14th International Conference on Structural Safety and Reliability - ICOSSAR'25
1-6 June, 2025 - Los Angeles California, USA
R. Ghanem, E. Johnson, S. Masri and A. Olivier (Eds)

HYPOTHESIS BASED RELIABILITY ANALYSIS

Marius Bittner[1a], **Konstantin Zuev**[2] **AND Michael Beer**[1,3,4]

[1] Institute for Risk and Reliability, Leibniz University Hannover
Callinstr. 34, 30167 Hannover
[a]bittner@irz.uni-hannover.de

[2] California Institute of Technology, Dep. of Computing and Mathematical Sciences
1200 E. California Blvd., MC 305-16
kostia@caltech.edu

[3] Department of Civil and Environmental Engineering, University of Liverpool
Liverpool L69 3GH, United Kingdom

[4] International Joint Research Center for Resilient Infrastructure & International Joint Research Center for Engineering Reliability and Stochastic Mechanics,
Tongji University, Shanghai 200092, China

Key words: Reliability Analysis, Hypothesis Testing, Small Failure Probabilities

Abstract. Estimating small failure probabilities of high-impact rare events on engineering systems is still a critical challenge in modern engineering and applied sciences. These probabilities are often expressed as high-dimensional integrals, making accurate estimation computationally demanding. However, in many practical applications, precise probability values are less important than determining whether they exceed a critical threshold. This study addresses this need by reformulating the reliability problem as a hypothesis-testing framework. By developing statistical methods for hypothesis-based reliability tests, this approach offers a computationally efficient alternative to traditional estimation methods. The proposed methodology has significant potential to streamline decision-making processes in reliability analysis.

1 INTRODUCTION

Hypothesis testing is a core concept in inferential statistics used to make decisions about a population based on sample data. It involves formulating two competing hypotheses: the null hypothesis (H_0), which typically represents a statement of no effect or no difference, and the alternative hypothesis (H_1), which represents what the researcher aims to support. Sample evidence is collected and analyzed using a test statistic to determine the probability (p-value) of observing such data. It is by default assumed that H_0 is true and evidence needs to be tested if this can be assumed or not. If the probability obtained by the p-value, which is part of the evidence, is below a predetermined significance level (α), the null hypothesis is rejected in favor of the alternative. This structured approach allows researchers to draw conclusions about population parameters with a known degree of uncertainty, building upon the formal mathematical framework for efficient statistical tests and decision-making established by Neyman and Pearson [1]. Although, it is worth mentioning that statistical tests

of such a kind can already be traced back in the 18th century by researchers as John Arbuthnot [2].

Reliability analysis and more partical reliability engineering is a quantitative field that applies probability theory and statistical methods to evaluate and improve the performance and longevity of engineering systems. At its core, it seeks to model and predict the likelihood of failure, understand the structure of failure processes, and optimize system design for durability and safety. The roots of modern reliability engineering lie not only in formal mathematical theory but also in the practical demands of high-stakes engineering projects. During and after World War II, large-scale initiatives like the Manhattan Project confronted engineers and scientists with unprecedented uncertainty in materials, systems, and outcomes — requiring careful management of safety margins and probabilistic risks. Similar challenges persisted into the Cold War era, particularly in aerospace and nuclear systems engineering. These practical needs catalyzed the development of formal reliability frameworks. By the 1960s, the theoretical foundation of the field was solidified through rigorous work on failure distributions, repair processes, and system reliability structures, such as the foundational work by Barlow and Proschan [3].

A perspective from the 2000s is given in [4], in which a comprehensive discussion of how uncertainties—both aleatory and epistemic—are modeled and managed in structural mechanics. It reviews probabilistic, non-probabilistic, and hybrid approaches for uncertainty quantification, highlighting their relevance for reliability analysis and structural safety assessment. It is stressed that computational efficiency and especially the treatment of the so-called "curse-of-dimensionality" can only be dealt with to a certain degree of accuracy and complexity.

In [5], recent advances in methods for estimating small failure probabilities in structural and multidisciplinary systems are presented. The authors also emphasize the computational challenges and practical relevance when it comes to the estimation of small failure probabilities. It categorizes existing approaches into sampling-based, surrogate-based, and extreme value-based methods, and provides guidance on selecting suitable techniques for both time-independent and time-dependent reliability problems.

