Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

DISTRIBUTIONAL DISCRIMINATION USING KOLMOGOROV-SMIRNOV STATISTICS AND KULLBACK-LEIBLER DIVERGENCE FOR GAMMA, LOG-NORMAL, AND WEIBULL DISTRIBUTIONS

Mario Andriulli
James K. Starling
Blake Schwartz

Center for Data Analysis and Statistics
Department of Mathematical Sciences
United States Military Academy
606 Thayer Road
West Point, NY 10996, USA

ABSTRACT

This research compares two methods of choosing a distribution to match sample data: Kullback-Leibler (KL) divergence and the Kolmogorov-Smirnoff (KS) statistic. We generate sample data from a known distribution (we used the gamma, log-normal, and Weibull distributions), find best matches to the data for each candidate distribution using maximum likelihood parameter estimation, then use KL divergence and the KS statistic to choose a best fit for the data. Using Monte-Carlo simulation, we estimate a probability of correct selection for KL divergence and the KS statistic by determining how frequently each method correctly selects the known underlying distribution. Results vary based on the data-generating distribution type, parameters, and sample size, but we find that KL divergence generally outperforms the KS statistic except in a few rare instances. This is an important result, as the two measures are not directly comparable, and are competing methods for measuring the distance between two distributions.

1 INTRODUCTION

1.1 Purpose

In modeling, selecting the distribution that most accurately reflects a real-world process is essential. But how essential? For some logistics applications, incorrect selection leads to poor life cycle scheduling, and increased operations and maintenance costs in best-case scenarios. In worst-case scenarios, incorrect selection can yield catastrophic results. To examine the impacts of poor distribution selection, this paper explores two methods of quantifying the difference between a forecast distribution and actual system performance when the underlying distribution of the system is not known. We will answer the question of quantifying these differences for varying sample sizes.

1.2 Motivation

Extensive research comparing any combination of two probability distributions is readily available. Evaluating the probability of correct selection among pairs of distributions leaves potential gaps in reliability analysis when considering the possibility that there may be other underlying distributions. Additionally, some research has also investigated variations in the KS statistic and KL divergence resulting from the probability of correct selection. However, this paper takes these evaluations further by analyzing three distributions commonly encountered in real-world applications. The two theoretical models applied in this

study are Kullback-Leibler divergence and the Kolmogorov-Smirnoff statistic test. Although both models are used to evaluate the best fit between sample and reference distributions, they are computationally significantly different. The KS statistic is a measure of the most significant distance between the two distributions when presented as cumulative distribution functions, whereas KL divergence measures the probability that the observed data represents the specified theoretical model. The KL divergence has also been termed relative entropy. It takes a more holistic measure of the difference in information between the proposed theoretical probability distribution and the sample distribution.

Further motivating this research is the consideration that using an incorrect distribution to model behavior may have significant ramifications. One such case involves reliability behavior for circuits, as described by Basavalingappa et al. (2017). Electromigration failure tests are used to determine expected lifetimes of integrated circuits, and log-normal or Weibull distributions are typically used to model circuit reliability. Because circuit life is generally quite long, reliability studies are typically conducted at accelerated conditions, with high current and temperatures, and the results are then extrapolated and scaled to reflect actual intended use conditions. After scaling and extrapolation, the Weibull- and log-normal-derived predictions differ from each other dramatically, so the choice of distribution is important to best characterize actual circuit behavior.

Another application where the correct distribution matters involves modeling reliability of bridges repaired or reinforced with fiber-reinforced polymer (FRP), described by Atedero et al. (2004). FRP composites are often used to repair or strengthen deteriorating bridges. Due to field conditions and the "wet layup process" used to apply FRP, there is a potential for vast variability in FRP strength and, thus, the reinforced bridge's reliability. Atedero et al. (2004) use multiple distributions (gamma, Gaussian, log-normal, and Weibull) to model the strength of FRP panels fabricated under field conditions and analyze the goodness-of-fit of the distributions using Pearson's Chi-square test. Overall, the Weibull best modeled the FRP data for strength and thickness, but the log-normal best modeled the tensile modulus data. The authors find that a reasonable estimate for each of these characteristics is the distribution mean plus or minus two standard deviations, so the choice of distributions significantly impacts repaired bridge reliability estimates.

Finally, in a third application, Basu et al. (2009) explore using the gamma, Gaussian, generalized exponential log-normal, and Weibull distributions to model strength reliability for brittle materials. While the Weibull distribution has been most commonly used for this purpose, the authors find that the gamma and log-normal distributions may better model strength data in some circumstances. They find the MLE for the gamma, Gaussian, generalized exponential log-normal, and Weibull distributions to fit known brittle material strength data and discuss different measures to select the best distribution, including maximum likelihood, KS statistic, and chi-square distance. Brittle materials have been in higher demand recently, as myriad new applications requiring high hardness, rigidity, and strength are needed at high temperatures. Unfortunately, there is significant variability in the strength of ceramics and other brittle materials, even among apparently identical samples. Efforts are underway to improve ceramic strength and better model the strength of brittle materials. Our research could be usefully applied here as well.

