Statistically significant journey: How to grow the economy and keep your hair

Stephen T. Ziliak[1]

Skeptical economists are frequently saying to Deirdre McCloskey and me, "before we join you in rejecting the mechanical *p*-value approach to statistical significance testing you will have to overturn a big result in economics." Skeptics inside of our own field of economics are claiming that Ziliak's and McCloskey's decades-long complaints about misused *p*-values, erroneous applications of Student's *t* and the rest are not worth their attention until a big influential model or economic fact is overturned by our arguments (see some of our replies in, for example, Ziliak and McCloskey 2013, 2004b).

Coming from a bunch of alleged scientists, it's a strange claim to hear. We don't much hear it from other social scientists. Nor from physicists or chemists or even professional statisticians—expert scientists and decision makers who appreciate more or less instantly the fundamental distinction between economic and statistical significance. In any case the economists are wrong. We don't think they're right, for a number of reasons. In science it is typically admitted that if a particular logic is wrong, then something—sometimes a *big* something—about the general argument is wrong. Likewise if the standard of quantitative judgment is wrong, the standardly made inference after Ronald A. Fisher is probably wrong, too (Ziliak 2008).

Take the leading case, the *p*-value. The *p*-value approach to significance testing, combined with a bright line rule of statistical significance such as $p < 0.05$, is

1 Stephen T. Ziliak is Professor of Economics and Faculty Member of the Social Justice Studies Program at Roosevelt University (Chicago) where as a faculty member he also served on the University Board of Trustees. He is also Conjoint Professor of Business and Law, University of Newcastle (Australia); Faculty Affiliate in the Graduate Program of Economics, Colorado State University and Faculty Member of the Angiogenesis Foundation (Cambridge).

indefensible on purely logical grounds beyond the missing economic "oomph" (our word for "magnitudes of economic or other substantive importance"). Besides lack of oomph another reason for doubting the test is that in its conventional formulation, the null hypothesis test procedure measured by a *p*-value, Fisher's test suffers from "the fallacy of the transposed conditional" (Ziliak and McCloskey, 2008, p. 17). The details are in *The Cult of Statistical Significance* but the fallacy comes down to this: the probability of gaining weight, given that you ate a full bag of Beloit turtles, is not the same as the probability of eating a full bag of Beloit turtles, given that you've gained weight. But the false equation—the fallacy of the transposed conditional—is made daily by significance testing economists and other scientists.

Little p-value—
What are you trying to say
of significance?

Besides, our critics exaggerate another fact. From studies of public employment programs to economic and legal scandals involving pain relief pills, Ziliak and McCloskey have overturned results in economics, medicine, pharmacology and other sciences, over and over again. The numerous examples are detailed in *The Cult of Statistical Significance* and in a couple dozen articles and book chapters. Still, the skeptics have a right to their opinion, however extreme.

In reply to the skeptics I would like to highlight two major—one could say "significant"—overturns in law and policy which are, I think you will agree, quite relevant to the theme of our conference on wealth and well-being: One, a 2011 U.S. Supreme Court case (Matrixx Initiatives v. Siracusano et al.) to which we were invited to contribute a technical brief of *amici curiae*; and, two, the 2016 American Statistical Association "Statement on *P*-values and Statistical Significance" which I influenced from behind the scenes in my role as lead author of the historic statement signed by the ASA Board of Directors. As you might imagine, there's a story here.

The Labor Department censors black unemployment rates

These overturns of "statistical significance" were not welcomed by most of our colleagues. In fact, if most had their druthers, these things wouldn't have happened at all, and I wouldn't be here.

My first discovery of the "significance" mistake—and especially of the miss-

ing economics and ethics from statistical tests—came as such things do, unexpectedly, in dealings with a bureaucrat. Many years ago, back in 1988, I received a phone call from a man in Gary, Indiana who was working on an economic development project. I was working at the time as a labor market analyst for the Indiana Department of Workforce Development, where half of my time was to be spent answering requests such as the one from the man in Gary. That's why he contacted me.

My former employer, the Indiana Department of Workforce Development, is the old "Department of Employment and Training Services"—that is, the State of Indiana branch of the federal labor exchange established by the Wagner-Peyser Act of 1933. We are the government agency charged with the task of, among other things, estimating and disseminating state and county level wages, employment, unemployment, future economic growth, that sort of thing.

The man in Gary was searching for estimates of unemployment rates for black youth workers, ages 16 to 21, in Indiana labor markets. That's why he called. Gary, Fort Wayne, Indianapolis and the other Metropolitan Statistical Areas (as they are known) down to Evansville in the southwest part of the state.

It seemed like the kind of request we could easily fill. I told him I'd be right back, and set the phone—a landline phone—down on the desk, with the man still on the line.

