

# bain: Bayesian informative hypotheses evaluation

H. Hoijtink<sup>1</sup>   C. van Lissa<sup>1</sup>   J. Mulder<sup>2</sup>   X. Gu<sup>3</sup>

<sup>1</sup>Department of Methodology and Statistics  
Utrecht University [H.Hoijtink@uu.nl](mailto:H.Hoijtink@uu.nl)

<sup>1</sup>Department of Methodology and Statistics  
Utrecht University [C.J.vanLissa@uu.nl](mailto:C.J.vanLissa@uu.nl)

<sup>2</sup>Department of Methodology and Statistics  
Tilburg University [J.Mulder3@uvt.nl](mailto:J.Mulder3@uvt.nl)

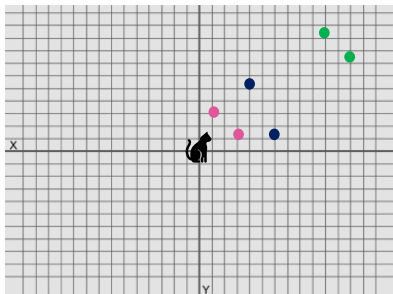
<sup>3</sup>Department of Educational Psychology  
East China Normal University [GuXin57@hotmail.com](mailto:GuXin57@hotmail.com)

# The Replication Crisis

# 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:

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

that is, the three means are equal

rendering

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

with

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

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

$$p\text{-value} = .79$$

with

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

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 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.

# 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 reproducibility 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>

## Discuss. First in Small Groups, then Plenary

What do you think are the causes of the replication crisis?

# Causes of the Replication Crisis

Masicampo and 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:

... and has real-life consequences

<b>p value scale</b>	*** .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 (pew)	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

# 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.

# Null Hypothesis Significance Testing

# 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.

Where 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: sloppy science, publication bias, ...

## P-values and Alpha Level

	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.

### Tests of Between-Subjects Effects

Dependent Variable: cs\_sbp

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	17754.778 <sup>a</sup>	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

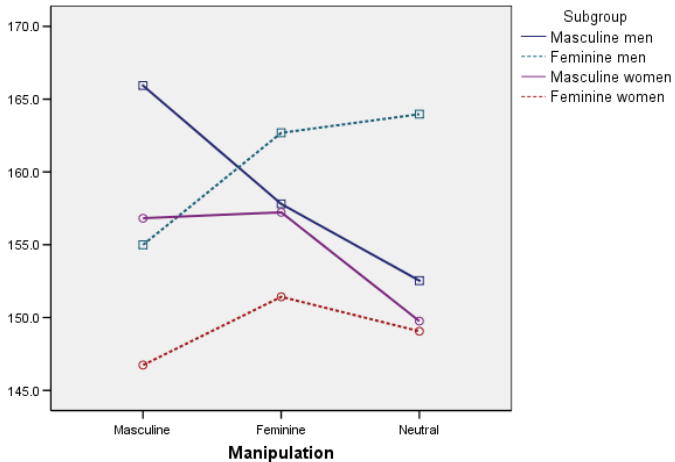
After observing ".06" (or .14 like on the previous slide) one *can not* update, that is, collect extra data and recompute the p-value. This procedure is called sequential data analysis. It has to be planned *before* the data is collected because it involves multiple evaluations of the hypotheses of interest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

## DEVELOPMENTAL PSYCHOLOGY



# bain: Bayesian Informative Hypotheses Evaluation

1. Hoijtink, H., Mulder, J., van Lissa, C., and Gu, X. (2018). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, 24, 539-556.



# 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

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

# 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

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

# 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

### Sex Mismatch Effect

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

# 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

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

# Bayes Factor

## Balancing Fit and Complexity

The Bayes factor quantifies the relative support in the data for two hypotheses, for example,

$$H_i : \mu_1 > \mu_2 > \mu_3$$

$$H_u : \mu_1, \mu_2, \mu_3$$

with

$$BF_{iu} = \frac{f_i}{c_i} = \frac{\text{fit } H_i}{\text{complexity } H_i}$$

that is, after observing the data  $H_i$  is  $BF_{iu}$  times as likely as  $H_u$ , for example, .2, 5, 10.