With these motivating applications in mind, the importance of correctly selecting a distribution that best fits sample data is apparent. This paper compares the differences between an MLE-calculated best fit distribution and known distributions, including the one from which the data was generated. The best measure of distance, however, is a contested topic. We compare the results using KS statistic and KL divergence to explore when each might be the preferred technique. To limit the scope of this study, we focus on the the three continuous distributions, gamma, log-normal, and Weibull, commonly referenced in the motivating research.

2 RELATED LITERATURE

In recent years, extensive research relevant to establishing the foundation for our analysis has been conducted. Each of the publications discussed expounds upon to a theory which we have considered in developing our

approach. All of the publications considered center on one or more research concepts, but a broad search did not uncover research that has evaluated the full breadth of topics discussed in this paper. The three topics underpinning this research are KS statistic, KL divergence, and probability of correct selection. Below, we summarize the relationships between some of the more insightful research papers in discrimination analysis and our research.

Kundu and Manglick (2004) consider the selection of gamma and log-normal distributions with unknown shape and scale. To achieve discrimination between these two distributions, the authors apply a ratio of maximum likelihoods (RML). In this paper, the probability of correct selection (PCS) is estimated using Monte Carlo simulations based on the distributions for an array of sample sizes, $n = 20, 40, 60, 80, 100$, and replicating the process 10,000 times. The Kundu and Manglick (2004) paper does not consider a comparison with KL Divergence.

An example was given in Peng et al. (2015) showing that the PCS can decrease in some situations where certain distributions are more heavily sampled. Additionally, the authors provide a general formulation of the probability of correct selection for k alternatives:

$$PCS \equiv P(\bar{X}_1 > \bar{X}_2, \cdots, \bar{X}_1 > \bar{X}_k), \tag{1}$$

where $\bar{X}_i$ is the sample mean of alternative i. We will describe how we modify this definition of PCS in Section 3.2 using the KL divergence and KS statistic where we will limit the alternatives to the gamma, log-normal, and Weibull distributions ($i \in \{$gamma, log-normal, and Weibull$\}$ with $k = 3$).

Dey and Kundu (2009) measured the discrimination between the log-normal, Weibull, and generalized exponential distributions using a discrimination process comparing the likelihood functions for each of the three distributions, combining both asymptotic results and Monte-Carlo simulation. They then use the Kolmogorov-Smirnov statistic for two distribution functions F and G to explain the results. Additionally, they continue to investigate instances where Type-I censoring occurs. Adopted from this paper is the use of a "scale" value of 1 for all distributions as well as the shapes used in varying the Weibull distribution, which are listed in section 4.1. While Dey and Kundu (2009) focus on the likelihood function, this research evaluates the probability of correct selection using the KS statistic and KL divergence.

Das and Park (2012) investigated several examples and considered the selection comparison between gamma and log-normal distributions using generalized linear models (GLM). This paper emphasized how incorrect model assumptions can mask significant factors, making them appear insignificant and resulting in serious error. Their example with heteroscedasticity demonstrates that modeled distributions should have a close enough fit, otherwise a result of "all models are wrong" can be achieved. Das and Park (2012) show how the GLM discrimination rule used in this paper can mislead a comparison for small sample sizes.

2.1 Mathematical Models

This paper assesses the performance of KL convergence and the KS statistic using the gamma, log-normal, and Weibull distributions, as these are commonly used in reliability engineering (Kapur and Pecht 2014). This section provides the distribution function, density function, and maximum likelihood estimators (if they exist) for each distribution. The specific parameterizations for the gamma, log-normal, and Weibull distributions used in this paper are shown in Appendix A.

2.1.1 Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence can measure the relative entropy, or information loss, between two distributions (Bromideh 2012; Bauckhage 2013; Bauckhage 2014). The KL divergence is defined as:

$$KL(F,G) = \int_0^\infty f(x) \ln \frac{f(x)}{g(x)} dx. \tag{2}$$

Since the Weibull distribution has support of $x \in [0,\infty)$, and the gamma and log-normal distributions have support of $x \in (0,\infty)$ we have limited the domain to $x \in [0,\infty)$. The Kullback-Leibler divergence can be thought of as the 'distance' between F and G or as the amount of information lost when using G to model F (Bromideh 2012). It is important to note that this is not a true distance, as generally speaking $KL(F||G) \neq KL(G||F)$.

2.1.2 Kolmogorov-Smirnov Statistic

The Kolmogorov-Smirnov (KS) statistic measures the maximum (vertical) difference between two cumulative distribution functions (CDFs). It is defined as

$$KS(F,G) = \sup_x |F(x) - G(x)| \tag{3}$$

for two distribution functions F and G (Kundu and Manglick 2005; Dey and Kundu 2009). They found instability in the results for KS statistic with small sets of observations using the maximum likelihood estimate. Additionally, censoring data further decreases the accuracy of correctly selecting the underlying distribution. The final significant observation related to this instability is that correctly discriminating distributions which are so similar may not provide additional functional benefit.

3 METHODS

In order to compare KS statistics and KL divergence, we sought to generate samples from known distributions, then calculate estimates of the distributions from the samples, and finally compare how KS statistics and KL divergence describe the distance between the estimated distributions and the known underlying distributions. The simulation developed to determine the probability of correct selection of one of the reference distributions corresponding to given shapes is outlined in Table 1. For each reference shape and iteration, we randomly sample from the distribution based on the required number of observations. With these samples, we estimate the best fit distribution using the maximum likelihood. The reference shape and scale is then compared to the estimated shape and scale by measuring both the KS statistic and KL divergence.

