Rose-Hulman Undergraduate Mathematics Journal

Non-parametric Statistics for Quantifying Differences in Discrete Spectra

Alexander M. DiBenedetto [a]

Volume 15, No. 1, Spring 2014

Sponsored by

Rose-Hulman Institute of Technology
Department of Mathematics
Terre Haute, IN 47803
Email: mathjournal@rose-hulman.edu
http://www.rose-hulman.edu/mathjournal

[a] University of Evansville, Department of Mathematics, Evansville IN 47722

Rose-Hulman Undergraduate Mathematics Journal
Volume 15, No. 1, Spring 2014

Non-parametric Statistics for Quantifying Differences in Discrete Spectra

Alexander M. DiBenedetto

Abstract. This paper introduces three statistics for comparing discrete spectra. Abstractly, a discrete spectrum (histogram with n bins) can be thought of as an ordered n-tuple. These three statistics are defined as comparisons of two n-tuples, representing pair-wise, ordered comparisons of bin heights. This paper defines all three statistics and formally proves the first one is a metric, while providing compelling evidence the other two are metrics. It shows that these statistics are gamma distributed, and for $n \geq 10$, approximately normally distributed. It also discusses a few other properties of all three associated metric spaces.

Acknowledgements: The author is grateful for the advice, help, and direction of advisor Clark Kimberling, as well as the invaluable input of Angela Reisetter, Dave Dwyer, and Robert Morse.

1 Introduction

Analyzing spectra is a major component of experimental physics. Spectra represent data points plotted against some dependent variable, such as time, or energy. After getting data, it is important to parameterize it in some meaningful way; often, this is an approximation to a continuous function, such as the example data histogram (spectra!) and function shown in Figure 1. This continuous function is usually one of many proposed parameterizations made by theoretical physicists, and helps select for the correct theory going forward. Since most data can be represented by a single continuous function, there are many methods for determining and quantifying differences in continuous spectra: moment-generating functions, the convolution, et cetera. The problem is that this process breaks down when there is no single theory underlying the data being collected. That is, when multiple physical processes are involved it can be difficult to tell the contributions made by each process. In order to tackle this problem, the discrete spectra will need to be compared directly.

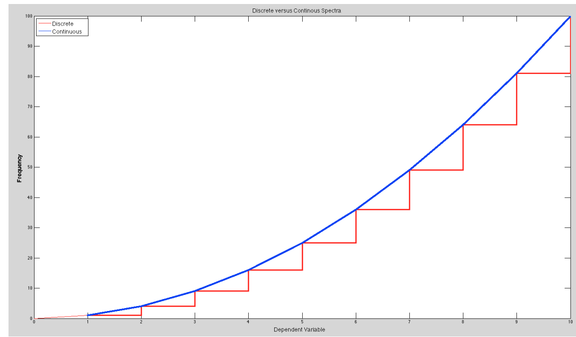


Figure 1: Illustrates the difference between a discrete (red) and a continuous (blue) spectra.

When comparing these spectra directly, there is a change made to one of the known, underlying variables. For example, in an analysis of a proposed next-generation dark matter experiment, these spectra are the energy spectra of incoming particles[1, 2]. The change between compared spectra was a change in the overburden (amount and density of rock the particles were to travel through), which affects the energy of the incoming muons[3]. The energy measurements are then histogrammed, using bins with a reasonable width. Despite the usefulness of log-log plots, like the one shown in Figure 2, it can still be unclear what degree of similarity exists between the shape and relative size of two discrete spectra. In order to bring some clarity to the comparisons of the muon energy spectra, the three $\mathcal{A}$ statistics were developed.

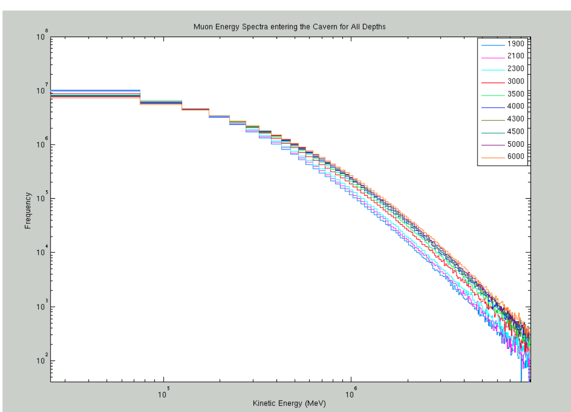


Figure 2: Muon energy spectra for a proposed next-generation dark matter experiment with ten different overburdens. Frequency against energy (MeV).

One existing method for comparing discrete spectra is the Kolmogorov-Smirnov test, which helps determine if there are any significant differences between two discrete spectra. However, this method does not provide a quantity to gauge how similar two given spectra are; a KS test just returns true or false[4]. The $\mathcal{A}$ statistics go further than a true or false by quantifying the difference between discrete spectra. Mathematically, they are metrics for quantifying the differences between two n-tuples with nonnegative, real components ($n > 0$); they essentially compare the bin-heights of these different histograms. The simplest metric, $\mathcal{A}_1$ is defined by

$$\mathcal{A}_1(X,Y) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\substack{i=0 \\ x_i+y_i>0}}^{n} \frac{|y_i - x_i|}{x_i + y_i} \tag{1}$$

where x_i and y_i are the i^{th} components of n-tuples X and Y with each x_i, $y_i \in \mathbb{R}_{\geq 0}$. $\mathcal{A}_2$ and $\mathcal{A}_3$ are defined in inequalities 8 and 9 in a similar fashion:

$$\mathcal{A}_2(X,Y) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\substack{i=0 \\ x_i+y_i>0}}^{n} \frac{\sqrt{|y_i - x_i|}}{\sqrt{x_i + y_i}} \tag{2}$$

$$\mathcal{A}_3(X,Y) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\substack{i=0 \\ x_i+y_i>0}}^{n} \frac{\sqrt{|y_i^2 - x_i^2|}}{\sqrt{x_i^2 + y_i^2}}. \tag{3}$$

Section 2 provides a proof that $\mathcal{A}_1$ is a metric. Since Monte-Carlo simulations and some contrived examples first suggested that these three $\mathcal{A}$ statistics are metrics, the end of section 2 discusses these simulations in more detail. As a result of these simulations, it is reasonable to believe that $\mathcal{A}_2$ and $\mathcal{A}_3$ are also metrics, but this has yet to be proven definitively. All three metrics have some additional, nice properties: they are closed under strictly-positive n-tuple multiplication and bounded between $[0, 1]$. However, they are not translation invariant or bi-variant. See section 3 for proof of these metric space properties.

