.

When the significance level α is prescribed, the log likelihood ratio, i.e., the test,

$$\Lambda_n = \log \frac{p(X_1)\cdots p(X_n)}{q(X_1)\cdots q(X_n)} = \sum_{j=1}^{n} \log \frac{p(X_j)}{q(X_j)} \tag{4.3}$$

is well-known to be the optimal test with respect to the significance level α where we have assumed that $X_1, X_2, \ldots, X_n$ are independent and identically distributed. When $p(X_j) > q(X_j)$ then the sample X_j is more likely to be associated with p since the ratio of $p(X_j)$ to $q(X_j)$ exceeds one so that the logarithm of the ratio is positive. Hence depending upon the sign of Λ_n, we'll accept or reject the null hypothesis.

If $f = p$, then the weak law of large numbers implies that the estimator

$$\hat{\Lambda}_n \coloneqq \frac{1}{n}\Lambda_n \to \kappa(p,q) \text{ in probability as } n \to \infty\,. \tag{4.4}$$

In words, the sample average of the random variables $\log\big(p(X_j)/q(X_j)\big)$ converges in probability[1] to the KL divergence of q from p. Recall that (4.1b) states that $\kappa(p,q)$ represents the average of the likelihood ratio. And so the weak law of large numbers asserts that with large probability the sample average $\hat{\Lambda}_n$ approximates the average of the likelihood ratio as the number of samples increases.

[1] More precisely, the weak law implies that for all positive ϵ

$$\lim_{n\to\infty} \mathbb{P}\big(|\hat{\Lambda}_n - \kappa(p,q)| > \epsilon\big) = 0$$

Because $\kappa(p,q)$ is zero only when $p = q$, the probability (the significance level) that the samples are distributed with respect to p is

$$\alpha_n = \mathbb{P}(\hat{\Lambda}_n < 0 \,|\, H_p \text{ is true}) \,. \tag{4.5}$$

Let's unpack this a bit. First,

$$\hat{\Lambda}_n < 0 \text{ if and only if } e^{\Lambda_n/n} < 1 \,. \tag{4.6a}$$

We then have by (4.3) that

$$\begin{aligned} \frac{\Lambda_n}{n} &= \log\left(\frac{p(X_1)\cdots p(X_n)}{q(X_1)\cdots q(X_n)}\right)^{1/n} \text{ so that when} \\ e^{\Lambda_n/n} &= \left(\frac{p(X_1)\cdots p(X_n)}{q(X_1)\cdots q(X_n)}\right)^{1/n} < 1 \\ &= \left(\frac{p(X_1)}{q(X_1)}\right)^{1/n} \cdots \left(\frac{p(X_n)}{q(X_n)}\right)^{1/n} < 1 \end{aligned} \tag{4.6b}$$

so that we may reexpress the type I error (4.5) as

$$\alpha_n = \mathbb{P}\left\{\left(\frac{p(X_1)}{q(X_1)}\right)^{1/n} \cdots \left(\frac{p(X_n)}{q(X_n)}\right)^{1/n} < 1 \,|\, H_p \text{ is true}\right\}. \tag{4.6c}$$

This indicates that a type I error occurs when the samples X_j are distributed according to p but

$$e^{\Lambda_n/n} < 1 \tag{4.6d}$$

holds.

For instance, consider the p and q in Figure 4-1a. A small probability event is when a sufficient number of the X_j's are within one standard deviation of μ_q so that (4.6d) holds. Because the two distributions in Figure 4-1a are so disparate, α_n is extremely small because the probability of such an event is extremely small. In contrast, Figure 4-2a depicts two distributions that are nearly the same so that n must be sufficiently large to achieve a prescribed level. However, in both cases, $\alpha_n > 0$, i.e., the significance level is positive so that false positives are inevitable. In fact, the limit (4.4) explains that as the number of samples increases without bound, the normalized log likelihood ratio converges to the relative entropy, a positive number unless $p = q$.

An important question is whether we can estimate α_n in terms of the number of samples. Let's once again reexpress the type I error (4.5) as

$$\alpha_n = \mathbb{P}\big(\kappa(p,q) - \hat{\Lambda}_n > \kappa(p,q) \,|\, H_p \text{ is true}\big) \,. \tag{4.6e}$$

We may now invoke Hoeffding's inequality to provide a non-asymptotic bound on the type I error

$$\alpha_n \leqslant \exp\big(-C\,n\,\kappa^2(p,q)\big) \to 0 \text{ as } n \to \infty \tag{4.6f}$$

for a positive constant[2] C. The value of C and $\kappa(p,q)$ determine the rate of decay, which is inevitable with an increasing number of samples. The significance level decreases when either C and $\kappa(p,q)$ increases. When C is not small, then the likelihood ratio p/q takes on a range of values and when the KL divergence of q from p is not small, sampling one of the distributions is typically not a random sample of the other distribution. Both cases confirm our intuition that a negative estimator $\hat{\Lambda}_n$ is a rare occurrence when the KL divergence of q from p or C are not small. However, when the product of the square of the divergence and C are small, a large number of samples may be necessary to achieve a prescribed significance level.

In a similar fashion, the probability of a type II error, (accepting the alternate hypothesis when it is false)

$$\begin{aligned}\beta_n &= \mathbb{P}(\hat{\Lambda}_n > 0 \,|\, H_q \text{ is true}) \\ &= \mathbb{P}(\hat{\Lambda}_n - \big(-\kappa(q,p)\big) > \kappa(q,p) \,|\, H_q \text{ is true}) \\ &\leqslant \exp\big(-C\,n\,\kappa^2(q,p)\big) \to 0 \text{ as } n \to \infty\,. \end{aligned} \tag{4.6g}$$

In contrast with the bound on the significance level (4.6f) that involves $\kappa(p,q)$, the type II error instead involves $\kappa(q,p)$.

Hoeffding's inequality is an example of a concentration inequality, the purpose of which is to quantify how a function of random variables cluster about its mean (in our case the sum of random variables $\hat{\Lambda}_n$ about the mean $\kappa(p,q)$ or $\kappa(q,p)$). The upper bound is non-asymptotic because it bounds the significance α_n or β_n for each n in contrast to the classical limit theorems of probability.

4.3. Example: Two normal distributions

Suppose we assume that $X_1, X_2, \ldots, X_n$ are normally distributed. The hypothesis test (4.2) then specializes to

$$\begin{aligned} H_p &: \mu = \mu_p\,, \sigma = \sigma_p \\ H_q &: \mu = \mu_q\,, \sigma = \sigma_q\,. \end{aligned} \tag{4.7}$$

Let's establish (4.4) for the case of p and q normal distributions. A tedious calculation establishes

$$\begin{aligned}\log\frac{p(X_1)\cdots p(X_n)}{q(X_1)\cdots q(X_n)} &= \log\frac{\sigma_q^n}{\sigma_p^n} - \frac{\sum_{i=1}^n (X_i-\mu_p)^2}{2\sigma_p^2} + \frac{\sum_{i=1}^n (X_i-\mu_q)^2}{2\sigma_q^2} \\ &= n\log\frac{\sigma_q}{\sigma_p} - n\bar{X}^2\left(\frac{1}{2\sigma_q^2} - \frac{1}{2\sigma_p^2}\right) + n\bar{X}\left(\frac{\mu_p}{\sigma_p^2} - \frac{\mu_q}{\sigma_q^2}\right) \\ &\quad + n\left(\frac{\mu_q^2}{2\sigma_q^2} - \frac{\mu_p^2}{2\sigma_p^2}\right) \end{aligned} \tag{4.8}$$

[2] If we assume that the n samples satisfy $m \leqslant X_j \leqslant M$, then $C = 4\log(M/m)$. We conclude that C quantifies the range of values that the likelihood ratio p/q achieves.

where $\bar{X}$ and $\bar{X}^2$ are the sample mean and sample second moment. We can then write the estimator

$$\begin{aligned}\hat{\Lambda} &= \log\frac{\sigma_q}{\sigma_p} - \bar{X}^2\left(\frac{1}{2\sigma_q^2} - \frac{1}{2\sigma_p^2}\right) + \bar{X}\left(\frac{\mu_p}{\sigma_p^2} - \frac{\mu_q}{\sigma_q^2}\right) + \frac{\mu_q^2}{2\sigma_q^2} - \frac{\mu_p^2}{2\sigma_p^2} \\ &\to \log\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_q - \mu_p)^2}{2\sigma_q^2} - \frac{1}{2} \text{ in probability as } n \to \infty \end{aligned} \tag{4.9}$$

where we used the two relations that $\bar{X} \to \mu_p$ and $\bar{X}^2 \to \sigma_p^2 + \mu_p^2$ in probability as $n \to \infty$. We established (4.4) since the limit is the relative entropy between two normal distributions.

