Title

Three common misuses of P values

Permalink
https://escholarship.org/uc/item/1kg520kr

Journal
Dental Hypotheses, 7(3)

ISSN
2155-8213

Authors
Kim, Jeehyoung
Bang, Heejung

Publication Date
2016

DOI
10.4103/2155-8213.190481

Peer reviewed

Published in final edited form as:

Dent Hypotheses. 2016 ; 7(3): 73–80. doi:10.4103/2155-8213.190481.

Three common misuses of P values

Jeehyoung Kim[1] and **Heejung Bang**[2]

[1] Department of Orthopedic Surgery, Seoul Sacred Heart General Hospital, Seoul, Korea

[2] Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, USA

Abstract

"Significance" has a specific meaning in science, especially in statistics. The p-value as a measure of statistical significance (evidence against a null hypothesis) has long been used in statistical inference and has served as a key player in science and research. Despite its clear mathematical definition and original purpose, and being just one of the many statistical measures/criteria, its role has been over-emphasized along with hypothesis testing. Observing and reflecting on this practice, some journals have attempted to ban reporting of p-values, and the American Statistical Association (for the first time in its 177 year old history) released a statement on p-values in 2016. In this article, we intend to review the correct definition of the p-value as well as its common misuses, in the hope that our article is useful to clinicians and researchers.

Keywords

confidence interval; multiple testing; p-value plot

> One sponsor reported to us that they had been looking at the data as each patient came in and stopped when the p-value was <0.05.... Peter A. Lachenbruch (2008)[1]

1. What is the p-value?

Let us revisit what the p-value (probability value) is in English, formula and graph. We believe that it is wise to learn what it is, before what it is not! We also think it is helpful to learn about null (H_0) vs. alternative (H_a) hypotheses and the two errors associated with these hypotheses. Commonly, we assume H_0: the difference is 0 vs. H_a: the difference is not 0 (or H_0 is not true), although one could take H_0 to represent a nonzero difference instead. Type I error (denoted by α) is the probability of rejecting H_0 when H_0 is true, i.e., false positive rate. Type II error (β) is the probability of not rejecting H_0 when H_a is true (or H_0 is false), i.e., false negative rate. α is also referred to as 'significance level'. [Note: R.A. Fisher and all others before J. Neyman and E. Pearson did not use an explicit alternative or talk of error rates. P-values can be defined and used without either concept.]

Corresponding Author: Jeehyoung Kim, MD, Department of Orthopedic Surgery, Seoul Sacred Heart General Hospital, Seoul, Korea, kjhnav@naver.com.

Conflicts of interest: Authors have no conflicts of interest.

The American Statistical Association (2016) provided this definition of the p-value:[2]

> Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value.

A more detailed definition may be: a p-value is the probability of obtaining a test statistic at least as extreme as the statistic *observed* under the H_0 (and every other assumption made), where the () part is important although it is often ignored or omitted. For 2 and 1-sided test, a mathematical formula and graphical depiction (Figure 1) can be provided as:

$$p-\text{value}=P\left(|test \quad statistic| \geq what \quad you \quad observed,\ under \quad H_0\right).$$

After setting up the test hypothesis and conducting a test (e.g., computing the p-value), we ordinarily compare the p-value with a *pre-specified* α, conventionally, 0.05. If the p-value is <0.05, we reject the H_0; if not, we do not reject the H_0.

An easy example would be a coin toss. If an unbiased coin is tossed properly many times, we would expect that about 50% of the time heads will face up. That is, if we toss 100 times, we will get heads about 50 times. But if we get heads 90% of the time, we would suspect this coin is biased or something strange is occurring. From this specific exercise, we can compute the p-value as the probability of heads facing up 90 times (i.e., what we observed) or more under the assumption that the truth is 50%, through a mathematical formula, and mark “x” n Figure 1.

2. Three common misuses

As documented elsewhere, there are many misuses of p-values and statistical procedures.[3] Here we focus on three common misuses.

2.1. Large p-value means no difference: Wrong

One property of the p-value is that it is a function of the sample size (N) (under the H_a). Thus, when N is large, the p-value is destined to be small; this feature can be a reward – acknowledging how hard it is to collect a large sample – but can cause other problems. Suppose that we observed the identical event rates (as well as risk difference (RD), odds ratio (OR) and risk ratio (RR) in the 2×2 tables in Figure 2). When N is doubled, the p-values can be meaningfully lower. Thus, statistical significance and acceptance vs. rejection of the H_0 could be different in these two scenarios.

If the p-value is above the pre-specified threshold (e.g., 0.05), we normally conclude that the H_0 is not rejected. However, it does not mean that the H_0 is true. The safer interpretation is that there is insufficient evidence to reject the H_0. Similarly, the ‘not H_0’ could suggest that something is wrong with the H_0, not necessarily that the H_a is right. It could be related to the assumptions of normality and independence, etc. which are often (unstated) parts of the H_0. A famous aphorism here is: *Absence of evidence is not evidence of absence*.