I pulled up my Lotus 1-2-3 spreadsheets (hurrah! Those were the good ole days) and couldn't find the data. I asked my boss and he couldn't find the data. We asked his boss and he couldn't find them. Finally we contacted the boss's boss's boss, that is, the head of labor market information for the U.S. Department of Labor, Chicago region.

For about an hour we waited in anxious curiosity. I remember the chit chat, the hand-rubbing, the pacing back and forth. "Isn't it strange," we said again and again, "that we can't find these unemployment rates?" Then the intranet email arrived, such as it was the mid and late 80s: blurry green and white blinking characters, almost at times resembling English. Reading email on line in the 1980s was like trying to read a book while playing a game of flashlight tag.

As part of the federal labor exchange, the Indiana Department of Workforce Development is overseen by the U.S. Department of Labor, which sets policy—including data dissemination and publication policy.

The Department of Labor policy was: do not distribute estimates of unemployment rates if estimates are "not statistically significant at the 0.10 level ($p < 0.10$)".

Understand: we had the data. We had the estimates. The employment office for the State of Indiana had estimates for the unemployment rates of black youth for all labor market areas. But we weren't allowed to release them. Not "significant". Not "statistically significant", they mean.

But statistical significance is not the same as economic significance. Statistical significance is not the same as economic or ethical or social justice significance.

Who would be surprised if the unemployment rate in some areas exceeded 40 or 50%? Nowadays, in some census tracts, how about 90%?

The upshot: black youth unemployment was hardly discussed in Indiana in the 1980s. Why? One big reason is because the unemployment rates were not entered into the public record!

Later that day I was embarrassed to return the Gary man's phone call to explain the official if terribly uncomfortable news.

I didn't know a fraction then of what I know now. (I had been alerted to the confusion of statistical with economic or other substantive significance by a single paragraph I had found in an introductory textbook by Wonnacott and Wonnacott [1982, p. 160].) But my moral outrage was greater than my scientific confusion, and I committed then and there to continue to study and fight against the illogical and dangerous policy of making decisions based on a bright-line rule of statistical significance. Bad decisions such as "do not discuss black unemployment".

Statistical fit—
Epistemological
strangling of wit!

Iowa professors and a "significant" article

Around that time I stumbled upon a little book called *The Rhetoric of Economics* (1985), by Donald N. McCloskey.

I was amazed to find there a chapter by an economist making some of the same points—and some different ones as well—about the censorious ritual of significance testing I had experienced at the Labor Department.

Turns out I was looking to join a good PhD program that could handle at least a little bit of this. McCloskey was a leading economic historian and philosopher and, despite our quite different politics, I took the bait: in the summer of 1991 my little family and I packed up and moved to Iowa. (At the time I was

not 100% sure where the state of Iowa is located. On the map I recall confusing it with Missouri.)

By '93 I had worked out a first draft of a controversial paper with McCloskey and in '95 our paper, called "The Standard Error of Regressions", was accepted for publication in—we could hardly believe it at the time—the *Journal of Economic Literature.*

The "J.E.L." is a blue ribbon journal, and our paper, overturning status quo beliefs about statistical significance, was going to be printed in it! As the official reference journal for the American Economic Association, the *JEL* is read and cited by economists all over the world. I, a graduate student, was understandably stoked. Ever since that phone call about black unemployment rates, back in '88, I wanted badly to do something more, and now I had.

Trouble is, not everyone was stoked. Not everyone was happy or impressed. In truth, that article earned me a lot of enemies in the academy, some of whom, twenty five years later, still hold a grudge.

In the autumn of 1995 the Iowa economics faculty organized a generous program to assist soon-to-be graduates and academic job-seekers such as me. The program was to simulate a 30 minute long job interview as typically conducted by economics faculty.

Understandably they wished to help us obtain our immediate and urgent dream, which for most of us was to get a job as an assistant professor of economics at a major university or good college. So two or three faculty members sat in a room across a table where the rising PhD and job candidate sat.

I was fortunate to be interviewed by two of the better professors, and to this day I am grateful for the tips they provided. But our meeting that day ended on a heated note.

During the question and review period at the end of the mock interview, one of the professors—an editor of *Econometrica*—suggested I remove "The Standard Error of Regressions" article from my CV.

"Why?" I asked. The other professor piped up, and answered: "It's too controversial, you won't get a job." The editor of *Econometrica* repeated, "It's too controversial for an assistant professor. It's for your own good."

Incensed, but also highly amused, I asked them: "If you had a *JEL* article, would you take it off of your CV?" (Answer: oh hell no!)

As I was walking back to my office in Brewery Square (yes, that's what it's called: on Linn Street, Iowa City) I realized I was in fact running. I ran as fast as I

could back to my office, opened up the Word Perfect file containing my CV, and started rearranging.

