Published in final edited form as:

Circ Res. 2021 May 28; 128(11): 1724–1727. doi:10.1161/CIRCRESAHA.120.317360.

Revisiting Some Useful Statistical Guidelines in *Circulation Research* in Response to a Changing Landscape

Heather M. Highland, **Eric R. Gamazon**, **Jennifer E. Below**

Keywords

Statements and Guidelines; Statistical analysis; statistics; guidelines

INTRODUCTION

In the 40 years since "Some Statistical Methods Useful in Circulation Research" was published, many of the same battery of statistical tests and concepts, such as *t*-tests, ANOVA, *p*-values, effect sizes, and standard errors, are still abundantly employed in hypothesis-driven research[1]. Newer methods, too, have emerged to address the challenges of big data analysis. Some methods now routinely employed to extract insights from data include regression analysis, supervised and unsupervised machine learning for clustering, density estimation, and dimensionality reduction (e.g., viSNE), as well as prediction modeling and enrichment analyses. Additionally, in basic science research, it is now common to encounter hypothesis-free analyses, in marked contrast to traditional statistical analyses that begin with an explicit hypothesis. To encourage reproducibility, rigor, interpretability, and transparency, many editorial teams, including those of *Circulation Research* and the AHA journals, have developed statistical guidelines for authors. Given the rapidly changing data landscape, such guidelines must extend beyond "What statistical test should I use?" (a question that often can be addressed by a decision tree diagram in applied statistical analysis textbooks), to address higher-level challenges that frequently face authors including multiple testing, standards of reporting, robustness to violations of assumptions, and the limitations of conventional measures of significance. To better support authors and readers, we have assembled some topics that warrant particular attention in basic and clinical scientific publications such as those published in *Circulation Research*. These guidelines are intended to complement those outlined by the American Heart Association's Statistical Taskforce in their concurrent "Guidelines for Statistical Reporting in Cardiovascular Medicine: A Special Report from the American Heart Association" [2].

Address correspondence to: Dr. Jennifer E. Below, Vanderbilt University Medical Center, Medical Genetics, 2215 Garland Ave, Light Hall #519B, Nashville, TN 37232, United States, Tel: 5102892578, jennifer.e.below@vanderbilt.edu.

DISCLOSURES

None.

REPORTING

Clear description of the approach used to generate each statistical value.

Key to interpretation and evaluation of any statistical analysis is clear and precise description of the methods used, the data under consideration, and the resulting parameters and test statistics. In the interest of reproducibility, authors must rigorously describe the approaches used to calculate summary values and all test statistics, which commonly include measures such as *p*-values, effect sizes or odds ratios, and z-scores. Often, providing the details of the specific statistical tests, the exact sample size, and reporting the exact value of the test statistic is sufficient to enable other researchers to reproduce the work. In principle, in statistical hypothesis testing, authors should specifically:

- Explicitly state the hypothesis (e.g., there is no difference in measure A between treatment groups)
- Describe the statistical assumptions concerning the samples (e.g., state whether the samples are assumed to be independent, or if the distribution they are drawn from is assumed to be normal)
- State the statistical test employed
- Give the specific test statistic for interpretation
- Describe the distribution of the test statistic under the null hypothesis
- State the predetermined significance threshold (alpha) that determines whether the null hypothesis is rejected
- Precisely report the resulting values and the determination on whether the null hypothesis is rejected as a consequence of the observations.

P-values.

Despite their importance in biomedical research, it is important to note what p-values are not. This has implications for what and how results should be reported:

1. A *p*-value is not a substitute for the size of an effect.
2. A *p*-value that meets a threshold (historically $p<0.05$) is often necessary but not sufficient to draw a particular conclusion.
3. A *p*-value is not interpretable without a transparent study design.

The American Statistical Association defines a *p*-value as the "probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value."[3] In hypothesis testing, a smaller *p*-value can indicate how incompatible the data are with the specific statistical model and is interpreted as stronger evidence in favor of the alternative hypothesis.

In reporting, authors occasionally prefer to present a *p*-value range (e.g., "$p < 0.001$"), rather than an exact value; however, a range does not provide sufficient information to interpret or

reproduce the presented work. Additionally, it is critical to report precise *p*-values for post-hoc multiple testing considerations and future meta-analysis. We recommend that *p*-values be reported to a minimum of two significant digits. We note that some software tools have limitations and cannot accommodate appropriate statistical tests or provide precise *p*-values. Such limitations should not dictate sound practice to ensure reproducibility and transparency. The number of hypotheses tested (e.g., multiple simultaneous comparisons within an experiment, each able to generate a "discovery") should be reported (see section on multiple test correction for additional information) and findings presented in an unbiased manner. For example, selective reporting of *p*-values should be based on predetermined criteria. All reported *p*-values should be from two-sided tests unless study design warrants a one-sided test.

Effect sizes and confidence intervals.

