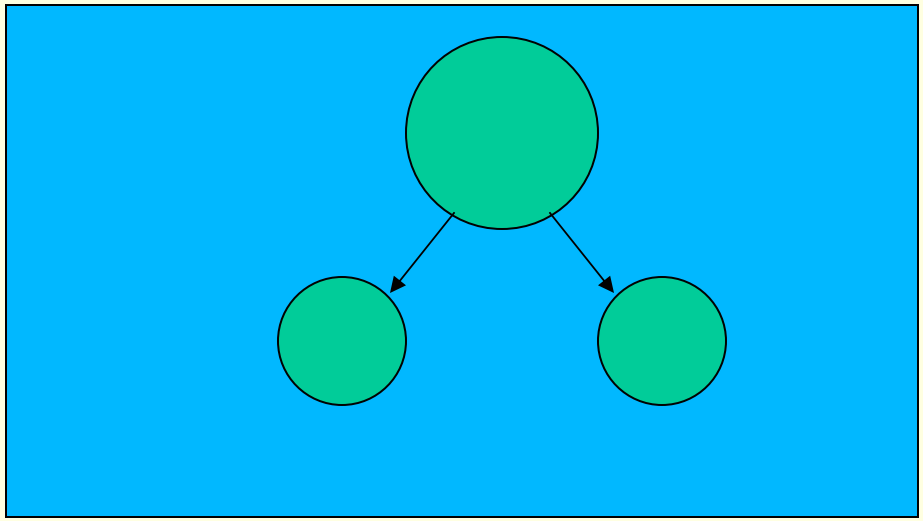
Abstract

Claims coming from human medical observational studies, when tested rigorously, most often fail to replicate. Whereas randomized clinical trials replicate over 80% of the time, medical observational studies replicate only 10 to 20% of the time. Multiple re-test studies reported JAMA failed to replicate. For example in the early 1990s, Vitamin E was reported to protect against heart attacks. Large, well-conducted randomized clinical trials did not replicate this claim. The claim that Type A Personality leads to heart attacks failed to replicate in two separate studies, yet the myth still lives. Clearly, there are systematic problems with how observational studies are conducted and analyzed that need to be identified and fixed. Edwards Deming, the most famous quality expert ever, says that any problem with a failed process is not the fault of the workers, scientists conducting observational studies, but of management. Funding agencies and journal editors need to fix a clearly broken process. Technical problems are identified. Tough management solution are proposed. A simple statistical analysis strategy is presented. Many human health problems can only be examined using observational data. Our proposals, technical and managerial, should lead to more reliable claims along with fair ways to judge their reliability.