These works, among others, state that to efficiently assess that a system is safe or unsafe is still posing a great challenge in reliability analysis problems, especially when estimating small failure probabilities for complex systems such as large networks, systems that exhibit strong non-linear behavior or systems including a large number of variables (especially random variables). However, in engineering design states, or also for stakeholders to make decisions that are in an uncertain and risky environment it is oftentimes not necessary to exactly know the true failure probability, but to have a guess if it is larger or smaller a specific critical probability. To achieve this we propose a reliability hypothesis testing method, which can deliver faster insights on failure probabilities for engineering designs or stakeholders. In [6], Zhang & Ni propose a Bayesian hypothesis testing framework for structural health monitoring. Their approach formulates damage detection as a statistical decision problem, allowing for the quantification of uncertainties and the incorporation of prior knowledge. This method enhances the reliability of damage detection in structures by systematically evaluating the probability of damage presence based on observed data. In our approach by combining existing well established statistical testing methods with a reliability based formulation for assessing failure probabilities, a simple decision making tool can be created.
The formal methodology is firstly presented followed by a very simple example to visualize the ideas and discuss further potential applications.

2 Methodology

We propose a hypothesis based reliability analysis for testing. In this let p_F be the true, but unknown failure probability of any imaginable system and let p_F^* be a preset critical failure probability. This leads to a formulation of the Null hypothesis to be

$$H_0 : p_F > p_F^* \rightarrow \text{“The system is not safe”},$$

and the alternative hypothesis

$$H_1 : p_F \leq p_F^* \rightarrow \text{“The system is safe”}.$$

This formulation is contradictory to the classical hypothesis formulation, where the H_0 would describe the “status-quo”.
In our approach defining H_0 as the undesired case, comes with specific characteristics that are:

1) Burden of proof lies on safety.
We are assuming the system is unsafe until proven safe. This is a conservative approach and aligns well with common engineering ethics and risk-averse decision-making. In hypothesis testing this approach might be similar to new drug approval in medicine: A new drug is assumed unsafe or ineffective (null) until there is statistical evidence of safety or efficacy. In a engineering reliability context this means: A design is assumed not acceptable until it is proven to meet reliability criteria with high confidence.

2) Type I error becomes “accepting an unsafe system”.
Type I error: Reject H_0 when it’s actually true $\rightarrow$ We assume the system is safe, but it’s not. Since this is the more dangerous error, we can choose a small significance level α in our test statistics to control this risk. This makes α directly interpretable as the maximum allowed probability of wrongly accepting a system that is unsafe.

3) Type II error becomes “rejecting a safe system”.
Type II error: Fail to reject H_0 when it’s false $\rightarrow$ We assume the system is unsafe, but it’s actually fine. This might lead to overdesign or unnecessary rejection, which can increase cost or delays. So, there’s a trade-off between safety and efficiency, managed by adjusting the power of the test. However, being conservative for safety critical systems aligns again with common engineering ethics.

4) Philosophical alignment with reliability goals.
This hypothesis structure reinforces a safety-first philosophy: Engineers and decision-makers are forced to prove safety, not merely fail to prove unsafety. It avoids the pitfall of taking non-rejection of H_0 (e.g., “failure probability is high”) as evidence of safety.

5) Statistical design becomes reliability-driven.
The hypothesis test directly evaluates: Whether observed or estimated p_F is significantly less than the threshold, whether model uncertainty and sample size support a reliable decision. This invites techniques like: One-sided confidence intervals for p_F, Sequential testing or Bayesian updating, Power analysis to ensure adequate sample size for detecting safety.

2.1 Test Design

In reliability analysis, the *failure probability* of a system is defined as

$$p_F = \Pr(x \in F) = \int_F \pi(x)\, dx, \tag{1}$$

where $\pi(x)$ is the joint probability density function of the input parameters, and F is the *failure domain*, determined via the limit state function $g(x)$, i.e.,

$$F = \{x : g(x) > y^*\}, \tag{2}$$

with y^* being a critical threshold indicating the boundary between safe and unsafe system behavior.

