1521-0103/351/1/200–205$25.00
THE JOURNAL OF PHARMACOLOGY AND EXPERIMENTAL THERAPEUTICS

http://dx.doi.org/10.1124/jpet.114.219170
J Pharmacol Exp Ther 351:200–205, October 2014

Commentary

Common Misconceptions about Data Analysis and Statistics

Harvey J. Motulsky

GraphPad Software Inc., La Jolla, California

Received August 8, 2014; accepted August 8, 2014

ABSTRACT

Ideally, any experienced investigator with the right tools should be able to reproduce a finding published in a peer-reviewed biomedical science journal. In fact, however, the reproducibility of a large percentage of published findings has been questioned. Undoubtedly, there are many reasons for this, but one reason may be that investigators fool themselves due to a poor understanding of statistical concepts. In particular, investigators often make these mistakes: 1) P-hacking, which is when you reanalyze a data set in many different ways, or perhaps reanalyze with additional replicates, until you get the result you want; 2) overemphasis on *P* values rather than on the actual size of the observed effect; 3) overuse of statistical hypothesis testing, and being seduced by the word "significant"; and 4) over-reliance on standard errors, which are often misunderstood.

Introduction

Ideally, any experienced investigator with the right tools should be able to reproduce a finding published in a peer-reviewed biomedical science journal. In fact, however, the reproducibility of a large percentage of published findings has been questioned. Investigators at Bayer Healthcare were reportedly able to reproduce only 20–25% of 67 preclinical studies (Prinz et al., 2011), and investigators at Amgen were able to reproduce only 6 of 53 studies in basic cancer biology despite often cooperating with the original investigators (Begley and Ellis, 2012). This problem has been featured in a cover story in The Economist (Anonymous, 2013) and has attracted the attention of the National Institutes of Health leaders (Collins and Tabak, 2014).

Why can so few findings be reproduced? Undoubtedly, there are many reasons. But in many cases, I suspect that investigators fooled themselves due to a poor understanding of statistical concepts (see Marino, 2014, for a good review of this topic). Here I identify five common misconceptions about statistics and data analysis, and explain how to avoid them. My recommendations are written for pharmacologists and other biologists publishing experimental research using commonly used statistical methods. They would need to be expanded for analyses of clinical or observational studies and for Bayesian analyses. This editorial is about analyzing and displaying data, and so does not address issues of experimental design.

My experience comes from basic pharmacology research conducted decades ago, followed by 25 years of answering e-mail questions from scientists needing help analyzing data with GraphPad Prism,[1] and authoring three editions of the text *Intuitive Biostatistics* (Motulsky, 2014a).

Misconception 1: P-Hacking Is OK

Statistical results can only be interpreted at face value when every choice in the data analysis was performed exactly as planned, and documented as part of the experimental design. From my conversations with scientists, it seems that this rule is commonly broken in reports of basic research. Instead, analyses are often performed as shown in Fig. 1. Collect and analyze some data. If the results are not statistically significant but show a difference or trend in the direction you expected, collect some more data and reanalyze. Or try a different way to analyze the data: remove a few outliers; transform to logarithms; try a nonparametric test; redefine the outcome by normalizing (say, dividing by each animal's weight); use a method to compare one variable while adjusting for differences in another; the list of possibilities is endless. Keep trying until you obtain a statistically significant result or until you run out of money, time, or curiosity.

The results from data collected this way cannot be interpreted at face value. Even if there really is no difference (or no effect), the chance of finding a "statistically significant"

This Commentary evolved from multiple conversations between the author and several editors of pharmacology journals. This work represents the opinions of the author and should not be attributed to the American Society for Pharmacology and Experimental Therapeutics or the *Journal of Pharmacology and Experimental Therapeutics* (*JPET*) editorial board. Publication of this article in *JPET* does not represent an endorsement of GraphPad Software. This article is being simultaneously published in *JPET*, *British Journal of Pharmacology*, *Pharmacology Research & Perspectives*, and *Naunyn-Schmiedeberg's Archives of Pharmacology* in a collaborative effort to help investigators and readers appropriately use and interpret statistical analyses in pharmacological research studies.

dx.doi.org/10.1124/jpet.114.219170.

[1]http://www.graphpad.com/prism.

