AN EMPIRICAL ANALYSIS FOR STUDY OF INFLUENCE FACTORS TO EFFECTS ON P-VALUES AND TESTS CRITERIA

Ishtiaq Ahmed[1], Dr. Talat Sharafat Rehmani[2] and Irfan Ali[3]

ABSTRACT

This paper aims to identify the key statistical factors that influence the threshold and interpretation of p-value significance levels. Our empirical analysis provides evidence to better understand the range of high significance, moderate significance, and non-significance levels. We thoroughly discuss several contributing factors, including sample size (large vs. small datasets), standard error, the quality and utility of statistical tests, statistical power, effect size, sampling procedures, data collection quality, and the use of both qualitative and quantitative research data. Consideration is also given to challenges presented by big data. Additionally, we review contributions from prominent researchers regarding factors that influence p-values. To support our analysis, we generate sets of random sample data (n = 10, 100, 500, 1000, 5000, and 10000) from well-known continuous and discrete distributions, along with observed secondary data, to examine the behavior of p-values under different conditions. This study also investigates the question: "Is p-value a reliable measure of unknown population characteristics?" We conclude with findings, recommendations, and possible remedies based on our empirical results. All analyses were conducted using the Statistical Package for Social Sciences (SPSS, version 24).

Keywords: *P-value, Influence factors, Test Statistics, Confidence Interval, Statistical Significance, Sample size, Standard Error, Good Point Estimators, Processes of Testing of Hypothesis, and Statistical Decision Making.*

1. INTRODUCTION

P-value plays a central role in statistical inference for researchers, clinical investigators, and scientists across various disciplines. It is widely used in hypothesis testing and is often viewed as a key metric for making decisions about population parameters. A sound p-value, however, is not an isolated outcome—it is the result of a rigorous research protocol and the appropriate application of statistical data analysis techniques.

[1]Senior Associate Professor, Bahria University, Karachi. Email: ishtiaqahmed.bukc@bahria.edu.pk
[2]Senior Assistant Professor, Bahria University, Karachi. Email: talat.bukc@bahria.edu.pk
[3]Ajjunct Professor, Bahria University, Karachi. Email: irfanso@yahoo.com

Researchers frequently rely on p-values to draw conclusions from their data, often attaching great significance to whether a result is "statistically significant" (typically, $p \leq 0.05$) or "not significant" ($p > 0.05$). However, such binary interpretations can be misleading and oversimplified. The reliance on p-values alone, without understanding their nuances, can result in flawed conclusions. Indeed, inferential decision-making should not be based solely on p-values.

This raises several important questions:

- Is p-value the only statistical metric that informs decision-making regarding unknown population parameters?
- Why has $p \leq 0.05$ become the conventional threshold for rejecting the null hypothesis?
- Can other statistical factors improve the interpret-ability and usefulness of p-values?

This paper seeks to explore the characteristics and influencing factors that affect p-values. Several key variables can impact the p-value, including:

- Effect size
- Sample size
- Margin of error
- Standard error
- Interval estimates
- Test statistics
- Statistical power
- Good Point Estimators
- Etc.

A common misconception in research is that a p-value ≤ 0.05 automatically implies a meaningful or practically significant result, while a p-value > 0.05 suggests no effect or difference. This binary view is problematic. As noted by Neham F. S. (Koreen J. Pain, 2017), *"P > 0.05 only means no evidence of difference. It does not mean evidence of no difference."* This distinction is critical: a non-significant result ($p > 0.05$) may be due to insufficient statistical power, small sample size, poor study design, or other methodological flaws. Similarly, a significant p-value ($p < 0.05$) does not validate a hypothesis or imply real-world importance.

P-values alone cannot confirm the correctness of a research argument. A statistically significant result does not guarantee scientific validity, and over reliance on p-values can distort decision-making and the interpretation of

findings. It is imperative to consider the broader context—including confidence intervals, effect sizes, and study design—when interpreting p-values.

1.1 Effect Size

Effect Size is a magnitude or strength of relationship between two variable and the difference between two groups. In statistical analysis, the effect size can be measured by several mathematical approaches:

1) Standardize mean difference
2) Odd ratio
3) Correlation Coefficient
4) Risk Ratio
5) Hazard Ratio
6) Type of statistical test being conducted

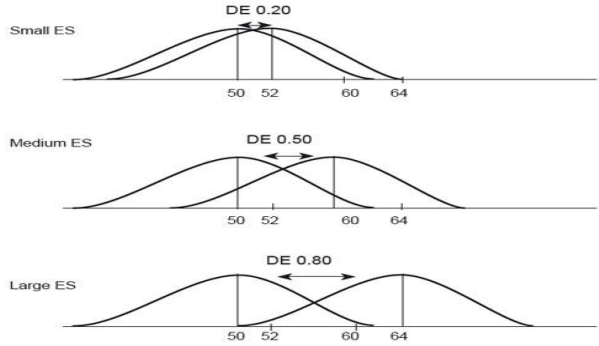
Effect Size can be measured in reference with the following formula

Eta Square = d = $\frac{(test\ statistic\ value)^2}{(n-1)+(test\ statistic\ value)^2}$

Cohan J. (1988), Statistical Power Analysis for behavioral Sciences NY: Routledge Academic summarized the range of Effect Size and Eta Square values 'd' as under:

Eta Square Value (d)	Effect Size
0.01	Small
0.06	Moderate
0.14≥	Large

Large effect size suggests stronger relationship or large difference between variables. Small effects size effect indicates a weaker relationship. Theoretical laws show that the relationship between effects size is directly proportional with p-value. This relationship is to be illustrated numerically in the statistical analysis section.