Figures 4-1a–4-1c depicts the use of the estimator

$$\hat{\Lambda}_n = \frac{1}{n}\sum_{j=1}^{n} \log\frac{p(X_j)}{q(X_j)}$$

where X_j are random numbers distributed with respect to p for the specific case

$$\begin{aligned} H_p &: \mu_p = 2, \sigma_p = 1/10 \\ H_q &: \mu_q = 1, \sigma_q = 2/10 . \end{aligned} \tag{4.10}$$

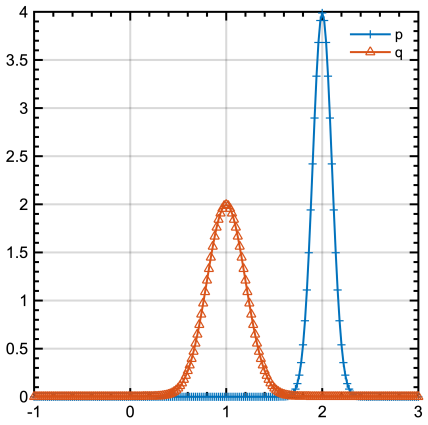
The type I and type II errors α_n are not displayed because they are extremely small and underflow the matlab computation. Let's instead modify the hypothesis test to be

$$\begin{aligned} H_p &: \mu_p = 1, \sigma_p = 1/10 \\ H_q &: \mu_q = 1, \sigma_q = 9/100 . \end{aligned} \tag{4.11}$$

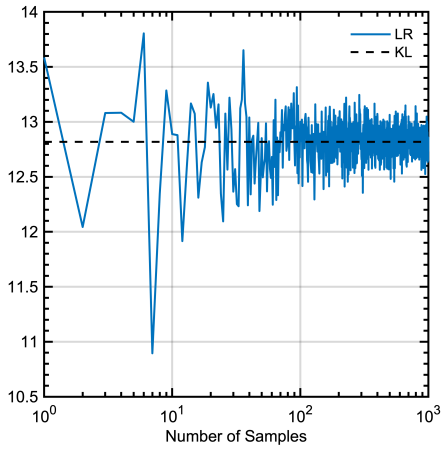
Figures 4-2a–4-2d provides an interesting contrast. In particular, Figure 4-2c depicts the convergence of the type I and type II errors α_n and β_n via the use of the bounds (4.6f) and (4.6g), respectively.

4.4. Notes and References

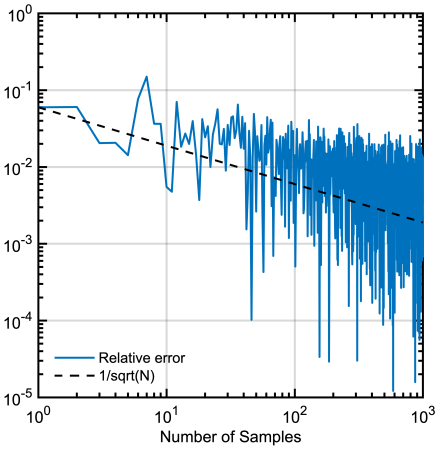
The discussion of this section is based upon the lecture note `https://nowak.ece.wisc.edu/ece830/ece830_spring15_lecture7.pdf` due to Robert Nowak.



(a) Depicting the normal distributions.

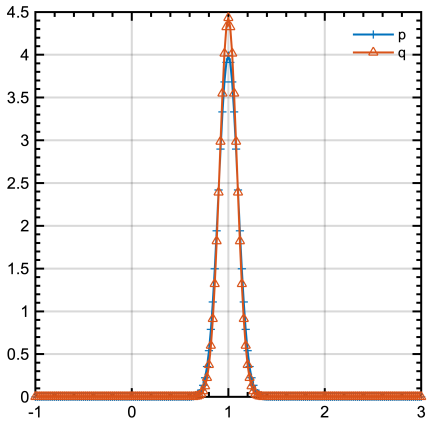


(b) Log likelihood convergence of Λ_n.

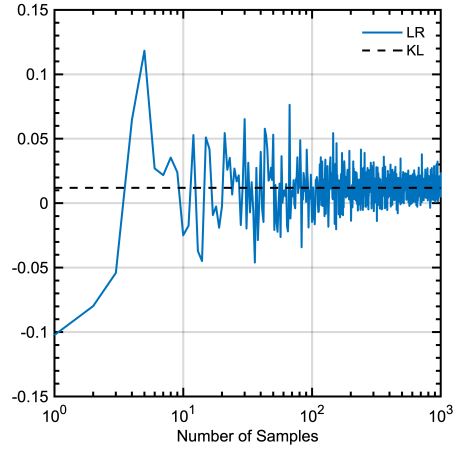


(c) Relative error of Λ_n including the convergence rate of Monte Carlo sampling.

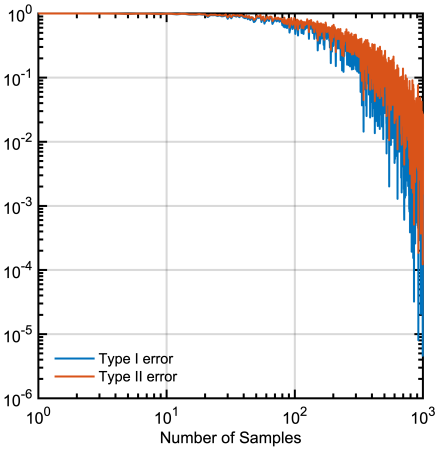
Figure 4-1. Relative entropy analysis on two normal distributions with parameters $\mu_p = 2, \sigma_p = 1/10$ and $\mu_q = 1, \sigma_q = 2/10$



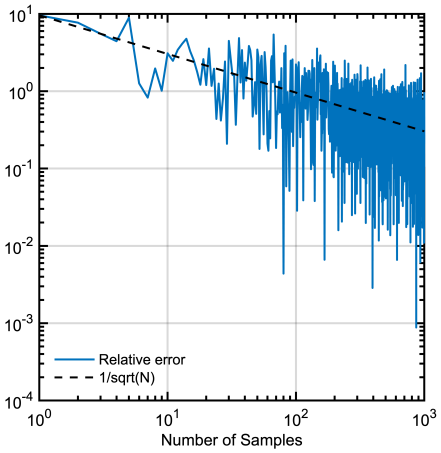
(a) Depicting the normal distributions.



(b) Log likelihood convergence of Λ_n.



(c) Convergence of the Type I and Type II errors (α_n, β_n).



(d) Relative error of Λ_n including the convergence rate of Monte Carlo sampling.

Figure 4-2. Entropy analysis on two normal distributions where p has parameters $\mu_p = 1, \sigma_p = 1/10$ and q has parameters $\mu_q = 1, \sigma_q = 9/100$.

5. FISHER INFORMATION

The *score* random variable is defined to be

$$s(\vartheta) = \nabla_\vartheta \log L(\vartheta) \tag{5.1a}$$

where ∇_ϑ is the gradient[1] and $L(\vartheta) = p(x;\vartheta)$ is the likelihood function. The Fisher information is defined to be the variance of the score

$$I(\vartheta) \coloneqq \mathbb{E}\, s^2(\vartheta) = \int \Big(\nabla_\vartheta \log p(x;\vartheta)\Big)^2 \, p(x;\vartheta)\, dx\,. \tag{5.1b}$$

Fisher information quantifies the relationship between a random variable X and the parameter ϑ of the log-likelihood function $L(\vartheta)$. While the score and Fisher information are functions of ϑ, the former holds fixed the sample x and the latter integrates over all possible samples x. The definition of the Fisher information entailed the introduction of two fundamental quantities—the score and the likelihood—in mathematical statistics; we'll return to the likelihood shortly. Finally note that the Fisher information is positive unless I does not depend upon ϑ. We'll explain in § 5.3, the Fisher information quantifies the mapping to the parameter from the sample. The intuition here is that the Fisher information increases with the amount of information in the sample to parameter mapping. When the Fisher information is small, many samples will be necessary to estimate the parameter.