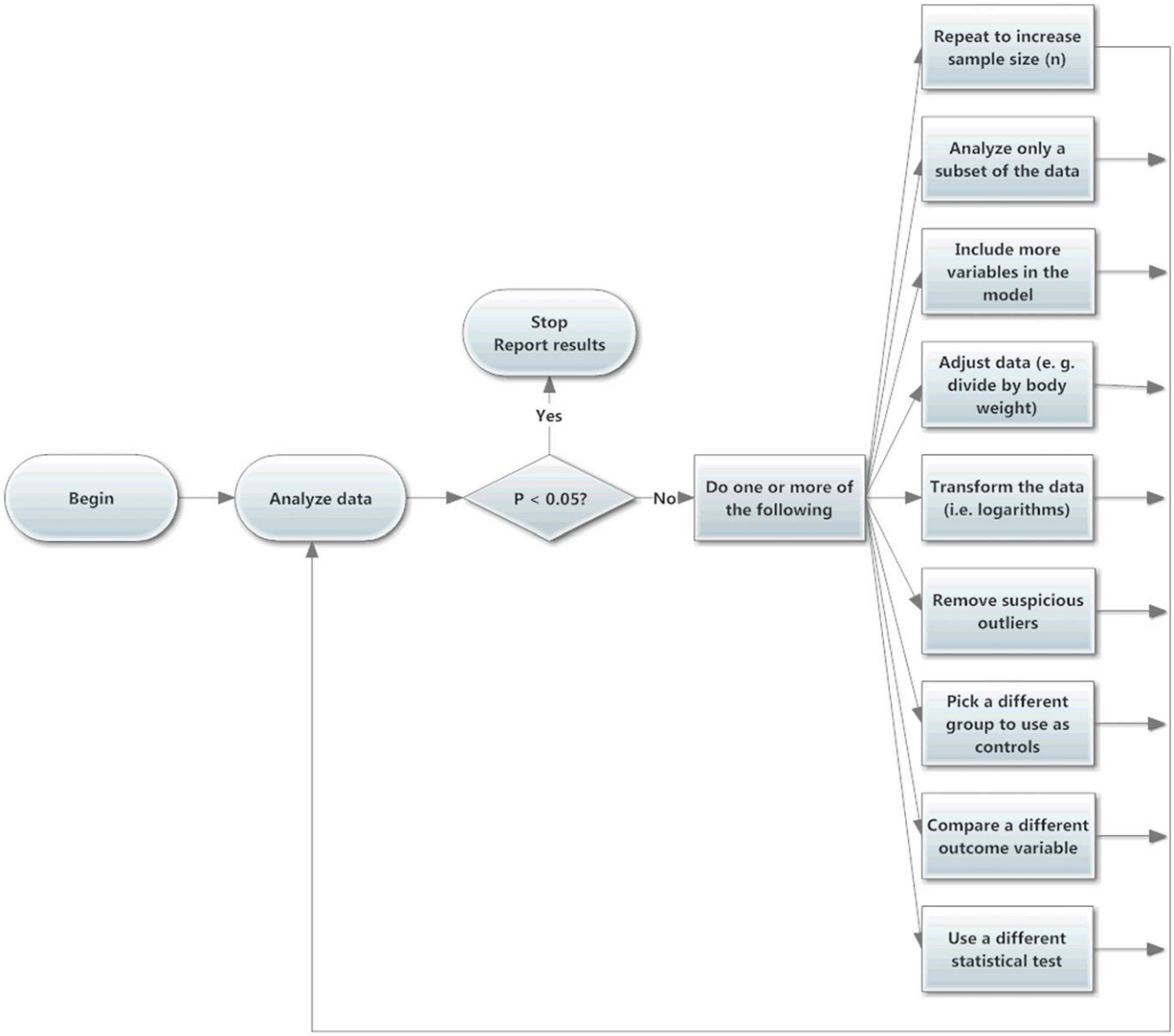

Fig. 1. The many forms of P-hacking. When you P-hack, the results cannot be interpreted at face value. Not shown in the figure is that, after trying various forms of P-hacking without getting a small *P* value, you will eventually give up when you run out of time, funds, or curiosity.

result exceeds 5%. The problem is that you introduce bias when you choose to collect more data (or analyze the data differently) only when the *P* value is greater than 0.05. If the *P* value was less than 0.05 in the first analyses, it might be larger than 0.05 after collecting more data or using an alternative analysis. But you would never see this if you only collected more data or tried different data analysis strategies when the first *P* value was greater than 0.05.

