## Introduction

It is important to confirm the validity and reliability of the results returned by the tool. Here, we devise experiments with a large set of instances comprised of different probability distributions and we also compare the results with benchmarks.

## Experiments

This section describes the tests performed to analyze the performance of the proposed method and compare it with a Johnson's method and Burr Type XII Distribution. The objective is to compare the performance of CP1 and CP2. Additionally this topic also gives an illustrative case study from electronic industry.

### *4.1 – Validation with the benchmark*

In order to test the methods, 5 instances with populations of 100000 values were created with the following features:

- Population 1: Normal distribution, with $\mu = 100.12\ and\ \sigma = 19.74$
- Population 2: Lognormal distribution, with $\mu = 100.12\ and\ \sigma = 20.12$
- Population 3: Lognormal distribution, with $\mu = 100.12\ and\ \sigma = 39.89$
- Population 4: Gamma distribution, with $\mu = 100.02\ and\ \sigma = 14.16$
- Population 5: Exponential distribution, with $\mu = 100.30$.
- Population 6: xxxx, with $\mu = 100.12\ and\ \sigma = 20.12$
- Population 7: xxx distribution, with $\mu = 100.12\ and\ \sigma = 39.89$
- Population 8: xxx distribution, with $\mu = 100.02\ and\ \sigma = 14.16$
- Population 9: xxxx distribution, with $\mu = 100.30$.

These populations were created through the following Matlab functions, respectively: *randn*, *lognrnd*, *lognrnd*, *gamrnd* and *exprnd*.

The accuracy of the calculation of the probability $P(X \leq x)$ is also related to the distance from $x$ to the mean, therefore each population is evaluated in 13 points: from the point $\mu - 3\sigma$ to the point $\mu + 3\sigma$ with increment of $0.5\sigma$. It is used 5 different sample sizes ($n$): 10, 20, 30, 50 and 100. Because we know the population, it is possible to compute the error for the probability calculations of the methods (Burr, Johnson, UPC). For each method, we perform 29250 calculations (9 populations, 5 sample sizes, 13 values for $x$, 50 replicas).

The figures 5 and 6 give the average errors for the PC1 and PC2 methods, respectively, for the all 5 populations together. The vertical axis gives the absolute percentage error, the horizontal axis gives the sample sizes, and for each sample size the average errors for the 50 runs performed for each $x$ value is presented. The values $x$ are presented here as the distance from the mean in terms of standard deviation, it means $x = (x' - \mu)/\sigma$. So, calling $x'$ the original value from the data set, $x$ is the transformed value of $x'$ and it goes from $-3\ to\ 3$.

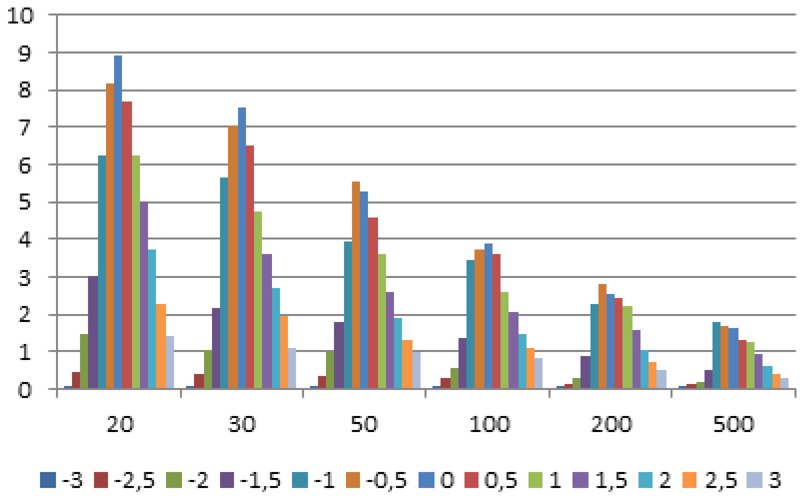

**FIGURE 5: Mean error for the proposed method (PC1)**

In the figure 5 it is possible to see how the errors get smaller as the sample size increases. For $n = 20$, the mean error for $x = 0$ is about 9%, and for $n = 500$,at the same point, the mean error for the same point is about 1.8%. Another observation from the figure 5 is that the error gets smaller close to the extremal points $x = -3\ and\ x = 3$, for these 2 points the errors were very small for all sample sizes.

