Cavalier Use of Inferential Statistics Is a Major Source of False and Irreproducible Scientific Findings

Leonid Hanin

Department of Mathematics and Statistics, Idaho State University, 921 S. 8th Avenue, Stop 8085, Pocatello, ID 83209-8085, USA; hanin@isu.edu; Tel.: +1-208-282-3293

Abstract: I uncover previously underappreciated systematic sources of false and irreproducible results in natural, biomedical and social sciences that are rooted in statistical methodology. They include the inevitably occurring deviations from basic assumptions behind statistical analyses and the use of various approximations. I show through a number of examples that (a) arbitrarily small deviations from distributional homogeneity can lead to arbitrarily large deviations in the outcomes of statistical analyses; (b) samples of random size may violate the Law of Large Numbers and thus are generally unsuitable for conventional statistical inference; (c) the same is true, in particular, when random sample size and observations are stochastically dependent; and (d) the use of the Gaussian approximation based on the Central Limit Theorem has dramatic implications for \( p \)-values and statistical significance essentially making pursuit of small significance levels and \( p \)-values for a fixed sample size meaningless. The latter is proven rigorously in the case of one-sided Z test. This article could serve as a cautionary guidance to scientists and practitioners employing statistical methods in their work.

Keywords: central limit theorem; distributional homogeneity; law of large numbers; probability metric; \( p \)-value; random sample size; reproducibility; statistical significance

1. Introduction

The last quarter century has witnessed an ever-growing alarm over the generation of false and irreproducible results in experimental sciences [1–4], commonly referred to as the reproducibility (or replication) crisis. The crisis has been discussed from many different angles [5–7], and a vast majority of scientists agree that it is very real [6]. Since the publication of a pioneering work [8] the most common explanation for the reproducibility crisis has been bias. It can result from fraudulent misrepresentation of research findings; conscious or unwitting manipulation of experimental design and data analysis procedures including selection of data subsets, explanatory variables, endpoints, data collection stopping times and statistical methods that produce more favorable outcomes; and selective publication of the most impressive results. Among the many social factors that made such a widespread bias possible are pervasive “publish or perish” mentality, warped incentives and performance criteria for scientists employed in academia, industry and the government, and the dominant climate in professional societies and funding agencies [9]. Reproducibility of research findings is also adversely affected by insufficient standardization of research protocols and inadequate reporting of experimental procedures [10].

On the statistical methods side, the reproducibility crisis was largely ascribed to widespread misuse, overuse, inappropriate dichotomization and misinterpretation of \( p \)-values; the use of null hypotheses beyond their intended purpose; failure to adequately account for multiple hypothesis testing; and misinterpretation of confidence intervals [11–17]. This largely came about as a result of uncritical adoption of Ronald Fisher’s views on hypothesis testing, statistical significance and \( p \)-values and their facile synthesis with an...
alternative philosophy of hypothesis testing developed by Jerzy Neyman and Egon Pearson. A historical background of how and why this happened is vividly described in [18]. Another contributing factor is the almost complete banishing of the history of statistics from the textbooks and classroom lectures.

Reproducibility of a research finding and its truth value are interrelated in a complex and subject-area-specific manner. In physics with its universal laws that have precise mathematical form, reproducibility of experimental results, broadly construed, is an empirical criterion of truth. In biomedical and social sciences, where no such laws have ever been discovered, the degree of separation between reproducibility and the validity of scientific results is much greater than in physical sciences. A typical research finding in biomedical and social sciences is a quantitative statement about a class of organisms or a human population including their responses to a defined condition or intervention. Because of substantial inter- and intra-subject variation, these statements are often formulated in statistical terms. Generation of a scientific result includes selection of a target population; sampling from this population; planning and conduction of an experiment or data collection; methods of statistical data analysis; and interpretation of the results. All these components contribute to the resulting reproducibility of research outcomes, but separation of their inputs is a daunting task. As a result, a reproducible finding may not necessarily be true; however, a finding that fails reproduction or replication under identical experimental conditions is most likely false. An additional factor operative in social sciences is the subjects’ beliefs and information available to them, which dilutes the concept of objective truth and exacerbates the epistemological divergence between reproducibility and validity of scientific results.

In this article, I identify a number of previously underappreciated systematic sources of false and irreproducible scientific results that are rooted in statistical methodology and are independent of fraud, bias, conflicts of interest, misplaced incentives, outdated educational practices and other human factors. These sources include ignoring deviations from basic assumptions underlying inferential statistical analyses (most notably, the random sample assumption) and the use of various approximations. The most common among the latter is the normal (Gaussian) approximation. Scientists commonly recognize, of course, that these assumptions cannot be exactly met in the real world and that approximations generate errors. However, a widely held belief is that their effects on the outcomes of statistical analyses are negligible. I show that this belief is unfounded. Specifically, I demonstrate that (a) arbitrarily small deviations from the random sample assumption can have arbitrarily large effects on the outcomes of statistical analyses; (b) the commonly occurring observations with random sample size may violate the Law of Large Numbers (LLN), which makes them unsuitable for conventional statistical inference; (c) the same is true when the sample size and the observations are stochastically dependent; and (d) the use of the Gaussian approximation through the Central Limit Theorem (CLT) may have dramatic effects on p-values and nominal statistical significance, essentially making pursuit of small p-values in CLT-based statistical tests with a fixed sample size meaningless.

Many aforementioned social, psychological, cognitive and educational sources of false and irreproducible knowledge generation are hard to detect and mitigate. By contrast, methodological roots of the reproducibility crisis discussed in this work are more tractable and amenable to change.

The conclusions of this work are based on five specific examples highlighting research questions that frequently occur in statistical data analysis. One of these examples, demonstrating the subtlety of the random sample assumption, is drawn from cancer research. The other four examples are mathematical and statistical in nature and contain rigorous proofs but do not pursue the greatest generality. To illustrate the points made in this article, I mostly refer to biomedical sciences; however, these points are applicable in equal measure to natural and social sciences.
I hope that this article will serve as a word of caution to applied statisticians, experimental and theoretical scientists, and practitioners who employ statistical methods in their work. Some related pitfalls of statistical methodology in the setting of clinical trials were discussed in [19].

