Journal of Quality Measurement and Analysis *JQMA* 22(SI) 2026, 65-78
e-ISSN: 2600-8602 *http://www.ukm.my/jqma*
https://doi.org/10.17576/jqma.22si.2026.05

A DYNAMIC FRAMEWORK FOR DETERMINING OPTIMAL EXPERIMENTAL REPETITION BASED ON DATA-DRIVEN CRITERIA

(*Rangka Kerja Dinamik bagi Menentukan Bilangan Optima Ulangan bagi Eksperimen Berpandukan Kriteria Data*)

AZHARUDIN MUKHTARUDDIN*, MOHD SYAZWAN MOHAMAD ANUAR & MAT KAMIL AWANG

ABSTRACT

Experimental research plays a crucial role in the scientific method as a means to acquire reliable knowledge. Ensuring the dependability and reproducibility of an experiment requires careful consideration of the number of repetitions or runs conducted. This study proposes a dynamic and data-driven framework to determine the optimal number of experimental repetitions based on the specific requirements of each experiment. The proposed steps begin with a normality test, followed by statistical comparison across data sets, identification of the required number of runs, and finally post-hoc confirmation. Several statistical techniques were employed, including Kruskal-Wallis test, power analysis, Sequential Probability Ratio Test, confidence interval estimation, and cumulative standard deviation to demonstrate the framework. Application of the framework to real-world electric field measurements obtained from a Giga-Hertz Transverse Electromagnetic (GTEM) cell demonstrates that the determination of an adequate number of runs is highly dependent on data behaviour and methodological assumptions. This open-ended yet structured approach empowers experimenters to design and analyze experiments with statistical rigor, enhancing both reliability and reproducibility.

Keywords: repeatability; experiment; framework; data-driven; statistical methods; Giga-Hertz Transverse Electromagnetic (GTEM); partial discharge

ABSTRAK

Penyelidikan eksperimen memainkan peranan yang penting dalam kaedah saintifik sebagai satu cara untuk memperoleh pengetahuan yang boleh dipercayai. Menjamin kebolehpercayaan dan kebolehulangan sesuatu eksperimen memerlukan pertimbangan yang teliti terhadap bilangan ulangan atau larian eksperimen yang dijalankan. Kajian ini mencadangkan satu rangka kerja yang dinamik dan berpandukan data bagi menentukan bilangan ulangan eksperimen yang optimum berdasarkan keperluan khusus setiap eksperimen. Langkah-langkah yang dicadangkan bermula dengan ujian kenormalan, diikuti oleh perbandingan statistik merentas set data, penentuan bilangan larian yang diperlukan, dan seterusnya pengesahan pasca-hoc terhadap bilangan larian tersebut. Beberapa teknik statistik telah digunakan, termasuk ujian Kruskal–Wallis, analisis kuasa, Ujian Nisbah Kebarangkalian Berjujukan (Sequential Probability Ratio Test), anggaran selang keyakinan, dan sisihan piawai kumulatif bagi menunjukkan keberkesanan rangka kerja ini. Aplikasi rangka kerja tersebut kepada data pengukuran medan elektrik dunia sebenar yang diperoleh daripada sel Giga-Hertz Transverse Electromagnetic (GTEM) menunjukkan bahawa penentuan bilangan larian yang mencukupi sangat bergantung kepada tingkah laku data dan andaian metodologi. Pendekatan yang terbuka namun berstruktur ini membolehkan penyelidik mereka bentuk dan menganalisis eksperimen dengan ketelitian statistik, sekali gus meningkatkan kebolehpercayaan dan kebolehulangan hasil eksperimen.

Katakunci: kebolehulangan; eksperimen; rangka kerja; berpandukan data; kaedah statistic; Giga-Hertz Transverse Electromagnetic (GTEM); discaj separa

1. Introduction

Repeatability is defined as having the same experiment done more than once by the same team with the same set-up (Gundersen 2021). Essentially this means the data generated from the experiment is stable from statistical point of view and this is an important component in defining experiment reproducibility (Wang *et al.* 2025). Gundersen (2021) defined reproducibility is the ability to draw the same conclusions from an experiment as described by the original experimenters. Reproducibility is the halmark in science as it provides the means to validate a claim (Peng 2011). In experimental research, determining the appropriate number of repetitions is critical to ensuring the validity, consistency, and reproducibility of results (Montgomery 2013). Repetition helps reduce random error and increases the statistical power of the findings (Field 2024). However, traditional approaches often rely on a fixed number of experimental runs, which may not be suitable across diverse experimental contexts. This can result in under-sampling or excessive repetitions, both of which reduce efficiency and may compromise data reliability. As modern experiments become more complex and data-intensive, there is a growing need for a more adaptive and data-driven approach to determine when enough repetitions have been achieved (Casella & Berger 2024).

This study introduces a dynamic framework for experimental repetition determination, which responds to the statistical characteristics of the collected data. The framework is a four-step process: 1) normality test, 2) asssess statistical differences between runs, 3) determine the number of required runs, and 4) Post-hoc confirmation. Several statistical tools were considered, including the Anderson-Darling test for normality assessment and the Kruskal-Wallis test for detecting distributional differences. However, note that those tools are selected based on their suitability rather than as fixed choices. The proposed framework does not introduce new statistical tests. Rather, its contribution lies in the structured integration of established statistical tools into a sequential, data-dependent decision workflow for determining the required number of experimental repetitions.

The application of the proposed framework was demonstrated using real experimental data involving electric field measurements in a Giga-Hertz Transverse Electromagnetic (GTEM) cell. This case study serves to demonstrate the operationalisation and practical feasibility of the framework. The outcomes highlight the importance of tailoring repetition decisions to the experiment's behaviour rather than applying fixed standards. Unlike conventional fixed-sample designs, the proposed framework makes the stopping decision explicit, transparent, and responsive to observed statistical behaviour.

