What Statistics 101 Doesn't Teach, But Should

Charles J. Geyer

School of Statistics
Minnesota Center for the Philosophy of Science
University of Minnesota

January 25, 2015

Impetus

The impetus for this talk was when BIG (biology interest group, Minnesota Center for the Philosophy of Science) read

Jonah Lehrer (2010).
The truth wears off: Is there something wrong with
the scientific method?
New Yorker, December 13 issue.
(not a research article).

John P. A. Ioannidis (2005).
Why most published research findings are false.
PLoS Medicine, 2, e124.
(not a research article).

The Decline Effect

J. B. Rhine, a parapsychologist, in the 1930's posited that effect sizes decrease as experiments are repeated and called this the *decline effect*.

Schooler, a psychologist, has recently been pushing this idea,

Jonathan Schooler (2011).
Unpublished results hide the decline effect.
Nature, 470, 437.
(not a research article).

and the *New Yorker* article makes much of this — that's what its title refers to.

So does the decline effect exist, and if so does it mean the "scientific method" is wrong?

My Rant

I delivered a rant in BIG saying

- The "decline effect" is mystical rubbish.
- The phenomenon is real but is a result of well-studied issues: multiple testing without correction and publication bias.
- These issues should be understood by every user of statistics.
- They should be taught in every statistics class, but aren't effectively taught. Students don't really understand them. Most scientists don't really understand them.

New Yorker Article

The *New Yorker* article actually presents essentially my argument in the middle.

> *Leigh Simmons, a biologist at the University of Western Australia . . . [said] "But the worst part was when I submitted these null results I had difficulty getting them published. The journals only wanted confirming data. It was too exciting an idea to disprove, at least back then." For Simmons, the steep rise and slow fall of fluctuating asymmetry is a clear example of a scientific paradigm, one of those intellectual fads that both guide and constrain research: after a new paradigm is proposed, the peer-review process is tilted toward positive results. But then, after a few years, the academic incentives shift — the paradigm has become entrenched — so that the most notable results now disprove the theory.*

New Yorker Article (cont.)

and

> *Michael Jennions, a biologist at Australian National University . . . similarly, argues that the decline effect is largely a product of publication bias, or the tendency of scientists and scientific journals to prefer positive data over null results*

New Yorker Article (cont.)

but then goes off the rails

> *While publication bias almost certainly plays a role in the decline effect, it remains an incomplete explanation. For one thing, it fails to account for the initial prevalence of positive results that never even get submitted to journals. [How do we know there is such a prevalence?] It also fails to explain the experience of people like Schooler, who have been unable to replicate their initial data despite their best efforts. [Wrong again. If the effect doesn't exist, then of course it can't be replicated.]*

New Yorker Article (cont.)

eventually getting mystical

> *Although such reforms [they will be mentioned later] would mitigate the dangers of publication bias and selective reporting, they still wouldn't erase the decline effect. This is largely because scientific research will always be overshadowed by a force that can't be curbed, only contained: sheer randomness. Although little research has been done on the experimental dangers of chance and happenstance, the research that exists isn't encouraging.*

Hypothesis Tests

A statistical hypothesis test compares two statistical models, one simple, the other more complex and containing the simple model as a special case.

The P-value is the probability, assuming the simple model is correct, of seeing data at least as favorable to the more complex model as are the observed data.

If the P-value is large, then the evidence in favor of the complex model is weak, and we say it fits the data no better than the simple model (any apparent better fit is not "statistically significant").

If the P-value is small, then the evidence in favor of the complex model is strong, and we say it fits the data better than the simple model (the apparent better fit is "statistically significant").

Hypothesis Tests (cont.)

What is "large" and "small"? Probabilities are between 0 and 1, so small is near 0 and large is near 1.

The traditional dividing line is 0.05.

$P < 0.05$ is "small" $=$ "statistically significant"

$P > 0.05$ is "large" $=$ "not statistically significant"

Hypothesis Tests (cont.)

The traditional 0.05 dividing line has nothing to recommend it other than that it is a round number.