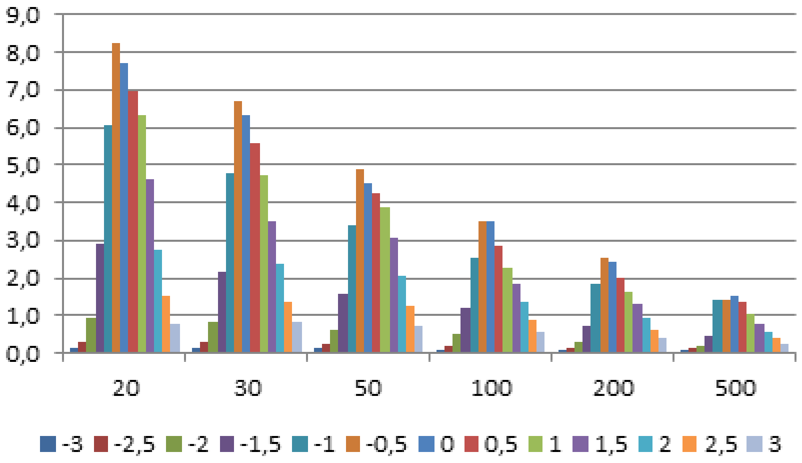

**FIGURE 6: Mean error for the benchmark (PC2)**

The figure 6 presents for $n = 20$ evaluated at $x = 0$ a mean error of 8.1%, smaller than the presented by figure 5 evaluated at the same point. From figures 5 and 6 it is possible to see that the behavior of the error over the sample sizes is similar. By these figures, it is not possible to see any significant difference between the 2 methods.

A test for equal variance was performed and the 95% Bonferroni Confidence Intervals for Standard Deviations is presented in the table 4. The null hypothesis assumes that all variances are equal and the significance level is $\alpha = 0.05$. [consider to do same test of the Juergen paper]. [Here I just wanted to show it is not possible to say one is better than other. Use tests from thesis paper + summary of the results for the error: mean, , 95th. [not max to not max user afraid, plot of the intervals]

**TABLE 4: Variance Test**

| Method | N | StDev | CI |
|---|---|---|---|
| PC1 | 19500 | 3.41 | (3.27; 3.55) |
| PC2 | 19500 | 3.10 | (2.99; 3.22) |

The table 4 shows that the intervals do not overlap indicating that the corresponding standard deviations are significantly different. This is confirmed by the results of two statistic tests, Multiple Comparisons and Levene, returning both p-value equal to zero.

To compare the means, the ANOVA was performed for different variances as confirmed previously. The interval plots are presented in the table 5. The null hypothesis assumes that the means are equal and the significance level is $\alpha = 0.05$.

**TABLE 5: ANOVA**

| Method | N | Mean | StDev | 95% CI |
|---|---|---|---|---|
| PC1 | 19500 | 2.38 | 3.41 | (2.32; 2.44) |
| PC2 | 19500 | 2.12 | 3.10 | (2.06; 2.17) |

The table 5 shows that the intervals do not overlap indicating that the mean is different. This is confirmed by Welch's Test that returns p-value equal to zero. Table 5 shows that the mean and standard deviation for the errors while comparing both methods present a small difference.

The figure 7 exhibits the interval plots comparing the mean error of PC1 and PC2 for the 5 instances, separately.

