Journal of Cybersecurity and Privacy

MDPI

Article

Challenging Assumptions of Normality in AES s-Box Configurations under Side-Channel Analysis

Clay Carper [1,*], Stone Olguin [1], Jarek Brown [1], Caylie Charlton [1] and Mike Borowczak [2]

[1] Department of Electrical Engineering and Computer Science, University of Wyoming, Laramie, WY 82071, USA; aolguin1@uwyo.edu (S.O.); jbrow125@uwyo.edu (J.B.); ccharlt1@uwyo.edu (C.C.)
[2] Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816, USA; mike.borowczak@ucf.edu
* Correspondence: ccarper2@uwyo.edu

Abstract: Power-based Side-Channel Analysis (SCA) began with visual-based examinations and has progressed to utilize data-driven statistical analysis. Two distinct classifications of these methods have emerged over the years; those focused on leakage exploitation and those dedicated to leakage detection. This work primarily focuses on a leakage detection-based schema that utilizes Welch's *t*-test, known as Test Vector Leakage Assessment (TVLA). Both classes of methods process collected data using statistical frameworks that result in the successful exfiltration of information via SCA. Often, statistical testing used during analysis requires the assumption that collected power consumption data originates from a normal distribution. To date, this assumption has remained largely uncontested. This work seeks to demonstrate that while past studies have assumed the normality of collected power traces, this assumption should be properly evaluated. In order to evaluate this assumption, an implementation of Tiny-AES-c with nine unique substitution-box (s-box) configurations is conducted using TVLA to guide experimental design. By leveraging the complexity of the AES algorithm, a sufficiently diverse and complex dataset was developed. Under this dataset, statistical tests for normality such as the Shapiro-Wilk test and the Kolmogorov-Smirnov test provide significant evidence to reject the null hypothesis that the power consumption data is normally distributed. To address this observation, existing non-parametric equivalents such as the Wilcoxon Signed-Rank Test and the Kruskal-Wallis Test are discussed in relation to currently used parametric tests such as Welch's *t*-test.

Keywords: test vector leakage assessment; side-channel analysis; AES encryption; side-channel; statistical methods

Citation: Carper, C.; Olguin, S.; Brown, J.; Charlton, C.; Borowczak, M. Challenging Assumptions of Normality in AES s-Box Configurations under Side-Channel Analysis. *J. Cybersecur. Priv.* **2023**, *3*, 844–857. https://doi.org/10.3390/jcp3040038

Received: 14 October 2023
Revised: 22 November 2023
Accepted: 26 November 2023
Published: 29 November 2023

1. Introduction

Computational systems all share at least one component, the consumption of electricity to facilitate computational tasks. When observed, this power consumption can be leveraged to exfiltrate information, leading to a violation of the confidentiality of the victimized system. Two distinct variants, leakage exploitation and leakage detection, serve as pillars of this subfield, referenced as power-based Side-Channel Analysis (SCA). In the twenty-four years since the first presentation by Kocher et al. [1], SCA has experienced a renaissance of development. Novel strategies for power-based cryptanalysis have continued to develop over the years, with targets focusing on both symmetric encryption [1–3] and asymmetric encryption [4] algorithms. Through power usage, it has readily been demonstrated that valid attack vectors may be leveraged against properly instrumented low-power embedded systems running cryptographic algorithms. While the world ushers in the age of quantum-resistant cryptography, low-power embedded systems, such as Internet of Things (IoT) devices, remain an attractive and lucrative target for adversaries to potentially exploit.

In general, SCA seeks to obtain the secret key used for cryptographic operations. For example, the Advanced Encryption Standard (AES) [5] is still widely used in the embedded

systems space. One common method for exploring SCA is to replicate known attacks on AES, which traditionally focus on examining what changes occur during the *SubBytes* step of encryption [5,6]. This step applies a nonlinear substitution to replace the input byte with an output byte corresponding to a set lookup table, commonly referred to as a substitution-box (s-box). It is important to note that this style of attack vector formation is typical when working within the realm of SCA. Valid cryptanalysis often targets a critical component of an algorithm, otherwise known as a subprocess. While both Differential Power Analysis (DPA) [1] and Correlation Power Analysis (CPA) [7] facilitate strong attack vectors, modern tooling has sought to bring more powerful analytics to fruition. Currently, the state-of-the-art techniques focus on Test Vector Leakage Assessment (TVLA), a versatile method used to evaluate whether there is information leakage under SCA; it has been invaluable for researchers in assessing whether power-based leakage constitutes a viable attack vector. TVLA utilizes Welch's *t*-test to determine whether there are meaningful differences in the means between two groups of collected data. When leakage of information has been detected, traditional side-channel attacks can then be leveraged to recover a key from the collected power usage data associated with a cryptographic operation. This collected data is often referred to as a power trace.

Focusing on key recovery has largely relied on statistical tests that assume power trace data is derived from a normal distribution. This assumption is wide-spread, affecting any work utilizing or building upon CPA or TVLA. Such an assumption not only limits the utility of classical SCA, but to a large extent, it may also be inappropriate. This body of work focuses on challenging the assumption of normality in power trace data by using TVLA on nine unique s-box configurations as the guiding framework. Each of these configurations introduce unique nonlinearity in their associated power trace data, giving a reliable, varied dataset. Using this collected data, the assumption of normality is assessed using the Shapiro-Wilk and Kolmogorov-Smirnov tests. From a statistical standpoint, the Wilcoxon Signed-Rank and Kruskal-Wallis tests are presented as potentially statistical alternatives to Welch's *t*-test.

The remainder of this document is as follows; Section 2 outlines the relevant previous work, Section 3 presents the strategies used in this study, Section 4 presents the statistical outcomes in the same order they were discussed previously, Section 5 offers commentary on the results, and finally, Section 6 presents the conclusion and a few possible future directions for this research.

