involve
a journal of mathematics

Benford’s law beyond independence: tracking Benford behavior in copula models

Rebecca F. Durst and Steven J. Miller

msp

2019 vol. 12, no. 7

INVOLVE 12:7 (2019)
dx.doi.org/10.2140/involve.2019.12.1193

Benford's law beyond independence: tracking Benford behavior in copula models

Rebecca F. Durst and Steven J. Miller

(Communicated by Stephan Garcia)

Benford's law describes a common phenomenon among many naturally occurring data sets and distributions in which the leading digits of the data are distributed with the probability of a first digit of d base B being $\log_B((d+1)/d)$. As it often successfully detects fraud in medical trials, voting, science and finance, significant effort has been made to understand when and how distributions exhibit Benford behavior. Most of the previous work has been restricted to cases of independent variables, and little is known about situations involving dependence. We use copulas to investigate the Benford behavior of the product of n dependent random variables. We develop a method for approximating the Benford behavior of a product of n dependent random variables modeled by a copula distribution C and quantify and bound a copula distribution's distance from Benford behavior. We then investigate the Benford behavior of various copulas under varying dependence parameters and number of marginals. Our investigations show that the convergence to Benford behavior seen with independent random variables as the number of variables in the product increases is not necessarily preserved when the variables are dependent and modeled by a copula. Furthermore, there is strong indication that the preservation of Benford behavior of the product of dependent random variables may be linked more to the structure of the copula than to the Benford behavior of the marginal distributions.

1. Introduction

Benford's law of digit bias applies to many commonly encountered data sets and distributions. A set of data $\{x_i\}_{i\in I}$ is said to be *Benford base B* if the probability of observing a value x_i in the set with the first digit d (where d is any integer from 1 to $B-1$) is given by the equation

$$\text{Prob(first digit of } \{x_i\}_{i\in I} \text{ is } d\text{) base } B = \log_B\left(\frac{d+1}{d}\right). \tag{1-1}$$

MSC2010: 11K99, 60E99.
Keywords: Benford's law, probability, theoretical statistics .

These probabilities monotonically decrease; e.g., in base 10 there is a leading digit of 1 about 30.103% of the time and a leading digit of 9 about 4.576% of the time.

Benford's law was discovered in 1881 by the astronomer-mathematician Simon Newcomb who, looking at his logarithm table, observed earlier pages were more heavily worn than later pages. As logarithm tables are organized by leading digit, this led him to conclude that values with leading digit 1 occurred more commonly than values with higher leading digits. These observations were mostly forgotten for fifty years, when Benford [1938] published his work detailing similar biases in a variety of settings. Since then, the number of fields where Benford behavior is seen has rapidly grown, including accounting, biology, computer science, economics, mathematics, physics and psychology to name a few; see [Benford 2009; Berger and Hill 2015; Miller 2015; Nigrini 1999; Raimi 1976] for a development of the general theory and many applications. This prevalence of Benford's law, particularly in naturally occurring data sets and common distributions, has allowed it to become a useful tool in detecting fraud. One notable example of this was its use in 2009 to find evidence suggesting the presence of fraud in the Iranian elections [Battersby 2009]. While Benford's law cannot prove that fraud happened, it is a useful tool for determining which sets of data are suspicious enough to merit further investigation (which is of great importance given finite resources); see for example [Nigrini and Mittermaier 1997; Singleton 2011].

To date, most of the work on the subject has involved independent random variables or deterministic processes (see though [Becker et al. 2018; Iafrate et al. 2015] for work on dependencies in partition problems). Our goal below is to explore dependent random variables through copulas, quantifying the connections between various relations and Benford behavior.

Copulas are multivariate probability distributions restricted to the unit hypercube by transforming the marginals into uniform random variables via the probability integral transform (see Section 2 for precise statements). The term copulas was first defined by Abe Sklar in 1959, when he published what is now known as Sklar's theorem (see Theorem 2.7), though similar objects were present in the work of Wassily Hoeffding as early as 1940. Sklar described their purpose as linking n-dimensional distributions with their one-dimensional margins. See [Nelsen 2006] for a detailed account of the presence and evolution of copulas.

Fisher [1997] writes, "Copulas [are] of interest to statisticians for two main reasons: Firstly, as a way of studying scale-free measures of dependence; and secondly, as a starting point for constructing families of bivariate distributions, sometimes with a view to simulation." More specifically, copulas are widely used in application in fields such as economics and actuarial studies; for example, [Kpanzou 2007] describes applications in survival analysis and extreme value theory, and [Wu et al. 2007] details the use of Archimedean copulas in economic modeling and risk management. Thus, as copulas are a convenient and useful way to model dependent random variables, they are often employed in fields relating to finance and economics. Since many of these areas are also highly susceptible to fraud, it is worth exploring connections between copulas and Benford's law, with the goal to develop data integrity tests.

Essentially, since so many dependencies may be modeled through copulas, it is natural to ask when and how often these structures will display Benford behavior. In this paper, we investigate when data modeled by a copula is close to Benford's law by developing a method for approximating Benford behavior. In Section 3, we develop this method for the product of n random variables whose joint distribution is modeled by the copula C. We then apply this method in Section 4 to directly investigate Benford behavior for various copulas and dependence parameters. We conclude that Benford behavior depends heavily on the structure of the copula. We use goodness of fit measures to show both numerically and graphically that the product of many random variables with dependence modeled by a copula will not necessarily level-off like products of independent random variables, the log of which we may expect to become more uniform as the number of variables increases. The results of this paper extend current techniques for testing Benford's law to situations where independence is not guaranteed, allowing analyses like that carried out in [Cuff et al. 2015] on the Weibull distribution and in [Durst et al. 2016] on the inverse gamma distribution to be conducted in the case of n dependent random variables. In Section 5, we restrict ourselves to n-tuples of random variables in which at least one is a Benford distribution and develop a concept of distance between our joint distribution and a Benford distribution, thus developing a concept of distance from a Benford distribution in order to understand how much deviation from Benford one might expect of a particular distribution. We then provide an upper bound for this distance using the L^1 norm of the function

$$N(u_1, u_2, \ldots, u_n) = 1 - \frac{\partial^n C(u_1, u_2, \ldots, u_n)}{\partial u_1 \partial u_2 \ldots \partial u_n}.$$

In doing so, we draw an interesting connection between the distance from a Benford distribution and a copula's distance from the space of copulas for which $C_{uv}(u, v) = 1$ for all u, v in [0, 1].

2. Terms and definitions

We abbreviate *probability density function* by PDF and *cumulative distribution function* by CDF, and assume all CDFs are uniformly or absolutely continuous. All results below are standard; see the references for proofs.

General mathematics and Benford's law.

Lemma 2.1 (Barbalat's lemma [Fontes and Magni 2004, Lemma 2.1]). *Let $t \mapsto F(t)$ be a differentiable function with a finite limit as $t \to \infty$. If F' is uniformly continuous, then $F'(t) \to 0$ as $t \to \infty$.*

Definition 2.2 (scientific notation). Any real number, x, can be written in the form

$$x = S_B(x) \cdot B^n, \tag{2-1}$$

where n is an integer and $S_B(x) < 10$. We call B the *base* and $S_B(x)$ the *significand.*

We define strong Benford's law base B; see, for example, [Berger and Hill 2015; Miller 2015]. This is the definition we primarily use in Section 3; strong indicates that we are studying the entire significand of the number and not just its first digit. In Section 5, we will provide insight into how one may define a weaker version of Benford's law that permits the probabilities to be within ϵ of the theoretical Benford probabilities.

Definition 2.3 (strong Benford's law [Miller 2015, Definition 1.6.1]). A data set satisfies the strong Benford's law base B if the probability of observing a leading digit of at most s in base B is $\log_B s$.

Theorem 2.4 (absorptive property of Benford's law [Tao 2010, page 56]). *Let X and Y be **independent** random variables. If X obeys Benford's law, then the product $W = XY$ obeys Benford's law regardless of whether or not Y obeys Benford's law.*

Copulas. All theorems and definitions in this section are from [Nelsen 2006] unless otherwise stated.

Remark 2.5. In [Nelsen 2006], functions are defined on the *extended real line*, $[-\infty, \infty]$; thus $f(t)$ is defined when $t = \pm\infty$. We use this notation in order to maintain consistency with that work, as it is one of the central texts in copula theory.

Definition 2.6 (n-dimensional copula). An *n-dimensional copula*, C, is a function satisfying the following properties:

(1) The domain of C is $[0, 1]^n$.

(2) (n-increasing) The n-th order difference of C is greater than or equal to zero.

(3) (grounded) $C(u_1, u_2, \ldots, u_n) = 0$ if $u_k = 0$ for at least one k in $\{1, 2, \ldots, n\}$.

(4) $C(1, 1, \ldots, 1, u_k, 1, , \ldots, 1) = u_k$ for some k in $\{1, 2, \ldots, n\}$.