An effect size is a quantitative value calculated from certain effect statistics (e.g., the value of the mean difference between groups, or the value of the correlation coefficient). A confidence interval (CI) is defined as the range of values that encompass the true value of the effect with a given probability, estimated by an effect statistic[4]. The value of an effect size and its corresponding CI can enable effective assessments of relationships within data and capture distinct information from statistical significance. We strongly recommend reporting effect sizes with standard errors/CIs for every test, and when reported, effects should always be accompanied by a measure of error. Qualitative claims of difference (e.g., 'increased', 'decreased', 'elevated', 'reduced', and so forth) should be quantified and statistically evaluated.

Multiple testing correction.

Multiple testing refers to the scenario in which multiple hypotheses are tested simultaneously, which can lead to increased type I error (i.e., rejection of a true null hypothesis). The more hypotheses are tested, and inferences made, the greater the possibility of incorrect inference. For example, when a family of ten hypotheses is tested with a significance threshold (alpha) of 0.05, the cumulative risk of observing at least one significant association when there is no real difference is as high as 40%[5]. To control the type I error rate, it is often appropriate when testing multiple hypotheses to set a stricter significance threshold than when making an individual comparison. A variety of methods exist to address the problem of multiple comparisons, such as Bonferroni adjustment to the significance threshold, *post hoc* adjustments of ANOVA tests, and Benjamini Hochberg False Discovery Rate (FDR) thresholds. These parameters must be justified, and should reflect the number of independent tests performed (it should be clear, for example, if you tested all possible pairs or only control versus each experimental condition). When many experiments are conducted within a family of hypotheses, the increased likelihood of an erroneous inference must be acknowledged, and this multiplicity problem should be addressed statistically. When reporting an adjusted *p*-value, authors must state the method used and the number of hypotheses adjusted for, as well as how many were actually tested.

Sample size.

For reproducibility, transparency, and interpretation, exact sample size must be clearly described for every test and group; it is not sufficient, generally, to give range values of sample size. Very small sample sizes are inadequate for establishing assumptions of normality for parametric statistical tests and offer limited statistical power, and, therefore, drawing strong conclusions should be avoided in such a scenario.

Power.

Power calculations should be considered a central element of study design, which may inform the necessary sample size to test a particular hypotheses. These calculations should account for anticipated multiple testing. The assumptions (e.g., regarding effect size) underlying the power analysis should be explicitly stated.

Correlation vs. causation.

Correlation and causation are not synonymous; tests of correlation are not sufficient to establish causation, especially in small sample sizes. Strong causal conclusions require multiple lines of evidence especially when sample sizes are small.

Data display.

Data visualization should support transparency[6]. For example, authors will often present "representative" images to illustrate a key finding. The procedure used to select such representative images for display from the multiple choices should be described, and ideally, images from all replicates should be included in the supplementary materials. Similarly, bar graphs should not be used to represent continuous data when dot plots and violin plots enable clearer representation of the data density and distribution.

Computational tests, code, pipelines, data sharing.

For transparency, to support reproducibility and downstream experiments, we advise making all scripts, pipelines, computations, and results available to readers. We also strongly encourage making raw or summary data available to readers.

DESIGN

Choice of statistical test or procedure.

The choice of a statistical test depends on the scientific question, the data type (e.g., continuous, binary, categorical), the data distribution (e.g., normal, Poisson, unknown), and the study design (e.g., paired or unpaired, independent or correlated). Violation of statistical test assumptions may lead to incorrect results, diminished power, or increased type I error. Common examples of an incorrectly applied statistical test or procedure include violations of assumptions (normality, independent sampling). Parametric tests (e.g., *t*-test, ANOVA) assume underlying statistical distributions in the data and can be biased when these assumptions are not met. The main tests for evaluation of normality include D'Agostino-Pearson, Kolmogorov-Smirnov (K-S) test and Shapiro-Wilk test. Although a formal test of normality is preferred, when sufficient data are available, visual assessment of normality

may suffice if it is provided in supplementary material. Non-independence of samples can occur when multiple samples are taken from the same organism (e.g., observing multiple tissue samples in the same mouse after treatment/sham).

Another example is the use of multiple pairwise comparisons (e.g., *t*-tests) when a multiway comparison (e.g., ANOVA) is more appropriate. Special consideration should be given to characteristics of the data, such as the presence of outliers or skewed data, to which some tests may be sensitive. In addition to evaluating whether outliers should be discarded, transformation of data or use of an alternative approach that is robust in the presence of outliers should be considered.

Control selection.

Selection of controls and the process of assigning samples to experimental groups is important to define. In human data, it is important to describe the recruitment and matching criteria and to show that controls are recruited from the same population as cases. If the cases and controls do not come from similar populations, effects can be detected due to a confounding factor (e.g., age) rather than the outcome of interest. For interventional studies, the process by which a subject (cellular, animal, or human) is assigned to a treatment or control group must be defined.

Claims of no difference or similarity between groups.

When testing a null hypothesis of no difference, a non-significant result implies that one cannot reject the null hypothesis, but this is not the same as being able to accept the null hypothesis or demonstrating statistically meaningful similarity. For example, in this situation it would be correct to report that “after X days of treatment, measurement Y is not significantly different between samples” and it would be inappropriate to state, “after X days of treatment, measurement Y is the same between samples”. Claims of similarity should be accompanied by statistical evidence, using an approach which tests for similarity, such as the two one-sided tests (TOST) for equivalence[7].