2. Related Works

Power-based side-channel analysis involves determining the relationships between the power consumption of a device and its operations. The most common targets of this analysis are cryptographic hardware devices. Randolph and Diehl [8] mention that the "most basic power side-channel attack" is Simple Power Analysis (SPA), or the direct observation of the power consumption of a device. Visual observation of the power traces associated with a device executing computations can demonstrate when rounds of AES encryption are computed, as demonstrated by Randolph and Diehl [8]. An example of SCA is given in Figure 1a, offering a visual representation of two traces sampled from an AES encryption. While subtle, close visual inspection illustrates small differences between the traces. However, visualizing two collected power traces does not provide an accurate representation of the data or underlying source algorithm. While SPA could lead to insights, or even a possible key extraction, such an exercise would require extensive knowledge of the underlying components of AES and the individual implementation. As a result, the assumption made when performing SPA is that the observer can determine useful leakage information from graphical representations of the power traces.

Despite the clear limitations associated with SPA, plotting the time-series of the power traces can yield useful information regarding the processes on a device, noise, and countermeasures such as masking or hiding [9]. Masking is the process of adding a random "masking value" to intermediate operations on the device to remove the correlation between

the power consumption of the cryptographic device and its secret cryptographic information. Hiding is the process of making power traces gathered using a cryptographic device appear to be random noise. This can be achieved by adding more noise to the operations on the device by performing non-operations (nops) or random process delays [9]. These countermeasures can make SPA difficult to perform, as any notable information from the time-series plots will appear to be random. Ultimately, relying on observations rather than statistical methods can lead to Type II errors, also known as false negatives. Drawing a false negative conclusion commonly occurs when an incorrect assumption leads to concluding that no information was leaked via the power usage of a computational system.

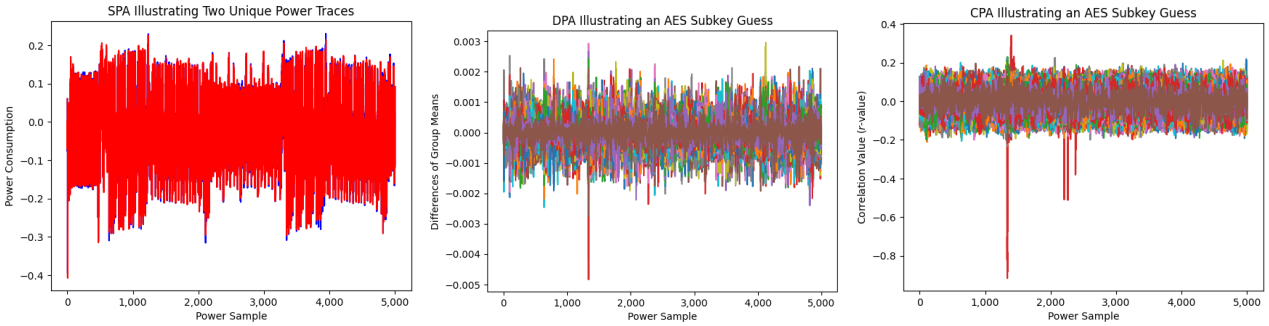
Kocher et al. proposed a new approach to power-based side-channel analysis called Differential Power Analysis (DPA) [1], which offered a novel, powerful method for exploiting power-based leakage. This approach sought to leverage statistical power, which measures the sensitivity or likelihood that a statistical measure detects an effect when it is actually present, to evaluate difficult-to-observe components in traces.

A secret key used by an encryption schema could then be exfiltrated from collected power traces. The statistical process used by Kocher et al. relies on separating the collected power traces into two groups based on whether a target's bit was set to 0 or 1. Within each group, the collection of power traces are then averaged based on each trace number to remove any collection-based noise in the time-series data, resulting in two master traces. While taking the average can be affected by outliers and skew, there are no assumptions made to perform the average. These refined traces lack the typically normally distributed perturbations present in electrically generated data. The denoising of the data was more descriptive than ever, giving a stronger power-based representation of the associated computational behavior.

By adding an elementary element of fundamental statistical analysis by taking the arithmetic mean, DPA allows for more concrete results in determining whether information was leaked. This development also produced new challenges associated with mitigation strategies against power-based side-channel attacks. As demonstrated by Clavier et al. [10], even when random process interrupts are implemented in a device to prevent against information leakage through power-based side-channel analysis, DPA can still be applied to gain information about the device. As such, DPA's usage of a statistic added resilience against countermeasures intended for side-channel analysis. However, it was limited in scope by its reliance on splitting the power traces based on their sensitivity of a bit at a single point in time. For example, a properly instrumented device may have an AES-key fully recovered using DPA. This attack-vector involves gathering an appropriate corpus of data and then examining each of the possible key-guesses for a difference in behavior using DPA. Once this operation is complete, the computed difference is evaluated for the largest difference between the two trace groups. For a visual representation of this process when applied to the AES algorithm, see Figure 1b.