Exploring your data can be a very useful way to generate hypotheses and make preliminary conclusions, but all such analyses need to be clearly labeled, and then retested with new data.

There are three related terms that describe this problem.

Ad-Hoc Sample Size Selection. This is when you did not choose a sample size in advance, but just kept going until you liked the results. Figure 2 demonstrates the problem with ad-hoc sample size determination. Distinguish unplanned ad-hoc sample size decisions from planned "adaptive" sample size methods that make you "pay" for the increased versatility in sample size collection by requiring a stronger effect to reach "significance" (Food and Drug Administration, 2010; Kairalla et al., 2012).

HARKing, or Hypothesizing after the Result Is Known (Kerr, 1998). This is when you analyze the data many different ways (say different subgroups), discover an intriguing relationship, and then publish the data so it appears that the hypothesis was stated before the data were collected (Fig. 3). This is a form of multiple comparisons (Berry, 2007). Kriegeskorte and colleagues (2009) call this *double dipping*, as you are using the same data both to generate a hypothesis and to test it.

P-Hacking. This is a general term that encompasses dynamic sample size collection, HARKing, and more. It was coined by Simmons et al. (2011), who also use the phrase "too many investigator degrees of freedom." P-hacking is especially misleading when it involves changing the actual values analyzed. Examples include ad-hoc sample size selection (see earlier

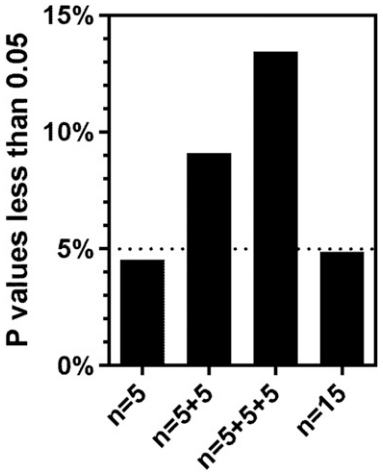

Fig. 2. The problem of ad-hoc sample size selection. I simulated 10,000 experiments sampling data from a Gaussian distribution with means of 5.0 and standard deviations of 1.0, and comparing two samples with $n = 5$ each using an unpaired t test. The first column shows the percentage of those experiments with a P value less than 0.05. Since both populations have the same mean, the null hypothesis is true, and so (as expected) about 5.0% of the simulations have P values less than 0.05. For the experiments where the P value was higher than 0.05, I added five more values to each group. The second column ($n = 5 + 5$) shows the fraction of P values where the P value was less than 0.05 either in the first analysis with $n = 5$ or after increasing the sample size to 10. For the third column, I added yet another 5 values to each group if the P value was greater than 0.05 for both of the first two analyses. Now 13% of the experiments (not 5%) have reached a P value less than 0.05. For the fourth column, I looked at all 10,000 of the simulated experiments with $n = 15$. As expected, very close to 5% of those experiments had P values less than 0.05. The higher fraction of "significant" findings in the $n = 5 + 5$ and $n = 5 + 5 + 5$ is due to the fact that I increased sample size only when the P value was high with smaller sample sizes. In many cases, when the P value was less than 0.05 with $n = 5$, the P value would have been higher than 0.05 with $n = 10$ or 15, but an experimenter seeing the small P value with the small sample size would not have increased sample size.

discussion), switching to an alternate control group (if you do not like the first results and your experiment involved two or more control groups), trying various combinations of independent variables to include in a multiple regression (whether the selection is manual or automatic), and analyzing various subgroups of the data. Reanalyzing a single data set in various ways is also P-hacking but will not usually mislead you quite as much.

My suggestions for authors are as follows:

- For each figure or table, clearly state whether the sample size was chosen in advance, and whether every step used to process and analyze the data was planned as part of the experimental protocol.
- If you use any form of P-hacking, label the conclusions as "preliminary."