In section 4, it is shown that these metrics are gamma distributed for all n and approximately normally distributed for $n \geq 10$. Since fewer than 10 bins is rarely used for energy spectra analysis (bin-widths are too wide), the bulk of section 4 is about their useful approximation to the standard normal distribution. The value of each $\mathcal{A}$ corresponds to a similarity percentile that is determined through a conversion to the standard normal. The relative similarity of two n-tuples is taken to be this similarity percentile, which is principally affected by differences in overall shape and relative size. The means and standard deviations needed for this conversion are found in Table 2. The purpose of section 5 is to highlight the varying sensitivity between these three metrics. It's important to have metrics of varying sensitivity because data does not always change in the same way. Some of the metrics are more sensitive to relative differences in bin-height, and others are more sensitive to changes in the shape. Section V also contains a topological representation of the $\mathcal{A}$ metrics for 1-tuples.

2 $\mathcal{A}$ Statistics as Metrics

The first four parts of this section provide a formal proof that $\mathcal{A}_1$ is a metric. The fifth part discusses why $\mathcal{A}_2$ and $\mathcal{A}_3$ are also believed to be metrics. Future work will include a more general proof for $\mathcal{A}_1$ that can be extended to include all three $\mathcal{A}$ statistics.

To be a metric, a function d (also called a distance function) is a function such that it satisfies the following four properties:

A. $d(X, Y) \geq 0$ for all X, Y. (Separation Property)
B. $d(X, Y) = 0$ if and only if $X = Y$. (Coincidence Property)
C. $d(X, Y) = d(Y, X)$ (Symmetric Property)
D. $d(X, Y) + d(Y, Z) \geq d(X, Z)$ (triangle inequality)[5].

Examples include the Euclidean metric, $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$, and the taxicab metric, $|x_1 - x_2| + |y_1 - y_2|$.

2.1 Proof of Separation Property

Let X and Y be arbitrary but particular n-tuples with nonnegative, real components. The definition of $\mathcal{A}_1(X, Y)$ is a summation: $\frac{1}{n} \sum\limits_{\substack{i=0 \\ x_i + y_i > 0}}^{n} \frac{|y_i - x_i|}{x_i + y_i}$. The quantity $|y_i - x_i|$ is always nonnegative by the definition of absolute value and $x_i + y_i$ is required to be strictly positive. Thus, $\mathcal{A}_1 \geq 0$ for any combination of elements in these arbitrary but particular n-tuples. Therefore, $d(X, Y) \geq 0$ for all X, Y and $\mathcal{A}_1$ satisfies the Separation Property.

2.2 Proof of Coincidence Property

Let X and Y be arbitrary but particular n-tuples with nonnegative, real components. Suppose $X = Y$. If $x_i = y_i$ for all $x_i \in X$ and $y_i \in Y$, then it is clear that $\mathcal{A}_1(X, Y) = \frac{1}{n} \sum\limits_{\substack{i=0 \\ y_i + y_i > 0}}^{n} \frac{|y_i - y_i|}{y_i + y_i}$. Now suppose that $\mathcal{A}_1 = 0$. Then there are only two possibilities to consider. Either $x_i = y_i = 0$, and nothing was added for that term because of the * condition, or $|y_i - x_i| = 0$. In either case, $x_i = y_i$ for all $x_i \in X$ and $y_i \in Y$. Thus, $X = Y$. Therefore, $d(X, Y) = 0$ if and only if $X = Y$, and $\mathcal{A}_1$ satisfies the Coincidence Property.

2.3 Proof of Symmetric Property

Let X and Y be arbitrary but particular n-tuples with nonnegative, real components. Then from the definition:

$$\mathcal{A}_1(X,Y) \;=\; \frac{1}{n} \sum_{\substack{i=0\\ x_i+y_i>0}}^{n} \frac{|y_i - x_i|}{x_i + y_i}.$$

By the definition of absolute value, $|y_i - x_i| = |x_i - y_i|$. And by the commutative property of addition, $x_i + y_i = y_i + x_i$. So

$$\mathcal{A}_1(X,Y) = \frac{1}{n} \sum_{\substack{i=0\\ x_i+y_i>0}}^{n} \frac{|x_i - y_i|}{y_i + x_i} \;=\; \mathcal{A}_1(Y,X).$$

Thus, $\mathcal{A}_1(X,Y) = \mathcal{A}_1(Y,X)$. Therefore $d(X,Y) = d(Y,X)$ and $\mathcal{A}_1$ satisfies the Symmetric Property.