It is a round number because humans have five fingers.

Imagine that! Crucial issues of scientific inference are decided by counting on our fingers.

I once wrote

> *Anyone who thinks there is an important difference between* $P = 0.049$ *and* $P = 0.051$ *understands neither science nor statistics.*

But my coauthor cut it — not because it was wrong but because it might offend.

Hypothesis Tests (cont.)

0.05 is a weak criterion

> *Recall Bernard Woolfe's definition of statistics as that branch of mathematics which enables a man to do twenty experiments a year and publish one false result in Nature.*

recounted by John Maynard Smith in an essay "Molecules are not Enough" collected in *Did Darwin Get it Right?*

A 95% confidence interval misses 5% of the time.

$P < 0.05$ announces "statistical significance" incorrectly (i. e., when the effect actually does not exist) 5% of the time.

Dogma

The central dogma of hypothesis testing.

- Do only one test.
- Choose what test will be done before the data are collected.
- Do it, and report it.

No one does this except the clinical trials people. The statistical analysis of a clinical trial is prescribed in the trial protocol, which describes what data will be collected and how it will be analyzed. Any paper will report this analysis.

Other analyses may also be done and spun wildly in the press conference, but the valid analysis is published for the scientifically and statistically astute to read.

Dogma (cont.)

The valid interpretation of $P < 0.05$, even if the dogma has been followed is

> *The complex model fits the data better than the simple model (the apparent better fit is "statistically significant").*

or, more tersely,

> *The null hypothesis (simple model) is false.*

But the following is not valid

> *The mechanistic, causal, scientific explanation I have for the simple model being incorrect has been proved by statistics.*

Publication Bias

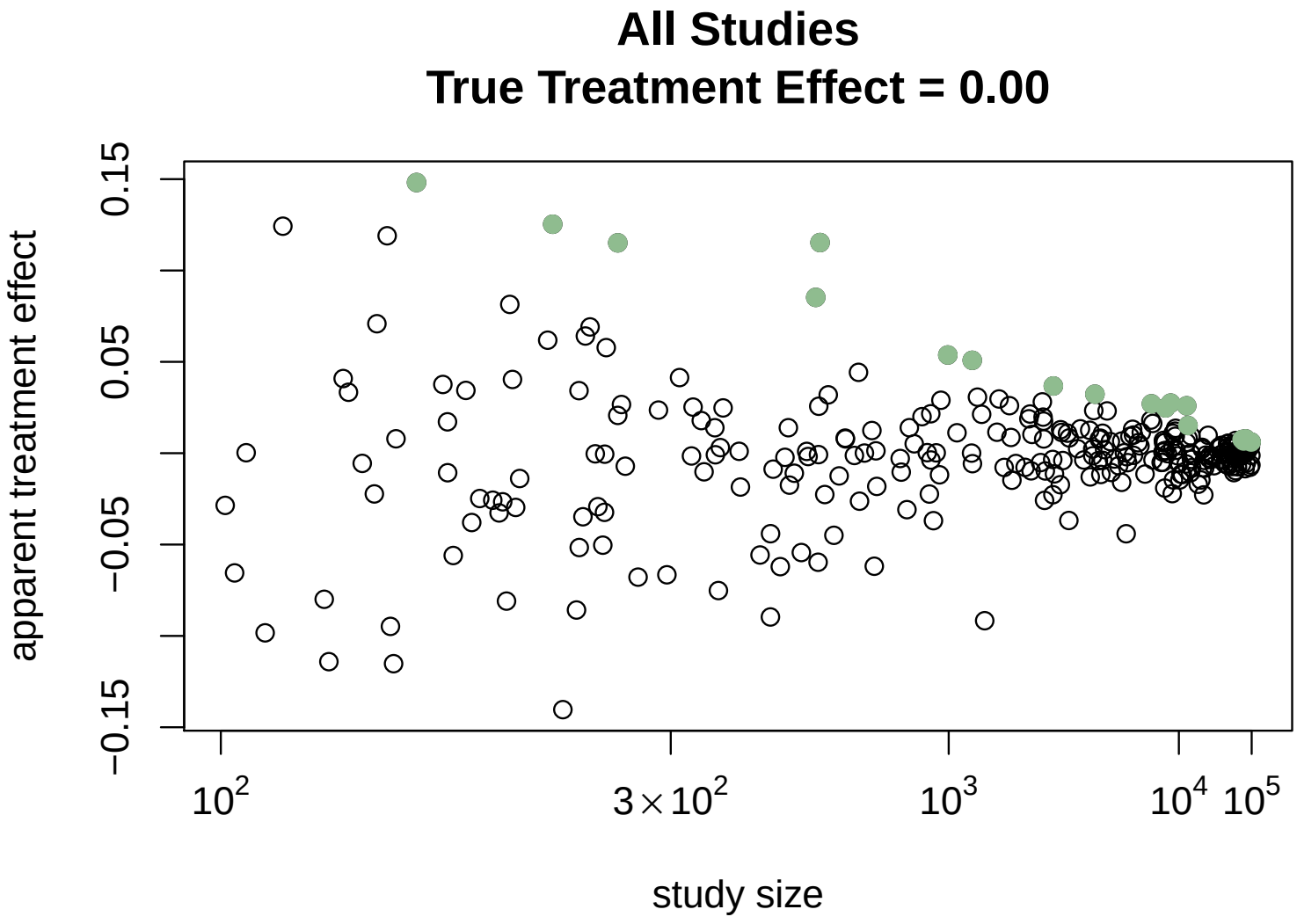
A. k. a. the *file drawer problem*.

