Estimation of Standardized Mutual Information

Zhiyi Zhang and Ann M. Stewart
Department of Mathematics and Statistics
University of North Carolina at Charlotte
Charlotte, NC 28223

June 28, 2016

Introduction

A question of considerable importance in statistics is that of the degree of dependence between two random variables, or the amount of information one random variable contains about another. Mutual information gives an answer to this question, however it can certainly be improved upon. Mutual information is always nonnegative, yet it does not have a uniform upper bound. This can make it difficult to interpret the strength of the association of the two random variables based on mutual information alone. It begs the question: just how high of a number must mutual information be for the two random variables to be considered to depend significantly, or even completely, on each other? In this paper we will consider a possible solution to this matter by defining a standardized mutual information κ that has the asset of being strictly between zero and one inclusive. This κ has the desirable trait of being equal to zero if and only if the two random variables are independent and equal to one if and only if the two random variable have a one-to-one correspondence. We will also consider the estimation of κ and the asymptotic properties of the estimators we develop.

Background

We begin by properly setting the scene and giving some necessary framework definitions.
Let X and Y be two random variables on the following finite alphabets:

$$\mathscr{X} = \{x_i; i = 1, \cdots, K_1\} \text{ and } \mathscr{Y} = \{y_j; j = 1, \cdots, K_2\} \tag{1}$$

with cardinalities $K_1 < \infty$ and $K_2 < \infty$ respectively. Also consider the Cartesian product $\mathscr{X} \times \mathscr{Y}$ with a corresponding joint probability distribution $\mathbf{p}_{\mathrm{X,Y}} = \{p_{i,j}\}$:

$$\mathscr{X} \times \mathscr{Y} = \{(x_i, y_j); i = 1, \cdots, K_1; j = 1, \cdots, K_2\} \tag{2}$$

Let the two marginal distributions be respectively denoted by

$$\mathbf{p}_{\mathrm{X}} = \left\{ p_{i,\cdot} = \sum_{j=1}^{K_2} p_{i,j}; i = 1, \cdots K_1 \right\} \tag{3}$$

and

$$\mathbf{p}_{\mathrm{Y}} = \left\{ p_{\cdot,j} = \sum_{i=1}^{K_1} p_{i,j}; j = 1, \cdots K_2 \right\} \tag{4}$$

For notation simplicity, $\sum_i = \sum_{i=1}^{K_1}$, $\sum_j = \sum_{j=1}^{K_2}$, and $\sum_{i,j} = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2}$ will be observed throughout the text unless otherwise specified.

Also consider the conditional probability distributions, respectively, of X on $\mathscr{X}$ given $Y = y_j$ where $j \geq 1$ is a specific index value, and of Y on $\mathscr{Y}$ given $X = x_i$ where $i \geq 1$ is a specific index value. In other words,

$$\mathbf{p}_{\mathrm{X}|\, y_j} = \left\{ p_{x_i|y_j} = \frac{p_{i,j}}{\sum_k p_{k,j}}; i \geq 1 \right\} \tag{5}$$

$$\mathbf{p}_{\mathrm{Y}|\, x_i} = \left\{ p_{y_j|x_i} = \frac{p_{i,j}}{\sum_k p_{i,k}}; j \geq 1 \right\} \tag{6}$$

In order to build toward the concept of the standardized mutual information κ, we introduce first the definitions of Kullback-Leibler Divergence, Shannon's Entropies, and Mutual Information, and also some relevant theorems.

Definition 1. *Shannon's entropies for $\mathscr{X}$, $\mathscr{Y}$, and $\mathscr{X} \times \mathscr{Y}$ are defined as*

$$\begin{aligned} G_{X}(\mathbf{v}) = H(X) &= -\textstyle\sum_i p_{i,\cdot} \ln p_{i,\cdot} \\ G_{Y}(\mathbf{v}) = H(Y) &= -\textstyle\sum_j p_{\cdot,j} \ln p_{\cdot,j} \\ G_{XY}(\mathbf{v}) = H(X,Y) &= -\textstyle\sum_i \sum_j p_{i,j} \ln p_{i,j} \\ G_{X|\, Y=\, y_j}(\mathbf{v}) = H(X \mid Y = y_j) &= -\textstyle\sum_i p_{x_i|y_j} \ln p_{x_i|y_j} \\ G_{Y|X=\, x_i}(\mathbf{v}) = H(Y \mid X = x_i) &= -\textstyle\sum_j p_{y_j|x_i} \ln p_{y_j|x_i}. \end{aligned} \tag{7}$$

Definition 2. *Given a joint probability distribution $\mathbf{p}_{X,Y}$ on $\mathscr{X} \times \mathscr{Y}$, the (expected) conditional entropy of Y given X is*

$$H(Y \mid X) = \sum_{i \geq 1} p_{i,\cdot} H(Y \mid X = x_i) \tag{8}$$

similarly, the (expected) conditional entropy of X given Y is

$$H(X \mid Y) = \sum_{j \geq 1} p_{\cdot,j} H(X \mid Y = y_j) \tag{9}$$

Lemma 1. *Given a joint probability distribution $\mathbf{p}_{X,Y} = \{p_{i,j}; i \geq 1, j \geq 1\}$ on $\mathscr{X} \times \mathscr{Y}$,*

1) $H(X,Y) = H(X) + H(Y|X)$, *and*

2) $H(X,Y) = H(Y) + H(X|Y)$

Corollary 1. *Given a joint probability distribution* $\mathbf{p}_{X,Y} = \{p_{i,j}; i \geq 1, j \geq 1\}$ *on* $\mathscr{X} \times \mathscr{Y}$,

1) $H(X) \leq H(X,Y)$

2) $H(Y) \leq H(X,Y)$, *and*

3) $H(X) + H(Y) - H(X,Y) \leq H(X,Y)$

Definition 3. *For two probability distributions* $\mathbf{p}$ *and* $\mathbf{q}$ *on the same alphabet* $\mathscr{X}$, *the relative entropy or the Kullback-Leibler divergence of* $\mathbf{p}$ *and* $\mathbf{q}$ *is defined as*

$$D(\mathbf{p}||\mathbf{q}) = \sum_{k=1}^{K} p_k \ln\left(\frac{p_k}{q_k}\right) \tag{10}$$

observing that, for each summand $p\ln(p/q)$,

1) *If* $p = 0$, $p\ln\left(\frac{p}{q}\right) = 0$, *and*

2) *If* $p > 0$ *and* $q = 0$, *then* $p\ln\left(\frac{p}{q}\right) = +\infty$.

Theorem 1. *Given two probability distributions* $\mathbf{p}$ *and* $\mathbf{q}$ *on a same alphabet* $\mathscr{X}$,

$$D(\mathbf{p}||\mathbf{q}) \geq 0 \tag{11}$$

Moreover, the equality holds if and only if $\mathbf{p} = \mathbf{q}$.

Kullback-Leibler divergence is a measure of the difference between two distributions on a common alphabet. So, one measure of the degree of dependence between two random variables X and Y with a joint distribution $\mathbf{p}_{\mathrm{X,Y}}$ is the Kullback-Leibler divergence between $\mathbf{p}_{\mathrm{X,Y}}$ and $\mathbf{p}_{\mathrm{XY}}$ on $\mathscr{X} \times \mathscr{Y}$. This is reasonable and intuitive because $D(\mathbf{p}||\mathbf{q}) = 0$ if and only if $\mathbf{p} = \mathbf{q}$ (by Theorem 1) and correspondingly, X and Y are independent if and only if $\mathbf{p}_{\mathrm{X,Y}} = \mathbf{p}_{\mathrm{XY}}$. We call this measure mutual information.

Definition 4. *The mutual information of random elements,* $(X,Y) \in \mathscr{X} \times \mathscr{Y}$ *with joint probability distribution* $\mathbf{p}_{X,Y}$ *is defined as*

$$MI(X,Y) = D(\mathbf{p}_{X,Y}||\mathbf{p}_{XY}) = \sum_{i\geq 1}\sum_{j\geq 1} p_{i,j} \ln\left(\frac{p_{i,j}}{p_{i,.}p_{.,j}}\right) \tag{12}$$

Lemma 2. *For any joint distribution* $\mathbf{p}_{X,Y}$ *on* $\mathscr{X} \times \mathscr{Y}$,

$$MI(X,Y) \geq 0 \tag{13}$$

Moreover, the equality holds if and only if X *and* Y *are independent.*

Lemma 3. *Suppose* $H(X,Y) < \infty$ *for a joint distribution* $\mathbf{p}_{X,Y}$ *on* $\mathscr{X} \times \mathscr{Y}$. *Then*

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \tag{14}$$

Standardized Mutual Information

At this point, the reader may be beginning to understand how mutual information is a meaningful tool to measure the degree of dependence between two random variables. However, as noted above, it may not be the best one. Observe that mutual information is unbounded above, since entropies can be arbitrarily large. This makes it difficult to convey the strength of the association between X and Y, even when it is known to be positive. We now have given enough background information to introduce the following alleviation of this issue: a standardized mutual information between X and Y, κ.

Definition 5. *Let X and Y be two random variables on finite alphabets $\mathscr{X}$ and $\mathscr{Y}$. Then the standardized mutual information is given by*

$$\kappa = \frac{MI(X,Y)}{H(X,Y)} = \frac{H(X)+H(Y)-H(X,Y)}{H(X,Y)} = \frac{H(X)+H(Y)}{H(X,Y)} - 1. \tag{15}$$

Definition 6. *Random elements $X \in \mathscr{X}$ and $Y \in \mathscr{Y}$ are said to have a one-to-one correspondence under a joint probability distribution $\mathbf{p}_{X,Y}$ on $\mathscr{X} \times \mathscr{Y}$, if*

1) *For every i satisfying $P(X = x_i) > 0$, there exists a unique j such that $P(Y = y_j|X = x_i) = 1$, and*

2) *For every j satisfying $P(Y = y_j) > 0$, there exists a unique i such that $P(X = x_i|Y = y_j) = 1$.*

Lemma 4. *Random elements $X \in \mathscr{X}$ and $Y \in \mathscr{Y}$ have a one-to-one correspondence under a joint probability distribution $\mathbf{p}_{X,Y}$ on $\mathscr{X} \times \mathscr{Y}$ if and only if $H(X) = H(Y) = H(X,Y)$.*

Proof. If X and Y have a one-to-one correspondence, then for each i satisfying $P(X = x_i) > 0$ there is a unique j such that $P(Y = y_j|X = x_i) = 1$. Let the unique corresponding j be denoted by j_i. Noting that, because X and Y have a one-to-one correspondence, $p_{i,j_i} = p_{i,\cdot}$ and therefore $p_{i,j_i}/p_{i,\cdot} = 1$,

$$\begin{aligned}
H(X,Y) &= -\sum_{i\geq 1, j\geq 1} p_{i,j} \ln p_{i,j} \\
&= -\sum_{j\geq 1}\sum_{i\geq 1} p_{i,\cdot}\left(\frac{p_{i,j}}{p_{i,\cdot}}\right) \ln\left[p_{i,\cdot}\left(\frac{p_{i,j}}{p_{i,\cdot}}\right)\right] \\
&= -\sum_{j\geq 1}\sum_{i\geq 1} p_{i,\cdot}\left(\frac{p_{i,j}}{p_{i,\cdot}}\right) \ln p_{i,\cdot} - \sum_{j\geq 1}\sum_{i\geq 1} p_{i,\cdot}\left(\frac{p_{i,j}}{p_{i,\cdot}}\right) \ln\left(\frac{p_{i,j}}{p_{i,\cdot}}\right) \\
&= -\sum_{i\geq 1} p_{i,\cdot}\left(\frac{p_{i,j_i}}{p_{i,\cdot}}\right) \ln p_{i,\cdot} - \sum_{i\geq 1} p_{i,\cdot}\left(\frac{p_{i,j_i}}{p_{i,\cdot}}\right) \ln\left(\frac{p_{i,j_i}}{p_{i,\cdot}}\right) \\
&= -\sum_{i\geq 1} p_{i,\cdot} \ln p_{i,\cdot} \\
&= H(X)
\end{aligned} \tag{16}$$

Similarly, $H(X,Y) = H(Y)$.
On the other hand, if $H(X) = H(X,Y)$ and $H(Y) = H(X,Y)$, Lemma 1 implies that $H(X|Y) = 0$ and $H(Y|X) = 0$.

