

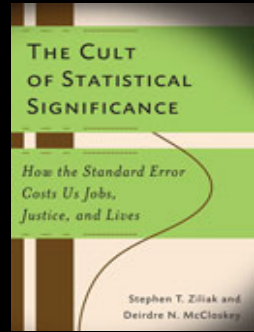
## The Cult of Statistical Significance

### How the Standard Error Costs Us Jobs, Justice, and Lives

By Stephen T. Ziliak and  
Deirdre N. McCloskey

joint statistical meetings 2009  
washington, d.c.

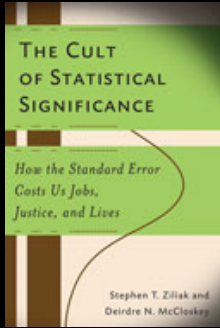
### Mindless “Significance Testing” is Costing Us Jobs, Justice, & Lives –



**The test of statistical significance** is the most important technique in the empirical branches of the life and human sciences, economics to medicine - and it is broken

**The main problem?**  
80-to-90% of scientists don't “test for” or “estimate” what we want, which is:  
**Oomph and its odds**  
(but Oomph, especially)

### Examples of Oomphless Science in Economics, Government, & Medicine



**Two diet pills, Oomph vs. Precision**  
which pill for Mom?

**Zero black unemployment rates for urban teens**  
why can't we find them?


**The 4,953+ cases of Vioxx**  
why insist on 19-to-1 odds?

**369 articles in the American Economic Review, 1980-1999**

### The “Significance” Mistake did not begin with “Student” – the inventor of $t$

A chemist by training, William Sealy Gosset (1876-1937), aka “Student”  
Learned statistics on his own—  
To solve inference problems in the Main & Experimental divisions of **Guinness's Brewery, Dublin**

Copyright: The Galton Laboratory, University College London



## “Student” was a Great Experimentalist (and a business person, too)

He invented or inspired half of modern statistics

He co-invented 3 barley varieties grown (by the 1920s) on 5 million acres, feeding breakfast eaters, beer drinkers, & other wild beasts

And he served at Guinness's as:  
Apprentice Brewer (1899-1906),  
Head Experimental Brewer (1907-1935),  
Head Statistician (c. 1922-35), and Head Brewer (Park Royal & Dublin, 1935-37)



## “Student” took an Economic Approach to the Logic of Uncertainty, 1904 to 1937

“Results are only valuable when the amount by which they probably differ from the truth is so small as to be insignificant for the purposes of the experiment.

What the odds should be depends:

1. On the degree of accuracy the experiment allows
2. On the importance of the issues at stake” (W. S. Gosset, 1904)



## “The Application of the Law of Error to the Work of the Brewery” (1904)

Gosset's report focused on MALT EXTRACT, measured in degrees saccharine per barrel of 168 lbs. malt

133° saccharine gave the targeted level of alcohol (excise tax was proportionate to alcohol level)

± .5° was an error that beer drinkers and tax payers could swallow

Source: Ziliak, "Guinnessometrics", 2008



“It might be maintained that malt extract should be [estimated] within ± .5° of the true result with a probability of 10 to 1.” He calculated extract means from a series of trials produced in the Main and Experimental Breweries.

Given the small samples he calculated the odds of observing the stipulated accuracy:

“Odds in favour of smaller error than .5

2 observations 4:1

3 “ 7:1

4 “ 12:1

5 “ 19:1

82 “ practically infinite”

**CONCLUSION** n=4 does the trick. Sort of. “How, in general, should one set the odds with small samples?”



**Fisher's campaign for a 5% Philosophy of Existence**  
**Fisher's re-formulation of "Student's" test**  
**is causing more than headaches**

Copyright: Joan Fisher Box



*Statistical Methods for  
Research Workers (1925)*

*Design of Experiments  
(1935)*

*Statistical Methods and  
Scientific Inference  
(1955/1956)*

*Statistical Tables for Bio.,  
Agri., and Medical Res.  
(with Yates, 1938)*

*And in scores of articles,  
letters, and speeches*

R.A. Fisher 1925 [1941], *Statistical Methods for  
Research Workers*, p. 42:

