Biostatistics Lecture 4

Mithat Gonen

Brenden Bready

Today's Lecture

- Review of Hypothesis Testing and p-values
- Type II errors and power
- Review of Goodman paper

Review of Hypothesis Testing

- Suppose we are designing a Phase II study
- Already decided that we will do a single-arm study and use response rate as the primary endpoint
- We think response rate for the standard of care in this population is 50% and our regimen needs to exceed this

Hypothesis Testing

- We can formulate this as follows
    - Hypothesis 0: Response rate is less than or equal to 50%
    - Hypothesis 1: Response rate greater than 50%
- Is the response rate referred to in the hypothesis statement a parameter or an estimate?

- We will use the data to discard one of these and retain the other

Type I and Type II Errors

- There are four things that can happen when we test a hypothesis
  - H0 is true. We discard H1 and retain H0. Correct decision (True Negative).
  - H0 is true. We discard H0 and retain H1. Type I Error (False Positive)
  - H1 is true. We discard H1 and retain H0. Type II Error (False Negative)
  - H1 is true. We discard H0 and retain H1. Correct decision (True Positive).
- It would be good if we have a method to test these hypothesis while controlling the probability of making the errors.

Neyman-Pearson Procedure

- Choose a Type I Error (almost always 0.05)
- Find yourself a test statistic. It will be a function of the data and the sample size. In other words, once I give you a data set you should be able to calculate the value of the test statistic from the data set. Ideally, the values the test statistic takes would be very different when H0 is true from when H1 is true.
    - There will be many test statistics that will be reasonable to use.
- Calculate the value of the test statistic in the data set at hand.

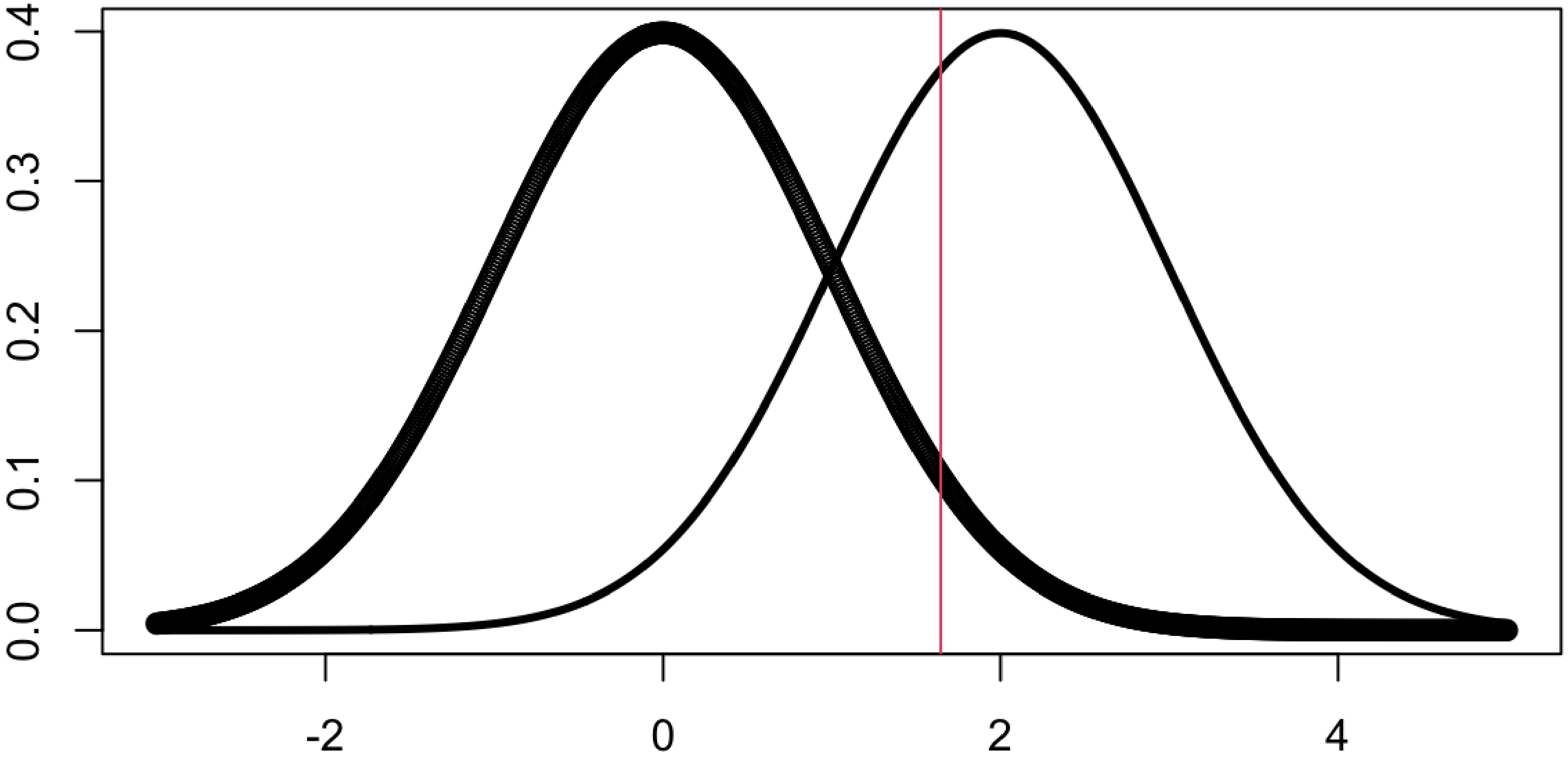
Neyman-Pearson Procedure (cont’d)

