Swiss TPH

Swiss Tropical and Public Health Institute
Schweizerisches Tropen- und Public Health-Institut
Institut Tropical et de Santé Publique Suisse

Confidence Intervals

Tracy Glass

Associate Professor, University of Basel

Conflicts of Interest

I have received funding for membership on Data Safety and Monitoring Boards and for the preparation of educational materials from:

Gilead Sciences

Novartis

Janssen-Cilag

Background

Although *P*-values are helpful in telling us which effects are likely to be real, and which are likely to be chance findings, they suffer from several limitations

The *P*-value by itself ***does not***:

- provide information about the size of an association
- Provide information about variability around the estimate
- Allow us to put findings in a clinical context

Outline

- Some limitations of *P*-values
- How can confidence intervals help?

Limitations of the p-value

Small changes in the data can switch the results from being non-significant to significant

	Good response		
	Yes	No	Total
A	11	25	36
B	45	42	87
Total	56	67	123

Chi-squared=3.79
P=0.0517

	Good response		
	Yes	No	Total
A	11	26	37
B	45	41	86
Total	56	67	123

Chi-squared=4.45
P=0.0348

Limitations of the p-value

Statistical significance ≠ clinical significance

If the study is large enough, results can be statistically significant even if not clinically important

	Good response		
	Yes	No	Total
A	750 (75%)	250 (25%)	1000
B	770 (77%)	230 (23%)	1000
Total	1520	480	2000

Chi-squared=0.99
P=0.32

	Good response		
	Yes	No	Total
A	7500 (75%)	2500 (25%)	10000
B	7700 (77%)	2300 (23%)	10000
Total	15200	4800	20000

Chi-squared=10.86
P=0.001

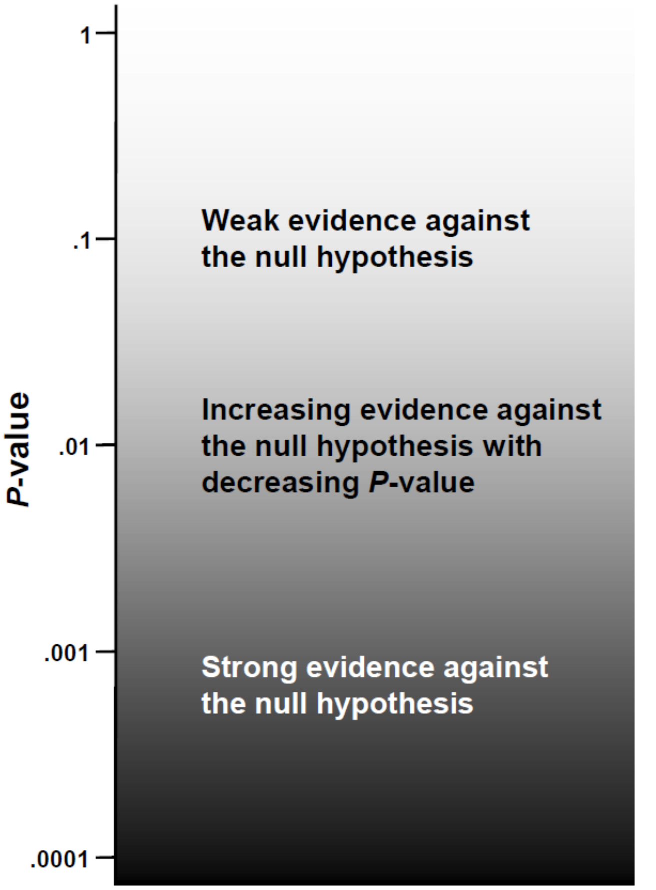
In contrast, in small studies that are insufficiently powered to detect clinically important differences, observed results with $p>0.05$ may result in the results being labeled 'non-significant' and ignored.

