Using Bayes Factors to test hypotheses in addiction science

Dr Emma Beard

Department of Epidemiology and Public Health, University College London

Department of Clinical, Educational and Health Psychology, University College London

Professor Robert West

Department of Epidemiology and Public Health, University College London, London

Professor Zoltan Dienes

School of Psychology, University of Sussex, Brighton

Dr Colin Muirhead

Institute of Health and Society, Newcastle University

Acknowledgements

- I have received unrestricted research funding from Pfizer for the Smoking Toolkit Study (www.smokinginengland)

- I am funded by Cancer Research UK and the National Institute for Health Research's School for Public Health Research (NIHR SPHR)

This is a partnership between:	
•The University of Sheffield •The University of Bristol •The University of Cambridge •University College London •The London School for Hygiene and Tropical Medicine •The University of Exeter Medical School	•The LiLaC collaboration between the Universities of Liverpool and Lancaster •Fuse; The Centre for Translational Research in Public Health, a collaboration between Newcastle, Durham, Northumbria, Sunderland and Teesside Universities

This is an outline of independent research funded by the National Institute for Health Research's School for Public Health Research (NIHR SPHR).

The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health

Limitations of p-values > 0.05

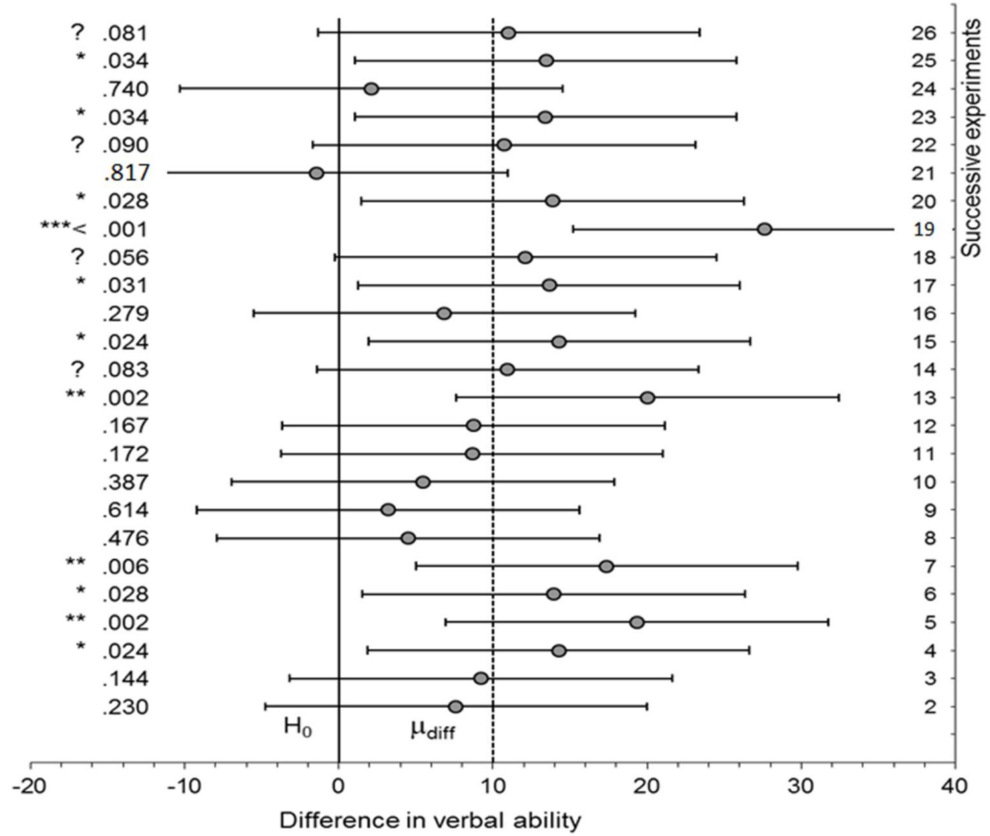
- No scientific conclusion follows automatically from *p*>0.05
- *A p-value is the probability of obtaining an effect <u>at least as extreme</u> as the one in your sample data, <u>assuming the null hypothesis is true</u>*
  - *p*>0.05 ≠ evidence for the null hypothesis
  - *p>0.05* = insufficient evidence to reject the null hypothesis

P(Observation | Hypothesis) ≠ P(Hypothesis | Observation)

Limitations of p-values > 0.05

- A p>0.05 could reflect either no evidence for an effect or data insensitivity (i.e. low power/high standard error)

- **Illustrative example:** *The dance of the p-value*



Cummings (2011)

Solution 1: Use power to determine data insensitivity

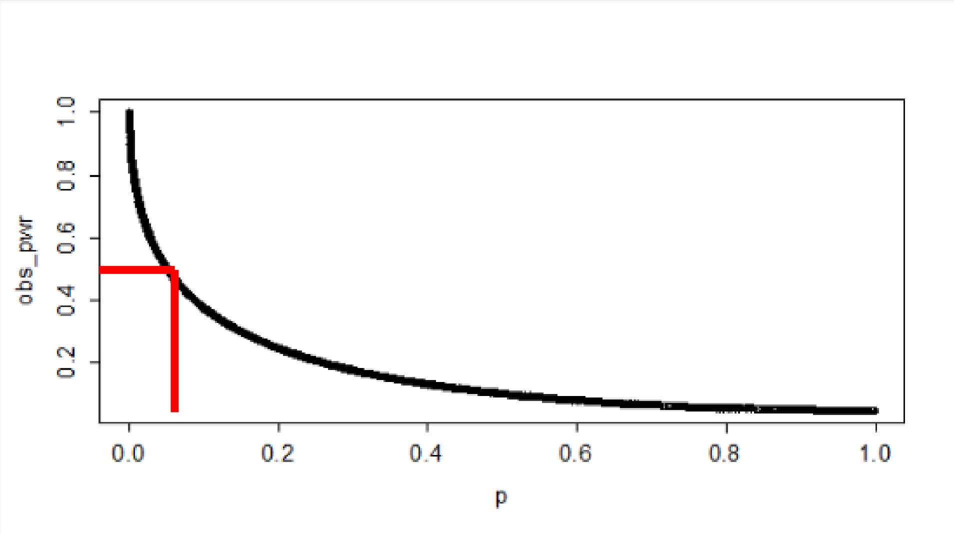
- When power is high we can be more confident that p>0.05 reflects no evidence for an effect