- Figure out what values would you have gotten if H0 is true.
    - You can represent these values with a distribution.
- Figure out which value of the test statistic would have given you a 5% Type I Error. This is your reference value.
- If the value of the test statistic exceeds the reference value then your test statistic is unusual with respect to this reference distribution, H0 is unlikely to be true. Reject it, retain H1.
- Else retain H0.

What is the reference distribution?

- It turns out that all Wald test statistics have the same reference distribution, “approximately”
  - Approximately is very critical. The approximation is better when n is large. How large depends on the actual test.
- Normal distribution with mean 0 and variance 1 (standard normal)
  - We did not exactly learn what this is but it is the usual bell shaped curve
- What is the reference value? There are tables for this but you can also use a software or ask Google or chatGPT
  - 1.645

Test Statistic



More on the reference distribution

- What the test statistic would look like if the null distribution is true?
- This has a “parallel universe” interpretation
- Suppose the true response rate is 0.5 and you repeat the trial, get a certain amount of responses and calculate the test statistic.
- Do it over and over a very large number of times
- Do a histogram of the test statistics you get --- and imagine its smooth version
- This is your reference distribution

Key Points

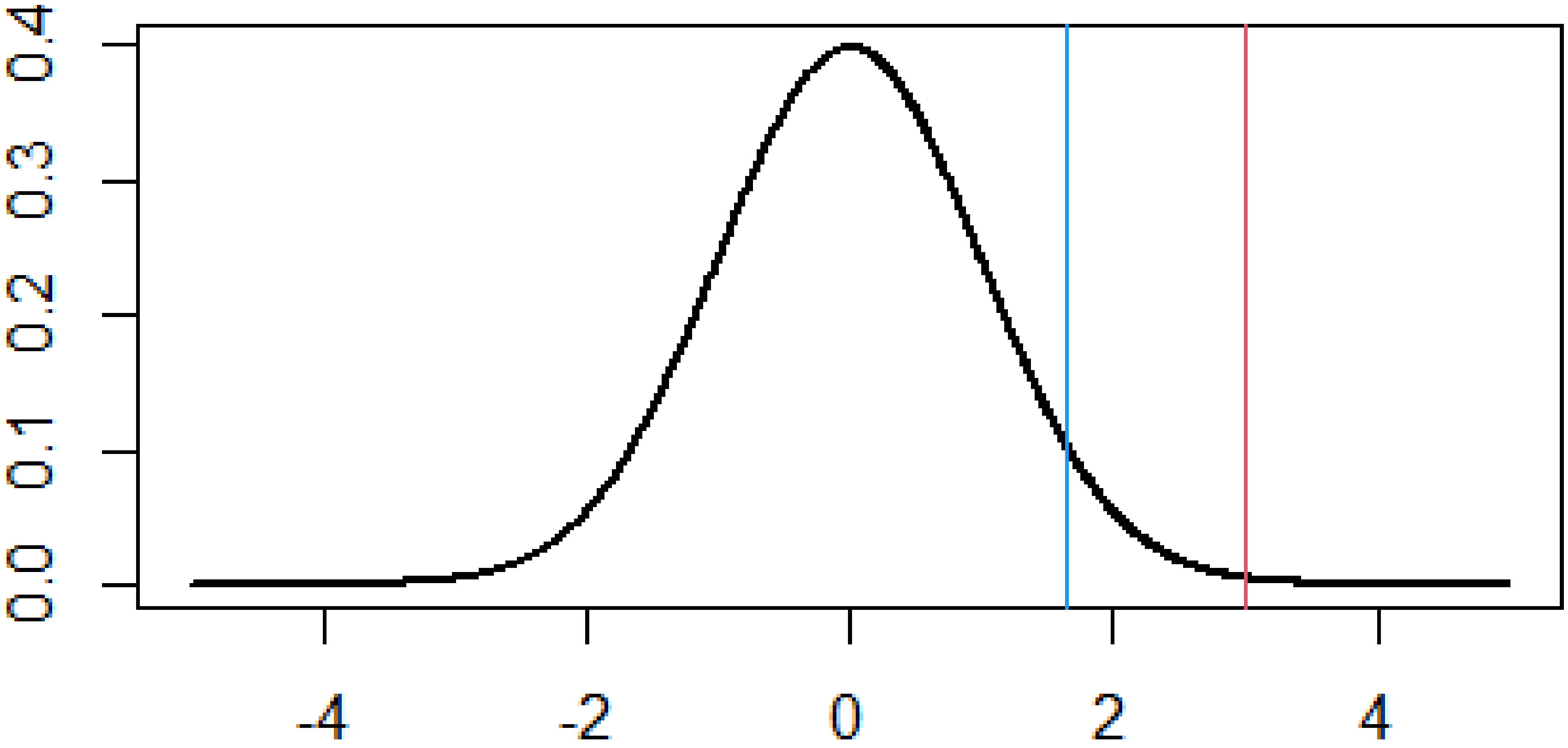
- What does the parallel universe ensure?
    - If H0 is true then this is what we expect to see
- What does 1.645 ensure?
    - If H0 is true then the value of the test statistic would exceed 1.645 only 5% of the time
- These two points together means Type I Error is controlled at 5%

