Advanced Data Analysis in Python

p-Values

David Carlson

October 7, 2021

The Definition of a p-Value

- The probability of observing the data or data more extreme given the null hypothesis is true

The Definition of a p-Value

- The probability of observing the data or data more extreme given the null hypothesis is true
- The null hypothesis (H_0) is generally that there is no relationship (correlation) between two variables

The Definition of a p-Value

- The probability of observing the data or data more extreme given the null hypothesis is true
- The null hypothesis (H_0) is generally that there is no relationship (correlation) between two variables
- Example of a null: There is no relationship between civil war onset and ethnic fractionalization

The Definition of a p-Value

- The probability of observing the data or data more extreme given the null hypothesis is true
- The null hypothesis (H_0) is generally that there is no relationship (correlation) between two variables
- Example of a null: There is no relationship between civil war onset and ethnic fractionalization
- The alternative hypothesis (H_1): There exists a relationship

The Definition of a p-Value

- The probability of observing the data or data more extreme given the null hypothesis is true
- The null hypothesis (H_0) is generally that there is no relationship (correlation) between two variables
- Example of a null: There is no relationship between civil war onset and ethnic fractionalization
- The alternative hypothesis (H_1): There exists a relationship
- p-value is used to determine the "significance" of the relationship

The Definition of a p-Value

- The probability of observing the data or data more extreme given the null hypothesis is true
- The null hypothesis (H_0) is generally that there is no relationship (correlation) between two variables
- Example of a null: There is no relationship between civil war onset and ethnic fractionalization
- The alternative hypothesis (H_1): There exists a relationship
- p-value is used to determine the "significance" of the relationship
- The smaller the p-value, the less likely the data comes from the null distribution

Probalistic Modus Tollens (Denying the Consequent)

- If A then B — If H_0 is true then the data will follow an expected pattern

Probalistic Modus Tollens (Denying the Consequent)

- If A then B — If H_0 is true then the data will follow an expected pattern
- Not B observed — The data do not follow the expected pattern

Probalistic Modus Tollens (Denying the Consequent)

- If A then B — If H_0 is true then the data will follow an expected pattern
- Not B observed — The data do not follow the expected pattern
- Therefore not A — Therefore H_0 is false

Probalistic Modus Tollens (Denying the Consequent) (cont.)

- If A then B is highly unlikely — If H_0 is true then the data are highly likely to follow an expected pattern

Probalistic Modus Tollens (Denying the Consequent) (cont.)

- If A then B is highly unlikely — If H_0 is true then the data are highly likely to follow an expected pattern
- Not B observed — The data do not follow the expected pattern

Probalistic Modus Tollens (Denying the Consequent) (cont.)

- If A then B is highly unlikely — If H_0 is true then the data are highly likely to follow an expected pattern
- Not B observed — The data do not follow the expected pattern
- Therefore A is highly unlikely — Therefore H_0 is highly unlikely

Probalistic Modus Tollens (Denying the Consequent) (cont.)

- If A then B is highly unlikely — If a person is an American then it is highly unlikely she is a member of Congress

Probalistic Modus Tollens (Denying the Consequent) (cont.)

- If A then B is highly unlikely — If a person is an American then it is highly unlikely she is a member of Congress
- Not B observed — The person is a member of Congress

Probalistic Modus Tollens (Denying the Consequent) (cont.)

- If A then B is highly unlikely — If a person is an American then it is highly unlikely she is a member of Congress
- Not B observed — The person is a member of Congress
- Therefore A is highly unlikely — Therefore it is highly unlikely she is an American

The Inverse Probability Problem

- Common belief: The smaller the p-value, the greater the probability that the null hypothesis is false

The Inverse Probability Problem

- Common belief: The smaller the p-value, the greater the probability that the null hypothesis is false
- Incorrect interpretation is that null hypothesis significance test produces $p(H_0|D)$

The Inverse Probability Problem

- Common belief: The smaller the p-value, the greater the probability that the null hypothesis is false
- Incorrect interpretation is that null hypothesis significance test produces $p(H_0|D)$
- Instead, produces $p(D|H_0)$

The Inverse Probability Problem

- Common belief: The smaller the p-value, the greater the probability that the null hypothesis is false
- Incorrect interpretation is that null hypothesis significance test produces $p(H_0|D)$
- Instead, produces $p(D|H_0)$
- $p(H_0|D) = \frac{p(H_0)}{p(D)} p(D|H_0)$