What if no paper is published unless $P < 0.05$?

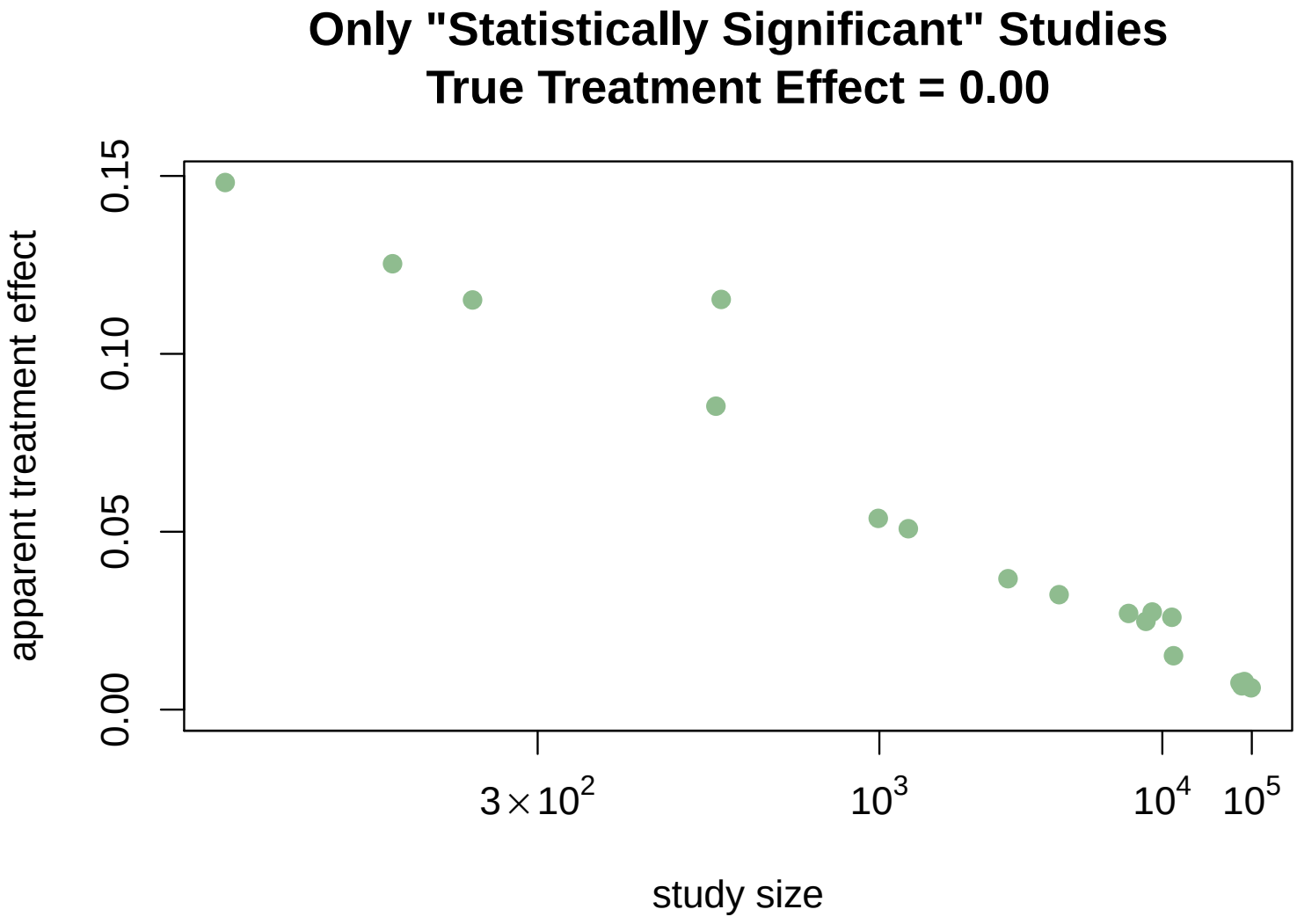
In small studies only large treatment effects are "statistically significant" and reported.

