Transparencia y reproducibilidad en el desarrollo de investigación en salud pública
Presented by Nerissa Nance y Mariana Morales
Grupo de trabajo de transparencia en la dirección de Economía Instituto de Salud Pública,
Cuernavaca, México
20 Septiembre 2017
SOME SONGS CAN CHANGE THE AGE OF THE LISTENER

Joseph P. Simmons et. Al., 2011

Study 1:
- Does listening to a children's song makes people feel old.
- 30 students from the University of Pennsylvania were randomly assigned to listen to a children's song or a song control.
- After listening to the song they answered a survey asking: "How old do you feel right now?", they reported the age of their father to control the variation in the age of the participants.
- A ANCOVA showed the expected effect: people felt older after hearing the children's song compared to listening to the song control (p = .033).

Study 2:
- Once it was proven that listening to children's songs makes people feel older, they investigated whether listening to a song about being older made people feel younger.
- Twenty students were asked to listen to "When I'm Sixty-Four" or a song control. They indicated their date of birth and the age of their parents.
- Variation in age of participants and the age of the parent were monitored.
- A ANCOVA showed the expected effect: people felt a year and a half younger after listening to "when I'm Sixty-Four" compared to those who heard the song control (p = .040).
P- Hacking (Degrees of freedom of the researcher):

It is the manipulation, unconscious or conscious, of statistical analyses and data eligibility specifications to obtain statistically significant results.

P-hacking increases the probability of having type 1 errors, that is to say, to find false positives by incorrectly rejecting the null hypothesis.

Can inflate the effect size found.

Joseph P. Simmons et al., 2011; Megan L. Head et al., 2015
How does P-hacking happen?

- Conducting exploratory research without a hypothesis
- Flexibility in when to stop collecting data
- Exclude certain observations
- Analyze different variables and only report the statistically significant ones
- Transforming the data
- Don’t ensure reproducibility
- Just report what’s convenient
- And many more…

Wicherts et al., 2016
<table>
<thead>
<tr>
<th>Code</th>
<th>Related</th>
<th>Type of degrees of freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hypothesizing</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>R6</td>
<td>Conducting explorative research without any hypothesis</td>
</tr>
<tr>
<td>T2</td>
<td></td>
<td>Studying a vague hypothesis that fails to specify the direction of the effect</td>
</tr>
<tr>
<td><strong>Design</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td>A8</td>
<td>Creating multiple manipulated independent variables and conditions</td>
</tr>
<tr>
<td>D2</td>
<td>A10</td>
<td>Measuring additional variables that can later be selected as covariates, independent variables, mediators, or moderators</td>
</tr>
<tr>
<td>D3</td>
<td>A5</td>
<td>Measuring the same dependent variable in several alternative ways</td>
</tr>
<tr>
<td>D4</td>
<td>A7</td>
<td>Measuring additional constructs that could potentially act as primary outcomes</td>
</tr>
<tr>
<td>D5</td>
<td>A12</td>
<td>Measuring additional variables that enable later exclusion of participants from the analyses (e.g., awareness or manipulation checks)</td>
</tr>
<tr>
<td>D6</td>
<td></td>
<td>Failing to conduct a well-founded power analysis</td>
</tr>
<tr>
<td>D7</td>
<td>C4</td>
<td>Failing to specify the sampling plan and allowing for running (multiple) small studies</td>
</tr>
<tr>
<td><strong>Collection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td></td>
<td>Failing to randomly assign participants to conditions</td>
</tr>
<tr>
<td>C2</td>
<td></td>
<td>Insufficient blinding of participants and/or experimenters</td>
</tr>
<tr>
<td>C3</td>
<td></td>
<td>Correcting, coding, or discarding data during data collection in a non-blinded manner</td>
</tr>
<tr>
<td>C4</td>
<td>D7</td>
<td>Determining the data collection stopping rule on the basis of desired results or intermediate significance testing</td>
</tr>
<tr>
<td><strong>Analyses</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td></td>
<td>Choosing between different options of dealing with incomplete or missing data on ad hoc grounds</td>
</tr>
<tr>
<td>A2</td>
<td></td>
<td>Specifying pre-processing of data (e.g., cleaning, normalization, smoothing, motion correction) in an ad hoc manner</td>
</tr>
<tr>
<td>A3</td>
<td></td>
<td>Deciding how to deal with violations of statistical assumptions in an ad hoc manner</td>
</tr>
<tr>
<td>A4</td>
<td></td>
<td>Deciding on how to deal with outliers in an ad hoc manner</td>
</tr>
<tr>
<td>A5</td>
<td>D3</td>
<td>Selecting the dependent variable out of several alternative measures of the same construct</td>
</tr>
<tr>
<td>A6</td>
<td></td>
<td>Trying out different ways to score the chosen primary dependent variable</td>
</tr>
<tr>
<td>A7</td>
<td>D4</td>
<td>Selecting another construct as the primary outcome</td>
</tr>
<tr>
<td><strong>Reporting</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td></td>
<td>Failing to assure reproducibility (verifying the data collection and data analysis)</td>
</tr>
<tr>
<td>R2</td>
<td></td>
<td>Failing to enable replication (re-running of the study)</td>
</tr>
<tr>
<td>R3</td>
<td></td>
<td>Failing to mention, misrepresenting, or misidentifying the study preregistration</td>
</tr>
<tr>
<td>R4</td>
<td></td>
<td>Failing to report so-called “failed studies” that were originally deemed relevant to the research question</td>
</tr>
<tr>
<td>R5</td>
<td></td>
<td>Mereporting results and p-values</td>
</tr>
<tr>
<td>R6</td>
<td>T1</td>
<td>Presenting exploratory analysis as confirmatory (HARKing)</td>
</tr>
</tbody>
</table>
Returning to studies 1 and 2