Example

- Assume that 2% of the US population are members of some right-wing militia group ($p(M) = 0.02$)

Example

- Assume that 2% of the US population are members of some right-wing militia group ($p(M) = 0.02$)
- A survey is 95% accurate on positive classification ($p(C|M) = 0.95$) and 97% accurate on negative classification ($p(C^C|M^C) = 0.97$)

Example

- Assume that 2% of the US population are members of some right-wing militia group ($p(M) = 0.02$)
- A survey is 95% accurate on positive classification ($p(C|M) = 0.95$) and 97% accurate on negative classification ($p(C^C|M^C) = 0.97$)
- Therefore, $p(M|C) = \frac{p(M)}{p(C)} p(C|M) = 0.38$

Example

- Assume that 2% of the US population are members of some right-wing militia group ($p(M) = 0.02$)
- A survey is 95% accurate on positive classification ($p(C|M) = 0.95$) and 97% accurate on negative classification ($p(C^C|M^C) = 0.97$)
- Therefore, $p(M|C) = \frac{p(M)}{p(C)} p(C|M) = 0.38$
- Probability of correctly classifying an individual as militia member given they are a member is 0.95, yet the probability that a person is a militia member given a positive classification is 0.38

Significance Through Sample Size

- Many have observed that a null hypothesis significance test based on a large sample size almost always results in statistical significance

Significance Through Sample Size

- Many have observed that a null hypothesis significance test based on a large sample size almost always results in statistical significance
- Statistical significance in a large sample study does not imply real world importance

Significance Through Sample Size

- Many have observed that a null hypothesis significance test based on a large sample size almost always results in statistical significance
- Statistical significance in a large sample study does not imply real world importance
- Researchers studying small sample events face substantial prejudice against their empirical findings when alpha levels appropriate to large sample research are applied

Significance Through Sample Size

- Many have observed that a null hypothesis significance test based on a large sample size almost always results in statistical significance
- Statistical significance in a large sample study does not imply real world importance
- Researchers studying small sample events face substantial prejudice against their empirical findings when alpha levels appropriate to large sample research are applied
- As sample size increases we are able to distinguish smaller population-effect sizes progressively

Significance Through Sample Size

- Many have observed that a null hypothesis significance test based on a large sample size almost always results in statistical significance
- Statistical significance in a large sample study does not imply real world importance
- Researchers studying small sample events face substantial prejudice against their empirical findings when alpha levels appropriate to large sample research are applied
- As sample size increases we are able to distinguish smaller population-effect sizes progressively
- Interested in magnitude of effect; making binary decisions about the existence of an effect is not particularly informative

Further Issues

- Arbitrariness of alpha: "... surely God loves .06 nearly as much as .05" (Rosnow and Rosenthal 1989)

Further Issues

- Arbitrariness of alpha: "... surely God loves .06 nearly as much as .05" (Rosnow and Rosenthal 1989)
- Replication fallacy: Probability of replication given a false null is actually the power of the test, not one minus alpha

Further Issues

- Arbitrariness of alpha: "... surely God loves .06 nearly as much as .05" (Rosnow and Rosenthal 1989)
- Replication fallacy: Probability of replication given a false null is actually the power of the test, not one minus alpha
- Asymmetry and accepting the null hypothesis: H_1 is held innocent until proven guilty, H_0 is held guilty until proven innocent

Pervasive Problem

- 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

Pervasive Problem

- 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
- We should also not conclude that two studies conflict because one had a statistically significant result and the other did not

Pervasive Problem

- 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
- We should also not 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

Example

- Consider a series of analyses of unintended effects of anti-inflammatory drugs

Example

- Consider a series of analyses of unintended effects of anti-inflammatory drugs
- Because their results were statistically non-significant, one set of researchers concluded that exposure to the drugs was "not associated" with new-onset atrial fibrillation (the most common disturbance to heart rhythm)

Example

- Consider a series of analyses of unintended effects of anti-inflammatory drugs
- Because their results were statistically non-significant, one set of researchers concluded that exposure to the drugs was "not associated" with new-onset atrial fibrillation (the most common disturbance to heart rhythm)
- Argued that the results stood in contrast to those from an earlier study with a statistically significant outcome

Example (cont.)