Then how to demonstrate ‘equality’ better? A more appropriate way is (bio)equivalence hypothesis testing, which is a norm in clinical trials (e.g., generic drug). Design and analysis should go hand in hand, whenever possible.

2.2. Multiple testing & 0.05

Let us assume that Thomas Edison, the Wright brothers, or you tried the same novel experiment 1000 times and finally succeeded. It is definitely a triumph. However, success out of one attempt and one success out of 50 attempts carry different meanings in terms of probability and statistics. For example, if you made a basket for the first time at your 50th basketball throw, it is important to disclose this. If you do not report 49 failures *intentionally or unintentionally*, one can misunderstand your performance. Similarly, if you try 50 different sports or conduct 50 laboratory experiments, a similar logic could apply. It is called the ‘multiple testing’ (‘multiple comparison’ or ‘multiplicity’) issue in statistics, with direct implications on α and the p-value. A simple and general – not perfect – rule of thumb solution is the Bonferroni adjustment, which is to use α=0.05/5=0.01 for 5 (independent) tests as a new threshold – or equivalently, inflate/adjust the observed p-values by multiplying by 5. One problem with this adjustment is that it drastically lowers the chance of detecting a real difference (the power of the test) if indeed there is one. As readers would be aware of, there are a variety of methods available for application in different contexts, each with slightly different properties, but having the same fundamental goal (e.g., Tukey's post-hoc; resampling version for correlated data; O'Brien-Fleming for interim analyses; and empirical Bayes methods).

The multiple testing issue often can take different forms in real life – sometimes in hidden and less clear manners – so that we do not realize we are running into multiple testing issues where a p-value adjustment (or at least some consideration thereof) might be warranted. For example, a study with multiple outcomes or multiple treatments/groups is quite clear. On the other hand, interim analyses/looks, subgroup analyses, multiple modeling, different categorizations of a variable (e.g., quartiles vs. quintiles), or searching for an optimal cutpoint (e.g., 3 or 4 cups of coffee) could be more or less subject to multiple testing issues.

The underlying mechanism of multiple testing may be well described as, “No Free Lunch,” “Fooled by Randomness,” “Forgone conclusion,” “File drawer problem,” “Leaving no trace,” and “If you torture the data enough, nature will always confess.” Acceptable solutions are 1) to designate a single Primary hypothesis (and outcome/parameter/method), while all others are secondary/sensitivity/confirmatory; 2) to reveal all analyses performed (under a given aim in one publication); 3) to present unadjusted vs. adjusted p-values side by side; or 4) a P-value Plot (when a number of p-values are computed). See Table 1 and Figure 3 as examples of 3) and 4).[4-6] Here, the underlying mathematical mechanism of the p-value plot is that p-values are uniformly distributed in 0-1 when the H_0 is true (and all other assumptions are met, regardless of N!).

A frequent and reasonable question from clinicians is, “Data do not change at all with or without multiple testing adjustment. Why should we care? Why don't you like presenting preliminary data for abstract submission?” Here are our answers: Multiple testing is more about ‘intention’ and the future likelihood of ‘replicability/reproducibility’ of the observed finding, rather than truth. Another way to view is: the data is the same, ordered or not. Suppose you compute 100 p-values, then order them from the smallest to the largest (as in the p-value plot). The smallest p-value is the 1^{st} order statistic. It is not a single random p-value. Suppose that you rank ordered the students in a school from the shortest to the tallest. The smallest student is not representative of students. In the case of p-values, there are ways to adjust for the fact that you are looking at the smallest p-value.

Interestingly, however, even leading statisticians do not fully agree concerning whether to adjust and how to adjust; thus, if you are against adjustment, you are not alone. Moreover, exploratory nature and serendipity in scientific discovery and advancement should *never* be undervalued. Indeed, some even do not think that the multiple testing problem really exists, asking how one can report thousands, millions, or billions of results? What we really have is a 'selective reporting problem'.[7] Most would agree that one out of one vs. one out of 50 can be interpreted differently, and readers/judges have a right to know this along with other details (e.g., method used). If authors/investigators are honest or willing, it is not difficult to do.

Another common practice we see is a disproportionate focus on 'false positive', compared to 'false negative.' Sometimes, the consequences of false negatives can be much greater than those of false positives, and the importance of α vs. β should be carefully considered and context-based, rather than handling/deciding them mechanically.[8]

Back to the original example, why do we not talk about multiple testing issues relative to Edison and the Wrights? Possibly because their experiments and successes would be replicated at the 1001th trial and beyond. Even if we applied multiple testing adjustment for their first 1000 trials (e.g., pilot) and they did not pass $p<0.05$, a 'new' rigorously designed protocol and experiment, including a priori hypothesis and N/power calculation, would easily pass $p<0.001$!