Remarks

- What about the Type II error?
  - This procedure has no assurance that we have a small Type II error
  - We will need to deal with it separately
- So the two hypotheses are not treated “symmetrically”
- This is why we put the hypothesis we want to reject in H0
  - If we reject H0, there is no concern about Type II error
  - We could not have a false negative

What is the p-value?

- It is the probability of observing a test statistic higher than the one you observed in your data set
- You can use the p-value to test a hypothesis
    - If p< =0.05 reject H0
    - Else do not reject H0.


0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4

Remarks About the P-Value

- It is not the Type I error !!!
- This cannot be over-emphasized
- P-VALUE IS NOT THE PROBABILITY OF TYPE I ERROR
- Type I error probability is not something you observe, it does not depend on the data. It is something you choose, ahead of time.
- Why do you have to choose it ahead of time?

More Remarks About the P-Value

- Think of the parallel universes
- If you do not choose the Type I error ahead of time, if you allow yourself to see the data first and then choose it ….
- Then these universes will not be parallel
- Parallel means identical procedures, rules and algorithms
- Parallel universes allow you to make a probability out of these parallel universes
- If they are not parallel you cannot calculate probabilities

More about p-values

- They are meant as a standardized way of measuring the strength of evidence against H0
- Smaller p, more evidence
- Test statistics measure that too but they are not necessarily standardized, and you cannot tell what is small or large

More about p-values

- They are strengths of evidence not effect
- You can have a small p-value with a small effect and a large sample size
- Think of the test statistic
- Numerator is the effect size
- It has sqrt(n) in the denominator of the denominator, i.e. n goes up denominator goes down test statistic goes up
- Same effect size but larger n means larger test statistic and smaller p-value

How Does One Test A Hypothesis Using a Confidence Interval?

- Produce a confidence interval and see if it is entirely contained in one of the hypothesized intervals or not.
- If it is then we rule in favor of that hypothesis
- In this example, confidence interval is 0.7 – 0.92, entirely contained within $r>0.5$, hence we conclude $r>0.5$
- What is the interval spanned both intervals (say it was 0.4 - 0.6)?

Asymmetry of hypothesis testing

- As long as our interval estimate contains a shred of the null region we cannot rule in favor of the alternative
    - For example, if the confidence interval here was 0.49-0.69 then we would not reject H0
- You get the same yes/no answer either way
    - Whether you use a p-value of a confidence interval
- Confidence interval has the effect size (center), strength of evidence (width) and the test (null value included or not)
- This is why you hear confidence intervals are better than tests

Why not use it all the time?

