### Statistical vs Clinical Significance

**Will G Hopkins**  
Auckland University of Technology  
Auckland, NZ

Other titles:
- Statistical vs clinical, practical, or mechanistic significance.
- A more meaningful way to make inferences from a sample.
- Statistical significance is unethical; clinical significance isn’t.
- What are the chances your finding is beneficial or harmful?
- Publishing without hypotheses and statistical significance.
- Non-significant effect? No problem!

### Background

- Most researchers and students misinterpret statistical significance and non-significance.
- Few people know the meaning of the P value that defines statistical significance.
- Reviewers and editors reject some papers with statistically non-significant effects that should be published.
- Use of confidence limits instead of a P value is only a partial solution to these problems.
- We’re trying to make inferences about a population from a sample.
- What’s missing is some way to make inferences about the clinical or practical significance of an effect.

### Making Inferences in Research

- We study a sample to get an observed value of a statistic representing an interesting effect, such as the relationship between physical activity and health or performance.
- But we want the true (= population) value of the statistic.
- The observed value and the variability in the sample allow us to make an inference about the true value.
- Use of the P value and statistical significance is one approach to making such inferences.
- Its use-by date was December 31, 1999.
- There are better ways to make inferences.

### P Values and Statistical Significance

- Based on notion that we can disprove, but not prove, things.
- Therefore, we need something to disprove.
- Let's assume the true effect is zero: the null hypothesis.
- 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) a probability or P value.
- $P < 0.05$ is regarded as unlikely enough to reject the null hypothesis (i.e., to conclude the effect is not zero).
  - 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 not enough evidence to reject the null.
  - We say the effect is statistically non-significant.
  - Some folks 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 in real life are never zero. Never.
- So, THE NULL HYPOTHESIS IS ALWAYS FALSE!
- Therefore, to assume that effects are zero until disproved is illogical, and sometimes impractical or even unethical.
- 0.05 is arbitrary.
- The answer? We need better ways to represent the uncertainties of real life:
  - Better interpretation of the classical P value
  - More emphasis on (im)precision of estimation, through use of confidence limits for the true value
  - Better types of P value, representing probabilities of clinical or practical benefit and harm
Better Interpretation of the Classical P Value

- P/2 is the probability that the true value is negative.
  - Example: P = 0.24

  ![](probability_distribution.png)

  - Easier to understand, and avoids statistical significance, but...
  - Problem: having to halve the P value is awkward, although we could use one-tailed P values directly.
  - Problem: focus is still on zero or null value of the effect.

Confidence (or Likely) Limits of the True Value

- These define a range within which the true value is likely to fall.
  - "Likely" is usually a probability of 0.95 (defining 95% limits).

  ![](confidence_limits.png)

  - Problem: 0.95 is arbitrary and gives an impression of imprecision.
  - 0.90 or less would be better.
  - Problem: still have to assess the upper and lower limits and the observed value in relation to clinically important values.

Clinical Significance

- Statistical significance focuses on the null value of the effect.
- More important is clinical significance defined by the smallest clinically beneficial and harmful values of the effect.
- These values are usually equal and opposite in sign.
- Example:

  ![](clinical_significance.png)

  - We now combine these values with the observed value to make a statement about clinical significance.

Choosing the Smallest Clinically Important Value

- If you can’t meet this challenge, quit the field.
- For performance in many sports, ~0.5% increases a top athlete’s chances of winning.
- The default for most other populations is Cohen’s set of smallest worthwhile effect sizes.
  - This approach applies to the smallest clinically, practically and/or mechanistically important effects.
  - Correlations: 0.10
  - Relative risks: ~1.2, depending on prevalence of the disease or other condition.
  - Changes or differences in the mean: 0.20 between-subject standard deviations.

  ![](smallest_clinically.png)

  - More on differences or changes in the mean...
    - Why the between-subject standard deviation is important:
      - Trivial effect (0.1x SD): females, males
      - Very large effect (3x SD): females, males

  ![](between_subject.png)

  - You must also use the between-subject standard deviation when analyzing the change in the mean in an experiment.
    - Many meta-analysts wrongly use the SD of the change score.
You should describe outcomes in plain language in your paper. Therefore you need to describe the probabilities that the effect is beneficial, trivial, and/or harmful.

Suggested schema:

### Interpreting the Probabilities

<table>
<thead>
<tr>
<th>Probability</th>
<th>Chances</th>
<th>Odds</th>
<th>The effect is...</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.01</td>
<td>&lt;1%</td>
<td>&lt;1:99</td>
<td>is not...</td>
</tr>
<tr>
<td>0.01–0.05</td>
<td>1–5%</td>
<td>1:99–1:19</td>
<td>is very unlikely to be...</td>
</tr>
<tr>
<td>0.05–0.25</td>
<td>5–25%</td>
<td>1:19–1:3</td>
<td>is unlikely to be...</td>
</tr>
<tr>
<td>0.25–0.75</td>
<td>25–75%</td>
<td>1:3–1:1</td>
<td>is possibly (not)...</td>
</tr>
<tr>
<td>0.75–0.95</td>
<td>75–95%</td>
<td>3:1–19:1</td>
<td>is likely to be...</td>
</tr>
<tr>
<td>0.95–0.99</td>
<td>95–99%</td>
<td>19:1–99:1</td>
<td>is very likely to be...</td>
</tr>
<tr>
<td>&gt;0.99</td>
<td>&gt;99%</td>
<td>&gt;99:1</td>
<td>is...</td>
</tr>
</tbody>
</table>