2.3. Smaller p-value is more significant? Not necessarily

We have discussed the well-known 'large N→small p' phenomenon. Below we illustrate that 'smaller p-value, smaller effect' can happen, when Ns are different. Another philosophical question may be: Which more strongly supports the effect, 'a large effect size from a small sample' vs. 'a small effect size from a large sample'? The answer can vary and may be not straightforward; yet we are easily convinced that 'sole reliance on p-values' can be problematic.

We assert that estimate (point and interval) and p-value can be complimentary, but each with advantages and disadvantages: the former better addresses clinical or practical significance and the latter addresses statistical significance, where clinical significance is a more important goal although it is not an easy concept or task.

3. Additional issues

3.1 Notable companion of the p-value: confidence interval (CI)

There are two ways to view a statistical hypothesis test: one is through a p-value (of the test) and the other is through a CI (of a parameter). Many busy clinicians use a simple rule, “If p<0.05, or the CI does not cover the null value, H_0 is rejected.” in practice. The p-value and CI are complementary while attempting to do the same/similar thing, where the p-value quantifies how ‘significant’ the association/difference is, while the CI quantifies how ‘precise’ the estimation is and what the plausible values are.

Ironically, however, another dominating measure in statistics, CI, does not have an easy definition. Perhaps, the shortest interpretation and definition of a 95% CI is: (a,b) is the set of all values with p>0.05 under the data-generation model. A more detailed definition may be: a 95% CI for a parameter (e.g., mean or OR) has the property that for *many* independent replications of the same experiment, approximately 95% of the CIs contain the true parameter. Here, the parameter is fixed and intervals are random! Thus (unfortunately), we need 1000 experiments under identical conditions in our brain when we try to understand CI properly. The following definition and its variants appear in top medical journals and editorials, often written by (bio)statisticians: “With 95% confidence, the population mean will lie in this interval.” This may be justified only in the sense of, “Perfect is the enemy of good.”

The point estimate plus or minus its ‘margin of error’ is a CI for the parameter of interest, where the margin is determined by the variability of the point estimate, so called via standard error, which decreases when N increases. Here, it is critical to know that standard error (margin of error, CI and p-value as well) accounts for ‘random sampling error’ only, not for other errors and numerous biases from other sources, including poorly worded questions, false answers, wrong/mis-specified model, and flawed/inadequate design in survey or experiment.

In the current literature (e.g., BigData or meta-analysis), you may find something like OR=3.11 [95% CI: 3.10-3.12], an extremely narrow CI. No one would believe the truth is really inside! Imagining *hypothetical* 1000 experiments and the margin of error in your mind, you would not be surprised by this interval, and you would also naturally understand the potential limitations of CI. With the pros and cons of each method, reporting all 3 (the point estimate, 95% CI, and p-value) would be advisable.

Another common practice in the use of CI and p-value is that readers often check if the two CIs do overlap in order to judge statistical significance. A rule of thumb is that non-overlapping CIs implies significant difference, but not the reverse: the two CIs may overlap and yet be significantly different, as long as each CI does not contain both point estimates. More on this topic can be found in a reference[9].

Note: Common technical mistakes in CIs are: 1) we want to attach the probability statement about CI. But, strictly speaking, we should not say, “95% likely or probable,” which is like saying, “95% chance of rain yesterday”; and 2) we tend to assume interval is fixed and truth is random or fixed. How to interpret Prob(89<true blood pressure<122)=0.95? This probability is 0 or 1.[10] In the classical, Frequentist approach the randomness comes from the repetition of experiments, while in the Bayesian approach the randomness comes from uncertainty about the value of the parameter, which could be more appealing and pertinent. Bayesian interval is often called, 'credible interval.' For rigorous definitions, properties and fallacies of CIs, see references.[3,11,12]

3.2. Reproducibility of the p-value

Today, the reproducibility of scientific finding under the Responsible Conduct of Research has become a component of various training programs because irreproducible/nonreplicable findings are unacceptably common (e.g., Random Medical News) in the competitive research arena. Actually, this is an old news.[13-17] Sadly, but unsurprisingly, statistics (and the p-value) is a big player there. In contrast to other statistical estimates, the p-value's sample-to-sample variability is not fully appreciated.[18] Related to reproducibility for future replicate p-values, it has been shown that p-values exhibit surprisingly large variability in typical data, and some call for lower p-value thresholds such as 0.005 or 0.001 (although randomized controlled trial (RCT) and laboratory science communities may be upset!).[19]

In addition, by definition p-value depends on effect size (e.g., observed difference and variability). In turn, observed difference and variability also depend on study design, sample selection, measurement, and method, among others. In an extreme scenario, if the two comparison groups do not overlap (e.g., cases vs. controls), we can get an impressively low p-value and perfect discrimination (AUC=1). This may indicate flawed design, such that the resulting p-values or comparison per se can be misleading or meaningless. If we use a more suitable design and sample even for the exactly same comparison, the previous small p-value would not be reproduced. In some sense, 'too good to be true' statistics (e.g., AUC≈1, p<0.0001) is a blessing, by effectively serving as an alarm to investigators as well as reviewers/editors, e.g., "Do not publish findings yet; more checking is needed." You may want to check out the *impressive* p-values in the famous 'vitamin C and terminal cancers' and 'vaccines and autism' papers in history[20,21]. If you were a reviewer and saw these p-values (and AUC≈1), what would you say even if you don't know underlying science well?