Reference: http://psychologyinrussia.com/volumes/pdf/2015_3/psychology_2015_3_3.pdf

1.2 Sample Size

It is technically an influence factor on p-value with increased or decreased samples size. The relationship of p-value is inversely proportional, because small sample size increases standard error and test statistic will be decreased, while large sample size reduces the standard error and test statistic will be increased. This is why it is important to choose an appropriate sample size when conducting hypothesis tests to ensure accurate and reliable results.

1.3 Test Statistic

Test statistic is the ratio between margin of error and standard error. Large computed test statistic impact positively on p-value. Mathematically, it can be computed by $\frac{observed-Hypothesied}{Statndard\ Error}$. It is very core benchmark statistical value to make decision about size of probability of p-value. A general proven result is test statistic is inversely proportional to the size of p-value and help statisticians to understand Type-I and Type-II statistical error. Several clinical trials make better decision in Type-I statistical error.

1.4 A Good Point Estimators

This is a necessary characteristic for a good point estimator. Parametric statistical analysis requires good point estimator for healthy, transparent, and acceptable decision making for unknown population. Computing of p-value relates maximum dependency at good point estimator. Non-biasedness, Consistency, Efficiency, and Sufficiency are four relative properties to identify a good point estimator. Margin of error fully based on characteristic of good point estimator and margin of error technically and statistically suggests suitable sample size determination for the quality of research activities. Commonly, method of moments and maximum likelihood are generally practiced to find out or suggest to be using a good point estimator.

1.5 Interval Estimate of Parameter

Interval estimate of parameter is a mathematical statement and range of unknown population parameters under given confidence level such as 95%, 99% or 96% etc. Mathematically, it can be stated as under:

(1 – α)% C. I for Population Mean (μ) = Pr. [$\mu_L \leq \mu \leq \mu_U$] or
(1 – α)% C. I for Population Proportion (p) = Pr. [$p_L \leq p \leq p_U$]
etc.

1.6 Types Of Statistical Error

Types of statistical error can be defined into two significant types such as type-I and type-II statistical errors. The Type-I statistical error explains about asset of researchers or data analyst. The Type-I statistical error is one good outcome and result for research investigators. In this error rejection region is going to reduce as compared to assumed level of significance. While Type-II statistical error may be increased rejection region. This paper also analyzes to critically observe what is the impact of sample size at types of statistical error?

1.7 Power of Statistical Test

Power of statistical test depend on p-value. In the case of type-I error power of statistical test will be increased and in the case of type-II error power of statistical test will be reduced at given confidence level. The power of statistical test provide performance of manufacture process of fast moving consumer goods (FMCG). Therefore statistics hypothesis testing is an important tools for quality production management.

1.8 The Central Limit Theorem

The central limit theorem emphasis and significantly supports that is sample size increases then a normal probability distribution approaches to approximately normal probability distribution regarding less shape of the distribution. According to the Central Limit Theorem *"The relation between population distribution and sampling distribution is called the Central Limit Theorem. In this relationship the shape of sampling distribution approaches to the approximate normal distribution, when sample size gradually increases. At large enough sample size will demonstrate the very close shape of approximate normal distribution of sampling distribution as shape of population distribution."*

1.9 The Null, Alternative Hypothesis and P-Value Approach

For the p-value approach the likelihood (*p*-value) of the numerical value of the test statistic is compared to the specified significance level (α) of the hypothesis test. The *p*-value corresponds to the probability of observing sample data at least as extreme as the actually obtained test statistic. Small *p*-values provide evidence against the null hypothesis. The smaller (closer to 0) the *p*-value, the stronger is the evidence against the null hypothesis. If the *p*-value is less than or equal to the specified significance level α, the null hypothesis is rejected. Otherwise, the null hypothesis is not rejected. If $p \leq \alpha$, reject H_0 otherwise, if $p > \alpha$, do not reject H_0. In consequence, by knowing the *p*-value at any desired level of significance may be assessed. For example, if the *p*-value of a hypothesis test is 0.01, the null hypothesis can be rejected at any significance level larger than or equal to 0.01. It is not rejected at any significance level smaller than 0.01. Thus, the *p*-value is commonly used to evaluate the strength of the evidence against the null hypothesis without reference to significance level. The following table provides guidelines for using the *p*-value to assess the evidence against the null hypothesis (Weiss, 2011):

P – Value	Evidence against H_o
p – value > 0.01	Weak or no evidence
0.05 < p-value ≤ 0.10	Moderate evidence
0.01 < p-value ≤ 0.05	Strong evidence
p-value ≤ 0.01	Very strong evidence

In this study, we will examine, review and investigate the significant features of numerical evidence to find out above highlighted factors' impact on p-value. It will also be suggested to make better infrastructure for desire statistical decision making with respect to p-value.

1.10 Effects of Outliers

Outliers in any data-set pose a significant challenge for statistical inference. The presence of outliers can lead to technical and interpretative problems, as they may inflate the standard error and distort estimates of central tendency and variability. This can result in misleading conclusions and reduced reliability of inferential outcomes.

Therefore, identifying and appropriately handling outliers is essential for statistical validity and compliance with analytical assumptions. In many cases, outliers originate from data collection errors, such as recording mistakes, instrument faults, or sampling anomalies. Mathematically it can be written as

Lower Outlier: Q1 - 1.5(Interquartile Range) i.e Interquartile Range=$Q_3 - Q_1$
Upper Outlier: Q3 + 1.5(Interquartile Range)

The Illustration of box-plots portrays a better picture of outliers, skewed types, and normality of data standardization.