2. Literature Review

In a study conducted to gauge perception of the importance of repeating the same experiment, it was found that students recognised it as a way to develop their technical skills and content knowledge (Wiggins *et al.* 2021). This fact is not surprising as repetition has long been regarded as a method of mastering knowledge or skills. However, the question is to determine how many repetitions are considered enough. Take an example of repetition of a full factorial experiment reported in Morovic *et al.* (2022). In the study, the number of repetitions for different samples were explicitly presented. However, the reason behind these numbers is not clear. A probable reason may be due to time and cost constraints (Moffat 1988). These could be the common limitations.

There is another type of experiment where repeating the same experiment could be done at minimal cost and time. Take the experiment in the study of partial discharge (PD) phenomena as an example. Partial discharge is a phenomenon that takes place when electrical insulation experiences imperfections, for example void, treeing, and pointy surface. As an example, the void inherently has a lower insulation characteristic from the rest of the material. The resulting electric field inside the void is higher than what is designed. As the result, there is a forced electron avalanche inside the void that caused discharge to develop. This phenomenon is known as PD because it takes place at a very short time before extinguishing itself and does not lead to the actual breakdown of the insulation. The discharge can be detected through several means and one of them is in the form of electromagnetic (EM) pulse signal. One of the characteristics of the pulse is it has a very short time usually in the range of nanoseconds to microseconds.

Research about the PD pulse characteristics is particularly important as it can reveal a lot of information about the source of the pulse. Hence, the pulse is commonly generated in laboratory using specific experiment set-up. Each pulse is considered as one sample and is repeated several times using identical experiment set-up. For example, an early PD experimenter repeated the production of the pulse 20,000 times (Bartnikas 1987). The same number of samples were also reported in Sriram *et al*. (2005) with 1 microsecond separation duration between two samples. No clear reason was given on the selection of the number. Authors in Bajwa *et al*. (2015) noted that number of samples should be selected based on the noise statistical characteristics. The study used wavelet transform (WT) and the noise was related to the process. The actual number of samples was also not explicitly mentioned in the report. However, in Yao *et al*. (2025), the authors applied machine learning (ML) to determine the smallest samples based on the final accuracy. They clearly stated the best accuracy was achieved at 200 iterations. Based on the studies, methods for deciding the number of repetitions were not explicitly discussed.

Documents that provide guidelines on related experimentation also leave the matter inconclusive. For example, a related good practice guidance produced by National Physical Laboratory, UK, does not specify how to determine the number of experiment repetition (Nothofer *et al.* 2003). Electromagnetic compatibility (EMC) testing and measurement techniques standard IEC 61000-4-20 focuses on emission and immunity test methods using transverse electromagnetic (TEM). Data used to demonstrate the proposed framework in this study came from an experiment where a generated PD pulse transmission was conducted in a variant of TEM, Giga-Hertz TEM (GTEM). The standard provides procedural guidance for GTEM validation, including field uniformity assessment and measurement configuration. However, it does not provide a statistical framework for determining the sufficient number of experimental repetitions required for data reliability (International Electrotechnical Commission 2022a). Subsequent article regarding the standard also does not include specific discussion about the matter (Burghardt *et al.* 2021).

Decision on the number of samples in ML is discussed comprehensively in (Batista *et al*. 2004; Rajput *et al*. 2023). In the latter study, apart from monitoring of the level of accuracies the authors also used changes in variances while adding more samples in the analysis. Crucially, statistical parameter was used as one of the indicators in determining the number of samples. An advanced study on data repeatability as in Wang *et al*. (2025) also used statistical reasoning. In addition, a new statistical-based algorithm was proposed in Petitjean and Musset (2025) in repeatability situation. This shows the applicability of statistical methods in guaranteeing the reliability of an experiment. In the same spirit, the current study also follows the steps in proposing a new framework for determining the number of repetitions of an experiment.

During the development of the proposed framework in this work, background studies were made about the data normality (Mishra *et al*. 2019a; Vaclavik *et al*. 2019) as well as on statistical properties differences for multiple data sets (Peck *et al*. 2020). Both reviews provided a foundation for the first and the second steps of the proposed framework which is selecting suitable methods to analyse statistical properties of dataset (Mishra *et al*. 2019b). For the third step, review on methods to determine required runs was made (Curtis *et al*. 2022; Antony 2023).

Since the third step is the core of the framework, further in-depth review was carried out. The determination of an adequate number of experimental repetitions has been addressed in the literature through multiple perspectives or classes including power-based design, precision-driven estimation, and practical resource considerations (Montgomery 2013). In addition to the three classes, Wald (2004) proposed sequential and trend-based approaches that relaxed the fixed-sample assumptions by allowing stopping decisions to be updated as data accumulate. Apart from that, stability- and convergence-based criteria assess adequacy by examining whether statistical properties stabilise with increasing repetitions (Hall & Martin 1988; Wang *et al*. 2025). Finally, practical experimental studies recognise that repetition is often constrained by time, cost, and instrumentation limitations, leading to resource-aware stopping considerations (Moffat 1988). However, this limitation is not part of the focus for this study. Table 1 tabulates the approaches related to step 3 and examples of methods that applicable to each approach. The selection of suitable approach and method is discussed as part of the Methodology section.