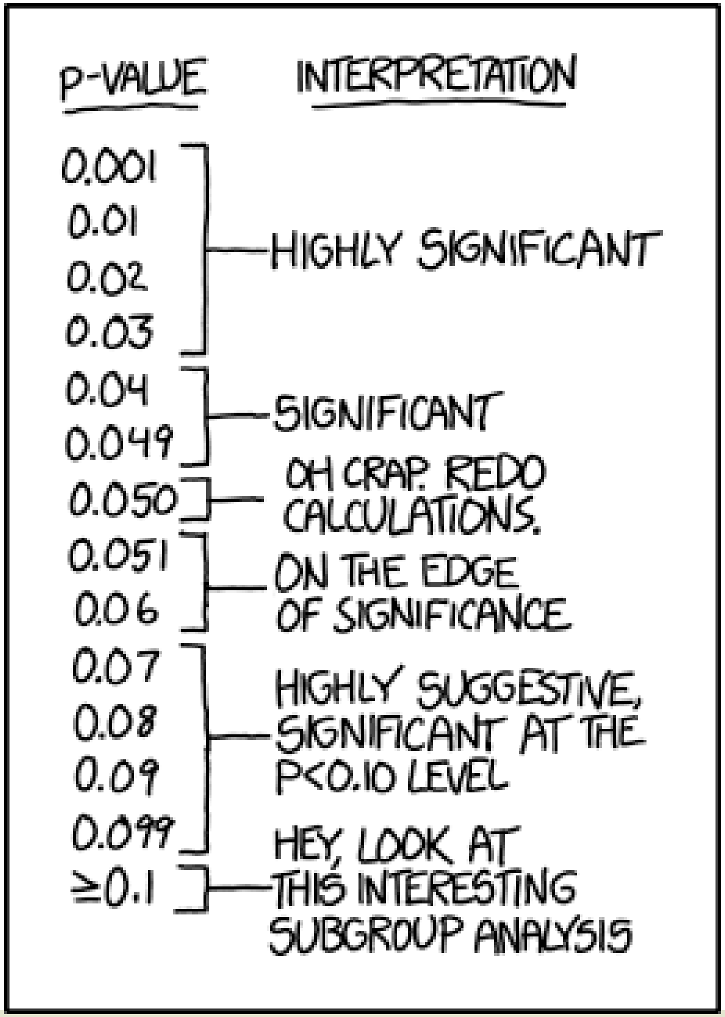
Limitations of the p-value

Threshold of 0.05 is rather arbitrary – what do you do if *P*=0.05?

Is this significant or non-significant?

Led to the re-interpretation of the p-value




P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Limitations of the p-value

All of this pressure to find a 'significant result' leads to ....

P-hacking (aka Data-dredging, fishing, significance chasing)

> "Manipulation of statistics such that the desired outcome assumes 'statistical significance', usually for the benefit of the study's sponsors."
>
> Urban dictionary

Which then leads to ....

Multiplicity

A *P*-value of 0.05 implies that there is a 5% probability that the results were due to chance

Probability that ≥1 of our results will be falsely significant increases exponentially as the number of tests performed increases

E.g. with 20 tests, the probability that *at least one of them* will have a *P*-value <0.05, even if there is no real effect, is almost 100%

In this case, we would conclude that there was a real effect even though no effect exists

This is a Type I error (a false positive finding)

Example

20 **trials**, each with 100 participants

Outcome: % viral load<50 cps/ml

Trial no.	Regimen		*P*-value
	A	B	
	N	N	
1	28/54	22/46	0.84
2	24/53	26/47	0.42
3	30/61	20/39	1.00
4	25/51	25/49	1.00
5	29/57	21/43	1.00
6	24/50	26/50	0.84
7	22/51	28/49	0.23
8	30/54	20/46	0.32
9	28/57	22/43	1.00
10	20/47	30/53	0.23

Trial no.	Regimen		*P*-value
	A	B	
	N	N	
11	29/59	21/41	1.00
12	20/47	30/53	0.23
13	23/51	27/49	0.42
14	22/40	28/60	0.54
15	16/45	34/55	0.02
16	26/54	24/46	0.84
17	24/49	26/51	1.00
18	28/53	22/47	0.69
19	25/42	25/58	0.16
20	22/47	28/53	0.69