A more common scenario is when testing the same hypothesis in different populations (e.g., low vs. high risk groups; American vs. Asian), the observed p-values can be vastly different even with the same N, which is natural due to the different effect sizes expected. Statistical inference is generally based on 'hypothetical' experiments (e.g., randomized, independent, sampling bias only) and mathematical formulations; to compare, real-world settings can be much more complex. Thus, limited reproducibility in p-values and varied performance of any model (e.g., prediction) in different settings/contexts are to be anticipated.

Related to reproducibility concerns and countless biases in practice, some people focus on large effect sizes (e.g., OR>2, in addition to or in place of p<0.05). Yet, we should not ignore small but real effects or rare cases/events, which may be potentially translated to large total (or cumulative) effect or expense at the population or society level. The p-value and CI cannot answer the meaningfulness and clinical or public health *significance* of 'losing 100 grams' and 'living 3 days longer after cancer screening,' which should be judged together with societal and individual perspectives and values (and possibly cost-effectiveness).

A possible solution for the irreproducibility crisis may be: In God we trust; all others must bring data (protocol and SAS output).

3.3. Large p with small N; post-hoc power to blame?

Let us imagine a common situation. When we finished data collection and analysis, we got p=0.2 for the primary hypothesis test. Naturally, we are disappointed after a long and hard work and tempted to find reasons, including anyone to blame? I guess low N (e.g., budget, boss's recommendation, wrong assumptions used in N/power calculation, etc.) and low observed power may be good victims for the post-hoc blame game because we already know "larger N → smaller p" when H_0 is false. There are debates regarding this issue – to compute post-hoc power or not. Recall power=1-β=1-Type II error, but post-hoc or observed power is not 1-β (say, 80%)! It is the same fallacy as for misinterpreting the CI: *After the data are in*, the CI either does or does not contain the true value (1 or 0, not 95%). In the same way, the Type II error after the analysis is either 1 or 0.

Somewhat depressingly, this is another controversial topic in statistics because widely used biostatistics textbooks and some instructors teach how to calculate post-hoc power. So again, you are not alone. Some recommend: once a study is over, we should focus on precision, rather than power, noting that, for any test, the observed power is a 1:1 function of the p-value.[22] Even if post-hoc power could be useless once one sees the p-value and CI, the results of a study can and should be used to design subsequent studies, since those results provide information about the crucial parameters used to estimate the N for subsequent studies (such as the size of the effect and the exposure or disease frequency one should expect). The CONSORT 2010 (item 17) also stated: there is little merit in a post-hoc calculation of statistical power using the results of a trial; the power is then appropriately indicated by CIs. [Remark: power should be precision here.]

3.4. Final decision always binary (i.e., p< vs. >0.05)?

A binary decision may be needed at courts, in sports, or on a job application; however, should scientific decision making always be 2 regions? We hope not! There are trinary decisions and hypotheses: accept (p<0.05); reject (p>0.2); and get more data (0.05<p<0.2). Indeed, some understand this as Fisher's original suggestion for using p-values. Adoption of trinary decisions instead of the current paradigm, e.g., "The earth is round (P<.05)." might be ideal in practice.[23]

4. Final thoughts and some recommendations

There are few numerical numbers/measures/tools both as common and controversial as the p-value in science and research (e.g., p-hacking, p-value chasing, fickle p-value, the cult of statistical significance). This one measure is often believed to govern our career (e.g., funding, publication) and the destinies of pharmaceutical companies. We view the main reasons for p-value's popularity and charm as being: an easily computed number, probability (within 0-1) with an interpretation of statistical significance, with high generality and universality, and solid mathematical foundations. Therefore, the p-value must be doing what it is supposed to do well; it is more likely that we misuse/abuse it for what it is not expected to do. If a tool has long been used by many, there are always reasons, generally more good than bad (but not necessarily so, as the history of practices like bloodletting reveal).