Table 1: Approach classes and example of methods to determine number of runs

Approach class	**Example methods**	**What the method assesses**	**Limitations**
Power-based / detectability-based	Power analysis, Sequential Probability Ratio Test (SPRT), likelihood ratio trend (Wald 2004; Montgomery 2013)	Whether the number of runs is sufficient to detect statistically meaningful differences	Requires effect size assumptions; often impractical with limited observations per run
Precision-based	Confidence interval width, margin-of-error estimation (Hahn & Meeker 2011; Montgomery 2013)	Whether estimates (e.g. mean or dispersion) have reached a desired precision	Requires user-defined precision thresholds; sensitive to early variability estimates
Stability- / convergence-based	Cumulative standard deviation (CSD), convergence of variance or summary statistics (Hall & Martin 1988; Montgomery 2013)	Whether statistical properties stabilise as more runs are added	Requires definition of stability criteria (tolerance, window length); thresholds may appear subjective
Sequential / trend-based	Moving-window analysis, trend monitoring of statistics (Wald 2004; Ghosh *et al.* 2011)	Whether changes between successive runs diminish over time	Requires careful interpretation; may be sensitive to local fluctuations

The final step of the proposed framework concerns post-hoc confirmation of stability. This step is motivated by the notion that reliable conclusions should be supported by evidence that observed statistical behaviour persists rather than fluctuates under continued repetition, as emphasised in empirical repeatability studies (Wang *et al*. 2025). In addition, the analysis adopts a temporal partitioning perspective, which is well established in data stream and temporal data analysis, where different portions of a time-ordered dataset are examined to assess persistence or change in behaviour (Gama 2010). The resulting confirmation procedure is introduced as a pragmatic validation step rather than as a formal statistical test.

3. Methodology

This section is divided into two parts. The first presents the experimental background that underpins the proposed framework, providing foundational context for its development. The second outlines the procedural steps of the framework itself.

3.1. *Experimental background*

A study to determine critical electrical parameters for a proposed antenna was designed with several supporting experiments. The antenna was designed to be used in detecting electromagnetic signal from partial discharge. One of the experiments was to study a shielded cell in which the antenna would be tested. This experiment was considered as part of the preliminary works to support the main study. Data from the experiment were used to demonstrate the framework proposed in this paper. The objective, methodology and expected outcome from the experiment is as listed below.

- **Objective**: the objective of this preliminary work was to determine the consistency of electric field (***E***) inside an enclosure known as Giga-Hertz Transverse Electromagnetic (GTEM) cell. The cell was developed and calibrated so that it became what is defined as shielded environment in standard IEEE 299.1 (Institute of Electrical and Electronics Engineers [IEEE] Standard Association 2013). Shielded environment means that only limited external electromagnetic (EM) field could infiltrate into the cell so that it could not significantly affect the internal EM field. The same goes for the opposite: only limited EM field is allowed to escape from the cell. The experiment to determine the uniformity of ***E*** inside the cell followed guidelines as specified in IEC 61000-4-3 (International Electrotechnical Commission [IEC] 2020)
- **Setup/Study**: a pulse generator was used to launch an electrical pulse to replicate partial discharge into the GTEM cell through a specific design of conductors inside the cell. The generator was calibrated so that each generated pulse was identical. The system of cables connecting the generator and the launch terminal of the GTEM were adequately protected from interference of external EM field. The sensor used to measure the data was an antenna with known electrical parameters. Five hundred data sets (or 500 runs) were collected with each run contained 53 observations. The setup of the experiment was identical throughout all runs.
- **Expected outcome**: The travelling pulse produced EM field inside the cell. The GTEM cell was designed to produce a uniform EM field inside the cell. In addition to that, the GTEM was designed to provide shielded environment, only insignificant EM field was expected to infiltrate. Because of the size of the cell and type of sensor used in detecting the field as well as stipulated in standard IEEE 299.1, only ***E*** was measured. By proving that the ***E*** was uniformed led to a conclusion that indeed the GTEM cell was able to fulfill the definition as shielded environment as outlined in standard IEEE 299.1.

3.2. *The experiment set-up*

The GTEM cell used in the experiment was bespoke designed and built to the specification as described in (Judd *et al.* 1998; Siegel *et al.* 2016; Judd *et al*. 2018). The design of the cell is as shown in Figure 1. On the top of the cell is an opening or aperture with the size of 0.3 m x 0.3 m and is covered with a movable aluminium cover. There is an insertion hole located in the middle of the cover where the cable connected to an antenna is lowered in to the cell. In actual esperiment, there are five insertion holes and the experiment was repeated each point. The antenna is connected to a digital storage oscilloscope (DSO) to capture and analyse signals captured inside the cell. Note that a set of septum (a set of copper wires) as seen from the opening.

Output from a function generator is connected to the launching unit or terminal. The terminal is essentially a connection point between the function generator and the septum. The signal from the function generator travelled in the septum, produced electrical and terminated at the tensioning mechanism. Suitable insulators are installed for the termination. In order to eliminate gaps between all connections, aluminium foil is used.

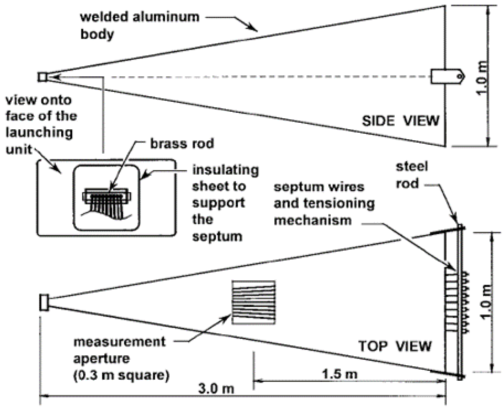

Figure 1: Structure and size of the wire septum GTEM cell (Judd *et al*. 1998)

The placement of the antenna is as shown in Figure 2. The figure shows part of the cell detailing the location of the antenna inside an imaginary boundary known as 'defined area' (DA). The DA is defined as a location in which the antenna can be placed and it must be only 1/3 of the height from the septum to the outer conductor. In turn, the distance from the DA to the septum must not exceeding 1/3 of the height from the septum to the outer conductor (Nothofer *et al.* 2003).