- Ignorance
  - Don’t know that it is better
- Habit
  - Editors and reviewers ask for a p-value
- Laziness
  - CI requires more intellectual work

Type I and Type II Errors

- There are four things that can happen when we test a hypothesis
  - H0 is true. We discard H1 and retain H0. Correct decision (True Negative).
  - H0 is true. We discard H0 and retain H1. Type I Error (False Positive)
  - H1 is true. We discard H1 and retain H0. Type II Error (False Negative)
  - H1 is true. We discard H0 and retain H1. Correct decision (True Positive).
- Neyman-Pearson procedure and $p<0.05$ focused on Type I Errors. At no point we were concerned with Type II errors

How to calculate the Type II error

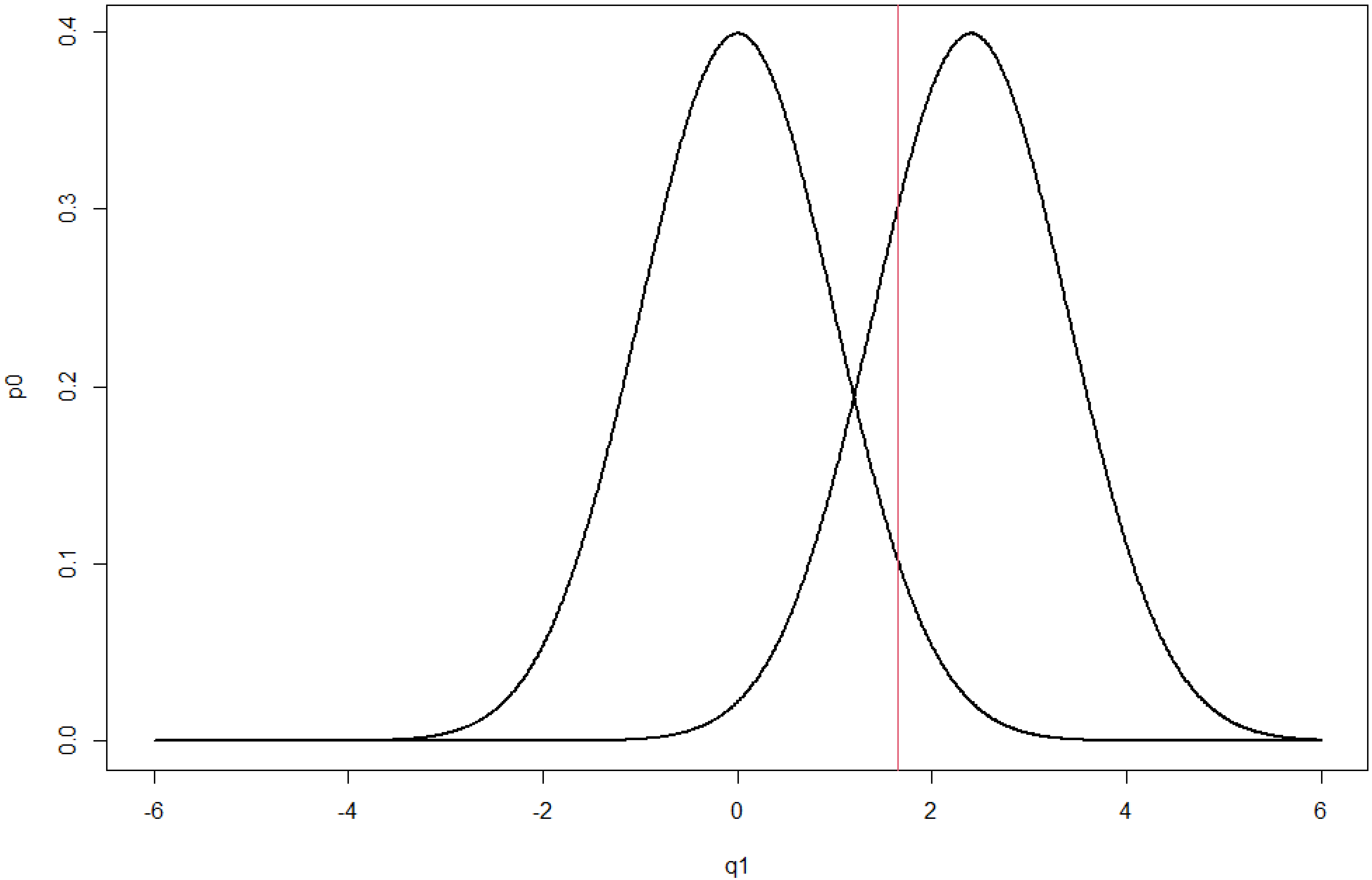
- Power = 1 – Prob(Type II Error)
- If H1 is true, what are the chances that we will not reject H0?
- Do you remember the reference distribution?
    - In parallel universes what would the test statistic look like if H0 is true
    - That was when H0 was true
- We need a distribution that represents the world under H1
    - H1: r > 50%
    - Many r’s here. Which one to use?
    - Power will depend on the value of r you use to calculate it.

