Biochemical Pharmacology 87 (2014) 78–92

Contents lists available at ScienceDirect

Biochemical Pharmacology

journal homepage: www.elsevier.com/locate/biochempharm

Review – Part of the Special Issue – Pharmacology in 21st Century Biomedical Research

The use and misuse of statistical methodologies in pharmacology research

Michael J. Marino

Merck Research Laboratories, Merck & Co., Inc., 770 Sumneytown Pike, West Point, PA 19486, United States

ARTICLE INFO

Article history:
Received 23 April 2013
Accepted 20 May 2013
Available online 4 June 2013

Keywords:
Parametric
Non-parametric
Exploratory data analysis
Power analysis
Statistical design

ABSTRACT

Descriptive, exploratory, and inferential statistics are necessary components of hypothesis-driven biomedical research. Despite the ubiquitous need for these tools, the emphasis on statistical methods in pharmacology has become dominated by inferential methods often chosen more by the availability of user-friendly software than by any understanding of the data set or the critical assumptions of the statistical tests. Such frank misuse of statistical methodology and the quest to reach the mystical $\alpha < 0.05$ criteria has hampered research via the publication of incorrect analysis driven by rudimentary statistical training. Perhaps more critically, a poor understanding of statistical tools limits the conclusions that may be drawn from a study by divorcing the investigator from their own data. The net result is a decrease in quality and confidence in research findings, fueling recent controversies over the reproducibility of high profile findings and effects that appear to diminish over time. The recent development of "omics" approaches leading to the production of massive higher dimensional data sets has amplified these issues making it clear that new approaches are needed to appropriately and effectively mine this type of data. Unfortunately, statistical education in the field has not kept pace. This commentary provides a foundation for an intuitive understanding of statistics that fosters an exploratory approach and an appreciation for the assumptions of various statistical tests that hopefully will increase the correct use of statistics, the application of exploratory data analysis, and the use of statistical study design, with the goal of increasing reproducibility and confidence in the literature.

Contents

E-mail address: michael_marino@merck.com.

0006-2952/$ – see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.bcp.2013.05.017

1. Introduction

The discipline of statistics provides a logical and mathematical framework for the collection, organization, analysis, interpretation, and presentation of experimental data. It is used to analyze experimental outcomes and determine the likelihood that an outcome in a sample population is predictive of the population from which the sample was derived, e.g. to establish the efficacy and safety of a new chemical entity (NCE) in a sample human test population that will allow its broader use in a general population.

There has been growing concern that much of what is published in both the preclinical and clinical literature is misleading, resulting in the creation of a "house of cards" that undermines the core values of the biomedical research enterprise through key findings that cannot be replicated [1,2]. While there are multiple deficiencies that may underlie these shortfalls, the current diminution in the understanding and appropriate use of statistical methodologies can only lead to additional problems. When used with appropriate insight and practical experience, statistics is the *sine qua non* of biomedical research. However, with well-meaning albeit misguided biomedical researchers driving toward the goal of statistical significance, often analyzing inappropriately derived data sets and switching between statistical tests until they get the expected "right" result, it is not surprising that there is a loss of faith in the literature.

Perhaps the greatest hurdle that must be overcome in restoring confidence in research findings is the misunderstanding of what statistics is. To some researchers [3], statistics is believed to be an ephemeral science and the rejection of the appropriate use of statistical design and analysis is defended with quotes like "There are three kinds of lies: lies, damned lies, and statistics" (attributed to both Benjamin Disraeli and Mark Twain). To others, statistics is the right tool for the wrong job being applied when convenient to support a favored theory. For these researchers, "statistics are like a bikini. What they reveal is suggestive, but what they conceal is vital" (attributed to Aaron Levenstein [4]). Rather than an ephemeral science or a convenient multi-tool, statistics is a practical science, encompassing not only data analysis but also the actual design of the experiments used to generate that data.

2. Background

The basic concept of the *scientific method*, a theoretical framework for conducting scientific inquiry, can be found as early as 400 BC in Greek and Chinese texts. In practice, the scientific method involves an iterative testing and modification of hypothesis in order to extract knowledge (Fig. 1). Hypotheses are generated based on previous investigation or knowledge of the subject under investigation, experiments are designed to test this hypothesis, and the results are interpreted and used to modify the original hypothesis generating a new hypothesis for further testing. Within the context of the scientific method, the experimental design process and the interpretation of results are the domain of statistics. *A statistic* is a quantity calculated from a set of data. For example the mean of a set of numbers is a statistic. Statistics (not to be confused with the plural of statistic) is a broad term that encompasses all quantitative aspects of data collection, interpretation, and presentation. The appropriate use of statistics is essential to insure the best methods are used to collect data in an

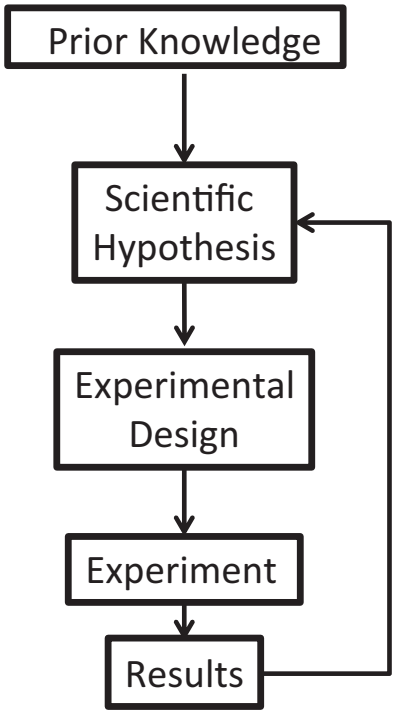


Fig. 1. Flow chart representing the primary steps in the iterative process known as the scientific method.

unbiased fashion and in manner that will allow for a transparent conclusion, and that the analysis of the collected data is done in a fashion that takes into account the properties of the underlying population and the various inherent assumptions regarding the data set. The scientific method contains a central philosophical assumption that a truth exists and that the truth is amenable to investigation. In other words, a well-designed and conducted scientific study should provide a solid confirmed fact that will stand up to tests of reproducibility within the limits of a scientist's ability to duplicate the experimental parameters. Unfortunately, several high profile reports in recent years have called this seemingly incontrovertible assumption into question [5–8].

A recent study conducted on several inbred strains of mice and a $5HT_{1B}$ receptor knockout mouse examined behavior in 6 tests across 3 laboratories and found that despite heroic attempts to standardize the experimental conditions there were significant differences in outcomes [9]. The findings were particularly disturbing in that no source of discrepancy could be identified and the difference in outcomes could only be attributed to very subtle unidentified differences between laboratories. The issue is not confined to preclinical research. Disturbing trends in the clinical testing of antipsychotic medications have found that the effect size of treatment has decreased over time while the placebo effect is increasing [10]. The loss of effect size and subsequent contradiction of findings extends to other areas of clinical research [6] with contradicted claims persisting in the literature despite obvious contrary findings [8].

While it is easy to ignore seemingly fantastic claims made in the popular press that "the truth wears off" [7], the more formal treatment explaining "why most published research findings are false" [5] is harder to dismiss. And to ignore these difficulties is to incur significant cost in wasted public funding of academic research and private support of drug discovery and costly clinical trials. As important is the irreplaceable cost of reduced confidence in the scientific literature. The ability of an investigator to develop new approaches and informed hypotheses based on published research findings is essential to progress in any scientific field. While replication of findings is a critical part of the scientific process, an apparent need to mistrust the literature places an undue burden on the individual investigator who may not have the resources to replicate every key piece of data that goes into the formulation of a new hypothesis.

While there are many factors contributing to the current state of affairs, one factor that is within the control of every scientist publishing data and available for criticism by every reviewer or consumer of scientific literature is the use of statistics in the design and interpretation of experiment. The trend in relegating a solid statistical education to the social sciences and requiring biologists and chemists to have only the most rudimentary of statistical training has taken a clear and significant toll on the quality of science. This effect has been further amplified by the development of a variety of high throughput "omics" approaches that generate data sets which most researchers are unable to interpret, and in some cases for which the statistical methods for inference have not yet been fully developed.

The aim of this commentary is to address two issues in the use and misuse of statistics that contribute to the problems of reproducibility and decreased confidence in the literature. The first is a frank misuse of statistical methodology often fueled by the ease of generating inferential statistics using modern software. Common examples of this sort of misuse are the repeated analysis of a data set with different inference tests until the desired result is obtained, and the use of multiple *t*-tests to evaluate significance in a complex multiple comparisons design without correcting for the multiple comparisons. The second issue has more to do with an overreliance on inferential statistics and the mystical $\alpha < 0.05$ criteria. It should be apparent to anyone with even a rudimentary understanding of statistical inference that there is nothing magical about the number 0.05 [11], and that biological relevance and statistical significance may not always be the same thing. A robust, albeit non-significant trend in the data may represent a biologically relevant effect, however the design of the experiment is such that even an appropriate test of inference will fail to demonstrate significance with $p < 0.05$. If, for example, a trend exists and $p = 0.06$, is reported as "no significant effect (data not shown)" it is not an accurate description of the biological relevance, but rather a misdirection for future investigators. While the fact remains that a negative result cannot be interpreted, it can certainly be examined.

In the following sections an attempt is made to provide a balance between technical information on statistical methodologies and the development of an intuitive understanding of statistical methods. For much of the basic computational methodologies or definition of basic terms (e.g. mean, variance, etc.) it is beyond the scope of this commentary to provide primary references, the reader is directed to any of several modern statistical texts for additional information [12–15]. By guiding the reader to a more intuitive understanding of the statistical methodologies employed in pharmacology research, it is anticipated that statistics will be used correctly more often and that this will enhance reproducibility and, as a result, confidence in the literature. It is also hoped that enhanced understanding will be applied to exploration of existing data sets and thoughtful design of future studies such that the right questions are being asked at the outset of the investigation. In the oft-cited words of John W. Tukey, "It is important to understand what you can do before you learn to measure how well you seem to have done it"[16].

