Food security and novel food solutions: implications for food choice, safety, and waste: 2nd June
Newcastle

Establishing an evidence-based approach to food security

Gavin Stewart
Direction

• Why do we need evidence?
• What does our evidence look like?
• The Replication crisis
  – The dance of the p values
  – Publication bias
  – Reporting and researcher degrees of freedom
  – Novelty and theory
• Possible Solutions
Why do we need evidence?

- The challenge of feeding Nine billion people
  - No more land, climate change, increasing variability

*Science* 327, 812 (2010)
What does our evidence look like?

- The replication crisis

## Empirical evidence

<table>
<thead>
<tr>
<th>Domain</th>
<th>Findings</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicine</td>
<td>Out of 49 highly cited papers, 45 claimed that studied therapy was effective. Of these studies, 16% were contradicted by subsequent studies, 16% had found stronger effects than did subsequent studies, 44% were replicated, and 24% remained largely unchallenged.</td>
<td>Ioannidis JA (13 July 2005). Contradicted and initially stronger effects in highly cited clinical research. <em>JAMA</em>. <strong>294</strong> (2): 218–228.</td>
</tr>
<tr>
<td>Psychology</td>
<td>Out of 100 studies from high-ranking journals only 36% had significant findings (<em>p</em> value below .05) compared to 97% of the original studies. The mean effect size in the replications was approximately half the magnitude of the effects reported in the original studies.</td>
<td>Collaboration, Open Science (2015). &quot;Estimating the reproducibility of psychological science&quot;. <em>Science</em>. <strong>349</strong> (6251): aac4716.</td>
</tr>
<tr>
<td></td>
<td>Questionable research practices (QRPs) have been identified as common in the field (majority of 2000 scientists confess to at least one of: e.g. selective reporting, p-hacking, nonpublication of data, post-hoc storytelling (framing exploratory analyses as confirmatory analyses), manipulation of outliers.</td>
<td>Leslie JK.; Loewenstein, GP, Drazen (2012). &quot;Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling&quot;. <em>Psychological Science</em>. <strong>23</strong> (5): 524–532</td>
</tr>
</tbody>
</table>
The classical P value: The probability of observing data at least as extreme as the actual data given infinite observations…. assuming the null hypothesis to be true.
The dance of the P values

<table>
<thead>
<tr>
<th>Strength of evidence</th>
<th>Significance language</th>
</tr>
</thead>
<tbody>
<tr>
<td>P&lt;0.001</td>
<td>Very highly Significant</td>
</tr>
<tr>
<td>P&lt;0.01</td>
<td>Highly significant</td>
</tr>
<tr>
<td>P&lt;0.05</td>
<td>Significant</td>
</tr>
<tr>
<td>P 0.05 to ?</td>
<td>Approaching Significant</td>
</tr>
<tr>
<td>P&gt;0.1</td>
<td>Non-significant</td>
</tr>
</tbody>
</table>

## The dance of the P values

<table>
<thead>
<tr>
<th>Strength of evidence</th>
<th>Significance language</th>
<th>Suggests Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P &lt; 0.001$</td>
<td>Very highly Significant</td>
<td>There is definitely an effect</td>
</tr>
<tr>
<td>$P &lt; 0.01$</td>
<td>Highly significant</td>
<td>There is an effect</td>
</tr>
<tr>
<td>$P &lt; 0.05$</td>
<td>Significant</td>
<td>Most likely there is an effect</td>
</tr>
<tr>
<td>$P 0.05$ to ?</td>
<td>Approaching Significant</td>
<td>Almost? Probably? (but low power)</td>
</tr>
<tr>
<td>$P &gt; 0.1$</td>
<td>Non-significant</td>
<td>No effect?</td>
</tr>
</tbody>
</table>

# The dance of the P values

<table>
<thead>
<tr>
<th>Strength of evidence</th>
<th>Significance language</th>
<th>Suggests Truth</th>
<th>Evokes emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>P&lt;0.001</td>
<td>Very highly Significant</td>
<td>There is definitely an effect</td>
<td>Elation, Exuberance, Smugness?</td>
</tr>
<tr>
<td>P&lt;0.01</td>
<td>Highly significant</td>
<td>There is an effect</td>
<td>Dancing, Drinking</td>
</tr>
<tr>
<td>P&lt;0.05</td>
<td>Significant</td>
<td>Most likely there is an effect</td>
<td>Relief, Cheerfulness</td>
</tr>
<tr>
<td>P 0.05 to ?</td>
<td>Approaching Significant</td>
<td>Almost? Probably? (but low power)</td>
<td>Frustration (if only)</td>
</tr>
<tr>
<td>P&gt;0.1</td>
<td>Non-significant</td>
<td>No effect?</td>
<td>Despair, depression</td>
</tr>
</tbody>
</table>

