Guidelines on design and statistics for Appetite

Geary, Nori; Higgs, Suzanne

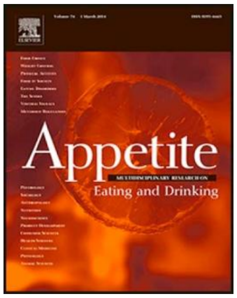
DOI:
10.1016/j.appet.2015.04.045

License:
Other (please specify with Rights Statement)

Document Version
Peer reviewed version

Citation for published version (Harvard):
Geary, N & Higgs, S 2015, 'Guidelines on design and statistics for Appetite', *Appetite*.
https://doi.org/10.1016/j.appet.2015.04.045

Accepted Manuscript



Title: Guidelines on design and Statistics for *appetite*

Author: Nori Geary, Suzanne Higgs

PII: S0195-6663(15)00191-9
DOI: http://dx.doi.org/doi:10.1016/j.appet.2015.04.045
Reference: APPET 2524

To appear in: *Appetite*

Guidelines on Design and Statistics for *Appetite*

Nori Geary, PhD
Professor (Retired)
Department of Psychiatry
Weill Medical College of Cornell University
New York, NY USA

Suzanne Higgs
School of Psychology
University of Birmingham
Edgbaston
Birmingham B15 2TT, UK
s.higgs.1@bham.ac.uk

ndg47@hotmail.com

21.1.2015

INTRODUCTION

Appetite strives to publish the highest quality science possible. To that end, we offer the following general guidelines on experimental design and statistical analysis. Authors are advised to consult statisticians for specific guidance concerning the design and analysis of their own work, especially if they have questions concerning the points raised in these guidelines. In addition, the development of many of the statistical approaches described here are themselves active research areas, which is another good reason to consult professionals.

GENERAL ASPECTS of GOOD STATISTICAL PRACTICE

Full data reporting*.* Disturbing issues in scientific integrity have been described in recent years, many of them having to do with statistical practice (Ioannidis, 2005; Landis et al., 2012; Simmons et al., 2011). *Appetite* seeks to minimize such issues. To this end, authors are urged to: [i] design the experiment, including the statistical approach, in advance; [ii] conduct the research – including the statistics – with integrity; and [iii] fully and clearly describe the design and execution of the experiments, including statistical methods, randomized or blinded aspects of the design, loss of data, etc.

Planning for meta-analyses. Scientific meaning is rarely established by a single study, but rather by the cumulative effect of many similar studies. The state-of-the-art for the quantitative integration of similar studies is meta-analysis (Borenstein et al., 2009; Cooper, 2010). Thus, a useful criterion for full data reporting is for authors to plan for the potential later inclusion of their work in meta-analyses, i.e., quantitative integration with other similar studies. To meet this criterion, all sample sizes, measures and outcome estimates (means, etc), and their variabilities should be reported. If this does not fit easily with the chosen style of presentation, it should be included as supplementary data.

Descriptive, exploratory and analytic statistics*.* Descriptive statistics summarize the data, and analytic statistics assist in making inferences about the meaning of data. Between the two lies exploratory data analysis or data mining, which refers to attempts to understand the collected data using a variety of descriptive approaches with the goal of discovering unexpected possibilities that could guide future experiments (Tufte, 2001; Tukey, 1977; Wainer, 1997; Gelman, 2003). Wainer and Velleman's (2008) exploration of blood glucose level graphing is an excellent example. Recently, nonparametric estimation methods have been used to quantify exploratory data analysis in novel ways (Harpole et al., 2014). Serendipity plays an important role in science. Exploratory analyses, however, should be clearly labeled as such and described separately from analytic statistics.

Analytic statistics. The two main parametric approaches to analytic statistics are based on the classical contributions to mathematical probability of Carl Friedrich Gauss (distributions of normal errors, least squares estimation, etc.) and Pierre-Simon LaPlace (central limit theorem, etc.) in the early 19th C. The first approach may be labeled the **statistical-significance approach** and the second, the **estimation approach**. Despite their common roots, these involve different analytic methods, different language, and different logical rules for data interpretation. For example, in the statistical significance approach, one tests hypotheses such as, "is there a difference between groups X and Y?," whereas in the estimation approach, rather than framing specific hypotheses, one asks, "how large is the difference between groups X and Y?" *Appetite* accepts both approaches. They should not, however, be mixed in a single experiment. Some aspects of each approach are described below. Finally, nonparametric tools have been developed for both the statistical-significance approach and the estimation approach (***nonparametric analytic statistics***).

