Case Studies in Bayesian Data Science

4: Optimal Bayesian Analysis in Digital Experimentation

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

Short Course (Day 5)
University of Reading (UK)

© *David Draper (all rights reserved)*

The Big Picture

- Problems addressed by the discipline of **statistics** typically have the following structure.
- You (Good 1950) [note the capital Y]: a generic person wishing to reason sensibly in the presence of uncertainty) are given a **problem** $\mathbb{P} = (\mathbb{Q}, \mathbb{C})$ involving **uncertainty** about θ, the unknown aspect of $\mathbb{P}$ of principal interest.
- Here $\mathbb{Q}$ identifies the main **questions** to be answered, and $\mathbb{C}$ represents the (real-world) **context** in which the questions are raised, instantiated through a finite set $\mathcal{B}$ of **(true/false) propositions**, all rendered true by problem context.
- You examine Your resources and find that it's possible to obtain a new **data set** D to decrease Your uncertainty about θ.
- In this setting, a **Theorem** due to Cox (1946) and Jaynes (2002) — recently rigorized and extended by Terenin and Draper (2015) — says that

The Big Picture (continued)

- *If You're prepared to specify two probability distributions — $p(\theta \mid \mathcal{B})$, encoding Your information about θ **external** to D, and $p(D \mid \theta \, \mathcal{B})$, capturing Your information about θ **internal** to D — then **optimal inference** about θ is based on the distribution $p(\theta \mid D \, \mathcal{B}) \propto p(\theta \mid \mathcal{B}) \, p(D \mid \theta \, \mathcal{B})$, and **optimal prediction** of new data D^* is based on the distribution $p(D^* \mid D \, \mathcal{B}) = \int_\Theta p(D^* \mid \theta D \, \mathcal{B}) \, p(\theta \mid D \, \mathcal{B}) \, d\theta$, where Θ is the set of possible values of θ (another part of the theorem covers **optimal decision-making**, but that's not relevant to this talk).*
- Let's agree to call $M = \{ p(\theta \mid \mathcal{B}), p(D \mid \theta \, \mathcal{B}) \}$ Your **model** for Your uncertainty about θ and D^*.
- The two main **practical challenges** in using this Theorem are
  - (technical) **Integrals** arising in **computing** the inferential and predictive distributions may be difficult to approximate accurately, and
  - (substantive) The mapping from $\mathbb{P}$ to $M = \{ p(\theta \mid \mathcal{B}), p(D \mid \theta \, \mathcal{B}) \}$ is rarely unique, giving rise to **model uncertainty**.

Optimal Model Specification

- $\boxed{\textbf{Definition:}}$ In model specification, **optimal** $=$ {conditioning only on propositions rendered true by the **context** of the problem and the design of the data-gathering process, while at the same time ensuring that the set of conditioning propositions includes **all relevant problem context**}.
- $\boxed{\textbf{Q:}}$ Is optimal model specification **possible**?
- $\boxed{\textbf{A:}}$ Yes, **sometimes**; for instance, **Bayesian non-parametric modeling** is an important approach to model specification optimality.
- $\boxed{\textbf{Example (part I of the talk):}}$ A/B **testing** (randomized controlled experiments) in **data science**.
  - **eCommerce** company X interacts with users through its **web site**; the company is constantly interested in **improving** its web experience, so (without telling the users) it **randomly assigns** them to **treatment** (A: a new variation on (e.g.) how information is presented) or **control** (B: the current best version of the web site) groups.

A/B Testing

- Let $\mathcal{P}$ be the **population** of company X users at time $(now + \Delta)$, in which Δ is fairly small (e.g., several months).
- In a typical A/B test, $(n^C + n^T)$ users are **sampled randomly** from a **proxy** for $\mathcal{P}$ — the population of company X users at time *now* — with n^C of these users **assigned at random** to C and n^T to T.
- The experimental users are **monitored** for k weeks (typically $2 \leq k \leq 6$), and a summary $y \in \mathbb{R}$ of their use of the web site (aggregated over the k weeks) is chosen as the **principal outcome variable**; often y is either **monetary** or measures **user satisfaction**; typically $y \geq 0$, which I assume in what follows.
- Let y_i^C be the **outcome value** for user i in C, and let y^C be the vector (of length n^C) of all C values; define y_j^T and y^T (of length n^T) analogously; Your **total data set** is then $D = (y^C, y^T)$.
- **Before** the data set arrives, Your **uncertainty** about the y_i^C and y_j^T values is **conditionally exchangeable** given the **experimental group indicators** $I = (1 \text{ if } T, 0 \text{ if } C)$.