Pre-lecture
Simple statistics

S. Stanley Young
National Institute of Statistical Sciences
Young@niss.org, 919 685 9328

P-value, t-test

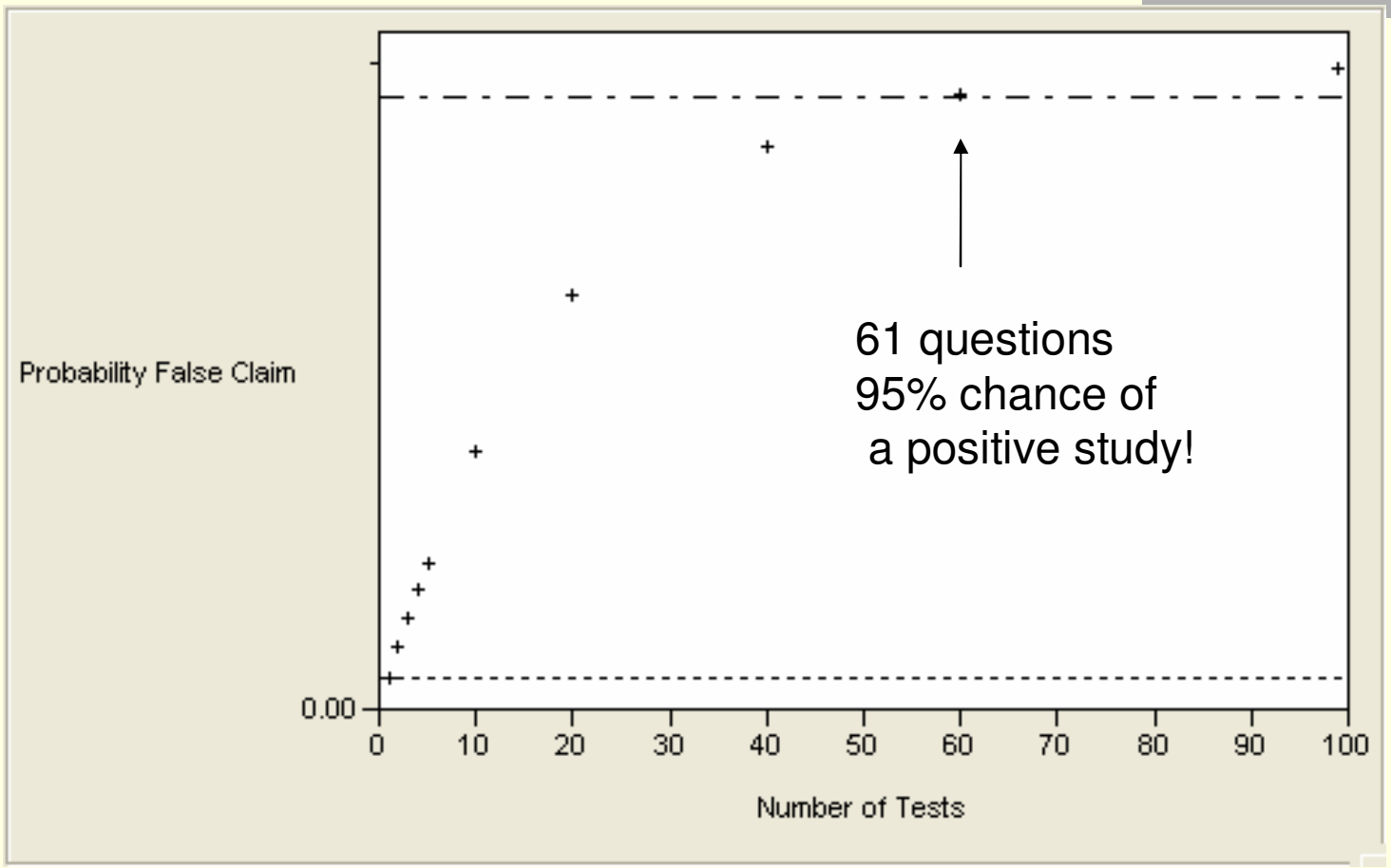


Population,
real or theoretical

Two samples,
random

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1 X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

How do you get a "p < 0.05"?
Answer: Ask lots of questions.



Let's run an epidemiology study!

