Worst Practices in Statistical Data Analysis

Best Practices in Statistical Data Analysis
Willem Heiser farewell symposium 30 January 2014

Richard Gill
http://www.math.leidenuniv.nl/~gill
A talk within a talk & some reflections on scientific integrity

Flashback: one year ago in Tilburg

Meeting of the social science section of the Dutch statistical society: *Statistiek uit de bocht* (Statistics round the bend)

Ideas stolen from Marcel van Assen, Jelte Wicherts, Frank van Kolfschooten, Han van der Maas, Howard Wainer …
Integrity or fraud …
or just questionable research practices?
… or … ?

Richard Gill

Original talk December 2012; updated July 2013

http://www.math.leidenuniv.nl/~gill
- Smeesters affair
- Geraerts affair

Smeesters: closed
Geraerts: open, controversial
August 2011: a friend draws attention of Uri Simonsohn (Wharton School, Univ. Penn.) to “The effect of color ... ” by D. Smeesters and J. Liu
Simonsohn does preliminary statistical analysis indicating results are “too good to be true”

Fig. 1. Number of correct answers as a function of color, prime, and dimension.

Hint: text mentions a number of within group SD’s; group sizes ≈ 14
$3 \times 2 \times 2 \times 2$ design,
$\approx 14$ subjects per group

- Outcome: # correct answers in 20 item multiple choice general knowledge quiz

- Three treatments:
  - Colour: red, white, blue
  - Stereotype or exemplar
  - Intelligent or unintelligent
<table>
<thead>
<tr>
<th>Unintelligent</th>
<th>Intelligent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exemplar</strong></td>
<td><strong>Stereotype</strong></td>
</tr>
<tr>
<td>Kate Moss</td>
<td>A supermodel</td>
</tr>
<tr>
<td>Albert Einstein</td>
<td>A professor</td>
</tr>
</tbody>
</table>
Priming

- Red makes one see differences
- Blue makes one see similarities
- White is neutral
- Seeing an intelligent person makes you feel intelligent if you are in a “blue” mood
- Seeing an intelligent person makes you feel dumb if you are in a “red” mood
- Effect depends on whether you see exemplar or stereotype
The theory predicts something very like the picture (an important three way interaction!)

Simonsohn does preliminary statistical analysis indicating “too good to be true”

September 2011: Simonsohn corresponds with Smeesters, obtains data, distribution-free analysis confirms earlier findings

Simonsohn discovers same anomalies in more papers by Smeesters, more anomalies

Smeesters’ hard disk crashes, all original data sets lost. None of his coauthors have copies. All original sources (paper documents) lost when moving office

Smeesters and Simonsohn report to authorities

June 2012: Erasmus CWI report published, Smeesters resigns, denies fraud, admits data-massage “which everyone does”
What did Simonsohn actually do?

- Erasmus report is censored, authors refuse to answer questions, Smeesters and Liu data is unobtainable, identity Simonsohn unknown

- Some months later: identity Simonsohn revealed, uncensored version of report published

- November 2012: Uri Simonsohn posts “Just Post it: The Lesson from Two Cases of Fabricated Data Detected by Statistics Alone”
  Two cases? Smeesters, Sanna; third case, inconclusive (original data not available)

- December 2012: original data still unavailable, questions to Erasmus CWI still unanswered

- March 2013: Simonsohn paper published, data posted
- Theory predicts that the 12 experimental groups can be split into two sets of 6
- Within each set, groups should be quite similar
- Smeesters & Liu report some of the group averages and some of the group SD’s
- Theory: variance of group average = within group variance divided by group size!
- The differences between group averages are too small compared to the within group variances!
• Simonsohn proposes ad-hoc (*) test-statistic (comparing between group to within group variance), null distribution evaluated using parametric bootstrap

• When original data is made available, can repeat with non-parametric bootstrap

• Alternative: permutation tests

• Note: to do this, he pools each set of six groups. “Assumption” that there is no difference between the groups within each of the two sets of six groups is conservative

(*) Fisher: use left tail-probability of F- test for testing too good to be true
A picture tells 1000 words

```r
sigma <- 2.9
pattern <- c(rep(c(1, 0), 3), rep(c(0, 1), 3))
means <- pattern
means[pattern == 1] <- 11.75
means[pattern == 0] <- 9.5
set.seed(2013)
par(mfrow = c(3,4), bty = "n", xaxt = "n", yaxt = "n")
for (i in 1:12) {
  averages <- rnorm(12, mean = means, sd = sigma/sqrt(14))
  dim(averages) <- c(2, 6)
  averages <- rbind(averages-6, 0)
  plot(c(0, 20), c(0, 7), xlab = "", ylab = "", type = "n")
  abline(h = 0:6)
  barplot(as.vector(averages), col = rep(c("black", "white", "white"), n = 6), add = TRUE)
}
```
Spot the odd one out!
Spot the odd one out!
Further analyses

