COGS 303

Gary Neels

UBC

Nov 3, 2023

Presentations

Presentations begin in two weeks (Nov 17, Nov 24, and Dec 1):

- If you haven't signed up for a presentation date yet, you should do so soon

Presentations

Presentations begin in two weeks (Nov 17, Nov 24, and Dec 1):

- If you haven't signed up for a presentation date yet, you should do so soon
- 8-10 minutes, with 2-4 minutes for Q & A

Presentations

Presentations begin in two weeks (Nov 17, Nov 24, and Dec 1):

- If you haven't signed up for a presentation date yet, you should do so soon
- 8-10 minutes, with 2-4 minutes for Q & A
- The standard is "work in progress"

Presentations

Presentations begin in two weeks (Nov 17, Nov 24, and Dec 1):

- If you haven't signed up for a presentation date yet, you should do so soon
- 8-10 minutes, with 2-4 minutes for Q & A
- The standard is "work in progress"
-

Component	Criteria	Points
Content	-Clear -Well-organized -Informative	7
Speech	-Correct length (8-10 minutes) -Polished -Flows well	4
Slides	-Relevant -Interesting -Aids the presentation	4

Quick recap from last week

We talked about the "replication crisis"

- How are we doing with respect to replicating scientific results?

Quick recap from last week

We talked about the “replication crisis”

- How are we doing with respect to replicating scientific results?
- Not great!

Quick recap from last week

We talked about the "replication crisis"

- How are we doing with respect to replicating scientific results?
- Not great!
- Does this mean the unreplicated results are false?

Quick recap from last week

We talked about the "replication crisis"

- How are we doing with respect to replicating scientific results?
- Not great!
- Does this mean the unreplicated results are false?
- Not necessarily, but it should undermine our confidence

Quick recap from last week

We talked about the “replication crisis”

- How are we doing with respect to replicating scientific results?
- Not great!
- Does this mean the unreplicated results are false?
- Not necessarily, but it should undermine our confidence
- So, what needs to change in order to improve the situation?

Sources of the problem

We identified two broad categories of sources of this problem:

- Systemic problems

Sources of the problem

We identified two broad categories of sources of this problem:

- Systemic problems
    - Publication bias

Sources of the problem

We identified two broad categories of sources of this problem:

- Systemic problems
  - Publication bias
  - **Over-reliance on statistical significance tests**

Sources of the problem

We identified two broad categories of sources of this problem:

- Systemic problems
  - Publication bias
  - **Over-reliance on statistical significance tests**
  - The incentive structure to publish innovative results

Sources of the problem

We identified two broad categories of sources of this problem:

- Systemic problems
  - Publication bias
  - **Over-reliance on statistical significance tests**
  - The incentive structure to publish innovative results
- Individual problems (Questionable Research Practices)

Sources of the problem

We identified two broad categories of sources of this problem:

- Systemic problems
  - Publication bias
  - **Over-reliance on statistical significance tests**
  - The incentive structure to publish innovative results
- Individual problems (Questionable Research Practices)
  - **P-hacking**

Sources of the problem

We identified two broad categories of sources of this problem:

- Systemic problems
    - Publication bias
    - **Over-reliance on statistical significance tests**
    - The incentive structure to publish innovative results
- Individual problems (Questionable Research Practices)
    - **P-hacking**
    - HARKing

Sources of the problem

We identified two broad categories of sources of this problem:

- Systemic problems
    - Publication bias
    - **Over-reliance on statistical significance tests**
    - The incentive structure to publish innovative results
- Individual problems (Questionable Research Practices)
    - **P-hacking**
    - HARKing
    - Outright fraud

Recap of Systemic Problems

- Publication bias

Recap of Systemic Problems

- Publication bias
  - Journals refusing to publish replication studies

Recap of Systemic Problems

- Publication bias
  - Journals refusing to publish replication studies
  - What's the incentive to perform a replication study if you can't get it published?

Recap of Systemic Problems

- Publication bias
  - Journals refusing to publish replication studies
  - What's the incentive to perform a replication study if you can't get it published?
- Bias toward innovation rather than rigorously testing current ideas

Recap of Systemic Problems

- Publication bias
  - Journals refusing to publish replication studies
  - What's the incentive to perform a replication study if you can't get it published?
- Bias toward innovation rather than rigorously testing current ideas
  - This suggests overconfidence in our current theories

Recap of Systemic Problems

- Publication bias
  - Journals refusing to publish replication studies
  - What's the incentive to perform a replication study if you can't get it published?
- Bias toward innovation rather than rigorously testing current ideas
  - This suggests overconfidence in our current theories
  - Innovation is exciting, but empirical testing is how we tell if our theories are likely

Recap of Individual Problems

- HARKing–Hypothesizing After Results are Known

Recap of Individual Problems

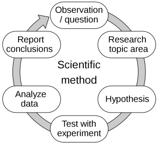
- HARKing–Hypothesizing After Results are Known
- Kerr told us HARKing was bad because prediction is more valuable than accommodation

Recap of Individual Problems

- HARKing–Hypothesizing After Results are Known
- Kerr told us HARKing was bad because prediction is more valuable than accommodation
- Is he right?

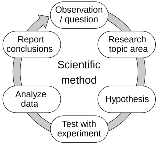
Recap of Individual Problems

- HARKing–Hypothesizing After Results are Known
- Kerr told us HARKing was bad because prediction is more valuable than accommodation
- Is he right?



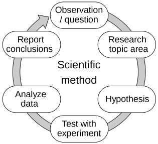
-

Recap of Individual Problems

- HARKing–Hypothesizing After Results are Known
- Kerr told us HARKing was bad because prediction is more valuable than accommodation
- Is he right?
- 

- Isn't all hypothesizing done after (at least some) results are known?

Recap of Individual Problems

- HARKing–Hypothesizing After Results are Known
- Kerr told us HARKing was bad because prediction is more valuable than accommodation
- Is he right?
- 

- Isn't all hypothesizing done after (at least some) results are known?
- The issue with HARKing is when the HARKed hypothesis is presented as if it has been tested and confirmed rather than merely suggested by the result of the test

More on HARKing

So, how do we prevent the bad sort of HARKing?

- Note how HARKing can lead to poor replication–if the HARKed hypothesis is just a quirk of the sample, that connection won't be seen in subsequent tests

More on HARKing

So, how do we prevent the bad sort of HARKing?

- Note how HARKing can lead to poor replication–if the HARKed hypothesis is just a quirk of the sample, that connection won't be seen in subsequent tests
- One innovation that has been put in place is "pre-registration"

More on HARKing

So, how do we prevent the bad sort of HARKing?

- Note how HARKing can lead to poor replication–if the HARKed hypothesis is just a quirk of the sample, that connection won't be seen in subsequent tests
- One innovation that has been put in place is "pre-registration"
- This practice fixes a few of the problems we've identified:

More on HARKing

So, how do we prevent the bad sort of HARKing?