# The dance of the P values

<table>
<thead>
<tr>
<th>Strength of evidence</th>
<th>Significance language</th>
<th>Suggests Truth</th>
<th>Evokes emotion</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>P&lt;0.001</td>
<td>Very highly Significant</td>
<td>There is definitely an effect</td>
<td>Elation, Exuberance</td>
<td>Nobel Prize Tenure Research Grant</td>
</tr>
<tr>
<td>P&lt;0.01</td>
<td>Highly significant</td>
<td>There is an effect</td>
<td>Dancing, Drinking</td>
<td>**** publication PhD</td>
</tr>
<tr>
<td>P&lt;0.05</td>
<td>Significant</td>
<td>Most likely there is an effect</td>
<td>Relief, Cheerfulness</td>
<td>*** publication</td>
</tr>
<tr>
<td>P 0.05 to ?</td>
<td>Approaching Significant</td>
<td>Almost? Probably? (but low power)</td>
<td>Frustration (if only)</td>
<td>Stress leave counselling</td>
</tr>
<tr>
<td>P&gt;0.1</td>
<td>Non-significant</td>
<td>No effect?</td>
<td>Despair, depression</td>
<td>Reconsider life goals</td>
</tr>
</tbody>
</table>

The Dance of the P values

• If P values are meaningful and represent the truth they should replicate...

• Lets run a simulation to see if they do...

https://www.youtube.com/watch?v=5OL1RqHrZQ8
Dance of the P values

- P values do not replicate
- (Over)reliance on P values has serious consequences for the rigour of our science…
And in Food

Publication bias

- Publication bias refers to bias that occurs when research found in the published literature is systematically unrepresentative of the population of studies (Rothstein et al., 2005).

- On average published studies have a larger mean effect size than unpublished studies, providing evidence for a publication bias (Lipsey and Wilson 1993).

- Also referred to as the ‘file drawer’ problem:
  “...journals are filled with the 5% of studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show non-significant (e.g. p < 0.05) results” (Rosenthal, 1979).

- Well-documented in different fields of research (biomedicine, public health, education, crime & justice, social welfare, ecology & evolution).

The funnel plot

A study

True effect from meta-analysis

Large studies close to true effect

95% of studies should be in the “funnel”

Small studies more variable

Low

Effect estimate

High

Standard error
Now with added publication bias

Studies missing from lower corner of funnel

Funnel is not symmetrical

Reporting and researcher degrees of freedom

• Do lots of things in different ways…and consciously or unconsciously introduce bias with selective reporting
• Develop an SEM with two different structures, split the data into male and female, analyse complete cases and imputed data…report only selected results (and worse selected methods)
• And just bad reporting of important information
In Food

• A systematic review of zinc biofortification

28 studies
Sample size reported for all
Only nine studies reported a measure of variance (standard error, standard deviation, confidence interval, variance)

M Barański, et al. British Journal of Nutrition 112 (05), 794-811
Novelty and theory

• Good research must be novel with sound theoretical underpinnings?
Good research updates our belief about evidence

Summary to date

- We’re BAD

- Over(reliance on p values)
- Publication bias
- Selective and poor reporting
- Fail to consider cumulative evidence appropriately
Solutions 1: P values

- Report and interpret effect sizes and confidence intervals (they convey much more information than p values)

- Establish universal reporting guidelines to enforce this [cf](https://www.equator-network.org)

- Some advocacy for banning p values altogether

Solution 2: Publication Bias

• Pre-registration

• TOP guidelines
  – Pre-registered
  – Open Data
  – Open Methods
Solution 3: selective and poor reporting

- See previous:
  - Less reliance on p values
  - Adherence to reporting guidelines
  - Pre-registration, open data, open methods
Solution 4: considering the cumulative evidence

- More high quality evidence synthesis
- Strength of evidence rather than novelty
- Systems approach to funding
  - Informed by ES and informing ES
  - Common outcomes rather than novelty
Acknowledgements

https://evidencesynthesisnewcastle.wordpress.com