### Table 1. Means (SD) for 12 conditions in Smeesters et al. (2011)

<table>
<thead>
<tr>
<th></th>
<th>Predicted low</th>
<th>Predicted high</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>(SD)</td>
<td>(SD)</td>
</tr>
<tr>
<td>Predicted low</td>
<td>9.07</td>
<td>9.43</td>
</tr>
<tr>
<td></td>
<td>(2.55)</td>
<td>(2.82)</td>
</tr>
<tr>
<td>Predicted high</td>
<td>11.43</td>
<td>11.71</td>
</tr>
<tr>
<td></td>
<td>(2.79)</td>
<td>(2.87)</td>
</tr>
</tbody>
</table>

Note: Summary statistics for number of correct answers (out of 20) in a general knowledge task taken by 169 participants assigned to 12 conditions, six conditions were predicted to have high means, the other low. Each condition had \(n=14\), except those with superscripts. \(^a\) \(n=16\), \(^b\) \(n=13\).
Further analyses

Simonsohn’s test-statistic is actually equivalent to standard ANOVA F-test of hypothesis “each of two groups of six conditions have the same mean” – except that we want to reject if the statistic is too small.
\begin{verbatim}
data <- data.frame(score = scores, colour = colour, prime = prime, 
dimension=dimension, pattern=pattern.long)

result.aov.full <- aov(score~colour*prime*dimension, data = data)
result.aov.null <- aov(score~(colour+prime+dimension)^2, data = data)
result.anova <- anova(result.aov.null, result.aov.full)
result.anova

result.aov.zero <- aov(score ~ pattern, data=data)
result.anova.zero <- anova(result.aov.zero, result.aov.full)

result.anova.zero$F[2]
pf(result.anova.zero$F[2], df1=10, df2=156)
\end{verbatim}

**Test of 3-way interaction**

> result.anova

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Model</th>
<th>Expression</th>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>score ~ (colour + prime + dimension)^2</td>
<td>159</td>
<td>1350.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>score ~ colour * prime * dimension</td>
<td>157</td>
<td>1299.6</td>
<td>2</td>
<td>51.155</td>
<td>3.0898</td>
<td>0.04829*</td>
</tr>
</tbody>
</table>

Smeesters and Liu (OK, except d.f.)

The same ANOVA on the number of correct answers yielded a significant three-way interaction between color, prime, and dimension, $F(1, 157) = 3.08, p < .05$ (see Fig. 1). We further analyzed this...
data <- data.frame(score = scores, colour = colour, prime = prime, 
                  dimension=dimension, pattern=pattern.long)

result.aov.full <- aov(score~colour*prime*dimension, data = data)
result.aov.null <- aov(score~(colour+prime+dimension)^2, data = data)
result.anova <- anova(result.aov.null, result.aov.full)
result.anova

result.aov.zero <- aov(score ~ pattern, data=data)
result.anova.zero <- anova(result.aov.zero, result.aov.full)

result.anova.zero$F[2]
pf(result.anova.zero$F[2], df1=10, df2=156)

> result.anova.zero$F[2]
[1] 0.0941672
> pf(result.anova.zero$F[2], df1=10, df2=156)
[1] 0.0001445605
Further analyses

• Scores (integers) appear *too uniform*

For example, the fourteen scores for one of the twelve conditions were:

[6, 7, 7, 8, 8, 9, 9, 10, 10, 10, 12, 12, 14, 15]. The mode here is 10 and it appears three times.

Across the twelve conditions nine had the mode appearing 3 times, and three just 2 times.

The sum of mode frequencies, \( F \), is hence \( F = 9 \times 3 + 3 \times 2 = 33 \).

