SUNSET SALVO: REPRISE

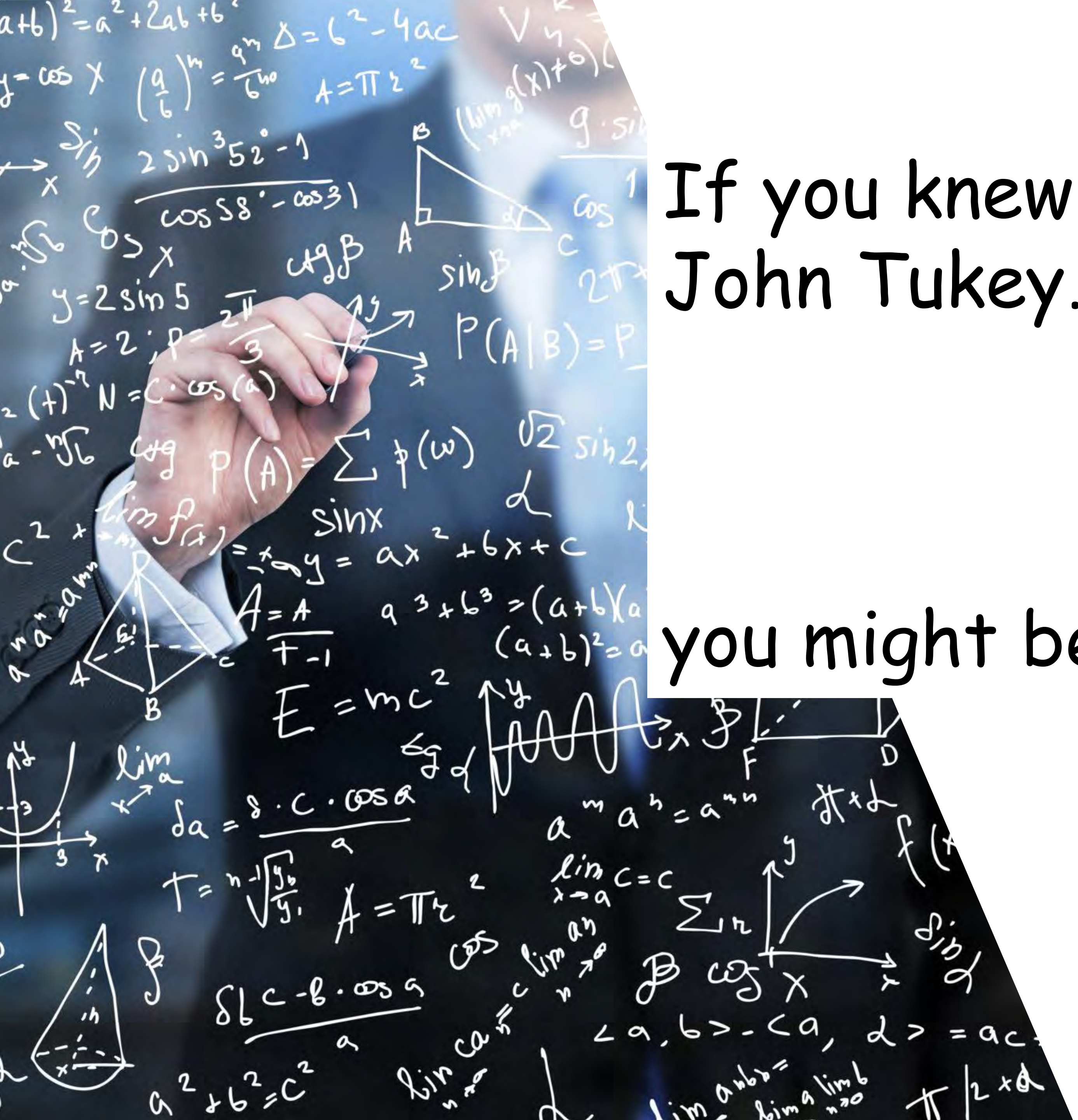
PAUL BIEMER

ppb consulting
Distinguished Fellow, Emeritus
RTI International
ppb.emeritus@rti.org

