On being Bayesian

Namur 13 October 2016

Stephen Senn

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Acknowledgements

Thank you for the kind invitation

This work is partly supported by the European Union's 7th Framework Programme for research, technological development and demonstration under grant agreement no. 602552. "IDEAL"

The work on historical placebos is joint with Olivier Collignon and Anna Schritz in my group

Basic Thesis

- The Bayesian approach holds out the promise of providing a principled way of synthesizing difference sources of information
- This is, however, more difficult than many suppose
- Key tasks are
  - Appropriate formulation of prior distributions
  - Establishing exactly what the objective content of such prior distributions is
  - Understanding what a prior distribution commits you to believe
  - Developing *insight* (mathematics is not enough)

Outline

- Brief, basic reminder as to how it works
  - Illustrated using a simple example
- What prior distributions have to reflect
- Some examples to check understanding
  - A simple binary outcome
  - Dawid's selection paradox
  - Historical placebos
- Some advice

Key features of Bayesian inference

1. Probability is given a personal and subjective interpretation
2. Probabilities do not have to be defined in terms of (theoretical) infinite repetitions
3. Probability statements about parameters and predictions become the goal of inference
4. There is nothing inherently special about a defined set-up for collecting data
5. To be fully Bayesian utilities should be considered also

An Example

My compact disc (CD) player* allowed me to press tracks in sequential order by pressing ***play*** or in random order by playing ***shuffle***.

One day I was playing the CD *Hysteria* by Def Leppard. This CD has 12 tracks.

I thought that I had pressed the *shuffle* button but the first track played was ‘women’, which is the first track on the CD.

Q. What is the probability that I did, in fact, press the *shuffle* button as intended?

*I now have an Ipod nano

A Bayesian Solution

We have two basic hypotheses:

1) I pressed *shuffle*.

2) I pressed *play*.

First we have to establish a so-called ***prior probability*** for these hypotheses: a probability before seeing the evidence.

Suppose that the probability that I press the *shuffle* button when I mean to press the shuffle button is 9/10. The probability of making a mistake and pressing the *play* button is then 1/10.

Next we establish probabilities of events *given* theories. These particular sorts of probabilities are referred to as ***likelihoods*** ,a term due to RA Fisher(1890-1962).

If I pressed *shuffle*, then the probability that the first track will be ‘women’ (W) is 1/12. If I pressed *play*, then the probability that the first track is W is 1.

For completeness (although it is not necessary for the solution) we consider the likelihoods had any other track apart from ‘women’ (say X) been played.

If I pressed *shuffle* then the probability of X is 11/12. If I pressed *play* then this probability is 0.

We can put this together as follows

Hypothesis	Prior Probability P	Evidence	Likelihood	P x L
Shuffle	9/10	W	1/12	9/120
Shuffle	9/10	X	11/12	99/120
Play	1/10	W	1	12/120
Play	1/10	X	0	0
TOTAL				120/120 = 1

After seeing (hearing) the evidence, however, only two rows remain

Hypothesis	Prior Probability P	Evidence	Likelihood	P x L
Shuffle	9/10	W	1/12	9/120
Shuffle	9/10	X	11/12	99/120
Play	1/10	W	1	12/120
Play	1/10	X	0	0
TOTAL				21/120

The probabilities of the two cases which remain do not add up to 1.

However, since these two cases cover all the possibilities which remain, their combined probability *must* be 1.

Therefore we rescale the individual probabilities to make them add to 1.

We can do this without changing their relative value by dividing by their total, 21/120.

This has been done in the table below.