Whereas DPA focuses on the significance of the presence of a single bit at a single fixed time, the next major advancement in side-channel analysis was motivated by the desire to exhaustively examine the contribution of each time step in a power trace. A new form of side-channel analysis, known as Correlation Power Analysis (CPA), was coined by Brier et al. [7]. Taking inspiration from DPA, Pearson's Correlation Coefficient was calculated by utilizing pairs consisting of a Hamming Weight, the number of 1's in a binary representation, relevant to plaintext data and a set of collected power traces. CPA diverges from the DPA's dependence on the mean power traces of two data groups. This marked the first time that side-channel analysis was heading towards more statistically backed conclusions to determine where significant differences occur, improving the exploitation of leakage. To be more precise, Brier et al. [7] concluded that while DPA can fail without knowledge of the underlying implementation of the cryptographic device, CPA can infer information relevant to the implementation. It is able to do all of this while requiring less power trace samples in comparison to DPA. However, Brier et al. also stated that CPA is vulnerable to the same countermeasures as DPA, since both procedures depend upon "side-

channel observability". While CPA was to be a step towards applying statistically focused methods to side-channel analysis, Pearson's Correlation Coefficient should only be used when assuming the normality of the samples, and when there exists a linear relationship between them [11]. Although it was initially used as an attack against AES, Xia et al. [12] demonstrate that CPA can be applied to the lightweight block cipher algorithm LiCi, but the assumptions from utilizing Pearson's Correlation Coefficient are still made, even on a different encryption scheme.



(**a**) (**b**) (**c**)

Figure 1. Visualizations for Simple Power Analysis (SPA), Differential Power Analysis (DPA), and Correlation Power Analysis (CPA) that accompany the methodological discussions in Section 2. (**a**) Sample SPA visualization. (**b**) Sample DPA visualization. (**c**) Sample CPA visualization.

The most recent noteworthy expansion to general power-based side-channel analysis was proposed by Goodwill et al. [13]. This development focused on detecting whether a device has power-based leakage directly in contrast to SPA, DPA, or CPA, all of which focus on exploiting possible side-channels. The novel methodology has since been commonly referred to as Test Vector Leakage Assessment (TVLA). This methodology acts as a successor to other attacks such as CPA while focusing on leakage detection rather than leakage exploitation. Where TVLA steps into new territory is with the separation methodology. To properly apply TVLA, two unique sets of power traces are generated, each of which have a meaningful, data-dependent difference. For example, when assessing whether an AES implementation exhibits power-based leakage, the two groups of data differ based on whether the plaintext provided fits a fixed or random pattern. A NIST specification for TVLA [13] was produced to ensure that the testing parameters are consistent, allowing for a rigid, verifiable procedure to be produced. When collecting samples for TVLA, it is imperative that both groups of data are sampled during a single experimental run and not two separate experiments. Following this procedural step ensures that no selection bias is introduced during data collection. Contrary to CPA, TVLA allows a user to test for differences between the two data groups using a number of statistical methods, such as Pearson's χ^2 test [14] or Welch's t-test [15]. However, utilizing the Pearson's χ^2 test assumes that each group is independent of each other [16], and the Welch's t-test assumes that the data are normally distributed [15]. To ensure that there is high confidence in detecting leakage, Goodwill et al. [13] chose a confidence level of $\alpha = 0.0001$ with its inversely related value, $C = 4.5$. While Goodwill et al. uses the variable C, the traditional usage of Welch's t-test refers to this value as t, representing the computed t-score. A t-value of 4.5 is the associated value with $\alpha = 0.0001$, such that if a computed t-value, $\hat{t}$, satisfies $|\hat{t}| > 4.5$, then there is 99.99% confidence that leakage was detected between the two groups. To extend this to using Pearson's χ^2 test, the p-value of testing the power traces must be less than 0.0001 to have the same confidence level as using Welch's t-test.

As outlined above, this process can be used to test implementations of AES for possible power-based leakage. One simple method to visualize the output from TVLA is given in Figure 2. It is important to note that the various horizontal lines illustrate a few possible $\hat{t}$-values. The overlapping elements are sourced from splitting both groups of power trace data into two equal-sized subsets and then using Welch's t-test to obtain a point-wise

$\hat{t}$-value. The two subsets are then plotted to generate a simplistic visual verification of potential leakage.

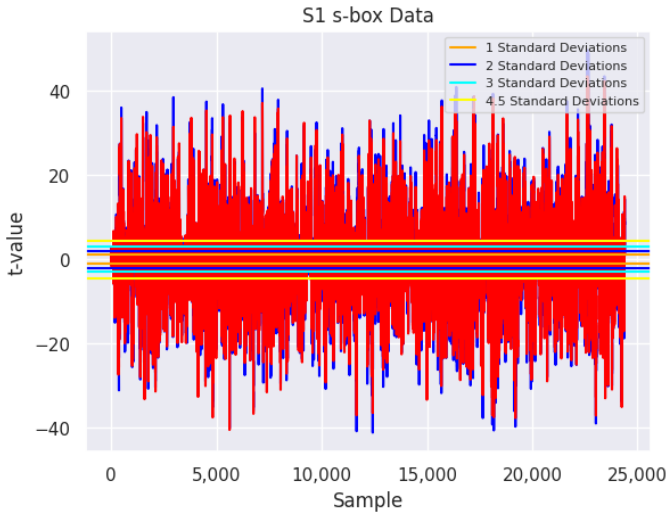


Figure 2. An example TVLA output for fixed and random plaintext outputs when run through Tiny-AES-c with the S1 s-box.

While they are tangential to this work, some alternative statistical methods have been examined. For example, Jaysena et al. [4] demonstrate an application of using the Kullback-Leibler test [17], which evaluates for significant divergence between two groups; this acts as a comparison to Welch's t-test. The assumptions of this test are that the two data groups have the same range of values which are positive, known as the support. The Kullback-Leibler test does not assume normality of the data, although it does have its own assumptions on the support of the data. This application of statistical testing is with regard to Register-Transfer Level (RTL) leakage exploitation-based attacks [18–20]. On the contrary, this work primarily focuses on leakage detection, leading to the exploration of alternate statistical methods.

