Informative hypotheses evaluation

Replication crisis

Rebecca M. Kuiper
(credits slides: Herbert Hoijtink and others)

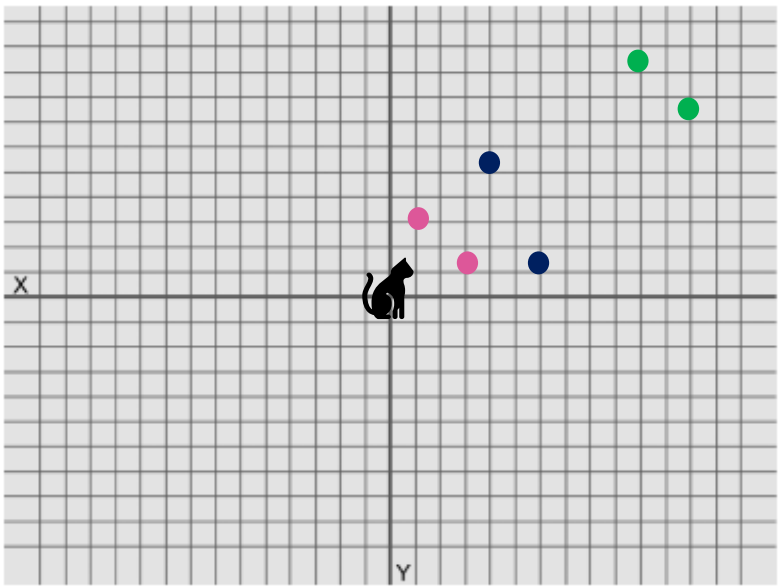
Department of Methodology & Statistics
Utrecht University

Table of Contents

A Research Project and Its Replication

An experiment with three conditions:

- The “close” condition
- The “intermediate” condition
- The “distant” condition



Participants Rated:

Attachment to:

- Siblings
- Parents
- Home-town

on a

1 (not at all strong) – 7 (extremely strong)
Likert scale

which are
averaged to obtain the dependent variable

The description given here is a modification of and inspired by the actual experiment executed by Williams, L.E. and Bargh, J.A. (2008). Keeping One’s Distance. The Influence of Spatial Distance Cues on Affect and Evaluation. *Psychological Science, 19,* 302-308.

The Main Research Outcomes

Williams and Bargh (2008) tested:

H_0: $\mu_{close} = \mu_{intermediate} = \mu_{distant}$,

that is, the three means are equal

rendering

p-value = .01, that is, smaller than .05, that is, the means are significantly different

with

$m_{close} = 5.61$, $m_{intermediate} = 5.23$, $m_{distant} = 4.86$

and

$\eta^2 = .11$, that is, the three conditions explain 11% of the variation in attachment, which is a medium to strong effect of condition

The replication by Joy-Gaba, Clay, and Cleary (2016) rendered

p-value = .79

with

$m_{close} = 5.44$, $m_{intermediate} = 5.31$, $m_{distant} = 5.31$

And

$\eta^2 = .00$

Joy-Gaba, J., Clay, R., and Cleary, H. (2016). Replication of keeping one's distance: The influence of spatial distance cues on affect and evaluation by Williams L.E. and Bargh J.A. (2008) *Psychological Science, 19,* 302-308). Retrieved from https://osf.io/a78bm/

The Replication Crisis

The Replication Crisis

This is only one of 100 psychological experiments of which only about 33% were successfully replicated (OSC, 2015).

This resulted in a reduced trust in science by scientists and society: The replication crisis was born.