2.4 Proof of Triangle Inequality

Let X, Y, and Z be arbitrary but particular n-tuples with nonnegative, real components. Since these are arbitrary but particular n-tuples, it is sufficient to show inequality 4 for a proof of the triangle inequality,

$$d(X,Y) + d(Y,Z) \geq d(X,Z). \tag{4}$$

To form a base case (for later induction on the length of the n-tuple), consider $n = 1$. The triangle inequality for a 1-tuple is shown in inequality 5:

$$\frac{|y_i - x_i|}{x_i + y_i} + \frac{|z_i - y_i|}{y_i + z_i} \geq \frac{|z_i - x_i|}{x_i + z_i}. \tag{5}$$

To start, consider the four cases where two or more of the 1-tuples have a value of zero:

i. $x_1 = y_1 = z_1 = 0$
ii. $x_1 = y_1 = 0$ and $z_1 > 0$
iii. $y_1 = z_1 = 0$ and $x_1 > 0$
iv. $x_1 = z_1 = 0$ and $y_1 > 0$

All four of these satisfy the triangle inequality:

i. $0 + 0 \geq 0$
ii. $0 + 1 \geq 1$
iii. $1 + 0 \geq 1$
iv. $1 + 1 \geq 0$

All remaining cases will have at most one 1-tuple with a value of zero, which guarantees that $x_i + y_i > 0$. There still remain eight possible arrangements of x_i, y_i, and z_i (called trichotomies), which are listed in Table 1. Two of the cases are self-contradictory, but the other six must all hold for this statistic to satisfy the triangle inequality for 1-tuples.

Using the conditions as specified in Table 1 and inequality 5, each of the eight cases are examined by removing the absolute value bars according to its definition. Cases 2 and 7 are

Table 1: Summary of Trichotomies of 1-tuple components (x_i, y_i, and z_i)

Case 1	$y_1 \geq x_1$	$z_1 \geq y_1$	$z_1 > x_1$
Case 2	$y_1 \geq x_1$	$z_1 \geq y_1$	$x_1 > z_1$
Case 3	$y_1 \geq x_1$	$y_1 \geq z_1$	$z_1 > x_1$
Case 4	$y_1 \geq x_1$	$y_1 \geq z_1$	$x_1 > z_1$
Case 5	$x_1 \geq y_1$	$z_1 \geq y_1$	$z_1 > x_1$
Case 6	$x_1 \geq y_1$	$z_1 \geq y_1$	$x_1 > z_1$
Case 7	$x_1 \geq y_1$	$y_1 \geq z_1$	$z_1 > x_1$
Case 8	$x_1 \geq y_1$	$y_1 \geq z_1$	$x_1 > z_1$

the self-contradictory cases, and they are removed form further discussion. In each of the remaining cases, removing the absolute value bars leads to a condition which must be true based on the assumptions of that particular case, as shown below.

Case 1

Using the conditions specified in Table 1 for Case 1, the following inequality must be true:

$$\frac{(y_1 - x_1)(z_1 - x_1)(z_1 - y_1)}{(x_1 + y_1)(x_1 + z_1)(y_1 + z_1)} \geq 0 \tag{6}$$

Through some algebra, this leads to inequality 7, which is a statement of the triangle inequality for Case 1:

$$\frac{y_1 - x_1}{x_1 + y_1} + \frac{z_1 - y_1}{y_1 + z_1} \geq \frac{z_1 - x_1}{x_1 + z_1} \tag{7}$$

Thus, the triangle inequality holds for Case 1.

Case 3

Using the conditions specified in Table 1 for Case 3, the following inequality must be true:

$$\frac{(y_1 - z_1)(x_1^2 + 3x_1y_1 + 3x_1z_1 + y_1z_1)}{(x_1 + y_1)(x_1 + z_1)(y_1 + z_1)} \geq 0. \tag{8}$$

Through some algebra, this leads to inequality 9, which is a statement of the triangle inequality for Case 3:

$$\frac{y_1 - x_1}{x_1 + y_1} + \frac{y_1 - z_1}{y_1 + z_1} \geq \frac{z_1 - x_1}{x_1 + z_1}. \tag{9}$$

Thus, the triangle inequality holds for Case 3.

Case 4

Using the conditions specified in Table 1 for Case 4, the following inequality must be true:

$$\frac{(y_1 - x_1)(x_1y_1 + 3x_1z_1 + 3y_1z_1 + z_1^2)}{(x_1 + y_1)(x_1 + z_1)(y_1 + z_1)} \geq 0. \tag{10}$$

Through some algebra, this leads to inequality 11, which is a statement of the triangle inequality for Case 4:

$$\frac{y_1 - x_1}{x_1 + y_1} + \frac{y_1 - z_1}{y_1 + z_1} \geq \frac{x_1 - z_1}{x_1 + z_1}. \tag{11}$$

Thus, the triangle inequality holds for Case 4.

Case 5

Using the conditions specified in Table 1 for Case 5, the following inequality must be true:

$$\frac{(x_1 - y_1)(x_1y_1 + 3x_1z_1 + 3y_1z_1 + z_1^2)}{(x_1 + y_1)(x_1 + z_1)(y_1 + z_1)} \geq 0. \tag{12}$$

Through some algebra, this leads to inequality 13, which is a statement of the triangle inequality for Case 5:

$$\frac{x_1 - y_1}{x_1 + y_1} + \frac{z_1 - y_1}{y_1 + z_1} \geq \frac{z_1 - x_1}{x_1 + z_1}. \tag{13}$$

Thus, the triangle inequality holds for Case 5.

Case 6

Using the conditions specified in Table 1 for Case 6, the following inequality must be true:

$$\frac{(z_1 - y_1)(x_1^2 + 3x_1y_1 + 3x_1z_1 + y_1z_1)}{(x_1 + y_1)(x_1 + z_1)(y_1 + z_1)} \geq 0. \tag{14}$$

Through some algebra, this leads to inequality 15, which is a statement of the triangle inequality for Case 6:

$$\frac{x_1 - y_1}{x_1 + y_1} + \frac{z_1 - y_1}{y_1 + z_1} \geq \frac{x_1 - z_1}{x_1 + z_1}. \tag{15}$$

Thus, the triangle inequality holds for Case 6.

Case 8

Using the conditions specified in Table 1 for Case 8, the following inequality must be true.

$$\frac{(x_1 - y_1)(x_1 - z_1)(y_1 - z_1)}{(x_1 + y_1)(x_1 + z_1)(y_1 + z_1)} \geq 0. \tag{16}$$

Through some algebra, this leads to inequality 17, which is a statement of the triangle inequality for Case 8:

$$\frac{x_1 - y_1}{x_1 + y_1} + \frac{y_1 - z_1}{y_1 + z_1} \geq \frac{x_1 - z_1}{x_1 + z_1}. \tag{17}$$

Thus, the triangle inequality holds for Case 8.

