Should we eliminate P-values or use more of them?: A discussion on the P-value controversy

Śaunak Sen

2019-05-07

Seminar series on statistical reasoning in biomedical research

- Apr 30: P-values: What they are and what they are not (Fridtjof Thomas, PhD)
- May 07: Should We eliminate P-Values or Use More of Them: A Discussion on the P-Value Controversy (Saunak Sen, PhD)
- May 14: The Bayesian Approach to Data Analysis (Fridtjof Thomas, PhD)
- May 21: Multiple Testing and the False Discovery Rate (Saunak Sen, PhD)
- May 28: The Perfect Doctor: An introduction to Causal Inference (Fridtjof Thomas, PhD)
- Jun 04: Enhancing Statistical Methods in Grants and Papers (Saunak Sen, PhD)

Outline

- History
- P-value definition and example
- Criticisms and debate
- Way forward

Controversy

In 2016 the American Statistical Association (ASA) released a statement on P-values.

- Siegfried, T. (2010), “Odds Are, It’s Wrong: Science Fails to Face the Shortcomings of Statistics,” Science News, 177, 26. Link
- Phys.org Science News Wire (2013), “The Problem With p Values: How Significant are They, Really?” Link html.
- Nuzzo, R. (2014), “Scientific Method: Statistical Errors,” Nature, 506, 150–152. Link
- Gelman, A., and Loken, E. (2014), “The Statistical Crisis in Science,” American Scientist, 102. Link
- Leek, J. (2014), “On the Scalability of Statistical Procedures: Why the p-Value Bashers Just Don’t Get It,” Simply Statistics Blog, Link
- Peng, R. (2015), “The Reproducibility Crisis in Science: A Statistical Counterattack,” Significance, 12, 30–32.

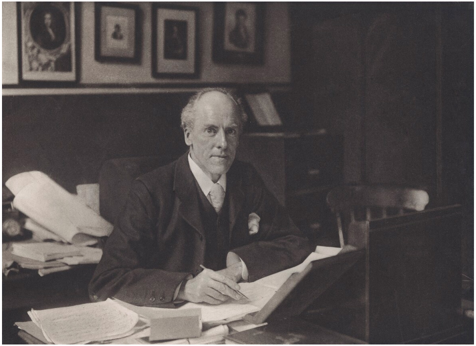
Controversy

In 2019 the ASA published a followup issue in the American Statistician

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond "$p < 0.05$", The American Statistician, 73:sup1, 1-19, DOI

- Nature editorial: It's time to talk about ditching statistical significance
- Call to retire statistical significance: Amrheim, Greenland, McShane and 800 signatories

History



Term coined by Karl Pearson as away of calibrating the results of a χ^2 goodness of fit test.

Pearson, Karl (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling" Philosophical Magazine. Series 5. 50 (302): 157–175.

History

Refined by R.A. Fisher to test more general hypotheses, and suggested the cutoff of 0.05 for “statistical significance.”

Fisher, Ronald (1925). Statistical Methods for Research Workers. Edinburgh, Scotland: Oliver & Boyd.

History

Jerzy Neyman and Egon Pearson formalized a decision-theoretic framework for statistical testing. Very influential in the second half of the 20th century.

Neyman, J.; Pearson, E. S. (1933-02-16). " On the problem of the most efficient tests of statistical hypotheses". Phil. Trans. R. Soc. Lond. A. 231 (694–706): 289–337.

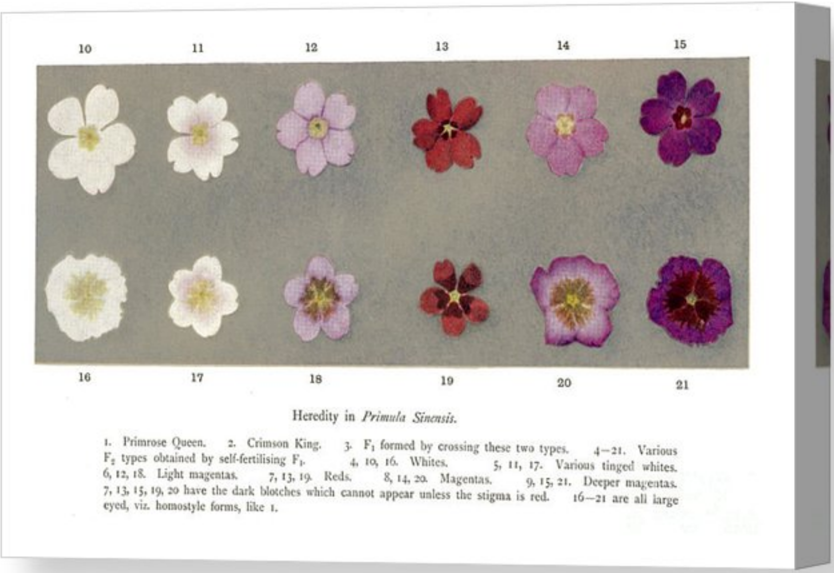
Primula sinensis data (de Winton and Bateson)

Do the eye type (Normal or Primrose Queen) and floral leaf type (Flar or Crimped) segregate independently in a 9:3:3:1 ratio?