2. LITERATURE REVIEW

Badenes R. L. et al. (2016) quoted clearly that "The "effect size" fallacy involves the belief that the p-value provides direct information about the effect magnitude (Gliner et al., 2001). In this way, the researchers believe that when p is smaller, the effect sizes are larger. Instead, the effect size can only be determined by directly estimating its value with the appropriate statistic and its confidence interval (*Cohen, 1994; Cumming, 2012; Kline, 2013*)".

0C0oncato J. et al. (2016) mentioned that "P ≤ 0.05 is often misunderstood as a rigid threshold, sometimes with a misguided 'win' ($p \leq 0.05$) or 'lose' ($p > 0.05$) approach. Also, in contemporary genomics studies, a threshold of $p \leq 10-8$ has been endorsed as a boundary for statistical significance when analyzing numerous genetic comparisons for each participant. A value of $p \leq 0.05$, or other thresholds".

Dahiru T. (2008) mentioned "while medical journals are florid of statement such as: "statistical significant", "unlikely due to chance", "not significant," "due to chance", or notations such as, "P > 0.05", "P < 0.05", the decision on whether to decide a test of hypothesis is significant or not based on P-value has generated an intense debate among statisticians". He also advocated "p < 0.05 (5%) significance as a standard level for concluding that there is evidence against the hypothesis tested, though not as an absolute rule. If p-values is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it's below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts".

Demidenko E. (2016) mentioned characteristics and relationship between null hypothesis and sample size that "The little-known fact among non-statisticians is that, with a large enough sample, n, the null hypothesis will always be rejected. This fact stems from the consistency of the test: With the sample size increasing to infinity, the power of the test approaches 1 even for alternatives very close to the null. In other words, with a large enough sample, the null hypothesis will always be rejected regardless of the Type I error, α. What kind of knowledge does statistical hypothesis testing give?, if it leads to only one answer, "Reject the null hypothesis".

Gelman A. and Carlin J. (1917) have argued that "the distinction between practical and statistical significance does not resolve the difficulties

with p-values. The problem is not so much with large samples and tiny but precisely-measured effects but rather with the opposite: large effect-size estimates that are hopelessly contaminated with noise. Consider an estimate of 30 with standard error 10, of an underlying effect that cannot realistically be much larger than 1. In this case the estimate is statistically significant and also practically significant but is essentially entirely the product of noise. This problem is central to the recent replication crisis in science (see Button et al., 2013, and Loken and Gelman, 2017) but is not at all touched by concerns of practical significance".

Greenland S. et.al (2016) indicate that "It is true that the smaller the P value, the more unusual the data would be if every single assumption were correct; but a very small P value does not tell us which assumption is incorrect. For example, the P value may be very small because the targeted hypothesis is false; but it may instead (or in addition) be very small because the study protocols were violated, or because it was selected for presentation based on its small size. Conversely, a large P value indicates only that the data are not unusual under the model, but does not imply that the model or any aspect of it (such as the targeted hypothesis) is correct; it may instead (or in addition) be large because (again) the study protocols were violated, o_r because it was selected for presentation based on its large size".

Kim J. and Bang H. (2016) mentioned that "There are two ways to view a statistical hypothesis test: one is through a p-value (of the test) and the other is through a CI (of a parameter). Many busy clinicians use a simple rule, "If $p < 0.05$ or the CI does not cover the null value, H_0 is rejected." in practice. The p-value and CI are complementary while attempting to do the same/similar thing, where the p-value quantifies how 'significant' the association/difference is, while the CI quantifies how 'precise' the estimation is and what the plausible values are".

Kwak S. (2023) mentioned that "In research, a topic to be identified is selected and hypotheses are established accordingly. In order to calculate the evidence to support this, related data is collected, and the collected data is analyzed using a statistical hypothesis test method suitable for the hypothesis. The method of statistical hypothesis testing is determined according to the type of data corresponding to whether the data is a quantitative variable or a qualitative variable, the research design and hypothesis, etc., but in the end, the hypothesis test is performed using the significance probability value calculated as a result of statistical analysis".

Leo D. G and Sardanelli F. (2020) mentioned that according to Loannidis, "moving the p-value threshold from .05 to .005 will shift about one-third of the statistically significant results of past biomedical literature to the

category of just suggestive". "We think that such a solution makes biomedical research harder and that, adopting this solution, an improvement in research quality is not granted. Lowering this way the p value threshold for significance is, at best, a palliative solution. Especially in clinical research, future trials would need to be larger, less feasible, and more expensive. Achieving 80% power with a threshold of 0.005, instead of 0.05, would require a 70% larger sample size for between subject study designs with two-sided tests (88% for one sided tests)".

Wasserstein R. L. et al. (2019) introduced that "Replace the 0.05 "statistical significance" threshold for claims of novel discoveries with a 0.005 threshold and refer to p-values between 0.05 and 0.005 as suggestive".

Some other eminent researchers have shown earlier in their research study in detail and findings about interpretation, misconception and contemporary understanding of p-values in their earlier research work, namely Hubbard R. et al. (2008), Lin M. et al. (2013), Kim J. et al. (2016), Marasini D. et al. (2016), Lin M. et al. (2013), Ghose A et al. (2011), and Nahm F. S. (2017), Gordon L. et al. (2010), Kline, R. B. (2013), Kwak S. (2023).