Scientists are alerted:

- Estimating the reproducibility of psychological science (OSC, 2015)
- An open investigation of the reproducibility of cancer biology research (Errington et al., 2014)

"Society" is alerted:

- Is psychology a real science? (Is psychologie wel een echte wetenschap, Volkskrant, 12-8-2016)
- Public Trust in Science (Rathenau Instituut, August 28, 2018)

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, 6251. https://osf.io/ezcuj/
Errington, T.M., Iorns, E., Gunn, W., Tan, F.E., Lomax, J., and Nosek, B.A. (2014). An open investigation of the reproducibilty of cancer biology research. *eLIFE, 3*, e04333. https://elifesciences.org/collections/9b1e83d1/reproducibility-project-cancer-biology
Volkskrant (2016). https://www.volkskrant.nl/columns-opinie/is-psychologie-wel-een-echte-wetenschap~b9978e6c
Rathenau Instituut (2018). Public Trust in Science. https://www.rathenau.nl/en/science-figures/impact/trust-science/public-trust-science

The p-value and The .05

p-value

The p-value is the probability of the observed data (or data that deviate more from H_0) assuming that H_0 is true.

.05

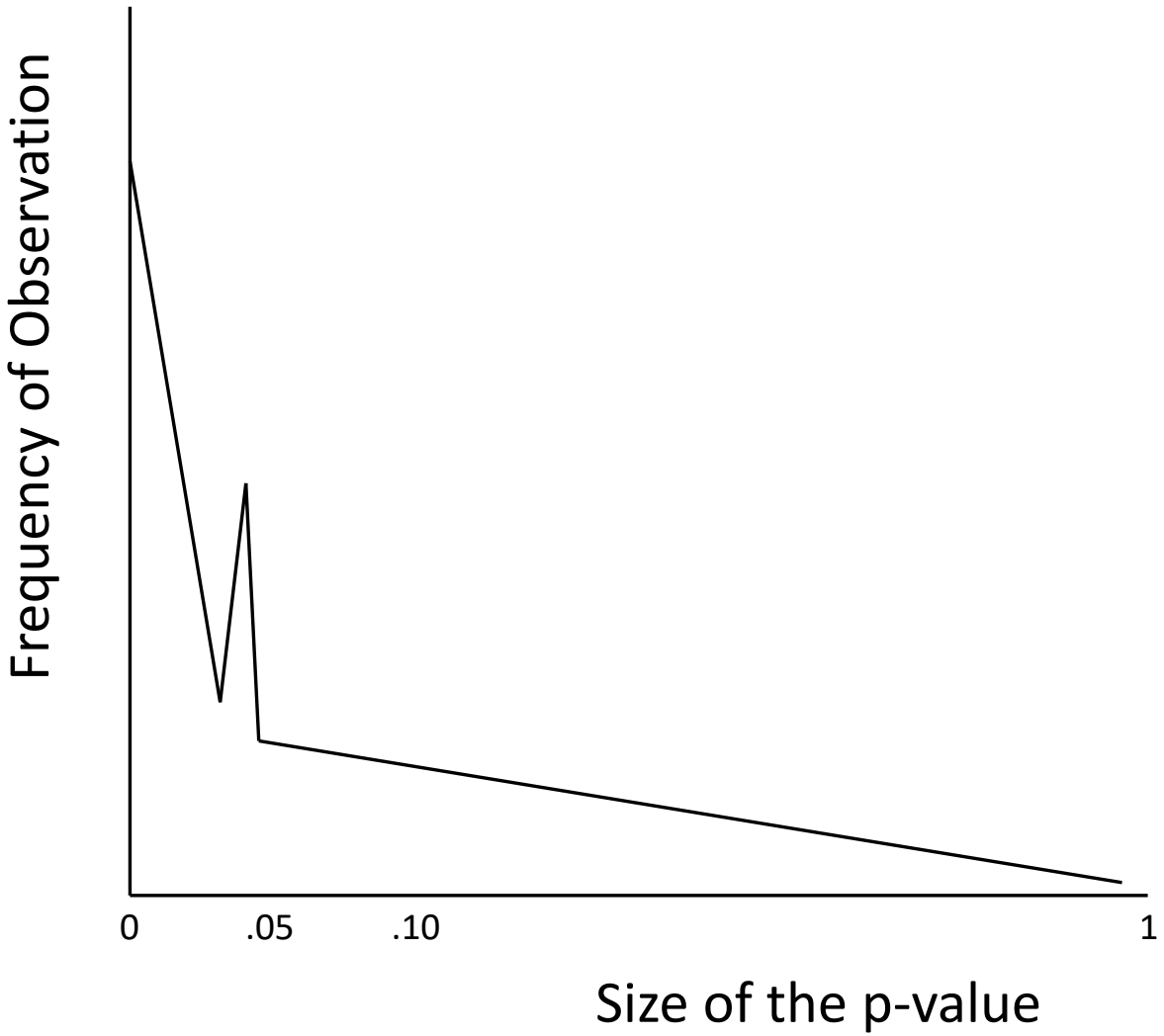
If the p-value is smaller than .05, it is considered to be so small that H_0 has to be rejected.

