Misuses and Misinterpretations of P-values

Claudio Fronterrè

SpatstatEpi Reading Group, 10 May 2017

Contents

History

Misinterpretations of single P-values

Misinterpretations of P-Value Comparisons

Conclusions

History

ASA Statement on P-values : Why?

1. 2010 On ScienceNews Siegfried writes "It's science's dirties secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation".
2. 2013 An article in Phys.org Science News Wire cited "numerous deep flaws" in null hypothesis significance testing.
3. 2014 Siegfried on ScienceNews said "statistical techniques for testing hypotheses ...have more flaws than Facebook's privacy policies."
4. 2014 Statistician and "Simply Statistics" blogger Jeff Leek responded. "The problem is not that people use P-values poorly, it is that the vast majority of data analysis is not performed by people properly trained to perform data analysis"
5. 2014 Statistician and science writer Regina Nuzzo published an article in Nature entitled "Scientific Method: Statistical Errors".

ASA Statement on P-values : Why?

1. 2010 On ScienceNews Siegfried writes “It’s science’s dirties secret: The ‘scientific method‘ of testing hypotheses by statistical analysis stands on a flimsy foundation”.
2. 2013 An article in Phys.org Science News Wire cited “numerous deep flaws” in null hypothesis significance testing.
3. 2014 Siegfried on ScienceNews said “statistical techniques for testing hypotheses ...have more flaws than Facebook’s privacy policies.”
4. 2014 Statistician and “Simply Statistics” blogger Jeff Leek responded. “The problem is not that people use P-values poorly, it is that the vast majority of data analysis is not performed by people properly trained to perform data analysis”
5. 2014 Statistician and science writer Regina Nuzzo published an article in Nature entitled “Scientific Method: Statistical Errors”.

ASA Statement on P-values : Why?

1. 2010 On ScienceNews Siegfried writes “It’s science’s dirties secret: The ‘scientific method‘ of testing hypotheses by statistical analysis stands on a flimsy foundation”.
2. 2013 An article in Phys.org Science News Wire cited “numerous deep flaws” in null hypothesis significance testing.
3. 2014 Siegfried on ScienceNews said “statistical techniques for testing hypotheses ...have more flaws than Facebook’s privacy policies.”
4. 2014 Statistician and “Simply Statistics” blogger Jeff Leek responded. “The problem is not that people use P-values poorly, it is that the vast majority of data analysis is not performed by people properly trained to perform data analysis”
5. 2014 Statistician and science writer Regina Nuzzo published an article in Nature entitled “Scientific Method: Statistical Errors”.

ASA Statement on P-values : Why?

1. 2010 On ScienceNews Siegfried writes "It's science's dirties secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation".

2. 2013 An article in Phys.org Science News Wire cited "numerous deep flaws" in null hypothesis significance testing.

3. 2014 Siegfried on ScienceNews said "statistical techniques for testing hypotheses ...have more flaws than Facebook's privacy policies."

4. 2014 Statistician and "Simply Statistics" blogger Jeff Leek responded. "The problem is not that people use P-values poorly, it is that the vast majority of data analysis is not performed by people properly trained to perform data analysis"

5. 2014 Statistician and science writer Regina Nuzzo published an article in Nature entitled "Scientific Method: Statistical Errors".

ASA Statement on P-values : Why?

1. 2010 On ScienceNews Siegfried writes "It's science's dirties secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation".

2. 2013 An article in Phys.org Science News Wire cited "numerous deep flaws" in null hypothesis significance testing.

3. 2014 Siegfried on ScienceNews said "statistical techniques for testing hypotheses ...have more flaws than Facebook's privacy policies."

4. 2014 Statistician and "Simply Statistics" blogger Jeff Leek responded. "The problem is not that people use P-values poorly, it is that the vast majority of data analysis is not performed by people properly trained to perform data analysis"

5. 2014 Statistician and science writer Regina Nuzzo published an article in Nature entitled "Scientific Method: Statistical Errors".

ASA Statement on P-values : Why?

