January 20, 2021

Advanced topics in Biostatistics 2020/2021: Introduction to Bayesian thinking

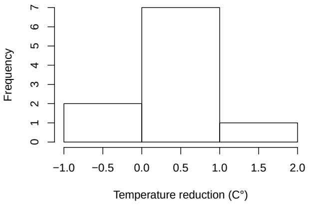
Silvia Calderazzo
Division of Biostatistics - C060
s.calderazzo@dkfz-heidelberg.de

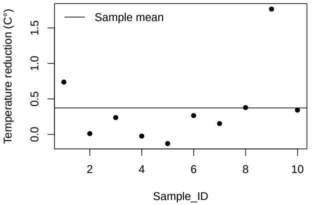
Why do we need Statistics in the first place

- Suppose we have collected a **set of measurements** from a population
- Measurements -> **random variability** (intrinsic variability / measurement error / model reduction)
- Often adequately **described by a probabilistic model**
- Aim: increase knowledge about the model
- Type of data models:
  - **Parametric**: infer the value of one or more parameters -> still need to choose a distribution (Normal, Poisson, Binomial...etc)
  - **Nonparametric**: more flexible but generally more complex/requires larger sample sizes

Example

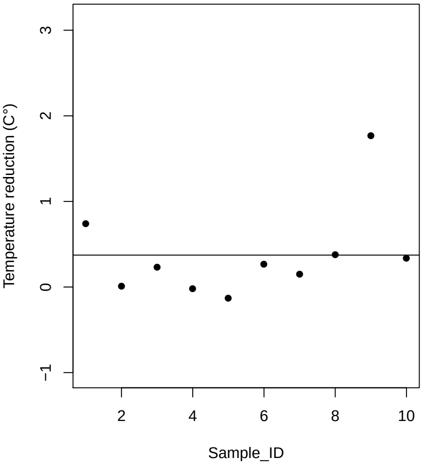
- Measurements: reduction in body temperature 1 hour after taking an anitpiretic drug for a (random, independent) sample of $n = 10$ individuals
- Sample mean: 0.37 C°
- Measurements are normally distributed (**assumption!**)
- A Normal distribution is uniquely identified by two parameters: mean (θ) and standard deviation (σ)
- Suppose we know $\sigma = 0.7$
- Aim: Inference about θ





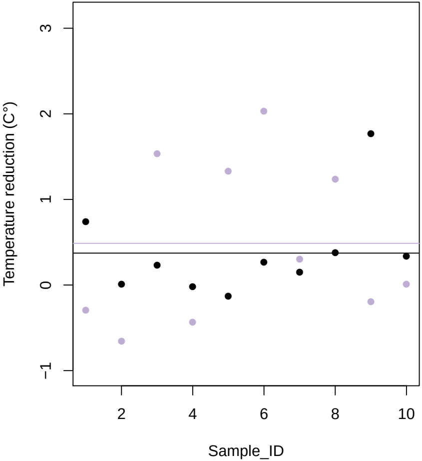
Back to our example

- Our 10 measurements have a mean of 0.37 C°
- This is one of many possible outcomes...



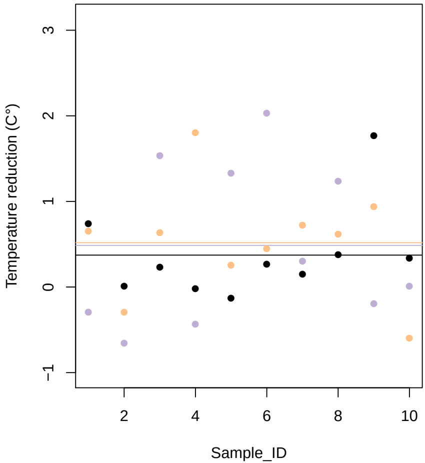
Back to our example

- This is one of many possible outcomes...