Its popularity and dominance in medicine is noteworthy, in part because simple and fast decision making (e.g., does this treatment work? is it best for this patient?) is needed on a daily basis. Regardless of its limitations, we expect the popularity of the p-value to continue. Even in the face of criticism, cynicism, or even banning of p-values, abandoning baseball statistics and the car because they are imperfect or throwing the baby out with the bathwater would be counter-productive or unwise.[24,25] Dennis Lindley, a leading Bayesian statistician in history, did not believe in significance testing, but he taught it at Cambridge![26] Better and feasible guidance on use and interpretation is of more use than outright rejection, which is virtually true for many other statistics, e.g., OR vs. RR, absolute vs. relative, kappa, and so on.

A recent survey of 1,576 researchers by *Nature* picked “Selective reporting” as #1 factor of irreproducible research and “Better understanding of statistics” as #1 solution.[7] We want to conclude our article with two suggestions for practitioners.

4.1 Better design, better data, better p-value

Design trumps Analysis and Experimentation trumps Observation in scientific research, and ‘Garbage in, Garbage out’ is so relevant to statistical analyses[27,28]. It is nearly impossible to expect valid statistical analysis (including p-values) from poor quality design and/or data. Fisher said, “To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.” More energy and emphasis should be given to more basic and fundamental components and steps: study design, measurement, data collection, blinding (if relevant), and protocol, including data analysis plan to help minimize retrospective rationalization and fishing expeditions. All of these are important prerequisites to valid p-values.

It is essential to remember that design can address causality, whereas p-value can only address numerical correlation/association in a given model/setting. In that sense, the old terms of ‘effect’ and ‘effect size’ (which the authors of this paper also used! And standard error is error, really?) could be the source of many wrongdoings. Also, similar to the *Almighty p-value*, the RCT is extremely beloved in comparative and evaluative research, namely, Trialism.[29,30] Yet it is crucial to understand the best gift that the RCT offers is *average* causal effect (of “intention”, again). Sick populations and sick individuals are not the same thing.[31] As Stephen Jay Gould said from his own long experience with cancer, perhaps “The median isn't be the message”[32]... Yet, if the median is accompanied by good intervals (e.g., 95% CI, inter-quartile range, min-max), it could carry useful messages, even for others than the Average Man.[33]

4.2. Beyond p-value and validation: toward total evidence

No single measure/method is perfect. Notwithstanding the pros and cons of p-value and CI, good scientists would look for both – and possibly more. Commonly utilized measures in biomedicine include: effect size (e.g., RD/OR/RR); point and interval estimates; statistical significance (p-value); discrimination (AUC); model determination (R^2); correlation (Pearson, Spearman); model quality (AIC/BIC), etc. For instance, it is not rare to encounter 'small p and small R^2', 'more predictors but lower AUC', or '$p<0.0001$ in one study and >0.05 in another'; these look ostensibly contradictory, but if we understand each method's capability, we would not be surprised, even if both studies were well done.

Furthermore, we should always seek non-statistical evidence, such as a theory, scientific explanations, qualitative evidence, and evidence from basic science (e.g., N=3) whenever available. Theory serves as a base for thinking and it helps us to understand what is really going on. Data have no scientific (or empirical) meaning without theory. As Charles Darwin said, without speculation there is no good and original observation. If you study the association of the number of refrigerators and the crime rate, the finding may be validated easily in different cities. Also, in-hospital mortality has been shown to be inversely related to the number of cardiovascular risk factors.[34] But for the both cases, we can look for possible explanations of such an association.

Through this editorial, we hope to help readers better understand and use the p-value. At the end of the day, we should wait for total evidence through (sensible) validation to lead us closer to an ultimate answer for a given setting, although it takes time and resources. If we use p-values correctly and wisely, we can shorten the time of this journey and save the resources. Maybe now is the time to move over "Publish or Perish" to "Validate or Vanish."

Acknowledgment

We thank Drs. Sander Greenland, Robert Elston, Stan Young, Dmitri Zaykin and Ross Gibson for reading our draft, providing valuable comments and criticisms, and/or educating authors.

Financial support: H. Bang was partly supported by the National Institutes of Health through grants UL1 TR000002 and P50 AR063043.

References

1. Lachenbruch, P. Some clinical trial design questions and answers Northeastern Illinois Chapter American Statistical Association Meeting. 2008. http://www.amstat.org/chapters/northeasternillinois/pastevents/fall08.htm
2. Wasserstein R, Lazar N. The ASA's statement on p-values: context, process, and purpose. The American Statistician. 2016
3. Greenland S, Senn S, Rothman K, et al. Statistical tests, p-values, confidence intervals, and power: a guide to misinterpretations. The American Statistician. 2016 Online Supplement:1-12.
4. Schweder T, Spjøtvoll E. Plots of P-values to evaluate many tests simultaneously. Biometrika. 1982; 69:493–502.
5. Young SS, Bang H, Oktay K. Cereal-induced gender selection? Most likely a multiple testing false positive. Proc Biol Sci. 2009; 276:1211–1212. [PubMed: 19141426]
6. Peto R, Emberson J, Landray M, et al. Analyses of cancer data from three Ezetimibe trials. NEJM. 2008; 359:1357–1366. [PubMed: 18765432]
7. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016; 533:452–454. [PubMed: 27225100]
8. Neyman, J. Frequentist probability and frequentist statistics. Springer; 1977.