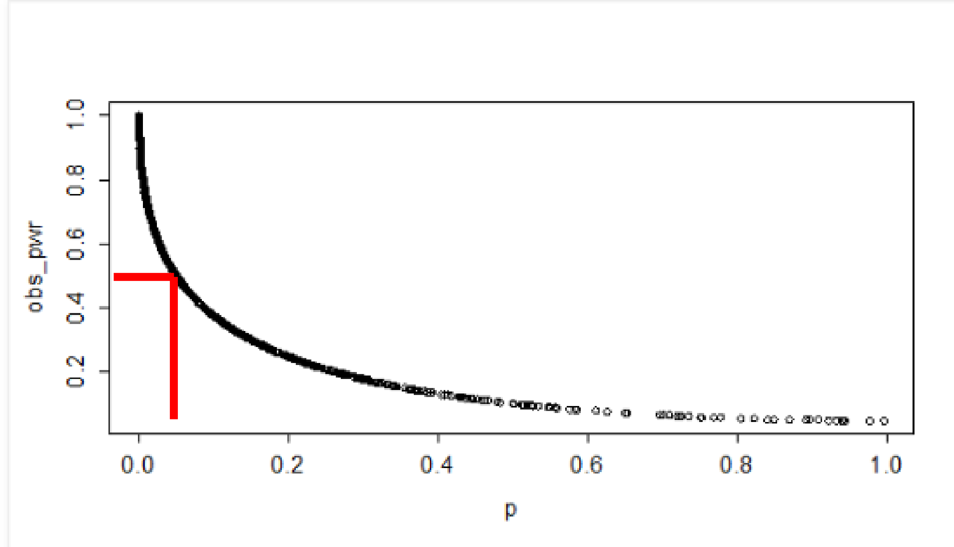
- When power is low there is a higher possibility of accepting the null when it is false i.e. that the data are insensitive

- If we have power of 80% then the chances of a type 2 error is 20%

- But . . . one needs to specify the <u>minimal interesting value</u> that is plausible . . . and power <u>cannot use the data themselves</u> in order to determine how sensitive the data are

Post-hoc power ←→ p-values





Plot of observed *p*-values and observed power for 10000 simulated studies with approximately **50%** power

Plot of observed *p*-values and observed power for 10000 simulated studies with approximately **90%** power

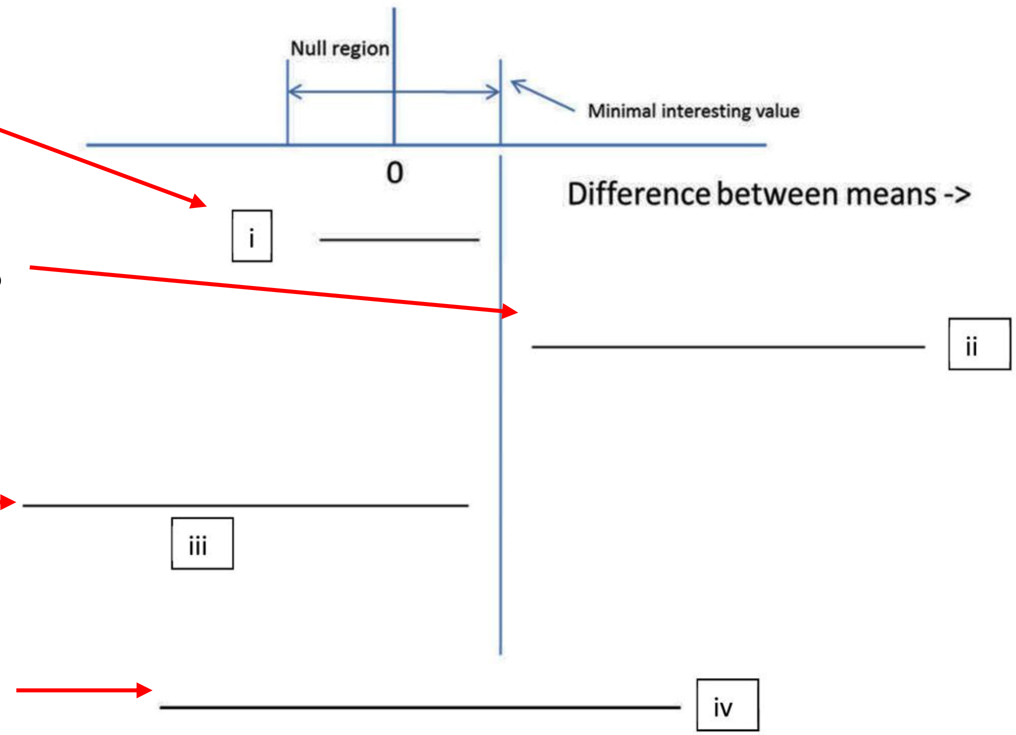
http://daniellakens.blogspot.co.uk

Solution 2: Use confidence intervals to determine data insensitivity

- Confidence intervals can indicate how sensitive the data are based on the very data themselves
- A confidence interval provides a set of possible population values consistent with the data (Cumming, 2011)
- When we specify a null hypothesis we can specify a null region rather than a point value
  - We can then draw four conclusions . . . .

