Significance Testing in the 19th Century

Glenn Shafer, Rutgers University

December 6, 2019

Measurement and Statistics Seminar

College of Education, Florida State University

Today statistics is misused and abused:

p-hacking, misuse of statistical significance, etc.

It all happened before—in the 19th century.

The 19th century story

1. Laplace solves his computational problem by discovering the central limit theorem (1810).
2. Laplace simplifies his Bayesian argument to a frequentist one, obtaining large-sample confidence intervals.
3. Fourier (1826), Poisson (1830), and Cournot (1843) make large-sample confidence intervals widely understood.
4. Significance testing and p-values emerge from confidence intervals for differences.
5. Abuses (p-hacking) emerge and are denounced by Cournot (1843).
6. Laplacean statistics (large-sample) and Gaussian engineering (best estimate) diverge.
7. Fierce argument for and against using probability in data analysis.
8. Laplacean statistics discredited and even forgotten in France by end of century.
9. Laplacean statistics revived in Britain by Karl Pearson around 1900. **Cycle repeats!**

How do we escape from the cycle?

My major conclusions:

1. Replacing p-values with confidence intervals is impossible. People invent p-values when they have confidence intervals for differences.

2. Banning the phrase “statistical significance” would accomplish nothing. The 19th century had the same problems before “statistical significance” was invented.

3. But p-values are too hard—too hard to teach, too hard to understand, too hard to remember. We need something completely different.

Inventing p-values from confidence intervals

Estimated percentage of boys:

- 53% in Paris
- 51% in the provinces

Observed difference = 2%.
Standard error = 0.9%.

Confidence intervals for true difference:

- 95% confidence interval = (0.24%,3.76%)
- 96% confidence interval = (0.15%,3.85%)
- 97% confidence interval = (0.05%,3.95%)
- 98% confidence interval = (-0.09%,4.09%)

You are 97% sure there is a difference.

The term “p-value” emerged in the 1930s,
but the idea was around in the 1830s.

My account of the history:

On the nineteenth-century origins of significance testing and p-hacking. http://www.probabilityandfinance.com/articles/55.pdf

My ideas for breaking the cycle:

The language of betting as a strategy for statistical and scientific communication. http://www.probabilityandfinance.com/articles/54.pdf

The 19th century literature is hard to read because it does not use standard deviations.

Instead of standard deviation (σ), they used probable error (0.675σ) or modulus ($\sqrt{2}\sigma$).

Bessel 1816

$$
\begin{array}{lll}
\alpha = 1 & \cdots\cdots & 1:1 \\
\alpha = 1.25 & \cdots\cdots & 1:1.505 \\
\alpha = 1.5 & \cdots\cdots & 1:2.209 \\
\alpha = 1.75 & \cdots\cdots & 1:3.204 \\
\alpha = 2 & \cdots\cdots & 1:4.638 \\
\alpha = 3 & \cdots\cdots & 1:30.51 \\
\alpha = 4 & \cdots\cdots & 1:142.36
\end{array}
$$

Gaussian school of least squares, which became dominant in astronomy, surveying, etc.

Fourier 1826

units of g	P	units of $g/\sqrt{2}$
0.47708	$\frac{1}{2}$	0.67
1.38591	$\frac{1}{20}$	1.96
1.98495	$\frac{1}{200}$	2.81
2.46130	$\frac{1}{2000}$	3.48
2.86783	$\frac{1}{20000}$	4.06

Used in testing for significant differences in social science and medicine, beginning in the 1830s.

Discredited by the end of 19th century. Revived in England.

Cournot's criticism of p-hacking in 1843

Clearly nothing limits the number of the aspects under which we can consider the natural and social facts to which statistical research is applied …

nor, consequently, the number of variables according to which we can distribute them into different groups or distinct categories.

Suppose we want to determine the chance of a masculine birth.

… in general it exceeds 1/2.

We can first distinguish between legitimate births and those outside marriage…

We can further distinguish between births in the countryside and births in the city...