Theorem 2.7 (Sklar's theorem [Nelsen 2006, Theorem 2.10.9]). *Let H be an n-dimensional distribution function with marginal CDFs $F_1, F_2, \ldots, F_n$. Then there exists an n-copula C such that for all $(x_1, x_2, \ldots, x_n)$ in $[-\infty, \infty]^n$,*

$$H(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)). \tag{2-2}$$

If all F_i are continuous, then C is unique; otherwise, C is uniquely determined on $\mathrm{Range}(F_1) \times \mathrm{Range}(F_2) \times \cdots \times \mathrm{Range}(F_n)$. *Conversely, if C is a copula and $F_1, F_2, \ldots, F_n$ are cumulative distribution functions, then the function H defined by* (2-2) *is a distribution function with marginal cumulative distribution functions $F_1, F_2, \ldots, F_n$.*

Theorem 2.8 (extension of [Nelsen 2006, Theorem 2.4.2]). *Let $X_1, X_2, \ldots, X_n$ be continuous random variables. Then they are independent if and only if their copula, $C_{X_1,X_2,\ldots,X_n}$, is given by $C_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \Pi(x_1, x_2, \ldots, x_n) = x_1 x_2 \cdots x_n$, where Π is called the* ***product copula****.*

Theorem 2.9 (extension of [Nelsen 2006, Theorem 2.4.3]). *Let $X_1, X_2, \ldots, X_n$ be continuous random variables with copula $C_{X_1,X_2,\ldots,X_n}$. If $a_1, a_2, \ldots, a_n$ are strictly increasing on* $\mathrm{Range}(X_1), \mathrm{Range}(X_2), \ldots, \mathrm{Range}(X_n)$, *respectively, then $C_{a_1(X_1),a_2(X_2),\ldots,a_n(X_n)} = C_{X_1,X_2,\ldots,X_n}$. Thus $C_{X_1,X_2,\ldots,X_n}$ is invariant under strictly increasing transformations of $X_1, X_2, \ldots, X_n$.*

Remark 2.10. For the following three definitions, see page 116 of [Nelsen 2006] for the 2-copula formulas and page 151 for the n-copula extension.

Definition 2.11 (Clayton family of copulas). An (n-dimensional) copula in the *Clayton family* is given by the equation

$$C(u_1, u_2, \ldots, u_n) = \max\{(u_1^{-\alpha} + u_2^{-\alpha} + \cdots + u_n^{-\alpha} + n - 1)^{-1/\alpha}, 0\}, \tag{2-3}$$

where $\alpha \in [-1, \infty) \setminus \{0\}$ is a parameter related to dependence, with $\alpha = 0$ as the independence case.

Definition 2.12 (Ali–Mikhail–Haq family of copulas). An (n-dimensional) copula in the *Ali–Mikhail–Haq family* is given by the equation

$$C(u_1, u_2, \ldots, u_n) = \frac{(1-\alpha)}{\left(\prod_{i=1}^{n}(1-\alpha(1-u_i))/u_i\right) - \alpha}, \tag{2-4}$$

where $\alpha \in [-1, 1)$ is a parameter related to dependence, with $\alpha = 0$ as the independence case.

Definition 2.13 (Gumbel–Barnett family of copulas). An (n-dimensional) copula in the *Gumbel–Barnett family* is given by the equation

$$C(u_1, u_2, \ldots, u_n) = \exp\frac{1-(1-\alpha\log u_1)(1-\alpha\log u_2)\cdots(1-\alpha\log u_n)}{\alpha}, \tag{2-5}$$

where $\alpha \in (0, 1]$ is a parameter related to dependence, with $\alpha = 0$ as the independence case.

3. Testing for Benford behavior of a product

We state the results below in arbitrary dimensions but for notational convenience give the proofs for just two dimensions as the generalization is straightforward.

Let $X_1, \dots, X_n$ be continuous random variables with CDFs $F_{X_1}(x_1), \dots, F_{X_n}(x_n)$. Let their joint PDF be $H_{X_1,\dots,X_n}(X_1, \dots, X_n)$. By Theorem 2.7, we know there exists a copula C such that

$$H_{X_1,\dots,X_n}(X_1, \dots, X_n) = C(F_{X_1}(X_1), \dots, F_{X_n}(X_n)). \tag{3-1}$$

Assume $X_1, \dots, X_n$ are such that their copula C is *absolutely continuous*. This allows us to define the joint probability density function [Nelsen 2006, page 27] by $\partial^n C/(\partial x_1 \cdots \partial x_n)$. Furthermore, we restrict ourselves to X_i such that all F_{X_i} are uniformly continuous, as this allows us to use Lemma 2.1 to later ensure that the PDFs approach zero in their right- and left-end limits.

From here we have the following lemma.

Lemma 3.1. *Given $X_1, \dots, X_n$ positive, continuous random variables with joint distribution modeled by the absolutely continuous copula C, let $U_i = \log_B X_i$ for all $i \le n$ and for some base B, and let the CDFs of each U_i be $F_i(u_i)$. Also, let $f_i(u_i)$ be the PDF of U_i for all i. Finally, let*

$$\boldsymbol{u}_0 = (u_1, \dots, u_{n-1}, s + k - (u_1 + \cdots + u_{n-1})).$$

Then

$$\operatorname{Prob}\left(\left(\sum_{i=1}^{n} U_i\right) \bmod 1 \le s\right)$$
$$= \int_0^s \sum_{k=-\infty}^{\infty} \int_{u_1=-\infty}^{\infty} \cdots \int_{u_{n-1}=-\infty}^{\infty} \frac{\partial^n C(F_1(u_1), \dots, F_{n-1}(u_{n-1}), F_n(u_n))}{\partial u_1 \cdots \partial u_n}\bigg|_{\boldsymbol{u}_0} du_1 \cdots du_{n-1}.$$

Therefore, the PDF of $(U + V) \bmod 1$ is given by

$$\sum_{k=-\infty}^{\infty} \int_{u_1=-\infty}^{\infty} \cdots \int_{u_{n-1}=-\infty}^{\infty} \frac{\partial^n C(F_1(u_1), \dots, F_{n-1}(u_{n-1}), F_n(u_n))}{\partial u_1 \cdots \partial u_n}\bigg|_{\boldsymbol{u}_0} du_1 \cdots du_{n-1}. \tag{3-2}$$

See Appendix 1 for the proof.

If (3-2) equals 1 for all s, then our product is Benford. If it is not identically equal to 1 for all s, then at each point we may assign a value ϵ_s that represents our distance from a Benford distribution. Thus we have

$$\epsilon_s = |1 - P|, \tag{3-3}$$

where P is the PDF given in (3-2). This formulation will form the basis of Section 5.

Unfortunately, the infinite sum and improper integral in (3-2) make it highly impractical to use in application unless we can determine a method to closely approximate them by a finite sum and finite integral. We note that (3-2) is a PDF, and so is $\partial^n C/(\partial x_1 \cdots \partial x_n)$, so we have the following properties (for notational convenience we state them in the two-dimensional case; similar results hold for n-dimensions).

(1) $\displaystyle\int_0^1 \Bigl(\sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} C_{u_1u_2}(F_1(u_1),\, F_2(s+k-u_1)) f_1(u_1) f_2(s+k-u_1)\, du_1 \Bigr) ds = 1.$

(2) $\displaystyle\sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} C_{u_1u_2}(F_1(u_1),\, F_2(s+k-u_1)) f_1(u_1) f_2(s+k-u_1)\, du_1 \geq 0$ for all s.

(3) $\displaystyle\int_{-\infty}^{\infty} C_{u_1u_2}(F_1(u_1),\, F_2(s+k-u_1)) f_1(u_1) f_2(s+k-u_1)\, du_1 \to 0$ as $k \to \pm\infty$.

(4) $C_{u_1u_2}(F_1(u_1),\, F_2(s+k-u_1)) f_1(u_1) f_2(s+k-u_1) \to 0$ as $u_1 \to \pm\infty$.

Property (1) is simply the definition of a PDF, and property (2) is a direct result of the fact that a PDF is always positive. Properties (3) and (4) are required, under Lemma 2.1, by the convergence of the integral in property (1) and by the convergence of the sum.

From properties (3) and (4) and the definition of convergence we obtain the following.

Lemma 3.2 (approximating the PDF). *Given $U_1, \ldots, U_n$ continuous random variables modeled by the copula C with marginal CDFs $F_1, \ldots, F_n$ and PDFs $f_1, \ldots, f_n$, there exist $a_1, \ldots, a_{n-1}$, $b_1, \ldots, b_{n-1}$, and c_1 and c_2 completely dependent on the F_i such that $a_i < b_i$ for all i and $c_1 < c_2$ and*

$$\sum_{k=-\infty}^{\infty} \int_{u_1=-\infty}^{\infty} \cdots \int_{u_{n-1}=-\infty}^{\infty} \frac{\partial^n C(F_1(u_1), \ldots, F_{n-1}(u_{n-1}), F_n(u_n))}{\partial u_1 \cdots \partial u_n}\bigg|_{\boldsymbol{u}_0} du_1 \cdots du_{n-1}$$
$$= \sum_{k=c_1}^{c_2} \int_{u_1=a_1}^{b_1} \cdots \int_{u_{n-1}=a_{n-1}}^{b_{n-1}} \frac{\partial^n C(F_1(u_1), \ldots, F_{n-1}(u_{n-1}), F_n(u_n))}{\partial u_1 \cdots \partial u_n}\bigg|_{\boldsymbol{u}_0} du_1 \cdots du_{n-1} + E_{a,b,c}(s), \quad (3\text{-}4)$$

where $E_{a,b,c}(s) \to 0$ as each a_i and c_1 go to $-\infty$ and each b_i and c_2 go to ∞. Thus, for any $\epsilon > 0$, there exists (for each i) $|a_i|$, $|b_i|$, $|c_1|$, and $|c_2|$ large enough such that $|E_{a,b,c}(s)| \leq \epsilon$.

