# Work-in-Progress: Why Statistical Power matters for Probabilistic Real-Time

Federico Reghenzani
DEIB - Politecnico di Milano
Milan, Italy
federico.reghenzani@polimi.it

Luca Santinelli
DTIS - Onera
Toulouse, France
luca.santinelli@onera.fr

William Fornaciari
DEIB - Politecnico di Milano
Milan, Italy
william.fornaciari@polimi.it

## ABSTRACT

The probabilistic approaches for real-time systems are based on the estimation of the probabilistic-WCET distribution. Such estimation is naturally subject to errors, caused by both systematic and estimation uncertainties. To solve this problem, statistical tests are applied on the resulting distribution to check whether such errors affect or not the output validity. In this paper, we show that the reliability of these tests depends on the statistical power that must be estimated in order to select the proper sample size. This a priori analysis is required to obtain a reliable result of the probabilistic-WCET.

**ACM Reference Format:**
Federico Reghenzani, Luca Santinelli, and William Fornaciari. 2019. Work-in-Progress: Why Statistical Power matters for Probabilistic Real-Time. In *2019 International Conference on Embedded Software (EMSOFT'19), October 13–18, 2019, New York, NY, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3349568.3351555

## 1 INTRODUCTION

Probabilistic real-time computing has been proposed [1] to overcome the issues of traditional Worst-Case Execution Time (WCET) analysis when modern platforms, such as multi-core processors, are considered. Traditional analyses rely on precise timing models of the instructions and the detailed knowledge of the workload control-flow graph. However, such analyses either require an unfeasible amount of computational power or produce too pessimistic values for the WCET. Probabilistic approaches, in particular the Measurement-Based Probabilistic Timing Analysis (MBPTA), are techniques that infer the statistical distribution of the WCET from the direct measurements of the task execution time [5]. Such resulting distribution is called *probabilistic-WCET* (pWCET). The complement of its Cumulative Distribution Function (CDF) represents the probability of observing execution times larger than a certain value $x$: $G(x) = 1 - F(x) = P(X > x)$. Provided that this distribution is correctly estimated, the resulting probability can be added to the fault-tree analysis of a safety-critical system. However, correctly estimating this distribution presents several challenges. The most critical problem is related to representativity of the inputs provided to the system when the time measurements are collected. In this paper, we do not tackle this problem, but we focus on the process used to estimate the pWCET and, in particular, on the statistical test procedures. This paper shows how confidence values on test results can be derived from statistical properties. Such confidence is needed to estimate the uncertainty of the distribution and, consequently, to move probabilistic real-time one step closer to certifiability. The next paragraph provides the necessary statistical background.

**Background**. To estimate the pWCET distribution, most of the literature uses the Extreme Value Theory (EVT) statistical tool. This theory, developed for natural disasters prediction, can be applied to time measurements to infer the probability of an extreme event to happen. The estimated pWCET is the statistical distribution that better approximates the tail of the original execution time measurement distribution. From the EVT theoretical result, this distribution is always a Generalized Extreme Value Distribution (GEVD) or a Generalized Pareto Distribution (GPD) according to the selected estimation method: in the former case the Block-Maxima (BM) filtering is applied, the latter case the Peak-over-Threshold (PoT) filtering is used. This short paper does not allow us to describe such techniques in details, however, several works are already available in literature for further reading. To be able to provide sound and correct results, EVT requires the satisfaction of some hypotheses. In particular, the input time measurements must be independent and identically distributed and the real (unknown) distribution must be in the domain-of-attraction of an extreme distribution. While the first is mainly dependent on the hardware and software, and it has been already described in previous works [3] [5], the latter is more a statistical detail, far from an easy interpretation on a real system. For this reason, this hypothesis is checked after the analysis, verifying if the distribution output of EVT process adheres, or not, to the tail of input measurements. This is possible thanks to the use of a Goodness-of-Fit (GoF) test, able to detect a violation of the domain-of-attraction condition of the estimated pWCET.

## 2 THE STATISTICAL POWER PROBLEM

The GoF test at the end of the EVT estimation process aims at finding any error in the estimation of the pWCET. This covers not only the previously cited domain-of-attraction hypothesis, but also estimation errors and uncertainties. The most famous GoF tests are the Chi-Squared (CS), Kolmogorov-Smirnov (KS), Cramer-von Mises (CvM), Anderson-Darling (AD) and the Modified Anderson-Darling (MAD) [2]. Such tests have the following hypothesis scheme:

- $H_0 : \ G(x) = F_n(x)$ (null hypothesis)
- $H_1 : \ G(x) \neq F_n(x)$ (alternative hypothesis)

where $G(x)$ is the CDF of the extreme value distribution output of the EVT estimator and $F_n(x)$ is the *Empirical CDF* built on the time samples after the BM/PoT filtering.


*EMSOFT'19, October 13–18, 2019, New York, NY, USA*

ACM ISBN 978-1-4503-6924-4/19/10…$15.00
https://doi.org/10.1145/3349568.3351555

**Hypothesis testing: what can actually do?** The testing procedures are built to detect any violation of the null hypothesis. Such detection can fail in two ways: the null hypothesis $H_0$ is not rejected when it is actually false (false-negative), or the null hypothesis $H_0$ is rejected, in favor of $H_1$, when it is actually true (false-positive). The false-positive error, called *Type I* error, is selected by the experimenter tuning the significance level $\alpha$. The false negative, or *Type II* error, depends on several factors and it cannot be easily controlled or estimated. It is referred by the letter $\beta$.

