bain: Bayesian informative hypotheses evaluation

H. Hoijtink[1] C. van Lissa[1] J. Mulder[2] X. Gu[3]

[1]Department of Methodology and Statistics
Utrecht University H.Hoijtink@uu.nl

[1]Department of Methodology and Statistics
Utrecht University C.J.vanLissa@uu.nl

[2]Department of Methodology and Statistics
Tilburg University J.Mulder3@uvt.nl

[3]Department of Educational Psychology
East China Normal University GuXin57@hotmail.com

Outline

The Replication Crisis

A Research Project

Mixed Emotions in the Face of Endings

On the day of their graduation, students of Stanford University either read "a text that reminds them an important period in their life is coming to an end" or "a text that does not". Subsequently, the degree to which they have "mixed emotions" (being both happy and sad) is measured.

1. Ersner-Hershfield, H., Mikels, J.A., Sullivan, S.J., and Carstensten, L.L. (2008). Poignancy: Mixed emotional experience in the face of meaningful endings. *Journal of Personality and Social Psychology, 94, 158-167. DOI: 10.1037/0022-3514.94.1.158*

Research Question

Do students that are reminded that their life as a student is ending have more or less mixed emotions than students that are not reminded that their life as a student is ending.

Variables

Each student is in one of two groups: "yes, reminded" or "no, not reminded".

The score on mixed emotions is the smallest of :

1. Happiness measured on a 7-point scale running from 1=not at all until 7=extremely
2. Sadness measured on a 7-point scale

The following hypotheses will be evaluated:

$H_0 : \mu_{yes} = \mu_{no}$

$H_a : \mu_{yes} \neq \mu_{no}$

The Main Research Outcomes

	de niet groep	de wel groep
N	59	51
m	3	3.75
sd	1.66	1.66
Cohen's d	.45	
p-value	.02	

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.

Let's Do It Again

The mixed emotions research project was replicated by Talhelm, Lee, and Eggleston (2015).

	origineel		replicatie	
	niet	wel	niet	wel
N	59	51	112	110
m	3	3.75	3.75	3.74
sd	1.66	1.66	1.00	1.00
Cohen's d	.45		.01	
p-value	.02		.94	

1. Talhelm, T., Lee, M., and Eggleston, C. (2015) Replication of Ersner-Hershfield, H., Mikels, J.A., Sullivan, S.J., and Carstensten, L.L. (2008) https://osf.io/fw6hv/

Discuss. First in Small Groups, then Plenary

What is going on? Why do the results from Ersner-Hershfield et al. (2008) not replicate in Talhelm et al. (2015). What does this mean for psychological science?

The Replication Crisis

The Open Science Collaboration has replicated 100 studies from the journals: Psychological Sciencs, Journal of Personality and Social Psychology, and Journal of Experimental Psychology: Learning, Memory, and Cognition.

Differences between means in the replication studies were usually half as large as differences in means in the original studies.

In almost all original studies the p-value was smaller than .05. In only 1/3 of the replication studies the p-value was smaller than .05.

1. Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science, 349, aac4716. http://dx.doi.org/10.1126/science.aac4716

Discuss. First in Small Groups, then Plenary

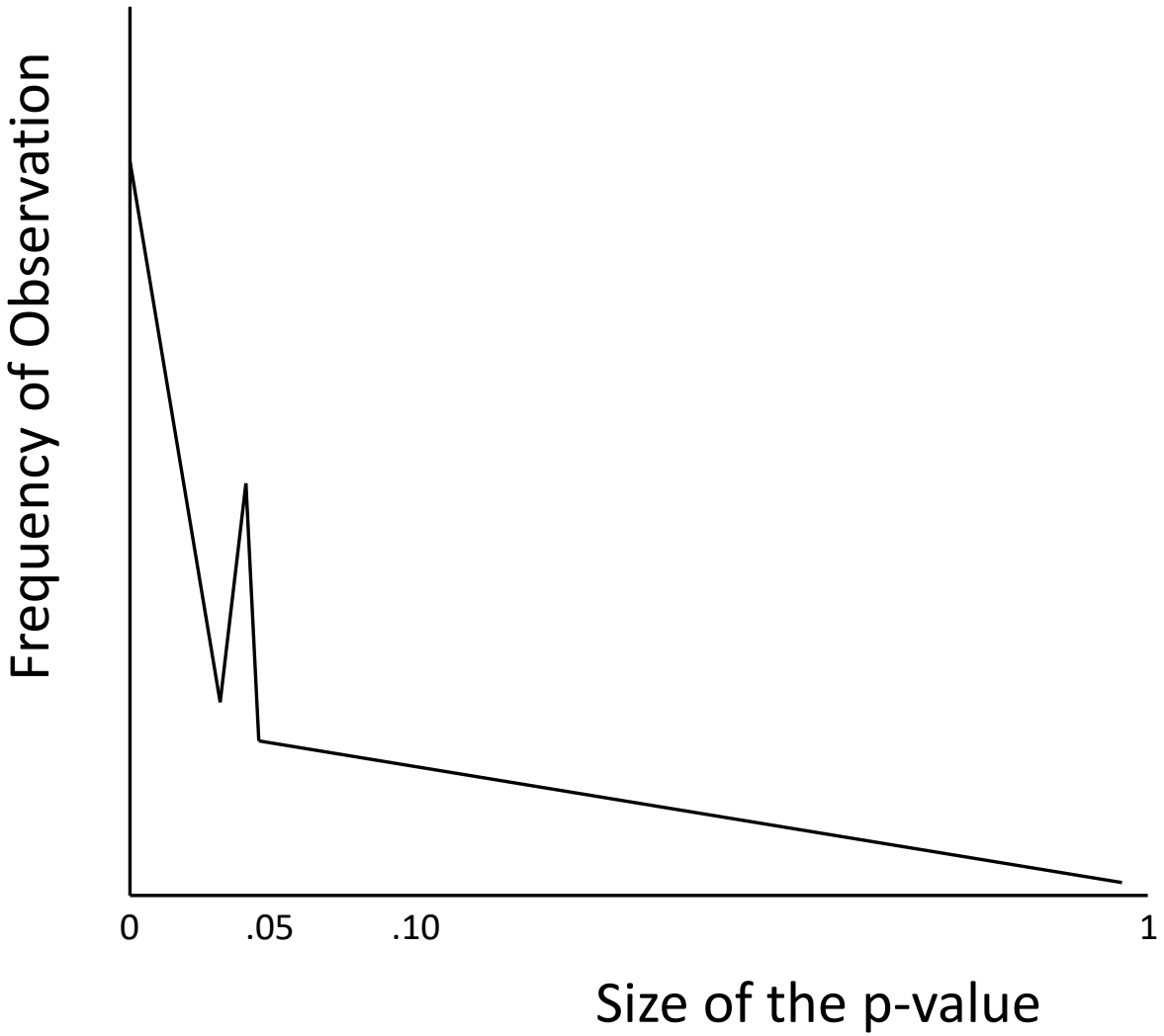
What do you think are the causes of the replication crisis?