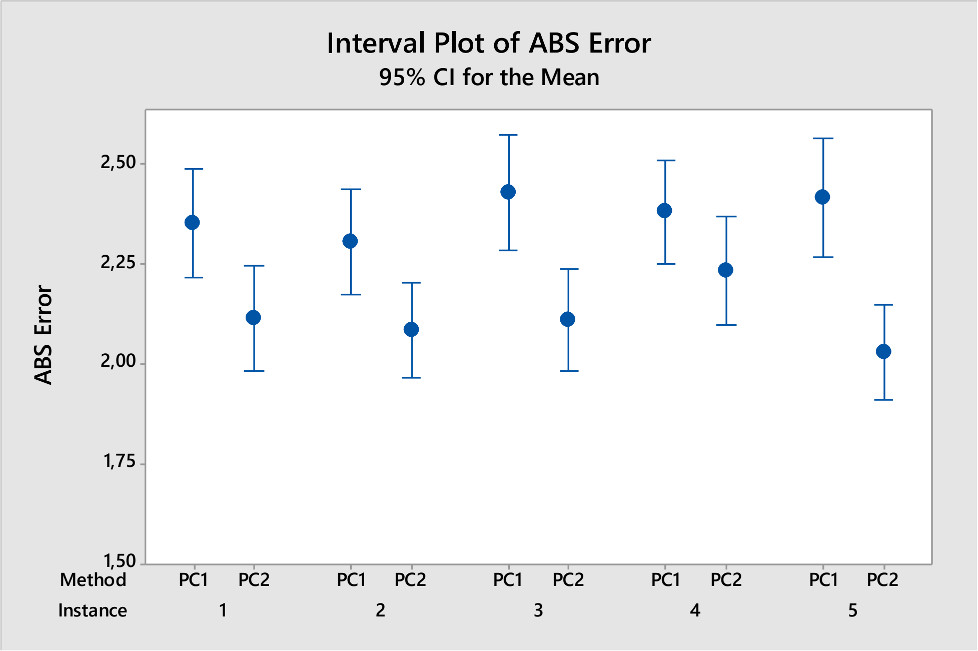

**FIGURE 7: Interval plot**

The figure 7 shows there is overlap between the 2 methods for a majority of the instances. It is not seen any big difference between the 2 methods. [maybe not necessary for UPC website]

Let's show more details regarding the population 5 (Exponential distribution) due to the fact this distributions presents a higher variation and asymmetry while compared with the others. The histogram of this population is presented in the figure 8.

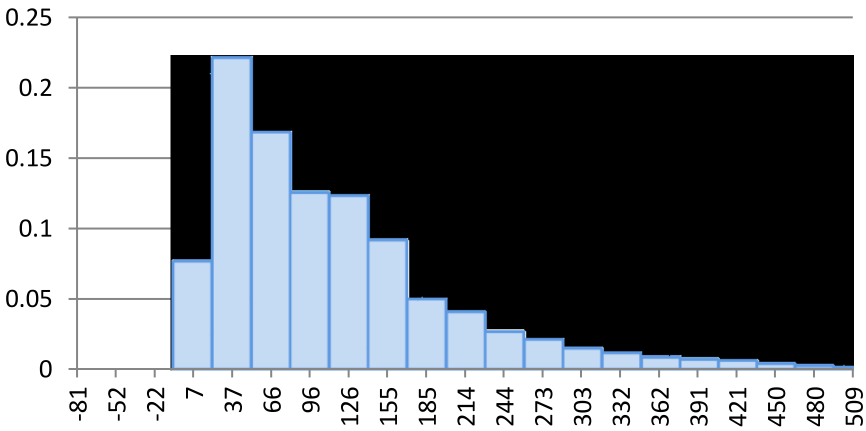

**FIGURE 8: Histogram for the population 5 (Exponential distribution)**

The figures 9 and 10 present the mean error for the population 5, with sample size of 50 and 100 respectively.