- The researchers describing their statistically non-significant results found a risk ratio of 1.2 (that is, a 20% greater risk in exposed patients relative to unexposed ones)

Example (cont.)

- The researchers describing their statistically non-significant results found a risk ratio of 1.2 (that is, a 20% greater risk in exposed patients relative to unexposed ones)
- They also found a 95% confidence interval that spanned everything from a trifling risk decrease of 3% to a considerable risk increase of 48% ($p = 0.091$)

Example (cont.)

- The researchers describing their statistically non-significant results found a risk ratio of 1.2 (that is, a 20% greater risk in exposed patients relative to unexposed ones)
- They also found a 95% confidence interval that spanned everything from a trifling risk decrease of 3% to a considerable risk increase of 48% ($p = 0.091$)
- The researchers from the earlier, statistically significant, study found the exact same risk ratio of 1.2

Example (cont.)

- The researchers describing their statistically non-significant results found a risk ratio of 1.2 (that is, a 20% greater risk in exposed patients relative to unexposed ones)
- They also found a 95% confidence interval that spanned everything from a trifling risk decrease of 3% to a considerable risk increase of 48% ($p = 0.091$)
- The researchers from the earlier, statistically significant, study found the exact same risk ratio of 1.2
- That study was simply more precise, with an interval spanning from 9% to 33% greater risk ($p = 0.0003$)

Example (cont.)

- Ludicrous to conclude that the statistically non-significant results showed "no association"

Example (cont.)

- Ludicrous to conclude that the statistically non-significant results showed "no association"
- Interval estimate included serious risk increases

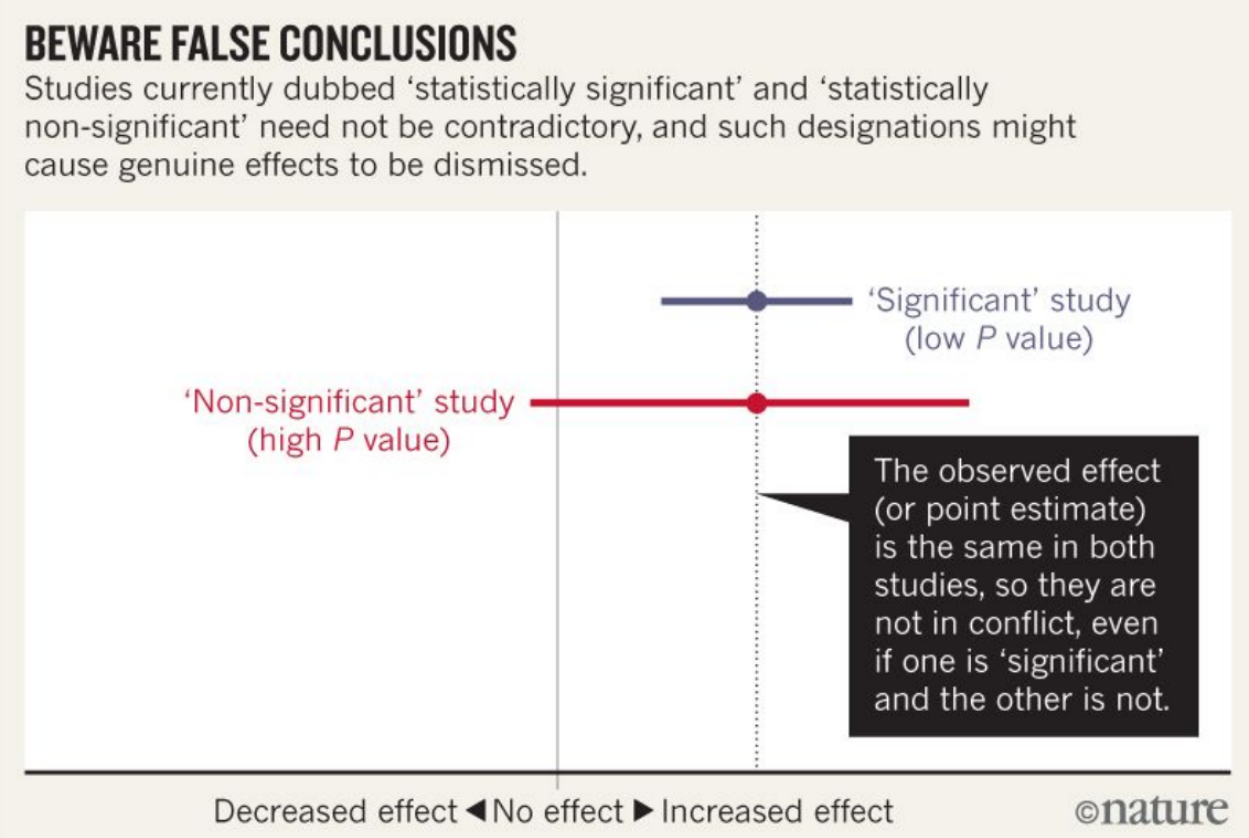
Example (cont.)

