Navigating the p-value swamp for meta-analysts

Edward Kroc* (ed.kroc@ubc.ca)

*Assistant Professor of Measurement, Evaluation, and Research Methodology
Department of Educational and Counselling Psychology, and Special Education
University of British Columbia, Vancouver, BC

March 27, 2019: UBC Department of Anesthesiology,
Pharmacology, and Therapeutics
In the past 10+ years, the usage of p-values to make inferential decisions has come under heavy criticism, e.g.

- J.P.A. Ioannidis (2005, 2016, etc.)
- Reproducibility Projects (Psychology 2012)
- ASA statement on p-values (Wasserstein et al. 2016)
P-values: a brief history

A brief history of the problem:

- **Pre-WWII:**
  - Fisher (p-values as strength of evidence)
  - Neyman, Pearson (null hypothesis significance testing)

- **Post-WWII**
  - Continued academic debate
  - Fisher-Neyman-Pearson hybrid becomes the norm via textbook industry
  - Bayesianism (Jeffreys, Lindley, et al.)
Formally, a **p-value** is defined as:

\[
p-value := \Pr(|T| \geq t_{obs} \mid H_0 \text{ true}),
\]

where \( T \) is our test statistic of choice and \( t_{obs} \) is the sample value of this statistic.
Informally, p-values take the following form:

\[ p\text{-value} \sim \Pr(\text{data} \mid H). \]

But what we are really interested in is usually the posterior probability:

\[ PP = \Pr(H \mid \text{data}). \]

These two quantities are related via conditional probability:

\[ \Pr(H \mid \text{data}) \propto \Pr(\text{data} \mid H) \cdot \Pr(H). \]

Notice that all-important second factor, \( \Pr(H) \): the prior probability of the hypothesis \( H \).
Thus, the relation

$$\Pr(H \mid \text{data}) \propto \Pr(\text{data} \mid H) \cdot \Pr(H)$$

requires consideration of the prior in order to translate information from a p-value into a statement about the believability of a hypothesis.

This is not *a priori* Bayesianism; this is just a consequence of conditional probability.

Even Fisher recognized this fact; thus, his focus on quality study design, targeted hypotheses, and *replication*. 
The role of prior probabilities

DID THE SUN JUST EXPLODE?
(IT's NIGHT, SO WE'RE NOT SURE.)

This neutrino detector measures whether the Sun has gone nova.

Then, it rolls two dice. If they both come up six, it lies to us. Otherwise, it tells the truth.

Let's try.

Detector! Has the Sun gone nova?

ROLL

YES.

XKCD 1132
The role of prior probabilities

DID THE SUN JUST EXPLODE?
(It’s night, so we’re not sure.)

This neutrino detector measures whether the sun has gone nova.

Then, it rolls two dice. If they both come up six, it lies to us. Otherwise, it tells the truth.

Let’s try. Detector! Has the sun gone nova?

Roll.

Yes.

Naive Analyst

Frequentist Statistician:
The probability of this result happening by chance is \( \frac{1}{36} = 0.027 \).
Since \( p < 0.05 \), I conclude that the sun has exploded.

Thoughtful Analyst

Bayesian Statistician:

BET YOU $50 IT HASN’T.

Replication?
What if we repeat this experiment 10 times and observe the same outcome each time?
The role of prior probabilities

What if we repeat this experiment 10 times and observe the same outcome each time? **Prior will still dominate.**
Frequentism, strong and weak

Implicit in the preceding is the role of replication:

- Vital to the frequentist paradigm
- Strong frequentism: poor model of reality?
  - Quantum indeterminancy
  - Dynamical systems
  - The “arrow of time"
- Weak frequentism: sufficient for most medical research
  - P-values, confidence intervals, NHST only make sense within the context of replication
What does a p-value = 0.04 represent?

- The probability that the null hypothesis is true is 0.04.
- The probability that the null hypothesis is true, given the data (sample value of the test statistic) is 0.04.
- Given a significance threshold of 0.05, it is likely that the null hypothesis is false.
- Given that the null hypothesis is true, the probability of observing a test statistic as or more extreme than the one we observed is 0.04.
- Given a significance threshold of 0.05, there is weak evidence against the null hypothesis.
- There is a 0.04 probability that the observed effect (test statistic) is due to random noise.
What does a p-value = 0.04 represent?