Therefore, the triangle inequality holds for $n = 1$, a base case. Now consider, as an inductive hypothesis, that the triangle inequality holds for a k-tuple. From the definition of $\mathcal{A}_1$:

$$\sum_{i=1}^{k} \frac{|y_i - x_i|}{x_i + y_i} + \sum_{i=1}^{k} \frac{|z_i - y_i|}{y_i + z_i} \geq \sum_{i=1}^{k} \frac{|z_i - x_i|}{x_i + z_i}. \tag{18}$$

To complete the inductive proof of the triangle inequality, it must be shown to be true for the $(k+1)$-tuple (inequality 19):

$$\sum_{i=1}^{k+1} \frac{|y_i - x_i|}{x_i + y_i} + \sum_{i=1}^{k+1} \frac{|z_i - y_i|}{y_i + z_i} \geq \sum_{i=1}^{k+1} \frac{|z_i - x_i|}{x_i + z_i}. \tag{19}$$

Let's start with inequality 20, which shows the triangle inequality for the $k+1$ element, a comparison of 1-tuples:

$$\frac{|y_{k+1} - x_{k+1}|}{x_{k+1} + y_{k+1}} + \frac{|z_{k+1} - y_{k+1}|}{y_{k+1} + z_{k+1}} \geq \frac{|z_{k+1} - x_{k+1}|}{x_{k+1} + z_{k+1}}. \tag{20}$$

We know this to be true since we proved the triangle inequality holds for 1-tuples (inequality 5). Adding inequality 20 to the inductive hypothesis (inequality 18):

$$\sum_{i=1}^{k} \frac{|y_i - x_i|}{x_i + y_i} + \frac{|y_{k+1} - x_{k+1}|}{x_{k+1} + y_{k+1}} + \sum_{i=1}^{k} \frac{|z_i - y_i|}{y_i + z_i} + \frac{|z_{k+1} - y_{k+1}|}{y_{k+1} + z_{k+1}} \geq \sum_{i=1}^{k} \frac{|z_i - x_i|}{x_i + z_i} + \frac{|z_{k+1} - x_{k+1}|}{x_{k+1} + z_{k+1}}.$$

This can be simplified to

$$\sum_{i=1}^{k+1} \frac{|y_i - x_i|}{x_i + y_i} + \sum_{i=1}^{k+1} \frac{|z_i - y_i|}{y_i + z_i} \geq \sum_{i=1}^{k+1} \frac{|z_i - x_i|}{x_i + z_i}. \tag{21}$$

Inequality 21 is identical to inequality 19, which means it has been shown that the triangle inequality holds for the $(k+1)$-tuple. Thus, these statistics satisfy inequality 4, the triangle inequality, which means they satisfy all of the criteria to be metrics.

2.5 Monte-Carlo simulations

Monte-Carlo simulations suggest $\mathcal{A}_2$ and $\mathcal{A}_3$ are metrics as well. In fact, the original reason for suspecting any of them were metrics is due to simulations which compared random distributions (with bin sizes of 1 through 1000) of random numbers (from a uniform random number generator). When it became evident that these may be metrics, some contrived examples were done to try and find a contradiction. None have been found. In addition, 100 billion simulations of these random distributions were performed for all three statistics and no contradictions to any of the four metric properties were detected. Future work will include an attempt to prove more generally that all $\mathcal{A}$ statistics are metrics.

3 Special Properties

The following are some proofs related to the three metric spaces.

3.1 Lack of Translation Invariance

Translation invariance, for a metric, is defined by the property

$$d(X,Y) = d(X+a, Y+a).$$

Assume these metrics are translation invariant. Then for each element of n-tuples X and Y,

$$\frac{|y_i - x_i|}{x_i + y_i} = \frac{|(y_i+a) - (x_i+a)|}{(x_i+a)+(y_i+a)}$$

$$= \frac{|y_i - x_i|}{x_i + y_i + 2a} \neq \frac{|y_i - x_i|}{x_i + y_i}.$$

This is a contradiction. Thus, $\mathcal{A}$ metrics are not translation invariant.

3.2 Lack of Bi-variance

Bi-variance is essentially the property that a metric be closed under group multiplication. That is, if X, Y, and Z are arbitrary but particular n-tuples, then

$$d(Y,Z) = d(XY, XZ) = d(YX, ZX).$$

This is true if, for each component of X, Y, and Z,

$$\frac{|z_i - y_i|}{y_i + z_i} = \frac{|(x_i)z_i - (x_i)y_i|}{(x_i)y_i + (x_i)z_i} = \frac{|z_i(x_i) - y_i(x_i)|}{y_i(x_i) + z_i(x_i)}.$$

The positive x_is cancel. However, whenever $x_i = 0$, they cannot cancel. And so these metrics are not bi-variant. They are, however, closed under strictly-positive n-tuple multiplication ($x_i \in X > 0$), since these x_is will cancel.

3.3 Proof of Boundedness

A metric is bounded if it satisfies $d(X, Y) \leq r$ for some r for all X, Y. Let X be the null-tuple, that is, $x_i = 0$ for all $x_i \in X$, because this is the smallest possible n-tuple. Let Y be an arbitrary but particular n-tuple with nonnegative, real components. Then for all positive entries

$$\frac{|y_i - 0|}{0 + y_i} = 1.$$

If an entry of y_i is 0 then $x_i = y_i = 0$ and $A_1 = 0$. Thus, the maximum the sum can be is n, which is divided out by n from the definition. The largest value this metric can take is 1. The smallest value it can take is 0, where every element in X and Y are the same. Since this metric has values between 0 and 1 inclusive, is has $d(X, Y) \leq 1$ for all X, Y. This satisfies the criterion for boundedness of a metric, where $r = 1$.