# Bayes Factor

## Balancing Fit and Complexity

A (very) loose interpretation of the meaning of fit

$$H_i : \mu_1 > \mu_2 > \mu_3$$

if  $\bar{x}_1 = 7$  &  $\bar{x}_2 = 4$  &  $\bar{x}_3 = 2$  the fit is good

if  $\bar{x}_1 = 2$  &  $\bar{x}_2 = 4$  &  $\bar{x}_3 = 7$  the fit is bad

# Bayes Factor

Balancing Fit and Complexity

A (very) loose interpretation of the meaning of complexity

$$H_1 : \mu_1 = \mu_2 = \mu_3$$

very parsimonious, the means have to be exactly equal

$$H_1 : \mu_1 > \mu_2 > \mu_3$$

one ordering of three means: 1-2-3, thus is parsimonious

$$H_2 : \mu_1 > (\mu_2, \mu_3)$$

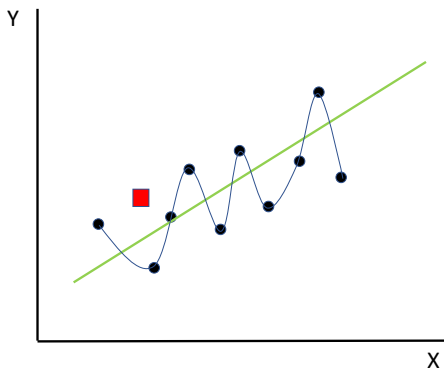
2 orderings of three means: 1-2-3 and 1-3-2, less parsimonious

$$H_u : \mu_1, \mu_2, \mu_3$$

contains all six possible orderings of three means, not

# Bayes Factor

## Balancing Fit and Complexity



The straight line results from a linear regression model with 3 parameters (intercept, slope, residual variance).

The other line results from a polynomial regression models with 11 parameters (intercept, nine slopes, residual variance).

The red square is a new observation that is added to the original 10 observations.

What is the predictive value of both models?



# Bayes Factor

## Balancing Fit and Complexity

Three forms of Hypotheses and Bayes factors involving

$$H_i : \mu_1 > \mu_2 > \mu_3$$

$BF_{iu}$  evaluating  $H_i$  versus  $H_u : \mu_1, \mu_2, \mu_3$

$BF_{ij'}$  evaluating  $H_i$  versus  $H_{j'} : \mu_1 = \mu_2 = \mu_3$

$BF_{ic}$  evaluating  $H_i$  versus  $H_c : \text{not } H_i$

# Bayes Factor

## Interpreting (the Size of) the Bayes Factor

1. Select the best of a set of hypotheses using  $BF_{iu}$
2. Compare two competing hypotheses using  $BF_{ij'}$
3. Compare "my theory" with "not my theory" using  $BF_{ic}$

	$f_i$	$c_i$	$BF_{iu}$	$BF_{ic}$
$H_1$ : Sex Match	.0039	.012	.32	.32
$H_2$ : Gender Role Match	.0725	.012	5.85	6.44
$H_3$ : Sex Mismatch	.0007	.012	.06	.06
$H_4$ : Gender Role Mismatch	.0001	.012	.01	.01

# Bayes Factor

## Effect Sizes and Descriptives

### Gender Role Match Effect

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

$$H_2 : (166, 163) > (158, 154, 155, 164) \text{ and}$$

$$(157, 152) > (157, 150, 143, 149)$$

### Gender Role Mismatch Effect

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

$$H_4 : (158, 155) > (166, 154, 163, 164) \text{ and}$$

$$(157, 143) > (157, 150, 152, 149)$$

# Bayes Factor

## Interpreting (the Size of) the Bayes Factor

1. The Bayes factor **is** a measure of support (also for the null-hypothesis)
2. The Bayes factor **can be indecisive**. A value around 1 denotes "the data don't tell us which hypothesis to prefer"
3. One **can update**, that is, collect more data and recompute the Bayes factor (see extra comments later on)
4. One **can compare** more than two hypotheses (see extra comments later on)
5. "Something is going on and **we do know what!**"
6. The Bayes factor **selects the best of the hypotheses under consideration**. Note that the "true" hypothesis may not be among them, and that all hypotheses may be "wrong"

# Bayes Factor

Interpreting (the Size of) the Bayes Factor

## When is the Bayes factor large enough?

1. Guidelines by Jeffreys (1969) and Kass and Raftery (1995), e.g.,  $< 3$  is ignorable,  $> 3$  is positive evidence,  $> 10$  is strong evidence ...
2. Will lead to a return of sloppy science and publication bias (when used without pre-registration or a pre-registered report)
3. Where does the 3 come from?

# Bayes Factor

Interpreting (the Size of) the Bayes Factor

## When is the Bayes factor large enough?

1. Before collecting or accessing the data, formulate informative hypotheses and decide how large you would like the Bayes factor to be.
2. Insert this information in a pre-registration or pre-registered report.
3. Collect data and evaluate hypotheses. Is one good and the best with a "large" Bayes factor: nice! Are the Bayes factors "not large enough": follow up research or updating is needed. Is none good: BIG news, well-constructed hypotheses have been rejected!

