AMS 206: Classical and Bayesian Inference

David Draper

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

draper@ucsc.edu
www.ams.ucsc.edu/∼draper

LECTURE NOTES (PART 1)

An Example, to Fix Ideas

Case Study 1. (**Krnjajić**, **Kottas**, **Draper** 2008): **In-home geriatric assessment (IHGA)**. In an **experiment** conducted in the **1980s** (**Hendriksen** et al., 1984), **572 elderly people**, **representative** of $\mathcal{P} = \{$all **non-institutionalized elderly people** in **Denmark**$\}$, were **randomized**, **287** to a **control** (C) group (who received **standard health care**) and **285** to a **treatment** (T) group (who received **standard care plus IHGA**: a kind of **preventive medicine** in which **each person's medical** and **social needs** were **assessed** and **acted upon individually**).

One **important outcome** was the **number of hospitalizations** during the **two-year** life of the study:

	Number of Hospitalizations						
Group	0	1	...	m	n	Mean	SD
Control	n_{C0}	n_{C1}	...	n_{Cm}	$n_C = 287$	$\bar{y}_C$	s_C
Treatment	n_{T0}	n_{T1}	...	n_{Tm}	$n_T = 285$	$\bar{y}_T$	s_T

Let μ_C and μ_T be the **mean hospitalization rates** (per two years) in $\mathcal{P}$ under the C and T **conditions**, respectively.

Here are **four statistical questions** that **arose** from **this study**:

The Four Principal Statistical Activities

$\boxed{Q_1:}$ Was the **mean number of hospitalizations per two years** in the **IHGA** group **different from** that in **control** by an **amount** that was **large** in **practical** terms? $\left[\right.$**description** involving $\left(\frac{\bar{y}_T - \bar{y}_C}{\bar{y}_C}\right)\left.\right]$

$\boxed{Q_2:}$ Did **IHGA (causally) change** the **mean number of hospitalizations per two years** by an **amount** that was **large** in **statistical** terms? $\left[\right.$**inference** about $\left(\frac{\mu_T - \mu_C}{\mu_C}\right)\left.\right]$

$\boxed{Q_3:}$ On the **basis** of **this study**, how **accurately** can You **predict** the **total decrease** in **hospitalizations** over a **period** of N years if **IHGA** were **implemented throughout Denmark**? **[prediction]**

$\boxed{Q_4:}$ On the **basis** of **this study**, is the **decision** to **implement IHGA throughout Denmark optimal** from a **cost-benefit point of view**? **[decision-making]**

These **questions encompass** almost all of the **discipline** of **statistics**: **describing** a **data set** D, **generalizing outward inferentially** from D, **predicting new data** D^*, and **helping** people **make decisions** in the **presence** of **uncertainty** (I include **sampling/experimental design** under **decision-making**; **omitted: data wrangling**, ...).

An Informal Axiomatization of Statistics

$\boxed{1}$ **(definition) Statistics** is the study of **uncertainty**: how to **measure it well**, and how to **make good choices** in the face of it.

$\boxed{2}$ **(definition) Uncertainty** is a state of **incomplete information** about something of interest to **You** (Good, 1950: a **generic person** wishing to **reason sensibly** in the presence of **uncertainty**).

$\boxed{3}$ **(axiom)** (**Your uncertainty** about) **"Something of interest to You"** can always be **expressed** in terms of **propositions**: **true/false** statements $A, B, \ldots$

Examples: You may be **uncertain** about the **truth status** of

- $A =$ (**Donald Trump** will be **re-elected U.S. President** in **2020**), or

- $B =$ (the **in-hospital mortality rate** for patients at **hospital** H admitted in **calendar 2010** with a principal diagnosis of **heart attack** was **between 5% and 25%**).

$\boxed{4}$ **(implication)** It **follows** from $\boxed{1}$–$\boxed{3}$ that **statistics** concerns **Your information** (**NOT** Your **beliefs**) about $A, B, \ldots$

Axiomatization (continued)

$\boxed{5}$ **(axiom)** But **Your information** cannot be **assessed** in a **vacuum**: all such **assessments** must be made **relative to (conditional on)** Your **background assumptions** and **judgments** about **how the world works** vis à vis $A, B, \ldots$.