6. 2014 George Cobb, Professor Emeritus of Mathematics and Statistics, posed these questions to an ASA discussion forum:
   - Q1 Why do so many colleges and grad schools teach $p = 0.05$?
   - A1 Because that's still what the scientific community and journal editors use.
   - Q2 Why do so many people still use $p = 0.05$?
   - A2 Because that's what they were taught in college or grad school.
7. 2015 Trafimow and Marks, editors of *Basic and Applied Social Psychology*, decided to ban p-values (null hypothesis significance testing)

ASA Statement on P-values : Why?

6. 2014 George Cobb, Professor Emeritus of Mathematics and Statistics, posed these questions to an ASA discussion forum:
   - Q1 Why do so many colleges and grad schools teach $p = 0.05$?
   - A1 Because that's still what the scientific community and journal editors use.
   - Q2 Why do so many people still use $p = 0.05$?
   - A2 Because that's what they were taught in college or grad school.
7. 2015 Trafimow and Marks, editors of *Basic and Applied Social Psychology*, decided to ban p-values (null hypothesis significance testing)

Misinterpretations of single P-values

Definition of P-value

General

A statistical summary of the compatibility between the observed data and what we would predict or expect to see if we knew the entire statistical model (all the assumptions used to compute the *P*-value) were correct.

1. The *P*-value tests the entire model, not just the targeted hypothesis it is supposed to test.

2. A very small *P*-value does not tell us which assumption is incorrect.

Definition of P-value

General

A statistical summary of the compatibility between the observed data and what we would predict or expect to see if we knew the entire statistical model (all the assumptions used to compute the P-value) were correct.

1. The P-value tests the entire model, not just the targeted hypothesis it is supposed to test.

2. A very small P-value does not tell us which assumption is incorrect.

Definition of P-value

General

A statistical summary of the compatibility between the observed data and what we would predict or expect to see if we knew the entire statistical model (all the assumptions used to compute the P-value) were correct.

1. The P-value tests the entire model, not just the targeted hypothesis it is supposed to test.

2. A very small P-value does not tell us which assumption is incorrect.

Misinterpretations of single P-values

1) The P-value is the probability that the test hypothesis is true; for example, if a test of the null hypothesis gave $P = 0.01$, the null hypothesis has only a 1% chance of being true; if instead it gave $P = 0.40$, the null hypothesis has a 40% chance of being true.

Misinterpretations of single P-values

2) The P-value for the null hypothesis is the probability that chance alone produced the observed association; for example, if the P-value for the null hypothesis is 0.08, there is an 8% probability that chance alone produced the association.

Misinterpretations of single P-values

3) A significant test result ($P \leq 0.05$) means that the test hypothesis is false or should be rejected.

$P \leq 0.05$ only means that a discrepancy from the hypothesis prediction (e.g., no difference between treatment groups) would be as large or larger than that observed no more than 5% of the time if only chance were creating the discrepancy.

4) A nonsignificant test result ($P > 0.05$) means that the test hypothesis is true or should be accepted.

Misinterpretations of single P-values

3) A significant test result ($P \leq 0.05$) means that the test hypothesis is false or should be rejected.

$P \leq 0.05$ only means that a discrepancy from the hypothesis prediction (e.g., no difference between treatment groups) would be as large or larger than that observed no more than 5% of the time if only chance were creating the discrepancy.

4) A nonsignificant test result ($P > 0.05$) means that the test hypothesis is true or should be accepted.

Misinterpretations of single P-values

> *3) A significant test result ($P \leq 0.05$) means that the test hypothesis is false or should be rejected.*

$P \leq 0.05$ only means that a discrepancy from the hypothesis prediction (e.g., no difference between treatment groups) would be as large or larger than that observed no more than 5% of the time if only chance were creating the discrepancy.

> *4) A nonsignificant test result ($P > 0.05$) means that the test hypothesis is true or should be accepted.*

Misinterpretations of single P-values

5) A large P-value is evidence in favor of the test hypothesis.

A large *P*-value oftend indicates only that the data are incapable of discriminating among many competing hypotheses.

6) A null-hypothesis P-value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated.

Misinterpretations of single P-values

5) A large P-value is evidence in favor of the test hypothesis.

A large *P*-value oftend indicates only that the data are incapable of discriminating among many competing hypotheses.

6) A null-hypothesis P-value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated.

Misinterpretations of single P-values

5) A large P-value is evidence in favor of the test hypothesis.

A large P-value oftend indicates only that the data are incapable of discriminating among many competing hypotheses.