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. http://dx.doi.org/10.1080/17470218.2012.711335

Causes of the Replication Crisis



Discuss. First in Small Groups, then Plenary

What can explain the unexpected large number of p-values just below .05?

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 mixed emotions from the "yes" group, the p-value becomes .04 :-))
2. After testing the null-hypothesis, the resulting p-value is .06 :-)) But after using a different measure of mixed emotions "the difference between Happiness and Saddness", the p-value becomes .04 :-))

Incentives for Questionable Research Practices

Found somewhere on the internet:

TS college 8 beamer.pdf

. . . and has real-life consequences

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

after Geoff Cumming, 2012

Prevalence of Questionable Research Practices

1. When explicitly asked if they ever fabricated or falsified research data, or if they altered or modified results to improve the outcome, between 0.3% and 4.9% of scientists replied affirmatively
2. Other questionable practices were admitted by up to 33.7% of respondents. Consistently across studies, scientists admitted more frequently to have "modified research results" to improve the outcome than to have reported results they "knew to be untrue".
3. When asked if they had personal knowledge of a colleague who fabricated or falsified research data, or who altered or modified research data between 5.2% and 33.3% of respondents replied affirmatively.

1. Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PloS ONE, 4, e5738. http://dx.doi.org/10.1371/journal.pone.0005738

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

Discuss. First in Small Groups, then Plenary

What can we, journal editors, researchers, students, the academic community as a whole, do to address the replication crisis?

How can the Replication Crisis be Addressed?

1. Always have a replication study execution by independent others and include the results in your research paper.
2. Pre-Registration and "Get Rid of the .05".
3. Multi-Group-Laboratory consortia investigating the same research questions and jointly publishing the research results.

As will be elaborated, bain can contribute to each of these three options.

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

P-values and Alpha Level

1. The p-value is *not* a measure of support for the null-hypothesis.
2. After observing ".06" 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.

P-values and Alpha Level

3 One can not compare more than two hypotheses, for example:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_{a1} : \mu_1 = \mu_2, \mu_3$$

$$H_{a1} : \mu_1 = \mu_3, \mu_2$$

$$H_{a1} : \mu_2 = \mu_3, \mu_1$$

$$H_a : \mu_1, \mu_2, \mu_3$$

P-values and Alpha Level

The traditional alpha level is .05. This is the cause of:

1. Sloppy sciencs (Ioannides, 2005)
2. Publication bias
3. "Surely God loves the .06 as much as the .05 ..." (Rosnow and Rosenthal, 1989). Dichotomous decisions are weird!! "... but he is completely infatuated with the .005."
4. Were does the .05 come from? Fisher used "no level", .05, .02 and was in no way married to the .05. And what about those that argue in favor of the .005 (Benjamin et al., 2017)?
5. Suppose the p-value is equal to .03, that is, smaller than .05: "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

	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.

