Prior Exposure 1: Bayes for beginners: **Introducing Bayesian data analysis**

Thom Baguley and Mark Andrews
Nottingham Trent University

Reprise: Bayesian data analysis in three simple steps

A Bayesian analysis involves three things:

1. A probability model for the data (the likelihood)

Reprise: Bayesian data analysis in three simple steps

A Bayesian analysis involves three things:

1. A probability model for the data (the likelihood)
2. A probability model of relevant prior information (the prior)

Reprise: Bayesian data analysis in three simple steps

A Bayesian analysis involves three things:

1. A probability model for the data (the likelihood)
2. A probability model of relevant prior information (the prior)
3. A probability model that *combines* the data and the prior information (the posterior)

Reprise: Bayesian data analysis in three simple steps

A Bayesian analysis involves three things:

1. A probability model for the data (the likelihood)
2. A probability model of relevant prior information (the prior)
3. A probability model that *combines* the data and the prior information (the posterior)

... this last step uses Bayes' theorem

Bayes' theorem

simple probability

$$\overbrace{Pr\left(A|B\right)}^{\text{Posterior}} = \frac{\overbrace{Pr\left(B|A\right)}^{\text{Evidence}} \quad \overbrace{Pr\left(A\right)}^{\text{Base rate (Prior)}}}{\underbrace{Pr\left(B\right)}_{\text{Probability of evidential event}}}$$

Bayes' theorem

Murder or SIDS: the Sally Clark case

What is the more likely cause of a cot death: Murder or SIDS?

$$\overbrace{Pr(A|B)}^{\text{Is it murder?}} = \frac{\overbrace{Pr(B|A)}^{\text{Prob. of death given murder}} \quad \overbrace{Pr(A)}^{\text{Prob. of murder}}}{\underbrace{Pr(B)}_{\text{Prob. of cot death}}}$$

Bayes' theorem

Murder or SIDS: the Sally Clark case

Trivially the probability of cot death given a murder is 1

$$\overbrace{Pr\left(A\,|\,B\right)}^{\text{Is it murder?}} = \frac{\overbrace{1}^{\text{Prob. of death given murder}}\;\overbrace{Pr\left(A\right)}^{\text{Prob. of murder}}}{\underbrace{Pr\left(B\right)}_{\text{Prob. of cot death}}}$$

Bayes' theorem

Murder or SIDS: the Sally Clark case

Homicide rates for infants are about 1 in 21,700 (Hill, 2004)

$$\overbrace{Pr(A|B)}^{\text{Is it murder?}} = \frac{\overbrace{1}^{\text{Prob. of death given murder}} \times \overbrace{1/21700}^{\text{Prob. of murder}}}{\underbrace{Pr(B)}_{\text{Prob. of cot death}}}$$

Bayes' theorem

Murder or SIDS: the Sally Clark case

Cot death rates are about 1 in 1,300 (Hill, 2004)

$$\overbrace{Pr(A|B)}^{\text{Is it murder?}} = \frac{\overbrace{1}^{\text{Prob. of death given murder}} \times \overbrace{1/21700}^{\text{Prob. of murder}}}{\underbrace{1/1300}_{\text{Prob. of cot death}}}$$

Bayes' theorem

Murder or SIDS: the Sally Clark case

So the probability that a cot death is murder is around 6%

$$\overbrace{1/16.7}^{\text{Is it murder?}} = \frac{\overbrace{1}^{\text{Prob. of death given murder}} \times \overbrace{1/21700}^{\text{Prob. of murder}}}{\underbrace{1/1300}_{\text{Prob. of cot death}}}$$