The flow of exposition in the article is as follows. In Sections 2–5 I discuss through a number of examples some specific sources of false and irreproducible scientific findings that are rooted in statistical methodology. Specifically, Section 2 is focused on the random sample assumption, that underlies most of the statistical methods and tests, and various reasons for its violation. In Section 3, I show that the standard problem of estimation of the expected value is ill-posed in the sense that arbitrarily small errors in specification of the distribution from which a random sample is drawn may lead to arbitrarily large deviations of the estimate from the true expected value. In Section 4, it is demonstrated that samples of random size and stochastic dependence between the sample size and random variables generating the random sample may lead to violation of the LLN, thus preventing conventional statistical inference. Section 5 reveals that pursuit of small significance levels and p-values for a fixed sample size in statistical tests based on the Gaussian approximation is erroneous. Finally, in Section 6 I discuss the findings of Sections 2–5 and possible mitigation measures against generation of false and irreproducible scientific results, while in Section 7 I take a look at the reproducibility crisis from a higher vantage point and make a few concluding remarks.

2. The Random Sample Assumption

Virtually all applications of statistics are based on, or eventually reduced to, the assumption that observations, say $x_1, x_2, \ldots, x_n$ form a random sample. By definition, this means that there exist jointly stochastically independent (i) and identically distributed (id) random variables (or, more generally, random vectors) $X_1, X_2, \ldots, X_n$ defined on a sample space $S$ such that $X_k(s) = x_k$ for some point $s \in S$ and all $k = 1, 2, \ldots, n$. Informally, the observations result from independent replications of the same random experiment. The iid assumption is an idealization that cannot be presumed to be exactly true in the real world. Furthermore, because each random variable $X_1, X_2, \ldots, X_n$ is represented by a single observation, the iid assumption is empirically untestable and therefore invariably taken on faith.

The strength of one’s faith in the iid assumption is setting-specific. For example, in cell biology, pre-existing or emerging genetic differences among cells, their differing spatial position within a cell population, and their distinct functional states at the time of sampling (e.g., proliferating vs. quiescent or stem vs. differentiated) may all make the iid assumption unrealistic. In particular, the range of the measured quantity for different cells may vary, which precludes the id property. For multicellular organisms, the possibilities for violation of the iid assumption are more abundant, even more so for humans with the enormous diversity of their biological and psychological traits. As one example, when observations are associated with a pathological condition in a group of subjects, substantial differences in the susceptibility of the subjects to this condition and its severity will make the underlying random variables non-id. Additionally, if some of the subjects are identical twins, siblings or relatives, or if some of them were exposed to similar environmental, professional or lifestyle hazards, both the independence and id assumptions may be violated. A similar argument applies to the effects and side effects of medical interventions. In the setting of clinical trials, stochastic dependence of clinical observations can also be induced by study entrance criteria and informed consent even when subjects were initially selected randomly. Finally, when sampling from a small population (e.g., when studying a rare disease or a rarely occurring combination of attributes), the mere fact of sampling from it without replacement makes observations stochastically dependent.

Determining whether a given set of observations is or is not a random sample can be quite subtle, as illustrated by the following example from cancer research.
Example 1. Let $x_1, x_2, \ldots, x_n$ be the volumes of metastases detected in a particular cancer patient in a given site (e.g., lungs, liver, brain, bones, lymph nodes or soft tissue). Is this a random sample?

Focusing on the data-generating process, let us assume for simplicity that metastases shed off the primary tumor and bound for the site of interest evolve independently, survive with the same probability, grow according to the same deterministic law and become detectable when their volumes reach a fixed threshold. Then, neglecting metastatic dormancy [20], one concludes that the volumes of detected metastases at the time of survey are a non-random transformation of their shedding times. Because the rate of shedding is time-dependent (e.g., it may depend on the primary tumor volume), the observed volumes of metastases cannot result from replications of the same random experiment. Thus, they do not form a random sample from any probability distribution even under the above simplifying assumptions. Consequently, conventional statistical data descriptors—such as sample mean and standard deviation—lose their inferential significance, a fact commonly neglected in volumetric studies of metastases.

An outcome of a statistical analysis is in the final count just a function of observations whose value is unaffected by the validity of the random sample and other assumptions or lack thereof. Thus, no red flags signaling their violation will ever appear in the process of data analysis, unless the validity of the underlying assumptions is deliberately investigated.

In experimental sciences, departure from the iid assumption is almost universally disregarded, likely based on the belief that the deviations are small and inconsequential. A fundamental question, then, is how to quantify these deviations and assess their impact on the outcome of statistical analysis.

3. The Lack of Robustness

Let us take a closer look at the iid assumption postulating that (let’s say, jointly independent) random variables $X_i, X_2, \ldots, X_n$ generating a given set of observations have the same probability distribution. Deviations from this assumption can be quantified by computing the maximum value, $\delta$, of pairwise distances $d(X_i, X_j), 1 \leq i < j \leq n$, where $d$ is a suitable probability metric such as the total variation, Kolmogorov-Smirnov (KS), Kantorovich, Lévy or Cramér-von Mises metrics [21]. The most convenient and widely used among them is the KS metric defined by

$$d_{KS}(X, Y) = \sup_{x \in \mathbb{R}} |F_x(x) - F_y(x)|,$$

where, $F_x(x) = \Pr(X \leq x)$, $x \in \mathbb{R}$, is the cumulative distribution function (cdf) of random variable $X$. The importance of the KS distance stems from the fact that if the KS distance between two random variables is small then the probabilities that these random variables take values from any given interval are uniformly close to each other. In fact, if $I = (a, b]$ then it follows from $\Pr(X \in I) = F_b(b) - F(a)$ that

$$|\Pr(X \in I) - \Pr(Y \in I)| \leq 2d_{KS}(X, Y).$$

Due to the continuity properties of the probability measure the same is also true for other intervals of the form $I = (a, b], [a, b)$ and $[a, b]$.

Let $\Delta$ be the absolute value of the difference between a statistical measure of interest (such as the point estimate of a parameter, the left or right endpoint of the associated confidence interval, the correlation coefficient, $p$-value for a given null hypothesis, statistical power for an alternative hypothesis, etc.) computed under the iid assumption and without it. Scientists commonly believe that $\Delta$ is small if $\delta$ is. That this can be false is demonstrated below.