$\boxed{6}$ **(axiom)** These **assumptions** and **judgments**, which are themselves a form of **information**, can always be **expressed** in a finite **set** $\mathcal{B} = \{B_1, \ldots, B_b\}$ of **propositions** (**examples** below).

$\boxed{7}$ **(definition)** Call the **"something of interest to You"** θ; in **applications** θ is often a **vector** (or **matrix**, or **array**) of **real numbers**, but **in principle** it could be **almost anything** (a **function**, an **image** of the surface of Mars, a **phylogenetic tree**, ...).

$\boxed{8}$ **(axiom)** There will typically be an **information source (data set)** D that You judge to be **relevant** to **decreasing** Your uncertainty about θ; in **applications** D is often again a **vector** (or **matrix**, or **array**) of **real numbers**, but **in principle** it too could be **almost anything** (a **movie**, the **words** in a **book**, ...).

Axiomatization (continued)

Examples of $\mathcal{B}$:

- If θ is the **mean survival time** for a **specified group of patients** (who are **alive** now), then $\mathcal{B}$ includes the **proposition** ($\theta \geq 0$).

- If D is the result of an **experiment** E, then $\mathcal{B}$ might include the **proposition** (Patients were **randomized** into one of two groups, **treatment (new drug)** or **control (current best drug)**).

$\boxed{9}$ **(implication)** The **presence** of D creates a **dichotomy**:

- **Your information** about θ {**internal, external**} to D.

(People often talk about a **different dichotomy**: **Your information** about θ {**before, after**} D arrives **(prior, posterior)**, but **temporal considerations** are actually **irrelevant**.)

$\boxed{10}$ **(implication)** It **follows** from $\boxed{1}$–$\boxed{9}$ that **statistics** concerns itself principally with **five things** (omitted: **description**, **data wrangling**, ...):

(1) **Quantifying Your information** about θ **internal** to D (given $\mathcal{B}$), and doing so **well** (this term is **not yet defined**);

Foundational Question

(2) **Quantifying Your information** about θ **external** to D (given $\mathcal{B}$), and doing so **well**;

(3) **Combining** these two **information sources** (and doing so **well**) to create a **summary** of **Your uncertainty** about θ (given $\mathcal{B}$) that includes **all available information** You judge to be **relevant** (this is **inference**);

and using **all Your information** about θ (given $\mathcal{B}$) to make

(4) **Predictions** about **future** data values D^* and

(5) **Decisions** about how to **act sensibly**, even though **Your information** about θ may be **incomplete**.

Foundational question: How should these tasks be **accomplished**?

This question has been addressed by **Bruno de Finetti**, in work he did from the 1920s through the 1970s, and by the American physicists **Richard T. Cox** (1946) and **Edwin T. Jaynes** (2002).

The Cox–Jaynes **Theorem** — recently rigorized and extended by Terenin and Draper (2015) — says that

The Big Picture (continued)

- If You're prepared to **uniquely** specify two probability distributions — $p(\theta \,|\, \mathcal{B})$, encoding Your information about θ **external** to D, and $p(D \,|\, \theta \, \mathcal{B})$, capturing Your information about θ **internal** to D — then
  - **optimal inference** about θ is based on the distribution

    $$p(\theta \,|\, D \, \mathcal{B}) \propto p(\theta \,|\, \mathcal{B}) \, p(D \,|\, \theta \, \mathcal{B}) \tag{1}$$

    (here **optimal** $=$ {**all** relevant information is **used appropriately**, and **no** other "information" is **inadvertently smuggled in**}), and
  - **optimal prediction** of new data D^* is based on the distribution

    $$p(D^* \,|\, D \, \mathcal{B}) = \int_\Theta p(D^* \,|\, \theta D \, \mathcal{B}) \, p(\theta \,|\, D \, \mathcal{B}) \, d\theta, \tag{2}$$

    where Θ is the set of possible values of θ;

Optimal Model Specification

- and if You're further prepared to **uniquely** specify two more ingredients — Your action space $a \in (\mathcal{A} \,|\, \mathcal{B})$ and Your utility function $U(a, \theta \,|\, \mathcal{B})$ — then **optimal decision-making** is attained by **maximizing expected utility**:

$$a^* = \operatorname*{argmax}_{a \in (\mathcal{A} \,|\, \mathcal{B})} \int_\Theta U(a, \theta \,|\, \mathcal{B}) \, p(\theta | D \, \mathcal{B}) \, d\theta \, . \tag{3}$$