If the normal leaf is dominant, and the flat leaf is dominant, in an F_2 cross, these characters are expected to segregate in a 9:3:3:1 ratio.

```
                   Eye
Leaves       Normal     Primrose Queen
 Flat           328             122
 Crimped         77              33
```

Primula sinensis



Heredity in *Primula Sinensis.*

1. Primrose Queen. 2. Crimson King. 3. F_1 formed by crossing these two types. 4–21. Various F_2 types obtained by self-fertilising F_1. 4, 10, 16. Whites. 5, 11, 17. Various tinged whites. 6, 12, 18. Light magentas. 7, 13, 19. Reds. 8, 14, 20. Magentas. 9, 15, 21. Deeper magentas. 7, 13, 15, 19, 20 have the dark blotches which cannot appear unless the stigma is red. 16–21 are all large eyed, viz. homostyle forms, like 1.

Primula sinensis

To test the null hypothesis (9:3:3:1 segregation, against the alternative hypothesis that is not the case, we can calculate the χ^2 statistic using the expected and observed counts.

```
                   Eye
Leaves      Normal     Primrose Queen
 Flat          328          122                OBSERVED
 Crimped        77           33
```

```
                   Eye
Leaves      Normal     Primrose Queen
 Flat          315          105                EXPECTED
 Crimped       105           35
```

The χ^2 statistic is 10.87, which has a χ^2 distribution with 3 degrees of freedom under the null. The chance that a χ^2 with 3 degrees of freedom will exceed the observed value, 10.87, is 0.012. This is the P-value.

Primula sinensis

Although there is a mention of the alternative hypothesis, it plays a secondary role in this case. In practice, this is how we use p-values which is very Fisherian.

The null hypothesis is actually a composite of three hypotheses: both characters are dominant (and segregate in a 3:1 ratio, and that they segregate independently (not linked).

Which hypotheses are true or false? This is not answered by the p-value.

Early debate (1935)

> *For example, we may wish to test whether a given sample differs significantly from a random sample from a normal population. Applying the χ^2 test, after finding the best fitting normal distribution, and using p $=$ 0.05, say, as the level of significance, we may find that our sample is just not significantly abnormal.*
> *The χ^2 criterion is perfectly justifiable up to this point. It is quite unjustifiable, however, to assert that the reverse hypothesis is true, namely, that the sample is likely to have come from a normal population, unless we have other reasons to believe this, in which case, of course, the chi^2 is not used as a criterion of the truth of the reverse hypothesis.*

If the p-value is less than 0.05 we cannot say that the null hypothesis is true.

Buchanan-Wollaston HJ (1935) Statistical Tests, Nature:136,182

Early debate (1942)

Berkson J (1942), "Tests of Significance Considered as Evidence," Journal of the American Statistical Association, 37, 325-335.

> *One of our most eminent members gave a paper presenting the application of the lambda test and used for illustration data designed to test a certain Mendelian hypothesis. The data having been examined and the test applied, a P of about 0.6 was found. "We can say therefore," he remarked, "that the results substantiate the hypothesis."*

Fisher responds:

> *It is not my purpose to make Dr. Berkson seem ridiculous, nor, of course, to prevent him from providing innocent amusement.*

Early debate (1942)

Berkson responds:

> *PROFESSOR R. A. FISHER in his note on my article "Tests of Significance Considered as Evidence" finds it pertinent to say, 'It is not my purpose to make Dr. Berkson seem ridiculous, nor, of course, to prevent him from providing innocent amusement,' and to make some biographical comments of a similar intention. This is no time to begin calling names across the sea, and however strongly I may differ with Professor Fisher in regard to scientific questions, I shall confine my differences to the subject matter discussed.*

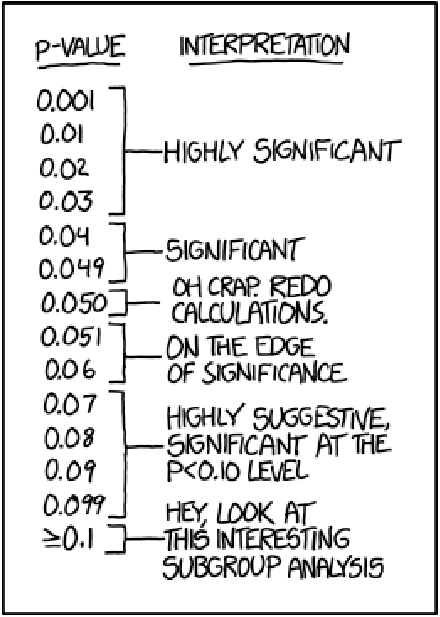
ASA statement on statistical significance and P-Values (2016)

1. P-values can indicate how incompatible the data are with a specified statistical model
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
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

ASA statement summary (2016)

Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.