p-value

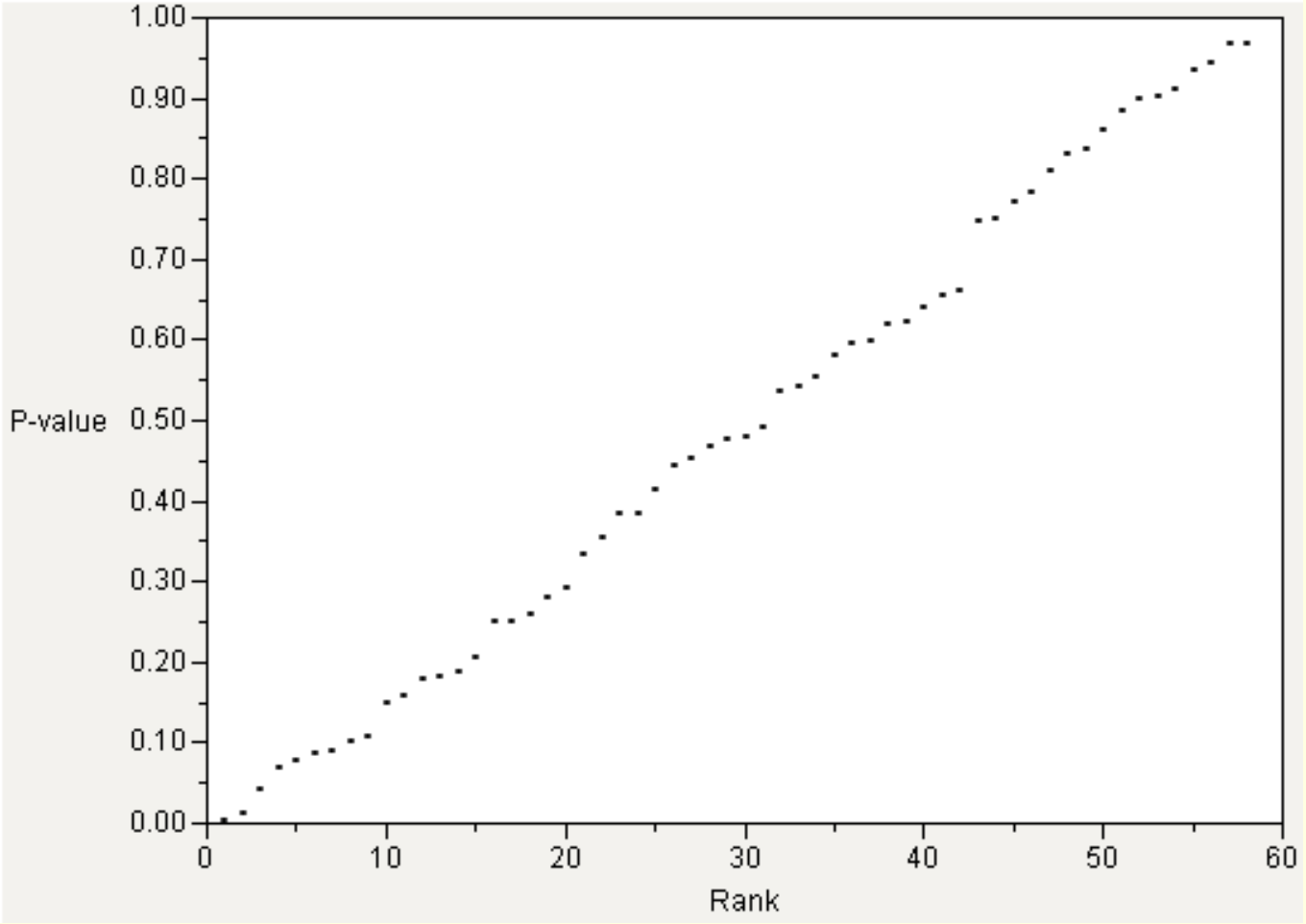
p-value = 0.046

10-sided dice simulation: Coffee causes X.

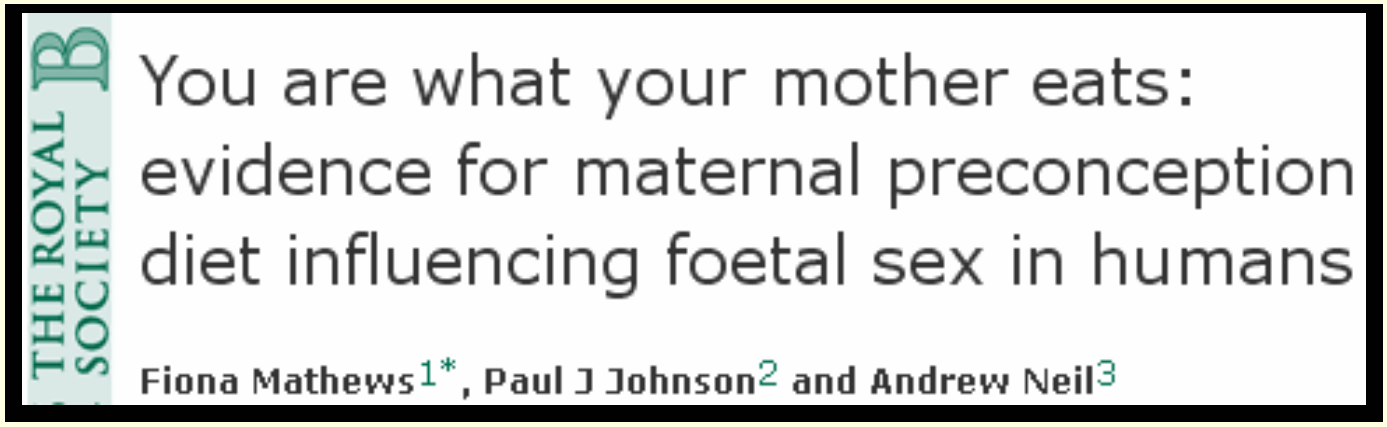
Work Sheet Stan Young, Simulation

MedCondition	YoungFemale	YoungMale	OldFemale	OldMale
1. Angina	.384	.660	.836	.067
2. Arthritis	.180	.251	.088	.451
3. Asthma	.205	.830	.258	.086
4. Cancer	.443	.641	.903	.491
5. C. Bronchitis	.810	.968	.076	.782
6. CHD	.599	.884	.280	.149
7. Emphysema	.100	.861	.107	.999
8. Heart Attack	.747	.543	.622	.158
9. Liver Disease	.183	.334	.596	.466
10. Stroke	.479	.013	.004	.999
11. Thyroid D.	.851	.935	.415	.042
12. Diabetes	.554	.654	.354	.772
13. H. LDL	.537	.383	.475	.900
14. L. HDL	.188	.618	.967	.293
15. C React Protein	.943	.910	.251	.750

P-value plot – 60 p-values.



Cereal determines human gender
Really?