In large studies both large and small treatment effects are "statistically significant" and reported.

Funnel Plot


All Studies
True Treatment Effect = 0.00
apparent treatment effect
0.15
0.05
−0.05
−0.15
10^2
3×10^2
10^3
10^4
10^5
study size

Funnel Plot (cont.)



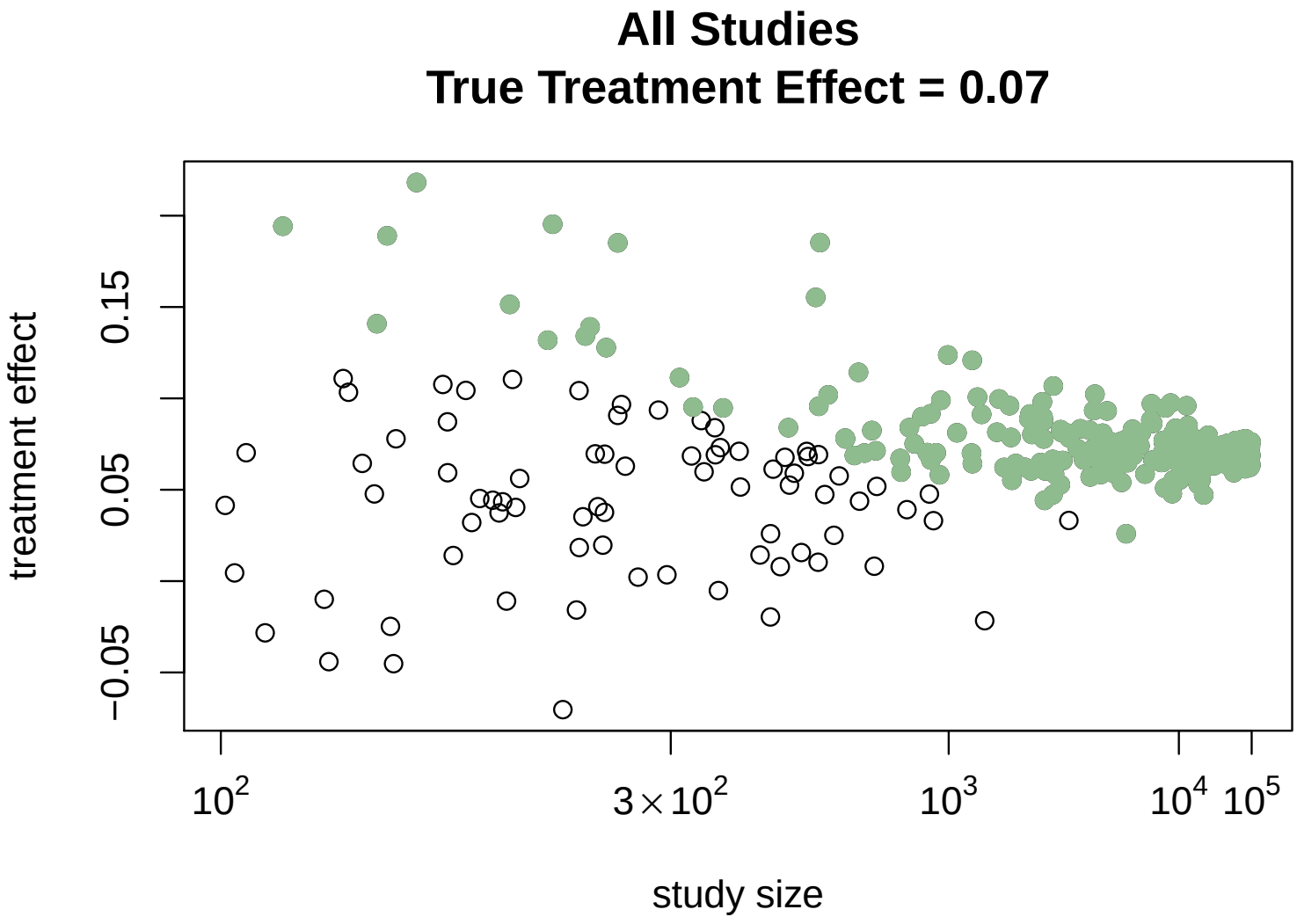
Publication Bias (cont.)

