Biostatistics
Lecture 5

Mithat Gonen

Brendon Bready

Hypothesis Testing

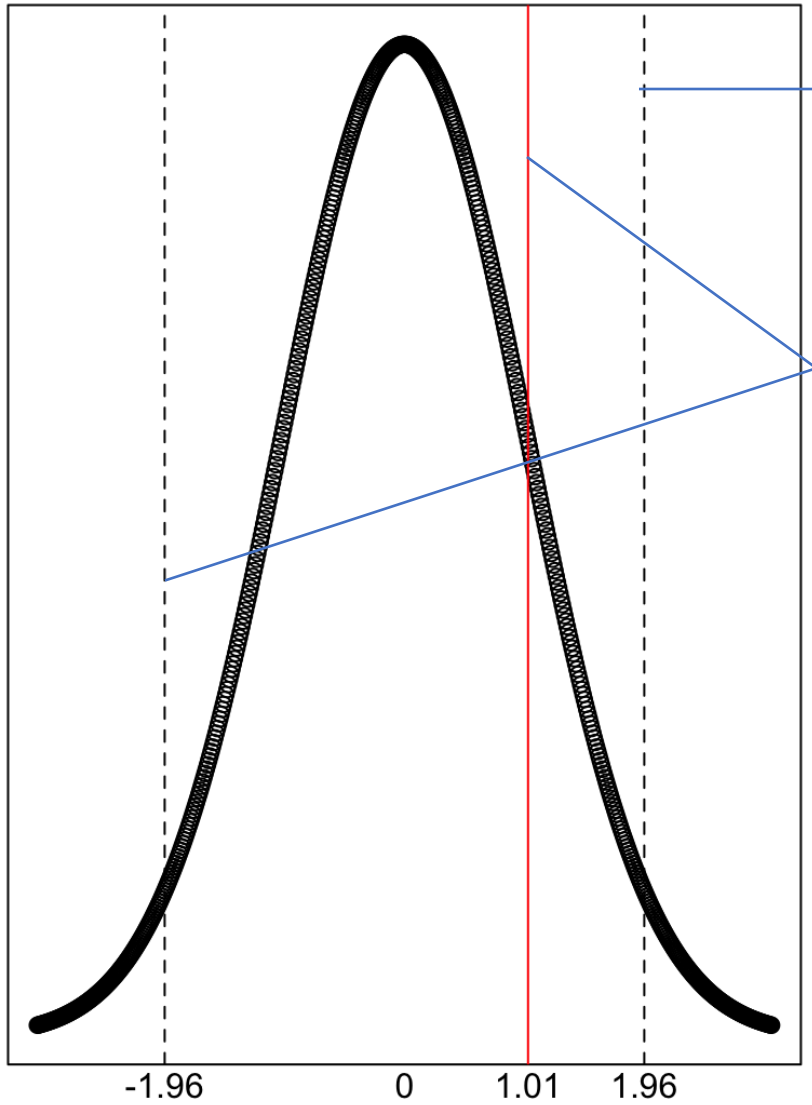
- We formulate a null hypothesis that reflects the state of the art and an alternative that includes an element of novelty.
- Decision: either reject or retain the null
- Reject the null when it is true
  - Type I error
- Retain the null when it is false
  - Type II error (= 1 – Power)
- Type I error is considered to be more important, hence we control it at a pre-specified level

Steps Involved

- Formulate a Null and an Alternative Hypoteheses
- Choose a Test Statistic
- Find a reference distribution for the test statistic (i.e. the distribution of the test statistic under the null)
- Choose a Type I Error and Find the Implied Threshold From the Reference Distribution
- Calculate the Value of of the Test Statistic from data and compare with the threshold
  - Test Statistic > Threshold → Reject the null
  - Test Statistic <= Threshold → Retain the null