title borrowed from John Tukey’s paper

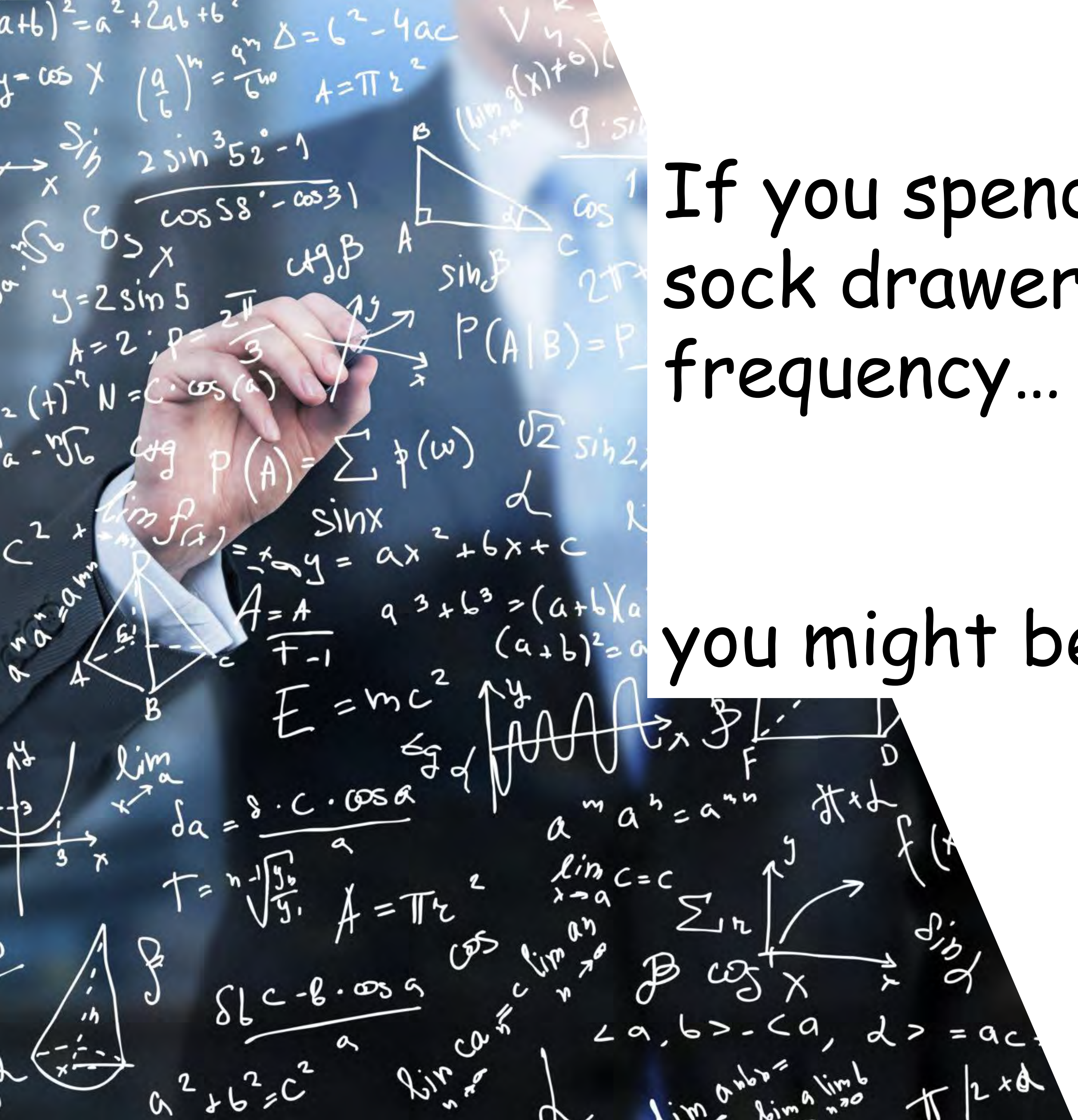
“SUNSET SALVO”
THE AMERICAN STATISTICIAN
FEBRUARY 1986

who also borrowed it from Estill Green’s (Bell Labs Exec VP) 1960 retirement speech.



If you knew I stole my title from John Tukey…

you might be a statistician.



If you spend hours arranging your sock drawer by color and frequency…

you might be a statistician.

If seeing outliers in the population affects your confidence level,...

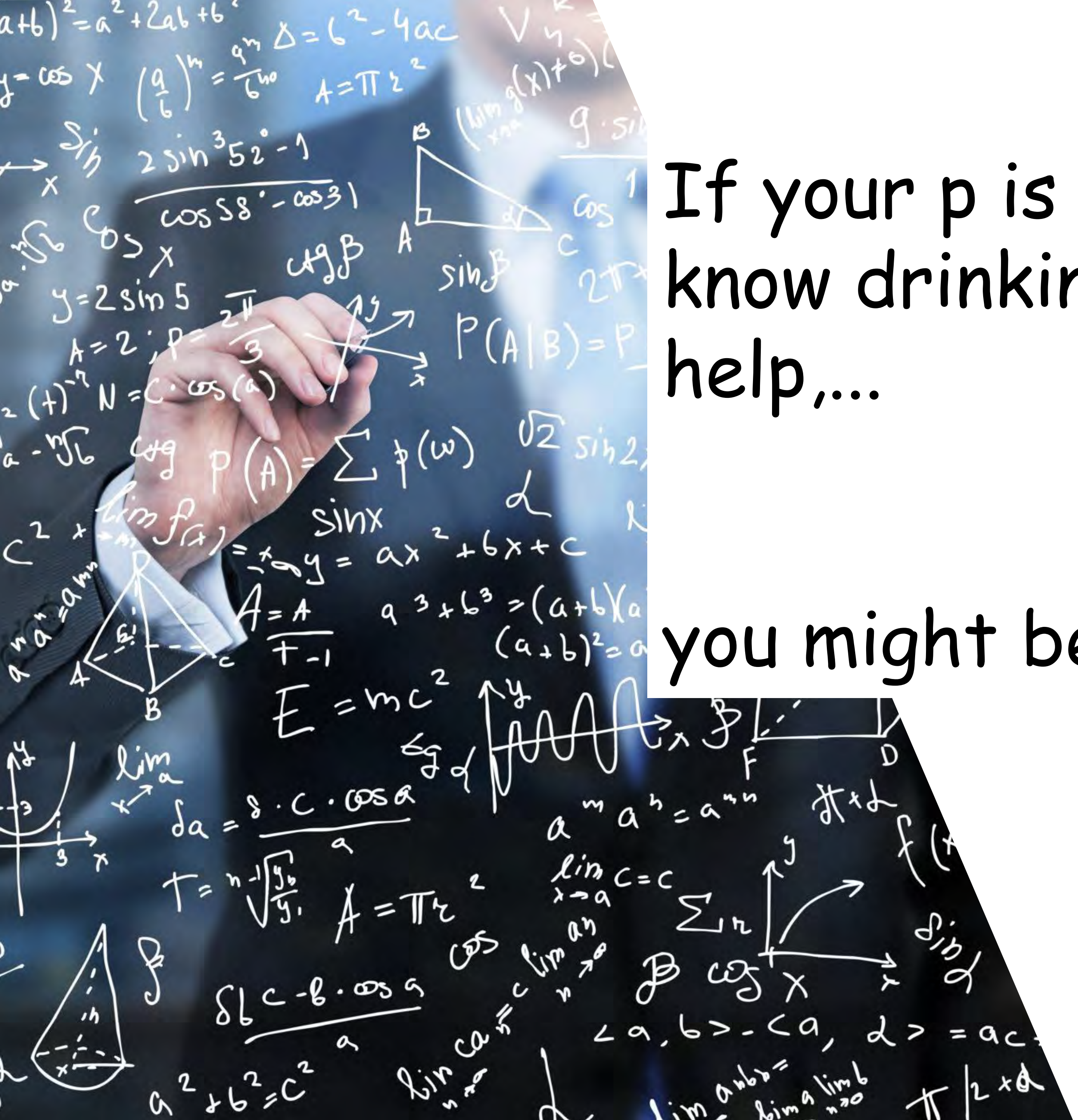
you might be a statistician.