The proof of this claim can be found in Appendix 1.

The specific values of c_1, c_2, and each a_i and b_i are best determined by numerically testing the errors caused by truncating the interval using either known error bounds or appropriate software. As seen in Appendix 2, the values for these constants are often quite reasonable, as long as the functions decay fast enough in the limit.

Because s only ranges from 0 to 1, we can always find a value of s that maximizes $E_{a,b,c}$ for any given set of a, b, and c and set this to be the maximum error. Furthermore, since all f_i should have similar tail-end behavior, we do not have to worry about the divergence of one canceling out the divergence of the other. Thus, for this analysis to work, it is sufficient to understand the tail-end behavior of only one of the marginals.

In Appendix 2, we provide several examples of this method for testing for Benford behavior computationally with two variables.

4. Testing For Benford behavior: examples

Now that an effective method for testing the Benford behavior of copulas has been established, we investigate how this behavior varies for specific copulas and marginals. In all χ^2 tests, we follow standard procedure for multiple comparison problems. We are sampling our distribution at 12 values of s, necessitating 11 degrees of freedom, and we impose a significance level of 0.005, meaning we only accept a 0.5% probability of false rejection. Thus, we reject the hypothesis, specifically *we reject that the distribution displays Benford behavior*, if the χ^2-value exceeds 2.6. Our main interest, however, is to observe how and if these values trend towards this critical value.

Please note that the α used in this section is the dependence parameter of the copula and does not represent the significance level, as traditionally seen in statistical analysis.

2-copulas with varying dependence parameter. The following figures display the nonerror values of (3-4) at various values of s for three different copulas. The line in each plot indicates the constant function $y = 1$, which will be achieved if the product XY is exactly Benford. For each copula, we test three different pairings of marginals:

(A) $\log X \sim \mathrm{N}(0, 1)$ and $\log Y \sim \mathrm{Exp}(1)$.

(B) $\log X \sim \mathrm{Pareto}(1)$ and $\log Y \sim \mathrm{N}(0, 1)$.

(C) $\log X \sim \mathrm{Pareto}(1)$ and $\log Y \sim \mathrm{Exp}(1)$.

In each case, we vary the dependence parameter α and compare the results to the case of independence. Our Pareto distribution has scale parameter $x_m = 1$ and shape parameter $\alpha_p = 2$. We note that in some cases the axes must be adjusted to be able to show any change in the Benford behavior.

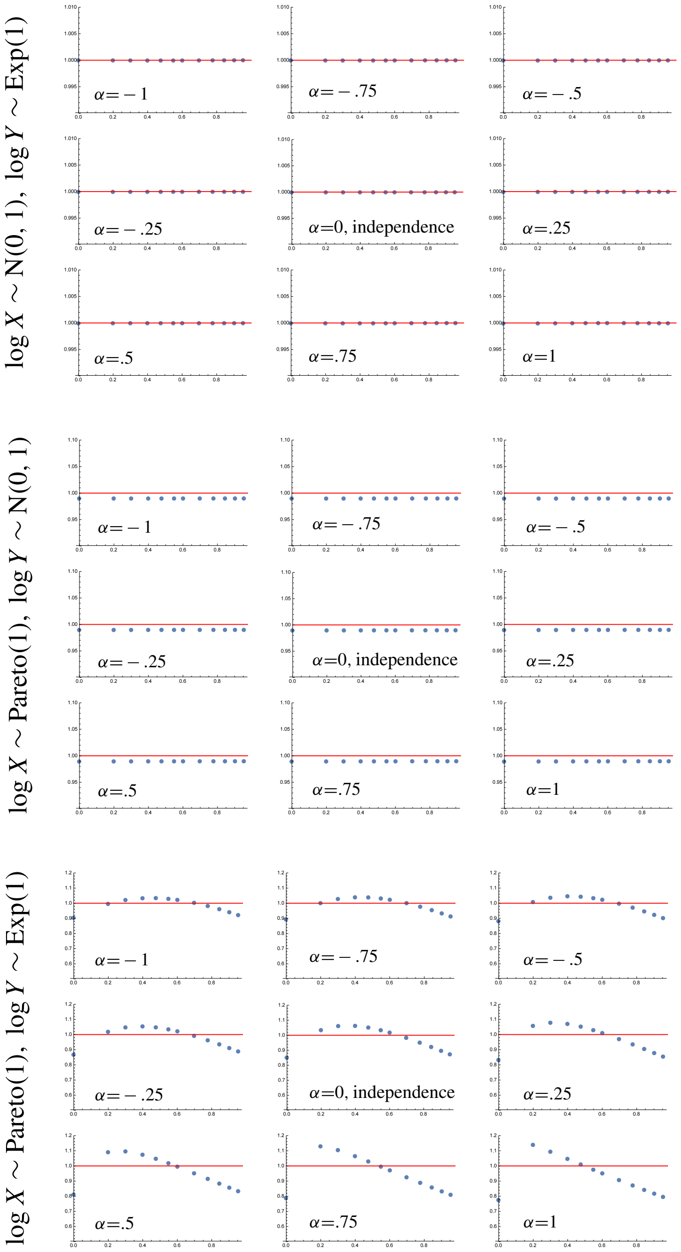

Figure 1. The Ali–Mikhail–Haq 2-copula (see Definition 2.12) modeled on three different sets of marginals with varying dependence parameter $\alpha \in [-1, 1)$. The y-axes of these plots represent the approximate values of the copula PDF of $\log_{10} XY \bmod 1$ at various values of $x \in [0, 1]$, where X and Y are the marginal distributions. The line represents the Benford distribution.

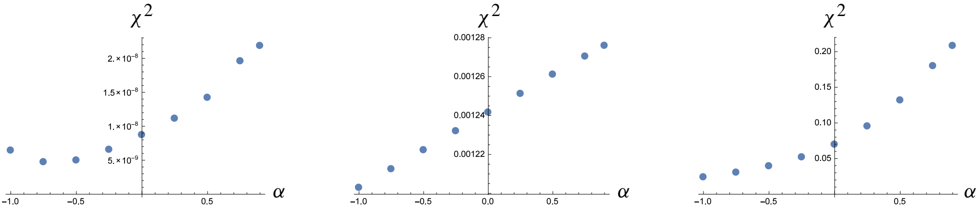

Figure 2. The χ^2-values associated to the plots in Figure 1 for the Ali–Mikhail–Haq copula for pairings (A), (B) and (C), shown from left to right. Each shows the comparison to Benford behavior as α increases. We have 11 degrees of freedom and a significance level of 0.005, so we reject the hypothesis if the value exceeds 2.6. Clearly, only case (C) comes within one order of magnitude of rejecting the hypothesis, so, in loose terms, it is the only case that "comes close" to rejecting the hypothesis. We have imposed a very strict significance level, but it can be clearly seen that a looser significance level of 0.05, perhaps, would likely cause us to reject case (C) entirely.

Ali–Mikhail–Haq copula. Considering the independence case, $\alpha = 0$, in Figure 1 we note that marginal pairings (A) and (B) have an approximately Benford product when independent. Pairing (C), however, does not. From these plots, it is evident that the Ali–Mikhail–Haq copula displays notably consistent Benford behavior, as each plot remains very close to the independence case as α moves over its full range. This is reinforced by the corresponding plots in Figure 2, which display the χ^2 values of each marginal pairing for each value of alpha. We point out that although each plot indicates a general trend away from Benford behavior (the constant function 1), the values for pairing (A) are all smaller than 10^{-7}, making them effectively 0. Similarly, the values for pairing (B) appear to increase linearly, but they are all of order of 10^{-6}. The values for pairing (B) vary from order 10^{-2} to order 10^{-1}, suggesting that the behavior is both significantly less Benford and more variable than the other two pairings.

Gumbel–Barnett copula. These plots suggest that the Gumbel–Barnett copula undergoes even less change over α than the Ali–Mikhail–Haq copula. For pairings (A) and (B), the range for the plots must be restricted to [0.9999, 1.0001] and [0.995, 1.010], respectively, in order to show any change at all. Pairing (C) is not nearly Benford, so its range is expected to vary (recall that the function described by each plot should integrate to 1 in the continuous case). We note, however, that the value at $s = 0$ in pairing (C) appears to vary over a range of 0.1 as α increases. The χ^2 plots in Figure 4 reinforce this interpretation, as in each case the values vary over a significantly small range.