3.1 Selected Parameter Values

The menu of shapes selected for the Monte Carlo simulation were compiled from values previously explored in similar research and are listed in Table 1. The shape value used for the Weibull distribution and standard deviation for the log-normal distribution are the same as used by Dey and Kundu (2009), while the shape values for the gamma distribution are similar to those used by Kundu and Manglick (2005). A scale of 1.0 was selected for all probability distributions ($\lambda = 1, \exp(\mu) = 1$ and $l = 1$ for gamma, log-normal, and Weibull distributions, respectively).

Table 1: Evaluated shapes (gamma and Weibull) and standard deviation (log-normal).

Distribution	**Parameters**							
gamma ($\eta =$)	2.0	4.0	6.0	8.0	10.0	12.0		
log-normal ($\sigma =$)	0.5	0.6	0.7	0.8	0.9	1.0	1.2	1.4
Weibull ($k =$)	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0

3.2 Simulation Algorithm

Algorithm 1 provides the framework we used to estimate the PCS among the three distributions. This example will assume a Weibull reference distribution with a sample size of N, where θ is the known parameters from the reference distribution consisting of the shape and scale.

Algorithm 1: Probability of Correct Selection algorithm

1 $klwin \leftarrow [\,]$
2 $kswin \leftarrow [\,]$
3 **for** $m = 1$ **to** M **do**
4 $\quad X \leftarrow Weib(\theta, N)$
5 $\quad A \leftarrow [\,]$
6 $\quad B \leftarrow [\,]$
7 $\quad$ **for** $i \in \{gamma, log\text{-}normal, Weibull\}$ **do**
8 $\quad\quad \hat{\theta}_i \leftarrow MLE(i|X)$
9 $\quad\quad A.append(KL(P(\theta), Q(\hat{\theta}_i)))$
10 $\quad\quad B.append(KS(F(\theta), G(\hat{\theta}_i)))$
11 $\quad$ **end**
12 $\quad klwin.append\left(\arg\min_i\{A_i\}\right)$
13 $\quad kswin.append\left(\arg\min_i\{B_i\}\right)$
14 **end**
15 $\hat{PCS}_{KL}(Weib, k) = \sum_{j=1}^{N} \frac{[klwin[j] = Weib]}{N}$
16 $\hat{PCS}_{KS}(Weib, k) = \sum_{j=1}^{N} \frac{[kswin[j] = Weib]}{N}$

In Algorithm 1, we initiate two lists to store the distribution with the smallest KL divergence values and KS statistic, respectively (lines 1 and 2) . We then iterate through M replications (line 3) and create N random variates from the reference distribution with reference parameters θ (line 4), and create two lists to store the KL divergence values and KS statistic (lines 5 and 6). Then, for each of the three estimated distributions (line 7), we estimate the parameters for the estimated distribution i, $\hat{\theta}_i$ (line 8). The KL divergence values (2) and KS statistic (3) are calculated using the reference parameters θ and the estimated parameters $\hat{\theta}_i$ (lines 9 and 10). The distribution with the best (smallest) KL divergence value and KS statistic is recorded (lines 12 and 13). The estimated probability of correct selection for the reference distribution with a sample size of k is calculated by counting the number of instances where the *klwin* and *kswin* is equal to the reference distribution for both the KL divergence and KS statistic (lines 15 and 16). The estimated probability of correct selection shown in line 15 is equivalent to the modifying equation (1) in the following manner: Let KL_δ be the KL divergence for distributions $\delta \in \{$gamma, log-normal, and Weibull$\}$, as calculated in line 9. Without loss of generality, we can define the estimated probability of correct selection for sample size k as and distribution indicator i

$$\hat{PCS}_{KL}(\delta = i, k) \approx P(KL_1 < KL_2, KL_1 < KL_3).$$

Substituting KS for KL will provide the estimated probability of correct selection shown in line 16 for the KS statistic.

We used Python 3.9.7 and Scipy 18.1 to simulate the values. Here we explain some of the details in executing Algorithm 1. The random variates (line 4) are created using the `scipy.stats.gamma.rvs`, `.lognorm.rvs`, or `.weibull_min.rvs` functions. The MLE values are calculated using the closed form estimates for the gamma (equations (4) and (5)) and log-normal distributions (equations (6) and (7)), while the MLE for the Weibull distribution use the `scipy.optimize.minimize function`. The KL divergence values (line 9) are calculated using the results found in Bauckhage (2013), Bauckhage (2014), and