IMHO publication bias accounts for the "decline effect."

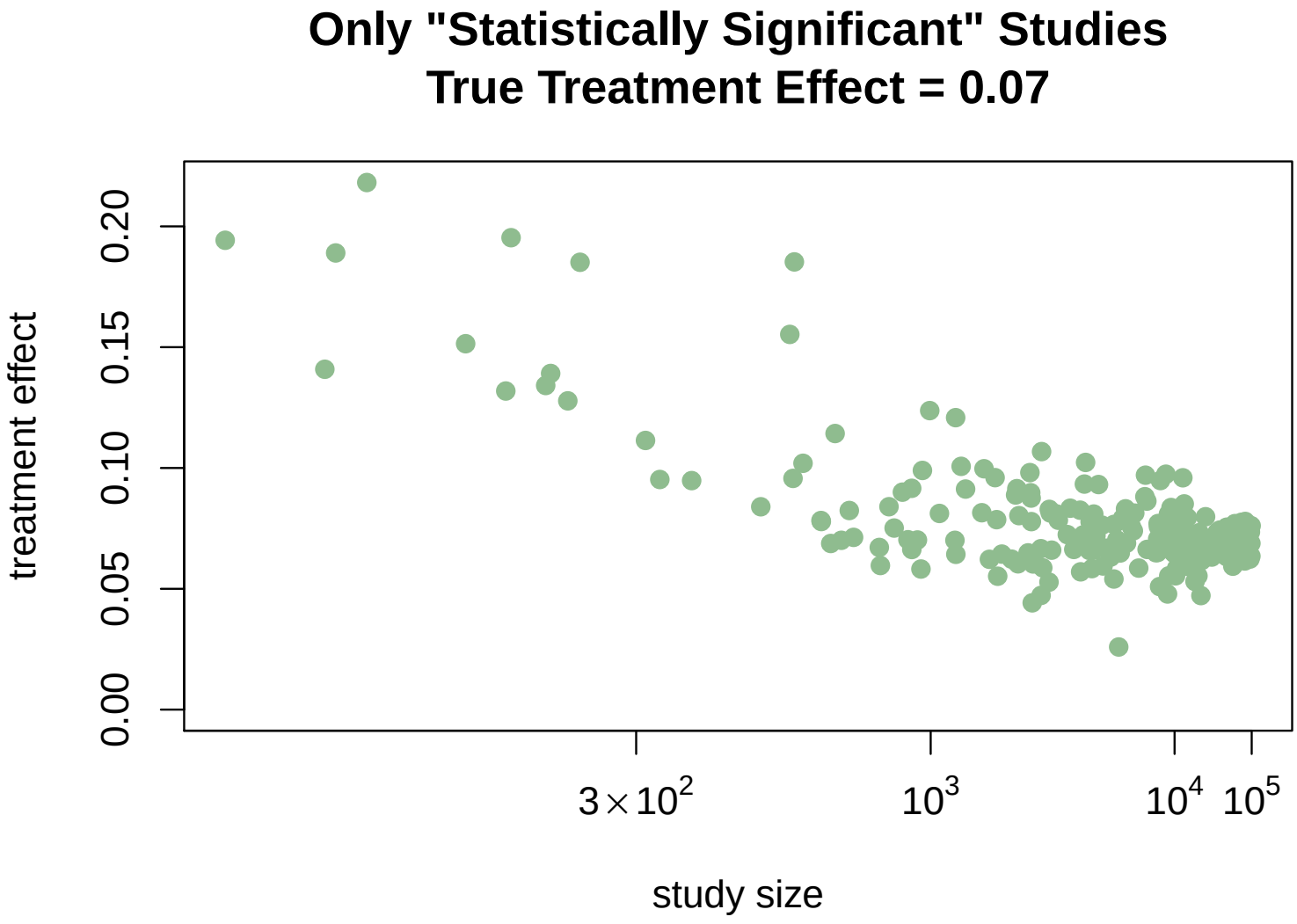
If larger, more expensive studies are done later, they will have smaller effect sizes if the effect doesn't really exist.