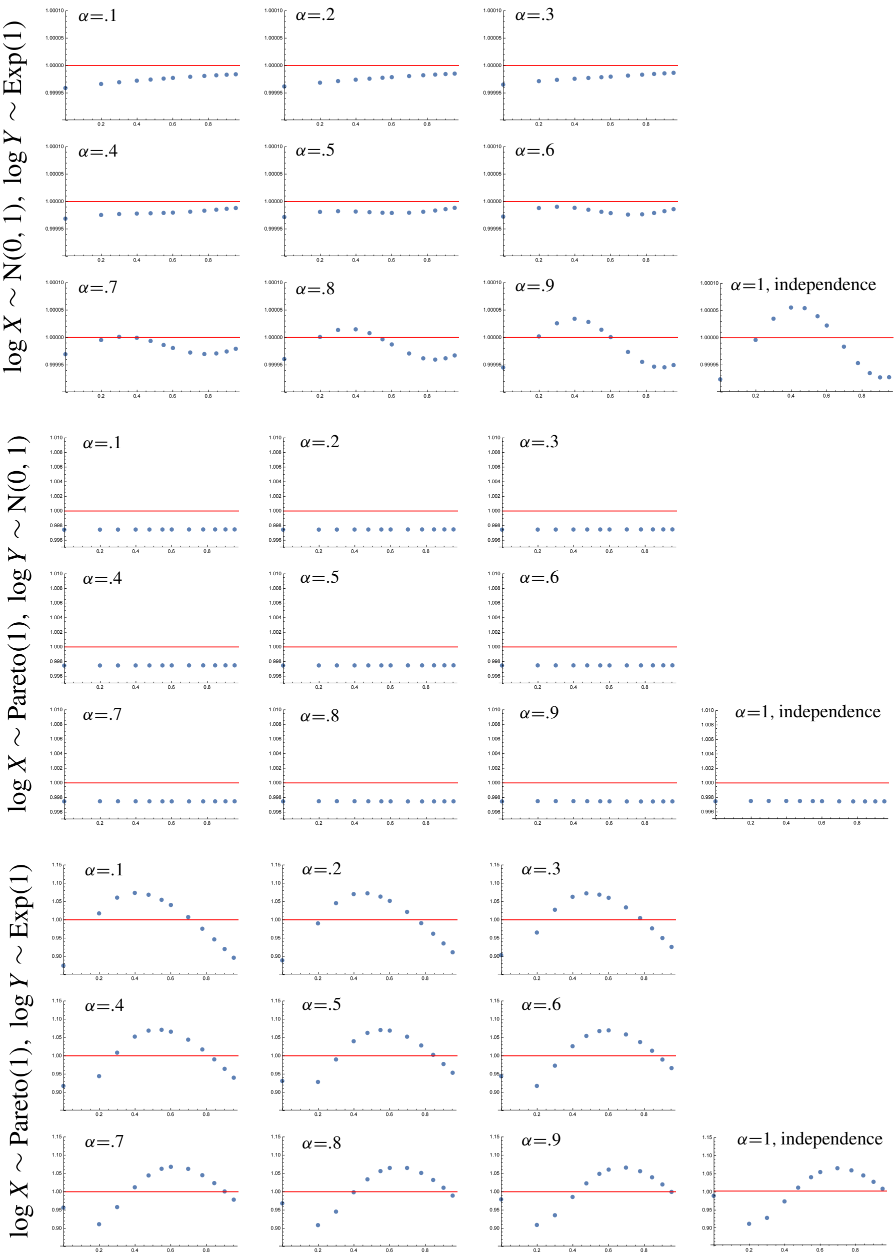

Figure 3. The Gumbel–Barnett 2-copula (see Definition 2.13) modeled on three different sets of marginals with varying dependence parameter $\alpha \in (0, 1]$. The y-axes of these plots represent the approximate values of the copula PDF of $\log_{10} XY \bmod 1$ at various values of $x \in [0, 1]$, where X and Y are the marginal distributions. The line represents the Benford distribution.

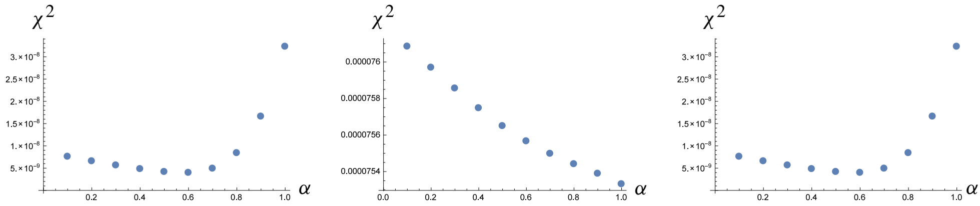

Figure 4. The χ^2-values associated to the plots in Figure 3 for the Gumbel–Barnett copula for pairings (A), (B) and (C), shown from left to right. Each shows the comparison to Benford behavior as α increases. We have 11 degrees of freedom and a significance level of 0.005, so we reject the hypothesis if the value exceeds 2.6. Despite the apparent variation, none of these cases approach the critical value.

This lack of variation is likely due to the actual formula of the copula,

$$C(x, y) = xye^{-\alpha xy}. \tag{4-1}$$

In this case, we have the independence copula, $C(x, y) = xy$ multiplied by a monotonic transformation of the independence copula, e^{-axy}. Thus, it is possible that one or both of these elements serves to preserve the Benford properties of the marginals.

Clayton copula. Unlike the previous two examples, the Clayton copula shows notable variance over α. Although it is not shown here, the independence case for Clayton copulas is $\alpha = 0$. For pairings (A) and (B), it appears that the plots diverge farther and farther away from $y = 1$ as α moves away from 0. For pairing (C), the plots appear to get more random as α grows, and there is no suggestion that Benford behavior may develop as we depart from independence. Furthermore, the plots in Figure 6 show χ^2-values that are significantly higher than those seen for the previous two copulas, suggesting that the dependence imposed by Clayton copula tends to heavily alter any Benford behavior of the marginals.

The results from these three copulas suggest that the preservation of Benford behavior relies more heavily on the underlying structure of the copula than on the Benford behavior of the marginals. Both the Ali–Mikhail–Haq copula and the Gumbel–Barnett copula formulas contain the independence copula, $C(x, y) = xy$. The Clayton copula, however, does not contain the independence copula and is also the only copula of the three to show noticeable variation as the dependence parameter changes.

***n*-copulas.** The previous results suggest that the underlying copula structure has a strong influence on the Benford behavior of 2-copulas. Thus the logical next step is to investigate whether this holds true as we increase the number of marginals.

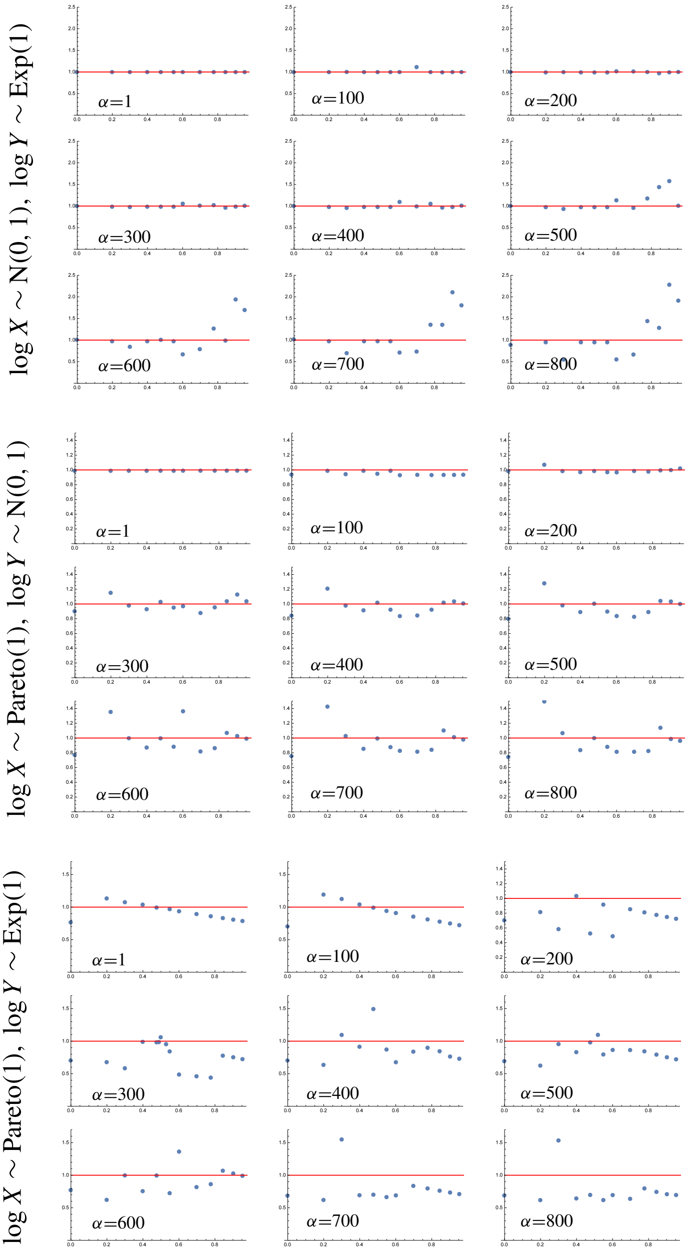

Figure 5. The Clayton 2-copula (see Definition 2.13) modeled on three different sets of marginals with varying dependence parameter $\alpha \in (0, 1]$. The y-axes of these plots represent the approximate values of the copula PDF of $\log_{10} XY \mod 1$ at various values of $x \in [0, 1]$, where X and Y are the marginal distributions. The line represents the Benford distribution.