3. The concept of sampling

Sampling is at the core of scientific investigation. Inherent is almost all experimental design is the reality that only a portion of the true population can be sampled, and that generalizations regarding the whole population must therefore be derived. On rare cases it may be possible to sample the entire population. For example, it is possible to determine a preference for wine or beer amongst the current members of the Pittsburgh Orchestra by canvassing every current member. It is then theoretically possible to determine the same preference amongst the current population of Pittsburgh, although highly impractical. Likewise, it would be impossible to determine the sensitivity of every dorsal root ganglion (DRG) cell to acetylcholine. However, a random selection of DRG cells may provide enough of a data set to derive a conclusion regarding the population of all dorsal root ganglion cells.

3.1. Sample bias

In taking a sample of the population, an obvious assumption is made that the resulting sample is representative of that population. *Sample bias* occurs when a sample is skewed away from the population either by errors inherent in the experimental design, or by hidden assumptions that require additional experimental work to uncover. To use the above example, the Pittsburgh Orchestra is certainly a sample of the population of Pittsburgh. Furthermore, it represents a very tractable sample for the proposed study since all of the members can be present in one room and can thus be easily polled. Despite that fact, very few would agree with the assertion that the preferences of the larger city population could be predicted from this clearly biased sample. A random sampling based on surveying representatives across the entire city would be required for any meaningful generalization of the findings. To return to the example of DRG cells, choosing cells based on ease of

dissection, somatic size, or viability in culture all may lead to a more tractable experimental design, but risks introducing a significant sampling bias that would be absent in a purely random sample.

3.2. Statistics and parameters

The above examples lead easily into the definition of population *parameters* and *statistics*. A parameter is measured across the whole population by examining every member. For example, measuring the locomotor response of every Sprague Dawley rat on the planet in response to a 1 mg/kg s.c. dose of amphetamine would allow the calculation of the average response of the population as a parameter. Parameters are a true characteristic value of the population and there is no error associated with the measure. Statistics are estimates of population parameters inferred from sampling the population (Fig. 2). It is very rare that a researcher actually has access to parameters (e.g. every Sprague Dawley rat on the planet cannot be tested on a routine basis), and in the field of pharmacology it is practically impossible to obtain parameters. Therefore, all conclusions drawn from pharmacological experiment are founded on an underlying assumption regarding the sample chosen to represent the population. Great care must be exercised in designing experiments to avoid sample bias and to choose appropriate statistical methods in order for inference regarding the population to be valid. This last point is of particular importance when considering studies that when retested, fail to replicate.

3.3. Sampling and non-sampling error

Since a *sample* is by definition not the whole population, it is unlikely that a sample statistic will be exactly equal to a parameter. To use the example above, since the response of every Sprague Dawley rat on the planet cannot be measured, the average response in a sample of the population must be used as an estimate of the whole population. The difference between the actual population parameter (e.g. response of every rat) and the measured statistic (response of every rat in the sample) is the *sampling error*. It should be apparent that sampling error cannot be exactly determined unless the whole population can be measured to determine the parameter. It is, however, possible to employ probabilistic modeling methods to estimate sampling error. For most day-to-day experimental design and statistical evaluation it is sufficient to understand that the sampling error exists. Sampling error occurs by chance as samples are drawn from the population. It cannot be avoided.

A distinction can be drawn between the concept of sampling error and *non-sampling error*. Non-sampling errors are not the result of chance, but are actual errors in study design. First and foremost among non-sampling errors is the issue of sample bias discussed above (Section 3.1). This can be more formally stated as the sample being a non-random selection of the population. The above example of the DRG cells chosen based on experimental tractability demonstrates how a seemingly random selection can in fact be made non-random by the introduction of an experimental design flaw. Non-sampling error can also be described as a systematic error arising from experimental technique (e.g. poorly calibrated pipettes) or from some more inscrutable source like a periodic fluctuation in the electrical current caused by a subway train passing by and placing a large load on the local portion of the electrical grid (an event actually encountered at Temple University in the early 1980s). While it is difficult to completely eliminate non-sampling error, and impossible to alter sampling error, any conscientious researcher needs to know enough about potential sources of error to minimize discrepancies between the population parameter and the derived statistic. As noted above, in almost all cases in pharmacology, the data set available is a sample since it is impossible to canvas any population in its entirety and establish a parameter. This reliance on statistics rather than parameters leads directly to a requirement for careful exploration of the available data to better understand the characteristics of the underlying population, and to the need for methods of inference to test hypothesis.

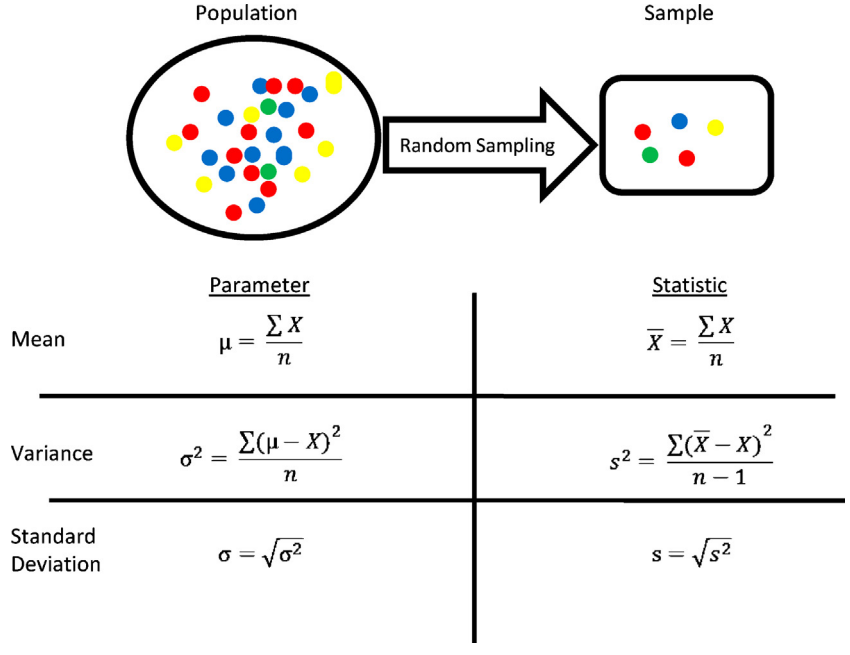


Fig. 2. Figure depicts an overview of the concept of sampling and the derivation of parameters and statistics. Random selection of a sample from the population provides a representative estimate of the true population parameters. Note the Bessel correction $(n-1)$ in the denominator of the sample variance equation which corrects the underestimate expected in the sample population.

3.4. Degrees of freedom

The concept of *degrees of freedom* (df) is one that many investigators find difficult to grapple with. Fortunately the problem is typically based in an expectation that the concept is more complex that it really is. Degrees of freedom are quite simply the number of values that can be chosen freely, or that are free to vary. It can also be thought of as the number of pieces of information available for use in estimating a population parameter.

A concrete example of df will facilitate understanding of this concept. Consider the choice of a set of 3 numbers, the sum of which is 100. The first number chosen can literally be any number as this value is free to vary. Assume the first number chosen is 75. The next number is again free to vary and can be chosen freely from the set of real numbers. However, once the second number is chosen, the third number is instantly fixed by the requirement that the sum equals 100. Choose 10, and the third number must be 15, choose 30, and the third number must be −5. Therefore, once the sum (and by extension the mean) of the sample is calculated, a df is lost.

In statistical estimation of parameters, the concept of df plays a central role in that each parameter estimation will lead to a loss of a df. For example, for a sample size of $n = 2$, the mean can be calculated, but it is not valid to calculate a variance or standard deviation because once the mean is determined only one value is free to vary. The importance of this will become evident during the discussion of the Bessel correction of variance and standard deviation below. The df value is also critical in that it is a characteristic value of some sampling distributions such at the t distribution.

4. Descriptive statistics

Once a sample is collected and exists as a finite string of numbers, methods are needed to summarize the sample population. Two types of values are needed, a *measure of central tendency* that indicates the approximate center of the sample distribution, and a *measure of dispersion* that indicates the degree to which individual members of the sample set depart from the central value.

4.1. Measures of central tendency

Measures of central tendency are primarily ways to represent a distribution of values (i.e. a frequency histogram) by a single typical value. Based on the above discussion around sampling it is obvious that any attempt to represent a sample population intended to approximate a population by a single statistic is futile since the sample statistic can only approximate the population parameter. However, a measure of central tendency will provide a characteristic of the sample population that is a typical value and is useful when provided with additional statistics such as a measure of dispersion discussed below. While the computational steps necessary to derive the various measures of central tendency are likely well known, a brief discussion and some examples are warranted to point out the shortcomings of these measures and to help develop a more intuitive understanding.