One-Sided vs Two-Sided Tests

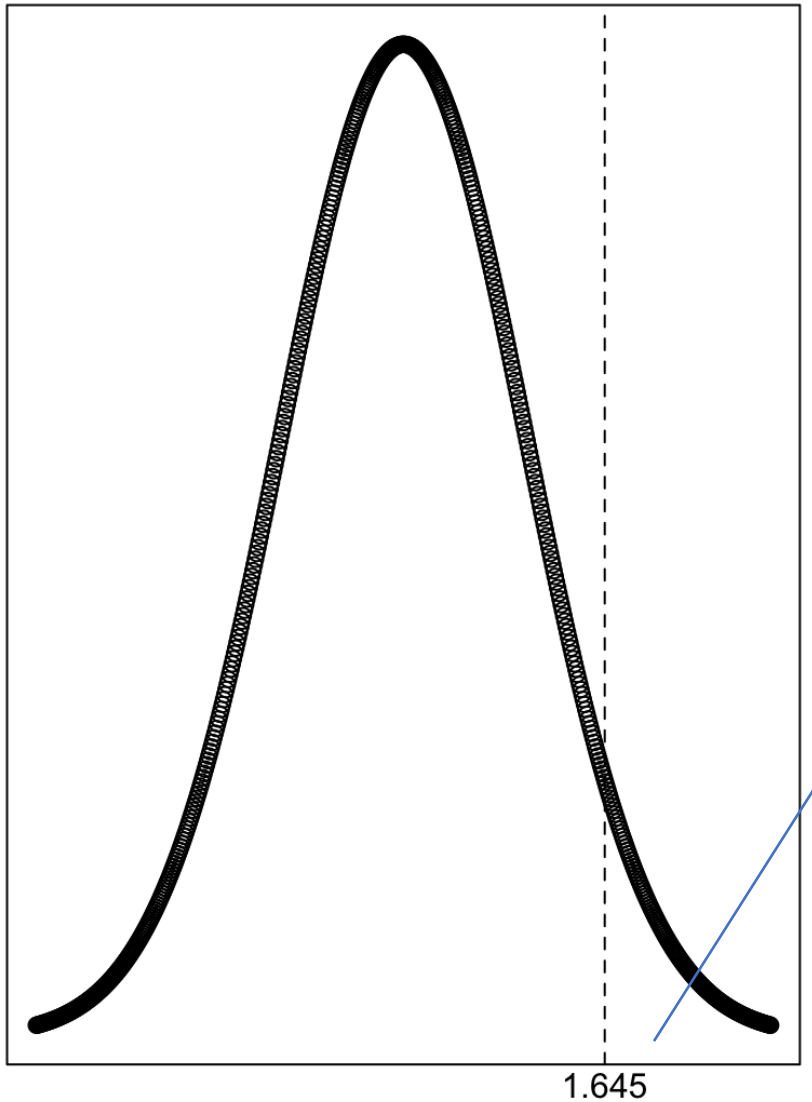
- It is really the hypothesis that is one or two-sided. Specifically, it is the alternative hypothesis
- Two-Sided
  - Null: New Drug = Old Drug
  - Alternative: New Drug ≠ Old Drug
- One-Sided
  - Null: New Drug <= Old Drug
  - Alternative: New Drug > Old Drug



Thresholds

Observed Test Statistic

Conclusion: Retain The Null



- This is for testing Null: New <= Old vs Alternative: New > Old
- The threshold have changed
- Because the null and alternative have changed
- This area is 0.05
- If the test statistic exceeds 1.645 we reject the null
- Else (including the case when the test statistic is less than -1.645) we retain the null
- What would this picture look like if the null is New > Old and the alternative is New <= Old?

Nonparametric Tests

- Takes up a large chunk of questions I receive from fellows
- T-test or Wilcoxon?
- What is the difference?
- T-test looks at the means in each group and their difference
- As a result it can be sensitive to outliers or skewness (situations where mean itself is not a good summary of the distribution)

How to Choose: T-Test or Wilcoxon

- In small data sets with outliers Wilcoxon test works better.
- In small data sets with no outliers t-test works better.
- In large data sets it does not matter a whole lot.
- Wilcoxon is always a safe bet, use that if not sure.

