Surprise, Surprise: It's a false positive

Jarrod Hadfield

University of Edinburgh

December 18, 2017

Surprise, Surprise: It's a false positive

Jarrod Hadfield

University of Edinburgh

December 18, 2017

- OK - this afternoon I'm going to get you to think a little bit about type I errors - findings that are deemed 'significant' but in fact turn out not to be true. A couple of years ago, I gave a very similar talk to this at one of our lab meetings, and quite a few people said that I should present this talk more widely, especially to PhD students.
- So I want to start by asking people what fraction of significant results do they think are real (true positives): so I don't necessarily mean your own work, but more generally; from the work you might read and cite: what fraction of significant results do you think are true. OK - now hold that number in your head - you're not allowed to change your mind when you see what other people go for.
- OK - so you are quite cynical, but what I want to make you think about is whether you're too cynical or not cynical enough. I also want you to think about all the times you've come up with an unexpected significant result, or someone at a lab meeting has presented something surprising. At least in our lab meetings I have never heard anybody say that they think someone else's significant result is a false positive, even for those bizzare three-way interactions we'd rather sit there for 20 minutes trying to come up with some semi-plausible biological reason to explain it. and so why is that, when most of you think that x% of results are false positives, why do I do never here anyone tell anyone else they think their result is a false positive? and the reason of course is that it would be embarrassing to say that.
- What I want to do in this talk is mainly to convince you to stop being embarrassed about false positives: by the very nature of what we do they have to be common, and we need to acknowledge that and stop believing and expecting our results to be definitive answers: they're very provisional. What I do want you to be embarrassed about is failing to protect yourself against false positives, and not correcting published positives when you later find them to be false.

Reproducibility in Medicine

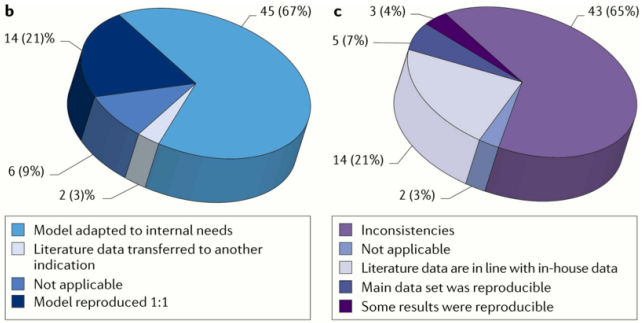
- Prinz, F. et al. (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 10

Reproducibility in Medicine
- Prinz, F. et al. (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 10

- Now people like Ben Goldacre have clearly been writing about this topic for a long period of time, but I think the watershed was really a few years ago, when two very influential papers were published.
- The first was published by an employee of Bayer HealthCare: which is a large German pharmaceutical company. Bayer HealthCare, and companies like it, often take promising results from the scientific literature and try to turn those results into a useful drug. and because a full-blown drug discovery and development programme is needed to turn these initial findings into a marketable drug, most pharmaceutical companies will run an in-house validation programme before putting forward the massive investments that are required.
- In 2011, faced with I guess disillusionment with current scientific practice, Bayer HealthCare decided to publish the outcomes for 67 of its validation programmes. The results were pretty devastating. So this figure on the right is the importat one. This light purple slice are those studies that could not be reproduced - 65% of studies were not reproducible. The dark blue and dark purple slices are those studies that are partly reproducible, and this grey slice are those studies that are completely reproducible - just 21%.
- All of the findings in this 65% were significant in the initial study, they were often high impact and published in high-impact journals, and in some cases had even spawned an entire field, with hundreds of secondary publications, without anyone going back to try and confirm the initial result.
- This second figure is also quite important. Often the original experimental set up was not followed exactly. In 21% of cases it was, but often the pharamaceutical company had to modify the experimental design to suit there own needs, a different cell line lets say. The key point is that reproducibility did not depend on whether the original experimental protocol was replicated exactly or some modified version was used. The results either seemed to be either wholly unrepeatable or repeatable under a moderate range of different conditions. This is an important point which I'll return to later.
- Following Bayer HealthCare's publication, Amgen (an american pharmaceutical company) followed suit. The outcome was even more depressing. For fifty-three landmark papers in cancer research only 11% could not be relpicated.
- A couple of years ago a large consortium of psychologists indepedently reran 100 classic experiments in psychology: 39% could be replicated. The papers from the pharmaceutical companies had such a big impact, ...

Reproducibility in Medicine

- Prinz, F. et al. (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 10

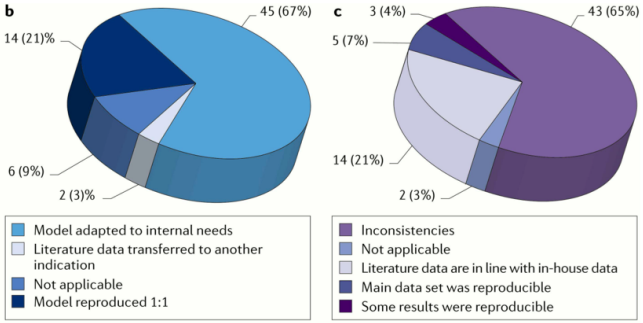


Reproducibility in Medicine
Prinz, F. et al. (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 10

- Now people like Ben Goldacre have clearly been writing about this topic for a long period of time, but I think the watershed was really a few years ago, when two very influential papers were published.
- The first was published by an employee of Bayer HealthCare: which is a large German pharmaceutical company. Bayer HealthCare, and companies like it, often take promising results from the scientific literature and try to turn those results into a useful drug. and because a full-blown drug discovery and development programme is needed to turn these initial findings into a marketable drug, most pharmaceutical companies will run an in-house validation programme before putting forward the massive investments that are required.
- In 2011, faced with I guess disillusionment with current scientific practice, Bayer HealthCare decided to publish the outcomes for 67 of its validation programmes. The results were pretty devastating. So this figure on the right is the importat one. This light purple slice are those studies that could not be reproduced - 65% of studies were not reproducible. The dark blue and dark purple slices are those studies that are partly reproducible, and this grey slice are those studies that are completely reproducible - just 21%.
- All of the findings in this 65% were significant in the initial study, they were often high impact and published in high-impact journals, and in some cases had even spawned an entire field, with hundreds of secondary publications, without anyone going back to try and confirm the initial result.
- This second figure is also quite important. Often the original experimental set up was not followed exactly. In 21% of cases it was, but often the pharamaceutical company had to modify the experimental design to suit there own needs, a different cell line lets say. The key point is that reproducibility did not depend on whether the original experimental protocol was replicated exactly or some modified version was used. The results either seemed to be either wholly unrepeatable or repeatable under a moderate range of different conditions. This is an important point which I'll return to later.
- Following Bayer HealthCare's publication, Amgen (an american pharmaceutical company) followed suit. The outcome was even more depressing. For fifty-three landmark papers in cancer research only 11% could not be relpicated.
- A couple of years ago a large consortium of psychologists indepedently reran 100 classic experiments in psychology: 39% could be replicated. The papers from the pharmaceutical companies had such a big impact, ...

Reproducibility in Medicine

- Prinz, F. et al. (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 10



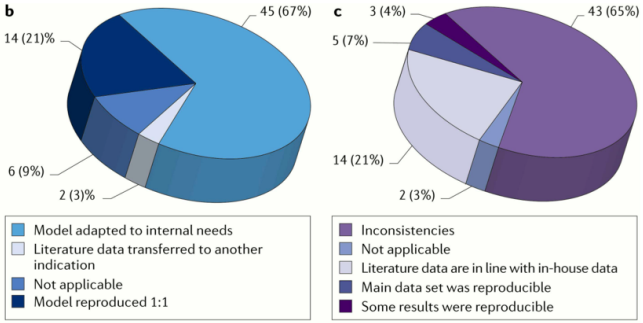
- Begley, CG. & Ellis, LM. (2012) Drug development: Raise standards for preclinical cancer research. Nature 483

Reproducibility in Medicine
Prinz, F. et al. (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 10
Begley, CG. & Ellis, LM. (2012) Drug development: Raise standards for preclinical cancer research. Nature 483

- Now people like Ben Goldacre have clearly been writing about this topic for a long period of time, but I think the watershed was really a few years ago, when two very influential papers were published.
- The first was published by an employee of Bayer HealthCare: which is a large German pharmaceutical company. Bayer HealthCare, and companies like it, often take promising results from the scientific literature and try to turn those results into a useful drug. and because a full-blown drug discovery and development programme is needed to turn these initial findings into a marketable drug, most pharmaceutical companies will run an in-house validation programme before putting forward the massive investments that are required.
- In 2011, faced with I guess disillusionment with current scientific practice, Bayer HealthCare decided to publish the outcomes for 67 of its validation programmes. The results were pretty devastating. So this figure on the right is the importat one. This light purple slice are those studies that could not be reproduced - 65% of studies were not reproducible. The dark blue and dark purple slices are those studies that are partly reproducible, and this grey slice are those studies that are completely reproducible - just 21%.
- All of the findings in this 65% were significant in the initial study, they were often high impact and published in high-impact journals, and in some cases had even spawned an entire field, with hundreds of secondary publications, without anyone going back to try and confirm the initial result.
- This second figure is also quite important. Often the original experimental set up was not followed exactly. In 21% of cases it was, but often the pharamaceutical company had to modify the experimental design to suit there own needs, a different cell line lets say. The key point is that reproducibility did not depend on whether the original experimental protocol was replicated exactly or some modified version was used. The results either seemed to be either wholly unrepeatable or repeatable under a moderate range of different conditions. This is an important point which I'll return to later.
- Following Bayer HealthCare's publication, Amgen (an american pharmaceutical company) followed suit. The outcome was even more depressing. For fifty-three landmark papers in cancer research only 11% could not be relpicated.
- A couple of years ago a large consortium of psychologists indepedently reran 100 classic experiments in psychology: 39% could be replicated. The papers from the pharmaceutical companies had such a big impact, ...

Reproducibility in Medicine

- Prinz, F. et al. (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 10



- Begley, CG. & Ellis, LM. (2012) Drug development: Raise standards for preclinical cancer research. Nature 483
- Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349

2017-12-18

Surprising?

└Reproducibility in Medicine

Reproducibility in Medicine
Prinz, F. et al. (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 10
Begley, CG. & Ellis, LM. (2012) Drug development: Raise standards for preclinical cancer research. Nature 483
Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349

- Now people like Ben Goldacre have clearly been writing about this topic for a long period of time, but I think the watershed was really a few years ago, when two very influential papers were published.
- The first was published by an employee of Bayer HealthCare: which is a large German pharmaceutical company. Bayer HealthCare, and companies like it, often take promising results from the scientific literature and try to turn those results into a useful drug. and because a full-blown drug discovery and development programme is needed to turn these initial findings into a marketable drug, most pharmaceutical companies will run an in-house validation programme before putting forward the massive investments that are required.
- In 2011, faced with I guess disillusionment with current scientific practice, Bayer HealthCare decided to publish the outcomes for 67 of its validation programmes. The results were pretty devastating. So this figure on the right is the importat one. This light purple slice are those studies that could not be reproduced - 65% of studies were not reproducible. The dark blue and dark purple slices are those studies that are partly reproducible, and this grey slice are those studies that are completely reproducible - just 21%.
- All of the findings in this 65% were significant in the initial study, they were often high impact and published in high-impact journals, and in some cases had even spawned an entire field, with hundreds of secondary publications, without anyone going back to try and confirm the initial result.
- This second figure is also quite important. Often the original experimental set up was not followed exactly. In 21% of cases it was, but often the pharamaceutical company had to modify the experimental design to suit there own needs, a different cell line lets say. The key point is that reproducibility did not depend on whether the original experimental protocol was replicated exactly or some modified version was used. The results either seemed to be either wholly unrepeatable or repeatable under a moderate range of different conditions. This is an important point which I'll return to later.
- Following Bayer HealthCare's publication, Amgen (an american pharmaceutical company) followed suit. The outcome was even more depressing. For fifty-three landmark papers in cancer research only 11% could not be relpicated.
- A couple of years ago a large consortium of psychologists indepedently reran 100 classic experiments in psychology: 39% could be replicated. The papers from the pharmaceutical companies had such a big impact, ...

Scientific Integrity and Transparency

- U.S. House of Representatives Hearing on 'Scientific Integrity and Transparency' (2013)

Scientific Integrity and Transparency
U.S. House of Representatives Hearing on 'Scientific Integrity and Transparency' (2013)

- that they motivated US congress to hold a hearing on 'Scientific Integrity and Transparency'. One of the witnesses was Bruce Alberts who at the time was editor of Science and who you might know from the text book "Molecular Biology of the Cell" if undergraduates still use this? You can find the transcript of this hearing on-line; he didn't actually have very much of interest to say but he did leave a couple of pithy quotes worth thinking about:
- *'Budding scientists must be taught technical skills, including statistics, and must be imbued with scepticism towards their own results and those of others. Researchers ought to be judged on the basis of the quality, not the quantity, of their work.'*
- and *'We need to develop a value system where simply moving on from one's mistakes without publicly acknowledging them severely damages, rather than protects, a scientific reputation.'*
- Albert's seems to believe a root cause is the cynicism of scientists and a distorting reward system ... maybe that's what being the editor of Science does to you.