…we could also classify births according to

- their order in the family,
- age, profession, wealth, and religion of the parents;
- first marriages vs second marriages,
- seasons of the year…

… as the number of groupings grows without limit, it is more and more likely *a priori* that merely as a result of chance at least one of the groupings will produce values appreciably different...

…for a statistician who undertakes a thorough investigation, the probability of a deviation of given size not being attributable to chance will have very different values depending on whether he has tried more or fewer groupings...

… usually the groupings that the experimenter went through leave no trace;

… the public only sees the result that seemed to merit being brought to its attention.

… an individual unacquainted with the system of groupings that preceded the result will have absolutely no fixed rule for betting on whether the result can be attributed to chance.

Airy's 1861 table

	Modulus.	Mean Error.	Error of Mean (σ) Square.	Probable Error.
In terms of Modulus	1.000000	0.564189	0.707107	0:476948
In terms of Mean Error	1.7724.54	1.000000	1.253314	0.845369
In terms of Error of Mean Square (σ)	1.414214	0.797885	1.000000	0.674506
In terms of Probable Error	2.096665	1.182916	1.482567	1.000000

modulus ≈ 2 probable errors

2 moduli ≈ 2.8 standard deviations
= practical certainty Gavarret 1830s ≈ 0.5%

2 standard deviations ≈ 3 probable errors
= Karl Pearson's probably significant ≈ 5%

4 standard deviations ≈ 6 probable errors ≈ 3 moduli
= Fourier's certainty 1826 & Karl Pearson's definitely significant ≈ 1/50,000

The English word “significance” was introduced into statistical testing by Francis Edgeworth in 1885.

- Edgeworth was repeating in English what Wilhelm Lexis had written in German.

- Lexis: A sufficiently large observed difference makes it *practically certain* (*praktisch gewiss*) that there is a real difference.

- Edgeworth’s translation: A sufficiently large observed difference *signifies* a real difference. It is *significant* of a real difference.

Karl Pearson adopted the Edgeworthian use of significant.

- Pearson and his followers talked about observed differences being
  --*probably significant*
  --*probably not significant*
  --*definitely significant*

- Usually they did not say the differences were “just barely significant” or “highly significant.

During the 1885 to 1915 period, significance was not a matter of degree. The probability of significance was a matter of degree.

This had changed by the time R. A. Fisher wrote Statistical Methods for Research Workers in 1925.

Karl Pearson's student Raymond Pearl, writing in 1923

There has grown up a certain conventional way of interpreting probable errors, which is accepted by many workers. It has been practically a universal custom among biometric workers to say that a difference (or a constant) which is smaller than twice its probable error is probably not significant, whereas a difference (or constant) which is three or more times its probable error is either "certainly," or at least "almost certainly," significant.

Edgeworth 1885

From p. 182: The science of Means comprises two main problems: 1. To find how far the difference between any proposed Means is accidental or indicative of a law? 2. To find what is the best kind of Mean; whether for the purpose contemplated by the first problem, the elimination of chance, or other purposes? ...The first problem investigates how far the difference between the average above stated and the results usually obtained in similar experience where pure chance reigns is a significant difference; indicative of the working of a law other than chance, or merely accidental. ...

...out of a set of (say) N statistical numbers which fulfil the law of error, we take one at random, it is exceedingly improbable that it will differ from the Mean to the extent of twice, and *à fortiori* thrice, the modulus.

From p. 188: ...we shall find that the observed difference between the proposed Means, namely about 2 (inches) far exceeds thrice the modulus of that curve, namely 0*2. The difference therefore "comes by cause."

Biometrika 1900-1920

In *Biometrika*'s first volume, we find "perhaps significant", "more probably significant", and "certainly significant". Here are some additional instances of the Edgeworthian "significant":