A careful reader might recall that the variance of a random variable subtracts the square of the random variable from the second moment of the random variable. Therefore the variance of the score is $\mathbb{E}\, s^2(\vartheta) - \big(\mathbb{E}\, s(\vartheta)\big)^2$. However by using the identity

$$p(x;\vartheta)\, \nabla_\vartheta \log p(x;\vartheta) = \nabla_\vartheta p(x;\vartheta)\,, \tag{5.1c}$$

we can show that the mean score is zero, i.e.,

$$\begin{aligned} \mathbb{E}\, s(\vartheta) &= \int_{-\infty}^{\infty} \nabla_\vartheta \log p(x;\vartheta)\, p(x;\vartheta)\, dx \\ &= \int_{-\infty}^{\infty} \nabla_\vartheta p(x;\vartheta)\, dx \\ &= \nabla_\vartheta \int_{-\infty}^{\infty} p(x;\vartheta)\, dx = \nabla_\vartheta 1 = 0 \end{aligned} \tag{5.1d}$$

[1] When ϑ is a vector of parameters then the gradient of the score is a vector and

$$\Big(\nabla_\vartheta \log p(x;\vartheta)\Big)^2 = \Big(\nabla_\vartheta \log p(x;\vartheta)\Big)^\top \nabla_\vartheta \log p(x;\vartheta) = \nabla_\vartheta \log p(x;\vartheta) \cdot \nabla_\vartheta \log p(x;\vartheta)\,.$$

where $\nabla_\vartheta = \begin{pmatrix} \frac{\partial}{\partial \vartheta_1} & \frac{\partial}{\partial \vartheta_1} & \cdots & \frac{\partial}{\partial \vartheta_k} \end{pmatrix}$

where we assumed that differentiation and integration can be interchanged. The aforementioned identity can also be used to show that

$$\mathbb{E}\Big(\nabla_\vartheta \log p(X;\vartheta)\Big)^2 = \mathbb{E}\Big(\frac{\nabla_\vartheta p(X;\vartheta)}{p(X;\vartheta)}\Big)^2 = -\mathbb{E}\,\nabla_\vartheta^2 \log p(X;\vartheta)\,, \tag{5.1e}$$

formulations you might encounter in the literature. The last equality, which we will use below, integrates by parts the first expression, assuming that two derivatives of $\log p(X;\vartheta)$ with respect to ϑ exist. If ϑ is a scalar, then $\nabla_\vartheta^2 = \Delta_\vartheta$, i.e., the Laplacian.

5.1. Information content

Fisher information maps ϑ to a non-negative real number. The mapping quantifies the variability between X and ϑ. If X, or equivalently p, is sensitive to changes in ϑ, then the Fisher information will be "large". If X is not sensitive to changes in ϑ, then the Fisher information will be "small". The derivative with respect to ϑ then encapsulates the sensitivity of X to ϑ. We have also learned that the differential entropy (2.1a) of X quantifies the average amount of surprise of X using the density p. So both notions are attempting to quantify the behavior of the random variable X via its density. Why do we need both notions of information? Let's first understand the question Fisher information helps to answer.

The likelihood, a fundamental quantity in mathematical statistics, was defined in (4.6b) and is used to emphasize that the parameter ϑ, which we'll assume is a scalar for simplicity of notation, is the variable of interest. Given independent samples $x_1, x_2, \ldots, x_m$ distributed according to $p(x;\theta)$, the log likelihood becomes

$$\log L(\theta) = \log \Pi_{i=1}^m p(x_i;\theta) = \sum_{i=1}^m \log p(x_i;\theta)\,. \tag{5.2}$$

But suppose we don't know the parameter value θ and denote this ignorance by using the variable ϑ. Now note that the log likelihood is also a random variable because it is a function of the samples and so the sample average is of interest. The law of large numbers then implies that

$$\frac{1}{m}\sum_{i=1}^m \log p(x_i;\vartheta) \to \mathbb{E}\log p(X;\vartheta) = \int \log p(X;\vartheta)\, p(x;\theta)\, dx \tag{5.3}$$

as $m \to \infty$.

So let's first understand the case of an extremely large number of samples m so that the sample average is an excellent approximation to the differential entropy for $p(x;\theta)$. Let's interchange expectation and the gradient followed by the use of the identity (5.1c) to obtain

$$\nabla_\vartheta\, \mathbb{E}\log p(X;\vartheta) = \mathbb{E}\nabla_\vartheta \log p(X;\vartheta) = \mathbb{E}\frac{\nabla_\vartheta p(X;\vartheta)}{p(X;\vartheta)}\,. \tag{5.4a}$$

Because the mean of the score is zero (recall (5.1d)), then

$$\mathbb{E}\frac{\nabla_\vartheta p(X;\vartheta)}{p(X;\vartheta)}\Big|_{\vartheta=\theta} = \mathbb{E}\frac{\nabla_\theta p(X;\theta)}{p(X;\theta)} = \int \nabla_\theta p(x;\theta)\,dx = 0 \tag{5.4b}$$

so that the parameter θ associated with the samples is a critical or stationary point for the expectation in (5.3). Now recall from elementary differential calculus that in order to show that the critical point θ is a maximum, we need to establish that the second derivative of the mean score is negative, i.e.,

$$\nabla_\vartheta^2 \mathbb{E}\log p(X;\vartheta)\Big|_{\vartheta=\theta} < 0\,. \tag{5.4c}$$

Assuming that we can interchange expectation with differentiation, the alternate formulations for the Fisher information (5.1e) imply that the critical point θ is a (local) maximum when the inequality

$$I(\theta) = \mathbb{E}\Big(\nabla_\theta \log p(X;\vartheta)\Big)^2\Big|_{\vartheta=\theta} > 0 \tag{5.4d}$$

is satisfied. Hence the amount of Fisher information explains the efficacy in estimating θ given samples of X that is postulated to be modeled by $p(x;\theta)$.

5.2. Entropy and Fisher information

The equality in (5.3) defines a mapping φ that satisfies

$$\varphi(\vartheta) = \mathbb{E}\log p(X;\vartheta) = \int \log p(x;\vartheta)\,p(x;\theta)\,dx\,. \tag{5.5}$$

This leads to a relationship between differential entropy and Fisher information because $\varphi(\theta) = \mathbb{E}\log p(X;\theta)$ is the differential entropy for the random variable X. Whereas the differential entropy quantifies the "average amount of surprise" in the random variable X by considering the value $\mathbb{E}\log p(X;\theta)$, the Fisher information describes the sensitivity of the mapping φ to changes in ϑ. This sensitivity is useful for statisticians who are typically interested in describing the population parameter θ given samples of X.

5.3. Maximum Likelihood

The Maximum Likelihood Estimate (MLE) is defined to be

$$\widehat{\vartheta}_{\text{MLE}} \coloneqq \arg\max_\vartheta \sum_{i=1}^{m} \log p(x_i;\vartheta)\,. \tag{5.6}$$

As its name suggests, the solution of this maximization problem results in an estimator that it is most "likely" for the samples. But can we explain why the MLE is a good estimate? In particular, because the MLE is a random variable, what is its distribution and variance?

In fact, a standard result is that

$$\widehat{\theta}_{\mathrm{MLE}} \sim N\Big(\theta, \frac{1}{m\, I(\theta)}\Big) \text{ as } m \to \infty\,. \tag{5.7}$$

In words, the MLE random variable is normally distributed about the parameter θ with a variance that decreases with the product of the Fisher information $I(\theta)$ and the number of samples. The estimate explains that the MLE is asymptotically unbiased, i.e.,

$$\mathbb{E}\,\widehat{\theta}_{\mathrm{MLE}} \to \theta\,. \tag{5.8}$$

Large Fisher information implies that the MLE is concentrated about the mean θ and the number of samples needed to reach a prescribed threshold on the variance is reduced. Conversely, small Fisher information implies that many samples are needed so that the MLE is concentrated about the mean θ. This discussion supports our contention that the Fisher information quantifies the mapping to the parameter from the sample. Recall that the Fisher information is positive; if larger than one, then it amplifies the amount of information per sample. Conversely, if less than one then the Fisher information constricts the amount of information per sample.