THE ROYAL SOCIETY B

You are what your mother eats: evidence for maternal preconception diet influencing foetal sex in humans

Fiona Mathews[1*], Paul J Johnson[2] and Andrew Neil[3]


PROCEEDINGS OF THE ROYAL SOCIETY B

FirstCite e-publishing

Proc. R. Soc. B
doi:10.1098/rspb.2008.1405
Published online

Comment

Cereal-induced gender selection? Most likely a multiple testing false positive

P-values for 262 statistical tests

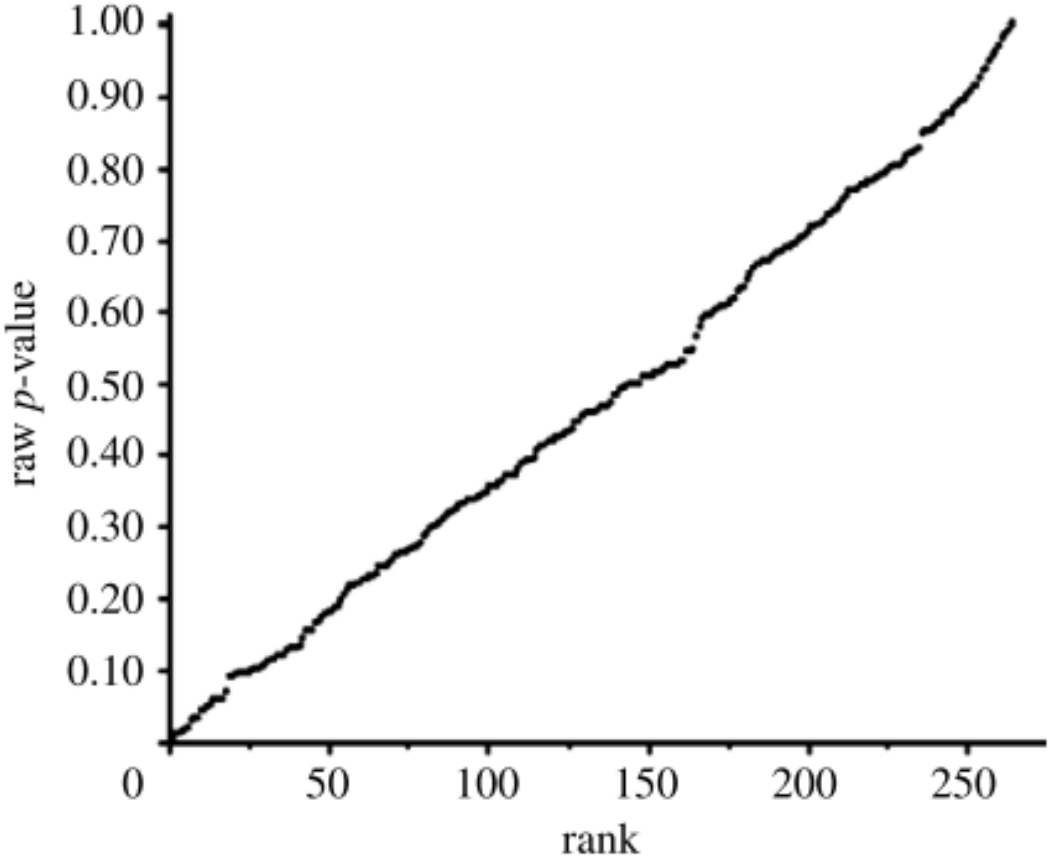


Figure 1. The *p*-value plot of 262 *p*-values.

Multiple testing, foods, multiple modeling, adjusting with covariates

Association Between More Frequent Chocolate Consumption and Lower Body Mass Index

Beatrice A. Golomb, MD, PhD
Sabrina Koperski, BS
Halbert L. White, PhD

Arch Intern Med 172 (NO. 6), Mar 26, 2012

Current multiple testing example

Multivitamins in the Prevention of Cancer in Men

The Physicians' Health Study II Randomized Controlled Trial

15 Questions (2x2x2x2 Factorial, 2^4-1=15)

21 Outcomes (mortality, multiple cancers)

315 Claims at issue (15x21 = 315)

Deming and statistical strategies to make observational studies more reliable

S. Stanley Young
National Institute of Statistical Sciences
Young@niss.org, 919 685 9328

Science point of view

What is the meaning of life?

What is real?

→ What is reproducible?

Fooled by randomness?

The Players

1. The workers – **scientists**, epidemiologists
2. The communicators –
   a. PR people
   b. *Bloggers*
   c. Reporters
   d. Science writers
3. The **consumers** – public, regulatory agencies, trial lawyers
4. The management – **funding agencies, journal editors**

The Worker is not the Problem.