- Note how HARKing can lead to poor replication–if the HARKed hypothesis is just a quirk of the sample, that connection won't be seen in subsequent tests
- One innovation that has been put in place is "pre-registration"
- This practice fixes a few of the problems we've identified:
  - The study is accepted for publication prior to the test being done–it's going to be published regardless of the result

More on HARKing

So, how do we prevent the bad sort of HARKing?

- Note how HARKing can lead to poor replication–if the HARKed hypothesis is just a quirk of the sample, that connection won't be seen in subsequent tests
- One innovation that has been put in place is "pre-registration"
- This practice fixes a few of the problems we've identified:
  - The study is accepted for publication prior to the test being done–it's going to be published regardless of the result
  - This removes the incentive/need to have some ground-breaking innovation to get published

More on HARKing

So, how do we prevent the bad sort of HARKing?

- Note how HARKing can lead to poor replication–if the HARKed hypothesis is just a quirk of the sample, that connection won't be seen in subsequent tests
- One innovation that has been put in place is "pre-registration"
- This practice fixes a few of the problems we've identified:
  - The study is accepted for publication prior to the test being done–it's going to be published regardless of the result
  - This removes the incentive/need to have some ground-breaking innovation to get published
  - It also means you can't revision the purpose of the test (as happens in HARKing)

More on HARKing

So, how do we prevent the bad sort of HARKing?

- Note how HARKing can lead to poor replication–if the HARKed hypothesis is just a quirk of the sample, that connection won't be seen in subsequent tests
- One innovation that has been put in place is "pre-registration"
- This practice fixes a few of the problems we've identified:
  - The study is accepted for publication prior to the test being done–it's going to be published regardless of the result
  - This removes the incentive/need to have some ground-breaking innovation to get published
  - It also means you can't revision the purpose of the test (as happens in HARKing)
  - If your test has a surprising result that suggests a new hypothesis, you can present that as a candidate for future testing in your conclusion (but not as having been confirmed by the test)

There are a number of connections between Statistical Analysis and the Replication Crisis.

- Some suggest that there is an over-reliance on statistical significance tests

There are a number of connections between Statistical Analysis and the Replication Crisis.

- Some suggest that there is an over-reliance on statistical significance tests
- Some suggest that the convention of treating $p < 0.05$ as the threshold for significance is too lax, and that $p < 0.01$ is more appropriate

There are a number of connections between Statistical Analysis and the Replication Crisis.

- Some suggest that there is an over-reliance on statistical significance tests
- Some suggest that the convention of treating $p < 0.05$ as the threshold for significance is too lax, and that $p < 0.01$ is more appropriate
- P-hacking

There are a number of connections between Statistical Analysis and the Replication Crisis.

- Some suggest that there is an over-reliance on statistical significance tests
- Some suggest that the convention of treating $p < 0.05$ as the threshold for significance is too lax, and that $p < 0.01$ is more appropriate
- P-hacking
- Some suggest that frequentist/classical statistical methods should be replaced by Bayesian methods

Bayesian vs Frequentist

Recall from Week 4 the distinction we made between objective and subjective interpretations of probability

- These different understandings of the meaning of probability lead to different methodologies with using statistical information in scientific reasoning

Bayesian vs Frequentist

Recall from Week 4 the distinction we made between objective and subjective interpretations of probability

- These different understandings of the meaning of probability lead to different methodologies with using statistical information in scientific reasoning
- Frequentists favour the objective interpretation

Bayesian vs Frequentist

Recall from Week 4 the distinction we made between objective and subjective interpretations of probability

- These different understandings of the meaning of probability lead to different methodologies with using statistical information in scientific reasoning
- Frequentists favour the objective interpretation
- Bayesians favour the subjective interpretation

Bayesian vs Frequentist

Recall from Week 4 the distinction we made between objective and subjective interpretations of probability

- These different understandings of the meaning of probability lead to different methodologies with using statistical information in scientific reasoning
- Frequentists favour the objective interpretation
- Bayesians favour the subjective interpretation
- Let's illustrate with an example

Suppose I'm about to flip a coin

- What's the probability of the toss resulting in heads?
  - Bayesian–0.5
  - Frequentist–0.5

Now suppose I have already flipped the coin, but haven't revealed the result to you...what's the probability of heads?

- Bayesian–0.5

Now suppose I have already flipped the coin, but haven't revealed the result to you...what's the probability of heads?

- Bayesian–0.5
- Frequentist–???

Now suppose I have already flipped the coin, but haven't revealed the result to you...what's the probability of heads?

- Bayesian–0.5
- Frequentist–???
- This question doesn't make sense to the frequentist

Now suppose I have already flipped the coin, but haven't revealed the result to you...what's the probability of heads?

- Bayesian–0.5
- Frequentist–???
- This question doesn't make sense to the frequentist
- It is what it is. The result is fixed

So, what do Bayesians and Frequentists mean when they say the probability is 0.5?

- Frequentist

So, what do Bayesians and Frequentists mean when they say the probability is 0.5?

- Frequentist
    - In the long run, repeating this type of exercise results in heads roughly 50% of the time

So, what do Bayesians and Frequentists mean when they say the probability is 0.5?

- Frequentist
  - In the long run, repeating this type of exercise results in heads roughly 50% of the time
  - What is "this type of exercise"?

So, what do Bayesians and Frequentists mean when they say the probability is 0.5?

- Frequentist
  - In the long run, repeating this type of exercise results in heads roughly 50% of the time
  - What is “this type of exercise”?
  - General coin flips (recall the discussion a few weeks ago about reference classes)

So, what do Bayesians and Frequentists mean when they say the probability is 0.5?

- Frequentist
  - In the long run, repeating this type of exercise results in heads roughly 50% of the time
  - What is "this type of exercise"?
  - General coin flips (recall the discussion a few weeks ago about reference classes)
- Bayesian

So, what do Bayesians and Frequentists mean when they say the probability is 0.5?

- Frequentist
  - In the long run, repeating this type of exercise results in heads roughly 50% of the time
  - What is “this type of exercise”?
  - General coin flips (recall the discussion a few weeks ago about reference classes)
- Bayesian
  - Here the probability is a reasonable expectation

So, what do Bayesians and Frequentists mean when they say the probability is 0.5?

- Frequentist
  - In the long run, repeating this type of exercise results in heads roughly 50% of the time
  - What is "this type of exercise"?
  - General coin flips (recall the discussion a few weeks ago about reference classes)
- Bayesian
  - Here the probability is a reasonable expectation
  - $\frac{heads}{possibilities}$ assuming the possibilities are equally likely

So, what do Bayesians and Frequentists mean when they say the probability is 0.5?