Bayes' theorem

probability distributions

... the same equation can be generalised to probability distributions (here Ω is the full set of parameters in the model including nuisance parameters)

$$\overbrace{\mathrm{P}(\theta|\mathcal{D},\Omega)}^{\text{Posterior}} = \frac{\overbrace{\mathrm{P}(\mathcal{D}|\theta)}^{\text{Likelihood}}\ \overbrace{\mathrm{P}(\theta|\Omega)}^{\text{Prior}}}{\underbrace{\int \mathrm{P}(\mathcal{D}|\theta)\mathrm{P}(\theta|\Omega)\,d\theta}_{\text{Marginal probability of data}}}$$

Marginal probability

... getting a posterior distribution for θ requires integrating out other parameters in the model

$$\int \mathrm{P}(\mathcal{D}|\theta)\mathrm{P}(\theta|\Omega)\,d\theta$$

Getting an analytic solution may be difficult or impossible

... but can also be very simple for some well-known problems

Normal distributions with known variance

$$\overbrace{\mathrm{P}(\mathcal{D}|\mu)}^{\text{Likelihood}} \sim N\left(\hat{\mu}, \hat{\sigma}^2_{\hat{\mu}}\right)$$

Normal distributions with known variance

$$\overbrace{\mathrm{P}(\mathcal{D}|\mu)}^{\text{Likelihood}} \sim N\left(\hat{\mu}, \hat{\sigma}^2_{\hat{\mu}}\right)$$

$$\overbrace{\mathrm{P}(\mu|\sigma^2)}^{\text{Prior}} \sim N\left(\mu_{prior}, \sigma^2_{prior}\right)$$

Normal distributions with known variance

$$\overbrace{\mathrm{P}(\mathcal{D}|\mu)}^{\text{Likelihood}} \sim N\left(\hat{\mu}, \hat{\sigma}^2_{\hat{\mu}}\right)$$

$$\overbrace{\mathrm{P}(\mu|\sigma^2)}^{\text{Prior}} \sim N\left(\mu_{prior}, \sigma^2_{prior}\right)$$

... together imply:

$$\overbrace{\mathrm{P}(\mu|\mathcal{D}, \sigma^2)}^{\text{Posterior}} \sim N\left(\mu_{post}, \sigma^2_{post}\right)$$

Normal distributions with known variance

Posterior variance

$$\sigma^2_{post} = \left(\frac{1}{\hat{\sigma}^2_{\hat{\mu}}} + \frac{1}{\sigma^2_{prior}} \right)^{-1}$$

Normal distributions with known variance

Posterior variance

$$\sigma^2_{post} = \left(\overbrace{\frac{1}{\hat{\sigma}^2_{\hat{\mu}}}}^{\text{precision}} + \overbrace{\frac{1}{\sigma^2_{prior}}}^{\text{precision}} \right)^{-1}$$

Normal distributions with known variance

Posterior mean

The posterior mean is, in effect, a combination of the observed mean and the prior weighted by their relative precision

$$\mu_{post} = \left(\frac{\sigma^2_{post}}{\hat{\sigma}^2_{\hat{\mu}}}\right)\hat{\mu} + \left(\frac{\sigma^2_{post}}{\sigma^2_{prior}}\right)\mu_{prior}$$

Normal distributions with known variance

Posterior mean

The posterior mean is, in effect, a combination of the observed mean and the prior weighted by their relative precision

$$\mu_{post} = \left(\frac{\sigma^2_{post}}{\hat{\sigma}^2_{\hat{\mu}}} \right) \hat{\mu} + \left(\frac{\sigma^2_{post}}{\sigma^2_{prior}} \right) \mu_{prior}$$

... this weighting is sometimes termed a *fully automatic Occam's razor* (Spegelhalter & Smith, 1980)

Normal distributions with known variance

Posterior probability intervals

Obtaining an interval estimate from the posterior distribution is now trivial

This is often known as a *credibility interval* ...

Normal distributions with known variance

Posterior probability intervals

Obtaining an interval estimate from the posterior distribution is now trivial

This is often known as a *credibility interval* ...