W. Edwards Deming,

the most visionary innovator ever on quality control, said

The worker is not the problem.
The problem is at the top! Management!

To Deming, blaming the workers—individual researchers—
is as incorrect as it is useless.

Bringing the system under control is the responsibility of those managing it.

Crisis in epidemiology? 1988

Scientific Standards in Epidemiologic Studies of the Menace of Daily Life

ALVAN R. FEINSTEIN

Science, 1988.

No Adjustments Are Needed for Multiple Comparisons

Kenneth J. Rothman

Editor, **Epidemiology**

Now: Ioannidis, JAMA, 2005

Contradicted and Initially Stronger Effects in Highly Cited Clinical Research

"Five of 6 highly-cited nonrandomized studies had been contradicted or had found stronger effects vs 9 of 39 randomized controlled trials."

		Failure to replicate
Observational	5/6	83.3%
RCT	9/39	23.1%

Crisis in science? 2011, 2012

Deming, data and observational studies

A process out of control and needing fixing

Significance, 2011

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Nature, 2012

Observational Studies

Significance, 2011

Pos	Neg	N	Treatment(s)	Reference
0	0	2	St. John's Wort	JAMA 2002;287:1807-1814
0	3	4	HRT	JAMA 2003;289:2651-2662; 2663-2672; 2673-2684
0	0	3	Vit E	JAMA 2005;293:1338-1347
0	0	3	Low Fat	JAMA. 2006;295:655-666
0	0	2	Low Fat	JAMA 2007;298:289-298
0	0	2	Ginkgo	JAMA 2008;300:2253–2262
0	0	12	Vit C, Vit E	JAMA 2008;300:2123-2133
0	0	3	Vit E, Selenium	JAMA 2009;301:39-51
0	0	12	Ginko2*	JAMA 2009;302:2663-2670
0	**3**	**43**		

Problems with observational studies "Everything is dangerous"

1. Data staging
2. No written analysis protocol
3. Multiple testing
4. Multiple modeling
5. Uncorrected bias
6. Self-serving paper writing
7. Self-serving press release
8. Actually believe the claims

Proof : Every study is positive

1. Data Staging

2. Bias

2. Multiple testing

3. Multiple model searching

Any or all will lead to essentially all observational studies being positive!

First, data staging

Stan:

Why do you think data staging is a big issue?

Because it can be done in myriad ways, is rarely documented, and is usually not reproducible?

David Madigan

Second, Bias

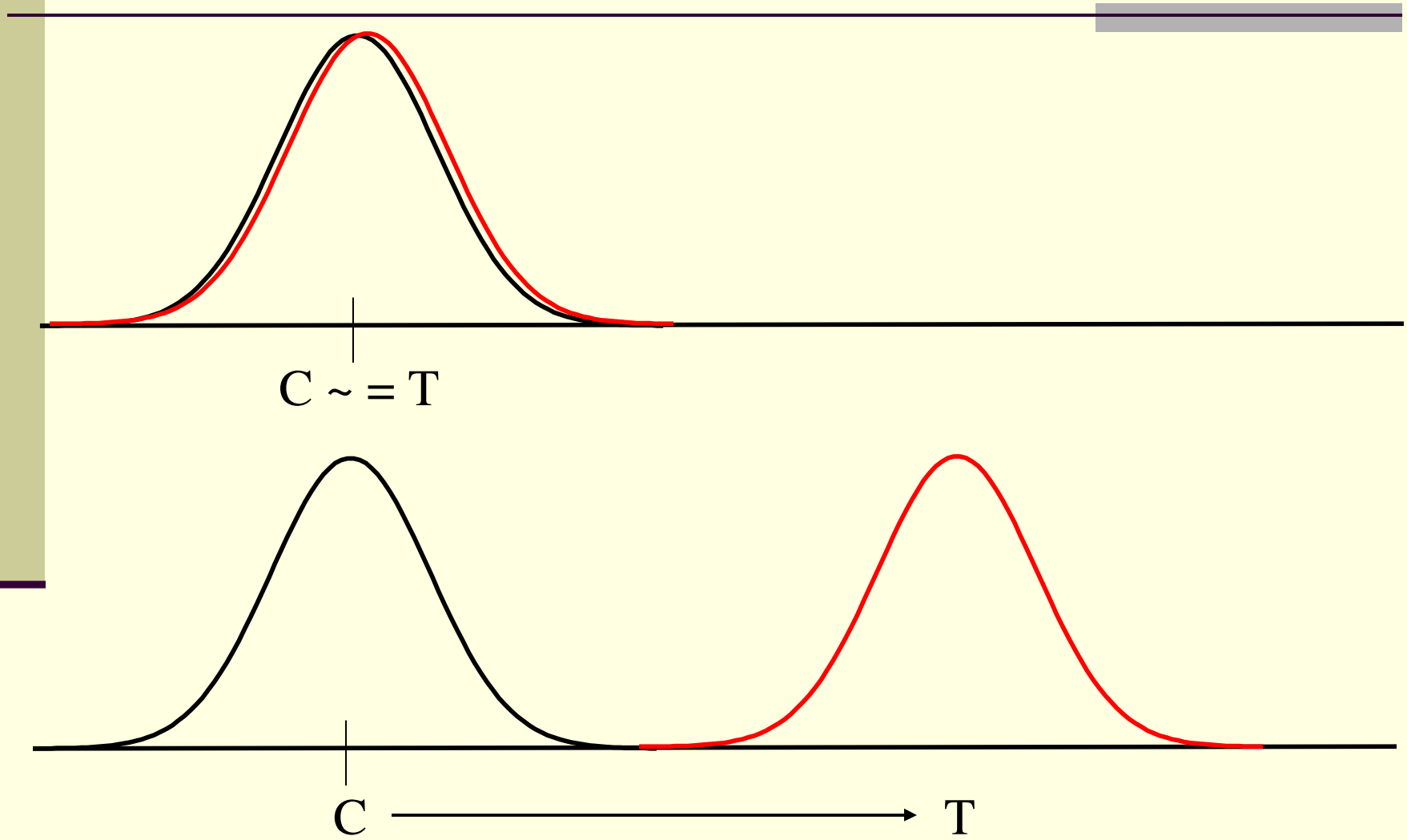
$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \ldots + \beta_p X_{pt} + \varepsilon$$