4.1.1. The mean

The *mean*, or average, is one of the most commonly used statistics in scientific research. It is simply the sum of all observations divided by the number of observations (Fig. 2). When thinking in terms of a frequency histogram, the mean can be visualized as the center of mass of the histogram, or the point (fulcrum) at which the histogram would balance (Fig. 3). The mean

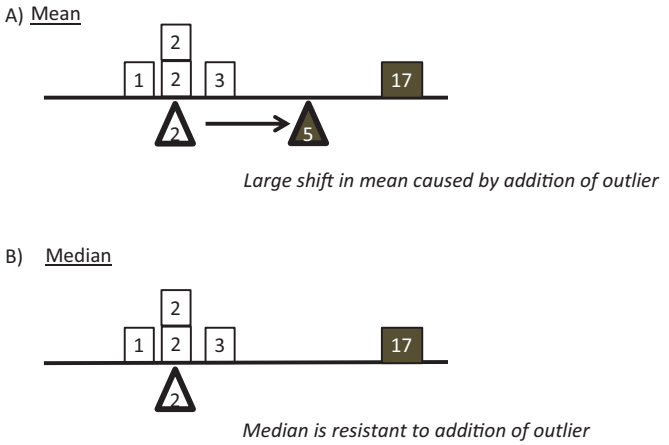


Fig. 3. A graphic depiction of the mean and median as measures of central tendency. (A) Note that the mean may be viewed as the balance point of the distribution (triangle). When an outlier is added to the distribution (shaded box) a relatively large shift in the mean is necessary to maintain balance (shaded triangle). (B) The median of the starting distribution is, in this case, identical to the mean. However, the addition of the same outlier produces no change in the median value. Therefore the median provides more of a measure of the center of the distribution than a balance point, and is less affected by outliers than the mean. It is also notable that the standard deviation of the original distribution is 0.8, but jumps to 6.7 with the addition of the outlier. The interquartile range is 1.75–2.25 for the original dataset and 2–3 after the outlier addition.

is the most commonly used measure of central tendency likely because it is a rather intuitive value. It is easy to imagine the values in a distribution balancing on some point, or to think about the center of a symmetrical distribution. However, the mean is not always the best choice for characterizing a distribution with a single number as it is highly sensitive to outliers. To think of this in physical terms, the line balancing on the mean point becomes a lever if a value is added far beyond the current mean, and the fulcrum point must move far from the original position to balance the histogram (Fig. 3). In this event, and assuming the previous mean value represented a close approximation of the population mean, many more observations would have to occur close to the population mean to restore the balance.

4.1.2. The median

The *median* is defined as the middle term in a data set arranged in rank order. It is used is far less commonly than the mean, however this is unfortunate in that the median can be quite useful especially in describing a data set containing outliers. The median is determined by rank ordering all the members in a data set and choosing the middle term for which 50% of the values lay above or below. This method provides a single number for data sets containing an odd number of values, and a pair of numbers that must be averaged in order to determine the median for a data set containing an even number of values. For example, the data set (1, 2, 5) has a median value of 2, while the data set (1, 3, 5, 6) has a median value of $(3 + 5)/2 = 4$.

Some wrongly view the median as a less rigorous measure of central tendency, in part because the computational process does not appear to take into account every sampled value in the distribution. However, by again taking a graphical approach, the median can be visualized as being the center of the histogram with an equal number of values on either side (Fig. 3). Therefore the location of every value in the histogram does come into play in determining the median.

Despite the superior resistance of the median to outliers, most common methods of inference rely on the mean as a measure of central tendency. However, the median provides an excellent

summary statistic. As discussed below under exploratory data analysis, the importance such tools that provide insight into the data behind the statistic should not be underestimated.

4.1.3. The mode

The mode is simply the most commonly occurring value in the sampled distribution. The mode is not often useful in pharmacological research but is mentioned for the sake of completing the list of measures of central tendency. One interesting fact regarding the mode is that while any distribution will have a single mean and a single median, it will not necessarily have a single mode. A distribution could actually have no mode if none of the values repeat, and could have more than one mode (e.g. a bimodal distribution).

4.2. Measures of dispersion

Collapsing a set of observations to a single value (i.e. determining the mean) provides a convenient summary of the distribution of values. It does not, however describe the distribution beyond the point of central tendency. If a given concentration of 5-HT produces on average a 50% decrease in cAMP accumulation in a given cell type, then it is expected that the next time this concentration of 5-HT is applied to that cell type the measured a response that is likely to be close to 50%. The key question however is how likely? How surprising would an observation of 70% be? And in the context of experimental design and inference, would a 30% response in the presence of a putative antagonist be interpretable as an inhibition, or be expected as normal variation around the mean? It is therefore essential that a measure of central tendency be coupled with a measure of dispersion which describes the spread of the sample values about the point of central tendency in order to provide an interpretable summary.

4.2.1. Range, variance, and standard deviation

The basic goal of calculating a measure of dispersion is to generate a description of how far the data are spread out about the mean. The simplest and most intuitive measure is the *range*, which is simply the difference between the largest and smallest value. While the range is simple to understand and easy to compute it carries the limitations of being based on only two numbers in the distribution and being highly sensitive to outliers. Therefore it is almost always better to employ a more robust measure of dispersion such as the variance or standard deviation.

The *variance* provides a reasonable estimate of the spread in a distribution, and the calculation of variance is relatively intuitive. If the mean represents the center of the sample distribution, the spread in the distribution can be measured by assessing how far each individual value differs from the mean. It would initially appear that an average difference would provide a useful measure of dispersion, however some thought and sample calculation will quickly demonstrate the problem inherent with this approach. The sum of differences will always be zero because the mean lies at the center of the distribution. This can be obviated by squaring the values of the individual differences and calculating a sum of squares which will then provide a measure of spread in the data. This sum of squares value is then used to calculate the variance – defined as the average of the squared differences from the mean.

Despite the simplicity of the concept of variance, there are different computational methods for calculating population and sample variance (Fig. 2). The difference being that when calculating sample variance (which is what is almost always done in pharmacology) a correction factor known as Bessel's correction must be applied such that the average is determined with a denominator of $n-1$. The easiest and perhaps most intuitive explanation for this correction factor is that one df has been used in calculating the mean, and this is accounted for by subtracting 1 from the sample size. This is the explanation most commonly found in introductory statistics texts, and while not incorrect, it is a simplification of the true reason which has to do with the desire of obtaining an unbiased estimate of the population variance. In general it is assumed that values in a sample are closer to the sample mean than to the population mean, and the variance would therefore be underestimated without correction. A complete explanation of the correction factor involves the concept of maximum likelihood estimates of the distribution parameters that is beyond the scope of the article. The df explanation is sufficient for an intuitive understanding of variance.

The sample variances provide an unbiased estimate of population variance, however it has units that are squared relative to the units of the mean value making it difficult to interpret. For example, measuring a mean distance and reporting the result as $125\ \mu m \pm 20\ \mu m^2$ does not intuitively provide an understanding of the accuracy of measurement. For this reason, the *standard deviation* (sd) has become the most commonly employed measure of dispersion, and it is simply defined as the square root of the variance (Fig. 2). In general, there is no correction of bias for the calculation of standard deviation. The sample standard deviation is calculated from the corrected sample variance; however the square root introduces nonlinearity, and therefore a bias into the calculation.

For non-parametric datasets where the median is used as a measure of central tendency, the interquartile ranges or hinges provide a useful non-parametric measure of dispersion. This is explained in more detail below under exploratory data analysis.

4.2.2. Standard error of the mean

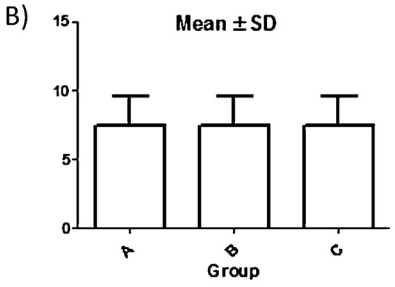
The *standard error of the mean* (SEM) is a very commonly used as a measure of dispersion in the biological sciences. Despite its common use, it is generally misunderstood and used more by convention that for logical reasons. In fact the SEM is NOT a measure of dispersion and does not tell the reader anything regarding the scatter of data about the mean. The fact that the SEM will always be smaller than the standard deviation may, in part, be responsible for its use in graphical representation of data providing an illusion that the measurements are more precise than they actually are, however this is clearly not a justifiable reason to choose the SEM over the standard deviation.

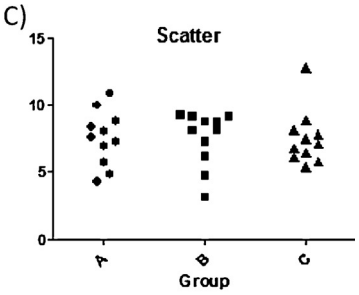
The concept of *sampling distributions* is central to an understanding of the SEM. As indicated above, population parameters are actual and exact values that are characteristic of a population. Statistics are estimates of parameters obtained by sampling the population. Therefore statistics behave as variables and exist in distributions known as a sampling distribution. If a population is repeatedly sampled with a constant samples size, the mean of each sample set would likely be different each time. In other words, when an experiment is run (i.e. sampled) 3 times, it would be unexpected to get the exact same value each time despite the fact that the population parameter that is being estimated with each run is an exact value. This collection of sample means derived from multiple runs of the same experiment represent a sampling distribution of values having a mean (the average of the sample means) and standard deviation (the standard deviation about the sample means). Importantly, this is true for any statistic, not just the mean. There is, for example, a sampling distribution of the standard deviations. For this reason, referring to the SEM by the colloquial "standard error" is incorrect for it does not indicate from which statistic the error is derived.

The SEM is a measure of how accurately the population mean has been estimated based on the central limit theorem (CLT). A full discussion of the CLT is beyond the scope of this article, and it is sufficient to understand that the CLT rigorously describes how sampling distributions behave relative to the actual population.

A)

	Group A	Group B	Group C
	4.26	3.10	5.39
	5.68	4.74	5.73
	7.24	6.13	6.08
	4.82	7.26	6.42
	6.95	8.14	6.77
	8.81	8.77	7.11
	8.04	9.14	7.46
	8.33	9.26	7.81
	10.84	9.13	8.15
	7.58	8.74	12.74
	9.96	8.10	8.84
Mean	7.50	7.50	7.50
SD	2.03	2.03	2.03





D)

	Group A	Group B	Group C
Maximum	10.84	9.26	12.74
Minimum	4.26	3.10	5.39
Upper Hinge	8.57	8.95	7.98
Median	7.58	8.14	7.11
Lower Hinge	6.32	6.70	6.25

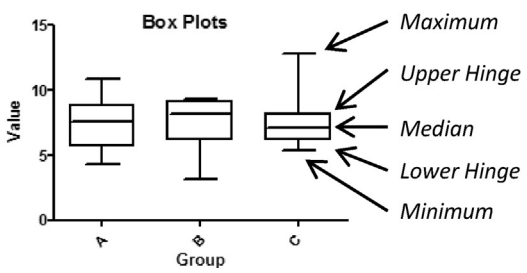


Fig. 4. Example data set for exploratory data analysis and descriptive statistics. (A) An Anscombe [50] example of 3 unique data sets which have identical means and standard deviations. (B) The standard bar graph (mean ± SD) summary of these 3 groups suggests that the groups are identical. Many researchers make the mistake of stopping at this point. (C) Simple scatter plots begin to reveal unnoticed differences in the data. (D) A 5 number summary makes it clear that there are differences in the distribution of the data despite the identical mean and standard deviation. (E) A box plot provides a graphical depiction of the 5 number summary and allows for a quick visual check on potential differences in the distribution. At this point in the analysis it is apparent that there is a skew in the distribution of the data in group B, and that one point in Group C may be an outlier.