If asked “Can you be discrete,” you answer:
“Very often”, “Often,” “Sometimes”, “Rarely,” or “Never,”...

you might be a statistician.

If you aren't surprised that someone could drown in a river with an average depth of only 3 feet,...

you might be a statistician.

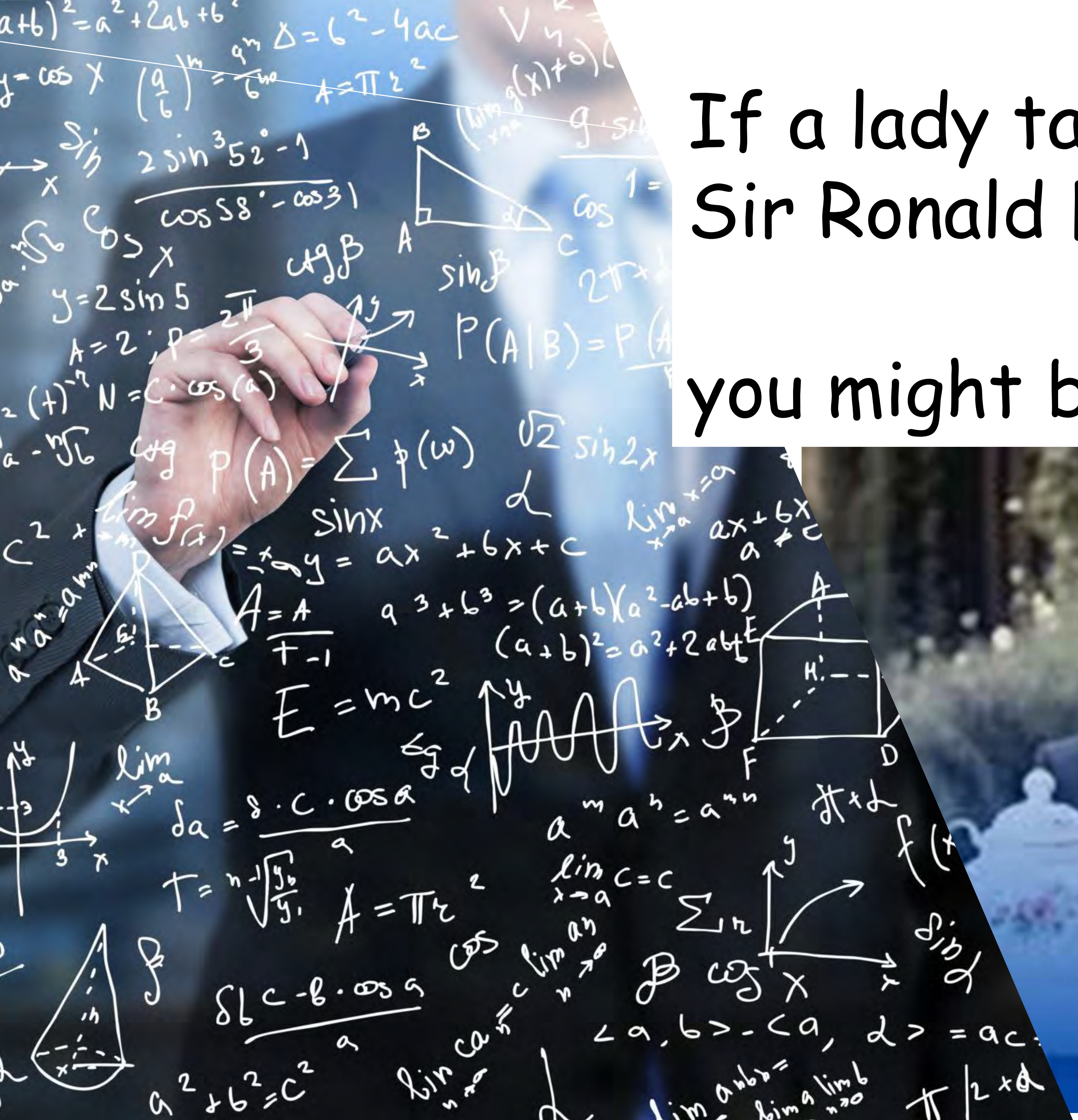


If your p is insignificant, and you know drinking more fluids won't help,...

you might be a statistician.