Scientific Integrity and Transparency

- U.S. House of Representatives Hearing on 'Scientific Integrity and Transparency' (2013)

  Bruce Alberts (then the editor of Science):

Scientific Integrity and Transparency
U.S. House of Representatives Hearing on 'Scientific Integrity and Transparency' (2013)
Bruce Alberts (then the editor of Science):

- that they motivated US congress to hold a hearing on 'Scientific Integrity and Transparency'. One of the witnesses was Bruce Alberts who at the time was editor of Science and who you might know from the text book "Molecular Biology of the Cell" if undergraduates still use this? You can find the transcript of this hearing on-line; he didn't actually have very much of interest to say but he did leave a couple of pithy quotes worth thinking about:
- *'Budding scientists must be taught technical skills, including statistics, and must be imbued with scepticism towards their own results and those of others. Researchers ought to be judged on the basis of the quality, not the quantity, of their work.'*
- and *'We need to develop a value system where simply moving on from one's mistakes without publicly acknowledging them severely damages, rather than protects, a scientific reputation.'*
- Albert's seems to believe a root cause is the cynicism of scientists and a distorting reward system ... maybe that's what being the editor of Science does to you.

Scientific Integrity and Transparency

- U.S. House of Representatives Hearing on 'Scientific Integrity and Transparency' (2013)

  Bruce Alberts (then the editor of Science):

  - *'Budding scientists must be taught technical skills, including statistics, and must be imbued with scepticism towards their own results and those of others. Researchers ought to be judged on the basis of the quality, not the quantity, of their work.'*

Scientific Integrity and Transparency
- U.S. House of Representatives Hearing on 'Scientific Integrity and Transparency' (2013)

Bruce Alberts (then the editor of Science):

- *'Budding scientists must be taught technical skills, including statistics, and must be imbued with scepticism towards their own results and those of others. Researchers ought to be judged on the basis of the quality, not the quantity, of their work.'*

- that they motivated US congress to hold a hearing on 'Scientific Integrity and Transparency'. One of the witnesses was Bruce Alberts who at the time was editor of Science and who you might know from the text book "Molecular Biology of the Cell" if undergraduates still use this? You can find the transcript of this hearing on-line; he didn't actually have very much of interest to say but he did leave a couple of pithy quotes worth thinking about:
- *'Budding scientists must be taught technical skills, including statistics, and must be imbued with scepticism towards their own results and those of others. Researchers ought to be judged on the basis of the quality, not the quantity, of their work.'*
- and *'We need to develop a value system where simply moving on from one's mistakes without publicly acknowledging them severely damages, rather than protects, a scientific reputation.'*
- Albert's seems to believe a root cause is the cynicism of scientists and a distorting reward system ... maybe that's what being the editor of Science does to you.

Scientific Integrity and Transparency

- U.S. House of Representatives Hearing on 'Scientific Integrity and Transparency' (2013)

  Bruce Alberts (then the editor of Science):

  - *'Budding scientists must be taught technical skills, including statistics, and must be imbued with scepticism towards their own results and those of others. Researchers ought to be judged on the basis of the quality, not the quantity, of their work.'*

  - *'We need to develop a value system where simply moving on from one's mistakes without publicly acknowledging them severely damages, rather than protects, a scientific reputation.'*

Scientific Integrity and Transparency
- U.S. House of Representatives Hearing on 'Scientific Integrity and Transparency' (2013)

Bruce Alberts (then the editor of Science):

- *'Budding scientists must be taught technical skills, including statistics, and must be imbued with scepticism towards their own results and those of others. Researchers ought to be judged on the basis of the quality, not the quantity, of their work.'*
- *'We need to develop a value system where simply moving on from one's mistakes without publicly acknowledging them severely damages, rather than protects, a scientific reputation.'*

- that they motivated US congress to hold a hearing on 'Scientific Integrity and Transparency'. One of the witnesses was Bruce Alberts who at the time was editor of Science and who you might know from the text book "Molecular Biology of the Cell" if undergraduates still use this? You can find the transcript of this hearing on-line; he didn't actually have very much of interest to say but he did leave a couple of pithy quotes worth thinking about:
- *'Budding scientists must be taught technical skills, including statistics, and must be imbued with scepticism towards their own results and those of others. Researchers ought to be judged on the basis of the quality, not the quantity, of their work.'*
- and *'We need to develop a value system where simply moving on from one's mistakes without publicly acknowledging them severely damages, rather than protects, a scientific reputation.'*
- Albert's seems to believe a root cause is the cynicism of scientists and a distorting reward system ... maybe that's what being the editor of Science does to you.

How Science Goes Wrong

How Science Goes Wrong

- A few months after the hearing the Economist ran a cover story on it. Now, the economist is one of the most, if not the most, pro-science, pro-technology newspapers you can buy. If its running a leader on scientific malpractice, then scientists, I think, need to start worrying. This is how that article ends - its perhaps a bit sensationalist, but the main article, if you can get a copy, is a brilliant piece of scientific journalism:

How Science Goes Wrong

'Science still commands enormous - if sometimes bemused - respect. But its privileged status is founded on the capacity to be right most of the time and to correct its mistakes when it gets things wrong. And it is not as if the universe is short of genuine mysteries to keep generations of scientists hard at work. The false trails laid down by shoddy research are an unforgivable barrier to understanding.'

How Science Goes Wrong

'Science still commands enormous - if sometimes bemused - respect. But its privileged status is founded on the capacity to be right most of the time and to correct its mistakes when it gets things wrong. And it is not as if the universe is short of genuine mysteries to keep generations of scientists hard at work. The false trails laid down by shoddy research are an unforgivable barrier to understanding.'

- A few months after the hearing the Economist ran a cover story on it. Now, the economist is one of the most, if not the most, pro-science, pro-technology newspapers you can buy. If its running a leader on scientific malpractice, then scientists, I think, need to start worrying. This is how that article ends - its perhaps a bit sensationalist, but the main article, if you can get a copy, is a brilliant piece of scientific journalism:

Surprise, Surprise: It's a false positive

- Background
- Problems
- Under-graduate Project
- Solutions
- Unacknowledged Issues

Surprise, Surprise: It's a false positive

Background
Problems
Under-graduate Project
Solutions
Unacknowledged Issues

- much of the background material I just spoke about was actually from this article
- A few weeks after reading it, I was asked to put forward an under-graduate project here in IEB and after badgering Nick and Andrew for a few days they agreed that I could make a zoology undergraduate do a statistics project. Unfortunately, she abandoned university a week after staring it - completely unrelated - but I reran the year after.
- Its mainly a bit of fun, but actually I think it makes for a nice case study, so I'll talk about it after discussing why false-positive rates are what they are, which hopefully should allow you to judge whether you are too cynical or not cynical enough.
- and I'll end, if I have time, with some recommendations on how you can reduce your false-positive rates.

Problems

- The reward system favours those that publish statistically significant high-profile work quickly without future correction/validation.

Problems
- The reward system favours those that publish statistically significant high-profile work quickly without future correction/validation.

- OK - so what are the problems - why do we have so many Type I errors. One of the main reasons put forward, as Alberts did in the congressional hearing, is that careerism amongst scientists is one of the biggest problems. I think it is a problem, but I honestly don't think its the biggest. Most people want to get the right answer most of the time.
- Inadequate peer review has been cited. Again, I think better peer review could help but I don't really think it would have a large impact: ultimately I think the responsibility for doing good science has to lie with the authors.
- The use of inappropriate statistical methods is another cited source of false-positives. And I think this is a big problem, but I also think that often its not because an inappropriate method has been used but because scientists tend to be really bad at understanding what statistical methods are telling them about uncertainty.
- Scientific research that is poorly thought through has been cited as contributor to high false-positive rates. Again, I think its true, but I think the main damage done by poorly thought through science really arises because of the statistical misinterpretation that accompanies it.

so I'm really going to focus on these last two points.

Problems

- The reward system favours those that publish statistically significant high-profile work quickly without future correction/validation.
- The quality of peer review is not adequate.

Problems
- The reward system favours those that publish statistically significant high-profile work quickly without future correction/validation.
- The quality of peer review is not adequate.

- OK - so what are the problems - why do we have so many Type I errors. One of the main reasons put forward, as Alberts did in the congressional hearing, is that careerism amongst scientists is one of the biggest problems. I think it is a problem, but I honestly don't think its the biggest. Most people want to get the right answer most of the time.
- Inadequate peer review has been cited. Again, I think better peer review could help but I don't really think it would have a large impact: ultimately I think the responsibility for doing good science has to lie with the authors.
- The use of inappropriate statistical methods is another cited source of false-positives. And I think this is a big problem, but I also think that often its not because an inappropriate method has been used but because scientists tend to be really bad at understanding what statistical methods are telling them about uncertainty.
- Scientific research that is poorly thought through has been cited as contributor to high false-positive rates. Again, I think its true, but I think the main damage done by poorly thought through science really arises because of the statistical misinterpretation that accompanies it.

so I'm really going to focus on these last two points.

Problems

- The reward system favours those that publish statistically significant high-profile work quickly without future correction/validation.
- The quality of peer review is not adequate.
- Statistical mistakes/misinterpretation are widespread.

Problems
- The reward system favours those that publish statistically significant high-profile work quickly without future correction/validation.
- The quality of peer review is not adequate.
- Statistical mistakes/misinterpretation are widespread.

- OK - so what are the problems - why do we have so many Type I errors. One of the main reasons put forward, as Alberts did in the congressional hearing, is that careerism amongst scientists is one of the biggest problems. I think it is a problem, but I honestly don't think its the biggest. Most people want to get the right answer most of the time.
- Inadequate peer review has been cited. Again, I think better peer review could help but I don't really think it would have a large impact: ultimately I think the responsibility for doing good science has to lie with the authors.
- The use of inappropriate statistical methods is another cited source of false-positives. And I think this is a big problem, but I also think that often its not because an inappropriate method has been used but because scientists tend to be really bad at understanding what statistical methods are telling them about uncertainty.
- Scientific research that is poorly thought through has been cited as contributor to high false-positive rates. Again, I think its true, but I think the main damage done by poorly thought through science really arises because of the statistical misinterpretation that accompanies it.

so I'm really going to focus on these last two points.

Problems

- The reward system favours those that publish statistically significant high-profile work quickly without future correction/validation.
- The quality of peer review is not adequate.
- Statistical mistakes/misinterpretation are widespread.
- A lot of scientific research is poorly thought through.

Problems
- The reward system favours those that publish statistically significant high-profile work quickly without future correction/validation.
- The quality of peer review is not adequate.
- Statistical mistakes/misinterpretation are widespread.
- A lot of scientific research is poorly thought through.

- OK - so what are the problems - why do we have so many Type I errors. One of the main reasons put forward, as Alberts did in the congressional hearing, is that careerism amongst scientists is one of the biggest problems. I think it is a problem, but I honestly don't think its the biggest. Most people want to get the right answer most of the time.
- Inadequate peer review has been cited. Again, I think better peer review could help but I don't really think it would have a large impact: ultimately I think the responsibility for doing good science has to lie with the authors.
- The use of inappropriate statistical methods is another cited source of false-positives. And I think this is a big problem, but I also think that often its not because an inappropriate method has been used but because scientists tend to be really bad at understanding what statistical methods are telling them about uncertainty.
- Scientific research that is poorly thought through has been cited as contributor to high false-positive rates. Again, I think its true, but I think the main damage done by poorly thought through science really arises because of the statistical misinterpretation that accompanies it.

so I'm really going to focus on these last two points.

How Science Goes Wrong

- Statistical mistakes/misinterpretation are widespread.
  - **Wacholder, S. et al.** (2004) *Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies.* Journal of the National Cancer Institute 96, 434-442
  - **Ioannidis, JPA.** (2005) *Why Most Published Research Findings Are False* PLoS Medicine 2 e124

How Science Goes Wrong
- Statistical mistakes/misinterpretation are widespread.
  - **Wacholder, S. et al.** (2004) *Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies.* Journal of the National Cancer Institute 96, 434-442
  - **Ioannidis, JPA.** (2005) *Why Most Published Research Findings Are False* PLoS Medicine 2 e124

- To start I'll cover some of the concepts discussed in these two papers. The concepts they discuss are pretty basic and were widely known but nevertheless these papers are very influential, the first paper because they explain the problems really well. The second paper took a different - and successful - tack where the author kicks the reader in the face about 20 times and then says, you need to listen to what I have to say.

False Positive Report Probability

- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)

False Positive Report Probability
- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)

- OK - we're going to have to do a bit of maths. To understand everything I'm going to tell you, you need to understand three quantities. The first is the significance-level, usually denoted as α. This is the probability of a statistically significant finding given that the null hypothesis is true: OK, its the chance of getting a false positive and you usually set it in advance. Typically, we use a significance-level of 0.05: if there's less than a 1 in 20 chance that the observed pattern could have been generated under the null model we would declare that a significant finding.
- The second quantity, which I'm sure you've all heard of, but your probably a little less familiar with is power. The power of a test is if the alternative hypothesis is true what is the probability that you will declare it significant. And unlike the significance-level, the power depends on your sample size and effect size. Power goes up if the effect size is large or the sample size is large: you don't need many mice and many elephants to say that elephants are probably bigger than mice, if you wanted to know whether voles were bigger than mice you'd probably need a larger sample size.
- The opposite of power is the Type II error rate usually denoted β: the chance that you declare something non-significant when in fact the association is real.
- Everybody happy with that? OK, Now for the slightly surprising result. Given that you have a significant result and you're very happy, what's the chance that this significant result represents a real association. This was the question I asked you at the start. Now if you asked most scientists this question they would immediately say well its 95%, because I set the false positive rate to 5%. They're telling you the rate of true negatives, but what you want to know
- is given I have a significant result what is the probability that it is a true positive. And this is the probability of a true positive, divided by the probability of obtaining a significant result.
- The probability that your significant result is true depends on power - it depends on the size of the effect you are trying to detect, and it also depends on your sample size. If the power to detect the effect was 5% then the chance that your significant result is true is only about 50%, no where near 95%.
- Unfortunately the reality is even worse than this. This equation assumes that the null hypothesis and the alternate hypothesis are equally likely. What we have to consider is the prior probability of the two hypotheses, something that may be hard to do other than qualitatively. We need to multiply these quantities by the prior probability that the association is real, and modify our equation.

