UoE Statistics reading group 13 May 2019, Finn Lindgren

To do or not to do statistical significance testing

Quoted material from

- Amrhein, Greenland, McShane (and 800 signatories), 2019, Nature, Scientists rise up against statistical significance
  https://www.nature.com/articles/d41586-019-00857-9
- Wasserstein, Schirm, Lazar, 2019, The American Statistician, Moving to a World Beyond "p<0.05"
  https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913
- Discussion on Andrew Gelman's blog
  https://statmodeling.stat.columbia.edu/2019/03/20/retire-statistical-significance-the-discussion/

Webpage: https://www.maths.ed.ac.uk/~flindgre/events/statistics-reading-group-arguments-about-statistical-significance-testing/

Amrhein et al:

> *When was the last time you heard a seminar speaker claim there was 'no difference' between two groups because the difference was 'statistically non-significant'?*
>
> *If your experience matches ours, there's a good chance that this happened at the last talk you attended. We hope that at least someone in the audience was perplexed if, as frequently happens, a plot or table showed that there actually was a difference.*
>
> *[...] we should never conclude there is 'no difference' or 'no association' just because a P value is larger than a threshold such as 0.05 or, equivalently, because a confidence interval includes zero. Neither should we conclude that two studies conflict because one had a statistically significant result and the other did not. These errors waste research efforts and misinform policy decisions.*

Amrhein et al:

> *When was the last time you heard a seminar speaker claim there was 'no difference' between two groups because the difference was 'statistically non-significant'?*
> *If your experience matches ours, there's a good chance that this happened at the last talk you attended. We hope that at least someone in the audience was perplexed if, as frequently happens, a plot or table showed that there actually was a difference.*
>
> *[...] we should never conclude there is 'no difference' or 'no association' just because a P value is larger than a threshold such as 0.05 or, equivalently, because a confidence interval includes zero. Neither should we conclude that two studies conflict because one had a statistically significant result and the other did not. These errors waste research efforts and misinform policy decisions.*

Amrhein et al:

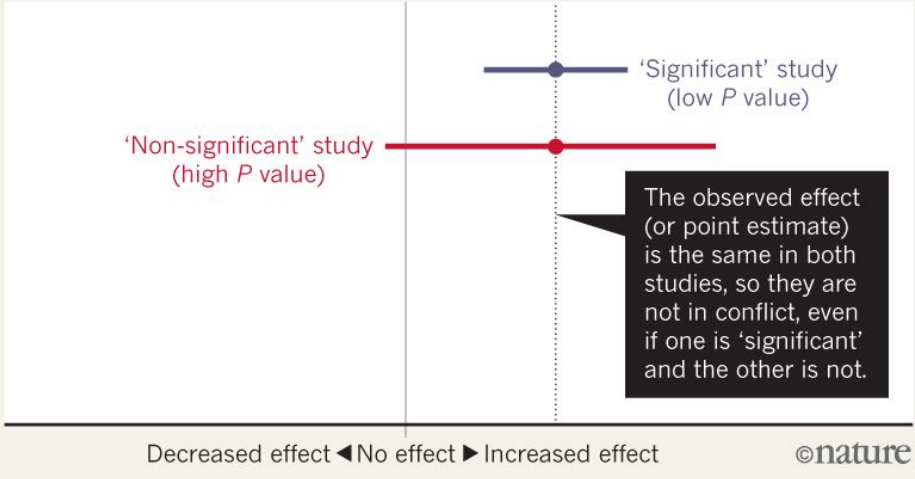
> *When was the last time you heard a seminar speaker claim there was 'no difference' between two groups because the difference was 'statistically non-significant'?*
> *If your experience matches ours, there's a good chance that this happened at the last talk you attended. We hope that at least someone in the audience was perplexed if, as frequently happens, a plot or table showed that there actually was a difference.*
>
> *[...] we should never conclude there is 'no difference' or 'no association' just because a P value is larger than a threshold such as 0.05 or, equivalently, because a confidence interval includes zero. Neither should we conclude that two studies conflict because one had a statistically significant result and the other did not. These errors waste research efforts and misinform policy decisions.*

Amrhein et al:

> *When was the last time you heard a seminar speaker claim there was 'no difference' between two groups because the difference was 'statistically non-significant'?*
> *If your experience matches ours, there's a good chance that this happened at the last talk you attended. We hope that at least someone in the audience was perplexed if, as frequently happens, a plot or table showed that there actually was a difference.*
>
> *[...] we should never conclude there is 'no difference' or 'no association' just because a P value is larger than a threshold such as 0.05 or, equivalently, because a confidence interval includes zero. Neither should we conclude that two studies conflict because one had a statistically significant result and the other did not. These errors waste research efforts and misinform policy decisions.*

BEWARE FALSE CONCLUSIONS

Studies currently dubbed 'statistically significant' and 'statistically non-significant' need not be contradictory, and such designations might cause genuine effects to be dismissed.