Circular analysis.

Feature selection and subsequent statistical inference on the same data can lead to “double dipping.” This problem may yield severely biased estimation and overly optimistic results. An example is the use of the entire dataset for training a model and then a subset of the same dataset for testing the model[8]. Another example is selecting a subset of data consistent with the effect to be identified. Methodological caution should be taken and may include partitioning the data into two independent samples, one for training or exploration and the other for testing. More generally, *k*-fold cross validation may be performed. To quantify the uncertainty in a test statistic, bootstrapping, i.e., resampling the data with replacement, may be used. In general, validation in an external and independent dataset will avoid the problem of circular analysis.

Omics data and analysis.

Omics data are typically generated from high-throughput (molecular) profiling technologies. In contrast with traditional (e.g., single gene) studies, omics data can be noisy and therefore

must typically undergo a series of extensive quality control steps. In addition, multiple hypothesis testing is particularly relevant for these datasets. In omics studies, analysis is often conducted on a large number of variables (e.g., variants, genes, proteins) measured in a much smaller number of samples (the so-called "n < p" problem). For example, a common goal of omics analysis is to identify a set of differentially expressed genes, from among many that are assessed, between two conditions. Given the number of statistical hypotheses being tested, an omics study should conduct power analyses. The goals of the study should be clearly defined and be realistic, as definitive conclusions may not be possible, and authors should acknowledge the often-severe limitations given the sample size and power.

Validation and replication.

There have been some concerns about the reproducibility of scientific findings[9]. To address this, authors must consider direct validation and/or replication of their findings. When a large number of tests are performed (e.g., in omics), we expect some false associations despite correcting for multiple testing. The results can be *replicated* using an independent sample from the same population or *validated* using a sample from a different population. In the case that additional data are not available, other sources of evidence should be considered to further strengthen the conclusion. Other sources of evidence may include other types of omics data, animal model databases, and functional follow up. When additional follow-up is beyond the scope of the current body of work, the need for further validation and replication should be acknowledged. Replication may improve robustness (i.e., stability of conclusion to a small change in one of the assumptions) and generalizability (i.e., the relevance of the conclusion in other contexts).

Conclusions.

The new statistical guidelines are based on the following key principles: reproducibility, rigor, interpretability, and transparency. Although these guidelines are focused on frequentist inference rather than Bayesian, Bayesian approaches have many benefits and are highly relevant. A checklist of specific suggestions for authors based on these guidelines, and addressing many issues that commonly arise in review, will be made available on the *Circulation Research* website (https://www.ahajournals.org/res/author-instructions). We emphasize that these are guidelines and are not the only acceptable designs and practices, and when appropriate will be adjusted to address additional statistical challenges commonly encountered in the broad range of impactful publications sent to *Circulation Research*.

Acknowledgments

SOURCES OF FUNDING

H.M.H. is supported by American Diabetes Association Grant #1-19-PDF-045 and the National Heart, Lung, and Blood Institute under Award Number R01HL142825. E.R.G. is supported by the National Human Genome Research Institute of the National Institutes of Health under Award Numbers R35HG010718 and R01HG011138. J.E.B. is supported by the National Institute of General Medical Science (R01GM133169), the National Institute on Aging (R01AG061351), the National Institute of Deafness and Other Communication Disorders (R01DC017175 and R01DC016977), the National Heart, Lung, and Blood Institute (R01HL142302), the National Institute of Dental and Craniofacial Research (R03DE027494), and the National Science Foundation (DUE1926794).

Nonstandard Abbreviations and Acronyms:

CI	confidence interval

REFERENCES

1. Wallenstein S, Zucker CL, Fleiss JL. Some statistical methods useful in circulation research. Circ Res. 1980;47:1–9. [PubMed: 7379260]
2. Althouse AD, Below JE, Claggett BL, et al.. Guidelines for Statistical Reporting in Cardiovascular Medicine: A Special Report From the American Heart Association. Circulation. 2021
3. Wasserstein RL, Lazar NA. The ASA Statement on p -Values: Context, Process, and Purpose. Am Stat. 2016;70:129–133.
4. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc. 2007;82:591–605. [PubMed: 17944619]
5. Harrington D, D'Agostino RB, Gatsonis C, Hogan JW, Hunter DJ, Normand S-LT, Drazen JM, Hamel MB. New guidelines for statistical reporting in the journal. N Engl J Med. 2019;381:285–286. [PubMed: 31314974]
6. Weissgerber TL, Winham SJ, Heinzen EP, Milin-Lazovic JS, Garcia-Valencia O, Bukumiric Z, Savic MD, Garovic VD, Milic NM. Reveal, don't conceal: transforming data visualization to improve transparency. Circulation. 2019;140:1506–1518. [PubMed: 31657957]
7. Lakens D Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. Soc Psychol Personal Sci. 2017;8:355–362. [PubMed: 28736600]
8. Button KS. Double-dipping revisited. Nat Neurosci 2019;22:688–690. [PubMed: 31011228]
9. Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? Sci Transl Med. 2016;8:1–6.