So we rescale by dividing by the total probability

Hypothesis	Prior Probability P	Evidence	Likelihood	P x L	Posterior Probability
Shuffle	9/10	W	1/12	9/120	(9/120)/(21/120) =9/21
Shuffle	9/10	X	11/12	99/120	
Play	1/10	W	1	12/120	(12/120)/(21/120) =12/21
Play	1/10	X	0	0	
TOTAL				21/120	21/21=1

The probability I pressed play is 9/21
This completes the Bayesian solution

Characteristics of prior distributions

- They must be what you would use to bet on in advance of getting any further data
- No amount of further data in any form should be capable of causing you to revise your prior distribution *qua* prior
  - Updating your prior distribution to become a posterior is quite another matter
  - Remember that to the extent defined by the model the prior distribution and the data are exchangeable
  - Wanting to change your prior is like wanting to change some data

An example to get you started

- You are proposing to estimate the probability θ of a binary event
  - E.g. cure/no cure
- You use a uniform prior on θ
- You now proceed to study 10,000 occurrences
- Which does your prior distribution say is more likely?
  - 10,000 successes
  - 5,000 successes 5,000 failures, *in any order*

Case Ascertainment

- One of the things we learn in statistics is that it matters how we ascertain cases
- The selection procedure affects our inferences
- We react differently if we learn that the results we are being shown are from a treatment that was one of many if it was chosen randomly or it was chosen as the best observed

A Selection Paradox of Dawid's

- Suppose that we estimate treatment means from a number of treatments in clinical research
- We use a standard conjugate prior
- Since Bayesian analysis is full conditioned on the data, then for any treatment the posterior mean will ***not*** depend on why we have chosen the treatment
  - At random
  - Because it gave the largest response

See DAWID, A. P. (1994), in *Multivariate Analysis and its Applications*, eds. T. W. Anderson, K. a.-t. a. Fang, & I. Olkin

A Simulation to Illustrate This

prior mean, $\theta = 0$

prior variance, $\tau^2 = 1.0$

data variance, $\sigma^2 = 4.0$

cluster size, $m = 10$

number of simulations = 500

μ = true mean

m = data mean

$\hat{\mu}_{Bayes}$ = posterior mean

- Simulate 10 true means
- For each true mean simulate observed value
- Now select in one of two ways
  1. Randomly choose one member from each group of 10
  2. Choose the member with the highest observed mean