Amrhein et al:

> *In 2016, the American Statistical Association released a statement in The American Statistician warning against the misuse of statistical significance and P values. [...]*
>
> *[In March 2019], a special issue in the same journal attempts to push these reforms further. It presents more than 40 papers on "Statistical inference in the 21st century: a world beyond P<0.05". The editors introduce the collection with the caution "don't say 'statistically significant'".*
>
> *Another article with dozens of signatories also calls on authors and journal editors to disavow those terms.*
>
> *We agree, and call for the entire concept of statistical significance to be abandoned.*
>
> *We are far from alone. [The 800 signatories] include statisticians, clinical and medical researchers, biologists and psychologists from more than 50 countries and across all continents except Antarctica.*

Amrhein et al:

> *In 2016, the American Statistical Association released a statement in The American Statistician warning against the misuse of statistical significance and P values. [...]*
>
> *[In March 2019], a special issue in the same journal attempts to push these reforms further. It presents more than 40 papers on "Statistical inference in the 21st century: a world beyond P<0.05". The editors introduce the collection with the caution "don't say 'statistically significant'".*
>
> *Another article with dozens of signatories also calls on authors and journal editors to disavow those terms.*
>
> *We agree, and call for the entire concept of statistical significance to be abandoned.*
>
> *We are far from alone. [The 800 signatories] include statisticians, clinical and medical researchers, biologists and psychologists from more than 50 countries and across all continents except Antarctica.*

Amrhein et al:

> *In 2016, the American Statistical Association released a statement in The American Statistician warning against the misuse of statistical significance and P values. [...]*
>
> *[In March 2019], a special issue in the same journal attempts to push these reforms further. It presents more than 40 papers on "Statistical inference in the 21st century: a world beyond P<0.05". The editors introduce the collection with the caution "don't say 'statistically significant'".*
>
> *Another article with dozens of signatories also calls on authors and journal editors to disavow those terms.*
>
> *We agree, and call for the entire concept of statistical significance to be abandoned.*
>
> *We are far from alone. [The 800 signatories] include statisticians, clinical and medical researchers, biologists and psychologists from more than 50 countries and across all continents except Antarctica.*

Amrhein et al:

> *In 2016, the American Statistical Association released a statement in The American Statistician warning against the misuse of statistical significance and P values. [...]*
>
> *[In March 2019], a special issue in the same journal attempts to push these reforms further. It presents more than 40 papers on "Statistical inference in the 21st century: a world beyond P<0.05". The editors introduce the collection with the caution "don't say 'statistically significant'".*
>
> *Another article with dozens of signatories also calls on authors and journal editors to disavow those terms.*
>
> *We agree, and call for the entire concept of statistical significance to be abandoned.*
>
> *We are far from alone. [The 800 signatories] include statisticians, clinical and medical researchers, biologists and psychologists from more than 50 countries and across all continents except Antarctica.*

Amrhein et al:

> *In 2016, the American Statistical Association released a statement in The American Statistician warning against the misuse of statistical significance and P values. [...]*
>
> *[In March 2019], a special issue in the same journal attempts to push these reforms further. It presents more than 40 papers on "Statistical inference in the 21st century: a world beyond P<0.05". The editors introduce the collection with the caution "don't say 'statistically significant'".*
>
> *Another article with dozens of signatories also calls on authors and journal editors to disavow those terms.*
>
> *We agree, and call for the entire concept of statistical significance to be abandoned.*
>
> *We are far from alone. [The 800 signatories] include statisticians, clinical and medical researchers, biologists and psychologists from more than 50 countries and across all continents except Antarctica.*

Amrhein et al:

Quit categorizing

The trouble is human and cognitive more than it is statistical:

bucketing results into 'statistically significant' and 'statistically non-significant' makes people think that the items assigned in that way are categorically different.

The same problems are likely to arise under any proposed statistical alternative that involves dichotomization, whether frequentist, Bayesian or otherwise.

Amrhein et al:

[...] we are not advocating a ban on P values, confidence intervals or other statistical measures — only that we should not treat them categorically. This includes dichotomization as statistically significant or not, as well as categorization based on other statistical measures such as Bayes factors.

Amrhein et al:

Quit categorizing

The trouble is human and cognitive more than it is statistical: bucketing results into 'statistically significant' and 'statistically non-significant' makes people think that the items assigned in that way are categorically different.

The same problems are likely to arise under any proposed statistical alternative that involves dichotomization, whether frequentist, Bayesian or otherwise.

Amrhein et al:

[...] we are not advocating a ban on P values, confidence intervals or other statistical measures — only that we should not treat them categorically. This includes dichotomization as statistically significant or not, as well as categorization based on other statistical measures such as Bayes factors.