- Let's agree to call $M = \{ p(\theta \,|\, \mathcal{B}), p(D \,|\, \theta \, \mathcal{B}) \}$ Your **model** for Your uncertainty about θ and D^*, and $M_d = \{ p(\theta \,|\, \mathcal{B}), p(D \,|\, \theta \, \mathcal{B}), (\mathcal{A} \,|\, \mathcal{B}), U(a, \theta \,|\, \mathcal{B}) \}$ Your **model** for Your decision uncertainty.
- The two main **practical challenges** in using this Theorem are
  - (technical) **Integrals** arising in **computing** the inferential and predictive distributions and the expected utility may be difficult to approximate accurately (and the action space may be difficult to **search** well), and
  - (substantive) The mapping from the problem $\mathbb{P} = (\mathbb{Q}, \mathbb{C})$ — $\mathbb{Q}$ = **questions**, $\mathbb{C}$ = **context** — to $M = \{ p(\theta \,|\, \mathcal{B}), p(D \,|\, \theta \, \mathcal{B}) \}$ and $M_d = \{ p(\theta \,|\, \mathcal{B}), p(D \,|\, \theta \, \mathcal{B}), (\mathcal{A} \,|\, \mathcal{B}), U(a, \theta \,|\, \mathcal{B}) \}$ is **rarely unique**, giving rise to **model uncertainty**.

Data-Science Example: A/B Testing

- **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**}.
- **Q:** Is optimal model specification **possible**?
- **A:** Yes, **sometimes**; for instance, **Bayesian non-parametric modeling** is an important approach to model specification optimality.
- **Case Study 2:** 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

$$\begin{array}{ccc|ccc} (F^C \mid \mathcal{B}) & \sim & p(F^C \mid \mathcal{B}) & (F^T \mid \mathcal{B}) & \sim & p(F^T \mid \mathcal{B}) \\ (y_i^C \mid F^C \, \mathcal{B}) & \stackrel{IID}{\sim} & F^C & (y_j^T \mid F^T \, \mathcal{B}) & \stackrel{IID}{\sim} & F^T \end{array} \quad (4)$$

- 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

$$\theta \triangleq \frac{\mu^T - \mu^C}{\mu^C} \, ; \quad (5)$$

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

Optimal Bayesian Model Specification

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

- 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**:

$$\begin{array}{rcl|rcl} (F^C \mid \mathcal{B}) & \sim & DP(\alpha^C, F_0^C) & (F^T \mid \mathcal{B}) & \sim & DP(\alpha^T, F_0^T) \\ (y_i^C \mid F^C \, \mathcal{B}) & \stackrel{IID}{\sim} & F^C & (y_j^T \mid F^T \, \mathcal{B}) & \stackrel{IID}{\sim} & F^T \end{array} \qquad (6)$$

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

$$(F^C \mid y^C \, \mathcal{B}) \sim DP\left(\alpha^C + n^C, \frac{\alpha^C \, F_0^C + n \, \hat{F}_n^C}{\alpha^C + n^C} \right) \qquad (7)$$

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.

$DP(n, \hat{F}_n)$

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

- 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

$$(F^C \,|\, y^C \, \mathcal{B}) \sim DP\left(n^C, \hat{F}_n^C\right) \quad (F^T \,|\, y^T \, \mathcal{B}) \sim DP\left(n^T, \hat{F}_n^T\right) . \quad (8)$$

In my view the $DP\left(n, \hat{F}_n\right)$ posterior should get **far more use** in **applied Bayesian work** 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 $y = (y_1, \ldots, y_n)$ 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 $(y_1, \ldots, y_n)$; 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 $(y_1, \ldots, y_n)$ and form the **empirical CDF** based on these draws.