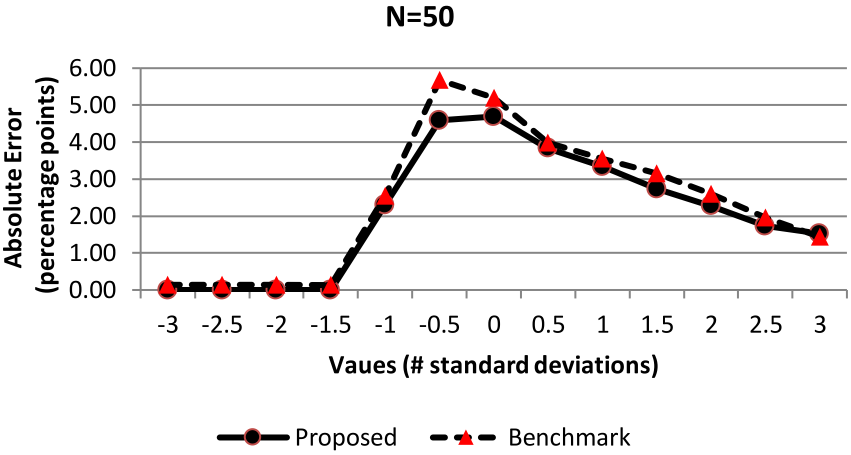

**FIGURE 9: Mean error for the Exponential Distribution with sample size 50**

The figure 9 shows the error is close to zero for the both method (proposed and benchmark) in the region from -3 to -1.5 standard deviations. It happens because in this region the values are in the range [-175.0,-39.9], that is the values are negative. After that, the error start growing and it achieves the maximum value approximately in region [-0.5, 0.0], then it starts to get smaller. For the 13 points, the proposed method presented smaller mean errors practically in all the 13 points, especially in the area with the biggest error, for the point -0.5, the mean error for the proposed method was 4.59 and for the benchmark was 5.68.

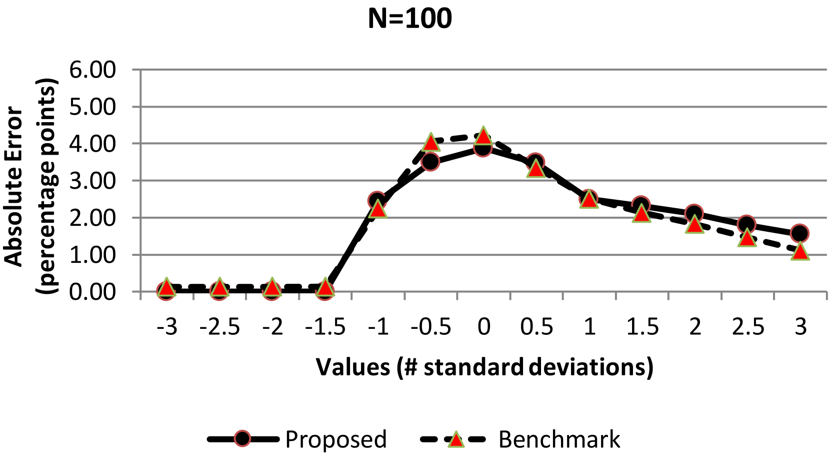

**FIGURE 10: Mean error for the Exponential Distribution with sample size 100**

Similar behavior is observed in the figure 10. Left side with errors close to zero, middle area with the biggest errors, and the right side with descending errors. One explanation for such behavior is that from the left side to the middle the exponential distribution grows quickly, so it is more difficult to capture this transition, resulting in bigger error. The exponential distribution presents a long tale to the right side; it means the values change smoothly in this zone, what is better captured by the regressions. Considering the results presented in the pictures 5,6,9 and 10, it is observed that the absolute error in terms of percentage points is descending in the right and left side of the distribution, therefore it is seen no reason to have a different behavior in areas without the range [-3, +3] standard deviations. Another point here is that once the cumulative function is built, it has a defined shape, and it is one reason the method is powerful. Naturally, the results presented here are valid for populations with similar features to the five populations used in the experiments. Still in the picture 10, for sample size 100, the results between the 2 methods were closer. The proposed method presented smaller errors in the middle zone and bigger error in the right side.

All the 39000 calculations performed in the experiments were also important to determine empirically the estimative of the confidence interval given by the proposed method. The table 6 shows the proportion of results that fell within an error interval of ± 5 percentage points. The rows give the behavior of the errors along different sample sizes and the columns along different evaluation points (as number of standard deviation from the true mean of the population). [maybe not necessary for UPC website]