9. Schenker N, Gentleman J. On judging the significance of differences by examining the overlap between confidence intervals. The American Statistician. 2001; 55:182–186.

10. Elston, R.; Johnson, W. Basic Biostatistics for Geneticists and Epidemiologists: A Practical Approach. Wiley; 2008.

11. Morey R, Hoekstra R, Rouder J, et al. The fallacy of placing confidence in confidence intervals. Psychonomic Bulletin & Review. 2016; 23:103–123. [PubMed: 26450628]

12. Greenland S, Poole C. Problems in common interpretations of statistics in scientific articles, expert reports, and testimony. Jurimetrics J. 2011; 51:113–129.

13. Shapiro S. Looking to the 21st century: have we learned from our mistakes, or are we doomed to compound them? Pharmacoepidemiology and drug safety. 2004; 13:257–265. [PubMed: 15255093]

14. Breslow N. Are statistical contributions to medicine undervalued? Biometrics. 2003; 59:1–8. [PubMed: 12762435]

15. Greenwood M. Is statistical method of any value in medical research? Lancet. 1924; 204:153–158.

16. Frey B. Publication as prostitution. Public Choice. 2003; 116:205–223.

17. Mantel N. How to guarantee significance. The American Statistician. 1976; 30:201–202.

18. Boos D, Stefanski L. P-Value precision and reproducibility. The American Statistician. 2011; 65:213–221. [PubMed: 22690019]

19. Johnson V. Revised standards for statistical evidence. PNAS. 2013; 110:19313–19317. [PubMed: 24218581]

20. Cameron E, Pauling L. Supplemental ascorbate in the supportive treatment of cancer: Prolongation of survival times in terminal human cancer. PNAS. 1976; 73:3685–3689. [PubMed: 1068480]

21. Wakefield A, Murch S, Anthony A, et al. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet. 1998; 351:637–641. [PubMed: 9500320]

22. Hoenig J, Heisey D. The abuse of power: the pervasive fallacy of power calculations for data analysis. The American Statistician. 2001; 55:19–24.

23. Cohen J. The earth is round (p<0.05). American Psychologist. 1994; 49:997–1003.

24. Woolston C. Psychology journal bans P values. Nature. 2015; 519:9.

25. Baker M. Statisticians issue warning over misuse of P values. Nature. 2016; 531:151. [PubMed: 26961635]

26. Zheng G, Li Z, Geller N. A conversation with Robert C. Elston. Statistical Science. 2015; 30:258–267.

27. Rubin DB. For objective causal inference, design trumps analysis. The Annals of Applied Statistics. 2008; 2:808–840.

28. Bang, H. Introduction to observational studies. In: Faries, D.; Leon, A.; Haro, J., et al., editors. Analysis of Observational Health-Care Data Using SAS. SAS Press Series; 2010.

29. Peck C. The almighty p-value or the significance of significance. Present Concepts Intern Med. 1971; 4:1021–1024.

30. Rimm A, Bortin M. Trialism: The belief in the Holy Trinity clinician—patient—biostatistician. Biomedicine Special Issue. 1978; 28:60–63.

31. Rose G. Sick individuals and sick populations. International Journal of Epidemiology. 1985; 14:32–38. [PubMed: 3872850]

32. Gould, S. The Median isn't the message. 2002. http://cancerguide.org/median_not_msg.html

33. Quetelet, A. Sur l'homme et le développement de ses facultés, ou Essai de physique sociale. Bachelier, imprimeur-libraire, quai des Augustins; Paris: 1835.

34. Canto J, Kiefe C, Rogers W, et al. Number of coronary heart disease risk factors and mortality in patients with first myocardial infarction. JAMA. 2011; 306:2120–2127. [PubMed: 22089719]