Causes of the Replication Crisis

Masicampo en Lalande (2012) collected the p-values published in the journals: Psychological Science, Journal of Personality and Social Psychology, and Journal of Experimental Psychology: General.

1. Masicampo, E.J. and Lalande, D.R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology, 65,* 2271-2279.

Causes of the Replication Crisis



Questionable Research Practices

1. After testing the null-hypothesis, the resulting p-value is .06. But after removing three persons with unexpected low scores on the dependent variable, the p-value becomes .04
2. After testing the effect of treatment/control on three operationalisations of depression, resulting in p-values of .04, .12, and .34, only "the significant" p-value is reported
3. Any other examples ...

Incentives for Questionable Research Practices

Found somewhere on the internet:

TS college 8 beamer.pdf

. . . and has real-life consequences

p value scale				
*** .001	very highly significant	there is an effect definitely for sure	elation exuberance smugness	nobel price tenur research grant
** .01	highly significant	there is an effect	great pleasure dancing drinking	phd price top publication
* .05	significant (phew)	most likely there is an effect	relief cheerfulness	consolation price fair publication
? .10	approaching significance	almost probably an effect but low power	frustration if only	counseling stress leave
	nonsignificant	no effect	despair depression	medication reconsider life goals

after Geoff Cumming, 2012

Prevalence of Questionable Research Practices

1. About 2% of scientists admits to having fabricated or falsified research data, or to have altered or modified results to improve the outcome
2. About 33% of scientists admits to having used questionable research practices
3. How about "me" and "you" ...

1. Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS ONE, 4,* e5738.
2. Ioannides, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124.

Publication Bias

1. In 1981, a psychologist investigated "feeling the future" ... The p-value for "H0: the choice is random" was .67. Paper was not published in a journal.
2. In 1991 ...
3. In 2001 ...
4. In 2011 Bem ... the resulting p-value was .015. Paper was published.
5. In 2012 Richie, Wiseman, and French replicated 3x with p-values of .15, .40, and .38. Paper was rejected by the original journal and accepted by another journal.

1. Bem, D.J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407-425. doi: 10.1037/a0021524
2. Ritchie, S.J., Wiseman, R., and French, C.C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *Plos One, 7*. doi: 10.1371/journal.pone.0033423

How can the Replication Crisis be Addressed?

Open Science

1. Pre-registration and pre-registered reports
2. Multiple lab and multiple cohort studies
3. Replication studies executed by the authors or independent others
4. Publish data and analyses
5. Open access publications

As will be elaborated, (Bayesian) evaluation of informative hypotheses can contribute to Open Science.

Table of Contents

The Traditional Null Hypothesis

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Cohen (1994) "The Earth is Round p <.05"

Royal (1997) "A power analysis should render $N = 0$"

Only use the null-hypothesis if it is a plausible representation of population of interest

1. Cohen, J. (1994). The earth is round, p<.05. *American Psychologist, 49,* 997-1003.
2. Royal, R. (1997). *Statistical Evidence. A Likelihood Paradigm.* New York: Chapman and Hall/CRC.

P-values and Alpha Level

p-value

The p-value is the probability of the observed data (or data that deviate more from H_0) assuming that H_0 is true.

The p-value is *not* a measure of support for the null-hypothesis, it is a measure of evidence *against* the null-hypothesis. It can therefore not be used to quantify the support in the data *for* the null-hypothesis.