6) A null-hypothesis P-value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated.

Misinterpretations of single P-values

7) Statistical significance indicates a scientifically or substantively important relation has been detected.

Stastical siginificane and scientific significance are two different things. The best practice is to look at the confidence interval to determine if the effect size is of scientific or other substantive importance. Moreover, any effect, no matter how tiny, can produce a small P-value if the sample size or measurament precision is high enough.

7) Statistical significance indicates a scientifically or substantively important relation has been detected.

Stastical siginificane and scientific significance are two different things. The best practice is to look at the confidence interval to determine if the effect size is of scientific or other substantive importance. Moreover, any effect, no matter how tiny, can produce a small P-value if the sample size or measurament precision is high enough.

Misinterpretations of single P-values

8) Lack of statistical significance indicates that the effect size is small.

Real large effects may be hidden by noise and thus fail to be detected by a statistical test if sample size is smal or measurements are imprecise.

Misinterpretations of single P-values

8) Lack of statistical significance indicates that the effect size is small.

Real large effects may be hidden by noise and thus fail to be detected by a statistical test if sample size is smal or measurements are imprecise.

Misinterpretations of single P-values

9) $P = 0.05$ and $P \leq 0.05$ mean the same thing.

10) P-values are properly reported as inequalities (e.g., report "$P < 0.02$" when $P = 0.015$ or report $P > 0.05$ when $P = 0.06$ or $P = 0.70$).

Misinterpretations of single P-values

9) $P = 0.05$ and $P \leq 0.05$ mean the same thing.

10) P-values are properly reported as inequalities (e.g., report "$P < 0.02$" when $P = 0.015$ or report $P > 0.05$ when $P = 0.06$ or $P = 0.70$).

Misinterpretations of P-Value Comparisons

Misinterpretations of P-Value Comparisons

1) When the same hypothesis is tested in different studies and none or a minority of the tests are statistically significant (all P > 0.05), the overall evidence supports the hypothesis.

If there were five studies each with P = 0.10, none would be significant at 0.05 level; but when these P-values are combined using the Fisher formula, the overall P-value would be 0.01. Thus, lack of statistical significance of individual studies should not be taken as implying that the totality of evidence supports no effect.

Misinterpretations of P-Value Comparisons

1) When the same hypothesis is tested in different studies and none or a minority of the tests are statistically significant (all P $>$ 0.05), the overall evidence supports the hypothesis.

If there were five studies each with P $= 0.10$, none would be significant at 0.05 level; but when these P-values are combined using the Fisher formula, the overall P-value would be 0.01. Thus, lack of statistical significance of individual studies should not be taken as implying that the totality of evidence supports no effect.

Misinterpretations of P-Value Comparisons

2) When the same hypothesis is tested in two different populations and the resulting P-values are on opposite sides of 0.05, the results are conflicting.

Suppose we had two randomized trials A and B of a treatment, identical except that trial A had a known standard error of 2 for the mean difference between treatment groups whereas trial B had a known standard error of 1 for the difference. If both trials observed a difference between treatment groups of exactly 3, the usual normal test would produce P = 0.13 in A but P = 0.003 in B. Despite their difference in P-values, the test of the hypothesis of no difference in effect across studies would have P = 1, reflecting the perfect agreement of the observed mean differences from the studies.

Misinterpretations of P-Value Comparisons

2) When the same hypothesis is tested in two different populations and the resulting P-values are on opposite sides of 0.05, the results are conflicting.

Suppose we had two randomized trials A and B of a treatment, identical except that trial A had a known standard error of 2 for the mean difference between treatment groups whereas trial B had a known standard error of 1 for the difference. If both trials observed a difference between treatment groups of exactly 3, the usual normal test would produce $P = 0.13$ in A but $P = 0.003$ in B. Despite their difference in P-values, the test of the hypothesis of no difference in effect across studies would have $P = 1$, reflecting the perfect agreement of the observed mean differences from the studies.

Misinterpretations of P-Value Comparisons

3) When the same hypothesis is tested in two different populations and the same P-values are obtained, the results are in agreement.

Suppose randomized experiment A observed a mean difference between treatment groups of 3.00 with standard error 1.00, while B observed a mean difference of 12.00 with standard error 4.00. Then the standard normal test would produce $P = 0.003$ in both; yet the test of the hypothesis of no difference in effect across studies gives $P = 0.03$, reflecting the large difference ($12.00 - 3.00 = 9.00$) between the mean differences.