By Definition 2, $H(Y|X) = 0$ suggests that $H(Y|X = x_i) = 0$ for every i satisfying $P(X = x_i) > 0$, which in turn implies that the conditional probability distribution of Y given $X = x_i$ puts a probability of mass one on a single point in $\mathscr{Y}$. By symmetry, $H(X|Y) = 0$ implies that the conditional probability distribution of X given $Y = y_j$ also puts a probability of mass one on a single point $\mathscr{X}$. Thus X and Y have a one-to-one correspondence. □

Theorem 2. *Let X and Y be two random variables on finite alphabets $\mathscr{X}$ and $\mathscr{Y}$. Then*

$$0 \leq \kappa \leq 1. \tag{17}$$

Moreover

1) $\kappa = 0$ *if and only if X and Y are independent, and*

2) $\kappa = 1$ *if and only if X and Y have a one-to-one correspondence.*

Proof. By Lemma 2 and Corollary 1,

$$0 \leq MI \leq H(X, Y) \tag{18}$$

Dividing all three parts above by $H(X, Y)$ gives (17). Since $\kappa = 0$ if and only if $MI = 0$, by Lemma 2, $\kappa = 0$ if and only if X and Y are independent.

Assume that X and Y have a one-to-one correspondence. By Lemma 4, $H(X) = H(Y) = H(X, Y)$ and therefore $H(X) + H(Y) = 2H(X, Y)$. This implies that

$$\kappa = \frac{H(X) + H(Y)}{H(X, Y)} - 1 = 2 - 1 = 1.$$

Now suppose that $\kappa = 1$. Then $H(X) + H(Y) = 2H(X, Y)$ and therefore

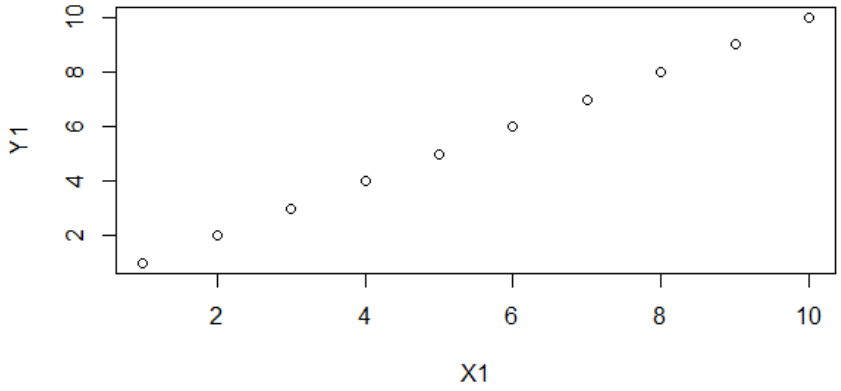
$$(H(X) - H(X, Y)) + (H(Y) - H(X, Y)) = 0.$$

However, by Corollary 1, both of the above additive terms are non-positive, which implies

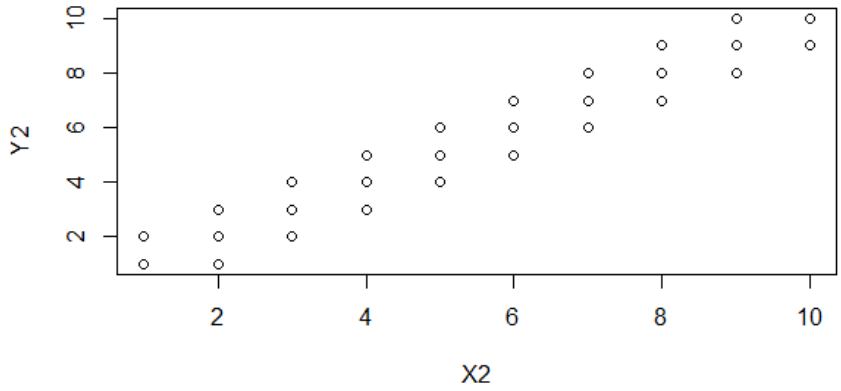
$$H(X) = H(X, Y) \text{ and } H(Y) = H(X, Y)$$

By Lemma 4, this implies that X and Y have a one-to-one correspondence. □

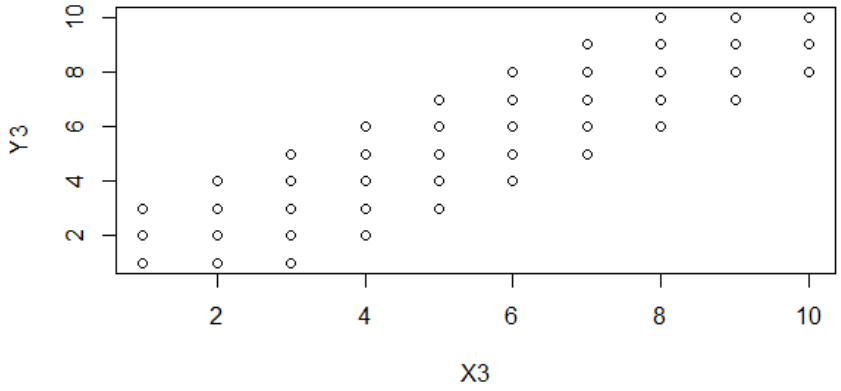
To get an intuitive feel of how specific values of κ may be interpreted, the following ten figures compare κ to ρ, Pearson's correlation coefficient. For each figure, the κ is calculated for a uniform distribution on $\mathscr{X} \times \mathscr{Y}$.



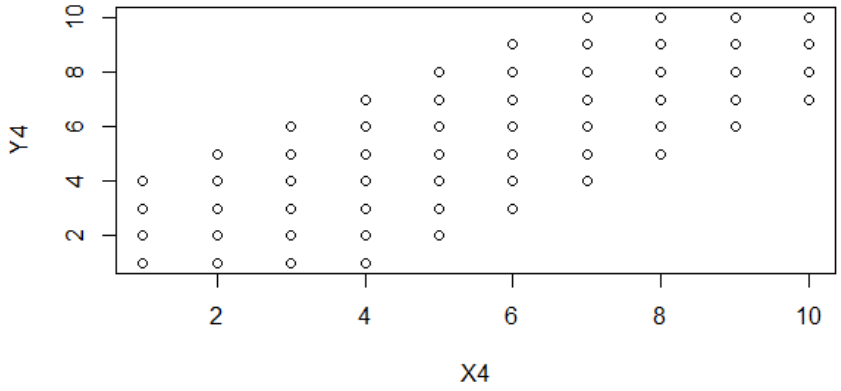
$\rho = 1$
$\kappa = 1$, $p_k = \frac{1}{10}$ for each k



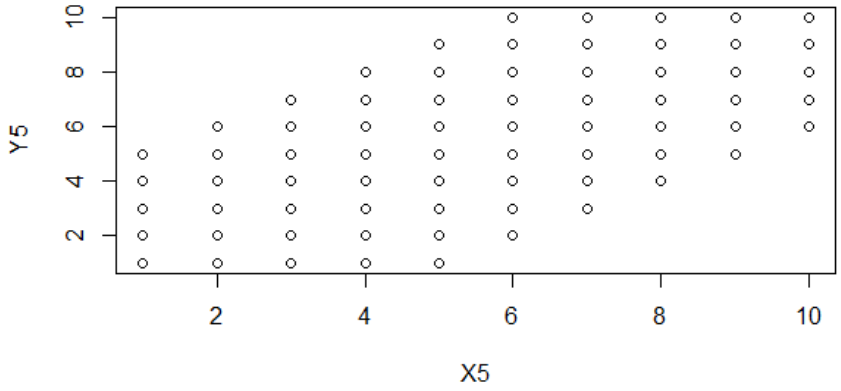
$\rho = 0.9565217$
$\kappa = 0.3753753$, $p_k = \frac{1}{28}$ for each k



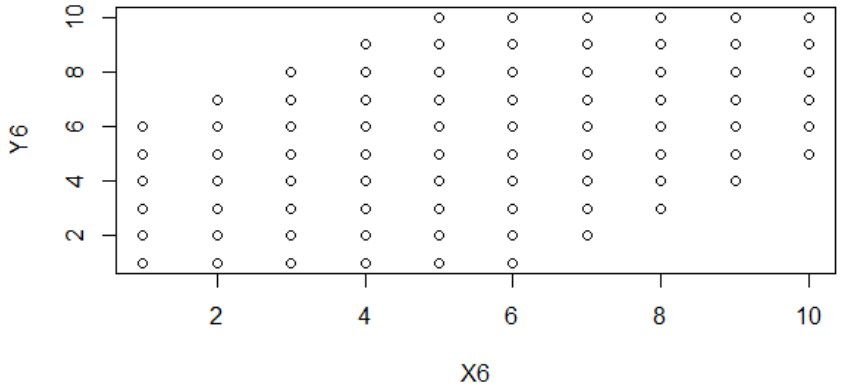
$\rho = 0.8664495$
$\kappa = 0.2076463$, $p_k = \frac{1}{44}$ for each k



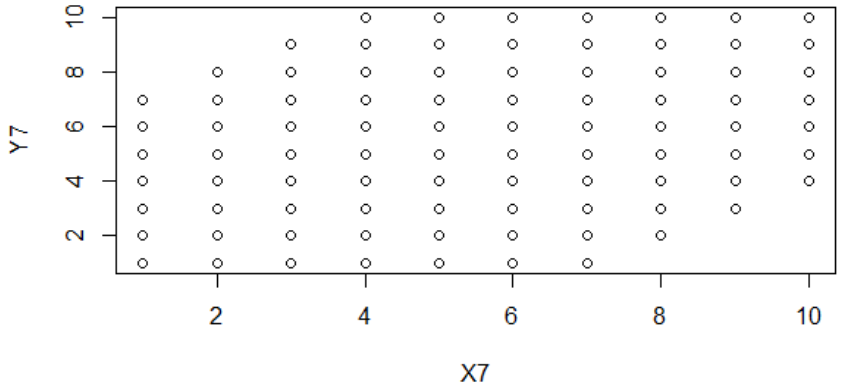
$\rho = 0.7363752$
$\kappa = 0.1238315$, $p_k = \frac{1}{58}$ for each k