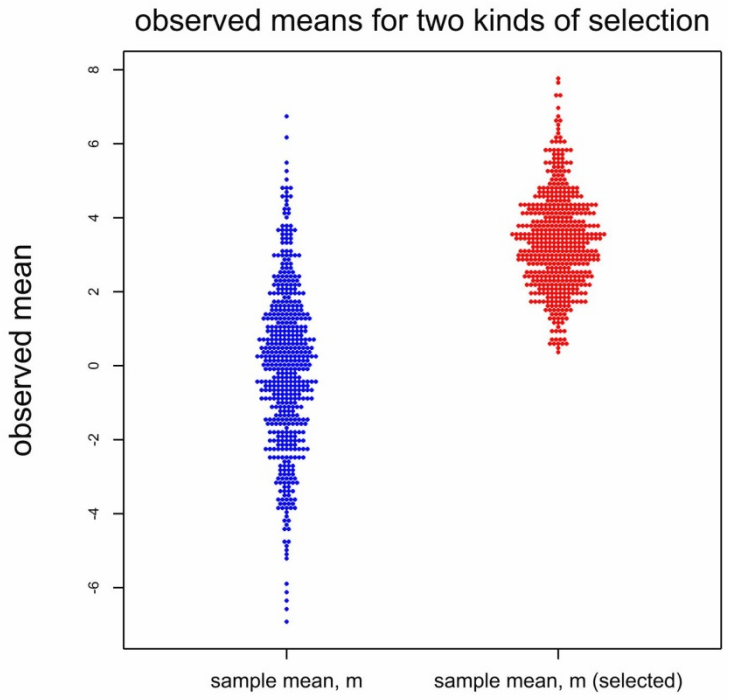
What the Bayesian theory says

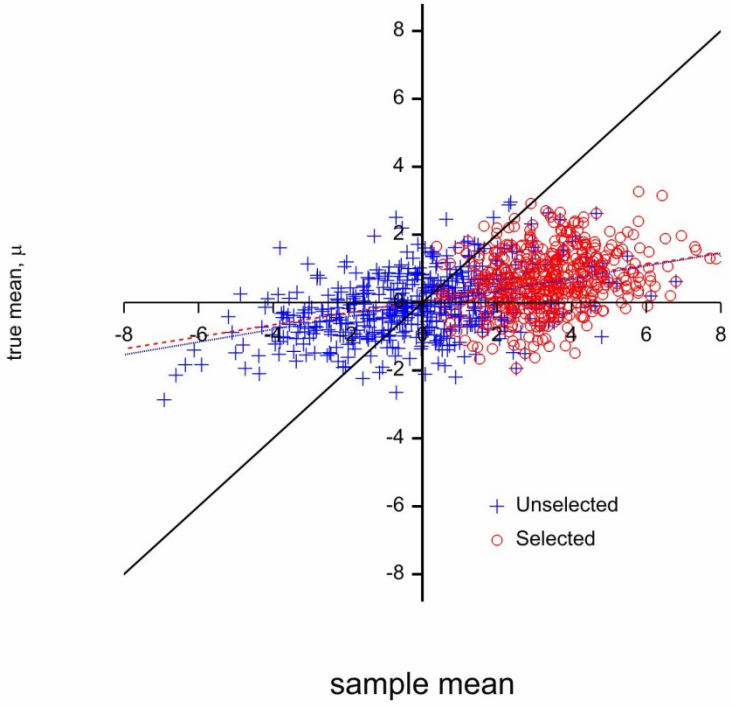
$$\hat{\mu}_{Bayes} = \frac{\left(\frac{1}{\tau^2} \times \theta\right) + \left(\frac{1}{\sigma^2} \times m\right)}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}}$$

$$\hat{\mu}_{Bayes} = \frac{\left(\sigma^2 \times \theta\right) + \left(\tau^2 \times m\right)}{\tau^2 + \sigma^2}$$

$$\hat{\mu}_{Bayes} = 0 + \frac{1}{1+4} m = 0 + 0.2m$$

What the simulation shows





Regression analysis

Estimates of parameters

Parameter	estimate	s.e.	t(498)	t pr.
Constant	-0.0409	0.0390	-1.05	0.294
sample mean, m	0.1869	0.0176	10.64	<.001