... but I prefer the more neutral *posterior probability interval* (or *probability interval* for short)

Normal distributions with known variance

Posterior probability intervals

Obtaining an interval estimate from the posterior distribution is now trivial

This is often known as a *credibility interval* ...

... but I prefer the more neutral *posterior probability interval* (or *probability interval* for short)

e.g., a 95% probability interval takes the form:

$$\mu_{post} \pm z_{.975}\sigma_{post} \approx \mu_{post} \pm 1.96\sigma_{post}$$

Normal distribution example

Standardized mean difference (Cohen's d)

A standardized mean difference is an example of a a normally distributed statistic with a known variance of 1

It measures the *detectability* or *discriminability* of an effect (not its size!)

Normal distribution example

Standardized mean difference (Cohen's d)

A standardized mean difference is an example of a a normally distributed statistic with a known variance of 1

It measures the *detectability* or *discriminability* of an effect (not its size!)

... this means we could use prior information to help counteract the upward bias of a published effect size statistic

Normal distribution example

Standardized mean difference (Cohen's *d*)

A standardized mean difference is an example of a a normally distributed statistic with a known variance of 1

It measures the *detectability* or *discriminability* of an effect (not its size!)

... this means we could use prior information to help counteract the upward bias of a published effect size statistic

This bias arises from multiple sources - notably from $p < .05$ acting as a filter that removes small effect sizes

Normal distribution example

Does a single exposure to an US flag influence voting intention?

Carter et al. (2011) report a priming study in which a single exposure to a US flag changes voting intentions (making them more Republican)

They report an 'effect size' of over a quarter of a standard deviation:

$$\hat{\mu}_E - \hat{\mu}_C = 0.142, \quad t_{181} = 2.02, \quad p = .04 \quad d = 0.298$$

Normal distribution example

Does a single exposure to an US flag influence voting intention?

Carter et al. (2011) report a priming study in which a single exposure to a US flag changes voting intentions (making them more Republican)

They report an 'effect size' of over a quarter of a standard deviation:

$$\hat{\mu}_E - \hat{\mu}_C = 0.142, \quad t_{181} = 2.02, \quad p = .04 \quad d = 0.298$$

... this struck many people as unrealistically large (being associated with a 10.7% swing)

Normal distribution example

Does a single exposure to an US flag influence voting intention?

Carter et al. (2011) report a priming study in which a single exposure to a US flag changes voting intentions (making them more Republican)

They report an 'effect size' of over a quarter of a standard deviation:

$$\hat{\mu}_E - \hat{\mu}_C = 0.142, \quad t_{181} = 2.02, \quad p = .04 \quad d = 0.298$$

... this struck many people as unrealistically large (being associated with a 10.7% swing)

N.B. Cohen's *d* is really a family of statistics. The statistic here is probably Hedges' g

Normal distribution example

The likelihood

The standard error of g is $\approx \left[(n_1+n_2)/(n_1 n_2) + \left(g^2/(2N-4)\right)\right]^{1/2}$

Normal distribution example

The likelihood

The standard error of g is $\approx \left[(n_1 + n_2)/(n_1 n_2) + \left(g^2/(2N-4)\right)\right]^{1/2}$

With $N = 183$ and roughly equal n per group gives g a standard error of 0.149

Normal distribution example

The likelihood

The standard error of g is $\approx \left[(n_1+n_2)/(n_1 n_2)+\left(g^2/(2N-4)\right)\right]^{1/2}$

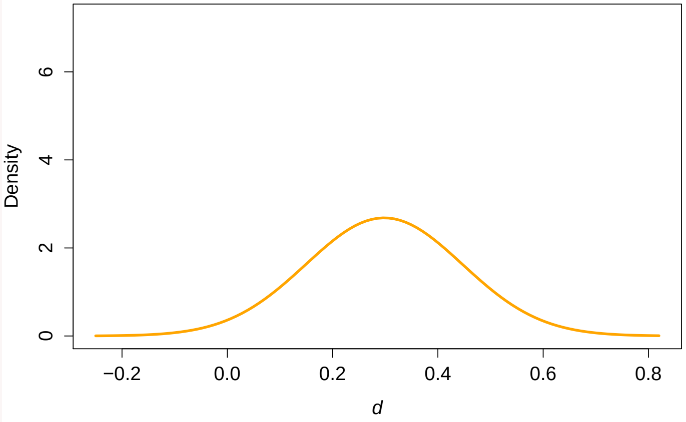
With $N = 183$ and roughly equal n per group gives g a standard error of 0.149