- Frequentist
  - In the long run, repeating this type of exercise results in heads roughly 50% of the time
  - What is "this type of exercise"?
  - General coin flips (recall the discussion a few weeks ago about reference classes)
- Bayesian
  - Here the probability is a reasonable expectation
  - $\frac{heads}{possibilities}$ assuming the possibilities are equally likely
- Where the Frequentist focusses on the result of repeating a type of exercise, the Bayesian focusses on the uncertainty involved in the exercise

These philosophical differences in understanding probabilities lead to differences in how statistical methods are used in research

- Bayesians and frequentists go about estimating parameters differently

These philosophical differences in understanding probabilities lead to differences in how statistical methods are used in research

- Bayesians and frequentists go about estimating parameters differently
  - For example, estimating the mean height of a population

These philosophical differences in understanding probabilities lead to differences in how statistical methods are used in research

- Bayesians and frequentists go about estimating parameters differently
    - For example, estimating the mean height of a population
- Bayesians and frequentists go about comparing hypotheses differently

These philosophical differences in understanding probabilities lead to differences in how statistical methods are used in research

- Bayesians and frequentists go about estimating parameters differently
  - For example, estimating the mean height of a population
- Bayesians and frequentists go about comparing hypotheses differently
  - Frequentists use significance tests

These philosophical differences in understanding probabilities lead to differences in how statistical methods are used in research

- Bayesians and frequentists go about estimating parameters differently
  - For example, estimating the mean height of a population
- Bayesians and frequentists go about comparing hypotheses differently
  - Frequentists use significance tests
  - Bayesians consider likelihood ratios

These philosophical differences in understanding probabilities lead to differences in how statistical methods are used in research

- Bayesians and frequentists go about estimating parameters differently
  - For example, estimating the mean height of a population
- Bayesians and frequentists go about comparing hypotheses differently
  - Frequentists use significance tests
  - Bayesians consider likelihood ratios
- Let's consider these in turn

Example–mean height of a population

Suppose we are interested in figuring out the mean height of a population. As Howson and Urbach mention, Bayesians and Frequentists go about this differently:

- Frequentists treat this unknown value as a **fixed** (ie non-random) quantity (call it μ)

Suppose we are interested in figuring out the mean height of a population. As Howson and Urbach mention, Bayesians and Frequentists go about this differently:

- Frequentists treat this unknown value as a **fixed** (ie non-random) quantity (call it μ)
- Fixed–it is what it is, like the coin that's been flipped

Suppose we are interested in figuring out the mean height of a population. As Howson and Urbach mention, Bayesians and Frequentists go about this differently:

- Frequentists treat this unknown value as a **fixed** (ie non-random) quantity (call it μ)
- Fixed–it is what it is, like the coin that's been flipped
- To estimate this value, we take a random sample of the population

Suppose we are interested in figuring out the mean height of a population. As Howson and Urbach mention, Bayesians and Frequentists go about this differently:

- Frequentists treat this unknown value as a **fixed** (ie non-random) quantity (call it μ)
- Fixed–it is what it is, like the coin that's been flipped
- To estimate this value, we take a random sample of the population
- The mean height of the sample group is a **random** variable

Suppose we are interested in figuring out the mean height of a population. As Howson and Urbach mention, Bayesians and Frequentists go about this differently:

- Frequentists treat this unknown value as a **fixed** (ie non-random) quantity (call it μ)
- Fixed–it is what it is, like the coin that's been flipped
- To estimate this value, we take a random sample of the population
- The mean height of the sample group is a **random** variable
- We then take it that $\mu = m$

Suppose we are interested in figuring out the mean height of a population. As Howson and Urbach mention, Bayesians and Frequentists go about this differently:

- Frequentists treat this unknown value as a **fixed** (ie non-random) quantity (call it μ)
- Fixed–it is what it is, like the coin that's been flipped
- To estimate this value, we take a random sample of the population
- The mean height of the sample group is a **random** variable
- We then take it that $\mu = m$
- There is, of course, uncertainty here...how confident should we be that our estimate is correct?

Suppose we are interested in figuring out the mean height of a population. As Howson and Urbach mention, Bayesians and Frequentists go about this differently:

- Frequentists treat this unknown value as a **fixed** (ie non-random) quantity (call it μ)
- Fixed–it is what it is, like the coin that's been flipped
- To estimate this value, we take a random sample of the population
- The mean height of the sample group is a **random** variable
- We then take it that $\mu = m$
- There is, of course, uncertainty here...how confident should we be that our estimate is correct?
- To answer this, frequentists point to the "confidence interval"

Confidence interval

- The confidence interval is a purely mathematical (objective) value

Confidence interval

- The confidence interval is a purely mathematical (objective) value
- Suppose we target a 95% level

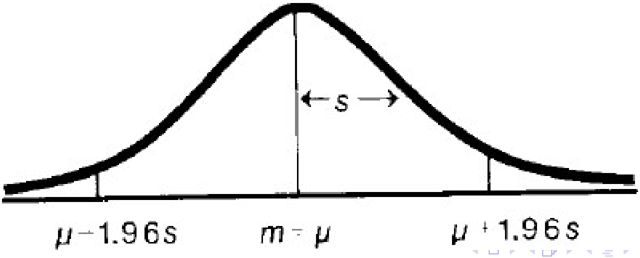
Confidence interval

- The confidence interval is a purely mathematical (objective) value
- Suppose we target a 95% level
- Let σ denote the standard deviation of heights in the population

Confidence interval

- The confidence interval is a purely mathematical (objective) value
- Suppose we target a 95% level
- Let σ denote the standard deviation of heights in the population
- We can use that value to calculate the standard deviation of the sample $s = \frac{\sigma}{\sqrt{n}}$ (where n is the sample size)

Confidence interval

- The confidence interval is a purely mathematical (objective) value
- Suppose we target a 95% level
- Let σ denote the standard deviation of heights in the population
- We can use that value to calculate the standard deviation of the sample $s = \frac{\sigma}{\sqrt{n}}$ (where n is the sample size)
- 

Is this confidence interval really objective?

- Howson & Urbach argue that it is not

Is this confidence interval really objective?

- Howson & Urbach argue that it is not
- First, 95% confidence interval does not mean that the probability of μ being in the range is 0.95 (because μ is not a random variable and therefore has no probability)

Is this confidence interval really objective?

- Howson & Urbach argue that it is not
- First, 95% confidence interval does not mean that the probability of μ being in the range is 0.95 (because μ is not a random variable and therefore has no probability)
- It may be tempting to think this way, but that is to shift from an objective to a subjective understanding of probability

Is this confidence interval really objective?

- Howson & Urbach argue that it is not
- First, 95% confidence interval does not mean that the probability of μ being in the range is 0.95 (because μ is not a random variable and therefore has no probability)
- It may be tempting to think this way, but that is to shift from an objective to a subjective understanding of probability
- Second, the validity of this calculation depends on the sample size n having the value that it has. But, why does it have that value?

Is this confidence interval really objective?

- Howson & Urbach argue that it is not
- First, 95% confidence interval does not mean that the probability of μ being in the range is 0.95 (because μ is not a random variable and therefore has no probability)
- It may be tempting to think this way, but that is to shift from an objective to a subjective understanding of probability
- Second, the validity of this calculation depends on the sample size n having the value that it has. But, why does it have that value?
- Howson & Urbach note that this depends on the (often private) intentions of the experimenter

Is this confidence interval really objective?