# Bayesian Error Probabilities

Posterior Model Probabilities, e.g.,  $PMP(H_i | \text{data})$  and  $PMP(H_c | \text{data})$  quantify the support in the data for each hypothesis.

$$\frac{PMP(H_i|\text{data})}{PMP(H_c|\text{data})} = \text{BF}_{ic} \times \frac{PRI(H_i)}{PRI(H_c)}, \quad (1)$$

where  $PRI(H_i)$  and  $PRI(H_c)$  denote the *prior* probabilities, that is, an evaluation of the support for the hypotheses *before* observing the data.

Usually equal prior model probabilities are used (which means that the PMP's convey the same information as the Bayes factors), but this is not a requirement. Consider, for example, "the Bem story" ...

# PMPs

PMPs can be interpreted as Bayesian error probabilities, that is, the Bayesian counterparts of the Type I and Type II errors.

	$f_i$	$c_i$	$BF_{iu}$	$PMP_i$	$PR_i$
$H_1$ : Sex Match	.0039	.012	.32	.04	1/5
$H_2$ : Gender Role Match	.0725	.012	5.85	.81	1/5
$H_3$ : Sex Mismatch	.0007	.012	.06	.01	1/5
$H_4$ : Gender Role Mismatch	.0001	.012	.01	.00	1/5
$H_u$ :				.14	1/5



# PMPs

Replacing  $H_u$  by  $H_c$

Soon to be implemented in `bain`

	$f_i$	$c_i$	$BF_{iu}$	$PMP_i$	$PRI_i$
$H_1$ : Sex Match	.0039	.012	.32	.04	1/5
$H_2$ : Gender Role Match	.0725	.012	5.85	.84	1/5
$H_3$ : Sex Mismatch	.0007	.012	.06	.00	1/5
$H_4$ : Gender Role Mismatch	.0001	.012	.01	.00	1/5
$H_c$ :	.9200	.9900	.93	.12	1/5

Where  $H_c$  denotes the complement  $H_1$  through  $H_4$ , that is, "not one of these four hypotheses". This will usually render smaller error probabilities.

# PMPs

## The Number of Hypotheses and PMPs

Look what happens if we compare many hypotheses, the PMPs become smaller and smaller, and thus the Bayesian error probabilities become larger and larger:

	$f_i$	$c_i$	$BF_{iu}$	$PMP_i$	$PRI_i$
$H_1$ : Sex Match	.0039	.012	.32	.013	1/13
$H_2$ : Gender Role Match	.0725	.012	5.85	.270	1/13
$H_3$ : Sex Mismatch	.0007	.012	.06	.003	1/13
$H_4$ : Gender Role Mismatch	.0001	.012	.01	.000	1/13
$H_5$ : Lets try this one too	.0521	.012	2.61	.180	1/13
...					
$H_{12}$ : Don't miss something	.0164	.012	1.36	.040	1/13
$H_u$ :				.047	1/13

# PMPs

## The Number of Hypotheses and PMPs

The same results as two slides up are in fact obtained by assigning PMPs of 0 to each hypothesis that is NOT considered:

	$f_i$	$c_i$	$BF_{iu}$	$PMP_i$	$PRI_i$
$H_1$ : Sex Match	.0039	.012	.32	.04	1/5
$H_2$ : Gender Role Match	.0725	.012	5.85	.81	1/5
$H_3$ : Sex Mismatch	.0007	.012	.06	.01	1/5
$H_4$ : Gender Role Mismatch	.0001	.012	.01	.00	1/5
$H_5$ : Lets try this one too	.0521	.012	2.61	.18	0
...					
$H_{12}$ : Don't miss something	.0164	.012	1.36	.04	0
$H_u$ :				.14	1/5

# Subjectivity of Bayesian Hypotheses Evaluation

1. Which hypotheses to evaluate?
2. How to formalize hypotheses? E.g.  $(\mu_1, \mu_2) > (\mu_3, \mu_4)$  or  $\mu_1 = \mu_2 > \mu_3 = \mu_4$
3. The (implicit) choice for equal prior model probabilities
4. The specification of the prior distribution (will be discussed lateron)

# Bayesian Updating

1. Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21, 301-308.

# Bayesian Updating

Repeated significance testing after increasing the sample requires "planning before the data collection has started" and "correction for multiple testing".