- Ludicrous to conclude that the statistically non-significant results showed "no association"
- Interval estimate included serious risk increases
- Equally absurd to claim these results were in contrast with the earlier results showing an identical observed effect

Example (cont.)

- Ludicrous to conclude that the statistically non-significant results showed "no association"
- Interval estimate included serious risk increases
- Equally absurd to claim these results were in contrast with the earlier results showing an identical observed effect
- Reliance on thresholds of statistical significance can mislead us

Example (cont.)



The Insignificance of Null Hypothesis Significance Testing

- Context: This article written 2 decades ago

The Insignificance of Null Hypothesis Significance Testing

- Context: This article written 2 decades ago
- Led in the social sciences by psychology, many are challenging the basic tenets of the way that nearly all social scientists are trained to develop and test empirical hypotheses

The Insignificance of Null Hypothesis Significance Testing

- Context: This article written 2 decades ago
- Led in the social sciences by psychology, many are challenging the basic tenets of the way that nearly all social scientists are trained to develop and test empirical hypotheses
- "Strangle-hold" (Rozenboom 1960)

The Insignificance of Null Hypothesis Significance Testing

- Context: This article written 2 decades ago
- Led in the social sciences by psychology, many are challenging the basic tenets of the way that nearly all social scientists are trained to develop and test empirical hypotheses
- "Strangle-hold" (Rozenboom 1960)
- "Deeply flawed or else ill-used by researchers" (Serlin and Lapsley 1993)

The Insignificance of Null Hypothesis Significance Testing

- Context: This article written 2 decades ago
- Led in the social sciences by psychology, many are challenging the basic tenets of the way that nearly all social scientists are trained to develop and test empirical hypotheses
- "Strangle-hold" (Rozenboom 1960)
- "Deeply flawed or else ill-used by researchers" (Serlin and Lapsley 1993)
- "A terrible mistake, basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology" (Meehl 1978)

The Insignificance of Null Hypothesis Significance Testing (cont.)

- "An instance of the kind of mindlessness in the conduct of research" (Bakan 1960)

The Insignificance of Null Hypothesis Significance Testing (cont.)

- "An instance of the kind of mindlessness in the conduct of research" (Bakan 1960)
- "Badly misused for a long time" (Cohen 1994)

The Insignificance of Null Hypothesis Significance Testing (cont.)

- "An instance of the kind of mindlessness in the conduct of research" (Bakan 1960)
- "Badly misused for a long time" (Cohen 1994)
- "Systematically retarded the growth of cumulative knowledge" (Schmidt 1996)

The Insignificance of Null Hypothesis Significance Testing (cont.)

- "An instance of the kind of mindlessness in the conduct of research" (Bakan 1960)
- "Badly misused for a long time" (Cohen 1994)
- "Systematically retarded the growth of cumulative knowledge" (Schmidt 1996)
- "The significance test as it is currently used in the social sciences just does not work" (Hunter 1997)

Pervasiveness

- These and similar errors are widespread

Pervasiveness

- These and similar errors are widespread
- Surveys of hundreds of articles have found that statistically non-significant results are interpreted as indicating 'no difference' or 'no effect' in around half

Pervasiveness

- These and similar errors are widespread
- Surveys of hundreds of articles have found that statistically non-significant results are interpreted as indicating 'no difference' or 'no effect' in around half
- In 2016, the American Statistical Association released a statement in *The American Statistician* warning against the misuse of statistical significance and p-values