Bayesian Non-Parametric Modeling

- Therefore, by **de Finetti's most important Representation Theorem**, Your **predictive uncertainty** about D is **expressible hierarchically** as

$$\begin{array}{ccc|ccc} (F^C \,|\, \mathcal{B}) & \sim & p(F^C \,|\, \mathcal{B}) & (F^T \,|\, \mathcal{B}) & \sim & p(F^T \,|\, \mathcal{B}) \\ (y_i^C \,|\, F^C \, \mathcal{B}) & \stackrel{IID}{\sim} & F^C & (y_j^T \,|\, F^T \, \mathcal{B}) & \stackrel{IID}{\sim} & F^T \end{array} \qquad (1)$$

- Here F^C is the **empirical CDF** of the y values You would see in

  *the population $\mathcal{P}$ to which You're interested in **generalizing inferentially***

  if all users in $\mathcal{P}$ were to receive the C version of the web experience, and F^T is the analogous empirical CDF if instead those same users were to **counterfactually** receive the T version.

- Assume that the means $\mu^C = \int y \, dF^C(y)$ and $\mu^T = \int y \, dF^T(y)$ **exist** and are **finite**, and define

$$\theta \triangleq \frac{\mu^T - \mu^C}{\mu^C} \, ; \qquad (2)$$

  in eCommerce this is referred to as the **lift** caused by the treatment.

Optimal Bayesian Model Specification

$$\begin{array}{ccc|ccc} (F^C \mid \mathcal{B}) & \sim & p(F^C \mid \mathcal{B}) & (F^T \mid \mathcal{B}) & \sim & p(F^T \mid \mathcal{B}) \\ (y_i^C \mid F^C \, \mathcal{B}) & \stackrel{IID}{\sim} & F^C & (y_j^T \mid F^T \, \mathcal{B}) & \stackrel{IID}{\sim} & F^T \end{array}$$

- I claim that this is an instance of **optimal Bayesian model specification**: this **Bayesian non-parametric (BNP) model** arises from **exchangeability** assumptions implied directly by **problem context**.
- I now **instantiate** this model with **Dirichlet process priors** placed directly on the **data scale**:

$$\begin{array}{ccc|ccc} (F^C \mid \mathcal{B}) & \sim & DP(\alpha^C, F_0^C) & (F^T \mid \mathcal{B}) & \sim & DP(\alpha^T, F_0^T) \\ (y_i^C \mid F^C \, \mathcal{B}) & \stackrel{IID}{\sim} & F^C & (y_j^T \mid F^T \, \mathcal{B}) & \stackrel{IID}{\sim} & F^T \end{array} \qquad (3)$$

- The usual **conjugate updating** produces the **posterior**

$$(F^C \mid y^C \, \mathcal{B}) \sim DP\left(\alpha^C + n^C, \frac{\alpha^C \, F_0^C + n \, \hat{F}_n^C}{\alpha^C + n^C}\right) \qquad (4)$$

and analogously for F^T, where $\hat{F}_n^C$ is the **empirical CDF** defined by the control group data vector y^C; these posteriors for F^C and F^T **induce posteriors** for μ^C and μ^T, and thus for θ.

$$(F^C \,|\, y^C \,\mathcal{B}) \sim DP\left(\alpha^C + n^C, \frac{\alpha^C F_0^C + n^C \hat{F}_n^C}{\alpha^C + n^C}\right) .$$

- How to **specify** $(\alpha^C, F_0^C, \alpha^T, F_0^T)$? In part 2 of the talk I'll describe a **method** for **incorporating** C information from other experiments; in eCommerce it's **controversial** to **combine information** across T groups; so here I'll present an analysis in which **little information external** to (y^C, y^T) is available.
- This **corresponds** to α^C and α^T values close to 0, and — with the **large** n^C and n^T values typical in A/B testing and $\alpha^C \doteq \alpha^T \doteq 0$ — it **doesn't matter** what You take for F_0^C and F_0^T; in the **limit** as $(\alpha^C, \alpha^T) \downarrow 0$ You get the posteriors

$$(F^C \,|\, y^C \,\mathcal{B}) \sim DP\left(n^C, \hat{F}_n^C\right) \quad (F^T \,|\, y^T \,\mathcal{B}) \sim DP\left(n^T, \hat{F}_n^T\right) . \quad (5)$$

In my view the $DP\left(n, \hat{F}_n\right)$ posterior should get **far more use** in **Bayesian data science** at **Big-Data scale** than it now does: it **arises directly from problem context** in many settings, and (next slide) is **readily computable**.