Exacerbating the publication bias effect is another: as time goes by and the effect gets "established" in the literature, it gets easier to publish $P > 0.05$ to debunk the conventional wisdom.

Funnel Plot (cont.)



Funnel Plot (cont.)



Registries

The solution? Publish every study.

In order to do a study one must before it is started register it in a permanent on-line database, saying what data will be collected and what statistical analysis will be done. Any ensuing paper must do that analysis and report it.

Other analyses may be done and spun any way the authors want, but the valid analysis is published for the scientifically and statistically astute to read.

If no paper is ever published, at least we can assume $P > 0.05$ for that study.

Registries (cont.)

Ioannidis (2005), Schooler (2011), and other authorities recommend such registries.

Some kinds of cancer trials currently have such registries.

I'm not holding my breath until other areas implement them.

The $P < 0.05$ Criterion

Another solution would be for editors and referees to stop using $P < 0.05$ as a criterion for publication.

If only evidence for an idea can be published – all evidence against an idea is suppressed – what kind of science is that?

But that is what the $P < 0.05$ standard does.

Of course, that is not the intended effect, but it is the effect!

Multiple Comparison

When the "do only one test" dogma is violated, it can be fixed up by considering multiple tests as a combined "omnibus" test.

The simplest such correction multiplies each P-value by the number of tests done (Bonferroni correction). If you do 20 tests, then you need $P < 0.0025$ on any one test to declare "statistical significance".

Other multiple comparison methodology exists, but is more complicated and only applies to special situations.

False Discovery Rate

Since the Bonferroni correction is so stringent, many people don't like it.

Recently, false discovery rate (FDR) correction has been recommended. This is (slightly) less stringent.

The first "discovery" (smallest P-value) uses the same criterion as Bonferroni correction, but successive "discoveries" (second smallest, third smallest, etc.) use (slightly) less stringent cutoffs.

The guarantee (under certain assumptions) is that only some proportion (usually 0.05) of the "discoveries" will be incorrect.

Like Bonferroni, FDR is much more stringent than multiple testing without correction.