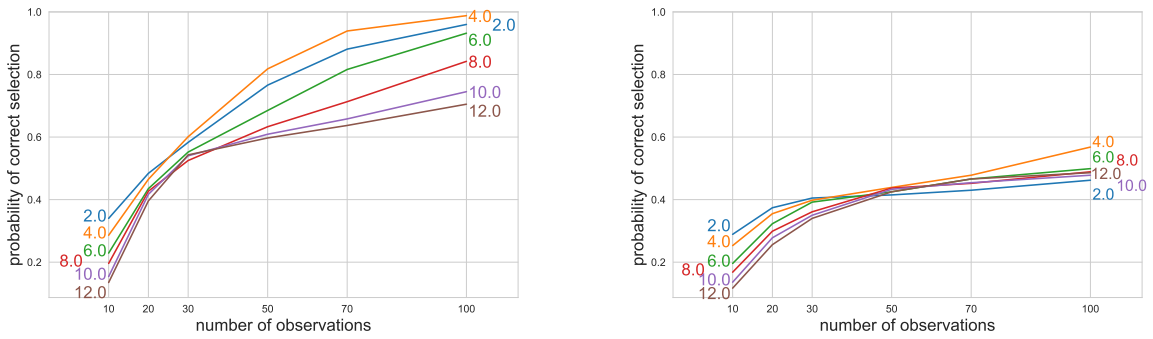
similar computations. The KS statistic (line 10) is calculated using the `scipy.stats.ks_2samp(X,Y)` function, where `X` and `Y` are created from appropriate combination of inverse cdf values (using the `scipy.stats.gamma.ppf`, `.lognorm.ppf`, or `.weibull_min.ppf` functions). Although this algorithm shows the PCS (lines 12, 13) for the reference distribution (Weibull in this example), the PCS for the other distributions are recorded and shown in Tables 2-4. All references to `scipy` can be referenced in Virtanen et al. (2020).

3.3 Simulation Results

Each simulation was completed with randomly generated samples starting with 10 observations, and increasing with each iteration to 20, 30, 50, 70, and 100 observations. Finally, 10,000 replications of each combination were performed.

3.3.1 Gamma Reference Distribution

The results for a gamma reference distribution with a scale parameter $\lambda = 1$ and shape parameters $\eta =$ 2.0, 4.0, 6.0, 8.0, 10.0, and 12.0 are shown in Figure 1 and Table 2. One can see that in general, as the sample size increases, the probability of choosing a gamma distribution with the KL divergence also increases (Figure 1a). The PCS of choosing gamma also increases when using the KS statistic albeit at a much slower rate. While the PCS of both metrics are comparable for smaller sample sizes ($n = 10, 20$), KL divergence yields a higher PCS in every case, with dramatically better results at larger sample sizes.



(a) KL divergence

(b) KS statistic

Figure 1: Probability of correct selection with gamma as reference distribution. The numbers listed on the figures represent the various shape parameter values for the gamma distribution.

Upon closer examination of Table 2, we see that the KL divergence with a smaller shape value ($\eta = 2.0$) and lower numbers of observations ($n = 10, 20$) has a higher percentage of incorrectly choosing Weibull at the smaller sample sizes compared to choosing the log-normal distribution, but has a higher percentage of incorrectly choosing the log-normal distribution with larger shape values and larger numbers of observations. We can see similar phenomena in the KS statistic tables, where the incorrect selection tends towards the log-normal distribution for larger shape values and sample sizes. A broad comparison of these results can be made to the results obtained in Dey and Kundu (2009). Although their analysis is not the same, they compare the PCS of the generalized exponential, log-normal, and Weibull distributions based on asymptotic distributions and obtain strikingly similar results.

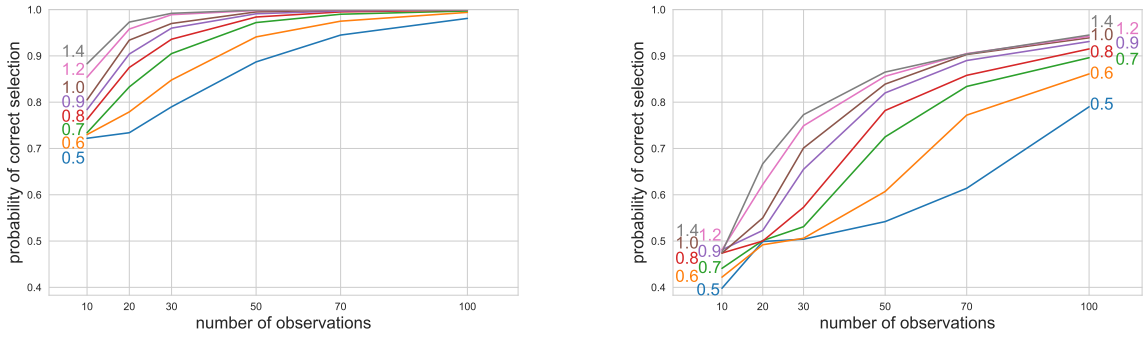
Table 2: Classification probabilities for various shape parameter values and sample sizes with gamma as reference distribution.