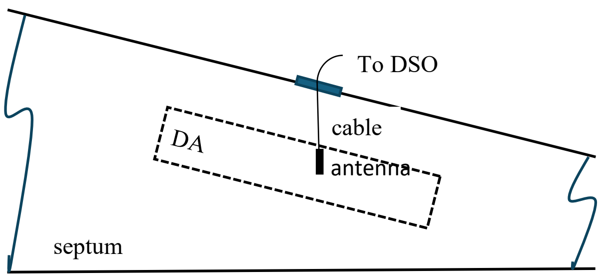

Figure 2: Side view showing the arrangement of the antenna inside thedefined area (DA)

As the result of the conditions, the placement of the antenna for the middle point is calculated to be at least 8.7 cm from the septum and must be at least 8.7 cm from the opening. This is due to the fact that the total distance from the opening to the septum is 26 cm. The antenna used in the experiment is a commercial dipole antenna (Link Technologies 2018).

3.3. *The proposed framework*

The original experiment was repeated 500 times (runs). In this study, the results from those runs were re-evaluated in order to determine how many repetitions are needed based on a framework that is supported with statistical methods. The proposed framework has the following flow as shown in Figure 3. Step 1 is normality test on a set of data (run). The purpose of the normality assessment in step 1 is not to enforce a distributional label, but to determine the class of statistical methods that can be validly applied in step 2 which is testing the statistical difference between two or more runs. Step 3 determines the adequate number of runs. This is followed by step 4 where the post-hoc confirmation about the decision in step 3 is made.

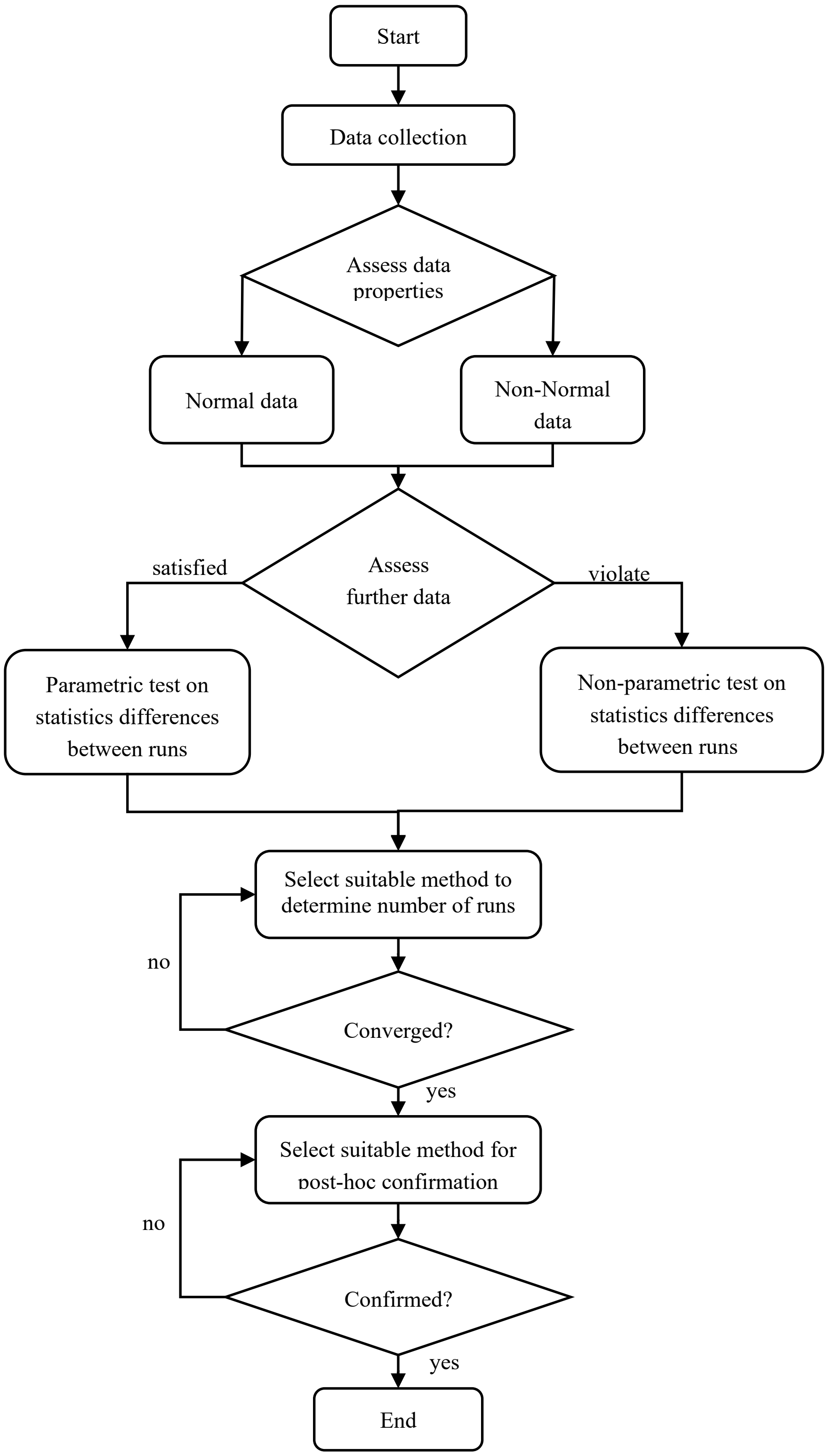

Figure 3: The proposed framework to determine the required number of runs

As mentioned before, Step 1 acts as a prerequisite that constrains and legitimises the statistical comparison conducted in Step 2. There are several methods available to analyse the data normality. The selection of the method must be based on the data properties such as the number of observations, measurement scale, and independence between runs. The objective of Step 2 is to determine whether repeated experimental runs have statistically significant differences between them. If no statistically meaningful differences are detected then each run behaves consistently. However, if statistically meaningful differences are present then run-to-run variability is non-negligible. Following the outcome in Step 1, the selection of method to test statistical difference in Step 2 can be made. Statistical tests are selected based on whether their assumptions are reasonably satisfied by the observed characteristics of the data. That can be done by for example inspecting the data distribution (from Step 1), sample size, independence (by design), and measurement scale. Table 2(a) and (b) shows the example of data properties and suitable methods for normal and non-normal data distribution respectively. If the data are non-normal, rank-based non-parametric methods such as the Kruskal–Wallis test provide an alternative for comparing multiple groups (Hollander *et al*. 2013).

