ASA Statement on P-Values

P. Swanson

April 1, 2016

American Statistical Association Statement on P-Values

- In March, the American Statistical Association (ASA) published a statement intended to clarify the meaning and use of p-values for non-statisticians [4, 13]
    - Definition + six principles to aid in understanding
    - 21 members of the committee were asked to write supplemental opinions which were published online www.asa.com
- The statement is the result of a yearlong discussion by a special panel set up by the ASA board
- 1st time in 177 year history ASA took an official position on a "matter of statistical practice"
- Excludes discussions of alternative hypothesis testing, error types, family-wise error rates, power, false discovery rates, multiple testing, etc.

Motivation for ASA Statement

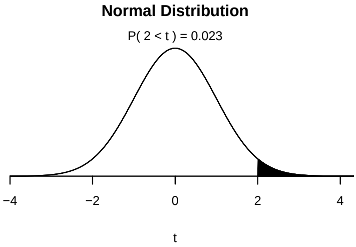
- Reproducibility Crisis in science
    - Bright lines (0.05)
    - P-hacking
    - File Drawer problem (publish or perish)
- The journal *Basic and Applied Social Psychology* banned p-values (NHST) [12]
- Supreme Court ruled that companies could not rely solely on statistical significance when deciding what to disclose to investors [6]
- Much older concerns about the misuse and misunderstanding of p-values
    - "The p-value was never intended to be a substitute for scientific reasoning. Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post $p < 0.05$ era.'" - R. Wasserstein, ASA President [4]

ASA Statement - Definition

ASA non-technical definition:
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. [13]

Background - Probability Distributions

- A random variable, T, follows a probability distribution, F
- If we know that distribution, we can say something about the probability of observing different values of T
- For example, let $T \sim N(0, 1)$. Say we have some observation $t = 2$ and we want to know the probability of observing a value of T at least as large as "2". Since the total area under the PDF is equal to 1, $P(T \geq 2) = 0.023$.



P-value Example

- Q: What is a P-value?
- A: $P(T \geq t_0 | H_0)$ [11]

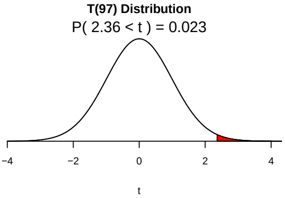
- Consider a regression example where we have specified the model $Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$
- Say we want to focus on variable X.
- IF
  1. The null hypothesis ($\beta_1 = 0$) is true
  2. All model assumptions are met
- THEN
  - We define a test statistic, $t_0 = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)}$, which in this example we know follows a T distribution $t_0 \sim T_{n-p-1}$

P-value (Simulated) Example

Say we have estimated the model, $Y = \beta_0 + \beta_1 X + \beta_2 Z$ on $n = 100$ observations.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.08	0.10	-0.81	0.42
x	0.22	0.09	2.36	0.02
z	0.33	0.09	3.72	0.00

- Our test statistic $t_0 = 2.36$.
- If our null hypothesis is true, $H_0 : \beta_1 = 0$, and the model is correctly specified, the probability of observing a test statistic at least as high as 2.36 is given by our p-value, $p = 0.02$.
- The p-value is equal to the area under the t_97 distribution to the right of the observed test statistic $t_0 = 2.36$.



Common Misinterpretations

- P-values can't say whether a hypothesis is true: $P(T \geq t_0 | H_0) \neq P(H_0 | T \geq t_0)$
  - "A p-value of 0.05 does not mean that there is a 95% chance that a given hypothesis is correct. Instead, it signifies that if the null hypothesis is true, and all other assumptions made are valid, there is a 5% chance of obtaining a result at least as extreme as the one observed." [5]
  - "Most scientists would look at his original P value of 0.01 and say that there was just a 1% chance of his result being a false alarm. But they would be wrong. The P value cannot say this: all it can do is summarize the data assuming a specific null hypothesis. It cannot work backwards and make statements about the underlying reality. That requires another piece of information: the odds that a real effect was there in the first place." [10]
  - "If you use p = 0.05 to suggest that you have made a discovery, you will be wrong at least 30% of the time. If, as is often the case, experiments are under powered, you will be wrong most of the time." [7]
- Nearly everything is significant in nature, it is just a matter of precision
  - The difference between 0.0 vs 0.0000000000001 is statistically significant
  - Problem becomes evident with larger datasets ($\uparrow n \propto \uparrow$ precision)
  - The math is correct, but p-values answer a question no one cares about
  - "[p-values are] at best uninformative and at worst seriously misleading" [9]

Misuse

- *P*-hacking & Multiple Testing
  - See [2] for a tutorial on *P*-hacking!
- Bright Lines (i.e. 0.05) used without context
  - There is no meaningful scientific basis for using 0.05 as a threshold. Fisher just said it was "convenient", then in the same book finds an effect with a $p < 0.05$ and dismisses it based on additional analysis and knowledge of the problem at hand.[8]
  - "The irony is that when UK statistician Ronald Fisher introduced the P value in the 1920s, he did not mean it to be a definitive test. He intended it simply as an informal way to judge whether evidence was significant in the old-fashioned sense: worthy of a second look."[10]
- *P*-values are not effect sizes
  - "...a *p*-value cannot indicate the importance of a finding..." [5]

Why All the Confusion?