The SEM is calculated by dividing the standard deviation of the mean of the sampling distribution of the means by the square root of the sample size. Therefore the SEM will decrease as the sample size increases and a more accurate estimate of the mean is obtained. In addition, the SEM will always be less than the standard deviation. Importantly, the SEM is never subjected to a Bessel correction (i.e. $n - 1$). This is a common error that unfortunately is rarely noticed by investigators and impossible to correct by reviewers.

Note that a predictable relation exists between the SEM and sample size in that as sample size increases, the SEM will always decrease independent of the true population variance. This is not true of the sample variance which is limited by the true population variance. Therefore it should be obvious that the SEM is not in any way a reliable estimate of dispersion. It can be argued that the use of SEM as an indication of the accuracy by which the mean has been determined means that using the SEM in a graphic depiction of data (e.g. a bar graph) allows for a visual estimate of significance. While there is some truth to this argument, it is far better to make such determinations based on probability values derived from inferential methods described below. The SEM is not a descriptive statistic and should not be used as one [17].

5. Exploratory data analysis

In his landmark book "Exploratory Data Analysis" John W. Tukey drew a sharp distinction between what he called confirmatory (inferential) and exploratory analysis and describes exploratory data analysis (EDA) as the detective work that uncovers the clues for the jury of confirmatory data analysis, or the process of finding the right question before looking for the right answer [16]. While Tukey's book is a somewhat dated reference (the focus on pencil and graph paper may seem quaint in the age of ubiquitous and inexpensive computers) it is very accessible and in emphasizing the need to spend time reviewing raw data to discover what the results actually indicate was both logical and prophetic.

While experimentation should typically follow the scientific method (i.e. be directed at making a planned comparison based on a clearly defined hypothesis), the common practice of dumping strings of numbers into a spreadsheet and extracting means, standard deviations, and *t*-test *p*-values, discards most of the information available in the dataset. Furthermore, this type of automated analysis promotes a distance from the data that makes it impossible to assess whether the assumptions of the various inferential methods discussed below have been met. Most importantly, a more considered approach to EDA coupled with a rediscovery of the lost art of running pilot studies can uncover otherwise unnoticed trends and relations in the data that may lead to new hypothesis and new research directions.

EDA is more of a conceptual or perhaps even philosophical approach to handling data than a strict subfield of statistics. The methods that make up the toolbox of EDA can be as simple as generating a scatter plot to running complex autocorrelation analysis on a time series. The goal of this section is to outline a few simple tools for looking at data sets in a pictorial fashion that provides additional insight. While these tools may appear familiar, they are unfortunately used infrequently being supplanted by bar graphs depicting the mean and standard deviation (or worse the SEM) that only reveal what the researcher already knew, making its value minimal compared to the visualizations described below. It is expected that the reader will discover or develop additional tools uniquely suited to their own work once the value of the EDA approach becomes evident.

5.1. Scatter plots

Other than a direct listing of values, there is no simpler representation of a data set than a *scatter plot*. Whether plotting observations against an independent variable, or simply plotting individual observations over time in a run sequence plot, the scatter plot can reveal structure in the data that indicate

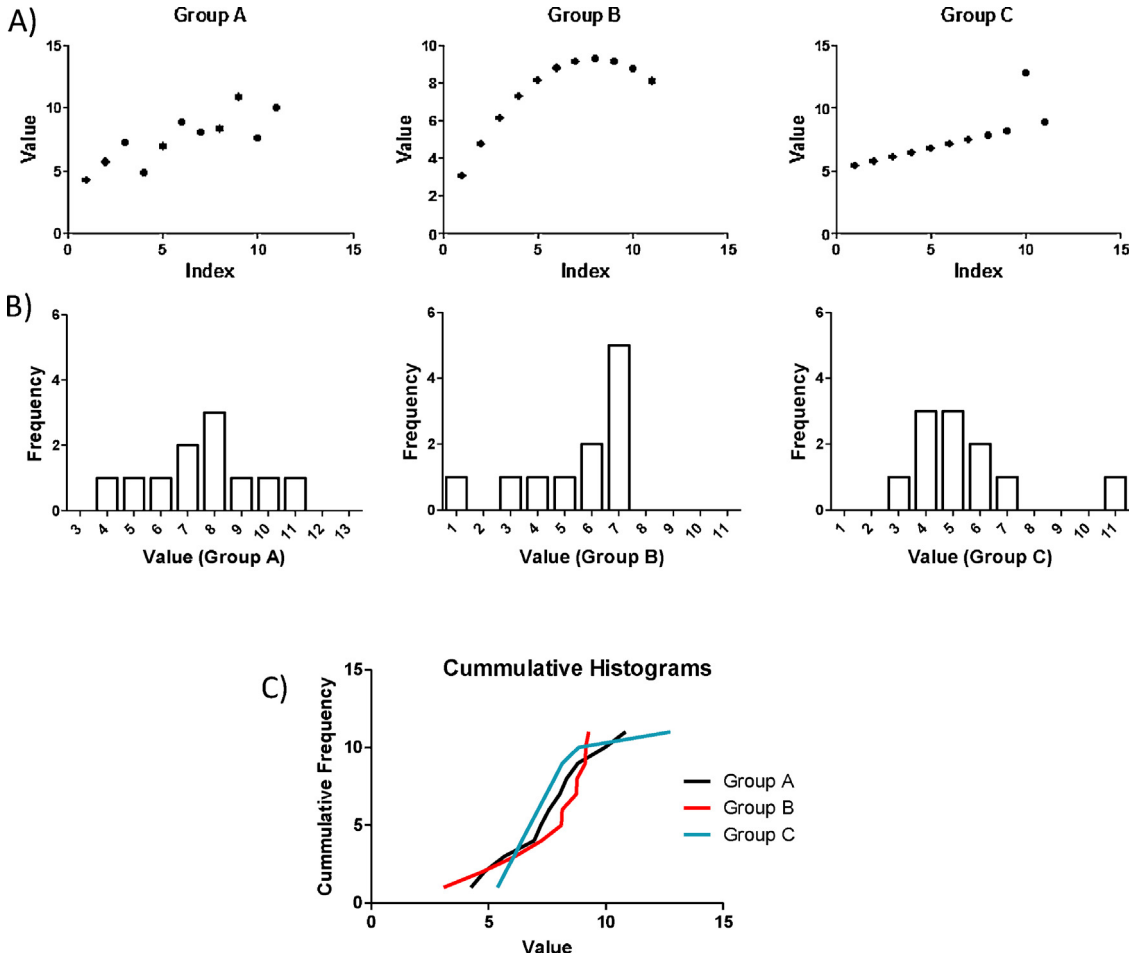


Fig. 5. Further exploratory analysis distinguishes features not apparent otherwise. (A) Run-sequence plots of the data from Fig. 4 reveal clear differences in the behavior of these data sets. Group A appears to follow a roughly linear increase, while Group B exhibits a nonlinear, non-monotonic rise in value. The outlier expected in group C based on scatter and box plots is now obvious. (B) Frequency histograms and (C) cumulative frequency histograms can provide additional visual insight into the likely hood of distribution being normally distributed, and can indicate otherwise subtle differences amongst the distributions.

associations and provide information regarding the distribution of variables. Side by side scatter plots can be quite valuable for comparing data sets and indicating biasing outliers that may lead to misleading summary statistics (Figs. 4c and 5a). All graphing software packages and spreadsheets are capable of producing scatter plots and the generation of these simple visual depictions of data should always be the first step in data analysis.

5.2. Frequency histograms

Frequency histograms are constructed from a table of frequencies by plotting the values obtained from observation on one axis, and the frequency of that observation on the other. Most graphing software provides the capability to generate frequency histograms in a bar format (Fig. 5b) and functions exist in spreadsheet programs that can be used to generate these plots in a few steps.

Frequency histograms also provide a convenient method of visualizing two data sets simultaneously. For example, observations from a vehicle and a compound treatment group can be plotted together to look for shifts in the distribution. For making this sort of comparison, a valuable variation on the frequency histogram called the *cumulative frequency histogram* can be used. The cumulative frequency histogram plots the summated frequency, or when frequencies are normalized to the total number of events, plots the cumulative probability. This type of visualization plotted in line format can identify subtle differences in distributions that may not be as evident in standard frequency histograms (Fig. 5c).

5.3. Five number summaries and box plots

As noted above (Section 4.1.1), the mean is the most commonly reported statistic and one that tells the least about the distribution. The five number summary (Fig. 4d) provides a concise and accurate description of a data set with information regarding the shape of the distribution using statistics that are easily calculated. The five numbers are the highest and lowest values, the median, and the upper and lower quartiles, or hinges. The quartiles are analogous to the median but taken at the 25% cut in the data set. So the lower quartile is the value that has 25% of the data below and 75% of the data above, while the upper quartile has 75% of the data below and 25% of the data above.

The box plot or stem and whisker plot provides a simple visual representation of the five number summary (Fig. 4e). The shape and spread in the distribution is immediately obvious in these plots, and when groups are plotted side by side, overlap in the distributions can be easily assessed. Most graphing packages will produce a box plot, and spreadsheet programs can easily be coerced into producing a five number summary. The value of these depictions of a data set should be intuitively obvious, and it is difficult to conceive of a situation where this type of visualization would not increase understanding of a data set.

