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 Universal Probability Calculator (UPC) when compared with Johnson's distribution and Burr Type XII distribution.

In order to test the methods, 9 instances with populations of 20000 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: Weibull, with $\mu = 100.10 \; and \; \sigma = 20.06$
- Population 7: Weibull distribution, with $\mu = 100.53 \; and \; \sigma = 49.46$
- Population 8: Logistic distribution, with $\mu = 99.78 \; and \; \sigma = 20.32$
- Population 9: Logistic distribution, with $\mu = 100.01 \; and \; \sigma = 58.84$

These populations were created through the following Matlab functions, *randn*, *lognrnd*, *gamrnd*, *exprnd*, wblrnd, makedist('Logistic'). We picked distributions that are more common to be found in the real world.

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 0.5σ. It is used 3 different sample sizes (n): 20, 30 and 50. Because we know the population, it is possible to compute the error of the probability calculations for the 3 methods (Burr, Johnson, UPC). For each method, we perform 7020 calculations (9 populations, 3 sample sizes, 13 values for x, 20 replicas).

Figures 1 shows histograms of the error for the probability calculation, for the 3 methods, where $Error = Actual\ Value - Calculated\ value$, using the same graph scale.

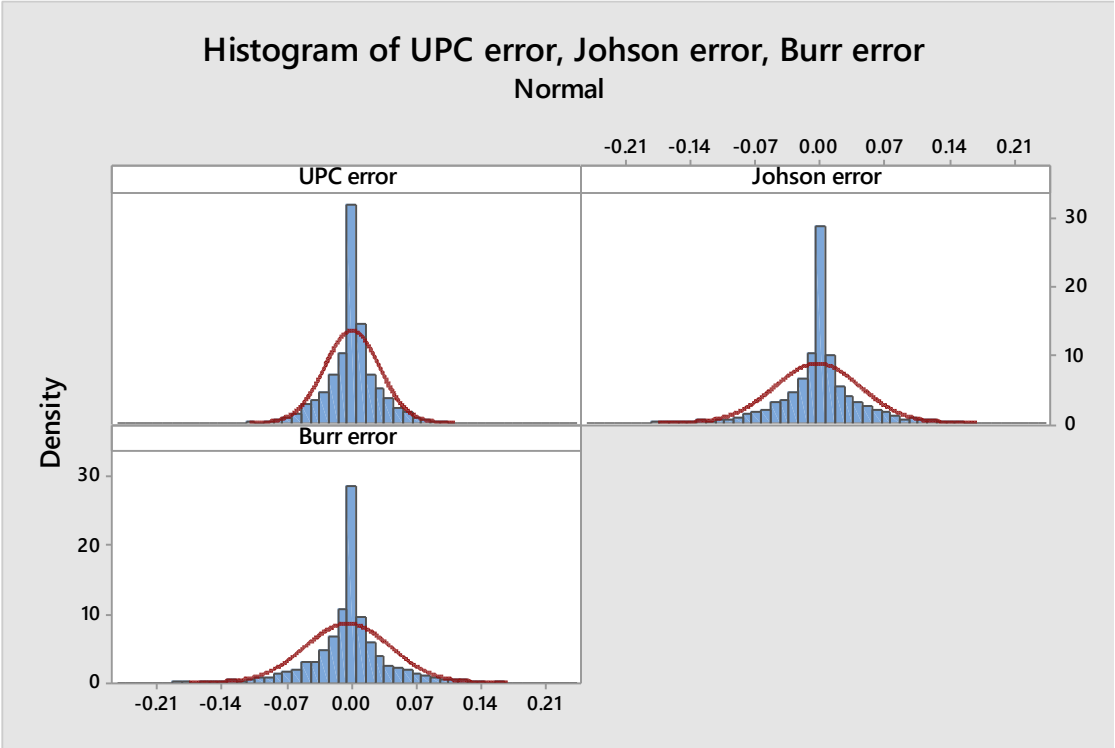

Figure 1: histograms of the error

In figure 1, the red line in each histogram represents a normal distribution with the same mean and standard deviation of the computed error, it helps to analyze how different the distribution is from a normal distribution. We see all histograms seem to be symmetric and centered close to the point zero indicating low skewness. We also see that the histograms have higher frequencies in the middle, indicating high kurtosis. Another observation is that the frequency close to the point zero is higher in UPC than the others, indicating that UPC presented more calculations with error zero than the other methods. Finally, it is seen that the UPC histogram is narrower than the others indicating a smaller variance.