The likelihood (our probability model for the data) is therefore:

$$\mathrm{P}(\mathcal{D}|\mu) \sim N\left(0.298, 0.149^2\right)$$

Normal distribution example

Graphing the likelihood



Normal distribution example

The prior

The prior is our model of relevant prior information

This could be

Normal distribution example

The prior

The prior is our model of relevant prior information

This could be

- uncertainty in our knowledge about g, or

Normal distribution example

The prior

The prior is our model of relevant prior information

This could be

- uncertainty in our knowledge about g, or
- uncertainty in our belief about g

Normal distribution example

The prior

The prior is our model of relevant prior information

This could be

- uncertainty in our knowledge about g, or
- uncertainty in our belief about g

How might we determine our prior here?

Normal distribution example

The prior

The prior is our model of relevant prior information

This could be

- uncertainty in our knowledge about g, or
- uncertainty in our belief about g

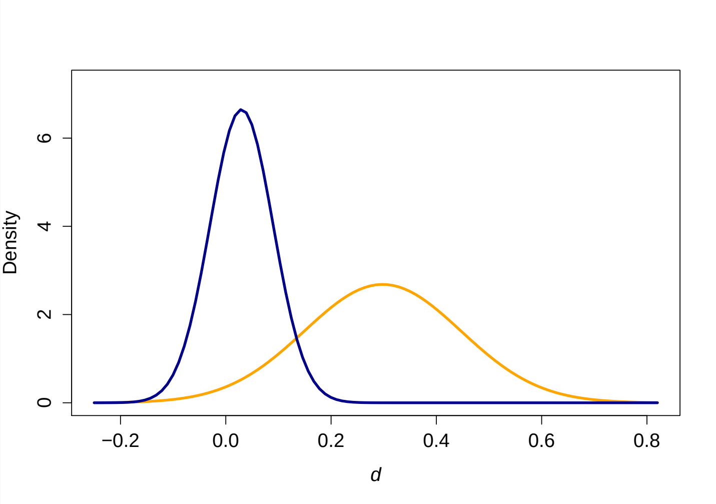
How might we determine our prior here?

A generous prior for the true effect size might be a 1% swing with a likely range of say $\pm 2\%$
If $g = 0.3$ corresponds to a 10% swing then this suggests the following normal prior:

$$\mathrm{P}(\mu|\sigma^2) \sim N\left(0.03, 0.06^2\right)$$

Normal distribution example

Adding the prior



Normal distribution example

Posterior mean and standard deviation

$$\sigma_{post} = \sqrt{\left(\frac{1}{0.149^2} + \frac{1}{0.06^2}\right)^{-1}} \approx 0.056$$

$$\mu_{post} = \left(\frac{0.056^2}{0.149^2}\right) 0.298 + \left(\frac{0.056^2}{0.06^2}\right) 0.03$$

$$= 0.14 \times 0.298 + .86 \times 0.03 \approx 0.067$$

Normal distribution example

Posterior mean and standard deviation

$$\sigma_{post} = \sqrt{\left(\frac{1}{0.149^2} + \frac{1}{0.06^2}\right)^{-1}} \approx 0.056$$