Background Analysis

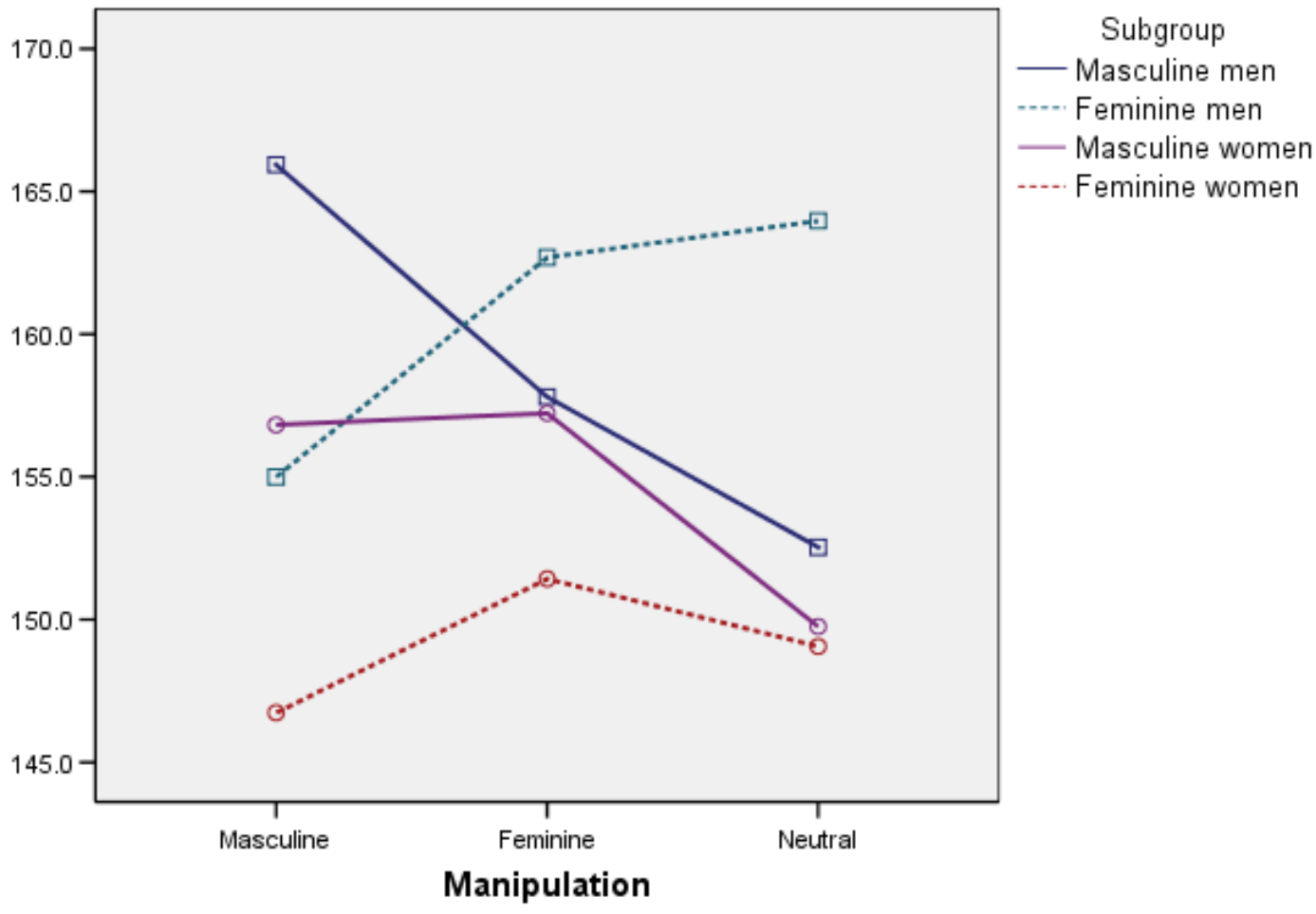
Tests of Between-Subjects Effects

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)

Background Algorithm



bain: Bayesian Informative Hypotheses Evaluation

Informative Hypotheses

Example 1: AN(C)OVA

	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

AN(C)OVA

	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

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

Informative Hypotheses

AN(C)OVA

	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

AN(C)OVA

	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

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

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

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

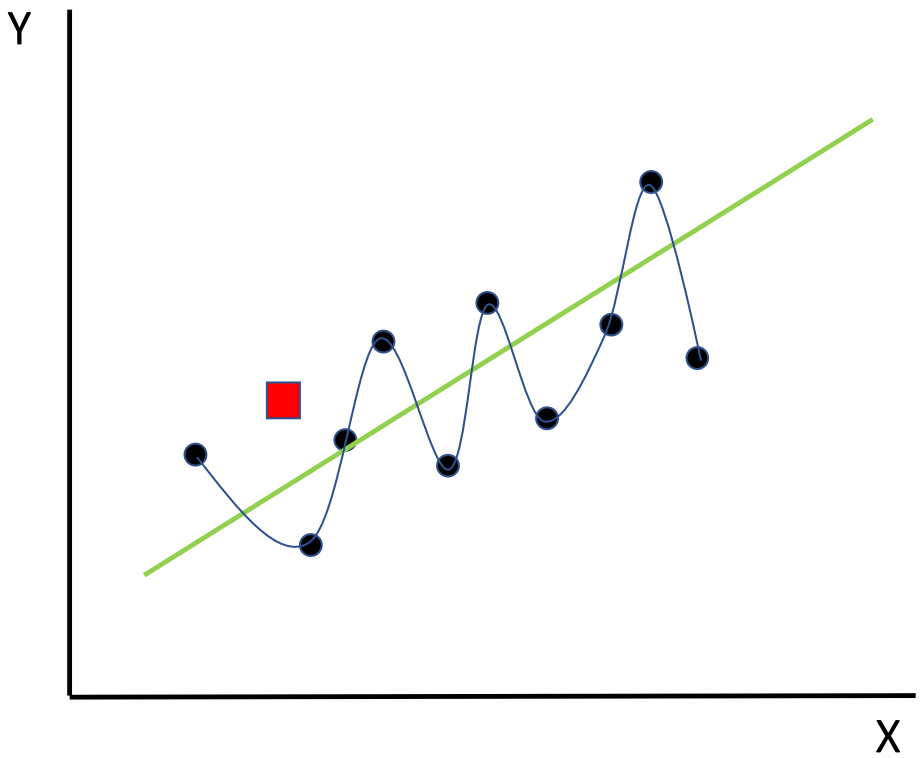
contains 2 orderings of three means, 1-2-3 and 1-3-2, thus is less parsimonious

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

contains all six possible orderings of three means, thus is not parsimonious

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

$$H_u : \mu_1, \mu_2, \mu_3 \qquad BF_{iu} = \frac{f_i}{c_i}$$

$$H_{i'} : \mu_1 = \mu_2 = \mu_3 \qquad BF_{ii'} = \frac{f_i}{c_i} / \frac{f_{i'}}{c_{i'}} = BF_{iu} / BF_{i'u}$$

$$H_c : \text{ not } H_i \qquad BF_{ic} = \frac{f_i}{c_i} / \frac{1 - f_i}{1 - c_i} = BF_{iu} / BF_{cu}$$

Bayes Factor

Interpreting (the Size of) the Bayes Factor

	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

Sex Match Effect

$H_1 : 166, 155 > 158, 154, 163, 164$ and $157, 152 > 157, 150, 143, 149$

Gender Role Match Effect

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

$H_2 : 166, 163 > 158, 154, 155, 164$ and $157, 152 > 157, 150, 143, 149$

Sex Mismatch Effect

$H_3 : 158, 163 > 166, 154, 155, 164$ and $157, 143 > 157, 150, 152, 149$

Gender Role Mismatch Effect

$H_4 : 158, 155 > 166, 154, 163, 164$ 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 ...
3. Will again lead to publication bias ...
4. "Surely God loves the 2.9 as much as the 3.1 and he is completely infatuated with the 10".
5. Were does the 3 come from?