3. DATA ANALYSIS

In this section, we present an empirical analysis of quantitative data to explore the relationship between sample size, test values, and their corresponding p-values. The analysis aims to investigate how p-values respond to varying sample sizes and test conditions, which is crucial for understanding the robustness of statistical inference. To achieve this, a series of statistical tables and graphs have been constructed to provide evidence supporting the central theme of this study. Specifically, test values of $\mu = 40$, 45, and 50 were analyzed across multiple sample sizes: n = 10, 100, 500, 1,000, 2,000, 5,000, and 10,000. The empirical results demonstrate how these variations affect the structure and behavior of p-values. The calculated values are in good agreement with the base statistical theory known as a bigger sample size corresponding to a lower p-value, whereas having a constant effect size in this estimate. This serves to verify the purpose of the study by demonstrating quantitatively the necessity of the sample size in the hypothesis testing.

The link between effect size (d), sample size (n), and the resulting p-values is shown in Tables 1, 2, and 3. These tables' numerical figures demonstrate a clear correlation between the two, with the p-value falling as the effect size rises. Likewise, the p-value and the effect size needed for significance tend to decline with increasing sample size (n).

These results provide empirical evidence that increasing the sample size enhances the sensitivity of hypothesis testing. In particular, larger sample

sizes improve the ability to detect small effects and reduce the likelihood of Type-1 errors when proper thresholds are maintained. This aligns with the preferences of many applied researchers, particularly in industrial and scientific contexts, where the reliability of statistical decisions is critical.

In summary, the construction of the test statistic, along with the roles of effect size (e) and sample size (n), plays a crucial part in shaping the p-value. Smaller p-values—achieved through appropriate sample sizing—support robust hypothesis testing and the identification of statistically significant outcomes.

Table-1 tells us in detail that sample size provide strong evidences to conclude sample size is an important factors to influence positively on sample means($\underline{x}$), standard deviation s, standard error $s_{\underline{x}}$, sampling error (e), Test Statistic, Confidence Interval of population parameters, and p-values. Our outcomes of data analysis of this paper p-value speedily decreasing under the sample size increases significantly, such as at n=10, 100, 500, 1000, 2000, 5000, and 10000 the p-value is 1.8%, 0.1%, 0.01%, 0.001%, 0.0001%, 0.00001%, and 0.000001 respectively. It is clearly observed that p-value significantly inspired at sample size (n). This outcome also supports Central Limited Theorem and theoretical concepts. If we take test value $= \mu = 40$ margin of error or sampling error emerge higher as compared with $n = 45$ and $n = 50$, it can be over-viewed in Table-2 and Table-3 critically. In all three cases at computed sample means $\underline{x}$ at prescribed sample sizes (n) are same numerical values i.e. 56.90 (at n = 10), 47,57 (at n = 100), 50.17 (n = 500), 50.23 (n =1,000), 49.65 (n = 2,000), 50.17 (n = 5,000), and 50.14 (n = 10,000).

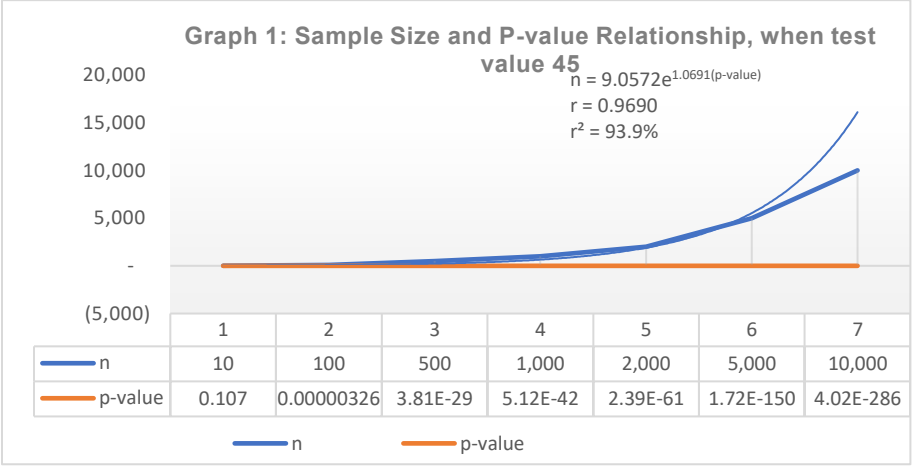
Table no.1: Effect Size (d), Standard Error ($S_{\underline{x}}$) Test Statistic, Confidence Interval if n increases at Test Statistics, $\mu = 40$

Test Value = $\mu = 40$							
Effect Size (d)	n	$\underline{x}$	s	$S_{\underline{x}}$	Test Statistic	95 % Confidence Interval	p-value
0.479254	10	56.90	18.568	5.872	2.878	$3.62 \leq \mu \leq 30.18$	0.018
0.174712	100	47.57	16.443	1.653	4.578	$4.29 \leq \mu \leq 10.85$	0.001
0.250871	500	50.17	17.595	0.787	12.927	$8.63 \leq \mu \leq 11.72$	0.0001
0.252337	1000	50.23	17.619	0.557	18.362	$9.14 \leq \mu \leq 11.32$	0.00001
0.232237	2000	49.65	17.556	0.393	24.590	$8.88 \leq \mu \leq 10.42$	0.000001
0.250416	5000	50.17	17.597	0.249	40.866	$9.68 \leq \mu \leq 10.66$	0.0000001
0.251597	10000	50.14	17.496	0.175	57.978	$9.80 \leq \mu \leq 10.44$	0.00000001