5.4. Outliers

While not strictly a component of EDA, the identification and assessment of *outliers* clearly grows from a visual depiction of datasets using EDA methods. An outlier is a data point that has an extreme separation from the rest of the data set. Unfortunately there is no clear-cut mathematical description of an outlier leaving the researcher with a subjective task in deciding whether or not to exclude a data point. The liberal approach of simply excluding any value that does not seem to "fit" clearly introduces a bias and produces an inaccurate if not dishonest summary of the results. However, outliers do exist. They may be technical outliers, the product of faulty method, equipment malfunction, etc., or they may be true statistical outliers. In a normally distributed population, approximately 68% of values lie within one standard deviation of the mean, 95% lie within 2 standard deviations from the mean, and 99.7% of values lie within 3 standard deviations from the mean. Therefore, there is a real probability (1 in 22) that a random observation will lie further than 2 standard deviations from the mean. If the dataset is composed of a relatively small number of observations, the presence of an outlier will have obvious consequences. Therefore, the conservative approach of never excluding outliers will produce as much bias and inaccuracy as the liberal approach. The only solution is to have a consistent and quantitatively justifiable method for identification and exclusion of outliers such as those described below.

Technical outliers arise from non-sampling systematic error in experimental technique. If the source of the systematic error is known, then technical outliers should be automatically excluded from further analysis. For example, a cohort of animals is found to respond differently to a compound treatment on a day when heavy construction is occurring outside the laboratory, or results in an *in vitro* study appear different after an enzyme preparation is improperly stored for several hours at room temperature before being used in the assay. The need to eliminate technical outliers highlights the need to take careful and comprehensive notes during experimentation. However, despite best efforts, at times technical errors may occur due to unknown or unidentified sources of systematic error. This type of outlier will be added to the set of statistical errors and must be dealt with by quantitative means as described below.

Statistical outliers are expected to occur at non-negligible frequency as described above. A number of methods have been suggested to identify statistical outliers including simple screens that eliminate any values more than 2 standard deviations from the mean, or other more complex statistical criteria [18], and statistical tests such as Grubbs' test [19] or Dixon's *Q* test [20]. Of these, Dixon's *Q* test is best suited for the relatively small data sets typical of pharmacology. This test makes fewer assumptions and is relatively easy to calculate. Regardless of the method chosen, it must be applied consistently across all data sets, must only be used once on a given data set, and should always be reported along with the excluded value when data are published.

6. Inference

The focus of this article thus far has been on descriptive statistics and the exploration of data sets. The scientific method is based on empirical observation and the testing of hypothesis in order to draw *inference* from the data sets. The remainder of this commentary will be focused on such methods of inference. One of the greatest misuses of statistical methods in pharmacology involves the choice of an inappropriate statistical method of inference (statistical test) based on a lack of understanding of the underlying assumptions of the methods.

	Fail to Reject H_0	Reject H_0
H_0 = True	Correct Choice	Type I Error α Error
H_0 = False	Type II Error β Error	Correct Choice

Fig. 6. This matrix defines the two types of errors that can be made in hypothesis testing. See text for further discussion. H_0 = Null hypothesis.

At the heart of inferences lies the concept of hypothesis testing with the goal of determining the confidence by which the *null hypothesis* may be rejected. The null hypothesis is a default statement that the manipulation or difference being investigated (compound treatment, temperature, age, ethnic background, etc.) has no effect. The *alternative hypothesis* is the opposite statement. At a very basic level, hypothesis testing involves the statement of the null, and alternative hypothesis, selection of a distribution (e.g. *t* or *F* distribution), determination of the rejection and non-rejection regions of the chosen distribution, calculation of a test statistic, and decision on whether or not to reject the null hypothesis. The primary goal is to minimize Type I (or α) and Type II (or β) errors (Fig. 6) with Type I errors representing false positives and Type II errors representing false negatives.

The bulk of most introductory statistics classes are dedicated to the concept of hypothesis testing with an emphasis on probability, probability density functions, and sampling distributions, a level of detail that lies well beyond the scope of this commentary and the reader is referred to any of the introductory texts [12–15] for a detailed exposition on this topic. Here the emphasis will be on the choice of statistical test, underlying assumptions, and common misuse.

6.1. Parametric and non-parametric tests

Parametric statistical methods (i.e. methods that rely on estimation of parameters) assume that the sample distribution is normally distributed. Parametric statistics include the most commonly employed tests such the *t*-test, and ANOVA and are often incorrectly used without consideration of the normality assumption. While a number of quantitative methods exist to test how well a data set can be modeled by a normal distribution [21] it is sometimes difficult to test for normality, especially with a relatively small sample size. Fortunately most quality statistical packages are capable of testing the assumption of normality.

A more important consideration is how experimental design may actually impose restrictions on observations which make normality impossible. For example, a classical measure of catalepsy in rodents involves placing the animal in an uncomfortable posture (e.g. front paws placed on an elevated bar) and measuring the latency to first movement. For convenience the investigator will often impose a cutoff time such that if the animal does not move in a set amount of time, the study is halted and the maximum time is reported. It is apparent then that treatments which produce a pronounced state of catalepsy will produce mean values close to the cutoff time with relatively small variance. However, the sampling distribution derived from these values is unlikely to be indicative of the actual distribution, and may be significantly skewed by the cutoff. Therefore the underlying assumption of normality is unlikely to be met for these data. In

cases such as this, a *non-parametric test* would be the most valid choice.

6.2. Correlation

Correlation methods allow for the assessment of *association* between measured variables in a data set. For example, a correlation can be measured between the concentration of an endogenous enzyme in a cell culture and the phosphorylation of a cellular protein substrate. While on the surface this would appear to be a relatively simple type of data set to interpret, misinterpretation is quite common. The *cum hoc*, *ergo propter hoc* fallacy represents the most serious pitfall in correlation analysis. The fact that two variables co-vary does not logically suggest that the changes in the two variables are causally connected, but merely implies some influence between the two variables or the existence of some common underlying cause. In the enzyme example given above, it may be tempting to conclude that the correlation between substrate phosphorylation and enzyme concentration demonstrates a causal relationship between the enzyme and the substrate. However, a role of the enzyme in the metabolic well-being of the cell altering basal ATP concentrations leading to a global decrease in protein phosphorylation could be an equally plausible explanation. Another example of this sort of faulty interpretation of correlation is described by Messerli in exploring the remarkable and significant correlation between a nations chocolate consumption and the likelihood of producing Nobel laureates [22].

It is also wrong to assume that a lack of correlation demonstrates a lack of association. Correlation analysis is designed to measure associations fitting a particular model (e.g. linear association). A significant higher order polynomial relation may exist between two variables which would not be detected by common linear correlation analysis. For example, it is well known that given an excess of nutrients bacteria will exhibit exponential, or log phase growth over time. If the variables of cell number and time were correlated using a linear model without first performing a log transform of the data, the correlation would be unlikely to yield a significant association between the two variables. As always, negative findings cannot be interpreted.

Importantly, correlation analysis should only be used when assessing correlation between 2 measured variables. It is inappropriate to employ correlation analysis in an experimental design where one variable is manipulated and another is measured (e.g. a concentration response study). Likewise, regression methods should not be confused with correlation analysis. Regression methods measure goodness of fit of a model to the data, not association between variables.

6.2.1. Correlation: experimental design, assumptions, and interpretation

Studies in which two variables are independently measured should be designed with a correlation analysis in mind. It is important to assess the range of values covered by both variables in order to avoid situations where a change covering several orders of magnitude is correlated with a variable covering a much smaller scale. The methods assume that the two variables are sampled from population distributions (an assumption violated if one of the variables is an independent variable). The output of a correlation analysis will include a correlation coefficient, and a p value. The correlation coefficient ranges from -1 to 1 and indicates the strength and direction of the association. The p-value indicates how likely it is that the association could be observed by chance.

6.2.2. Parametric correlation

The Pearson product-moment correlation coefficient (r) [23] is employed to assess association between two variables. It assesses linear dependence, and assumes that both variables are sampled from populations which are normally distributed. This assumption is particularly important for small data sets (e.g. $n < 10$). In addition to the requirement for normality, r is also sensitive to outliers.

The output of a correlation analysis using the Pearson method should include the correlation coefficient, a p value that tests the null hypothesis that $r = 0$, and the coefficient of determination (r^2). Interpretation of the value of r is not strictly quantitative; however the sign of r indicates the direction of association. In general the further the r values lay from 0, the stronger the association. The value of r provides a measure of covariance between the two variables normalized to the product of their standard deviations, and squaring r to obtain the coefficient of determination provides a measure of the proportion of the variance that can be accounted for by the association of the two variables. For example, an r^2 value of 0.8 indicates that 80% of the variance is shared by the two variables.

6.2.3. Non-parametric correlation

The *Spearman rank correlation coefficient* (r_s) [24] measures association between two variables, but differs from the Pearson product-moment correlation coefficient in that it assesses monotonic (i.e. not necessarily linear) dependence. In practice, r_s is calculated as r but between the ranked variables. The Spearman method makes no assumptions regarding the distribution of the populations from which the variables are drawn and can assess nonlinear monotonic relationships. Therefore, it is the preferred measure for small data sets, when the assumption of normality is violated or not assessed, or when the relationship appears to be nonlinear. The interpretation of r_s and the p-value are similar to that discussed above for r. Because ranked values are used in the calculation of r_s, the coefficient of determination cannot be calculated for r_s, therefore, only the value of r_s and not r_s^2 should be reported.

6.3. Hypothesis testing

The choice of a statistical test of significance is not something to be done at the end of the experiment. This approach often leads to overly complex design and frank misuse and misinterpretation of results. If the researcher begins with a toolbox of statistical methods and designs the experiments with these methods in mind, the results are easier to interpret and assumptions are less likely to be violated.