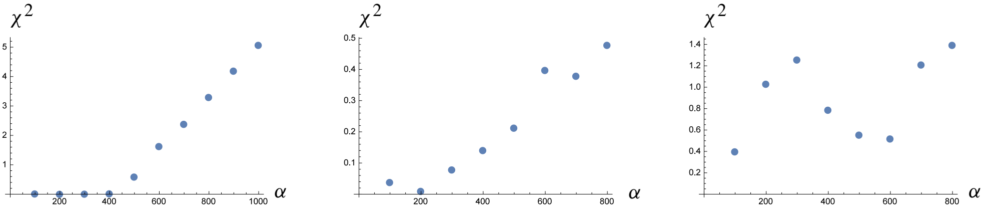

Figure 6. The χ^2 values associated to the plots in Figure 5 for the Clayton copula for pairings (A), (B) and (C), shown from left to right. Each shows the comparison to Benford behavior as α increases. We have 11 degrees of freedom and a significance level of 0.005, so we reject the hypothesis if the value exceeds 2.6. Unlike the previous two copulas, only cases (B) and (C) stay below the critical value. However, the behavior of the plots suggests they will quickly surpass the critical value as α continues to increase.

For all χ^2-tests, we have 8 degrees of freedom and again take a significance level of 0.005. In practice, this means we reject the hypothesis if the value exceeds 1.3.

We consider the most stable of the three previous copulas, the Gumbel–Barnett copula. We fix $\alpha = 0.1$ and set the log, base 10, of all marginals to be identically distributed according to the normal distribution with mean 0 and variance 1, our most Benford-like marginal. We then consider cases where the copula has 2 to 7 marginals. We can see from Figure 7 that the Benford behavior of the Gumbel–Barnett copula begins to fall apart as marginals are added. This is in direct contrast to what would be expected from a central-limit-type property, which should become increasingly more uniform as variables are added. This is further reinforced by the χ^2-values in Figure 8 and suggests that the dependence structure imposed by the copula prevents any leveling-off from happening.

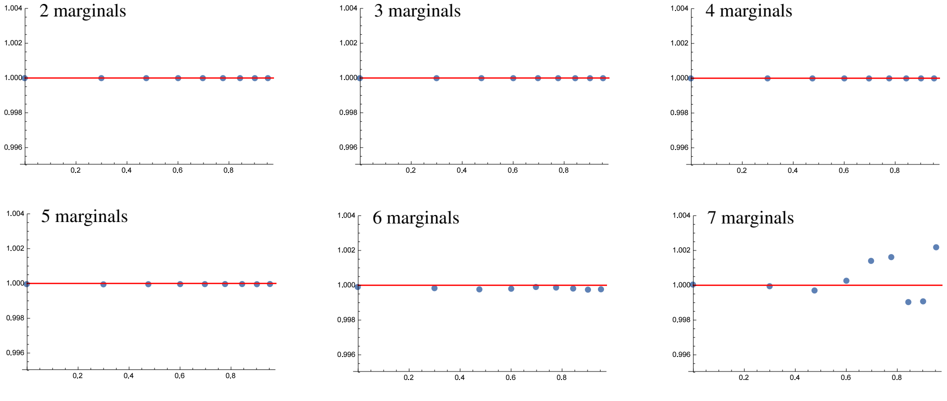

Figure 7. Gumbel–Barnett copula with two to seven marginals.

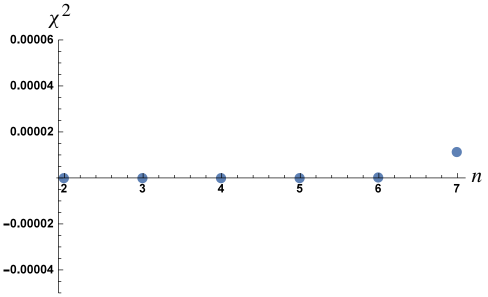

Figure 8. The χ^2-values comparing the behavior of the product to a Benford PDF as the number of marginals increases. We have 8 degrees of freedom and a significance level of 0.005, so we reject the hypothesis if the value exceeds 1.3.

5. Benford distance

Now that we know that we can test for Benford behavior of a product, regardless of dependence, it would be prudent to know how often this behavior is expected to show up. In order to do this, we investigate if the absorptive property of Benford products is common in dependent random variables, or if its presence relies on some sort of proximity to independence.

To get an idea of this, let $\mathcal{W}$ be the space of all n-tuples of continuous random variables $(X_1, X_2, \dots, X_n)$ for which at least one is Benford. Now let us assume that our set of marginals, $(X_1, X_2, \dots, X_n)$, form an element in $\mathcal{W}$. Then we know that their product, assuming independence, will always be Benford.

From this, we can restrict our Benford distance, (3-3), to $\mathcal{W}$ and define it as

$$\epsilon_{s,W} = \left| \sum_{k=-\infty}^{\infty} \int_{u_1=-\infty}^{\infty} \cdots \int_{u_{n-1}=-\infty}^{\infty} (N(u_1, \dots, u_n)|_{\boldsymbol{u}_0}\, du_1 \cdots du_{n-1}) \right|, \tag{5-1}$$

where

$$N(u_1, \dots, u_n) = 1 - \frac{\partial^n C(F_1(u_1), \dots, F_{n-1}(u_{n-1}), F_n(u_n))}{\partial u_1 \cdots \partial u_n}$$

and $\boldsymbol{u}_0$ is defined as in Lemma 3.1. Therefore, our problem becomes minimizing the value of $\epsilon_{s,W} = 0$, as proximity to 0 should indicate proximity to a Benford distribution.

Cases that are ϵ away from Benford. Rather than directly calculating the value of $\epsilon_{s,W}$, it may often be more convenient to provide a bound that depends only on the copula C. If the value of $\partial^n C(F_1(u_1), \dots, F_{n-1}(u_{n-1}), F_n(u_n))/(\partial u_1 \cdots \partial u_n)$ is identically 1 for all values of $(u_1, \dots, u_n)$, then the value of $\epsilon_{s,W}$ will be identically 0 and our product will be Benford. Even though this case does not cover all situations in which our product will be Benford, it suggests that a product's distance from Benford may be related to the distance between the function $\partial^n C(F_1(u_1), \ldots, F_{n-1}(u_{n-1}), F_n(u_n))/(\partial u_1 \cdots \partial u_n)$ and the constant function 1. This brings us to the main result of this section.

Theorem 5.1. *Suppose that $X_1, \ldots, X_n$ are continuous random variables where $(X_1, \ldots, X_n) \in \mathcal{W}$. Assume also that they are jointly described by a copula C, where the function*

$$N(u_1, u_2, \ldots, u_n) = 1 - \frac{\partial^n C(F_1(u_1), \ldots, F_{n-1}(u_{n-1}), F_n(u_n))}{\partial u_1 \cdots \partial u_n}$$

is in $L^1(\mathbb{R}^n)$. Let $U_i = \log_B X_i$ for each i and some base, B, and let F_i be the CDFs of U_i for each i. Then the ***L^1 distance from Benford***, *defined by*

$$\int_0^1 \left| \sum_{k=-\infty}^{\infty} \int_{u_1=-\infty}^{\infty} \cdots \int_{u_{n-1}=-\infty}^{\infty} (N(u_1, \ldots, u_n)|_{\boldsymbol{u}_0})\, du_1 \cdots du_{n-1} \right| ds \tag{5-2}$$

is bounded above by the L^1 norm of N. In other words

$$\int_0^1 \left| \sum_{k=-\infty}^{\infty} \int_{u_1=-\infty}^{\infty} \cdots \int_{u_{n-1}=-\infty}^{\infty} (1-N(u_1, \ldots, u_n)|_{\boldsymbol{u}_0})\, du_1 \cdots du_{n-1} \right| ds \\ \leq \|N(u_1, \ldots, u_n)\|_{L^1}. \tag{5-3}$$

We prove this for the two-dimensional case, as the results in n dimensions proceed similarly. We need the following result (see Appendix 1 for a proof).