"Bayesian updating" is simply recomputing the evidence presented by all the data that are currently available. This can both be done using the Bayes factor "what is the relative support in the available data for this pair of hypotheses" and/or the PMPs "what is the support in the available data for this hypothesis".

# Bayesian Updating

Bayes factors	<i>N</i> per group			
	8	+8	+12	+5
$H_1$ : Sex Match	.32	.24	.12	.02
$H_2$ : Gender Role Match	5.85	7.12	9.23	11.82
$H_3$ : Sex Mismatch	.06	.02	.00	.00
$H_4$ : Gender Role Mismatch	.01	.00	.00	.00

PMPs	<i>N</i> per group			
	8	+8	+12	+5
$H_1$ : Sex Match	.04	.03	.01	.00
$H_2$ : Gender Role Match	.84	.86	.91	.93
$H_3$ : Sex Mismatch	.00	.00	.00	.00
$H_4$ : Gender Role Mismatch	.00	.00	.00	.00
$H_c$	.12	.11	.08	.07

# ANOVA and Beyond, Example Analyses with R and JASP



# Informative Hypotheses

## Example 1: ANOVA

What is the relation between "knowledge of numbers after watching Sesame Street for a year"

and

site from which the child originates (1 = disadvantaged inner city, 2 = advantaged suburban , 3 = advantaged rural, 4 = disadvantaged rural, 5 = disadvantaged Spanish speaking).

# Informative Hypotheses

## Example 1: ANOVA

```
library(bain)
sesamesim$site <- as.factor(sesamesim$site)
anov <- lm(postnumb~site-1,sesamesim)
coef(anov)
set.seed(100)
results <- bain(anov,
                 "site1=site2=site3=site4=site5;
                 site2>site5>site1>site3>site4")
print(results)
summary(results, ci = 0.95)
```

# Informative Hypotheses

## Example 1: ANOVA

`coef(anov)` renders

	site1	site2	site3	site4	site5
	29.66667	38.98182	23.18750	25.32558	31.72222

`summary(results)` renders

	Parameter	n	Estimate	lb	ub
1	site1	60	29.66667	26.82991	32.50343
2	site2	55	38.98182	36.01892	41.94472
3	site3	64	23.18750	20.44082	25.93418
4	site4	43	25.32558	21.97466	28.67650
5	site5	18	31.72222	26.54303	36.90141

# Informative Hypotheses

## Example 1: ANOVA

The main output is

	Fit	Com	BF.u	BF.c	PMPa	PMPb	PMPc
H1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
H2	0.121	0.008	14.559	16.428	1.000	0.936	0.943
Hu						0.064	
Hc	0.879	0.992	0.886				0.057

Hypotheses:

H1: site1=site2=site3=site4=site5

H2: site2>site5>site1>site3>site4

# Informative Hypotheses

## Example 1: ANOVA

The screenshot shows the JASP software interface for a Bain ANOVA analysis. The left sidebar contains a list of variables: sex, setting, viewenc, age, peabody, prenumb, funumb, Bb, Bl, Bf, Bn, Br, Bc, Ab, Al, Af, and An. The main panel is titled "Bain ANOVA" and shows the following settings:

- Dependent Variable: postnumb
- Fixed Factors: site
- Tables: ☒ Descriptives, ☐ Bayes factor matrix, ☐ Posterior probabilities, ☐ Descriptives plot
- Additional Options: Seed 100
- Model Constraints: (empty)

Below the settings, there is a text box for hypotheses:

Place each hypothesis on a new line. For example:

```
factorLow = factorMed = factorHigh
factorLow < factorMed < factorHigh
```

where factor is the factor name and Low/Med/High are the factor level names. Read the help file for further instructions.

The bottom of the interface shows the following hypotheses:

```
site1 = site2 = site3 = site4 = site5
site2 > site5 > site1 > site3 > site4
```

The right sidebar displays the "Results" section, titled "Bain ANOVA". It includes a "Hypothesis Legend" table:

Hypothesis	
H1	site1 = site2 = site3 = site4 = site5
H2	site2 > site5 > site1 > site3 > site4

Below the legend is a "Bain ANOVA" table:

	BF <sub>c</sub>	PMP <sub>a</sub>	PMP <sub>b</sub>
H1	1.151e-11	7.338e-13	6.899e-13
H2	17.784	1.000	0.940
Hu			0.060

Note: BF<sub>c</sub> denotes the Bayes factor of the hypothesis in the row versus its complement. Posterior model probabilities (a: excluding the unconstrained hypothesis, b: including the unconstrained hypothesis) are based on equal prior model probabilities.