4 Distribution Characteristics

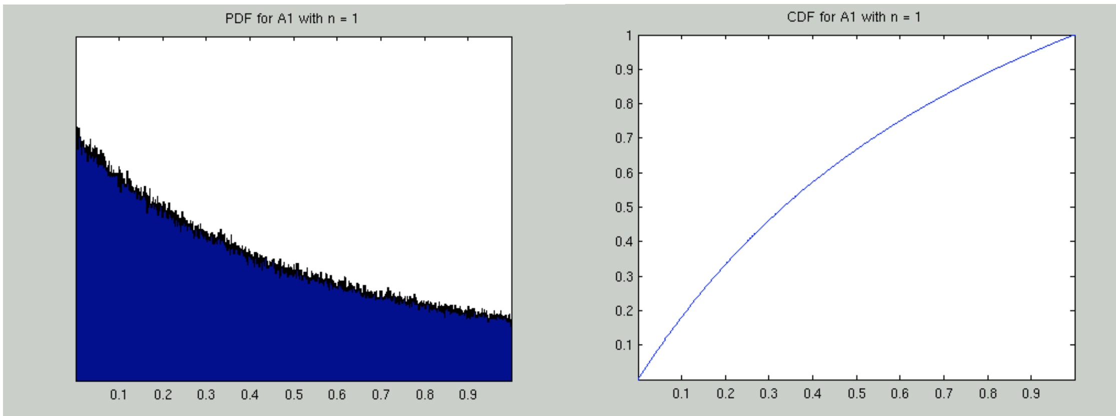
Section 4 discusses the connection between the $\mathcal{A}$ metrics and the standard normal distribution (for $n \geq 10$), the means and standard deviations for all three metrics, and their associated gamma distribution parameters, k and θ.

4.1 Monte-Carlo simulation of $\mathcal{A}$ metric characteristics

Figures 3 and 4 show $\mathcal{A}_1$ is gamma distributed for various values of n. Simulations were performed 1 billion times for each n value in order to plot the probability density function (pdf) and cumulative density function (cdf) for this metric as well as determine its mean and standard deviation. The n-tuples used for these comparisons were randomly sized tuples of random numbers ranging from 0 to a randomly high value (capped at 1 million). The mean was determined to be $\mu = .3863 \pm .0001$ for all values of n.

Figures 5 and 6 indicate that for $n \geq 10$, $\mathcal{A}_1$ is approximately normally distributed. Gamma distributions being approximately normally distributed for integer parameter $k \geq 10$ is more fully explored by Sadoulet, et. al.[6].

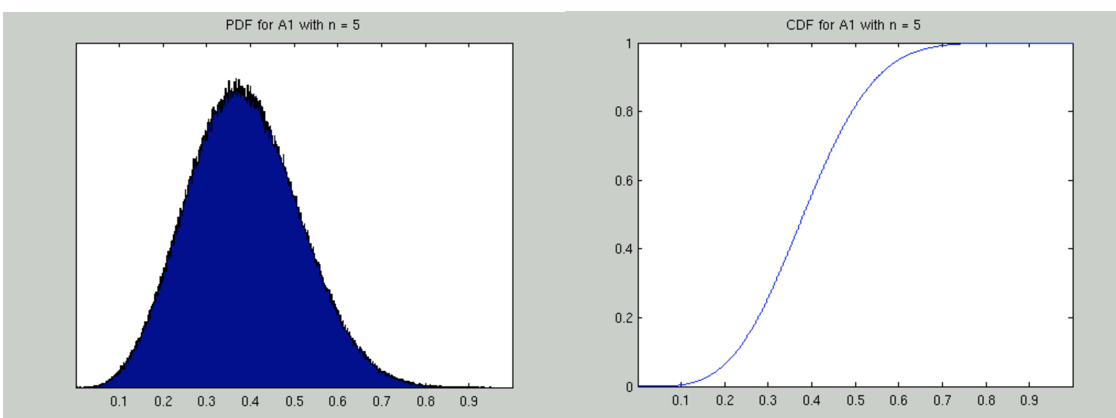
After Monte-Carlo simulations were used to find the mean and standard deviation for a variety of n values, it was determined that the standard deviation, σ, is parameterized by



(a) probability density function (b) cumulative density function

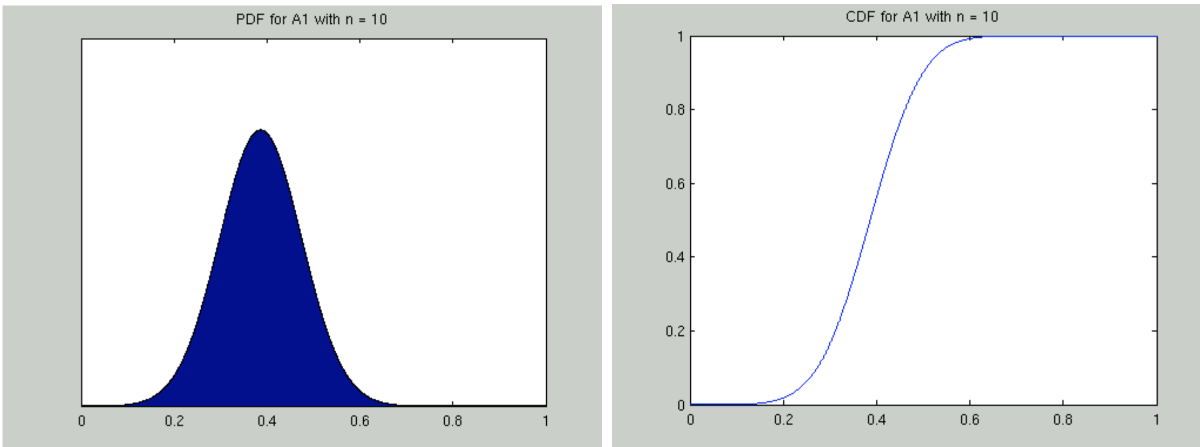
Figure 3: The distribution functions of $\mathcal{A}_1$ for $n = 1$

.