Misinterpretations of P-Value Comparisons

3) When the same hypothesis is tested in two different populations and the same P-values are obtained, the results are in agreement.

Suppose randomized experiment A observed a mean difference between treatment groups of 3.00 with standard error 1.00, while B observed a mean difference of 12.00 with standard error 4.00. Then the standard normal test would produce $P = 0.003$ in both; yet the test of the hypothesis of no difference in effect across studies gives $P = 0.03$, reflecting the large difference $(12.00 - 3.00 = 9.00)$ between the mean differences.

Misinterpretations of P-Value Comparisons

4) If one observes a small P-value, there is a good chance that the next study will produce a P-value at least as small for the same hypothesis.

If one observes $P = 0.03$, the chance that the new study will show $P \leq 0.03$ is only 3%; thus the chance the new study will show a P-value as small or smaller (the "replication probability") is exactly the observed P-value!

Misinterpretations of P-Value Comparisons

4) If one observes a small P-value, there is a good chance that the next study will produce a P-value at least as small for the same hypothesis.

If one observes $\mathrm{P} = 0.03$, the chance that the new study will show $\mathrm{P} \leq 0.03$ is only 3%; thus the chance the new study will show a P-value as small or smaller (the "replication probability") is exactly the observed P-value!

Conclusions

Guidelines

- Correct and careful interpretation of statistical tests demands examining the sizes of effect estimates and confidence limits, as well as precise P-values.
- Critical examination of the assumptions and conventions used for the statistical analysis.
- Interval estimates aid in evaluating whether the data are capable of discriminating among various hypotheses about effect sizes.
- Correct statistical evaluation of multiple studies requires a pooled analysis or meta-analysis that deals correctly with study biases.
- Any opinion offered about the probability, likelihood, certainty, or similar property for a hypothesis cannot be derived from significance tests and confidence intervals alone.

Guidelines

- Correct and careful interpretation of statistical tests demands examining the sizes of effect estimates and confidence limits, as well as precise P-values.
- Critical examination of the assumptions and conventions used for the statistical analysis.
- Interval estimates aid in evaluating whether the data are capable of discriminating among various hypotheses about effect sizes.
- Correct statistical evaluation of multiple studies requires a pooled analysis or meta-analysis that deals correctly with study biases.
- Any opinion offered about the probability, likelihood, certainty, or similar property for a hypothesis cannot be derived from significance tests and confidence intervals alone.

Guidelines

- Correct and careful interpretation of statistical tests demands examining the sizes of effect estimates and confidence limits, as well as precise P-values.
- Critical examination of the assumptions and conventions used for the statistical analysis.
- Interval estimates aid in evaluating whether the data are capable of discriminating among various hypotheses about effect sizes.
- Correct statistical evaluation of multiple studies requires a pooled analysis or meta-analysis that deals correctly with study biases.
- Any opinion offered about the probability, likelihood, certainty, or similar property for a hypothesis cannot be derived from significance tests and confidence intervals alone.

Guidelines

- Correct and careful interpretation of statistical tests demands examining the sizes of effect estimates and confidence limits, as well as precise P-values.
- Critical examination of the assumptions and conventions used for the statistical analysis.
- Interval estimates aid in evaluating whether the data are capable of discriminating among various hypotheses about effect sizes.
- Correct statistical evaluation of multiple studies requires a pooled analysis or meta-analysis that deals correctly with study biases.
- Any opinion offered about the probability, likelihood, certainty, or similar property for a hypothesis cannot be derived from significance tests and confidence intervals alone.

Guidelines

- Correct and careful interpretation of statistical tests demands examining the sizes of effect estimates and confidence limits, as well as precise P-values.
- Critical examination of the assumptions and conventions used for the statistical analysis.
- Interval estimates aid in evaluating whether the data are capable of discriminating among various hypotheses about effect sizes.
- Correct statistical evaluation of multiple studies requires a pooled analysis or meta-analysis that deals correctly with study biases.
- Any opinion offered about the probability, likelihood, certainty, or similar property for a hypothesis cannot be derived from significance tests and confidence intervals alone.