Example 2. Let $0 < \varepsilon < 1$ and $X_1, X_2, \ldots, X_n$ be independent random variables such that $n$ of them have the same distribution $P$ with finite unknown expected value $\mu$ and the other $n$ have distribution $Q = (1- \varepsilon)P + \varepsilon\mathcal{N}(\varepsilon)$, where $\mathcal{N}(\varepsilon) \neq P$ is a distribution whose expected value, $\mathcal{N}(\varepsilon)$, has the
property that $\varepsilon r(\varepsilon) \to 0$ as $\varepsilon \to 0$. Examples of such distribution $R(\varepsilon)$ include the exponential distribution $\text{Exp}(\varepsilon^0)$ and the uniform distribution $U(0, \varepsilon^0)$ with $\theta > 1$. Suppose our goal is to estimate parameter $\mu$.

Observe that the KS distance between the distributions $P$ and $Q$ does not exceed $\varepsilon$. Neglecting the difference between the distributions $P$ and $Q$, i.e., invoking the iid assumption, makes our $2n$ observations a random sample from the distribution $P$. The estimator of $\mu$, then, is the sample mean of $2n$ observations, which according to the LLN is almost surely close to $\mu$ for large $n$. However, without the iid assumption, the sample mean for large $n$ is close to

$$0.5[\mu + (1-\varepsilon)\mu + \varepsilon r(\varepsilon)] = \mu + 0.5\varepsilon[r(\varepsilon) - \mu].$$

Notice that the departure, $\Delta$, of this estimate from $\mu$ is arbitrarily large for sufficiently small $\varepsilon$. Thus, arbitrarily small deviations from the iid assumption, as measured by the KS metric, may produce arbitrarily large errors in the estimates of parameter $\mu$.

The same effect can be achieved with an arbitrarily small fraction of observations whose distribution $Q$ deviates from $P$. Furthermore, if the distribution $P$ has a long (in particular, infinitely long) tail, then looking for outliers will not help to distinguish observations drawn from $P$ from those originating from $Q$.

An essential and almost universal feature of the applications of statistics in sciences is that the distribution $P$ that underlies a random sample of observations is unknown. A small modification of Example 2 would show that even a minor misspecification of the distribution $P$ (resulting e.g., from assuming normality or any other convenient distributional form) may have a devastating effect on the estimation of its expected value. In fact, if the unknown “true” distribution $P$ is replaced with distribution $Q$ from Example 2, then the mean for a large sample drawn from $Q$ would deviate from $\mu$ by an amount close to $\varepsilon[r(\varepsilon) - \mu]$. Thus, the deviation becomes arbitrarily large for small $\varepsilon$.

Establishing robustness of a statistical analysis depends critically on availability of a tight, readily computable upper bound for $\Delta$ expressed in terms of $\varepsilon$ such that the bound approaches zero as $\varepsilon \to 0$. Example 2 shows that selection of appropriate probability metric $\delta$ is essential for obtaining such a bound. If the latter is beyond reach for all suitable probability metrics or if available bounds do not have the above “continuity at zero” property or are too conservative to be practical, there is no guarantee that the statistical analysis won’t produce a false and/or irreproducible result.

Similar sensitivity analysis would account for the deviations from joint independence. Such analysis, however, is greatly complicated by the difficulty finding a simple quantitative measure of such deviation. Note that covariance or its normalized version, the correlation coefficient, are unfit for this purpose, for zero correlation does not imply independence of two random variables while pairwise independence of random variables does not entail their joint independence.

4. Samples of Random Size

Yet another reason why the site-specific volumes of detectable metastases in Example 1 do not form a random sample is that their number, $n$, is random. This is so because many aspects of metastasis—shedding times, survival of metastases, duration of their dormancy, and rates of growth—all depend on chance [22]. Interestingly, if metastasis shedding off the primary tumor is governed by a Poisson process with an arbitrary time-dependent rate then the observed volumes of metastases can naturally be associated with a probability distribution. Specifically, in this case the vector of ordered volumes of detected metastases is equidistributed, conditional on $n$, with the vector of order statistics for a random sample of size $n$ drawn from an absolutely continuous probability distribution which is independent of $n$ and can be computed in closed form [23,24].

Although the assumption of fixed sample size is ubiquitous in statistical analyses of experimental data, it is quite often violated in practice. In particular, samples of random
size arise when reaching a predetermined number of observations proves impossible or impractical due to limited resources or constraints on the observation time. Another systematic source of random sample sizes is the researchers’ natural desire to collect observations that are sufficiently representative of the studied phenomenon. Finally, random sample size results from study designs that require a sufficiently large number of events of interest randomly occurring in tested subjects.

Importantly, samples of random size cannot be handled with standard statistical tools because even the most basic results in probability that form the bedrock of statistics, such as the LLN and CLT, may fail. Below we give specific examples illustrating the failure of the LLN.

Recall that the LLN states that if \( \{X_n: n = 1, 2, 3, \ldots \} \) is a sequence of iid random variables with finite expected value \( \mu \) then

\[
(X_1 + X_2 + \ldots + X_n)/n \to \mu \quad \text{as} \quad n \to \infty \tag{1}
\]

in probability or, in the case of strong LLN, almost surely (i.e., with probability 1). The LLN asserts the consistency of the sample mean and leads to the frequentist interpretation of probability; it also shows that fluctuations of averaged iid random variables around their mean asymptotically cancel out.

To formulate a natural extension of the classic LLN to the case of random sample size, consider a sequence \( \{N_n: n = 1, 2, 3, \ldots \} \) of positive integer-valued random variables (sample sizes) \( N_n \) with the expected values \( \mu_n \to \infty \). Then the above sequences \( \{X_n\} \) and \( \{N_n\} \) satisfy the LLN if

\[
Y_n = (X_1 + X_2 + \ldots + X_{N_n})/N_n \to \mu \quad \text{as} \quad n \to \infty \tag{2}
\]

in probability (or almost surely in the case of strong LLN). Clearly, if \( N_n = n \) with probability 1 for all \( n \) then (2) is identical to (1).

**Example 3.** Let \( \{X_n: n = 1, 2, 3, \ldots \} \) be iid random variables with the expected value \( \mu = 0 \) and a distribution \( P \) other than the unit mass at 0. Suppose that for each \( n \geq 1 \) random variables \( N_n \) and \( X_k, 1 \leq k \leq n \), are stochastically independent and that for some fixed natural number \( M \) and \( p \in (0, 1) \)

\[
\Pr(1 \leq N_n \leq M) \geq p \quad \text{for all} \quad n \geq 1. \tag{3}
\]

If one conducts \( r \) replications of an experimental study with random sample sizes \( N_1, N_2, \ldots, N_r \), then if we denote by \( M \) the largest among the observed sample sizes \( n_1, n_2, \ldots, n_r \), then assumption (3) is satisfied for \( n = 1, 2, \ldots, r \) with \( p = \min\{\Pr(N_k = n_k): 1 \leq k \leq r\} > 0 \). This, together with practical limitations on the sample size, makes assumption (3) plausible.

