The Tyranny of Statistical Significance Testing

Why I think we should stop quoting P values ...

May 28, 2013

Motivation and Conclusion in one page

- P values are frequently abused and more commonly misunderstood.
- As commonly practices, statistical significance testing is flawed.
- The 'framework' of statistical significance testing is uniquely suited to sow confusion amongst its practicioners (and consumers).
- P values have minimal useful information content and they don't answer the questions you want answers to.

Why Focus on Significance Testing? What about the Rest?

In this talk we seemingly only deal with comparing two populations, or rather, their means and are trying to find out whether these measured means stem from different populations.

P values can be determined for many other things:

- paired comparisons,
- two receiver-operating curves etc.

The mechanics will change but, ultimately, *the P value issue and the statistical significance testing problem remain*.

Acknowledgement and Sources

Stang, Poole, Kuss (2010)

The ongoing tyranny of statistical significance testing in biomedical research.

Eur J Epidem 25(4): 225–230.

Goodman (2008)

A dirty dozen: Twelve P-Value Misconceptions.

Sem Hemat 45: 135–140.

Gigerenzer (2004).

The Null Ritual: What You Always Wanted to Know About Significance Testing but Were Afraid to Ask.

published in: D.Kaplan (Ed.) 2004 *The Sage Handbook of quant. method. for the social sciences* (pp. 391–408)

Goodman (1999)

Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy.
Toward Evidence-Based Medical Statistics. 2: The Bayes Factor.

Ann Intern Med 130: 995–1013.

What is a P Value?

The P Value

In statistical significance testing the P value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.[4]

What is a P Value?

The P Value

In statistical significance testing the P value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.[4]

A test statistic:

- is a function of the sample
- is a numerical summary reducing data to one value that can be used to perform a hypothesis test.
- is typically selected to quantify, within observed data, behaviours that would distinguish the null from the alternative hypothesis
- shares some of the same qualities of a descriptive statistic

Quiz: What Does a Significant Result Mean?

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a simple independent means t-test and your result is significant ($t = 2.7, \text{df} = 18, p = .01$).

True (A) or False (B)?

1. You have absolutely disproved the null hypothesis (i.e., there is no difference between the population means).

Quiz: What Does a Significant Result Mean?

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a simple independent means t-test and your result is significant ($t = 2.7, \text{df} = 18, p = .01$).

True (A) or False (B)?

2. You have found the probability of the null hypothesis being true.

Quiz: What Does a Significant Result Mean?

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a simple independent means t-test and your result is significant ($t = 2.7, \text{df} = 18, p = .01$).

True (A) or False (B)?

3. You have absolutely proved your experimental hypothesis (that there is a difference between the population means).

Quiz: What Does a Significant Result Mean?

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a simple independent means t-test and your result is significant ($t = 2.7, \text{df} = 18, p = .01$).

True (A) or False (B)?

4. You can deduce the probability of the experimental hypothesis being true.

Quiz: What Does a Significant Result Mean?

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a simple independent means t-test and your result is significant ($t = 2.7, \text{df} = 18, p = .01$).

True (A) or False (B)?

5. You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.

Quiz: What Does a Significant Result Mean?

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a simple independent means t-test and your result is significant ($t = 2.7, \text{df} = 18, p = .01$).

True (A) or False (B)?

6. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

Here is how your peers did on those questions ...

Table 1
Percentages of False Answers (i.e., Statements Marked as True) in the Three Groups of Figure 1

	Germany 2000			*United Kingdom 1986*
Statement (abbreviated)	*Psychology students*	*Professors and lecturers: not teaching statistics*	*Professors and lecturers: teaching statistics*	*Professors and lecturers*
1. H_0 is absolutely disproved	34	15	10	1
2. Probability of H_0 is found	32	26	17	36
3. H_1 is absolutely proved	20	13	10	6
4. Probability of H_1 is found	59	33	33	66
5. Probability of wrong decision	68	67	73	86
6. Probability of replication	41	49	37	60

Note. For comparison, the results of Oakes' (1986) study with academic psychologists in the United Kingdom are shown in the right column.

from Gigerenzer's "The Null Ritual..." [3].