6.3.1. Experimental design for two independent groups

The simplest experimental design is the comparison of means between two independent groups. For example, a single concentration of a test compound is applied to cells in culture and the resultant effect on cell viability is measured. The second group is comprised of cultures treated with a control vehicle. The null hypothesis is stated as the application of test compound produces no effect on cell viability, and the test is designed to assess how likely it is that the observed difference in cell viability can be accounted for by chance.

6.3.1.1. Parametric two independent groups: assumptions and interpretation. Student's t-test [25] is the first choice for comparing two independent groups. The t-test has several underlying assumptions that are critical to understand. It assumes that the means were drawn from normally distributed populations and are truly independent (i.e. no repeated measures). The test also assumes *equal variance* exists between the two groups. This assumption is often violated measuring effects that are subject to a "ceiling" or "basement." For example, a treatment that moves a

necessarily positive value toward zero may also produce a smaller variance by virtue of the fact that the value cannot extend below zero. Fortunately an adaptation of Student's *t*-test that is robust in the face of unequal variance called Welch's *t*-test [26] can be used in these circumstances. A final and critical assumption of the *t*-test that is very commonly violated is that of *independence of error*. As described above (Section 3.3), sampling and non-sampling errors are the causes for a given data point to be different from the population mean. The assumption of independence of error requires that all of the factors contributing to these errors be independent for each data point. Therefore, triplicate measurements from a single 96 well plate or triplicate lanes from a single gel cannot be used in a *t*-test as an $n = 3$ case because the factors contributing to the non-sampling error for each set of triplicates are not independent. This necessitates running the study 3 times, and averaging each of the triplicates which would provide 9 observations, but the triplicate values would be averaged to produce a final $n = 3$. It will become obvious throughout the remainder of this section that the assumption of independence of error is important for all hypothesis testing and should never be violated despite the use of modern high throughput technologies.

The calculations involved in performing the *t*-test will provide a *t*-ratio and df. These values are then used to obtain a *p*-value. The *t*-test can be run as either a 1 tailed or 2 tailed test with the two tailed being more conservative. When choosing between 1 and 2 tailed tests, one must decide is there is reason to believe that the treatment could only move the mean in one direction. If so, a one tailed test is appropriate. The *p*-value is interpreted against an α value (typically but not necessarily 0.05) and can be interpreted as the probability of rejecting the null hypothesis when it is in fact true (i.e. making a type I error).

6.3.1.2. Non-parametric two independent groups: assumptions and interpretation. The Mann–Whitney *U* test (also called the Wilcoxon rank-sum test) [27,28] is the most common non-parametric test used for the comparison of two independent groups. The Mann–Whitney *U* test does not assume that the samples were drawn from normal distributions, but it does assume that the distributions have the same basic shape. The assumption of independence of errors is maintained by the Mann–Whitney *U* test, so it requires that the observations must be truly independent. The essence of the Mann Whitney *U* statistic is a comparison of mean ranks, with the null hypothesis being that the distribution of values is not different. The resultant *p*-value provides the probability that the mean ranks would be found to be different based on chance alone.

6.3.2. Experimental design for two groups, repeated measures

The two groups, repeated measures design, or more generally *paired design* occurs when there is reason to pair off the individual subjects in the two groups being compared. Examples of this design include electrophysiological recording of synaptic transmission before and after compound application to provide baseline and compound treated values from the same cell, or the use of a crossover design where every animal in a study receives both a vehicle and compound treatment. In each case, there is reason to pair the values in the first group (baseline or vehicle response) with values in the second group (compound treatment). Pairing assumes that some of the variance is shared between the two groups, and by focusing on the difference between pairs allows for this variance to be blocked. Paired approaches can therefore offer greater statistical power.

6.3.2.1. Parametric two groups, repeated measures: assumptions and interpretation. The paired version of Student's *t*-test is the best parametric choice for this design. While formally similar to the unpaired *t*-test, the paired *t*-test is performed on difference values. It therefore assumes that the difference values are normally distributed. Because of the reliance on the difference between pairs, the paired *t*-test does not assume equal variance between the two samples. The paired *t*-test also assumes, perhaps obviously so, that there is a dependence between the paired values. The assumption of independent error discussed above for the unpaired *t*-test is retained by the paired *t*-test and must not be violated.

The formal statement of the null hypothesis for the paired *t*-test takes the form of the difference between the two groups is equal to zero. Therefore the *p*-value indicates the probability that the differences would be obtained by chance alone if the actual difference was zero. Many statistical software packages will also provide a Pearson correlation coefficient and related *p*-value to assess the effect of pairing. By measuring this association it is possible to test the assumption of dependence between the paired values. For example, in a crossover design it is expected that variance in the vehicle group and variance in the treatment group is somewhat shared as the same subjects are included in both groups. The Pearson correlation analysis rigorously tests this assumption by quantifying the association.

6.3.2.2. Non-parametric two groups, repeated measures: assumptions and interpretation. The Wilcoxon signed-rank test [29] provides a non-parametric alternative to the paired *t*-test for the evaluation of matched samples. This test does not assume that difference values are normally distributed; however it does assume that the distribution of differences is symmetrical about the median. It does retain the assumptions of dependence between paired values and independence of error.

The null hypothesis for the Wilcoxon signed-rank test state that the median differences between pairs is zero. The *p*-value is therefore interpreted as the probability that the median difference could be observed by chance alone if the actual difference were zero.

6.3.3. Experimental design for greater than two independent groups

With complex experimental designs involving more than two groups, the issue of *multiple comparisons* must be considered. The simplest way to understand the problem involved with making multiple comparisons is to consider the experimental design and the interpretation of the *p*-value obtained using a two-sample statistic like the *t*-test. For example, consider an experiment in which a compound is administered to a subject and a behavior is measured at multiple points in time (repeated measures design) or several behaviors are measured at a single point in time (planned multiple comparisons). Assume that 3 measures are obtained in either type of study. The first comparison by *t*-test produces a *p*-value of 0.05. This is interpreted as a probability of obtaining this difference based on chance alone is equal to 5%. Therefore it would be expected that if the study was run 100 times, this difference would be obtained 5 times even if there were no real difference between the groups. If a second comparison is now made, it would seem obvious that the *p*-value would need to be adjusted because the probability of obtaining a difference by chance alone is increased by making another observation (again from a theoretical population of 100). In other words, the more comparisons that are made, the higher the likelihood is of witnessing the occurrence of a rare event. Therefore, each comparison increases the chance of making a type I error, and confidence in the result is greatly reduced. A corollary to this indicates that all comparisons made should be planned prior to running the study, all planned comparisons are made, and all the comparisons made are reported.

Methods have been developed to compensate for the loss of statistical power produced by making multiple comparisons, however it should be noted that no method is superior to a simpler 2 group design. The Bonferroni correction [30] is the simplest and

most conservative method employed to deal with the problem of multiple comparisons. In practice the α value is simply divided by the number of comparisons to make the Bonferroni correction, and the p-value is interpreted against this corrected α value. This leads to an overly conservative estimate of significance and greatly increases the possibility of making a type II error when dealing with large numbers of comparisons, an issue that has become increasingly important with the development of imaging and gene association type studies (see below). Therefore, the most commonly used approach to multiple comparisons is to first use an overall test for the difference in means which will only indicate that a difference exists. This is then followed by a post hoc test to determine which of the comparisons, if any, achieve significance.

6.3.3.1. Parametric greater than two independent groups: assumptions and interpretation. The *analysis of variance* (ANOVA) provides a statistical test of whether 3 or more means differ significantly [31]. This is done by dividing the observed variance into different components (e.g. within groups, between groups). ANOVA is actually a collection of statistical methods and models the details of which are beyond the scope of this commentary. The ANOVA assumes normality in all groups. This is obviously a more critical assumption in ANOVA given that more groups are involved. There is also an assumption of equal variance across groups, and independence of error.

In practice, most statistical software packages will output a number of results many of which are critical to the interpretation of the ANOVA and should be reported. The ANOVA partitions variability into two components. First, variability due to treatment, which is the variability measured among the group means. The second variability assessed is the variability within the groups, also referred to as the residual variability. These are expressed as mean square values, and the ratio of these mean square values provides the F ratio. The F ratio and the df from each of the variability estimates are then used to obtain a p-value. The null hypothesis can be stated as the treatment does not produce any change in the mean values, and in general, an F ratio close to 1 indicates that the null hypothesis should be retained. The p-value is the probability that the observed differences would be expected to occur by chance alone.

The results of an ANOVA should always be reported in full as follows: $F(df_{bg},df_{wg}) = F$ ratio, $p = p$-value, where df_{bg} = the degrees of freedom in the estimate of the between groups variability which forms the numerator of the F ratio, df_{wg} = the degrees of freedom in the estimate of the within groups or residual variability which forms the denominator of the ratio, F ratio = the calculated F ratio, and the p-value = the obtained p-value.

The ANOVA only indicates that there is a difference among the means, but does not indicate which means are different. Assuming the ANOVA does not allow the null hypothesis to be retained; subsequent post hoc test can be performed to determine which means are significantly different. It is important to note that an overall significance from an ANOVA does not indicate that any of the individual means will be significant on post hoc testing. In cases where the overall ANOVA is significant, but the post hoc testing does not indicate significance, the interpretation is that treatment did produce a significant effect on the means, however that effect cannot be pinpointed to any singular pair of means. This is still valuable information and should not be reported as a lack of significance because it indicates that better designed follow up studies are necessary.

Choice of a post hoc test is a complex subject, and all too often the choice is made from a drop down menu in the statistical software based on little more than a familiarity with the name of the test. The factors for consideration include whether the comparisons were planned, if all pairwise comparisons are to be made or if comparisons should be restricted to assessing differences from a control group, and how conservative the test will tend to be (i.e. how likely a type II error may be). For planned comparisons between a subset of means, the Bonferroni correction remains the only choice, however as indicated above it is a highly conservative test. When making all pairwise comparisons of means, the Tukey honestly significant test (Tukey HSD) also known as the Tukey Kramer method [32] is often the most reasonable choice and provides an adequate compensation for multiple comparisons. When comparing against a control mean the Dunnett test [33] is often the best choice and like the Tukey HSD provides a correction for multiple comparisons. Scheffe's method [34] provides a means of evaluating contrast among the factor levels which are comparison involving multiple groups. While Scheffe's method is not commonly used, it has substantial power to detect differences and is the only post hoc test that will always find significance in a group of means that have been found different by ANOVA.