False Positive Report Probability

- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)

	Significant	Not Signficant
No Association	[False Positive] α	[True Negative] $(1-\alpha)$

False Positive Report Probability
• α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
No Association | Significant [False Positive] α | Not Significant [True Negative] $(1-\alpha)$

- OK - we're going to have to do a bit of maths. To understand everything I'm going to tell you, you need to understand three quantities. The first is the significance-level, usually denoted as α. This is the probability of a statistically significant finding given that the null hypothesis is true: OK, its the chance of getting a false positive and you usually set it in advance. Typically, we use a significance-level of 0.05: if there's less than a 1 in 20 chance that the observed pattern could have been generated under the null model we would declare that a significant finding.
- The second quantity, which I'm sure you've all heard of, but your probably a little less familiar with is power. The power of a test is if the alternative hypothesis is true what is the probability that you will declare it significant. And unlike the significance-level, the power depends on your sample size and effect size. Power goes up if the effect size is large or the sample size is large: you don't need many mice and many elephants to say that elephants are probably bigger than mice, if you wanted to know whether voles were bigger than mice you'd probably need a larger sample size.
- The opposite of power is the Type II error rate usually denoted β: the chance that you declare something non-significant when in fact the association is real.
- Everybody happy with that? OK, Now for the slightly surprising result. Given that you have a significant result and you're very happy, what's the chance that this significant result represents a real association. This was the question I asked you at the start. Now if you asked most scientists this question they would immediately say well its 95%, because I set the false positive rate to 5%. They're telling you the rate of true negatives, but what you want to know
- is given I have a significant result what is the probability that it is a true positive. And this is the probability of a true positive, divided by the probability of obtaining a significant result.
- The probability that your significant result is true depends on power - it depends on the size of the effect you are trying to detect, and it also depends on your sample size. If the power to detect the effect was 5% then the chance that your significant result is true is only about 50%, no where near 95%.
- Unfortunately the reality is even worse than this. This equation assumes that the null hypothesis and the alternate hypothesis are equally likely. What we have to consider is the prior probability of the two hypotheses, something that may be hard to do other than qualitatively. We need to multiply these quantities by the prior probability that the association is real, and modify our equation.

False Positive Report Probability

- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true (β is the Type II error rate)

	Significant	Not Signficant
No Association	[False Positive] α	[True Negative] $(1 - \alpha)$

False Positive Report Probability
- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true (β is the Type II error rate)

	Significant	Not Significant
No Association	[False Positive] α	[True Negative] $(1 - \alpha)$

- OK - we're going to have to do a bit of maths. To understand everything I'm going to tell you, you need to understand three quantities. The first is the significance-level, usually denoted as α. This is the probability of a statistically significant finding given that the null hypothesis is true: OK, its the chance of getting a false positive and you usually set it in advance. Typically, we use a significance-level of 0.05: if there's less than a 1 in 20 chance that the observed pattern could have been generated under the null model we would declare that a significant finding.
- The second quantity, which I'm sure you've all heard of, but your probably a little less familiar with is power. The power of a test is if the alternative hypothesis is true what is the probability that you will declare it significant. And unlike the significance-level, the power depends on your sample size and effect size. Power goes up if the effect size is large or the sample size is large: you don't need many mice and many elephants to say that elephants are probably bigger than mice, if you wanted to know whether voles were bigger than mice you'd probably need a larger sample size.
- The opposite of power is the Type II error rate usually denoted β: the chance that you declare something non-significant when in fact the association is real.
- Everybody happy with that? OK, Now for the slightly surprising result. Given that you have a significant result and you're very happy, what's the chance that this significant result represents a real association. This was the question I asked you at the start. Now if you asked most scientists this question they would immediately say well its 95%, because I set the false positive rate to 5%. They're telling you the rate of true negatives, but what you want to know
- is given I have a significant result what is the probability that it is a true positive. And this is the probability of a true positive, divided by the probability of obtaining a significant result.
- The probability that your significant result is true depends on power - it depends on the size of the effect you are trying to detect, and it also depends on your sample size. If the power to detect the effect was 5% then the chance that your significant result is true is only about 50%, no where near 95%.
- Unfortunately the reality is even worse than this. This equation assumes that the null hypothesis and the alternate hypothesis are equally likely. What we have to consider is the prior probability of the two hypotheses, something that may be hard to do other than qualitatively. We need to multiply these quantities by the prior probability that the association is real, and modify our equation.

False Positive Report Probability

- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true (β is the Type II error rate)

	Significant	Not Signficant
No Association	[False Positive] α	[True Negative] $(1 - \alpha)$
True Association	[True Positive] $(1 - \beta)$	[False Negative] β

False Positive Report Probability
- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true (β is the Type II error rate)

	Significant	Not Significant
No Association	[False Positive] α	[True Negative] $(1 - \alpha)$
True Association	[True Positive] $(1 - \beta)$	[False Negative] β

- OK - we're going to have to do a bit of maths. To understand everything I'm going to tell you, you need to understand three quantities. The first is the significance-level, usually denoted as α. This is the probability of a statistically significant finding given that the null hypothesis is true: OK, its the chance of getting a false positive and you usually set it in advance. Typically, we use a significance-level of 0.05: if there's less than a 1 in 20 chance that the observed pattern could have been generated under the null model we would declare that a significant finding.
- The second quantity, which I'm sure you've all heard of, but your probably a little less familiar with is power. The power of a test is if the alternative hypothesis is true what is the probability that you will declare it significant. And unlike the significance-level, the power depends on your sample size and effect size. Power goes up if the effect size is large or the sample size is large: you don't need many mice and many elephants to say that elephants are probably bigger than mice, if you wanted to know whether voles were bigger than mice you'd probably need a larger sample size.
- The opposite of power is the Type II error rate usually denoted β: the chance that you declare something non-significant when in fact the association is real.
- Everybody happy with that? OK, Now for the slightly surprising result. Given that you have a significant result and you're very happy, what's the chance that this significant result represents a real association. This was the question I asked you at the start. Now if you asked most scientists this question they would immediately say well its 95%, because I set the false positive rate to 5%. They're telling you the rate of true negatives, but what you want to know
- is given I have a significant result what is the probability that it is a true positive. And this is the probability of a true positive, divided by the probability of obtaining a significant result.
- The probability that your significant result is true depends on power - it depends on the size of the effect you are trying to detect, and it also depends on your sample size. If the power to detect the effect was 5% then the chance that your significant result is true is only about 50%, no where near 95%.
- Unfortunately the reality is even worse than this. This equation assumes that the null hypothesis and the alternate hypothesis are equally likely. What we have to consider is the prior probability of the two hypotheses, something that may be hard to do other than qualitatively. We need to multiply these quantities by the prior probability that the association is real, and modify our equation.

False Positive Report Probability

- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true (β is the Type II error rate)

	Significant	Not Signficant
No Association	[False Positive] α	[True Negative] $(1 - \alpha)$
True Association	[True Positive] $(1 - \beta)$	[False Negative] β

Probability the alternative hypothesis is true *given* a significant result:

False Positive Report Probability

- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true (β is the Type II error rate)

	Significant	Not Significant
No Association	[False Positive] α	[True Negative] $(1 - \alpha)$
True Association	[True Positive] $(1 - \beta)$	[False Negative] β

Probability the alternative hypothesis is true given a significant result:

- OK - we're going to have to do a bit of maths. To understand everything I'm going to tell you, you need to understand three quantities. The first is the significance-level, usually denoted as α. This is the probability of a statistically significant finding given that the null hypothesis is true: OK, its the chance of getting a false positive and you usually set it in advance. Typically, we use a significance-level of 0.05: if there's less than a 1 in 20 chance that the observed pattern could have been generated under the null model we would declare that a significant finding.
- The second quantity, which I'm sure you've all heard of, but your probably a little less familiar with is power. The power of a test is if the alternative hypothesis is true what is the probability that you will declare it significant. And unlike the significance-level, the power depends on your sample size and effect size. Power goes up if the effect size is large or the sample size is large: you don't need many mice and many elephants to say that elephants are probably bigger than mice, if you wanted to know whether voles were bigger than mice you'd probably need a larger sample size.
- The opposite of power is the Type II error rate usually denoted β: the chance that you declare something non-significant when in fact the association is real.
- Everybody happy with that? OK, Now for the slightly surprising result. Given that you have a significant result and you're very happy, what's the chance that this significant result represents a real association. This was the question I asked you at the start. Now if you asked most scientists this question they would immediately say well its 95%, because I set the false positive rate to 5%. They're telling you the rate of true negatives, but what you want to know
- is given I have a significant result what is the probability that it is a true positive. And this is the probability of a true positive, divided by the probability of obtaining a significant result.
- The probability that your significant result is true depends on power - it depends on the size of the effect you are trying to detect, and it also depends on your sample size. If the power to detect the effect was 5% then the chance that your significant result is true is only about 50%, no where near 95%.
- Unfortunately the reality is even worse than this. This equation assumes that the null hypothesis and the alternate hypothesis are equally likely. What we have to consider is the prior probability of the two hypotheses, something that may be hard to do other than qualitatively. We need to multiply these quantities by the prior probability that the association is real, and modify our equation.

False Positive Report Probability

- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true (β is the Type II error rate)

	Significant	Not Signficant
No Association	[False Positive] α	[True Negative] $(1 - \alpha)$
True Association	[True Positive] $(1 - \beta)$	[False Negative] β

Probability the alternative hypothesis is true *given* a significant result:

$$\frac{(1 - \beta)}{(1 - \beta) \quad + \alpha}$$

False Positive Report Probability

- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true (β is the Type II error rate)

	Significant	Not Signficant
No Association	[False Positive] α	[True Negative] $(1 - \alpha)$
True Association	[True Positive] $(1 - \beta)$	[False Negative] β

Probability the alternative hypothesis is true given a significant result:

$$\frac{(1 - \beta)}{(1 - \beta) \quad + \alpha}$$

- OK - we're going to have to do a bit of maths. To understand everything I'm going to tell you, you need to understand three quantities. The first is the significance-level, usually denoted as α. This is the probability of a statistically significant finding given that the null hypothesis is true: OK, its the chance of getting a false positive and you usually set it in advance. Typically, we use a significance-level of 0.05: if there's less than a 1 in 20 chance that the observed pattern could have been generated under the null model we would declare that a significant finding.
- The second quantity, which I'm sure you've all heard of, but your probably a little less familiar with is power. The power of a test is if the alternative hypothesis is true what is the probability that you will declare it significant. And unlike the significance-level, the power depends on your sample size and effect size. Power goes up if the effect size is large or the sample size is large: you don't need many mice and many elephants to say that elephants are probably bigger than mice, if you wanted to know whether voles were bigger than mice you'd probably need a larger sample size.
- The opposite of power is the Type II error rate usually denoted β: the chance that you declare something non-significant when in fact the association is real.
- Everybody happy with that? OK, Now for the slightly surprising result. Given that you have a significant result and you're very happy, what's the chance that this significant result represents a real association. This was the question I asked you at the start. Now if you asked most scientists this question they would immediately say well its 95%, because I set the false positive rate to 5%. They're telling you the rate of true negatives, but what you want to know
- is given I have a significant result what is the probability that it is a true positive. And this is the probability of a true positive, divided by the probability of obtaining a significant result.
- The probability that your significant result is true depends on power - it depends on the size of the effect you are trying to detect, and it also depends on your sample size. If the power to detect the effect was 5% then the chance that your significant result is true is only about 50%, no where near 95%.
- Unfortunately the reality is even worse than this. This equation assumes that the null hypothesis and the alternate hypothesis are equally likely. What we have to consider is the prior probability of the two hypotheses, something that may be hard to do other than qualitatively. We need to multiply these quantities by the prior probability that the association is real, and modify our equation.

False Positive Report Probability

- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true (β is the Type II error rate)
- π is the prior probability that the alternative hypothesis is true

	Significant	Not Signficant
No Association	[False Positive] α	[True Negative] $(1 - \alpha)$
True Association	[True Positive] $(1 - \beta)$	[False Negative] β

Probability the alternative hypothesis is true *given* a significant result:

$$\frac{(1 - \beta)}{(1 - \beta) \quad + \alpha}$$

False Positive Report Probability

- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true (β is the Type II error rate)
- π is the prior probability that the alternative hypothesis is true

	Significant	Not Significant
No Association	[False Positive] α	[True Negative] $(1 - \alpha)$
True Association	[True Positive] $(1 - \beta)$	[False Negative] β

Probability the alternative hypothesis is true given a significant result:

$$\frac{(1 - \beta)}{(1 - \beta) \quad + \alpha}$$

- OK - we're going to have to do a bit of maths. To understand everything I'm going to tell you, you need to understand three quantities. The first is the significance-level, usually denoted as α. This is the probability of a statistically significant finding given that the null hypothesis is true: OK, its the chance of getting a false positive and you usually set it in advance. Typically, we use a significance-level of 0.05: if there's less than a 1 in 20 chance that the observed pattern could have been generated under the null model we would declare that a significant finding.
- The second quantity, which I'm sure you've all heard of, but your probably a little less familiar with is power. The power of a test is if the alternative hypothesis is true what is the probability that you will declare it significant. And unlike the significance-level, the power depends on your sample size and effect size. Power goes up if the effect size is large or the sample size is large: you don't need many mice and many elephants to say that elephants are probably bigger than mice, if you wanted to know whether voles were bigger than mice you'd probably need a larger sample size.
- The opposite of power is the Type II error rate usually denoted β: the chance that you declare something non-significant when in fact the association is real.
- Everybody happy with that? OK, Now for the slightly surprising result. Given that you have a significant result and you're very happy, what's the chance that this significant result represents a real association. This was the question I asked you at the start. Now if you asked most scientists this question they would immediately say well its 95%, because I set the false positive rate to 5%. They're telling you the rate of true negatives, but what you want to know
- is given I have a significant result what is the probability that it is a true positive. And this is the probability of a true positive, divided by the probability of obtaining a significant result.
- The probability that your significant result is true depends on power - it depends on the size of the effect you are trying to detect, and it also depends on your sample size. If the power to detect the effect was 5% then the chance that your significant result is true is only about 50%, no where near 95%.
- Unfortunately the reality is even worse than this. This equation assumes that the null hypothesis and the alternate hypothesis are equally likely. What we have to consider is the prior probability of the two hypotheses, something that may be hard to do other than qualitatively. We need to multiply these quantities by the prior probability that the association is real, and modify our equation.