"The value for which  $P=.05$ , or 1 in 20, is 1.96 or  
nearly 2; *it is convenient to take this point as a  
limit in judging whether a deviation is to be  
considered significant or not. Deviations  
exceeding twice the standard deviation are thus  
formally regarded as significant.*"

R.A. Fisher 1926, "Arrangement of Field  
Experiments," p. 504

"Personally, the writer prefers to set a low  
standard of significance at the 5 per cent  
point, and *ignore entirely all results which  
fail to reach this level.*"

R.A. Fisher 1935 [1960], *The Design of  
Experiments*, p. 13:

*"It is usual and convenient for experimenters to  
take 5 per cent. as a standard level of  
significance, in the sense that they are prepared  
to ignore all results which fail to reach this  
standard, and, by this means, to eliminate from  
further discussion the greater part of the  
fluctuations which chance causes have  
introduced into their experimental results."*

R.A. Fisher 1955, "Statistical Methods and Scientific Induction," p. 75

*"Finally, in inductive inference we introduce no cost functions for faulty judgments . . . In fact, scientific research is not geared to maximize the profits of any particular organization . . . We make no attempt to evaluate these consequences, and do not assume that they are capable of evaluation in any currency."*

Compare "Student's" "*Original question and its modified form*" (1905)

"When I first reported on the subject [of "The Application of the 'Law of Error' to the Work of the Brewery"], I thought that perhaps there might be some degree of probability which is conventionally treated as sufficient in such work as ours and I advised that some outside authority in mathematics [such as Karl Pearson] should be consulted as to what certainty is required to aim at in large scale work. *However it would appear that in such work as ours the degree of certainty to be aimed at must depend on the pecuniary advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment.* This is one of the points on which I should like advice."

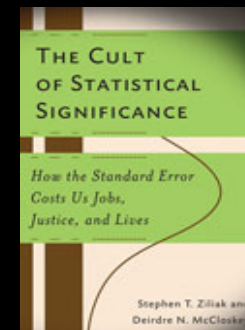
Source: W. S. Gosset to Karl Pearson, c. April 1905, in E. S. Pearson 1939, pp. 215-216; first italics in original

"Student" - as Head Brewer of Guinness - did not find "significance" to be profitable

"[O]bviously the important thing . . . is to have a low real error [said "Student" to Egon Pearson], not to have a "significant" result at a particular station. The latter ["Student" said] seems to me to be nearly valueless in itself. . . . Experiments at a single station [that is, tests of statistical significance on a single set of data] are almost valueless. . . . What you really want is a low real error. You want to be able to say not only "We have significant evidence that if farmers in general do this they will make money by it", but also "we have found it so in nineteen cases out of twenty and we are finding out why it doesn't work in the twentieth." To do that you have to be as sure as possible which is the 20th—your real error must be small."

Source: "Student" to E. S. Pearson 1937, in Pearson 1939, p. 244. (Egon was the editor of *Biometrika*, in the era before David Cox)

**If we stop doing 5% philosophy, what will we do?**

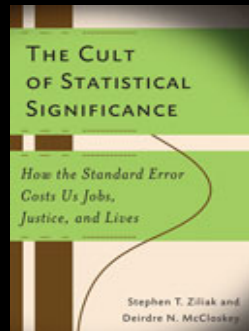


**Ask to see the Oomph, loss function, and power** of a scientific model, variate or stimulus package

**Demand to know the expected value** of a cancer treatment or diet pill, subject to "real error"

**Listen to a Guinness brewer** & don't settle for small beer - **go for substantive significance** - not mere statistical significance!

## And, of course, write more Haiku



Statistical fit:  
epistemological  
strangling, of wit!

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

## Copyright and References

"The Cult of Statistical Significance"  
was presented by Stephen T. Ziliak  
at the Joint Statistical Meetings  
(JSM), Aug. 3rd, 2009, in  
Washington, D.C.

The contents of these slides—  
including the Gosset and Fisher  
quotes—are from S.T. Ziliak's and  
D.N. McCloskey's *The Cult of  
Statistical Significance: How the  
Standard Error Costs Us Jobs,  
Justice, and Lives* (2008, University  
of Michigan Press) and from  
Ziliak's "Guinnessometrics: The  
Economic Foundation of  
'Student's'  $t$ ," *Journal of Economic  
Perspectives* (Fall 2008).

**Copyright: 2009  
Stephen T. Ziliak**