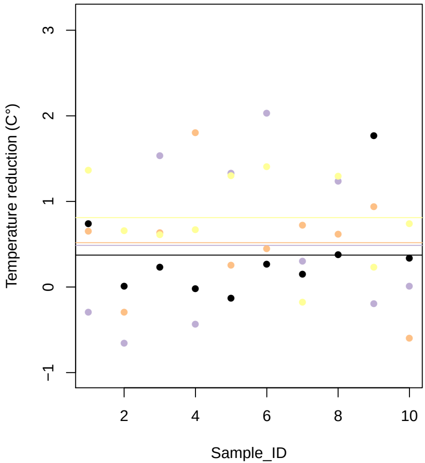
Back to our example

- This is one of many possible outcomes...



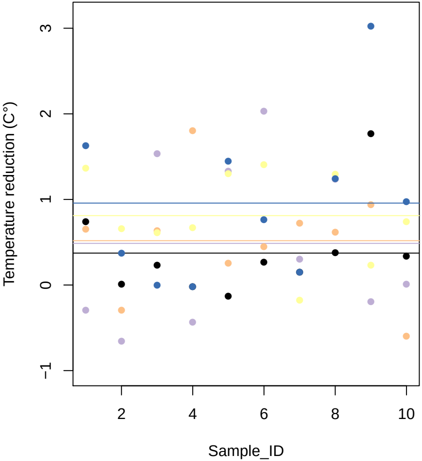
Back to our example

- This is one of many possible outcomes...



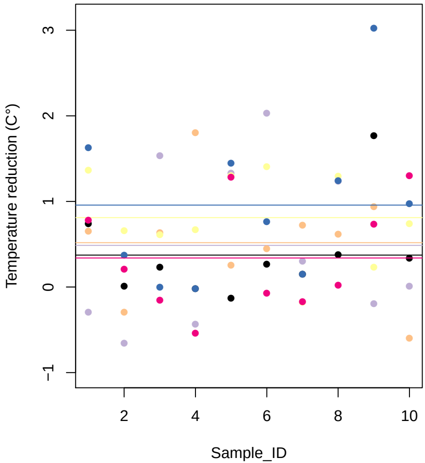
Back to our example

- This is one of many possible outcomes...



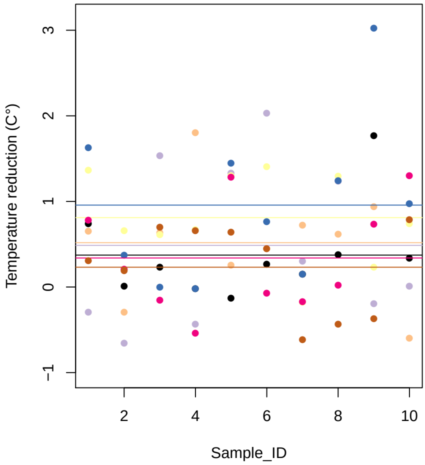
Back to our example

- This is one of many possible outcomes...



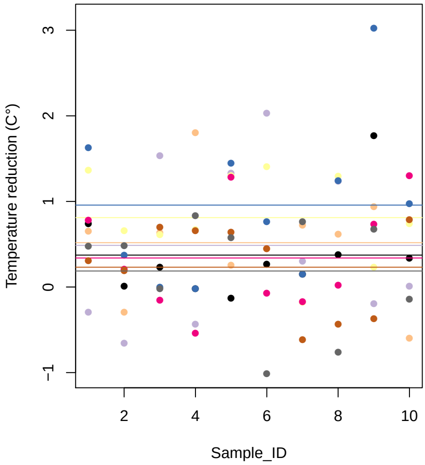
Back to our example

- This is one of many possible outcomes...



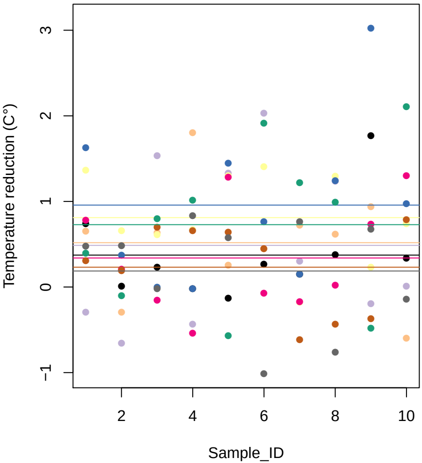
Back to our example

- This is one of many possible outcomes...