Multiple Testing

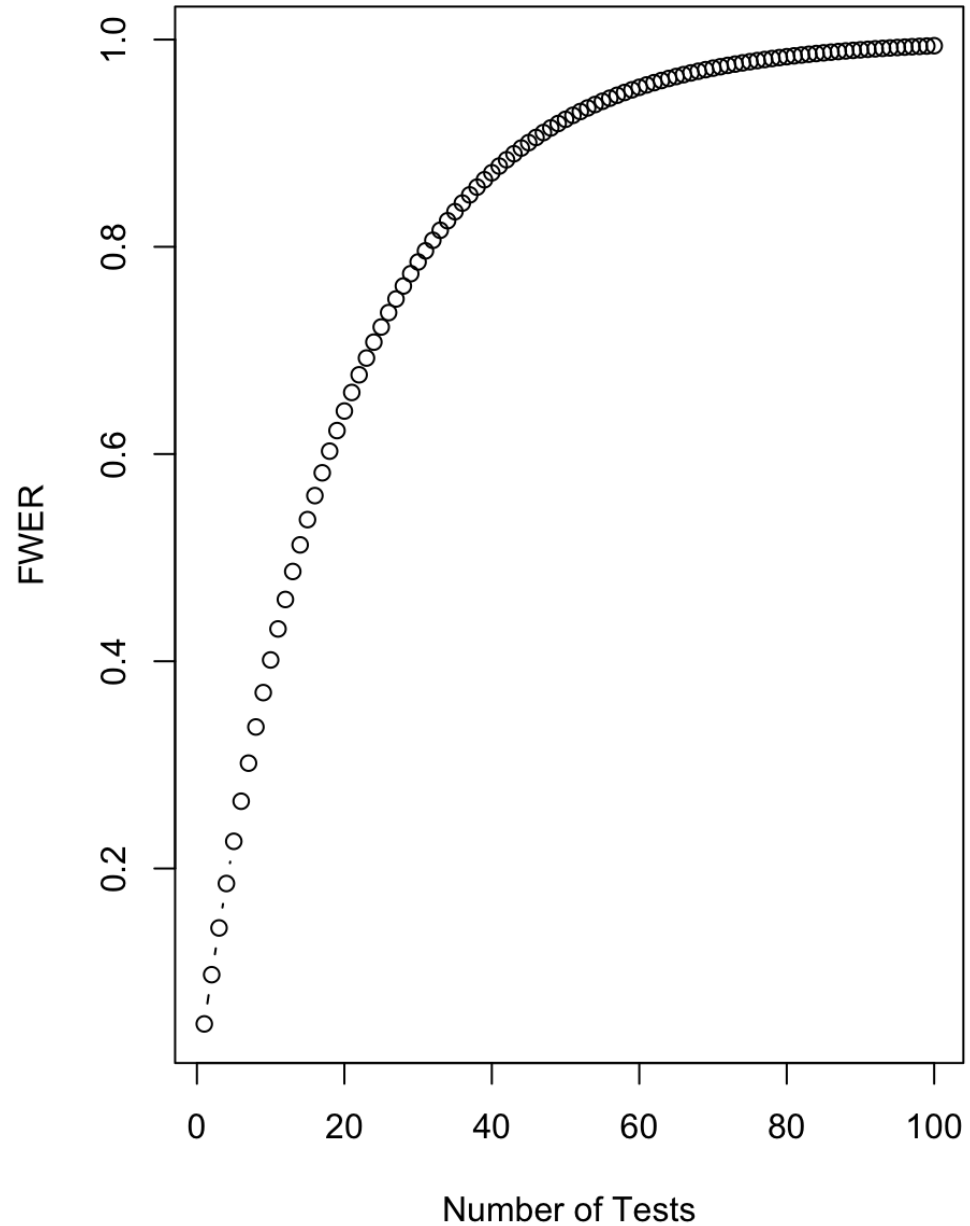
- What happens when you simultaneously test many hypotheses?
- Suppose we have two tests, T1 and T2
- What is the null? Are there two nulls?
- Or is the null Null(T1) and Null (T2)?