- P-values are hard to explain
  - “Try to distill the p-value down to an intuitive concept and it loses all its nuances and complexity, said science journalist Regina Nuzzo, a statistics professor at Gallaudet University. “Then people get it wrong, and this is why statisticians are upset and scientists are confused.” You can get it right, or you can make it intuitive, but it’s all but impossible to do both.” [1]
- P-values are misrepresented in applied stats books, Centers for Public Health, CDC website, ... (see [3] for examples)
- Circularity à la George Cobb
  - Q: Why do so many colleges and grad schools teach $p = .05$?
  - A: Because that’s still what the scientific community and journal editors use.
  - Q: Why do so many people still use $p = .05$?
  - A: Because that’s what they were taught in college or grad school.

ASA Statement - Definition + Example

ASA non-technical definition:
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. [13]

Specified Model:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.08	0.10	-0.81	0.42
x	0.22	0.09	2.36	0.02
z	0.33	0.09	3.72	0.00

Under our null hypothesis, $\beta_1 = 0$, we define a test statistic which follows a t_{n-k-1} distribution.

$$\frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)} \sim T_{n-k-1}$$

$$P(T \geq t_0 | H_0)$$

ASA Statement - Principle 1

Principle 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.

ASA Statement - Principle 2

Principle 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.**

ASA Statement - Principle 3

Principle 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.**

ASA Statement - Principle 4

Principle 4: Proper inference requires full reporting and transparency

- *P*-values and related analyses should not be reported selectively. **Conducting 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 based on p-values and related statistics 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.

ASA Statement - Principle 5

Principle 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.

ASA Statement - Principle 6

Principle 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.

Implications for Predictive Modelers

- Confirmation
  - Predictive modelers have always put a heightened emphasis on predictive power rather than blindly following *p*-values. Out-of-sample testing, backtesting, and sensitivity analysis provide more rigorous model validation.
  - Subject Matter Experts are brought in early and consulted often as a way to identify and confirm model inputs.
- The ASA did not recommend abandoning *p*-values
  - "In some sense it offers a first line of defense against being fooled by randomness, separating signal from noise, because the models it requires are simpler than any other statistical tool needs."[7]
- Alternatives & Additional Analyses
  - Business knowledge / practical considerations
  - Effect sizes & sensitivity analysis
  - Backtesting
  - Confidence/credibility/prediction intervals
  - MC integration to better understand complex processes
  - Bayesian methods, $P(H|X)$

Conclusion

- *p*-values are often misused and misunderstood
- *p*-values are still useful, but should not be used alone
- It is important that we understand their limitations and explain them rather than resorting to oversimplifying statements like “a *p*-value is the probability that the null hypothesis is correct”. This can lead to a false sense of certainty about our models
- ASA Statement confirms predictive modeling’s focus on alternatives to *p*-values

Bibliography

[1] C. Aschwanden. Not even scientists can easily explain p-values. *FiveThirtyEight. com, Nov*, 24:2015, 2015.

[2] C. Aschwanden. Science isn't broken. *URL: http://www. fivethirtyeight. com/features/science-isnt-broken/#part1*, 8 2015.

[3] C. Aschwanden. Statisticians found one thing they can agree on: Its time to stop misusing p-values, 3 2016.

[4] A. S. Association. American statistical association releases statement on statistical significance and p-values: Provides principles to improve the conduct and interpretation of quantitative science. *URL: https://www. amstat. org/newsroom/pressreleases/P-ValueStatement. pdf*, 2016.

[5] M. Baker et al. Statisticians issue warning on p values. *Nature*, 531(7593):151, 2016.

[6] C. Bialik. Making a stat less significant. *The Wall Street Journal*, 4 2011.

[7] D. Colquhoun. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society open science*, 1(3):140216, 2014.

[8] R. Fisher. Statistical methods for research workers. edinburgh: Oliver and boyd, 1925. *Google Scholar*, 1932.

[9] N. Matloff. From algorithms to z-scores: Probabilistic and statistical modeling in computer science. *Creative Commons License*, 2009.

[10] R. Nuzzo. Scientific method: statistical errors. *Nature News*, 506(7487):150, 2014.

[11] J. Rice. *Mathematical Statistics and Data Analysis*. Cengage Learning, 2006.

[12] D. Trafimow and M. Marks. Editorial. *Basic and Applied Social Psychology*, 37:1–2, 2015.

[13] R. L. Wasserstein, N. A. Lazar, et al. The asa statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.