Negative data. *Appetite* recognizes the need to publish well designed experiments that address interesting questions but fail to result in convincing outcomes. Not to do so inflates the meaning of positive reports and invalidates future meta-analyses. Negative data are rarely considered suitable for publication, however, if the experimental design does not include a suitable power analysis.

It is crucial to understand that "negative data" does not mean that the statistics show that there is no difference. Rather, it means only that the statistics failed to demonstrate evidence of a difference, which is very different. As has been pointed out repeatedly, "absence of evidence is not evidence of absence" (Alderson, 2004; Altman and Bland, 1995; Bramness et al., 2008; Hartung et al., 1985). Negative data should be described with this in mind.

***Statistics and meaning**.* Authors are encouraged to bear in mind that statistics are not in themselves the meaning of experiments, but are merely guides to meaning. Balance and perspective are necessary in interpreting statistical results. As already mentioned, individual experiments should not be over-interpreted. In addition, authors should realize that, paradoxically, rigorous design and statistics may obscure meaning. Corning and Tukey (1956) explained this with a metaphor. Imagine that the experimenter is on one side of the river, that meaning is on the other side, and that a bridge with two spans connects the former with the latter. The first span signifies the experimental design and statistical outcome, and the second span signifies the gulf between the experimental outcome and actual phenomena of interest. Now consider maneuvers that make statistical outcomes clearer, such as adding exclusion criteria to reduce sample variability or choosing certain statistical models over others. These may shorten the first span, i.e., increase the precision of the experimental result, but they do not affect the width of the river and, therefore, lengthen the second span, i.e., increase the gulf between the outcome and its meaning.

EXPERIMENTAL DESIGN.

Prespecification. The experimental design should be specified in advance. This also applies to the statistical approach. Not to do so leads to false-positive results (Simmons et al., 2011). Therefore, analyses that are not prespecified should be identified and discussed as provisional.

Measurement. Most experiments result in numerical measurements. Statistics should be appropriate for the scale of measurement used. For example, arithmetic means and most parametric analytic statistics are not appropriate for data derived from ordinal scales of measurement. If the underlying scale of a measurement is unclear, as is often the case with psychological rating schemes, data should be assumed to ordinal. Clinical scales that are validated only for detection of clinically relevant vs clinically irrelevant scores may be best considered categorical.

Mathematical transformations (multiplying values by a constant, taking logs, etc.) can change measurements in ways distorting their meaning. Thus, only "permissible" transformations of data, i.e., those that do not distort the underlying scale, are recommended for the purpose of meeting the requirements of ANOVA or other statistical procedures. Transformations into percentages of baseline values are especially troublesome. These can render small measured absolute differences larger than large measured absolute differences. Analysis of covariance is usually a better strategy to integrate baseline data (see ***Correlational approaches***).

Sarle (1997) gives an excellent introduction to these issues, including several examples from psychology experiments.

Meta-Analyses. Authors performing meta-analyses are encouraged to adhere to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; www.prisma-statement.org).

Clinical trials. Authors are encouraged to adhere to the Consolidated Standards of Reporting Trials (CONSORT; www.consort-statement.org).

Robust statistical methods. Computers have permitted the development of a variety of novel and powerful statistical methods, commonly known as robust statistics (Wilcox, 2003). Two useful methods are computerized resampling and bootstrapping methods (Kirby & Gerlanc, 2013). Authors are encouraged to consider these alternatives.

Extreme values / Outliers. Robust statistical methods to detect and exclude extreme values are appropriate and often especially useful in small-sample studies. A simple method is to compute the probability of suspected extreme values using median standard scores:

$$(x - \text{group median})/1.48\ \text{MAD},$$

where x is the suspect datum, MAD is the median absolute deviate ($|x_i -$ group median$|$ for each x_i in the group; note: 1.48 MAD ≈ the group's standard deviation). Other robust methods are described by Rousseeuw and Croux, 1993.

***Reporting**.* Authors should clearly describe the design and execution of the experiments, including all measures, data manipulations, and data exclusions. All randomized or blinded aspects of the design should be mentioned.