Amrhein et al:

Quit categorizing

The trouble is human and cognitive more than it is statistical: bucketing results into 'statistically significant' and 'statistically non-significant' makes people think that the items assigned in that way are categorically different.

The same problems are likely to arise under any proposed statistical alternative that involves dichotomization, whether frequentist, Bayesian or otherwise.

Amrhein et al:

[...] we are not advocating a ban on P values, confidence intervals or other statistical measures — only that we should not treat them categorically. This includes dichotomization as statistically significant or not, as well as categorization based on other statistical measures such as Bayes factors.

Amrhein et al:

> **Quit categorizing**
>
> *The trouble is human and cognitive more than it is statistical: bucketing results into 'statistically significant' and 'statistically non-significant' makes people think that the items assigned in that way are categorically different.*
>
> *The same problems are likely to arise under any proposed statistical alternative that involves dichotomization, whether frequentist, Bayesian or otherwise.*

Amrhein et al:

> *[...] we are not advocating a ban on P values, confidence intervals or other statistical measures — only that we should not treat them categorically. This includes dichotomization as statistically significant or not, as well as categorization based on other statistical measures such as Bayes factors.*

Wasserstein et al:

> *[...] the problem is not that of having only two labels. Results should not be trichotomized, or indeed categorized into any number of groups, based on arbitrary p-value thresholds. Similarly, we need to stop using confidence intervals as another means of dichotomizing (based, on whether a null value falls within the interval). And, to preclude a reappearance of this problem elsewhere, we must not begin arbitrarily categorizing other statistical measures (such as Bayes factors).*

Partial countercomment (Lindgren):

- The theoretical equivalence between testing and confidence intervals isn't a practical equivalence
- NHST considers only a *single* null hypothesis
- A confidence interval describes a *collection* of null hypotheses
- The confidence interval holds information beyond a dichotomy; Amrhein et al argue for a name change: *compatibility intervals*

Wasserstein et al:

> *[...] the problem is not that of having only two labels. Results should not be trichotomized, or indeed categorized into any number of groups, based on arbitrary p-value thresholds. Similarly, we need to stop using confidence intervals as another means of dichotomizing (based, on whether a null value falls within the interval). And, to preclude a reappearance of this problem elsewhere, we must not begin arbitrarily categorizing other statistical measures (such as Bayes factors).*

Partial countercomment (Lindgren):

- The theoretical equivalence between testing and confidence intervals isn't a practical equivalence
- NHST considers only a *single* null hypothesis
- A confidence interval describes a *collection* of null hypotheses
- The confidence interval holds information beyond a dichotomy; Amrhein et al argue for a name change: *compatibility intervals*

Wasserstein et al:

> *[...] the problem is not that of having only two labels. Results should not be trichotomized, or indeed categorized into any number of groups, based on arbitrary p-value thresholds. Similarly, we need to stop using confidence intervals as another means of dichotomizing (based, on whether a null value falls within the interval). And, to preclude a reappearance of this problem elsewhere, we must not begin arbitrarily categorizing other statistical measures (such as Bayes factors).*

Partial countercomment (Lindgren):

- The theoretical equivalence between testing and confidence intervals isn't a practical equivalence
- NHST considers only a *single* null hypothesis
- A confidence interval describes a *collection* of null hypotheses
- The confidence interval holds information beyond a dichotomy; Amrhein et al argue for a name change: *compatibility intervals*

Wasserstein et al:

> *[...] the problem is not that of having only two labels. Results should not be trichotomized, or indeed categorized into any number of groups, based on arbitrary p-value thresholds. Similarly, we need to stop using confidence intervals as another means of dichotomizing (based, on whether a null value falls within the interval). And, to preclude a reappearance of this problem elsewhere, we must not begin arbitrarily categorizing other statistical measures (such as Bayes factors).*

Partial countercomment (Lindgren):

- The theoretical equivalence between testing and confidence intervals isn't a practical equivalence
- NHST considers only a *single* null hypothesis
- A confidence interval describes a *collection* of null hypotheses
- The confidence interval holds information beyond a dichotomy; Amrhein et al argue for a name change: *compatibility intervals*

Wasserstein et al:

- *Don't base your conclusions solely on whether an association or effect was found to be "statistically significant" [...]*
- *Don't believe that an association or effect exists just becauseit was statistically significant.*
- *Don't believe that an association or effect is absent justbecause it was not statistically significant.*
- *Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.*
- *Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).*

But also, and at least as important,

"Don't" is not enough

Wasserstein et al has a long list of "Do" suggestions in Section 7

Wasserstein et al:

- *Don't base your conclusions solely on whether an association or effect was found to be "statistically significant" [...]*
- *Don't believe that an association or effect exists just becauseit was statistically significant.*
- *Don't believe that an association or effect is absent justbecause it was not statistically significant.*
- *Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.*
- *Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).*

But also, and at least as important,

"Don't" is not enough

Wasserstein et al has a long list of "Do" suggestions in Section 7

Gelman:

> *I agree with [Daniel Lakeland's] statement:*
> *I just don't think that "teaching the true meaning of p values" is an important part of the path.*
> *Part of this is just that class time is precious so why waste it on a method that is not relevant to most applied questions (other than the question, "How can I get my noisy study accepted in Psychological Science?").*
> *[...]*
> *Pretty much the main point of teaching the true meaning of p-values would be to convince people not to use them. And that's not really where i want to spend most of my time as a teacher, telling people what not to do. It's just kind of demoralizing for all concerned. Maybe we should be teaching that way, but I think it's a hard sell for teachers and for students alike.*

Gelman:

> *I agree with [Daniel Lakeland's] statement:*
> *I just don't think that "teaching the true meaning of p values" is an important part of the path.*
> *Part of this is just that class time is precious so why waste it on a method that is not relevant to most applied questions (other than the question, "How can I get my noisy study accepted in Psychological Science?").*
> *[...]*
> *Pretty much the main point of teaching the true meaning of p-values would be to convince people not to use them. And that's not really where i want to spend most of my time as a teacher, telling people what not to do. It's just kind of demoralizing for all concerned. Maybe we should be teaching that way, but I think it's a hard sell for teachers and for students alike.*

Gelman:

> *I agree with [Daniel Lakeland's] statement:*
> *I just don't think that "teaching the true meaning of p values" is an important part of the path.*
> *Part of this is just that class time is precious so why waste it on a method that is not relevant to most applied questions (other than the question, "How can I get my noisy study accepted in Psychological Science?").*
> *[...]*
> *Pretty much the main point of teaching the true meaning of p-values would be to convince people not to use them. And that's not really where i want to spend most of my time as a teacher, telling people what not to do. It's just kind of demoralizing for all concerned. Maybe we should be teaching that way, but I think it's a hard sell for teachers and for students alike.*

Peter Dorman:

> *A statistical decision rule is a coordination equilibrium in a very large game with thousands of researchers, journal editors and data users. Perhaps once upon a time such a rule might have been proposed on scientific grounds alone (rightly or wrongly), but now the rule is firmly in place with each use providing an incentive for additional use. That's why my students [...] set aside what I taught in my stats class and embraced NHST. The research they rely on uses it, and the research they hope to produce will be judged by it. That matters a lot more to them than what I think.*
>
> *That's why mass signatures make sense. It is not mob rule in the sociological sense; we signers are not swept up in a wave of transient hysterical solidarity. Rather, we are trying to dent the self-fulfilling power of expectations that locks NHST in place. 800 is too few to do this, alas, but it's worth a try to get this going.*

Peter Dorman:

> *A statistical decision rule is a coordination equilibrium in a very large game with thousands of researchers, journal editors and data users. Perhaps once upon a time such a rule might have been proposed on scientific grounds alone (rightly or wrongly), but now the rule is firmly in place with each use providing an incentive for additional use. That's why my students [...] set aside what I taught in my stats class and embraced NHST. The research they rely on uses it, and the research they hope to produce will be judged by it. That matters a lot more to them than what I think.*
>
> *That's why mass signatures make sense. It is not mob rule in the sociological sense; we signers are not swept up in a wave of transient hysterical solidarity. Rather, we are trying to dent the self-fulfilling power of expectations that locks NHST in place. 800 is too few to do this, alas, but it's worth a try to get this going.*

Peter Dorman:

> *A statistical decision rule is a coordination equilibrium in a very large game with thousands of researchers, journal editors and data users. Perhaps once upon a time such a rule might have been proposed on scientific grounds alone (rightly or wrongly), but now the rule is firmly in place with each use providing an incentive for additional use. That's why my students [...] set aside what I taught in my stats class and embraced NHST. The research they rely on uses it, and the research they hope to produce will be judged by it. That matters a lot more to them than what I think.*
>
> *That's why mass signatures make sense. It is not mob rule in the sociological sense; we signers are not swept up in a wave of transient hysterical solidarity. Rather, we are trying to dent the self-fulfilling power of expectations that locks NHST in place. 800 is too few to do this, alas, but it's worth a try to get this going.*

Discussion comments (Lindgren)

- The statistical landscape is changing; the failure to practically account for the fact that the world of science doesn't obey idealised assumptions is more broadly noticed
- Some of the criticisms of NHST are addressed by formal decision theory
- Prediction vs inference; prediction is not new in statistics, but is often claimed to be a special feature of machine learning
- How to take these issues into account in teaching? Long process, but some change is necessary!