The Frequentist Bootstrap in BNP Calculations

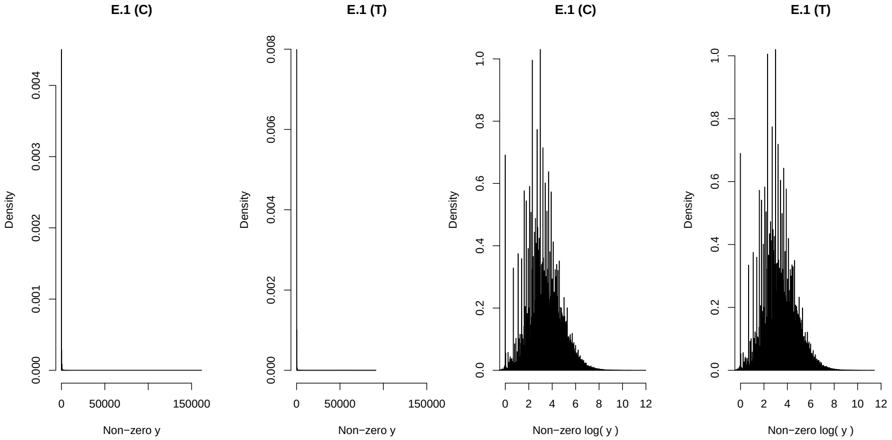
- **This is precisely the frequentist bootstrap** (Efron 1979), which turns out to be about **30 times faster** than stick-breaking and is **embarrassingly parallelizable** to boot (e.g., Alex Terenin tells me that this is **ludicrously easy** to implement in `MapReduce`).
- Therefore, to **simulate**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



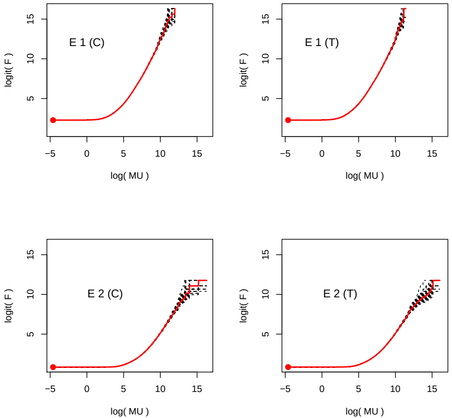
Numerical Summaries of E_1 and E_2

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., 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?

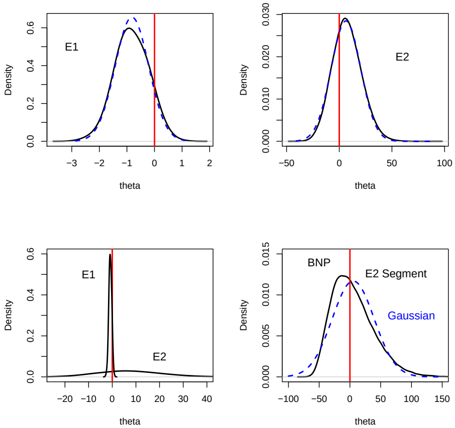
- $\hat{\theta}$ is driven by the **sample means** $\bar{y}^C$ and $\bar{y}^T$, so with **large enough sample sizes** the posterior for θ will be **close to Gaussian** (by the Bayesian CLT), rendering the **bootstrapping unnecessary**, but the **skewness** and **kurtosis** values for the outcome y are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

$$\text{skewness}(\bar{y}_n) = \frac{\text{skewness}(y_1)}{\sqrt{n}} \quad \text{and} \quad \text{kurtosis}(\bar{y}_n) = \frac{\text{kurtosis}(y_1)}{n} \,. \qquad (9)$$

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

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**)

$$SD(\theta \mid y^T \, y^C \, \mathcal{B}) \doteq \sqrt{\frac{\bar{y}_T^2 \, s_C^2}{\bar{y}_C^4 \, n_C} + \frac{s_T^2}{\bar{y}_C^2 \, n_T}} \, . \tag{10}$$

- **Extension:** **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 $j = 1, \ldots, n_{group}$,

$$\begin{array}{ccc} (F^T \mid \mathcal{B}) & \sim & DP(\alpha^T, F_0^T) \\ (y_j^T \mid F^T \, \mathcal{B}) & \stackrel{IID}{\sim} & F^T \end{array} \quad \left| \quad \begin{array}{ccc} (F_0^C \mid \mathcal{B}) & \sim & DP(\gamma, G) \\ (F^{C_i} \mid F_0^C \, \mathcal{B}) & \stackrel{IID}{\sim} & DP(\alpha^C, F_0^C) \\ (y_j^{C_i} \mid F^{C_i} \, \mathcal{B}) & \stackrel{IID}{\sim} & F^{C_i} \end{array} \right. \qquad (11)$$