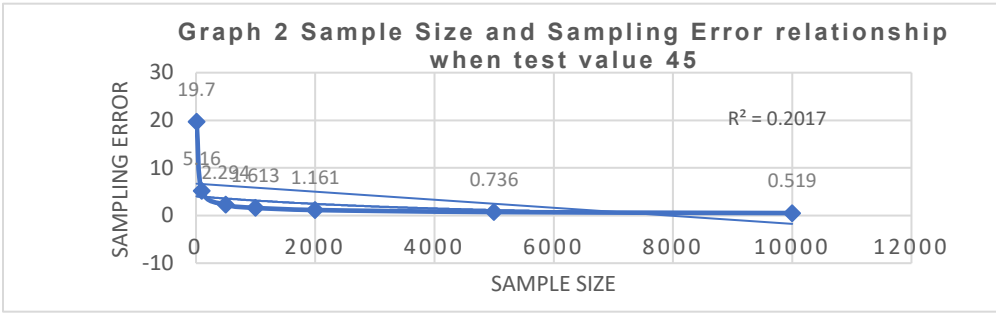
Table no.2: Effect Size (d), Standard Error ($S_{\underline{x}}$) Test Statistic, Confidence Interval if n increases at Test Statistics, $\mu = 45$

Test Value = $\mu = 45$							
Effect Size (d)	n	$\underline{x}$	s	$s_{\underline{x}}$	Test Statistic	95 % Confidence Interval	p-value
0.313435	10	56.90	18.57	5.87	2.03	$-1.38 \leq \mu \leq 25.18$	0.073
0.023782	100	47.57	16.44	1.65	1.55	$-0.71 \leq \mu \leq 5.85$	0.124
0.079683	500	50.17	17.60	0.79	6.57	$3.63 \leq \mu \leq 6.72$	0.0001
0.081071	1000	50.23	17.62	0.56	9.39	$4.14 \leq \mu \leq 6.32$	0.00001
0.065677	2000	49.65	17.56	0.39	11.85	$3.88 \leq \mu \leq 5.42$	0.000001
0.079476	5000	50.17	17.60	0.25	20.78	$4.68 \leq \mu \leq 5.66$	0.0000001
0.079567	10000	50.14	17.50	0.18	29.40	$4.80 \leq \mu \leq 5.49$	0.00000001

Graph-1 shows that association of sample size and p-value make an exponential curve with large sample size and p-value will approach to negligible. Association(r) and coefficient of determination (r^2) among sample size and p-value are r=0.9690 and 93.9%, which are recognized statistically strong relationship among them. This results help us to recommend sample size is a robust factor to reduce p-value gradually.



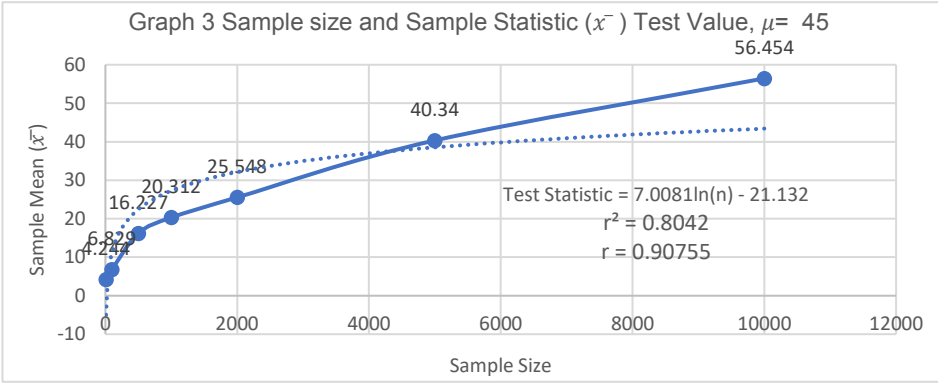
Another significant observation is noted from the data analysis that effect size of all three tables remain unchanged at prescribed same sample size and test values $\mu = 40, 45, \& \ 50$. The standard error ($s_{\underline{x}}$) and sampling error (e) is also decreasing with reference to sample size(n) and p-values, while population standard deviation s gradually approaches constant around a fixed numerical value at increasing sample sizes.

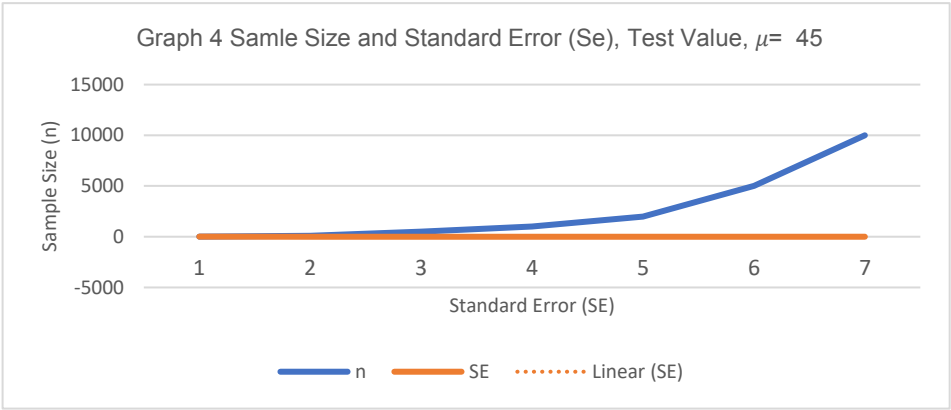


The test statistic value increases at prescribed sample sizes. The test statistic is directly proportion with sample size categorically in three cases. The test values $\mu = 40, 45, \& 50$ do not impact dramatically to change in effect size, population standard deviation, standard error, test statistic, sample means, sampling error and p-values.