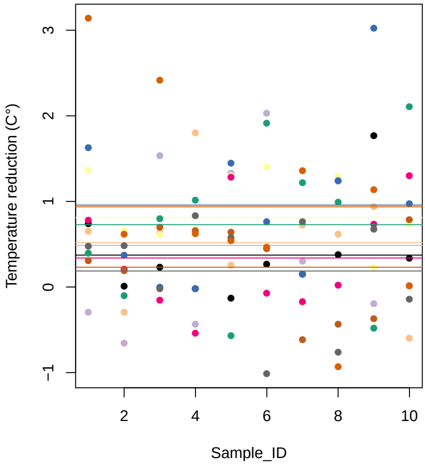
Back to our example

- This is one of many possible outcomes...



Back to our example

- This is one of many possible outcomes...
- How can I formalise this uncertainty and use it for inference?



Frequentist approach: in a nutshell

The outcome of my experiment is one of many possible outcomes I could get if I were to repeat my experiment many times.

Randomness is only associated with the data outcomes, while the parameter is a fixed, although unknown, quantity.

Inference?

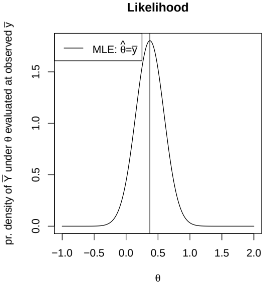
- **(Point estimate)** Most likely value which generated the observed data: maximum likelihood estimate
- **(Variability)** Variability of the point estimate based on the variability of the data outcomes
- **(Confidence interval)** Interval which would contain the true value of the parameter $(1 - \alpha) * 100\%$ times under (potential) repetition of my experiment
- **(Hypothesis test - Fisher)** *p*-value as the probability of observing a (potential) data outcome as or more extreme than the observed one under the null hypothesis.

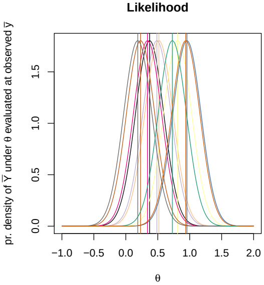
Frequentist inference on temperature reduction (1)

Recall that $\bar{y} = (1/n)\sum_{i=1}^{n} y_i = 0.37$, and $y_i \sim N(\theta, \sigma = 0.7)$

- **(Point estimation)** what is the most likely value of θ? -> a 'good' estimator for θ is the maximum likelihood estimator: $\hat{\theta} = \bar{y} = 0.37$

- **(Variability)** what is the variance of $\hat{\theta}$? -> $Var(\hat{\theta}) = Var(\bar{y})$, then (leap of faith / basic lecture series) $Var(\bar{y}) = \sigma^2/n = 0.049$ -> $SD(\bar{y})$=0.22.





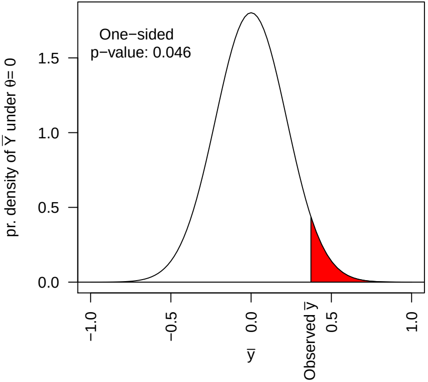
- **(Confidence interval)** An interval which would contain the true θ on average 95% of the times?
-> let $z = \frac{\bar{y}-\theta}{\sigma/\sqrt{n}}$, then (leap of faith / basic lecture series) $z \sim N(0,1)$ which implies
$Pr[-1.96 < z < 1.96] = 0.95 \iff$
$Pr[\bar{y} - 1.96 * \sigma/\sqrt{n} < \theta < \bar{y} + 1.96 * \sigma/\sqrt{n}] = 0.95$ -> 95% CI = [-0.06, 0.81]

Frequentist inference on temperature reduction (2)

Recall that $\bar{y} = (1/n) \sum_{i=1}^{n} y_i = 0.37$, and $y_i \sim N(\theta, \sigma = 0.7)$

(Hypothesis test)?