In the version of the CV I gave to the Iowa professors, my "Work Experience" was listed at the top of the CV and "Publications" at the bottom. (I had two publications when I graduated. At the time that was two pubs above the norm among economics grad students.)

Instead of removing "The Standard Error of Regressions" I inverted the arrangement of my CV so that the article appeared on top! For my own good.

The Iowa professors were well-intentioned but wrong. In the event, I was invited for eleven job interviews at ten excellent schools and one branch of the Federal Reserve Bank (Dallas). In a cohort of seven rising graduates from Iowa's economics PhD program, I was the first one to land a job.

The Iowa professors were in addition not the only people wondering if I was going to cut my hair before hitting the job market. My family wondered, too. One aunt went so far as to ask, "when are you going to cut that crap?" I replied, "My advisor has just changed gender in full public view. And you're worried about my hair?" I had spent many years cultivating my own version of the bohemian-philosopher look. I had no passion for the bourgeoisie and couldn't see the point of conforming to their fashion, either. I told Deirdre, the philosopher of bourgeois virtue, around that same time: "Thank you for that—I've got it made now!" And I still have my hair, that which hasn't fallen out.

Welcome, but two little things you shouldn't mention

In grad school my main field of study was economic history. And I am proud to say that in February 1996 I was offered and I accepted the one and only position in economic history which was advertised by our professional job magazine, "Job Openings for Economists" or "JOE" as it's known.

Though I got a big lump in my throat when I saw the "downtown" of the tiny Midwestern city where I would eventually work and live—at mid-day the parking lot of the Big Boy restaurant was jam packed—the job fit seemed promising. I could overlook the quality of the restaurants. But again, I met a major—one could say again, "significant"—stumbling block or choke hold if you will.

Soon after I arrived on the job my several new friends, the colleagues who recruited me, sat me down for a talk about department politics and how I should behave in the classroom and in my research leading up to tenure.

The upshot? "Don't mention rhetoric, that isn't respected by half of the tenured faculty. [In addition to the PhD in Economics I earned at Iowa a PhD Certificate in the Rhetoric of the Human Sciences, a program co-founded by Deirdre.) "And whatever you do," they said to me, the author of the *JEL* article, "don't mention your complaints about statistical significance. The faculty won't like it. You won't get tenure."

And that was coming from my friends! Just looking out, I know. Just like the Iowa professors.

After three mixed years—what do you expect to happen when you cut two of the major strings in the bow?—I was happy to get out and go elsewhere. Turns out, neither were the Georgia Tech professors ready for the long haired statistician (the students were and are).

Statistical significance stinks, says the U.S. Supreme Court

Since our first large-scale survey of best practice significance testing in economics, covering the 1980s in the *American Economic Review*, the significance mistake has gotten worse, not better. That is what we showed in "Size Matters" and again in *The Cult of Statistical Significance* (Ziliak and McCloskey 2004a, 2008). Eight or nine of every ten articles published in the leading journals of science are making the significance mistake.

Fortunately a new rule of law, handed down in 2011 by the Supreme Court of the United States, ought to help (Supreme Court of the United States, 2011a). In Matrixx Initiatives v. Siracusano et al. (2011) the high court ruled that companies can no longer conceal from investors relevantly bad news about their products by claiming that the adverse effects are not "statistically" significant at the 0.05 or any other level (note: this section borrows heavily from Ziliak and McCloskey, 2016).

Companies must consider the human meaning of the consequences, not merely the estimated probability, of biomedical results. Statistical significance without a loss function is no longer the rule of securities law. Substance, magnitude, oomph, risk of loss, is. On March 22, 2011, in *Matrixx Initiatives, Inc. v. Siracusano*, No. 09-1156, the Supreme Court rejected Fisher's rule by a 9–0 vote.

The case involved a homeopathic medicine called Zicam, a zinc-based cold remedy produced by Matrixx Initiatives. When swabbed or sprayed in the nose, the drug is expected to reduce incipient colds. But it also causes some users to lose

permanently their sense of smell (and thus of taste), a condition called *anosmia*. The loss function here is a function, then, of a high probability of stopping a cold balanced against a low probability of losing all taste of food and not smelling the flowers or your lover ever again.

(Incidentally, the econometrician Bob Elder of Beloit College has made a persuasive case in these *Proceedings* that the *F*-statistic, when derived from a loss function, can itself be used as a loss function for comparing losses. The *F*-test, he claims, does not have to be seen as a simple multi-variate *t*-test subject to the same old "sizeless stare of statistical significance" (Ziliak and McCloskey 2008, 2004a). In haiku form, Elder writes:

The F-statistic
is simply a loss function
for comparing losses.

Elder agrees with the empirical finding of our large scale survey, however, showing that more than 90% of all *F*-tests are computed without a loss function.)