• Permutation test: \( p\text{-value} = 0.00002 \)
Linking thought suppression and recovered memories of childhood sexual abuse

Elke Geraerts
Maastricht University, the Netherlands, and Harvard University, Cambridge, MA, USA

Richard J. McNally
Harvard University, Cambridge, MA, USA

Marko Jelicic, Harald Merckelbach, and Linsey Raymaekers
Maastricht University, the Netherlands

There are two types of recovered memories: those that gradually return in recovered memory therapy and those that are spontaneously recovered outside the context of therapy. In the current study, we employed a thought suppression paradigm, with autobiographical experiences as target thoughts, to test whether individuals reporting spontaneously recovered memories of childhood sexual abuse (CSA) are more adept at suppressing positive and anxious autobiographical thoughts, relative to individuals reporting CSA memories recovered in therapy, relative to individuals with continuous abuse memories, and relative to controls reporting no history of abuse. Results showed that people reporting spontaneously recovered memories are superior in suppressing anxious autobiographical thoughts, both in the short term and long term (7 days). Our findings may partly explain why people with
Reduced Meta-Consciousness of Intrusions as an Explanation for Recovered Memory Reports

Elke Geraerts¹,²*, Richard J. McNally³, Harald Merckelbach², Anne-Laura van Harmelen⁴, Linsey Raymaekers², & Jonathan W. Schooler⁵

¹School of Psychology, University of St. Andrews, United Kingdom
²Department of Clinical Psychological Science, Maastricht University, The Netherlands
³Department of Psychology, Harvard University, United States of America
⁴Department of Psychology, Leiden University, The Netherlands
⁵Department of Psychology, University of California, Santa Barbara, United States of America

Word Count: 3,404

*To whom correspondence should be addressed: Elke Geraerts, E-mail: elke.geraerts@st-andrews.ac.uk

Abstract

People with spontaneously recovered memories of childhood sexual abuse (CSA) have been shown to be especially susceptible to underestimating their prior remembering of the abuse events. The current study examined whether this may be explained by a reduced “meta-consciousness” of their intrusions related to those events: That is, are these individuals failing to notice that memories of abuse do come to mind, thereby producing the illusion that they repressed the abuse events for many years? We used an adapted thought-suppression paradigm...
Senior author Merckelbach becomes suspicious of data reported in papers 1 and 2

He can’t find “Maastricht data” among Geraerts combined “Maastricht + Harvard” data set for paper 2 (JAb: Journal of Abnormal Psychology)
Figure 1. Summation of self-reported and probe-reported negative intrusions across the suppression and expression periods.

Too good to be true?
Curiouser and curiouser:

Self-rep + Probe-rep (Spontaneous) = idem (Others)
Self-rep (Spontaneous) = Probe-rep (Others)

Samples matched (on sex, age education), analysis does not reflect design

Figure 1. Summation of self-reported and probe-reported negative intrusions across the suppression and expression periods.
Merckelbach reports Geraerts to Maastricht and to Rotterdam authorities

Conclusion: (Maastricht) some carelessness but no fraud; (Rotterdam) no responsibility

Merckelbach and McNally request editors of “Memory” to retract their names from joint paper

The journalists love it (NRC; van Kolfschooten …)
### TABLE 1
Mean frequencies (SD) of target thoughts during suppression period

<table>
<thead>
<tr>
<th></th>
<th>Anxious event</th>
<th>Positive event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spontaneously recovered</td>
<td>1.27 (0.98)</td>
<td>3.17 (5.05)</td>
</tr>
<tr>
<td>Recovered in therapy</td>
<td>3.97 (3.14)</td>
<td>3.57 (2.75)</td>
</tr>
<tr>
<td>Continuous</td>
<td>3.10 (4.09)</td>
<td>3.77 (4.89)</td>
</tr>
<tr>
<td>Controls</td>
<td>3.50 (3.04)</td>
<td>4.13 (4.61)</td>
</tr>
</tbody>
</table>

Mean frequencies (and standard deviations) of target thoughts for anxious and positive autobiographical target events during the suppression period reported by the four groups (each n = 30).

### TABLE 2
Post-suppression rebound effect

<table>
<thead>
<tr>
<th></th>
<th>Anxious event</th>
<th>Positive event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spontaneously recovered</td>
<td>0.47 (2.32)</td>
<td>2.97 (5.07)</td>
</tr>
<tr>
<td>Recovered in therapy</td>
<td>4.37 (3.20)</td>
<td>2.76 (5.70)</td>
</tr>
<tr>
<td>Continuous</td>
<td>3.57 (2.97)</td>
<td>2.93 (6.74)</td>
</tr>
<tr>
<td>Controls</td>
<td>4.10 (5.64)</td>
<td>2.47 (5.00)</td>
</tr>
</tbody>
</table>

Mean change (and standard deviations) in frequencies of target thoughts from suppression to expression periods (i.e., post-suppression rebound effect).