Data ordinarily should be reported in the form measured, using SI units (Le Système International d'Unités) where possible and clearly defined units otherwise. Care should be taken to report only significant figures; i.e., figures that reflect the precision of the measurements. Data shown in figures or tables should not be described in the text. If the data are in non-natural units (i.e., not g, J, etc.), not related to such units (e.g., effects measured with visual-analog scales are often relatable to amount eaten), or not in units with known biological or clinical meaning, then effect sizes, such as Cohen's δ (Cohen, 1988, 1992), are accepted indices of meaningfulness.

The measurement scale determines the form in which the data should be reported. To describe central tendency, means are appropriate for ratio or interval scales, medians for ordinal scales, and modes for nominal scales. To describe spread, standard deviations and related measures (see the **Reporting** sections in **ANALYTIC STATISTICS**) are appropriate for ratio or interval scales, and the index of dispersion is appropriate for interval or nominal scales. The index of dispersion (D) is defined:

$$D = k(n^2 - \text{sum}\ f_i^2) / n^2 (k - 1),$$

where k is the number of categories or intervals, n is the number of data points, and f is the number of data points in each of the categories, i = 1 to k. Many texts recommend ranges for interval data, but this is incorrect because ranges are differences between data points, which are not meaningful for interval-scale measures.

SOME ISSUES in ANALYTIC STATISTICS

The statistical-significance approach

Categorical parametric approaches: t-tests and ANOVA. The most familiar analytic statistics, t-tests and analysis of variance (ANOVA) are considered categorical parametric statistics; *categorical* because the independent variable is different levels of some nominal or categorical measure (e.g., two sexes) rather than a continuous dimension (e.g., age), and *parametric* because they are based on mathematics assuming the Gaussian (normal) distribution. Such statistics require interval- or ratio-scale measurements. They are designed to test whether there are differences in the means of two (t-tests) or more (ANOVA) groups.

ANOVA approaches are applicable to a number of designs, including factorial designs, multivariate ANOVA, analysis of covariance, etc.

Assumptions. For t-tests, computer modeling has demonstrated that the assumption that the data are drawn from Gaussian distributions is not crucial; there is little risk of error as long as the distributions are unimodal and fairly symmetric. This is not the case for ANOVA. Rather, the distributions of all groups should be approximately Gaussian unless sample size is about 30 or more, their variances should be similar, groups sizes should be nearly equal (this is not crucial for one-way ANOVA), and, for repeated measures designs, the sphericity criterion should be met. Most computer statistics packages include tests of these criteria.

Note that ANOVA designs in which the independent variable arises from an ordinal, interval or ratio scale of measurement may also be analyzed with correlational approaches. The choice of which is more appropriate usually depends on the specific hypotheses being tested. Usually only one or the other type of analysis should be presented.

If the assumptions of parametric categorical approaches are not met, non-parametric approaches are called for (see **Nonparametric approaches).**

Interaction effects. Factorial ANOVA are almost universally analyzed by partitioning the variance among main effects, interaction effects and error, although it is entirely possible to partition variance without interactions. The choice whether to include interaction effects should be an educated one: [i] because interactions are defined as departures from additivity, unless the factors are themselves additive, the interaction makes little sense; [ii] similarly, if the independent variable is truly categorical, whether its levels are additive or not is impossible to determine; and [iii] any transform of the data can produce, or prevent, interactions. These issues are especially problematic in analyses of synergy (Caudle and Williams, 1993; Geary, 2013; Winer, 1971).

ANOVA follow-up. ANOVA and related approaches to determine statistical significance in experiments involving more than two groups are known as omnibus procedures because they yield overall estimates of statistical significance. These usually require follow-up tests to identify the specific source(s) of significance. A crucial aspect of these follow-ups is that they must protect the experiment- or analysis-wide α (see ***Multiplicity***).

Multiplicity. If several measures are used to test a single hypothesis (for example, different measures of the same underlying process), these should be regarded as a single family of tests, and it is necessary to maintain or protect the family-wide type 1 error rate (α, the probability of obtaining statistical significance when in fact there is no effect). In the absence of a hypothesis, descriptive statistics are preferred.