### Publishing the Outcome

Example:

<p>| TABLE 2. Differences in improvements in kayaking performance between the slow, explosive and control training groups, and chances that the differences are substantial (greater than the smallest worthwhile change of 0.5%) for a top kayaker. |</p>
<table>
<thead>
<tr>
<th>Mean improvement (%) and 90% confidence limits</th>
<th>Chances (% and qualitative) of substantial improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slow - control</td>
<td>3.1; ±1.6</td>
</tr>
<tr>
<td>Explosive - control</td>
<td>2.0; ±1.2</td>
</tr>
<tr>
<td>Slow - explosive</td>
<td>1.1; ±1.4</td>
</tr>
</tbody>
</table>

Chances of substantial decline in performance all <5% (very unlikely).

### Summary

When you report your research...

- Show the observed magnitude of the effect.
- Attend to precision of estimation by showing 90% confidence limits of the true value.
- Show the P value if you must, but do not test a null hypothesis and do not mention statistical significance.
- Attend to clinical, practical or mechanistic significance by stating the smallest worthwhile value then showing the 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.

Examples showing use of the spreadsheet and the clinical importance of p=0.20

More examples on supplementary slides at end of slideshow.

Supplementary slides:

- Original meaning of P value
- More examples of clinical significance

This presentation is available from:

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

See Sportscience 6, 2002
Traditional Interpretation of the P Value

- Example: P = 0.20 for an observed positive value of a statistic
  - If the true value is zero, there is a probability of 0.20 of observing a more extreme positive or negative value.

<table>
<thead>
<tr>
<th>Probability distribution of observed value</th>
<th>if true value = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of effect statistic</td>
<td>observed value</td>
</tr>
<tr>
<td>Probability of positive value</td>
<td>0.1</td>
</tr>
<tr>
<td>Probability of negative value</td>
<td>0.1</td>
</tr>
<tr>
<td>Total probability</td>
<td>0.2</td>
</tr>
</tbody>
</table>

• Problem: huh? (Hard to understand.)
• Problem: everything that’s wrong with statistical significance.

More Examples of Clinical Significance

- Example 5–clinically trivial, statistically significant (publishable rare outcome that can arise from a large sample size; usually misinterpreted as a worthwhile effect):
  - The observed effect of the treatment was 1.1 units (90% likely limits 0.4 to 1.8 units; P = 0.007).
  - The chances that the true effect is practically beneficial/trivial/harmful are 99/1/0%.

Example 6–clinically trivial, statistically non-significant (publishable, but sometimes not submitted or accepted):
- The observed effect of the treatment was 0.3 units (90% likely limits –1.7 to 2.3 units; P = 0.80).
- The chances that the true effect is practically beneficial/trivial/harmful are 89/8/3%.

Example 3–clinically unclear, statistically non-significant (the worst kind of outcome, due to small sample or large error of measurement; usually rejected, but could/should be published to contribute to a future meta-analysis):
- The observed effect of the treatment was 2.7 units (90% likely limits –5.9 to 11 units; P = 0.60).
- The chances that the true effect is practically beneficial/trivial/harmful are 55/26/18%.

Example 4–clinically unclear, statistically significant (good publishable study; true effect is on the borderline of beneficial):
- The observed effect of the treatment was 1.9 units (90% likely limits 0.4 to 3.4 units; P = 0.04).
- The chances that the true effect is practically beneficial/trivial/harmful are 46/54/0%.

Example 1–clinically beneficial, statistically non-significant (inappropriately rejected by editors):
- The observed effect of the treatment was 6.0 units (90% likely limits –1.8 to 14 units; P = 0.20).
- The chances that the true effect is practically beneficial/trivial/harmful are 80/15/5%.

Example 2–clinically beneficial, statistically significant (no problem with publishing):
- The observed effect of the treatment was 3.3 units (90% likely limits 1.3 to 5.3 units; P = 0.007).
- The chances that the true effect is practically beneficial/trivial/harmful are 87/13/0%.

Example 5–clinically trivial, statistically non-significant (publishable rare outcome that can arise from a large sample size; usually misinterpreted as a worthwhile effect):
- The observed effect of the treatment was 1.1 units (90% likely limits 0.4 to 1.8 units; P = 0.007).
- The chances that the true effect is practically beneficial/trivial/harmful are 99/1/0%.

Example 6–clinically trivial, statistically non-significant (publishable, but sometimes not submitted or accepted):
- The observed effect of the treatment was 0.3 units (90% likely limits –1.7 to 2.3 units; P = 0.80).
- The chances that the true effect is practically beneficial/trivial/harmful are 89/8/3%.