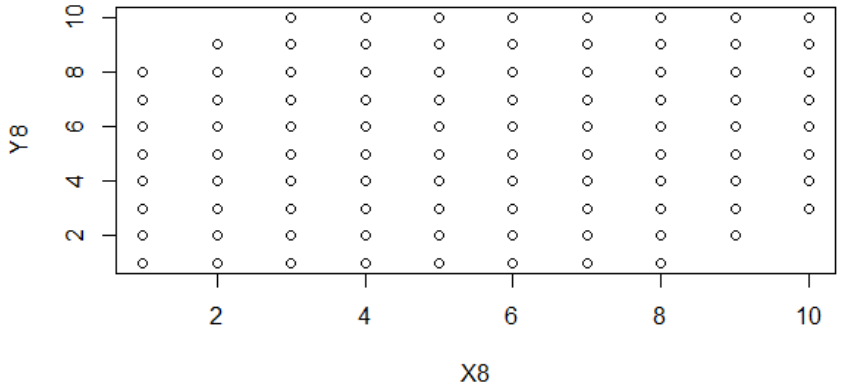
$\rho = 0.5811518$
$\kappa = 0.07423127$, $p_k = \frac{1}{70}$ for each k



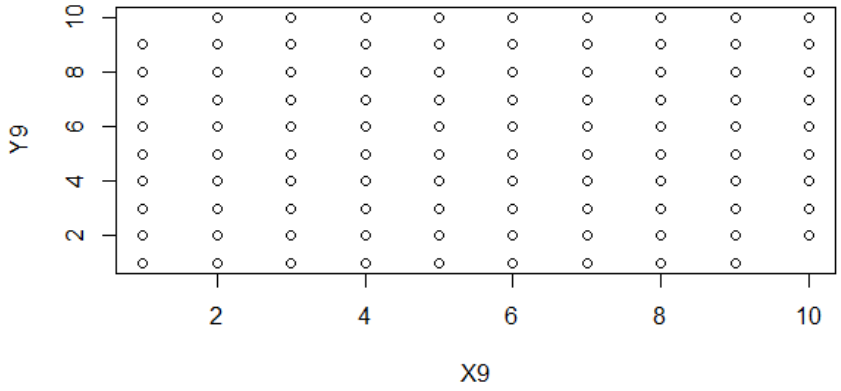
$\rho = 0.4196429$
$\kappa = 0.04372635$, $p_k = \frac{1}{80}$ for each k



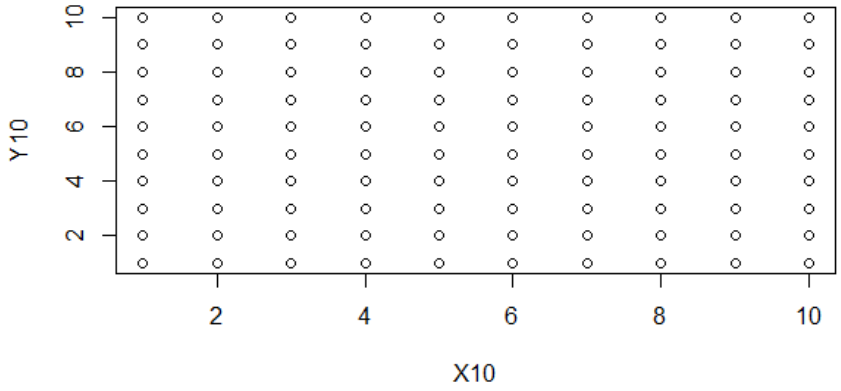
$\rho = 0.2694704$
$\kappa = 0.02454567$, $p_k = \frac{1}{88}$ for each k



$\rho = 0.1438499$
$\kappa = 0.01198232$, $p_k = \frac{1}{94}$ for each k



$\rho = 0.05162524$
$\kappa = 0.004035177$, $p_k = \frac{1}{98}$ for each k



$\rho = 0$
$\kappa = 0$, $p_k = \frac{1}{100}$ for each k

It may be of interest to note that the decrease in κ is very steep compared to that of ρ. When using κ as a measure in applied settings, the fact that κ decreases very sharply may be helpful to keep in mind.

The κ described above is not the only measure that has such desirable properties. For finite alphabets, $H(X,Y) < \infty$, so the following may also be considered as standardized mutual information

$$\kappa_1 = \frac{MI(X,Y)}{\min\{H(X), H(Y)\}}$$

$$\kappa_2 = \frac{MI(X,Y)}{\sqrt{H(X)H(Y)}}$$

$$\kappa_3 = \frac{MI(X,Y)}{(H(X)+H(Y))/2}$$

$$\kappa_4 = \frac{MI(X,Y)}{\max\{H(X),H(Y)\}}.$$

Lemma 5. *Given that $H(X,Y) < \infty$ for finite alphabets,*

$$0 \leq MI(X,Y) \leq \min\{H(X),H(Y)\} \leq \sqrt{H(X)H(Y)} \leq \frac{H(X)+H(Y)}{2} \leq \max\{H(X),H(Y)\} \leq H(X,Y)$$

Proof. By Lemma 2, Lemma 3, and Corollary 1, we have

$$0 \leq MI(X,Y) = H(X) + H(Y) - H(X,Y) \leq H(X) + H(Y) - H(X) = H(Y)$$

and

$$0 \leq MI(X,Y) = H(X) + H(Y) - H(X,Y) \leq H(X) + H(Y) - H(Y) = H(X)$$

Thus

$$0 \leq MI(X,Y) \leq \min\{H(X),H(Y)\}. \tag{19}$$

Note that

$$0 \leq MI(X,Y) \leq \min\{H(X),H(Y)\} \leq H(X)$$

and

$$0 \leq MI(X,Y) \leq \min\{H(X),H(Y)\} \leq H(Y)$$

And therefore

$$(\min\{H(X),H(Y)\})^2 \leq H(X)H(Y)$$

which implies that

$$\min\{H(X),H(Y)\} \leq \sqrt{H(X)H(Y)}. \tag{20}$$

To show that $\sqrt{H(X)H(Y)} \leq (H(X)+H(Y))/2$, consider the following. The first inequality is clearly always true. Each of the subsequent inequalities follows directly from the previous one.

$$\begin{aligned}
0 &\leq (H(X)-H(Y))^2 \\
0 &\leq H^2(X) - 2H(X)H(Y) + H^2(Y) \\
0 &\leq H^2(X) + 2H(X)H(Y) + H^2(Y) - 4H(X)H(Y) \\
4H(X)H(Y) &\leq (H(X)+H(Y))^2 \\
2\sqrt{H(X)H(Y)} &\leq H(X)+H(Y) \\
\sqrt{H(X)H(Y)} &\leq (H(X)+H(Y))/2
\end{aligned}$$

Thus we have

$$\sqrt{H(X)H(Y)} \leq \frac{H(X)+H(Y)}{2} \tag{21}$$

Finally, suppose that $\max\{H(X),H(Y)\} = H(Y)$, in other words $H(X) \leq H(Y)$. Then $H(X) + H(Y) \leq 2H(Y)$, which implies that

$$\frac{H(X)+H(Y)}{2} \leq H(Y) = \max\{H(X), H(Y)\}.$$

A similar argument would show that

$$\frac{H(X)+H(Y)}{2} \leq \max\{H(X), H(Y)\}. \tag{22}$$

if $H(X)$ is the maximum. This proves our last inequality.

□

Corollary 2. *Assuming that $H(X,Y) < \infty$, for $i = 1, 2, 3, 4$,*

$$0 \leq \kappa_i \leq 1$$

And

1) *For $i = 1, 2, 3, 4$, $\kappa_i = 0$ if and only if X and Y are independent, and*

2) *For $i = 2, 3, 4$, $\kappa_i = 1$ if and only if X and Y have a one-to-one correspondence.*

Proof. For $i = 1, 2, 3, 4$, $\kappa_i = 0$ if and only if $MI = 0$. So by (2), for each i, $\kappa_i = 0$ if and only if X and Y are independent.

By Lemma 5,

$$0 \leq MI(X,Y) \leq \min\{H(X), H(Y)\}$$

and so

$$0 \leq \frac{MI(X,Y)}{\min\{H(X), H(Y)\}} \leq 1$$

ie,

$$0 \leq \kappa_1 \leq 1.$$

Next we consider κ_2. By Lemma 5,

$$0 \leq MI(X,Y) \leq \sqrt{H(X)H(Y)}$$

and so

$$0 \leq \frac{MI(X,Y)}{\sqrt{H(X)H(Y)}} \leq 1$$

ie,

$$0 \leq \kappa_2 \leq 1.$$

By Lemma 4, if X and Y have a one-to-one correspondence, $H(X) = H(Y) = H(X,Y)$ and thus

$$\kappa_2 = \frac{H(X)+H(Y)-H(X,Y)}{\sqrt{H(X)H(Y)}} = \frac{H(X,Y)}{\sqrt{H(X,Y)H(X,Y)}} = \frac{H(X,Y)}{H(X,Y)} = 1.$$

Now suppose $\kappa_2 = 1$. Then $H(X)+H(Y)-H(X,Y) = \sqrt{H(X)H(Y)}$. Squaring both sides, we obtain

$$H^2(X)+H^2(Y)+H^2(X,Y)+2H(X)H(Y)-2H(X)H(X,Y)-2H(Y)H(X,Y) = H(X)H(Y)$$

Doing some algebra, we find that this is equivalent to

$$0 = (H(X) + H(Y) - H(X,Y))(H(X) - H(X,Y)) + H(Y)(H(Y) - H(X,Y))$$

By Corollary 1, $H(X) - H(X,Y) \leq 0$ and $H(Y) - H(X,Y) \leq 0$. By Lemma 2 and the fact that $MI(X,Y) \neq 0$ if $\kappa_2 = 1$, $H(X) + H(Y) - H(X,Y) = MI(X,Y) > 0$. It is safe to assume that $H(X) \neq 0$, and thus that $H(X) > 0$. This means that each additive term is non-positive, and since the sum is equal to 0, this implies that each term must be zero. Hence $H(X) - H(X,Y) = H(Y) - H(X,Y) = 0$, and therefore $H(X) = H(X,Y)$ and $H(Y) = H(X,Y)$. By Lemma 4, this implies that X and Y have a one-to-one correspondence.