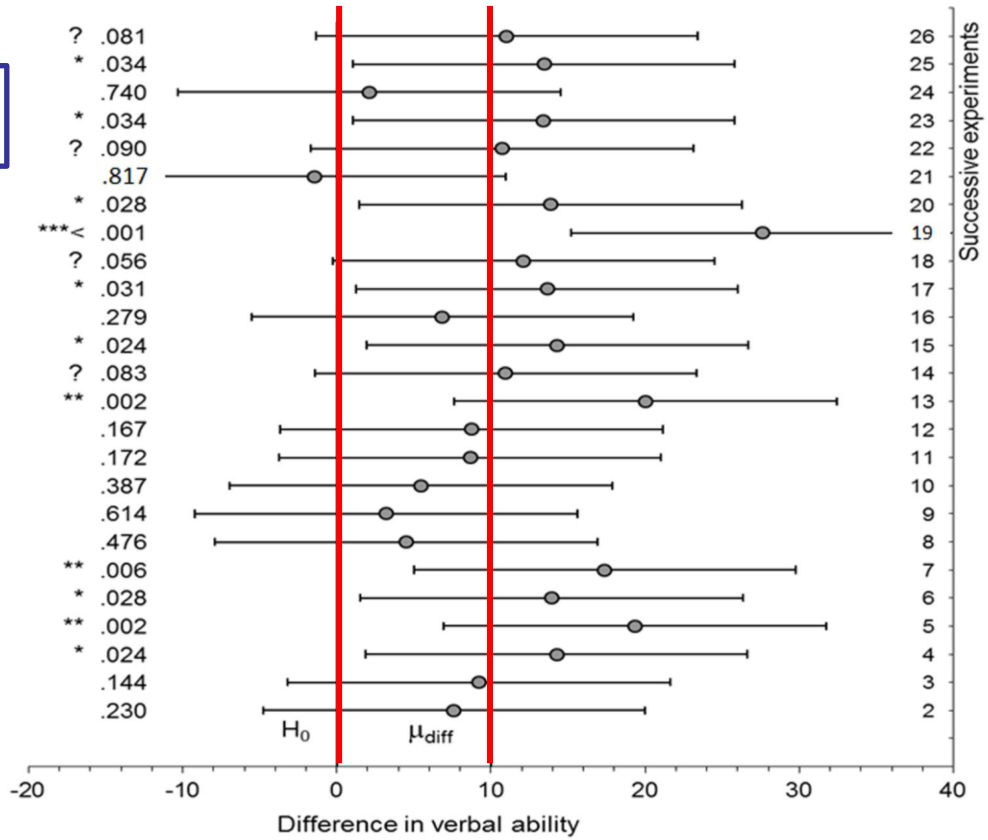
Four principles of inference by intervals (Dienes, 2014)

1. Interval contained in the null region → accept the null region hypothesis
2. Interval outside of the null region → reject the null region hypothesis
3. Upper limit of the interval is below the upper limit of the null region hypothesis → reject positive difference
4. Interval contains both null and theoretically interesting values → data are insensitive



Solution 2: Use confidence intervals to determine data insensitivity

Null region 0-10



Solution 3: Bayes Factors

- Bayes Factors (named after Thomas Bayes 1701-1761) Indicate the relative strength of evidence for two theories

$$Bayes\ factor = \frac{likelihood\ of\ data\ given\ H_1}{likelihood\ of\ data\ given\ H_0} = \frac{P(D|H_1)}{P(D|H_0)}$$

Solution 3: Bayes Factors

- **Interpretation:** the data are B times more likely under the alternative than under the null

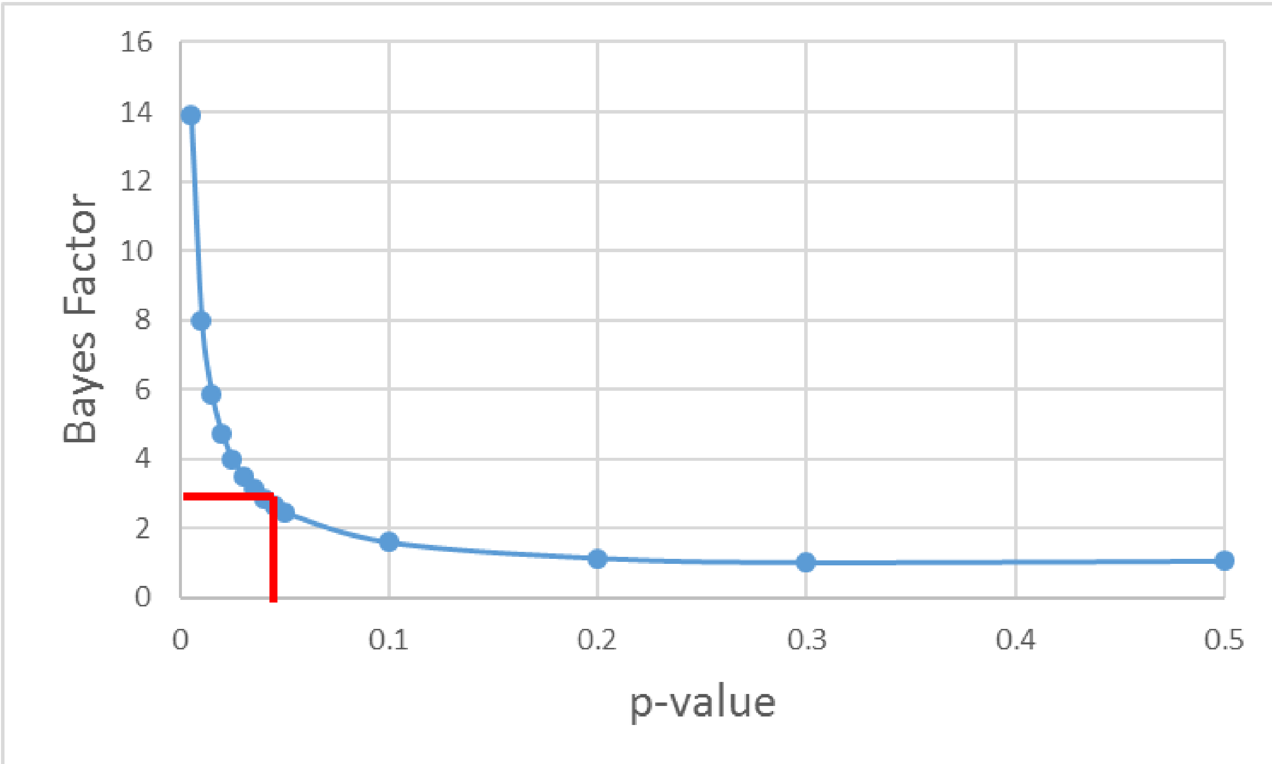
- B can range from 0 to ∞ and there are conventional cut-offs (Jeffreys et al, 1961; Dienes, 2014)
  - >3 evidence for the alternative hypothesis
  - <1/3rd evidence for the null hypothesis
  - >1/3rd and <3 data are insensitive

Calculating a Bayes Factor

- Many software packages (e.g. R)
- Online calculators (e.g. Zoltan Dienes (http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm)
- Bayes Factor bound

Bayes Factor Bound

- The *largest* Bayes factor in favour of H_1 that is possible (under reasonable assumptions) (Sellke, Bayarri, & Berger, 2001 and Vovk, 1993).