When the product of the number of samples and the Fisher information is sufficiently large, then $\widehat{\theta}_{\mathrm{MLE}} \approx \theta$ so that the conclusions reached in (5.4) suggest that

$$\mathbb{E}\nabla_\vartheta \log p(X;\vartheta)\Big|_{\vartheta=\widehat{\theta}_{\mathrm{MLE}}} \approx 0 \tag{5.9a}$$

and

$$I(\vartheta)\Big|_{\vartheta=\widehat{\theta}_{\mathrm{MLE}}} > 0 \tag{5.9b}$$

are plausible. This is useful because the parameter θ is unknown and so $I(\widehat{\theta}_{\mathrm{MLE}})$ can be used to construct confidence intervals.

5.4. Cramer-Rao bound

The asymptotic distribution for the MLE (5.7) quantifies that the Fisher information is inversely proportional to the variance. This begs the questions of whether there are better estimators and what is the best possible variance given a sample. The latter question is addressed by the Cramér-Rao bound, which in the case of a scalar parameter states that

$$\mathbb{E}(\theta - \widehat{\theta})^2 \geqslant \frac{1}{I(\theta)} \tag{5.10}$$

for an unbiased estimator $\widehat{\theta}$ for θ. The latter implies that $\mathbb{E}(\theta - \widehat{\theta}) = 0$ so that the inequality (5.10) explains that the variance of the estimator $\widehat{\theta}$ is bounded from below by the inverse of the Fisher information $I(\theta)$. Hence a large Fisher information explains that the variance of an unbiased estimator may be small. Whether an estimator can be determined satisfying the bound is another matter but no unbiased estimator can be smaller than the inverse of the Fisher information.

Given m independent samples, the maximum likelihood estimate (5.7) satisfies the lower bound

$$\mathbb{E}(\theta - \widehat{\theta}_{\mathrm{MLE}})^2 \geqslant \frac{1}{mI(\theta)} \tag{5.11}$$

which explains that the bound for the MLE estimator decreases with the number of samples.

5.5. Fisher information and relative entropy

How close is the density $p(x;\vartheta)$ to $p(x;\theta)$? One way to quantify this closeness is via the relative entropy $\psi(\vartheta) = \kappa(p(x;\vartheta), p(x;\theta))$. If we assume that the relative entropy is a differentiable function of ϑ then

$$\psi(\theta) = 0 = \psi'(\theta)$$

By (4.1a), the first equality follows since the relative entropy is only zero when the two densities are the same, i.e., $\vartheta = \theta$, so that the function ψ is minimized at $\vartheta = \theta$, i.e., the second equality. Hence the Taylor series expansion

$$\begin{aligned}\psi(\vartheta) &= \psi(\theta) + \psi'(\theta)(\vartheta - \theta) + \frac{1}{2}\psi''(\theta)(\vartheta - \theta)^2 + O(\vartheta - \theta)^3 \\ &= \frac{1}{2}\psi''(\theta)(\vartheta - \theta)^2 + O(\vartheta - \theta)^3\end{aligned}$$

grants

$$\begin{aligned}\kappa\big(p(x;\vartheta), p(x;\theta)\big) &= \frac{1}{2}\frac{\partial^2}{\partial\vartheta^2}\kappa\big(p(x;\vartheta), p(x;\theta)\big)\Big|_{\vartheta=\theta}(\vartheta - \theta)^2 + O(\vartheta - \theta)^3 \\ &= \frac{1}{2}\kappa\Big(\frac{\partial^2}{\partial\vartheta^2}p(x;\vartheta)\Big|_{\vartheta=\theta}, p(x;\theta)\Big)\Big|_{\vartheta=\theta}(\vartheta - \theta)^2 + O(\vartheta - \theta)^3 \\ &= I(\theta)\frac{(\vartheta - \theta)^2}{2} + O(\vartheta - \theta)^3\end{aligned}$$

where the second and third equalities follow by (5.1e) and (5.1b). In words, the relative entropy

$$\kappa\big(p(x;\vartheta), p(x;\theta)\big) = I(\theta)\,\frac{(\vartheta - \theta)^2}{2} + O(\vartheta - \theta)^3$$

is linearly approximated by the Fisher information for ϑ sufficiently close to θ. This is an instance of the well-understood relationship between an f-divergence (e.g., KL-divergence) and the Fisher information. As with all Taylor series approximations, the analysis is formal since we have no idea how close ϑ needs to be to θ nor whether any of the higher order derivatives of ψ are bounded as ϑ approaches θ. The importance, though, is to emphasize the primacy of the Fisher information.

6. MUTUAL DEPENDENCE

The mutual information of two random variables quantifies the dependence between the two variables. A related quantity is the concordance of random variables and is typically conflated with dependence. The latter concept relates the functional relationship between two random variables while the former concept measures the presence of positive or negative co-movement. Correlation is properly understood as a concordance and its absolute value is widely used as a measure of dependence despite its limitations.

Let's start to understand what mutual information quantifies by comparing it to the linear correlation between the variables. This helps to clarify the meaning of the important concept of *dependence* between random variables and draw a contrast with concordance.

Recall that the correlation of two random variables X and Y is

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{6.1}$$

where

$$\begin{aligned}
\text{Cov}(X,Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y) \\
&= \int\int p_{XY}(x,y)xy\,dx\,dy - \int p_X(x)x\,dx \int p_Y(y)y\,dy \\
&= \int\int \big(p_{XY}(x,y) - p_X(x)p_Y(y)\big)\,xy\,dx\,dy
\end{aligned} \tag{6.2}$$

and

$$p_X(x) = \int p_{XY}(x,y)\,dy \text{ and } p_Y(y) = \int p_{XY}(x,y)\,dx \tag{6.3}$$

are the marginal pdfs. In contrast the mutual information of two random variables X and Y is

$$\begin{aligned}
I(X;Y) &= \mathbb{E}\left(\ln \frac{p_{XY}(x,y)}{p_Y(x)p_X(y)}\right) \\
&= \int\int \big(\ln p_{XY}(x,y) - \ln p_X(x)p_Y(y)\big)\,p_{XY}(x,y)\,dx\,dy
\end{aligned} \tag{6.4}$$

6.1. Example

Suppose that X is a finite variance random variable symmetrically distributed about the origin so that $\mathbb{E}_X(X) = 0$ (think of a mean zero normal random variable) and $Y = X^{2k}$ for a positive integer

k. Then

$$\begin{aligned}\mathbb{E}(XY) &= \mathbb{E}_{XY}(XY)\\ &= \int\int x\,y\,p_{XY}(x,y)\,dx\,dy\\ &= \int x^{2k+1}\int p_{XY}(x,y)\,dy\,dx\\ &= \int x^{2k+1}p_X(x)\,dx = \mathbb{E}_X(X^{2k+1}) = 0\end{aligned} \tag{6.5a}$$

where we used our assumption that X is symmetrically distributed about the origin, i.e., $p_X(-x) = p_X(x)$ is an even function. Recall that X and Y are independent if and only if $p_{X,Y}(x,y) = p_X(x)\,p_Y(y)$. In other words the joint density $p_{X,Y}$ of independent random variables is a product of their marginal densities. If the random variables are dependent, then the joint density involves more than the marginal densities. And so a limitation of the correlation is revealed because in passing to the fourth equality, the dependence between the random variable is absorbed during the integration over y that results in the marginal density of X. Hence

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y} = \frac{\mathbb{E}_{XY}(XY) - \mathbb{E}_X(X)\mathbb{E}_Y(Y)}{\sigma_X\,\sigma_Y} = 0\,. \tag{6.5b}$$

This example demonstrates that the random variables X and $Y = X^{2k}$ have zero correlation even though they are related in a nonlinear fashion. However, only in the special case that X and $Y = X^{2k}$ are *jointly* normal can we also conclude that the pair is independent.

6.2. Copula

How do we generate a $p_{X,Y}$ that is *not* a product of its marginal pdfs p_X and p_X, i.e., how do we generate a function $q \neq 1$ satisfying

$$p_{XY}(x,y) = q(x,y)\,p_X(x)\,p_Y(y) \tag{6.6}$$

so that X and Y are dependent random variables? This is not so simple as it appears since the product of the q selected and the marginals must be a function that can be identified with a joint pdf.