Next, consider κ_3. By Lemma 5,

$$0 \leq MI(X,Y) \leq \frac{H(X)H(Y)}{2}$$

and so

$$0 \leq \frac{MI(X,Y)}{(H(X)H(Y))/2} \leq 1$$

ie,

$$0 \leq \kappa_3 \leq 1.$$

By Lemma 4, if X and Y have a one-to-one correspondence, $H(X) = H(Y) = H(X,Y)$ and thus

$$\kappa_3 = \frac{H(X) + H(Y) - H(X,Y)}{(H(X) + H(Y))/2} = \frac{2(H(X,Y) + H(X,Y) - H(X,Y))}{H(X,Y) + H(X,Y)} = \frac{2H(X,Y)}{2H(X,Y)} = 1.$$

Now suppose $\kappa_3 = 1$. Then

$$\begin{aligned} 2(H(X) + H(Y) - H(X,Y)) &= H(X) + H(Y) \\ H(X) + H(Y) - 2H(X,Y) &= 0 \\ H(X) + H(Y) &= 2H(X,Y) \end{aligned}$$

By the same argument used in the proof of Theorem 2, $H(X) + H(Y) = 2H(X,Y)$ implies that $H(X) = H(Y) = H(X,Y)$. By Lemma 4, this means that X and Y have a one-to-one correspondence.

Lastly, consider κ_4. By Lemma 5,

$$0 \leq MI(X,Y) \leq \max\{H(X), H(Y)\}$$

and so

$$0 \leq \frac{MI(X,Y)}{\max\{H(X), H(Y)\}} \leq 1$$

ie,

$$0 \leq \kappa_4 \leq 1.$$

By Lemma 4, if X and Y have a one-to-one correspondence, $H(X) = H(Y) = H(X,Y)$ and thus

$$\kappa_4 = \frac{H(X) + H(Y) - H(X,Y)}{\max\{H(X), H(Y)\}} = \frac{H(X,Y)}{H(X,Y)} = 1.$$

If $\kappa_4 = 1$, then $H(X) + H(Y) - H(X,Y) = \max\{H(X), H(Y)\}$. Without loss of generality, assume that the maximum is $H(Y)$, ie that $H(X) \leq H(Y)$. Then

$$\begin{aligned} H(X) + H(Y) - H(X,Y) &= H(Y) \\ H(X) - H(X,Y) &= 0 \\ H(X) &= H(X,Y) \end{aligned}$$

By Corollary 1, $H(Y) \leq H(X,Y)$. By assumption and the above equation, $H(Y) \geq H(X) = H(X,Y)$. Thus $H(Y) = H(X,Y)$. Since we already have $H(X) = H(X,Y)$, by Lemma 4, X and Y have a one-to-one correspondence. By symmetry, the argument also holds if the maximum is $H(X)$.

□

Estimation and Asymptotic Normality

In this section, we consider the plug-in estimation of the standardized mutual information indices κ, κ_1, κ_2, κ_3, and κ_4, and explore their asymptotic distributions.

For every pair of (i,j), let $f_{i,j}$ be the observed frequency of the random pair (X,Y) taking value (x_i, y_j) for all $1 \leq i \leq K_1$ and all $1 \leq j \leq K_2$ in an iid sample of size n from $\mathscr{X} \times \mathscr{Y}$. Let $\hat{p}_{i,j} = f_{i,j}/n$ be the corresponding relative frequency. Thus we have $\hat{\mathbf{p}}_{\mathrm{X,Y}} = \{\hat{p}_{i,j}\}$, $\hat{\mathbf{p}}_{\mathrm{X}} = \{\hat{p}_{i,\cdot}\}$, and $\hat{\mathbf{p}}_{\mathrm{Y}} = \{\hat{p}_{\cdot,j}\}$ as the sets of observed joint and marginal relative frequencies.

Let $K = K_1 K_2$ be the number of positive joint probabilities in $\{p_{i,j}\}$ for every pair of (i,j) satisfying $1 \leq i \leq K_1$ and $1 \leq j \leq K_2$. That is,

$$K = \sum_{i,j} 1[p_{i,j} > 0] \tag{23}$$

First consider the case of $K = K_1 K_2$, which means $p_{i,j} > 0$ for every pair (i,j) for all $1 \leq i \leq K_1$ and all $1 \leq j \leq K_2$.

Let $\mathbf{p}$ be a specifically arranged $p_{i,j}$ as follows.

$$\mathbf{p} = (p_{1,1}, p_{1,2}, \cdots, p_{1,K_2}, p_{2,1}, p_{2,2}, \cdots, p_{2,K_2}, \cdots, p_{K_1,1}, \cdots, p_{K_1,K_2-1})^\tau.$$

Accordingly let

$$\hat{\mathbf{p}} = (\hat{p}_{1,1}, \hat{p}_{1,2}, \cdots, \hat{p}_{1,K_2}, \hat{p}_{2,1}, \hat{p}_{2,2}, \cdots, \hat{p}_{2,K_2}, \cdots, \hat{p}_{K_1,1}, \cdots, \hat{p}_{K_1,K_2-1})^\tau.$$

For notation convenience, for that specific arrangement of $p_{i,j}$s in $\mathbf{p}$, we may also re-enumerate them by a single index k, as in

$$\mathbf{v} = (p_1, \cdots, p_{K-1})^\tau \tag{24}$$

and

$$\hat{\mathbf{v}} = (\hat{p}_1, \cdots, \hat{p}_{K-1})^\tau \tag{25}$$

where $K = K_1 K_2$.

By the multivariate central limit theorem we know that

$$\sqrt{n}\,(\hat{\mathbf{v}} - \mathbf{v}) \xrightarrow{L} MVN(0, \Sigma) \tag{26}$$

where

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{K-1} \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_{K-1} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ -p_{K-1}p_1 & -p_{K-1}p_2 & \cdots & p_{K-1}(1-p_{K-1}) \end{pmatrix}_{(K-1)\times(K-1)} . \tag{27}$$

Let $\hat{H}(X)$, $\hat{H}(Y)$ and $\hat{H}(X,Y)$ be the plug-in estimators of $H(X)$, $H(Y)$ and $H(X,Y)$ respectively:

$$\hat{H}(X) = -\sum_i \hat{p}_{i,\cdot} \ln \hat{p}_{i,\cdot} \tag{28}$$

$$\hat{H}(Y) = -\sum_j \hat{p}_{\cdot,j} \ln \hat{p}_{\cdot,j} \tag{29}$$

$$\hat{H}(X,Y) = -\sum_i \sum_j \hat{p}_{i,j} \ln \hat{p}_{i,j} = -\sum_k \hat{p}_k \ln \hat{p}_k \tag{30}$$

The respective gradients for $G_{\mathrm{X}}(\mathbf{v})$, $G_{\mathrm{Y}}(\mathbf{v})$ and $G_{\mathrm{XY}}(\mathbf{v})$, as in (7) are

$$g_{\mathrm{X}}(\mathbf{v}) = \bigtriangledown G_{\mathrm{X}}(\mathbf{v}) = \left(\frac{\partial}{\partial p_1} G_{\mathrm{X}}(\mathbf{v}), \cdots, \frac{\partial}{\partial p_{K-1}} G_{\mathrm{X}}(\mathbf{v}) \right)^\tau$$

$$g_{\mathrm{Y}}(\mathbf{v}) = \bigtriangledown G_{\mathrm{Y}}(\mathbf{v}) = \left(\frac{\partial}{\partial p_1} G_{\mathrm{X}}(\mathbf{v}), \cdots, \frac{\partial}{\partial p_{K-1}} G_{\mathrm{Y}}(\mathbf{v}) \right)^\tau$$

$$g_{\mathrm{XY}}(\mathbf{v}) = \bigtriangledown G_{\mathrm{XY}}(\mathbf{v}) = \left(\frac{\partial}{\partial p_1} G_{\mathrm{XY}}(\mathbf{v}), \cdots, \frac{\partial}{\partial p_{K-1}} G_{\mathrm{XY}}(\mathbf{v}) \right)^\tau$$

For every k, $1 \le k \le K-1$ such that $p_k = p_{i,j}$, the following facts can be verified.

$$\frac{\partial}{\partial p_k} G_{\mathrm{X}}(\mathbf{v}) = \frac{\partial H(X)}{\partial p_k} = \ln p_{K_1,\cdot} - \ln p_{i,\cdot} \tag{31}$$

$$\frac{\partial}{\partial p_k} G_{\mathrm{Y}}(\mathbf{v}) = \frac{\partial H(Y)}{\partial p_k} = \ln p_{\cdot,K_2} - \ln p_{\cdot,j} \tag{32}$$

$$\frac{\partial}{\partial p_k} G_{\mathrm{XY}}(\mathbf{v}) = \frac{\partial H(X,Y)}{\partial p_k} = \ln p_K - \ln p_k \tag{33}$$

Let

$$A = \begin{pmatrix} \frac{\partial}{\partial p_1} G_{\mathrm{X}}(\mathbf{v}) & \cdots & \frac{\partial}{\partial p_{K-1}} G_{\mathrm{X}}(\mathbf{v}) \\ \frac{\partial}{\partial p_1} G_{\mathrm{Y}}(\mathbf{v}) & \cdots & \frac{\partial}{\partial p_{K-1}} G_{\mathrm{Y}}(\mathbf{v}) \\ \frac{\partial}{\partial p_1} G_{\mathrm{XY}}(\mathbf{v}) & \cdots & \frac{\partial}{\partial p_{K-1}} G_{\mathrm{XY}}(\mathbf{v}) \end{pmatrix}_{3\times(K-1)} \tag{34}$$

An application of the multivariate delta method yields the following lemma.

Lemma 6. *Let $\mathscr{X}$ and $\mathscr{Y}$ be as in (1), let $\mathbf{p}_{X,Y} = \{p_{i,j}\}$ be a joint probability distribution on $\mathscr{X} \times \mathscr{Y}$, and let $\mathbf{v}$ and $\hat{\mathbf{v}}$ be as in (24) and (25).*

$$\sqrt{n}\left[\begin{pmatrix} \hat{H}(X) \\ \hat{H}(Y) \\ \hat{H}(X,Y) \end{pmatrix} - \begin{pmatrix} H(X) \\ H(Y) \\ H(X,Y) \end{pmatrix}\right] \xrightarrow{L} MVN(0, \Sigma_H) \tag{35}$$

where

$$\Sigma_H = A\Sigma A^\tau$$

Σ is as in (27) and A is as in (34).

More generally, we can consider the case where $K \leq K_1 K_2$, which by (23) means that $p_{i,j}$ may be zero for some pairs of (i, j). We will derive a result corresponding to Lemma 6 for this case. For any arbitrary but fixed re-enumeration of the K positive probabilities in $\{p_{i,j}\}$, denoted as

$$\{p_k; k = 1, \cdots, K\} \tag{36}$$

consider the following two partitions

$$\{S_1, \cdots, S_{K_1}\} \qquad \text{and} \qquad \{T_1, \cdots, T_{K_2}\}$$

of the index set $\{1, 2, \cdots, K\}$ such that

1. $\{p_k; k \in S_s\}$ is the collection of all positive probabilities in $\{p_{i,j}; i = s\}$ for each $s, s = 1, \cdots, K_1$; and
2. $\{p_k; k \in T_t\}$ is the collection of all positive probabilities in $\{p_{i,j}; j = t\}$ for each $t, t = 1, \cdots, K_2$.