		KL divergence						KS statistic					
shape	dist	10	20	30	50	70	100	10	20	30	50	70	100
2	**gamma**	0.34	0.484	0.583	0.766	0.881	0.96	0.289	0.374	0.405	0.415	0.43	0.462
	lognorm	0.233	0.11	0.057	0.015	0.003	0	0.22	0.112	0.064	0.06	0.068	0.067
	weibull	0.428	0.406	0.36	0.219	0.116	0.039	0.491	0.514	0.531	0.526	0.502	0.471
4	**gamma**	0.286	0.465	0.601	0.818	0.939	0.988	0.253	0.355	0.398	0.439	0.478	0.568
	lognorm	0.389	0.34	0.28	0.134	0.05	0.011	0.327	0.285	0.24	0.168	0.115	0.055
	weibull	0.325	0.195	0.119	0.047	0.01	0.002	0.42	0.36	0.361	0.393	0.407	0.377
6	**gamma**	0.229	0.435	0.552	0.685	0.816	0.932	0.196	0.323	0.392	0.425	0.466	0.499
	lognorm	0.379	0.427	0.398	0.3	0.18	0.068	0.264	0.336	0.338	0.338	0.295	0.236
	weibull	0.392	0.138	0.05	0.015	0.003	0.001	0.54	0.341	0.27	0.237	0.239	0.264
8	**gamma**	0.196	0.427	0.525	0.633	0.713	0.842	0.168	0.299	0.361	0.437	0.452	0.489
	lognorm	0.34	0.411	0.439	0.359	0.285	0.158	0.19	0.285	0.342	0.393	0.399	0.363
	weibull	0.464	0.162	0.036	0.008	0.001	0	0.642	0.416	0.297	0.17	0.149	0.147
10	**gamma**	0.154	0.417	0.54	0.609	0.658	0.745	0.136	0.278	0.35	0.433	0.454	0.478
	lognorm	0.303	0.369	0.425	0.387	0.341	0.255	0.126	0.22	0.287	0.393	0.44	0.432
	weibull	0.544	0.214	0.035	0.005	0.001	0	0.738	0.502	0.363	0.174	0.106	0.09
12	**gamma**	0.135	0.396	0.543	0.597	0.637	0.705	0.117	0.256	0.34	0.425	0.466	0.486
	lognorm	0.258	0.337	0.416	0.399	0.362	0.295	0.082	0.172	0.25	0.363	0.444	0.455
	weibull	0.607	0.267	0.041	0.003	0.001	0	0.8	0.572	0.41	0.212	0.09	0.059

3.3.2 Log-Normal Reference Distribution

The results for a log-normal reference distribution with a scale parameter $\exp(\mu) = 1$ and various standard deviations σ are shown in Figure 2 and Table 3. Comparing the KL divergence results and the KS statistic results, one can see that the KL divergence correctly identifies the log-normal distribution at a higher probability for every σ and sample size.

In general, as the sample size increases, the probability of choosing a gamma distribution with the either the KL divergence or the KS statistic also increases, as expected (Figure 1a).



(a) KL divergence

(b) KS statistic

Figure 2: Probability of correct selection with log-normal as reference distribution. The numbers listed on the figures represent the various standard-deviation parameter values for the log-normal distribution.

Further examination of Table 3 shows that the KL divergence achieves a generally increasing probability of correct selection. Additionally, the KL divergence more accurately selects log-normal distributions at a significantly higher rate than the it did for the gamma distribution. However, with the KS statistic

for standard deviation values of ($\sigma = 1.2, 1.4$) and lower numbers of observations ($n = 10, 20$) there is a temporary spike in the percentage of correctly choosing log-normal distribution. For these two standard deviation values, there is some turbulence in reliably selecting the correct distribution for higher sample sizes. Comparing these results to Table 2, we see that the KL statistic provides a higher PCS for log-normal reference distributions for all the tested standard deviation parameters. The final note for this assessment is that when log-normal failed to be selected, the gamma distribution was incorrectly selected at a much higher rate than Weibull. Interestingly, this phenomenon appeared to reverse with standard deviation values of ($\sigma = 1.2, 1.4$).

Table 3: Classification probabilities for various standard-deviation parameter values and sample sizes with log-normal as reference distribution.