We now focus on the absolute error to analyze the methods. Figure 2 has the same design of Figure 1, now measuring the absolute errors.

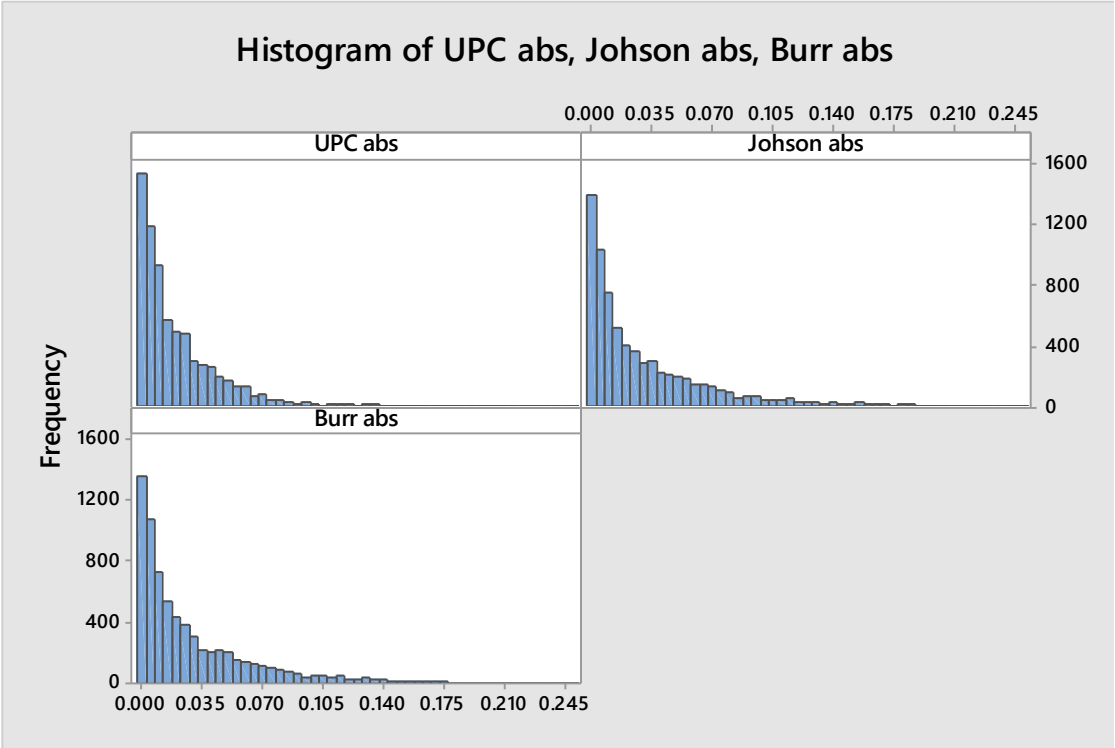

Figure 2: histograms of the absolute error

In Figure 2, we see that most of the errors are close to zero. The UPC errors are little bit more concentrated to left and its tale is shorter than the others. This indicates a smaller average and variance, which leads to Table 1. In this table we have the mean of the error, the max error among all 7020 calculations and the 95th percentile. It is seen that all metrics for UPC outperforms the others, and similar results for Johnson and Burr.

Table 1: absolute errors

Method	Mean Error	Max Error	95th Error
UPC	1.96%	21.26%	6.38%
Johnson	2.97%	24.34%	10.66%
Burr	2.95%	24.79%	10.70%

We complement this analysis checking if there is statistical difference for the mean of the hypervolume among the three methods. We initially considered to perform a pairwise t-test but because the normality test rejected the hypothesis of normal distribution for the difference of the pairs, we apply Wilcoxon signed rank test (which focus on the medians), with 95% confidence level (α=0.05), testing for no difference in the null hypothesis. The results for the statistical analysis are showed in Table 2 and Figure 3.