By construction of the partitions,

$$\textstyle\sum_{k \in S_i} p_k = p_{i,\cdot} \qquad \text{and} \qquad \sum_{k \in T_j} p_k = p_{\cdot,j}.$$

Without loss of generality, it may be assumed that $K \in S_{K_1} \cap T_{K_2}$. If not, then $K \in S_{i_0} \cap T_{j_0}$ for some i_0 and j_0, by a re-arrangement of the indices (i, j), $K \in S_{K_1} \cap T_{K_2}$ will be true. Letting

$$\mathbf{v} = (p_1, \cdots, p_{K-1})^\tau \qquad \text{and} \qquad \hat{\mathbf{v}} = (\hat{p}_1, \cdots, \hat{p}_{K-1})^\tau \tag{37}$$

an application of multivariate central limit theorem gives

$$\sqrt{n}\,(\hat{\mathbf{v}} - \mathbf{v}) \xrightarrow{L} MVN(0, \Sigma) \tag{38}$$

where Σ is the covariance matrix given by

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{K-1} \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_{K-1} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ -p_{K-1}p_1 & -p_{K-1}p_2 & \cdots & p_{K-1}(1-p_{K-1}) \end{pmatrix}_{(K-1)\times(K-1)} . \tag{39}$$

Referring to (7), the following facts can be verified.

$$\frac{\partial G_{\mathrm{X}}}{\partial p_k} = \frac{\partial H(X)}{\partial p_k} = \begin{cases} \ln p_{K_1,\cdot} - \ln p_{i,\cdot} & \text{if } k \in S_i \neq S_{K_1} \\ 0 & \text{if } k \in S_{K_1} \end{cases} \tag{40}$$

$$\frac{\partial G_{\mathrm{Y}}}{\partial p_k} = \frac{\partial H(Y)}{\partial p_k} = \begin{cases} \ln p_{\cdot,K_2} - \ln p_{\cdot,j} & \text{if } k \in T_j \neq T_{K_2} \\ 0 & \text{if } k \in T_{K_2} \end{cases} \tag{41}$$

$$\frac{\partial G_{\mathrm{XY}}}{\partial p_k} = \frac{\partial H(X,Y)}{\partial p_k} = \ln p_K - \ln p_k, \text{ for } 1 \leq k \leq K-1. \tag{42}$$

For any arbitrary but fixed enumeration of the positive terms of $\{p_{i,j}\}$ in (36), we can apply the multivariate delta method based on (38) to give the following lemma.

Lemma 7. *Let $\mathscr{X}$ and $\mathscr{Y}$ be as in (1), let $\mathbf{p}_{X,Y} = \{p_{i,j}\}$ be a joint probability distribution on $\mathscr{X} \times \mathscr{Y}$, and let v and $\hat{v}$ be as in (37).*

$$\sqrt{n}\left[\begin{pmatrix} \hat{H}(X) \\ \hat{H}(Y) \\ \hat{H}(X,Y) \end{pmatrix} - \begin{pmatrix} H(X) \\ H(Y) \\ H(X,Y) \end{pmatrix}\right] \xrightarrow{L} MVN(0, \Sigma_H) \tag{43}$$

where

$$\Sigma_H = A\Sigma A^\tau \tag{44}$$

with Σ as in (39) and A is as in (34) according to (40), (41) and (42).

Note that

$$MI(X,Y) = (1,1,-1)\begin{pmatrix} H(X) \\ H(Y) \\ H(X,Y) \end{pmatrix}$$

and let

$$\widehat{MI}(X,Y) = (1,1,-1)\begin{pmatrix} \hat{H}(X) \\ \hat{H}(Y) \\ \hat{H}(X,Y) \end{pmatrix} \tag{45}$$

be the plug-in estimator for mutual information. Then an application of the multivariate delta method based on (43) gives the following theorem.

Theorem 3. *Suppose that*

$$\sigma^2 = (1,1,-1)\Sigma(1,1,-1)^\tau > 0 \tag{46}$$

where Σ is as in (39). Then under the conditions of Lemma 7,

$$\sqrt{n}(\widehat{MI} - MI) \xrightarrow{L} N(0, \sigma^2) \tag{47}$$

Let

$$\hat{\Sigma} = \Sigma(\hat{\mathbf{v}}) = \begin{pmatrix} \hat{p}_1(1-\hat{p}_1) & -\hat{p}_1\hat{p}_2 & \cdots & -\hat{p}_1\hat{p}_{K-1} \\ -\hat{p}_1\hat{p}_2 & \hat{p}_2(1-\hat{p}_2) & \cdots & -\hat{p}_2\hat{p}_{K-1} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ -\hat{p}_{K-1}\hat{p}_1 & -\hat{p}_{K-1}\hat{p}_2 & \cdots & \hat{p}_{K-1}(1-\hat{p}_{K-1}) \end{pmatrix}_{(K-1)\times(K-1)} \tag{48}$$

$$\hat{A} = A(\hat{\mathbf{v}}) = \begin{pmatrix} \frac{\partial}{\partial p_1} G_{\mathrm{X}}(\hat{\mathbf{v}}) & \cdots & \frac{\partial}{\partial p_{K-1}} G_{\mathrm{X}}(\hat{\mathbf{v}}) \\ \frac{\partial}{\partial p_1} G_{\mathrm{Y}}(\hat{\mathbf{v}}) & \cdots & \frac{\partial}{\partial p_{K-1}} G_{\mathrm{Y}}(\hat{\mathbf{v}}) \\ \frac{\partial}{\partial p_1} G_{\mathrm{XY}}(\hat{\mathbf{v}}) & \cdots & \frac{\partial}{\partial p_{K-1}} G_{\mathrm{XY}}(\hat{\mathbf{v}}) \end{pmatrix}_{3\times(K-1)} \tag{49}$$

$$\hat{\Sigma}_H = \hat{A}\hat{\Sigma}\hat{A}^{\tau} \tag{50}$$

$$\hat{\sigma}^2 = (1,1,-1)\hat{\Sigma}_H(1,1,-1)^{\tau} \tag{51}$$

Having shown the above lemma and theorem, we are now at the point where we can show the asymptotic normality of the estimators for κ, κ_2 and κ_3. For this purpose, consider the following functions, with the same domain: $0 < x_1 < \infty$, $0 < x_2 < \infty$ and $0 < x_3 < \infty$.

$$\begin{aligned} \kappa(x_1,x_2,x_3) =& \frac{x_1+x_2}{x_3} - 1, \\ \kappa_2(x_1,x_2,x_3) =& \frac{x_1+x_2-x_3}{\sqrt{x_1x_2}}, \\ \kappa_3(x_1,x_2,x_3) =& 2\left(1 - \frac{x_3}{x_1+x_2}\right). \end{aligned}$$

The gradients of these functions are, respectively

$$\begin{aligned} g_{\kappa}(x_1,x_2,x_3) &= \left(\frac{1}{x_3}, \frac{1}{x_3}, -\frac{x_1+x_2}{x_3^2}\right)^{\tau}, \\ g_{\kappa_2}(x_1,x_2,x_3) &= \left(\frac{1}{(x_1x_2)^{1/2}} - \frac{x_2(x_1+x_2-x_3)}{2(x_1x_2)^{3/2}}, \frac{1}{(x_1x_2)^{1/2}} - \frac{x_1(x_1+x_2-x_3)}{2(x_1x_2)^{3/2}}, -\frac{1}{\sqrt{x_1x_2}}\right)^{\tau}, \\ g_{\kappa_3}(x_1,x_2,x_3) &= \left(-\frac{2x_3}{(x_1+x_2)^2}, -\frac{2x_3}{(x_1+x_2)^2}, -\frac{2}{x_1+x_2},\right)^{\tau}. \end{aligned}$$

Let the following be the plug-in estimators (MLEs) of κ, κ_2 and κ_3.

$$\hat{\kappa} = \frac{\hat{H}(X)+\hat{H}(Y)}{\hat{H}(X,Y)} - 1 = \frac{\hat{H}(X)+\hat{H}(Y)-\hat{H}(X,Y)}{\hat{H}(X,Y)}, \tag{52}$$

$$\hat{\kappa}_2 = \frac{\hat{H}(X)+\hat{H}(Y)-\hat{H}(X,Y)}{\sqrt{\hat{H}(X)\hat{H}(Y)}}, \tag{53}$$

$$\hat{\kappa}_3 = 2\left(1 - \frac{\hat{H}(X,Y)}{\hat{H}(X)+\hat{H}(Y)}\right) = \frac{\hat{H}(X)+\hat{H}(Y)-\hat{H}(X,Y)}{(\hat{H}(X)+\hat{H}(Y))/2}. \tag{54}$$

Here, we can apply the multivariate delta method directly using (43) to obtain the following theorem.

Theorem 4. *Under the conditions of Lemma 7,*

1. $\sqrt{n}(\hat{\kappa}-\kappa) \xrightarrow{L} N(0,\sigma^2_{\hat{\kappa}})$,
2. $\sqrt{n}(\hat{\kappa}_2-\kappa_2) \xrightarrow{L} N(0,\sigma^2_{\hat{\kappa}_2})$,
3. $\sqrt{n}(\hat{\kappa}_3-\kappa_3) \xrightarrow{L} N(0,\sigma^2_{\hat{\kappa}_3})$,

where

$$\begin{aligned}
\sigma^2_{\hat{\kappa}} &= g^\tau_\kappa(H(X),H(Y),H(X,Y))\Sigma_H g_\kappa(H(X),H(Y),H(X,Y)),\\
\sigma^2_{\hat{\kappa}_2} &= g^\tau_{\kappa_2}(H(X),H(Y),H(X,Y))\Sigma_H g_{\kappa_2}(H(X),H(Y),H(X,Y)),\\
\sigma^2_{\hat{\kappa}_3} &= g^\tau_{\kappa_3}(H(X),H(Y),H(X,Y))\Sigma_H g_{\kappa_3}(H(X),H(Y),H(X,Y)).
\end{aligned}$$

with Σ_H is as in (44).

