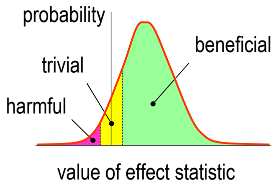
Making Inferences: Clinical vs Statistical Significance

Will G Hopkins (will@clear.net.nz)
AUT University
Auckland, NZ

- Hypothesis testing, p values, statistical significance
- Confidence limits
- Chances of benefit/harm



- This slideshow is an updated version of the slideshow in:
  - Batterham AM, Hopkins WG (2005). Making meaningful inferences about magnitudes. Sportscience 9, 6-13. See link at sportsci.org.
- Other resources:
  - Hopkins WG (2007). A spreadsheet for deriving a confidence interval, mechanistic inference and clinical inference from a p value. Sportscience 11, 16-20. See sportsci.org.
  - Hopkins WG, Marshall SW, Batterham AM, Hanin J (2009). Progressive statistics for studies in sports medicine and exercise science. Medicine and Science in Sports and Exercise 41, 3-12. (Also available at sportsci.org: Sportscience 13, 55-70, 2009.)

Background

- A major aim of research is to make an inference about an effect in a population based on study of a sample.
- Null-hypothesis testing via the P value and statistical significance is the traditional but flawed approach to making an inference.
- Precision of estimation via confidence limits is an improvement.
- But what's missing is some way to make inferences about the clinical, practical or mechanistic significance of an effect.
- I will explain how to do it via confidence limits using values for the smallest beneficial and harmful effect.
- I will also explain how to do it by calculating and interpreting chances that an effect is beneficial, trivial, and harmful.

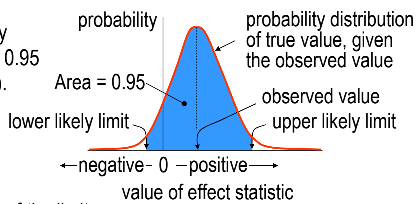
Hypothesis Testing, P Values and Statistical Significance

- Based on the notion that we can disprove, but not prove, things.
- Therefore, we need a thing to disprove.
- Let's try the null hypothesis: the population or true effect is zero.
- If the value of the observed effect is unlikely under this assumption, we reject (disprove) the null hypothesis.
- *Unlikely* is related to (but not equal to) the P value.
- $P < 0.05$ is regarded as unlikely enough to reject the null hypothesis (that is, to conclude the effect is not zero or null).
  - We say *the effect is statistically significant at the 0.05 or 5% level.*
  - Some folks also say *there is a real effect*.
- $P > 0.05$ means there is not enough evidence to reject the null.
  - We say *the effect is statistically non-significant.*
  - Some folks also accept the null and say *there is no effect*.

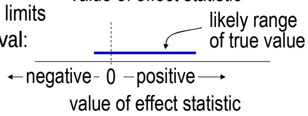
- Problems with this philosophy…
  - We can disprove things only in pure mathematics, not in real life.
  - Failure to reject the null doesn't mean we have to accept the null.
  - In any case, true effects are always "real", never zero. So…
  - The null hypothesis is always false!
  - Therefore, to assume that effects are zero until disproved is illogical and sometimes impractical or unethical.
  - 0.05 is arbitrary.
  - The P value is not a probability of anything in reality.
  - Some useful effects aren't statistically significant.
  - Some statistically significant effects aren't useful.
  - *Non-significant* is usually misinterpreted as *unpublishable*.
  - So good data don't get published.
- Solution: clinical significance or magnitude-based inferences via confidence limits and chances of benefit and harm.
  - Statistical significance = null-based inferences.

Clinical Significance via Confidence Limits

- Start with confidence limits, which define a range within which we infer the true, population or large-sample value is likely to fall.
  - *Likely* is usually a probability of 0.95 (for 95% limits).