What r to use?

- You should use an r that makes “clinical sense” – a value of response rate that would be make you and others think it is worth taking this regimen forward (or changing practice, depending on the context).
- Achilles heel
  - Not so easy to agree on r but power can be very sensitive to this choice

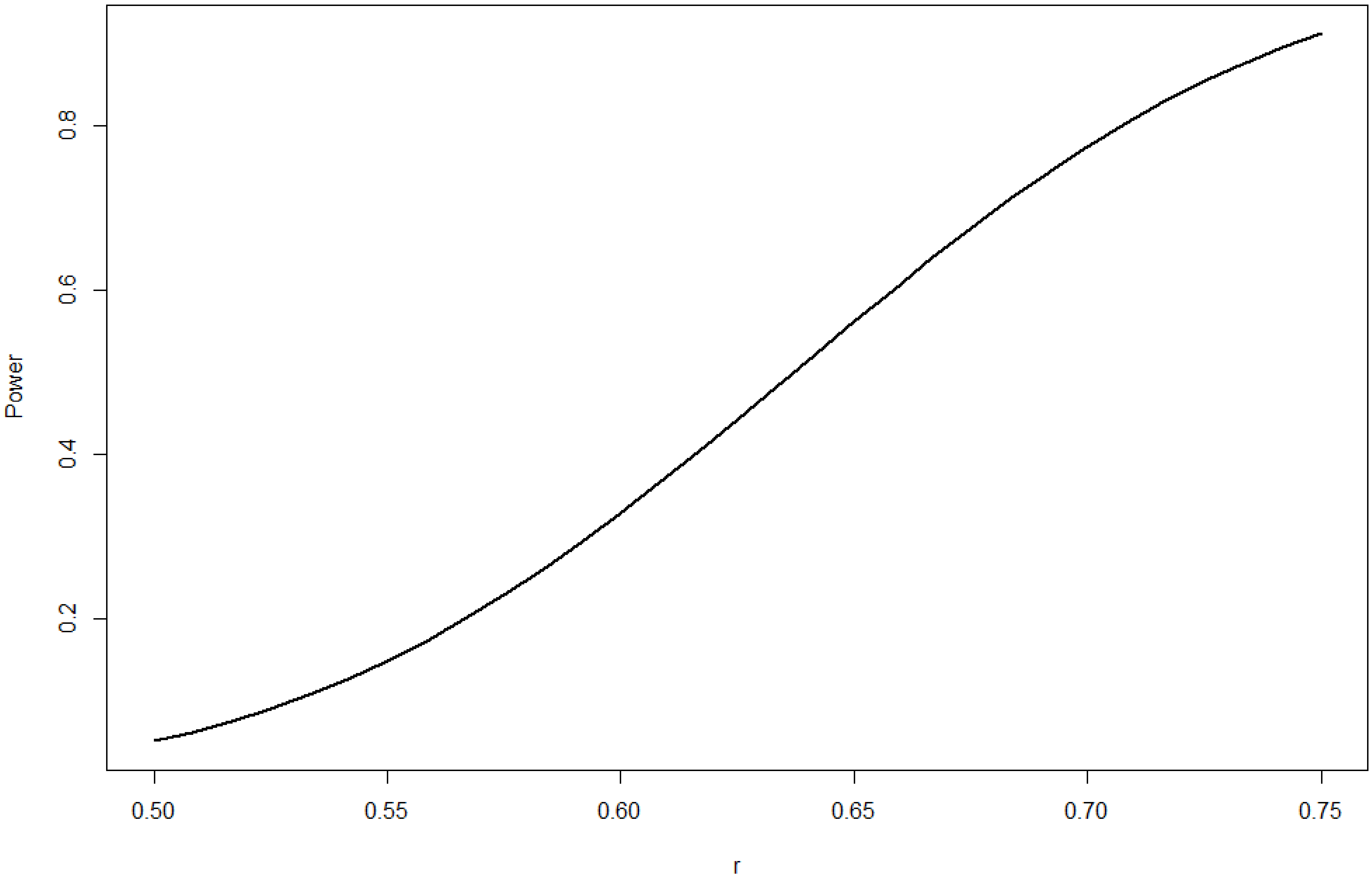
What do you do then?