- Howson & Urbach argue that it is not
- First, 95% confidence interval does not mean that the probability of μ being in the range is 0.95 (because μ is not a random variable and therefore has no probability)
- It may be tempting to think this way, but that is to shift from an objective to a subjective understanding of probability
- Second, the validity of this calculation depends on the sample size n having the value that it has. But, why does it have that value?
- Howson & Urbach note that this depends on the (often private) intentions of the experimenter
- So, there's some obstacles to understanding confidence intervals as purely objective

Bayesian estimation

So, how might the Bayesian statistician estimate μ?

- Unlike the frequentist, Bayesians treat μ as a random variable

Bayesian estimation

So, how might the Bayesian statistician estimate μ?

- Unlike the frequentist, Bayesians treat μ as a random variable
- They begin with an initial guess about the distribution of μ

Bayesian estimation

So, how might the Bayesian statistician estimate μ?

- Unlike the frequentist, Bayesians treat μ as a random variable
- They begin with an initial guess about the distribution of μ
- As data comes in, they update that distribution according to Bayes' rule

Bayesian estimation

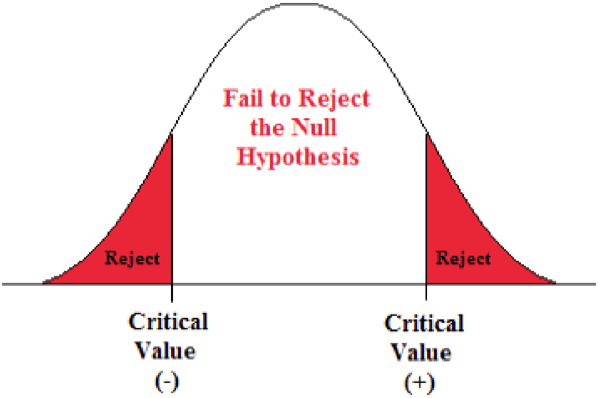
So, how might the Bayesian statistician estimate μ?

- Unlike the frequentist, Bayesians treat μ as a random variable
- They begin with an initial guess about the distribution of μ
- As data comes in, they update that distribution according to Bayes' rule
- As the sample size increases, the effect of the prior distribution is minimized

Bayesian estimation

So, how might the Bayesian statistician estimate μ?

- Unlike the frequentist, Bayesians treat μ as a random variable
- They begin with an initial guess about the distribution of μ
- As data comes in, they update that distribution according to Bayes' rule
- As the sample size increases, the effect of the prior distribution is minimized
- That means that as evidence accumulates, two researchers with very different prior distributions will come to the same posterior distribution


Fail to Reject the Null Hypothesis
Reject
Reject
Critical Value (-)
Critical Value (+)

What is significance testing?

- Significance tests are a frequentist method for testing a hypothesis with a statistical result

What is significance testing?

- Significance tests are a frequentist method for testing a hypothesis with a statistical result
- They are used to assess how well a particular sample of statistical evidence supports some hypothesis about the population being studied

What is significance testing?

- Significance tests are a frequentist method for testing a hypothesis with a statistical result
- They are used to assess how well a particular sample of statistical evidence supports some hypothesis about the population being studied
- The hypothesis is about some feature of the population, such as the proportion that have some trait or the mean (average) value of some trait in the population

What is significance testing?

- Significance tests are a frequentist method for testing a hypothesis with a statistical result
- They are used to assess how well a particular sample of statistical evidence supports some hypothesis about the population being studied
- The hypothesis is about some feature of the population, such as the proportion that have some trait or the mean (average) value of some trait in the population
- The result of the test is a probability of the result given the "null hypothesis" (H_0)

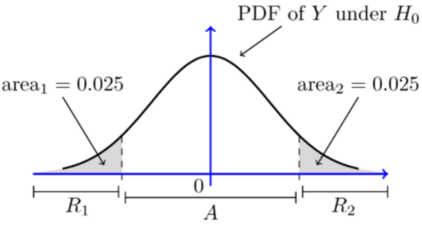
What is significance testing?

- Significance tests are a frequentist method for testing a hypothesis with a statistical result
- They are used to assess how well a particular sample of statistical evidence supports some hypothesis about the population being studied
- The hypothesis is about some feature of the population, such as the proportion that have some trait or the mean (average) value of some trait in the population
- The result of the test is a probability of the result given the "null hypothesis" (H_0)
- The null hypothesis is the hypothesis that the study will not demonstrate an effect

What is significance testing?

- Significance tests are a frequentist method for testing a hypothesis with a statistical result
- They are used to assess how well a particular sample of statistical evidence supports some hypothesis about the population being studied
- The hypothesis is about some feature of the population, such as the proportion that have some trait or the mean (average) value of some trait in the population
- The result of the test is a probability of the result given the "null hypothesis" (H_0)
- The null hypothesis is the hypothesis that the study will not demonstrate an effect
- It contrasts with the "alternative hypothesis" (H_a) which is the hypothesis we are actually interested in

Fisher

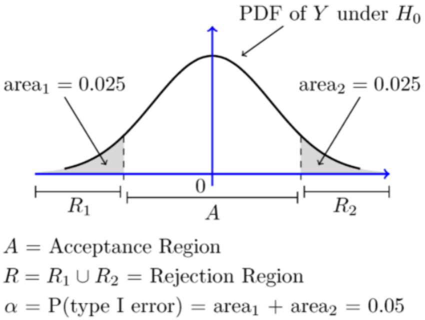


$$A = \text{Acceptance Region}$$
$$R = R_1 \cup R_2 = \text{Rejection Region}$$
$$\alpha = \text{P(type I error)} = \text{area}_1 + \text{area}_2 = 0.05$$

Source: Probability Course (see syllabus for link)

- The x-axis represents all the possible results of the test

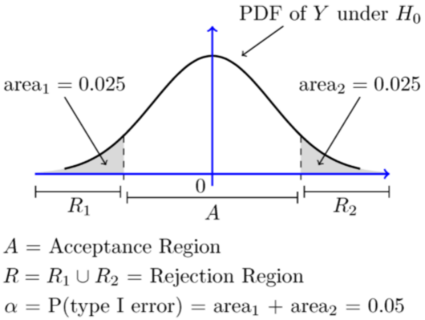
Fisher



Source: Probability Course (see syllabus for link)

- The x-axis represents all the possible results of the test
- The curve represents the probability density function for test results given H_0

Fisher



Source: Probability Course (see syllabus for link)

- The x-axis represents all the possible results of the test
- The curve represents the probability density function for test results given H_0
- The tail ends of the curve represent test values that are significant (ie. mathematically unlikely if H_0 were true)

What is the logic of significance testing?

Why not just test H_a directly?

- To answer this, we need to consider the logic behind significance testing

What is the logic of significance testing?

Why not just test H_a directly?

- To answer this, we need to consider the logic behind significance testing
- While it uses probabilities, the logic is deductive

What is the logic of significance testing?

Why not just test H_a directly?