Bayes Factor

Interpreting (the Size of) the Bayes Factor

When is the Bayes factor large enough?

1. Design your research project and formulate informative hypotheses
2. Send to journal - good design and interesting hypotheses leads to accepted paper
3. Collect data and evaluate hypotheses. Is one good and the best - nice! Is none good - BIG news, well-constructed hypotheses that are supported by our peers have been rejected.

Bayesian Error Probabilities

The ratio of two Posterior Model Probabilities (the posterior odds) can be computed using the BF and the prior odds via:

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

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" ...

Bayesian Error Probabilities

The Posterior Model Probabilities $PMP(H_i \mid \text{data})$ and $PMP(H_c \mid \text{data})$ quantify the support in the data for each hypothesis.

In case of the comparison of H_i with H_c:

$PMP(H_i \mid \text{data})$ can be seen as the Bayesian *error* probability when H_c is selected as the preferred hypothesis

$PMP(H_c \mid \text{data})$ is the Bayesian *error* probability when H_i is selected as the preferred hypothesis.

Bayesian Error Probabilities

The Posterior Model Probability (PMP) is not a probability it is a measure of support in the data accounting for the fit and complexity of a hypothesis on a 0-1 scale.

The sum of the PMPs for a set of hypotheses is always 1.0.

It is **not** a frequentist probability, subjectivity enters in three ways:

1. The choice of which and how many hypotheses to evaluate
2. The choice of the prior distribution (not yet discussed, is coming up later)
3. The choice of the prior model probabilities.

Bayes Factor

The Number of Hypotheses and PMPs

	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_u :				.14	1/5

Bayes Factor

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				.180	1/13
...					
H_{12} : Don't miss something				.040	1/13
H_u :				.047	1/13

Bayes Factor

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				0	0
...					
H_{12} : Don't miss something				0	0
H_u :				.14	1/5

Bayesian Updating

Bayes Factor

Bayesian Updating

Repeated significance testing after increasing the sample implies "planning ahead" and "correction for capitalization on chance"

"Bayesian updating" is simply recomputing the evidence presented by all the data that are currently available

	N 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

A Closer Look at the Bayes Factor: Approximation and Sensitivity

A Closer Look at the Bayes Factor: Three Simple Hypotheses

Consider the hypotheses:

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

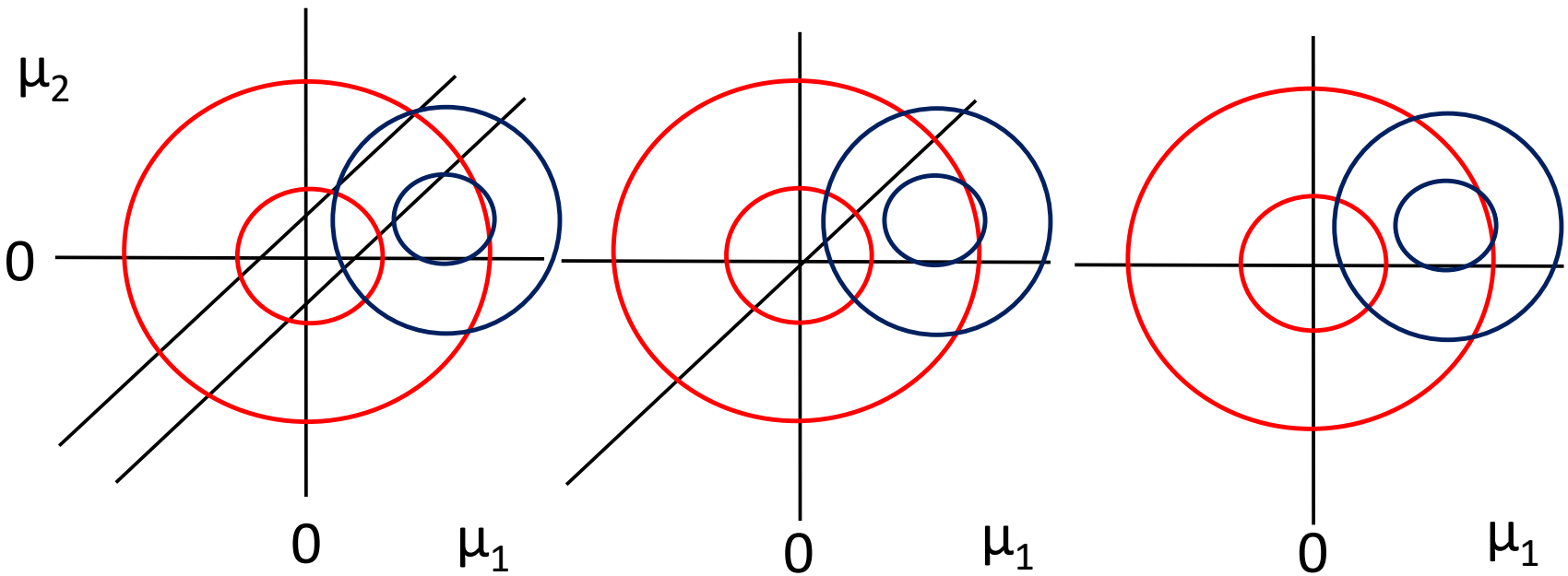
$$H_2 : \mu_1 > \mu_2$$

$$H_3 : \mu_1, \mu_2$$

A Closer Look at the Bayes Factor

Posterior Distribution, Prior Distribution, and Hypotheses

$H_1: \mu_1 \approx \mu_2$ $H_2: \mu_1 > \mu_2$ $H_u: \mu_1 , \mu_2$



$BF_{1u} = f_1/c_1 = .1/.2 = .5$ $BF_{2u} = f_2/c_2 = .9/.5 = 1.8$

$BF_{21} = 1.8/.5 = 3.6$

A Closer Look at the Bayes Factor

Properties of the Posterior Distribution

1. The mean of the posterior distribution is given bij the estimate of μ_1 and μ_2 in the data at hand, that is, the sample means.
2. For each mean the variance of the posterior distribution is the residual variance σ^2 divided by the sample size of the group at hand (the squared standard error of the mean).
3. The fit of an hypothesis is the proportion of the posterior distribution in agreement with the hypothesis.
4. Note that, bain uses a normal approximation of the posterior distribution of μ_1 and μ_2. The true posterior is a t-distribution.