Does there exist appropriate functions q? A theorem due to Skalar [1959] explains that at least one $q(x,y) = c(u,v)$ exists. The function c is called the copula density where

$$c(u,v) = \frac{\partial^2}{\partial u\,\partial v}C(u,v) \tag{6.7}$$

and C is the copula, by definition, a joint cumulative distribution function with uniform marginal distributions. Hence we can select $u(x) = F_X(x)$ and $v(y) = F_Y(y)$ the cdfs for p_X and q_X, respectively, then

$$\begin{aligned}c(x,y) &= \frac{\partial^2}{\partial u\,\partial v}C\big(F_X(x),F_Y(y)\big)\,\frac{\partial F_X(x)}{\partial x}\,\frac{\partial F_Y(y)}{\partial y}\\ &= \frac{\partial^2}{\partial u\,\partial v}C\big(F_X(x),F_Y(y)\big)\,p_X(x)\,p_Y(y)\end{aligned} \tag{6.8}$$

where we've exploited the result that the cdfs F_X and F_Y are uniformly distributed random variables; see the discussion following (2.5). We can now write

$$p_{XY}(x,y) = c(x,y)\, p_X(x)\, p_Y(y)\,. \tag{6.9}$$

Roughly, this statement states that the randomness in the joint probability density is a product of three randomness—the two marginals and the copula. Figure 6-1 shows two marginal distributions linked by different copulas. How do we quantify how much information is associated with the marginals and the copula? Given our interest with information theory, mutual information proves crucial.

6.3. Mutual independence

Suppose that X and Y are dependent (so that $c \neq 1$). Then by (6.9)

$$0 < I(X;Y) = \mathbb{E}\left(\ln \frac{p_{XY}(x,y)}{p_X(x)\, p_Y(y)} \right) = \mathbb{E} \ln c(u_X, v_Y) = -h\big(c(X,Y)\big) \tag{6.10}$$

Recall that the mutual information of two random variables quantifies the *dependence* between them, a term that wasn't precisely defined. We now see that dependence is encapsulated by the random variable distributed with respect to copula density $c(x,y)$.

Now recall the example in the previous section: X is a finite variance random variable symmetrically distributed about the origin so that $\mathbb{E}_X(X) = 0$ and $Y = X^{2k}$ for a positive integer k. In contrast to correlation, the mutual information is zero if and only if $c = 1$, i.e., when X and Y are independent. It is in this sense that mutual information detects the dependence between X and Y that is not possible with the correlation.

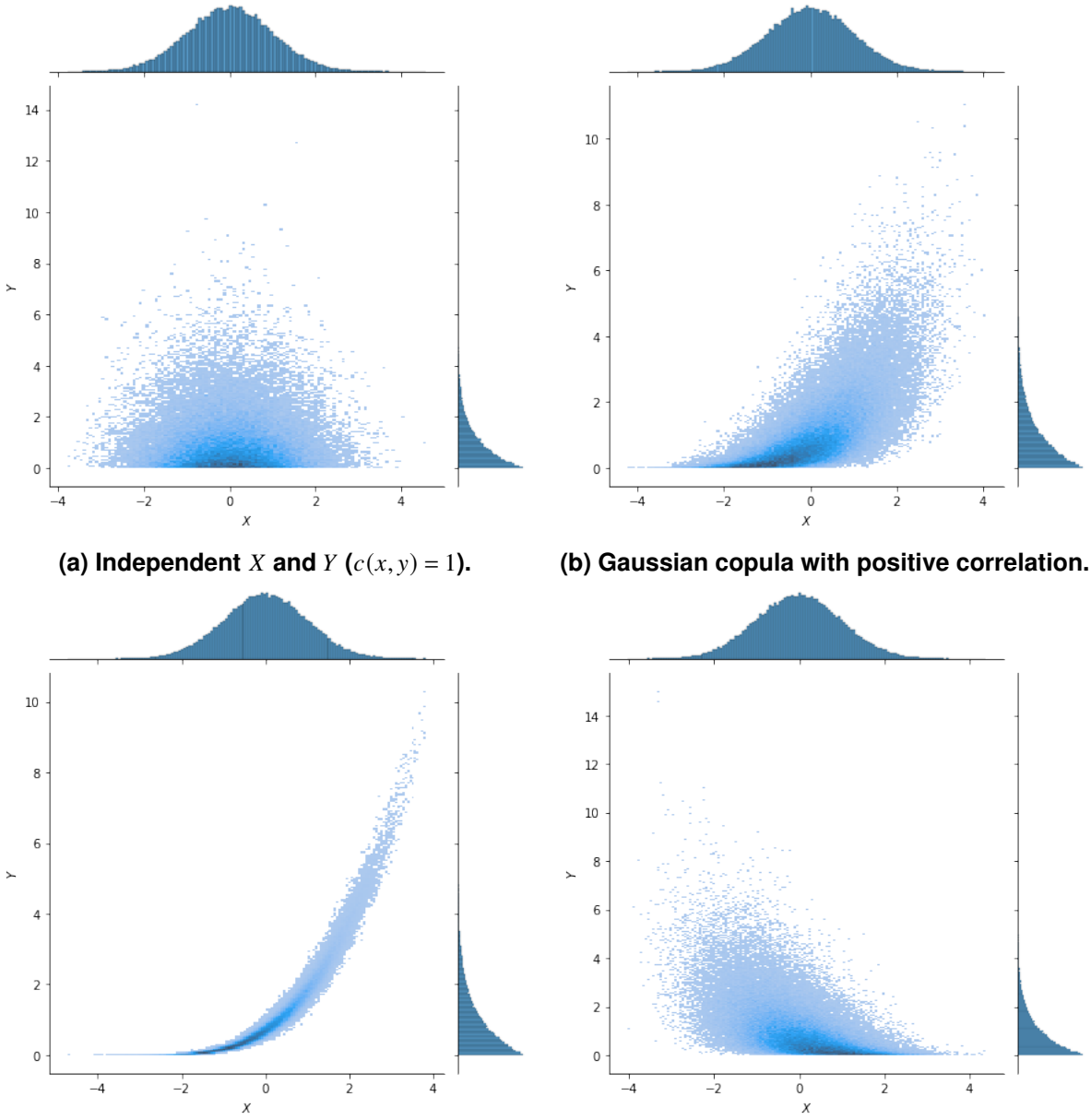
Let's test our understanding. Suppose that X and Y are independent. Then $p_{XY}(x,y) = p_X(x)\, p_Y(y)$ so that $I(X;Y) = 0$. Suppose $Y = \alpha X$ for $\alpha \neq 0$ so that Y has the same distribution as αX. What is $I(X;\alpha X)$? Because $p_Y(y) = p_X(x/\alpha)(1/\alpha)$ we have

$$I(X;\alpha X) = \mathbb{E}\left(\ln \frac{p_{XY}(x,y)}{p_X(x)\, p_Y(y)} \right) = \mathbb{E}\left(\ln \frac{p_X(x)\, p_X(x/\alpha)(1/\alpha)}{p_X(x)\, p_X(x/\alpha)(1/\alpha)} \right) = 0\,.$$

Although knowledge of either marginal density is informative for the other marginal, the copula determines their dependence. If we know that X and Y are dependent, then $c \neq 1$, then their mutual entropy describes the amount of dependence and the negative describes the average amount of surprise.

6.4. Mutual information is unbounded

Note that $I(X;Y) = I(Y;X)$ and $I(X;Y) \geqslant 0$ using Jensen's inequality and $I(X;Y) = 0$ only when X and Y are independent. And so how large can the mutual information between random variables



(a) Independent X and Y ($c(x, y) = 1$).

(b) Gaussian copula with positive correlation.

(c) Gaussian copula with positive correlation.

(d) Gaussian copula with negative correlation.