- 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:

$$(\mu^T \mid y^T \, \mathcal{B}) \stackrel{\cdot}{\sim} N\left[\bar{y}^T, \frac{(s^T)^2}{n^T}\right] \quad \text{and} \quad (\mu^{C_i} \mid y^{C_i} \, \mathcal{B}) \stackrel{\cdot}{\sim} N\left[\bar{y}^{C_i}, \frac{(s^{C_i})^2}{n^{C_i}}\right] . \qquad (12)$$

Approximate BNP With 100 Million Observations

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

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**):

$$\begin{array}{rcl} (\mu^T \mid \mathcal{B}) & \propto & 1 \\ (\bar{y}^T \mid \mu^T \, \mathcal{B}) & \sim & N\left[\mu^T, \frac{(s^T)^2}{n^T}\right] \end{array} \quad \Bigg| \quad \begin{array}{rcl} (\sigma \mid \mathcal{B}) & \sim & U(0, A) \\ (\mu^C \mid \sigma \, \mathcal{B}) & \propto & 1 \\ (\mu^{C_i} \mid \mu^C \, \sigma \, \mathcal{B}) & \stackrel{IID}{\sim} & N(\mu^C, \sigma^2) \\ (\bar{y}^{C_i} \mid \mu^{C_i} \, \mathcal{B}) & \sim & N\left[\mu^{C_i}, \frac{(s^{C_i})^2}{n^{C_i}}\right] \end{array} \qquad (13)$$

- 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

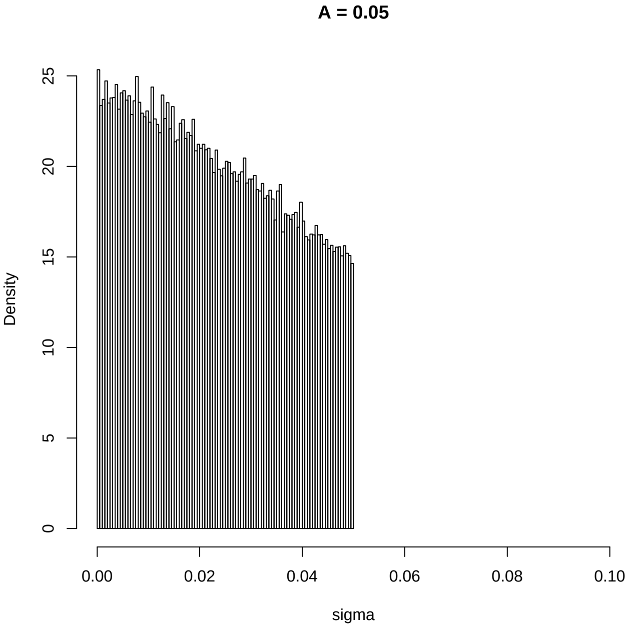
- Start with **one** C **group**: **simulated data** similar to E_1 in part I but with a **bigger treatment effect** — total sample size **24.5 million**, $\bar{y}^T = 9.286$, $\bar{y}^C = 9.203$, $\hat{\theta} =$**+0.9%** with posterior SD **0.6%**, **posterior probability of positive effect 0.93**.

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

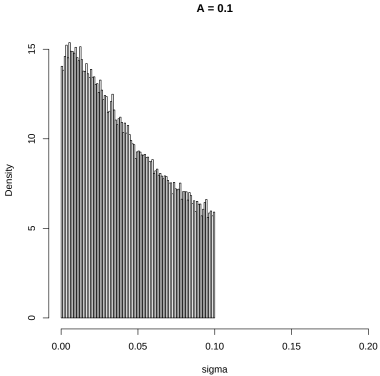


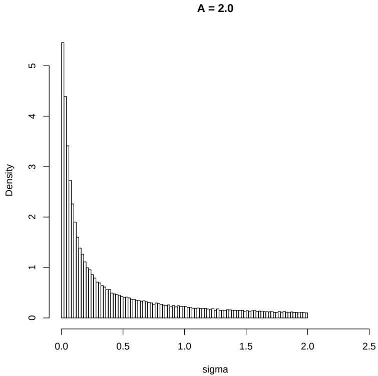
$A = 0.1$: Borrowing Strength Seems to Disappear

		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)

group	n	y mean	y sd	mu mean	mu sd	theta mean	theta sd	theta 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**.