What is the effect of the long-run view of the experimenter?

You are comparing drug A to drug B. Your experiment is to try them head to head and see which is better. Here is the data you got: A, A, A, A, A, B

What is p?

Exp Paradigm One: You do six experiments, you count how many times A or B is the winner.

Exp Paradigm Two: You do no more than six experiments, each time you check whether B is better than A - if it is, you stop; if not, you keep going for a maximum of six experiments.

Conclusion?

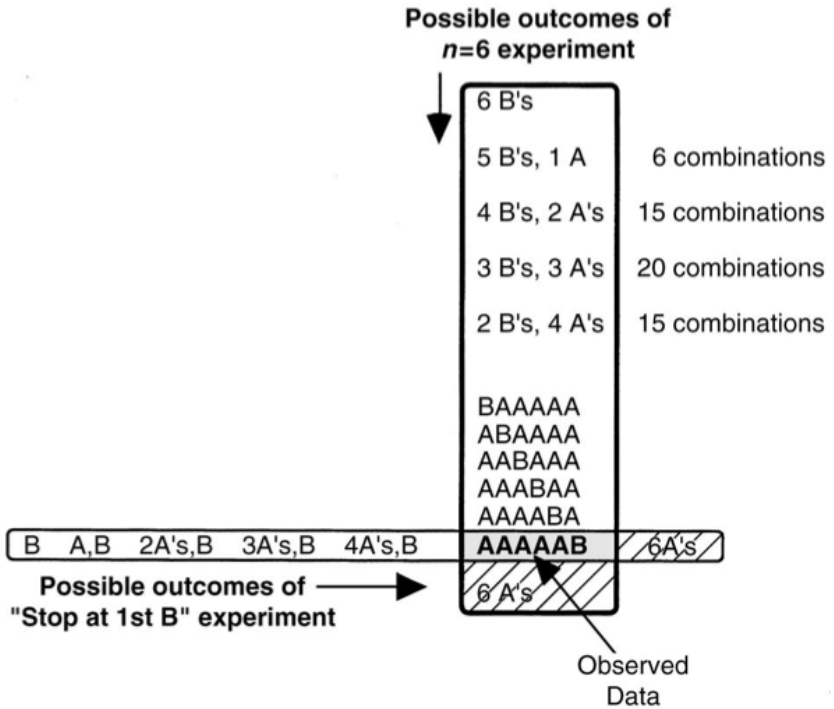

Figure 4. Possible outcomes of two hypothetical trials in six patients (Appendix). The only possible overlapping results are the observed data and the result in which treatment A was preferred in all patients.

Goodman [4]

The P Value depends on:

- Your mental model of the long run. (Different models of the experiment but **same data** leads to different P values ?!)
- The effect size (if your means are very different p will be smaller)
- The sample size (if you have measured the means to a small uncertainty, your p will be smaller)

Conclusion?