Corollary 3. *Under the conditions of Theorem 4,*

1. $\frac{\sqrt{n}(\hat{\kappa}-\kappa)}{\hat{\sigma}_{\hat{\kappa}}} \xrightarrow{L} N(0,1)$,
2. $\frac{\sqrt{n}(\hat{\kappa}_2-\kappa_2)}{\hat{\sigma}_{\hat{\kappa}_2}} \xrightarrow{L} N(0,1)$,
3. $\frac{\sqrt{n}(\hat{\kappa}_3-\kappa_3)}{\hat{\sigma}_{\hat{\kappa}_3}} \xrightarrow{L} N(0,1)$,

where $\hat{\Sigma}_H$ is such that every $p_{i,j}$ in Σ_H is substituted by $\hat{p}_{i,j}$ and

$$\begin{aligned}
\hat{\sigma}^2_{\hat{\kappa}} &= g^\tau_\kappa(\hat{H}(X),\hat{H}(Y),\hat{H}(X,Y))\hat{\Sigma}_H g_\kappa(\hat{H}(X),\hat{H}(Y),\hat{H}(X,Y)),\\
\hat{\sigma}^2_{\hat{\kappa}_2} &= g^\tau_{\kappa_2}(\hat{H}(X),\hat{H}(Y),\hat{H}(X,Y))\hat{\Sigma}_H g_{\kappa_2}(\hat{H}(X),\hat{H}(Y),\hat{H}(X,Y)),\\
\hat{\sigma}^2_{\hat{\kappa}_3} &= g^\tau_{\kappa_3}(\hat{H}(X),\hat{H}(Y),\hat{H}(X,Y))\hat{\Sigma}_H g_{\kappa_3}(\hat{H}(X),\hat{H}(Y),\hat{H}(X,Y)).
\end{aligned}$$

Finding the asymptotic distributions for κ_1 and κ_4 turns out to be a bit more complex than in the other cases. This is because of the sharp edge that occurs for both of these indices where $H(X) = H(Y)$, and so the respective gradients are different depending on whether $H(X) < H(Y)$ or $H(X) > H(Y)$. Let us define the following functions for x_1, x_2, and x_3 with the constraints that $0 < x_1 \le x_3 < \infty$ and $0 < x_2 \le x_3 < \infty$.

$$\kappa_1(x_1,x_2,x_3) = \begin{cases} \dfrac{x_1+x_2-x_3}{x_1} & x_1 < x_2 \\ \dfrac{x_1+x_2-x_3}{x_2} & x_1 > x_2 \\ \dfrac{x_1+x_2-x_3}{(x_1+x_2)/2} = 2\left(1-\dfrac{x_3}{x_1+x_2}\right) & x_1 = x_2 \end{cases} \tag{55}$$

$$\kappa_4(x_1,x_2,x_3)=\begin{cases}\dfrac{x_1+x_2-x_3}{x_1} & x_1>x_2\\[2ex] \dfrac{x_1+x_2-x_3}{x_2} & x_1<x_2\\[2ex] \dfrac{x_1+x_2-x_3}{(x_1+x_2)/2}=2\left(1-\dfrac{x_3}{x_1+x_2}\right) & x_1=x_2\end{cases} \tag{56}$$

Note that (55) and (56) are equivalent to the following:

$$\kappa_1(x_1,x_2,x_3)=\frac{x_1+x_2-x_3}{\min\{x_1,x_2\}}$$

$$\kappa_4(x_1,x_2,x_3)=\frac{x_1+x_2-x_3}{\max\{x_1,x_2\}}$$

The gradients of these functions are, respectively

$$g_{\kappa_1,x_1<x_2}(x_1,x_2,x_3)=\left(\frac{x_3-x_2}{x_1^2},\frac{1}{x_1},-\frac{1}{x_1}\right)^\tau$$

$$g_{\kappa_1,x_1>x_2}(x_1,x_2,x_3)=\left(\frac{1}{x_2},\frac{x_3-x_1}{x_2^2},-\frac{1}{x_2}\right)^\tau$$

$$g_{\kappa_4,x_1<x_2}(x_1,x_2,x_3)=\left(\frac{1}{x_2},\frac{x_3-x_1}{x_2^2},-\frac{1}{x_2}\right)^\tau$$

$$g_{\kappa_4,x_1>x_2}(x_1,x_2,x_3)=\left(\frac{x_3-x_2}{x_1^2},\frac{1}{x_1},-\frac{1}{x_1}\right)^\tau$$

Let the following be the plug-in estimators (MLEs) of κ_1 and κ_4.

$$\hat{\kappa}_1=\frac{\hat{H}(X)+\hat{H}(Y)-\hat{H}(X,Y)}{\min\{\hat{H}(X),\hat{H}(Y)\}}, \tag{57}$$

$$\hat{\kappa}_4=\frac{\hat{H}(X)+\hat{H}(Y)-\hat{H}(X,Y)}{\max\{\hat{H}(X),\hat{H}(Y)\}}, \tag{58}$$

Theorem 5. *Suppose $H(X)\neq H(Y)$. Then under the conditions of Lemma 7,*

1. $\sqrt{n}(\hat{\kappa}_1-\kappa_1)\xrightarrow{L}N(0,\sigma^2_{\hat{\kappa}_1})$,
2. $\sqrt{n}(\hat{\kappa}_4-\kappa_4)\xrightarrow{L}N(0,\sigma^2_{\hat{\kappa}_4})$

where, if $H(X)<H(Y)$,

$$\sigma^2_{\hat{\kappa}_1} = g^{\tau}_{\kappa_1, x_1<x_2}(H(X), H(Y), H(X,Y))\Sigma_H g_{\kappa_1, x_1<x_2}(H(X), H(Y), H(X,Y)),$$
$$\sigma^2_{\hat{\kappa}_4} = g^{\tau}_{\kappa_4, x_1<x_2}(H(X), H(Y), H(X,Y))\Sigma_H g_{\kappa_4, x_1<x_2}(H(X), H(Y), H(X,Y))$$

and if $H(X) > H(Y)$*, then*

$$\sigma^2_{\hat{\kappa}_1} = g^{\tau}_{\kappa_1, x_1>x_2}(H(X), H(Y), H(X,Y))\Sigma_H g_{\kappa_1, x_1>x_2}(H(X), H(Y), H(X,Y)),$$
$$\sigma^2_{\hat{\kappa}_4} = g^{\tau}_{\kappa_4, x_1>x_2}(H(X), H(Y), H(X,Y))\Sigma_H g_{\kappa_4, x_1>x_2}(H(X), H(Y), H(X,Y))$$

with Σ_H *is as in (44).*

Corollary 4. *Under the conditions of Theorem 5,*

1. $\frac{\sqrt{n}(\hat{\kappa}_1 - \kappa_1)}{\hat{\sigma}_{\hat{\kappa}_1}} \xrightarrow{L} N(0,1)$,
2. $\frac{\sqrt{n}(\hat{\kappa}_4 - \kappa_4)}{\hat{\sigma}_{\hat{\kappa}_4}} \xrightarrow{L} N(0,1)$

where $\hat{\Sigma}_H$ *is such that every* $p_{i,j}$ *in* Σ_H *is substituted by* $\hat{p}_{i,j}$*. If* $H(X) < H(Y)$*,*

$$\hat{\sigma}^2_{\hat{\kappa}_1} = g^{\tau}_{\kappa_1, x_1<x_2}(\hat{H}(X), \hat{H}(Y), \hat{H}(X,Y))\hat{\Sigma}_H g_{\kappa_1, x_1<x_2}(\hat{H}(X), \hat{H}(Y), \hat{H}(X,Y)),$$
$$\hat{\sigma}^2_{\hat{\kappa}_4} = g^{\tau}_{\kappa_4, x_1<x_2}(\hat{H}(X), \hat{H}(Y), \hat{H}(X,Y))\hat{\Sigma}_H g_{\kappa_4, x_1<x_2}(\hat{H}(X), \hat{H}(Y), \hat{H}(X,Y)).$$

and if $H(X) > H(Y)$*,*

$$\hat{\sigma}^2_{\hat{\kappa}_1} = g^{\tau}_{\kappa_1, x_1>x_2}(\hat{H}(X), \hat{H}(Y), \hat{H}(X,Y))\hat{\Sigma}_H g_{\kappa_1, x_1>x_2}(\hat{H}(X), \hat{H}(Y), \hat{H}(X,Y)),$$
$$\hat{\sigma}^2_{\hat{\kappa}_4} = g^{\tau}_{\kappa_4, x_1>x_2}(\hat{H}(X), \hat{H}(Y), \hat{H}(X,Y))\hat{\Sigma}_H g_{\kappa_4, x_1>x_2}(\hat{H}(X), \hat{H}(Y), \hat{H}(X,Y)).$$

Estimation and Asymptotic Normality from Turing's Perspective

As an alternative to the plug-in estimators, there are a set of estimators of κ, κ_1, κ_2, κ_3 and κ_4 that approach the problem through the perspective of Alan Turing.
Let

$$Z_{1,v} = \frac{n^{1+v}[n-(1+v)]!}{n!} \sum_{k \le 1} \left[\hat{p}_k \prod_{j=0}^{v-1} \left(1 - \hat{p}_k - \frac{j}{n}\right) \right]$$

Then an estimator of entropy based on Turing's perspective is

$$\hat{H}_z = \sum_{v=1}^{n-1} \frac{1}{v} Z_{1,v} \tag{59}$$

This naturally extends to an estimator of mutual information based on Turing's perspective.

$$\widehat{MI}_z(X,Y) = \hat{H}_z(X) + \hat{H}_z(Y) - \hat{H}_z(X,Y) = \sum_{v=1}^{n-1} \frac{1}{v} \left\{ \frac{n^{v+1}[n-(v+1)]!}{n!} \sum_{i=1}^{K_1} \left[\hat{p}_{i,\cdot} \prod_{k=0}^{v-1} \left(1 - \hat{p}_{i,\cdot} - \frac{k}{n} \right) \right] \right\}$$
$$+ \sum_{v=1}^{n-1} \frac{1}{v} \left\{ \frac{n^{v+1}[n-(v+1)]!}{n!} \sum_{j=1}^{K_2} \left[\hat{p}_{\cdot,j} \prod_{k=0}^{v-1} \left(1 - \hat{p}_{\cdot,j} - \frac{k}{n} \right) \right] \right\}$$
$$- \sum_{v=1}^{n-1} \frac{1}{v} \left\{ \frac{n^{v+1}[n-(v+1)]!}{n!} \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \left[\hat{p}_{i,j} \prod_{k=0}^{v-1} \left(1 - \hat{p}_{i,j} - \frac{k}{n} \right) \right] \right\} \tag{60}$$

Corollary 5. *Under the conditions of Lemma 7,*

$$\sqrt{n} \left[\begin{pmatrix} \hat{H}_z(X) \\ \hat{H}_z(Y) \\ \hat{H}_z(X,Y) \end{pmatrix} - \begin{pmatrix} H(X) \\ H(Y) \\ H(X,Y) \end{pmatrix} \right] \xrightarrow{L} MVN(0, \Sigma_H) \tag{61}$$

where

$$\Sigma_H = A \Sigma A^\tau$$

with Σ as in (39) and A is as in (34) according to (40), (41) and (42).

Lemma 8. *Suppose $\{p_k\}$ is a non-uniform distribution on $\mathscr{L} = \{\ell_k\}$. Then*

$$\sqrt{n} \left(\hat{H}_z - \hat{H} \right) \xrightarrow{p} 0 \tag{62}$$

From this we can also derive the asymptotic normality of $\sqrt{n} \left(\widehat{MI}_z - MI \right)$. Note that

$$\sqrt{n} \left(\widehat{MI}_z - \widehat{MI} \right) = \sqrt{n} \left(\hat{H}_z(X) - \hat{H}(X) \right)$$
$$+ \sqrt{n} \left(\hat{H}_z(Y) - \hat{H}(Y) \right)$$
$$- \sqrt{n} \left(\hat{H}_z(X,Y) - \hat{H}(X,Y) \right)$$

By Lemma 8, each of the three additive terms in the above equation converge to zero in probability. This means that $\sqrt{n} \left(\widehat{MI}_z - \widehat{MI} \right) \xrightarrow{p} 0$. Applying Slutsky's Theorem and Theorem 3 immediately gives the following theorem.