Table no.3: Effect Size (d), Standard Error ($S_{\underline{x}}$) Test Statistic, Confidence Interval if n increases at Test Statistics, $\mu = 50$

Test Value, $\mu =$ 50							
Effect Size (d)	n	$\underline{x}$	s	$s_{\underline{x}}$	Test Statistic	95 %Confidence Interval	p-value
0.133983	10	56.90	18.57	5.87	1.18	$-6.38 \leq \mu \leq 20.18$	0.270
0.021361	100	47.57	16.44	1.65	-1.47	$-5.71 \leq \mu \leq 0.85$	0.144
9.7E-05	500	50.17	17.59	0.79	0.22	$-1.37 \leq \mu \leq 1.72$	0.827
0.000177	1000	50.23	17.62	0.56	0.42	$-0.86 \leq \mu \leq 1.32$	0.679
0.000387	2000	49.65	17.56	0.39	-0.88	$-1.12 \leq \mu \leq 0.42$	0.378
9.25E-05	5000	50.17	17.60	0.25	0.68	$-0.32 \leq \mu \leq 0.66$	0.494
6.72E-05	10000	50.14	17.50	0.18	0.82	$-0.20 \leq \mu \leq 0.49$	0.411





Similarly, we can clearly observe that Graph 3, and Graph 4 show the behavior among simple size & sample mean ($\underline{x}$), and sample size and Standard Error when test statistic $\mu = 45$

Most researchers aim to draw inferences about the mathematical characteristics of parameters from large and often unknown populations. Since studying an entire population is typically impractical, researchers rely on samples. Therefore, it is essential that these samples are properly representative and free from bias to ensure the validity of the conclusions. This issue must be addressed during the research design phase to enhance the quality of statistical decision-making. A well-designed sampling strategy leads to more accurate and reliable statistical conclusions, particularly when evaluating whether an outcome is statistically significant or not. Such conclusions should be based on appropriate threshold probability values (e.g., $\alpha = 0.05$), supported by strong numerical evidence.

Problems often arise in tests of statistical significance because researchers typically work with samples, not entire populations. Since conclusions are generalized from sample data to the broader population, it is crucial that the sample be representative. If the sample is biased, it can lead to incorrect or misleading results.

To ensure valid inferences, the sample must accurately reflect the population's characteristics. This is particularly important in fields such as economics, social sciences, and biomedical research, where decisions are often based on probabilistic reasoning.

In most scientific disciplines, a result is considered statistically significant if it meets a confidence level of 95% ($p < 0.05$) or, in more rigorous cases, 99% ($p < 0.01$). These thresholds help determine whether observed effects are likely due to chance or represent real population-level patterns.

3.1 Is P-Value A Better Statistical Measure For Statistical Tests?

P-values by themselves are insufficient for sound statistical judgment. Even though the p-value is frequently employed in hypothesis testing to evaluate statistical significance, it is neither the sole nor always the most trustworthy metric for determining unknown population parameters. P-values have several drawbacks, including their susceptibility to sample size and their incapacity to accurately represent the magnitude or practical significance of an impact. Therefore, in order to fully comprehend their data and guarantee more trustworthy results, researchers had to take into account complementary statistical tools. In statistical analysis, the following are some significant substitutes and supplements to the p-value:

- Confidence Intervals (CI)
- Effect Size
- Bayesian Statistics
- Likelihood Ratios
- Akaike Information Criterion (AIC) / Bayesian Information Criterion (BIC)
- Decision Trees / Classification and Regression Trees (CART)
- ROC Curves and AUC (Area Under Curve)
- Power Analysis
- Descriptive Statistics (Mean, Median, Variance, etc.)

3.2 Why P-Value ≤ 0.05 Is A Threshold Value To Reject The Null Hypothesis?

The commonly used threshold of $p \leq 0.05$ in hypothesis testing is a convention, not a strict scientific rule. It is widely accepted across many disciplines as a benchmark for statistical significance, implying that there is a 5% probability (or less) that the observed results could have occurred under the assumption that the null hypothesis is true. However, this threshold should not be treated as a "magic number." Relying solely on it can lead to misinterpretation and oversimplification of statistical evidence. Instead, the p-value should be considered in the broader context of effect size, study design, replication, and practical significance. It includes some historical work as under:

1. Historical Origins:
   - The 0.05 threshold was popularized by Ronald Fisher in the 1920s. He suggested it as a convenient cutoff in his early work on statistical significance.
   - It struck a balance between being too lenient (e.g., 0.10) and too strict (e.g., 0.01), making it a practical choice for early statisticians. Definition of the p-value:

- The p-value measures the probability of observing data at least as extreme as the sample, assuming the null hypothesis is true.
- A $p \leq 0.05$ means there is a 5% or lower chance that the observed result (or one more extreme) would occur if the null hypothesis were true.

2. Error Rates:
   - Using 0.05 as the significance level (α) implies that researchers accept a 5% chance of a Type I error—rejecting a true null hypothesis.
   - It helps control the false-positive rate in scientific research.
3. Tradition and Standardization:
   - Over time, 0.05 became a widely accepted standard, making it easier for researchers to interpret and compare results across studies.
4. Important Caveats:
   - Arbitrary: There is nothing inherently special about 0.05. In many fields, other thresholds (like 0.01 or 0.10) are used depending on the context.
   - Context Matters: In high-stakes testing (like medicine or aerospace), more stringent thresholds are often used.
   - Misinterpretation: A p-value ≤ 0.05 does *not* prove the alternative hypothesis is true or to that the effect is meaningful—it just suggests that the observed data is unlikely under the null hypothesis.