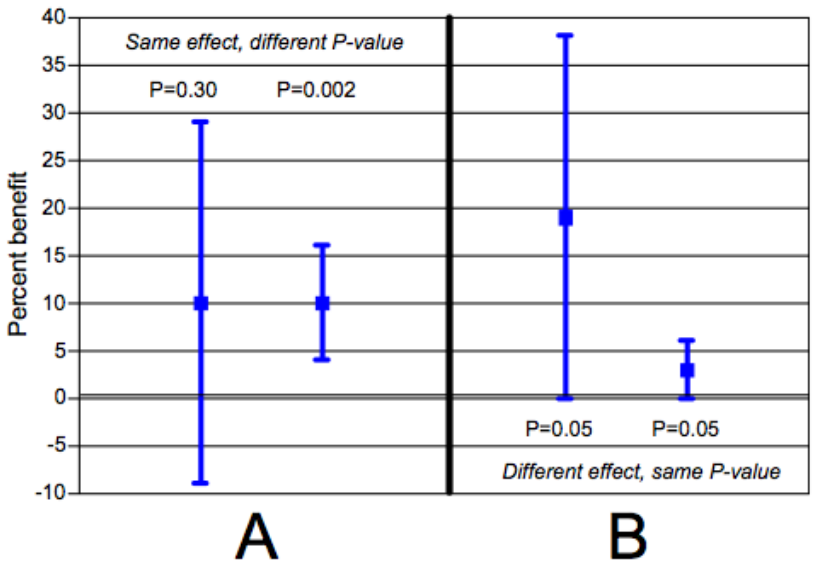

Figure 2 Figure showing how the *P* values of very different significance can arise from trials showing the identical effect with different precision (A, Misconception #4), or how same *P* value can be derived from profoundly different results (B, Misconception #5).

Goodman [4]

The P Value depends on:

- Your mental model of the long run. (Different models of the experiment but **same data** leads to different P values ?!)
- The effect size (if your means are very different p will be smaller)
- The sample size (if you have measured the means to a small uncertainty, your p will be smaller)

What is the problem to be solved?