Online calculator (Dienes)

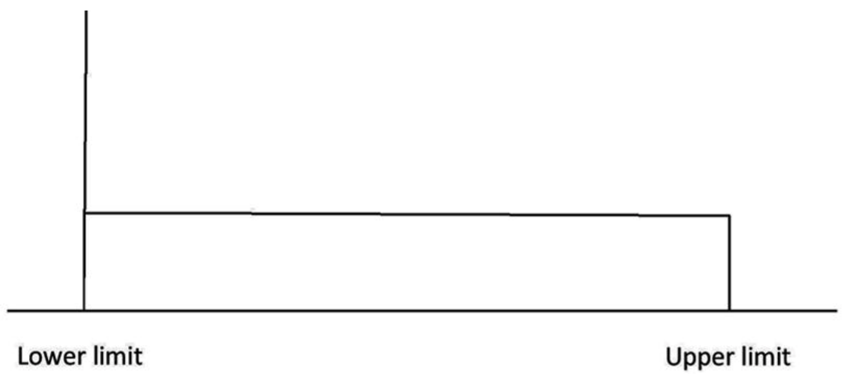
1. Published effect size
2. Standard error of the published parameter
3. Specify the effects which are consistent with your theory
   - Maximum plausible effect
   - Plausible predicted effect
4. Choose your distribution → normal, half-normal or uniform

- NOTE: Sampling distribution of the parameter estimate is distributed normally → log odds instead of odds ratios
- Specific to that calculator and not to Bayes generally

Calculating a Bayes Factor

If you can specify a maximum plausible effect

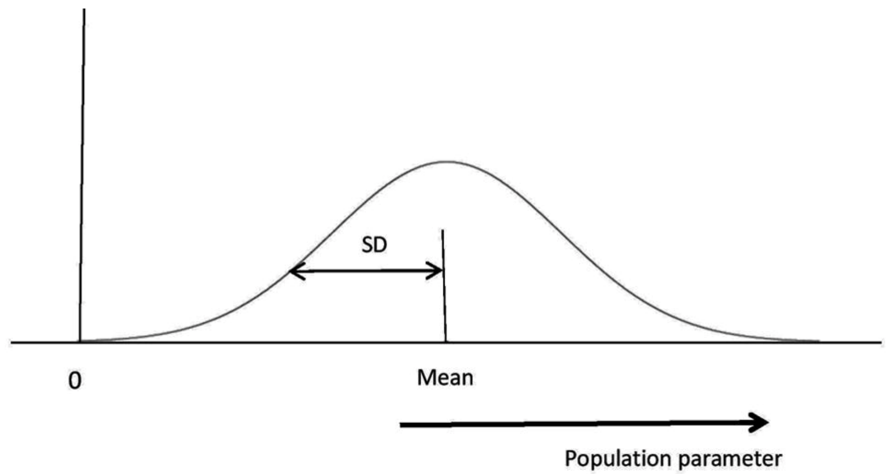
- ‘Uniform distribution’
  - Between 0 (or a minimally clinically significant value) and a plausible upper bound
  - Useful when there are constraints on measurements (e.g. Likert scale)



Calculating a Bayes Factor

If you can specify a plausible predicted effect P and make a non-directional prediction

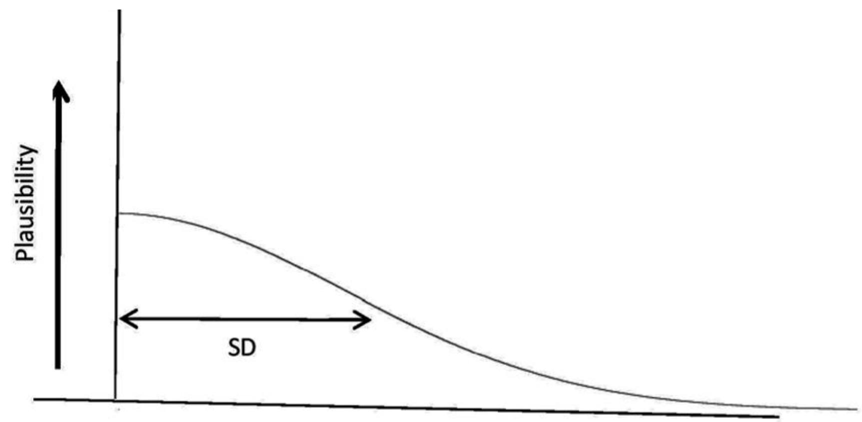
- ‘Normal distribution’
  - Population parameter values close to the mean are more plausible than others
  - SD default is P/2



Calculating a Bayes Factor

If you can specify a plausible predicted effect P and make a directional prediction [Most conservative → default]

- 'Half normal distribution'
  - Peak at 0 (no effect) with values close to 0 being plausible
  - SD is typically estimated using the effect size
  - Population values less than 0 are ruled out



Calculating a Bayes Factor - example

- Okuyemi et al (2013) motivational interviewing (MI) counselling plus nicotine patch versus nicotine patch
  - **Outcome:** verified seven-day abstinence rates
  - **Results**: week 26 non-significant difference (OR 1.33; 95% CI=0.88, 2.02; p= 0.17).
  - **Conclusion**: “Adding motivational interviewing counselling to nicotine patch did not significantly increase smoking rate at 26-week follow-up for homeless smokers”.