- To answer this, we need to consider the logic behind significance testing
- While it uses probabilities, the logic is deductive
- Recall Popper's schema:

What is the logic of significance testing?

Why not just test H_a directly?

- To answer this, we need to consider the logic behind significance testing
- While it uses probabilities, the logic is deductive
- Recall Popper's schema:
    1. Theory T implies Observation O

What is the logic of significance testing?

Why not just test H_a directly?

- To answer this, we need to consider the logic behind significance testing
- While it uses probabilities, the logic is deductive
- Recall Popper's schema:
    1. Theory T implies Observation O
    2. O is not observed

What is the logic of significance testing?

Why not just test H_a directly?

- To answer this, we need to consider the logic behind significance testing
- While it uses probabilities, the logic is deductive
- Recall Popper's schema:
  1. Theory T implies Observation O
  2. O is not observed
  3. Therefore, T is false

What is the logic of significance testing?

Why not just test H_a directly?

- To answer this, we need to consider the logic behind significance testing
- While it uses probabilities, the logic is deductive
- Recall Popper's schema:
  1. Theory T implies Observation O
  2. O is not observed
  3. Therefore, T is false
- Significance testing is an instance of this schema:

What is the logic of significance testing?

Why not just test H_a directly?

- To answer this, we need to consider the logic behind significance testing
- While it uses probabilities, the logic is deductive
- Recall Popper's schema:
    1. Theory T implies Observation O
    2. O is not observed
    3. Therefore, T is false
- Significance testing is an instance of this schema:
    1. H_0 predicts a non-significant result

What is the logic of significance testing?

Why not just test H_a directly?

- To answer this, we need to consider the logic behind significance testing
- While it uses probabilities, the logic is deductive
- Recall Popper's schema:
    1. Theory T implies Observation O
    2. O is not observed
    3. Therefore, T is false
- Significance testing is an instance of this schema:
    1. H_0 predicts a non-significant result
    2. A significant result is observed

What is the logic of significance testing?

Why not just test H_a directly?

- To answer this, we need to consider the logic behind significance testing
- While it uses probabilities, the logic is deductive
- Recall Popper's schema:
    1. Theory T implies Observation O
    2. O is not observed
    3. Therefore, T is false
- Significance testing is an instance of this schema:
    1. H_0 predicts a non-significant result
    2. A significant result is observed
    3. Therefore, H_0 is rejected

What is the logic of significance testing?

Why not just test H_a directly?

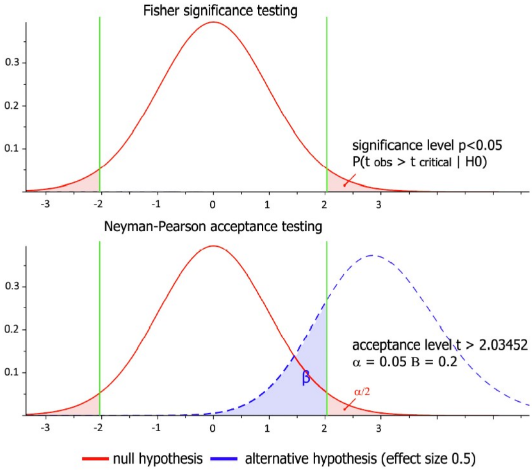
- To answer this, we need to consider the logic behind significance testing
- While it uses probabilities, the logic is deductive
- Recall Popper's schema:
  1. Theory T implies Observation O
  2. O is not observed
  3. Therefore, T is false
- Significance testing is an instance of this schema:
  1. H_0 predicts a non-significant result
  2. A significant result is observed
  3. Therefore, H_0 is rejected
- If H_0 is rejected, then this corroborates H_a

What is the logic of significance testing?

Why not just test H_a directly?

- To answer this, we need to consider the logic behind significance testing
- While it uses probabilities, the logic is deductive
- Recall Popper's schema:
  1. Theory T implies Observation O
  2. O is not observed
  3. Therefore, T is false
- Significance testing is an instance of this schema:
  1. H_0 predicts a non-significant result
  2. A significant result is observed
  3. Therefore, H_0 is rejected
- If H_0 is rejected, then this corroborates H_a
- But it does not confirm H_a! (Nothing does, because this framework is Popperian)

Fisher vs Neyman-Pearson



Source: Cyril Pernet

But what actually is "significance"?

- Significance is a measure of how (im)probable the result would be if H_0 were true

But what actually is “significance”?

- Significance is a measure of how (im)probable the result would be if H_0 were true
- A significant result is one that is mathematically unlikely if H_0 were true

But what actually is "significance"?

- Significance is a measure of how (im)probable the result would be if H_0 were true
- A significant result is one that is mathematically unlikely if H_0 were true
- In terms of probability, a significant result is one in which $P(R|H_0)$ is very low

But what actually is “significance”?

- Significance is a measure of how (im)probable the result would be if H_0 were true
- A significant result is one that is mathematically unlikely if H_0 were true
- In terms of probability, a significant result is one in which $P(R|H_0)$ is very low
- Of course, we know that this is not the same as $P(H_0|R)$

But what actually is "significance"?

- Significance is a measure of how (im)probable the result would be if H_0 were true
- A significant result is one that is mathematically unlikely if H_0 were true
- In terms of probability, a significant result is one in which $P(R|H_0)$ is very low
- Of course, we know that this is not the same as $P(H_0|R)$
- Note, Travers makes this mistake! When describing what $p = 0.26$ means: "In other words...there is a 26 percent probability that the null hypothesis is true given these results" (p.214)

Rejecting H_0

- Significance is measured in p-values

Rejecting H_0

- Significance is measured in p-values
- The lower the $p-$value, the higher the significance of the result of the test

Rejecting H_0

- Significance is measured in p-values
- The lower the $p-$value, the higher the significance of the result of the test
- We can set a critical value α to determine the threshold for significance (ie. where $p < 0.05$)

Rejecting H_0

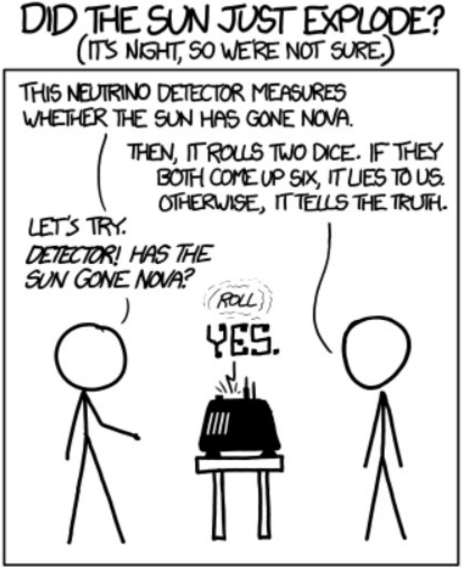
- Significance is measured in p-values
- The lower the $p-$value, the higher the significance of the result of the test
- We can set a critical value α to determine the threshold for significance (ie. where $p < 0.05$)
- Often, you can look these critical values up in pre-calculated tables, based on the structure of the test