3. Methods

To adequately address the assumption of normality in side-channel analysis strategies, it is paramount to understand how data was collected. In Section 2, the evolution of side-channel analysis strategies was outlined, providing commentary on how such approaches could be applied to key recovery under AES encryption. Details of any underlying computational device were omitted to reduce the overhead of understanding SCA tooling. Going forward, the ChipWhisperer Lite (CWLite) build system, consisting of a multi-use capture unit and a target board [21], is utilized. An STM32F303 microcontroller occupies the target board, commonly referred to as the device under test (DUT). This device was selected for 10-bit precision with a sampling rate of 105 Mega samples per second (MS/s). These ADC capture capabilities exceed the TVLA sampling requirement of an 8-bit precision on measurements [13]. To accommodate the requirement for the number of samples, n, to be at least greater than 5000 [13], 24,400 samples were gathered for each trace.

A suitable AES implementation is the next requirement that must be satisfied. Tiny AES in C (Tiny-AES-c) is a well-known, widely used implementation of AES in C [22], with full CWLite integration across many DUTs, including the STM23F303. A suite of ChipWhisperer tools, such as the DPA ChipWhisperer tutorial [23], in tandem with Tiny-AES-c provides a framework for data collection. To add variability to this study, nine s-box configurations were utilized. The first of these s-boxes is included in Tiny-AES-C and will be referred to as the *Default* s-box. The remaining eight s-boxes, referenced as S1 through S8, were sourced from Siddiqui et al. [24] and were selected due to their consistent measure of nonlinearity, defined by Hua et al. [25], which is identical to the Default s-box. Maintaining

the same level of nonlinearity reduces the risk of SCA's more powerful statistical methods producing biased results relating to different levels of information-theoretic entropy [26].

For each s-box, an instance of Tiny-AES-C was created. This instance was then used to generate power consumption data based on the NIST TVLA guidelines, with approximately 500 traces being gathered for each of the two groups of data, with a total of 1000 traces being collected. The variance in group sizes originates from the fact that the plaintext generation is pseudorandom. From experimental testing, the number of traces in a group varied by at most 2.5%. Two groups of data are required for TVLA; one with a known plaintext, with the associated power traces being referred to hereafter as the *fixed data*, and the other group consisting of power traces associated with random plaintexts, which will be addressed as *random data*. It is important to note here that the key used with both groups remains fixed throughout the data collection process. Using the CWLite system, twenty sets of fixed and random data were collected for each of the nine s-boxes. Before any further analysis was performed, master traces were computed for each pair of data and compared against their corresponding counterpart. No discernible outliers were detected within each of the twenty collections for a given s-box, signifying no issues with the data collection process. For the remainder of this analysis and without the loss of generality, the thirteenth collection of data for each of the nine s-boxes was used. Unless otherwise noted, all tests were run on the full sets of traces and not their corresponding master traces.

As outlined in Section 2, the first step is establishing a baseline analysis using traditional TVLA. This analysis focuses on applying Welch's t-test, a statistical measure used in the seminal TVLA paper [13]. Welch's t-test evaluates the difference between two groups' means while also assuming that the two groups are observation-independent, contain no significant outliers, and that each group's data are normally distributed [15]. Welch's t-test is applied to the data rather than Student's t-test, since each group of power traces' variances are not assumed to be equal. The null hypothesis for Welch's t-test states that the two groups' difference in means is 0, while the two-tailed alternative hypothesis is that the difference between the means is not 0. A significance level, α, is selected before performing the test and is used to determine whether there is any statistically significant difference between the two groups' means. To follow Goodwill et al.'s chosen significance level, α is chosen to be 0.0001, which has an associated t-statistic of 4.5. The calculation of the samples' t-statistic, $\hat{t}$, is defined as:

$$\hat{t} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}},$$

where $\bar{X}_i$ is the ith sample mean, s_i is the sample standard deviation, and N_i is the sample size. Under Welch's t-test, values satisfying $|\hat{t}| > 4.5$ provide sufficient evidence, with a 99.99% confidence interval, that there is a significant difference between the two data groups. This is consistent with the past experimental testing of AES under TVLA. In order to make critical comparisons to later evaluations, a separate Welch's t-test was run to have a base-line p-value for each of the nine datasets.

Whether a dataset was sampled from a normal distribution can be assessed using the Shapiro-Wilk test [27]. This test was developed by S. Shapiro and M. Wilk to determine the normality of a sample using analysis of variance. If the p-value returned from this test is less than the desired confidence level α, then there is significant evidence to support the claim that the data is not normal at a $(\alpha * 100)\%$ significance level. The Shapiro-Wilk test for normality has an implementation in R, `shapiro.test(x)` [28]. However, `shapiro.test(x)` from the base R package only functions for between 3 and 5000 input variables. To accommodate this restriction, each dataset was separated into sequential subsets consisting of approximately 2400 data points. Approximation is used to slightly vary the number of traces present in the fixed and random groups for each s-box configuration. By splitting the data, there is no need to conduct sufficient random sampling across the entire dataset at one time. The Shapiro-Wilk test calculates a W-statistic as follows:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} \left(x_i \bar{x}\right)^2},$$

where $x_{(i)}$ is the ith order statistic, $\bar{x} = (x_1 + \cdots, +x_n)$, and the coefficient a_i are given by $(a_1, \cdots, a_n) = \frac{m^T V^{-1}}{C}$, where C is a vector norm $C = || \ V^{-1}m \ || = (m^T V^{-1} V^{-1} m)^{1/2}$ for a vector $m = (m_1, \cdots, m_n)^T$.