A Closer Look at the Bayes Factor

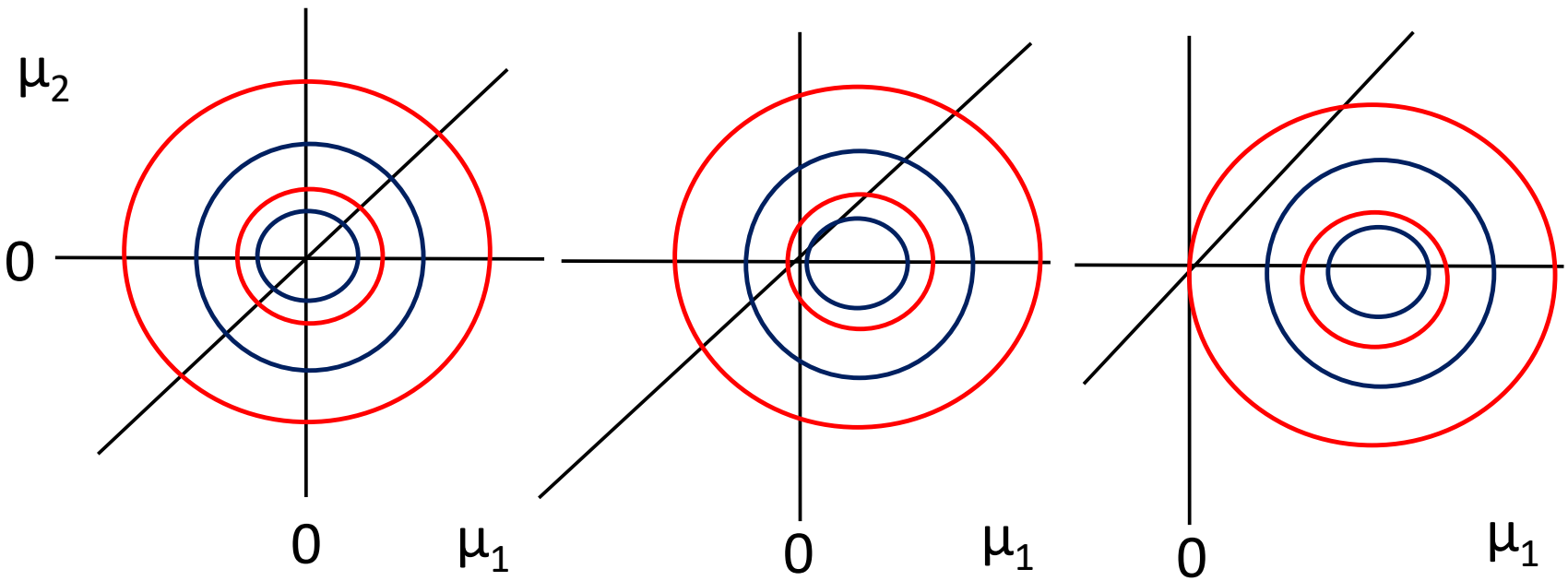
Properties of the Prior Distribution

1. The mean of the prior distribution is set at 0,0. If this is not done, the resulting Bayes factor will be inconsistent (see later).
2. For each mean, the variance of the prior distribution is $2 \times \sigma^2$, and is obtained using a so-called, minimal fraction $1/(2N)$ of the information in the data to compute the variance. Note that N denotes the sample size in each of the groups.
3. The complexity of an hypothesis is the proportion of the prior distribution in agreement with the hypothesis.

A Closer Look at the Bayes Factor

Motivating the Adjusted Mean of the Prior Distribution

$H_2: \mu_1 > \mu_2$



$BF_{2u} = f_2/c_2 = .5/.5 = 1$ $BF_{2u} = .8/.8 = 1$ $BF_{2u} = .95/.95 = 1$

A Closer Look at the Bayes Factor

Prior Sensitivity

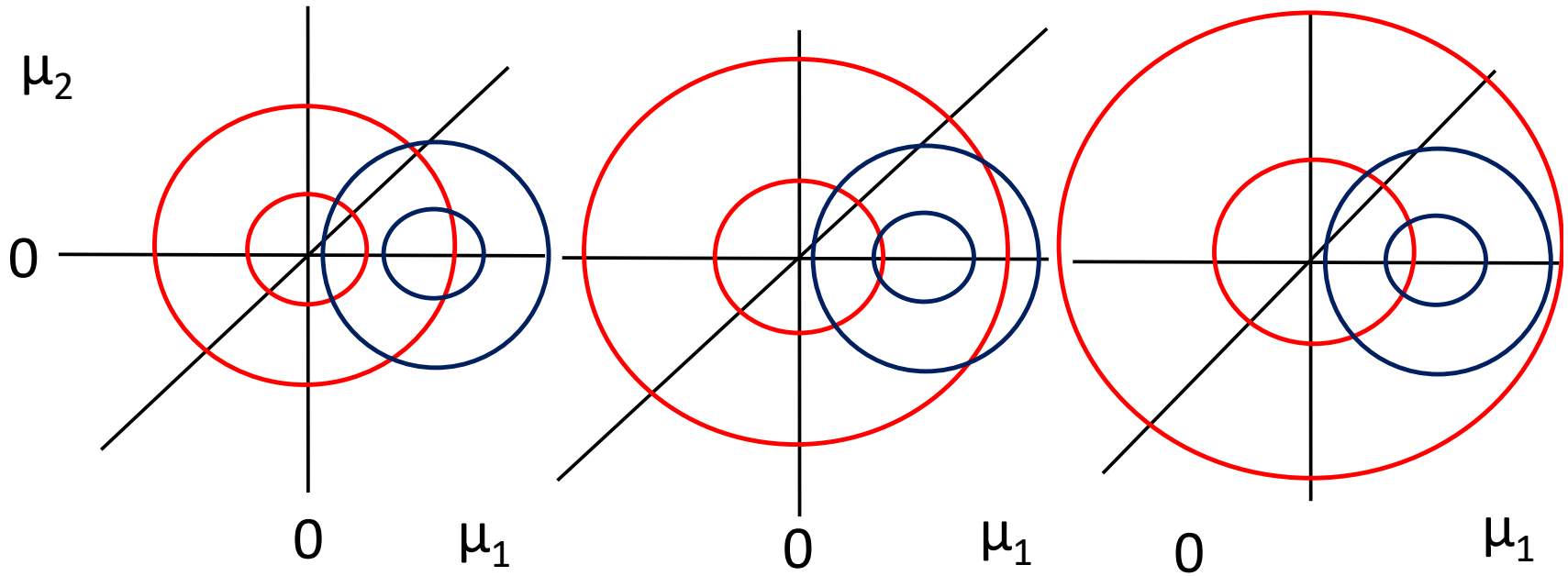
If an hypothesis is formulated using ONLY inequality constraints (that is, not equality or range constraints), the Bayes factor is independent of the variance of the prior distribution.