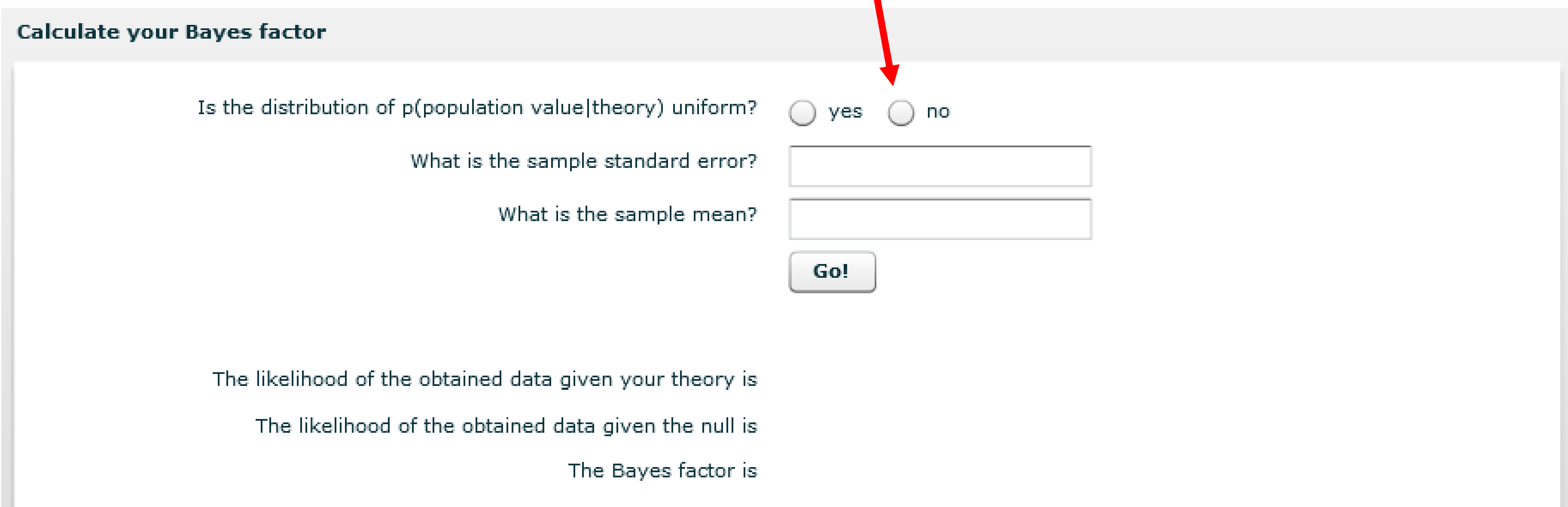
Calculating a Bayes Factor - example

- Transform odds ratio and SE to natural logarithmic scale
  - LN(1.33)= 0.29 (2 dp)
  - [LN(2.02)-LN(0.88)]/3.92= 0.21 (2 dp)

- Choose the half-normal distribution
  - Meta-analysis of the use of MI for smoking cessation (Hettema et al, 2010)
    - OR for long-term follow-up =1.35 (log odds ratio of 0.30)

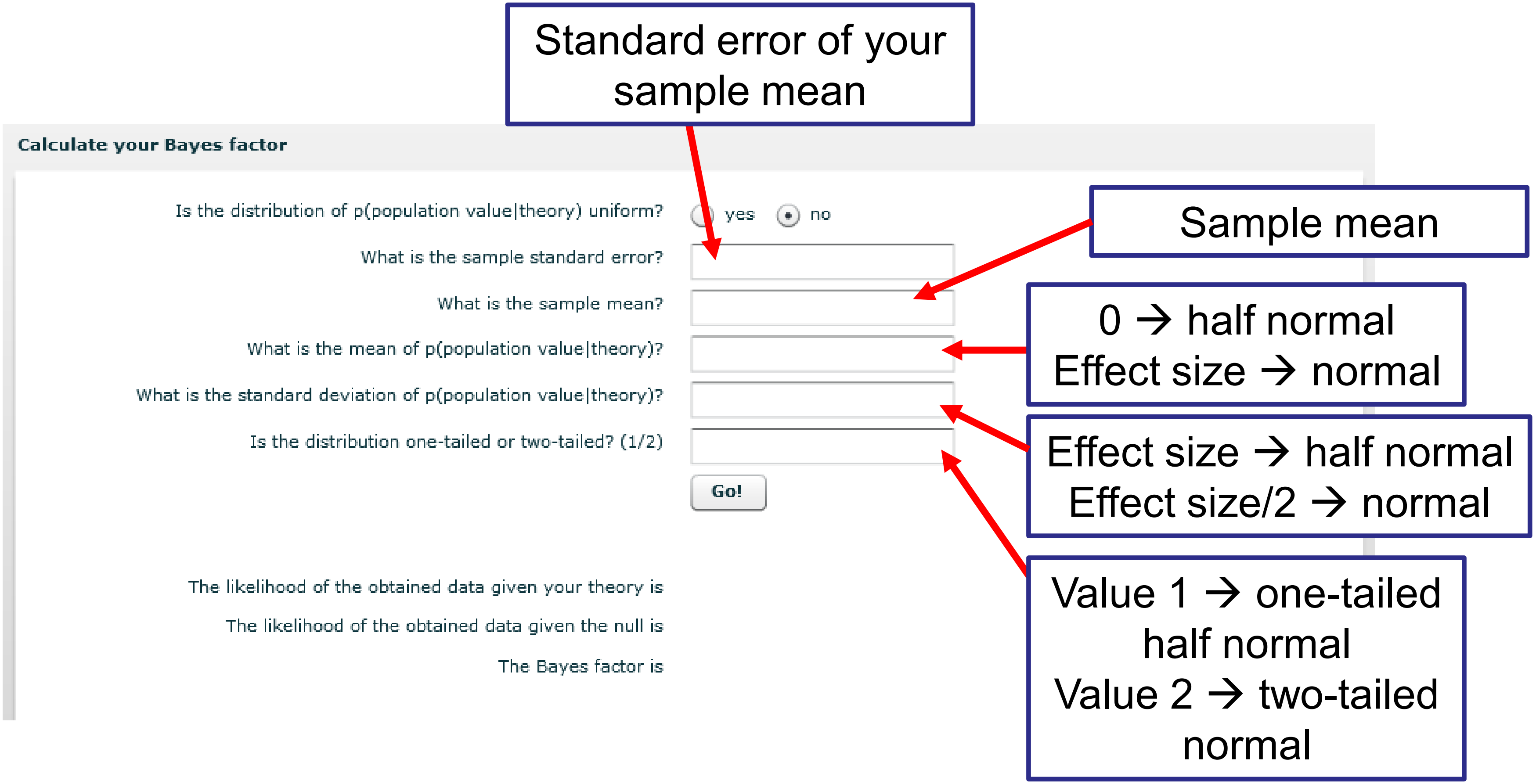
Calculating a Bayes Factor - example

- First mark the box ‘no’ next to ‘Is the distribution of p(population value|theory) uniform?’