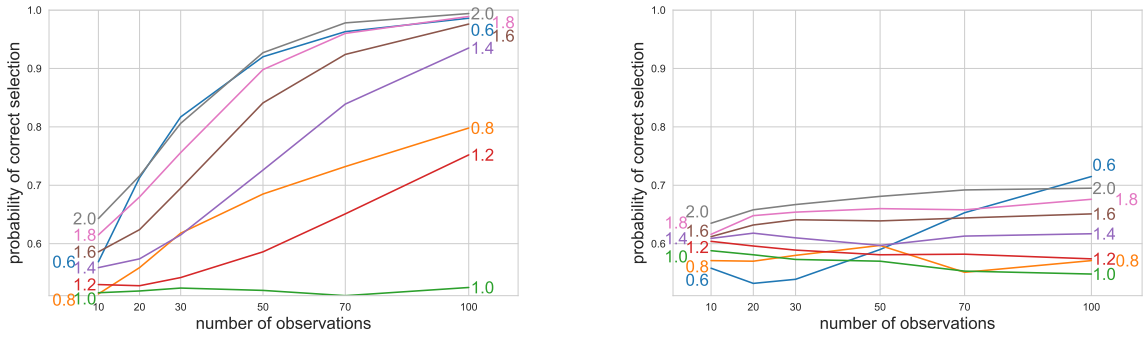
		KL divergence						KS statistic					
stdev	dist	10	20	30	50	70	100	10	20	30	50	70	100
0.5	gamma	0.264	0.265	0.21	0.113	0.055	0.019	0.281	0.398	0.409	0.394	0.332	0.174
	lognorm	0.722	0.734	0.79	0.887	0.945	0.981	0.398	0.499	0.504	0.542	0.614	0.79
	weibull	0.014	0.001	0	0	0	0	0.321	0.103	0.087	0.064	0.053	0.036
0.6	gamma	0.26	0.22	0.152	0.059	0.025	0.006	0.324	0.4	0.396	0.316	0.172	0.108
	lognorm	0.73	0.779	0.848	0.941	0.975	0.994	0.422	0.492	0.506	0.607	0.772	0.861
	weibull	0.01	0.001	0	0	0	0	0.254	0.108	0.097	0.076	0.055	0.031
0.7	gamma	0.254	0.166	0.095	0.028	0.01	0.003	0.347	0.384	0.364	0.191	0.112	0.077
	lognorm	0.734	0.833	0.905	0.972	0.99	0.997	0.441	0.501	0.531	0.725	0.834	0.896
	weibull	0.012	0.001	0	0	0	0	0.212	0.115	0.105	0.084	0.054	0.028
0.8	gamma	0.226	0.124	0.064	0.016	0.005	0.001	0.352	0.372	0.312	0.122	0.085	0.059
	lognorm	0.763	0.875	0.936	0.984	0.995	0.999	0.474	0.5	0.573	0.782	0.858	0.915
	weibull	0.011	0.001	0	0	0	0	0.174	0.127	0.115	0.096	0.057	0.025
0.9	gamma	0.204	0.095	0.04	0.009	0.003	0	0.357	0.339	0.221	0.086	0.047	0.038
	lognorm	0.784	0.904	0.96	0.991	0.997	1	0.481	0.523	0.655	0.82	0.89	0.931
	weibull	0.012	0.001	0	0	0	0	0.162	0.138	0.124	0.094	0.063	0.032
1	gamma	0.182	0.065	0.029	0.005	0.001	0	0.361	0.304	0.169	0.061	0.029	0.018
	lognorm	0.805	0.934	0.97	0.995	0.999	1	0.474	0.55	0.701	0.839	0.903	0.939
	weibull	0.014	0.001	0	0	0	0	0.166	0.147	0.131	0.1	0.068	0.043
1.2	gamma	0.124	0.037	0.01	0.002	0	0	0.34	0.215	0.106	0.035	0.013	0.004
	lognorm	0.854	0.958	0.989	0.998	1	1	0.481	0.622	0.749	0.856	0.905	0.942
	weibull	0.022	0.004	0.001	0.001	0	0	0.179	0.164	0.145	0.108	0.081	0.054
1.4	gamma	0.088	0.018	0.004	0	0	0	0.336	0.159	0.075	0.029	0.015	0.007
	lognorm	0.883	0.973	0.992	0.999	1	1	0.476	0.667	0.773	0.865	0.904	0.945
	weibull	0.029	0.009	0.003	0	0	0	0.188	0.174	0.152	0.106	0.08	0.048

3.3.3 Weibull Reference Distribution

The results shown in Figure 3 and Table 4 for the Weibull distribution are of particular interest. The multi-color lines represent results for shape parameters $\eta = 0.6$, 1.0, 1.2, 1.4, 1.6, 1.8, and 2.0. Inspecting Figure 3 shows that using the KL divergence can result in reliably selecting the correct distribution at least half of the time. As the shape k "moves further away" from 1.0, and the number of observations increases, the PCS also increases. With the shape $k = 1.0$, KL divergence consistently resulted in a correct selection for approximately half of the simulations, regardless of the number of observations. This is because with shape $k = 1.0$, the Weibull is an exponential distribution. The gamma distribution with shape parameter 1 is also exponential, so the MLE fitting can yield the same exponential distribution by starting with the gamma. Note that the algorithm "incorrectly" chooses gamma half of the time, in fact finding the correct exponential distribution. With the shape close to 1, at $k = 0.8$ and $k = 1.2$, we see a similar pattern. While KL divergence yields a PCS that increases with sample size, it appears to under-perform when compared to

other reference distributions. When the wrong distribution is selected in these cases, it is almost exclusively chooses the gamma distribution.

The KS statistic correctly selected the Weibull distribution more than half the time, but did not improve much with increased sample size. The gamma distribution is incorrectly chosen much more frequently than the log-normal, which may be expected because the gamma and Weibull are in the exponential family of distributions.



(a) KL divergence

(b) KS statistic

Figure 3: Probability of correct selection with Weibull as reference distribution. The numbers listed on the figures represent the various shape parameter values for the Weibull distribution.

Table 4: Classification probabilities for various shape parameter values and sample sizes with Weibull as reference distribution.