Type-1 error rates increase exponentially with the number of tests of the hypothesis (n). This is easily calculated by subtracting the probability of for making no type-1 errors from 1:

$$P[1 \text{ or more type-1 errors}] = 1 - (1 - \alpha)^n.$$

For example, if a brain-imaging study tests the hypothesis that a manipulation will increase neural activity in the limbic system, and 13 limbic areas are measured, then P[1 or more type-1 errors] $>$ 0.50.

There are two strategies to deal with the problem of multiplicity: [i] to maintain (or "protect") the experiment (or analysis)-wide type 1 error rate (α) or [ii] to maintain the false-discovery rate.

A number of follow-up tests have been developed in order to maintain (or "protect") the experiment (or analysis)-wide type-1 error rate have been derived. Some of these, however, have been determined to be defective and should not be used; these include multiple t-tests, (Fisher's) LSD test, and Dunnett's test. Others are valid, but unnecessarily "conservative," i.e., have poor power. This is the case for both the Tukey HSD test and the Bonferroni-corrected t-test. Hochberg (1988) and Rom (2013) describe simple, more powerful variations of Bonferroni-corrected t-tests.

Controlling the false-discovery rate, rather than the type-1 error rate, is a powerful and increasingly popular approach to the multiplicity problem. Curran-Everett (2000) provides an introduction.

It is important to note that all the Bonferroni variations and the false-discovery-rate strategies can be applied to both parametric and nonparametric analyses. Finally, it is important to appreciate the difference between simple and complex follow-up tests: the former are valid only to test individual group means; the latter must be used to test combinations of means, an issue that arises frequently (see ***Interactions & complex follow-up tests***). Note that computerized statistical packages offer only simple follow-up tests.

Interactions & complex follow-up tests. Interaction tests arise in designs comparing in two or more experimental effects. These situations require an explicit test of the difference in the two effects; it does not suffice to show that one effect is significant and the other is not. The appropriate test is an example of a complex follow-up test because it involves comparing two combinations of means ("the difference of differences"), i.e., comparison of two effects, where each effect is a difference between two means, such as control manipulation vs. test manipulation (for discussion, see: Nieuwenhuis et al., 2011). Computerized statistical packages do not offer such tests. They can be done with the methods mentioned under ***Multiplicity***.

Correlational approaches. Correlational or dimensional analyses are ordinarily the most appropriate approach for bivariate and multivariate data. As for categorical analyses, tools for both the statistical-significance approach (described here) and the estimation approach are available.

If both the independent and dependent variables are generated from interval- or ratio-scale measurements (see ***Measurement***), Pearson correlational analysis or corresponding multiple regression approaches should be used. If the design includes baseline measurements, these usually should be included as a covariate in an analysis of covariance. Logistic regression enables correlational analysis when the dependent variable is dichotomous, and Poisson regression, when the data describe the rate of occurrences of events in time.

Multiple groups should not be included in a single correlation unless each group appears to have the same slope and intercept as the overall correlation.

Collapsing dimensional data into categories to enable categorical analysis approaches (e.g., ANOVA) should be avoided.

Planned comparisons. Typical ANOVA follow-up tests for differences between pairs (Tukey's HSD test, etc.) often involve a large number of (if there are k groups in the ANOVA, there are C(k, 2) = k! / [2 (k-2)!] pairwise contrasts). Protecting the analysis-wide α leads to each comparison having rather low power. If several of these differences are not of interest, planned comparisons provide a more powerful alternative. A simple and adaptable planned-comparison method is to design the necessary comparisons and test them using, for example, the Hochberg or Rom variations of the Bonferroni method (see ***Multiplicity***). In the planned-comparisons approach, ANOVA is used simply to generate an experiment-wide SED, not to assess overall significance, according to the formula:

$$\mathrm{SED} = [\ 2\ \mathrm{MS}_{\mathrm{error}} / n]^{1/2} \quad (n = n/\mathrm{group}).$$

Power. Power refers to the probability of detecting an effect of a certain size. In the statistical-significance approach, power is defined as 1 - β, where β is the probability of a type 2 error, i.e., not detecting a significant effect when there is a group difference. Experiments should be designed with adequate power. Underpowered experiments reduce the probability both [i] true effects will be detected, and [ii] that significant results reflect true effects (Button et al., 2013). Note also that replicating significant results is expected to require larger sample sizes than used in the original study (Button et al., 2013).