Lemma 5.2. *Given C_{uv}, $F(u)$, and $G(v)$ as defined before, we have*

$$\|1 - C_{uv}(u, v)\|_{L^1} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(u) g(v) |1 - C_{uv}(F(u), G(v))|\, du\, dv. \tag{5-4}$$

Proof of Theorem 5.1. From the positivity of f and g we have

$$\int_0^1 \left| \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} f(u) g(s+k-u)(1 - C_{uv}(F(u), G(s+k-u)))\, du \right| ds \\ \leq \int_0^1 \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} f(u) g(s+k-u) |1 - C_{uv}(F(u), G(s+k-u))|\, du\, ds. \tag{5-5}$$

We investigate exactly what region (5-5) covers. The lines shown in Figure 9 are the sets $A_k = \{(u, v) : v = s + k - u\}$. We integrate

$$f(u) g(s+k-u)(1 - C_{uv}(F(u), G(s+k-u)))$$

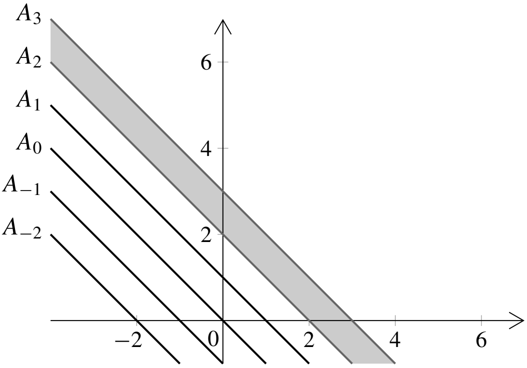

Figure 9. The plane broken up into a few of the sections A_k.

along each of these lines and sum the results over k. The shaded region shows the area covered when A_2 is integrated over s from 0 to 1.

As all of our sums and integrals converge absolutely, by Fubini's theorem we may switch our sum and integral in (5-5) and get

$$\int_0^1 \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} f(u)g(s+k-u)|1-C_{uv}(F(u), G(s+k-u))|\,du\,ds$$

$$= \sum_{k=-\infty}^{\infty} \int_0^1 \int_{-\infty}^{\infty} f(u)g(s+k-u)|1-C_{uv}(F(u), G(s+k-u))|\,du\,ds. \tag{5-6}$$

From this, we can quickly see that for any k,

$$\int_0^1 \int_{-\infty}^{\infty} f(u)g(s+k-u)|1-C_{uv}(F(u), G(s+k-u))|\,du\,ds \tag{5-7}$$

is the integral of $f(u)g(s+k-u)|1-C_{uv}(F(u), G(s+k-u))|$ over a region inbetween and including A_k and A_{k+1}, just like the shaded region in Figure 9. Therefore, (5-6) is the sum of the integrals of $f(u)g(s+k-u)|1-C_{uv}(F(u), G(s+k-u))|$ over all of these (disjoint) regions (over all k), which is equivalent to integrating over all of $\mathbb{R}^2$, giving us

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(u)g(v)|1-C_{uv}(F(u), G(v))|\,du\,dv. \tag{5-8}$$

Finally, from Lemma 5.2, we know that this is equal to $\|1-C_{uv}(u, v)\|_{L^1}$. □

Consequences of an L^1 bound in $\mathbb{R}^2$. What Theorem 5.1 provides is a way to understand the behavior of our probabilities. To see this, let $\mathcal{S} \subset [0, 1]$ be the region over which $\epsilon_{s,W} > \epsilon_N$. If $\epsilon_{s,W}$ is large on $\mathcal{S}$, then the measure of $\mathcal{S}$ must be small in order to conform to (5-3), which requires that if $\|1 - C_{uv}(u, v)\|_{L^1} \leq \epsilon_N$, then $\int_0^1 \epsilon_{s,W}\, ds \leq \epsilon_N$ as well. In fact, the following corollary proves that Theorem 5.1 provides useful information regarding how large $|\mathcal{S}|$ can be.

Corollary 5.3. *Let $\mathcal{S} \subset [0, 1]$ be the set $\{s : \epsilon_{s,W} \geq \epsilon\}$. Then*

$$|\mathcal{S}| \leq \frac{\|1 - C_{uv}(u, v)\|_{L^1}}{\epsilon}. \tag{5-9}$$

Proof. This result comes directly from Markov's inequality:

$$|\{s : \epsilon_{s,W} \geq \epsilon\}| \leq \frac{1}{\epsilon} \int_0^1 \epsilon_{s,W} \leq \frac{\|1 - C_{uv}(u, v)\|_{L^1}}{\epsilon}. \qquad \square$$

6. Applications, future work, and conclusion

Fitting copulas. The results of Section 3 allow us to determine the Benford behavior of the product of n distributions jointly modeled by a specific copula. However, we may wish to go in the other direction and, instead, find a copula that best fits n correlated data sets. Statisticians have several methods for testing the goodness-of-fit to find the best choice of copula in these situations (see [Genest et al. 2006] for some examples and an analysis of several forms of goodness-of-fit tests), but it is not known whether or not these goodness-of-fit tests take Benford behavior into account. That is to say, will the prescribed copula mimic the Benford behavior observed in the data?

The results of Section 4 have shown us that the product of the same set of marginals will not display the same Benford behavior when modeled by different copulas. Thus, Benford behavior is not guaranteed. A natural next step is to investigate how the goodness-of-fit of a copula may or may not be correlated with how well it preserves the expected Benford behavior of the product of two or more marginals. A comparison between the L^1 norm and well-known goodness of fit tests would enable us to see whether or not a strong Benford fit corresponds to a well-fit distribution as a whole. Furthermore, if a stronger Benford fit may be shown to correspond to a smaller L^1 bound, then we may be able to define this bound as a new goodness of fit test for distributions with one or more Benford marginals.

With these results, it is now reasonable to begin searching for specific situations where this analysis of dependence structure would prove useful. As Benford analysis for single-variate distributions has already proven useful in a variety of situations, it is reasonable to assume that the multivariate analysis will be similarly useful. Thus future work may also be directed towards investigating the various applications of these results and how they may be used to improve current practices.

Conclusion. In fields such as actuarial sciences and statistics Benford's law is useful for fraud detection. Furthermore, copulas are a highly effective tool for modeling systems with dependencies. In Section 3 we demonstrated that Benford behavior for dependent variables modeled by a copula may be detected and therefore analyzed to investigate the product of the variables. Thus these results indicate that the Benford's law methods used by professionals on single-variate and/or independent data sets are now at the disposal of individuals who wish to model dependent data via a copula. We then applied these results in Section 4, where we observed that the preservation of Benford behavior appears to rely more heavily on the structure of the copula than on the marginals.

Essentially, the results of Section 3 permit analyses like those carried out in [Cuff et al. 2015; Durst et al. 2016] in which a known distribution, in these cases the Weibull and the inverse-gamma distributions, is analyzed to determine the conditions under which Benford behavior should arise. Once these conditions are established, any non-Benford data set which is expected to come from such a distribution may be considered suspicious enough to warrant a fraud investigation. In the case of copulas, the results of Section 3 allow one to conduct this exact method of analysis on the product of n random variables jointly modeled by a copula C.

Finally, in Section 5 we encountered a useful consequence of considering a distribution's L^1 distance from a Benford distribution to determine a useful bound for this Benford distance. We determined that the Benford distance of a product of n random variables will always be bounded above by the distance between the copula PDF and the class of copulas whose PDFs are identically 1.

Appendix A: Proofs for supporting lemmas and theorems

Lemma 3.1. *Given X and Y positive, continuous random variables with joint distribution modeled by the absolutely continuous copula C, let $U = \log_B X$ and $V = \log_B Y$ for some base, B, and let the (marginal) CDFs of U and V be $F(u)$ and $G(v)$, respectively. Also, let $f(u)$ and $g(v)$ be the PDFs of U and V, respectively. Then*

$$\text{Prob}((U+V)\ \mathit{mod}\ 1 \le s) = \int_0^s \biggl(\sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} C_{uv}(F(u), G(s+k-u)) f(u) g(s+k-u)\, du \biggr). \tag{A-1}$$

Therefore, the PDF of $(U+V)$ mod 1 *is given by*

$$\sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} C_{uv}(F(u), G(s+k-u)) f(u) g(s+k-u)\, du. \tag{A-2}$$

Proof. By the invariance of copulas under monotonically increasing functions (Theorem 2.9), we know that the joint CDF of U and V is given by the same copula as X and Y. Thus, the joint CDF of U and V is given by

$$C(F(U), G(V)). \tag{A-3}$$

Then, by definition, the joint PDF of U and V is given by the mixed partial derivative.

$$\begin{aligned}\frac{\partial}{\partial v}\frac{\partial}{\partial u}C(F(u), G(v)) &= C_{uv}(F(u), G(v))f(u)g(v) + C_u(F(u), G(v))\frac{\partial}{\partial v}f(u)\\ &= C_{uv}(F(u), G(v))f(u)g(v). \qquad \text{(A-4)}\end{aligned}$$

Note that we assume that $du/dv = 0$ since all dependence between U and V is modeled by C.