If a lady tasting tea reminds you of Sir Ronald Fisher,…

you might be a statistician



ITSEW'S FAR-FLUNG LOCATIONS

2024	**Washington, DC – "Understanding Error…in a Blended World"**
2021	**Virtual** because of COVID with a theme of "Survey Quality"
2019	**University of Bergamo, Italy,** with a theme on "Integration of surveys and alternative data sources"
2018	**Duke University**, exploring "Approaches for Mitigating Total Survey Error (TSE) and Its Effects"
2017	**Nuremberg, Germany**, discussing "Total Survey Error: Combined data products from a TSE perspective"
2016	**Sydney, Australia**, pondering whether "Total Survey Error Will Save Survey Science?"
2015	**Baltimore, Maryland**, focused on "Improving Data Quality in the Era of Big Data" (resulting in the book "Total Survey Error in Practice")1.
2014	**Washington, DC**, delving into "Total Survey Error: Fundamentals and Frontiers"
2013	**Iowa State University** as the 7th International Total Survey Error Workshop.
2012	**Sanpoort, Netherlands**, reflecting on "Total Survey Error: Past, Present, and Future".
2011	**Québec, Canada**, known as the "International Total Survey Error Workshop 2011".
2010	**Stowe, Vermont**, explored "The Ongoing Evolution of Survey Methodology and the Impact on Total Survey Error".
2009	**Tällberg, Sweden**, focused on "The Total Survey Error Concept: Uses and Abuses".
2008	**Research Triangle Park**, North Carolina, discussed "Multiple Sources of Error and Their Interaction".
2005	**Washington, DC**, centered around "Latent Variable Models in the Social Sciences".

My all-time favorite ITSEW!

WHAT'S SO SPECIAL ABOUT THE YEAR 2005?

First International Total Survey Error Workshop

Thursday, March 17, 2005 - 8:15am to Friday, March 18, 2005 - 3:15pm

Sponsored by the National Institute of Statistical Sciences
in conjunction with the SAMSI program on Latent Variable Models in the Social Sciences

Organizers

Paul Biemer
919-541-6056
ppb@rti.org

Jerry Reiter
919-668-5227
jerry@stat.duke.edu

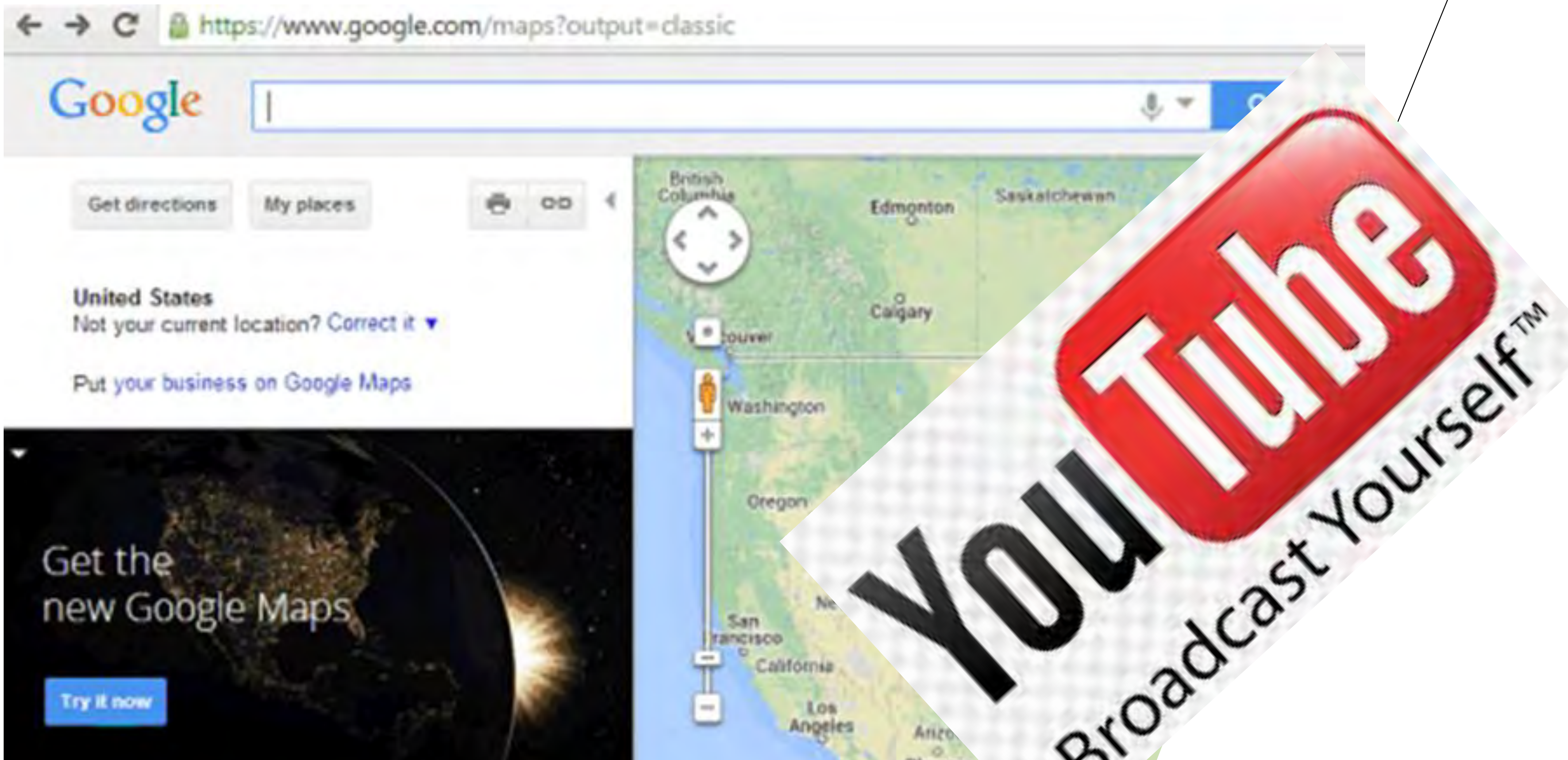
The goal of the workshop is to bring together researchers from federal agencies, academia, and survey organizations to discuss methods for measuring nonsampling errors. The workshop will include invited presentations by distinguished researchers in survey methodology, particularly those specializing in the

LOCATION
Bureau of Labor Statistics, Washington DC
United States

POLICY
Reimbursement Form
Refund Policy

Iraq’s first ever parliamentary election

Launch of Google Maps and YouTube



THERE WAS BAD NEWS TOO…

London bombings – 4 coordinated attacks killing 52 and injuring 700

....AND WORSE NEWS...