When a doctor appeared on the *Good Morning America* television show in 2004 explaining the dangers of zinc-based treatments, Matrixx stock price plummeted. The company replied, though, that the adverse effect reports were not statistically significant. The company assured investors that revenue from Zicam, a hundred million dollar a year seller, was expected to grow vastly—by "50 and then 80 percent" (Supreme Court, 2011*b*: p. 3).

In the January 10, 2011, oral arguments before the Supreme Court, Justice Sotomayor chastised counsel for the petitioners (petitioning, that is, to have an appeals-court ruling against Matrixx reversed [Supreme Court, 2011*a*]).

"Mr. Hacker" was chastised for neglecting to respond to technical briefs on the subject that had been authored and filed by *amici* of the court. Many of the friends of the court, the Justice said, "did a wonderful job." (Full disclosure: we were two of the *amici* [McCloskey and Ziliak, 2010]. As is common in such matters, though, the "wonderful job" was mostly done by Allan Ingraham, an economist who drafted the brief for a New York law firm on the basis of our writings.)

Investors in Matrixx stock had filed suit against the company in a federal district court. They told the court that the company had failed to disclose the bad news it had received from expert nose doctors. But the district court dismissed the suit on the basis that investors did not prove "materiality," which meant, un-

der then-existing precedents, statistical significance. Statistical significance had long since become part of securities law: if it is statistically "insignificant" then, however illogical, it is materially insignificant, too. The Court of Appeals for the Ninth Circuit then reversed the district court's decision, reasoning in a narrow fashion "that whether facts are statistically significant, and thus [under the then-existing rule of law] material, is a question of fact that should ordinarily be left to the trier of fact—usually the jury."

The Justices went deeper. They disagreed with the definition of materiality invoked by the district court in the first place. The Justices said that the district court "erred when it took liberties in making that determination on its own."

"Something more is needed," Justice Sotomayor wrote for the unanimous Court, and the something, she said, should address the "source, content, and context" of the bad news. *Matrixx v. Siracusano* presented the Court with the question whether plaintiffs can sustain a claim of securities fraud against a company neglecting to warn investors about bad news that is *not* statistically significant. Nine to zero it ruled that they can.

The Court is not well known for economic or statistical sophistication. But, in this case, it got it right. The precedent, now the law of the land, should be followed, we believe, for all statistical reporting, nine to zero, from climate change research to randomized field experiments in developing nations. In other words, loss functions matter. Oomph is what we seek. And oomph, not the level of Student's t, is the new rule of law.

"Student" himself, by the way, that is William Sealy Gosset (1876-1937), must be dancing in his grave. Student's day job was running experiments on Guinness beer and the inputs to the beer. Student was a pen name which the publishing Mr. Gosset used to protect the brewery's proprietary rights. He rose to Head Brewer of the-then largest brewery in the world, persuading the Guinness Board with his experimental economic approach to the logic of uncertainty. As I've shown in archival work at the Guinness Archives and elsewhere, the inventor of Student's t did not put much weight on statistical significance!

What a reasonable investor might say

The Court examined the expectations of a "reasonable investor." Would undisclosed bad news be likely to negatively affect the "total mix" of information considered by a reasonable investor? If yes, then the report must be disclosed,

regardless of statistical significance or insignificance. Sotomayor wrote for the Court (Supreme Court of the United States, 2011a),

medical professionals and researchers do not limit the data they consider to the results of randomized clinical trials or to statistically significant evidence [unhappily the movement for "evidence-based medicine" may falsify her claim]…. The FDA similarly does not limit the evidence it considers for purposes of assessing causation and taking regulatory action to statistically significant data. In assessing the safety

risk posed by a product, the FDA considers factors such as "strength of the association," "temporal relationship of product use and the event," "consistency of findings across available data sources," "evidence of a dose-response for the effect," "biologic plausibility," "seriousness of the event relative to the disease being treated," "potential to mitigate the risk in the population," "feasibility of further study using observational or controlled clinical study designs," and "degree of benefit the product provides, including availability of other therapies."… [The FDA] does not apply any single metric for determining when additional inquiry or action is necessary.

To the theory of the attorneys for Matrixx that statistical significance set the standard for disclosure, over and above "background noise," Justice Breyer (Supreme Court, 2011*a*: 22) replied to a Mr. Hacker, "Oh, no, it can't be. I mean, all right—I'm sorry. I don't mean to take a position yet." [Laughter.]

JUSTICE BREYER. But, look—I mean, Albert Einstein had the theory of relativity without any empirical evidence, okay? So we could get the greatest doctor in the world, and he has dozens of theories, and the theories are very sound, and all that fits in here is an allegation he now has learned that it's the free zinc ion that counts.

MR. HACKER. But….

JUSTICE BREYER. And that could be devastating to a drug even though there isn't one person yet who has been hurt.