Table 2(a): Example of statistical methods for assumptions compliance

Data property	**Observed characteristic**	**Example methods**
Sample size per run	Moderate to large	ANOVA, t-test
Variance across runs	Variance homogeneity assumed	One-way ANOVA
Variance across runs	Variance homogeneity not assumed	Welch's ANOVA

Table 2(b): Example of statistical methods for assumptions violation

Data property	**Observed characteristic**	**Example methods**
Sample size per run	Small to moderate	Kruskal–Wallis test
Variance across runs	Unequal or unstable	Kruskal–Wallis test
Measurement scale	Ordinal or discretised	Kruskal–Wallis test

The third step is the actual determination of the number of runs. Several methods were tried in determining the required number of runs or repetitions. For example, the first method was the power analysis. The purpose of this method is to ensure that a sufficient number of runs are conducted to reliably detect true differences between datasets, thereby confirming that the observed variations stem from inherent statistical properties rather than random chance. However, this method failed to reach the required power due to low observations in each run. Sequential Likelihood Ratio Test (Wald's SPRT) and Likelihood Ratio Trends methods confirmed the lack of convergence. There was no convergence using both methods. The subsequent method employed in the study was the confidence intervals comparison. The purpose of this method was to determine the number of runs needed to reach stable mean estimates.

Given the limitations observed in power-based and confidence-interval-based approaches, the cumulative standard deviation (CSD) was adopted as the primary criterion for determining the adequate number of runs. The threshold specifies the maximum allowable absolute change in CSD between successive evaluations that will be regarded as practically negligible. The accumulation level specifies the number of consecutive runs for which the CSD change must remain within this threshold before stability is declared. These parameters are intentionally experimenter-defined, allowing the stopping rule to be adjusted according to the desired strictness, experimental constraints, and risk tolerance – consistent with the dynamic and human-centric nature of the proposed framework.

The fourth step is the post-hoc test to confirm the determined number of runs calculated previously. Once again, there are several methods available for the purpose. In this study, an early-runs and late-runs (ERLR) comparison was employed to assess whether summary statistics, including mean and dispersion, remain consistent between the initial and final portions of the run sequence.

As described, the framework consists of four steps. Each step has different objectives and methods. The selection of method in each step is based on the properties and assumptions of the data. Consequently, the framework is dynamic, reflecting both the observed data behaviour and experimenter-defined decision criteria. This reflects that adequacy of repetition is context-dependent and must balance statistical stability against constraints such as cost, time, and risk tolerance. Rather than having rigid stopping rules, the proposed framework makes these decisions explicit and transparent, allowing the stopping criterion to be adjusted according to the objectives of the study. Importantly, sensitivity to these parameters is demonstrated in the results by showing how stricter criteria lead to increased required runs, thereby ensuring that conclusions are not contingent on a single arbitrary setting.

The proposed framework is intentionally designed to be flexible and adaptive, but its correct application requires adequate understanding of statistical methods and descriptive data analysis. This requirement cannot be understated. Looking back at each step, there are a multitude of challenges such as ability to acknowledge type of and properties of the data that one is dealing with and deciding on suitable statistical tools for a given situation.

4. Results

This section presents results from the application of the proposed framework on the prescribed data. The results are presented according to the steps described in the methodology section as the following:

- Subsection 4.1: the normality test.
- Subsection 4.2: testing the statistical difference between runs.
- Subsection 4.3: determine the required experiment runs.
- Subsection 4.4: post-hoc confirmation on the number of experiment runs.

4.1. *Normality test*

The Anderson-Darling test on normality was applied to all 500 data sets of which only 42 were found to exhibit normal distribution with $p > 0.05$. Table 3 shows the sets that passed the test.

Table 3: Datasets that pass the Anderson-Darling test

Datasets that pass the Anderson-Darling test
10, 20, 41, 44, 48, 101, 121, 129, 133, 134, 138, 143, 152, 160, 162, 165, 179, 184, 204, 213, 220, 225, 234, 240, 246, 270, 274, 286, 322, 335, 343, 346, 353, 358, 372, 395, 404, 439, 448, 458, 462, 488

Based on the result, there was no discernible pattern among the sets.

Figure 4 is the visualisation of the first 12 data sets (marks with column) including set number 10 that passed the normality test. The x-axis for each bar chart is the response magnitude in volt (V) bin and the y-axis in the count for each bin. Visually, it is understandable why the normality test returned a result where majority of the data sets were non-normal with p > 0.05.

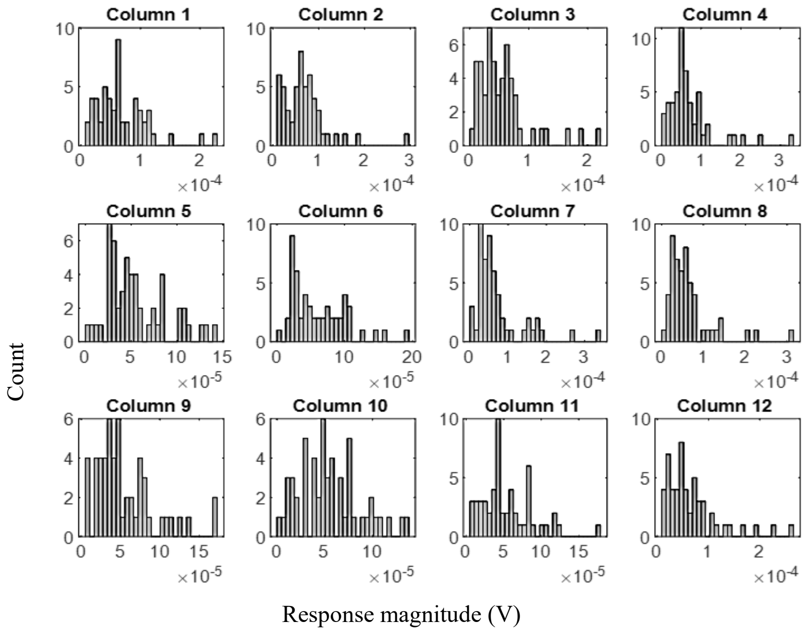

Figure 4: Histogram charts for the first twelve data sets (count versus response magnitude volt (V))