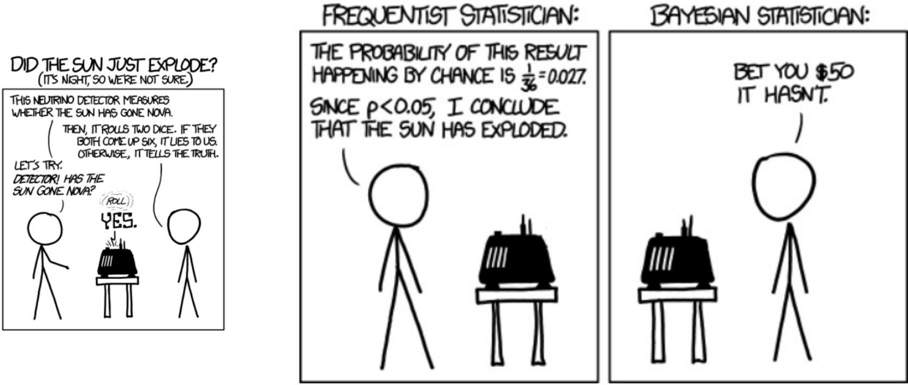
Rejecting H_0

- Significance is measured in p-values
- The lower the $p-$value, the higher the significance of the result of the test
- We can set a critical value α to determine the threshold for significance (ie. where $p < 0.05$)
- Often, you can look these critical values up in pre-calculated tables, based on the structure of the test
- Again, what does significance tell us? What does it mean if our result passes the threshold for significance?

Rejecting H_0

- Significance is measured in p-values
- The lower the $p-$value, the higher the significance of the result of the test
- We can set a critical value α to determine the threshold for significance (ie. where $p < 0.05$)
- Often, you can look these critical values up in pre-calculated tables, based on the structure of the test
- Again, what does significance tell us? What does it mean if our result passes the threshold for significance?
- That the result would be unlikely if H_0 were true


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.



Source: xkcd.com

P-hacking

This refers to a collection of less-than-honest practices (fudging):

- Stop collecting data once a result that is statistically significant has been obtained

P-hacking

This refers to a collection of less-than-honest practices (fudging):

- Stop collecting data once a result that is statistically significant has been obtained
- Checking statistical significance in order to decide whether to collect more data

P-hacking

This refers to a collection of less-than-honest practices (fudging):

- Stop collecting data once a result that is statistically significant has been obtained
- Checking statistical significance in order to decide whether to collect more data
- Not reporting on the effect of removing outliers from the data set

P-hacking

This refers to a collection of less-than-honest practices (fudging):

- Stop collecting data once a result that is statistically significant has been obtained
- Checking statistical significance in order to decide whether to collect more data
- Not reporting on the effect of removing outliers from the data set
- Deciding to remove outliers only after checking the effect this has on statistical significance

P-hacking

This refers to a collection of less-than-honest practices (fudging):

- Stop collecting data once a result that is statistically significant has been obtained
- Checking statistical significance in order to decide whether to collect more data
- Not reporting on the effect of removing outliers from the data set
- Deciding to remove outliers only after checking the effect this has on statistical significance
- These practices raise the likelihood of making a Type 1 error

P-hacking

This refers to a collection of less-than-honest practices (fudging):

- Stop collecting data once a result that is statistically significant has been obtained
- Checking statistical significance in order to decide whether to collect more data
- Not reporting on the effect of removing outliers from the data set
- Deciding to remove outliers only after checking the effect this has on statistical significance
- These practices raise the likelihood of making a Type 1 error
- Why? What does that mean?

P-hacking


CUT CORNERS. NOT CUT CORNERS.
CUT CORNERS. I WILL NOT CUT CORNERS.
CUT CORNERS. I WILL NOT CUT CORNERS.

Should significance be the "gold standard"?

- Significance is not the only relevant statistical measure

Should significance be the "gold standard"?

- Significance is not the only relevant statistical measure
  - Statistical power–the probability of correctly rejecting H_0 (eg. Consider a sample size of 50, and an effect that occurs in 1% of cases. This is a case of low statistical power.)

Should significance be the "gold standard"?

- Significance is not the only relevant statistical measure
  - Statistical power–the probability of correctly rejecting H_0 (eg. Consider a sample size of 50, and an effect that occurs in 1% of cases. This is a case of low statistical power.)
  - Effect size–how great a difference the intervention makes (eg. Both alcohol and tobacco have been shown to have a statistically significant increase in cancer risk. But alcohol makes it 2-3 times more likely to get cancer, while smoking makes it 15-30 times more likely.)

Should significance be the "gold standard"?

- Significance is not the only relevant statistical measure
  - Statistical power–the probability of correctly rejecting H_0 (eg. Consider a sample size of 50, and an effect that occurs in 1% of cases. This is a case of low statistical power.)
  - Effect size–how great a difference the intervention makes (eg. Both alcohol and tobacco have been shown to have a statistically significant increase in cancer risk. But alcohol makes it 2-3 times more likely to get cancer, while smoking makes it 15-30 times more likely.)
- Statistically insignificant findings can also be of interest. For example, a famous study out of Denmark in 2003 found no significant correlation between thimerasol exposure (in vaccines) and autism diagnoses. (In fact, they were slightly negatively correlated–as thimerasol use went down, the rate of autism diagnoses went up.)

Strengths and Weaknesses of NHST

- Strengths

Strengths and Weaknesses of NHST

- Strengths
  - It gives a standardized framework for testing hypotheses

Strengths and Weaknesses of NHST

- Strengths
  - It gives a standardized framework for testing hypotheses
  - Anyone who has been trained in statistics can verify the conclusion just using math

Strengths and Weaknesses of NHST

- Strengths
  - It gives a standardized framework for testing hypotheses
  - Anyone who has been trained in statistics can verify the conclusion just using math
  - It is a useful tool to evaluate claims about about whether an intervention will have the effect we are interested in (eg. oral reading fluency intervention)

Strengths and Weaknesses of NHST

- Strengths
  - It gives a standardized framework for testing hypotheses
  - Anyone who has been trained in statistics can verify the conclusion just using math
  - It is a useful tool to evaluate claims about about whether an intervention will have the effect we are interested in (eg. oral reading fluency intervention)
- Weaknesses

Strengths and Weaknesses of NHST

- Strengths
  - It gives a standardized framework for testing hypotheses
  - Anyone who has been trained in statistics can verify the conclusion just using math
  - It is a useful tool to evaluate claims about about whether an intervention will have the effect we are interested in (eg. oral reading fluency intervention)
- Weaknesses
  - It is easy to misunderstand

Strengths and Weaknesses of NHST

- Strengths
  - It gives a standardized framework for testing hypotheses
  - Anyone who has been trained in statistics can verify the conclusion just using math
  - It is a useful tool to evaluate claims about about whether an intervention will have the effect we are interested in (eg. oral reading fluency intervention)
- Weaknesses
  - It is easy to misunderstand
  - Choosing $p < 0.05$ is a bit arbitrary and inflexible. What about the consequences of rejecting/failing to reject the H_0?