To Hacker's argument that statistical "significance" is the way to truth and justice in biomedical suits and cases of securities fraud, Breyer snorted, "This statistical significance always works and always doesn't work." In the same session, Sotomayor (citing *amici*) said that what counts as "statistical importance can't be a measure because it depends on the nature of the study." Justices Kagan and Ginsberg argued that small numbers of humanly meaningfully large effects can be materially relevant, independent of the level of statistical signif-

icance. Thus, the loss function. Loss of smell is bad enough, but suppose (a small number of) people died? Kagan referred to a situation in which a small number of instances of blindness were known to be associated with the use of a contact lens solution. The FDA, she noted, would not wait around for statistical significance to make a determination or to investigate further into the facts of such black swans.

Chief Justice Roberts sympathized with the test of expectations of a "reasonable investor," concluding that statistical significance was not necessary for establishing causation or belief in association. Sotomayor, in the Court's decision again (Supreme Court, 2011*b*: 1–2, 11): "We conclude that the materiality of adverse event reports cannot be reduced to a bright-line rule. Although in many cases reasonable investors would not consider reports of adverse events to be material information, respondents have alleged facts plausibly suggesting that reasonable investors would have viewed these particular reports as material…. Matrixx's argument rests on the premise that statistical significance is the only reliable indication of causation. This premise is flawed."

Statistical Significance is not material oomph

The *Matrixx* decision is consistent with the high court's prior rejection of a bright-line rule in a fact-finding and economically important situation. Citing *Basic v. Levinson* (1976), a case involving a bright-line definition for what is meant by "merger negotiations," Justice Sotomayor argued (Supreme Court, 2011*b*) that "we observed [in Basic] that 'any approach that designates a single fact or occurrence as always determinative of an inherently fact specific finding such as materiality, must necessarily be overinclusive or underinclusive.'"

Consider a pill that is thought to be effective at relieving pain but at the cost of an increased risk of heart attack. Suppose a well-designed experiment is conducted on a sample of adult humans: half taking the drug, the other half taking another and competing drug. The significance tester—in search of a single, determinative fact—then poses the question: "Assuming there is no real difference between the two pills, what is the chance that the data—showing some amount of difference—will be observed?" If the chance of seeing a difference in adverse effect larger than the one observed is less than or equal to 5 percent, it is declared to be statistically significantly different from the null hypothesis of "no difference"—without saying how much that difference is or how one should view

it. But it is an ethically flawed procedure, and before the Justices spoke it was accepted by American law.

In the early 2000s, around the time that Matrixx and Zicam were getting into trouble, a much larger producer, Merck, a pharmaceutical company, got into billions of dollars of trouble with their Vioxx pill. Vioxx-takers began to die from heart disease and heart attacks. In a clinical trial, the Merck scientists reported that Vioxx takers risk a big adverse effect—death. Yet the *p*-value came in at 0.20, meaning that a 4:1 or higher odds of experiencing a major

cost (such as death) is not worthy of policy consideration because it is not "statistically" significant at $p = .05$, or higher than 19:1 odds (see Ziliak and McCloskey, 2008, chapter 3). Therefore the company neglected the adverse outcomes. Therefore they committed the error of under-inclusiveness, a deathblow to science and lives, an error caused by unnecessary adherence to a bright-line rule of statistical significance.

Something more is needed

What the Supreme Court did not say is that the test of significance gives us the *wrong* information, period. The test gives a probability of finding a larger difference than that observed in the sample on offer, assuming that treatment and control drugs are actually the same. But that is "the fallacy of the transposed conditional" (Ziliak and McCloskey, 2008, chapters 13-16). What we really want to know is the probability of a hypothesis being true (or at least practically useful), given all the data we've got—not the other way around. We want to know the probability that the two drugs are *different* and by how much, given the available evidence. The significance test—based as it is on Fisher's fallacy of the transposed conditional—does not and cannot tell us that probability. The power function, the expected loss function, and many other decision-theoretic and Bayesian methods descending from William S. Gosset aka "Student", Harold Jeffreys, and others, now widely available, do.

A "significant" result does not in any way answer the How Much question, the question of how much or how valuable the difference in magnitude is (such as loss of smell or sight, or relief from pain, or nipping a cold in the bud). The significant result cannot demonstrate economic, medical, or any other importance for the obvious reason that it does not address it. In other words, we want to know the probability of detecting a *large and practically important difference* when

the difference is truthfully there. We need exploratory methods, a power function, an expected loss function, and, ideally speaking, a series of independently repeated experiments controlling for random and real error.

On writing the ASA Statement on Statistical Significance and P-Values

By autumn 2014 the complaints about statistical significance and its role in the crisis of reproducibility (sometimes called the "crisis of replication") had bubbled up to the top of the American Statistical Association.