- The probability that the null hypothesis is true is 0.04.
- The probability that the null hypothesis is true, given the data (sample value of the test statistic) is 0.04.
- Given a significance threshold of 0.05, it is likely that the null hypothesis is false.
- Given that the null hypothesis is true, the probability of observing a test statistic as or more extreme than the one we observed is 0.04.
- Given a significance threshold of 0.05, there is weak evidence against the null hypothesis, assuming sufficient power.
- There is a 0.04 probability that the observed effect (test statistic) is due to random noise.
What does a 95% confidence interval of the mean represent?

- There is a 95% chance that the true (population) mean lies inside the 95% confidence interval.
- About 95% of the data lie within the 95% confidence interval.
- If we repeated the same experiment many times, about 95% of those experiments would produce sample means that fall inside our original experiment’s 95% confidence interval.
- If we repeated the same experiment many times and calculated a 95% confidence interval for the mean each time, then about 95% of those confidence intervals would contain the true population mean.
What does a 95% confidence interval of the mean represent?

- There is a 95% chance that the true (population) mean lies inside the 95% confidence interval.
- About 95% of the data lie within the 95% confidence interval.
- If we repeated the same experiment many times, about 95% of those experiments would produce sample means that fall inside our original experiment’s 95% confidence interval.
- If we repeated the same experiment many times and calculated a 95% confidence interval for the mean each time, then about 95% of those confidence intervals would contain the true population mean.
P-values and confidence intervals

The precise meanings of p-value and confidence interval are not useful absent replication.

- But we (usually) have to interpret p-values, CIs without the benefit of direct replications.

So instead, we have the following heuristic interpretations:

- Given a significance threshold of $\alpha$, a p-value $< \alpha$ is weak evidence against the null hypothesis, assuming sufficient power.

- A confidence interval for any point estimate (e.g. mean) is a rough quantification of the sample uncertainty of that point estimate.
Statistical power, the ability to detect non-null effects, is a function of many things, e.g.

- Study design
- Choice of test hypothesis (effect of interest)
- Choice of test procedure
- Sample size
- Sampling variance
- Measurement quality
- Targeted or true effect size
- Distribution of test statistic (distribution of data)
Statistical power, the ability to detect non-null effects, is a function of many things, e.g.

- Study design
- Choice of test hypothesis (effect of interest)
- Choice of test procedure
- Sample size
- Sampling variance
- Measurement quality
- Targeted or true effect size
- Distribution of test statistic (distribution of data)
Recall the standard $t$-test:

- If “large” sample sizes, then no problem (CLT)
- If “small” sample sizes, then data should be approx. normal
Normality assumptions

Recall the standard $t$-test:

- If “large” sample sizes, then no problem (CLT)
- If “small” sample sizes, then data should be approx. normal

 Effects of non-normality for small samples?
Normality assumptions

Power of the t-test for effect size $r=0.5$

Sample Size

Distribution
- NonNormal
- Normal

Graphic courtesy of O.L.O. Astivia
"My love-hate relationship with G*Power,"
psychometroscar.com
Normality assumptions

Recall the typical random (mixed) effect MA-model:

\[ \hat{\theta}_k = \theta_{RE} + \beta X_k + u_k + \varepsilon_k, \]

where:

- \( \hat{\theta}_k \) is the estimated effect of interest from study \( k \),
- \( \theta_{RE} \) is the (grand) mean of the true effects,
- \( \beta X_k \) is a (possibly absent) fixed effect(s),
- \( u_k \sim N(0, \tau^2) \) captures the random distribution of true effects about their mean,
- \( \varepsilon_k \sim N(0, \sigma_k^2) \) captures the sample uncertainty of study \( k \).
Two normality assumptions in play:

- $\varepsilon_k \sim N(0, \sigma_k^2)$: sometimes justifiable by an appeal to a CLT;
- $u_k \sim N(0, \tau^2)$: true effects in the meta-population are normally distributed. What if this fails?