6.3.3.2. Non-parametric greater than two independent groups: assumptions and interpretation. The Kruskal–Wallis test [35] is a non-parametric test that can be used to compare 3 or more groups. It has the same assumptions as the Mann–Whitney U test in that it does not assume that the samples were drawn from normal distributions, but it does assume that the distributions have the same basic shape and has the standard assumption of independence of error. As a rank sum test, the Kruskal–Wallis tests against a null hypothesis that the sum of ranks are not different from each other. The resulting p-value is the probability that the difference is due to chance.

As with the parametric ANOVA, the Kruskal Wallis test only indicates that a difference exists among the means but does not indicate which means are different. Post hoc testing involves the use of the Mann–Whitney U test with Bonferroni correction, sometimes referred to as Dunn's test, or more simply termed the post test [12].

6.3.4. Experimental design for repeated measures on greater than two groups, one way repeated measures

A one way repeated measures design occurs in studies where more than one measurement is taken across a group. Importantly, the one way repeated measures design involves only one factor. For example, a study looking at the effect of increasing concentration of an agonist on a response, or the response of an agonist followed by an antagonist. As with the two groups paired design discussed above, the repeated measures design assumes that variance is shared by the individuals across measures and therefore allows for this variance to be factored out in making statistical comparisons.

6.3.4.1. Parametric repeated measures on greater than two groups, one way repeated measures: assumptions and interpretation. The repeated measures ANOVA is an extension of the ANOVA to account for matched groups. The assumptions of the one way repeated measures ANOVA are the same as the ANOVA with the addition of an assumption of sphericity. Sphericity is an extension of the equal variance assumption that states that the variance of the differences between all combinations of related groups is equal. Repeated measures ANOVA is particularly sensitive to violations of sphericity, and violations of sphericity are fairly common in pharmacology research. For example, the practice of generating cumulative dose response measurements by administering consecutively higher doses to a subject and recording effect after each dose is a common repeated measures design. The design implicitly assumes that the effect measured after the third dose is independent of the effect measured after the first dose. It is easy to

see that any residual effect of preceding doses could impact on the variance observed for effects at higher doses and therefore would violate sphericity. Most software packages that provide repeated measured ANOVA will perform tests of sphericity, and these tests should be used. Violations of sphericity can be somewhat corrected for by using the Greenhouse–Geisser [36] or Huynh–Feldt [37] corrections. However, rerunning the study with better experimental design including randomization of treatments is a more desirable approach. For example, avoiding a cumulative dose response design, or allowing for full reversal of the effect of each dose before beginning the next.

The null hypothesis for a repeated measures ANOVA design states that all groups have the same mean. The *p*-value is interpreted as the probability that any observed differences can be attributed to chance alone. Reporting the *F* ratio and *p*-value should follow the same format as described above for the one-way ANOVA. Analogous to the one-way ANOVA, the repeated measures ANOVA tests for significance among groups, but does not indicate which groups are different. Post hoc procedures are essentially the same as those employed after the one-way ANOVA.

6.3.4.2. Non-parametric greater than two independent groups, one way repeated measures: assumptions and interpretation. For situations where the normality assumption is violated in a repeated measures design involving 3 or more groups, the Friedman test [38], a rank non-parametric version of the analysis of variance can be used. The Friedman test is an extension of the Wilcoxon signed-rank test and carries all of the assumptions of that test as described above. In addition, the experimental design should be one that avoids violations of the assumption of sphericity. The null hypothesis for the Friedman test states that all groups have the same median value and the *p*-value is interpreted as the probability that differences in the median can be attributed to chance alone. As with the Wilcoxon signed-rank test, Dunn's test can be used as a post hoc analysis to determine which groups are significantly different.

6.3.5. Multifactorial designs

The above discussion on multiple comparison designs can be extended to multifactorial designs where the effect of more than one factor is assessed using the same logical approach as described for one factor ANOVA. These experimental designs may examine the impact of a manipulation across time or other grouping factors such as age, genetic strain, or sex. For example, a study aimed at measuring the effect of 3 concentrations of a glutamate receptor antagonist on apoptosis measured in cortical pyramidal neurons cultured from wildtype mice and mice carrying a mutation in γ-secretase. Concentration and genetic strain would both be factors in the design. Of course the issues described above regarding multiple comparisons and statistical power are even more apparent in these more complex multifactorial designs. Furthermore, it becomes difficult to state a concise null hypothesis to guide such studies. Two factor designs are common and the appropriate statistical tools to evaluate such studies are implemented in most quality software packages. Similar considerations and assumptions outlined above for single factor analysis are required in the implementation of these methods. While the statistical methods exist to analyze designs with more than 2 factors, there is rarely a justifiable reason for designing a study that complex.

7. Analysis of power

The above discussion of statistical inference centered on the probability of rejecting the null hypothesis when it is in fact true (i.e. making a type I error). When designing an experiment, the investigator should be most interested in maximizing the probability that the null hypothesis will be rejected when it is in fact false (i.e. avoiding a type II error) which is done by designing the experiment to have appropriate power. The *power* of a statistical test is defined as the probability that the test will lead to the appropriate rejection of the null hypothesis if it is in fact false [39].

Statistical power is influenced by (i) the reliability of results, (ii) the size of the expected effect, (iii) the size of the sample population (i.e. *n* value), and (iv) the predetermined significance level (α). The latter value is usually, although not necessarily set at an arbitrary conservative value such as 0.05. The reliability of results has to be determined from either historical control data or from pilot experiments to determine the variance in the population. The size of the expected effect may also be determined in a pilot study, or may be the subject of an educated guess. These values are then used in a calculation along with tables to estimate the power of the proposed statistical test for a given sample size. Unfortunately, the calculations are complex and most basic statistical packages either fail to include or poorly implement power analysis. It is possible with some effort to produce a spreadsheet that will calculate power values for a range of sample sizes based on published descriptions and tables [40], however for designs more complex than one way ANOVA this quickly becomes cumbersome. Despite these difficulties, beginning any set of experiments without first performing an analysis of power means proceeding with uncertainty regarding the probability of making a type II error. The impact of this should not be underestimated. A failure to replicate published results could be due to either the original report containing a type I error, or the efforts at replication containing a type II error. It is typical to assume the former case, but this assumption is without foundation in the absence of a power analysis. Therefore, the investment in a comprehensive statistical software package, the development of a spreadsheet to perform the analysis, or collaboration with a statistician is essential.

Once a power analysis is completed, a choice is made to either proceed with the experiment as designed using the appropriate sample size identified by the analysis, or to revisit the design in an attempt to either increase the reliability of measurement or increase the effect size. Therefore, the analysis of power plays a central role in experimental design.

7.1. Misuse of power analysis

There are a few possible uses for power analysis in a post hoc case, all of which are a misuse of the method [41–43]. In light of a statistical test not reaching significance, there are two possible interpretations: (i) the null hypothesis is correct, or (ii) the test did not have sufficient power to reject the null hypothesis. A post hoc analysis of power is sometimes wrongly used to indicate that null hypothesis is true because there was adequate statistical power to prove otherwise. However, this is a circular argument as power is defined as the probability of correctly rejecting the null hypothesis. If the null hypothesis is actually true, then statistical power is irrelevant. Another misuse that is particularly worrisome is the use of a post hoc analysis of power to determine how many subjects to add to the data set in order to achieve significance. Since power relies only on variance, significance criteria, and effect size, it should be obvious that any effect size will be significant at a given criteria given enough observations. The practice of post hoc power analysis and subsequent increase in sample size to achieve significance greatly increases the probability of making a type I error.

Perhaps the greatest and most common misuse of the analysis of power is the failure to perform the analysis prior to beginning an

experiment. Since to goal of most studies is to minimize the probability of making a type 1 error, to not assess the probability if frankly irrational.

8. A discourse on large data sets, "modern approaches" and Bayseian methods

This commentary has focused on the use and misuse of statistics in classical hypothesis driven scientific investigation. For centuries, this has been the primary mode of scientific investigation because it provides a clear and rational framework for designing and interpreting studies. However, the reliance on the scientific method can also be somewhat attributed to the technical limitations that have existed for centuries on data acquisition. The problems of multiple comparisons and subsequent loss of power remained relatively tenable not only because scientists have been careful to design studies with statistical methods in mind, but also because it has been technically challenging to conduct studies with massive parallel comparisons and still maintain reasonable control of sources of non-sampling error. This has changed with the development of "omics" approached and imaging techniques which can produce thousands and even millions of data points in a single study leading to a situation where the time required to carefully analyze the resultant data far exceeds the time needed to acquire the data. The ability to use imaging methods to look for variations across the entire brain in a disease population for example has led to a decreased reliance on hypothesis testing and a more "shotgun" look-and-see approach termed "connectomics". This sort of approach driven by remarkable technical advances continues to gain prominence under the poorly defined umbrella term of systems biology at the cost of well-defined hypothesis driven research, a process described by Brenner as "low input, high throughput, no output science" [44].

Despite the issues with loosing site of the guiding light of scientific method, it would seem foolish to simply ignore the massive data sets available through modern techniques. Certainly there is value in being able to profile variations across the entire genome between a control and a diseased population (assuming the disease has a true genetic component). But as discussed above, this sort of study design is a massive multiple comparisons exercise. In the field of genome wide association studies (GWAS) millions of genetic markers measured simultaneously leads to vanishingly small α values with any of the corrections for multiple comparisons describe above, a fact that likely attributes to low rates of reproducibility [45]. One possible framework for drawing value out of this sort of study is to treat the data set as pilot data and employ an EDA approach to try and understand where future studies should be directed. The development of bioinformatics approaches such as pathway analysis [46] should provide valuable tools to explore these large data sets, however the outcome of these analysis must be thought of in terms of hypothesis and not considered tested until well designed follow-up studies are conducted.