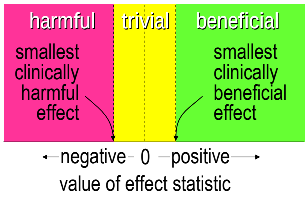


  - Representation of the limits as a confidence interval:



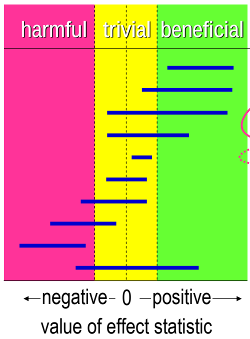
  - Caution: the confidence interval is not a range of responses!

- For clinical significance, we interpret confidence limits in relation to the smallest clinically beneficial and harmful effects.
  - These are usually equal and opposite in sign.
    - Harm is the opposite of benefit, not side effects.
  - They define regions of beneficial, trivial, and harmful values:



- The next slide is the key to clinical or practical significance.
  - All you need is these two things: the confidence interval and a sense of what is important (e.g., beneficial and harmful).

- Put the confidence interval and these regions together to make a decision about clinically significant, clear or decisive effects.

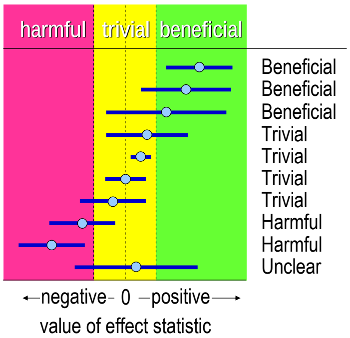


Clinically decisive?	Statistically significant?
Yes: use it.	Yes
Yes: use it.	Yes
Yes: use it.	No
Yes: depends.	No
Yes: don't use it.	Yes
Yes: don't use it.	No
Yes: don't use it.	No
Yes: don't use it.	Yes
Yes: don't use it.	Yes
No: need more research.	No

Why hypothesis testing is unethical and impractical!

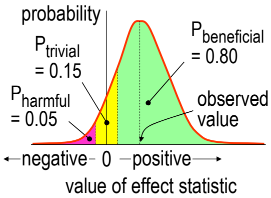
- UNDERSTAND THIS SLIDE!

- Making a crude call on magnitude.
  - Declare the observed magnitude of clinically clear effects.



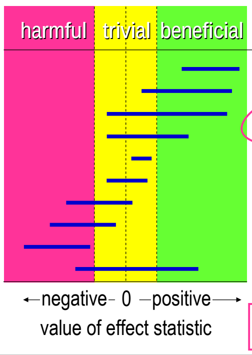
Clinical Significance via Clinical Chances

- We calculate probabilities that the true effect could be clinically beneficial, trivial, or harmful ($P_{beneficial}$, $P_{trivial}$, $P_{harmful}$).
- These Ps are NOT the proportions of positive, non- and negative responders in the population.
- Calculating the Ps is easy.
  - Put the observed value, smallest beneficial/harmful value, and P value into a spreadsheet at newstats.org.



- The Ps allow a more detailed call on magnitude, as follows…

- Making a more detailed call on magnitudes using chances of benefit and harm.



Chances (%) that the effect is harmful / trivial / beneficial	
0.01/0.3/99.7	Most likely beneficial
0.1/7/93	Likely beneficial
2/33/65	Possibly beneficial
1/59/40	Clinical: unclear (Mechanistic: possibly +ive)
0.2/97/3	Very likely trivial
2/94/4	Likely trivial
28/70/2	Possibly harmful
74/26/0.2	Possibly harmful
97/3/0.01	Very likely harmful
9/60/31	Mechanistic and clinical: unclear

Risk of harm >0.5% is unacceptable, unless chance of benefit is high enough.

- Use this table for the plain-language version of chances:

Probability	Chances	Odds	The effect… beneficial/trivial/harmful
<0.005	<0.5%	<1:199	is almost certainly not…
0.005–0.05	0.5–5%	1:999–1:19	is very unlikely to be…
0.05–0.25	5–25%	1:19–1:3	is unlikely to be…, is probably not…
0.25–0.75	25–75%	1:3–3:1	is possibly (not)…, may (not) be…
0.75–0.95	75–95%	3:1–19:1	is likely to be…, is probably…
0.95–0.995	95–99.5%	19:1–199:1	is very likely to be…
>0.995	>99.5%	>199:1	is almost certainly…

- An effect should be almost certainly not harmful (<0.5%) and at least possibly beneficial (>25%) before you decide to use it.
  - But you can tolerate higher chances of harm if chances of benefit are much higher: e.g., 3% harm and 76% benefit = clearly useful.
  - I use an odds ratio of benefit/harm of >66 in such situations.