To determine if the system complies with a target safety requirement, we define a *hypothesis test* that compares the true (but unknown) failure probability p_F with a threshold (critical value) p_F^*. The test is formulated as

$$H_0 : p_F > p_F^* \quad \text{vs.} \quad H_1 : p_F \leq p_F^*, \tag{3}$$

where the null hypothesis H_0 assumes that the system is not sufficiently safe. Rejecting H_0 provides statistical evidence that the failure probability does not exceed the threshold.

To estimate p_F, we can employ methods such as a Monte Carlo (MC) simulation. A sample $X_i \sim \pi(x)$ is drawn and evaluated through the model $g(x)$, yielding

$$X_i = \left(x_1^{(i)}, \ldots, x_d^{(i)}\right), \quad Y_i = g(X_i), \quad Z_i = I_F(Y_i), \tag{4}$$

where I_F is the indicator function:

$$Z_i = \begin{cases} 1, & \text{if } X_i \in F, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

This leads to the model:

$$Z_1, \ldots, Z_n \sim \mathcal{B}(p_F), \tag{6}$$

i.e., the binary responses are i.i.d. Bernoulli random variables with parameter p_F.

We define the test statistic as the inverse of the number of observed failures:

$$s(Z) = \frac{1}{\sum_{i=1}^n Z_i}. \tag{7}$$

The null hypothesis is rejected if the test statistic exceeds a critical value c, which is equivalent to observing too few failures:

$$\text{Reject } H_0 \Leftrightarrow s(Z) \geq c \quad \Leftrightarrow \quad \sum_{i=1}^n Z_i \leq c^{-1}. \tag{8}$$

To control the *Type I error* (false acceptance of a system as safe), we select c such that the test has a desired significance level α:

$$\alpha = \sup_{p_F > p_F^*} \mathbb{P}(s(Z) \geq c \mid p_F) = \sup_{p_F > p_F^*} \mathbb{P}\left(\sum_{i=1}^n Z_i \leq c^{-1} \mid p_F\right). \tag{9}$$

As $\sum_{i=1}^{n} Z_i \sim \text{Bin}(n, p_F)$, this probability can be expressed using the binomial CDF F_{Bin}:

$$\alpha = \sup_{p_F > p_F^*} F_{\text{Bin}}(c^{-1} \mid n, p_F) = F_{\text{Bin}}(c^{-1} \mid n, p_F^*). \tag{10}$$

Solving for the critical value yields:

$$c = \frac{1}{F_{\text{Bin}}^{-1}(\alpha \mid n, p_F^*)}, \tag{11}$$

and the final test decision rule becomes:

$$\text{Reject } H_0 \Leftrightarrow \sum_{i=1}^{n} Z_i \leq F_{\text{Bin}}^{-1}(\alpha \mid n, p_F^*). \tag{12}$$

Lastly, we compute the *p-value* of the observed data as

$$p = F_{\text{Bin}}\left(\sum_{i=1}^{n} Z_i \mid n, p_F^*\right), \tag{13}$$

which represents the probability, under the assumption $p_F = p_F^*$, of observing a failure count as small or smaller than that observed.

2.2 Bootstrap Method for Hypothesis Based Reliability Analysis

In the hypothesis testing procedure for assessing whether the failure probability p_F exceeds a critical threshold p_F^*, we use a test statistic based on n independent binary failure indicators $Z_i \sim \mathcal{B}(p_F)$, as already stated in Eq. (6). The associated p-value is computed via a binomial distribution under the assumption that $p_F = p_F^*$ via the CDF in Eq. (13).

Since this p-value is computed from a random sample, it is itself a random quantity subject to variation. This leads to the fact that when repeating the same test with different samples (of size n), p-values underlay a variation. To quantify this variability, we employ the *bootstrap method*.