Hurricane Katrina hits the US killing 1,392

WHAT ELSE IS SPECIAL ABOUT THE YEAR 2005?

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key

…tors that influence this problem and …e corollaries thereof.

…deling the Framework for False …sitive Findings

…eral methodologists have …nted out [9–11] that the high … of nonreplication (lack of …firmation) of research discoveries … consequence of the convenient, ill-founded strategy of claiming …clusive research findings solely on … basis of a single study assessed by …mal statistical significance, typically a *p*-value less than 0.05. Research …ot most appropriately represented … summarized by *p*-values, but, …ortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let *R* be the ratio of the number of "true relationships" to "no relationships"

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1-\beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that *c* relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2 × 2 table, one gets $PPV = (1-\beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124

Abbreviation: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

"Why Most Published Research Findings Are False"

Published by John Ioannidis in *PLOS Medicine*

Most referenced technical paper from 2005 –

Cited 13,258 times and counting!

WHY WAS THIS PAPER SO POPULAR?

- **Title is very provocative**
  - Essentially says that you can't believe what you read in scientific journals!
- **Conclusions apply to all fields of science and all types of data**
- **Simple concepts and mathematics used to support conclusions.**
  - Paper is accessible to just about anyone with an elementary background in statistics

PAPER IS VERY RELEVANT FOR ITSEW BECAUSE…

1. **Focuses on error (researcher bias) and how it affects data analysis and reporting.**
2. **Does not consider other sources of nonsampling error. Doing so could make the results even more applicable and impactful.**
3. **ITSEW community has previously made important contributions to this topic.**
4. **Remainder of my presentation summarizes key results and suggests some extensions.**

BASIC NOTATION AND CONCEPTS

c = number of hypotheses to be tested

α = Pr(Type I error)

β = Pr(Type II error)

d_s = specified effect size

T = proportion of tests truly having effect size $\geq d_s$

PPV = positive predictive value

FDR = false discovery rate (1-PPV)

Study designed to detect minimum effect size, d_s with power $(1-\beta)(100\%)$ and significance level $\alpha(100)\%$. c hypotheses to be tested. Assume for some proportion, T, of these tests, the true effect size $d \geq d_s$.

BASIC NOTATION AND CONCEPTS

c = number of hypotheses to be tested

α = Pr(Type I error)

β = Pr(Type II error)

d_s = specified effect size

T = proportion of tests truly having effect size $\geq d_s$

PPV = positive predictive value

FDR = false discovery rate (1-PPV)

Study designed to detect minimum effect size, d_s with power $(1-\beta)(100\%)$ and significance level $\alpha(100)\%$. c hypotheses to be tested. Assume for some proportion, T, of these tests, the true effect size $d \geq d_s$.

Research Finding	True Relationship		
	Yes	No	Total
Yes	$c(1-\beta)T$	$c\alpha(1-T)$	$c[(1-\beta)T+\alpha(1-T)]$
No	$c\beta T$	$c(1-\alpha)(1-T)$	$c[\beta T+(1-\alpha)(1-T)]$
Total	cT	$c(1-T)$	c

c = number of hypotheses to be tested

α = Pr(Type I error)

β = Pr(Type II error)

d_s = specified effect size

T = proportion of tests truly having effect size $\geq d_s$

PPV = positive predictive value

FDR = false discovery rate (1-PPV)

POSITIVE PREDICTIVE VALUE
the proportion of significant findings that are real (also 1-FDR)

Research Finding	True Relationship		
	Yes	No	Total
Yes	$c(1-\beta)T$		$c[(1-\beta)T+\alpha(1-T)]$
No	$c\beta T$	$c(1-\alpha)(1-T)$	$c[\beta T+(1-\alpha)(1-T)]$
Total	cT	$c(1-T)$	c

divided by

$$\text{PPV} = \frac{c(1-\beta)T}{c\left[(1-\beta)T+\alpha(1-T)\right]} = \frac{(1-\beta)}{(1-\beta)+\alpha\frac{1-T}{T}}$$

c = number of hypotheses to be tested

α = Pr(Type I error)

β = Pr(Type II error)

d_s = specified effect size

T = proportion of tests truly having effect size $\geq d_s$

PPV = positive predictive value

FDR = false discovery rate (1-PPV)

ILLUSTRATION (THE PERFECT STUDY)

Suppose

c = 500 hypotheses will be tested

T = 0.20, or 20% (i.e., 100/500) of these truly have an effect size, $d \geq d_s$