Pervasiveness

- These and similar errors are widespread
- Surveys of hundreds of articles have found that statistically non-significant results are interpreted as indicating 'no difference' or 'no effect' in around half
- In 2016, the American Statistical Association released a statement in *The American Statistician* warning against the misuse of statistical significance and p-values
- A special issue presents more than 40 papers on 'Statistical inference in the 21st century: a world beyond $p < 0.05$', cautioning "don't say 'statistically significant"'

Pervasiveness

- These and similar errors are widespread
- Surveys of hundreds of articles have found that statistically non-significant results are interpreted as indicating 'no difference' or 'no effect' in around half
- In 2016, the American Statistical Association released a statement in *The American Statistician* warning against the misuse of statistical significance and *p*-values
- A special issue presents more than 40 papers on 'Statistical inference in the 21st century: a world beyond $p < 0.05$', cautioning "don't say 'statistically significant'"
- The authors call for the entire concept of statistical significance to be abandoned

Receptivity

- Authors invited others to read a draft of this comment and sign their names if they concurred with the message

Receptivity

- Authors invited others to read a draft of this comment and sign their names if they concurred with the message
- 250 did so within the first 24 hours

Receptivity

- Authors invited others to read a draft of this comment and sign their names if they concurred with the message
- 250 did so within the first 24 hours
- A week later, more than 800 signatories — all checked for an academic affiliation or other indication of present or past work in a field that depends on statistical modelling

Receptivity

- Authors invited others to read a draft of this comment and sign their names if they concurred with the message
- 250 did so within the first 24 hours
- A week later, more than 800 signatories — all checked for an academic affiliation or other indication of present or past work in a field that depends on statistical modelling
- More than 50 countries and across all continents except Antarctica

Receptivity

- Authors invited others to read a draft of this comment and sign their names if they concurred with the message
- 250 did so within the first 24 hours
- A week later, more than 800 signatories — all checked for an academic affiliation or other indication of present or past work in a field that depends on statistical modelling
- More than 50 countries and across all continents except Antarctica
- One advocate called it a "surgical strike against thoughtless testing of statistical significance" and "an opportunity to register your voice in favour of better scientific practices"

Recommendations

- Use confidence intervals; more informative

Recommendations

- Use confidence intervals; more informative
- Bayesian analysis; calculate $p(\theta|D, H_0)$

Recommendations

- Use confidence intervals; more informative
- Bayesian analysis; calculate $p(\theta|D, H_0)$
- Meta-analysis: Offers attractive proposition that the accumulation of knowledge on some research question can be compared and combined in a single procedure

Discussion Questions

- This discussion is not new; why have we been unable to change practices?

Discussion Questions

- This discussion is not new; why have we been unable to change practices?
- Should there be a decision criteria; what would it be?

Discussion Questions

- This discussion is not new; why have we been unable to change practices?
- Should there be a decision criteria; what would it be?
- Should we care more about p-values or magnitudes?

Discussion Questions

- This discussion is not new; why have we been unable to change practices?
- Should there be a decision criteria; what would it be?
- Should we care more about p-values or magnitudes?
- Why is the null hypothesis significance test pervasive?

Discussion Questions

- This discussion is not new; why have we been unable to change practices?
- Should there be a decision criteria; what would it be?
- Should we care more about p-values or magnitudes?
- Why is the null hypothesis significance test pervasive?
- Have you seen a study that might speak to these issues?

Discussion Questions

- This discussion is not new; why have we been unable to change practices?
- Should there be a decision criteria; what would it be?
- Should we care more about p-values or magnitudes?
- Why is the null hypothesis significance test pervasive?
- Have you seen a study that might speak to these issues?
- Are there any alternatives not mentioned that you can think of?

Pre-registration

- The pre-registration of studies and a commitment to publish all results of all analyses can do much to mitigate these issues

Pre-registration

- The pre-registration of studies and a commitment to publish all results of all analyses can do much to mitigate these issues
- Even results from pre-registered studies can be biased by decisions invariably left open in the analysis plan

Pre-registration

- The pre-registration of studies and a commitment to publish all results of all analyses can do much to mitigate these issues
- Even results from pre-registered studies can be biased by decisions invariably left open in the analysis plan
- This occurs even with the best of intentions

Avoiding Dichotomous Engagement

- All statistics, including p-values and confidence intervals, naturally vary from study to study, and often do so to a surprising degree