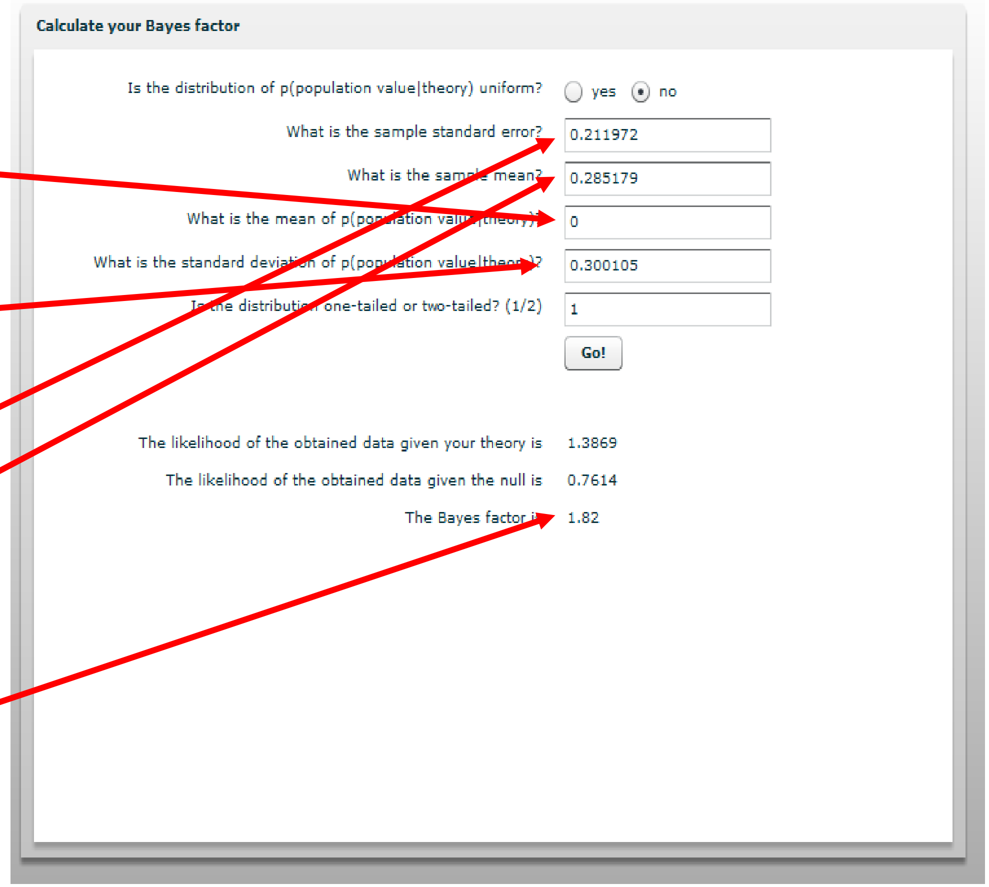
Calculating a Bayes Factor - example

- You will then see a new screen with additional boxes



Calculating a Bayes Factor - example

- We set mean to 0
- SD to our plausible expected value
- We must also enter the standard error and mean of our sample
- Bayes Factor = 1.82
  - The data are ‘insensitive’



Do Bayes factors aid interpretation?

- Review of RCTs reported in Addiction between Jan and June 2013 (Beard et al, 2016)

- 75 effect sizes and their standard errors were extracted from 12 trials
  - 73% (n=55) were non-significant ($p>0.05$)
  - 22% (n=20) were significant ($p<0.05$)

- Bayes Factor was calculated using a population effect derived from previous research

Do Bayes factors aid interpretation?

- 76.4% of non-significant findings had Bayes Factors between 1/3rd and 3 → data insensitive

- 20% of non-significant findings had Bayes Factors <1/3rd → support for the null hypothesis

Authors either decided not to discuss results where P > 0.05, to report them as non-significant and/or to state that no association was found

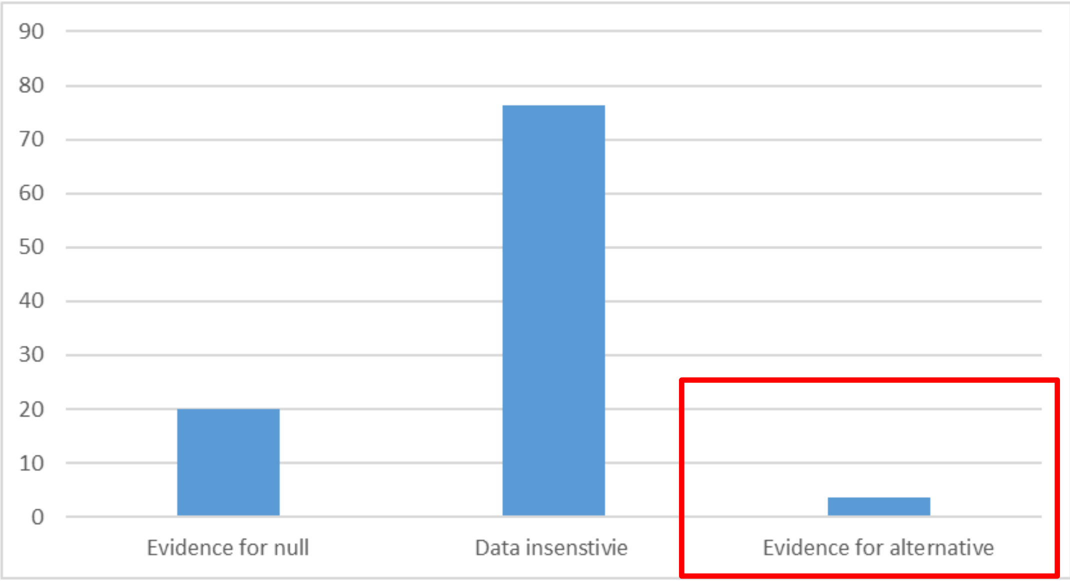


Figure 1: Conclusions of Bayes Factors for non-significant findings

Do Bayes factors aid interpretation?