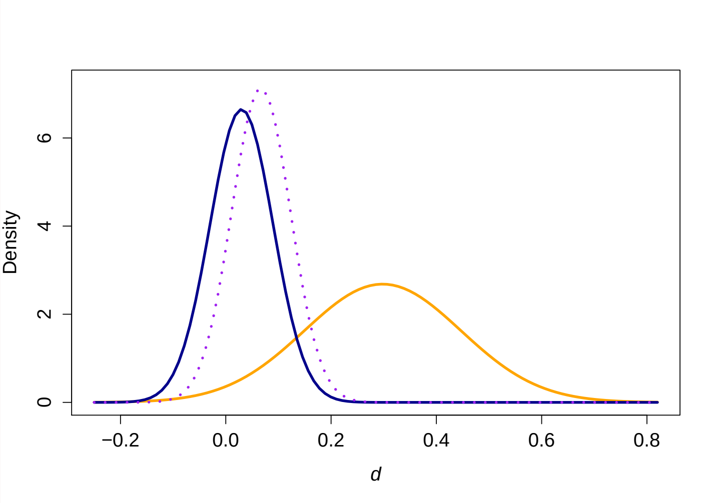
$$\mu_{post} = \left(\frac{0.056^2}{0.149^2}\right) 0.298 + \left(\frac{0.056^2}{0.06^2}\right) 0.03$$

$$= 0.14 \times 0.298 + .86 \times 0.03 \approx 0.067$$

$$\mathrm{P}(\mu|\mathcal{D}, \sigma^2) \sim N\left(0.067, 0.056^2\right)$$

Normal distribution example

Calculating the posterior



Normal distribution example

Evaluating the model

A 95% probability interval for the posterior distribution of the flag effect is:

$$0.067 \pm 1.96 \times 0.056 = [-0.042, 0.177]$$

Normal distribution example

Evaluating the model

A 95% probability interval for the posterior distribution of the flag effect is:

$$0.067 \pm 1.96 \times 0.056 = [-0.042, 0.177]$$

Not only has my prior has greatly reduced the estimate of the effect but they effect could now plausibly be a 1.5% swing away from or a 6% swing to the Republicans

Normal distribution example

Evaluating the model

A 95% probability interval for the posterior distribution of the flag effect is:

$$0.067 \pm 1.96 \times 0.056 = [-0.042, 0.177]$$

Not only has my prior has greatly reduced the estimate of the effect but they effect could now plausibly be a 1.5% swing away from or a 6% swing to the Republicans

... but is my prior a good one?

Selecting a prior

Subjective versus objective priors

One option is to seek 'objective' prior distributions that can be used in situations where judgemental input is supposed to be minimized, such as in scientific publications. [...]

Selecting a prior

Subjective versus objective priors

> *One option is to seek 'objective' prior distributions that can be used in situations where judgemental input is supposed to be minimized, such as in scientific publications. [...]*
>
> *A complete alternative is the fully subjectivist position, which compels one to elicit priors on all parameters based on the personal judgement of appropriate individuals.*

Spiegelhalter & Rice (2009)

Subjective Bayes

Eliciting a prior

A key task for a subjective Bayesian is to elicit a prior distribution:

> *... successful elicitation faithfully represents the opinion of the person being elicited. It is not necessarily "true" in some objectivistic sense, and cannot be judged that way. We see elicitation as simply part of the process of statistical modeling.*

Garthwaite et al. (2005)

Subjective Bayes

Typical characteristics of subjective priors

- priors represent personal judgement or belief

Subjective Bayes

Typical characteristics of subjective priors

- priors represent personal judgement or belief
- priors are generally elicited from 'experts'

Subjective Bayes

Typical characteristics of subjective priors

- priors represent personal judgement or belief
- priors are generally elicited from 'experts'
- different people can legitimately have different priors

Subjective Bayes

Typical characteristics of subjective priors

- priors represent personal judgement or belief
- priors are generally elicited from 'experts'
- different people can legitimately have different priors
- priors can be elicited from individuals or groups