I now proceed with disproving, under the above conditions, the version of LLN given by (2). Denote \( Z_k = (X_1 + X_2 + \ldots + X_k)/k \), \( k \geq 1 \). For \( \varepsilon > 0 \), set

\[
A(\varepsilon) = \min\{\Pr(|Z_k| \geq \varepsilon): 1 \leq k \leq M\}.
\]

**Proposition 1.** There exists \( \varepsilon > 0 \) for which \( A(\varepsilon) > 0 \).

**Proof.** Suppose \( A(\varepsilon) = 0 \) for all \( \varepsilon > 0 \). Consideration of \( \varepsilon = 1/m \), where \( m \) is a natural number, leads to a conclusion that there exists \( k, 1 \leq k \leq M \), such that \( \Pr(|Z_k| \geq 1/m) = 0 \) for infinitely many \( m \). This implies that \( Z_k = 0 \), where this and other similar equalities in the next two sentences occur with probability 1. In the case \( k = 1 \) this means that \( X_1 = 0 \), which contradicts the assumed non-degeneracy of the distribution \( P \). For \( k > 1 \), the fact that \( Z_k = 0 \) amounts to \( X_1 + X_2 + \ldots + X_{k-1} = 0 \) or equivalently to the equality \( X_k = -(X_1 + X_2 + \ldots + X_{k-1}) \). Since random variables \( X_i, i \geq 1 \), are independent, so are \( X_i \) and \( X_1 + X_2 + \ldots + X_{k-1} \). This implies that random variable \( X_k \) is independent with itself, hence \( P(X_k \in B) = 0 \) or 1 for any Borel subset \( B \) of \( \mathbb{R} \). Since \( P(X_k \in \mathbb{R}) = 1 \) there is a finite closed interval \( I = [a, b] \) such that \( P(X_k \in \mathbb{R}) = 1 \).
\[ \begin{align*}
\in I) = 1. \text{ Then the same is true for exactly one of the two subintervals } [a, (a + b)/2] \text{ and } ((a + b)/2, b]. \text{ Denote the closure of such interval } I. \text{ Continuation of this process produces a decreasing sequence } (I) \text{ of closed intervals such that } P(X \in I) = 1 \text{ for all } j. \text{ Because } \text{diam}(I) \to 0, \text{ the intersection of these intervals is just one point, say } c. \text{ The continuity of probability implies that } P(X_\in = c) = 1, \text{ and in view of the assumption that } E_X = \mu = 0, \text{ one must have } c = 0. \text{ This contradicts, yet again, the assumption that distribution } P \text{ is different from } \delta_0. \square
\end{align*}\]

The argument that the LLN in the form (2) fails can now be completed as follows. By the formula of total probability and in view of the assumption that, for each } n, \text{ sample size } N_n \text{ is independent of random variables } X_k, 1 \leq k \leq n, \text{ we have}

\[
\text{Pr}(Y_n \geq \epsilon) = \sum_{k=1}^{\infty} \text{Pr}(|Y_n| \geq \epsilon|N_n = k) \text{ Pr}(N_n = k) = \sum_{k=1}^{\infty} \text{Pr}(|Z_k| \geq \epsilon) \text{ Pr}(N_n = k)
\]

\[
\geq \sum_{k=1}^{M} \text{Pr}(|Z_k| \geq \epsilon) \text{ Pr}(N_n = k) \geq A(\epsilon) \sum_{k=1}^{M} \text{Pr}(N_n = k) = A(\epsilon) \text{ Pr}(1 \leq N_n \leq M) \geq p A(\epsilon).
\]

Thus, for the number } \epsilon > 0 \text{ specified in Proposition 1, } \text{Pr}(1 Y_n \geq \epsilon) \text{ does not converge to } 0 \text{ as } n \to \infty. \text{ Therefore, the sequence of random variables } \{Y_n\} \text{ does not converge to } 0 \text{ in probability, hence also almost surely.}

The next example addresses the case where random sample sizes } N_n \text{ and random variables } X_1, X_2, \ldots, X_n \text{ are stochastically dependent.

**Example 4.** Let, as in Example 3, } \{X_n: n = 1, 2, 3, \ldots\} \text{ be iid random variables with the expected value } \mu = 0 \text{ and a distribution } P \text{ other than the unit mass at } 0. \text{ Define random variables } N_N \text{ as follows: } N_1 = 1 \text{ with probability } 1 \text{ and for } n \geq 2 \text{ set } N_n = 1 \text{ if } X_1 \geq 0 \text{ and } N_n = n \text{ if } X_1 < 0.

Let } \lambda = \text{Pr}(X_1 \geq 0), \text{ then } 0 < \lambda < 1. \text{ For the expected value of } N_n \text{ we have: } \mu_n = E N_n = \lambda + (1-\lambda) n \to \infty \text{ as } n \to \infty. \text{ In the case } n \geq 2 \text{ random variables } Y_n \text{ defined in (2) are given by } Y_n = X_1 \text{ with probability } \lambda \text{ and } Y_n = X_n \text{ with probability } 1 - \lambda. \text{ Therefore, according to the standard LLN the almost sure limit in (2) is a non-constant random variable rather than } \mu = 0.

Note that an example somewhat similar to Example 4 was considered in [25] to show that the CLT may fail when random variables } \{X_n: n \geq 1\} \text{ and random sample sizes } N_n \text{ are stochastically dependent. For a review of the limiting behavior of the sum of random number of random variables, see [26].

5. Effects of the Gaussian Approximation on P-Values and Statistical Significance

One systematic source of Gaussian distributions in statistics is the CLT that approximates (in the KS metric) the distribution of the normalized sum of iid random variables by the standard Gaussian distribution. The following example below demonstrates that the error resulting from such an approximation may have dramatic consequences for p-values and statistical significance of research findings.