Table 1: Jeffreys' Bayes Factor cut-offs

Evidence for alternative hypothesis

Bayes Factor	Interpretation
>100	Extreme evidence for the experimental hypothesis
30-100	Very strong evidence for the experimental hypothesis
10-30	Strong evidence for the experimental hypothesis
3-10	Moderate evidence for the experimental hypothesis
1-3	Anecdotal evidence for the experimental hypothesis
1	No evidence
1/3-1	Anecdotal evidence for the null hypothesis
1/3-1/10	Moderate evidence for the null hypothesis
1/10-1/30	Strong evidence for the null hypothesis
1/30-1/100	Very strong evidence for the null hypothesis
<1/100	Extreme evidence for the null hypothesis

Data insensitive

Evidence for the null hypothesis

Do Bayes factors aid interpretation?

Table 1: Jeffreys' Bayes Factor cut-offs

Bayes Factor	Interpretation
>100	Extreme evidence for the experimental hypothesis
30-100	Very strong evidence for the experimental hypothesis
10-30	Strong evidence for the experimental hypothesis
3-10	Moderate evidence for the experimental hypothesis
1-3	Anecdotal evidence for the experimental hypothesis
1	No evidence
1/3-1	Anecdotal evidence for the null hypothesis
1/3-1/10	Moderate evidence for the null hypothesis
1/10-1/30	Strong evidence for the null hypothesis
1/30-1/100	Very strong evidence for the null hypothesis
<1/100	Extreme evidence for the null hypothesis

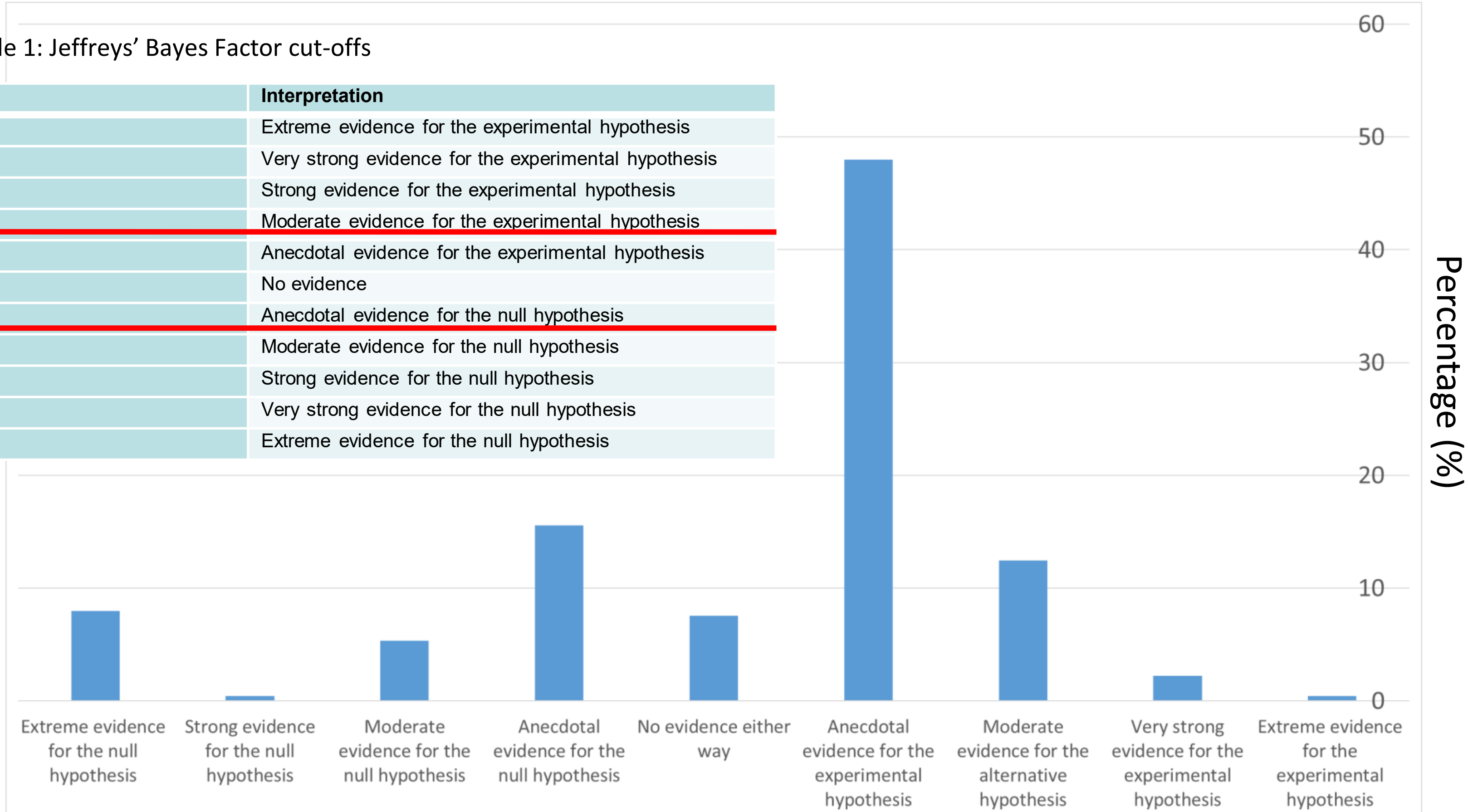


Figure 2: Conclusions of Bayes Factors for significant and non-significant findings

Conclusion

- A sensitive result is never guaranteed with high power
  - Power is helpful in finding rough No. of observations needed

- Sensitivity can be guaranteed with **intervals and Bayes factors**
  - Collect data until:
    - a) The interval is smaller than the null region and is either in or out of the null region
    - b) Until the Bayes factor is either >3 or <1/3rd
  - Bad practice to not have fixed stopping rules in Frequentist statistics

Conclusion

- Bayes Factors are most sensitive to the **maximum**, which could be specified reasonably objectively.
- Inference by intervals is completely dependent on specification of the **minimum**, which is often hard to specify objectively.