Multiple Testing

- Imaging tossing two coins at the same time, heads means reject the null incorrectly, tails retain correctly
- What is the probability that I will reject at least one of them (two heads)
  - 1 – Probability of Retaining Both = 1 – 0.5*0.5 = 0.75
- Further imagine coins are biased, P(heads) = 0.05. What is the probability now?
  - 1 – 0.95*0.95 = 0.0975
- Analogous to a Type I error calculation with two tests

Why?

- Under the “joint null” (i.e. both null(T1) and null(T2) are true)
- Probability of at least one false rejection = Type I Error for multiple testing = Familywise Error Rate (FWER)
- If you are performing k tests, each at 5%
  - Type I Error = $1 – 0.95^k$


FWER
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Number of Tests

Family of Tests to Adjust For

- F in FWER
- Which nulls should be included in the family
  - All tests in a table?
  - All tests in a paper?
  - All tests in one's lifetime?
- No right or wrong answer here

Some Guidelines For Multiple Testing

- When you are conducting a really large number of tests
  - Like gene expression analysis
- When it is plausible that all your nulls can be true
  - Table 1 in a retrospective study is not the place to correct for multiple testing
- When you are ready to make big claims
  - That you have discovered the cure for COVID (or cancer)

Sidebar: Clinical Trials

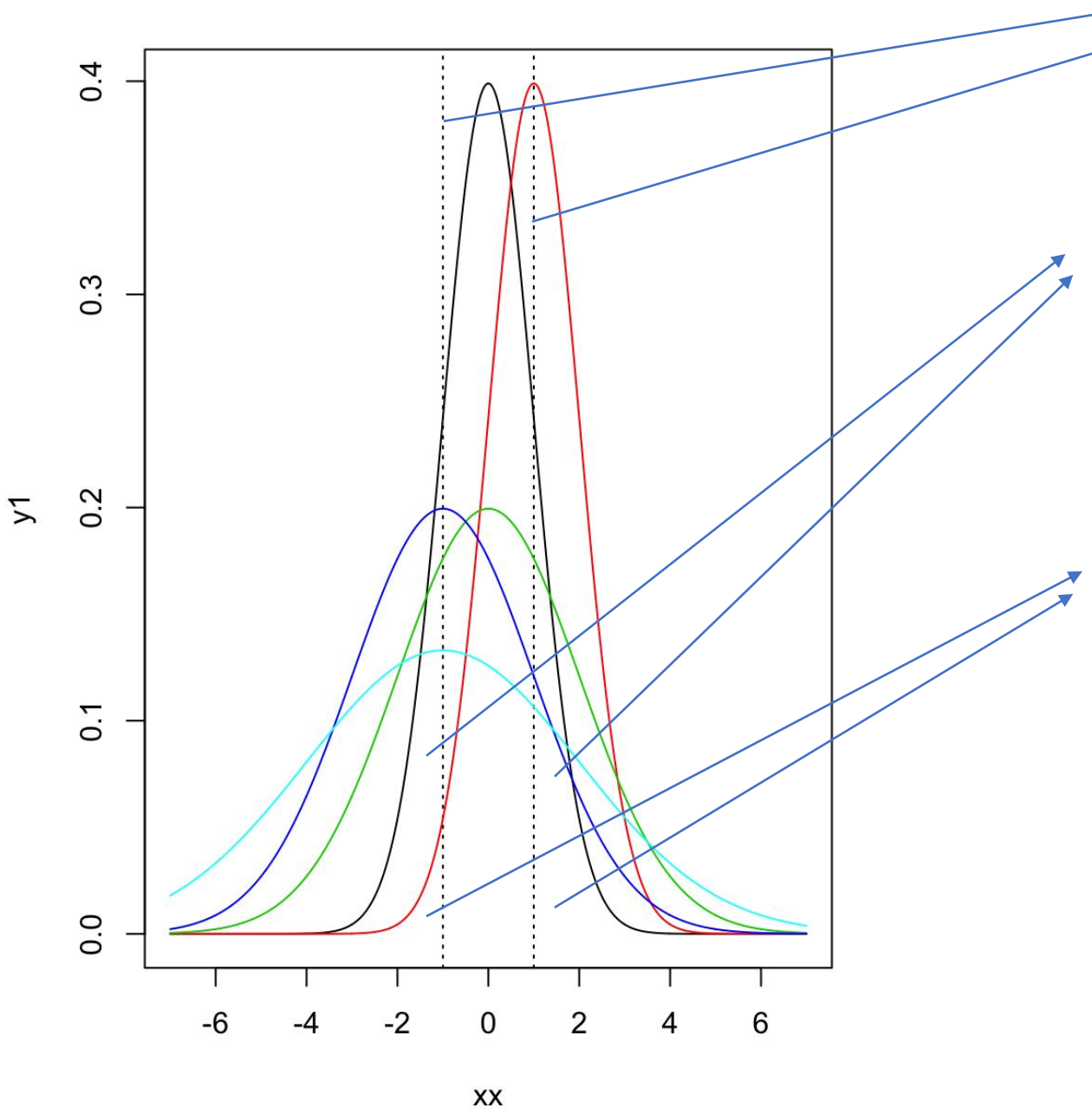
- This is the reason (or one of the major reasons) why one chooses a primary endpoint
- Everything else is secondary
- Avoids multiple testing issues