Figure 6-1. Plots of samples from joint distributions of random variables X and Y linked by different copulas. The marginal distributions for X (distributed normally) and Y (distributed exponentially) can be seen along the top and right of each plot.

be? Let c be the copula density for a uniformly distributed joint random variable on the square $(0, \sqrt{a}) \times (0, \sqrt{a})$ for $0 < a \leqslant 1$ so that by (6.10),

$$\begin{aligned} I(X;Y) &= -h\big(c(X,Y)\big) \\ &= \frac{1}{a} \ln \frac{1}{a} \int_0^{\sqrt{a}} \int_0^{\sqrt{a}} du\, dv \\ &= -\log a \to \begin{cases} \infty & \text{as } a \to 0^+ \\ 0 & \text{as } a \to 1^- \end{cases} \end{aligned} \tag{6.11}$$

Consider the latter case when $a \to 1^-$ so that $c \to 1$ over the square $(0, \sqrt{a}) \times (0, \sqrt{a})$. Hence $p_{XY}(x, y) \to p_X(x)\, p_Y(y)$ so X and Y are approaching independent random variables. In contrast, as $a \to 0^+$, then $c \to 0$ or roughly, the copula c is approaching the product of Dirac delta function $\delta(x)\, \delta(y)$ located at the origin. Hence $p_{XY}(x, y) = 0$ over the domain $(\sqrt{a}, 1) \times (\sqrt{a}, 1)$ so that the mutual information between X and Y increases (without bound) as a decreases to zero. Because the negative of the mutual information I is the negative of the entropy of c, the mutual information is large precisely when c conveys little, if any, average surprise. And so as a decreases to zero, the corresponding copula is associated with a random variable that is nearly deterministic, which implies that the entropy is much less than zero. As a increases towards one, then the entropy of the copula is nearly 0. An important conclusion is that entropy of a copula is bounded from above by zero in contrast to the entropy of the marginal distributions, which are unbounded. In other words, the entropy of the copula contains at most zero average surprise.

6.5. Entropic decomposition

Let's summarize what we've learned via several remarks.

- The copula density c quantifies the dependence between X and Y. The mutual information $I(X;Y)$ is the negative of the entropy of the copula. Large mutual information corresponds to little, if any, average surprise in X given knowledge of Y or in Y given knowledge of X. In contrast, as the mutual information decreases to zero the average surprise in X given knowledge of Y or in Y given knowledge of X increases. However, the largest the amount of surprise is zero.

- $c = 1$ if and only if $I(X;Y) = 0$ (can you support this?). In other words, X and Y are independent. The statisticians draw comfort. The information for a random variable distributed with respect to the joint probability density only depends upon the marginal densities. Their is no dependence between X and Y.

- As c approaches the product of Dirac delta measures (within the unit square) then the mutual information between X and Y increases without bound. In other words, the dependence information between them increases as c decreases to a point mass. This is consistent with our intuition that such a copula characterizes a nearly deterministic variable, e.g., one where the randomness is negligible.

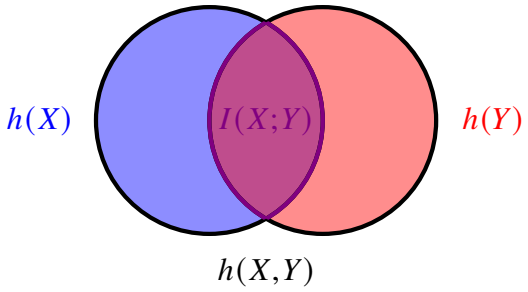


Figure 6-2. A Venn diagram relating the joint entropy $h(X,Y)$ to $h(X)$, $h(Y)$, and $I(X;Y)$.

The important conclusion is that the dependence between X and Y is distinct from the individual attributes of X and Y. Can we support this conjecture by considering the information associated with each of these three random variables?

An elementary identity is the entropic decomposition

$$h(X,Y) = h(X) + h(Y) + h\big(c(X,Y)\big)\,. \tag{6.12}$$

In words, the (differential) entropy for the pair X and Y is the sum of the (differential) entropies of the two marginal densities and copula density. The product of the three densities in (6.9) and their relative importance is given in terms of the information of each. The information, or average surprise for the pair X and Y is the average surprise for X, Y and the dependence $h\big(c(X,Y)\big) = -I(X;Y)$. Since the mutual information I is non-negative, dependence between X and Y reduces the average amount of surprise. When X and Y are independent then the entropy of each defines the entropy of the pair. A common illustration for the entropic composition identity is shown in Figure 6-2.

Note that the (differential) entropy $h(X,Y) \ll 0$ when an appropriate combination of marginal densities and copula density is nearly deterministic, i.e., has little surprise. The entropy $h(X,Y) = 0$ when all three densities are uniform on the unit square. The entropy increases with the standard deviation for a finite variance random variable. This explains that as the differential entropy increases over the interval $(-\infty,\infty)$ the average surprise increases. See the discussion following the entropy definitions (2.1).

7. WASSERSTEIN METRIC

Chapter 4 reviewed relative entropy as a way to compare two distributions. Though the relative entropy is not a distance (since relative entropy is asymmetric under an interchange of p and q), the relationship with hypothesis testing explains in what sense the two distributions can be compared. Chapter 6 explained that mutual information between random variables depends on more than their marginal distributions. The joint pdf (or pmf) considers all possible outcomes of pairs so that the copula and marginal distributions define the joint pdf. Our current chapter reviews a distance that involves the copula.

The ℓ^{th} Wasserstein distance between two probability measures p_X and p_Y is defined in terms of the joint density p_{XY}

$$\begin{aligned} W_\ell(p_X, p_Y) &\coloneqq \left(\inf_{p_{XY} \in P_{XY}(X,Y)} \int_{\mathbb{R}\times\mathbb{R}} d^\ell(x,y) \, \mathrm{d}p_{XY}(x,y) \right)^{1/\ell} \\ &= \left(\inf_{p_{XY} \in P_{XY}(X,Y)} \mathbb{E}[d^\ell(X,Y)] \right)^{1/\ell} \end{aligned} \tag{7.1}$$

where $P_{XY}(X,Y)$ is the collection of all joint probability densities over $\mathbb{R}\times\mathbb{R}$ with marginals p_X, p_Y and d is a cost, a non-negative function. The copula equality (6.9) implies that the collection of joint probabilities densities can be identified with the collection of copulas. In contrast to relative entropy W_ℓ is a distance function. More significantly, however, is that W_ℓ implicitly involves the mutual information (or entropy of the copula c).

In this chapter, we will focus on the first Wasserstein distance $\ell = 1$ unless otherwise specified. The first Wasserstein distance is related to the earth mover's distance, which can roughly be thought of as the minimum cost to transform one pile of dirt (or probability distribution) to another. Unless otherwise specified, we use Euclidean distance $d^1(x,y) = \sqrt{(x-y)^2} = |x-y|$ as the cost function in our computations of Wasserstein distance.

7.1. Comparing the Wasserstein distance and the KL Divergence

7.1.1. Probability mass functions

Some pairs of probability distributions have the same KL divergence but different Wasserstein distance. Other pairs can have the same Wasserstein distance but different KL divergences. Consider,

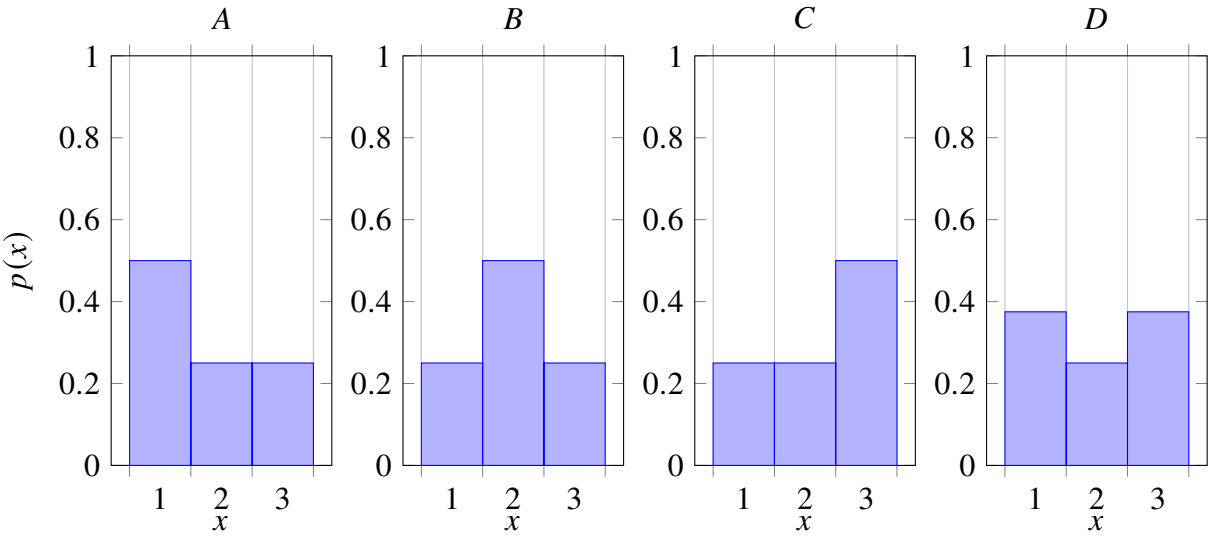


Figure 7-1. An example of four different probability distributions. Note that $\kappa(p_A, p_B) = \kappa(p_A, p_C)$ and $W_1(p_A, p_B) < W_1(p_A, p_C)$. Also, $\kappa(p_B, p_C) > \kappa(p_B, p_D)$ and $W_1(p_B, p_C) = W_1(p_B, p_D)$.