Strengths and Weaknesses of NHST

- Strengths
  - It gives a standardized framework for testing hypotheses
  - Anyone who has been trained in statistics can verify the conclusion just using math
  - It is a useful tool to evaluate claims about about whether an intervention will have the effect we are interested in (eg. oral reading fluency intervention)
- Weaknesses
  - It is easy to misunderstand
  - Choosing $p < 0.05$ is a bit arbitrary and inflexible. What about the consequences of rejecting/failing to reject the H_0?
  - It is tempting to see significance as a pass/fail measure of whether the test was successful. But that's overly simplistic

How does NHST relate to replication?

- Some suggest that $p < 0.05$ is too relaxed of a threshold, and we should change the benchmark $p < 0.01$ to improve replication

How does NHST relate to replication?

- Some suggest that $p < 0.05$ is too relaxed of a threshold, and we should change the benchmark $p < 0.01$ to improve replication
  - This could actually make the problem worse!

How does NHST relate to replication?

- Some suggest that $p < 0.05$ is too relaxed of a threshold, and we should change the benchmark $p < 0.01$ to improve replication
  - This could actually make the problem worse!
  - Take the set of statistically significant findings: this is a mixture of honest findings and fraudulent ones

How does NHST relate to replication?

- Some suggest that $p < 0.05$ is too relaxed of a threshold, and we should change the benchmark $p < 0.01$ to improve replication
  - This could actually make the problem worse!
  - Take the set of statistically significant findings: this is a mixture of honest findings and fraudulent ones
  - What happens to that ratio if we reduce the amount of honest findings?

How does NHST relate to replication?

- Some suggest that $p < 0.05$ is too relaxed of a threshold, and we should change the benchmark $p < 0.01$ to improve replication
  - This could actually make the problem worse!
  - Take the set of statistically significant findings: this is a mixture of honest findings and fraudulent ones
  - What happens to that ratio if we reduce the amount of honest findings?
- Should there be a “one-size fits all” approach?

How does NHST relate to replication?

- Some suggest that $p < 0.05$ is too relaxed of a threshold, and we should change the benchmark $p < 0.01$ to improve replication
  - This could actually make the problem worse!
  - Take the set of statistically significant findings: this is a mixture of honest findings and fraudulent ones
  - What happens to that ratio if we reduce the amount of honest findings?
- Should there be a "one-size fits all" approach?
  - $p < 0.05$ is considered significant by convention...but it's a bit arbitrary

How does NHST relate to replication?

- Some suggest that $p < 0.05$ is too relaxed of a threshold, and we should change the benchmark $p < 0.01$ to improve replication
  - This could actually make the problem worse!
  - Take the set of statistically significant findings: this is a mixture of honest findings and fraudulent ones
  - What happens to that ratio if we reduce the amount of honest findings?
- Should there be a "one-size fits all" approach?
  - $p < 0.05$ is considered significant by convention...but it's a bit arbitrary
  - Should there even be a single standard value?

How does NHST relate to replication?

- Some suggest that $p < 0.05$ is too relaxed of a threshold, and we should change the benchmark $p < 0.01$ to improve replication
  - This could actually make the problem worse!
  - Take the set of statistically significant findings: this is a mixture of honest findings and fraudulent ones
  - What happens to that ratio if we reduce the amount of honest findings?
- Should there be a "one-size fits all" approach?
  - $p < 0.05$ is considered significant by convention...but it's a bit arbitrary
  - Should there even be a single standard value?
  - What would Heather Douglas say?

How does NHST relate to replication?

- Others suggest we need to pay more attention to other statistical factors

How does NHST relate to replication?

- Others suggest we need to pay more attention to other statistical factors
  - Effect size

How does NHST relate to replication?

- Others suggest we need to pay more attention to other statistical factors
    - Effect size
    - Statistical power

How does NHST relate to replication?

- Others suggest we need to pay more attention to other statistical factors
  - Effect size
  - Statistical power
- "File-drawer effect"–publication bias in favour of significant results (at the expense of insignificant results)

How does NHST relate to replication?

- Others suggest we need to pay more attention to other statistical factors
  - Effect size
  - Statistical power
- "File-drawer effect"–publication bias in favour of significant results (at the expense of insignificant results)
- We mentioned pre-registration earlier. Here is another way that pre-registration is useful (as mentioned in the Romero & Sprenger article)

How does NHST relate to replication?

- Others suggest we need to pay more attention to other statistical factors
    - Effect size
    - Statistical power
- "File-drawer effect"–publication bias in favour of significant results (at the expense of insignificant results)
- We mentioned pre-registration earlier. Here is another way that pre-registration is useful (as mentioned in the Romero & Sprenger article)
- What do you think?

Can Bayesian analysis help replication?

One proposal has been that we should switch from frequentism (NHST) to Bayesian analysis

- One note: Bayesian analyses often make use of the “Bayes Factor”

Can Bayesian analysis help replication?

One proposal has been that we should switch from frequentism (NHST) to Bayesian analysis

- One note: Bayesian analyses often make use of the "Bayes Factor"
- (Recall from Week 5) The Bayes Factor is a likelihood ratio

$$\frac{P(R|H_a)}{P(R|H_0)}$$

Can Bayesian analysis help replication?

One proposal has been that we should switch from frequentism (NHST) to Bayesian analysis

- One note: Bayesian analyses often make use of the "Bayes Factor"
- (Recall from Week 5) The Bayes Factor is a likelihood ratio

$$\frac{P(R|H_a)}{P(R|H_0)}$$

- The Bayes Factor is a contrastive measure of how strongly the evidence favours H_a over H_0

Can Bayesian analysis help replication?

One proposal has been that we should switch from frequentism (NHST) to Bayesian analysis

- One note: Bayesian analyses often make use of the "Bayes Factor"
- (Recall from Week 5) The Bayes Factor is a likelihood ratio

$$\frac{P(R|H_a)}{P(R|H_0)}$$

- The Bayes Factor is a contrastive measure of how strongly the evidence favours H_a over H_0
- If $BF = 1$, then the evidence is neutral (supports each hypothesis equally)

Can Bayesian analysis help replication?

One proposal has been that we should switch from frequentism (NHST) to Bayesian analysis

- One note: Bayesian analyses often make use of the "Bayes Factor"
- (Recall from Week 5) The Bayes Factor is a likelihood ratio

$$\frac{P(R|H_a)}{P(R|H_0)}$$

- The Bayes Factor is a contrastive measure of how strongly the evidence favours H_a over H_0
- If $BF = 1$, then the evidence is neutral (supports each hypothesis equally)
- If $BF > 1$, then the evidence supports H_a

Can Bayesian analysis help replication?