Below the table is a "Descriptive Statistics" table:

	N	Mean	SD	SE	95% Credible Interval	
					Lower	Upper
site1	60	29.667	11.427	1.447	26.830	32.503
site2	55	38.982	12.991	1.512	36.019	41.945
site3	64	23.188	11.361	1.401	20.441	25.934
site4	43	25.326	8.941	1.710	21.975	28.677
site5	18	31.722	8.512	2.642	26.543	36.901

# Informative Hypotheses

## Example 2: ANOVA Interaction Effect

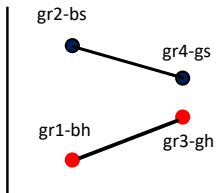
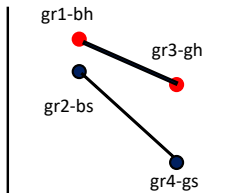
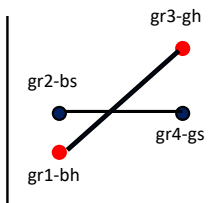
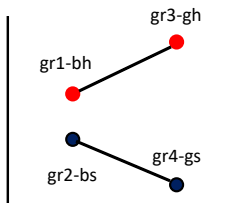
Dependent variable: Knowledge of numbers.

Factors: sex (boy, girl) and setting (watching at home, watching at school).

Gr: 1=boyhome, 2= boyschool, 3= girlhome, 4=girlschoo.

# Informative Hypotheses

## Example 2: ANOVA Interaction Effect



*H<sub>i</sub> :*

$$gr2 - gr1 > gr4 - gr3$$

and

$$gr2 > gr1$$

$$gr2 > gr4$$

# Informative Hypotheses

## Example 2: ANOVA Interaction Effect

```
sesamesim$gr <- as.factor(sesamesim$gr)
anov <- lm(postnumb~gr-1,sesamesim)
results <- bain(anov,
"gr2 - gr1 > gr4 - gr3 & gr2 > gr1 & gr2 > gr4")
```



# Informative Hypotheses

## Example 2: ANOVA Interaction Effect

The main output is

	Fit	Com	BF.u	BF.c	PMPa	PMPb	PMPc
H1	0.922	0.283	3.262	29.984	1.000	0.765	0.968
Hu						0.235	
Hc	0.078	0.717	0.109				0.032

Hypotheses:

H1:  $\text{gr2} - \text{gr1} > \text{gr4} - \text{gr3} \& \text{gr2} > \text{gr1} \& \text{gr2} > \text{gr4}$

# Informative Hypotheses

How to write down an hypothesis

Linear combinations can handle hypotheses build using constraints on (linear combinations) of parameters. Suppose the parameter names are "a", "b", "c".

Step 1: Construct the elements of the linear combination. E.g. "a" or "a + 2" or "3 \* a" or "2 \* a + 4"

Step 2: Constrain the resulting elements. E.g.  $a > b > c$

or  $a > b + 2 \ \& \ b > c + 2$

or  $2 * a > b + c \ \& \ b > 0 \ \& \ c > 0$

or  $a > (b, c) \ \& \ b - c > 0$

# Informative Hypotheses

## Example 3: Repeated Measures

Development of depression				
	Measurement			
	8 years	12 years	16 years	20 years
Men	$\mu_1$	$\mu_2$	$\mu_3$	$\mu_4$
Women	$\mu_5$	$\mu_6$	$\mu_7$	$\mu_8$

$$H_1 : \mu_5 - \mu_1 > \mu_6 - \mu_2 > \mu_7 - \mu_3 < \mu_8 - \mu_4$$

$$H_2 : \mu_6 - \mu_5 < \mu_7 - \mu_6 > \mu_8 - \mu_7$$

# Informative Hypotheses

## Example 4: Multiple Regression

$$\text{postnumb}_i = \beta_0 + \beta_1 \times \text{age}_i + \beta_2 \times \text{prenumb}_i + \epsilon_i$$

$$H_1 : \beta_1 > 0, \beta_2 > 0, \beta_1 < \beta_2$$

Note:  $\beta_1$  and  $\beta_2$  are only comparable if age and prenumb are standardized

# Informative Hypotheses

## Example 4: Multiple Regression

**Regression**

Dependent Variable: postnumb

Covariates: age, prenumb

**Results**

**Regression**

Hypothesis Legend

	Hypothesis
H1	age = 0 & prenumb = 0
H2	age > 0 & prenumb > 0 & prenumb > age

Bain Linear Regression

	BF.c	FMP a	FMP b
H1	7.909e-83	6.426e-84	5.943e-84
H2	108.494	1.000	0.925
Hu			0.075

Note: BF.c denotes the Bayes factor of the hypothesis in the row versus its complement. Posterior model probabilities (a: excluding the unconstrained hypothesis, b: including the unconstrained hypothesis) are based on equal prior model probabilities.