- Do you remember the threshold (1.645) – your decision rule?
- Once you choose an r and calculate the distribution under H1 then you need to find out how often in the parallel universes you will exceed the threshold
    - i.e. how often you will reject H0
- Let’s assume we chose r = 0.7 for this


0.0
0.1
0.2
0.3
0.4
p0
-6
-4
-2
0
2
4
6
q1

In this case

- Power = 79%
- 20% Type I error is the maximum acceptable
- Phase III trials often use 10%
- What if the r I wanted to use was 0.6 instead of 0.7
- Power would be 33% instead of 79%
- Remember this for the same sample size of 36
- If you want 80% power for r=0.6 you need to raise your sample size to 144! Quadruple.



Power
0.2
0.4
0.6
0.8
0.50
0.55
0.60
0.65
0.70
0.75
r

Power ←→Sample Size

- Same coin two different sides
- Only this way because we set up the procedure to favor Type I errors
- This is how you set up the sample size for your study
- More on this later

Accept H0 or Do not Reject H0

- P = 0.06, do not reject H0
- Why do we not come out and say accept it
- Because we do not know if have enough power just from looking at a p-value that.
- If we do not have enough power then our conclusion has a serious chance of being a Type II error
- P=0.04, reject H0. We are also looking at only p-value. How can we say reject here?

Type I error is fixed

- It is because we fixed the Type I error at 5%. Made sure our procedure protects it
- Again, no symmetry in hypothesis testing
- So “do not reject H0” language is boilerplate because we do not know the power of a test reported in a paper
- Exception clinical trials. This is why we can say “no effect” in a well-powered randomized trial.

One-Sided vs Two-Sided Tests

- It is really the hypothesis that is one or two-sided. Specifically, it is the alternative hypothesis
- Two-Sided
  - Null: New Drug = Old Drug
  - Alternative: New Drug ≠ Old Drug
- One-Sided
  - Null: New Drug <= Old Drug
  - Alternative: New Drug > Old Drug

When To Use Which?

- Source of confusion
- Most of the time one-sided seems appropriate since we do not care if the new drug is worse or if it is simply not different
    - Although you could argue you want to know
- When you have multiple primary endpoints it could be confusing to have one-sided tests
- Then there are the journal policies

JCO

- JCO
  - https://ascopubs.org/jco/authors/journal-policies
  - “For randomized clinical trials, report two-sided *P* values.”
- NEJM
  - https://www.nejm.org/author-center/new-manuscripts
  - Unless one-sided tests are required by study design, such as in noninferiority clinical trials, all reported P values should be two-sided.