Theorem 6. *Suppose that X and Y are not independent. Under the conditions of Theorem 3,*

$$\sqrt{n} \left(\widehat{MI}_z - MI \right) \xrightarrow{L} N(0, \sigma^2) \tag{63}$$

where σ is as in (46).

We can directly apply Theorem 6 and Slutsky's Theorem to give the following corollary.

Corollary 6. *Under the conditions of Theorem 6,*

$$\frac{\sqrt{n}\left(\widehat{MI}_z - MI\right)}{\hat{\sigma}} \xrightarrow{L} N(0,1) \tag{64}$$

where $\hat{\sigma}$ is as given in (51).

Let $\hat{H}_z(X)$, $H_z(Y)$ and $\hat{H}_z(X,Y)$ be as in (59). Let

$$\hat{\kappa}_z = \frac{\hat{H}_z(X) + \hat{H}_z(Y) - \hat{H}_z(X,Y)}{\hat{H}_z(X,Y)} \tag{65}$$

$$\hat{\kappa}_{1z} = \frac{\hat{H}_z(X) + \hat{H}_z(Y) - \hat{H}_z(X,Y)}{\min\{\hat{H}_z(X), \hat{H}_z(Y)\}} \tag{66}$$

$$\hat{\kappa}_{2z} = \frac{\hat{H}_z(X) + \hat{H}_z(Y) - \hat{H}_z(X,Y)}{\sqrt{\hat{H}_z(X)\hat{H}_z(Y)}} \tag{67}$$

$$\kappa_{3z} = \frac{\hat{H}_z(X) + \hat{H}_z(Y) - \hat{H}_z(X,Y)}{(\hat{H}_z(X) + \hat{H}_z(Y))/2} \tag{68}$$

$$\hat{\kappa}_{4z} = \frac{\hat{H}_z(X) + \hat{H}_z(Y) - \hat{H}_z(X,Y)}{\max\{\hat{H}_z(X), \hat{H}_z(Y)\}} \tag{69}$$

Theorem 7. *Under the conditions of Corollary 5,*

1. $\sqrt{n}(\hat{\kappa}_z - \kappa) \xrightarrow{L} N(0, \sigma^2_{\hat{\kappa}_z})$,
2. $\sqrt{n}(\hat{\kappa}_{2z} - \kappa_2) \xrightarrow{L} N(0, \sigma^2_{\hat{\kappa}_{2z}})$,
3. $\sqrt{n}(\hat{\kappa}_{3z} - \kappa_3) \xrightarrow{L} N(0, \sigma^2_{\hat{\kappa}_{3z}})$,

where

$$\begin{aligned}
\sigma^2_{\hat{\kappa}_z} =& g^\tau_\kappa(H(X), H(Y), H(X,Y))\Sigma_H g_\kappa(H(X), H(Y), H(X,Y)),\\
\sigma^2_{\hat{\kappa}_{2z}} =& g^\tau_{\kappa_2}(H(X), H(Y), H(X,Y))\Sigma_H g_{\kappa_2}(H(X), H(Y), H(X,Y)),\\
\sigma^2_{\hat{\kappa}_{3z}} =& g^\tau_{\kappa_3}(H(X), H(Y), H(X,Y))\Sigma_H g_{\kappa_3}(H(X), H(Y), H(X,Y)).
\end{aligned}$$

with Σ_H is as in (44).

Corollary 7. *Under the conditions of Theorem 7,*

1. $\frac{\sqrt{n}(\hat{\kappa}_z - \kappa)}{\hat{\sigma}_{\hat{\kappa}_z}} \xrightarrow{L} N(0,1)$,
2. $\frac{\sqrt{n}(\hat{\kappa}_{2z} - \kappa_2)}{\hat{\sigma}_{\hat{\kappa}_{2z}}} \xrightarrow{L} N(0,1)$,
3. $\frac{\sqrt{n}(\hat{\kappa}_{3z} - \kappa_3)}{\hat{\sigma}_{\hat{\kappa}_{3z}}} \xrightarrow{L} N(0,1)$,

where $\hat{\Sigma}_H$ *is such that every* $p_{i,j}$ *in* Σ_H *is substituted by* $\hat{p}_{i,j}$ *and*

$$\begin{aligned}
\hat{\sigma}^2_{\hat{\kappa}_z} =& g^{\tau}_{\kappa}(\hat{H}_z(X), \hat{H}_z(Y), \hat{H}_z(X,Y))\hat{\Sigma}_H g_{\kappa}(\hat{H}_z(X), \hat{H}_z(Y), \hat{H}_z(X,Y)),\\
\hat{\sigma}^2_{\hat{\kappa}_{2z}} =& g^{\tau}_{\kappa_2}(\hat{H}_z(X), \hat{H}_z(Y), \hat{H}_z(X,Y))\hat{\Sigma}_H g_{\kappa_2}(\hat{H}_z(X), \hat{H}_z(Y), \hat{H}_z(X,Y)),\\
\hat{\sigma}^2_{\hat{\kappa}_{3z}} =& g^{\tau}_{\kappa_3}(\hat{H}_z(X), \hat{H}_z(Y), \hat{H}_z(X,Y))\hat{\Sigma}_H g_{\kappa_3}(\hat{H}_z(X), \hat{H}_z(Y), \hat{H}_z(X,Y)).
\end{aligned}$$

Theorem 8. *Suppose* $H(X) \neq H(Y)$*. Then under the conditions of Corollary 5,*

1. $\sqrt{n}(\hat{\kappa}_{1z} - \kappa_1) \xrightarrow{L} N(0, \sigma^2_{\hat{\kappa}_{1z}})$,
2. $\sqrt{n}(\hat{\kappa}_{4z} - \kappa_4) \xrightarrow{L} N(0, \sigma^2_{\hat{\kappa}_{4z}})$

where, if $H(X) < H(Y)$*,*

$$\begin{aligned}
\sigma^2_{\hat{\kappa}_{1z}} =& g^{\tau}_{\kappa_1, x_1<x_2}(H(X), H(Y), H(X,Y))\Sigma_H g_{\kappa_1, x_1<x_2}(H(X), H(Y), H(X,Y)),\\
\sigma^2_{\hat{\kappa}_{4z}} =& g^{\tau}_{\kappa_4, x_1<x_2}(H(X), H(Y), H(X,Y))\Sigma_H g_{\kappa_4, x_1<x_2}(H(X), H(Y), H(X,Y))
\end{aligned}$$

and if $H(X) > H(Y)$*, then*

$$\begin{aligned}
\sigma^2_{\hat{\kappa}_{1z}} =& g^{\tau}_{\kappa_1, x_1>x_2}(H(X), H(Y), H(X,Y))\Sigma_H g_{\kappa_1, x_1>x_2}(H(X), H(Y), H(X,Y)),\\
\sigma^2_{\hat{\kappa}_{4z}} =& g^{\tau}_{\kappa_4, x_1>x_2}(H(X), H(Y), H(X,Y))\Sigma_H g_{\kappa_4, x_1>x_2}(H(X), H(Y), H(X,Y))
\end{aligned}$$

with Σ_H *is as in (44).*

Corollary 8. *Under the conditions of Theorem 8,*

1. $\frac{\sqrt{n}(\hat{\kappa}_{1z} - \kappa_1)}{\hat{\sigma}_{\hat{\kappa}_{1z}}} \xrightarrow{L} N(0,1)$,
2. $\frac{\sqrt{n}(\hat{\kappa}_{4z} - \kappa_4)}{\hat{\sigma}_{\hat{\kappa}_{4z}}} \xrightarrow{L} N(0,1)$

where $\hat{\Sigma}_H$ *is such that every* $p_{i,j}$ *in* Σ_H *is substituted by* $\hat{p}_{i,j}$*. If* $H(X) < H(Y)$*,*

$$\begin{aligned}
\hat{\sigma}^2_{\hat{\kappa}_{1z}} =& g^{\tau}_{\kappa_1, x_1<x_2}(\hat{H}_z(X), \hat{H}_z(Y), \hat{H}_z(X,Y))\hat{\Sigma}_H g_{\kappa_1, x_1<x_2}(\hat{H}_z(X), \hat{H}_z(Y), \hat{H}_z(X,Y)),\\
\hat{\sigma}^2_{\hat{\kappa}_{4z}} =& g^{\tau}_{\kappa_4, x_1<x_2}(\hat{H}_z(X), \hat{H}_z(Y), \hat{H}_z(X,Y))\hat{\Sigma}_H g_{\kappa_4, x_1<x_2}(\hat{H}_z(X), \hat{H}_z(Y), \hat{H}_z(X,Y)).
\end{aligned}$$

and if $H(X) > H(Y)$*,*

$$\begin{aligned}
\hat{\sigma}^2_{\hat{\kappa}_{1z}} =& g^{\tau}_{\kappa_1, x_1>x_2}(\hat{H}_z(X), \hat{H}_z(Y), \hat{H}_z(X,Y))\hat{\Sigma}_H g_{\kappa_1, x_1>x_2}(\hat{H}_z(X), \hat{H}_z(Y), \hat{H}_z(X,Y)),\\
\hat{\sigma}^2_{\hat{\kappa}_{4z}} =& g^{\tau}_{\kappa_4, x_1>x_2}(\hat{H}_z(X), \hat{H}_z(Y), \hat{H}_z(X,Y))\hat{\Sigma}_H g_{\kappa_4, x_1>x_2}(\hat{H}_z(X), \hat{H}_z(Y), \hat{H}_z(X,Y)).
\end{aligned}$$

Hypothesis Testing

By Theorem 3, a necessary condition for the asymptotic normality of mutual information is that

$$\sigma^2 = (1, 1, -1)\Sigma(1, 1, -1)^\tau > 0 \tag{70}$$

Since the asymptotic normality of κ, κ_1, κ_2, κ_3, and κ_4 are derived from the asymptotic normality of mutual information, (70) is therefore a necessary condition for the asymptotic normality of these as well. If X and Y are independent random elements on the joint alphabet $\mathscr{X} \times \mathscr{Y}$, then $MI(X,Y) = \kappa = \kappa_2 = \kappa_3 = 0$, and (70) is not satisfied. If (70) holds, then the asymptotic normality is satisfied and a large sample hypothesis test is valid for κ, κ_2 and κ_3.