Bootstrap Procedure

Given a single set of observed binary outcomes $\{Z_1, Z_2, \ldots, Z_n\}$, as from Eq. (5) and subsequently Eq. (6), we compute another test statistic $k(Z) = \sum_{i=1}^{n} Z_i$ and corresponding p-value:

$$p = F_{\text{Bin}}(k(Z) \mid n, p_F^*). \tag{14}$$

To estimate the uncertainty in this p-value, we apply the bootstrap as follows:

1. **Resample:** From the observed data $\{Z_1, \ldots, Z_n\}$, generate n_B bootstrap samples by sampling with replacement.
2. **Recalculate:** For each bootstrap sample $b = 1, \ldots, n_B$, compute the test statistic $k^{(b)}$ and corresponding bootstrap p-value:
$$p^{(b)} = F_{\text{Bin}}(k^{(b)} \mid n, p_F^*). \tag{15}$$
3. **Estimate Distribution:** From the bootstrap distribution $\{p^{(1)}, \ldots, p^{(B)}\}$, estimate a confidence interval by computing the empirical quantiles:
$$\text{BCI}_{\text{lower}} = Q_\alpha(p^{(b)}), \quad \text{BCI}_{\text{upper}} = Q_{1-\alpha}(p^{(b)}) \tag{16}$$
where $Q_\alpha(x)$ is the α-quantile of x.

Bootstrap Stopping Criterion

Rather than relying solely on a single p-value compared to a fixed significance level α, we monitor the entire bootstrap confidence interval:

$$\text{Reject } H_0 \quad \Leftrightarrow \quad \text{BCI}_{\text{upper}} < \alpha, \tag{17}$$

$$\text{Accept } H_0 \quad \Leftrightarrow \quad \text{BCI}_{\text{lower}} > 1 - \alpha. \tag{18}$$

This approach provides a more robust decision rule by ensuring that the uncertainty in the p-value is fully considered. If the bootstrap confidence interval is not fully inside the threshold α, the test remains inconclusive, and additional samples should be collected to reduce uncertainty. Using the bootstrap method helps mitigate the risk of random fluctuations in the data leading to incorrect conclusions, particularly in borderline cases where the true failure probability p_F is close to the threshold p_F^*. This enhances the reliability of statistical conclusions in safety-critical engineering contexts.

3 Numerical Tests

With the introduction of the Binomial test in Eq. (10), which in the realm of reliability analysis is a versatile first approach, we can already perform some simple numerical tests. These are shown here in the following. In Fig. 1 different cases and different trials are carried out, for the simple Binomial test, these trials already appear quite robust. Please note that the number of assessed samples n is adjusted depending on the values in p_F and p_F^*.