The critical point is not necessarily to change the chosen cutoff of $p \leq 0.05$—as there is no universally better alternative for most contexts. Rather, the key is for readers to recognize that 0.05 is an arbitrary threshold, and more importantly, to look beyond p-values when evaluating the validity of an experiment and the biological or practical significance of the results. It is often more informative to report the exact p-value rather than simply stating $p \leq 0.05$. For example, a result with $p = 0.049$ is roughly three times more likely to have occurred by chance than one with $p = 0.016$, yet both are commonly reported as $p \leq 0.05$. This oversimplification can mask important differences in statistical significance and reliability of findings.

Moreover, merely noting outcomes like $p \leq 0.05$ does not give sufficient context regarding the extent of the effect or the accuracy of the estimations. When analyzing the results, it is more instructive to incorporate contextual elements (such sample size and study design), effect sizes, and confidence ranges. More solid and significant findings concerning the research are produced by this more comprehensive technique. Instead of depending only on the p-value threshold, which frequently results in inaccurate conclusions, this change encourages researchers to submit comprehensive statistical information.

3.3 Can Some Statistical Factors Better, Can Improve Results For P-Values?

There are various statistical techniques that can enhance the reliability and clarity of p-values in hypothesis testing. These approaches can address prevalent issues like multiple comparisons, limited sample sizes, and misunderstanding of statistical significance, resulting in more robust research findings. Below is a polished version of our paragraph, accompanied by an explanation of essential strategies.

Various statistical techniques can improve the reliability and clarity of p-values in hypothesis testing. These techniques seek to tackle prevalent problems such as multiple comparisons, limited sample sizes, and the incorrect interpretation of statistical significance. Implementing these techniques enables researchers to make more precise statistical choices and reach trustworthy, valid conclusions for both unknown and specified populations. Below are some essential methods to enhance the understanding of p-values and statistical decision-making:

- Adjusting for Multiple Comparisons
- Addressing Small Sample Sizes
- Reporting Effect Size and Confidence Intervals
- Utilizing Bayesian Methods
- Improving Data Quality

By adopting these techniques, researchers can improve the reliability and clarity of p-values, resulting in more dependable statistical evaluations and well-informed decision-making. This methodology guarantees that conclusions are not just statistically sound but also significant and pertinent to the context, thereby enhancing the quality and influence of the research.

3.4 Limitation of P-Value

A major drawback of utilizing p-values is that they do not convey any details regarding the practical significance or real-life relevance of the findings. They solely indicate whether the results are statistically significant, which may not accurately represent the magnitude or importance of the effect observed. To evaluate the quality and credibility of an analysis, it is crucial to also take into account additional factors, including sample size, effect size, confidence intervals, and possible sources of error or bias.

Additionally, p-values are frequently misinterpreted or misapplied by both researchers and readers, resulting in erroneous or misleading inferences. Common misunderstandings include:

- The p-value represents the likelihood that the null hypothesis is accurate.
- This is a crucial misconception—p-values indicate the probability of obtaining the observed data, assuming the null hypothesis is true, not the likelihood that the null hypothesis itself is valid.
- The p-value indicates the magnitude of the effect or the strength of the association.
- P-values do not give any insight into the size of the observed effect. To assess this, researchers should utilize measures of effect size. The p-value is a measure of the explicit-ability or reliability of the results.
- The p-value does not indicate the extent to which the data can elucidate or replicate the results, nor does it offer any indication of the experiment's reliability. The p-value reflects the importance or relevance of the findings.
- The p-value fails to consider the real-world consequences of the findings. It serves to highlight the significance or relevance of the findings. However, a statistically significant outcome does not inherently imply that the effect is of practical importance or relevance.

These misunderstandings occur because the p-value is determined exclusively by the data and the statistical test applied, without taking into account any prior knowledge or beliefs regarding the hypotheses. Additionally, it does not indicate the direction or size of the effect, nor does it assess the clarity or significance within the study's context. This highlights the limitations of p-values and underscores the necessity of incorporating alternative statistical metrics and contextual considerations for a more thorough interpretation of research findings. We are interested in delving deeper into these misconceptions or exploring potential strategies to mitigate them in research methodologies.

4. CONCLUSION

P-Values serve as a valuable instrument for comparing outcomes across various studies and variables, allowing analysts to determine which results are statistically significant. In disciplines such as medicine, economics, and engineering, p-values are crucial for making informed decisions based on data, assisting researchers in steering clear of conclusions derived from anecdotal or biased information. The p-value measures the probability of observing the given data (or more extreme results) assuming the null hypothesis holds true. By establishing a standardized threshold (commonly set at 0.05), p-values provide a definitive method for deciding whether to reject

the null hypothesis. This standardization fosters consistency across studies, enabling researchers to make trustworthy comparisons and draw conclusions across diverse research contexts and fields.