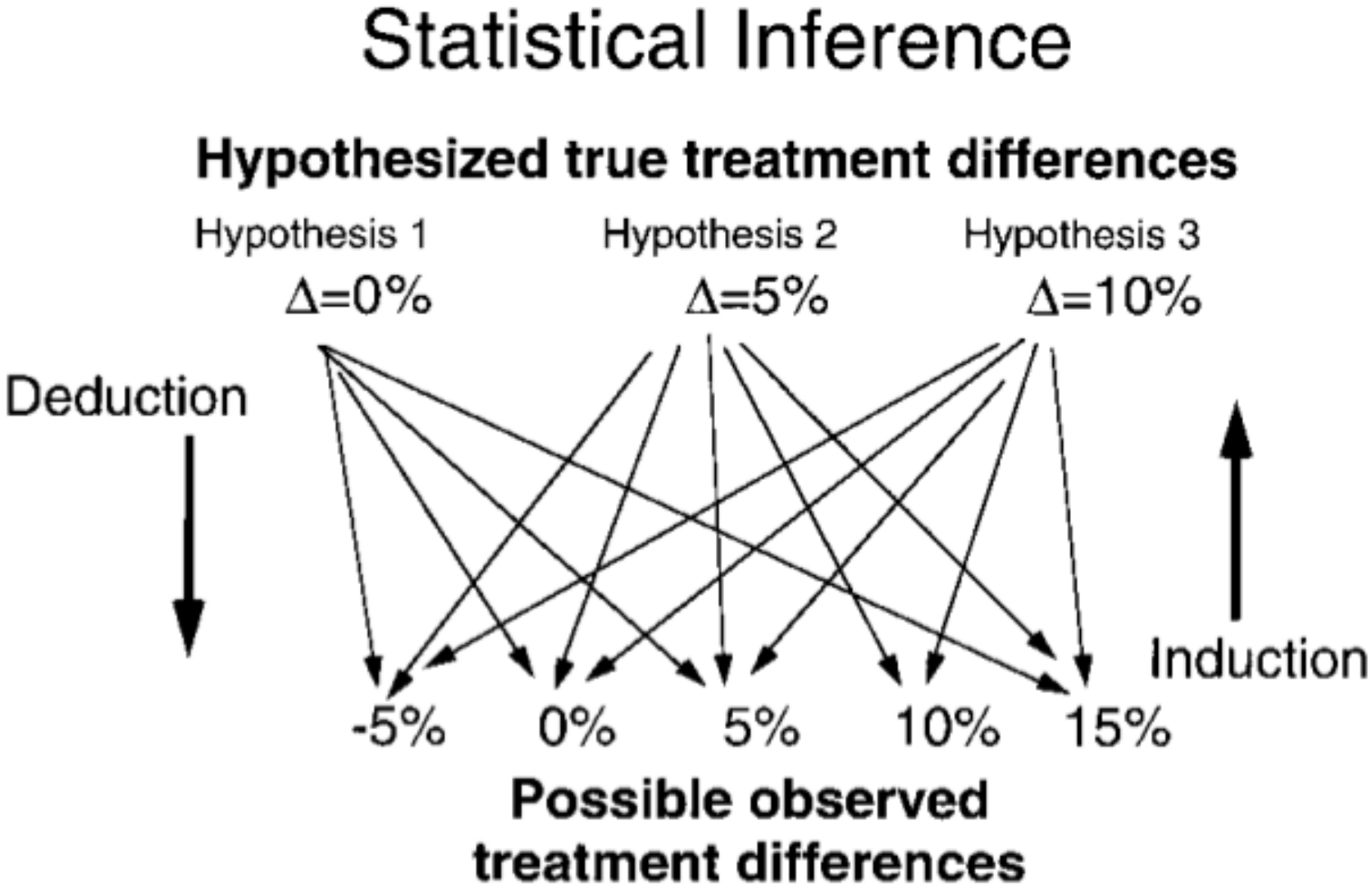

Fig 1 in Goodman (1999) "Toward Evidence-Based Medical Statistics." [4].

A bit of history : Fisher

Fisher's Null Hypothesis Testing

- Formulate Null Hypothesis, H_0 (often but not necessarily the "no-effect hypothesis")
- Observe data and report its probability under the assumption of H_0

Issues to note:

- there is no alternative hypothesis
- hence no Type 2 error ("failure to reject a false null hypothesis")
- there is no statistical power
- p describes the strength of evidence against H_0 but, a small p warrants further (repeated!) study...

A bit of history : Neyman and Pearson

N&P's Decision Theory approach

Neyman & Pearson criticized the utility of Fisher's approach.

- objective is to mostly make the right *decision* wrt various hypothesis.
- formulate H_0 and H_A
- set error rates (α for Type I error and β for Type II error), sample size ahead of performing experiment
- report acceptance/rejection of hypothesis

Issues to note:

- There is no measure for the strength of evidence ("don't quote p")
- There is no conclusion from the particular data at hand — you behave as if H_0 is true (or not) and move on.

A bit of history: The Unholy Mix

The methods by Fisher and by Neyman/Pearson

- are individually sound but don't (and can't) fulfil the desire of inferring truth from one experiment.
- are not compatible, but have been mixed (around 1940 AD) into an incoherent framework posing as a unified received method (aka "The Tyranny of Statistical Significance Testing")

What alternatives to SST are there? Feeling cheated by today's talk?

What is an honest, well meaning researcher to do?

"[...] appreciate [...] that there is no number generated by standard methods that tells us the probability that a given conclusion is right or wrong.[...] The second principle is that the size of an effect matters, and that the entire confidence interval should be considered as an experiments result, more so than the P value or even the effect estimate. " Goodman[2]

Present your work using

- confidence intervals
- show me the data
- Bayes Factor (= Likelihood Ratio)