- In the very first issue, from Pearson's close collaborator W. F. R. Weldon [125, p. 119]: "With probable errors of the order indicated by Tables I. and II., it is unlikely that any of these differences are significant. Even in the case of the last pair of entries the difference, although it is considerable (0·0229 mm.), is less than twice the probable error of the determination."
- In the second issue, from Oswald H. Latter [74, p. 167]: "To test whether any deviation is significant, M_r is taken as the mean of the whole race of Cuckoos and M_s the mean of Cuckoo's eggs found in the nest of any one species of foster-parent: the standard deviation (σ_s) of such eggs is also ascertainied. The value of $M_r - M_s$ is then compared with that of $0 \cdot 67449\sqrt{\frac{\sigma_r^2}{n_1} + \frac{\sigma_s^2}{n_2}}$, where n_1 = total number of Cuckoo's eggs and n_2 = the number of Cuckoo's eggs in the nests of the species in question, which is the probable error of $M_r - M_s$ due to random sampling. If the value of $M_r - M_s$ be not at least 1.5 to 3 times as great as the value of the other expression the difference of M_r and M_s is not definitely significant."
- In a 1912 article co-authored by Pearson himself [8, p. 301]: "Hence the difference is more than three times the probable error and likely to be significant."

Anders Hald (1990, 1998) reviews the earliest work on significance tests:

- Study of the sex ratio by Arbuthnott (1712), Nicholas Bernoulli (1713), Daniel Bernoulli (1770)
- De Moivre's calculation of the probabilities of coincidences
- Daniel Bernoulli's test of the randomness of planetary orbits (1734)
- Michell's test for the random distribution of stars (1767)
- Lambert's remarks on checking the reliability of weather forecasts (1773)
- Laplace's test of the randomness of planetary orbits (1776 and 1810)

Widespread use and abuse of significance testing derives, however, from large-sample confidence intervals based on Laplace's central limit theorem.

Laplace originally derived his large sample confidence intervals, in 1810, by a Bayesian argument, but very quickly he and others began using the simpler "frequentist" arguments for them.

My personal view on how to break the cycle

The 19th century history shows that p-values arise naturally from confidence intervals.

But p-values are too hard—too hard to teach, too hard to understand, too hard to remember.

Not just one confidence interval (or test).

A whole family of confidence intervals (or tests) with different significant levels.

We need something completely different.

I suggest replacing p-values with the outcomes of bets.

Alice announces probabilities for sports events.

- One week she does a tennis tournament, assigning each player a probability of winning.
- The next week she gives probabilities for a soccer game—probabilities for Real Madrid winning, for Barcelona winning, and for a tie.
- Then she announces a probability distribution for the winning point spread between the Nets and the 76ers.
- And so on.

How could you test Alice?

One way is to try to make money at the odds she offers.

Can you think of any other way?

Testing by betting

Bet by buying a random variable for its expected value.

- Bob challenges Alice's prowess as a probability forecaster by betting at the odds Alice announces.
- If Bob begins with $1, risks no more than this, and walks away with $100 after a year of betting, he will have put a big dent in Alice's reputation as a forecaster.
- Alice may plead that she was merely unlucky, but she cannot claim success as a forecaster.

Bob is a frequentist, not a Bayesian.
Bayesians make bets on hypotheses—bets that are never settled.
Bob makes bets that are settled, and uses the outcomes like p-values.

Bob can challenge and discredit Alice without giving alternative probabilities.

Maybe he does not believe that there are meaningful or reliable probabilities for the events in question.

Bob does not need to risk real money.

He can bet with play money.

His goal is to make a point, not to get rich.

When he uses play money, he does not need a counterparty to his bets.

So Alice is not risking real money either; she is risking only her reputation as a forecaster.

Alice may know more about the sports and the competitors than Bob.

If Bob makes money on her forecasts, then perhaps Alice's additional information is not very relevant.

Bob may know more about the sports and the competitors than Alice.

If Alice has a good reputation as a sports forecaster, and yet Bob makes money on her forecasts, then information known to Bob but not to Alice may be relevant.

If Bob does not make money betting against Alice's probabilities, then we have no evidence against them.

If we know Bob to be very clever and very knowledgeable about the events in question, then we have evidence that Alice is doing her job well.

The idea of testing by betting can be used in standard statistical problems, where the betting can be thought of as buying likelihood ratios.

This also leads to more flexible methods of meta-analysis.

See my working paper “The language of betting as a strategy for statistical and scientific communication”, http://www.probabilityandfinance.com/articles/54.pdf, and my new book with Vovk.