If an hypothesis is formulated using equality (or range) constraints, the Bayes factor is SENSITIVE to the choice of the prior distribution. It is then recommended to execute a sensitivity analysis. In bain this is achieved by not only using the minimal fraction of information to specify the variance of the prior distribution, but also smaller and larger fractions.

A Closer Look at the Bayes Factor

Prior In-Sensitivity for > < Constrained Hypotheses

$H_2: \mu_1 > \mu_2$

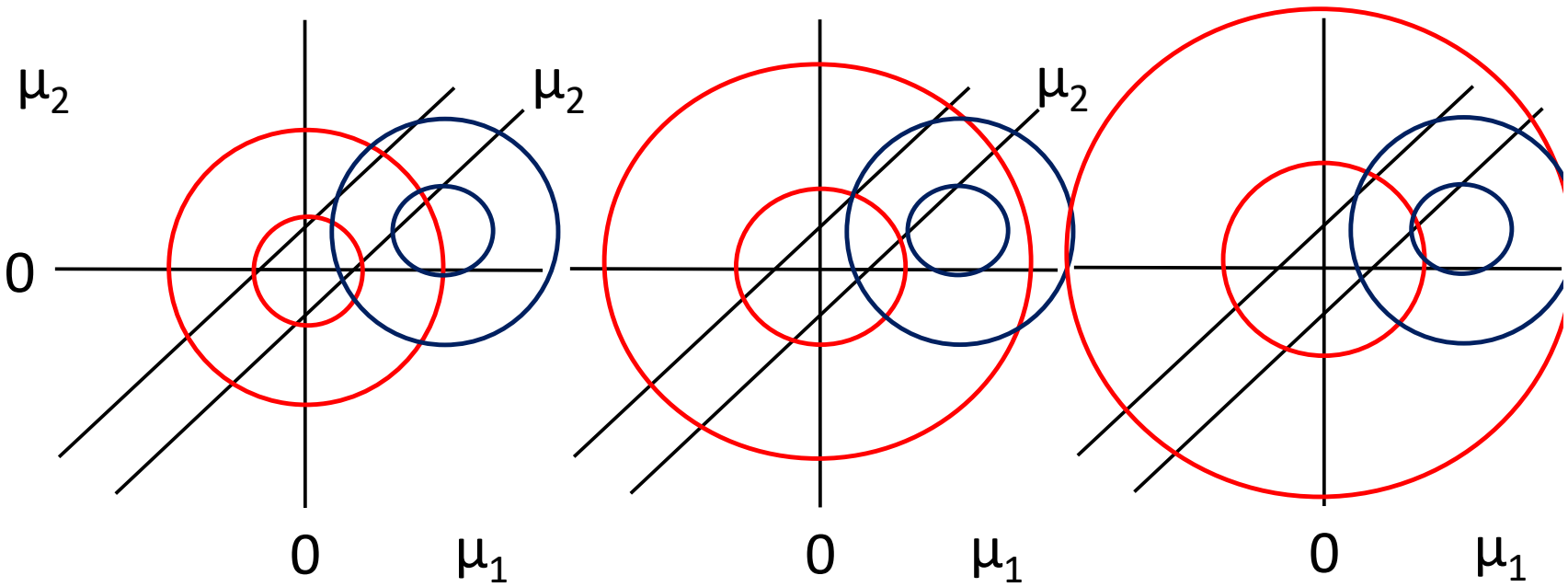


$BF_{2u} = f_2/c_2 = .9/.5 = 1.8$ $BF_{2u} = .9/.5 = 1.8$ $BF_{2u} = .9/.5 = 1.8$

A Closer Look at the Bayes Factor

Prior Sensitivity for = Constrained Hypotheses

$H_1: \mu_1 \approx \mu_2$



$BF_{1u} = f_1/c_1 = .2/.2 = 1$ $BF_{1u} = .2/.05 = 4$ $BF_{1u} = .2/.01 = 20$

Beyond ANOVA

Informative Hypotheses

Example 2: Repeated Measures

Development of depression

	Measurement			
	8 years	12 years	16 years	20 years
Men	μ_1	μ_2	μ_3	μ_4
Women	μ_5	μ_6	μ_7	μ_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 3: Multiple Regression

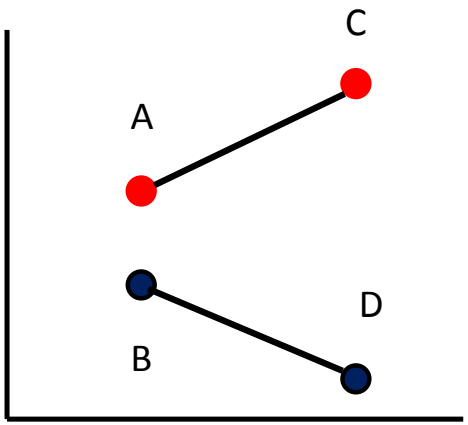
$$\text{Income}_i = \beta_0 + \beta_1 \times \text{IQ}_i + \beta_2 \times \text{SES}_i + \epsilon_i$$