With more than 19,000 members and growing, the ASA is the largest, most influential professional association of statisticians in the world. Past presidents and officers of the ASA include some of the world's most influential scientists, representing fields from economics and agriculture to physics and biology.

A number of past presidents of the ASA had spoken out against mindless significance testing, some of them while still in office. Wallis, Kruskal, Zellner and quite a few others up to and including David Morganstein and Jessica Utts, who presided over the ASA in 2015 and 2016. But compared to artists and English professors, statisticians as a group tend to fall on the conservative side of the policy activism/I'm-going-to-tell-you-what-to- do spectrum. And prior to 2015 the [second] oldest national level professional association in the country (established in 1839) had never taken a policy position on "methods," not once we believe.

So I was understandably delighted when I was tapped by Executive Director Ron Wasserstein to join a platinum team of experts, charged with making a statement about *what statistical significance, Student's t, and Fisher's p cannot do, what tests of significance do not mean and don't imply.*

The list of names on the team of around two dozen reads like a 20th and 21st Century Statistical Hall of Fame: Don Rubin, Rod Little, Don Berry, Ken Rothman, Andrew Gelman, Val Johnson, Sander Greenland, Stephen Senn, Brad Carlin, and others. Twenty four, I believe, in total. We met in person in Alexandria, VA, at ASA headquarters, on two unseasonably cold days in the middle of October 2015.

Twenty years after the Iowa professors and my first academic job, I was— though a Full Professor— a bit anxious (as was Ron Wasserstein) about the meeting. The world class statisticians would be examining line by line a draft statement and supplementary working paper drafted by me mainly, with light editing

by Ron Wasserstein and several other committee members.

There was no public hanging. On the contrary. Though the conversation grew heated on a number of points after two days our committee was able to emerge with near consensus on at least six principles related to statistical significance testing. The outcome of our work, spanning nearly a whole year, is the "ASA Statement on Statistical Significance and *P*-Values" that is now world famous, or as close to famous as such things ever get.

Here are excerpts from the Statement which was approved and signed in March 2016 by the ASA Board of Directors:

"ASA Statement on Statistical Significance and *P*-values

Introduction

. . . The validity of scientific conclusions, including their reproducibility, depends on more than the statistical methods themselves. Appropriately chosen techniques, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that uncertainty surrounding them is represented properly.

Underpinning many published scientific conclusions is the concept of "statistical significance," typically assessed with an index called the *p*-value. While the *p*-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of *p*-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since *p*-values were first introduced.

In this context, the American Statistical Association (ASA) believes that the scientific community could benefit from a formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the *p*-value. The issues touched on here affect not only research, but research funding, journal practices, career advancement, scientific education, public policy, journalism, and law. This statement does not seek to resolve all the issues relating to sound statistical practice, nor to settle foundational controversies. Rather, the statement articulates in non-technical terms a few select principles that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community.

What is a *p*-value?

Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

Principles

1. *P-values can indicate how incompatible the data are with a specified statistical model.*

A *p*-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The most common context is a model, constructed under a set of assumptions, together with a so-called "null hypothesis." Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. The smaller the *p*-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the *p*-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*

Researchers often wish to turn a *p*-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The *p*-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.

3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*

Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision-making. A conclusion does not immediately become "true" on one side of the divide and "false" on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, "yes-no" decisions, but this does not mean that *p*-values alone can ensure that a decision is correct or incorrect. The widespread use of "statistical significance" (generally interpreted as "$p \leq 0.05$") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.

4. *Proper inference requires full reporting and transparency*

P-values and related analyses should not be reported selectively. Conduct-

ing multiple analyses of the data and reporting only those with certain *p*-values (typically those passing a significance threshold) renders the reported *p*-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference and "p-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided. One need not formally carry out multiple statistical tests for this problem to arise: Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted and all *p*-values computed. Valid scientific conclusions cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including *p*-values) were selected for reporting.

5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*

Statistical significance is not equivalent to scientific, human, or economic significance. Smaller *p*-values do not necessarily imply the presence of larger or more important effects, and larger *p*-values do not imply a lack of importance or even lack of effect. Any effect, no matter how tiny, can produce a small *p*-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive *p*-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different *p*-values if the precision of the estimates differs.

6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

Researchers should recognize that a *p*-value without context or other evidence provides limited information. For example, a *p*-value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large *p*-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a *p*-value when other approaches are appropriate and feasible.