To ensure that there was no biasing in the subset samplings while testing for normality, the Kolmogorov–Smirnov test [29] from the base R package [28], was utilized. This test accommodates input data regardless of size, and so leveraging this removes the need for processing subsets of the power traces to test for normality. Both the one-sample and two-sample variants of the Kolmogorov-Smirnov test are utilized in this analysis. The two-sample version tests whether a given pair of traces from the fixed and random data originate from the same distribution. The one-sample version tests if a group of trace data comes from the normal distribution. To accommodate the one-sample test, both the fixed and random data are stored in separate columns of a shared R DataFrame, while the two-sample test accepts two separate DataFrames, with each containing either fixed or random data. While both the one-sample and two-sample variants are utilized in this paper, only the one-sample variant is defined due to its relevance in examining the normality of data. The one-sample Kolmogorov-Smirnov test is defined as follows:

$$F_n(x) = \frac{(\text{number of elements in the sample} \leq x)}{n} = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, x]}(X_i),$$

where $1_{(-\infty,x]}(X_i)$ is the indicator function, equal to 1 if $X_i \leq x$ and equal to 0 otherwise. The Kolmogorov-Smirnov statistic for a given cumulative distribution function $F(x)$ is given by:

$$D_n = \sup_x |F_n(x) - F(x)|$$

where $\sup_x$ is the supremum of the set of distances.

To test if two traces are from a distribution that is symmetric about a mean of $\mu = 0$, the Wilcoxon Signed-Rank Test [30] was performed using the base R package's function `wilcox.test()` with a parameter `paired = TRUE` [28]. The `paired` parameter specifies that the time-series components of each power trace are considered when calculating differences. This test determines whether two samples have a statistically different mean and whether there is a pairing between the two sets of data. The Wilcoxon Signed-Rank Test is performed on pairs of power traces from each of the two groups. The results from this test are interpreted to determine whether the mean of each group of power traces is statistically significant from each other.

To round out the data analysis, Kruskal-Wallis [31] was utilized. This test offers a generalization of the Wilcoxon Signed-Rank test by allowing for testing between n samples [32]. An implementation of Kruskal-Wallis is available in the base R stats package [28] via the `kruskal.test()` function. This test offers a measure of how much variance there is in two populations. Due to the large number of measurements in the power traces, two instances of Kruskal-Wallis were run. The first focused on the full power traces for each group, while the second computed the H-statistic for each ten-quantile of each group in a pairwise fashion. The Kruskal-Wallis test was applied to both the full trace data and the master trace data for each s-box.

4. Results

When applying traditional TVLA, significant data leakage was detected in all nine s-box configurations. From Welch's t-test applied in R, the computed values are summarized in Figure 3. A t-value having a magnitude greater than 4.5 is equivalent to a p-value of less than 0.0001.

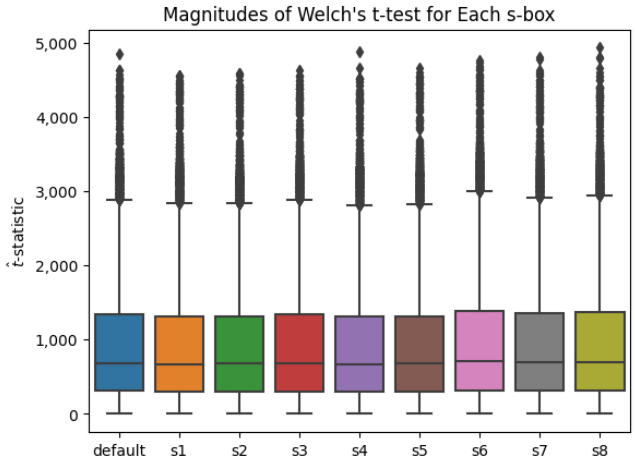


Figure 3. An example t-test output showing the $\hat{t}$-values based on the magnitudes of their values.

To test for normality, the Shapiro-Wilk test was run on each of the nine data groups. Across all nine configurations, the largest p-value reported across all traces was 1.443×10^{-31}. The one-sample Kolmogorov-Smirnov test produced a maximal p-value of 2×10^{-16} for each point in each power trace, indicating that the null hypothesis should be rejected, meaning that the data is not normally distributed. The two-sample variant of the test reported high p-values of at least 0.9, so that it cannot be said that the two data sources originate from different distributions. The approximation of p-values is due to a limitation in the implementation of these tests in base R.

In order to evaluate the dataset for variance, the Wilcoxon Signed-Rank test was utilized, with the results summarized in Figure 4. The results of the Wilcoxon Signed-Rank test results in a W-statistic. The W-statistic behaves similarly to the t-statistic in that it is inversely related to a p-value calculation based on the statistic. That is, a higher W-statistic is associated with a lower p-value, inviting the conclusion that there are significant differences when our p-value is less than the chosen $\alpha = 0.0001$ significance level.

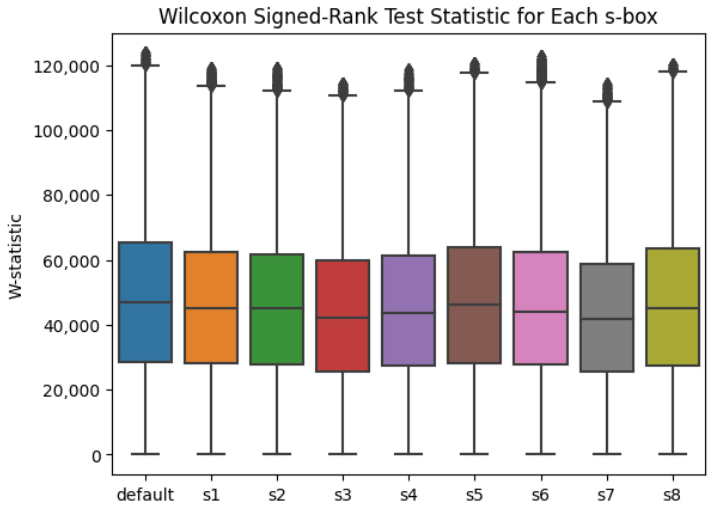


Figure 4. Boxplot describing the W-values calculated by the Wilcoxon Signed-Rank test.