Stargazing

Louis Guttman (1985).
The illogic of statistical inference for cumulative science.
Applied Stochastic Models and Data Analysis, 1, 3–10

> *Very few researchers are devoted to testing a single hypothesis or estimating a single parameter. The mathematicians seem to have forgotten that we are in the age of the computer, which spews forth dozens and hundreds of statistics from a single study. The mishmash of stars and double stars in textbooks and journal articles throughout the social sciences, as well as in other sciences, show something lacking in the teaching of statistical inference. Almost all the presumed 'probabilities' published are wrong, yet the teachers of declarative inference remain silent, and refuse to revise their curriculum.*

Electric Power Lines and Cancer

Nancy Wertheimer and Ed Leeper (1979).
Electrical wiring configurations and childhood cancer.
American Journal of Epidemiology, 109, 273–284.

Found an association between (some forms of) childhood cancer and living close to electric power lines.

A case-control study, multiple testing was done without correction, and electric and magnetic field levels were estimated rather than measured.

A typical paper in a respected refereed journal. Multiple later studies seemed to confirm their findings.

Electric Power Lines and Cancer (cont.)

Paul Brodeur (1989-1990).
Annals of Radiation. The Hazards of electromagnetic fields: I, Power lines, II, Something is happening, III, Video-display terminals. Calamity on Meadow Street.
New Yorker, June 12, 19, and 26, 1989 and July 9, 1990.

Paul Brodeur (1989).
Currents of Death.
New York: Simon and Schuster.

Paul Brodeur (1993).
The Great Power-Line Cover-Up: How the Utilities and the Government Are Trying to Hide the Cancer Hazard Posed by Electromagnetic Fields.
New York: Little-Brown.

Electric Power Lines and Cancer (cont.)

Committee on the Possible Effects of Electromagnetic Fields on Biologic Systems, National Research Council (1997).
Possible Health Effects of Exposure to Residential Electric and Magnetic Fields
Washington: National Academies Press.
(379 pages)

Electric Power Lines and Cancer (cont.)

The NRC report found that the link between electric and magnetic fields and cancer or other biologic effects had not been established. It highlighted three issues.

- There was no plausible physical mechanism.
- There was no reproducible evidence from studies in animals, bacteria, and tissue cultures.
- Most of the epidemiological studies did not directly measure magnetic field strength in the home and the few that did had null results.

But it also mentions that some scientists argue that in the epidemiological studies "proper adjustment has not been made for multiple comparisons".

Electric Power Lines and Cancer (cont.)

Martha S. Linet, Elizabeth E. Hatch, Ruth A. Kleinerman, Leslie L. Robison, William T. Kaune, Dana R. Friedman, Richard, K. Severson, Carol M. Haines, Charleen T. Hartsock, Shelley Niwa, Sholom Wacholder, and Robert E. Tarone (1997).
Residential exposure to magnetic fields and acute lymphoblastic leukemia in children.
The New England Journal of Medicine, 337, 1–7.
(research article)

Edward W. Campion (1997).
Power lines, cancer, and fear.
The New England Journal of Medicine, 337, 44–46.
(editorial)

Electric Power Lines and Cancer (cont.)

The NEJM research article did one test for association with one cancer (acute lymphoblastic leukemia, ALL). Magnetic field strength was measured by "blinded" workers (who did not know whether the resident of the house was a case or control). No association was found. Not even close to statistical significance.

> *The odds ratio for ALL was 1.24 (95 percent confidence interval, 0.86 to 1.79) at exposures of 0.200 μT or greater as compared with less than 0.065 μT.*

The NEJM editorial repeats the points made by the NRC report, including that the epidemiological studies did "huge numbers of comparisons with selective emphasis on those that were positive."

Electric Power Lines and Cancer (cont.)

Another large, well designed, well executed study also showed no effect.

UK Childhood Cancer Study Investigators (1999).
Exposure to power-frequency magnetic fields and the risk of childhood cancer.
Lancet, 354, 1925–1931.

Electric Power Lines and Cancer (cont.)

All the published studies that did multiple testing without correction found the link between electric power lines and cancer (not always the same form of cancer).

All of the studies that obeyed the "only one test" dogma had negative results.