Place each hypothesis on a new line. For example:

```
age = length = weight
age < length < weight,
```

where age, length and weight are the names of the predictors.  
Read the help file for further instructions.

```
age = 0 & prenumb = 0
age > 0 & prenumb > 0 & prenumb > age
```

# Informative Hypotheses

## Example 5: About Equality Constraints

Is the difference in number knowledge relevantly different between boys and girls?

# Informative Hypotheses

## Example 5: About Equality Constraints

```
sesamesim$sex <- as.factor(sesamesim$sex)
anov <- lm(postnumb~sex-1,sesamesim)
results <- bain(anov, "-2 < sex1 - sex2 < 2")
```

# Informative Hypotheses

## Example 5: About Equality Constraints

```
sex1      sex2
30.09565  28.85600
```

	Fit	Com	BF.u	BF.c	PMPa	PMPb	PMPc
H1	0.664	0.091	7.304	19.735	1.000	0.880	0.952
Hu						0.120	
Hc	0.336	0.909	0.370				0.048

Hypotheses:

H1:  $-2 < \text{sex1} - \text{sex2} < 2$



# Informative Hypotheses

## Example 6: Structural Equation Modelling

```
library(bain)
library(lavaan)

model <- '
  A  =~ Ab + Al + Af + An + Ar + Ac
  B  =~ Bb + Bl + Bf + Bn + Br + Bc
  A  ~ B + age + peabody'
fit <- sem(model, data = sesamesim, std.lv = TRUE)

hypotheses <- "A~B = A~peabody = A~age = 0;
               A~B > A~peabody > A~age = 0"
set.seed(100)
y1 <- bain(fit, hypotheses, standardize = TRUE)
```

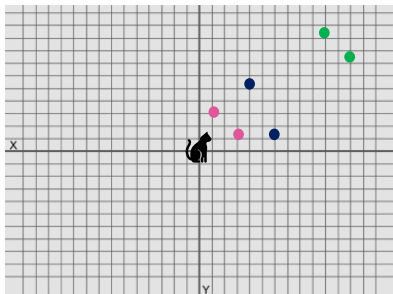
# Bayesian Evaluation of Replication Studies

1. Leplaa, H., Rietbergen, C., and Hoijtink, H. (unpublished). Bayesian Evaluation of Replication Studies.

# 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:

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

that is, the three means are equal

rendering

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

with

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

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

$$p\text{-value} = .79$$

with

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

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/>

# Replication Research

## Replication Hypotheses Derived from the Original Study

$$\mu_{close} > \mu_{intermediate} > \mu_{distant}$$

$$\mu_{close} > \mu_{intermediate} + .2sd \text{ \& } \mu_{intermediate} > \mu_{distant} + .2sd$$

# Replication Research

## Replication Hypotheses Derived from the Original Study

	Fit	Com	BF.u	BF.c	PMPa	PMPb	PMPc
H1	0.298	0.168	1.780	2.112	1.000	0.640	0.680
Hu						0.360	
Hc	0.702	0.832	0.843				0.320

Hypotheses:

H1: Close>Interm>Dist

# Replication Research

## Replication Hypotheses Derived from the Original Study

	Fit	Com	BF.u	BF.c	PMPa	PMPb	PMPc
H1	0.025	0.168	0.147	0.125	1.000	0.128	0.111
Hu						0.872	
Hc	0.975	0.832	1.171				0.899

Hypotheses:

H1:Close>Interm+.20&Interm>Dist+.20

# bain Evidence Synthesis

1. Kuiper, R., Buskens, V., Raub, W., and Hoijtink, H. (2012). Combining statistical evidence from studies: A method using Bayesian updating and an example from research on trust problems in social and economic exchange. *Sociological Methods & Research*, 42, 60-81.
2. Zondervan-Zwijnenburg, M.A.J., Veldkamp, S.A.M., Nelemans, S.A., Neumann, A., Barzeva, S., Branje, S. J. T., van Beijsterveldt C.E.M., Meeus, W.H.J., Tiemeier, H., Hoijtink, H., Oldehinkel, A.J., and Boomsma, D.I. (unpublished). Parental age and offspring childhood mental health: A multi-cohort, population-based investigation.



# tain Evidence Synthesis

Researchers from three different cohort studies in the Netherlands (TRAILS, NTR, GEN-R, and RADAR) combined forces to investigate **one** research question using the data from the three cohorts.