Finally, the results of applying the Kruskal-Wallis test, focusing on the variance within the data, is provided in Table 1.

Table 1. The p-values calculated using the Kruskal-Wallis test, associated with each of the nine s-boxes. Values that are in bold are statistically significant for $\alpha = 0.0001$.

Data	**K-W p-Value**
(a) Default s-box	
Full Default	0.9178411
Quantile 1	0.7973211
Quantile 2	0.9492303
Quantile 3	0.4865965
Quantile 4	0.003551987
Quantile 5	0.4060664
Quantile 6	0.2482653
Quantile 7	0.03989401
Quantile 8	$\mathbf{4.704231 \times 10^{-5}}$
Quantile 9	$\mathbf{1.787921 \times 10^{-28}}$
Quantile 10	0.0002557376
(b) S1 s-box	
Full S1	0.8889263
Quantile 1	0.9047302
Quantile 2	0.9721271
Quantile 3	0.5588611
Quantile 4	0.01427206
Quantile 5	0.5762426
Quantile 6	0.1448012
Quantile 7	0.1547279
Quantile 8	$\mathbf{1.198919 \times 10^{-7}}$
Quantile 9	$\mathbf{6.674018 \times 10^{-24}}$
Quantile 10	$\mathbf{6.324433 \times 10^{-6}}$
(c) S2 s-box	
Full S2	0.915456
Quantile 1	0.8438535
Quantile 2	0.9775268
Quantile 3	0.5287456
Quantile 4	0.01392253
Quantile 5	0.5347051
Quantile 6	0.1436789
Quantile 7	0.1304291
Quantile 8	$\mathbf{1.721812 \times 10^{-7}}$
Quantile 9	$\mathbf{3.478914 \times 10^{-25}}$
Quantile 10	$\mathbf{5.773031 \times 10^{-6}}$
(d) S3 s-box	
Full S3	0.9500724
Quantile 1	0.9117693
Quantile 2	0.9705669
Quantile 3	0.7504782
Quantile 4	0.04621217
Quantile 5	0.9294591
Quantile 6	0.03665044
Quantile 7	0.8746878
Quantile 8	$\mathbf{1.247497 \times 10^{-10}}$
Quantile 9	$\mathbf{3.693989 \times 10^{-11}}$
Quantile 10	$\mathbf{2.67089 \times 10^{-11}}$

Table 1. *Cont.*

Data	K-W *p*-Value
(e) S4 s-box	
Full S4	0.6587587
Quantile 1	0.8206295
Quantile 2	0.9661761
Quantile 3	0.6459478
Quantile 4	0.01331219
Quantile 5	0.6543537
Quantile 6	0.140855
Quantile 7	0.1583064
Quantile 8	$\mathbf{1.341678 \times 10^{-7}}$
Quantile 9	$\mathbf{1.222115 \times 10^{-23}}$
Quantile 10	$\mathbf{7.782813 \times 10^{-6}}$
(f) S5 s-box	
Full S5	0.8426999
Quantile 1	0.8273909
Quantile 2	0.9961258
Quantile 3	0.6122431
Quantile 4	0.01023039
Quantile 5	0.7160496
Quantile 6	0.1471896
Quantile 7	0.1200313
Quantile 8	$\mathbf{1.044246 \times 10^{-7}}$
Quantile 9	$\mathbf{3.630053 \times 10^{-24}}$
Quantile 10	$\mathbf{5.205475 \times 10^{-6}}$
(g) S6 s-box	
Full S6	0.7960942
Quantile 1	0.8861468
Quantile 2	0.9273349
Quantile 3	0.5546987
Quantile 4	0.00573446
Quantile 5	0.3858158
Quantile 6	0.1886492
Quantile 7	0.03089781
Quantile 8	$\mathbf{3.049533 \times 10^{-5}}$
Quantile 9	$\mathbf{3.701519 \times 10^{-27}}$
Quantile 10	0.0003635786
(h) S7 s-box	
Full S7	0.7562423
Quantile 1	0.836804
Quantile 2	0.8870617
Quantile 3	0.671789
Quantile 4	0.05251282
Quantile 5	0.923315
Quantile 6	0.05537985
Quantile 7	0.8434102
Quantile 8	$\mathbf{1.351392 \times 10^{-10}}$
Quantile 9	$\mathbf{3.106407 \times 10^{-11}}$
Quantile 10	$\mathbf{4.581774 \times 10^{-11}}$

Table 1. *Cont.*

Data	K-W p-Value
(i) S8 s-box	
Full S8	0.629495
Quantile 1	0.838456
Quantile 2	0.9678031
Quantile 3	0.6064738
Quantile 4	0.010838
Quantile 5	0.7013192
Quantile 6	0.158997
Quantile 7	0.1423515
Quantile 8	$\mathbf{7.853093 \times 10^{-8}}$
Quantile 9	$\mathbf{4.059782 \times 10^{-24}}$
Quantile 10	$\mathbf{3.801423 \times 10^{-6}}$

5. Discussion

With regard to traditional TVLA testing, it is clear that the application of Welch's t-test is intended as a method for establishing differences between two groups of data with some amount of statistical certainty. Cases where the magnitude of a $\hat{t}$-value is greater than 4.5, represented by the blue line in Figure 3, have traditionally formed the basis of measuring leakage in power-based side-channel analysis. For all nine s-box configurations, significant power-based leakage is measured, forming the basis of a known, readily verifiable result. This result will serve as a baseline for comparison later in this section.