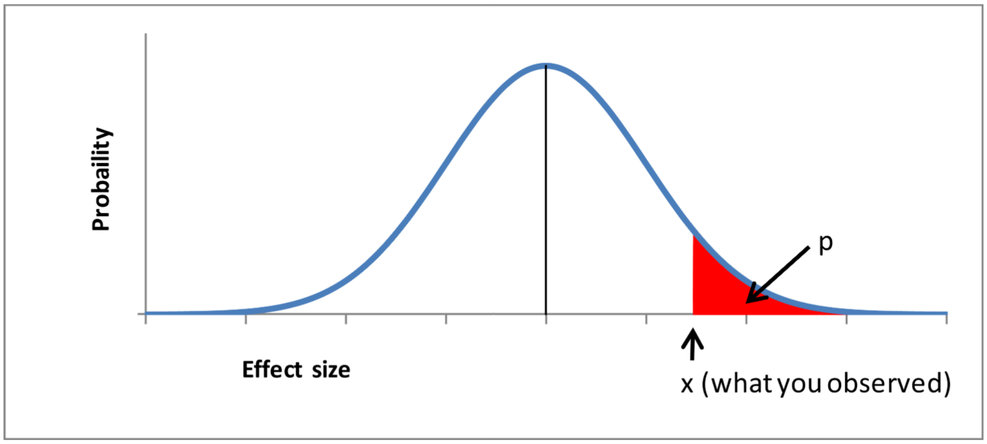
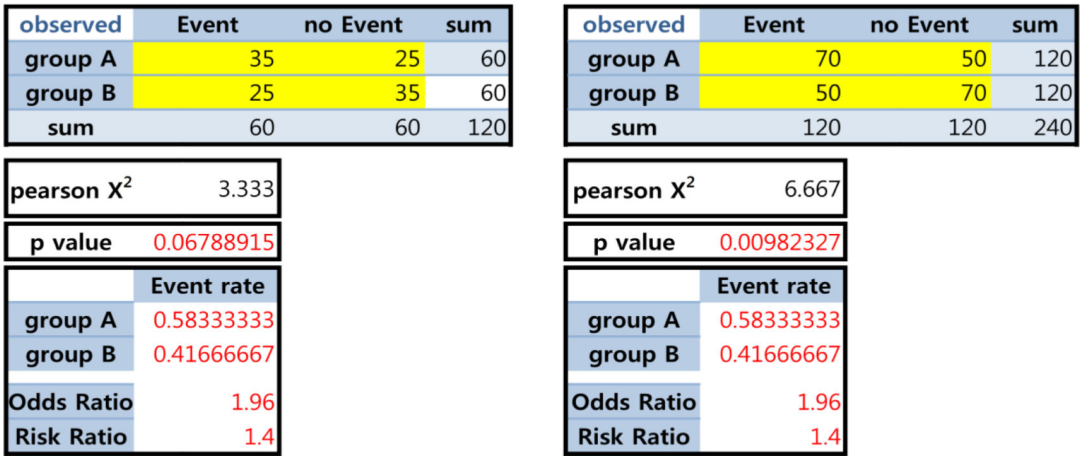


Figure 1.

One-sided p-value for H_0: effect no greater than θ_0



observed	Event	no Event	sum
group A	35	25	60
group B	25	35	60
sum	60	60	120

pearson X^2	3.333
p value	0.06788915

	Event rate
group A	0.58333333
group B	0.41666667
Odds Ratio	1.96
Risk Ratio	1.4

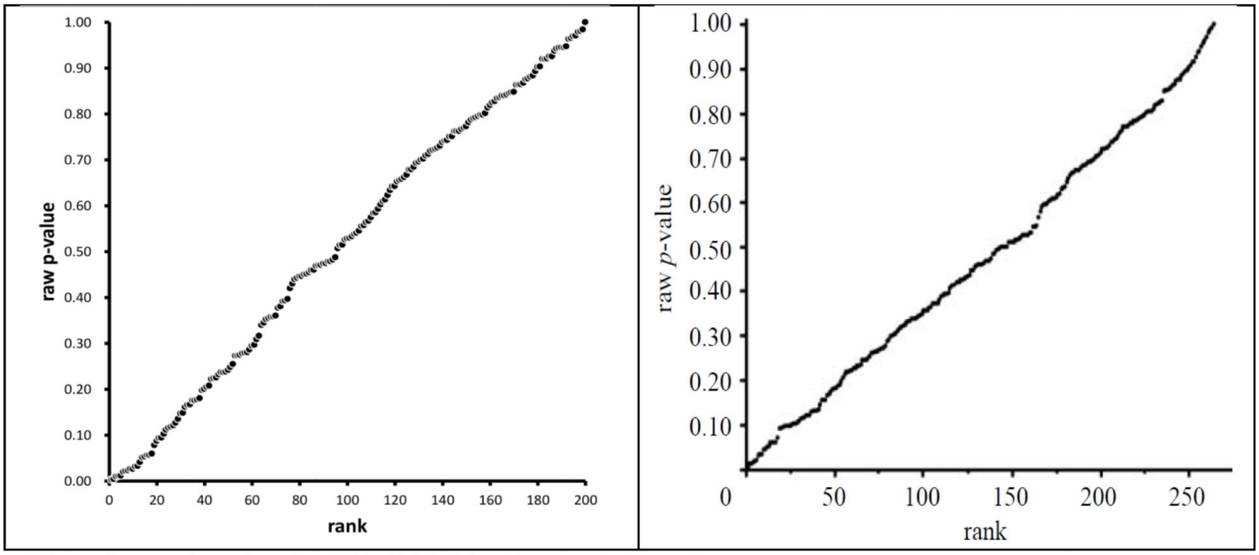
observed	Event	no Event	sum
group A	70	50	120
group B	50	70	120
sum	120	120	240

pearson X^2	6.667
p value	0.00982327

	Event rate
group A	0.58333333
group B	0.41666667
Odds Ratio	1.96
Risk Ratio	1.4

Figure 2.