Conclusion

- $p>0.05$ and $B > 0.33$ → avoid use of terms such as 'no difference' or 'lack of association'
- $p>0.05$ and $B < 0.33$ → can use terms such as 'no difference' or 'lack of association'
- If you do not calculate a B → 'The findings were inconclusive as to whether or not a difference/association was present'
- Should pre-register analysis plan with effect size (e.g., Open Science Framework)

Things to note

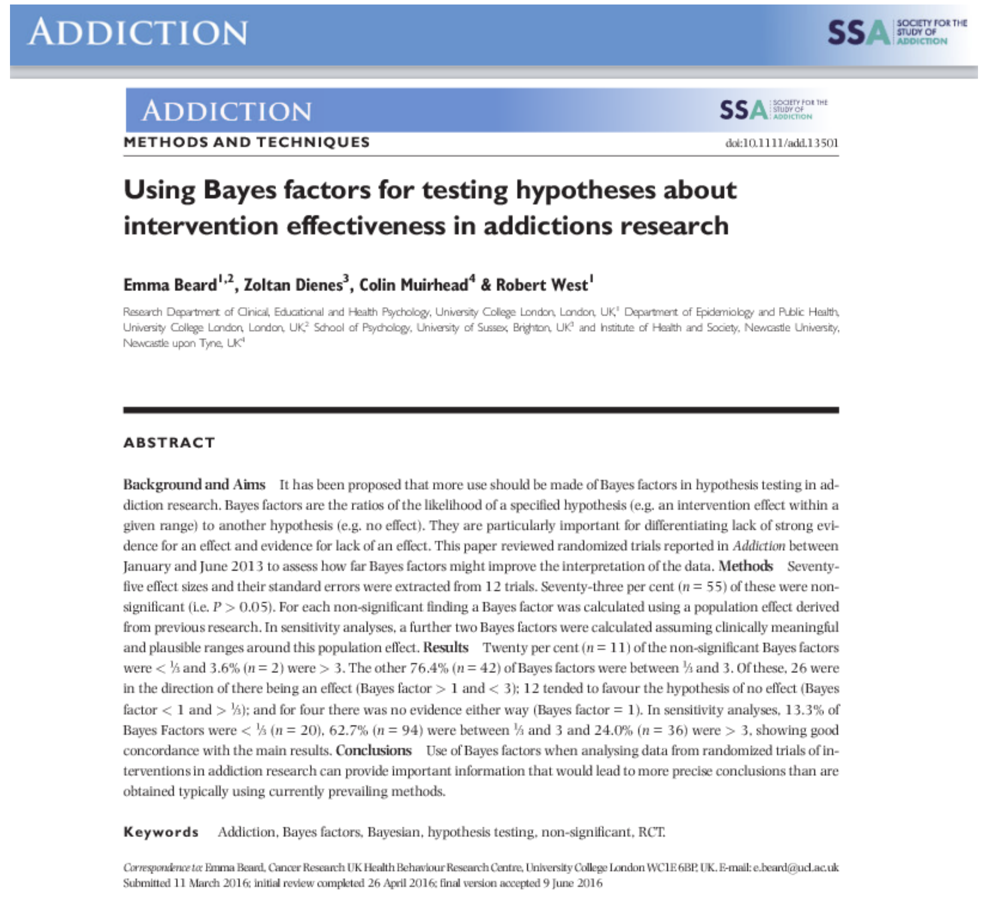
- Bayes can be criticized for being too subjective as it relies on "priors"
  - *Posterior odds* = *BF* × *prior odds*
  - We have lifted the Bayes factor out of full Bayesian schema → represents a measure of strength of evidence

- There are many ways of being a Bayesian and they are not exclusive (e.g. Kruschke (2010) & Lee and Wagenmakers (2014))
  - Aim here is to make the minimal changes to current practice

Thank you

For further details:
e.beard@ucl.ac.uk

Special thanks to my co-authors:
Professor West, Professor Dienes and Dr Muirhead



ADDICTION

ADDICTION

METHODS AND TECHNIQUES doi:10.1111/add.13501

Using Bayes factors for testing hypotheses about intervention effectiveness in addictions research

Emma Beard[1,2], Zoltan Dienes[3], Colin Muirhead[4] & Robert West[1]

Research Department of Clinical, Educational and Health Psychology, University College London, London, UK,[1] Department of Epidemiology and Public Health, University College London, London, UK,[2] School of Psychology, University of Sussex, Brighton, UK[3] and Institute of Health and Society, Newcastle University, Newcastle upon Tyne, UK[4]

ABSTRACT

Background and Aims It has been proposed that more use should be made of Bayes factors in hypothesis testing in addiction research. Bayes factors are the ratios of the likelihood of a specified hypothesis (e.g. an intervention effect within a given range) to another hypothesis (e.g. no effect). They are particularly important for differentiating lack of strong evidence for an effect and evidence for lack of an effect. This paper reviewed randomized trials reported in *Addiction* between January and June 2013 to assess how far Bayes factors might improve the interpretation of the data. **Methods** Seventy-five effect sizes and their standard errors were extracted from 12 trials. Seventy-three per cent ($n = 55$) of these were non-significant (i.e. $P > 0.05$). For each non-significant finding a Bayes factor was calculated using a population effect derived from previous research. In sensitivity analyses, a further two Bayes factors were calculated assuming clinically meaningful and plausible ranges around this population effect. **Results** Twenty per cent ($n = 11$) of the non-significant Bayes factors were < ⅓ and 3.6% ($n = 2$) were > 3. The other 76.4% ($n = 42$) of Bayes factors were between ⅓ and 3. Of these, 26 were in the direction of there being an effect (Bayes factor > 1 and < 3); 12 tended to favour the hypothesis of no effect (Bayes factor < 1 and > ⅓); and for four there was no evidence either way (Bayes factor = 1). In sensitivity analyses, 13.3% of Bayes Factors were < ⅓ ($n = 20$), 62.7% ($n = 94$) were between ⅓ and 3 and 24.0% ($n = 36$) were > 3, showing good concordance with the main results. **Conclusions** Use of Bayes factors when analysing data from randomized trials of interventions in addiction research can provide important information that would lead to more precise conclusions than are obtained typically using currently prevailing methods.

Keywords Addiction, Bayes factors, Bayesian, hypothesis testing, non-significant, RCT.

Correspondence to: Emma Beard, Cancer Research UK Health Behaviour Research Centre, University College London WC1E 6BP, UK. E-mail: e.beard@ucl.ac.uk
Submitted 11 March 2016; initial review completed 26 April 2016; final version accepted 9 June 2016