- Define a null hypothesis H_0 and an alternative hypothesis H_1
- Suppose (one-sided testing) $H_0 : \theta \leq 0$ vs $H_1 : \theta > 0$
- What is the probability of observing a result as or more extreme than 0.37 under H_0? (p-value)
- Test decision: to control type I error rate at level α, reject if the p-value $< \alpha$
- Recall:
  - type I error rate (α): Pr(rejecting H_0 given H_0 is true)
  - type II error rate (β): Pr(keeping H_0 given H_1 is true)



Do we really need anything else?

- Frequentist statistics is concerned about **long run properties**:
  - Confidence intervals contain the true θ on average $(1-\alpha) * 100\%$ of the times
  - The null hypothesis falsely rejected on average $\alpha * 100\%$ of the times
  - ...
- Properties actually meant to hold across different experiments/true parameter values
- But they do not describe **probabilities about the parameters based on the current experiment**:
  - The current true θ may or may not belong to the current confidence interval
  - Type I error rate describes the probability of rejecting H_0 given it is true -> says nothing about the probability of H_0 being true given that it is rejected!
- Also, is the data the only information we have (and willing to consider) about the parameter?

Prior information: Is it important?

Think of two situations [1]

- A music expert claims to be able to distinguish a page of Haydn score from a page of Mozart score
- A drunken friend says he can predict the outcome of a flip of a fair coin

Suppose that in both cases 10 trials are conducted, all successful.

Null hypothesis: the person is guessing (-> observations Bin(10,θ_0=0.5))
In both situations we have the same empirical evidence, so same *p*-value=0.5^{10}

What would be your conclusions?

[1] Savage (1961), see Berger (1985)

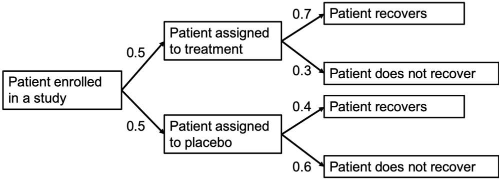
The Bayesian approach

The parameter - although still a fixed unknown quantity - is uncertain. This uncertainty can be modelled in the form of a probability distribution.

The key tool is Bayes' theorem

$$Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)} = \frac{Pr(A|B)Pr(B)}{Pr(A)}$$

Bayes' theorem: a simple example



Basic rule: $\Pr(A \cap B) = Pr(A|B)Pr(B)$

Note: if A and B are independent, Pr(A|B)=Pr(A) and Pr(B|A)=Pr(B)

Example:
A=Patient recovers, B=Patient assigned to treatment,
A^C=Patient does not recover, B^C=Patient assigned to placebo

$Pr(A \cap B) = Pr(A|B)Pr(B) = 0.7 * 0.5$
$Pr(A) = Pr(A|B)Pr(B) + Pr(A|B^C)Pr(B^C) = 0.7 * 0.5 + 0.4 * 0.5$

$$Pr(B|A) = \frac{Pr(A|B)P(B)}{Pr(A)} = \frac{0.7 * 0.5}{0.7 * 0.5 + 0.4 * 0.5} \approx 0.64$$

The Bayes' rule

Assume now that

- 'B' : parameter θ -> Pr(B) replaced by the **prior distribution** $\pi(\theta)$
- 'A' : data y -> Pr(A|B) replaced by the **likelihood** $\pi(y|\theta)$

The **posterior** distribution describes the updated knowledge about θ once the data have been observed:

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)} \propto \pi(y|\theta)\pi(\theta)$$