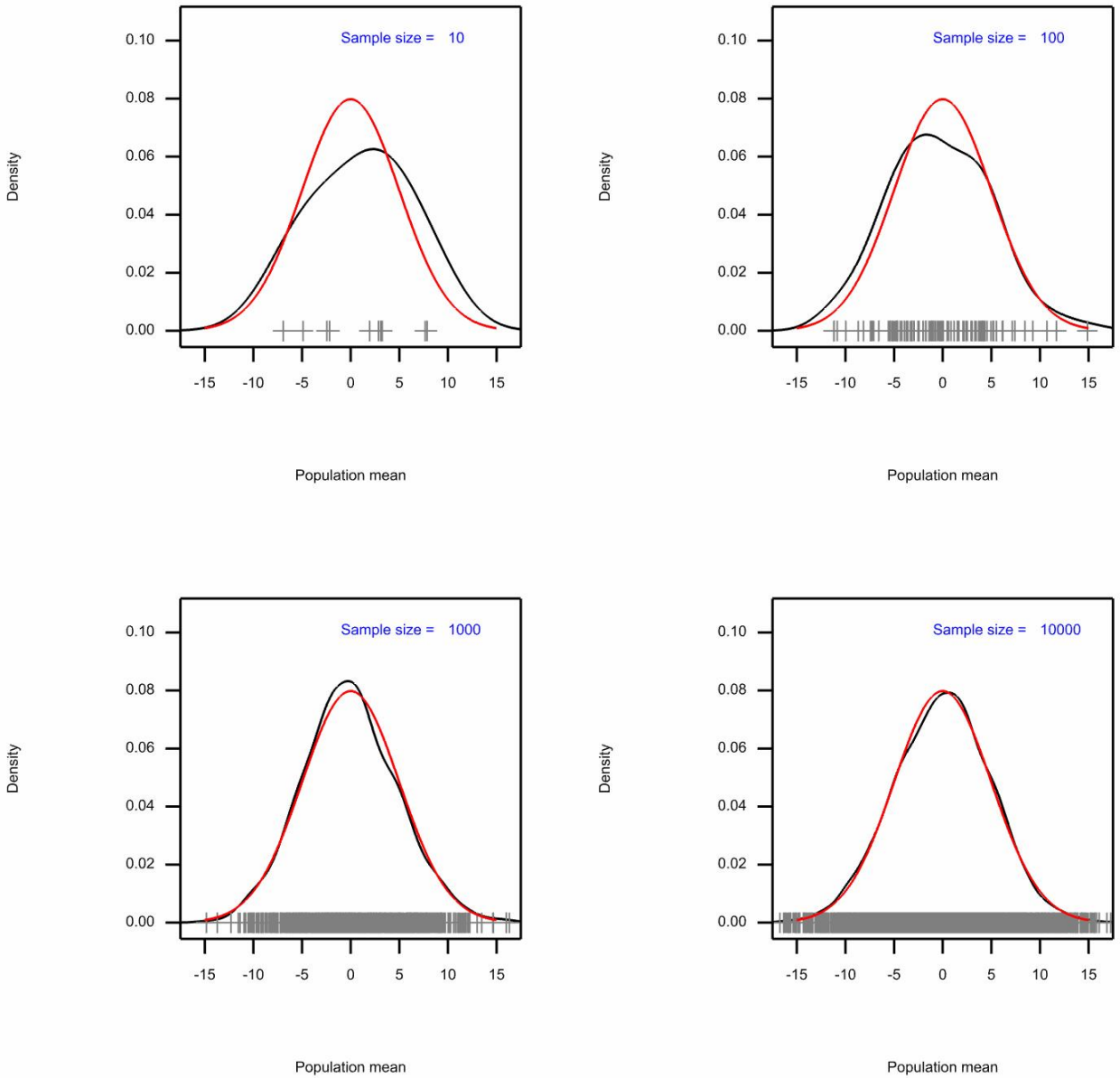
Theory says 0.2

Estimates of parameters

Parameter	estimate	s.e.	t(498)	t pr.
Constant	0.052	0.106	0.49	0.625
sample mean, m (selected)	0.1769	0.0294	6.02	<.001

Does this mean the frequentist intuition is wrong?

- Not necessarily
- One needs to think carefully about what the prior distribution implies
- Actually, even if the prior variance were large the prior distribution would be *very* informative about two things
  - Normality
  - Conditional independence


Sample size = 10
Sample size = 100
Sample size = 1000
Sample size = 10000
Density
Population mean
0.00
0.02
0.04
0.06
0.08
0.10
-15
-10
-5
0
5
10
15

The explanation of the paradox

- Having a Normal prior is equivalent to having seen *thousands* of true means
- Furthermore, *a priori*, the true mean of any value in your sample of ten is exchangeable with any one of these thousands of means
- Why should the fact that it is locally the highest have any effect on your Bayesian calibration?
- Now let us see what happened when we no longer make the means exchangeable