***Reporting**.* Results of categorical statistical tests should be reported in standard detail; i.e. for ANOVA, report the F value, degrees of freedom, and probability: F(x,x,) = x.xx, P = 0.xxx. A precision of 0.001 ordinarily suffices for reporting statistics. Sample sizes should be given, for example in figure captions. If tables of statistical outcomes are appropriate, these should be given as supplementary data.

Reporting the exact P value rather than $P < 0.05$ (if 0.05 is the α level) is preferred because it provides more information. However, if a sequentially rejective approach such as the Hochberg-Bonferroni procedure (mentioned in ***Multiplicity***) is used, then $P < 0.05$ rather than exact P values should be reported. The reason for this is that in these procedures the differences are ordered by their magnitudes, which changes their probabilities of exceeding α levels, so that for all but the smallest difference, the α associated with the specific comparison is less than 0.05.

Pearson-type correlations should be reported with r or r^2 as well as the intercept and the signed slope, with its standard error (SE), and significance. Multiple regressions should be reported with both unstandardized slopes, each with its SE, and standardized (β) slopes as well as the significance.

Reporting variability brings several choices. The best choice is to report both the standard deviation as a measure of population spread and 95% confidence interval (assuming $\alpha = 0.05$) as a measure of the accuracy of the estimation of the mean. Carter (2012) describes the advantages of the 95% confidence interval over the standard error of the mean (SEM). Note that if data derive from repeated-measures designs, both SD and the usual SEM or 95% confidence interval conflate within- and between-subject variability; in such cases, standard errors of the difference (SED) or repeated-measures confidence intervals are more meaningful. Confidence intervals and standard errors of the estimate (SEE) are useful measures of the variability of correlated data.

The estimation approach.

Point and interval estimates. In the estimation approach is based on estimates of the values of the important experimental outcomes and their precision, i.e., the probability that the estimates fall in a certain range (the confidence interval; typically the 95% confidence interval). These two statistics are often called point and interval estimates (Cumming, 2012, 2014). Often the parameter estimated is the effect size, a dimensionless statistic that ranges from 0 to 1, with 0.2, 0.5, and 0.8 generally considered small, medium, and large effects, respectively (Cohen, 1988, 1992).

The estimation approach requires larger sample sizes to function than the statistical significance approach. Confidence intervals become quite large in small-sample experiments. Cumming (2014) states that if $n < 10$, confidence intervals are usually so large as to not be interpretable.

Classical probability theory enters the estimation approach in the computation of confidence intervals. Thus, the two methods are mathematically interconvertible (although this is often not trivial). Altman and Bland (2011) described some methods for this. As described in ***Interpretation***, however, the significance approach and the estimation approach are not epistemologically interconvertible.

Interpretation. In recognition that a single result is unlikely to be dispositive as to meaning, inferences are based on estimates and their precisions in a continuous way. Both the point and the interval estimates should be included in the interpretation. The underlying assumptions are [i] that the particular outcome of the experiment is just one of an infinite number of outcomes from the underlying sampling distribution, and [ii] that the best use of the data is in a future meta-analysis. Statistical significance is not assessed, and no particular importance is given to outcomes that would be statistically significant. Graphical displays are often especially effective, such as the two-dimensional cat's eye representation combining the length of the confidence interval and the shape of its sampling distribution.

Repeated-measures designs**.** Confidence intervals for repeated measures designs should be computed separately from those of the individual groups. Blouin and Riopelle (2005) and Masson and Loftus (2003) describe methods.

Power. Estimation approaches do not involve α, so there is no β and statistical power cannot be calculated. Instead, one specifies the size of the maximum confidence interval desired and uses the expected variance of the sample to calculate the sample size required to yield it (Maxwell et al., 2008; Cummings, 2014).

Reporting. In the estimation approach, point estimates (i.e., the sample means, etc.) and interval estimates (usually 95% confidence intervals) are reported. Group standard deviations and sample sizes should also be reported.

Nonparametric approaches

Nonparametric approaches, i.e., those that are not based on non-Gaussian probability models, are used for categorical- (nominal-) or ordinal-scale data or for interval-scale data that fail to meet the assumptions for parametric tests. Nonparametric tests are generally less susceptible to type-1 errors, but more susceptible to type-2 errors.