4.2. *Inspecting statistical differences between data sets*

The presence of so many non-normal data sets left the Kruskal-Wallis test (KWt) as the choice to analyse whether samples originated from the same distribution. The outcome of the test showed that there was significant difference detected among the 500 runs ($p < 0.05$; $p = 7.3054 \times 10^{-9}$). Post-hoc test to determine which pairwise had significant difference was not required. This was because the result from the KWt was sufficient to conclude that there were data sets that were statistically different. Hence, it is not safe to conduct the experiment with only one run, analyse the data and produce a conclusion. This is despite all preventive measures designed to produce identical response.

4.3. *Determination of required experiment runs*

Using Wald's SPRT with $H_0{:}p=0.5$, $H_1{:}p=0.505$, $\alpha=0.05$, and $\beta=0.20$, the likelihood ratio remained between 0.4127 and 1.0393 across 500 runs and did not cross the decision thresholds (0.2105, 16); hence, the test remained in the continuation region and did not terminate with a decision (i.e., no evidential convergence toward H_0 or H_1 within the available runs). A fixed-sample power analysis for detecting a small directional deviation ($p_1=0.505$) at $\alpha=0.05$ with 500 runs yielded an achieved power of approximately 0.06, indicating that the data are insufficiently powered to detect such small deviations even under an optimal fixed-sample test. A 95% Wilson confidence interval for the run-level probability yielded $p \in [0.454, 0.542]$, which includes both the null value (p=0.5) and the SPRT alternative (p=0.505), indicating that the data at 500 runs are insufficient to resolve such small deviations.

To find the probable adequate runs of experiment, cumulative trend of standard deviation (CSD) was used. By setting the accumulation at 5 consecutive runs with a threshold of 1.5% (0.015), the probable adequate runs were found to be 26. The following Table 4 shows the CSD for each run and absolute changes between each run. Notice that from run 22 to 26 the absolute change is less than 1.5%. A lower threshold of 1% was not achievable over five consecutive runs; therefore, 1.5% was selected at 5 consecutive runs.

Table 4: CSD and absolute change between runs

Runs	22	23	24	25	26
CSD	2.6910×10^{-5}	2.6795×10^{-5}	2.6439×10^{-5}	2.6690×10^{-5}	2.6903×10^{-5}
Absolute change	0.0011	0.0043	0.0135	0.0094	0.0079

The overall change is presented in Figure 5. Note that at run = 1, there is no change since it is the initial data. Change at run = 2 is for the change of standard deviation between point 2 and point 1. Further inspection shows that beyond run 26, the change between consecutive runs 28 and 27 is 5.37% and between consecutive 39 and 38 runs is -2.04%. Other pairwise data sets with change higher than 1.5% are 41-40 (-1.80%) and 70-69 (1.65%). The rest of the pairwise are less than 1.5%. However, since the required number of consecutive runs was set to 5, the probable adequate run stays at 26. Different settings may result in different required runs. For example, if the requirement is changed to 10 cumulative runs with threshold of 1%, it can only be satisfied at run 91 to 100. Hence, the result is dynamic based on the requirement.

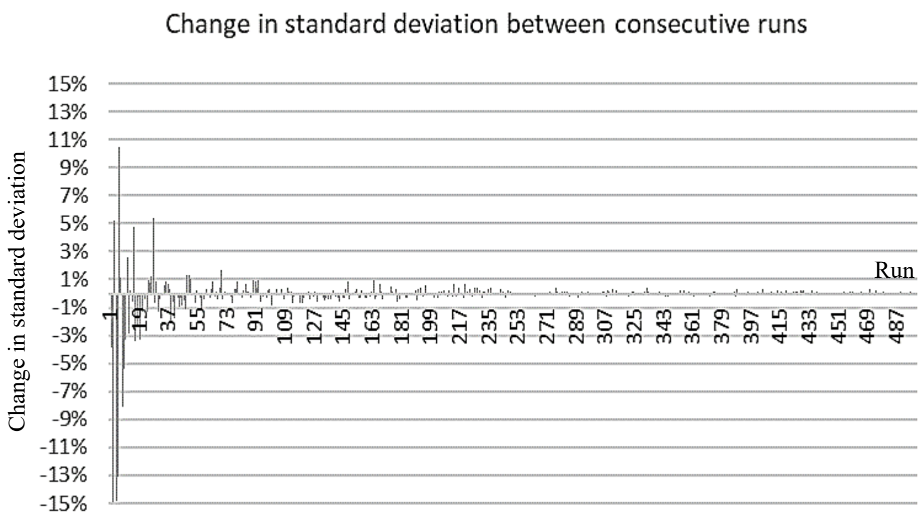

Figure 5: Change in standard deviation between two consecutive runs versus run

Note that the selection of threshold in this study is solely to demonstrate how the framework is working. In the methodology section, is stated that experiment to study the electric field uniformity inside the cell is related to latter experiment to determine the shielding effectiveness (SE) of the cell. The specification of SE itself varies according to industries and frequencies in which the intended products are expected to function. For example, for military application the SE range is 70 to 100 dB (Office of the Assistant Secretary Defence (Supply and Logistic)

1956). Eventhough the set-up of the experiment as defined in the standard has been replaced by IEEE 299.1, the latter does not specify the SE range. For the other products such as medical, the IEC 60601-2-33:2022 (International Electrotechnical Commission 2022b) provides an indicative SE figure which is at least 90 dB while for consumer products IEC 61000-5-7:2008 (International Electrotechnical Commission 2001) states typical enclosure must provide shielding effectiveness in the range of 10 dB to 30 dB. Thus, the selection of the threshold is based on a lot of factors that must be technically studied and developed.