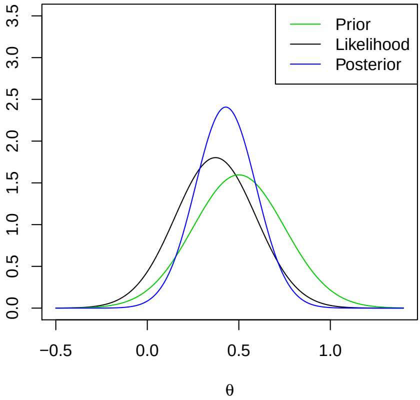
The denominator $\pi(y)$ is often omitted as it does not depend on θ.

Back to the temperature example

- Suppose $\pi(\theta)$ is a $N(\mu_\theta = 0.5, \sigma_\theta = 0.25)$ **(prior assumption)**
- Recall $\pi(\bar{y}|\theta)$ is a $N(\theta, \sigma = 0.22)$ **(data evidence)**

The **posterior** $\pi(\theta|\bar{y})$ in this case is also normal, with

- Posterior mean= 0.43
- Posterior st. deviation= 0.17

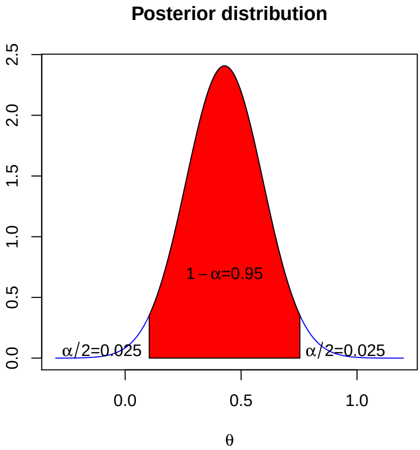


Bayesian inference

- $\pi(\theta|y)$ contains all the available information concerning θ -> most comprehensive output to report
- More synthetic measures derived from $\pi(\theta|y)$:
  - **(Point estimate)**: e.g. posterior mean, mode, median
  - **(Variability)**: for the post. mean as point estimate it is the post. variance
  - **(Credible interval)**: a region of the parameter space which has $1-\alpha$ probability
  - **(Hypothesis test decision)**: taken according to the posterior odds or the Bayes Factor
- In non-standard problems, the most complex part (numerically) is to obtain the posterior distribution.
- Derived quantities are generally quite straightforward.

Bayesian inference for temperature reduction

- **(Point estimate)** Posterior mean= 0.43
- **(Variability)** Posterior st. deviation= 0.17
- **(Interval estimation)** A 95% symmetric credible interval results: [0.1,0.75]



Testing

Probabilities of the hypotheses

Set of hypotheses: H_0: $\theta \leq \theta_0$ vs H_1: $\theta > \theta_0$

We can now talk about the **probability of the hypotheses**

Prior probabilities of the hypotheses are based on the prior distribution $\pi(\theta)$

- Prior probability of H_0: $\Pr(\theta \leq \theta_0)$
- Prior probability of H_1: $\Pr(\theta > \theta_0)=1-\Pr(\theta \leq \theta_0)$

Posterior probabilities of the hypotheses are based on the posterior distribution $\pi(\theta|y)$

- Posterior probability of $H_0|y$: $\Pr(\theta \leq \theta_0|y)$
- Posterior probability of $H_1|y$: $\Pr(\theta > \theta_0|y)=1-\Pr(\theta \leq \theta_0|y)$

Test decision

Decision-theoretic solution

Set of hypotheses: H_0: $\theta \leq \theta_0$ vs H_1: $\theta > \theta_0$

Test decision?
Decision theoretic solution:

- **Cost ratio** c_0/c_1: ratio between the cost of making a type II error (c_0) and the cost of making a type I error (c_1)
- Reject H_0 if

$$c_1 Pr(H_0|y) < c_0 Pr(H_1|y) \iff \underbrace{\frac{Pr(H_1|y)}{Pr(H_0|y)}}_{\text{Posterior odds}} > \underbrace{\frac{c_1}{c_0}}_{\text{Costs ratio}}$$