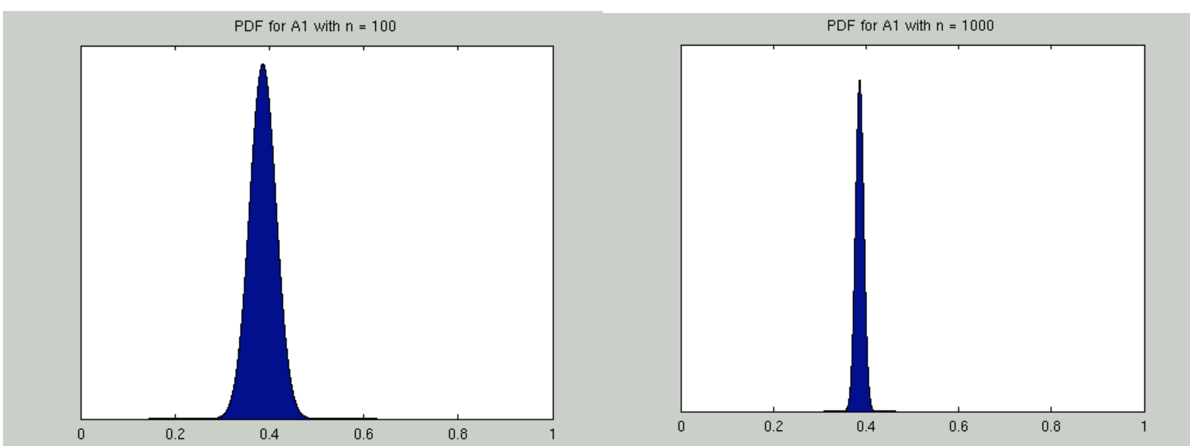
(a) probability density function (b) cumulative density function

Figure 4: The distribution functions of $\mathcal{A}_1$ for $n = 5$



(a) probability density function (b) cumulative density function

Figure 5: The distribution functions of $\mathcal{A}_1$ for $n = 10$



(a) $n = 100$ (b) $n = 1000$

Figure 6: The probability density functions of $\mathcal{A}_1$ for larger n values.

n as shown in Figure 7. This relationship and the value of the mean will allow the gamma distribution parameters to be determined for the $\mathcal{A}$ metrics.

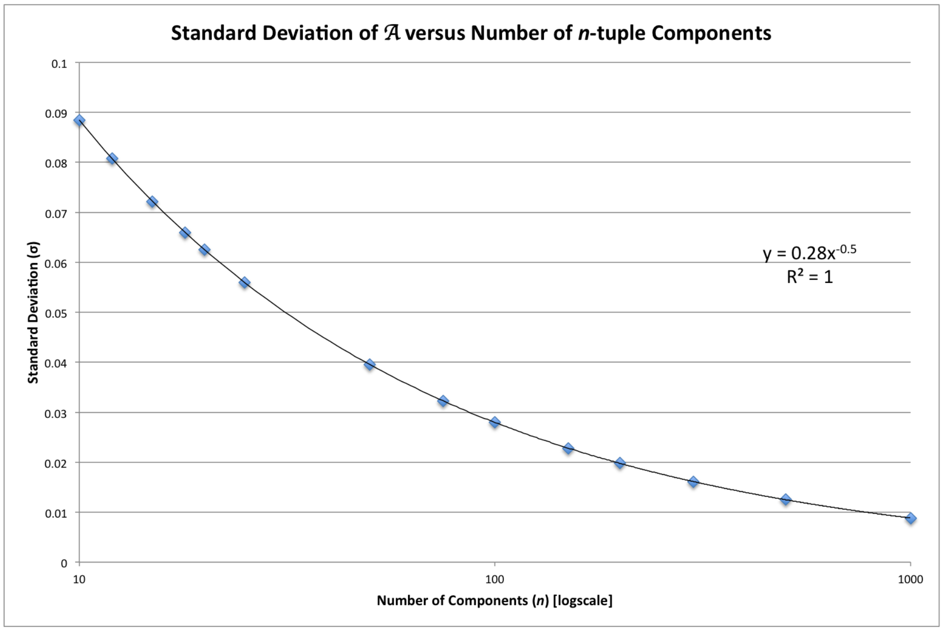


Figure 7: The relationship between standard deviation and number of components for $\mathcal{A}_1$. The least-squares best-fit for $f(n) = an^b$ was found from the plotted data with extremely high correlation.

The relationship derived in Figure 7 can be stated as

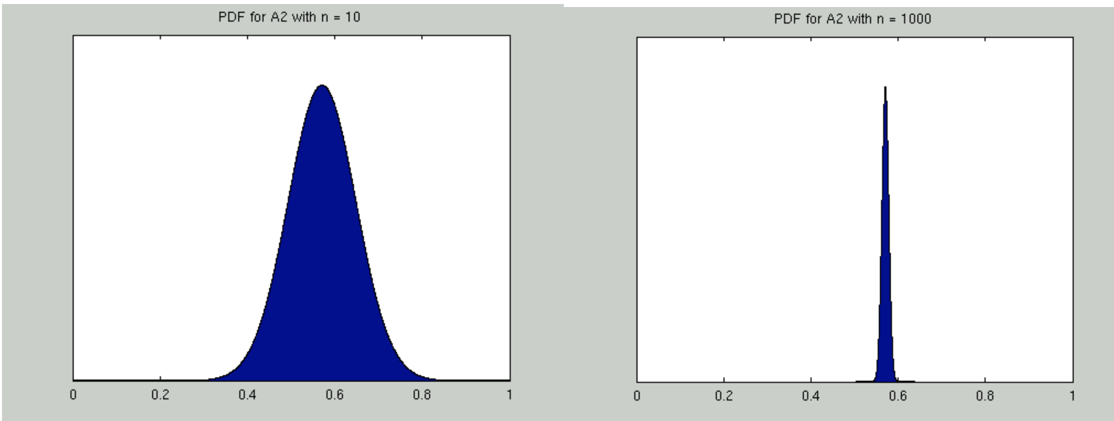
$$\sigma = \frac{.280 \ \pm .001}{\sqrt{n}}. \tag{22}$$

$\mathcal{A}_2$ and $\mathcal{A}_3$ have similar characteristics: they are also gamma distributed for all n and approximately normally distributed for $n \geq 10$. Figures 8 and 9 show $\mathcal{A}_2$ and $\mathcal{A}_3$ for some n values.