Fast DP Posterior Simulation at Large Scale

$$(F^C \mid y^C \mathcal{B}) \sim DP\left(n^C, \hat{F}_n^C\right) \quad (F^T \mid y^T \mathcal{B}) \sim DP\left(n^T, \hat{F}_n^T\right).$$

- How to **quickly simulate** F draws from $DP\left(n, \hat{F}_n\right)$ when n is large (e.g., $O\left(10^7\right)$ or more)? You can of course use **stick-breaking** (Sethuramen 1994), but this is **slow** because the size of the next stick fragment **depends sequentially** on how much of the stick has already been allocated.

- Instead, use the **Pólya Urn representation** of the **DP predictive distribution** (Blackwell and MacQueen 1973): having observed $y = (y_1, \dots, y_n)$ from the model $(F \mid \mathcal{B}) \sim DP(\alpha, F_0)$, $(y_i \mid F \mathcal{B}) \stackrel{IID}{\sim} F$, by **marginalizing** over F You can show that to make a **draw** from the **posterior predictive** for y_{n+1} You just sample from $\hat{F}_n$ with probability $\frac{n}{\alpha+n}$ (and from F_0 with probability $\frac{\alpha}{\alpha+n}$); as $\alpha \downarrow 0$ this becomes simply **making a random draw** from $(y_1, \dots, y_n)$; and it turns out that, to make an F draw from $(F \mid y \, \mathcal{B})$ that **stochastically matches** what You would get from stick-breaking, You just make n IID draws from $(y_1, \dots, y_n)$ and form the **empirical CDF** based on these draws.

The Frequentist Bootstrap in BNP Calculations

- **This is precisely the frequentist bootstrap** (Efron 1979), which turns out to be about **30 times faster** than stick-breaking and is **embarrassingly parallelizable** to boot (e.g., Alex Terenin tells me that this is **ludicrously easy** to implement in `MapReduce`).

- Therefore, to **simulate** from the **posterior** for θ in this model: for large M

  (1) Take M independent **bootstrap samples** from y^C, calculating the **sample means** μ_*^C of each of these bootstrap samples;

  (2) **Repeat** (1) on y^T, obtaining the vector μ_*^T of length M; and

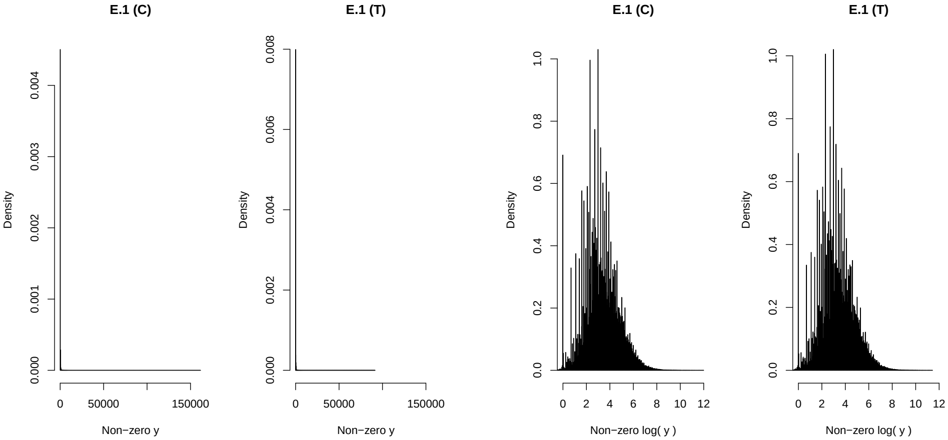
  (3) Make the **vector calculation** $\theta_* = \frac{\mu_*^T - \mu_*^C}{\mu_*^C}$.

- I claim that this is an **essentially optimal Bayesian analysis** (the only assumption not driven by **problem context** was the choice of the **DP prior**, when other BNP priors are available).

- $\boxed{\textbf{Case Studies:}}$ **Two experiments** at company X, conducted a few years ago; E_1 involved about **24.5 million users**, and E_2 about **257,000 users**; in both cases the outcome y was **monetary**, expressed here in **Monetary Units (MUs)**, a **monotonic increasing transformation** of US$.

Visualizing E_1

- In both C and T in E_1, **90.7%** of the users had $y = \mathbf{0}$, but the remaining **non-zero values** ranged up to **162,000**.