Test decision

Bayes factor

Set of hypotheses:
H_0: $\theta \leq \theta_0$ vs H_1: $\theta > \theta_0$

Test decision?
If smaller influence of the prior is desired, use **Bayes Factor**:

$$\mathrm{BF} = \underbrace{\frac{Pr(H_1|y)}{Pr(H_0|y)}}_{\text{Posterior odds}} \Big/ \underbrace{\frac{Pr(H_1)}{Pr(H_0)}}_{\text{Prior odds}}$$

- Describes the change of belief provided by the experiment
- Guidance about critical values of the BF is available

Test decision

Bayes factor guidance

Kass & Raftery's scale of evidence

BF	Evidence against H_0
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
More than 150	Very strong

Note: for evidence against H_1, same interpretation applies to 1/BF

Bayesian testing for temperature reduction

$H_0 : \theta \leq 0$ vs $H_1 : \theta > 0$

Prior probabilities:

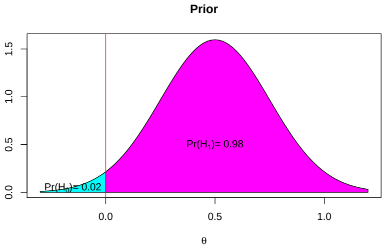
- H_0: $\Pr(\theta \leq 0)$=0.02
- H_1: $\Pr(\theta > 0)$=0.98

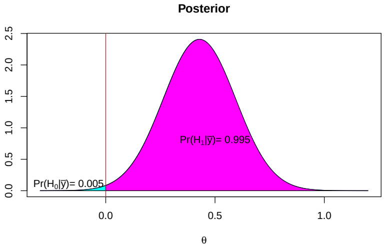
Posterior probabilities:

- H_0: $\Pr(\theta \leq 0|y)$=0.005
- H_1: $\Pr(\theta > 0|y)$=0.995

Posterior odds = 205 (reject H_0 if type I error is less than 205 times more costly than type II error)

Bayes factor = 4.78 (Positive evidence against H_0)





Type I error rate and the probability of the null

A digression

From Bayes' theorem:

$$Pr(H_0 \text{ true} \mid \text{reject } H_0) = \underbrace{Pr(\text{reject } H_0 | H_0 \text{ true})}_{\text{Type I error rate}} \frac{Pr(H_0 \text{ true})}{Pr(\text{reject } H_0)}$$

Example:

- Suppose Pr(H_0 true)=Pr(H_1 true)=0.5
- Suppose type I error rate α=0.05 and type II error rate β=0.9 (power = 0.1)
- Recall: Pr(reject H_0)=
  Pr(reject H_0 | H_0 true) Pr(H_0 true)+Pr(reject H_0 | H_1 true) Pr(H_1 true)=
  $\alpha * 0.5 + (1 - \beta) * 0.5 = 0.075$
- -> Pr(H_0 true | reject H_0) $\approx$ 0.33!

Prior choice

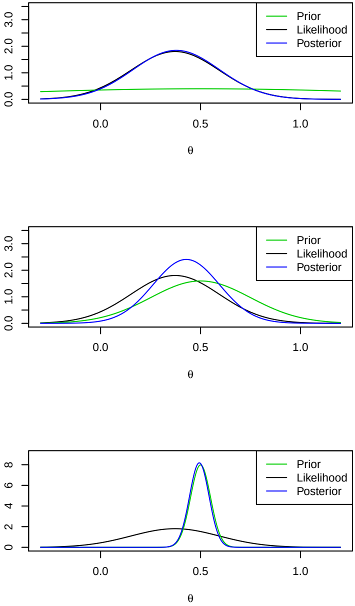
The Bayesian view

- Bayesian approach requires elicitation of a prior distribution for the parameter
- A prior can be informed by past experiments or expert beliefs
- Rules have been developed to elicit ‘objective’ priors if no prior knowledge is available
- *Any* prior is somewhat informative (Robert, 2007)

Prior choice

Why and when is it important?