Each cohort study tracks the development of thousands of children by repeatedly collecting data from them, their parents, their teachers etc., while they are growing up.

## bain Evidence Synthesis

One of the questions was whether age of the mother could be used to predict externalizing problem behavior (rated by the mother using the CBCL child behavior checklist) of children around the age of 11. Large samples were available for TRAILS (N=1955), NTR (N=21921), and GEN-R (N=4549).

# bain Evidence Synthesis

Subsequently, the following steps will be presented:

1. Randomly divide the data of each cohort into an exploratory and confirmatory part.
2. Use the exploratory data of the three cohorts to construct informative hypotheses with respect to the relation between mother age and externalizing problem behavior at the age of 11.
3. Use the confirmatory data of the three cohorts to evaluate the informative hypotheses using Bayes factors and the associated posterior model probabilities.
4. Combine the results obtained for the three cohorts into one overall conclusion (Bayesian research synthesis).

# Gain Evidence Synthesis

## Step 1

After randomly choosing 50% of each data set (the exploration set) the following results were obtained for each cohort:

Cohort	$\beta_1$	p-val	$\beta_2$	p-val	R2
Gen-R	-.10	<.001	.02	<.001	.02
NTR	-.11	<.001	.06	<.001	.02
TRAILS	-.13	<.001	.06	.06	.02

where the model was:

$$\text{CBCL} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2 + \text{error} \quad (2)$$

# bain Evidence Synthesis

## Step 2

The results were translated into following set of competing informative hypotheses

$$H_3 : \beta_1 < 0 \ \& \ \beta_2 > 0,$$

that is, the older the mothers the less externalizing problems occur, and, the rate of decrease "decreases" with age.

Competing hypotheses were

$$H_1 : \beta_1 = 0 \ \& \ \beta_2 = 0,$$

that is, age can not be used to predict externalizing problems,

$$H_2 : \beta_1 < 0 \ \& \ \beta_2 = 0,$$

that is, there is only a linear effect of age.

# Gain Evidence Synthesis

## Steps 3 and 4

Using the second the other 50% of the data of each of the three cohorts (the confirmation set) the following posterior model probabilities were obtained:

Cohort	PMP $H_1$	PMP $H_2$	PMP $H_3$	PMP $H_a$
Gen-R	.82	.04	.10	.05
NTR	.00	.97	.02	.01
TRAILS	.00	.88	.09	.03
All	.00	.99	.01	.00

# Random Evidence Synthesis

## Conclusion

1. If a research question can be translated into informative hypotheses, random evidence synthesis can be used to combine the information of multiple studies into one over-all conclusion with respect to the informative hypotheses under consideration.
2. Open question: how to deal with a situation in which not each study has (about) the same support for the informative hypotheses under consideration?

# A Closer Look at the Bayes Factor



# A Closer Look at the Bayes Factor

## Three Simple Hypotheses

Consider the hypotheses:

$$H_1 : \mu_1 \approx \mu_2, \text{ that is, } |\mu_1 - \mu_2| < .1$$

$$H_2 : \mu_1 > \mu_2$$

$$H_3 : \mu_1, \mu_2$$

# A Closer Look at the Bayes Factor

Information in the Data about the Two Means

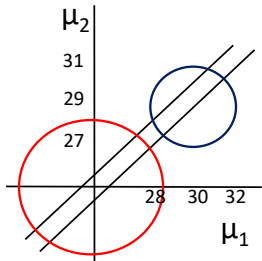
	N	Mean	SD	SE	95% Credible Interval	
					Lower	Upper
sex1	115	30.096	13.058	1.175	27.793	32.398
sex2	125	28.856	12.162	1.127	26.647	31.065

$$g(\mu_1, \mu_2 \mid \text{data}) \approx \mathcal{N} \left( \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} se_1^2 = \frac{SD_1^2}{N_1} & 0 \\ 0 & se_2^2 = \frac{SD_2^2}{N_2} \end{bmatrix} \right),$$