### TABLE 3
Mean frequency (SD) of intrusions

<table>
<thead>
<tr>
<th></th>
<th>Anxious event</th>
<th>Positive event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spontaneously recovered</td>
<td>1.50 (1.94)</td>
<td>2.40 (1.07)</td>
</tr>
<tr>
<td>Recovered in therapy</td>
<td>5.57 (1.38)</td>
<td>2.60 (1.10)</td>
</tr>
<tr>
<td>Continuous</td>
<td>5.40 (1.67)</td>
<td>2.63 (1.13)</td>
</tr>
<tr>
<td>Controls</td>
<td>5.53 (1.83)</td>
<td>2.57 (1.04)</td>
</tr>
</tbody>
</table>

Mean frequency (and standard deviations) of intrusions over 7 days for anxious and positive autobiographical target events.
> results
[1,]  0.13599556  0.37733885
[2,]  0.01409201  0.25327297
[3,]  0.15298798  0.08453114
> sum(-log(results))
[1] 12.95321
> pgamma(sum(-log(results)), 6, lower.tail = FALSE)
[1] 0.01106587

■ Parametric analysis of Memory tables confirms, esp. on combining results from 3×2 analyses (Fisher comb.)

■ For the JAb paper I received the data from van Kolfschooten

■ Parametric analysis gives same result again (4×2)

■ Distribution-free (permutation) analysis confirms! (though: permutation p-value only 0.01 vs normality +independence 0.0002)
Scientific = Reproducible: Data preparation and data analysis are integral parts of experiment

Keeping proper log-books of all steps of data preparation, manipulation, selection/exclusion of cases, makes the experiment reproducible

Sharing statistical analyses over several authors is almost necessary in order to prevent errors

These cases couldn’t have occurred if all this had been standard practice
The morals of the story (2)

- Data collection protocol should be written down in advance in detail and followed carefully
- Exploratory analyses, pilot studies … also science
- Replicating others’ experiments: also science
- It's easy to make mistakes doing statistical analyses: the statistician needs a co-pilot
- Senior co-authors co-responsible for good scientific practices of young scientists in their group
- *These cases couldn’t have occurred if all this had been standard practice*
Memory affair postscript

- Geraerts is forbidden to talk to Gill
- Erasmus University Psychology Institute asks Han van der Maas (UvA) to investigate “too good to be true” pattern in “Memory” paper
- Nonparametric analysis confirms my findings
- Recommendations: 1) the paper is retracted ✓; 2) report is made public □; 3) data-set idem □

Obtaining the data “for peer review”: send request to secretariaatpsychologie@fsw.eur.nl
Main findings

- No proof of fraud (fraud = intentional deception)
- Definite evidence of errors in data management
- Un-documented, unreproducible reduction from 42 + 39 + 47 + 33 subjects to 30 + 30 + 30 + 30

Together, mega-opportunities for Questionable Research Practice number 7: deciding whether or not to exclude data after looking at the impact of doing so on the results

(Estimated prevalence near 100%, estimated acceptability rating near 100%)
Remarks

• A balanced design looks more scientific but is an open invitation to QRP 7

• Identical “too good to be true” pattern is apparent in an earlier published paper; the data has been lost

The latest developments

• I finally got the data from Geraerts (extraordinary confidentiality agreement)
• But you can read it off the pdf pictures in the report!
• So let’s take a look…
Suppression, Neg – Pos

Expression, Neg – Pos

Diary, Neg – Pos
The latest developments

- I cannot identify Maastricht subset in this data
- The JAb paper does not exhibit any of these anomalies!
- All data of the third paper with same “too good to be true” is lost
- A number of psychologists also saw “too good to be true” w.r.t. content (not statistics)
The next case?

Spot the difference between the following two slides
Exhibit A
Exhibit B
Notes

- Exhibit A is based on several studies in one paper
- Exhibit B is based on similar studies published in several papers of other authors
- More papers of author of Exhibit A exhibit same pattern
- His/her graphical displays are bar charts, ordered Treatment 1 (High) : Treatment 2 (Low) : Control
- Displays here based on the ordering Low : Medium (control) : High
Remarks

• IMHO, “scientific anomalies” should in the first place be discussed openly in the scientific community; not investigated by disciplinary bodies (CWI, LOWI, …)

• Data should be published and shared openly, as much as possible

• Never forget Hanlon’s razor

• I never met a science journalist who knew the “1 over root n” law

• I never met a psychologist who knew the “1 over root n” law

• Corollary: statisticians must participate in social debate, must work with the media
Exercises

- How/why could faked data exhibit this latest “too good to be true” pattern?
- Develop model-free statistical test for “excess linearity” [no replicate groups, so no permutation test]
- How could Geraerts’ data ever get the way it is? In three different papers, different anomalies, same overall “too good to be true” pattern?
- What do you consider could be the moral of this story?