- Prior has an impact on the posterior -> generally stronger the smaller is the sample size
- If the prior is very 'concentrated' around specific values, a large number of observations needed to 'overtake' prior assumptions
- The prior can and should in principle be useful!



Prior choice

Handle with care

"It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so." - (attributed to) M. Twain

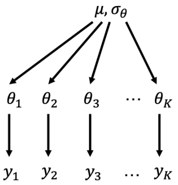
Solutions?

- The prior should be appropriate for the analysis at hand (!)
- 'Appropriateness' should be checked 'a priori' but also after seeing the data: to detect potential **prior-data conflict**
- Sensitivity analyses
- Robust prior choices and robust Bayesian analysis
- The information contained in the prior should generally not be stronger than information provided by the data

Hierarchical modelling

- Hierarchical modelling when parameters are related
- Parameters from a **common prior distribution**

Normal hierarchical model (variances known)



Hyperprior: $\mu \sim N(\phi, \sigma_\mu)$

Hyperparameters

Prior: $\theta_k \sim N(\mu, \sigma_\theta)$

Data: $y_k \sim N(\theta_k, \sigma_y)$

- Exchangeability assumption:
  no information (apart from the data itself) is available to make the parameters distinguishable (Gelman et al., 2013)
- E.g. measurements from different labs/hospitals/test centres with no prior information on whether locations may differ systematically

Hierarchical modelling: notes

- Hierarchical modelling compromise between pooling all data together, and analyse them separately
- It can '**borrow strength**' from all data sources when estimating the parameters
- The choice of the hyperprior/hyperparameters is generally important, particularly with few groups
- 'Empirical Bayes' estimates hyperparameters from the data (no hyperprior) (Berger, 1985):
    - Computationally more convenient, but ignores uncertainty about the estimates
    - With a large number of groups, close to estimation under the full hierarchical model

Computation

- Computation is generally more time-consuming with Bayesian approaches
- The main difficulty is in computing the posterior
- Exception: simple models with ‘conjugate priors’ -> analytic solutions available
- When posterior not known analytically: **draw samples** from it and/or **approximate** it
- Variety of algorithms
  - targeting the exact posterior (e.g. Markov Chain Monte Carlo - MCMC, Hamiltonian Monte Carlo - HMC)
  - targeting an approximation (e.g. Variational Bayes, Integrated nested Laplace approximation - INLA, Approximate Bayesian Computation - ABC)
- Variety of packages - e.g. STAN (newer), JAGS, INLA...

Main sources & further reading

- Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science Business Media.
- Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science Business Media.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Conclusions & take home message

- The Bayesian approach has been proven useful when **incorporation of prior/external information** is desired...
- ...And to obtain probabilistic statements about the parameters in the current experiment.
- It is also convenient when fitting particularly **complex** models, as the only added difficulty is computational (obtain the posterior distribution)
- Hierarchical modelling has several advantages as it ‘automatically’ compromises between pooling all data and performing separate analyses
- The choice of a prior distribution is often challenging, and although there are proposals for ‘standard’ priors to use when no knowledge is available, any prior is somewhat informative
- Long run properties are less of a concern in Bayesian analyses -> they may be satisfactory, but if they are strictly required, it may be better to just do a frequentist analysis
- In general, eliciting a **good data model** and collecting **good data** is more crucial than choosing a paradigm

...Up next

27th January: Issues in Statistical Practice: Errors, missing data, and reproducible research (Dr. Manuel Wiesenfarth)

References

M. J. Bayarri and J. O. Berger. The Interplay of Bayesian and Frequentist Analysis. *Statistical Science*, 19(1):58–80, 2004. ISSN 08834237. URL http://www.jstor.org/stable/4144373.

J. O. Berger. *Statistical decision theory and Bayesian analysis; 2nd ed.* Springer Series in Statistics. Springer, New York, 1985.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

C. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.