As a prerequisite to this example, several results about the support of probability distributions and an equivalent representation of the KS distance have to be established. Recall that an open interval } I \text{ is a null set for a probability distribution } P \text{ on } R \text{ if } P(I) = 0. \text{ Let } G \text{ be the union of all null intervals for } P \text{ or, in other words, the largest open null set for } P. \text{ The complement of } G \text{ is called the support of } P \text{ and denoted } S(P). \text{ If } P \text{ is the distribution of a random variable } X, \text{ then notations } S(P) \text{ and } S(X) \text{ will be used interchangeably. For example, the support of the distributions } U(a, b), \text{ Exp}(\lambda) \text{ and } N(\mu, \sigma^2) \text{ are } [a, b], [0, \infty) \text{ and } R, \text{ respectively. Recall that distribution } P \text{ is continuous if } P(|X|) = 0 \text{ for all } x \in R. \text{ Alternatively, a continuous distribution can be defined as a distribution whose cdf is a continuous function. The significance of the notion of support for statistics lies in the fact that if number } x \text{ is drawn at random from a continuous distribution } P \text{ then } x \in S(P).

**Proposition 2.** Suppose } X \text{ and } Z \text{ are continuous random variables and } S(Z) = R. \text{ Then}
\[ d_{KS}(X, Z) = \sup \{|F_X(x) - F_Z(x)| : x \in S(X)\}. \] (4)

**Proof.** Let \( G \) be the complement of \( S(X) \). Denote by \( d \) the right-hand side of (4). We have to show that \(|F_X(x) - F_Z(x)| \leq d \) for each \( x \in G \). Recall that \( G \) can be represented as a finite or countable disjoint union of open intervals whose endpoints belong to \( S(X) \) or are \( -\infty \) or \( \infty \). Let \( I = (a, b) \) be one such interval. By definition of the support of \( X \), function \( F_X \) is constant on \( I \): \( F_X(x) = c \) for all \( x \in I \). Then due to the continuity of the cdf \( F_X \) we have \( F_X(a) = F_X(b) = c. \) Note that function \( F_Z \) is strictly increasing on \( R \). Therefore, if \( F_Z(b) \leq c \) then

\[ \sup \{|F_X(x) - F_Z(x)| : x \in I\} = c - F_Z(a) = |F_X(a) - F_Z(a)|. \]

Similarly, if \( F_Z(a) \geq c \) then

\[ \sup \{|F_X(x) - F_Z(x)| : x \in I\} = F_Z(b) - c = |F_X(b) - F_Z(b)|. \]

In the remaining case where \( F_Z(a) < c < F_Z(b) \) there is a unique point \( \alpha \in (a, b) \) such that \( F_Z(\alpha) = c. \) Then, the function \( f(x) = |c - F_Z(x)|, x \in I, \) is decreasing on \([a, \alpha)\) and increasing on \([\alpha, b)\) so that

\[ \sup \{|F_X(x) - F_Z(x)| : x \in I\} = \max \{|F_X(a) - F_Z(a)|, |F_X(b) - F_Z(b)|\} \leq d. \]

Thus, the inequality \( \sup \{|F_X(x) - F_Z(x)| : x \in I\} \leq d \) is true in all cases including \( a = -\infty \) and \( b = \infty. \) \( \square \)

To formulate the next result, denote by \( \text{cl}(A) \) the closure of a subset \( A \) of \( R \), that is, the set of all its limit points.

**Proposition 3.** Let \( X_1, X_2, \ldots, X_n \) be independent random variables. Then \( S(X_1 + X_2 + \ldots + X_n) \) is contained in \( \text{cl}[S(X_1) + S(X_2) + \ldots + S(X_n)] \).

**Proof.** Denote by \( P_1, P_2, \ldots, P_n \) the respective distributions of random variables \( X_1, X_2, \ldots, X_n \). We first consider the case \( n = 2 \). Since random variables \( X_1 \) and \( X_2 \) are independent, the distribution of \( X_1 + X_2 \) is the convolution \( P_1 \ast P_2 \) defined as the probability measure such that

\[
\int_R f(x) d(P_1 \ast P_2)(x) = \int\int_{R^2} f(x + y) dP_1(x) dP_2(y)
\]

or equivalently

\[
\int_{S(P_1 \ast P_2)} f(x) d(P_1 \ast P_2)(x) = \int\int_{S(P_1) \times S(P_2)} f(x + y) dP_1(x) dP_2(y)
\] (5)

for all continuous functions \( f \) on \( R \) with zero limit at \( \pm \infty \). To show that \( A = S(P_1 \ast P_2) \) is a subset of \( B = \text{cl}[S(P_1) + S(P_2)] \), pick a point \( x \) that does not belong to \( B \). The open set \( B^c \) can be represented as a disjoint union of open intervals; one of them, say \( I \), contains \( x \). Then, for every non-negative continuous function \( f \) that vanishes outside of \( I \), the integral on the right-hand side of (5) equals 0. Therefore, so is the integral on the left-hand side of (5), which implies that \( (P_1 \ast P_2)(I) = 0 \). Thus, \( I \) is a null set for \( P_1 \ast P_2 \) and hence \( x \) does not belong to \( A \). This establishes the required inclusion in the case \( n = 2 \). The rest of the proof proceeds by induction in \( n \) combined with monotonicity of the closure operation with respect to containment of sets and the fact that there is no change in \( \text{cl}[U + \text{cl}(V)] = \text{cl}(U + V) \) for all subsets \( U, V \) of \( R \). \( \square \)

We are now ready for the next example.
Example 5. Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous (but not necessarily absolutely continuous) probability distribution $P$ with an unknown expected value $\mu$, known variance $\sigma^2 > 0$ and a finite third absolute central moment $\beta$. Suppose our goal is to compute the $p$-value for the null hypothesis $\mu = \mu_0$ from a sample of observations $x_1, x_2, \ldots, x_n$. According to the CLT, under the null hypothesis the distribution of the random variable

$$Y_n = (X_1 + X_2 + \ldots + X_n - n\mu_0)/(n^{1/2} \sigma)$$

can be approximated in the KS metric by the standard Gaussian distribution represented by a random variable $Z$. Denote

$$y_n = (x_1 + x_2 + \ldots + x_n - n\mu_0)/(n^{1/2} \sigma)$$

Let $\pi_n$ the “true” one-sided $p$-value, that is, $\Pr(Y_n \geq y_n)$ if $y_n \geq 0$ and $\Pr(Y_n \leq y_n)$ if $y_n < 0$. Because the distribution of random variable $Y_n$ is typically unknown, it is a common practice to compute the approximate one-sided $p$-value, $p_n$, based on the standard Gaussian approximation to the distribution of $Y_n$. Thus, $p_n$ is given by the probability $\Pr(Z \geq y_n)$ if $y_n \geq 0$ and $\Pr(Z \leq y_n)$ if $y_n < 0$. How large is the maximum value, $\Delta_n$, of the error $|p_n - \pi_n|$ in $p$-value determination (with respect to data variation) due to the Gaussian approximation? This question is answered by the following theorem.