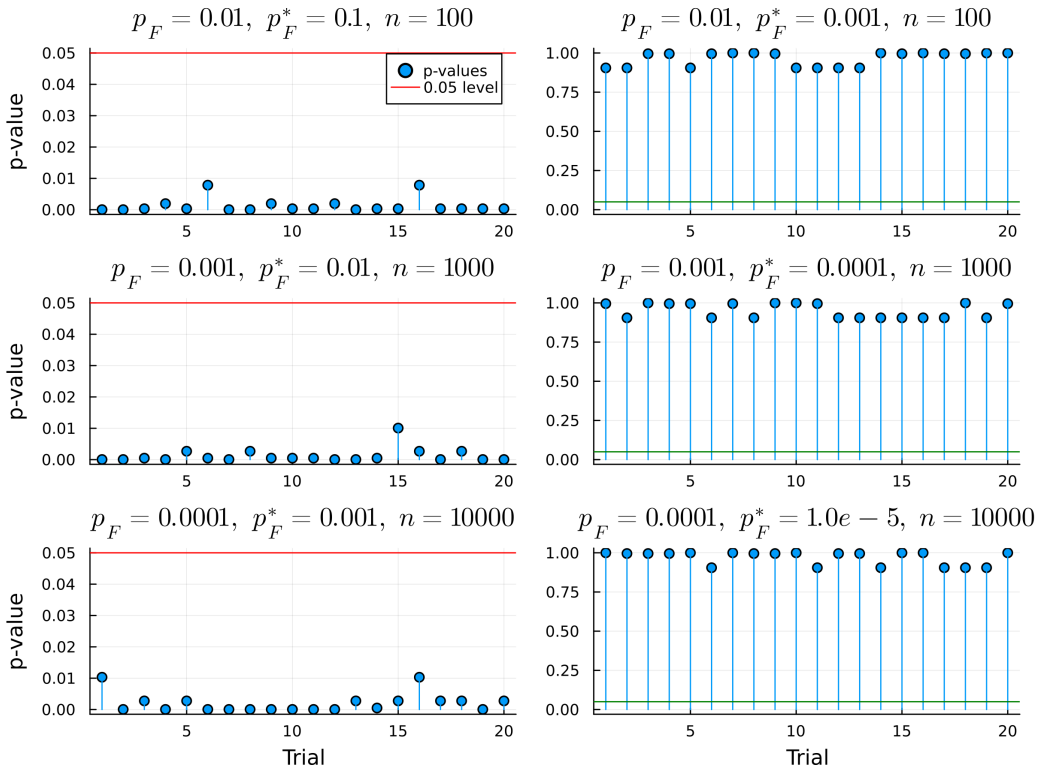

Figure 1: Binomial test example p-value results for different p_F, p_F^* and n.

Since the number of needed samples and the true failure probability is usually not know a-prior, a convergency analysis regarding the sample size n should be carried out. These results can be seen Fig. 2. The test result in this case, seems to fluctuate, which leaves us still with inconclusive information. However, it can be seen that with a sufficiently large sample size n, the test seems to be robust.

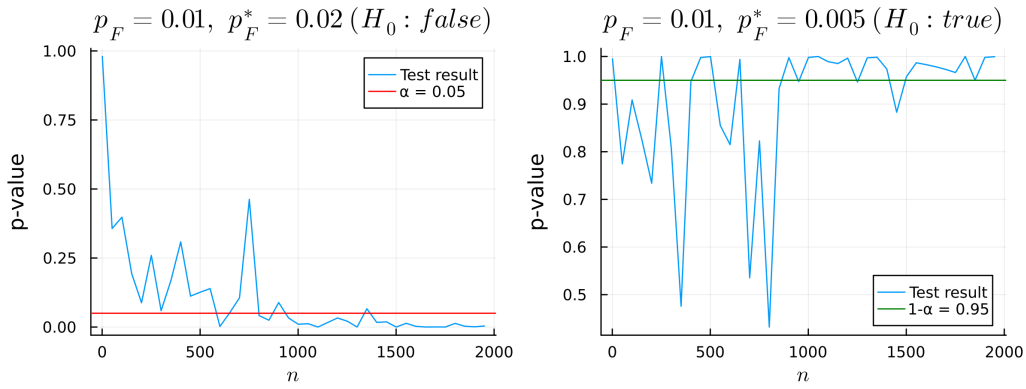

Figure 2: Binomial test example p-value results for two cases of p_F, p_F^* and a simple convergency analysis regarding n.

The robustness becomes clearer, when performing an MC simulation for each of the test sample sizes. In Fig. 3, for each configuration of n an additional MC simulation is carried out to estimate several p-values, then the mean of these p- values is taken. The p-value is converging. However, adding MC simulation samples, for real engineering applications increases the needed computational effort. To overcome this, the bootstrap method as already introduced in Section 2.2 is employed.

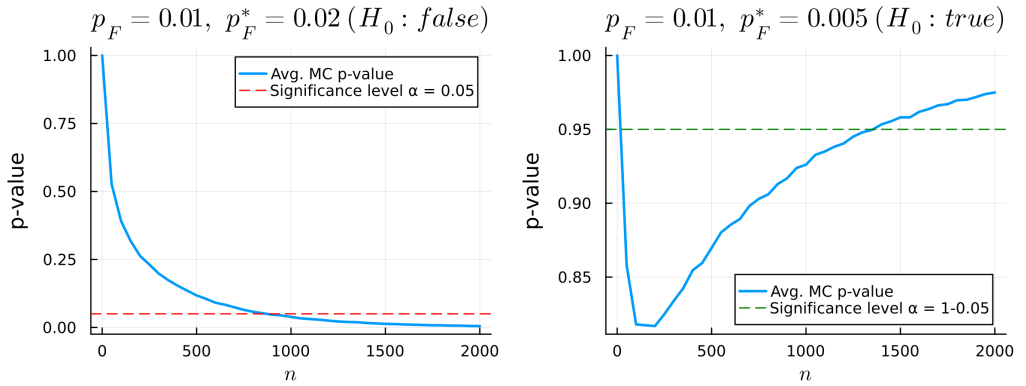

Figure 3: Averaged p-values of different MC simulations (1000 MC samples), for two cases of p_F, p_F^* and a simple convergency analysis regarding n.