Avoiding Dichotomous Engagement

- All statistics, including p-values and confidence intervals, naturally vary from study to study, and often do so to a surprising degree
- Random variation alone can easily lead to large disparities in p-values, far beyond falling just to either side of the 0.05 threshold

Avoiding Dichotomous Engagement

- All statistics, including p-values and confidence intervals, naturally vary from study to study, and often do so to a surprising degree
- Random variation alone can easily lead to large disparities in p-values, far beyond falling just to either side of the 0.05 threshold
- Even if researchers could conduct two perfect replication studies of some genuine effect, each with 80% power (chance) of achieving $p < 0.05$, it would not be very surprising for one to obtain $p < 0.01$ and the other $p > 0.30$

Avoiding Dichotomous Engagement

- All statistics, including p-values and confidence intervals, naturally vary from study to study, and often do so to a surprising degree
- Random variation alone can easily lead to large disparities in p-values, far beyond falling just to either side of the 0.05 threshold
- Even if researchers could conduct two perfect replication studies of some genuine effect, each with 80% power (chance) of achieving $p < 0.05$, it would not be very surprising for one to obtain $p < 0.01$ and the other $p > 0.30$
- Whether a p-value is small or large, caution is warranted

Uncertainty

- We must learn to embrace uncertainty

Uncertainty

- We must learn to embrace uncertainty
- One practical way to do so is to rename confidence intervals as 'compatibility intervals' and interpret them in a way that avoids overconfidence

Uncertainty

- We must learn to embrace uncertainty
- One practical way to do so is to rename confidence intervals as 'compatibility intervals' and interpret them in a way that avoids overconfidence
- Recommend that authors describe the practical implications of all values inside the interval, especially the observed effect (or point estimate) and the limits

Uncertainty

- We must learn to embrace uncertainty
- One practical way to do so is to rename confidence intervals as 'compatibility intervals' and interpret them in a way that avoids overconfidence
- Recommend that authors describe the practical implications of all values inside the interval, especially the observed effect (or point estimate) and the limits
- Remember that all the values between the interval's limits are reasonably compatible with the data, given the statistical assumptions used to compute the interval

Uncertainty

- We must learn to embrace uncertainty
- One practical way to do so is to rename confidence intervals as 'compatibility intervals' and interpret them in a way that avoids overconfidence
- Recommend that authors describe the practical implications of all values inside the interval, especially the observed effect (or point estimate) and the limits
- Remember that all the values between the interval's limits are reasonably compatible with the data, given the statistical assumptions used to compute the interval
- Singling out one particular value (such as the null value) in the interval as 'shown' makes no sense

Uncertainty

- We must learn to embrace uncertainty
- One practical way to do so is to rename confidence intervals as 'compatibility intervals' and interpret them in a way that avoids overconfidence
- Recommend that authors describe the practical implications of all values inside the interval, especially the observed effect (or point estimate) and the limits
- Remember that all the values between the interval's limits are reasonably compatible with the data, given the statistical assumptions used to compute the interval
- Singling out one particular value (such as the null value) in the interval as 'shown' makes no sense
- An interval that contains the null value will often also contain non-null values of high practical importance

Compatibility Intervals

- Just because the interval gives the values most compatible with the data, given the assumptions, it doesnt mean values outside it are incompatible; they are just less compatible

Compatibility Intervals

- Just because the interval gives the values most compatible with the data, given the assumptions, it doesnt mean values outside it are incompatible; they are just less compatible
- Values just outside the interval do not differ substantively from those just inside the interval

Compatibility Intervals

- Just because the interval gives the values most compatible with the data, given the assumptions, it doesnt mean values outside it are incompatible; they are just less compatible
- Values just outside the interval do not differ substantively from those just inside the interval
- It is thus wrong to claim that an interval shows all possible values

Compatibility Intervals (cont.)

- Not all values inside are equally compatible with the data, given the assumptions

Compatibility Intervals (cont.)

- Not all values inside are equally compatible with the data, given the assumptions
- The point estimate is the most compatible, and values near it are more compatible than those near the limits

Compatibility Intervals (cont.)

- Not all values inside are equally compatible with the data, given the assumptions
- The point estimate is the most compatible, and values near it are more compatible than those near the limits
- Discuss the point estimate, even when they have a large p-value or a wide interval, as well as discussing the limits of that interval

Compatibility Intervals (cont.)