False Positive Report Probability

- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true (β is the Type II error rate)
- π is the prior probability that the alternative hypothesis is true

	Significant	Not Signficant
No Association	[False Positive] $\alpha(1-\pi)$	[True Negative] $(1-\alpha)(1-\pi)$
True Association	[True Positive] $(1-\beta)\pi$	[False Negative] $\beta\pi$

Probability the alternative hypothesis is true *given* a significant result:

$$\frac{(1-\beta)}{(1-\beta) \quad + \alpha}$$

False Positive Report Probability

- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true (β is the Type II error rate)
- π is the prior probability that the alternative hypothesis is true

	Significant	Not Significant
No Association	[False Positive] $\alpha(1-\pi)$	[True Negative] $(1-\alpha)(1-\pi)$
True Association	[True Positive] $(1-\beta)\pi$	[False Negative] $\beta\pi$

Probability the alternative hypothesis is true given a significant result:

$$\frac{(1-\beta)}{(1-\beta) \quad + \alpha}$$

- OK - we're going to have to do a bit of maths. To understand everything I'm going to tell you, you need to understand three quantities. The first is the significance-level, usually denoted as α. This is the probability of a statistically significant finding given that the null hypothesis is true: OK, its the chance of getting a false positive and you usually set it in advance. Typically, we use a significance-level of 0.05: if there's less than a 1 in 20 chance that the observed pattern could have been generated under the null model we would declare that a significant finding.
- The second quantity, which I'm sure you've all heard of, but your probably a little less familiar with is power. The power of a test is if the alternative hypothesis is true what is the probability that you will declare it significant. And unlike the significance-level, the power depends on your sample size and effect size. Power goes up if the effect size is large or the sample size is large: you don't need many mice and many elephants to say that elephants are probably bigger than mice, if you wanted to know whether voles were bigger than mice you'd probably need a larger sample size.
- The opposite of power is the Type II error rate usually denoted β: the chance that you declare something non-significant when in fact the association is real.
- Everybody happy with that? OK, Now for the slightly surprising result. Given that you have a significant result and you're very happy, what's the chance that this significant result represents a real association. This was the question I asked you at the start. Now if you asked most scientists this question they would immediately say well its 95%, because I set the false positive rate to 5%. They're telling you the rate of true negatives, but what you want to know
- is given I have a significant result what is the probability that it is a true positive. And this is the probability of a true positive, divided by the probability of obtaining a significant result.
- The probability that your significant result is true depends on power - it depends on the size of the effect you are trying to detect, and it also depends on your sample size. If the power to detect the effect was 5% then the chance that your significant result is true is only about 50%, no where near 95%.
- Unfortunately the reality is even worse than this. This equation assumes that the null hypothesis and the alternate hypothesis are equally likely. What we have to consider is the prior probability of the two hypotheses, something that may be hard to do other than qualitatively. We need to multiply these quantities by the prior probability that the association is real, and modify our equation.

False Positive Report Probability

- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true (β is the Type II error rate)
- π is the prior probability that the alternative hypothesis is true

	Significant	Not Signficant
No Association	[False Positive] $\alpha(1-\pi)$	[True Negative] $(1-\alpha)(1-\pi)$
True Association	[True Positive] $(1-\beta)\pi$	[False Negative] $\beta\pi$

Probability the alternative hypothesis is true *given* a significant result:

$$\frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

False Positive Report Probability
- α is the probability of a statistically significant finding, given that the null hypothesis is true (α is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true (β is the Type II error rate)
- π is the prior probability that the alternative hypothesis is true

	Significant	Not Significant
No Association	[False Positive] $\alpha(1-\pi)$	[True Negative] $(1-\alpha)(1-\pi)$
True Association	[True Positive] $(1-\beta)\pi$	[False Negative] $\beta\pi$

Probability the alternative hypothesis is true given a significant result:

$$\frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- OK - we're going to have to do a bit of maths. To understand everything I'm going to tell you, you need to understand three quantities. The first is the significance-level, usually denoted as α. This is the probability of a statistically significant finding given that the null hypothesis is true: OK, its the chance of getting a false positive and you usually set it in advance. Typically, we use a significance-level of 0.05: if there's less than a 1 in 20 chance that the observed pattern could have been generated under the null model we would declare that a significant finding.
- The second quantity, which I'm sure you've all heard of, but your probably a little less familiar with is power. The power of a test is if the alternative hypothesis is true what is the probability that you will declare it significant. And unlike the significance-level, the power depends on your sample size and effect size. Power goes up if the effect size is large or the sample size is large: you don't need many mice and many elephants to say that elephants are probably bigger than mice, if you wanted to know whether voles were bigger than mice you'd probably need a larger sample size.
- The opposite of power is the Type II error rate usually denoted β: the chance that you declare something non-significant when in fact the association is real.
- Everybody happy with that? OK, Now for the slightly surprising result. Given that you have a significant result and you're very happy, what's the chance that this significant result represents a real association. This was the question I asked you at the start. Now if you asked most scientists this question they would immediately say well its 95%, because I set the false positive rate to 5%. They're telling you the rate of true negatives, but what you want to know
- is given I have a significant result what is the probability that it is a true positive. And this is the probability of a true positive, divided by the probability of obtaining a significant result.
- The probability that your significant result is true depends on power - it depends on the size of the effect you are trying to detect, and it also depends on your sample size. If the power to detect the effect was 5% then the chance that your significant result is true is only about 50%, no where near 95%.
- Unfortunately the reality is even worse than this. This equation assumes that the null hypothesis and the alternate hypothesis are equally likely. What we have to consider is the prior probability of the two hypotheses, something that may be hard to do other than qualitatively. We need to multiply these quantities by the prior probability that the association is real, and modify our equation.

False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

False Positive Report Probability
$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$

- OK, this is the probability that a significant result is true given the two possible outcomes. It doesn't have a name, but 1 minus this is called the false positive report probability. Now the question is, what sort of numbers are we typically dealing with.
- α we know, we usually set it to 0.05
- and lets assume for now our two hypotheses are equally likely.
- Power is a bit more tricky because it depends on sample size and the effect size of the alternate hypothesis.
- The general recommendation is that a study should be designed so it has a power of 0.8 - in 80% of cases the test would be significant if the effect size you wish to detect is the true effect size. Under these conditions the probability that your significant result is in fact real is about 94% - close to people's knee-jerk value of 95%
- So 80% power is what is recommended, but is this actually met? Jacob Cohen was a statistician and psychologist who did a lot pioneering work in meta-analysis, and he categorised effect sizes into small, medium and large. If you measure the association as a correlation between two continuous variables this would be a correlation of 0.1, 0.3, or 0.5. If you have two treatment groups then you would measure it as the difference in the means of the two groups: 0.2 standard deviations, 0.5 standard deviations and 0.8 standard deviations.
- I often here people say that as a rule of thumb you should aim for 30 replicates in a two-way experiment: 15 in each group. For small effect sizes the power is appalling an 8% chance of detection, and even if you do detect something there's a 40% chance its a false positive. It gets a little better if the effect size is medium.

False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.

False Positive Report Probability
$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$
- $\alpha = 0.05$ is usually fixed.

- OK, this is the probability that a significant result is true given the two possible outcomes. It doesn't have a name, but 1 minus this is called the false positive report probability. Now the question is, what sort of numbers are we typically dealing with.
- α we know, we usually set it to 0.05
- and lets assume for now our two hypotheses are equally likely.
- Power is a bit more tricky because it depends on sample size and the effect size of the alternate hypothesis.
- The general recommendation is that a study should be designed so it has a power of 0.8 - in 80% of cases the test would be significant if the effect size you wish to detect is the true effect size. Under these conditions the probability that your significant result is in fact real is about 94% - close to people's knee-jerk value of 95%
- So 80% power is what is recommended, but is this actually met? Jacob Cohen was a statistician and psychologist who did a lot pioneering work in meta-analysis, and he categorised effect sizes into small, medium and large. If you measure the association as a correlation between two continuous variables this would be a correlation of 0.1, 0.3, or 0.5. If you have two treatment groups then you would measure it as the difference in the means of the two groups: 0.2 standard deviations, 0.5 standard deviations and 0.8 standard deviations.
- I often here people say that as a rule of thumb you should aim for 30 replicates in a two-way experiment: 15 in each group. For small effect sizes the power is appalling an 8% chance of detection, and even if you do detect something there's a 40% chance its a false positive. It gets a little better if the effect size is medium.

False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.

False Positive Report Probability
$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$
- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.

- OK, this is the probability that a significant result is true given the two possible outcomes. It doesn't have a name, but 1 minus this is called the false positive report probability. Now the question is, what sort of numbers are we typically dealing with.
- α we know, we usually set it to 0.05
- and lets assume for now our two hypotheses are equally likely.
- Power is a bit more tricky because it depends on sample size and the effect size of the alternate hypothesis.
- The general recommendation is that a study should be designed so it has a power of 0.8 - in 80% of cases the test would be significant if the effect size you wish to detect is the true effect size. Under these conditions the probability that your significant result is in fact real is about 94% - close to people's knee-jerk value of 95%
- So 80% power is what is recommended, but is this actually met? Jacob Cohen was a statistician and psychologist who did a lot pioneering work in meta-analysis, and he categorised effect sizes into small, medium and large. If you measure the association as a correlation between two continuous variables this would be a correlation of 0.1, 0.3, or 0.5. If you have two treatment groups then you would measure it as the difference in the means of the two groups: 0.2 standard deviations, 0.5 standard deviations and 0.8 standard deviations.
- I often here people say that as a rule of thumb you should aim for 30 replicates in a two-way experiment: 15 in each group. For small effect sizes the power is appalling an 8% chance of detection, and even if you do detect something there's a 40% chance its a false positive. It gets a little better if the effect size is medium.

False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

False Positive Report Probability
$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$
- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

- OK, this is the probability that a significant result is true given the two possible outcomes. It doesn't have a name, but 1 minus this is called the false positive report probability. Now the question is, what sort of numbers are we typically dealing with.
- α we know, we usually set it to 0.05
- and lets assume for now our two hypotheses are equally likely.
- Power is a bit more tricky because it depends on sample size and the effect size of the alternate hypothesis.
- The general recommendation is that a study should be designed so it has a power of 0.8 - in 80% of cases the test would be significant if the effect size you wish to detect is the true effect size. Under these conditions the probability that your significant result is in fact real is about 94% - close to people's knee-jerk value of 95%
- So 80% power is what is recommended, but is this actually met? Jacob Cohen was a statistician and psychologist who did a lot pioneering work in meta-analysis, and he categorised effect sizes into small, medium and large. If you measure the association as a correlation between two continuous variables this would be a correlation of 0.1, 0.3, or 0.5. If you have two treatment groups then you would measure it as the difference in the means of the two groups: 0.2 standard deviations, 0.5 standard deviations and 0.8 standard deviations.
- I often here people say that as a rule of thumb you should aim for 30 replicates in a two-way experiment: 15 in each group. For small effect sizes the power is appalling an 8% chance of detection, and even if you do detect something there's a 40% chance its a false positive. It gets a little better if the effect size is medium.

False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

Scenario	Effect Size	Power	1-FPRP
Ideal		0.80	0.94

False Positive Report Probability
$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$
- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

Scenario	Effect Size	Power	1-FPRP
Ideal		0.80	0.94

- OK, this is the probability that a significant result is true given the two possible outcomes. It doesn't have a name, but 1 minus this is called the false positive report probability. Now the question is, what sort of numbers are we typically dealing with.
- α we know, we usually set it to 0.05
- and lets assume for now our two hypotheses are equally likely.
- Power is a bit more tricky because it depends on sample size and the effect size of the alternate hypothesis.
- The general recommendation is that a study should be designed so it has a power of 0.8 - in 80% of cases the test would be significant if the effect size you wish to detect is the true effect size. Under these conditions the probability that your significant result is in fact real is about 94% - close to people's knee-jerk value of 95%
- So 80% power is what is recommended, but is this actually met? Jacob Cohen was a statistician and psychologist who did a lot pioneering work in meta-analysis, and he categorised effect sizes into small, medium and large. If you measure the association as a correlation between two continuous variables this would be a correlation of 0.1, 0.3, or 0.5. If you have two treatment groups then you would measure it as the difference in the means of the two groups: 0.2 standard deviations, 0.5 standard deviations and 0.8 standard deviations.
- I often here people say that as a rule of thumb you should aim for 30 replicates in a two-way experiment: 15 in each group. For small effect sizes the power is appalling an 8% chance of detection, and even if you do detect something there's a 40% chance its a false positive. It gets a little better if the effect size is medium.