**TABLE 6: Confidence Level for ± 5% CI**

| St.Dev. | n=20 | n=30 | n=50 | n=70 | n=100 | n=200 | n=500 |
|---|---|---|---|---|---|---|---|
| -3 | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| -2,5 | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| -2 | 99% | 100% | 100% | 100% | 100% | 100% | 100% |
| -1,5 | 87% | 94% | 97% | 97% | 99% | 99% | 100% |
| -1 | 67% | 76% | 85% | 84% | 90% | 94% | 98% |
| -0,5 | 63% | 72% | 80% | 88% | 88% | 92% | 96% |
| 0 | 53% | 65% | 77% | 85% | 90% | 94% | 98% |
| 0,5 | 55% | 68% | 71% | 78% | 83% | 86% | 89% |
| 1 | 69% | 76% | 84% | 88% | 90% | 93% | 97% |
| 1,5 | 90% | 94% | 98% | 98% | 99% | 100% | 100% |
| 2 | 97% | 99% | 99% | 100% | 100% | 100% | 100% |
| 2,5 | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| 3 | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

In the table 6, it is clear to notice that the higher the sample size the higher the confidence level. And also, the close to ±3 standard deviations, the higher the confidence level.

### *4.2 – Application of the proposed method for real field data* [maybe not necessary for UPC website, examples in another file]

In order to illustrate the application of the method, real field data from the electronic industry was chosen as a case study. Data from the manufacturing plant is gathered and analyzed. The company has an assembly line of one specific model of sensor used in refrigerators. That is a new model with no historic data. The sensor should be activated when the temperature is ≥ 80.4 degree Celsius (°C). An analyst collected a sample of 20 units and the manager wants to know what is the probability of taking a sensor that will be activated without the specification range; it means $P(X \leq 80.4\ )$. The analyst has no idea of the shape of the distribution and no statistical knowledge to go deeper into this analysis.

The machine is able to reject automatically the sensors activated without the specification. It is important to estimate the yield of this model because it defines the expected level of rework the operation will have to do, affecting the cost and the planning of the operation. The data with the values of the sample collected by the analyst is in the table 7 and its histogram in the figure 11.

**TABLE 7: Sample** (degree Celsius °C)

| | | | |
|---|---|---|---|
| 82.00 | 96.48 | 84.51 | 119.75 |
| 112.69 | 95.12 | 115.74 | 107.86 |
| 101.35 | 82.05 | 128.18 | 103.89 |
| 96.26 | 84.15 | 89.32 | 105.60 |
| 105.83 | 80.94 | 138.56 | 101.02 |

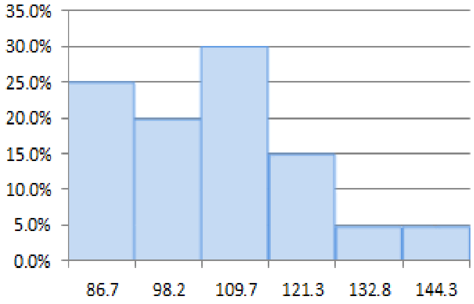

**FIGURE 11: Histogram**

Considering all samples had values greater than the specified value, a very basic analysis indicates that $P(X < 80.4) = 0/20 = 0\%$. The histogram shows a significant skewness to the left, indicating the probability distribution might not be normal. The table 8 gives the results using the proposed method and the benchmark. During 1 month the analyst counted the number of rejected and approved sensors. After this time, 1534 units were produced, 339 rejected, so the actual yield was 22.1%.

**TABLE 8: Calculation**

| Method | Calculated | Error |
|---|---|---|
| Direct | 0% | 22.10% |
| PC1 | 18.7% | 3.40% |
| Benchmark | 16.3% | 5.80% |

The table 8 shows the probability calculated and the errors based on the actual rejection. Naturally, the yield during the month depends on others variables such as raw-material, equipment maintenance, setup of the machine by the user and others, but it is a reference to analyze how accurate was the probability calculation. Another point is the small value of the sample, just 20. For this sample size, and considering the point $x = 80.5$ is about $-1$ standard deviation from the mean, the table 6 gives 68% confidence level for $\pm 5\%$ $CI$. In this context, it is plausible to say that the results are coherent.

## 5 – Concluding Remarks and Further Research

This research proposes a practical method to calculate probabilities. The method was validated with a benchmark, tested with different probability distributions, and finally it was applied to a real case.