Note, also, that $\text{Prob}(XY \le 10^s) = \text{Prob}((U+V) \le s)$. Thus we have

$$\begin{aligned}&\text{Prob}((U+V) \bmod 1 \le s)\\ &\qquad = \sum_{k=-\infty}^{\infty}\int_{u=-\infty}^{\infty}\int_{v=k-u}^{s+k-u} C_{uv}(F(u), G(v))f(u)g(v)\,dv\,du. \qquad \text{(A-5)}\end{aligned}$$

If XY is Benford, then (A-5) will equal s for all s. It is, however, easier to test the PDF than the CDF. So we differentiate with respect to s. Let $C_1(u, v)$ be the antiderivative of $C_{uv}(F(u), G(v))f(u)g(v)$ with respect to v. Then

$$\begin{aligned}&\frac{\partial}{\partial s}\sum_{k=-\infty}^{\infty}\int_{u=-\infty}^{\infty}\int_{v=k-u}^{s+k-u} C_{uv}(F(u), G(v))f(u)g(v)\,dv\,du\\ &\qquad = \frac{\partial}{\partial s}\sum_{k=-\infty}^{\infty}\left(\int_{u=-\infty}^{\infty}(C_1(u, s+k-u) - C_1(u, k-u))\right)du\\ &\qquad = \sum_{k=-\infty}^{\infty}\int_{-\infty}^{\infty} C_{uv}(F(u), G(s+k-u))f(u)g(s+k-u)\,du. \qquad \square\end{aligned}$$

Lemma 3.2. *Given U and V, continuous random variables modeled by the copula C with marginals F and G, respectively, there exist a_1, a_2, b_1, and b_2 completely dependent on F or G such that $a_1 < a_2$ and $b_1 < b_2$, and*

$$\begin{aligned}&\sum_{k=-\infty}^{\infty}\int_{-\infty}^{\infty} C_{uv}(F(u), G(s+k-u))f(u)g(s+k-u)\,du\\ &\qquad = \sum_{k=b_1}^{b_2}\int_{a_1}^{a_2} C_{uv}(F(u), G(s+k-u))f(u)g(s+k-u)\,du + E_{a,b}(s), \qquad \text{(A-6)}\end{aligned}$$

where $E_{a,b}(s) \to 0$ as $a_1, b_1 \to -\infty$ and $a_2, b_2 \to \infty$. Thus, for any $\epsilon > 0$, there exists $|a_1|$, $|a_2|$, $|b_1|$, and $|b_2|$ large enough such that $|E_{a,b}(s)| \le \epsilon$.

Proof. Since both the sum and the integral are convergent, the proofs for a_1, a_2 and b_1, b_2 are nearly identical, so we only provide the work here for a_1 and a_2. The same steps may be used in the proof for b_1 and b_2. We also know that $C_{uv}(F(u), G(s+k-u)) f(u) g(s+k-u)$ must go to 0 as u goes to $\pm\infty$ because of this convergence. Thus we choose to prove the case where f and/or g converge faster than C_{uv}. If C_{uv} were to converge faster, the results derived here would still suffice. We prove that for any $\epsilon > 0$ we can find a_1 and a_2 such that, for all $u \le a_1$ and all $u \ge a_2$, we have

$$|C_{uv}(F(u), G(s+k-u)) f(u) g(s+k-u)| \le \epsilon.$$

Let $\epsilon > 0$, set s and k to be constant, and assume C_{uv} is nonzero everywhere. If C_{uv} is zero at any point, then we have a trivial case. Because F and G are CDFs, we know that $f \to 0$ as $u \to \pm\infty$ and $g \to 0$ as $-u \to \pm\infty$; thus, we may choose a_{f1}, a_{f2}, a_{g1}, and a_{g2} such that, for all $u \le a_{f1}$ and all $u \ge a_{f2}$, we have

$$f(u) \le \sqrt{\frac{\epsilon}{C_{uv}(F(u)G(s+k-u))}}. \tag{A-7}$$

The same can be done for g such that, for all $u \ge a_{g1}$ and all $u \le a_{g2}$, we have

$$g(s+k-u) \le \sqrt{\frac{\epsilon}{C_{uv}(F(u)G(s+k-u))}}. \tag{A-8}$$

Thus, we let $a_1 = \min\{a_{f1}, a_{g1}\}$ and $a_2 = \max\{a_{f2}, a_{g2}\}$. Then, for all $u \le a_1$ and all $u \ge a_2$, we have

$$|C_{uv}(F(u), G(s+k-u)) f(u) g(s+k-u)| \le \epsilon. \qquad \square$$

Lemma 5.2. *Given C_{uv}, $F(u)$, and $G(v)$ as defined in Theorem 5.1, we have*

$$\|1 - C_{uv}(u, v)\|_{L^1} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(u) g(v) |1 - C_{uv}(F(u), G(v))| \, du \, dv. \tag{A-9}$$

Proof. We know that u and v are defined on $[0, 1]$. Thus,

$$\|1 - C_{uv}(u, v)\|_{L^1} = \int_0^1 \int_0^1 |1 - C_{uv}(u, v)| \, du \, dv. \tag{A-10}$$

However, by a simple change of variables $u \to F(u)$, $v \to G(v)$ (defined as CDFs, just like before, so their derivatives are $f(u)$ and $g(v)$, both of which are greater than or equal to 0), we get

$$\|1 - C_{uv}(u, v)\|_{L^1} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u)g(v)|1 - C_{uv}(F(u), G(v))|\,du\,dv. \qquad \square$$

Appendix B: Computationally testing for Benford behavior

In this section, we use Clayton copulas (see Definition 2.11) to determine the Benford behavior of different combinations of marginals. We specifically look at marginals of the form $X = 10^U$ and $Y = 10^V$, where U and V are N(0, 1) or Exp(1). In all analyses, we let $\alpha = 2$ and $B = 10$. We also provide the independence case for each set of marginals to allow for comparison.

All numerical results and coding were done using Wolfram Mathematica, version 10.1 or later.

Case 1: U and $V \sim N(0, 1)$. Given our definition of X and Y, (3-2) we first determine acceptable values for a_1, b_1, a_2, and b_2 by using an error analysis to test whether or not -10 and 10 should be acceptable values for a_1 and a_2.

We generated a list of the errors caused by truncating the integral at these values for various values of s; the first value of each triple in the list is s, the second is the lower error and the third is the upper error:

```
In[262]:= errorsb =
 Table[{N[Log[10, s]], ea[Log[10, s]], eb[Log[10, s]]}, {s, 1, 9}]

Out[262]= {{0., 6.86784*10^-22, 1.28213*10^-22},
           {0.30103, 9.38169*10^-24, 1.28257*10^-22},
           {0.477121, 3.03058*10^-25, 1.28274*10^-22},
           {0.60206, 2.74232*10^-26, 1.28266*10^-22},
           {0.69897, 4.3443*10^-27, 1.28249*10^-22},
           {0.778151, 9.77379*10^-28, 1.28234*10^-22},
           {0.845098, 2.79567*10^-28, 1.28223*10^-22},
           {0.90309, 9.52164*10^-29, 1.28216*10^-22},
           {0.954243, 3.70245*10^-29, 1.28213*10^-22}}
```

To determine the error caused by truncating the integral, we used the approximation method detailed in Section 3. As the list shows, the error is on the order of 10^{-22} or smaller, indicating that our selections for a_1 and a_2 are good bounds. We took the sum from $k = -20$ to $k = 20$ because we know this will be sufficient, as indicated by the convergence in Figure 10 below.

We now plot in Figure 10 the value of our truncated form of our PDF for different values of s. The line $y = 1$ is included to demonstrate how close to 1 our PDF is for all values of s, suggesting that the product of X and Y with joint PDF modeled by a Clayton copula with $\alpha = 2$ should display Benford behavior.

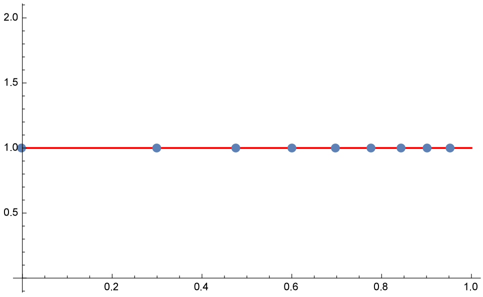

Figure 10. $U \sim \mathrm{N}(0, 1)$ and $V \sim \mathrm{N}(0, 1)$.

Case 2: $U \sim N(0, 1)$ and $V \sim \mathrm{Exp}(1)$. A similar analysis as before was conducted on this new set of variables. Through an identical analysis, we defined the bounds for our integral to be $a = -5$ and $b = 10$, and provide the accumulated errors in the code below, where the first term in each pair is s and the second and third are the lower and upper errors, respectively:

```
In[419]:= Table[{N[Log[10, s]], ea2[Log[10, s]], eb2[Log[10, s]]},
                {s, 1, 9}]

Out[419]= {{0., 3.30411*10^-21, 1.23628*10^-22},
           {0.30103, 2.43577*10^-21, 1.27151*10^-22},
           {0.477121, 2.03887*10^-21, 1.31758*10^-22},
           {0.60206, 1.79746*10^-21, 1.32924*10^-22},
           {0.69897, 1.63021*10^-21, 1.32387*10^-22},
           {0.778151, 1.50526*10^-21, 1.31045*10^-22},
           {0.845098, 1.40717*10^-21, 1.2933*10^-22},
           {0.90309, 1.32741*10^-21, 1.27456*10^-22},
           {0.954243, 1.26084*10^-21, 1.25536*10^-22}}
```

As we can see, the errors are still very, very small.