Theorem 1. $\Delta_n = d_{KS}(Y_n, Z)$.

Proof. Set $U_i = (X_i - n\mu_0)/(n^{1/2} \sigma)$, $1 \leq i \leq n$. Observe that random variables $U_1, U_2, \ldots, U_n$ are iid, have a continuous distribution that we denote by $Q$, and $Y_n = U_1 + U_2 + \ldots + U_n$. Then $y_n$ defined in (6) can be represented in the form $y_n = u_1 + u_2 + \ldots + u_n$, where $u_1, u_2, \ldots, u_n$ is a random sample from $Q$. Therefore, the set $S_n$ of all possible values of $y_n$ is the sum of $n$ copies of $S(Q)$.

The continuity of random variables $Y_n$ and $Z$ implies that for $y_n \geq 0$

$$|p_n - \pi_n| = |\Pr(Y_n \geq y_n) - \Pr(Z \geq y_n)| = |\Pr(Y_n < y_n) - \Pr(Z < y_n)| = |\Pr(Y_n \leq y_n) - \Pr(Z \leq y_n)|$$

while in the case $y_n < 0$ one has $|p_n - \pi_n| = |\Pr(Y_n \leq y_n) - \Pr(Z \leq y_n)|$. Therefore,

$$\Delta_n = \sup\{|p_n - \pi_n|: y_n \in S_n\} = \sup\{|F_{Y_n}(y_n) - F_Z(y_n)|: y_n \in S_n\} = d_{KS}(Y_n, Z)$$

On the other hand, according to Propositions 2 and 3,

$$d_{KS}(Y_n, Z) = \sup\{|F_{Y_n}(x) - F_Z(x)|: x \in S(Y_n)\} \leq \sup\{|F_{Y_n}(x) - F_Z(x)|: x \in cl(S_n)\} = \sup\{|F_{Y_n}(x) - F_Z(x)|: x \in S_n\}.$$ 

Thus, $\Delta_n = d_{KS}(Y_n, Z)$, as claimed. $\square$

Note also that for the two-sided $p$-value $\Delta_n \leq 2d_{KS}(Y_n, Z)$ but the equality is generally not guaranteed.

Denote $\delta_n = d_{KS}(Y_n, Z)$. The fact that $\delta_n = \Delta_n$ has important implications for $p$-values and the nominal statistical significance level, $\alpha$, when the one-sided Z-test is employed to test the null hypothesis that $\mu = \mu_0$ based on a random sample of size $n$. According to Berry-Esseen theorem [27-29], $\delta_n \leq 0.5C/n^{1/2}$, where $C = \beta/\sigma^3 \geq 1$. Importantly, this is not true if the coefficient 0.5 is replaced with 0.4. Specifically, it was shown in [30] that $\delta_n > 0.4C/n^{1/2}$ for a discrete distribution $P$ with two probability masses (and hence for infinitely many other distributions including continuous ones). Therefore, for any such distribution we have $\Delta_n > 0.4/n^{1/2}$. This implies that there exist infinitely many continuous probability distributions $P$ for which the inequality $|p_n - \pi_n| > 0.4/n^{1/2}$ holds for some random samples of size $n$ drawn from $P$.

This inequality sheds light on the range of significance levels $\alpha$ that can be used in the one-sided Z-test. Recall that the standard practice is to reject the null hypothesis if $p_n \leq \alpha$. This inequality is a surrogate for the finite-sample inequality of interest $\pi_n \leq \alpha$, which is typically hard or impossible to verify directly. Of course, one should only consider such
values of $\alpha$ that for any distribution $P$ both inequalities $p_0 \leq \alpha$ and $\pi_0 \leq \alpha$ hold for at least one random sample drawn from $P$. I will call such values of $\alpha$ admissible. Because $p_0$, $\pi_0 \geq 0$ and for some distributions $P$ and random samples drawn from $P$ one has $|p_0 - \pi_0| > 0.4/n^{1/2}$, any significance level $\alpha \leq 0.4/n^{1/2}$ is inadmissible. In other words, any significance level $\alpha$ admissible for the one-sided $Z$-test must satisfy the inequality $\alpha > 0.4/n^{1/2}$.

Applying the inequality $\alpha > 0.4/n^{1/2}$ to $\alpha = 0.05$, we conclude that this standard significance level for the one-sided $Z$-test is admissible only for sample sizes $n \geq 65$ ($n \geq 257$ for the two-sided test), and these are just conservative lower bounds based on the smallest possible value $C = 1$, which occurs only for discrete distributions with two equal probability masses! In a recent article [31], 72 distinguished statistical scientists advocated a blanket reduction of the nominal significance threshold to $\alpha = 0.005$. In the case of $Z$-test, this is feasible only if $n \geq 6401$ ($n \geq 25,601$ for the two-sided test). Clearly, sample sizes in the vast majority of empirical studies do not meet these constraints.

The above lower bounds for the maximum error in $p$-value computation for a fixed sample size $n$ can, at least in principle, be obtained for other statistical tests that use the CLT-based normal approximation. As shown above, such a lower bound, $L(n)$, also serves as a lower bound for the nominal significance level $\alpha$ and implies a lower bound for the sample size given $\alpha$. For all such tests, the use of significance levels $\alpha \leq L(n)$ is indefensible. The same applies to $p$-values $p_0 \leq L(n)$, for this would imply, for certain random samples, that the “true” $p$-value $\pi_0$ satisfies the inequality $\pi_0 > p_0 + L(n)$. This renders pursuit of small $p$-values $p_0$ for a fixed sample size $n$ meaningless.

6. Discussion and Recommendations

It is a basic tenet of engineering that the design for a load-bearing structure or device will not be commissioned or certified unless it can withstand a load far exceeding the one that would occur under the most unfavorable operation conditions. Likewise, no numerical method would be considered satisfactory without hard estimates of the effects of systematic and round-off errors. In contrast to this precautionary ethos, scientific researchers and practitioners often accept the results of statistical analysis based on empirically untestable assumptions that are never exactly met and various approximations without paying heed to the magnitude of associated errors. Such a contrast is especially troubling given that statistical methods are often used for reaching important scientific conclusions and making critical policy decisions.