The impossibility to control $\beta$ forces the statisticians to say that a statistical test can never "accept" the null hypothesis. This is because, if the hypothesis is not rejected and the value of $\beta$ is unknown, no conclusions could be drawn. In the pWCET world, this means that if a GoF test rejects a pWCET distribution, we are sure with $(1 - \alpha)$ confidence that the actual pWCET distribution is wrong. In this case, the pWCET estimation stops[1] and the safety is guaranteed. On the other hand, if the pWCET is not rejected nothing can be said about its validity. This is a safety problem: even if there are no clear evidences that the estimated pWCET is invalid, we have no confidence bound on this statement, making the GoF test completely useless in terms of reliability. A possible solution could be to invert the $H_0$ and $H_1$ hypotheses, however, no statistical test is known to exist with such hypothesis scheme.

**The role of the statistical power**. The statistical power can solve this uncertainty problem of the test result. The statistical power is a scalar value defined defined as:

$$W = P(\text{not reject } H_0 | H_0 \text{ is false}) = 1 - \beta$$

The statistical power $W$ of a goodness-of-fit test depends on several factors: 1) the significance level $\alpha$, 2) the sample size $n$, 3) on the test procedure itself, 4) on the shape of the real (unknown) distribution. The statistical power increases when $\alpha$ or $n$ increases, i.e. we can decrease the rate of false negative results against an increasing of false positive results, or we can decrease the rate of false negative results increasing the number of samples required for testing. The last observation is exactly the point we want to exploit: determine the minimum sample size that is required to reach a certain confidence on the test result. Such *power analysis*, for the best of our knowledge, has never been performed in probabilistic real-time context. The peculiar characteristics of probabilistic real-time, with its high level of confidence required, made necessary to write a dedicated Monte Carlo algorithm to obtain the statistical power estimation. These results, related to the statistical part, have been published in a previous work [4]. As already justified in such dataset, we excluded CS and CvM tests because the first is not adequate to probabilistic real-time, the second because it has been already proved to have less statistical power that other tests. For this reason, we estimated the statistical power for the AD, KS and MAD tests.

The availability of the statistical power allows us to provide a value to the confidence of the null hypothesis non-rejection, in turn, on the pWCET distribution confidence. This result is an enabler for the estimation of the overall reliability of the pWCET distribution.

[1]How to deal this situation is out of scope of this paper. Typical solutions include an in-depth analysis of the causes and the re-tuning of the EVT process parameters.

| $n$ | 50 | 100 | 200 | 500 | 1000 |
|---|---|---|---|---|---|
| **KS** | 0.03 | 0.29 | 0.74 | $1 - 10^{-3}$ | $1 - 10^{-7}$ |
| **AD** | 0.49 | 0.86 | 0.99 | $1 - 10^{-8}$ | $> 1 - 10^{-9}$ |
| **MAD** | 0.26 | 0.80 | 0.95 | $1 - 10^{-8}$ | $> 1 - 10^{-9}$ |

**Table 1: Minimum achieved statistical power among all the dataset scenarios for $\alpha = 0.05$.**

## 3 PRELIMINARY COMPARISON OF GOF TESTS

From the published dataset [4] it is possible to compare the three tests and provide a minimum number of samples required to obtain a large confidence on the result of the GoF test. From the original dataset we computed the statistical powers presented in Table 1 varying the number of samples $n$. It should be noted that the number of samples refers to the input of the GoF test, thus after the BM/PoT filtering and not to the original time measurements. From these results it is possible to conclude the following statements: (1) KS has in general less statistical power than the other two tests, while AD and MAD presents similar behaviour for large number of samples; (2) To obtain a very high confidence on the result of the tests, it is necessary to use a sample of at least 1000 time measurements. This size refers to the sample of observed time measurements in the tail of the distribution, i.e. after applying the BM/PoT filter.

Most of the previous works on probabilistic real-time empirically selected a number of samples, usually not higher than 100, making the non-rejecting result of the domain-of-attraction test very untrustworthy and with low significance. The experimenter should select the sample size depending on the desired confidence level and he or she should pay attention if KS is used, since the number of samples required to obtain a sufficient confidence could be very large. It should be noted that (M)AD tests relies on a internal Monte Carlo estimation, thus they have an intrinsic uncertainty on the test result, that should also be considered and properly analyzed.

## 4 CONCLUSIONS

In this short paper we described the problem of statistical power and how this could affect the reliability of the pWCET. Running a statistical GoF test, for what concern the pWCET safety, is useless if no information is available on the statistical power, that in turn corresponds to the probability of accepting a potentially unsafe pWCET distribution. The minimum number of sample size to obtained a certain confidence required by three tests – KS, AD, and (M)AD – has been presented together with the problematics in the application of such tests for pWCET estimation.

## REFERENCES

[1] G. Bernat, A. Colin, and S. M. Petters. 2002. WCET analysis of probabilistic hard real-time systems. In *23rd IEEE Real-Time Systems Symposium*. 279–288.
[2] J.H. Heo, H. Shin, W. Nam, J. Om, and C. Jeong. 2013. Approximation of modified Anderson–Darling test statistics for extreme value distributions with unknown shape parameter. *Journal of Hydrology* 499 (2013), 41–49.
[3] F. Reghenzani, G. Massari, W. Fornaciari, and A. Galimberti. 2019. Probabilistic-WCET Reliability: On the Experimental Validation of EVT Hypotheses. In *Proceedings of the International Conference on Omni-Layer Intelligent Systems (COINS '19)*. ACM, New York, NY, USA, 229–234. https://doi.org/10.1145/3312614.3312660
[4] F. Reghenzani, G. Massari, L. Santinelli, and W. Fornaciari. 2019. Statistical power estimation dataset for external validation GoF tests on EVT distribution. *Data in Brief* 25 (2019), 104071. https://doi.org/10.1016/j.dib.2019.104071
[5] L. Santinelli, F. Guet, and J. Morio. 2017. Revising Measurement-Based Probabilistic Timing Analysis. In *IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*. 199–208.