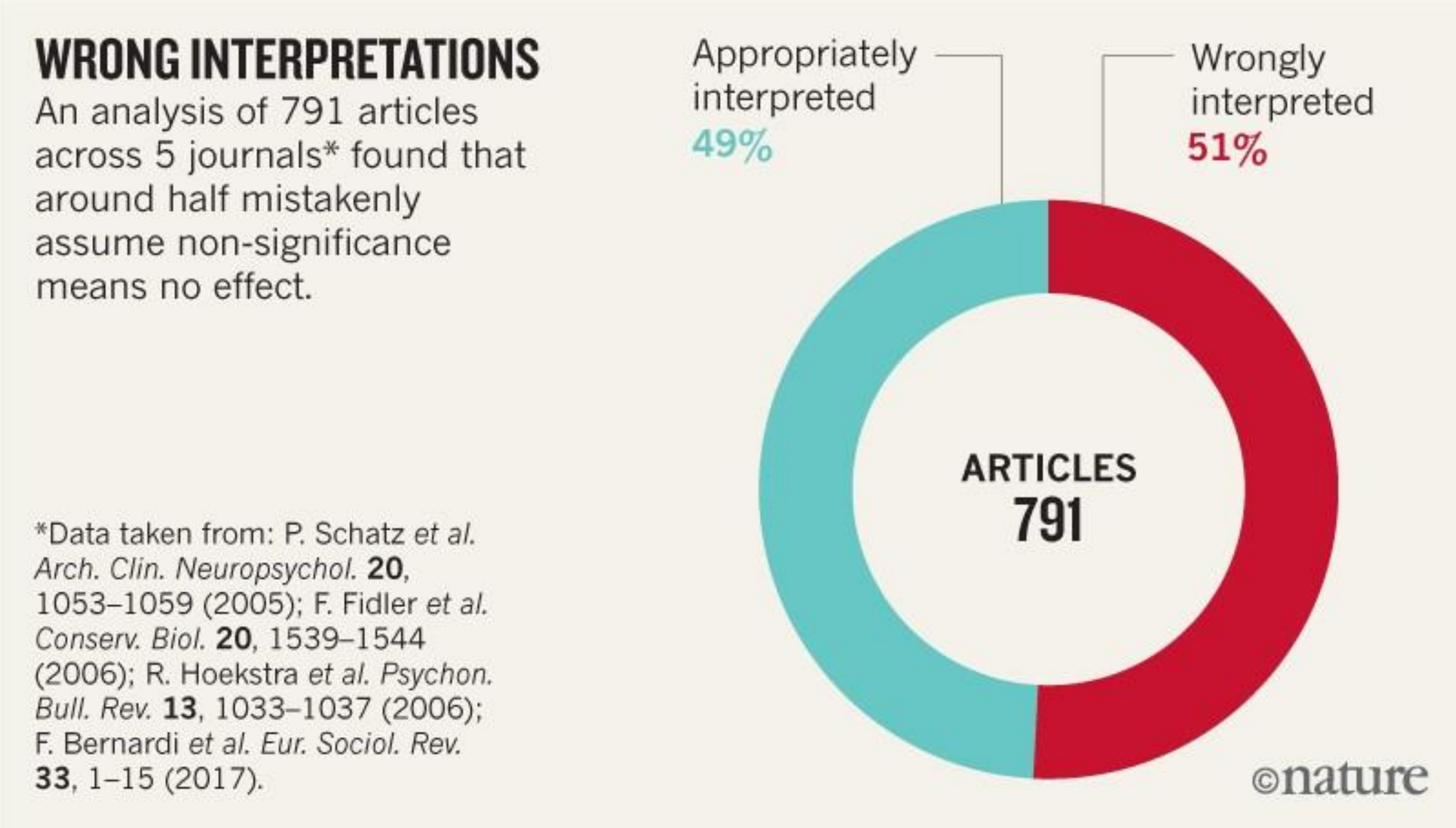
Goodman Article

- https://pubmed.ncbi.nlm.nih.gov/10383371/

Scientists rise up against statistical significance

- When was the last time you heard a seminar speaker claim there was ‘no difference’ between two groups because the difference was ‘statistically non-significant’?
  - Can you think of examples?
- We have some proposals to keep scientists from falling prey to these misconceptions.
  - Let’s be clear about what must stop: we should never conclude there is ‘no difference’ or ‘no association’ just because a *P* value is larger than a threshold such as 0.05 or, equivalently, because a confidence interval includes zero.
  - Neither should we conclude that two studies conflict because one had a statistically significant result and the other did not. These errors waste research efforts and misinform policy decisions.

We call for the entire concept of statistical significance to be abandoned.



WRONG INTERPRETATIONS

An analysis of 791 articles across 5 journals* found that around half mistakenly assume non-significance means no effect.

*Data taken from: P. Schatz *et al.* *Arch. Clin. Neuropsychol.* **20**, 1053–1059 (2005); F. Fidler *et al.* *Conserv. Biol.* **20**, 1539–1544 (2006); R. Hoekstra *et al.* *Psychon. Bull. Rev.* **13**, 1033–1037 (2006); F. Bernardi *et al.* *Eur. Sociol. Rev.* **33**, 1–15 (2017).

©nature

What do you think these percentages are for your fields?

We are not calling for a ban on *P* values

- We are not advocating for an anything-goes situation, in which weak evidence suddenly becomes credible.
- Rather, and in line with many others over the decades, we are calling for a stop to the use of *p*-values in the conventional, dichotomous way — to decide whether a result refutes or supports a scientific hypothesis

Bias and Validity

- Statistically significant estimates are biased upwards in magnitude and potentially to a large degree, whereas statistically non-significant estimates are biased downwards in magnitude. Consequently, any discussion that focuses on estimates chosen for their significance will be biased.
  - This is true only if you have a number of estimates and choose to report only the significant ones
  - It is not true if you report all analyses, or if you have a pre-planned single primary analysis

Bias and Validity

- Statistically significant estimates are biased upwards in magnitude and potentially to a large degree, whereas statistically non-significant estimates are biased downwards in magnitude. Consequently, any discussion that focuses on estimates chosen for their significance will be biased.
- On top of this, the rigid focus on statistical significance encourages researchers to choose data and methods that yield statistical significance for some desired (or simply publishable) result, or that yield statistical non-significance for an undesired result, such as potential side effects of drugs — thereby invalidating conclusions.