P-values and Alpha Level

alpha level / Type I error

The probability of incorrectly rejecting the null-hypothesis. The "usual" value is .05.

Were does the .05 come from? Fisher used "no level", .05, .02, .10, .01, and was in no way married to the .05.

Consequences of the .05 (or any other number): sloppy science, publication bias, ...

P-values and Alpha Level

Example

3 factors with corresponding group numbers:

	Condition		
	Masculine	Feminine	Neutral
Masculine Men	1	2	3
Feminine Men	4	5	6
Masculine Women	7	8	9
Feminine Women	10	11	12

Van Well, S., Kolk, A.M., Klugkist, I. (2008). Effects of Sex, Gender Role Identification, and Gender relevance of Two Types of Stressors on Cardiovascular and Subjective Responses: Sex and Gender Match/Mismatch Effects. Behavior Modification, 32, 427 - 449.

P-values and Alpha Level

Example: $3 \times 2 \times 2$ ANOVA

Tests of Between-Subjects Effects

Dependent Variable: cs_sbp

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	17754.778[a]	12	1479.565	11.207	.000
Intercept	2145861.954	1	2145861.954	16253.783	.000
Baseline SBP	13049.137	1	13049.137	98.840	.000
Sekse	1339.880	1	1339.880	10.149	.002
GRI	76.680	1	76.680	.581	.448
Manipulation	180.911	2	90.456	.685	.507
Sekse*Manipulation	290.301	1	290.301	2.199	.142
Sekse*GRI	40.979	2	20.489	.155	.857
GRI*Manipulation	929.848	2	464.924	3.522	.034
Sekse*GRI*Manipulation	179.114	2	89.557	.678	.510
Error	10693.807	81	132.022		
Total	2280649.278	94			
Corrected Total	28448.586	93			

a. R Squared = .624 (Adjusted R Squared = .568)

P-values and Alpha Level

After observing ".06" (or .14 like on the previous slide) one *can not* update, that is, collect extra data and recompute the p-value. This procedure is called sequential data analysis. It has to be planned *before* the data is collected because it involves multiple evaluations of the hypotheses of interest and therefore the alpha level has to be corrected.

On top of that, how to deal with the fact that "the .05" is applied multiple times on the previous slide? How to correct for multiple hypotheses testing?

P-values and Alpha Level

Also, using an alpha level of .20 :-), we find four significant results.
It is clear that 'Something is going on, but we don't know what!'
And here we go eye-balling the data and effect sizes to interpret the results.