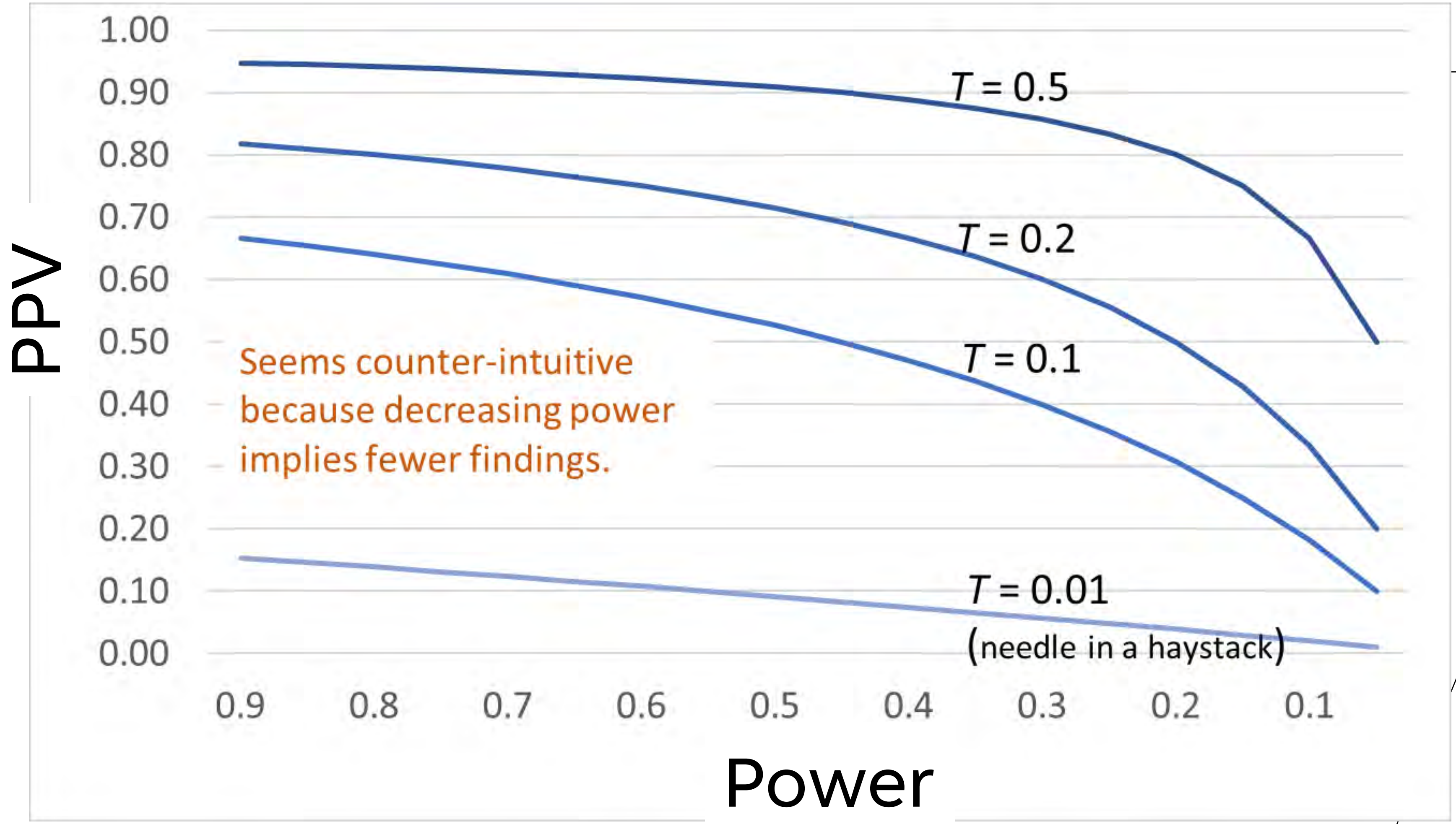
α = 0.05

β = 0.20 i.e, power is 0.80

Research Finding	True Relationship		
	Yes	No	Total
Yes	80	20	100
No	20	380	400
Total	100	400	500

PPV = 80/100 = 0.80

AS POWER OR *T* DECREASES, SO DOES PPV
(I.E., LOWER POWER → MORE FALSE FINDINGS)



IOANNIDIS SUPPOSES THE RESEARCHER IS BIASED?

"Let u be the proportion of probed analyses that would not have been research findings but nevertheless end up presented and reported as such, because of bias." – Ioannidis, 2005

Why would the researcher be biased?

- Publish or perish pressures at universities and grant funding agencies.
- Personal beliefs regarding the existence (or nonexistence) of an effect.
- Distrust of significant tests.
- Other financial and personal reasons.

IOANNIDIS SUPPOSES THE RESEARCHER IS BIASED?

"Let *u* be the proportion of probed analyses that would not have been research findings but nevertheless end up presented and reported as such, because of bias." – Ioannidis, 2005

Why would the researcher be biased?

- Publish or perish pressures at universities and grant funding agencies.
- Personal beliefs regarding the existence (or nonexistence) of an effect.
- Distrust of significant tests.
- Other financial and personal reasons.

What are the consequences of the researcher's bias?

- Statistically insignificant effects are reported as "findings" and PPV is lowered
- Research findings cannot be reproduced
- Research consumers are misled by false findings
- The reverse may also be true, i.e., significant effects are sometimes ignored

c = number of hypotheses to be tested

α = Pr(Type I error)

β = Pr(Type II error)

d_s = specified effect size

u = *proportion non-findings misclassified as findings*

T = proportion of hypotheses truly having effect size $\geq d_s$

PPV = positive predictive value

FDR = false discovery rate (1-PPV)

IOANNIDIS ASSUMES THAT SOME FRACTION, *u*, OF NULL FINDINGS ARE MISREPORTED AS REAL RESEARCH FINDINGS

Research Finding	True Relationship: Yes	True Relationship: No	Total
Yes	$c(1-\beta)T$	$c\alpha(1-T)$	$c[(1-\beta)T+\alpha(1-T)]$
No	$c\beta T$ **(×u)**	$c(1-\alpha)(1-T)$ **(×u)**	$c[\beta T+(1-\alpha)(1-T)]$
Total	cT	$c(1-T)$	c

Move *u*×100% of the "no" findings in row 2 to "yes" findings in row 1

c = number of hypotheses to be tested

α = Pr(Type I error)

β = Pr(Type II error)

d_s = specified effect size

u = *proportion non-findings misclassified as findings*

T = proportion of hypotheses truly having effect size $\geq d_s$

PPV = positive predictive value

FDR = false discovery rate (1-PPV)