4.4. *Post-hoc test to confirm the required experiment runs*

To confirm that run 26 is the correct point to terminate the repetition, an early-runs and late-runs (ERLR) stability test was used. From the test, results were calculated and presented in Figure 5. Both mean and variance differences between the early runs (mean: 6.2055×10^{-5}, variance: 2.1650×10^{-9}) and late runs (mean: 6.2321×10^{-5}, variance: 2.0672×10^{-9}) are small. Hence no major drift in the data is expected from the means analysis. The analysis returned p-value of 0.8167 which indicated that no statistically significant difference between early and late runs can be supported. Variability has not significantly changed across runs. The p-value using Levene's test is 0.7088, hence no significant difference in the variances.

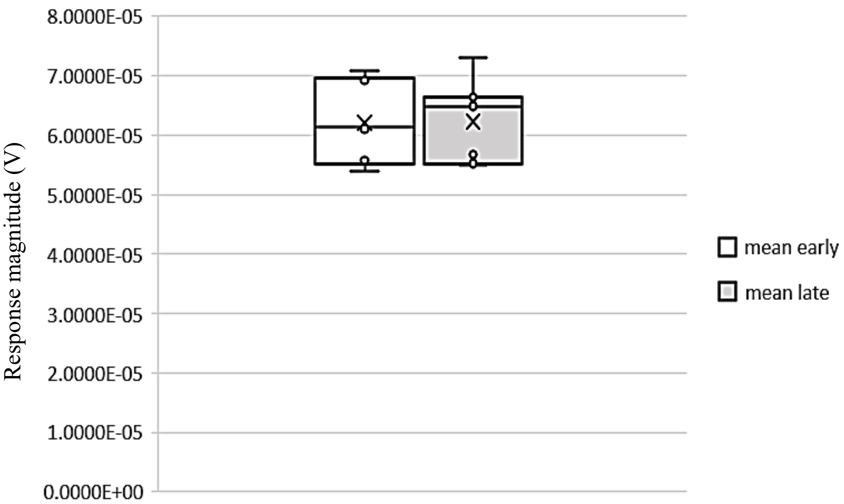

Figure 5: Box plot for early runs mean and late runs mean

5. Conclusion

This study presents a flexible, data-centric approach to determining the requisite number of experimental repetitions, moving beyond conventional fixed-count methodologies. By integrating statistical analyses that respond to the data's inherent properties, the proposed framework ensures that experimental conclusions are both reliable and reproducible. The proposed framework is made up of four distinct steps: 1) testing data normality, 2) statistical differences across multiple runs, 3) determining the required number of repetitions, and 4) confirming the number of repetitions. The application of this methodology to GTEM cell measurements underscores its practical utility and potential for broader adoption across various experimental disciplines. For instance, violation of distributional assumptions in the first step restricted subsequent analysis in step 2 to only non-parametric methods. In addition to that, characteristics of the data rendered several methods insufficient to determine the required number of runs in step 3. Furthermore, human decision is critical in setting the level of accumulation of runs and threshold in step 3. All these observations point to a framework that is data-driven where at each step the characteristics of the data are weighted. As part of the framework, certain decisions, such as the targeted accuracy, must be defined by the experimenters. Thus the framework is also human-centric and dynamic. Future work may explore the integration of this framework with automated systems and AI-driven decision-making processes to further enhance experimental design and efficiency.

References

Antony J. 2023. *Design of Experiments for Engineers and Scientists*. Oxford, UK: Elsevier.

Bajwa A.A., Habib S. & Kamran M. 2015. An investigation into partial discharge pulse extraction methods. *International Journal of Electrical Power & Energy Systems* **73**: 964–982.

Bartnikas R. 1987. A commentary on partial discharge measurement and detection. *IEEE Transactions on Electrical Insulation* **22**(5): 629–653.

Batista G.E.A.P.A., Prati R.C. & Monard M.C. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter* **6**(1): 20–29.

Burghardt F. & Garbe H. 2021. Validation of GTEM cells according to IEC 61000-4-20: Real world field measurements. *IEEE Letters on Electromagnetic Compatibility Practice and Applications* **3**(3): 88–91.

Casella G. & Berger R.L. 2024. *Statistical Inference*. 2nd Ed. Boca Raton, FL: CRC Press.

Curtis M.J., Alexander S.P.H., Cirino G., George C.H., Kendall D.A., Insel P.A., Izzo A.A., Ji Y., Panettieri R.A. & Patel H.H., Sobey C.G., Stanford S.C., Stanle P., Stefanska B., Stephens G.J., Teixeira M.M., Vergnolle N. & Ahluwalia A. 2022. Planning experiments: Updated guidance on experimental design and analysis and their reporting III. *British Journal of Pharmacology* **179**(15): 3907–3913.

Field A. 2024. *Discovering Statistics using IBM SPSS Statistics*. 6th Ed. London, UK: Sage Publications Limited.

Gama J. 2010. *Knowledge Discovery from Data Streams*. Boca Raton, FL: CRC Press.

Ghosh M., Mukhopadhyay N. & Sen P.K. 2011. *Sequential Estimation*. Hoboken, NJ: John Wiley & Sons.

Gundersen O.E. 2021. The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A* **379**(2197): 20200210.

Hall P. & Martin M.A. 1988. On bootstrap resampling and iteration. *Biometrika* **75**(4): 661–671.

Hahn G.J. & Meeker W.Q. 2011. *Statistical Intervals: A Guide for Practitioners (Vol. 92)*. Hoboken, NJ: John Wiley & Sons.