Errors We Can Make

- Reject the null when it is true
  - Type I error
- Retain the null when it is false
  - Type II error (= 1 – Power)
- Where is the Type II Error?
- We have not studied it yet

Reference Distribution

- It represents the distribution of values the test statistic will take if were to sample again and again under the null hypothesis (when the two distributions are the same)
  - Sometimes called the null distribution
- To Study Type II Error we need the distribution of the test statistic under the alternative hypothesis



Thresholds (defined by Type I Error)

Area in the tails under the null distribution (black line): Type II error

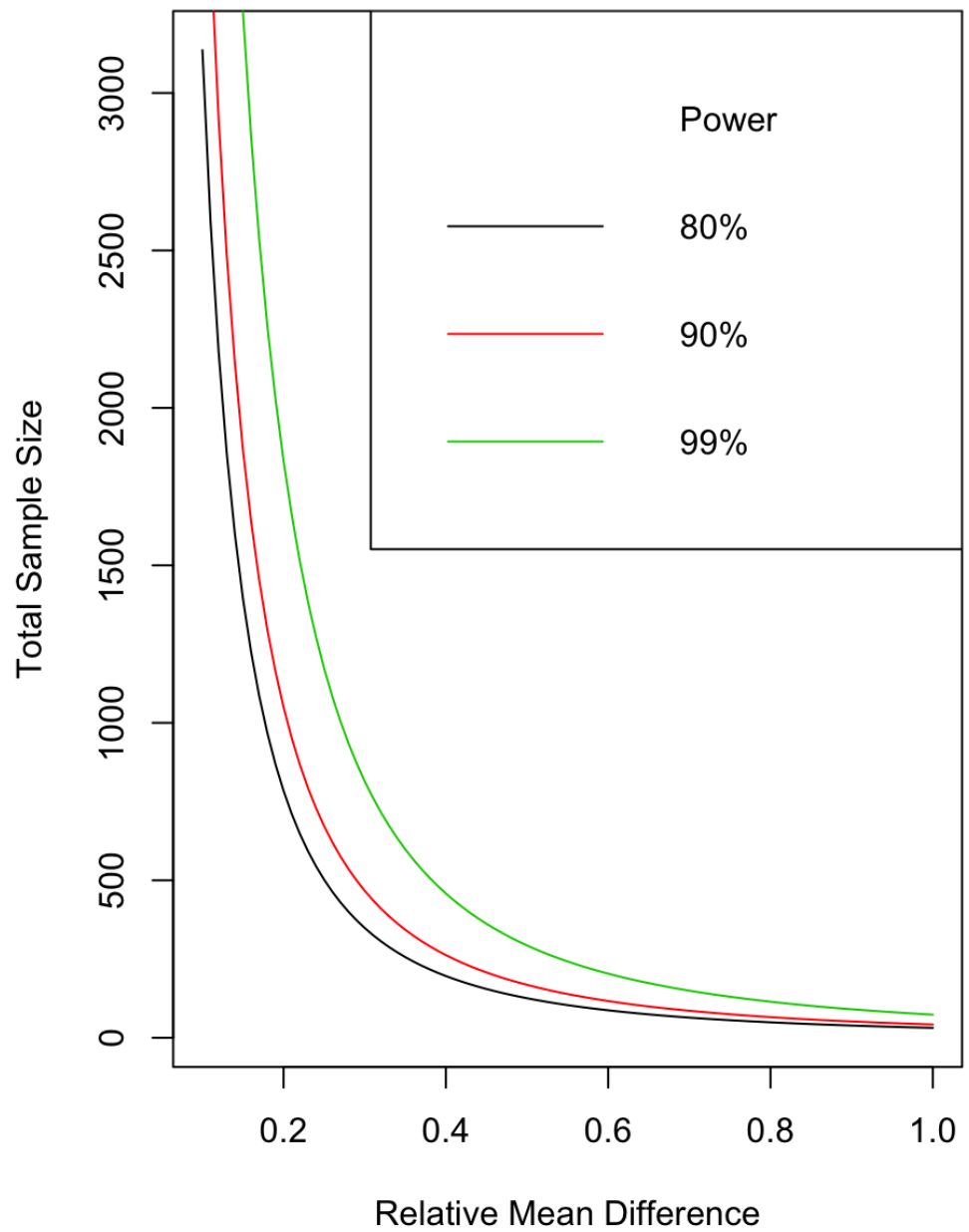
Area in the tails under the reference distribution (red line): Type II error