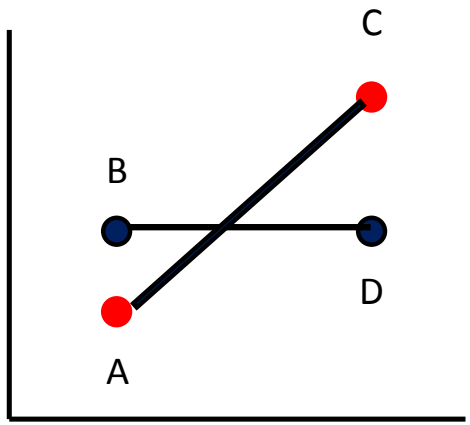
$$H_1 : \beta_1 > 0, \beta_2 > 0, \beta_1 > \beta_2$$

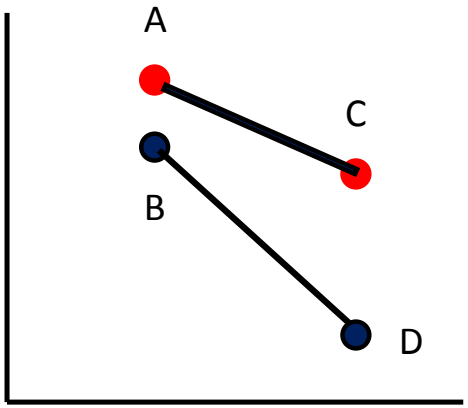
Note: β_1 and β_2 are only comparble if IQ and SES are standardized (either data standardization or parameter standardization)

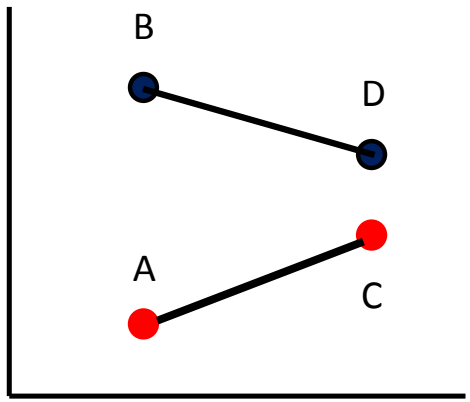
Informative Hypotheses

Example 4: ANOVA Interaction Effect









$$H_i : (A-B) < (C-D)$$

to identify add

$$A > B$$

$$A < C$$

Informative Hypotheses

Example 5: About Equality Constraints

$$H_0 : \mu_1 = \mu_2$$

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

Results could be
$p \in \{.01, .05, .30\}$

And then what?

$$H_0 : (\mu_1 - \mu_2) > .2sd$$

$$H_0 : (\mu_1 - \mu_2) < .2sd$$

Results could be
$BF_{12} \in \{.2, 1, 10\}$

Note subjective choice of "relevance" interval in terms of Cohen's d =.2 in terms of observed standard deviation

Informative Hypotheses

Example 6: The General Form of Informative Hypotheses

bain can handle hypotheses build using constraints on linear combinations of parameters. For example (parameters are named a, b, c, etc.:

1. a-b > b-c > c-d < d-e = e-f
2. 2*a + 5*b - 5 < b - 3*c + 7 > 5

Carefully read the bain help file to learn what is and is not possible with respect to the specification of informative hypotheses

Informative Hypotheses

Input for Single and Multiple Group Analyses

bain makes a distinction between single and multiple group analyses.

1. A multiple regression is a single group analysis
2. An ANOVA is a multiple group analysis

The bain instruction differs for single and multiple group analyses. Read the bain help for further instructions. Note, if each group has (about) the same sample size, a multiple group analysis can (about) be treated as a single group analysis.

bain Evaluation of Replication Studies

Replication Research

Monin, Sawyer, and Marquez (2008) investigate the attraction to "moral rebels", that is, persons that take an unpopular morally laudable stand. Imagine that you are in a group (all the others in the group are actors) and that the atmosphere in the group is that criminal behavior is linked to having an African American background.

1. You publicly have to rate your attraction to a person in a video.
2. This is repeated using the same group of actors with you replaced by another person, that is, there are more participants in the experiment that have to rate the attraction to a person in a video.
3. There are three experimental conditions (see the next slide).

There are three conditions in the experiment:

1. Condition 1 participants rate the attraction to a person that is "obedient" and selects an African American person from a police line up of three.
2. Condition 2 participants rate a moral rebel (a person not selecting the African American person) after executing a self-affirmation task intended to boost their self-confidence.
3. Condition 3 participants rate a moral rebel after executing a bogus writing task.

This study was replicated by Holubar (2015).

Replication Research

Descriptives obtained for the Monin data

group	n	mean	sd
1	19	1.88	1.38
2	19	2.54	1.95
3	29	0.02	2.38

Replication Research

Hypotheses evaluated for the Monin data

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_{a1} : \mu_1 = \mu_2, \mu_3$$
$$H_{a2} : \mu_1 = \mu_3, \mu_2$$
$$H_{a3} : \mu_2 = \mu_3, \mu_1$$
$$H_a : \mu_1, \mu_2, \mu_3,$$

Replication Research

Bayes factors and PMPs obtained for the Monin data

```
Hypothesis testing result
      f= f><|=    c= c><|=     f     c  BF1c   PMPb
H0       0     1 0.015     1     0 0.015 0.001  0
Ha1  0.367     1 0.114     1 0.367 0.114 3.216  0.754
Ha2  0.005     1 0.114     1 0.005 0.114 0.045  0.011
Ha3      0     1 0.114     1     0 0.114 0.001  0
Ha       .     .     .     .     .     .     .  0.235
```

Replication Research

Evaluating Replication Studies by means of Bayesian Hypotheses Evaluation

Step 1. Translate results of the original study into an informative hypothesis H_i.