Same event rate but different p-values



264 p-values are based on the 132 individual food items at 2 time points.

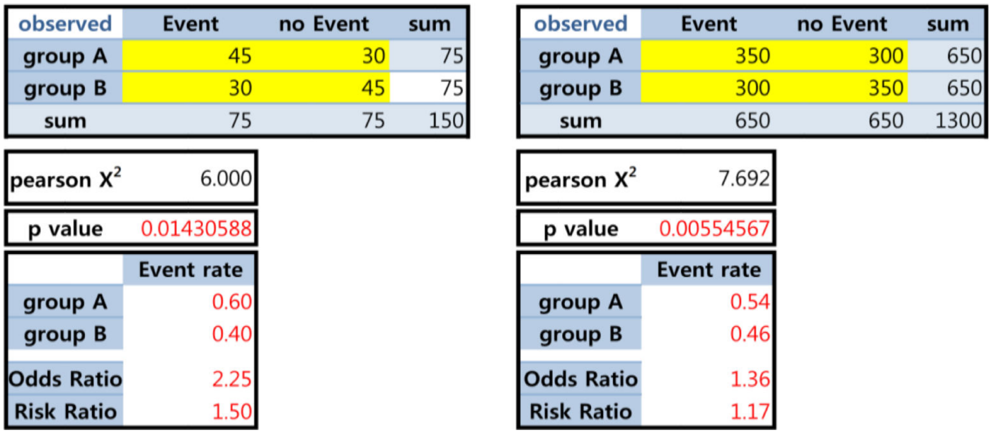
Permission to reproduce this figure was granted by the publisher, the Royal Society.

Figure 3.

P-value plots

Left: 200 p-values from t-tests for randomly generated data under H_0

Right: 264 p-values from t-tests for real nutritional data (Young et al. 2009)[5]



observed	Event	no Event	sum
group A	45	30	75
group B	30	45	75
sum	75	75	150

pearson X^2	6.000
p value	0.01430588

	Event rate
group A	0.60
group B	0.40
Odds Ratio	2.25
Risk Ratio	1.50

observed	Event	no Event	sum
group A	350	300	650
group B	300	350	650
sum	650	650	1300

pearson X^2	7.692
p value	0.00554567

	Event rate
group A	0.54
group B	0.46
Odds Ratio	1.36
Risk Ratio	1.17

Figure 4.

Smaller effect size with smaller p-value

Table 1

Unadjusted vs. adjusted p-values side by side

Numbers of Persons with Onset of Fatal or Nonfatal Cancer in the SEAS Trial and in SHARP and IMPROVE-IT.

Value	SEAS Trial				SHARP and IMPROVE-IT			
	Active Treatment (N = 944)	Control (N = 929)	Uncorrected P Value	Corrected P Value*	Active Treatment (N = 10,319)	Control (N = 10,298)	Uncorrected P Value	Corrected P Value*
Total follow-up for cancer (person-yr)	3810	3826			18,246	18,255		
Any cancer								
No.	101	65	0.006†	—	313	326	0.61	—
Percent per yr	2.7	1.7			1.7	1.8		
Site of cancer (no. of persons)‡								
Lip, mouth, pharynx, or esophagus	1	1	1.00	1.00	16	14	0.86	1.00
Stomach	5	1	0.23	1.00	6	9	0.60	1.00
Large bowel or intestine	9	8	1.00	1.00	36	39	0.81	1.00
Pancreas	3	1	0.63	1.00	5	7	0.77	1.00
Liver, gallbladder, or bile ducts	2	3	1.00	1.00	10	11	1.00	1.00
Lung	7	10	0.60	1.00	33	28	0.61	1.00
Other respiratory site	1	0	1.00	1.00	4	2	0.68	1.00
Skin	18	8	0.08	0.80	74	89	0.27	1.00
Breast	8	5	0.60	1.00	21	19	0.88	1.00
Prostate	21	13	0.24	1.00	25	36	0.20	1.00
Kidney	2	2	1.00	1.00	25	11	0.03	0.48
Bladder	7	7	1.00	1.00	18	20	0.87	1.00
Genital site	4	4	1.00	1.00	6	5	1.00	1.00
Hematologic site	7	5	0.79	1.00	19	19	1.00	1.00
Other known site	3	1	0.63	1.00	11	11	1.00	1.00
Unspecified	9	6	0.63	1.00	20	18	0.88	1.00

Peto et al. (2008). Permission to reproduce this table was obtained from NEJM/Massachusetts Medical Society.