Where is my sample size

- Remember the test statistic?
- Mean difference divided by the standard error of the mean
- Standard Error of The Mean (SEM) = SD/sqrt(n)
- The test statistic is a function of the sample size
- Larger sample size, smaller standard error
- Smaller standard error, smaller area in the tail, smaller Type II Error
- More power!

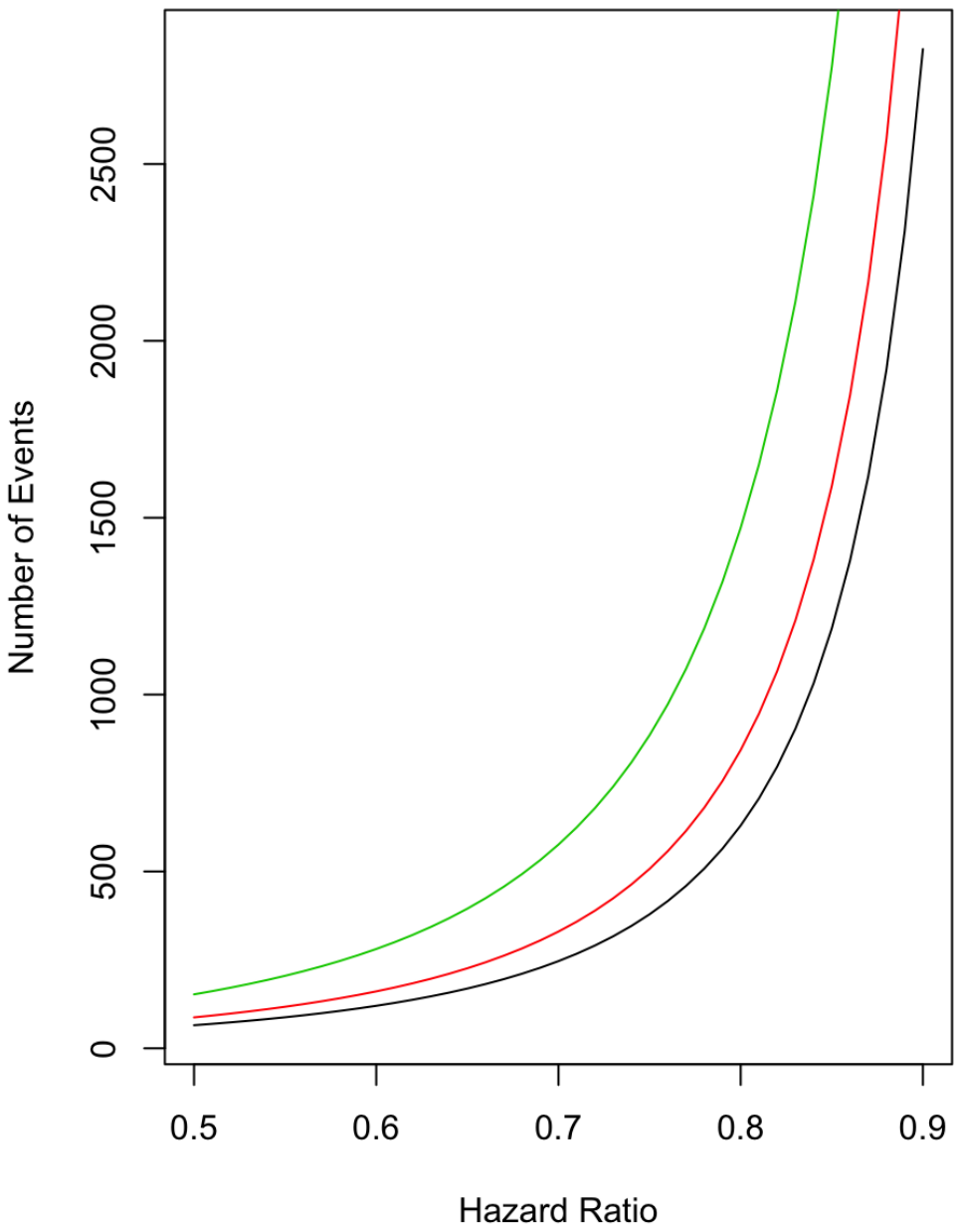
The formula!

- (total) Sample size = $(2*(z_1+z_2)/d)^2$
- d is the relative mean difference
  - Mean difference divided by the standard error
- z_1 is derived from Type I error
  - 1.96 for 5%
- z_2 is derived from Type II error
  - 0.84 for 20%, i.e. 80% power
  - 1.28 for 10%, i.e. 90% power


Power
80%
90%
99%
Total Sample Size
0
500
1000
1500
2000
2500
3000
0.2
0.4
0.6
0.8
1.0
Relative Mean Difference

The formula for a survival outcome

- (total) Number of events required = $(2*(z_1+z_2)/d)^2$
- d is the log of the hazard ratio you chose from the universe of alternatives



HR	Number of Events
0.5	65
0.6	120
0.7	247
0.8	630
0.9	2825

Sample Size Samba (1)

- Clinician: How many patients do I need for this?
- Statistician: What is the difference you want to detect?
- C: [looks confused] What does that even mean?
- S: [feeling superior] It is the distance between two probability distributions defined on a space of Borel sets
- C: [picks up the phone] Let me call employee mental health for you

Sample Size Samba (2)

- Clinician: How many patients do I need for this?
- Statistician: What do you think the HR will be ?
- C: [looks confused] I dunno. If I did why would I do this trial?
- S: [frustrated] I can’t help you then
- C: [picks up the phone] Let me call your service chief

Sample Size Samba (3)

- Clinician: How many patients do I need for this?
- Statistician: What do you think the HR will be ?
- C: [looks confused] 0.8
- S: [frustrated] You need 630 events, assuming this much follow-up, translates to 1000 patients
- C: [picks up the phone] Are you for real? It will take me 10 years to enroll