- Can deflate power
- Can distort estimate of mean true effect $\widehat{\theta}_{RE}$
- Can seriously distort the prediction interval

Interestingly, lack of normality of random effect $u_k$ does not seem to greatly affect the quality of the fixed effect estimates $\beta$. 
Ex: Gamma distributed effect sizes

Histogram of true effects from 37 simulated studies
Ex: Gamma distributed effect sizes

Histogram of observed effects from 37 simulated studies
Ex: Gamma distributed effect sizes

| Study 1 | -4.52 [-5.49, -3.56] |
| Study 2 | -0.74 [-2.87, 1.39] |
| Study 3 | -3.13 [-4.75, -1.51] |
| Study 4 | -2.29 [-3.81, -0.77] |
| Study 5 | -7.55 [-9.18, -5.91] |
| Study 6 | -1.43 [-4.03, 1.18] |
| Study 7 | 2.00 [-0.71, 4.71] |
| Study 8 | -5.93 [-8.50, -3.36] |
| Study 9 | -10.13 [-11.85, -8.41] |
| Study 10 | -1.68 [-2.77, -0.58] |
| Study 11 | -12.46 [-13.24, -11.67] |
| Study 12 | -8.54 [-11.30, -5.79] |
| Study 13 | -4.20 [-6.49, -1.91] |
| Study 14 | -8.27 [-10.19, -6.34] |
| Study 15 | -1.28 [-2.15, -0.40] |
| Study 16 | -1.50 [-3.62, 0.62] |
| Study 17 | -4.49 [-5.60, -3.35] |
| Study 18 | -6.58 [-9.25, -3.91] |
| Study 19 | -2.18 [-4.44, 0.07] |
| Study 20 | -9.64 [-12.08, -7.20] |
| Study 21 | -2.66 [-4.71, -0.60] |
| Study 22 | -5.08 [-6.27, -3.89] |
| Study 23 | -2.65 [-4.26, -1.04] |
| Study 24 | -6.55 [-8.12, -4.98] |
| Study 25 | -5.54 [-7.46, -3.61] |
| Study 26 | -3.89 [-6.47, -1.31] |
| Study 27 | -2.64 [-4.53, -1.16] |
| Study 28 | -3.34 [-5.37, -1.31] |
| Study 29 | -0.77 [-3.06, 1.52] |
| Study 30 | 0.94 [-0.20, 2.07] |
| Study 31 | -6.05 [-8.49, -3.61] |
| Study 32 | -4.29 [-6.90, -1.68] |
| Study 33 | -3.17 [-4.61, -1.73] |
| Study 34 | -7.45 [-10.15, -4.75] |
| Study 35 | -1.72 [-2.64, -0.80] |
| Study 36 | -10.11 [-12.32, -7.89] |
| Study 37 | -4.12 [-6.45, -1.76] |