$$Y_c = \beta_0 + \beta_1 X_{1c} + \beta_2 X_{2c} + \beta_3 X_{3c} + \beta_4 X_{4c} + \ldots + \beta_p X_{pc} + \varepsilon$$

$$\Delta_{t\text{-}c} = \left(\bar{Y}_t - \bar{Y}_c\right) = \beta_1\left(\bar{X}_{1t} - \bar{X}_{1c}\right) + \beta_2\left(\bar{X}_{2t} - \bar{X}_{2c}\right) + \ldots + \beta_p\left(\bar{X}_{pt} - \bar{X}_{pc}\right) + \left(\bar{\varepsilon}_t - \bar{\varepsilon}_c\right)$$

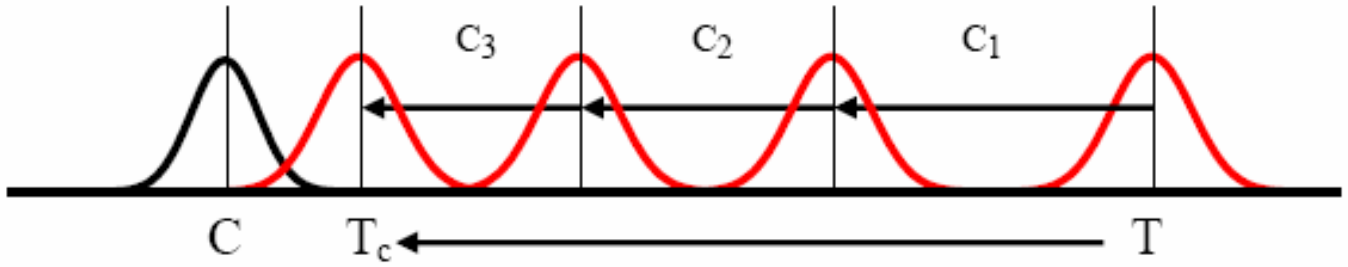
$$\Delta_{t\text{-}c} - [\text{known confounders}] = \beta_1 + [\text{unknown confounders}]$$