Misconception 2: *P* Values Convey Information about Effect Size

To compute a P value, you first must clearly define a null hypothesis—usually that two means (or proportions or EC_{50} values, etc.) are identical. Given some assumptions, the P value is the probability of seeing an effect as large as or larger than you observed in the current experiment if in fact the null hypothesis was true. But note that the P value gives you no information about how large the difference (or effect) is. Figure 4 demonstrates this point by plotting the P values

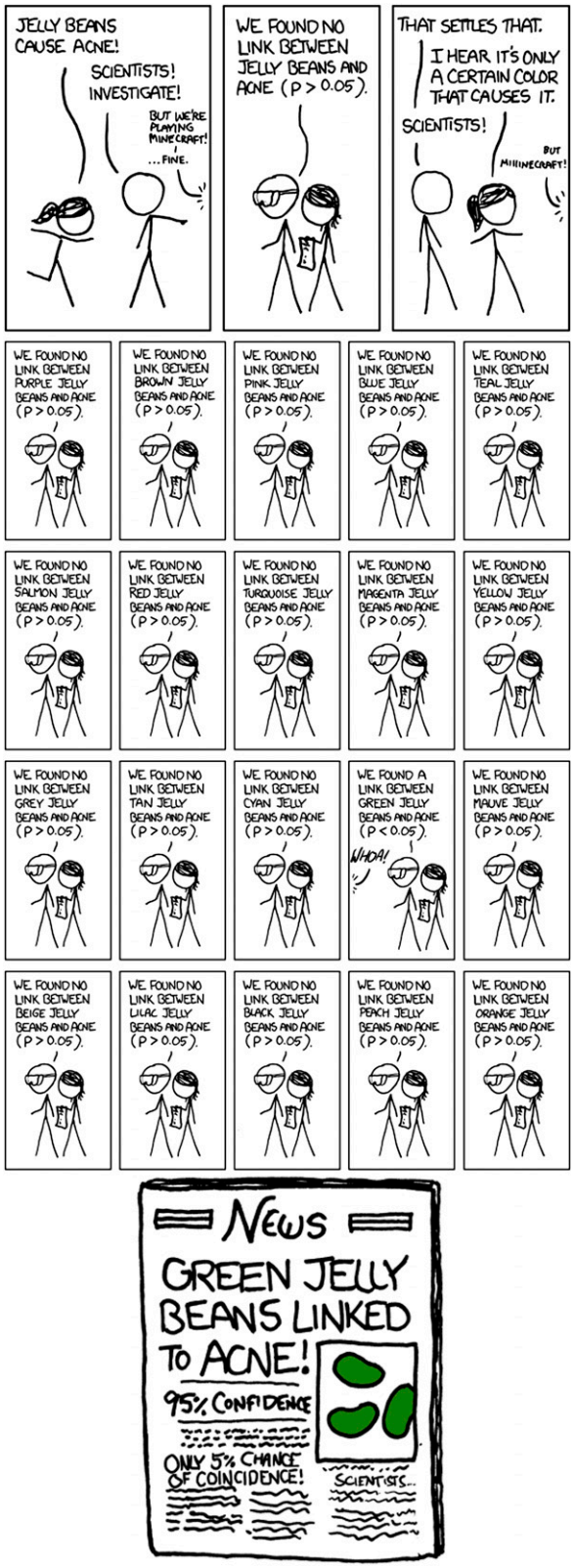

Fig. 3. The problem of HARKing. (Reprinted from http://xkcd.com/882 under the CC BY-NC 2.5 license.)

that result from comparing two samples in experiments with different sample sizes. Even though the means and standard deviations are identical for each simulated experiment, the P values are far from identical. With $n = 3$ in each group, the

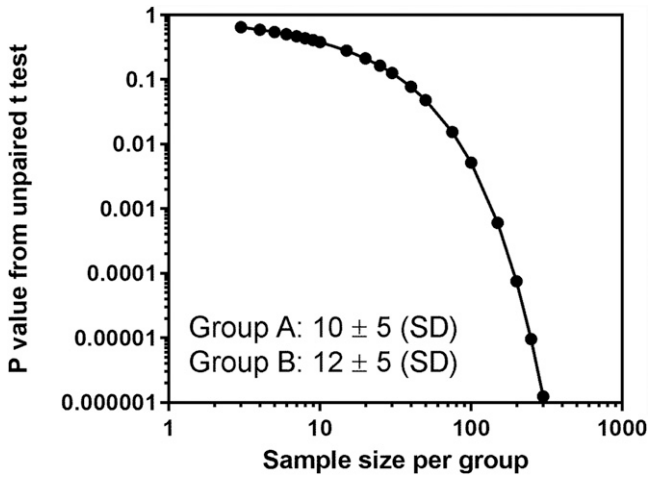

Fig. 4. *P* values depend upon sample size. This graph shows *P* values computed by unpaired *t* tests comparing two sets of data. The means of the two samples are 10 and 12. The S.D. of each sample is 5.0. I computed a *t* test using various sample sizes plotted on the *x*-axis. You can see that the *P* value depends on sample size. Note that both axes use a logarithmic scale.