RE Model: -4.41 [-5.48, -3.35]
Ex: Gamma distributed effect sizes

<table>
<thead>
<tr>
<th>Study</th>
<th>Effect Size [Lower, Upper]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>-4.52 [-5.49, -3.56]</td>
</tr>
<tr>
<td>Study 2</td>
<td>-0.74 [-2.87, 1.39]</td>
</tr>
<tr>
<td>Study 3</td>
<td>-3.13 [-4.75, -1.51]</td>
</tr>
<tr>
<td>Study 4</td>
<td>-2.29 [-3.81, -0.77]</td>
</tr>
<tr>
<td>Study 5</td>
<td>-7.55 [-9.18, -5.91]</td>
</tr>
<tr>
<td>Study 6</td>
<td>-1.43 [-4.03, 1.18]</td>
</tr>
<tr>
<td>Study 7</td>
<td>2.00 [-0.71, 4.71]</td>
</tr>
<tr>
<td>Study 8</td>
<td>-5.93 [-8.50, -3.36]</td>
</tr>
<tr>
<td>Study 9</td>
<td>-10.13 [-11.85, -8.41]</td>
</tr>
<tr>
<td>Study 10</td>
<td>-1.68 [-2.77, -0.58]</td>
</tr>
<tr>
<td>Study 11</td>
<td>-12.46 [-13.24, -11.67]</td>
</tr>
<tr>
<td>Study 12</td>
<td>-8.54 [-11.30, -5.79]</td>
</tr>
<tr>
<td>Study 13</td>
<td>-4.20 [-6.49, -1.91]</td>
</tr>
<tr>
<td>Study 14</td>
<td>-8.27 [-10.19, -6.34]</td>
</tr>
<tr>
<td>Study 15</td>
<td>-1.28 [-2.15, -0.40]</td>
</tr>
<tr>
<td>Study 16</td>
<td>-1.50 [-3.62, 0.62]</td>
</tr>
<tr>
<td>Study 17</td>
<td>-4.49 [-5.60, -3.35]</td>
</tr>
<tr>
<td>Study 18</td>
<td>-6.58 [-9.25, -3.91]</td>
</tr>
<tr>
<td>Study 19</td>
<td>-2.18 [-4.44, 0.07]</td>
</tr>
<tr>
<td>Study 20</td>
<td>-9.64 [-12.08, -7.20]</td>
</tr>
<tr>
<td>Study 21</td>
<td>-2.66 [-4.71, -0.60]</td>
</tr>
<tr>
<td>Study 22</td>
<td>-5.08 [-6.27, -3.89]</td>
</tr>
<tr>
<td>Study 23</td>
<td>-2.65 [-4.26, -1.04]</td>
</tr>
<tr>
<td>Study 24</td>
<td>-6.55 [-8.12, -4.98]</td>
</tr>
<tr>
<td>Study 25</td>
<td>-5.54 [-7.46, -3.61]</td>
</tr>
<tr>
<td>Study 26</td>
<td>-3.89 [-6.47, -1.31]</td>
</tr>
<tr>
<td>Study 27</td>
<td>-2.84 [-4.53, -1.16]</td>
</tr>
<tr>
<td>Study 28</td>
<td>-3.34 [-5.37, -1.31]</td>
</tr>
<tr>
<td>Study 29</td>
<td>-0.77 [-3.06, 1.52]</td>
</tr>
<tr>
<td>Study 30</td>
<td>0.94 [-0.20, 2.07]</td>
</tr>
<tr>
<td>Study 31</td>
<td>-6.05 [-8.49, -3.61]</td>
</tr>
<tr>
<td>Study 32</td>
<td>-4.29 [-6.90, -1.68]</td>
</tr>
<tr>
<td>Study 33</td>
<td>-3.17 [-4.61, -1.73]</td>
</tr>
<tr>
<td>Study 34</td>
<td>-7.45 [-10.15, -4.75]</td>
</tr>
<tr>
<td>Study 35</td>
<td>-1.72 [-2.64, -0.80]</td>
</tr>
<tr>
<td>Study 36</td>
<td>-10.11 [-12.32, -7.89]</td>
</tr>
<tr>
<td>Study 37</td>
<td>-4.12 [-6.46, -1.78]</td>
</tr>
</tbody>
</table>

RE Model: -4.41 [-5.48, -3.35]
Non-normal distribution of effects can wreak particular havoc when MAing only a few studies:

Five studies, with one ostensible “outlier”.
Non-normal distribution of effects can wreak particular havoc when MAing only a few studies:

![Gamma(3,1) distribution](image)

Five studies, with one ostensible “outlier".
Non-normal distribution of effect sizes

Problem:
- MAs often contain too few studies to sufficiently assess the normality assumption.
- Negative effects of non-normality are amplified for small sample situations.

Advice:
- Change the assumed distribution of the random effect $u_k$.
- Do not discount studies that appear as outliers: downweight if you must.
- If still using a normal random effect, consider the profile likelihood approach for MAs with high heterogeneity (see Kontopantelis & Reeves, 2010).
Statistical power, the ability to detect non-null effects, is a function of many things, e.g.

- Study design
- Choice of test hypothesis (effect of interest)
- Choice of test procedure
- Sample size
- Sampling variance
- Measurement quality
- Targeted or true effect size
- Distribution of test statistic (distribution of data)
The precise meanings of p-value and confidence interval are not useful absent replication.

- But we (usually) have to interpret p-values, CIs without the benefit of direct replications.