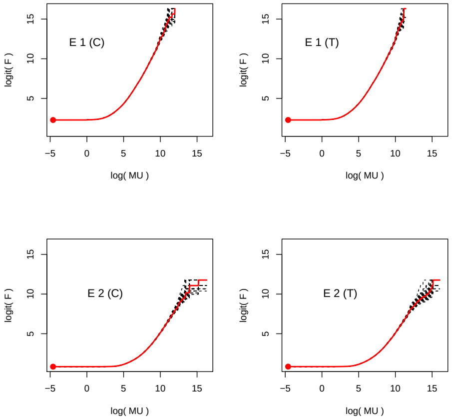
Numerical Summaries of E_1 and E_2

Descriptive summaries of a monetary outcome y measured in two A/B tests E_1 and E_2 at eCommerce company X; SD = standard deviation.

Experiment	n	% 0	MU Mean	MU SD	Skewness	Kurtosis
E_1: T	12,234,293	90.7	9.128	129.7	157.6	59,247
E_1: C	12,231,500	90.7	9.203	147.8	**328.9**	**266,640**
E_2: T	128,349	70.1	**1,080.8**	**33,095.8**	205.9	52,888
E_2: C	128,372	70.0	**1,016.2**	**36,484.9**	289.1	92,750

- The outcome y in C in E_1 had **skewness 329** (Gaussian 0) and **kurtosis 267,000** (Gaussian 0); the noise-to-signal ratio (SD/mean) in C in E_2 was **36**.
- The **estimated lift** in E_1 was $\hat{\theta} = \frac{9.128 - 9.203}{9.203} \doteq$ **−0.8%** (i.e., if anything T **made things worse**); in E_2, $\hat{\theta} = \frac{1080.8 - 1016.2}{1016.2} \doteq$ **+6.4%** (**highly promising**), but the **between-user variability** in the outcome y in E_2 was **massive** (SDs in C and T on the order of **36,000**).

Sampling from The Posteriors For F^C and F^T



In E_1, with $n =$ **12 million** in each group, posterior uncertainty about F **does not begin to exhibit itself** (reading left to right) **until about** $e^9 \doteq 8{,}100$ MUs, which corresponds to the $\text{logit}^{-1}(\ 10\ ) =$ **99.9995th percentile**; but with the **mean at stake** and **violently skewed and kurtotic distributions**, **extremely high percentiles** are precisely the distributional locations of **greatest leverage**.

What Does The Central Limit Theorem Have To Say?

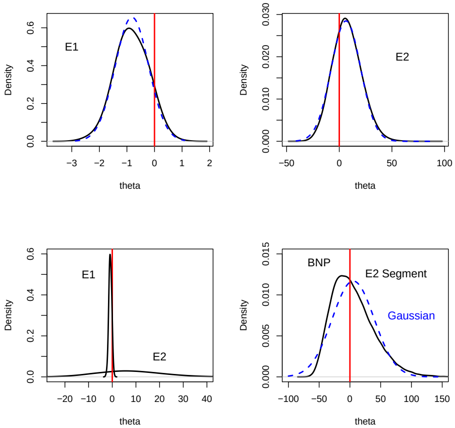
- $\hat{\theta}$ is driven by the **sample means** $\bar{y}^C$ and $\bar{y}^T$, so with **large enough sample sizes** the posterior for θ will be **close to Gaussian** (by the Bayesian CLT), rendering the **bootstrapping unnecessary**, but the **skewness** and **kurtosis** values for the outcome y are **large**; when does the **CLT kick in**?
- **Not-widely-known fact**: under **IID sampling**,

$$\text{skewness}(\bar{y}_n) = \frac{\text{skewness}(y_1)}{\sqrt{n}} \quad \text{and} \quad \text{kurtosis}(\bar{y}_n) = \frac{\text{kurtosis}(y_1)}{n} . \tag{6}$$

E_1 (C)

n	skewness($\bar{y}_n$)	kurtosis($\bar{y}_n$)
1	328.9	266,640.0
10	104.0	26,664.0
100	32.9	2,666.4
1,000	10.4	266.6
10,000	3.3	26.7
100,000	1.0	2.7
1,000,000	0.3	0.3
10,000,000	0.1	0.0

Exact and Approximate Posteriors for θ



BNP posterior distributions (solid curves) for the **lift** θ in E_1 (upper left) and E_2 (upper right), with **Gaussian approximations** (dotted lines) superimposed; lower left: the θ **posteriors** from E_1 and E_2 on the same graph, to give a sense of **relative information content** in the two experiments; lower right: BNP and approximate-Gaussian posteriors for θ in a **small subgroup (segment)** of E_2.

eCommerce Conclusions

BNP inferential summaries of lift in the two A/B tests E_1 and E_2.