for example, random variables A, B, C and D with pmfs

$$p_A(x) = \begin{cases} \frac{1}{2} & x = 1 \\ \frac{1}{4} & x = 2 \\ \frac{1}{4} & x = 3 \\ 0 & \text{else} \end{cases}, \quad p_B(x) = \begin{cases} \frac{1}{4} & x = 1 \\ \frac{1}{2} & x = 2 \\ \frac{1}{4} & x = 3 \\ 0 & \text{else} \end{cases}, \quad p_C(x) = \begin{cases} \frac{1}{4} & x = 1 \\ \frac{1}{4} & x = 2 \\ \frac{1}{2} & x = 3 \\ 0 & \text{else} \end{cases}, \quad \text{and} \quad p_D(x) = \begin{cases} \frac{3}{8} & x = 1 \\ \frac{1}{4} & x = 2 \\ \frac{3}{8} & x = 3 \\ 0 & \text{else} \end{cases}$$

as shown in Figure 7-1. In this example $\kappa(p_A, p_B) = \kappa(p_A, p_C) = \frac{1}{4}\log(2)$, and $\frac{1}{4} = W_1(p_A, p_B) < W_1(p_A, p_C) = \frac{1}{2}$. In contrast, $\kappa(p_B, p_C) > \kappa(p_B, p_D) = \log(2) - \frac{1}{2}\log(3)$ and $W_1(p_B, p_C) = W_1(p_B, p_D)$.

7.1.2. Normal Random Variables

The first Wasserstein distance between two normal random variables X and Y with $\sigma = \sigma_X = \sigma_Y$ is

$$W_1(p_X, p_Y) = |\mu_X - \mu_Y|$$

and the relative entropy is

$$\kappa(p_X, p_Y) = \frac{(\mu_X - \mu_Y)^2}{2\sigma^2}.$$

where we used (4.9). The KL divergence depends upon σ while the Wasserstein distance does not. The relationship with hypothesis testing provides a clue. The distinction between samples of p_X or p_Y decreases with increasing variance when the difference between the means is fixed. Figure 7-2 plots the KL divergence and the Wasserstein distance when $\mu_Y = 0$. The plots display the linear and quadratic dependence of W_1 and κ upon μ_X and illustrates that κ flattens out as the variance increases.

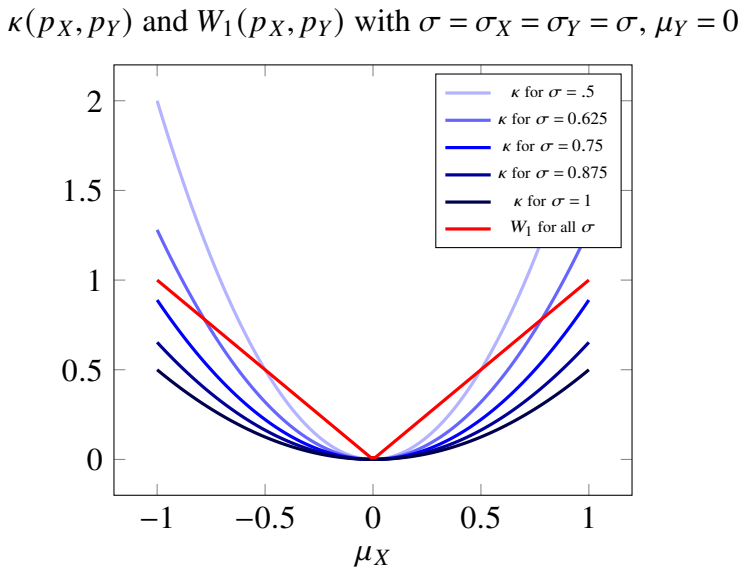


Figure 7-2. A comparison of the KL divergence (blue) and Wasserstein distance (red) among various pairs of normal distributions with the same standard deviation.

7.2. Estimating KL Divergence and Wasserstein Distance from Samples

We can also compare Monte Carlo approximations of KL divergence and Wasserstein distance. A practical consideration emerges when computing a Monte Carlo approximation to the KL divergence. Recall that $\kappa(p_X, p_Y)$ is defined only if $p_Y(x) = 0$ implies $p_X(x) = 0$. If we use histograms to approximate $p_X \approx \hat{p}_X$ and $p_Y \approx \hat{p}_Y$ based on samples, there may be values where $\hat{p}_Y(x) = 0$ and $\hat{p}_X(x) \neq 0$ even if $p_X(x) = 0$ whenever $p_Y(x) = 0$. One strategy to allow for the estimation of $\kappa(p_X, p_Y)$ is to discard all bins in the histogram where $\hat{p}_Y(x) = 0$. Another possibility regularizes the estimated probabilities. That is, for each bin in the histogram, a small positive number is added to the estimated probability for that bin, and renormalize the estimated probability.

Monte Carlo approximations to the KL divergence and W_1 using samples of two normally distributed random variables X and Y with the same standard deviation are shown in Figure 7-3. An immediate observation is that the error in the Wasserstein approximation is small regardless of the difference in the means. In contrast, the error in the KL divergence approximation increases with the difference in means.

Recall that the definition of Wasserstein distance involves joint probability densities, and the definition of KL divergence does not. In our experiments, we observe that the copula used to link two random variables X and Y impacts the estimates of the Wasserstein distance. More specifically, Monte Carlo approximations of $W_1(X, Y)$ appear to be more consistent across different trials when X and Y are linked by a Gaussian copula with a positive correlation. In comparison, the estimates of $\kappa(X, Y)$ are spread out even as the correlation increases. An example for two normally distributed random variables is show in Figure 7-4.

One question that may arise concerns the difference between Wasserstein distance estimates for the