So instead, we have the following heuristic interpretations:

- Given a significance threshold of $\alpha$, a p-value $< \alpha$ is weak evidence against the null hypothesis, assuming sufficient power.

- A confidence interval for any point estimate (e.g. mean) is a rough quantification of the sample uncertainty of that point estimate.
Significance bias

- It has been well established that statistically significant results are far more likely to be published (and at a greater rate) than studies that do not find statistical significance (see Dickersin 2005).

- This drives “p-hacking", ad hoc hypothesis selection, and other abuses of statistics.
Significance bias

Evidence for reporting bias

Proportion of studies not published

Effect size bias

- It has also been well established that it is easier to publish results with larger estimated effect sizes.

- This phenomenon is particularly pronounced in studies with small sample sizes. Why?
  - Answer: Type M error (Gelman & Carlin, 2014)
  - Mathematically, if your study is underpowered, then if you find a statistically significant result, the corresponding effect size must be exaggerated.
Type M error

Cohen's $d = 0.5$

Power = 8%
Type S = 9%
Type M = 4.8
Type M error

Cohen's $d = 1$

Power = 17%
Type S < 1%
Type M = 2.5
Type M error

Cohen's $d = 2$

Power = 52%
Type S < .01%
Type M = 1.4
Type M error

Cohen's $d = 4.5$

Power = 99%
Type S ~ 0%
Type M ~ 1
Type M error

In low-powered studies:

- Searching for significance will actually hurt your estimates and inferences.

- Significant results produce estimates that are wildly inaccurate.

- Seemingly small things like measurement error, sampling variability, or minor experimental imperfections become magnified.

- Results are often entirely driven by “noise”.

Kroc (UBC)
Weighting for Type M error

Upshot for meta-analysis: **downweight studies with low power!**

- Recall: inverse-variance weighting only downweights w.r.t. **sample size and sampling variance ...**

- ... but power depends on much more!

- What to do? Use a **different weighting!**
Weighting for Type M error

Upshot for meta-analysis: **downweight studies with low power!**

- Recall: inverse-variance weighting only downweights w.r.t. sample size and sampling variance . . .

- . . . but power depends on much more!

What to do? Use a different weighting!

For example, can specify weights that explicitly encode Type M error information:

$$w_k = \frac{1}{ER \cdot (\sigma^2_k + \tau^2)}$$

where $ER$ is the *exaggeration ratio* quantifying the Type M error.
Weighting for Type M error

Major challenge: Type M error calculation requires a hypothesized true effect size.

- Reenter the prior believability of a hypothesis (effect size)
- Must use the previous literature to determine this
- Could use other effect sizes to be MAed as a baseline
- Ideally, then conduct sensitivity analysis

Remember: *p-values and confidence intervals are only as reliable as the underlying power of their associated tests.*
Ex: small sample MA with one low-powered study

RE-MA of 5 studies (fit with DL method), using traditional inverse-variance weighting:

<table>
<thead>
<tr>
<th>Study</th>
<th>Effect Size</th>
<th>p-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>11.59%</td>
<td>12.21</td>
<td>[5.17, 19.25]</td>
</tr>
<tr>
<td>Study 2</td>
<td>19.45%</td>
<td>0.12</td>
<td>[-4.43, 4.67]</td>
</tr>
<tr>
<td>Study 3</td>
<td>16.86%</td>
<td>1.03</td>
<td>[-4.19, 6.25]</td>
</tr>
<tr>
<td>Study 4</td>
<td>26.37%</td>
<td>2.21</td>
<td>[-0.89, 5.31]</td>
</tr>
<tr>
<td>Study 5</td>
<td>25.74%</td>
<td>3.41</td>
<td>[0.19, 6.63]</td>
</tr>
<tr>
<td>RE Model</td>
<td>100.00%</td>
<td>3.07</td>
<td>[0.20, 5.95]</td>
</tr>
</tbody>
</table>

Observed Outcome
Ex: small sample MA with one low-powered study

RE-MA of 5 studies (fit with DL method), using augmented Type M + inverse-variance weighting, assuming maximum believable effect size of twice the largest reliable observed effect: $ER(1) = 1.44$