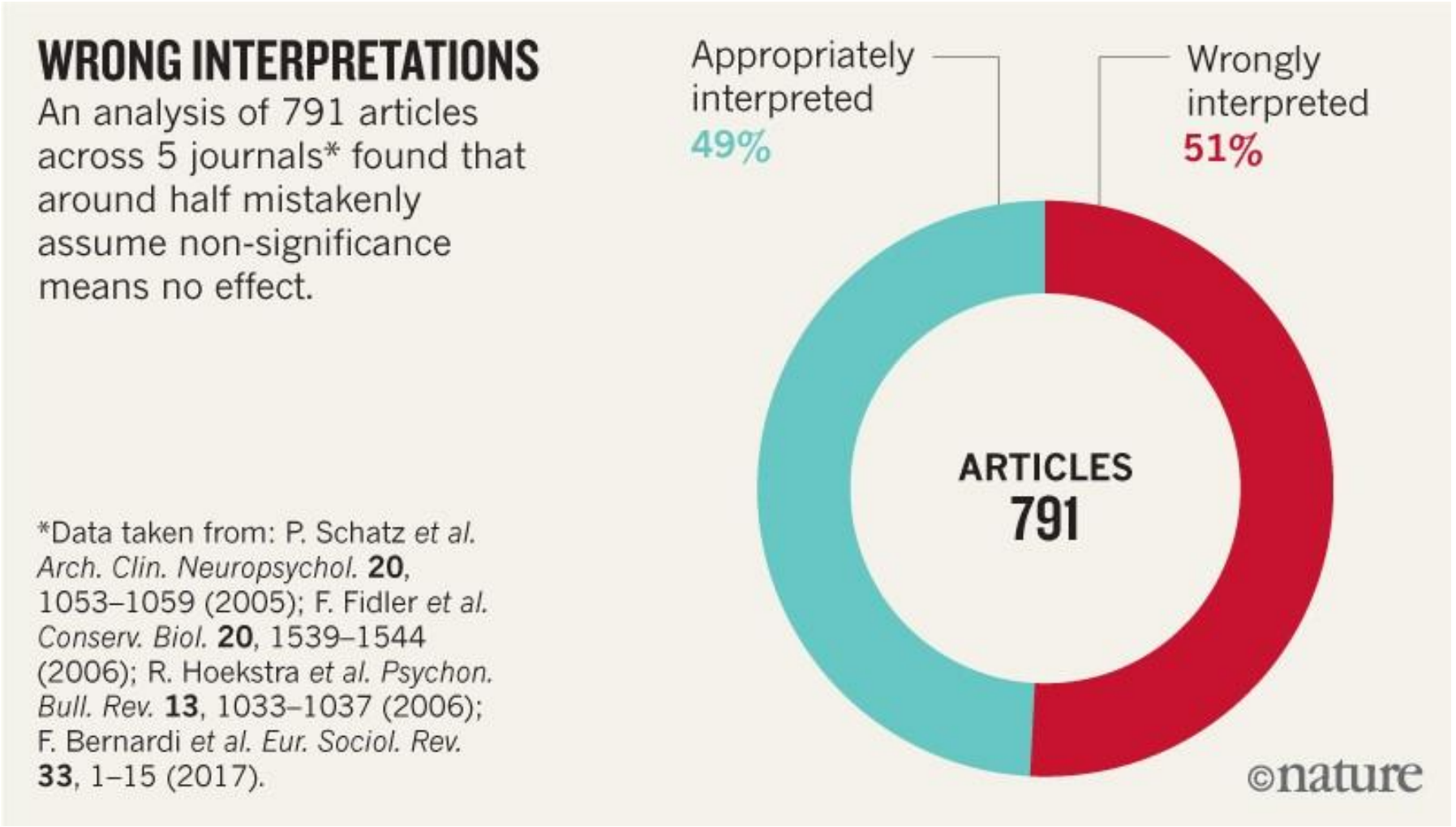
Sample Size Samba (4)

- Clinician: How many patients do I need for this?
- Statistician: What do you think the HR will be ?
- C: [looks confused] 0.8
- S: [frustrated] You need 630 events, assuming this much follow-up, translates to 1000 patients
- C: [picks up the phone] What about 0.7?
- S: about 500
- C: what about 0.6
- … and the samba continues

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.



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.