No bias: Randomized Clinical Trial

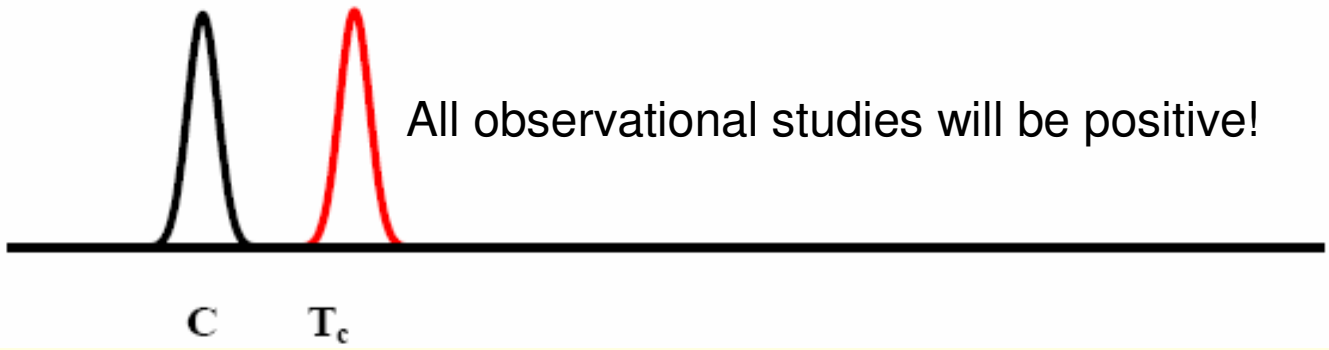


Residual bias: observational studies

(a) Use confounding variables to reduce bias.



(b) As n get large the standard error of the mean gets small.



All observational studies will be positive!

Bias

Observational studies are likely to have residual bias.

As the sample size gets large, residual bias will likely lead to “statistical significance”.

Bias is not expected to go to Zero as sample size increases.

Third: multiple testing

Multiple testing is covered in pre-lecture.

Asking hundreds of questions and not adjusting the analysis can be viewed as deceiving the consumer of the paper.

Where are the editors and referees?

Fourth: model uncertainty

Model uncertainty and health effect studies for particulate matter

Merlise Clyde*,†

ENVIRONMETRICS
Environmetrics 2000; **11**: 745–763

"Because of the large number of potential variables, model selection is often used to find a parsimonious model. Different model selection strategies may lead to very different models and conclusions for the same set of data. As variable selection may involve numerous test of hypotheses, the resulting significance levels may be called into question, and there is a concern that the positive associations are the result of multiple testing."

Algebra, again

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \ldots + \beta_p X_{pt} + \varepsilon$$

$$Y_c = \beta_0 + \beta_1 X_{1c} + \beta_2 X_{2c} + \beta_3 X_{3c} + \beta_4 X_{4c} + \ldots + \beta_p X_{pc} + \varepsilon$$

$$\Delta_{t\text{-}c} = \left(\bar{Y}_t - \bar{Y}_c\right) = \beta_1\left(\bar{X}_{1t} - \bar{X}_{1c}\right) + \beta_2\left(\bar{X}_{2t} - \bar{X}_{2c}\right) + \ldots + \beta_p\left(\bar{X}_{pt} - \bar{X}_{pc}\right) + \left(\bar{\varepsilon}_t - \bar{\varepsilon}_c\right)$$

$$\Delta_{t\text{-}c} - [\text{known confounders}] = \beta_1 + [\text{unknown confounders}]$$

A multiple testing/modeling train wreck

Association of Urinary Bisphenol A Concentration With Medical Disorders and Laboratory Abnormalities in Adults

1. 275 chemicals
2. 32 medical outcomes
3. 10 demographic covariates

275 x 32 = 8800 x 2^{10} = **~9 million**

A CDC "systems" train wreck in progress!

*Maverick Solitaire

Maverick Solitaire.

Given a normal 52-card deck of playing cards, shuffle, and then deal 25 cards.

Set aside the rest of the deck.

Attempt to arrange the 25 cards into five hands of five cards each,
such that each hand is “pat”, a flush, a straight, a full house,
or four of a kind.

In simulations the win rate was 98% on first 100 deals.

If a scientist gets to stage the data, do multiple tries at analysis,
he can almost always get statistical significance.

End of proof

Combination of data staging,
residual bias,
multiple testing
multiple analysis
means that

You are a winner – every study is positive!

If you are a consumer,
observational studies are not dependable.

Leaving no trace

Usually these attempts through which the experimenter passed, don't leave any traces; the public will only know the result that has been found worth pointing out; and as a consequence, someone unfamiliar with the attempts which have led to this result completely lacks a clear rule for deciding whether the result can or can not be attributed to chance.

Shaffer, 2007

One irate study evaluator, 2012

Criminals in the Citadel and Deceit all Along the Watchtower: Irresponsibility, Fraud, and Complicity in the Search for Scientific Truth

*Prathap Tharyan**

Mens Sana Monograph, 2012

Suggestions for effective management of observational studies

No funding / publication without:

1. Public posting protocol before study initiation.

2. Public posting of data set on publication.

3. Clear statement of questions under consideration.

4. Conform to “Reproducible Research” guidelines.

5. Any claims must be independently replicated.

Aggressive validation strategy, under control of funding agency.