One proposal has been that we should switch from frequentism (NHST) to Bayesian analysis

- One note: Bayesian analyses often make use of the "Bayes Factor"
- (Recall from Week 5) The Bayes Factor is a likelihood ratio

$$\frac{P(R|H_a)}{P(R|H_0)}$$

- The Bayes Factor is a contrastive measure of how strongly the evidence favours H_a over H_0
- If $BF = 1$, then the evidence is neutral (supports each hypothesis equally)
- If $BF > 1$, then the evidence supports H_a
- If $BF < 1$, then the evidence supports H_0

Can Bayesian analysis help replication?

One proposal has been that we should switch from frequentism (NHST) to Bayesian analysis

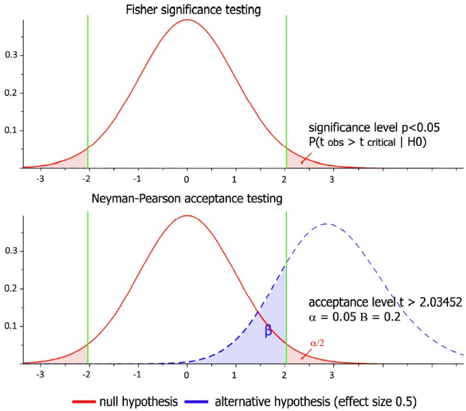
- One note: Bayesian analyses often make use of the "Bayes Factor"
- (Recall from Week 5) The Bayes Factor is a likelihood ratio

$$\frac{P(R|H_a)}{P(R|H_0)}$$

- The Bayes Factor is a contrastive measure of how strongly the evidence favours H_a over H_0
- If $BF = 1$, then the evidence is neutral (supports each hypothesis equally)
- If $BF > 1$, then the evidence supports H_a
- If $BF < 1$, then the evidence supports H_0
- By convention: "Inconclusive" is when $\frac{1}{3} < BF < 3$

Inconclusive Evidence

Where we would see $\frac{1}{3} < BF < 3$ on either of these graphs?



Source: Cyril Pernet

Can Bayesian analysis help replication?

Romero & Sprenger: There is less of a "file-drawer" effect with Bayesian analysis

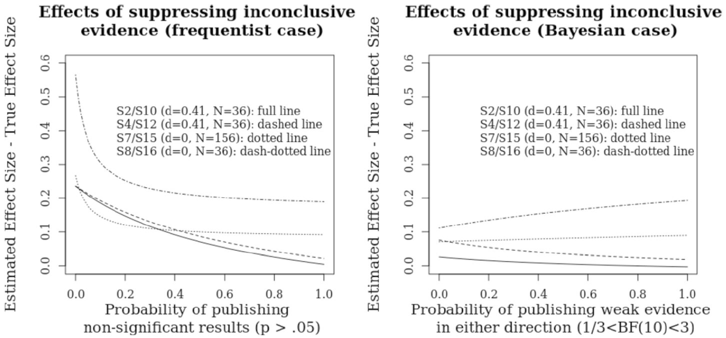


Fig. 4 Difference between estimated and true effect size as a function of the probability of suppressing inconclusive evidence, that is, the prevalence of the file drawer effect. Left graph = frequentist analysis, right graph = Bayesian analysis

- This lends some support to the claim that Bayesian analysis can improve replication

- This lends some support to the claim that Bayesian analysis can improve replication
- Where resources are limited, and where inconclusive evidence is suppressed, Bayesian analysis can be more accurate

- This lends some support to the claim that Bayesian analysis can improve replication
- Where resources are limited, and where inconclusive evidence is suppressed, Bayesian analysis can be more accurate
- But, Bayesian analysis is imperfect as well (eg. it tends to underestimate effect size)

- This lends some support to the claim that Bayesian analysis can improve replication
- Where resources are limited, and where inconclusive evidence is suppressed, Bayesian analysis can be more accurate
- But, Bayesian analysis is imperfect as well (eg. it tends to underestimate effect size)
- Summing up:

- This lends some support to the claim that Bayesian analysis can improve replication
- Where resources are limited, and where inconclusive evidence is suppressed, Bayesian analysis can be more accurate
- But, Bayesian analysis is imperfect as well (eg. it tends to underestimate effect size)
- Summing up:
    - NHST is not the cause of the replication crisis

- This lends some support to the claim that Bayesian analysis can improve replication
- Where resources are limited, and where inconclusive evidence is suppressed, Bayesian analysis can be more accurate
- But, Bayesian analysis is imperfect as well (eg. it tends to underestimate effect size)
- Summing up:
    - NHST is not the cause of the replication crisis
    - Misapplication of it (through suppression of inconclusive results) does contribute

- This lends some support to the claim that Bayesian analysis can improve replication
- Where resources are limited, and where inconclusive evidence is suppressed, Bayesian analysis can be more accurate
- But, Bayesian analysis is imperfect as well (eg. it tends to underestimate effect size)
- Summing up:
  - NHST is not the cause of the replication crisis
  - Misapplication of it (through suppression of inconclusive results) does contribute
  - A switch to Bayesian analysis *might* help improve that problem

- This lends some support to the claim that Bayesian analysis can improve replication
- Where resources are limited, and where inconclusive evidence is suppressed, Bayesian analysis can be more accurate
- But, Bayesian analysis is imperfect as well (eg. it tends to underestimate effect size)
- Summing up:
    - NHST is not the cause of the replication crisis
    - Misapplication of it (through suppression of inconclusive results) does contribute
    - A switch to Bayesian analysis *might* help improve that problem
    - But other changes like publishing data sets (facilitating meta-analysis) and pre-accepting studies for publication (to avoid suppressing valuable data) should also be encouraged

Summing up

- There isn't a single cause for the replication crisis

Summing up

- There isn't a single cause for the replication crisis
- We shouldn't expect a single cure

Summing up

- There isn't a single cause for the replication crisis
- We shouldn't expect a single cure
- A varied approach is needed

Summing up

- There isn't a single cause for the replication crisis
- We shouldn't expect a single cure
- A varied approach is needed
    - Funding for replications

Summing up

- There isn't a single cause for the replication crisis
- We shouldn't expect a single cure
- A varied approach is needed
    - Funding for replications
    - Pre-registration/pre-acceptance for publication

Summing up

- There isn't a single cause for the replication crisis
- We shouldn't expect a single cure
- A varied approach is needed
  - Funding for replications
  - Pre-registration/pre-acceptance for publication
  - Better use of statistical methods

Summing up

- There isn't a single cause for the replication crisis
- We shouldn't expect a single cure
- A varied approach is needed
  - Funding for replications
  - Pre-registration/pre-acceptance for publication
  - Better use of statistical methods
  - Publication of all data (to facilitate better meta-analysis)

Summing up

- There isn't a single cause for the replication crisis
- We shouldn't expect a single cure
- A varied approach is needed
    - Funding for replications
    - Pre-registration/pre-acceptance for publication
    - Better use of statistical methods
    - Publication of all data (to facilitate better meta-analysis)
    - Other?

Readings for next week

- Readings for next week
  - Dentith and Keeley "The Applied Epistemology of Conspiracy Theories: an Overview"
  - Grimes "On the Viability of Conspiratorial Beliefs"
  - O'Connor and Weatherall "The Social Network"