Due to various assumptions and approximations, some of which were discussed in this article, the “standard” distributions of test statistics invariably deviate from their true distributions. As shown above, arbitrarily small departures from these assumptions and the use of the Gaussian approximation may have a dramatic impact on the outcomes of statistical analysis. Furthermore, some experimental settings and designs (e.g., when sample size is random or when the random sample size and the data are stochastically dependent) render conventional statistical methodology inapplicable. The frequent failure to identify and grapple with these challenges is one of the major causes of false and irreproducible scientific findings.

What can be done to mitigate this problem, beyond improving the design of empirical studies, avoiding conflicts of interest and various biases, and enforcing standardization and sharing of experimental protocols?

One remedy would be to change the culture of applied statistical research. As a system of knowledge, statistics holds a special place among mathematical sciences in that understanding and correct application of even the most basic tools of inferential statistics requires extensive technical knowledge of many areas of mathematics such as real, complex and functional analysis (including measure theory and Lebesgue integration) and advanced probability. Every statistical argument, method and test rests on the foundation of analysis, probability and mathematical statistics, see e.g., Section 5. Combining a creative use of this theoretical arsenal with additional features built into a particular empirical setting can lead to an extension of statistical methods beyond the standard assumptions.
and approximations and to rigorous estimation of associated errors. This will enrich statistical methodology, generate better and more reproducible research outcomes, and make statistical analysis of a particular scientific problem a unique and exciting endeavor. However, if in spite of all the effort the challenges are too daunting, or if the error estimates prove too conservative to be practical, there is nothing wrong about dropping inferential statistical analysis altogether.

Paradoxically, such rigorous application of statistical methods is rarely attempted. All too often, the practice of statistics devolves into ritualistic application of standard statistical models and tests, $p$-values and confidence intervals unmoored from their mathematical roots and insufficiently integrated with the scientific problems they help to solve.

Another part of the solution needs to come from broadening graduate education. Evaluation of empirically untestable assumptions built into statistical data analysis requires keen understanding of a scientific problem from both subject-specific and formal mathematical viewpoints; however, to be successful, the synthesis must take place in the same head! Even greater mathematical proficiency including intimate knowledge of theory and proofs is required for extending statistical methods beyond their standard premises and assessing their robustness to errors. This calls for more dual master’s and Ph.D. programs that will graduate scientists who are equally well trained in mathematics (with concentration in analysis, probability and mathematical statistics) and in a broadly defined subject area.

One area of statistics that has a special importance for such an extended graduate education, continued education of applied statisticians and scientists, and more generally mitigation of the reproducibility crisis is robust statistics. It has developed statistical estimators and methods that are resistant to outliers, the presence of a limited fraction of missing or misspecified data, moderate contamination of the underlying probability distributions with distributions that belong to the same or other well-defined parametric families, and certain other minor violations of the standard assumptions [32–34]. In the univariate setting, a number of convenient quantitative measures of robustness have been proposed [32,33]. Although the foundations of robust statistics were laid in the 1970s–1980s, its methodology has only gained a limited traction in the wider community of applied statisticians and experimental scientists. The reasons, I believe, are two-fold. First, application of robust statistics can hardly be done in a cookbook, entirely algorithmic way. Rather, it involves creative application, and possibly modification, of many approaches and comparison of the results. This requires a deep understanding of both motivational ideas behind robust statistics and its difficult technical aspects, which is barely accessible to those practitioners who are not well versed in probability and analysis. Second, the problems arising from modern science are typically multivariate. The methods of robust statistics, that were originally developed with the univariate setting in mind, do not generalize in a straightforward manner to the multivariate case. Furthermore, the methods of multivariate robust statistics are very complex and still an ongoing project.

In spite of a considerable advantage of the methods of robust statistics over the classical ones, they are still based on certain untestable assumptions and rely on various approximations. Additionally, their practical use requires selection of various thresholds. Most importantly for the subject of this article, robust statistics does not lead to hard estimates of the errors arising from the violation of its assumptions and the use of relevant approximations.

Yet another promising direction in statistics that reduces the reliance of statistical methods on asymptotic results including the CLT is the use of finite-sample asymptotic formulas for the probability density functions of various statistics, parameter estimators and confidence intervals [35]. These results depend heavily on the methods of complex analysis [35]. An additional benefit can be derived from the so-called “pre-limit theorems.” These theorems are based on the discovery that the distribution of certain statistics (e.g., normalized sums of iid random variables or their maxima) for intermediate sample sizes are close to the distributions that belong to well-known classes and are different from
the asymptotic distributions of these statistics arising when the sample size increases indefinitely [36].

7. Concluding Remarks

False and irreproducible scientific findings strike at the very heart of science, contravening its goal of uncovering the truth about the natural world and human society and undermining the credibility of this arguably most important human endeavor. Although human and societal factors undoubtedly contribute to the generation of false and irreproducible knowledge, another source of this phenomenon is undue elevation of the importance of p-values combined with their ritualistic use, misuse and misinterpretation. In spite of numerous objections and warnings issued over the last century, p-values, especially the magic \( p \leq 0.05 \) criterion, have become the single most important filter of empirical scientific knowledge employed by both scientists and journals that publish their work. As one example, a massive meta-analysis study [37] found that, among almost 2 million biomedical papers published between 1990 and 2015, 96% appealed to \( p \)-value \( \leq 0.05 \) to claim significance of their results. In reality, \( p \)-values have little to do with the validity or importance of scientific findings, and the glaring discrepancy between their publicly accepted and real value has led to fraud, abuse and, most unfortunately, to the growing distrust of statistical methods by scientists including, in extreme cases, their rejection in toto. Belatedly, the American Statistical Association has issued in 2016 a formal position statement on \( p \)-values [16] followed in 2019 by an editorial and a broad discussion in The American Statistician of how to transition to the world beyond \( p < 0.05 \) [17]. The 43 contributions to [17] showed no clear consensus regarding the path forward.

The \( p \)-value crisis is only the tip of an iceberg and a symptom of deeper problems with statistical methodology. Some of these problems and possible remedies were discussed in this article. Specifically, I argued, both theoretically and using specific examples, that the validity and reproducibility of scientific findings based on traditional statistical data analysis alone is not assured. To be a valid and reliable methodology for natural, biomedical and social sciences, conventional statistics should be augmented by rigorous analysis and hard estimates of various errors resulting from the violation of underlying assumptions and the use of various approximations.