P value is 0.65. When $n = 300$ in each group, the *P* value is 0.000001.

The dependence of *P* values on sample size can lead to two problems.

A Large *P* Value Is Not Proof of No (or Little) Effect. The top two rows of Table 1 presents the results of two simulated experiments. The two *P* values are both about 0.6, but the two experiments lead to very different conclusions.

In experiment A (from Table 1), the difference between means in the experimental sample is 10, so the difference equals 1% of the mean of treatment 1. Assuming random sampling from Gaussian populations, the 95% confidence interval for the difference between the two population means ranges from −30 to 50. In other words, the data are consistent (with 95% confidence) with a decrease of 3%, an increase of 5%, or anything in between. The interpretation depends on the scientific context and the goals of the experiment, but in most contexts these results can be summarized simply: the data are consistent with a tiny decrease, no change, or a tiny increase. These are solid negative data.

Experiment B is very different. The difference between means is larger, and the confidence interval is much wider (because the sample size is so small). Assuming random sampling from Gaussian populations, the data are consistent (with 95% confidence) with anything between a decrease of 18% and an increase of 28%. The data are consistent with a large decrease, a small decrease, no difference, a small increase, or a large increase. These data lead to no useful conclusion at all. An experiment like this should not be published.

A Small *P* Value Is Not Proof of a Large Effect. The bottom two rows of Table 1 presents the results of two simulated experiments where both *P* values are 0.001, but again two experiments lead to very different conclusions.

In experiment C (from Table 1), the difference between means in the experimental sample is only 2 (so the difference equals 2% of the mean of treatment 1). Assuming random sampling from Gaussian populations, the 95% confidence interval for the difference between the two population means ranges from 0.8 to 3.2. In other words, the data are consistent (with 95% confidence) with anything between an increase of 0.8% and an increase of 3.2%. How to interpret that depends on the scientific context and the goals of the experiment, but in most contexts this can be summarized simply: the data clearly demonstrate an increase, but that increase is tiny.

Experiment D is very different. The difference between means is 35 (so 35% of the control mean), and the confidence interval extends from an increase of 23.7% to an increase of 46.3%. The data clearly demonstrate that there is an increase that is (with 95% confidence) substantial.

My suggestions for authors are as follows:

- Always show and emphasize the effect size (as difference, percent difference, ratio, or correlation coefficient) along with its confidence interval.
- Consider omitting the reporting of *P* values.

Misconception 3: Statistical Hypothesis Testing and Reports of Statistical Significance Are Necessary in Experimental Research

Statistical hypothesis testing is a way to make a crisp decision from one analysis. If the *P* value is less than a preset value (usually 0.05), the result is deemed "statistically significant" and you make one decision. Otherwise, the result is deemed "not statistically significant" and you make the other decision. This is helpful in quality control and some clinical studies. It also is useful when you rigorously compare the fits of two scientifically sensible models to your data, and choose one to guide your interpretation of the data and to plan future experiments.

Here are five reasons to avoid use of statistical hypothesis testing in experimental research:

TABLE 1

Identical *P* values with very different interpretations

Experiments A and B have identical *P* values, but the scientific conclusion is very different. The interpretation depends upon the scientific context, but in most fields experiment A would be solid negative data proving that there either is no effect or that the effect is tiny. In contrast, experiment B has such a wide confidence interval as to be consistent with nearly any hypothesis. Those data simply do not help answer your scientific question. Similarly, experiments C and D have identical *P* values, but should be interpreted differently. In most experimental contexts, experiment C demonstrates convincingly that, although the difference is not zero, it is quite small. Experiment D provides convincing evidence that the effect is large.