0. Data are made publicly available on publication

1. Data staging and analysis are separate

2. Split sample: A, modeling; and B, holdout (testing)
3. Analysis plan is written, based only on A X's
4. Written protocol publicly posted
5. Analysis of A only data set
6. Journal accepts paper based on A only

7. Analysis of B data set gives => Addendum

Well-conducted study, Young

1. Statistical protocol is posted before data is examined.

2. The number of questions at issue are clearly stated in the paper.

3. There is adjustment for multiple testing.

4. There is adjustment for multiple modeling.

5. The data set and analysis code are e-available.

What to do? Ioannidis

Improving Validation Practices in "Omics" Research
John P. A. Ioannidis and Muin J. Khoury
Science **334**, 1230 (2011);

Analytic validity	Do different labs, techniques, and platforms measure the same thing?
Repeatability	Can other scientists access the data and protocols, repeat the analyses, and get the same results?
Replication	Do many different data sets and their combination (meta-analysis) get consistent results?
External validation	Do different data sets by different teams, preferably prospectively and with large-scale evidence, get consistent results?
Clinical validity	Does the discovered information predict clinical outcomes?
Clinical utility	Does the use of the discovered information improve clinical outcomes?

Can other scientists get the data…

Volume 329:1753-1759 December 9, 1993 Number 24 Next ▸

An Association between Air Pollution and Mortality in Six U.S. Cities

Douglas W. Dockery, C. Arden Pope, Xiping Xu, John D. Spengler, James H. Ware, Martha E. Fay, Benjamin G. Ferris, and Frank E. Speizer

1. Key environmental pollution paper.
2. Analysis changed from city to city.
3. Essentially the data is private.
4. Similar studies have been refuted.

What can journal editors do?

Quality by inspection, p-value < 0.05, is not working.
(The workers are gaming the system.)

Management needs to re-design the system to build quality into the product.

Papers following good manufacturing procedures and addressing important questions, should be accepted without regard to statistical significance.

Require data used in publication be posted on publication.

Congressional Management: True Science Transparency Act

Any federal agency proposing rule-making or legislation shall specifically name each document used to support the proposed rule-making or legislation and provide all data used in said document for viewing by the public.

Agency Management: Federal Study Transparency Act

If federal funds are provided for a study, all data relating to the reporting of results of said study must be provided for scrutiny by the public at the time of publication.

Data is deposited on publication.

What can you, the consumer, do?

1. Be skeptical of observational study claims.
2. Read the actual paper.
3. Count the claims under consideration.
4. Ask for the data set.
5. Letter to editor : voodoo stats and trust me science. (Educate editors.)
6. Write to funding agency.
7. Write to congressman.

Researcher Incentives and Empirical Methods

by

Edward L. Glaeser

The solution to this problem is not to expect a mass renunciation of data mining, selective data cleaning or opportunistic methodology selection, but rather ...in designing and using techniques that anticipate the behavior of optimizing researchers.

Put indelicately: We need methods to thwart data staging and analysis manipulation.

Bottom line

1. Trust no claims from observational studies.

2. If multiple testing is an issue, write editor.

3. If data not public, write funding agency/congressman.

Contact Information

Stan Young

National Institute of Statistical Sciences

www.niss.org

young@niss.org

919 685 9328

Post processing

From WITCHES, FLOODS, AND WONDER DRUGS: HISTORICAL PERSPECTIVES ON RISK MANAGEMENT, by William C. Clark. “For several centuries spanning the Renaissance and Reformation, societal risk assessment meant witch hunting. Contemporary accounts record wheat inexplicably rotting in the fields, sheep dying of unknown causes, vineyards smitten with unseasonable frost, human disease and impotence on the rise. In other words, a litany of life's sorrows not very different from those which concern us today.

The institutionalized expertise of that earlier time resided with the Church. Then, as now, the experts were called upon to provide explanation of the unknown and to mitigate its undesirable consequences. Rather than seek particular sources of particular evils, rather than acknowledge their own limitations and ignorance, these experts assigned the generic name of "witchcraft" to the phenomenology of the unknown. Having a name, they proceeded to found a new professional interest dedicated to its investigation and control.

As the true magnitude of the witch problem became more apparent, the Church enlisted the Inquisition, an applied institution specifically designed to address pressing social concerns. The Inquisition became the growth industry of the day, offering exciting work, rapid advancement, and wide recognition to its professional and technical workers. Its creative and energetic efforts to create a witch-free world unearthed dangers in the most unlikely places; the rates of witch identification, assessment and evaluation soared. By the dawn of the En1ightenment, witches had been virtually eliminated from Europe and North America. Crop failures, disease, and general misfortune had not. And more than half a million people had been burned at the stake, largely ‘for crimes they committed in someone else's dreams’". (People are deluded in groups and come to their senses as individuals.)