<table>
<thead>
<tr>
<th>Study</th>
<th>Effect Size</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>8.35%</td>
<td>12.21 [5.17, 19.25]</td>
</tr>
<tr>
<td>Study 2</td>
<td>20.16%</td>
<td>0.12 [-4.43, 4.67]</td>
</tr>
<tr>
<td>Study 3</td>
<td>17.47%</td>
<td>1.03 [-4.19, 6.25]</td>
</tr>
<tr>
<td>Study 4</td>
<td>27.34%</td>
<td>2.21 [-0.89, 5.31]</td>
</tr>
<tr>
<td>Study 5</td>
<td>26.68%</td>
<td>3.41 [0.19, 6.63]</td>
</tr>
<tr>
<td>RE Model</td>
<td>100.00%</td>
<td>2.74 [-0.15, 5.63]</td>
</tr>
</tbody>
</table>
Ex: small sample MA with one low-powered study

RE-MA of 5 studies (fit with DL method), using augmented Type M + inverse-variance weighting, assuming maximum believable effect size of 1.5 times the largest reliable observed effect: $ER(1) = 1.82$

<table>
<thead>
<tr>
<th>Study</th>
<th>p-value (%)</th>
<th>Effect Size</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>6.71%</td>
<td>12.21</td>
<td>[5.17, 19.25]</td>
</tr>
<tr>
<td>Study 2</td>
<td>20.52%</td>
<td>0.12</td>
<td>[-4.43, 4.67]</td>
</tr>
<tr>
<td>Study 3</td>
<td>17.79%</td>
<td>1.03</td>
<td>[-4.19, 6.25]</td>
</tr>
<tr>
<td>Study 4</td>
<td>27.83%</td>
<td>2.21</td>
<td>[-0.89, 5.31]</td>
</tr>
<tr>
<td>Study 5</td>
<td>27.16%</td>
<td>3.41</td>
<td>[0.19, 6.63]</td>
</tr>
</tbody>
</table>

RE Model: 100.00% 2.57 [-0.34, 5.47]

Observed Outcome

Kroc (UBC) p-values for meta-analysts March 27, 2019 46/49
Ex: small sample MA with one low-powered study

RE-MA of 5 studies (fit with DL method), using augmented Type M + inverse-variance weighting, assuming maximum believable effect size of max. value in 95% PI of mean discounting Study 1: \( ER(1) = 2.24 \)

<table>
<thead>
<tr>
<th>Study</th>
<th>Effect Size (%)</th>
<th>p-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>4.77%</td>
<td>0.034</td>
<td>[ 5.17, 19.25]</td>
</tr>
<tr>
<td>Study 2</td>
<td>20.95%</td>
<td>0.12</td>
<td>[-4.43, 4.67]</td>
</tr>
<tr>
<td>Study 3</td>
<td>18.15%</td>
<td>1.03</td>
<td>[-4.19, 6.25]</td>
</tr>
<tr>
<td>Study 4</td>
<td>28.40%</td>
<td>2.21</td>
<td>[-0.89, 5.31]</td>
</tr>
<tr>
<td>Study 5</td>
<td>27.72%</td>
<td>3.41</td>
<td>[ 0.19, 6.63]</td>
</tr>
<tr>
<td>RE Model</td>
<td>100.00%</td>
<td>2.37</td>
<td>[-0.57, 5.31]</td>
</tr>
</tbody>
</table>

Observed Outcome

Kroc (UBC)
In summary

- P-values and CIs are built to be interpreted in the context of direct replications.

- They do not directly tell us about the believability of a hypothesis, given the data.

- Proper interpretation requires careful consideration of the prior believability of the target hypotheses.

- Furthermore, without direct replication, estimates/tests need sufficient power to yield reliable p-values and CIs.

- *Always* assess the power and prior believability of reported effects/tests.

- In MA, build these assessments explicitly into the weights of your estimator.
Thank you!

Edward Kroc
Assistant Professor of Measurement, Evaluation, and Research Methodology
Department of ECPS, UBC

Office: Scarfe 2526
Phone: (604)-822-8671
Email: ed.kroc@ubc.ca
IG: @ed_kroc
Website: ekroc.weebly.com

Thank you!