However, large sample hypothesis tests of the form $H_0 : \kappa = \epsilon$, $H_0 : \kappa_2 = \epsilon$, and $H_0 : \kappa_3 = \epsilon$ can be performed. We must have $\epsilon > 0$ because when $\epsilon = 0$, the underlying asymptotic normality does not hold. Under the previous null hypotheses, the following test statistics are approximately standard normal random variables and may be used:

$$Z = \frac{\sqrt{n}(\hat{\kappa} - \epsilon)}{\hat{\sigma}_{\hat{\kappa}}} \quad Z = \frac{\sqrt{n}(\hat{\kappa}_2 - \epsilon)}{\hat{\sigma}_{\hat{\kappa}_2}} \quad Z = \frac{\sqrt{n}(\hat{\kappa}_3 - \epsilon)}{\hat{\sigma}_{\hat{\kappa}_3}}$$

The corresponding test statistics from Turing's perspective may also be used, and are approximately distributed standard normal under each the null hypotheses:

$$Z = \frac{\sqrt{n}(\hat{\kappa}_z - \epsilon)}{\hat{\sigma}_{\hat{\kappa}_z}} \quad Z = \frac{\sqrt{n}(\hat{\kappa}_{2z} - \epsilon)}{\hat{\sigma}_{\hat{\kappa}_{2z}}} \quad Z = \frac{\sqrt{n}(\hat{\kappa}_{3z} - \epsilon)}{\hat{\sigma}_{\hat{\kappa}_{3z}}}$$

If one is interested in testing whether or not $\kappa = \kappa_2 = \kappa_3 = 0$, by Theorem 2 this is a test of independence of X and Y on $\mathscr{X} \times \mathscr{Y}$. Since the asymptotic normality does not hold, one must conduct another test for independence. The Pearson chi-square statistic for independence in two-way contingency tables is

$$Q = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \frac{(F_{i,j} - n\hat{p}_{i,\cdot}\hat{p}_{\cdot,j})^2}{n\hat{p}_{i,\cdot}\hat{p}_{\cdot,j}}$$

where $F_{i,j}$ is the observed frequency of $(x_i, y_j) \in \mathscr{X} \times \mathscr{Y}$ in an *iid* sample of size n; $\hat{p}_{i,j} = F_{i,j}/n$, $\hat{p}_{i,\cdot} = \sum_{j=1}^{K_2} \hat{p}_{i,j}$, and $\sum_{i=1}^{K_1}$. Under $H_0 : \kappa = \kappa_2 = \kappa_3 = 0$, Q is asymptotically a chi-square random variable with degrees of freedom $(K_1 - 1)(K_2 - 1)$. One would reject H_0 if Q takes a large value, say greater than $\chi^2_\alpha((K_1 - 1)(K_2 - 1))$, the $(1 - \alpha) \times 100$th quantile of the chi-square distribution with degrees of freedom $(K_1 - 1)(K_2 - 1)$.

Another possible test of hypotheses that may be of interest is that of whether or not the standardized mutual information κ of a population changes from one time period to another. This would mean testing

$$H_0 : \kappa_{t1} - \kappa_{t2} = 0 \text{ vs. } H_a : \kappa_{t1} - \kappa_{t2} \neq 0$$

where κ_{t1} is standardized mutual information at time $t1$ and κ_{t2} is standardized mutual information at time $t2$. Because the two samples would be independent from each other, κ_{t1} and κ_{t2} are independent. In order to find the test statistic for this hypothesis, we need to mention a corollary.

Corollary 9. *If κ_{t1} is standardized mutual information at time $t1$ and κ_{t2} is standardized mutual information at time $t2$, and as long as $\kappa_{t1} \neq 0$ and $\kappa_{t2} \neq 0$, then*

$$\frac{\sqrt{n}((\hat{\kappa}_{t1} - \hat{\kappa}_{t2}) - (\kappa_{t1} - \kappa_{t2}))}{\sqrt{\hat{\sigma}^2_{\hat{\kappa}_{t1}} + \hat{\sigma}^2_{\hat{\kappa}_{t2}}}} \xrightarrow{L} N(0,1) \tag{71}$$

Thus, the appropriate test statistic for this hypothesis test would be

$$Z = \frac{\sqrt{n}(\hat{\kappa}_{t1} - \hat{\kappa}_{t2})}{\sqrt{\hat{\sigma}^2_{\hat{\kappa}_{t1}} + \hat{\sigma}^2_{\hat{\kappa}_{t2}}}} \tag{72}$$

and is approximately distributed standard normal under H_0 when n is sufficiently large.

It is to be noted that, with κ_1 and κ_4, the asymptotic normality only holds if $H(X) \neq H(Y)$, and the convergence tends toward two different distributions, dependent on whether $H(X) < H(Y)$ or $H(X) > H(Y)$. The question of which of these, or $H(X) = H(Y)$, is true, is impossible to determine unless the underlying distributions of X and Y are completely known. In this case there is no purpose in hypothesis testing, and so for κ_1 and κ_4 large sample hypothesis testing is not viable.

An Illustrative Example

The Religious Landscape Study done by the Pew Research Center in 2014 consisted of a nationally representative sample of adults in the United States, of size $n = 33,538$. For each individual in the selected group, among other things, the religion and the ethnicity was noted. There were $K_1 = 12$ choices for religion, and $K_2 = 5$ choices for ethnicity. The survey resulted in the data set in Table 1. Let X be the religion and Y be the ethnicity of a randomly selected individual. Then i takes values in $\{1, \cdots, 12 = K_1\}$ corresponding to religions Buddhist through Unaffiliated in the given vertical order in Table 1, and j takes values in $\{1, \cdots, 5 = K_2\}$ corresponding to ethnic groups White to Other/Mixed in the given horizontal order of the same table. We also have $K = K_1 K_2 = 12 * 5 = 60$, assuming $p_{i,j} > 0$ for every pair (i, j).

In the same configuration as that of Table 1, the observed point and marginal distributions are given in Table 2.

According to (28), (29), (30) and (45),

$$\begin{gathered}\hat{H}(X) = 1.817290539452137 \\ \hat{H}(Y) = 1.0233818418002496 \\ \hat{H}(X,Y) = 2.6385505772306845 \\ \widehat{MI}(X,Y) = 0.202121804\end{gathered}$$

and according to (59) and (60),

$$\begin{gathered}\hat{H}_z(X) = 1.8174545984818056 \\ \hat{H}_z(Y) = 1.0234414810331058 \\ \hat{H}_z(X,Y) = 2.6394424808941004 \\ \widehat{MI}_z(X,Y) = 0.201453599\end{gathered}$$

	White	Black	Asian	Latino	Other/Mixed	Total
Buddhist	115	8	87	31	21	262
Catholic	4197	213	213	2419	71	7113
Evangelical Protestant	6444	509	169	933	424	8479
Hindu	8	4	179	2	4	197
Historically Black Protestant	38	1798	1	57	19	1913
Jehovah's Witness	86	64	1	77	14	242
Jewish	754	17	17	33	17	838
Mainline Protestant	5156	180	60	359	240	5995
Mormon	558	7	6	52	33	656
Muslim	88	64	64	9	7	232
Orthodox Christian	149	15	5	11	4	184
Unaffiliated	5050	669	371	966	371	7427
Total	22,643	3548	1173	4949	1225	33,538

Table 1: Frequency Data of Ethnicity and Religion

Also, by (52), (57), (53), (54), (58) we have

$$\begin{aligned}\hat{\kappa} &= 0.076603346\\ \hat{\kappa}_1 &= 0.197503801\\ \hat{\kappa}_2 &= 0.148211577\\ \hat{\kappa}_3 &= 0.142305607\\ \hat{\kappa}_4 &= 0.111221513.\end{aligned}$$

and by (65), (66), (67), (68) and (69)

$$\begin{aligned}\hat{\kappa}_z &= 0.0763243\\ \hat{\kappa}_{1z} &= 0.196839392\\ \hat{\kappa}_{2z} &= 0.147710625\\ \hat{\kappa}_{3z} &= 0.141823983\\ \hat{\kappa}_{4z} &= 0.110843814\end{aligned}$$

$\hat{p}_{i,j}$	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$\hat{p}_{i,\cdot}$
$i=1$	0.003428946	0.000238535	0.002594072	0.000924325	0.000626155	0.007812034
$i=2$	0.12514163	0.006351005	0.006351005	0.072127139	0.002117002	0.212087781
$i=3$	0.192140259	0.015176814	0.00503906	0.02781919	0.012642376	0.252817699
$i=4$	0.000238535	0.000119268	0.005337229	0.000059633848	0.000119268	0.005873934
$i=5$	0.001133043	0.05361083	0.000029816924	0.001699565	0.000566522	0.057039776
$i=6$	0.002564255	0.001908283	0.000029816924	0.002295903	0.000417437	0.007215696
$i=7$	0.022481961	0.000506888	0.000506888	0.000983958	0.000506888	0.024986582
$i=8$	0.153736061	0.005367046	0.001789015	0.010704276	0.007156062	0.17875246
$i=9$	0.016637844	0.000208718	0.000178902	0.00155048	0.000983958	0.019559902
$i=10$	0.002623889	0.001908283	0.001908283	0.000268352	0.000208718	0.006917526
$i=11$	0.004442722	0.000447254	0.000149085	0.000327986	0.000119268	0.005486314
$i=12$	0.150575467	0.019947522	0.011062079	0.028803149	0.011062079	0.221450295
$\hat{p}_{\cdot,j}$	0.675144612	0.105790447	0.034975252	0.147563957	0.036525732	1

Table 2: Relative Frequency Data of Ethnicity and Religion

References

[1] Antos, A. and Kontoyiannis, I. (2001). Convergence Properties of Functional Estimates for Discrete Distributions. *Random Structures and Algorithms*, 19, 163-193.

[2] Blyth, C.R. (1959). Note on Estimating Information. *Annals of Mathematical Statistics*, 30, 71-79.

[3] Harris, B. (1975). The Statistical Estimation of Entropy in the Non-Parametric case. *Topics in Information Theory*, edited by I. Csiszar, Amsterdam: North-Holland, 323-355.

[4] Kullback, S. and Leibler, R.A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1), 79-86.

[5] Paninski, L. (2003). Estimation of Entropy and Mutual Information. *Neural Computation*. 15, 1191-1253.

[6] Pearson, K. (1900). On a Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling. *Philosophical Magazine*, Series 5, 50, 157-175. (Reprinted in 1948 in Karl Pearson's Early Statistical Papers, ed by E.S. Pearson, Cambridge: Cambridge University Press.)

[7] Pearson, K. (1922). On the Chi Square Test of Goodness of Fit. *Biometrika*, 9, 22-27.

[8] Shannon, C.E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27, 379-423 and 623-656.

[9] Vinh, N.X., Epps, J. and Bailey, J. (2010). Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11, 2837-2854.

[10] Yao, Y.Y. (2003). Information-Theoretic Measures for Knowledge Discovery and Data Mining. *Entropy Measures, Maxmimum Entropy Principle and Emerging Applications*, Karmeshu (ed.), Springer, 115-136.

[11] Zhang, Z. (2012). Entropy Estimation in Turing's Perspective. *Neural Computation*, 24(5), 1368-1389.

[12] Zhang, Z. (2013b). Asymptotic Normality of an Entropy Estimator with Exponentially Decaying Bias. *IEEE Transactions on Information Theory*, 59(1), 504-508.

[13] Zhang, Z. and Zhang, X. (2012). A Normal Law for the Plug-in Estimator of Entropy. *IEEE Transactions on Information Theory*, 58(5), 2745-2747.

[14] Zhang, Z. and Zheng, L. (2015). A Mutual Information Estimator with Exponentially Decaying Bias. *Statistical Applications in Genetics and Molecular Biology*, 14(3), 243-252.

[15] Zhang, Z. and Zhou, J. (2010). Re-Parameterization of Multinomial Distribution and Diversity Indices. *Journal of Statistical Planning and Inference*, 140(7), 1731-1738.