A hierarchical simulation

prior mean of all clusters, $\theta = 0$

prior variance of cluster means, $\tau^2 = 0.5$

within cluster variance, $\gamma^2 = 0.5$

data variance, $\sigma^2 = 4.0$

cluster size, $m = 10$

number of simulations = 500

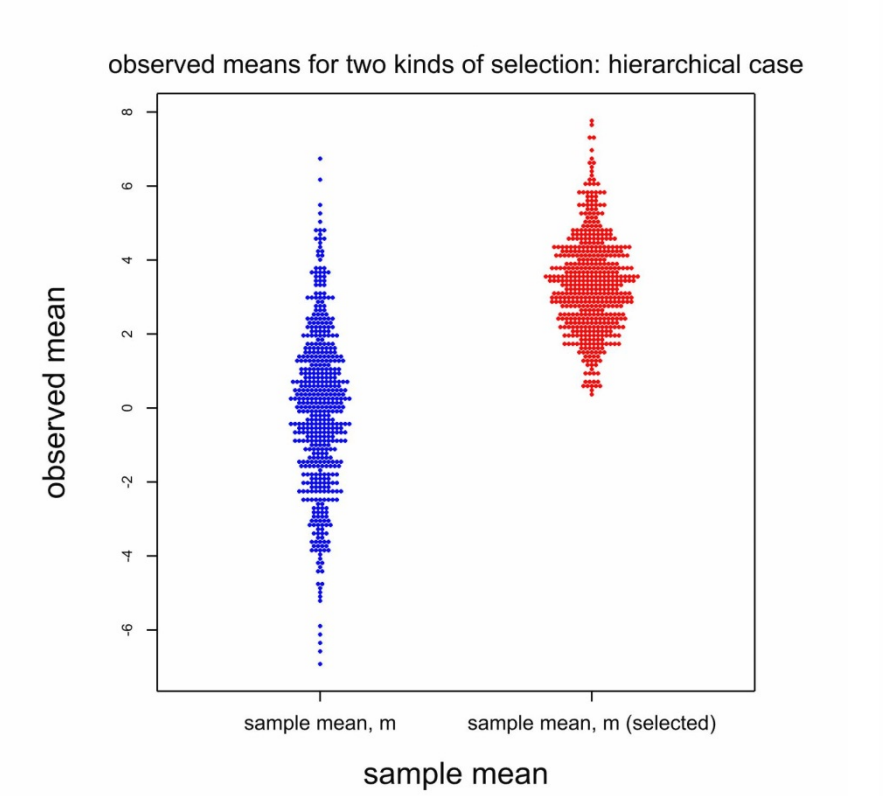
μ = true mean

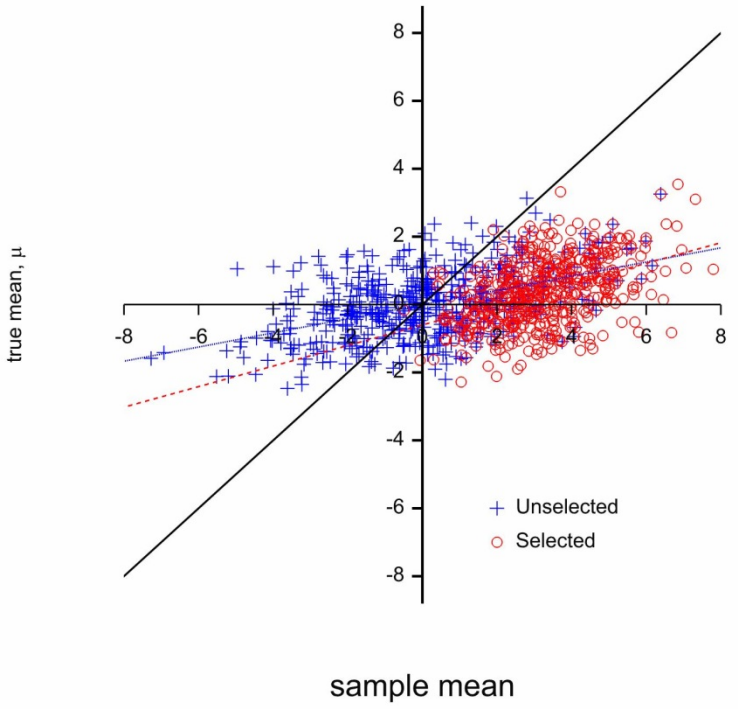
m = data mean

$\hat{\mu}_{Bayes}$ = posterior mean

- Simulate cluster mean
- Then simulate for cluster members
- Simulation run two ways
  1. Randomly choose one member from each group of 10
  2. Choose the member with the highest observed mean

What the simulation shows





The regression equations are now quite different depending on how the means were chosen

Regression analysis

Estimates of parameters

Parameter	estimate	s.e.	t(498)	t pr.
Constant	0.0011	0.0390	0.03	0.978
sample mean, m	0.2087	0.0171	12.21	<.001