The application of the bootstrap method with the stopping criterion can be seen in Fig. 4. Here a robust decision for rejecting H_0 was made after $n = 900$ model evaluations, in this case for accepting H_0, $n = 1500$ model evaluations were needed. This aligns with the results from the MC simulation. But please note, since for the bootstrap method no new model evaluations are needed, this approach is computationally more efficient.

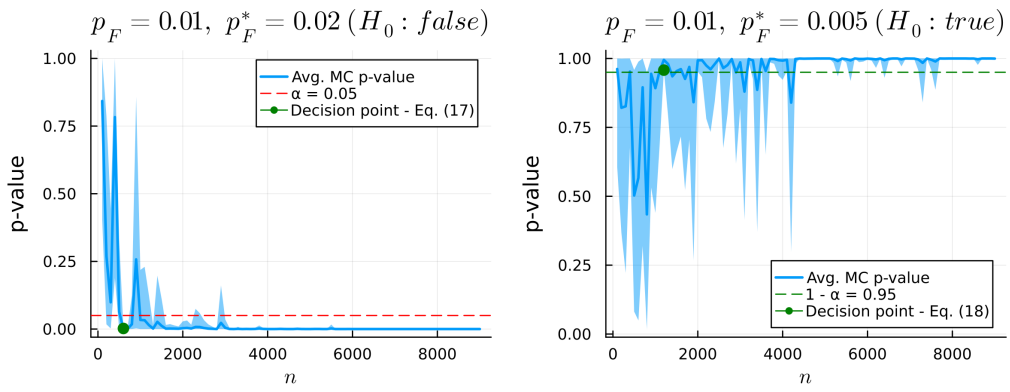

Figure 4: Bootstrap approach for the confidence intervals of the hypothesis based reliability test.

4 Conclusion

Hypothesis-based reliability analysis offers a rigorous statistical framework for decision-making under uncertainty. By formulating engineering safety questions as hypothesis tests—e.g., "Is the failure probability below an acceptable threshold?"—engineers can apply well-established statistical tools to assess safety with quantifiable confidence. This approach naturally enforces a safety-first philosophy, requiring strong evidence before declaring a system safe, and avoiding the misinterpretation of non-rejection as confirmation. Moreover, the bootstrap method enhances this process by providing confidence intervals for p-values without relying on repeated model-based simulations. Especially for complex engineering systems, where each simulation can be computationally expensive, the bootstrap allows efficient uncertainty quantification based solely on observed or simulated samples. This can significantly reduce computational cost while maintaining robustness in safety assessments. However, additionally to these very simple examples, more complex and realistic examples should be analyzed. Also numerous extensions are imaginable.

5 Acknowledgments

M. Bittner would like to thank K. Zuev & M. Beer for this opportunity, for the patience, trust, insights and guidance. The authors also acknowledge the Humboldt-Foundation.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used GPT-4o in order to increase readability, flow of text and check for grammar/spelling issues. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

REFERENCES

[1] Neyman, J. and Pearson, E.S.; On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* (1933), Vol. 231, No. 694–706, pp. 289–337.

[2] Heyde, C.C.; *Statisticians of the Centuries*. Springer Science, New York, (2001).

[3] Barlow, R.E. and Proschan, F. *Mathematical Theory of Reliability*. Wiley, (1965).

[4] Schuëller, G.I., On the treatment of uncertainties in structural mechanics and analysis, *Computers & Structures* (2007), Vol. 85

[5] Lee, I. and Lee, U. and Ramu, P. and Yadav, D. and Bayrak, G. and Acar, E., Small failure probability: principles, progress and perspectives, *Structural and Multidisciplinary Optimization* (2022), Vol. 65, No. 11

[6] Zhang Q., Ni Y. A Bayesian hypothesis testing-based statistical decision philosophy for structural damage detection. *Structural Health Monitoring* (2022), Vol. 4