		KL divergence						KS statistic					
shape	dist	10	20	30	50	70	100	10	20	30	50	70	100
0.6	gamma	0.312	0.239	0.163	0.075	0.036	0.014	0.319	0.298	0.294	0.265	0.228	0.19
	lognorm	0.119	0.048	0.02	0.004	0.001	0	0.122	0.17	0.168	0.146	0.119	0.095
	weibull	0.569	0.713	0.817	0.92	0.963	0.986	0.558	0.532	0.539	0.59	0.653	0.715
0.8	gamma	0.382	0.399	0.366	0.311	0.267	0.202	0.339	0.32	0.299	0.296	0.351	0.35
	lognorm	0.104	0.041	0.016	0.004	0.001	0	0.089	0.109	0.122	0.107	0.098	0.08
	weibull	0.514	0.559	0.618	0.685	0.732	0.798	0.571	0.57	0.58	0.597	0.551	0.571
1	gamma	0.394	0.448	0.465	0.479	0.489	0.475	0.322	0.355	0.359	0.358	0.384	0.403
	lognorm	0.09	0.034	0.011	0.001	0	0	0.09	0.063	0.068	0.072	0.063	0.048
	weibull	0.516	0.519	0.524	0.52	0.511	0.525	0.588	0.581	0.573	0.57	0.553	0.548
1.2	gamma	0.376	0.446	0.449	0.412	0.349	0.248	0.288	0.366	0.364	0.376	0.378	0.396
	lognorm	0.094	0.026	0.009	0.001	0	0	0.108	0.038	0.047	0.043	0.04	0.031
	weibull	0.53	0.528	0.542	0.586	0.651	0.752	0.604	0.596	0.589	0.581	0.582	0.574
1.4	gamma	0.346	0.4	0.377	0.274	0.161	0.065	0.256	0.34	0.36	0.375	0.361	0.357
	lognorm	0.095	0.025	0.009	0	0	0	0.134	0.042	0.03	0.028	0.025	0.026
	weibull	0.559	0.574	0.615	0.726	0.839	0.935	0.609	0.618	0.61	0.597	0.613	0.617
1.6	gamma	0.315	0.35	0.298	0.158	0.076	0.024	0.239	0.319	0.334	0.34	0.335	0.33
	lognorm	0.1	0.025	0.006	0	0	0	0.15	0.048	0.025	0.021	0.022	0.02
	weibull	0.586	0.624	0.695	0.841	0.924	0.976	0.612	0.632	0.641	0.639	0.644	0.651
1.8	gamma	0.281	0.295	0.238	0.102	0.04	0.011	0.214	0.293	0.317	0.321	0.325	0.308
	lognorm	0.105	0.025	0.006	0	0	0	0.17	0.059	0.029	0.019	0.016	0.016
	weibull	0.615	0.68	0.756	0.898	0.96	0.989	0.616	0.648	0.654	0.66	0.658	0.676
2	gamma	0.241	0.256	0.187	0.073	0.022	0.006	0.179	0.272	0.302	0.306	0.297	0.291
	lognorm	0.116	0.028	0.007	0.001	0	0	0.186	0.07	0.031	0.014	0.01	0.014
	weibull	0.643	0.716	0.806	0.927	0.978	0.994	0.635	0.658	0.667	0.681	0.692	0.695

An interesting observation from Table 4 for the Weibull distribution is that for $k = 1.0$ the KS statistic incorrectly indicates the gamma distribution at a rate increasing from 64 percent at 10 observations increasing to around 80 percent for higher numbers of observations. The data in Table 4 shows that overall, when an incorrect selection was made, the selected distribution was commonly gamma. Considering that both Weibull and gamma simplify to exponential distributions when their shape parameter is one, the difficulty in correctly choosing the Weibull reference distribution when $k = 1.0$ is reasonable.

4 DISCUSSION

The log-normal distribution overall achieved the highest rate of successful correct selection, with PCS increasing as the number of observations increased. The first significant observation resulting from this research is that the KL divergence is almost universally more likely to result in correctly selecting the underlying distribution from a randomly sampled set of observations. The low PCS obtained from the KS statistic are not surprising because it relies on evaluating the greatest distance between the estimated and reference distributions. This observation further reinforces past research by Dey and Kundu (2009) showing that the KS statistics from reference distributions to their estimated distributions tend to be very close, resulting in difficulty in choosing the correct distribution. Additionally, increasing sample size does not necessarily increase the probability of correct selection for certain shape values. When the reference distribution is Weibull, both statistics become more accurate the farther the shape is from 1.0.

The most important result of this work is the finding that under all explored scenarios the KL divergence outperformed the KS statistic. As our research objective was to determine circumstances under which one measure outperformed the other to aid practitioners in choosing the best distribution for reliability applications, these results are very powerful. Further research is warranted into applications for higher-dimentional problems and other distributions not tested here, as well as other methods of distribution fitting (such as the likelihood function method).

A APPENDIX

A.1 Gamma Distribution

The gamma probability density function (PDF) is given by

$$f(x|\eta,\lambda) = \frac{1}{\Gamma(\eta)\lambda^{\eta}} x^{\eta-1} \exp\left[-\left(\frac{x}{\lambda}\right)\right],$$

and cumulative distribution function (CDF)

$$F(x) = \frac{1}{\Gamma(\eta)} \gamma\left(\eta, \frac{x}{\lambda}\right)$$

where $\eta > 0$ is the shape parameter, $\lambda > 0$ is the scale parameter, and $\gamma\left(\eta, \frac{x}{\lambda}\right)$ is the lower incomplete gamma function.