RESEARCH FINDING VS TRUE RELATIONSHIP ASSUMING RESEARCHER BIAS, *u*

Research Finding	**True Relationship**		
	Yes	**No**	**Total**
Yes	$c(1-\beta)T +$ $uc\beta T$	$c\alpha(1-T) +$ $uc(1-\alpha)(1-T)$	$c[(1-\beta)T + \alpha(1-T)] +$ $uc\beta T + uc(1-\alpha)(1-T)$
No	$c\beta T -$ $uc\beta T$	$c(1-\alpha)(1-T) -$ $uc(1-\alpha)(1-T)$	$c[\beta T + (1-\alpha)(1-T)] -$ $uc\beta T - uc(1-\alpha)(1-T)$
Total	cT	$c(1-T)$	c

c = number of hypotheses to be tested

α = Pr(Type I error)

β = Pr(Type II error)

d_s = specified effect size

u = proportion non-findings misclassified as findings

T = proportion of hypotheses truly having effect size $\geq d_s$

PPV = positive predictive value

FDR = false discovery rate (1-PPV)

RESEARCH FINDING VS TRUE RELATIONSHIP ASSUMING RESEARCHER BIAS, *u*

Research Finding	**True Relationship**		
	Yes	**No**	**Total**
Yes	$c(1\text{-}\beta)T$ + $uc\beta T$	$c(1\text{-}\alpha)(1\text{-}T)$ + $uc(1\text{-}\alpha)(1\text{-}T)$	$c[(1\text{-}\beta)T + \alpha(1\text{-}T)]$ + $uc\beta T$ + $uc(1\text{-}\alpha)(1\text{-}T)$
No	$c\beta T$ – $uc\beta T$	$c(1\text{-}\alpha)(1\text{-}T)$ – $uc(1\text{-}\alpha)(1\text{-}T)$	$c[\beta T + (1\text{-}\alpha)(1\text{-}T)]$ – $uc\beta T$ - $uc(1\text{-}\alpha)(1\text{-}T)$
Total	cT	$c(1\text{-}T)$	c

divided by

$$\mathrm{PPV} = \frac{(1-\beta)+u\beta}{(1-\beta)+\alpha\frac{1-T}{T}+u\left[\beta+(1-\alpha)\frac{1-T}{T}\right]}$$

c = number of hypotheses to be tested

α = Pr(Type I error)

β = Pr(Type II error)

d_s = specified effect size

u = *proportion non-findings misclassified as findings*

T = proportion of hypotheses truly having effect size $\geq d_s$

PPV = positive predictive value

FDR = false discovery rate (1-PPV)

REPEAT PREVIOUS ILLUSTRATION, BUT ASSUME u=0.20

Research Finding	True Relationship		
	Yes	No	Total
Yes	82	58	140
No	18	342	360
Total	100	400	500

PPV = 82/140 = 0.59

i.e., PPV drops from 80% to around 60%

ACCORDING TO IOANNIDIS, THE SITUATION IS **MUCH** WORSE

- **Power < 80%. If power is 0.50, PPV < 49%, even for an otherwise "perfect" study**
- **$T < 0.20$. Exploratory studies (surveys) have 1000's of variables and T is very small (<1%)**
  - **e.g., if $T = 0.10$, power is 0.50, $u = 0.10$, PPV < 30%**
- **$u > 0.20$ according to Ioannidis who considers $0.3 \leq u \leq 0.8$**
  - **e.g., if $u = 0.3$, PPV < 20%; if $u = 0.8$, PPV = 10%**
- **Other sources of error - nonresponse, coverage error, measurement error - can exacerbate the problem, but Ioannidis doesn't discuss these.**

THE EFFECTS OF NONSAMPLING VARIANCE

Measurement variance (i.e., poor reliability, *R*)…

reduces the power **which in turn** **reduces PPV**

Power* when Reliability (R) =			
$R = 1$	$R = 0.80$	$R = 0.65$	$R = 0.50$
0.8	0.71	0.62	0.51
0.7	0.60	0.52	0.42
0.6	0.51	0.43	0.35
0.5	0.42	0.35	0.28
0.4	0.33	0.28	0.23

*Assumes T = 0.10;
(see Biemer & Trewin, 1997)

PPV* when Reliability (R) =			
$R = 1$	$R = 0.80$	$R = 0.65$	$R = 0.50$
0.64	0.61	0.58	0.53
0.61	0.57	0.53	0.48
0.57	0.53	0.49	0.43
0.53	0.48	0.44	0.39
0.47	0.42	0.38	0.33

*Assumes power is given by table to the left

THE EFFECTS OF NONSAMPLING BIAS

Nonsampling bias changes the effect size, *d*.

$$d = \frac{|\mu_1 - \mu_0|}{\sigma}$$

$$d' = \frac{|(\mu_1 + \text{Bias}) - \mu_0|}{\sigma}$$

For testing $\text{H}_0: \mu_0 = 0 \text{ vs. } \text{H}_a: \mu_0 \neq 0$

$$d' = d\,|1 + \text{RB}|$$

Relative Bias defined as $\text{RB} = \frac{\text{Bias}}{\mu_1}$

Nonsampling bias alters the distribution of effect sizes for the study

$$d' = d\,|1 + \mathrm{RB}| \Rightarrow$$

$$\mathrm{Var}(d') = \mathrm{Var}\left[(1+\mathrm{RB})d\right] = (1+\mathrm{RB})^2\,\mathrm{Var}(d)$$

$$\Rightarrow \ \mathrm{RB} > 0 \text{ increases the variance}$$

$$\mathrm{RB} < 0 \text{ decreases the variance}$$

Let *T'* denote the proportion of effect sizes, *d'*, that exceed the specified effect size, d_s. Note that:

$T' > T$ if RB > 0

$T' < T$ if RB < 0

hypotheses to be tested

α = Pr(Type I error)