Table 3. Study 2: Original Report (in Bold Text) and the Requirement-Compliant Report (With Addition of Gray Text)

| Using the same method as in Study 1, we asked 29 University of Pennsylvania undergraduates to listen only to either “When I’m Sixty-Four” by The Beatles or “Kalimba” or “Hot Potato” by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. Then, in an ostensibly unrelated task, they indicated only their birth date (mm/dd/yyyy) and how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with “computers are complicated machines,” their father’s age, their mother’s age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as “the good old days,” and their gender. We used father’s age to control for variation in baseline age across participants. An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to “When I’m Sixty-Four” (adjusted \( M = 20.1 \) years) rather than to “Kalimba” (adjusted \( M = 21.5 \) years), \( F(1, 17) = 4.92, p = .040 \). Without controlling for father’s age, the age difference was smaller and did not reach significance (\( Ms = 20.3 \) and 21.2, respectively), \( F(1, 18) = 1.01, p = .33 \). |
HOW BAD CAN IT BE?

Degrees of Freedom analyzed:
• Flexibility by choosing dependent variables
• Flexibility in sample size
• Flexibility in the use of covariates
• Flexibility reporting subsets of experimental conditions

Table 1. Likelihood of Obtaining a False-Positive Result

<table>
<thead>
<tr>
<th>Researcher degrees of freedom</th>
<th>p &lt; .1</th>
<th>p &lt; .05</th>
<th>p &lt; .01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Situation A: two dependent variables (r = .50)</td>
<td>17.8%</td>
<td>9.5%</td>
<td>2.2%</td>
</tr>
<tr>
<td>Situation B: addition of 10 more observations per cell</td>
<td>14.5%</td>
<td>7.7%</td>
<td>1.6%</td>
</tr>
<tr>
<td>Situation C: controlling for gender or interaction of gender with treatment</td>
<td>21.6%</td>
<td>11.7%</td>
<td>2.7%</td>
</tr>
<tr>
<td>Situation D: dropping (or not dropping) one of three conditions</td>
<td>23.2%</td>
<td>12.6%</td>
<td>2.8%</td>
</tr>
<tr>
<td>Combine Situations A and B</td>
<td>26.0%</td>
<td>14.4%</td>
<td>3.3%</td>
</tr>
<tr>
<td>Combine Situations A, B, and C</td>
<td>50.9%</td>
<td>30.9%</td>
<td>8.4%</td>
</tr>
<tr>
<td>Combine Situations A, B, C, and D</td>
<td>81.5%</td>
<td>60.7%</td>
<td>21.5%</td>
</tr>
</tbody>
</table>

Simmons et al., 2011
If the published data is skewed there will be problems with:

- The inference
- Difficulties in trying to replicate findings from studies with false positives
- Some journals will not publish null results
- Waste of resources: inspire investment in fruitless research programs and therefore may have ineffective policy changes.
- Loss of credibility for fields of research that are known to publish false positives.