8.1. Bayseian inference

An alternate approach to drawing value from large datasets is to dispense with the above described methods of inference termed "frequentists" methods, and embrace *Bayseian inference*. The number of journal articles employing Bayseian methods has increased significantly with the advent of "omics" technologies. Unfortunately, the understanding of these methods by journal referees has not kept pace. Many find it difficult to grasp the fundamental concepts of Bayseian inference, perhaps because it is foreign to the way people draw inference in day to day life [47]. At the heart of Bayseian methods is the concept that the probability of an event occurring can be informed by prior knowledge. Thus conditional probabilities come into play and prior data can be used to inform statistical models. A common example is guessing a playing card drawn at random from a well-shuffled deck. The probability of correctly guessing any one card (e.g. the king of spades), is 1 in 52. However, if you know that a face card has been drawn (prior knowledge) the probability of correctly guessing the card is raised to 1/12. The new probability is conditional on the prior knowledge. If a similar approach is taken to inference from a large dataset, prior knowledge about the population or experimental outcomes can be translated into a probabilistic form and combined with new results to increase statistical power and draw out differences without the difficulties of dimensionality encountered in classical multiple comparison approaches. A full treatment of Bayseian inference can be found in several introductory texts [48,49].

The development of new statistical approaches to the treatment of large higher dimensional data sets is an exciting advancement in the field, and the difficulties experienced to date should not be viewed as an overall indictment of these approaches. However, the fact must be recognized that the field of pharmacology does not yet have clear quantitative framework with which reliable inference can be drawn for this type of study. Therefore caution must be exercised in the interpretation of results and independent confirmation of findings using well designed null hypothesis driven approaches would seem to be a requirement before using such data as a foundation for future work.

9. Summary

The two primary issues laid out in the introduction to this commentary, frank misuse of statistical methods and failure to understand data before employing methods of inference, along with an appreciation of the methods and assumptions discussed above can be used as a foundation for developing a practical and formal approach to embedding appropriate statistical design and analysis into the framework of the scientific method (Fig. 7). Hypotheses are generated based on current knowledge and rigorously formulated in a null hypothesis framework. Pilot studies are performed (when necessary) and EDA methods are used to further refine the hypothesis and inform the experimental

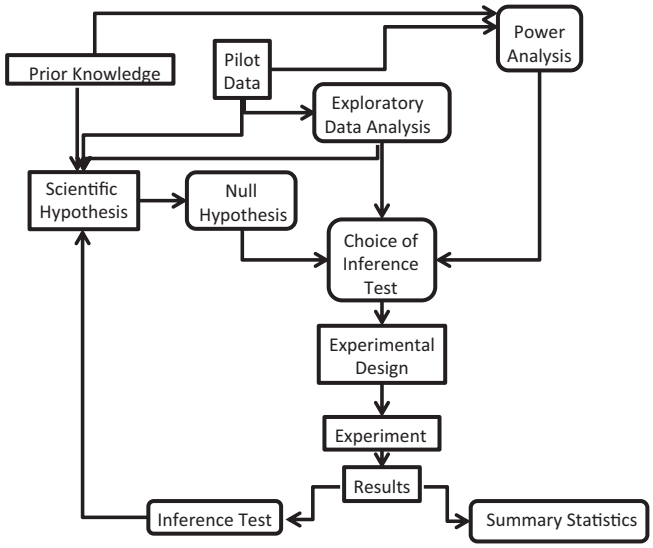


Fig. 7. A restatement of the scientific method to include the role of descriptive and inferential statistics and, exploratory data analysis in the overall process. See text for additional details.

design and subsequent inferential methods. Prior to beginning experiment, an inferential method should be chosen based on an assessment of assumptions from the pilot data or prior knowledge (e.g. repeated measures or single observation, parametric or nonparametric, etc.), and an analysis of power is conducted to determine the appropriate sample size. After collecting the data, EDA methods are again employed to visualize results, check assumptions, and evaluate outliers before proceeding to the appropriate pre-determine inference test. Once the test is complete, the results are reported along with all appropriate summary statistics and a full accounting of the significance test (e.g. p values, sample size, df, etc.). The results are then fed back to inform modification of the original hypothesis.

If these steps appear to be extreme, it is not because they are irrational but because the misuse of statistical methodology has become so common as to appear rational. It is impossible to logically conduct scientific research without taking the steps necessary to insure confidence in experimental results.

Acknowledgement

The author would like to thank Dr. David C. Wood for teaching the importance of experimental design and exploration of data.

References

[1] Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. Nature 2012;483:531–3.
[2] Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 2011;10:712.
[3] Kitchen I. Statistics and pharmacology: the bloody obvious test. Trends in Pharmacological Sciences 1987;8:252–3.
[4] Murray A. The wall street journal essential guide to management: lasting lessons from the best leadership minds of our time. New York: HarperBusiness; 2010.
[5] Ioannidis JP. Why most published research findings are false. PLoS Medicine 2005;2:e124.
[6] Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005;294:218–28.
[7] Lehrer J. The truth wears off. New Yorker 2010;13(December):2010, http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer.
[8] Tatsioni A, Bonitsis NG, Ioannidis JP. Persistence of contradicted claims in the literature. JAMA 2007;298:2517–26.
[9] Crabbe JC, Wahlsten D, Dudek BC. Genetics of mouse behavior: interactions with laboratory environment. Science 1999;284:1670–2.
[10] Kemp AS, Schooler NR, Kalali AH, Alphs L, Anand R, Awad G, et al. What is causing the reduced drug-placebo difference in recent schizophrenia clinical trials and what can be done about it? Schizophrenia Bulletin 2010;36:504–9.
[11] Cohen J. The earth is round ($p < 0.05$). American Psychologist 1994;49:997–1003.
[12] Daniel WW. Applied nonparametric statistics. Pacific Grove, CA: Duxbury/Thomson Learning; 2000.
[13] Degroot MH, Schervis MJ. Probability and statistics. Boston, MA: Addison-Wesley; 2002.
[14] Rowntree D. Statistics without tears: an introduction for non-mathematicians. Toronto: Penguin Books; 2000.
[15] Kirk RE. Statistics: an introduction. Belmont, CA: Wadsworth Publishing; 2007
[16] Tukey JW. Exploratory data analysis. Reading, MA: Addison Wesley; 1977.
[17] Barde MP, Barde PJ. What to use to express the variability of data: Standard deviation or standard error of mean? Perspectives in Clinical Research 2012;3:113–6.
[18] Peirce B. Criterion for the rejection of doubtful observations. Astronomical Journal 1852;2:162–3.
[19] Grubbs FE. Procedures for detecting outlying observations in samples. Technometrics 1969;11:1–21.
[20] Dean RB, Dixon WJ. Simplified statistics for small numbers of observations. Analytical Chemistry 1951;23:636–8.
[21] D'Agostino RB. Tests for the normal distribution. Goodness-of-fit techniques, New York: Marcel Dekker; 1986.
[22] Messerli FH. Chocolate consumption, cognitive function, and Nobel laureates. New England Journal of Medicine 2012;367:1562–4.
[23] Rodgers JL, Nicewander WA. Thirteen ways to look at the correlation coefficient. American Statistician 1988;42:59–66.
[24] Spearman C. The proof and measurement of association between two things. American Journal of Psychology 1904;15:72–101.
[25] Livingston EH. Who was student and why do we care so much about his t-test? Journal of Surgical Research 2004;118:58–65.
[26] Welch BL. The generalisation of student's problems when several different population variances are involved. Biometrika 1947;34:28–35.
[27] Fay MP, Proschan MA. Wilcoxon–Mann–Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Statistics Surveys 2010;4:1–39.
[28] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics 1947;18:50–60.
[29] Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin 1945;1:80–3.
[30] Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998;316:1236–8.
[31] Fisher RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1970.
[32] Kramer CY. Extension of multiple range tests to group means with unequal numbers of replications. Biometrics 1956;12:307–10.
[33] Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association 1955;50:1096–121.
[34] Scheffe H. The analysis of variance. New York: Wiley-Interscience Publication; 1999.
[35] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 1952;47:583–621.
[36] Greenhouse FW, Geisser S. On methods in the analysis of profile data. Psychometrika 1959;24:95–112.
[37] Huynh H, Feldt LS. Estimation of the Box correction for degrees of freedom from sample data in randomized block and split = plot designs. Journal of Educational Statistics 1976;1:69–82.
[38] Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 1937;32:675–701.
[39] Cohen J. Statistical power analysis for the behavioral sciences. New York: Psychology Press; 1988.
[40] Cohen J. A power primer. Psychological Bulletin 1992;112:155–9.
[41] Levine M, Ensom MH. Post hoc power analysis: an idea whose time has passed? Pharmacotherapy 2001;21:405–9.
[42] Thomas L. Retrospective power analysis. Conservation Biology 1997;11:276–80.
[43] Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. American Statistician 2001;55:1–6.
[44] Brenner S. An interview with... Sydney Brenner. Interview by Errol C. Friedberg. Nature Reviews Molecular Cell Biology 2008;9:8–9.
[45] Ioannidis JP. Why most discovered true associations are inflated. Epidemiology 2008;19:640–8.
[46] Kelder T, Conklin BR, Evelo CT, Pico AR. Finding the right questions: exploratory pathway analysis to enhance biological discovery in large datasets. PLoS Biology 2010;8.
[47] Sedlmeier P, Gigerenzer G. Teaching Bayesian reasoning in less than two hours. Journal of Experimental Psychology General 2001;130:380–400.
[48] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. Boca Raton, FL: Chapman and Hall; 1995.
[49] MacKay DJC. Information theory, inference, and learning algorithims. Cambridge, UK: Cambridge University Press; 2003.
[50] Anscombe FJ. Graphs in statistical analysis. American Statistician 1973;27:17–21.