The primary focus of this body of work has been to establish an experimental set-up with sufficient complexity to determine whether the assumption of normality of power trace data associated with AES is well-founded. Due to how Shapiro-Wilk was applied to each dataset, the subsetting could lead to a biased outcome. To combat potential issues from random sampling, the Kolmogorov-Smirnov test was utilized as a secondary test for normality. Examining the results of both the Shapiro-Wilk and the one-sample Kolmogorov-Smirnov tests, there was sufficient evidence to reject the null hypothesis with a 99.99% level of confidence. This indicates that each of the nine datasets are unlikely to be normally distributed. The null hypothesis for a two-sample Kolmogorov-Smirnov states that the two given sets of data come from the same distribution. The results of the two-sample tests indicate that there is no significant evidence to reject the null hypothesis for each of the nine AES datasets. Therefore, the two groups of experimental data for each s-box are sourced from the same distribution with a 99.99% confidence. Under the TVLA specifications, this conclusion is reasonable; it is expected that data samples collected from a single device with an identical firmware share similarities.

It is important to examine different statistical tools that do not require assumptions about the distribution of the data, referred to as non-parametric methods. A non-parametric alternative to Welch's t-test is the Wilcoxon Signed-Rank test. Examining Figure 4, there are similar trends to those in Figure 3. Despite Welch's t-test and the Wilcoxon Signed-Rank test computing a $\hat{t}$-statistic and W-statistic respectively, both give an equivalent, comparable p-value. Thus, if a p-value associated with a calculated W-statistic is less than $\alpha = 0.0001$, the significance between the means is equivalent to having the magnitude of a $\hat{t}$-statistic as greater than or equal to 4.5. This evenly balanced range of values is reminiscent of the sinusoidal nature of power trace data, and warrants further study in power-based SCA.

Additional exploration was conducted using the Kruskal-Wallis test, a non-parametric method for examining the variance between two populations. Two variants were tested, the first of which focused on examining each group of traces for an s-box configuration as a whole. The results here are neither surprising nor particularly interesting. It is expected there is little variance when two traces are compared as singular entities; there are components of AES that operate independently of the s-box configuration. The overall power trace should reflect such a property. This variant of applying Kruskal-Wallis confirms

that there is no statistically significant difference between traces from the two data groups for a given s-box configuration when examined as a whole. However, the second test, which focused on comparing pairs of ten-quantiles, had a different result. As seen in Table 1, quantiles 8, 9, and 10, except for the Default and S6 s-box configurations, show statistically significant differences with respect to the variance between the fixed and random data. The differing behavior in the Default and S6 s-box configurations could be attributed to noise in either the fixed or random data samples, or it could indicate a possible issue with the underlying data sampling process. Further testing is required to verify which, if any, of these are the cause.This suggests that when the plaintext is randomly varied, there is a measurable effect within subsets of the data, which is consistent with previous applications of parametric statistics in SCA.

6. Conclusions

Under the lens of examining data sampled using traditional TVLA, power trace data associated with AES s-box configurations most likely do not originate from a normal distribution. Classical SCA strategies rely on an assumption of normality, as has been made evident by the use of parametric statistical tests (i.e., Welch's *t*-test). This assumption, based on this study, may be a severe limitation to power-based side-channel analysis. While non-parametric statistical methods have been widely available, they have remained untapped and underused in SCA to this point. As the field continues to evolve and move towards more complex and varied analysis, equivalent non-parametric approaches should be examined for viability in both leakage exploitation and leakage detection related to power-based side-channel analysis. One natural extension of this work lies in applying parametric and non-parametric statistical methods to a larger set of s-box configurations with varying nonlinearity measures to establish a baseline of performance for each class of tests. Another possible direction would encompass an analysis centered on different microprocessor architectures.

Author Contributions: Conceptualization, C.C. (Clay Carper) and S.O.; Methodology, C.C. (Clay Carper) and J.B.; Software, C.C. (Clay Carper) and J.B.; Validation, C.C. (Clay Carper), S.O. and J.B.; Formal analysis, C.C. (Clay Carper) and J.B.; Investigation, C.C. (Clay Carper) and J.B.; Resources, C.C. (Clay Carper) and S.O.; Data curation, C.C. (Clay Carper) and C.C. (Caylie Charlton); Writing—original draft, C.C. (Clay Carper) and S.O.; Writing—review & editing, C.C. (Clay Carper), S.O., J.B., C.C. (Caylie Charlton) and M.B.; Visualization, C.C. (Clay Carper) and S.O.; Supervision, C.C. (Clay Carper) and M.B.; Project administration, C.C. (Clay Carper) and M.B.; Funding acquisition, M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in FigShare at 10.6084/m9.figshare.24650373.

Acknowledgments: The research team would like to acknowledge and thank the Secure Systems Collaborative for their assistance in the revisions and presentation of this information.

Conflicts of Interest: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any agency, sponsor, or corporate entity.