Funding: The author confirms that no funding associated with this manuscript was provided.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The author is grateful to Mark Hanin for his helpful comments and suggestions on the style and substance of this article.

Conflicts of Interest: The author declares no conflict of interest.

References
1. Altman, D.G. The scandal of poor medical research. *Br. Med. J.* 1994, 308, 283–284.
2. Ioannidis, J.P.A. Contradicted and initially stronger effects in highly cited clinical research. *J. Am. Med. Assoc.* 2005, 294, 218–228.
3. Horton, R. What is medicine’s 5 sigma? *Lancet* 2015, 385, 1380.
4. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 2015, 349, aac4726.
5. Branch, M.N. The "Reproducibility Crisis:" Might the methods used frequently in behavior-analysis research help? *Perspect. Behav. Sci.* 2018, 42, 77–89.
6. França, T.F.A.; Monserrat, J.M. Reproducibility crisis in science or unrealistic expectations? *EMBO Rep.* 2018, 19, e46008.
7. Romero, F. Philosophy of science and the replicability crisis. *Philos. Compass* 2019, 14, e12633.
8. Ioannidis, J.P.A. Why most published research findings are false. *PLoS Med.* 2005, 2, e124.
9. Goodman, S.N. Why is getting rid of \( p \)-values so hard? Musings on science and statistics. *Am. Stat.* 2019, 73, 26–30.
10. Goodman, S.N.; Fanelli, D.; Ioannidis, J.P.A. What does research reproducibility mean? *Sci. Transl. Med.* 2016, 8, 341ps12.
11. Goodman, S.N. Toward evidence-based medical statistics, 1: The P value fallacy. *Ann. Intern. Med.* **1999**, *130*, 995–1004.
12. Goodman, S.N. A dirty dozen: Twelve p-value misconceptions. *Semin. Hematol.* **2008**, *45*, 135–140.
13. Ziliak, S.; McCloskey, D. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*; University of Michigan Press: Ann Arbor, MI, USA, 2008.
14. Gelman, A. P values and statistical practice. *Epidemiology* **2013**, *24*, 69–72.
15. Greenland, S.; Senn, S.J.; Rothman, K.J.; Carlin, J.B.; Poole, C.; Goodman, S.N.; Altman, D.J. Statistical tests, P-values, confidence intervals, and power: A guide to misinterpretations. *Am. Stat.* **2016**, *70*, 1–12.
16. Wasserstein, R.L.; Lazar, N.A. The ASA’s statement on p-values: Context, process, and purpose. *Am. Stat.* **2016**, *70*, 129–133.
17. Wasserstein, R.L.; Schirm, A.L.; Lazar, N.A. Moving to a world beyond “p < 0.05”. *Am. Stat.* **2019**, *73*, 1–19.
18. Kennedy-Shaffer, L. Before p < 0.05 to Beyond p < 0.05: Using history to contextualize p-values and significance testing. *Am. Stat.* **2019**, *73*, 82–90.
19. Hanin, L. Why statistical inference from clinical trials is likely to generate false and irreproducible results. *BMC Med. Res. Methodol.* **2017**, *17*, 127.
20. Demicheli, R. Tumour dormancy: Findings and hypotheses from clinical research on breast cancer. *Semin. Cancer Biol.* **2001**, *11*, 297–306.
21. Rachev, S.T.; Klebanov, L.; Stoyanov, S.V.; Fabozzi, F. *The Methods of Distances in the Theory of Probability and Statistics*; Springer: New York, NY, USA, 2013.
22. Kendall, W.S. Chance mechanisms affecting the burden of metastases. *BMC Cancer* **2006**, *5*, 138–146.
23. Hanin, L.; Rose, J.; Zaider, M. A stochastic model for the sizes of detectable metastases. *J. Theor. Biol.* **2006**, *243*, 407–417.
24. Hanin, L.; Rose, J. Suppression of metastasis by primary tumor and acceleration of metastasis following primary tumor resection: A natural law? *Bull. Math. Biol.* **2018**, *80*, 519–539.
25. Rényi, A. On the central limit theorem for the sum of a random number of independent random variables. *Acta Math. Acad. Sci. Hungar.* **1960**, *11*, 97–102.
26. Galambos, J. Random sample sizes: Limit theorems and characterizations. In *Probability Theory and Applications*; Galambos, J., Katali, I., Eds.; Mathematics and Its Applications; Springer: New York, NY, USA, 1992; pp. 107–123.
27. Berry, A.C. The accuracy of the Gaussian approximation to the sum of independent variates. *Trans. AMS* **1941**, *49*, 122–136.
28. Shiryaev, A.N. *Probability*; Graduate Texts in Mathematics; Springer: New York, NY, USA, 1996.
29. Tyurin, I.S. An improvement of upper estimates of the constants in the Lyapunov theorem. *Russ. Math. Surv.* **2010**, *65*, 201–202.
30. Esseen, C.-G. A moment inequality with an application to the central limit theorem. *Scand. Actuar. J.* **1956**, *39*, 160–170.
31. Benjamin, D.J.; Berger, J.O.; Johannesson, M.; Nosek, B.A.; Wagenmakers, E.-J.; Berk, R.; Bollen, K.A.; Brembs, B.; Brown, L.; Camerer, C.; et al. Redefine statistical significance. *Nat. Hum. Behav.* **2017**, *2*, 6–10, doi:10.1038/s41562-017-0189-z.
32. Huber, P.J. *Robust Statistics*; Wiley: New York, NY, USA, 1981.
33. Maronna, R.A.; Martin, R.D.; Yohai, V.J. *Robust Statistics: Theory and Methods*; Wiley: New York, NY, USA, 2006.
34. Morgenthaler, S. A survey of robust statistics. *Stat. Methods Appl.* **2007**, *15*, 271–293; Erratum in *2007*, *16*, 171–172.
35. Ronchetti, E. Small sample asymptotics: A review with applications to robust statistics. *Comput. Stat. Data Anal.* **1990**, *10*, 207–223.
36. Klebanov, L.B.; Rachev, S.T.; Szekely, G.J. Pre-limit theorems and their applications. *Acta Appl. Math.* **1999**, *58*, 159–174.
37. Chavarialias, D.; Wallach, J.D.; Li, A.H.T.; Ioannidis, J.P.A. Evolution of reporting P values in the biomedical literature, 1990–2015. *J. Am. Med. Assoc.* **2016**, *315*, 1141–1148.