Monte-Carlo simulations were also used to find the mean and standard deviations for $\mathcal{A}_2$ and $\mathcal{A}_3$ for ns ranging from 10 to 1000, with 1 billion simulations each. These values are also listed in Table 2.

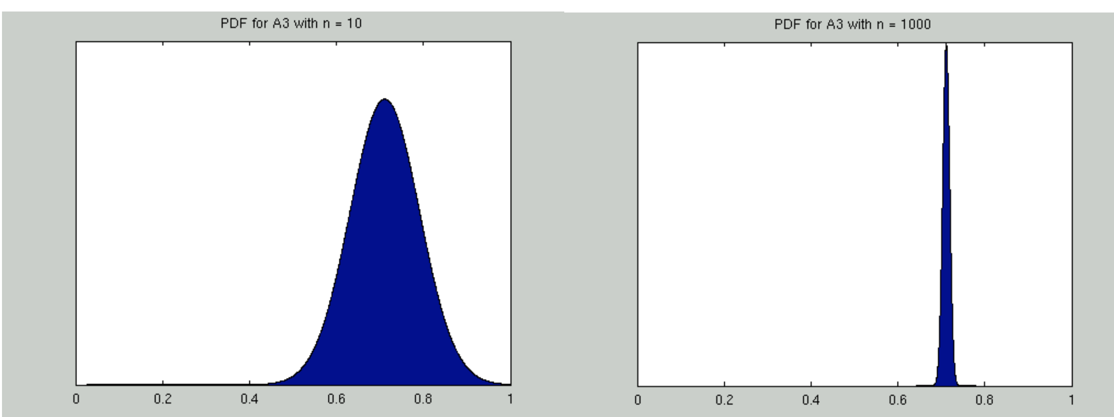
Table 2: Estimated means and standard deviations for the $\mathcal{A}$ metrics.

$\mathcal{A}$	Mean	Standard Deviation
1	$.3863 \pm .0001$	$\frac{.280 \pm .003}{\sqrt{n}}$
2	$.5708 \pm .0001$	$\frac{.246 \pm .002}{\sqrt{n}}$
3	$.7122 \pm .0001$	$\frac{.253 \pm .002}{\sqrt{n}}$



(a) $n = 10$ (b) $n = 1000$

Figure 8: The probability density functions of $\mathcal{A}_2$ for some n values.



(a) $n = 10$ (b) $n = 1000$

Figure 9: The probability density functions of $\mathcal{A}_3$ for some n values.

4.2 Gamma distribution parameters

After using these Monte-Carlo simulations to find the mean and standard deviation, the parameters for a gamma distribution can be determined. The probability density function for a gamma distribution is given by

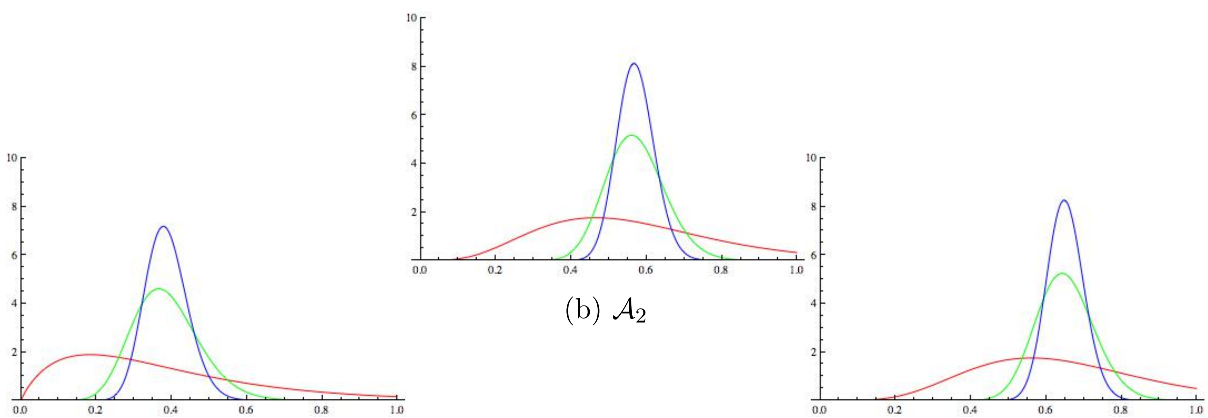
$$PDF(k,\theta) \stackrel{\text{def}}{=} \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}} \tag{23}$$

with expected value (mean) $\mathrm{E}[X] = k\theta$ and standard deviation $\mathrm{Std}[X] = \sqrt{k}\theta$. Table 3 gives the values of k and θ for each of the $\mathcal{A}$ metrics using the estimated means and standard deviations from Table 2.

Table 3: Estimated gamma distribution parameters, k and θ for the $\mathcal{A}$ metrics. These parameters were estimated using the means and standard deviations from Table 2.

$\mathcal{A}$	k	θ
$\mathcal{A}_1$	$1.9034n$	$\frac{0.20295}{n}$
$\mathcal{A}_2$	$5.3839n$	$\frac{0.10602}{n}$
$\mathcal{A}_3$	$7.9243n$	$\frac{0.089875}{n}$

It's interesting to note that the distribution parameters both depend on the size of the n-tuple, but the mean does not. To further illustrate this point, Figure 10 shows the gamma distributions for the $\mathcal{A}$ metrics plotted for a variety of n values. These plots are nearly identical to the ones produced by Monte-Carlo simulation earlier in this section.