Other approaches

In view of the prevalent misuses of and misconceptions concerning *p*-values, some statisticians prefer to supplement or even replace *p*-values with other ap-

proaches. These include methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates. All of these measures and approaches rely on assumptions, but they more directly address the size of an effect and its associated uncertainty, or the likelihood that a hypothesis is true. "

You get the idea. The leading statisticians of the leading statistical association have now said in public that they agree with us. The ASA agrees. The Supreme Court of the United States, agrees. And before them, several Nobel laureates and other leading lights of econometrics, from Clive Granger and Graham Elliott to Arnold Zellner, Edward Leamer, Joel Horowitz (formerly of Iowa!) and Jeffrey Wooldridge said in print and in public at a big plenary on our work at the 2004 American Economic Association meetings: yes, we agree. We agree, the eminent theorists told an assembly of 350 economists and journalists, that (1) economic significance is not the same thing as statistical significance, and (2) most economists devote too much attention to statistical significance and not enough to economic significance (Ziliak and McCloskey 2004b; The Economist, 2004; see also Schelling 2004).

Twenty years ago, even twelve years ago, it was culturally speaking quite easy to ignore the arguments and facts against the significance mistake. Most, including most economists at the Department of Labor, did.

You'd be surprised how many smart people still Don't Get It. Steve Levitt of *Freakonomics* fame, and co-author John List, for example: clueless (Ziliak 2014). It will be interesting to see how earlier critics—the last defenders of the old status quo—Kevin Hoover, Aris Spanos, Deborah Mayo and others—reply to the new rule of law and ASA Statement. If history is any guide, probably with more defensive if erroneous arguments.

Health economists reject bright-line rules of significance

Or perhaps they will do as the health economists have done and collectively join together to reject erroneous uses of statistical significance and insignificance. In May 2015, almost a full year before publication of the ASA Statement, the editors and editorial boards of eight different journals of health economics banded together to publish the following statement, influenced by (we've been told by one of the editors) the Ziliak-McCloskey research:

"EDITORIAL STATEMENT ON NEGATIVE FINDINGS

The Editors of the health economics journals named below believe that well-designed, well-executed empirical studies that address interesting and important problems in health economics, utilize appropriate data in a sound and creative manner, and deploy innovative conceptual and methodological approaches compatible with each journal's distinctive emphasis and scope have potential scientific and publication merit regardless of whether such studies' empirical findings do or do not reject null hypotheses that may be specified. As such, the Editors wish to articulate clearly that the submission to our journals of studies that meet these standards is encouraged.

We believe that publication of such studies provides properly balanced perspectives on the empirical issues at hand. Moreover, we believe that this should reduce the incentives to engage in two forms of behavior that we feel ought to be discouraged in the spirit of scientific advancement:

1. Authors withholding from submission such studies that are otherwise meritorious but whose main empirical findings are highly likely "negative" (e.g., null hypotheses not rejected).

2. Authors engaging in "data mining," "specification searching," and other such empirical strategies with the goal of producing results that are ostensibly "positive" (e.g., null hypotheses reported as rejected).

Henceforth, we will remind our referees of this editorial philosophy at the time they are invited to review papers. As always, the ultimate responsibility for acceptance or rejection of a submission rests with each journal's Editors.

[Signed, THE EDITORS of] American Journal of Health Economics; European Journal of Health Economics; Forum for Health Economics & Policy; Health Economics Policy and Law; Health Economics Review; Health Economics; International Journal of Health Economics and Management; Journal of Health Economics"

Statistical significance is a type of scientific misconduct

The job is not complete. Practice has not changed (Ziliak and Teather-Posadas 2016). The revolution has not been televised. The next essential step or "overturn" at the national level is in my view to officially regulate and penalize deliberate misuse of statistical significance. The slower journals might wait for threats

of real loss, money and ranking. Why they'd wait is anyone's guess. Perhaps they need to feel for themselves that misuse of statistical significance is costly to science and lives. As a nation and community of ethical scientists we have to call it and penalize it for what it is: a species of scientific misconduct.

For starters, impact factors ought to be supplemented by—or even partially computed by—a numerical scale of substantive, scientific significance. A scale which penalizes the journal and article for misuse of statistical significance and rewards them for calculations of economic or other substantive significance, drawn along the lines of our surveys and principles. Today an "A"-list journal can remain A-list, despite earning an "F" grade in statistical significance. Some scale!

In a special issue on ethics and economics published in the *Review of Social Economy,* I noted that several of the major institutions for the advancement of science in the United States—from the National Institutes of Health and National Science Foundation to the American Association for the Advancement of Science itself—have sought to define and to enforce national standards for research integrity and ethical scientific conduct (Ziliak 2016). Statistical significance is not on their list of standards. Fabrication or falsification of data, deceitful manipulation, and plagiarism, I observed, are the most commonly cited forms of misconduct named and pursued. Although gross misuse of statistical significance has led to approval of faulty medical therapies which cause harm to real people—and, in some cases, such as the Vioxx debacle, even death—the scientific community has not added misuse of statistical significance to the list of scientific misconduct.