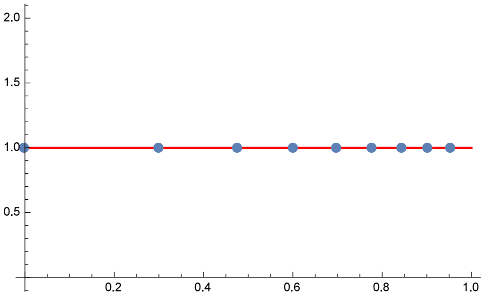

Figure 11. $U \sim N(0, 1)$ and $V \sim \mathrm{Exp}(1)$.

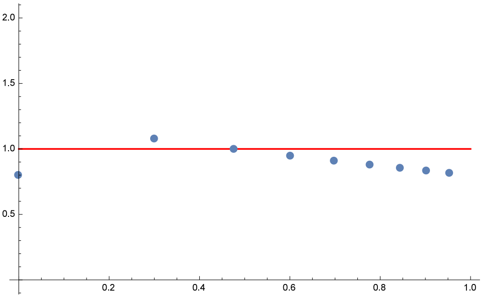

Figure 12. U and $V \sim \mathrm{Exp}(1)$.

We now plot in Figure 11 the value of our truncated form of our PDF for various s. We again note how close the PDF remains to 1 for all values of s, suggesting that the product of X and Y, with joint PDF modeled by a Clayton copula with $\alpha = 2$ should display Benford behavior.

<u>Case 3</u>: $U \sim \mathrm{Exp}(1)$ and $V \sim \mathrm{Exp}(1)$.

Finally, we conduct our analysis on the case of two exponentials. Our error terms for $a = 25$ are generated in the code below (By inspection, we can tell that $C_{uv}(F(u), G(s+k-u))f(u)g(s+k-u)$ will be zero for negative values of u). Again we choose k from 0 to 50, and the first term in each pair is s:

```
In[363]:= Table[{N[Log[10, s]], N[eb1[Log[10, s]]]}, {s, 1, 9}]

Out[363]= {{0., 5.57839*10^-11}, {0.30103, 5.73736*10^-11},
           {0.477121, 5.94524*10^-11}, {0.60206, 5.99786*10^-11},
           {0.698970, 5.97362*10^-11}, {0.778151, 5.91306*10^-11},
           {0.845098, 5.83566*10^-11}, {0.90309, 5.75112*10^-11},
           {0.954243, 5.66447*10^-11}}
```

Now that we know $a = 25$ provides a small enough error, we plot, once again, the PDF for various values of s, as shown in Figure 12. We quickly see that the PDF does not converge to 1 and actually changes for each value of s. Even though we only take our sum out to $k = \pm 50$, this is enough to suggest that Benford behavior is unlikely.

<u>Checking the marginals</u>: To understand why this might be the case, we took a look at the marginal distributions. We note that $X = 10^U$, where $U \sim N[0, 1]$ is a closely Benford distribution with $\chi^2 \approx 0.9918$, but $Y = 10^V$, where $V \sim \mathrm{Exp}[1]$ is not, with $\chi^2 \approx 0.7084$. Thus, in the independent case we would expect that two variables modeled like X, or any product with X, should yield a Benford distribution. The product of two variables modeled like Y, however, should not be Benford.

Acknowledgements

The authors were partially supported by NSF grants DMS-1347804, DMS-1265673 and DMS-1561945, and by Williams College. We thank our colleagues from the Williams College SMALL REU program for many helpful discussions, and the referee for many suggestions which improved the exposition of the paper.

References

[Battersby 2009] S. Battersby, “Statistics hint at fraud in Iranian election”, *New Scient.* **202**:2714 (2009), 10.

[Becker et al. 2018] T. Becker, D. Burt, T. C. Corcoran, A. Greaves-Tunnell, J. R. Iafrate, J. Jing, S. J. Miller, J. D. Porfilio, R. Ronan, J. Samranvedhya, F. W. Strauch, and B. Talbut, “Benford’s law and continuous dependent random variables”, *Ann. Physics* **388** (2018), 350–381. MR Zbl

[Benford 1938] F. Benford, “The law of anomalous numbers”, *Proc. Amer. Philos. Soc.* **78**:4 (1938), 551–572. Zbl

[Benford 2009] A. Berger, T. P. Hill, and E. Rogers, “Benford Online Bibliography”, database, 2009, available at http://www.benfordonline.net.

[Berger and Hill 2015] A. Berger and T. P. Hill, *An introduction to Benford’s law*, Princeton Univ. Press, 2015. MR Zbl

[Cuff et al. 2015] V. Cuff, A. Lewis, and S. J. Miller, “The Weibull distribution and Benford’s law”, *Involve* **8**:5 (2015), 859–874. MR Zbl

[Durst et al. 2016] R. F. Durst, C. Huynh, A. Lott, S. J. Miller, E. A. Palsson, W. Touw, and G. Vriend, “The inverse gamma distribution and Benford’s law”, preprint, 2016. arXiv

[Fisher 1997] N. I. Fisher, “Copulas”, pp. 159–163 in *Encylopedia of statistical sciences, Update Vol. 1*, edited by S. Kotz et al., Wiley, New York, 1997.

[Fontes and Magni 2004] F. A. C. C. Fontes and L. Magni, “A generalization of Barbalat’s lemma with applications to robust model predictive control”, pp. art. id. 395 in *Proceedings sixteenth International Symposium on Mathematical Theory of Networks and Systems* (Leuven, Belgium, 2004), edited by B. De Moor et al., Katholieke Univ. Leuven, 2004.

[Genest et al. 2006] C. Genest, J.-F. Quessy, and B. Rémillard, “Goodness-of-fit procedures for copula models based on the probability integral transformation”, *Scand. J. Statist.* **33**:2 (2006), 337–366. MR Zbl

[Iafrate et al. 2015] J. R. Iafrate, S. J. Miller, and F. W. Strauch, “Equipartitions and a distribution for numbers: a statistical model for Benford’s law”, *Phys. Rev. E* (3) **91**:6 (2015), art. id. 062138. MR

[Kpanzou 2007] T. A. Kpanzou, “Copulas in statistics”, preprint, African Inst. Math. Sci., 2007, available at https://tinyurl.com/kpancop.

[Miller 2015] S. J. Miller (editor), *Benford’s law: theory and applications*, Princeton Univ. Press, 2015. MR Zbl

[Nelsen 2006] R. B. Nelsen, *An introduction to copulas*, 2nd ed., Springer, 2006. MR Zbl

[Nigrini 1999] M. J. Nigrini, “I’ve got your number”, *J. Accountancy* **187**:5 (1999), 79–83.

[Nigrini and Mittermaier 1997] M. Nigrini and L. J. Mittermaier, “The use of Benford’s law as an aid in analytical procedures”, *Auditing* **16**:2 (1997), 52–67.

[Raimi 1976] R. A. Raimi, “The first digit problem”, *Amer. Math. Monthly* **83**:7 (1976), 521–538. MR Zbl

[Singleton 2011] T. W. Singleton, "Understanding and applying Benford's law", *ISACA J.* **3** (2011), 1–4.

[Tao 2010] T. Tao, *An epsilon of room, II: Pages from year three of a mathematical blog*, Grad. Studies in Math. **117**, Amer. Math. Soc., Providence, RI, 2010. MR Zbl

[Wu et al. 2007] F. Wu, E. Valdez, and M. Sherris, "Simulating from exchangeable Archimedean copulas", *Comm. Statist. Simulation Comput.* **36**:5 (2007), 1019–1034. MR Zbl

Received: 2019-01-16 Revised: 2019-04-11 Accepted: 2019-04-15

rfd1@williams.edu *Department of Mathematics and Statistics, Williams College,*
Williamstown, MA, United States

Current address: *Division of Applied Mathematics, Brown University,*
Providence, RI, United States

sjm1@williams.edu *Department of Mathematics and Statistics, Williams College,*
Williamstown, MA, United States

Current address: *Department of Mathematics, Carnegie Mellon University,*
Pittsburgh, PA, United States

involve

msp.org/involve

INVOLVE YOUR STUDENTS IN RESEARCH

Involve showcases and encourages high-quality mathematical research involving students from all academic levels. The editorial board consists of mathematical scientists committed to nurturing student participation in research. Bridging the gap between the extremes of purely undergraduate research journals and mainstream research journals, *Involve* provides a venue to mathematicians wishing to encourage the creative involvement of students.

Cover: Alex Scorpan

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2019 is US $195/year for the electronic version, and $260/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFLOW® from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

involve

2019 vol. 12 no. 7