(a) $\mathcal{A}_1$ (b) $\mathcal{A}_2$ (c) $\mathcal{A}_3$

Figure 10: Probability Density Function from computed distribution parameters for the $\mathcal{A}$ metrics. The n values shown are: 1 (red), 10 (green), and 25 (blue).

These variations in the gamma distribution parameters leads directly to the variations in sensitivity of the three metrics. Section V further explores this varying sensitivity among the three metrics.

4.3 Similarity percentile by conversion to the standard normal distribution

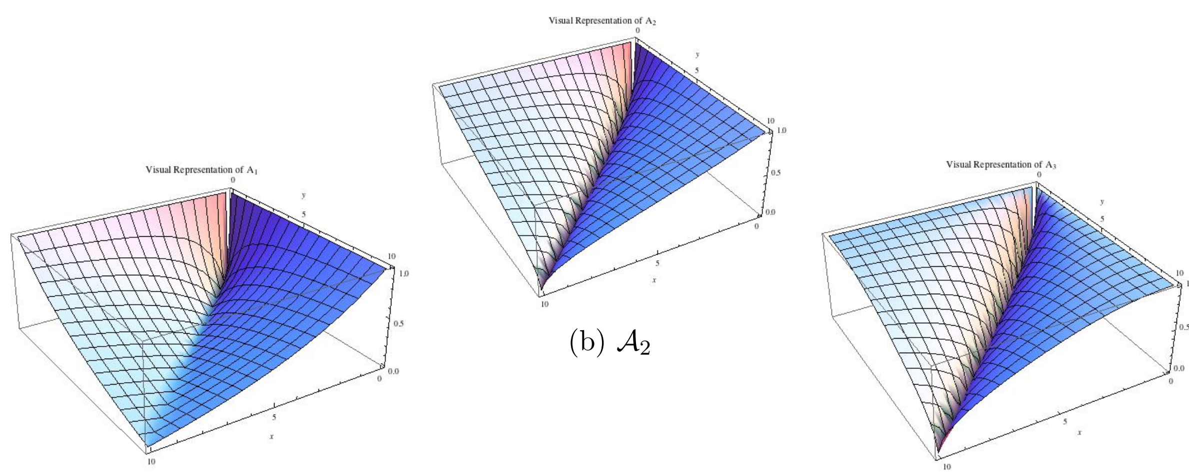
Gamma distribution, for $n \geq 10$, are approximately normally distributed. Since most of the time the discrete spectra being analyzed satisfy this requirement, it is useful to convert each $\mathcal{A}$ metric to the standard normal distribution[7] using the relationship

$$z \stackrel{\text{def}}{=} \frac{\mathcal{A} - \mu}{\sigma}. \tag{24}$$

Be sure to use the μ and σ that correspond to the correct statistic by referencing Table 2. Use the z-score from equation 24 to find the percentile of dissimilarity by looking the value up in a one-sided z-score table. Take one-hundred minus the percentile of dissimilarity to find the percentile of similarity. This similarity percentile is what quantifies the differences in the discrete spectra, and provides crucial information about the relative similarity of two discrete spectra.

5 Topological Plots for $n = 1$

Figure 11 shows the topology of these metrics for 1-tuples. They show the varying levels of sensitivity between $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$. Particularly, $\mathcal{A}_1$ is much broader than either $\mathcal{A}_2$ or $\mathcal{A}_3$.



(a) $\mathcal{A}_1$

(b) $\mathcal{A}_2$

(c) $\mathcal{A}_3$

Figure 11: Topological Plots of the $\mathcal{A}$ metrics

The varying sensitivity stems from differences in both the shape parameter, k, and the scale parameter, θ. $\mathcal{A}_1$ has a smaller shape parameter but a larger scale parameter, so it is more sensitive to changes in the relative size of the discrete spectra. $\mathcal{A}_3$ has a much larger shape parameter but a smaller scale parameter, so it is more sensitive to differences in the overall shape than $\mathcal{A}_1$. $\mathcal{A}_2$ lays in between the two in terms of the two types of sensitivity, so it is best for getting some of both sensitivity types. It's ultimately useful to have three statistics with varying sensitivity because discrete spectra vary in different ways: sometimes the shape is different and sometimes the relative size is different. These three metrics allow for a variance in the shape and scale parameter for quantifying comparisons of discrete spectra.

6 Conclusion

The $\mathcal{A}$ metrics provide a quantifiable way to study the similarities in discrete spectra. These metrics are supplementary statistical information that is particularly useful for data sets that have visually hidden fluctuations in shape. The fact that they are metrics is meaningful because they ignore units; the size of the input has no bearing on the range of the output. Lastly, they are useful because they make no assumptions about the data being analyzed. Their quantification of similarity is not an approximation to a continuous function or parameterized by a theoretical distribution. They return a gamma distributed number, between 0 and 1, which can be normally approximated for $n \geq 10$, and quantify the similarity between two discrete spectra.

References

[1] www.geant4.cern.ch

[2] http://cdms.berkeley.edu

[3] DiBenedetto, Alexander M. *Estimating Systematic Uncertainties in the Cosmogenic Background for Next-generation SuperCDMS*. In preparation to be submitted to the *Journal of Undergraduate Research in Physics*.

[4] Panchenko, Dmitry. Section 13 *Kolomogorov-Smirnov Test*, 83-89 (2006).

[5] Venema, Gerard A. *Foundations of Geometry*, 60 (2002).

[6] Sadoulet, B. et. al. *Statistical Methods in Experimental Physics*, 75-76 (1971).

[7] Miller, Irwin, Marylees Miller, and John E. Freund. *John E. Freund's Mathematical Statistics with Applications*, 77-139 (2004).