XKCD


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Moving beyond p<0.05 (Wasserstein, Schirm, Lazar, 2019)

1. Don’t say “statistically significant”
2. Accept uncertainty
3. Be thoughtful (look at the big picture, consider context and prior knowledge, consider alternatives to P-values)
4. Be open (to transparency, role of expert judgement, and in communication)
5. Editorial, educational, and other institutional practices will have to change

Proposal: More stringent cutoffs

Valen Johnson - Revise standards for statistical evidence: Proposed a cutoff of 0.005 as standard.

A modeling of the consequences showed that:

> *Published results may be more reliable, but publication rates would go down.*
> *There may be a large cost in terms of fewer true breakthrough discoveries.*
> *We found that the impacts of adopting a sample size requirement policy are similar to the impacts of lowering the α.*

Harlan Campbell & Paul Gustafson The World of Research Has Gone Berserk: Modeling the Consequences of Requiring "Greater Statistical Stringency" for Scientific Publication

Proposal: Ban p-values

In 2016, the journal Basic and Applied Social Psychology (BASP) banned p-values.

> *In this article, we assess the 31 articles published in Basic and Applied Social Psychology (BASP) in 2016, which is one full year after the BASP editors banned the use of inferential statistics. We discuss how the authors collected their data, how they reported and summarized their data, and how they used their data to reach conclusions. We found multiple instances of authors overstating conclusions beyond what the data would support if statistical significance had been considered. Readers would be largely unable to recognize this because the necessary information to do so was not readily available.*

Ronald D. Fricker Jr., Katherine Burke, Xiaoyan Han & William H. Woodall Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their p-Value Ban

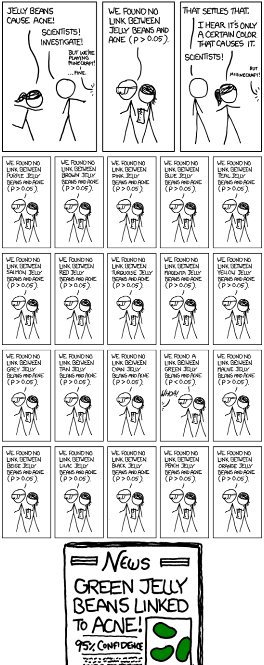
Proposal: Use confidence intervals

Another proposal is to use confidence intervals whenever possible.

Estruch et. al. (2018) Primary Prevention of Cardiovascular Disease with a Mediterranean Diet Supplemented with Extra-Virgin Olive Oil or Nuts, NEJM

This also relies on an arbitrary level of significance for the confidence interval, but may be preferable to using the p-value alone.

XKCD


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE

Other proposals

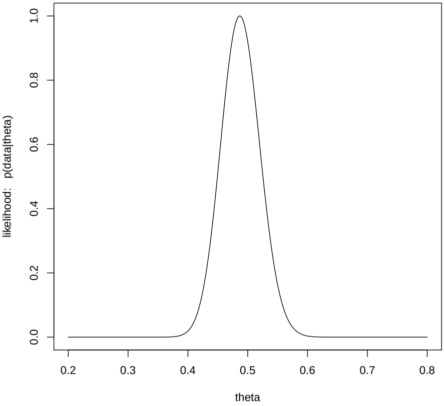
Use Bayesian statistics (see next lecture by Dr. Thomas). It has a logically coherent theory of how information accumulates. This requires expert knowledge, and there is potential of misuse.

Provide context about all tests used (see lecture on multiple comparisons in two weeks): Need to be honest about the process by which discovery was made.

Use causal models (see lecture on causal inference by Dr. Thomas in three weeks): By thinking more carefully about the process by which data was obtained, and the underlying causal mechanisms more thoughtful conclusions can be made.

Primula data

We can formulate separate hypotheses. For example, if they are linked with a recombination fraction θ the expected proportions are like this. The evidence can be laid out by a likelihood function which is $p(data|\theta)$. Traits likely not linked.



Primula data

We can also test the hypotheses that the two traits are dominant.

The 95% confidence interval for proportion of normal eye is (0.68 0.76). The p-value for the test of 3:1 segregation is 0.1571.

The 95% confidence interval for proportion of flat leaves is (0.77 0.84). The p-value for the test of 3:1 segregation is 0.004.

There is evidence of some segregation distortion, but it is not very big and there is considerable uncertainty.

The segregation distortion is the likely cause of the small χ^2 test p-value.

Future seminars

- May 14: The Bayesian Approach to Data Analysis (Fridtjof Thomas, PhD)
- May 21: Multiple Testing and the False Discovery Rate (Saunak Sen, PhD)
- May 28: The Perfect Doctor: An introduction to Causal Inference (Fridtjof Thomas, PhD)
- Jun 04: Enhancing Statistical Methods in Grants and Papers (Saunak Sen, PhD)

Slides at https://tnctsi.uthsc.edu