- Not all values inside are equally compatible with the data, given the assumptions
- The point estimate is the most compatible, and values near it are more compatible than those near the limits
- Discuss the point estimate, even when they have a large p-value or a wide interval, as well as discussing the limits of that interval
- The authors above could have written: 'Like a previous study, our results suggest a 20% increase in risk of new-onset atrial fibrillation in patients given the anti-inflammatory drugs. Nonetheless, a risk difference ranging from a 3% decrease, a small negative association, to a 48% increase, a substantial positive association, is also reasonably compatible with our data, given our assumptions.'

Compatibility Intervals (cont.)

- Not all values inside are equally compatible with the data, given the assumptions
- The point estimate is the most compatible, and values near it are more compatible than those near the limits
- Discuss the point estimate, even when they have a large p-value or a wide interval, as well as discussing the limits of that interval
- The authors above could have written: 'Like a previous study, our results suggest a 20% increase in risk of new-onset atrial fibrillation in patients given the anti-inflammatory drugs. Nonetheless, a risk difference ranging from a 3% decrease, a small negative association, to a 48% increase, a substantial positive association, is also reasonably compatible with our data, given our assumptions.'
- Interpreting the point estimate, while acknowledging its uncertainty, will keep you from making false declarations of 'no difference,' and from making overconfident claims

Compatibility Intervals (cont.)

- Like the 0.05 threshold from which it came, the default 95% used to compute intervals is itself an arbitrary convention

Compatibility Intervals (cont.)

- Like the 0.05 threshold from which it came, the default 95% used to compute intervals is itself an arbitrary convention
- It is based on the false idea that there is a 95% chance that the computed interval itself contains the true value, coupled with the vague feeling that this is a basis for a confident decision

Compatibility Intervals (cont.)

- Like the 0.05 threshold from which it came, the default 95% used to compute intervals is itself an arbitrary convention
- It is based on the false idea that there is a 95% chance that the computed interval itself contains the true value, coupled with the vague feeling that this is a basis for a confident decision
- Interval estimates can perpetuate the problems of statistical significance when the dichotomization they impose is treated as a scientific standard

Compatibility Intervals (cont.)

- Be humble: compatibility assessments hinge on the correctness of the statistical assumptions used to compute the interval

Compatibility Intervals (cont.)

- Be humble: compatibility assessments hinge on the correctness of the statistical assumptions used to compute the interval
- In practice, these assumptions are at best subject to considerable uncertainty

Compatibility Intervals (cont.)

- Be humble: compatibility assessments hinge on the correctness of the statistical assumptions used to compute the interval
- In practice, these assumptions are at best subject to considerable uncertainty
- Make these assumptions as clear as possible and test the ones you can, for example by plotting your data and by fitting alternative models, and then reporting all results

Compatibility Intervals (cont.)

- Be humble: compatibility assessments hinge on the correctness of the statistical assumptions used to compute the interval
- In practice, these assumptions are at best subject to considerable uncertainty
- Make these assumptions as clear as possible and test the ones you can, for example by plotting your data and by fitting alternative models, and then reporting all results
- Whatever the statistics show, it is fine to suggest reasons for your results, but discuss a range of potential explanations, not just favoured ones

Compatibility Intervals (cont.)

- Be humble: compatibility assessments hinge on the correctness of the statistical assumptions used to compute the interval
- In practice, these assumptions are at best subject to considerable uncertainty
- Make these assumptions as clear as possible and test the ones you can, for example by plotting your data and by fitting alternative models, and then reporting all results
- Whatever the statistics show, it is fine to suggest reasons for your results, but discuss a range of potential explanations, not just favoured ones
- Inferences should be scientific, and that goes far beyond the merely statistical

Compatibility Intervals (cont.)

- Be humble: compatibility assessments hinge on the correctness of the statistical assumptions used to compute the interval
- In practice, these assumptions are at best subject to considerable uncertainty
- Make these assumptions as clear as possible and test the ones you can, for example by plotting your data and by fitting alternative models, and then reporting all results
- Whatever the statistics show, it is fine to suggest reasons for your results, but discuss a range of potential explanations, not just favoured ones
- Inferences should be scientific, and that goes far beyond the merely statistical
- Factors such as background evidence, study design, data quality and understanding of underlying mechanisms are often more important than statistical measures such as p-values or intervals