Estimates of parameters

Parameter	estimate	s.e.	t(498)	t pr.
Constant	-0.6004	0.0973	-6.17	<.001
sample mean, m (selected)	0.3027	0.0275	11.01	<.001

Lessons

- As soon as you replace the conjugate prior with a hierarchical one you get very different results according to selection
- Be *very careful* to establish what your prior implies
- True Bayesian inference does not necessarily give you the license to ignore frequentist lessons you might think

Historical control

- In many indications, the same treatment is often used as a control
  - Either a placebo
  - Or a standard treatment
- This means that when a new treatment is trialled there will be a lot of information from previous trials on the control being used
- Since Bayes is supposed to be a way of synthesizing all information, how would we do this?

Problem

- Obviously a historical control is not worth the same as a concurrent control
- How should we deal with this?
- Ask the following question
- Given a choice between an infinite number of historical controls and *n* concurrent controls how large does *n* have to be before I prefer the latter?

Model (frequentist formulation)

$$\mu_i \sim N\left(\mathrm{M}, \gamma^2\right)$$

$$Y_{ic} \sim N\left(\mu_i, \sigma^2\right)$$

$$Y_{it} \sim N\left(\mu_i + \tau, \sigma^2\right)$$

$$Var\left(\bar{Y}_{it} - \bar{Y}_{ic}\right) = \sigma^2\left(\frac{1}{n_{it}} + \frac{1}{n_{ic}}\right), Var\left(\bar{Y}_{it} - \bar{Y}_{jc}\right) = \sigma^2\left(\frac{1}{n_{it}} + \frac{1}{n_{jc}}\right) + 2\gamma^2, i \neq j$$

$$\sigma^2\left(\frac{1}{n_{it}} + \frac{1}{n_{jc}}\right) + 2\gamma^2 = \frac{\sigma^2}{n_{it}} + 2\gamma^2$$

$$n_{jc} \to \infty$$

Hence, $2\gamma^2 = \frac{\sigma^2}{n_{ic}^*}$, where $\mathrm{n}_{\mathrm{ic}}^*$ is number of concurrent controls you would prefer to infinitely many historical ones

But….

- When you start thinking like this you begin to wonder
- Is it really the number of historical control patients that I have that is important?
- Or should I really be thinking about the data in some other way?
- What do the data really represent?

The TARGET study

- One of the largest studies ever run in osteoarthritis
- 18,000 patients
- Randomisation took place in two sub-studies of equal size
  - Lumiracoxib versus ibuprofen
  - Lumiracoxib versus naproxen
- Purpose to investigate cardiovascular and gastric tolerability of lumiracoxib
  - That is to say side-effects on the heart and the stomach

Baseline Demographics

	Sub-Study 1		Sub Study 2	
Demographic Characteristic	**Lumiracoxib n = 4376**	**Ibuprofen n = 4397**	**Lumiracoxib n = 4741**	**Naproxen n = 4730**
Use of low-dose aspirin	975 (22.3)	966 (22.0)	1195 (25.1)	1193 (25.2)
History of vascular disease	393 (9.0)	340 (7.7)	588 (12.4)	559 (11.8)
Cerebro-vascular disease	69 (1.6)	65 (1.5)	108 (2.3)	107 (2.3)
Dyslipidaemias	1030 (23.5)	1025 (23.3)	799 (16.9)	809 (17.1)
Nitrate use	105 (2.4)	79 (1.8)	181 (3.8)	165 (3.5)

Baseline Chi-square P-values

	Model Term		
Demographic Characteristic	**Sub-study (DF=1)**	**Treatment given Sub-study (DF=2)**	**Treatment (DF=2)**
Use of low-dose aspirin	< 0.0001	0.94	0.0012
History of vascular disease	< 0.0001	0.07	<0.0001
Cerebro-vascular disease	0.0002	0.93	0.0208
Dyslipidaemias	<0.0001	0.92	<0.0001
Nitrate use	< 0.0001	0.10	<0.0001