False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

Scenario	Effect Size	Power	1-FPRP
Ideal		0.80	0.94

- Cohen (1988) effect size: small (r=0.1, d=0.2) medium (r=0.3, d=0.5) large (r=0.5, d=0.8)

False Positive Report Probability
$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$
- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

Scenario	Effect Size	Power	1-FPRP
Ideal		0.80	0.94

- Cohen (1988) effect size: small (r=0.1, d=0.2) medium (r=0.3, d=0.5) large (r=0.5, d=0.8)

- OK, this is the probability that a significant result is true given the two possible outcomes. It doesn't have a name, but 1 minus this is called the false positive report probability. Now the question is, what sort of numbers are we typically dealing with.
- α we know, we usually set it to 0.05
- and lets assume for now our two hypotheses are equally likely.
- Power is a bit more tricky because it depends on sample size and the effect size of the alternate hypothesis.
- The general recommendation is that a study should be designed so it has a power of 0.8 - in 80% of cases the test would be significant if the effect size you wish to detect is the true effect size. Under these conditions the probability that your significant result is in fact real is about 94% - close to people's knee-jerk value of 95%
- So 80% power is what is recommended, but is this actually met? Jacob Cohen was a statistician and psychologist who did a lot pioneering work in meta-analysis, and he categorised effect sizes into small, medium and large. If you measure the association as a correlation between two continuous variables this would be a correlation of 0.1, 0.3, or 0.5. If you have two treatment groups then you would measure it as the difference in the means of the two groups: 0.2 standard deviations, 0.5 standard deviations and 0.8 standard deviations.
- I often here people say that as a rule of thumb you should aim for 30 replicates in a two-way experiment: 15 in each group. For small effect sizes the power is appalling an 8% chance of detection, and even if you do detect something there's a 40% chance its a false positive. It gets a little better if the effect size is medium.

False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

Scenario	Effect Size	Power	1-FPRP
Ideal		0.80	0.94
Experiment (n=30)	Small	0.08	0.61
Experiment (n=30)	Medium	0.26	0.84

- Cohen (1988) effect size: small (r=0.1, d=0.2) medium (r=0.3, d=0.5) large (r=0.5, d=0.8)

False Positive Report Probability
$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$
- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

Scenario	Effect Size	Power	1-FPRP
Ideal		0.80	0.94
Experiment (n=30)	Small	0.08	0.61
Experiment (n=30)	Medium	0.26	0.84

- Cohen (1988) effect size: small (r=0.1, d=0.2) medium (r=0.3, d=0.5) large (r=0.5, d=0.8)

- OK, this is the probability that a significant result is true given the two possible outcomes. It doesn't have a name, but 1 minus this is called the false positive report probability. Now the question is, what sort of numbers are we typically dealing with.
- α we know, we usually set it to 0.05
- and lets assume for now our two hypotheses are equally likely.
- Power is a bit more tricky because it depends on sample size and the effect size of the alternate hypothesis.
- The general recommendation is that a study should be designed so it has a power of 0.8 - in 80% of cases the test would be significant if the effect size you wish to detect is the true effect size. Under these conditions the probability that your significant result is in fact real is about 94% - close to people's knee-jerk value of 95%
- So 80% power is what is recommended, but is this actually met? Jacob Cohen was a statistician and psychologist who did a lot pioneering work in meta-analysis, and he categorised effect sizes into small, medium and large. If you measure the association as a correlation between two continuous variables this would be a correlation of 0.1, 0.3, or 0.5. If you have two treatment groups then you would measure it as the difference in the means of the two groups: 0.2 standard deviations, 0.5 standard deviations and 0.8 standard deviations.
- I often here people say that as a rule of thumb you should aim for 30 replicates in a two-way experiment: 15 in each group. For small effect sizes the power is appalling an 8% chance of detection, and even if you do detect something there's a 40% chance its a false positive. It gets a little better if the effect size is medium.

False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

Scenario	Effect Size	Power	1-FPRP
Ideal		0.80	0.94
Experiment (n=30)	Small	0.08	0.61
Experiment (n=30)	Medium	0.26	0.84

- Cohen (1988) effect size: small (r=0.1, d=0.2) medium (r=0.3, d=0.5) large (r=0.5, d=0.8)
- Moller & Jennions (2002) and Jennions & Moller (2003): report average effect size and power in ecological and evolutionary studies.

False Positive Report Probability
$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$
- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

Scenario	Effect Size	Power	1-FPRP
Ideal		0.80	0.94
Experiment (n=30)	Small	0.08	0.61
Experiment (n=30)	Medium	0.26	0.84

- Cohen (1988) effect size: small (r=0.1, d=0.2) medium (r=0.3, d=0.5) large (r=0.5, d=0.8)
- Moller & Jennions (2002) and Jennions & Moller (2003): report average effect size and power in ecological and evolutionary studies.

- OK, this is the probability that a significant result is true given the two possible outcomes. It doesn't have a name, but 1 minus this is called the false positive report probability. Now the question is, what sort of numbers are we typically dealing with.
- α we know, we usually set it to 0.05
- and lets assume for now our two hypotheses are equally likely.
- Power is a bit more tricky because it depends on sample size and the effect size of the alternate hypothesis.
- The general recommendation is that a study should be designed so it has a power of 0.8 - in 80% of cases the test would be significant if the effect size you wish to detect is the true effect size. Under these conditions the probability that your significant result is in fact real is about 94% - close to people's knee-jerk value of 95%
- So 80% power is what is recommended, but is this actually met? Jacob Cohen was a statistician and psychologist who did a lot pioneering work in meta-analysis, and he categorised effect sizes into small, medium and large. If you measure the association as a correlation between two continuous variables this would be a correlation of 0.1, 0.3, or 0.5. If you have two treatment groups then you would measure it as the difference in the means of the two groups: 0.2 standard deviations, 0.5 standard deviations and 0.8 standard deviations.
- I often here people say that as a rule of thumb you should aim for 30 replicates in a two-way experiment: 15 in each group. For small effect sizes the power is appalling an 8% chance of detection, and even if you do detect something there's a 40% chance its a false positive. It gets a little better if the effect size is medium.

False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

Scenario	Effect Size	Power	1-FPRP
Ideal		0.80	0.94
Experiment (n=30)	Small	0.08	0.61
Experiment (n=30)	Medium	0.26	0.84
Average	Small	0.20	0.80
Average	Medium	0.50	0.91

- Cohen (1988) effect size: small (r=0.1, d=0.2) medium (r=0.3, d=0.5) large (r=0.5, d=0.8)
- Moller & Jennions (2002) and Jennions & Moller (2003): report average effect size and power in ecological and evolutionary studies.

False Positive Report Probability
$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$
- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

Scenario	Effect Size	Power	1-FPRP
Ideal		0.80	0.94
Experiment (n=30)	Small	0.08	0.61
Experiment (n=30)	Medium	0.26	0.84
Average	Small	0.20	0.80
Average	Medium	0.50	0.91

- Cohen (1988) effect size: small (r=0.1, d=0.2) medium (r=0.3, d=0.5) large (r=0.5, d=0.8)
- Moller & Jennions (2002) and Jennions & Moller (2003): report average effect size and power in ecological and evolutionary studies.

- Studies of this size are actually a little smaller than the average study in ecology and evolution. Moller and Jennions looked at the average power of studies to detect small and medium effects and the power was a little better: not quite the recommended 80% but still better. The chance that a significant result indicates a true relationship for these studies is about 4 out of 5 or 9 out of 10.
- This is a lot better than what you cynics believe, and a lot better than what replcation studies suggest. But, you are thinking, are the null and alternate hypotheses equally likely? My PhD was on blue tit plumage coloration, and just before I started a nature paper came out showing that if you change the colour of a male blue tit's head his mate will produce less sons - the argument being that if you are mated to an ugly male you don't want to produce ugly sons. That was the general conclusion, but it wasn't actually that clear cut; it turns out that the treatment only 'works' for males that naturally had very blue heads before you painted them and you also need to control for what type of woodland they are in and his age, and her age. So what's the prior probability of that? Its hard to know, but we can probably break it down. What's the chance that blue tits can even facultatively bias their sex ratio? 50:50, 1 in 10, 1 in a 100 any one want to hazard a guess? Lets be kind: 1 in 10. (I think its probably closer to one in a thousand). Given they can facultatively adjust their sex what is the chance that they would do so if the colour of their mate was experimentally changed. Perhaps quite high actually, let's say 0.5, and given they do this what is the chance that they would only do so if the male was originally bright? Again we can be kind, lets say 1 in 5. So altogether I would say the prior probability that their finding is true is about 1 in a 100 (and I am being generous).

False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

Scenario	Effect Size	$\pi = 0.5$
Ideal		0.94
Experiment (n=30)	Small	0.61
Experiment (n=30)	Medium	0.84
Average	Small	0.80
Average	Medium	0.91

False Positive Report Probability
$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$
Scenario | Effect Size | $\pi = 0.5$
Ideal | | 0.94
Experiment (n=30) | Small | 0.61
Experiment (n=30) | Medium | 0.84
Average | Small | 0.80
Average | Medium | 0.91

- This was the probability that a significant result was true when the odds of the two hypotheses were equal. The blue tit experiment was small, and its hard to believe that the effect would be anything other than small.
- If the prior odds were one in 10 most likely the result is not true - a 15% chance.
- If the odds are one in hundred the probability that it was a true positive starts is just a handful of percent.
- and these are best case scenarios. You've designed an experiment and you've performed a single test. Now its clear from the blue tit paper that the intention was to see if males had less sons if you changed the colour of their head, and yet this isn't what they find: they find that on average there's no effect but there is an interaction, only males that were originally very blue are affected by the treatment. This is, I think, clearly post-hoc.
- Now lets say that the authors strategy was to try out lets say 5 different interactions, and if one turned out to be significant that is the interaction they would publish and talk about. Now if they do this α is no longer 0.05, what is the probability of getting at least one false positive if, in reality, all five null hypotheses are true: its about 23%. There's a 23% chance you will get a significant result just by chance and what then: the chance it is true is then much less than 1%.
- If you keep digging you will find something

False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

Scenario	Effect Size	$\pi = 0.5$	$\pi = 0.1$
Ideal		0.94	0.64
Experiment (n=30)	Small	0.61	0.15
Experiment (n=30)	Medium	0.84	0.37
Average	Small	0.80	0.31
Average	Medium	0.91	0.53

False Positive Report Probability
$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$
Scenario | Effect Size | $\pi = 0.5$ | $\pi = 0.1$
Ideal | | 0.94 | 0.64
Experiment (n=30) | Small | 0.61 | 0.15
Experiment (n=30) | Medium | 0.84 | 0.37
Average | Small | 0.80 | 0.31
Average | Medium | 0.91 | 0.53

- This was the probability that a significant result was true when the odds of the two hypotheses were equal. The blue tit experiment was small, and its hard to believe that the effect would be anything other than small.
- If the prior odds were one in 10 most likely the result is not true - a 15% chance.
- If the odds are one in hundred the probability that it was a true positive starts is just a handful of percent.
- and these are best case scenarios. You've designed an experiment and you've performed a single test. Now its clear from the blue tit paper that the intention was to see if males had less sons if you changed the colour of their head, and yet this isn't what they find: they find that on average there's no effect but there is an interaction, only males that were originally very blue are affected by the treatment. This is, I think, clearly post-hoc.
- Now lets say that the authors strategy was to try out lets say 5 different interactions, and if one turned out to be significant that is the interaction they would publish and talk about. Now if they do this α is no longer 0.05, what is the probability of getting at least one false positive if, in reality, all five null hypotheses are true: its about 23%. There's a 23% chance you will get a significant result just by chance and what then: the chance it is true is then much less than 1%.
- If you keep digging you will find something

False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

Scenario	Effect Size	$\pi = 0.5$	$\pi = 0.1$	$\pi = 0.01$
Ideal		0.94	0.64	0.14
Experiment (n=30)	Small	0.61	0.15	0.02
Experiment (n=30)	Medium	0.84	0.37	0.05
Average	Small	0.80	0.31	0.04
Average	Medium	0.91	0.53	0.09

False Positive Report Probability
$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$
Scenario | Effect Size | $\pi = 0.5$ | $\pi = 0.1$ | $\pi = 0.01$
Ideal | | 0.94 | 0.64 | 0.14
Experiment (n=30) | Small | 0.61 | 0.15 | 0.02
Experiment (n=30) | Medium | 0.84 | 0.37 | 0.05
Average | Small | 0.80 | 0.31 | 0.04
Average | Medium | 0.91 | 0.53 | 0.09

- This was the probability that a significant result was true when the odds of the two hypotheses were equal. The blue tit experiment was small, and its hard to believe that the effect would be anything other than small.
- If the prior odds were one in 10 most likely the result is not true - a 15% chance.
- If the odds are one in hundred the probability that it was a true positive starts is just a handful of percent.
- and these are best case scenarios. You've designed an experiment and you've performed a single test. Now its clear from the blue tit paper that the intention was to see if males had less sons if you changed the colour of their head, and yet this isn't what they find: they find that on average there's no effect but there is an interaction, only males that were originally very blue are affected by the treatment. This is, I think, clearly post-hoc.
- Now lets say that the authors strategy was to try out lets say 5 different interactions, and if one turned out to be significant that is the interaction they would publish and talk about. Now if they do this α is no longer 0.05, what is the probability of getting at least one false positive if, in reality, all five null hypotheses are true: its about 23%. There's a 23% chance you will get a significant result just by chance and what then: the chance it is true is then much less than 1%.
- If you keep digging you will find something

False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

Scenario	Effect Size	$\pi = 0.5$	$\pi = 0.1$	$\pi = 0.01$	$\alpha = 0.23$
Ideal		0.94	0.64	0.14	0.034
Experiment (n=30)	Small	0.61	0.15	0.02	0.003
Experiment (n=30)	Medium	0.84	0.37	0.05	0.011
Average	Small	0.80	0.31	0.04	0.009
Average	Medium	0.91	0.53	0.09	0.021

False Positive Report Probability
$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$
Scenario | Effect Size | $\pi = 0.5$ | $\pi = 0.1$ | $\pi = 0.01$ | $\alpha = 0.23$
Ideal | | 0.94 | 0.64 | 0.14 | 0.034
Experiment (n=30) | Small | 0.61 | 0.15 | 0.02 | 0.003
Experiment (n=30) | Medium | 0.84 | 0.37 | 0.05 | 0.011
Average | Small | 0.80 | 0.31 | 0.04 | 0.009
Average | Medium | 0.91 | 0.53 | 0.09 | 0.021