Effect of Moderate and High Alcohol Consumption

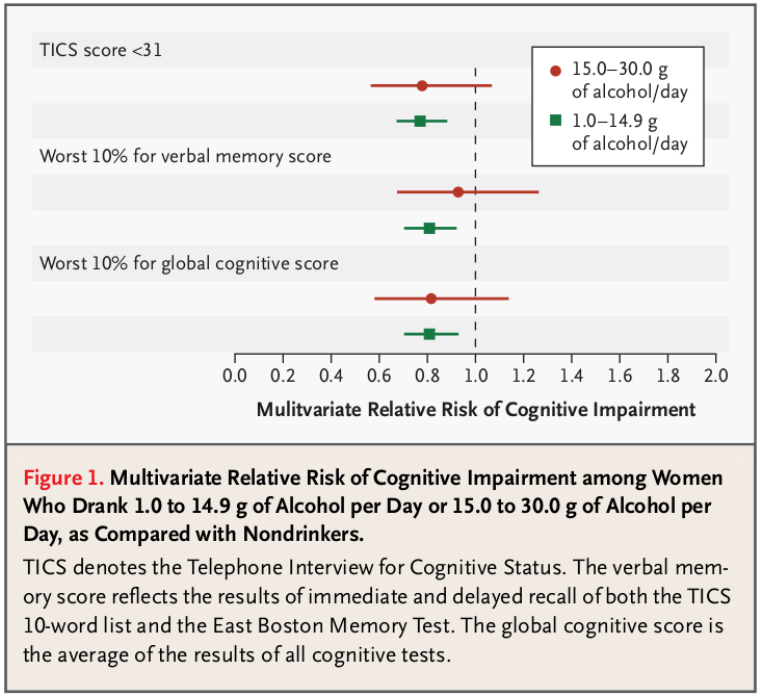

Figure 1. Multivariate Relative Risk of Cognitive Impairment among Women Who Drank 1.0 to 14.9 g of Alcohol per Day or 15.0 to 30.0 g of Alcohol per Day, as Compared with Nondrinkers.

TICS denotes the Telephone Interview for Cognitive Status. The verbal memory score reflects the results of immediate and delayed recall of both the TICS 10-word list and the East Boston Memory Test. The global cognitive score is the average of the results of all cognitive tests.

Stampfer, NEJM (2005) 352(3) 245.

- The authors claim there is 'no association' for high consumption
- confidence intervals show a different story
- The problem is the low prevalence of high consumption in the study group and, hence, a high uncertainty in the outcome measures.

Illustration of the Bayes Factor

Is this coin fair?

Observed data: H H H T

What can you say about the null hypothesis, H_0, that the coin is fair (i.e. $p_T = \frac{1}{2}$)?

What can you say about the best supported alternative, H_A, $p_T = \frac{1}{4}$?

Evaluate p under the assumption of H_0.

What does the p under this null hypothesis say about the probability of the coin being fair?

Illustration of the Bayes Factor

Bayes Theorem

$$\text{Posterior Odds}(H_0|\text{data}) = \text{Prior Odds}(H_0|\text{data}) \times \text{Bayes factor}$$

where $\quad \text{Bayes' factor} = \frac{p(data|H_0)}{p(data|H_A)} \quad$ and $\quad \text{Odds} = \frac{p}{1-p}$

Illustration of the Bayes Factor

Bayes Theorem

$$\text{Posterior Odds}(H_0|\text{data}) = \text{Prior Odds}(H_0|\text{data}) \times \text{Bayes factor}$$

where $\quad \text{Bayes' factor} = \frac{p(data|H_0)}{p(data|H_A)} \quad$ and $\quad \text{Odds} = \frac{p}{1-p}$

The issue of Bayes factor is complex in the case of composite hypothesis (e.g. H_A = non-zero difference of means...)

Illustration of the Bayes Factor

Bayes Theorem

$$\text{Posterior Odds}(H_0|\text{data}) = \text{Prior Odds}(H_0|\text{data}) \times \text{Bayes factor}$$

where $\quad \text{Bayes' factor} = \frac{p(data|H_0)}{p(data|H_A)} \quad$ and $\quad \text{Odds} = \frac{p}{1-p}$

in our case:

$p(data|H_0) = (\frac{1}{2})^3 \cdot \frac{1}{2} = \frac{1}{16}$ and $p(data|H_A) = (\frac{3}{4})^3 \cdot \frac{1}{4} = \frac{27}{256}$
resulting in a Bayes' factor of $\frac{16}{27} \approx .59$

We can state that the evidence for H_0 ($p_T = \frac{1}{2}$) is only $\approx 59\%$ as strong as the evidence that supports H_A (meaning $p_T = \frac{1}{4}$).