Subjective Bayes

Typical characteristics of subjective priors

- priors represent personal judgement or belief
- priors are generally elicited from 'experts'
- different people can legitimately have different priors
- priors can be elicited from individuals or groups
- priors are open to scrutiny

Subjective Bayes

Typical characteristics of subjective priors

- priors represent personal judgement or belief
- priors are generally elicited from 'experts'
- different people can legitimately have different priors
- priors can be elicited from individuals or groups
- priors are open to scrutiny
- ... but models don't (in general) need to be checked

Subjective Bayes

Typical characteristics of subjective priors

- priors represent personal judgement or belief
- priors are generally elicited from 'experts'
- different people can legitimately have different priors
- priors can be elicited from individuals or groups
- priors are open to scrutiny
- ... but models don't (in general) need to be checked

Objective Bayes

What is objectivity in Bayesian data analysis

... I take objectivity to mean that given the same data and the same assumptions regarding the model, different researchers will arrive at the same conclusions.

Wagenmakers (2007, Appendix)

Objective Bayes

Typical characteristics of objective priors

- priors do not reflect personal judgement or belief

Objective Bayes

Typical characteristics of objective priors

- priors do not reflect personal judgement or belief
- a prior is a probability model of known or agreed prior information

Objective Bayes

Typical characteristics of objective priors

- priors do not reflect personal judgement or belief
- a prior is a probability model of known or agreed prior information
- priors don't need to be elicited for each parameter

Objective Bayes

Typical characteristics of objective priors

- priors do not reflect personal judgement or belief
- a prior is a probability model of known or agreed prior information
- priors don't need to be elicited for each parameter
- priors can be viewed as default assumptions

Objective Bayes

Typical characteristics of objective priors

- priors do not reflect personal judgement or belief
- a prior is a probability model of known or agreed prior information
- priors don't need to be elicited for each parameter
- priors can be viewed as default assumptions
- models need to be checked

Objective Bayes

Typical characteristics of objective priors

- priors do not reflect personal judgement or belief
- a prior is a probability model of known or agreed prior information
- priors don't need to be elicited for each parameter
- priors can be viewed as default assumptions
- models need to be checked

That wretched prior

A viewpoint from subjective Bayes

The objectivity of the likelihood function is often an illusion as the choice of different models testifies. Objectivity is merely subjectivity when nearly everyone agrees.

Lindley (2004)

The objectivity of the likelihood function is often an illusion as the choice of different models testifies. Objectivity is merely subjectivity when nearly everyone agrees.

Lindley (2004)

That wretched prior

A viewpoint from objective Bayes

... it [objective Bayes] simultaneously achieves what should be a major goal of Bayesianism - ensuring that answers are conditional on the data actually obtained - while at the same time respecting the frequentist notion that the methodology must ensure success in repeated usage by scientists.

Berger (2006)

Reprise: Why Bayesian data analysis (BDA)?

- efficient and *coherent* use of available information

Reprise: Why Bayesian data analysis (BDA)?

- efficient and *coherent* use of available information
- The right tool for certain problems

  (e.g., realistic assessment of effect size)

Reprise: Why Bayesian data analysis (BDA)?

- efficient and *coherent* use of available information
- The right tool for certain problems

  (e.g., realistic assessment of effect size)

- less reliance on *ad hoc* procedures

  (e.g., adjustment of α)

Reprise: Why Bayesian data analysis (BDA)?

- efficient and *coherent* use of available information
- The right tool for certain problems

  (e.g., realistic assessment of effect size)

- less reliance on *ad hoc* procedures

  (e.g., adjustment of α)

- greater emphasis on transparency in modelling

Reprise: Why Bayesian data analysis (BDA)?

- efficient and *coherent* use of available information
- The right tool for certain problems

  (e.g., realistic assessment of effect size)

- less reliance on *ad hoc* procedures

  (e.g., adjustment of α)

- greater emphasis on transparency in modelling
- more flexibility in modelling complex, messy (real world) data sets