References

1. Kocher, P.; Jaffe, J.; Jun, B. Differential power analysis. In Proceedings of the Annual International Cryptology Conference, Santa Barbara, CA, USA, 15–19 August 1999; Springer: Berlin/Heidelberg, Germany, 1999; pp. 388–397.
2. Soares, R.; Lima, V.; Lellis, R.; Finkenauer, P., Jr.; Camargo, V. Hardware Countermeasures against Power Analysis Attacks: A Survey from Past to Present. *J. Integr. Circuits Syst.* **2021**, *16*, 1–12. [CrossRef]

3. Wang, X.; Zheng, J.; Wu, L.; Zhu, J.; Hu, W. A Correlation Fault Attack on Rotating S-Box Masking AES. In Proceedings of the 2021 Asian Hardware Oriented Security and Trust Symposium (AsianHOST), Shanghai, China, 16–18 December 2021; pp. 1–6. [CrossRef]
4. Jayasena, A.; Andrews, E.; Mishra, P. TVLA*: Test Vector Leakage Assessment on Hardware Implementations of Asymmetric Cryptography Algorithms. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2023**, *31*, 1269–1279. [CrossRef]
5. National Institute of Standards and Technology. Advanced Encryption Standard. *NIST FIPS PUB 197*. 2001. Available online: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.197.pdf (accessed on 29 September 2023)
6. Daemen, J.; Rijmen, V. *The Design of Rijndael*; Information Security and Cryptography; Springer: Berlin/Heidelberg, Germany, 2002. [CrossRef]
7. Brier, E.; Clavier, C.; Olivier, F. Correlation Power Analysis with a Leakage Model. In *Cryptographic Hardware and Embedded Systems-CHES 2004*; Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., et al., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; Volume 3156, pp. 16–29.
8. Randolph, M.; Diehl, W. Power side-channel attack analysis: A review of 20 years of study for the layman. *Cryptography* **2020**, *4*, 15. [CrossRef]
9. Mangard, S.; Oswald, E.; Popp, T. *Power Analysis Attacks*; Springer: Boston, MA, USA, 2007.
10. Clavier, C.; Coron, J.S.; Dabbous, N. Differential power analysis in the presence of hardware countermeasures. In Proceedings of the International Workshop on Cryptographic Hardware and Embedded Systems, Worcester, MA, USA, 17–18 August 2000; Springer: Berlin/Heidelberg, Germany, 2000; pp. 252–263.
11. Kirch, W. (Ed.) Pearson's Correlation Coefficient. In *Encyclopedia of Public Health*; Springer: Dordrecht, The Netherlands, 2008; pp. 1090–1091. [CrossRef]
12. Xia, X.; Chen, B.; Zhong, W. Correlation Power Analysis of Lightweight Block Cipher Algorithm LiCi. *J. Phys. Conf. Ser.* **2021**, *1972*, 012055. [CrossRef]
13. Goodwill, G.; Jun, B.; Jaffe, J.; Rohatgi, P. *A Testing Methodology for Side-Channel Resistance Validation*; Cryptography Research Inc.: San Francisco, CA, USA, 2011; p. 15.
14. Pearson, K.X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1900**, *50*, 157–175. [CrossRef]
15. Welch, B.L. The Generalization of 'Student's' Problem When Several Different Population Variances are Involved. *Biometrika* **1947**, *34*, 28–35. [CrossRef] [PubMed]
16. McHugh, M.L. The Chi-square test of independence. *Biochem. Medica* **2013**, *23*, 143–149. [CrossRef] [PubMed]
17. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
18. He, M.; Park, J.; Nahiyan, A.; Vassilev, A.; Jin, Y.; Tehranipoor, M. RTL-PSC: Automated power side-channel leakage assessment at register-transfer level. In Proceedings of the 2019 IEEE 37th VLSI Test Symposium (VTS), Monterey, CA, USA, 23–25 April 2019; pp. 1–6.
19. Pundir, N.; Park, J.; Farahmandi, F.; Tehranipoor, M. Power Side-Channel Leakage Assessment Framework at Register-Transfer Level. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2022**, *30*, 1207–1218. [CrossRef]
20. Zhang, T.; Park, J.; Tehranipoor, M.; Farahmandi, F. PSC-TG: RTL Power Side-Channel Leakage Assessment with Test Pattern Generation. In Proceedings of the 2021 58th ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 5–9 December 2021; pp. 709–714. [CrossRef]
21. NewAE Technology. CW1173: Chipwhisperer-Lite. 2018. Available online: https://media.newae.com/datasheets/NAE-CW1173_datasheet.pdf (accessed on 29 September 2023).
22. kokke. tiny-AES-c: A Small Portable AES128/192/256 in C. Available online: https://github.com/kokke/tiny-AES-c (accessed on 29 September 2023).
23. Inc., NewAE Technology. DPA on Firmware Implementation of AES. Available online: https://github.com/newaetech/chipwhisperer-jupyter (accessed on 29 September 2023).
24. Siddiqui, N.; Yousaf, F.; Murtaza, F.; Ehatisham-ul Haq, M.; Ashraf, M.U.; Alghamdi, A.M.; Alfakeeh, A.S. A highly nonlinear substitution-box (S-box) design using action of modular group on a projective line over a finite field. *PLoS ONE* **2020**, *15*, e0241890. [CrossRef] [PubMed]
25. Hua, Z.; Li, J.; Chen, Y.; Yi, S. Design and application of an S-box using complete Latin square. *Nonlinear Dyn.* **2021**, *104*, 807–825. [CrossRef]
26. Heys, H.M. A tutorial on linear and differential cryptanalysis. *Cryptologia* **2002**, *26*, 189–221. [CrossRef]
27. Shapiro, S.S.; Wilk, M.B. An analysis of variance test for normality (complete samples). *Biometrika* **1965**, *52*, 591–611. [CrossRef]
28. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
29. Conover, W. *Chapter 6: Statistics of the Kolmogorov-Smirnov Type. Practical Nonparametric Statistics*; John Wiley & Sons: New York, NY, USA, 1971.
30. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80. [CrossRef]

31. Kruskal, W.H.; Wallis, W.A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621. [CrossRef]
32. Forthofer, R.N.; Lee, E.S.; Hernandez, M. 9-Nonparametric Tests. In *Biostatistics*, 2nd ed.; Forthofer, R.N., Lee, E.S., Hernandez, M., Eds.; Academic Press: San Diego, CA, USA, 2007; pp. 249–268. [CrossRef]