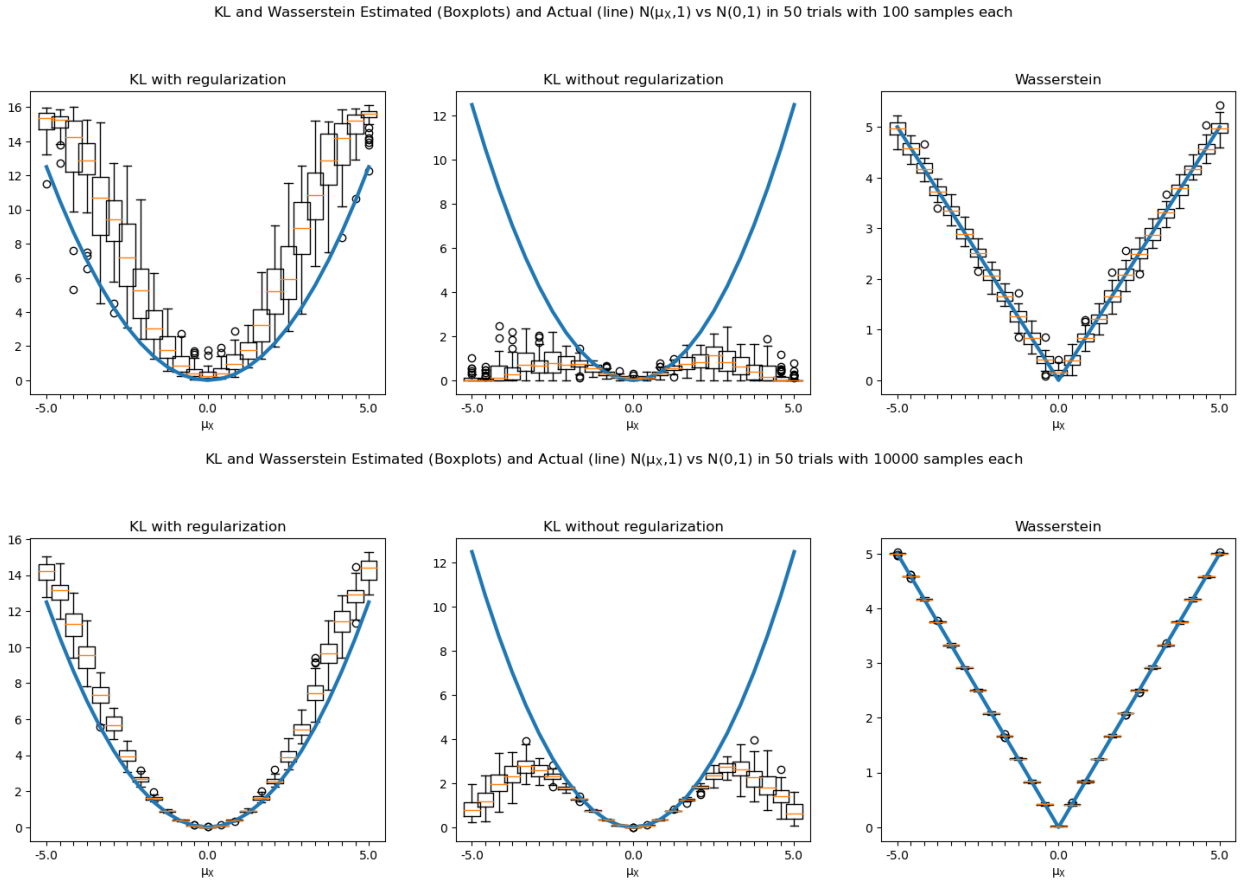


Figure 7-3. Estimated KL divergence and first Wasserstein distance based on samples of normally distributed random variables X (with mean μ_X and standard deviation 1) and Y (with mean 0 and standard deviation 1). The boxplots show estimates from 50 trials with (top) 100 samples each or (bottom) 10,000 samples each. The lines show the actual KL divergence and Wasserstein distance between p_X and p_Y. The estimates use numpy histograms with 'auto' bins.

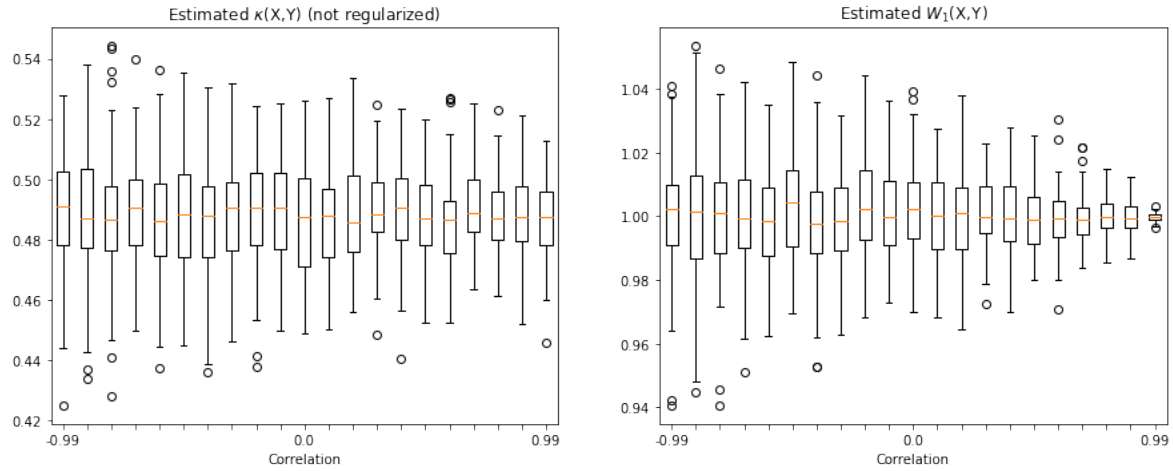


Figure 7-4. Estimated KL divergence and first Wasserstein distance for random variables $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(1,1)$ linked by Gaussian copulas with different correlations. The box plots show results from 100 trials of 10,000 samples from each distribution. From the equations for the KL divergence and Wasserstein distance of normal distributions, we know that $\kappa(X,Y) = \frac{1}{2}$ and $W_1(X,Y) = 1$. The estimates use numpy histograms with 'auto' bins.

random variables linked by copulas with positive and negative correlations. After all, shouldn't a large negative correlation tell us as much about the relationship between X and Y as a large positive correlation?

The Wasserstein distance measures a cost of transforming one distribution to another. Consider $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(1,1)$. If X and Y are connected by a copula with large positive correlation, then a sample distance $|X-Y|$ tends to be close to the distance $|\mu_X - \mu_Y|$. This can be seen in the rightmost plot of Figure 7-5, where samples of (X,Y) are clustered around the line $y = \mu_Y - \mu_X + x$. In our observations (x,y) of (X,Y), the x and y tend to be $\mu_Y - \mu_X$ apart. If, instead, X and Y are connected by a copula with large negative correlation, then samples of $|X-Y|$ are more spread out, i.e., the entropy of the random variable $|X-Y|$ is larger. This can be seen in the leftmost plot of Figure 7-5, where samples of (X,Y) are clustered about the line $y = \mu_Y - \mu_X - x$. Regardless of the correlation, with enough samples, the average distance between x and y in observations (x,y) of (X,Y) approaches $|\mu_X - \mu_Y|$. However, more samples are needed to approach $|\mu_X - \mu_Y|$ as the correlation decreases.

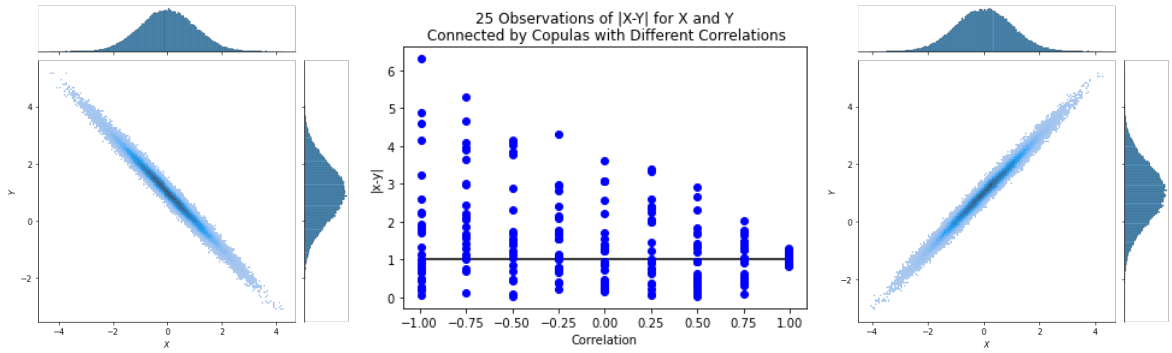


Figure 7-5. Samples from normal distributions $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(1,1)$ linked by Gaussian copulas with (left) negative correlation or (right) positive correlation. (Center) Some observations of $|X-Y|$ at different correlations. The black line shows $|\mu_X - \mu_Y|$.

REFERENCES

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2012.

R.A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 7th ed., rev. and enl. edition, 1925.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214/aoms/1177729694.

Duncan J Murdoch, Yu-Ling Tsai, and James Adcock. P-values are random variables. *The American Statistician*, 62(3):242–245, 2008. doi: 10.1198/000313008X332421. URL https://doi.org/10.1198/000313008X332421.

J. Neyman and E. S. Pearson. IX. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A.*, 231:289–337, 1933.

Abe Skalar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

DISTRIBUTION

Email—Internal

Name	Org.	Sandia Email Address
Technical Library	1911	sanddocs@sandia.gov

Hardcopy—Internal

Number of Copies	Name	Org.	Mailstop
1	L. Martin, LDRD Office	1910	0359

Hardcopy—External

Number of Copies	Name(s)	Company Name and Company Mailing Address

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.