	Treatment 1	Treatment 2	Difference between Means	*P* Value	95% CI of the Difference between Means
	mean ± S.D. (n)				
Experiment A	1000 ± 100 (50)	990.0 ± 100 (50)	10	0.6	−30 to 50
Experiment B	1000 ± 100 (3)	950.0 ± 100 (3)	50	0.6	−177 to 277
Experiment C	100 ± 5.0 (135)	102 ± 5.0 (135)	2	0.001	0.8 to 3.2
Experiment D	100 ± 5.0 (3)	135 ± 5.0 (3)	35	0.001	24 to 46

CI, confidence interval.

- The need to make a crisp decision based on one analysis is rare in basic research. A decision about whether to place an asterisk on a figure does not count. If you are not planning to make a crisp decision, the whole idea of statistical hypothesis testing is not helpful.
- Statistical hypothesis testing "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does" (Cohen, 1994). Statistical hypothesis testing has even been called a cult (Ziliak and McCloskey, 2008). The question we want to answer is: Given these data, how likely is the null hypothesis? The question that a *P* value answers is: Assuming the null hypothesis is true, how unlikely are these data? These two questions are distinct, and so have distinct answers.
- Scientists who intend to use statistical hypothesis testing often end up not using it. If the *P* value is just a bit larger than 0.05, scientists often avoid the strict use of hypothesis testing and instead apply the "time-honoured tactic of circumlocution to disguise the non-significant result as something more interesting" (M. Hankins, http://mchankins.wordpress.com/2013/04/21/still-not-significant-2/). They do this by using terms such as "almost significant," "bordered on being statistically significant," "a statistical trend toward significance," or "approaching significance." Hankins lists 468 such phrases found in published papers.
- The 5% significance threshold is often misunderstood. If you use a *P* value to make a decision, of course it is possible that you will make the wrong decision. In some cases, the *P* value will be tiny just by chance, even though the null hypothesis of no difference is actually true. In these cases, a conclusion that a finding is statistically significant is a *false positive*, and you will have made what is called a *type I error*.[2] Many scientists mistakenly believe that the chance of making a false-positive conclusion is 5%. In fact, in many situations, the chance of making a type I false-positive conclusion is much higher than 5% (Colquhoun, 2014). For example, in a situation where you expect the null hypothesis to be true 90% of the time (say you are screening lightly prescreened compounds, so expect 10% to work), you have chosen a sample size large enough to ensure 80% power, and you use the traditional 5% significance level, the false discovery rate is not 5% but rather is 36% (the calculations are shown in Table 2). If you only look at experiments in which the *P* value is just a tiny bit less than 0.050, the probability of a false positive rises to 79% (H. J. Motulsky, http://www.graphpad.com/support/faqid/1923/). Ioannidis (2005) used calculations such as these (and other considerations) to argue that most published research findings are probably false.
- The word "significant" is often misunderstood. The problem is that "significant" has two distinct meanings in science (Motulsky, 2014b). One meaning is that a *P* value is less than a preset threshold (usually 0.05). The other meaning of "significant" is that an effect is large enough to have a substantial physiologic or clinical impact. These two meanings are completely different, but are often confused.

My suggestions for authors are as follows:

- Only report statistical hypothesis testing (and place significance asterisks on figures) when you will make a decision based on that one analysis.
- Never use the word "significant" in a scientific paper. If you use statistical hypothesis testing to make a decision, state the *P* value, your preset *P* value threshold, and your decision. When discussing the possible physiologic or clinical impacts of a finding, use other words.

[2]In contrast, a type II error, or false-negative, is when there really is a difference but the result in your experiment is not statistically significant.