- This was the probability that a significant result was true when the odds of the two hypotheses were equal. The blue tit experiment was small, and its hard to believe that the effect would be anything other than small.
- If the prior odds were one in 10 most likely the result is not true - a 15% chance.
- If the odds are one in hundred the probability that it was a true positive starts is just a handful of percent.
- and these are best case scenarios. You've designed an experiment and you've performed a single test. Now its clear from the blue tit paper that the intention was to see if males had less sons if you changed the colour of their head, and yet this isn't what they find: they find that on average there's no effect but there is an interaction, only males that were originally very blue are affected by the treatment. This is, I think, clearly post-hoc.
- Now lets say that the authors strategy was to try out lets say 5 different interactions, and if one turned out to be significant that is the interaction they would publish and talk about. Now if they do this α is no longer 0.05, what is the probability of getting at least one false positive if, in reality, all five null hypotheses are true: its about 23%. There's a 23% chance you will get a significant result just by chance and what then: the chance it is true is then much less than 1%.
- If you keep digging you will find something

How Science Goes Wrong

How Science Goes Wrong

- This is Henry Stanhope from Shropshire. He's 14, and after digging up serveral thousand potatoes he eventually found one that looked like a teddy bear. He looks pretty happy with it, but do you think he should try and publish his finding in Nature? Should he waste your time and intelectual effort trying to understand what this potato means.
- No. Jospeh Simmons published this influential article in Psychology, making the simple and obvious point that the greater the fexibility you give people to 'find' significant results, the more likely they are to do so with a concomitant rise in the type I error rate. The paper coined the phrase 'Researcher degrees of freedom': basically how many substantive decisions were made during the course of data collection, analysis, presentation and publication. Their recommendation was that these decisions should be minimised and reported, and the case for publication should depend on them.
- And they're not just talking about major decisions like shall we change our hypothesis, but also little decsisions like should I include moderator variables like age, should I fit interactions, should I keep block effects even if they are not siginigicant and so on.

How Science Goes Wrong

- **Simmons, JP.** (2011) *False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant* Psychological Science 22 1359-1366
- *Researcher degrees of freedom*: how many decisions were made during the course of data collection, analysis, presentation and publication.

How Science Goes Wrong

- **Simmons, JP.** (2011) *False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant* Psychological Science 22 1359-1366
- *Researcher degrees of freedom*: how many decisions were made during the course of data collection, analysis, presentation and publication.

- This is Henry Stanhope from Shropshire. He's 14, and after digging up serveral thousand potatoes he eventually found one that looked like a teddy bear. He looks pretty happy with it, but do you think he should try and publish his finding in Nature? Should he waste your time and intelectual effort trying to understand what this potato means.
- No. Jospeh Simmons published this influential article in Psychology, making the simple and obvious point that the greater the fexibility you give people to 'find' significant results, the more likely they are to do so with a concomitant rise in the type I error rate. The paper coined the phrase 'Researcher degrees of freedom': basically how many substantive decisions were made during the course of data collection, analysis, presentation and publication. Their recommendation was that these decisions should be minimised and reported, and the case for publication should depend on them.
- And they're not just talking about major decisions like shall we change our hypothesis, but also little decsisions like should I include moderator variables like age, should I fit interactions, should I keep block effects even if they are not siginigicant and so on.

Under-graduate project: Gary Cameron

Under-graduate project: Gary Cameron

- These are big problems and the question is how are we going to fix it. The first issue, I think, is to get people to listen and understand the depth of the problem. So as a bit of fun and hopefully to grab people's attention I designed an under graduate project, which a student called Gary Cameron did in 2014-2015.
- The aim was to see whether qualitative assessments of prior probabilities are useful in predicting false positive rates
- The strategy was reasonably straightforward: Gary picked the 12 journals ranked most highly in zoology, and within each he found eight articles where the authors had claimed in the abstract that one of their results was surprising or unexpected. As we have just seen the word surprising or unexpected should raise alarm bells - the author is inadvertently telling you not to trust their result.
- He then recorded the sample size of that test and the reported p-value
- and then went to the next article in that journal and found an equivalent result that was not reported as surprising or unexpected, and reported its sample size and p-value.
- what you would expect is that if surprising results are all false positives then the p-value shouldn't depend on sample size. If they did depend on sample size the whole of statistics would be worthless: just by increasing sample size you could improve your chances of getting a significant result even under the null-hypothesis!
- As you increase the number of true positives then the p-value should start to depend on sample-size: if there really are true effects then you should start to detect them, and you should start to detect them with greater probability as you increase sample size and therefore power.
- If non-surprising results have a greater proprtion of true positives - because those results are a priori more likely to be true - then the slope should be steeper. Its a simple test.
- So Gary collected the data and looked at the relationship between the two, on the log-log scale because they were both highly skewed. And there we go: there's no relationship between p-value and sample size for surprising results suggesting they're mainly false positives, but there is a negative relationship for non-surpring results suggesting that a greater number of them are real: the difference in the slopes is significant. Gary was pretty chuffed, but if you look there doesn't seem to be enough data points - we should have 96 in each group. Well the issue is that many papers don't actually report their p-values, many just say whether it was under some threshold, below 0.05 or below 0.01, and on careful consideration Gary though it is probably best to omit these studies.
- But what happens if we add them back in: not quite so convincing
- and indeed the slopes differ in the predicted way but it is not significant. This is how easy it is: Gary was a smart student, he'd read alot about false positives and researcher degree's of freedom, he'd got pretty fired up about scientific bad practice, but even then when faced with a dissapointing result like this the temptation to explore was too great.

Under-graduate project: Gary Cameron

Are qualitative assessments of π useful?

Under-graduate project: Gary Cameron
Are qualitative assessments of π useful?

- These are big problems and the question is how are we going to fix it. The first issue, I think, is to get people to listen and understand the depth of the problem. So as a bit of fun and hopefully to grab people's attention I designed an under graduate project, which a student called Gary Cameron did in 2014-2015.
- The aim was to see whether qualitative assessments of prior probabilities are useful in predicting false positive rates
- The strategy was reasonably straightforward: Gary picked the 12 journals ranked most highly in zoology, and within each he found eight articles where the authors had claimed in the abstract that one of their results was surprising or unexpected. As we have just seen the word surprising or unexpected should raise alarm bells - the author is inadvertently telling you not to trust their result.
- He then recorded the sample size of that test and the reported p-value
- and then went to the next article in that journal and found an equivalent result that was not reported as surprising or unexpected, and reported its sample size and p-value.
- what you would expect is that if surprising results are all false positives then the p-value shouldn't depend on sample size. If they did depend on sample size the whole of statistics would be worthless: just by increasing sample size you could improve your chances of getting a significant result even under the null-hypothesis!
- As you increase the number of true positives then the p-value should start to depend on sample-size: if there really are true effects then you should start to detect them, and you should start to detect them with greater probability as you increase sample size and therefore power.
- If non-surprising results have a greater proprtion of true positives - because those results are a priori more likely to be true - then the slope should be steeper. Its a simple test.
- So Gary collected the data and looked at the relationship between the two, on the log-log scale because they were both highly skewed. And there we go: there's no relationship between p-value and sample size for surprising results suggesting they're mainly false positives, but there is a negative relationship for non-surpring results suggesting that a greater number of them are real: the difference in the slopes is significant. Gary was pretty chuffed, but if you look there doesn't seem to be enough data points - we should have 96 in each group. Well the issue is that many papers don't actually report their p-values, many just say whether it was under some threshold, below 0.05 or below 0.01, and on careful consideration Gary though it is probably best to omit these studies.
- But what happens if we add them back in: not quite so convincing
- and indeed the slopes differ in the predicted way but it is not significant. This is how easy it is: Gary was a smart student, he'd read alot about false positives and researcher degree's of freedom, he'd got pretty fired up about scientific bad practice, but even then when faced with a dissapointing result like this the temptation to explore was too great.

Under-graduate project: Gary Cameron

Are qualitative assessments of π useful?

- 12 top zoology journals.

Under-graduate project: Gary Cameron
Are qualitative assessments of π useful?
12 top zoology journals.

- These are big problems and the question is how are we going to fix it. The first issue, I think, is to get people to listen and understand the depth of the problem. So as a bit of fun and hopefully to grab people's attention I designed an under graduate project, which a student called Gary Cameron did in 2014-2015.
- The aim was to see whether qualitative assessments of prior probabilities are useful in predicting false positive rates
- The strategy was reasonably straightforward: Gary picked the 12 journals ranked most highly in zoology, and within each he found eight articles where the authors had claimed in the abstract that one of their results was surprising or unexpected. As we have just seen the word surprising or unexpected should raise alarm bells - the author is inadvertently telling you not to trust their result.
- He then recorded the sample size of that test and the reported p-value
- and then went to the next article in that journal and found an equivalent result that was not reported as surprising or unexpected, and reported its sample size and p-value.
- what you would expect is that if surprising results are all false positives then the p-value shouldn't depend on sample size. If they did depend on sample size the whole of statistics would be worthless: just by increasing sample size you could improve your chances of getting a significant result even under the null-hypothesis!
- As you increase the number of true positives then the p-value should start to depend on sample-size: if there really are true effects then you should start to detect them, and you should start to detect them with greater probability as you increase sample size and therefore power.
- If non-surprising results have a greater proprtion of true positives - because those results are a priori more likely to be true - then the slope should be steeper. Its a simple test.
- So Gary collected the data and looked at the relationship between the two, on the log-log scale because they were both highly skewed. And there we go: there's no relationship between p-value and sample size for surprising results suggesting they're mainly false positives, but there is a negative relationship for non-surpring results suggesting that a greater number of them are real: the difference in the slopes is significant. Gary was pretty chuffed, but if you look there doesn't seem to be enough data points - we should have 96 in each group. Well the issue is that many papers don't actually report their p-values, many just say whether it was under some threshold, below 0.05 or below 0.01, and on careful consideration Gary though it is probably best to omit these studies.
- But what happens if we add them back in: not quite so convincing
- and indeed the slopes differ in the predicted way but it is not significant. This is how easy it is: Gary was a smart student, he'd read alot about false positives and researcher degree's of freedom, he'd got pretty fired up about scientific bad practice, but even then when faced with a dissapointing result like this the temptation to explore was too great.

Under-graduate project: Gary Cameron

Are qualitative assessments of π useful?

- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.

Under-graduate project: Gary Cameron

Are qualitative assessments of π useful?

- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.

- These are big problems and the question is how are we going to fix it. The first issue, I think, is to get people to listen and understand the depth of the problem. So as a bit of fun and hopefully to grab people's attention I designed an under graduate project, which a student called Gary Cameron did in 2014-2015.
- The aim was to see whether qualitative assessments of prior probabilities are useful in predicting false positive rates
- The strategy was reasonably straightforward: Gary picked the 12 journals ranked most highly in zoology, and within each he found eight articles where the authors had claimed in the abstract that one of their results was surprising or unexpected. As we have just seen the word surprising or unexpected should raise alarm bells - the author is inadvertently telling you not to trust their result.
- He then recorded the sample size of that test and the reported p-value
- and then went to the next article in that journal and found an equivalent result that was not reported as surprising or unexpected, and reported its sample size and p-value.
- what you would expect is that if surprising results are all false positives then the p-value shouldn't depend on sample size. If they did depend on sample size the whole of statistics would be worthless: just by increasing sample size you could improve your chances of getting a significant result even under the null-hypothesis!
- As you increase the number of true positives then the p-value should start to depend on sample-size: if there really are true effects then you should start to detect them, and you should start to detect them with greater probability as you increase sample size and therefore power.
- If non-surprising results have a greater proprtion of true positives - because those results are a priori more likely to be true - then the slope should be steeper. Its a simple test.
- So Gary collected the data and looked at the relationship between the two, on the log-log scale because they were both highly skewed. And there we go: there's no relationship between p-value and sample size for surprising results suggesting they're mainly false positives, but there is a negative relationship for non-surpring results suggesting that a greater number of them are real: the difference in the slopes is significant. Gary was pretty chuffed, but if you look there doesn't seem to be enough data points - we should have 96 in each group. Well the issue is that many papers don't actually report their p-values, many just say whether it was under some threshold, below 0.05 or below 0.01, and on careful consideration Gary though it is probably best to omit these studies.
- But what happens if we add them back in: not quite so convincing
- and indeed the slopes differ in the predicted way but it is not significant. This is how easy it is: Gary was a smart student, he'd read alot about false positives and researcher degree's of freedom, he'd got pretty fired up about scientific bad practice, but even then when faced with a dissapointing result like this the temptation to explore was too great.

Under-graduate project: Gary Cameron

Are qualitative assessments of π useful?

- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test

Under-graduate project: Gary Cameron
Are qualitative assessments of π useful?
- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test

- These are big problems and the question is how are we going to fix it. The first issue, I think, is to get people to listen and understand the depth of the problem. So as a bit of fun and hopefully to grab people's attention I designed an under graduate project, which a student called Gary Cameron did in 2014-2015.
- The aim was to see whether qualitative assessments of prior probabilities are useful in predicting false positive rates
- The strategy was reasonably straightforward: Gary picked the 12 journals ranked most highly in zoology, and within each he found eight articles where the authors had claimed in the abstract that one of their results was surprising or unexpected. As we have just seen the word surprising or unexpected should raise alarm bells - the author is inadvertently telling you not to trust their result.
- He then recorded the sample size of that test and the reported p-value
- and then went to the next article in that journal and found an equivalent result that was not reported as surprising or unexpected, and reported its sample size and p-value.
- what you would expect is that if surprising results are all false positives then the p-value shouldn't depend on sample size. If they did depend on sample size the whole of statistics would be worthless: just by increasing sample size you could improve your chances of getting a significant result even under the null-hypothesis!
- As you increase the number of true positives then the p-value should start to depend on sample-size: if there really are true effects then you should start to detect them, and you should start to detect them with greater probability as you increase sample size and therefore power.
- If non-surprising results have a greater proprtion of true positives - because those results are a priori more likely to be true - then the slope should be steeper. Its a simple test.
- So Gary collected the data and looked at the relationship between the two, on the log-log scale because they were both highly skewed. And there we go: there's no relationship between p-value and sample size for surprising results suggesting they're mainly false positives, but there is a negative relationship for non-surpring results suggesting that a greater number of them are real: the difference in the slopes is significant. Gary was pretty chuffed, but if you look there doesn't seem to be enough data points - we should have 96 in each group. Well the issue is that many papers don't actually report their p-values, many just say whether it was under some threshold, below 0.05 or below 0.01, and on careful consideration Gary though it is probably best to omit these studies.
- But what happens if we add them back in: not quite so convincing
- and indeed the slopes differ in the predicted way but it is not significant. This is how easy it is: Gary was a smart student, he'd read alot about false positives and researcher degree's of freedom, he'd got pretty fired up about scientific bad practice, but even then when faced with a dissapointing result like this the temptation to explore was too great.

Under-graduate project: Gary Cameron

Are qualitative assessments of π useful?

- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test

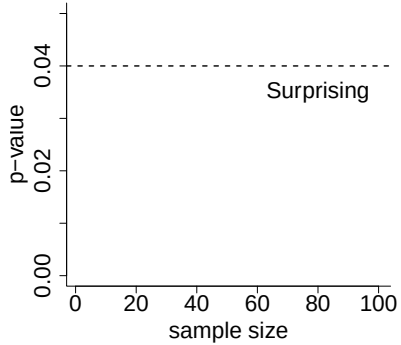
Under-graduate project: Gary Cameron
Are qualitative assessments of π useful?
- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test

- These are big problems and the question is how are we going to fix it. The first issue, I think, is to get people to listen and understand the depth of the problem. So as a bit of fun and hopefully to grab people's attention I designed an under graduate project, which a student called Gary Cameron did in 2014-2015.
- The aim was to see whether qualitative assessments of prior probabilities are useful in predicting false positive rates
- The strategy was reasonably straightforward: Gary picked the 12 journals ranked most highly in zoology, and within each he found eight articles where the authors had claimed in the abstract that one of their results was surprising or unexpected. As we have just seen the word surprising or unexpected should raise alarm bells - the author is inadvertently telling you not to trust their result.
- He then recorded the sample size of that test and the reported p-value
- and then went to the next article in that journal and found an equivalent result that was not reported as surprising or unexpected, and reported its sample size and p-value.
- what you would expect is that if surprising results are all false positives then the p-value shouldn't depend on sample size. If they did depend on sample size the whole of statistics would be worthless: just by increasing sample size you could improve your chances of getting a significant result even under the null-hypothesis!
- As you increase the number of true positives then the p-value should start to depend on sample-size: if there really are true effects then you should start to detect them, and you should start to detect them with greater probability as you increase sample size and therefore power.
- If non-surprising results have a greater proprtion of true positives - because those results are a priori more likely to be true - then the slope should be steeper. Its a simple test.
- So Gary collected the data and looked at the relationship between the two, on the log-log scale because they were both highly skewed. And there we go: there's no relationship between p-value and sample size for surprising results suggesting they're mainly false positives, but there is a negative relationship for non-surpring results suggesting that a greater number of them are real: the difference in the slopes is significant. Gary was pretty chuffed, but if you look there doesn't seem to be enough data points - we should have 96 in each group. Well the issue is that many papers don't actually report their p-values, many just say whether it was under some threshold, below 0.05 or below 0.01, and on careful consideration Gary though it is probably best to omit these studies.
- But what happens if we add them back in: not quite so convincing
- and indeed the slopes differ in the predicted way but it is not significant. This is how easy it is: Gary was a smart student, he'd read alot about false positives and researcher degree's of freedom, he'd got pretty fired up about scientific bad practice, but even then when faced with a dissapointing result like this the temptation to explore was too great.

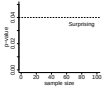
Under-graduate project: Gary Cameron

Are qualitative assessments of π useful?

- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test



Under-graduate project: Gary Cameron
Are qualitative assessments of π useful?
12 top zoology journals.
Find 8 articles in each that self report a surprising result in the abstract.
Record sample size and p-value for that test
Go to the next paper in the journal and find an equivalent result.
Record sample size and p-value for that test

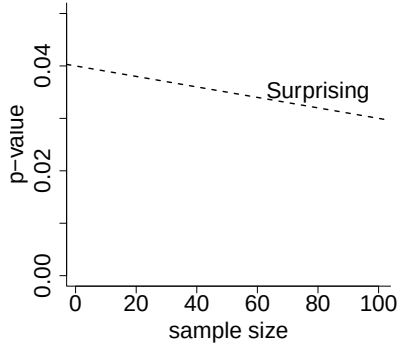


- These are big problems and the question is how are we going to fix it. The first issue, I think, is to get people to listen and understand the depth of the problem. So as a bit of fun and hopefully to grab people's attention I designed an under graduate project, which a student called Gary Cameron did in 2014-2015.
- The aim was to see whether qualitative assessments of prior probabilities are useful in predicting false positive rates
- The strategy was reasonably straightforward: Gary picked the 12 journals ranked most highly in zoology, and within each he found eight articles where the authors had claimed in the abstract that one of their results was surprising or unexpected. As we have just seen the word surprising or unexpected should raise alarm bells - the author is inadvertently telling you not to trust their result.
- He then recorded the sample size of that test and the reported p-value
- and then went to the next article in that journal and found an equivalent result that was not reported as surprising or unexpected, and reported its sample size and p-value.
- what you would expect is that if surprising results are all false positives then the p-value shouldn't depend on sample size. If they did depend on sample size the whole of statistics would be worthless: just by increasing sample size you could improve your chances of getting a significant result even under the null-hypothesis!
- As you increase the number of true positives then the p-value should start to depend on sample-size: if there really are true effects then you should start to detect them, and you should start to detect them with greater probability as you increase sample size and therefore power.
- If non-surprising results have a greater proprtion of true positives - because those results are a priori more likely to be true - then the slope should be steeper. Its a simple test.
- So Gary collected the data and looked at the relationship between the two, on the log-log scale because they were both highly skewed. And there we go: there's no relationship between p-value and sample size for surprising results suggesting they're mainly false positives, but there is a negative relationship for non-surprising results suggesting that a greater number of them are real: the difference in the slopes is significant. Gary was pretty chuffed, but if you look there doesn't seem to be enough data points - we should have 96 in each group. Well the issue is that many papers don't actually report their p-values, many just say whether it was under some threshold, below 0.05 or below 0.01, and on careful consideration Gary though it is probably best to omit these studies.
- But what happens if we add them back in: not quite so convincing
- and indeed the slopes differ in the predicted way but it is not significant. This is how easy it is: Gary was a smart student, he'd read alot about false positives and researcher degree's of freedom, he'd got pretty fired up about scientific bad practice, but even then when faced with a dissapointing result like this the temptation to explore was too great.

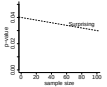
Under-graduate project: Gary Cameron

Are qualitative assessments of π useful?

- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test



Under-graduate project: Gary Cameron
Are qualitative assessments of π useful?
- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test

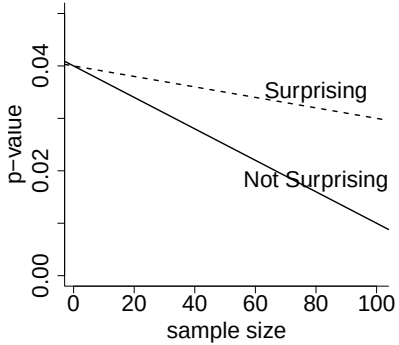


- These are big problems and the question is how are we going to fix it. The first issue, I think, is to get people to listen and understand the depth of the problem. So as a bit of fun and hopefully to grab people's attention I designed an under graduate project, which a student called Gary Cameron did in 2014-2015.
- The aim was to see whether qualitative assessments of prior probabilities are useful in predicting false positive rates
- The strategy was reasonably straightforward: Gary picked the 12 journals ranked most highly in zoology, and within each he found eight articles where the authors had claimed in the abstract that one of their results was surprising or unexpected. As we have just seen the word surprising or unexpected should raise alarm bells - the author is inadvertently telling you not to trust their result.
- He then recorded the sample size of that test and the reported p-value
- and then went to the next article in that journal and found an equivalent result that was not reported as surprising or unexpected, and reported its sample size and p-value.
- what you would expect is that if surprising results are all false positives then the p-value shouldn't depend on sample size. If they did depend on sample size the whole of statistics would be worthless: just by increasing sample size you could improve your chances of getting a significant result even under the null-hypothesis!
- As you increase the number of true positives then the p-value should start to depend on sample-size: if there really are true effects then you should start to detect them, and you should start to detect them with greater probability as you increase sample size and therefore power.
- If non-surprising results have a greater proprtion of true positives - because those results are a priori more likely to be true - then the slope should be steeper. Its a simple test.
- So Gary collected the data and looked at the relationship between the two, on the log-log scale because they were both highly skewed. And there we go: there's no relationship between p-value and sample size for surprising results suggesting they're mainly false positives, but there is a negative relationship for non-surprising results suggesting that a greater number of them are real: the difference in the slopes is significant. Gary was pretty chuffed, but if you look there doesn't seem to be enough data points - we should have 96 in each group. Well the issue is that many papers don't actually report their p-values, many just say whether it was under some threshold, below 0.05 or below 0.01, and on careful consideration Gary though it is probably best to omit these studies.
- But what happens if we add them back in: not quite so convincing
- and indeed the slopes differ in the predicted way but it is not significant. This is how easy it is: Gary was a smart student, he'd read alot about false positives and researcher degree's of freedom, he'd got pretty fired up about scientific bad practice, but even then when faced with a dissapointing result like this the temptation to explore was too great.

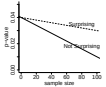
Under-graduate project: Gary Cameron

Are qualitative assessments of π useful?

- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test



Under-graduate project: Gary Cameron
Are qualitative assessments of π useful?
- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test

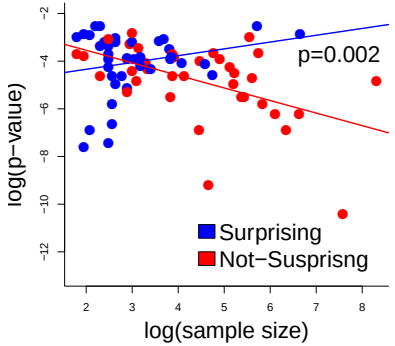


- These are big problems and the question is how are we going to fix it. The first issue, I think, is to get people to listen and understand the depth of the problem. So as a bit of fun and hopefully to grab people's attention I designed an under graduate project, which a student called Gary Cameron did in 2014-2015.
- The aim was to see whether qualitative assessments of prior probabilities are useful in predicting false positive rates
- The strategy was reasonably straightforward: Gary picked the 12 journals ranked most highly in zoology, and within each he found eight articles where the authors had claimed in the abstract that one of their results was surprising or unexpected. As we have just seen the word surprising or unexpected should raise alarm bells - the author is inadvertently telling you not to trust their result.
- He then recorded the sample size of that test and the reported p-value
- and then went to the next article in that journal and found an equivalent result that was not reported as surprising or unexpected, and reported its sample size and p-value.
- what you would expect is that if surprising results are all false positives then the p-value shouldn't depend on sample size. If they did depend on sample size the whole of statistics would be worthless: just by increasing sample size you could improve your chances of getting a significant result even under the null-hypothesis!
- As you increase the number of true positives then the p-value should start to depend on sample-size: if there really are true effects then you should start to detect them, and you should start to detect them with greater probability as you increase sample size and therefore power.
- If non-surprising results have a greater proprtion of true positives - because those results are a priori more likely to be true - then the slope should be steeper. Its a simple test.
- So Gary collected the data and looked at the relationship between the two, on the log-log scale because they were both highly skewed. And there we go: there's no relationship between p-value and sample size for surprising results suggesting they're mainly false positives, but there is a negative relationship for non-surprising results suggesting that a greater number of them are real: the difference in the slopes is significant. Gary was pretty chuffed, but if you look there doesn't seem to be enough data points - we should have 96 in each group. Well the issue is that many papers don't actually report their p-values, many just say whether it was under some threshold, below 0.05 or below 0.01, and on careful consideration Gary though it is probably best to omit these studies.
- But what happens if we add them back in: not quite so convincing
- and indeed the slopes differ in the predicted way but it is not significant. This is how easy it is: Gary was a smart student, he'd read alot about false positives and researcher degree's of freedom, he'd got pretty fired up about scientific bad practice, but even then when faced with a dissapointing result like this the temptation to explore was too great.

Under-graduate project: Gary Cameron

Are qualitative assessments of π useful?

- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test



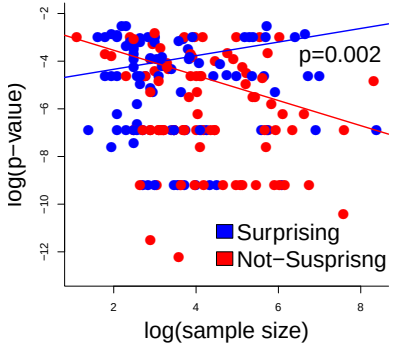
Under-graduate project: Gary Cameron
Are qualitative assessments of π useful?
12 top zoology journals.
Find 8 articles in each that self report a surprising result in the abstract.
Record sample size and p-value for that test
Go to the next paper in the journal and find an equivalent result.
Record sample size and p-value for that test
p=0.002
Surprising
Not−Susprisng
log(p−value)
log(sample size)

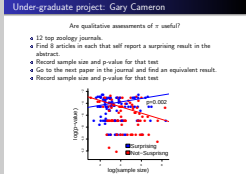
- These are big problems and the question is how are we going to fix it. The first issue, I think, is to get people to listen and understand the depth of the problem. So as a bit of fun and hopefully to grab people's attention I designed an under graduate project, which a student called Gary Cameron did in 2014-2015.
- The aim was to see whether qualitative assessments of prior probabilities are useful in predicting false positive rates
- The strategy was reasonably straightforward: Gary picked the 12 journals ranked most highly in zoology, and within each he found eight articles where the authors had claimed in the abstract that one of their results was surprising or unexpected. As we have just seen the word surprising or unexpected should raise alarm bells - the author is inadvertently telling you not to trust their result.
- He then recorded the sample size of that test and the reported p-value
- and then went to the next article in that journal and found an equivalent result that was not reported as surprising or unexpected, and reported its sample size and p-value.
- what you would expect is that if surprising results are all false positives then the p-value shouldn't depend on sample size. If they did depend on sample size the whole of statistics would be worthless: just by increasing sample size you could improve your chances of getting a significant result even under the null-hypothesis!
- As you increase the number of true positives then the p-value should start to depend on sample-size: if there really are true effects then you should start to detect them, and you should start to detect them with greater probability as you increase sample size and therefore power.
- If non-surprising results have a greater proprtion of true positives - because those results are a priori more likely to be true - then the slope should be steeper. Its a simple test.
- So Gary collected the data and looked at the relationship between the two, on the log-log scale because they were both highly skewed. And there we go: there's no relationship between p-value and sample size for surprising results suggesting they're mainly false positives, but there is a negative relationship for non-surprising results suggesting that a greater number of them are real: the difference in the slopes is significant. Gary was pretty chuffed, but if you look there doesn't seem to be enough data points - we should have 96 in each group. Well the issue is that many papers don't actually report their p-values, many just say whether it was under some threshold, below 0.05 or below 0.01, and on careful consideration Gary though it is probably best to omit these studies.
- But what happens if we add them back in: not quite so convincing
- and indeed the slopes differ in the predicted way but it is not significant. This is how easy it is: Gary was a smart student, he'd read alot about false positives and researcher degree's of freedom, he'd got pretty fired up about scientific bad practice, but even then when faced with a dissapointing result like this the temptation to explore was too great.

Under-graduate project: Gary Cameron

Are qualitative assessments of π useful?

- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test



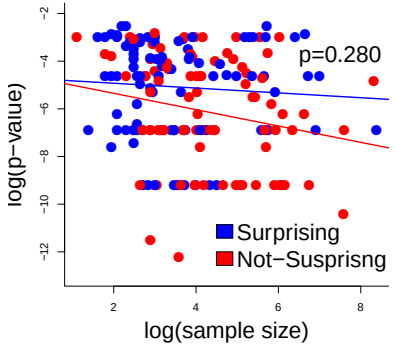


- These are big problems and the question is how are we going to fix it. The first issue, I think, is to get people to listen and understand the depth of the problem. So as a bit of fun and hopefully to grab people's attention I designed an under graduate project, which a student called Gary Cameron did in 2014-2015.
- The aim was to see whether qualitative assessments of prior probabilities are useful in predicting false positive rates
- The strategy was reasonably straightforward: Gary picked the 12 journals ranked most highly in zoology, and within each he found eight articles where the authors had claimed in the abstract that one of their results was surprising or unexpected. As we have just seen the word surprising or unexpected should raise alarm bells - the author is inadvertently telling you not to trust their result.
- He then recorded the sample size of that test and the reported p-value
- and then went to the next article in that journal and found an equivalent result that was not reported as surprising or unexpected, and reported its sample size and p-value.
- what you would expect is that if surprising results are all false positives then the p-value shouldn't depend on sample size. If they did depend on sample size the whole of statistics would be worthless: just by increasing sample size you could improve your chances of getting a significant result even under the null-hypothesis!
- As you increase the number of true positives then the p-value should start to depend on sample-size: if there really are true effects then you should start to detect them, and you should start to detect them with greater probability as you increase sample size and therefore power.
- If non-surprising results have a greater proprtion of true positives - because those results are a priori more likely to be true - then the slope should be steeper. Its a simple test.
- So Gary collected the data and looked at the relationship between the two, on the log-log scale because they were both highly skewed. And there we go: there's no relationship between p-value and sample size for surprising results suggesting they're mainly false positives, but there is a negative relationship for non-surprising results suggesting that a greater number of them are real: the difference in the slopes is significant. Gary was pretty chuffed, but if you look there doesn't seem to be enough data points - we should have 96 in each group. Well the issue is that many papers don't actually report their p-values, many just say whether it was under some threshold, below 0.05 or below 0.01, and on careful consideration Gary though it is probably best to omit these studies.
- But what happens if we add them back in: not quite so convincing
- and indeed the slopes differ in the predicted way but it is not significant. This is how easy it is: Gary was a smart student, he'd read alot about false positives and researcher degree's of freedom, he'd got pretty fired up about scientific bad practice, but even then when faced with a dissapointing result like this the temptation to explore was too great.

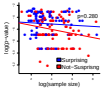
Under-graduate project: Gary Cameron

Are qualitative assessments of π useful?

- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test

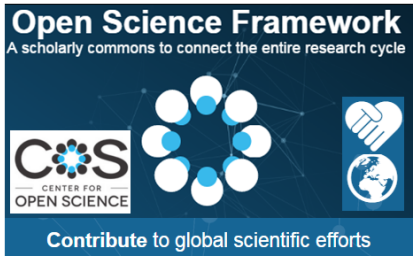


Under-graduate project: Gary Cameron
Are qualitative assessments of π useful?
12 top zoology journals.
Find 8 articles in each that self report a surprising result in the abstract.
Record sample size and p-value for that test
Go to the next paper in the journal and find an equivalent result.
Record sample size and p-value for that test



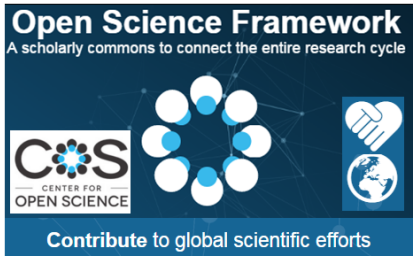
- These are big problems and the question is how are we going to fix it. The first issue, I think, is to get people to listen and understand the depth of the problem. So as a bit of fun and hopefully to grab people's attention I designed an under graduate project, which a student called Gary Cameron did in 2014-2015.
- The aim was to see whether qualitative assessments of prior probabilities are useful in predicting false positive rates
- The strategy was reasonably straightforward: Gary picked the 12 journals ranked most highly in zoology, and within each he found eight articles where the authors had claimed in the abstract that one of their results was surprising or unexpected. As we have just seen the word surprising or unexpected should raise alarm bells - the author is inadvertently telling you not to trust their result.
- He then recorded the sample size of that test and the reported p-value
- and then went to the next article in that journal and found an equivalent result that was not reported as surprising or unexpected, and reported its sample size and p-value.
- what you would expect is that if surprising results are all false positives then the p-value shouldn't depend on sample size. If they did depend on sample size the whole of statistics would be worthless: just by increasing sample size you could improve your chances of getting a significant result even under the null-hypothesis!
- As you increase the number of true positives then the p-value should start to depend on sample-size: if there really are true effects then you should start to detect them, and you should start to detect them with greater probability as you increase sample size and therefore power.
- If non-surprising results have a greater proprtion of true positives - because those results are a priori more likely to be true - then the slope should be steeper. Its a simple test.
- So Gary collected the data and looked at the relationship between the two, on the log-log scale because they were both highly skewed. And there we go: there's no relationship between p-value and sample size for surprising results suggesting they're mainly false positives, but there is a negative relationship for non-surprising results suggesting that a greater number of them are real: the difference in the slopes is significant. Gary was pretty chuffed, but if you look there doesn't seem to be enough data points - we should have 96 in each group. Well the issue is that many papers don't actually report their p-values, many just say whether it was under some threshold, below 0.05 or below 0.01, and on careful consideration Gary though it is probably best to omit these studies.
- But what happens if we add them back in: not quite so convincing
- and indeed the slopes differ in the predicted way but it is not significant. This is how easy it is: Gary was a smart student, he'd read alot about false positives and researcher degree's of freedom, he'd got pretty fired up about scientific bad practice, but even then when faced with a dissapointing result like this the temptation to explore was too great.

Solutions



- So how do we fix the problem. A couple of years ago I was at a meeting hosted by the center for open science in the US. They brought together editors in chief of many evolution and eclogy journals and a handful of expert witnesses like myself. The objective was to change journal policies in order to reduce the rate of false positives. The meeting was a mixed success: it definitely raised awareness of the sort of issues I've talked about, but sadly it seems that most journals are reluctant to do anything substantive about it. So I think any change, at least initially, is going to be from individual scientists like yourselves. Particularly you. The older generation have their head in the sand are up to their neck in false positives, they are probably a lost cause.
- And so what I would advise you to do is pre-register your study at its earliest stages. Write down what you want to test, how you are going to collect the data and how you are going to anlyse it. You're not obliged to stick to pre-registered plans of course, but it does mean you have a record of intent which you can go back to and judge how your decsions impacted the final outcome. I wish I had known about pre-registration during my PHD: it would have stopped me deluding yourself, it would have saved me a lot of time 'exploring' data, and it would have stopped me agonising over whether my significant results really were signifidcant or not. My guess is that some supervisors and bosses will grumble about pre-registration - but its your time their bad science is wasting - so I would say just go ahead and do it anyway: you don't want to spend four years looking at tealeaves in the bottom of a teacup.
- If you want you can stick your preregitstration plans on an online-repository like the one hosted by the open science framework, and it will be made public once you publish the work. And I think people that do this will see a big benefit: as a reviwer, or editor or reader faced with a p-value of 0.049, I would be much more likely to accept it at face value if a study was pregistered.

Solutions



- Digitally pre-register study hypotheses, data collection and analysis plans as a record of intent.

Solutions
Open Science Framework
A scholarly commons to connect the entire research cycle
Contribute to global scientific efforts
Digitally pre-register study hypotheses, data collection and analysis plans as a record of intent.

- So how do we fix the problem. A couple of years ago I was at a meeting hosted by the center for open science in the US. They brought together editors in chief of many evolution and eclogy journals and a handful of expert witnesses like myself. The objective was to change journal policies in order to reduce the rate of false positives. The meeting was a mixed success: it definitely raised awareness of the sort of issues I've talked about, but sadly it seems that most journals are reluctant to do anything substantive about it. So I think any change, at least initially, is going to be from individual scientists like yourselves. Particularly you. The older generation have their head in the sand are up to their neck in false positives, they are probably a lost cause.
- And so what I would advise you to do is pre-register your study at its earliest stages. Write down what you want to test, how you are going to collect the data and how you are going to anlyse it. You're not obliged to stick to pre-registered plans of course, but it does mean you have a record of intent which you can go back to and judge how your decsions impacted the final outcome. I wish I had known about pre-registration during my PHD: it would have stopped me deluding yourself, it would have saved me a lot of time 'exploring' data, and it would have stopped me agonising over whether my significant results really were signifidcant or not. My guess is that some supervisors and bosses will grumble about pre-registration - but its your time their bad science is wasting - so I would say just go ahead and do it anyway: you don't want to spend four years looking at tealeaves in the bottom of a teacup.
- If you want you can stick your preregitstration plans on an online-repository like the one hosted by the open science framework, and it will be made public once you publish the work. And I think people that do this will see a big benefit: as a reviwer, or editor or reader faced with a p-value of 0.049, I would be much more likely to accept it at face value if a study was pregistered.

Unacknowledged problems

- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?

Unacknowledged problems
- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?

Unacknowledged problems

- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?

Should we bring (qualitative) priors into it?

Unacknowledged problems
Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?
Should we bring (qualitative) priors into it?

Unacknowledged problems

- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?

Should we bring (qualitative) priors into it?

- Yes - a surprising result is less likely to be true.

Unacknowledged problems
Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?
Should we bring (qualitative) priors into it?
Yes - a surprising result is less likely to be true.

Unacknowledged problems

- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?

Should we bring (qualitative) priors into it?

- Yes - a surprising result is less likely to be true.
- No - fair and precise assessments are not possible

Unacknowledged problems
- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?

Should we bring (qualitative) priors into it?
- Yes - a surprising result is less likely to be true.
- No - fair and precise assessments are not possible

Unacknowledged problems

- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?

Should we bring (qualitative) priors into it?

- Yes - a surprising result is less likely to be true.
- No - fair and precise assessments are not possible
- Not Sure - the expected information content of a true positive $= -log(\pi)$. Surprising results that are true are worth more.

Unacknowledged problems
- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?
Should we bring (qualitative) priors into it?
- Yes - a surprising result is less likely to be true.
- No - fair and precise assessments are not possible
- Not Sure - the expected information content of a true positive $= -log(\pi)$. Surprising results that are true are worth more.

Unacknowledged problems

- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?

Should we bring (qualitative) priors into it?

- Yes - a surprising result is less likely to be true.
- No - fair and precise assessments are not possible
- Not Sure - the expected information content of a true positive $= -log(\pi)$. Surprising results that are true are worth more.

Unacknowledged problems
- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?
Should we bring (qualitative) priors into it?
- Yes - a surprising result is less likely to be true.
- No - fair and precise assessments are not possible
- Not Sure - the expected information content of a true positive $= -log(\pi)$. Surprising results that are true are worth more.