Yet researchers engaged in similar types of manipulation or questionable research practices have been penalized by those same agencies. For example, a University of Oregon researcher was recently penalized by the U.S. Department of Health and Human Services for publishing "knowingly falsified data by removing outlier values or replacing outliers with mean values to produce results that conform to predictions" (Office of Research Integrity 2015).

Misuse of statistical significance fabricates results in a similar manner and not only by dropping "insignificant" adverse results in the high-pressure drug industry. The significance mistake is undesirable, inefficient, and, in most cases—philosophers agree—unethical (see Ziliak and Teather-Posadas [2016] for theoretical discussion of ethics in empirical economics including drugs and field experiments in development economics). But the significance mistake seems to be outside the bounds of the current definition of scientific misconduct used by government agencies, research universities, and—with the extraordinary excep-

tion of the Matrixx v. Siracusano case decided by the U.S. Supreme Court—the legal process when such matters get litigated in a court of law. If we are going to stem and finally stop altogether the widespread misuse of statistical significance, we must begin to get the incentives right and in more than improved publication style and journal editorial policy.

Whatever my critics decide, I doubt I'll cut my hair.

References

American Statistical Association. 2016. "ASA Statement on Statistical Significance and *P*-Values," *The American Statistician, March issue (on line). http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108*

McCloskey, D. and S. Ziliak. 1996. "The Standard Error of Regressions." Journal of Economic Literature 34(March):97–114.

McCloskey, D. and S. Ziliak. 2010. *Brief of Amici Curiae Statistics Experts Professors Deirdre N. McCloskey and Stephen T. Ziliak in Support of Respondents*, *Matrixx Initiatives Inc. et al. v. Siracusano et al.* (vol. No. 09-1156, pp. 22). Washington, DC: Supreme Court of the United States. (Ed.) Edward Labaton et al. Counsel of Record.

Office of Research Integrity. 2015. "Findings of Research Misconduct," U.S. Department of Health and Human Services, Office of the Secretary, Washington DC: https://ori.hhs.gov/content/case-summary-anderson-david

Schelling, T. 2004. Correspondence [on Ziliak's and McCloskey's "Size Matters"] *Econ Journal Watch 1 (3*): 539-545.

Supreme Court of the United States. 2011a. "Matrixx Initiatives, Inc., et al., No. 09-1156, Petitioner v. James Siracusano et al.," *On Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit,* March 22nd, 25 pp., syllabus.

Supreme Court of the United States. 2011b. No. 09-1156, Matrixx Inititiaves, Inc., et al., Petitioner v. James Siracusano et al., On Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit, oral arguments, January 10, 2011. http://www.supremecourt.gov/oral_arguments/argument_transcripts/09-1156.pdf

The Economist. 2004. "Signifying Nothing: Too Many Economists Misuse Statistics," *The Economist,* January 29th. http://www.economist.com/node/2384590

Wonnacott, R. and R. Wonnacott. 1982. *Statistics: Discovering Its Power* (New York: Wiley).

Ziliak, S. 2008. "Guinnessometrics: The Economic Foundation of "Student's" *t*," *Journal of Economic Perspectives* 22 (4, Fall): 199-216.

Ziliak, S. 2014. "Balanced versus Randomized Field Experiments in Economics: Why W.S. Gosset Matters," *Review of Behavioral Economics* 1 (1-2): 167-208.

Ziliak, S. 2016. "Statistical Significance and Scientific Misconduct: Improving the Style of the Published Research Paper", *Review of Social Economy* 74 (1): 83-97.

Ziliak, S. and D. McCloskey. 2004a. "Size Matters: The Standard Error of Regressions in the *American Economic Review*," *The Journal of Socio-Economics* 33 (5): 527-546.

Ziliak, S. and D. McCloskey. 2004b. "Significance Redux [reply to Granger, Elliot, Zellner, Leamer, Wooldridge, Horowitz, Thorbecke, Lunt, Gigerenzer and others]" *The Journal of Socio-Economics* 33 (5): 665–675

Ziliak, S. and D. McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor: University of Michigan Press.

Ziliak, S. and D. McCloskey. 2013. "We Agree That Statistical Significance Proves Essentially Nothing: A Rejoinder to Thomas Mayer," *Econ Journal Watch* 10 (1, Jan.): 97-107.

Ziliak, S. and D. McCloskey. 2016. "Lady Justice v. Cult of Statistical Significance: Oomph-less Science and the New Rule of Law," in G. DeMartino and D. McCloskey (eds.) *Oxford Handbook of Professional Economic Ethics,* Oxford: Oxford University Press: 352-364.

Ziliak, S. and E. Teather-Posadas. 2016. "The Unprincipled Randomization Principle in Economics and Medicine," in G. DeMartino and D. McCloskey (eds.) *Oxford Handbook of Professional Economic Ethics,* Oxford: Oxford University Press: 423-452.