- Two examples of use of the spreadsheet for clinical chances:

P value	value of statistic	Conf. level (%)	deg. of freedom	Confidence limits lower	Confidence limits upper	threshold values for clinical chances positive	threshold values for clinical chances negative
0.03	1.5	90	18	0.4	2.6	1	-1
0.20	2.4	90	18	-0.7	5.5	1	-1

Both these effects are clinically decisive, clear, or significant.

Chances (% or odds) that the true value of the statistic is					
clinically positive		clinically trivial		clinically negative	
prob (%)	odds	prob (%)	odds	prob (%)	odds
78	3:1	22	1:3	0	1:2071
likely, probable		unlikely, probably not		almost certainly not	
78	3:1	19	1:4	3	1:30
likely, probable		unlikely, probably not		very unlikely	

- How to Publish Clinical Chances
  Example of a table from a randomized controlled trial:

TABLE 1–Differences in improvements in kayaking sprint speed between slow, explosive and control training groups.

Compared groups	Mean improvement (%) and 90% confidence limits	Qualitative outcome[a]
Slow - control	3.1; ±1.6	Almost certainly beneficial
Explosive - control	2.6; ±1.2	Very likely beneficial
Slow - explosive	0.5; ±1.4	Unclear

[a] with reference to a smallest worthwhile change of 0.5%.

- Problem: what's the smallest clinically important effect?
  - If you can't answer this question, quit the field.
  - This problem applies also with hypothesis testing, because it determines sample size you need to test the null properly.
- Example: in many solo sports, ~0.5% change in power output changes substantially a top athlete's chances of winning.
- The default for most other populations and effects is Cohen's set of smallest values.
  - These values apply to **clinical**, **practical** and/or **mechanistic** importance…
  - Standardized changes or differences in the mean: 0.20 of the between-subject standard deviation.
    - In a controlled trial, it's the SD of all subjects in the **pre-test**, not the SD of the change scores.
  - Correlations: 0.10.
  - Injury or health risk, odds or hazard ratios: 1.1-1.3.

- Problem: these new approaches are not yet mainstream.
  - Confidence limits at least are coming in, so look for and interpret the importance of the lower and upper limits.
  - You can use a spreadsheet to convert a published P value into a more meaningful magnitude-based inference.
    - If the authors state "P<0.05" you can't do it properly.
    - If they state "P>0.05" or "NS", you can't do it at all.
- Problem: these approaches, and hypothesis testing, deal with uncertainty about an effect in a **population**.
  - But effects like risk of injury or changes in physiology or performance can apply to **individuals**.
  - Alas, more information and analyses are needed to make inferences about effects on individuals.
    - Researchers almost always ignore this issue, because…
    - they don't know how to deal with it, and/or…
    - they don't have enough data to deal with it properly.

Summary

- Show the observed magnitude of the effect.
- Attend to precision of estimation by showing 90% confidence limits of the true value.
- Do NOT show P values, do NOT test a hypothesis and do NOT mention statistical significance.
- Attend to clinical, practical or mechanistic significance by…
  - stating, with justification, the **smallest worthwhile effect**, then…
  - interpreting the **confidence limits** in relation to this effect, or…
  - estimating **probabilities** that the true effect is beneficial, trivial, and/or harmful (or substantially positive, trivial, and/or negative).
- Make a qualitative statement about the clinical or practical significance of the effect, using *unlikely, very likely,* and so on.
  - Remember, it applies to populations, not individuals.

For related articles and resources:

SPORTSCIENCE sportsci.org
A Peer-Reviewed Site for Sport Research

A New View of Statistics newstats.org
SUMMARIZING DATA — GENERALIZING TO A POPULATION
Simple & Effect Statistics, Dimension Reduction, Precision of Measurement, Confidence Limits, Statistical Models, Sample-Size Estimation