Experiment	Total n	Posterior for θ (%) Mean	SD	$P(\theta > 0 \mid y^T \, y^C \, \mathcal{B})$ BNP	Gaussian
E_1	24,465,793	-0.818	0.608	0.0894	0.0892
E_2 full	256,721	$+6.365$	14.01	0.6955	0.6752
E_2 segment	23,674	$+5.496$	34.26	0.5075	0.5637

The **bottom row** of this table presents the **results** for a **small subgroup** (known in eCommerce as a **segment**) of users in E_2, identified by a particular set of **covariates**; the combined sample size here is "only" about **24,000**, and the **Gaussian approximation** to $P(\theta > 0 \mid y^T \, y^C \, \mathcal{B})$ is **too high by more than 11%**.

From a **business perspective**, the **treatment intervention** in E_1 was demonstrably a **failure**, with an estimated lift that represents a **loss** of about **0.8%**; the treatment in E_2 was **highly promising** — $\hat{\theta} \doteq$ **+6.4%** — but (with an outcome variable this **noisy**) the total sample size of "only" about **257,000** was **insufficient** to demonstrate its effectiveness **convincingly**.

Combining Information Across Similar Control Groups

$\boxed{\textbf{NB}}$ In the **Gaussian approximation**, the posterior for θ is Normal with mean $\hat{\theta} = \frac{\bar{y}^T - \bar{y}^C}{\bar{y}^C}$ and (by **Taylor expansion**)

$$SD(\theta \,|\, y^T \, y^C \, \mathcal{B}) \doteq \sqrt{\frac{\bar{y}_T^2 \, s_C^2}{\bar{y}_C^4 \, n_C} + \frac{s_T^2}{\bar{y}_C^2 \, n_T}} \, . \tag{7}$$

- $\boxed{\textbf{Example (part II of the talk):}}$ **Borrowing strength across similar control groups**.
- In practice **eCommerce company** X runs a number of experiments **simultaneously**, making it possible to consider a **modeling strategy** in which T data in experiment E is compared with a **combination** of $\{C$ data from E plus data from **similar** C groups in **other experiments**$\}$.
- **Suppose therefore** that You **judge** control groups $(C_1, \ldots, C_N)$ **exchangeable** — not directly **poolable**, but **like random draws** from a **common** C **reservoir** (as with **random-effects hierarchical models**, in which **between-group heterogeneity** among the C_i is **explicitly acknowledged**).

BNP For Combining Information

- An **extension** of the **BNP modeling** in part I to accommodate this new **borrowing of strength** would look like this: for $i = 1, \ldots, N$ and $j = 1, \ldots, n_{group}$,

$$
\begin{array}{ccc}
(F^T \mid \mathcal{B}) & \sim & DP(\alpha^T, F_0^T) \\
(y_j^T \mid F^T \, \mathcal{B}) & \stackrel{IID}{\sim} & F^T
\end{array}
\quad \Bigg| \quad
\begin{array}{ccc}
(F_0^C \mid \mathcal{B}) & \sim & DP(\gamma, G) \\
(F^{C_i} \mid F_0^C \, \mathcal{B}) & \stackrel{IID}{\sim} & DP(\alpha^C, F_0^C) \\
(y_j^{C_i} \mid F^{C_i} \, \mathcal{B}) & \stackrel{IID}{\sim} & F^{C_i}
\end{array}
\qquad (8)
$$

- The **modeling** in the C groups is an example of a **hierarchical Dirichlet process** (Teh, Jordan, Beal and Blei 2005).
- I've not yet **implemented** this model; with the **large sample sizes** in eCommerce, $DP\left(n, \hat{F}_n\right)$ will again be **central**, and some version of **frequentist bootstrapping** will again do the calculations **quickly**.
- **Suppose** for the rest of the talk that the **sample sizes** are large enough for the **Gaussian approximation** in part I to hold:

$$
(\mu^T \mid y^T \, \mathcal{B}) \,\dot{\sim}\, N\left[\bar{y}^T, \frac{(s^T)^2}{n^T}\right] \quad \text{and} \quad (\mu^{C_i} \mid y^{C_i} \, \mathcal{B}) \,\dot{\sim}\, N\left[\bar{y}^{C_i}, \frac{(s^{C_i})^2}{n^{C_i}}\right] . \qquad (9)
$$

Approximate BNP With 100 Million Observations

$$(\mu^T \mid y^T\,\mathcal{B}) \mathrel{\dot\sim} N\left[\bar{y}^T, \frac{(s^T)^2}{n^T}\right] \quad \text{and} \quad (\mu^{C_i} \mid y^{C_i}\,\mathcal{B}) \mathrel{\dot\sim} N\left[\bar{y}^{C_i}, \frac{(s^{C_i})^2}{n^{C_i}}\right]$$

With n^T and the $n^{C_i} \doteq$ **10 million** each and (e.g.) $N \doteq 10$, the above equation represents a **fully efficient summary** of an **approximate BNP analysis** of $O($**100 million**$)$ observations.