Outcome Variables

All four groups

	Sub-Study 1		Sub Study 2	
Outcome Variables	**Lumiracoxib n = 4376**	**Ibuprofen n = 4397**	**Lumiracoxib n = 4741**	**Naproxen n = 4730**
Total of discontinuations	1751 (40.01)	1941 (44.14)	1719 (36.26)	1790 (37.84)
CV events	33 (0.75)	32 (0.73)	52 (1.10)	43 (0.91)
At least one AE	699 (15.97)	789 (17.94)	710 (14.98)	846 (17.89)
Any GI	1855 (42.39)	1851 (42.10)	1785 (37.65)	1988 (21.87)
Dyspepsia	1230 (28.11)	1205 (27.41)	1037 (21.87)	1119 (23.66)

Deviances and P-Values

Lumiracoxib only fitting Sub-study

	Statistic	
Outcome Variables	**Deviance**	**P-Value**
Total of discontinuations	13.61	0.0002
CV events	2.92	0.09
At least one AE	1.73	0.19
Any GI	21.31	<0.0001
Dyspepsia	47.34	< 0.0001

A Simple Model

An unrealistic balanced trial

n patients per arm, c centres in total with p patients per centre

$$2n = pc, \qquad n = \frac{pc}{2}$$

Between-centres variance is γ^2 within-centre variance is σ^2.

Design	Variance of Treatment Contrast
Completely randomised	$4\frac{(\gamma^2+\sigma^2)}{cp}$
Randomised blocks (centre blocks)	$4\frac{\sigma^2}{cp}$
Cluster randomised	$4\frac{\left(\gamma^2+\frac{\sigma^2}{p}\right)}{c}$

When using external controls we have *at least* the variability of a cluster randomised trial

Lessons from TARGET

- If you want to use historical controls you will have to work *very* hard
- You need at least two components of variation in your model
  - Between centre
  - Between trial
- And possibly a third
  - Between eras
- What seems like a lot of information may not be much
- Concurrent control and randomisation seems to work well
- Moral for any Bayesian: find out as much as possible about any data you intend to use

That example revisited

The question

- You are proposing to estimate the probability θ of a binary event
  - E.g. cure/no cure
- You use a uniform prior on θ
- You now proceed to study 10,000 occurrences
- Which does your prior distribution say is more likely?
  - 10,000 successes
  - 5,000 successes 5,000 failures, *in any order*

The solution

- You started with an 'uninformative prior
- After 10,000 trials the observed proportion must be pretty much what you believe is the true probability
- But you said every true probability is equally likely
- Therefore 5,000 success in any order is just as likely as 10,000 success

Advice

- Think hard about any prior distribution
- Try to establish the objective content of any prior distribution
- Uninformative prior distributions are not appropriate for nuisance parameters
- Be prepared to think hierarchically
- Check that
  - The prior distribution states your current belief
  - No data in any shape of form would cause you to abandon it
- If the result seems to contradict frequentist wisdom think carefully why
- Develop statistical insight – understand what being Bayesian *means*
- As in any statistically system, ask yourself the question
  - How did I get to see what I see?

References

1 Senn, S. J. Bayesian, likelihood and frequentist approaches to statistics. *Applied Clinical Trials* ***12, 35-38 (2003).***

2 Senn, S. J. Trying to be precise about vagueness. *Statistics in Medicine* ***26, 1417-1430 (2007).***

3 Senn, S. A note concerning a selection "Paradox" of Dawid's. *American Statistician* ***62, 206-210, doi:10.1198/000313008x331530 (2008).***

4 Senn, S. J. Comment on article by Gelman. *Bayesian analysis* ***3, 459-462 (2008).***

5 Senn, S. J. Comment on "Harold Jeffreys's Theory of Probability Revisited". *Statistical Science* ***24, 185-186 (2009).***

6 Senn, S. J. You may believe you are a Bayesian but you are probably wrong. *Rationality, Markets and Morals* ***2, 48-66 (2011).***