Hollander M., Wolfe D.A. & Chicken E. 2013. *Nonparametric Statistical Methods*. 3rd Ed. Hoboken, NJ: John Wiley & Sons.

IEEE Standard Association. 2013. *IEEE standard method for measuring the shielding effectiveness of enclosures and boxes having all dimensions between 0.1 m and 2 m (IEEE 299.1)*. IEEE.

International Electrotechnical Commission. 2001. *Electromagnetic compatibility (EMC) – Part 5-7: Installation and mitigation guidelines - Degrees of protection provided by enclosures against electromagnetic disturbances (EM code) (IEC 61000-5-7)*. IEC.

International Electrotechnical Commission. 2020. *Electromagnetic compatibility (EMC) – Part 4-3: Testing and Measurement Techniques – Radiated, radio-frequency, electromagnetic field immunity test (IEC 61000-4-3)*. IEC.

International Electrotechnical Commission. 2022a. *Electromagnetic compatibility (EMC) – Part 4-20: Testing and Measurement Techniques – Emission and Immunity Testing in Transverse Electromagnetic (TEM) Waveguides (IEC 61000-4-20)*. IEC.

International Electrotechnical Commission. 2022b. *Medical electrical equipment – Part 2-33: Particular requirements for the basic safety and essential performance of magnetic resonance equipment for medical diagnosis (IEC 60601-2-33:2022)*. IEC.

Judd M.D. & Farish O. 1998. A pulsed GTEM system for UHF sensor calibration. *IEEE Transactions on Instrumentation and Measurement* **47**(4): 875–880.

Judd M., Siegel M. & Coenen S. 2018. UHF PD sensor characterisation using GTEM cells. *Proceedings of the VDE High Voltage Technology 2018; ETG-Symposium*, pp. 1–6.

Linx Technologies. 2018. ANT-LTE-HDP-2000-SMA. Remote Adhesive-Mount LTE Antenna (linxtechnologies.com). Link Technologies.

Mishra P., Pandey C. M., Singh U., Gupta A., Sahu C. & Keshri A. 2019a. Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia* **22**(1): 67–72.

Mishra P., Pandey C. M., Singh U., Keshri A. & Sabaretnam M. 2019b. Selection of appropriate statistical methods for data analysis. *Annals of Cardiac Anaesthesia* **22**(3): 297–301.

Moffat R.J. 1988. Describing the uncertainties in experimental results. *Experimental Thermal and Fluid Science* **1**(1): 3–17.
Montgomery D.C. 2013. *Design and Analysis of Experiments*. 8th Ed. Hoboken, NJ: John Wiley & Sons.
Morovic L., Kritikos M., Sobrino D.R.D., Bilik J., Sobota R. & Kapustova M. 2022. A statistical approach in the analysis of geometrical product specification during the cold tube drawing process. *Applied Sciences* **12**(2): 676.
Nothofer A., Alexander M., Bozec D., Marvin A. & McCormack L. 2003. *The use of GTEM cells for EMC measurements: Measurement Good Practice Guide No. 65*. Teddington, Middlesex, United Kingdom: National Physical Laboratory.
Office of the Assistant Secretary Defence (Supply and Logistic). 1956. *Attenuation measurements for enclosures, electromagnetic shielding, for electronic test purposes, method of (MIL-STD-285)*. U.S. Department of Defence.
Peck R., Short T. & Olsen C. 2020. *Introduction to Statistics and Data Analysis*. 6th Ed. Boston, MA: Cengage Learning.
Peng R.D. 2011. Reproducible research in computational science. *Science* **334**(6060): 1226–1227.
Petitjean A. & Musset O. 2025. Application for determining repeatability and reproducibility in an Honest Gauge R&R study. *Quality and Reliability Engineering International* **41**(5): 1953–1970.
Rajput D., Wang W.J. & Chen C.C. 2023. Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics* **24**(1): 48.
Siegel M., Beltle M., & Tenbohlen S. 2016. Characterization of UHF PD sensors for power transformers using an oil-filled GTEM cell. *IEEE Transactions on Dielectrics and Electrical Insulation* **23**(3): 1580–1588.
Sriram S., Nitin S., Prabhu K.M.M. & Bastiaans M.J. 2005. Signal denoising techniques for partial discharge measurements. *IEEE Transactions on Dielectrics and Electrical Insulation* **12**(6): 1182–1191.
Vaclavik M., Sikorova Z. & Barot T. 2019. Particular analysis of normality of data in applied quantitative research. In Silhavy R., Silhavy P. & Prokopova Z. (eds.). *Computational and Statistical Methods in Intelligent Systems*: 353–365. Cham: Springer.
Wald A. 2004. *Sequential Analysis*. New York, NY: Courier Corporation.
Wang Z., Bridgeford E., Wang S., Vogelstein J.T. & Caffo B. 2025. Statistical analysis of data repeatability measures. *International Statistical Review* **93**(3): 517–543.
Wiggins B.L., Sefi-Cyr H., Lily L.S. & Dahlberg C.L. 2021. Repetition is important to students and their understanding during laboratory courses that include research. *Journal of Microbiology & Biology Education* **22**(2): e00158–21.
Yao Y., Gu J., Li T., Zhang Y., Jia Z., Zhao Q. & Zhang J. 2025. Improved firefly algorithm-optimized ResNet18 for partial discharge pattern recognition within small-sample scenarios. *Processes* **13**(6): 1764.

Faculty of Engineering
Universiti Pertahanan Nasional Malaysia
57000 Sungai Besi
Kuala Lumpur, MALAYSIA
E-mail: azharudin@upnm.edu.my, kamil@upnm.edu.my*

Defence Foundation Center
Universiti Pertahanan Nasional Malaysia
57000 Sungai Besi
Kuala Lumpur, MALAYSIA
E-mail: m.syazwan@upnm.edu.my

Received: 30 June 2025
Accepted: 24 February 2026

*Corresponding author