Getting carried away

- For example, even if researchers could conduct two perfect replication studies of some genuine effect, each with 80% power (chance) of achieving $P < 0.05$, it would not be very surprising for one to obtain $P < 0.01$ and the other $P > 0.30$. Whether a P value is small or large, caution is warranted.
  - In a study powered 80%, the probability that the resulting p-value will be less than 0.01 is 60%
  - In a study powered 80%, the probability that the resulting p-value will be greater than 0.30 is 3.5%

Compatibility interval

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

How would you interpret these differences in RECIST response rates?

- The numbers are response rate with drug 1 minus with drug 2 in a randomized study. Point estimate (95% Conf int)
  - 20% (-20%, 60%)
  - 20% (-1%, 41%)
  - 20% (1%, 39%)
  - 20% (5%, 35%)
  - 20% (15%, 25%)
- One immediate effect is that small studies will not be as much penalized as they are now.

More time thinking

- What will retiring statistical significance look like? We hope that methods sections and data tabulation will be more detailed and nuanced. Authors will emphasize their estimates and the uncertainty in them — for example, by explicitly discussing the lower and upper limits of their intervals … Decisions to interpret or to publish results will not be based on statistical thresholds. People will spend less time with statistical software, and more time thinking.
    - People will spend less time with statistical software and more time wordsmithing the desired conclusions into their interpretations. The only way to get people thinking is to teach them statistics by way of thinking

Misuse

- The misuse of statistical significance has done much harm to the scientific community and those who rely on scientific advice. *P* values, intervals and other statistical measures all have their place, but it’s time for statistical significance to go.
    - The misuse of statistics has done much harm to the scientific community and those who rely on scientific advice. This is due to lack of building statistical literacy in scientific training. Banning significance will not change this.

The Importance of Predefined Rules and Prespecified Statistical Analyses

- Behind the so-called war on significance lie fundamental issues about the conduct and interpretation of research that extend beyond (mis)interpretation of statistical significance.
- These issues include what effect sizes should be of interest, how to replicate or refute research findings, and how to decide and act based on evidence.
- Dichotomous decisions are the rule in medicine and public health interventions. An intervention, such as a new drug, will either be licensed or not and will either be used or not.

Skeptics and enthusiasts

- Some scientists may be skeptical about some research questions and enthusiastic about others.
- The suggestion to abandon statistical significance espouses the perspective of enthusiasts: it raises concerns about unwarranted statements of “no difference” and unwarranted claims of refutation but does not address unwarranted claims of “difference” and unwarranted denial of refutation.

Objectively Assessed Evidence

- Interpretations go beyond statistics. They also vary depending on what other (eg, mechanistic) evidence is considered relevant. However, determination of the relevance of qualitative or triangulating types of evidence can be substantially subjective.
- The statistical data analysis is often the only piece of evidence processing that has a chance of being objectively assessed before experts, professional societies, and governmental agencies begin to review the data and make recommendations.
- This means that, ideally, the statistical analysis should use carefully prethought, rigorous probes …When the analyses are preplanned, clear, and followed carefully, such tests are useful. Interpretation of any result is far more complicated than just significance testing, but it is a starting point.

Gatekeeper

- The proposal to entirely remove the barrier does not mean that scientists will not often still wish to interpret their results as showing important signals and fit preconceived notions and biases.
- With the gatekeeper of statistical significance, eager investigators whose analyses yield, for example, $P = .09$ have to either manipulate their statistics to get to $P < .05$ or add spin to their interpretation to suggest that results point to an important signal through an observed "trend."
- When that gatekeeper is removed, any result may be directly claimed to reflect an important signal or fit to a preexisting narrative. Moreover, refutation of an early study by a subsequent replication effort can always be denied.

Refutation

- Many fields of investigation (ranging from bench studies and animal experiments to observational population studies and even clinical trials) have major gaps in the ways they conduct, analyze, and report studies and lack protection from bias.
- Potential for falsification is a prerequisite for science. Fields that obstinately resist refutation can hide behind the abolition of statistical significance but risk becoming self-ostracized from the remit of science.

Pre-specified conclusions

- In a recent survey completed by 390 consulting statisticians, a large percentage perceived that they had received inappropriate requests from investigators to analyze data in ways that obtain desirable results.
- Studies have shown that unless an analysis is prespecified, analytical choice (eg, different adjustments for covariates in nonrandomized studies) may allow obtaining a wide range of results.

Statistical Numeracy, Statistical Anarchy

- The statistical numeracy of the scientific workforce requires improvement. Banning statistical significance while retaining *P* values (or confidence intervals) will not improve numeracy and may foster statistical confusion and create problematic issues with study interpretation, a state of statistical anarchy.