- Now simply **turn** the above Gaussian relationships **around** to **induce** the **likelihood function** in a **hierarchical Gaussian random-effects model** (the **sample sizes** are **so large** that the within-groups **sample SDs** (e.g., s^T) can be regarded as **known**):

$$\begin{array}{rcl} (\mu^T \mid \mathcal{B}) & \propto & 1 \\ (\bar{y}^T \mid \mu^T\,\mathcal{B}) & \sim & N\left[\mu^T, \frac{(s^T)^2}{n^T}\right] \end{array} \quad \Bigg| \quad \begin{array}{rcl} (\sigma \mid \mathcal{B}) & \sim & U(0, A) \\ (\mu^C \mid \sigma\,\mathcal{B}) & \propto & 1 \\ (\mu^{C_i} \mid \mu^C\,\sigma\,\mathcal{B}) & \stackrel{IID}{\sim} & N(\mu^C, \sigma^2) \\ (\bar{y}^{C_i} \mid \mu^{C_i}\,\mathcal{B}) & \sim & N\left[\mu^{C_i}, \frac{(s^{C_i})^2}{n^{C_i}}\right] \end{array} \qquad (10)$$

- The **Uniform**$(0, A)$ **prior** on the between-C-groups SD σ has been shown (e.g., Gelman 2006) to have **good calibration** properties (choose A just large enough to **avoid likelihood truncation**).

In Spiegelhalter's Honor

```
{

  eta.C ~ dflat( )
  sigma.mu.C ~ dunif( 0.0, A )
  mu.T ~ dflat( )

  y.bar.T ~ dnorm( mu.T, tau.mu.T )

  for ( i in 1:N ) {

    y.bar.C[ i ] ~ dnorm( mu.C[ i ], tau.y.bar.C[ i ] )
    mu.C[ i ] ~ dnorm( eta.C, tau.mu.C )

  }

  tau.mu.C <- 1.0 / ( sigma.mu.C * sigma.mu.C )

  theta <- ( mu.T - eta.C ) / eta.C
  theta.positive <- step( theta )

}
```

One C Group First

```
list( A = 0.001,
      y.bar.T = 9.286,
      tau.mu.T = 727.28,
      N = 1,
      y.bar.C = c( 9.203 ),
      tau.y.bar.C = c( 559.94 )
    )

list( eta.C = 9.203,
      sigma.mu.C = 0.0,
      mu.T = 9.286
    )
                          y              mu                  theta
group         n  mean    sd   mean      sd       mean       sd positive

    T  12234293 9.286 129.7  9.286 0.03708
    C  12231500 9.203 147.8  9.203 0.04217  0.008904 0.006165   0.9276
```

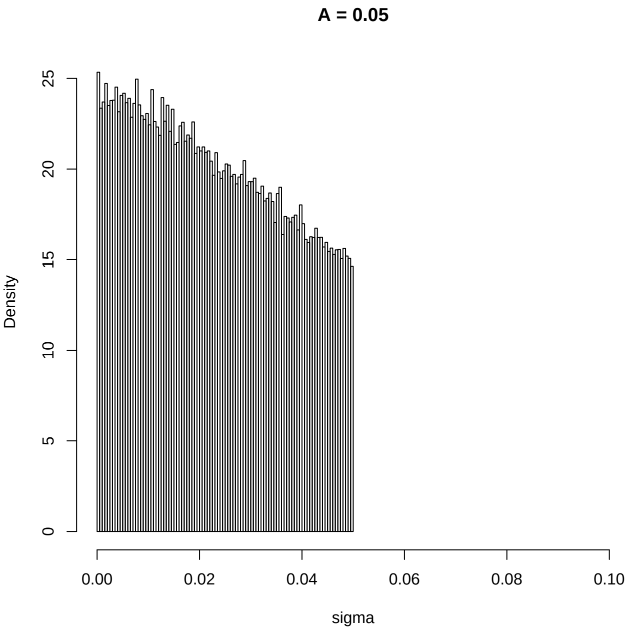
- Start with **one** C **group**: **simulated data** similar to E_1 in part I but with a **bigger treatment effect** — total sample size **24.5 million**, $\bar{y}^T = 9.286$, $\bar{y}^C = 9.203$, $\hat{\theta} =$**+0.9%** with posterior SD **0.6%**, **posterior probability of positive effect 0.93**.