β = Pr(Type II error)

d_s = specified effect size

T = proportion of hypotheses truly having effect size $\geq d_s$

T' = proportion of hypothesis with $d' \geq d_s$

PPV = positive predictive value

FDR = false discovery rate (1-PPV)

EFFECT OF NONSAMPLING BIAS ON PPV

Research Finding	True Relationship		
	Yes	No	Total
Yes ($T'<T$)	$cT'(1-\beta) + c(T-T')\alpha$	$c(1-T)\alpha$	$cT'(1-\beta) + c(1-T')\alpha$
Yes ($T'>T$)	$cT(1-\beta)$	$c(T'-T)(1-\beta) + c(1-T')\alpha$	$cT'(1-\beta) + c(1-T')\alpha$

hypotheses to be tested

α = Pr(Type I error)

β = Pr(Type II error)

d_s = specified effect size

T = proportion of hypotheses truly having effect size $\geq d_s$

T' = proportion of hypothesis with $d' \geq d_s$

PPV = positive predictive value

FDR = false discovery rate (1-PPV)

EFFECT OF NONSAMPLING BIAS ON PPV

Research Finding	True Relationship		
	Yes	No	Total
Yes ($T'<T$)	$cT'(1-\beta) + c(T-T')\alpha$	divided by →	$cT'(1-\beta) + c(1-T')\alpha$
Yes ($T'>T$)	$cT(1-\beta)$	$c(T'-T)(1-\beta) + c(1-T')\alpha$	$cT'(1-\beta) + c(1-T')\alpha$

$$PPV_{T'<T} = \frac{(1-\beta) + \frac{(T-T')}{T'}\alpha}{(1-\beta) + \frac{(1-T')}{T'}\alpha}$$

$$PPV_{T'>T} = \frac{(1-\beta)}{\frac{T'}{T}(1-\beta) + \frac{(1-T')}{T}\alpha}$$

hypotheses to be tested

α = Pr(Type I error)

β = Pr(Type II error)

d_s = specified effect size

T = proportion of hypotheses truly having effect size $\geq d_s$

T' = proportion of hypothesis with $d' \geq d_s$

PPV = positive predictive value

FDR = false discovery rate (1-PPV)

EFFECT OF NONSAMPLING BIAS ON PPV

Research Finding	True Relationship		
	Yes	No	Total
Yes ($T'<T$)	$cT'(1-\beta) + c(T-T')\alpha$	$c(1-T)\alpha$	$cT'(1-\beta) + c(1-T')\alpha$
Yes ($T'>T$)	$cT(1-\beta)$	$(T'-T)(1-\beta) + c(1-T')\alpha$	$cT'(1-\beta) + c(1-T')\alpha$

divided by

$$PPV_{T'<T} = \frac{(1-\beta) + \dfrac{(T-T')}{T'}\alpha}{(1-\beta) + \dfrac{(1-T')}{T'}\alpha}$$

$$PPV_{T'>T} = \frac{(1-\beta)}{\dfrac{T'}{T}(1-\beta) + \dfrac{(1-T')}{T}\alpha}$$

ILLUSTRATION (RB IS 10%)

Assumptions:

1. **$d_s = 0.15$ (small effect size)**
2. ***T* = 0.20 ; i.*e*., *T* = *1*-Pr(-0.15 ≤ *d* ≤ 0.15*)* = 0.20**
3. ***d* ~ N(0,δ) ==> δ =0.118 (by assumption *2*)**
4. **RB = 0.10**

It follows that

$$\begin{aligned} T' &= 1 - \Pr(-0.15 \le (1+RB) \times d \le 0.15) \\ &= 1 - \Pr(\frac{-0.15}{1.10} \le d \le \frac{0.15}{1.10}) \\ &= \mathbf{0.25} \end{aligned}$$

Because T' > *T*, PPV is given by

$$PPV_{T'>T} = \frac{(1-\beta)}{\frac{T'}{T}(1-\beta) + \frac{(1-T')}{T}\alpha} = \frac{(1-0.2)}{\frac{0.25}{0.20}(1-0.20) + \frac{0.75}{0.20}0.05} = \mathbf{0.68}$$

Compare to PPV = **0.80** for the perfect study

FINALLY, SUPPOSE RELATIVE BIAS IS 10% AND RELIABILITY IS 70% RELIABILITY

Power drops from 80% to 51%

PPV drops from 0.68 to 0.43!

This assumes *u* = 0!

As Ioannidis noted, the effects of *u* can be devastating on their own.

RB = 0.10 is considered small; is much larger for many surveys.

SUMMARY

Ioannidis' claim that - "Most research findings are false for most research designs and for most fields" has merit, but perhaps not for the reasons he states.

- That 30% - 80% of research findings are inadvertently or deliberately falsified by the researcher seems far-fetched.
- However, nonsampling errors are real and can be estimated. Unlike *u*, no speculation gauging the size of *reliability and nonsampling bias.*
- By themselves, nonsampling errors can have devastating effects on the PPV.
- Remedies such as Bonferroni and other family-wise α-level adjustments won't address these problems.

SUMMARY

Ioannidis' claim that - "Most research findings are false for most research designs and for most fields" has merit, but perhaps not for the reasons he states.

- That 30% - 80% of research findings are inadvertently or deliberately falsified by the researcher seems far-fetched.
- However, nonsampling errors are real and can be estimated. Unlike *u*, no speculation gauging the size of *reliability and nonsampling bias.*
- By themselves, nonsampling errors can have devastating effects on the PPV.
- Remedies such as Bonferroni and other family-wise α-level adjustments won't address these problems.

Other conclusions:

- 2005 was a very interesting year!
- And so ends this presentation as well as my 48-year career as a statistician.
- Maybe I'll see you at the beach!

FARE THEE WELL