Statistical-significance approach. For categorical (nominal) measurements, the variations on the chi-squared test are usually the best choice: [i] determine whether there is are differences among the expected frequencies and the observed frequencies in one or more categories related to a single independent variable; [ii] McNemar chi-squared test for differences among the expected frequencies and the observed frequencies if there are paired categories, again with one independent variable; [iii] the Mantel-Haenszel chi-squared for differences among the expected frequencies and the observed frequencies in one or more categories related to two independent variables. These tests break down if the expected or observed frequencies in individual cells are $<$ 6. In this situation, Fisher's exact test can replace the chi-squared test.

The chi-squared distribution upon which the test is based comes up in many more contexts; for example, the expected value of sample variances follows the chi-squared distribution. Thus, the F distribution, which is the basis of ANOVA, is the ratio of two chi-squared distributions.

For ordinal (ranked) data, the Mann–Whitney–Wilcoxon test is an appropriate nonparametric version of t-tests for both independent and non-independent samples. It tests for differences in the central tendency (not means) of two groups. This test can be more powerful than the t-test if, for example, the data include extreme values. It is important to note that not all computerized statistics packages compute this statistic accurately (Bergmann et al., 2000). The Kruskal-Wallis and Friedman and tests are appropriate nonparametric versions of one-way ANOVA for independent samples and repeated-measures samples of ranked data, respectively.

Nonparametric approaches also require protection of the experiment- or analysis-wide α (see ***Multiplicity***).

Spearman's rho is an appropriate nonparametric measure of association if one or both variables is an ordinal-scale measurement.

Reporting. Because non-parametric tests use the ordinal structure of the data, central tendency should be reported with medians and are used. If the data are interval or ratio scale measures, spread may be reported with the MAD (see ***Extreme values***) or ranges, usually the semi-interquartile range.

The chi-squared test is an oddity: both the degrees of freedom and the sample size are required to specify the probability level, so both should be reported, along with the value of the test statistic and its probability. The Mann–Whitney–Wilcoxon, Kruskal-Wallis and Friedman tests depend only on the group sizes, so these should be reported together with and the test statistics and their probabilities. The significance of Spearman's rho is tested with a t-test and reported as described above.

Estimation approach. Nonparametric estimation methods are not as advanced as the nonparametric significance tests described above, although a number are under development (Brown and Levine, 2007; Powell, 2003, 2003; Soltanian and Hossein, 2012; Wang et al., 2012). Methods based on kernel-density estimation (Parzen, 1962; Rosenblatt, 1956) are beginning to appear more often in both exploratory data analysis (e.g., Harpole et al., 2014) and in analytic statistics (e.g., Miladinovic et al., 2014).

REFERENCES

Alderson, P. Absence of evidence is not evidence of absence. BMJ 328: 476-477, 2004.

Altman, D.G., Bland, J.M. Absence of evidence is not evidence of absence. *BMJ 311*: 485, 1995.

Altman, D.G., Bland, J.M. How to obtain the confidence interval from a P value. *BMJ 343*: d2090, 2011.

Bergmann R et al. Different outcomes of the Wilcoxon—Mann—Whitney test from different statistics packages. *Am Stat 54:* 72, 2000.

Blouin, D. C., Riopelle, A. J. On confidence intervals for within-subjects designs. *Psychological Methods, 10*: 397–412, 2005.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. *Introduction to meta-analysis*. Chichester, England: Wiley, 2009.

Bramness, J.G., Skurtveit, S., Gustavsen, I., Mørland. J. The absence of evidence is not the same as evidence for absence! Addiction 103: 513-514, 2008.

Brown, L.D., Levine, M. Variance estimation in nonparametric regression via the differences sequence method. *Annals of Statistics 35*, 2219–2232, 2007.

Button KS et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Rev Neurosci 14*: 365, 2013.

Carter, R.E. A standard error: distinguishing standard deviation from standard error. *Diabetes 62*: e15, 2013.

Caudle RM, Williams GM. The misuse of analysis of variance to detect synergy in combination drug studies. *Pain 55*: 313, 1993.

Cohen, J. A power primer. *Psychological Bulletin 112*: 155-159, 1992.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum, 1988.

Cooper, H. M. *Research synthesis and meta-analysis: A step-by-step approach* (4th edition). Thousand Oaks, CA: Sage, 2010.