Multiple critical elements affect the p-value and its effectiveness in enhancing the quality of data analysis and decision-making. It is crucial to comprehend and control these elements to achieve precise, dependable, and significant statistical results. These include:

1. **The size of the sample (n):** It has a direct impact on the accuracy of the estimate. A larger sample size can identify even minor, potentially insignificant effects and produce a statistically significant p-value, whereas a smaller sample may overlook important differences..
2. **Effect Size**: The p-value is affected by the strength of the effect or association under examination. More substantial effects tend to produce smaller p-values, whereas minor effects may fail to achieve significance unless the sample size is adequately large..
3. **Variance in Data**: High variability or noise within the data can elevate the standard error and distort the p-value, thereby complicating the identification of a genuine effect.
4. **Significance Level (α)**: The selected significance threshold, typically set at 0.05, establishes the criterion for determining when results are deemed statistically significant. It is important to note that this threshold is arbitrary and should be adjusted according to the context of the study and the implications of potential decision errors..
5. **The design and methodology of a study:** These are crucial; inadequate or biased designs can result in systematic errors that compromise the validity of p-values. Effective randomization, blinding, and controlling for confounding variables are vital for drawing accurate conclusions.
6. **Multiple Comparisons**: When multiple hypotheses are tested simultaneously, the chance of a false positive (Type I error) increases. Without proper adjustments (e.g., Bonferroni or FDR correction), researchers risk drawing incorrect conclusions.
7. **Data Cleaning and Pre-processing**: Outliers, missing values, and data entry errors can distort results. Careful pre-processing ensures that p-values reflect genuine patterns in the data, not artifacts.
8. **Underlying Assumptions of the Test**: P-values are only valid if the assumptions of the statistical test (e.g., normality, homoscedasticity, independence) are met. Violations can lead to incorrect inferences.

Despite its widespread use, the p-value is often misinterpreted. A p-value does not:

- Represent the probability that the null hypothesis is true.
- Indicate the size or importance of an effect.
- Reflect the reliability of the data or the study design.
- Account for bias or confounding.

Rather, the p-value is the probability of observing data as extreme as (or more extreme than) the actual data, assuming the null hypothesis is true. It measures the compatibility of the observed data with the null hypothesis—not its truth. Moreover, there is a common tendency to oversimplify p-values into binary outcomes of "significant" or "not significant" based on whether they cross the 0.05 threshold. This practice can be highly misleading, especially when:

- Small but clinically trivial effects appear significant due to large sample size.
- Large and meaningful effects are dismissed due to underpowered studies.

There is also the temptation for "p-hacking" or fishing expeditions, where researchers test numerous variables until something meets the $p < 0.05$ criterion. Without correction for multiple testing, the likelihood of false discoveries increases sharply.

In conclusion, while p-values remain a useful tool, their interpretation must be contextualized with other statistical measures (such as confidence intervals and effect sizes) and guided by strong study design and methodological rigor. Sole reliance on p-values—especially without understanding their assumptions and limitations—can lead to misleading conclusions and poor scientific practices.

Similarly above factors (discussed in Introduction section) are involved to effect poor or strong achievement of p-values. Therefore, researchers and investigators cannot ignore or missing them during their research design processes.

REFERENCES

Badenes, R. L., Frias, N. D., Lotti, B., Bonilla, C. A., & Longobardi, C. (2016). Misconceptions of the P-value among Chilean and Italian academic psychologists. Frontiers in Psychology, 7, 1247.

Barnett, M. L., & Mathisen, A. (1997). Tyranny of the p-value: The conflict between statistical significance and common sense (Editorial). Journal of Dental Research, 76(2), 152–154.

Clarue, G. M., & Kempson, R. E. (1997)**.** Introduction to the design and analysis of experiments. Arnold.

Concato, J., & Hartigan, J. A. (2016). P-values: From suggestion to superstition. Journal of Investigative Medicine, 64(11), 1166–1170.

Dahiru, T. (2008). P-value, a true test of statistical significance? A cautionary note. Annals of Ibadan Postgraduate Medicine, 6(1), 21–26.

Demidenko, E. (2016). The p-value you can't buy. The American Statistician, 7(1), 33–38.

Feinstein, A. R. (1998). P-values and confidence intervals: Two sides of the same unsatisfactory coin. Journal of Clinical Epidemiology, 51(3), 355–360.

Gelman, A., & Carlin, J. (2017). Some natural solutions to the *p*-value communication problem—and why they won't work. Journal of the American Statistical Association, 112(519), 899–901.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. **G.** (2016). Statistical tests, P-values, confidence intervals, and power: A guide to misinterpretations. European Journal of Epidemiology, 31(4), 337–350.

Hubbard, R., & Lindsay, R. M. (2008). Why P-values are not a useful measure of evidence in statistical significance testing. Theory & Psychology, 18(1), 69–88.

Kim, J., & Bang, H. (2016). Three common misuses of P-values. Dental Hypotheses, 7(3), 73–80.

Kwak, S. (2023). Are only *p*-values less than 0.05 significant? A *p*-value greater than 0.05 is also significant! Journal of Lipid and Atherosclerosis, 12(2), 89–95.

Lin, M., Lucas, H. C., & Shmueli, G. (2013). Too big to fail: Large samples and the *p*-value problem. Information Systems Research, 24(4), 906–917.

Leo, D. G., & Sardanelli, F. (2020). Statistical significance: *p* value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. European Radiology Experimental, 4(1), 2–8.

Marasini, D., Quatto, P., & Ripamonti, E. (2016). The use of *p*-values in applied research: Interpretation and new trends. Statistica, 76(4), 315–325.

Nahm, F. S. (2017). What the P-values really tell us. Korean Journal of Pain, 30(4), 241–242.

Ramp, W. K., & Yancey, J. M**.** (1991). P-values and their problems. Bone and Mineral, 13, 163–165.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. The American Statistician, 70(2), 129-133.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". The American Statistician, 73(Suppl. 1), 1–19.