P-values and Alpha Level

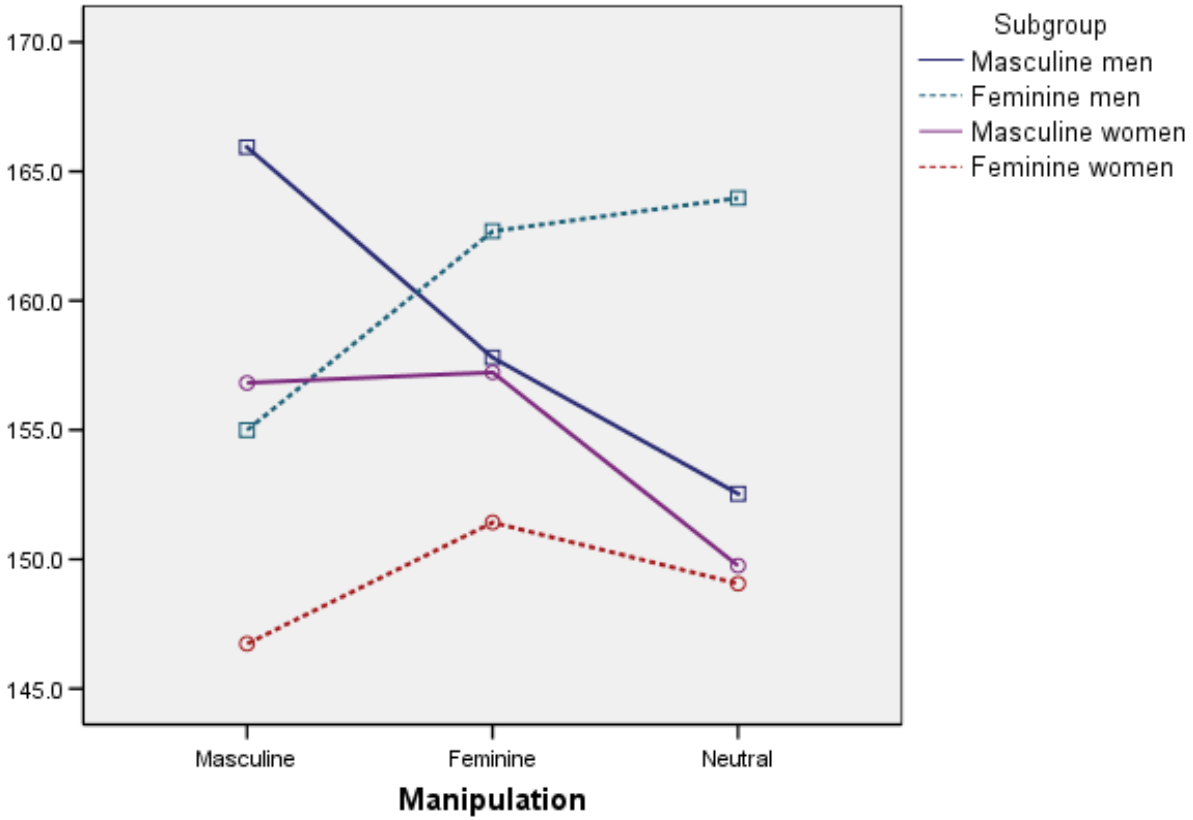


Table of Contents

Informative Hypotheses

Example 1: ANOVA

	Condition		
	Masculine	Feminine	Neutral
Masculine Men	1	2	3
Feminine Men	4	5	6
Masculine Women	7	8	9
Feminine Women	10	11	12

Sex Match Effect

$H_1 : (\mu_1, \mu_4) > (\mu_2, \mu_3, \mu_5, \mu_6)$ and $(\mu_8, \mu_{11}) > (\mu_7, \mu_9, \mu_{10}, \mu_{12})$

Informative Hypotheses

ANOVA

	Condition		
	Masculine	Feminine	Neutral
Masculine Men	1	2	3
Feminine Men	4	5	6
Masculine Women	7	8	9
Feminine Women	10	11	12

Gender Role Match Effect

$H_2 : (\mu_1, \mu_7) > (\mu_2, \mu_3, \mu_8, \mu_9)$ and $(\mu_5, \mu_{11}) > (\mu_4, \mu_6, \mu_{10}, \mu_{12})$

Informative Hypotheses

ANOVA

	Condition		
	Masculine	Feminine	Neutral
Masculine Men	1	2	3
Feminine Men	4	5	6
Masculine Women	7	8	9
Feminine Women	10	11	12

Sex Mismatch Effect

$H_3 : (\mu_2, \mu_5) > (\mu_1, \mu_3, \mu_4, \mu_6)$ and $(\mu_7, \mu_{10}) > (\mu_8, \mu_9, \mu_{11}, \mu_{12})$

Informative Hypotheses

ANOVA

	Condition		
	Masculine	Feminine	Neutral
Masculine Men	1	2	3
Feminine Men	4	5	6
Masculine Women	7	8	9
Feminine Women	10	11	12

Gender Role Mismatch Effect

$H_4 : (\mu_4, \mu_{10}) > (\mu_5, \mu_6, \mu_{11}, \mu_{12})$ and $(\mu_2, \mu_8) > (\mu_1, \mu_3, \mu_7, \mu_9)$

Table of Contents

What's next

- NHST versus evaluating informative hypotheses
- Model selection using information criteria
- Bayesian model selection
- Evidence synthesis / Support aggregation