Simmons et al., 2011
HOW TO AVOID P-HACKING?

- Registration of studies
- Pre-analysis plans
Avoid publishing bias

- Register the study before its implementation, this increases the transparency of the work.
- It's a way to publish and keep it accessible to other researchers, the hypothesis, data collection plan and the way in which the data will be analyzed.

- The type of registration is dependent on the type of study.
- Some magazines request that registration is completed with specific sites in order to be published in their magazine.

- Registration should be submitted before beginning data collection.
Decide with the principal investigator which registration option is best

Options:

- **Clinical Trials**: Administered by the US National Institute of Health. Principle uses is to register clinical trials. This site is a good choice for studies that require randomization of participants. INSP has an account with this institute.
- **Registry for International Impact Evaluations (RIDIE)**: Specifically for impact evaluations and quasi experimental experiments.
- **Open Science Framework (OSF)**: A tool to help you register your studies, upload pre-analysis plans and link your study to relevant websites. Any type of studies are allowed.
One of the best ways to reduce the researcher's degrees of freedom in an analysis is to make a pre-analysis plan. This plan summarizes your hypothesis, data collection plan, and data analysis plan.
PRE-ANALYSIS PLAN TEMPLATE

I. General description
   - Study design
   - Study population
   - Arm/interaction group description
   - Number of participants at each arm
   - Inclusion and exclusion criteria
   - Study period
   - Analysis objectives/research questions
   - Hypotheses

II. Data collection
   - How will data be collected?
   - Sources used
   - Data collection time periods
   - Primary and secondary outcomes:
     - Name and definition of the specific outcome measure, including the titles of any categories into which the outcome measure data are aggregated.
     - Description of the metric and the unit of measure used to characterize the specific outcome measure.
     - Time point(s) at which the measurement will be assessed for the specific metric.
   - Covariates:
     - Name, definition and titles of any categories
     - Description of the metric and the unit of measure used to characterize the specific outcome measure
     - Time point(s) at which the measurement will be assessed for the specific metric.

III. Analysis
   - The effect measures and statistical methods used to address each objective
   - Hypothesis test used and p-value
   - Model specification
     - Methods of handling missing data:
       - Which confounders will be controlled and how they will be addressed?
       - Methods for assessing the level of confounding adjustment achieved
     - Any sensitivity analyses or sub-analyses
   - Handling of missing data:
     - How missing data will be handled
     - Methods of imputation
     - Sensitivity analyses for handling missing data

General description

Data collection

Analysis
GENERAL INFORMATION

- Design of the study
- Study population
- Description of the intervention
  - Number of participants in each group
  - Inclusion and exclusion criteria
- Time period
- Objectives
- Hypothesis
DATA COLLECTION (I)

• How to plan your data collection
• Sources of data
• Time period for data
• Results (primary and secondary)
  • Name and definition of each result
  • Points of time in which measurements will be taken, and the results that will be measured.
DATA COLLECTION (II)

• Covariates
  • Name, definitions, and titles of any categories.
  • Description of the metric and unit of measurement used for each covariate.
  • Points in time in which you are going to measure the variable.
ANALYSIS (I)

• Measures of effect and methods used to address each objective.
• Hypothesis tests using p-values
• Model specifications
• Methods to address "confounding", such as:
  • What are considered confounding variables and how will they be defined?
  • Methods to quantify "confounding" and how they will adjusted
ANALYSIS (II)

- Sensibility analysis and sub analysis
- Approaches to attrition
- Approaches to missing data
  - How to report missing data
  - Imputation methods
  - Sensibility analysis
THE MOST IMPORTANT STEP...

Upload to a repository!

So your methods can be seen by the scientific community
HOW TO USE A RAP?

A RAP can be useful for both intervention studies and writing and uploading the data before analyzing.
THANK YOU
TRENDINSP
Transparencia y reproducibilidad en el desarrollo de investigación en salud pública