Two C Groups

		y		mu		theta		
group	n	mean	sd	mean	sd	mean	sd	positive
T	12234293	9.286	129.7	9.286	0.03704			
C1	12231500	9.203	147.8	9.203	0.03263			
C2	12232367	9.204	140.1	9.204	0.03196			
C	24463867	---	---	9.204	0.03458	0.008973	0.005538	0.9487

- Now **two** C **groups**, chosen to be **quite homogeneous** (group means 9.203 and 9.204, simulated from $\sigma = \mathbf{0.01}$) — with **truncation point** $A = 0.05$ in the **Uniform prior** for σ, the **posterior mean** for θ is **about the same** as before (**+0.9%**) but the posterior SD has **dropped** from **0.61%** to **0.55%** (**strength is being borrowed**), and the **posterior probability** of a **positive effect** has risen to **95%**.
- However, has $A = 0.05$ **inadvertently truncated the likelihood** for σ?

Yes

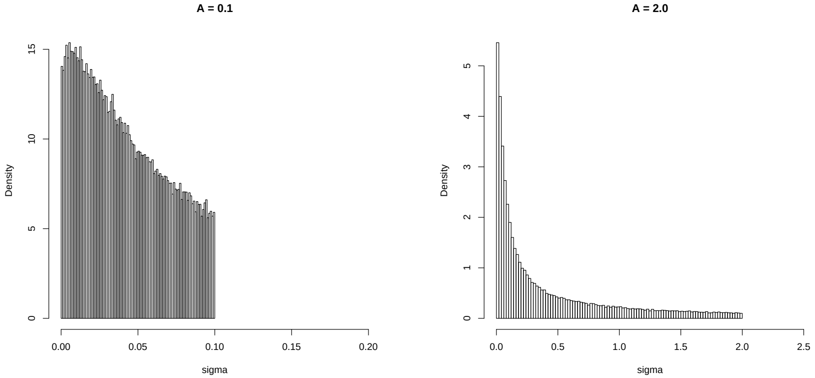


$A = 0.1$: Borrowing Strength Seems to Disappear

		y		mu		theta		
group	n	mean	sd	mean	sd	mean	sd	positive
T	12234293	9.286	129.7	9.286	0.03704			
C1	12231500	9.203	147.8	9.203	0.03535			
C2	12232367	9.204	140.1	9.204	0.03426			
C	24463867	---	---	9.203	0.04563	0.009011	0.006434	0.9231

- With $A = 0.1$, the **posterior SD** for θ rises to **0.64%**, and the posterior probability of a positive lift (**92%**) is now **smaller than when only one** C **group was used** — the borrowing of strength **seems to have disappeared**.
- Moreover, $A = 0.1$ **still leads to truncation**; exploration reveals that **truncation** doesn't start to become **negligible** until $A \geq$ **2.0** (and remember that the **actual value** of σ in this simulated data set was **0.01**).

You Can Get Anything You Want ...



		y		mu		theta		
group	n	mean	sd	mean	sd	mean	sd	positive
T	12234293	9.286	129.7	9.286	0.03704			
C1	12231500	9.203	147.8	9.203	0.03981	(this is with A = 2.0)		
C2	12232367	9.204	140.1	9.204	0.03794			
C	24463867	---	---	9.204	0.4691	0.01164	0.05475	0.7341

Between-C-Groups Heterogeneity

- The **right way** to set A (I haven't done this yet) is via **inferential calibration** on the **target quantity** of interest θ: create a **simulation environment** identical to the real-world setting ($n^T =$ 12,234,293; $n^{C_1} =$ 12,231,500; $n^{C_2} =$ 12,232,367; $s^T = 0.03704$; $s^{C_1} = 0.03981$; $s^{C_2} = 0.03794$) except that $(\mu^T, \mu^C, \theta, \sigma)$ are **known** to be $(9.286; 9.203; 0.90\%; 0.01)$ — now **simulate many data sets** from the **hierarchical model** in equation (10) on page 19 and **vary** A until the $100(1-\eta)\%$ **posterior intervals** for θ include the **right answer** about $100(1-\eta)\%$ of the time for a **broad range** of η values.