Corning, J., Tukey, J.W. Average values of mean squares in factorials. *Annals of Mathematical Statistics 27:* 907-949, 1956.

Cumming, G. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge, 2012.

Cumming, G. The new statistics: when and why. *Psychological Science 25*: 7–29, 2014.

Curran-Everett D. Multiple comparisons: philosophies and illustrations. *Am J Physiol 279*: R1, 2000.

Geary, N. Understanding synergy. *Am J Physiol 304*: E237, 2013.

Gelman, A. Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review 73*: 369-382, 2003.

Harpole, J.K., Woods, C.M., Rodebaugh, T.L., Levinson, C.A., Lenze, E.J. How bandwidth selection algorithms impact exploratory data analysis using kernel density estimation. *Psychol Methods 19*: 428-443, 2014.

Hartung, J., Cottrell, J.E., Giffin, J.P. Absence of evidence is not evidence of absence. *Anesthesiology 58*: 298-300, 1983.

Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika 75*: 800, 1988.

Ioannidis, J. P. A. Why most published research finding are false. *PLoS Medicine 2:* e124, 2005.

Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Darnell RB, Ferrante RJ, Fillit H, Finkelstein R, Fisher M, Gendelman HE, Golub RM, Goudreau JL, Gross RA, Gubitz AK, Hesterlee SE, Howells DW, Huguenard J, Kelner K, Koroshetz W, Krainc D, Lazic SE, Levine MS, Macleod MR, McCall JM, Moxley RT 3rd, Narasimhan K, Noble LJ, Perrin S, Porter JD, Steward O, Unger E, Utz U, Silberberg SD. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature 490*: 187-191, 2012.

Masson, M. E. J., Loftus, G. R. Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale 57*: 203–220, 2003.

Maxwell, S. E., Kelley, K., & Rausch, J. R. Sample-size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*: 537–563, 2008.

Miladinovic, B., Kumar, A., Mhaskar, R., Djulbegovic, B. Benchmarks for detecting 'breakthroughs' in clinical trials: empirical assessment of the probability of large treatment effects using kernel density estimation. *British Medical Journal Open 4*:e005249, 2014.

Nieuwenhuis S et al. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neurosci 14*: 1105, 2011.

Parzen, E. On estimation of a probability density function and mode. *Annals of Mathematical Statistics 33*: 1065, 1962.

Powell, J.L. Notes On Nonparametric Density Estimation. University of California, Berkely, 2003. http://eml.berkeley.edu/~powell/e241a_sp10/ndnotes.pdf

Powell, J.L. Master class in semi- and non-parametric econometrics. Economic and Social Research Council, 2003. http://www.cemmap.ac.uk/resource/id/41

Rom DM. An improved Hochberg procedure for multiple tests of significance. *Br J Math Stat Psychol 66*: 189, 2013.

Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics 27*: 832, 1956.

Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. *J Am Stat Assoc 88*: 1273, 1993.

Sarle WS. *Measurement theory: Frequently asked questions*, Version 3 (SAS Institute, Cary NC USA, 1997; ftp.sas.com/pub/neural/measurement.html).

Simmons JP et al. False positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci 22*: 1359, 2011.

Soltanian, A.R., Hossein, M. A non-parametric method for hazard rate estimation in acute myocardial infarction patients: kernel smoothing approach. *J Res Health Sci 12*: 19-24, 2012.

Tufte, E.R. The Visual Display of Quantitative Information (2nd Ed.) Cheshire, CT, Graphics Press, 2001.

Tukey, J. W. *Exploratory data analysis*. Reading, MA: Addison-Wesley, 1977.

Wang, Q., Dinse, G.E., Liu, C. Hazard function estimation with cause-of-death data missing at random. *Ann Inst Stat Math 64*: 415-438, 2012.

Wainer, H. *Visual Revelations*. New York, NY. Copernicus – Springer, 2007.

Wainer, H., Velleman. P. Looking at blood sugar. *Chance 21*: 56-61, 2008.

Wilcox R. *Applying Contemporary Statistical Techniques.* Elsevier: Gulf Professional Publishing, Houston, TX, 2003.

Winer, B.J. Statistical principles in experimental design. McGraw-Hill, New York NY, 1971.