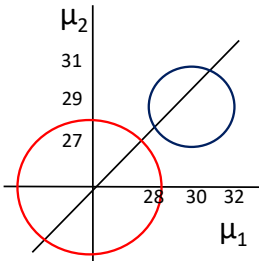
# A Closer Look at the Bayes Factor

Posterior Distribution, Prior Distribution, and Hypotheses

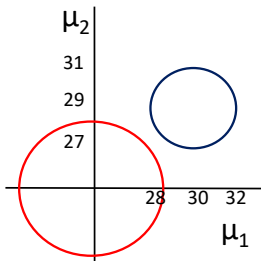
$$H_1: \mu_1 \approx \mu_2$$



$$H_2: \mu_1 > \mu_2$$



$$H_u: \mu_1, \mu_2$$



$$BF_{1u} = f_1/c_1 = .25/.05 = 5 \quad BF_{2u} = f_2/c_2 = .75/.5 = 1.5$$

$$BF_{12} = 5/1.5 = 3.33$$

# A Closer Look at the Bayes Factor

## Fit and Complexity

1. The fit of a hypothesis is the proportion of the posterior distribution in agreement with the hypothesis.
2. The complexity of a hypothesis is the proportion of the prior distribution in agreement with the hypothesis.

# A Closer Look at the Bayes Factor

## The Prior Distribution

$$h(\mu_1, \mu_2 \mid \text{data}) \approx \mathcal{N} \left( \begin{bmatrix} m \\ m \end{bmatrix}, \begin{bmatrix} \frac{SD_1^2}{.5*J} & 0 \\ 0 & \frac{SD_2^2}{.5*J} \end{bmatrix} \right),$$

Where  $\mu_1$  and  $\mu_2$  have the same prior mean  $m$ , and where  $J$  denotes the size of the training sample.

Choices for  $J$  for the example at hand:

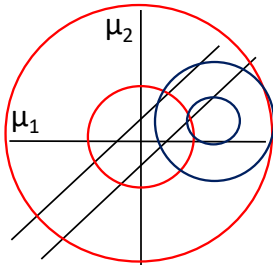
- Default in `bain`: number of independent constraints, that is, 1, this is a conservative choice
- Minimal training sample size, that is, 4, because four observations are needed to estimate two means and variances
- $J_{ref}$ , which renders  $BF_{0u} = 19$  if the effect size in the sample equals 0

# A Closer Look at the Bayes Factor

Prior Sensitivity for = Constrained Hypotheses

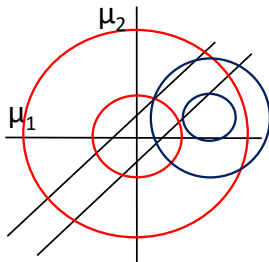
$$H_1: \mu_1 \approx \mu_2$$

$J = 1$



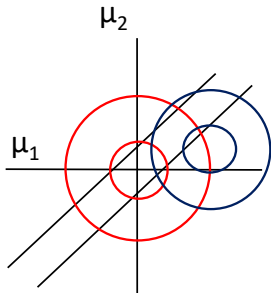
$$BF_{1u} = .2/.01 = 20$$

$J = 2$



$$BF_{1u} = .2/.05 = 4$$

$J = 3$



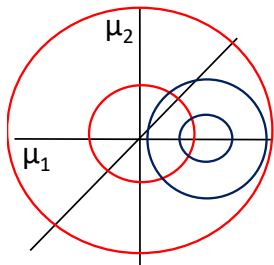
$$BF_{1u} = .2/.2 = 1$$

# A Closer Look at the Bayes Factor

Prior In-Sensitivity for  $> <$  Constrained Hypotheses

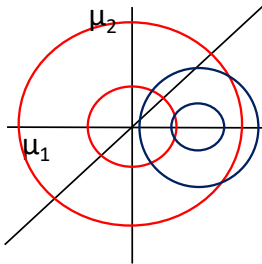
$$H_2: \mu_1 > \mu_2$$

$J=1$



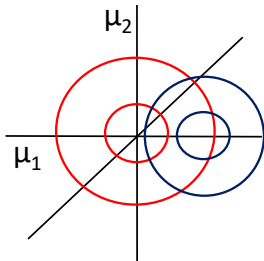
$$BF_{2u} = .9/.5 = 1.8$$

$J=2$



$$BF_{2u} = .9/.5 = 1.8$$

$J=3$



$$BF_{2u} = .9/.5 = 1.8$$

# Lab Meeting

## Exercises

Install R and RStudio. Install bain from CRAN using RStudio

Download BFtutorial.pdf and BFtutorial.R from  
<https://informative-hypotheses.sites.uu.nl/software/bain/>

Execute the following steps from BFtutorial.R (all these and subsequent steps are discussed in BFtutorial.pdf) for a first analysis: 1-5.

Then execute from the following what has your interest:  
Bayesian updating, Steps 6-7; Sensitivity analysis, Steps 8A en 8B; the effect of outliers, Step 9; Evaluating Informative Hypotheses, Steps 10-11; Evaluating a replication study, Steps 12A, 12B, 12C.