Component 1. $\mu_1 = \mu_2$
Component 2. $\mu_1 > \mu_2$
Component 3. $\mu_1 > \mu_2 + .2sd$

Step 2. Choose as competing hypotheses
H_0 : all the means are equal and H_a or H_c

Replication Research

The study by Monin, Sawyer, and Marquez (2008) seems to imply H_i presented below:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_i : \mu_1 = \mu_2 > \mu_3$$
$$H_a : \mu_1, \mu_2, \mu_3$$

Replication Research

Replicating Monin, Sawyer, and Marquez (2008) using the Holubar data

```
Hypothesis testing result
      f= f><|=    c= c><|=     f     c  BF1c  PMPa  PMPb
H0 0.111     1 0.022     1 0.111 0.022 5.023 0.816 0.702
Hi  0.12 0.655 0.138   0.5 0.079 0.069 1.134 0.184 0.158
Ha     .     .     .     .     .     .     .     .  0.14

BF-matrix
      H0    Hi
H0 1.000 4.428
Hi 0.226 1.000
```

bain Evidence Synthesis

bain Evidence Synthesis

Researchers from four different cohort studies in the Netherlands (TRAILS, NTR, GEN-R, and RADAR) decided to combine forces to investigate **one** research question using the data from the four 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. These data were available for TRAILS (N=1955), NTR (N=21921), and GEN-R (N=4549).

What follows is a presentation of the methodology (Kuiper et al., 2012; Zondervan-Zwijnenburg et al., unpublised) used to answer this research question.

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).

bain Evidence Synthesis

Step 1

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

Cohort	β_1	p-val	β_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

Informative hypotheses represent the expectations of researchers by translating them into restrictions on parameters of the statistical model of interest. Examples are:

1. $H_1 : \mu_{xanax} < \mu_{placebo} < \mu_{control}$ (when the goal is to treat depression)
2. $H_2 : \beta_{IQ} > 0, \beta_{SES} > 0, \beta_{IQ} > \beta_{SES}$ (when the goal is to predict income using multiple regression)

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, and,

$$H_a : \text{ none of the above.}$$

bain Evidence Synthesis

Step 3

1. For each of H_1, H_2, H_3 the Bayes factor versus H_a is computed. The Bayes factor evaluates the *fit* and *complexity* of each hypothesis. If, for example, $BF_{1a} = 5$, this implies that the support in the data is 5 times larger for H_1 than for H_a.
2. The information in the resulting Bayes factors are translated into, so-called, posterior model probabilities. These are numbers on a 0-1 scale that quantify the relative support in the data for each of the hypotheses under consideration (again accounting for the fit and complexity of each hypothesis).

bain 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

bain Evidence Synthesis

Step 4

Bayes theorem is used to update the posterior model probabilities resulting from Gen-R using the PMPs resulting from NTR and TRAILS (combining the results from different studies). This consists of the following steps:

1. The PMPs resulting from Gen-R are called P_1, P_2, P_3, and P_a
2. These can be updated using PMP_1, PMP_2, PMP_3, and PMP_a, from NTR into PU_1, PU_2, PU_3, and PU_a

bain Evidence Synthesis

Step 4

3. An example of the formula used is:

$$\mathrm{PU}_1 = \frac{\mathrm{PMP}_1 \times \mathrm{P}_1}{\mathrm{PMP}_1 \times \mathrm{P}_1 + \mathrm{PMP}_2 \times \mathrm{P}_2 + \mathrm{PMP}_3 \times \mathrm{P}_3 + \mathrm{PMP}_a \times \mathrm{P}_a}.$$

This is an application of Bayes theorem because

$$P(H_1|NTR) = \frac{P(NTR|H_1) \times P(H_1)}{P(NTR|H_1) \times P(H_1) + \ldots + P(NTR|H_a) \times P(H_a)},$$

or, in "highschool notation":

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}$$

bain Evidence Synthesis

Step 4

4. Analogously, the PU's are then updated using the PMPs resulting from TRAILS. Let us return to the last line of the table above for the "combined" result.

bain Evidence Synthesis

Conclusion

1. If a research question can be translated into informative hypotheses, bain 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?

Lab Meeting

bain

bain can within Rstudio be installed from CRAN. Carefully read the help file before using bain.

A part of bain is iimplemented in JASP https://jasp-stats.org/. Carefully read the help file before using the bain-JASP implementation.

Further information can be found at https://informative-hypotheses.sites.uu.nl/software/bain/

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.

References

1. Benjamin, D.J., ..., Hoijtink, H. ... (2017). Redefine statistical significance. *Nature Human Behavior*
2. Cohen, J. (1994). The earth is round, p<.05. *American Psychologist, 49,* 997-1003. Hoijtink, H., Mulder, J., van Lissa, C., and Gu, X. (2018). A tutorial on testing hypotheses using the Bayes factor. Psychological Methods. DOI: 10.1037/met0000201
3. Holubar, T. (2015). Replication of "The rejection of moral rebels", study 4 by Monin, Sawyer, and Marques (2008, JPSP). https://osf.io/ezcuj/
4. Ioannides, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124.
5. Jeffreys, H. (1961). *Theory of Probability.* Oxford: Clarendon Press.
6. Kass, R.E. and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*, 773-795.
7. 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.
8. Monin, B, Sawyer, P.J., and Marquez, M.J. (2008). The rejection of moral rebels: resenting those who do the right thing. *Journal of Personality and Social Psychology, 95,* 76-93.

References

9. Rosnow, R.L. and Rosenthal, R. (1989). Statistical procedures and the justifications of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.
10. Royal, R. (1997). *Statistical Evidence. A Likelihood Paradigm.* New York: Chapman and Hall/CRC.
11. Van Well, S., Kolk, A.M., and 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.
12. 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.