The maximum likelihood estimators for the gamma distribution can be estimated by (Ye and Chen 2017)

$$\begin{aligned}
\psi &= N \sum_{i=1}^{N} x_i \ln(x_i) - \sum_{i=1}^{N} \ln(x_i) \sum_{i=1}^{N} x_i \\
\hat{\eta} &= \frac{N \sum_{i=1}^{N} x_i}{\psi} \\
\hat{\lambda} &= \frac{\psi}{N^2},
\end{aligned}$$

where N is the number of observations or sample size. Unbiased estimators are calculated using (Louzada, Ramos, and Ramos 2019)

$$\tilde{\eta} = \hat{\eta} - \frac{1}{N}\left(3\hat{\eta} - \frac{2}{3}\left(\frac{\hat{\eta}}{1+\hat{\eta}}\right) - \frac{4}{5}\frac{\hat{\eta}}{(1+\hat{\eta})^2}\right) \tag{4}$$

$$\tilde{\lambda} = \frac{N}{N-1}\hat{\lambda}. \tag{5}$$

A.2 Log-Normal Distribution

The log-normal PDF is given by

$$f(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left[\left(-\frac{1}{2}\right)\left(\frac{\ln x - \mu}{\sigma}\right)^2\right],$$

and CDF

$$F(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\ln(x)-\mu}{\sigma\sqrt{2}}\right)\right],$$

where $\mu \in (-\infty,\infty), \sigma > 0$ are the mean and standard deviation of the natural logarithm of the variable x, respectively, and erf is the error function. We will use the commonly used parameterization of $scale = \exp(\mu)$.

The unbiased maximum likelihood estimators can be estimated using

$$\hat{\mu} = \frac{\sum_n \ln x_n}{n}, \tag{6}$$

$$\hat{\sigma}^2 = \frac{\sum_n (\ln x_n - \hat{\mu})^2}{n}. \tag{7}$$

A.3 Weibull Distribution

The Weibull probability density function (PDF) is given by

$$f(x|k,l) = \frac{k}{l}\left(\frac{x}{l}\right)^{k-1}\exp\left[-\left(\frac{x}{l}\right)^k\right],$$

and cumulative distribution function (CDF)

$$F(x) = 1 - \exp\left[\left(\frac{x}{l}\right)^k\right],$$

where $k > 0$, $l > 0$ are the shape and scale parameters, respectively. Unlike the gamma and log-normal distributions, the MLE for the Weibull distribution does not have an explicit expression and must be estimated numerically.

REFERENCES

Atedero, R., L. Lee, and V. Karbhari. 2004. "Consideration of Material Variability in Reliability Analysis of FRP Strengthened Bridge Decks". *Composite Structures* 70:430–443.

Basavalingappa, A., J. M. Passage, M. Y. Shen, and J. Lloyd. 2017. "Electromigration: Lognormal Versus Weibull Distribution". In *2017 IEEE International Integrated Reliability Workshop*. South Lake Tahoe, CA, Oct 8th-12th, Institute of Electrical and Electronics Engineers, Inc., 1-4.

Basu, B., D. Tiwari, D. Kundu, and R. Presad. 2009. "Is Weibull Distribution the Most Appropriate Statistical Strength Distribution For Brittle Materials?". *Ceramics International* 35(1):237–246.

Bauckhage, C. 2013. "Computing the Kullback-Leibler Divergence Between Two Weibull Distributions". *arXiv preprint arXiv:1310.3713*.

Bauckhage, C. 2014. "Computing the Kullback-Leibler Divergence Between Two Generalized Gamma Distributions". *arXiv preprint arXiv:1401.6853*.

Bromideh, A. A. 2012. "Discriminating Between Weibull and Log-Normal Distributions Based on Kullback-Leibler Divergence". *Ekonometri ve İstatistik e-Dergisi* 0(16):44–54.

Das, R. N., and J.-S. Park. 2012. "Discrepancy in Regression Estimates Between Log-Normal and Gamma: Some Case Studies". *Journal of Applied Statistics* 39(1):97–111.

Dey, A. K., and D. Kundu. 2009. "Discriminating Among the Log-Normal, Weibull, and Generalized Exponential Distributions". *IEEE Transactions on Reliability* 58(3):416–424.

Kapur, K. C., and M. Pecht. 2014. *Reliability Engineering*, Volume 86. Hoboken, New Jersey: John Wiley & Sons.

Kundu, D., and A. Manglick. 2004. "Discriminating Between the Weibull and Log-Normal Distributions". *Naval Research Logistics* 51(6):893–905.

Kundu, D., and A. Manglick. 2005. "Discriminating Between The Log-normal and Gamma Distributions". *Journal of the Applied Statistical Sciences* 14:175–187.

Louzada, F., P. L. Ramos, and E. Ramos. 2019. "A Note on Bias of Closed-Form Estimators for the Gamma Distribution Derived from Likelihood Equations". *The American Statistician* 73(2):195–199.

Peng, Y., C.-H. Chen, M. C. Fu, and J.-Q. Hu. 2015. "Non-Motonicity of Probability of Correct Selection". In *Preceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 3678–3689. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". *Nature Methods* 17:261–272.

Ye, Z.-S., and N. Chen. 2017. "Closed-Form Estimators for the Gamma Distribution Derived from Likelihood Equations". *The American Statistician* 71(2):177–181.

AUTHOR BIOGRAPHIES

MARIO ANDRIULLI is an assistant professor at the United States Military Academy. His research interests include observability, controllability, and sustainable energy systems. His email address is mario.c.andriulli.mil@army.mil.

JAMES K. STARLING is an assistant professor at the United States Military Academy. His research interests include obsolescence management, optimization, simulation, and military applications. His email address is james.starling@westpoint.edu.

BLAKE SCHWARTZ is an assistant professor at the United States Military Academy, and the director of USMA's Center for Data Analysis and Statistics. His research interests include stochastic network modeling, machine learning for military applications, and robust optimization. His e-mail address is blake.e.schwartz.mil@army.mil.