TABLE 2

The false discovery rate when $P < 0.05$

This table tabulates the theoretical results of 1000 experiments where the prior probability that the null hypothesis is false is 10%, the sample size is large enough so that the power is 80%, and the significance level is the traditional 5%. In 100 of the experiments (10%), there really is an effect (the null hypothesis is false), and you will obtain a "statistically significant" result ($P < 0.05$) in 80 of these (because the power is 80%). In 900 experiments, the null hypothesis is true, but you will obtain a statistically significant result in 45 of them (because the significance threshold is 5%, and 5% of 900 is 45). In total, you will obtain 80 + 45 = 125 statistically significant results, but 45/125 = 36% of these will be false positive. The proportion of conclusions of "statistical significance" that are false discoveries or false positives depends on the context of the experiment, as expressed by the prior probability (here, 10%). If you do obtain a small *P* value and reject the null hypothesis, you will conclude that the values in the two groups were sampled from different distributions. As noted earlier, there may be a high chance that you made a false-positive conclusion due to random sampling. But even if the conclusion is "true" from a statistical point of view and not a false positive due to random sampling, the effect may have occurred for a reason different from the one you hypothesized. When thinking about *why* an effect occurred, ignore the statistical calculations, and instead think about blinding, randomization, positive controls, negative controls, calibration, biases, and other aspects of experimental design.

	$P < 0.05$	$P > 0.05$	Total
Really is an effect	80	20	100
No effect (null hypothesis true)	45	855	900
Total	125	875	1000

Misconception 4: The Standard Error of the Mean Quantifies Variability

Pharmacology journals are full of graphs and tables showing the mean and the S.E.M.

Here is a quick review. The S.D. quantifies variation among a set of values, but the S.E.M. does not. The S.E.M. is computed by dividing the S.D. by the square root of the sample size. With large samples, the S.E.M. will be tiny even if there is a lot of variability.

One problem with plotting or displaying the mean ± S.E.M. is that some people viewing the graph or table may mistakenly think that the error bars show the variability of the data. A second problem with reporting means with S.E.M. is that the range mean ± S.E.M. cannot be rigorously interpreted. The S.E.M. gives information about how precisely you have determined the population mean. So the range mean ± S.E.M. is a confidence interval, but the confidence level depends on the sample size. With large samples, that range is a 68% confidence interval of the mean. When $n = 3$, that range is only a 58% confidence interval.[3]

My suggestions for authors are as follows:

- If you want to display the variability among the values, show raw data (which is not done often enough, in my opinion). If showing the raw data would make the graph hard to read, show instead a box-whisker plot, a frequency distribution, or the mean and S.D.
- If you want readers to see how precisely you have determined the mean, report a confidence interval (95% confidence intervals are standard). Figure 5 shows a data set plotted using all of these methods.
- When reporting results from regression, show the 95% confidence interval of each parameter rather than standard errors.

[3]Computed using this Excel formula: =(1-T.DIST.2T(1.0,2)). The first argument (1.0) is the number of S.E.Ms. (in each direction) included in the confidence interval, and the second argument (2) is the number of degrees of freedom, which equals $n-1$.

Misconception 5: You Do Not Need to Report the Details

The methods section of every paper should report the methods with enough detail that someone else could reproduce your work. This applies to statistical methods just as it does to experimental methods.

My suggestions for authors are as follows:

- When reporting a sample size, explain exactly what you counted. Did you count replicates in one experiment (technical replicates), repeat experiments, the number of studies pooled in a meta-analysis, or something else?
- If you eliminated any outliers, state how many outliers you eliminated, the rule used to identify them, and whether this rule was chosen before collecting data.
- If you normalized data, explain exactly how you defined 100% and 0%.
- When possible, report the *P* value up to at least a few digits of precision, rather than just stating whether the *P* value is less than or greater than an arbitrary threshold. For each *P* value, state the null hypothesis it tests if there is any possible ambiguity.
- When reporting a *P* value that compares two groups, state whether the *P* value is one- or two-tailed. If you report a one-tailed *P* value, state that you recorded a prediction for the direction of the effect (for example, increase or decrease) before you collected any data and what this prediction was. If you did not record such a prediction, report a two-tailed *P* value.
- Explain the details of the statistical methods you used. For example, if you fit a curve using nonlinear regression, explain precisely which model you fit to the data and whether (and how) data were weighted. Also, state the full version number and platform of the software you use.
- Consider posting files containing both the raw data and the analyses so other investigators can see the details.

Summary

The physicist E. Rutherford supposedly said, "If your experiment needs statistics, you ought to have done a better experiment."[4] There is a lot of truth to that statement when you are working in a field with a very high signal-to-noise ratio. In these fields, statistical analysis may not be necessary. But if you work in a field with a lower signal-to-noise ratio, or are trying to compare the fits of alternative models that do not differ all that much, you need statistical analyses to properly quantify your confidence in your conclusions.