Example dealing with multiple testing

- Planned multiple testing can be adjusted for to control the overall Type 1 error rate
- ACTG 5142 trial – comparison of three HAART regimens:
  EFV+2NRTIs; LPV/r+2NRTIs; LPV/r+EFV+2NRTIs
- Three primary comparisons of interest
- Three planned interim analyses

- "The overall type I error rate was 0.05, with 0.017 (0.05 ÷ 3) allocated to each pairwise comparison between study groups; after adjustment for interim analyses (3, each at 0.001), the final type I error rate was 0.014. Thus, only P values of less than 0.014 were considered to have statistical significance in the analyses of primary objectives."

Riddler SA et al. *NEJM* (2008); **358**: 2095-106

Outline

- Some limitations of *P*-values
- How can confidence intervals help?

Treatment effects

Instead of just providing a P-value, should provide an estimate of the effect of interest (i.e. some comparative effect)

The ‘treatment effect’ is the <u>additional benefit</u> that the new drug/regimen provides compared to the ‘standard of care’

Example:

1.	Drug A (standard of care)	63% response
2.	Drug B (new regimen)	71% response

The treatment effect = 71% - 63% = 8%

For every 100 patients treated with regimen B, expect 8 more patients would respond compared to the number that would be expected to respond had they been treated with regimen A

Confidence intervals

Estimate of treatment effect was 8%, a point estimate; this is our 'best guess' as to the true treatment effect, but it gives no indication of variability around this guess

What is a 95% confidence interval?

An interval around our treatment effect for which we are 95% confident includes the *true treatment effect*

Confidence intervals provide a range of additional plausible values that are supported by the results of the study.

They also indicate the precision of the estimate

Example

	Drug				
Trial number	A		B		
	n	n (%) responding	n	n (%) responding	Difference (B – A)
1	50	34 (68)	50	40 (80)	12%

We estimate that drug B is 12% more effective than Drug A

The 95% CI for this estimate is: -5.0% to +29.0%

Drug B could be up to 5% *less effective* than drug A

or up to 29% *more effective* than drug A

What are your views about drug B?

Example

	Drug				
Trial number	A		B		
	n	n (%) responding	n	n (%) responding	Difference (B – A)
1	150	102 (68)	150	120 (80)	12%

We estimate that drug B is 12% more effective than Drug A

The 95% CI for this estimate is: +2.2% to +21.8%

Drug B could be up to 2% *more effective* than drug A

or up to 22% *more effective* than drug A

What are your views about drug B?

Precision of the CI

First confidence interval was too wide to allow us to judge whether drug B was better, worse or the same as drug A

Second confidence interval was narrower, allowing us to conclude that drug B was likely to be better than drug A

The confidence interval from this trial was more precise

How do you obtain a narrower CI?

Assume that 68% of patients on drug A and 80% of patients on drug B respond to therapy....

Number in each group	Treatment 'effect'	95% CI for treatment effect
50	12.0%	-5.0%, +29.0%
100	12.0%	-0.0%, +24.0%
150	12.0%	+2.2%, +21.8%
200	12.0%	+3.5%, +20.1%
300	12.0%	+5.1%, +19.0%
500	12.0%	+6.6%, +17.4%

Major determinant of width of CI is the sample size

Although we have focused on confidence intervals for the difference in two proportions, they can be generated for almost every statistic

Calculations may be tricky, but most statistical packages will generate them automatically

Most journals now require that confidence intervals are provided for all treatment effects reported in a paper

Summary

We use *P*-values to judge whether any effects we see are bigger than would be expected by chance

However, they suffer from a number of limitations and should not be interpreted in isolation

Any comparison should always be accompanied by some measure of effect size (e.g. the difference in proportions with a virological response) and a confidence interval for this effect

For some types of RCT, such as equivalence or non-inferiority trials, confidence intervals are even *more* important