Table 2: Wilcoxon Signed Rank Conf. Interval

Scenario	Estimated Median	Achieved Confidence	Confidence Interval Lower	Upper
UPC-Johson	-0.00497	95%	-0.00548	-0.00449
UPC-Burr	-0.00462	95%	-0.00514	-0.00413
Johson-Burr	0.00009	95%	0.00001	0.00019

In Table 2, we see the Estimated Median of the difference of the values from the 2 methods of the respective pair. Note that if the difference is negative, the first method of the pair presents smaller values for the error than the second method. For the 3 pairs, the confidence intervals of the differences do not have zero, therefore we can reject the null hypothesis of no difference and assume there is statistical difference for the medians of all pairs. We see this difference is higher for UPC-Johson and UPC-Burr, and smaller for Johson-Burr.

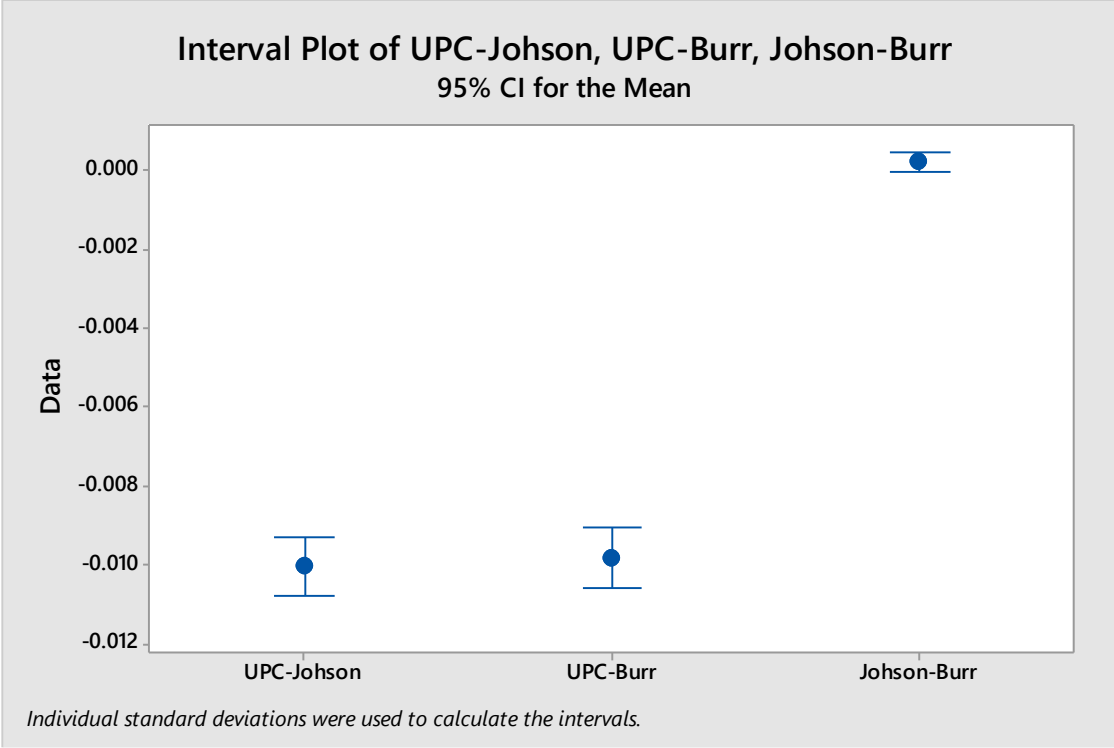

Figure 3: interval plot for the means

In Figure 3 we have the interval plot for the means. It is very clear the difference in the mean when comparing UPC with Johson and Burr, reinforcing that the UPC presented smaller errors than Burr and Johson.

Concluding Remarks

We performed experiments to compare the performance of the Universal Probability Calculator (UPC) with two well-known benchmarks: Johnson's distribution and Burr Type XII distribution. We designed an experiment using a set of different probability distributions, computing values close and far from the average, and with replicas to improve the validity and reliability of the results.

By the results, we found that UPC presented significant smaller errors than both benchmarks, while the difference between the benchmarks was very small.