I suspect that one of the reasons that the results reported in many papers cannot be reproduced is that statistical analyses are often performed as a quick afterthought, with the goal to sprinkle asterisks on figures and the word "significant" on conclusions. The suggestions I propose in this Commentary can all be summarized simply: If you are going to analyze your data using statistical methods, plan the methods carefully, do the analyses seriously, and report the data, methods, and results completely.

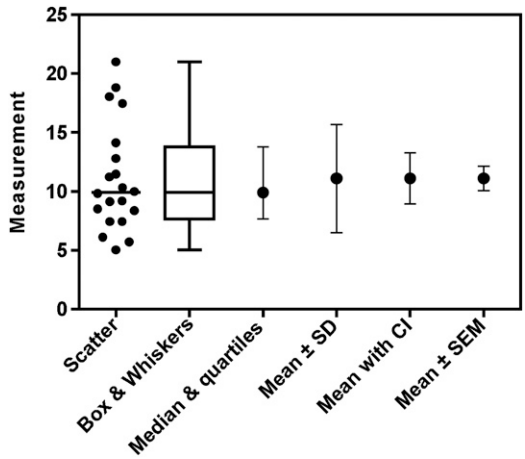

Fig. 5. Standard error bars do not show variability and do a poor job of showing precision. The figure plots one data set six ways. The left-most lane shows a scatter plot of every value, and so is the most informative. The next lane shows a box-and-whisker plot showing the range of the data, the quartiles, and the median (whiskers can be plotted in various ways, and do not always show the range). The third lane plots the median and quartiles. This shows less detail, but still demonstrates that the distribution is a bit asymmetric. The fourth lane plots mean with error bars showing plus or minus one standard deviation. Note that these error bars, by definition, are symmetrical and so give no hint about the asymmetry of the data. The next two lanes are different from the others as they do not show scatter. Instead they show how precisely we know the population mean, accounting for scatter and sample size. The fifth lane shows the mean with error bars showing the 95% confidence interval (CI) of the mean. The sixth (right-most) lane plots the mean plus or minus one standard error of the mean, which does not show variation and does a poor job of showing precision.

References

Anonymous (2013). Trouble at the lab. *The Economist* **409**:23–27.
Collins FS and Tabak LA (2014) Policy: NIH plans to enhance reproducibility. *Nature* **505**:612–613.
Begley CG and Ellis LM (2012) Drug development: Raise standards for preclinical cancer research. *Nature* **483**:531–533.
Berry DA (2007) The difficult and ubiquitous problems of multiplicities. *Pharm Stat* **6**:155–160.
Cohen J (1994) The earth is round ($p < 0.05$). *Am Psychol* **49**:997–1003.
Colquhoun D (2014) An investigation of the false discovery rate and the misinterpretation of P values [article online]. *arXiv*:1407.5296 (available from http://arxiv.org/abs/1407.5296).
Food and Drug Administration (2010) Guidance for industry: adaptive design clinical trials for drugs and biologics. Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD.
Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* **2**:e124.
Kairalla JA, Coffey CS, Thomann MA, and Muller KE (2012) Adaptive trial designs: a review of barriers and opportunities. *Trials* **13**:145.
Kerr NL (1998) HARKing: hypothesizing after the results are known. *Pers Soc Psychol Rev* **2**:196–217.
Kriegeskorte N, Simmons WK, Bellgowan PSF, and Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci* **12**:535–540.
Marino MJ (2014) The use and misuse of statistical methodologies in pharmacology research. *Biochem Pharmacol* **87**:78–92.
Motulsky HJ (2014a) *Intuitive Biostatistics*, 3rd ed, Oxford University Press, New York.
Motulsky HJ (2014b) Opinion: never use the word "significant" in a scientific paper. *Advances Regenerative Biol* **1**:25155, in press DOI: 10.3402/arb.v1.25155.
Prinz F, Schlange T, and Asadullah K (2011) Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* **10**:712–713.
Simmons JP, Nelson LD, and Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* **22**:1359–1366.
Ziliak S and McCloskey DN (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, University of Michigan Press, Ann Arbor, MI.

[4]The quotation is widely attributed to this famous physicist, but I cannot find an actual citation.

Address correspondence to: Harvey J. Motulsky, 1018 Yale Street, Santa Monica, CA 90403. E-mail: hmotulsky@graphpad.com