The results presented a small difference between the proposed method and the benchmark for the accuracy of the probability calculations. The method has showed to have enough flexibility to deal with different density functions and the empirical confidence interval table gives important and positive information about the quality of the calculations. Even for the Exponential distribution with its peculiar features in term of variation and asymmetry, the results were competitive.

The application of the method in real case was able to illustrate how simple and useful the method can be on the field.

The method is easy to be implemented and it does not require statistical knowledge from the user. It is possible not only to evaluate the probability at any point, but also have an estimative of the quality of the result through the confidence interval table. Basically, if the user wants to increase the confidence level of the results, it is necessary to increase the sample size.

Considering the main goal of this paper was to develop a simple and effective method to calculate probabilities, the paper achieved its objective.

## 8 – References

Abbasi, B. ; Hosseinifard, S.Z. ; Coit, D.W. A neural network applied to estimate Burr XII distribution parameters. Reliability Engineering and System Safety, June 2010, Vol.95(6), pp.647-654

Alexopoulos, C., D. Goldsman, J. Fontanesi, D. Kopald, and J. R. Wilson. (2008). Modeling patient arrivals in community clinics. Omega 36: 33-43.

Burr, Irving W. (1941). "Cumulative frequency functions." The Annals of Mathematical Statistics, Vol. 13, Number 2. pp. 215–232.

Farnum, N. R. (1996). Using Johnson curves to describe non-normal process data. Quality Engineering 9(2): 329-336.

Flynn, M. R. (2007). Analysis of exposure–biomarker relationships with the Johnson SBB distribution. Ann. Occup. Hyg. 51(6): 533–541.

Freedman, David; Diaconis, P. (1981). "On the histogram as a density estimator: L2 theory". Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 57 (4): 453–476.

George, F. & Ramachandran, K. M. (2011) Estimation of Parameters of Johnson's System of Distributions. Journal of Modern Applied Statistical Methods. Vol. 10, No. 2, 494-504

Hahn J. Gerald and Shapiro S. Samuel (1967). Statistical models in Engineering, John Wiley and Sons.

Hill, I. D. (1976). Algorithm AS 100: Normal-Johnson and Johnson-Normal Transformations. Journal of the Royal Statistical Society. Series C (Applied Statistics) 25(2): 190-192.

Hill, I. D., R. Hill, and R. L. Holder. (1976). Algorithm AS 99: Fitting Johnson curves by moments. Journal of the Royal Statistical Society. Series C (Applied Statistics) 25(2): 180-189.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. Biometrika 36: 180-189.

Jones, D. L. (2014). The Johnson Curve Toolbox for Matlab: analysis of non-normal data using the Johnson system of distributions. College of Marine Science, University of South Florida.

Matthews, J. L., E. K. Lada, L. M. Weiland, R. C. Smith, and D. J. Leo. (2006). Monte Carlo simulation of a solvated ionic polymer with cluster morphology. Smart Mater. Struct. 15: 187–199.

Nadarajah, Saralees. 2012. On the characteristic function for Burr distributions. Statistics., Vol. 46 Issue 3, p419-428. 10

Paranaíba, Patrcia F. ; Ortega, Edwin M.M. ; Pescim, Rodrigo R. ; Cordeiro, Gauss M. 2011. The beta Burr XII distribution with application to lifetime data. Computational Statistics and Data Analysis, 1 February 2011, Vol.55(2), pp.1118-1136.

Kovarick M. & Sarga L. (2014). Process Capability Indices for Non-Normal Data. WSEAS TRANSACTIONS on BUSINESS and ECONOMICS. Volume 11.

Simonato, J. G. (2011). The performance of Johnson distributions for value at risk and expected shortfall computation. Journal of Derivatives 19: 7-24.

Tadikamalla, Pandu R (1980). A Look at the Burr and Related Distributions International Statistical Review, Vol. 48, No. 3 (Dec., 1980), pp. 337-344

van den Oord, E. J. C. G. (2005). Estimating Johnson curve population distributions in MULTILOG. Applied Psychological Measurement 29(1): 45–64.

Venables W. N. and Ripley, B. D. (2002) "Density Estimation", Modern Applied Statistics with S (2002), Springer, 4th edition

Wold, S. (1974). Spline functions in data analysis. Technometrics 16:1-11