- Even when A has been **correctly calibrated**, when the **number** N of C groups being combined is **small** it doesn't take much **between-group heterogeneity** for the model to tell You that **You have more uncertainty** about θ with 2 control groups than with 1.

Between-C-Groups Heterogeneity (continued)

		y		mu		theta		
group	n	mean	sd	mean	sd	mean	sd	positive
T	12234293	9.286	129.7	9.286	0.03704			
C1	12231500	9.203	147.8	9.203	0.03263	(here sigma = 0.01)		
C2	12232367	9.204	140.1	9.204	0.03196			
C	24463867	---	---	9.204	0.03458	0.008973	0.005538	0.9487
C1	12231500	9.203	147.8	9.209	0.03542			
C2	12232367	9.222	140.1	9.217	0.03426	(here sigma = 0.015)		
C	24463867	---	---	9.213	0.04543	0.007976	0.006391	0.8983

- In the **top part** of the table above with $\sigma = \mathbf{0.01}$, **borrowing strength decreased the posterior SD** from its value with only 1 C group, but in the **bottom part** of the table — with σ only slightly larger at **0.015** — there was enough **heterogeneity** to **drop** the tail area from **92.8%** (1 C group) to **89.8%**.

$N = 10$ C Groups, Small Heterogeneity

		y		mu		theta		
group	n	mean	sd	mean	sd	mean	sd	positive
T	12234293	9.286	129.7	9.286	0.03708			
C	12231500	9.203	147.8	9.203	0.04217	0.008904	0.006165	0.9276
C1	12232834	9.193	144.6	9.202	0.01823			
C2	12233905	9.204	141.4	9.204	0.01807			
C3	12232724	9.191	143.9	9.202	0.01817			
C4	12232184	9.222	139.7	9.205	0.01821			
C5	12231697	9.206	139.3	9.204	0.01803			
C6	12231778	9.191	144.0	9.202	0.01825			
C7	12232383	9.208	130.1	9.204	0.01769	(here sigma = 0.01)		
C8	12232949	9.211	138.3	9.204	0.01805			
C9	12233349	9.209	143.0	9.204	0.01808			
C10	12232636	9.197	142.2	9.203	0.01811			
C	122326439	---	---	9.203	0.01391	0.008974	0.004299	0.9817

- Here with $N = 10$ C **groups** and a **small amount** of between–C–groups **heterogeneity** ($\sigma = 0.01$), borrowing strength leads to a **substantial sharpening** of the T versus C comparison (the problem of setting A **disappears**, because the posterior for σ is now **quite concentrated**) (NB **total sample size** is now **135 million**).

$N = 10$ C Groups, Large Heterogeneity

		y		mu		theta		
group	n	mean	sd	mean	sd	mean	sd	positive
T	12234293	9.286	129.7	9.286	0.03708			
C	12231500	9.203	147.8	9.203	0.04217	0.008904	0.006165	0.9276
C1	12232834	9.082	144.6	9.094	0.03996			
C2	12233905	9.211	141.4	9.210	0.03867			
C3	12232724	9.048	143.9	9.063	0.03984			
C4	12232184	9.437*	139.7	9.416	0.03981			
C5	12231697	9.235	139.3	9.232	0.03818			
C6	12231778	9.050	144.0	9.065	0.03996			
C7	12232383	9.260	130.1	9.255	0.03592	(here sigma = 0.125)		
C8	12232949	9.300*	138.3	9.291	0.03818			
C9	12233349	9.274	143.0	9.267	0.03911			
C10	12232636	9.133	142.2	9.140	0.03888			
C	122326439	---	---	9.203	0.04762	0.009052	0.006589	0.9195

- With $N = 10$ it's possible to **"go backwards"** in apparent information about θ because of **large heterogeneity** ($\sigma = 0.125$ above), but only by making the heterogeneity **so large** that the exchangeability judgment is **questionable** (the 2 C groups marked $*$ actually had means that were **larger** than the T mean).

Conclusions in Part II

- With **large sample sizes** it's straightforward to use **hierarchical random-effects Gaussian models** — as good **approximations** to a **full BNP analysis** — in combining C groups to **improve accuracy** in estimating T effects, but
  - When the number N of C groups to be combined is **small**, the results are **extremely sensitive** to Your prior on the between–C–groups SD σ, and it doesn't take much heterogeneity among the C means for the model to tell You that **You know less about θ than when there was only 1 C group**, and
  - With a **larger** N there's **less sensitivity** to the prior for σ, and **borrowing strength** will generally **succeed** in sharpening the comparison unless the **heterogeneity** is so large as to make the **exchangeability judgment** that led to the C–group combining **questionable**.