An Example, to Fix Ideas

Case Study 1. (**Krnjajić**, **Kottas**, **Draper** 2008): **In-home geriatric assessment (IHGA)**. In an **experiment** conducted in the **1980s** (**Hendriksen** et al., 1984), **572 elderly people**, **representative** of $\mathcal{P} =$ {all **non-institutionalized elderly people** in **Denmark**}, were **randomized**, **287** to a **control** (C) group (who received **standard health care**) and **285** to a **treatment** (T) group (who received **standard care plus IHGA**: a kind of **preventive medicine** in which **each person's medical** and **social needs** were **assessed** and **acted upon individually**).

One **important outcome** was the **number of hospitalizations** during the **two-year** life of the study:

	Number of Hospitalizations										
Group	0	1	2	3	4	5	6	7	n	Mean	SD
Control	138	77	46	12	8	4	0	2	$n_C = 287$	0.944	1.239
Treatment	147	83	37	13	3	1	1	0	$n_T = 285$	0.768	1.008

Let μ_C and μ_T be the **mean hospitalization rates** (per two years) in $\mathcal{P}$ under the C and T **conditions**, respectively.

Here are **four statistical questions** that **arose** from **this study**:

Bayesian Qual/Quant Inference

Recall from our earlier discussion that if I judge **binary** $(y_1, \ldots, y_n)$ to be part of **infinitely exchangeable sequence**, to be **coherent** my joint predictive distribution $p(y_1, \ldots, y_n)$ must have simple **hierarchical** form

$$\begin{aligned} \theta & \sim p(\theta) \\ (y_i|\theta) & \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta), \end{aligned}$$

where $\theta = P(y_i = 1) =$ **limiting value of mean of** y_i in infinite sequence.

Writing $s = (s_1, s_2)$ where s_1 and s_2 are the **numbers of 0s and 1s**, respectively in $(y_1, \ldots, y_n)$, this is **equivalent** to the model

$$\begin{aligned} \theta_2 & \sim p(\theta_2) \\ (s_2|\theta_2) & \sim \text{Binomial}(n, \theta_2), \end{aligned} \tag{1}$$

where (in a slight change of notation) $\theta_2 = P(y_i = 1)$; i.e., in this simplest case the form of the **likelihood function** ($\text{Binomial}(n, \theta_2)$) is determined by **coherence**.

The **likelihood function** for θ_2 in this model is

$$l(\theta_2|y) = c\,\theta_2^{s_2}(1-\theta_2)^{n-s_2} = c\,\theta_1^{s_1}\theta_2^{s_2}, \tag{2}$$

from which it's evident that the **conjugate prior** for the **Bernoulli/Binomial likelihood** (the choice of prior having the property that the **posterior** for θ_2 has the same **mathematical form** as the **prior**) is the family of **Beta**(α_1, α_2) densities

$$p(\theta_2) = c\,\theta_2^{\alpha_2-1}(1-\theta_2)^{\alpha_1-1} = c\,\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}. \tag{3}$$

for some $\alpha_1 > 0, \alpha_2 > 0$.

Bayesian Qual/Quant Inference

With this prior the **conjugate updating rule** is evidently

$$\left\{ \begin{array}{c} \theta_2 \sim \text{Beta}(\alpha_1, \alpha_2) \\ (s_2 | \theta_2) \sim \text{Binomial}(n, \theta_2) \end{array} \right\} \rightarrow (\theta_2 | y) \sim \text{Beta}(\alpha_1 + s_1, \alpha_2 + s_2), \tag{4}$$

where s_1 (s_2) is the **number of 0s (1s)** in the data set $y = (y_1, \ldots, y_n)$.

Moreover, given that the **likelihood** represents a **(sample) data set** with s_1 0s and s_2 1s and a **data sample size** of $n = (s_1 + s_2)$, it's clear that

(a) the **Beta**(α_1, α_2) prior acts like a **(prior) data set** with α_1 0s and α_2 1s and a **prior sample size** of $(\alpha_1 + \alpha_2)$, and

(b) to achieve a relatively **diffuse (low-information-content**) prior for θ_2 (if that's what **context** suggests I should aim for) I should try to specify α_1 and α_2 **not far from 0**.

Easy **generalization** of all of this: suppose the y_i take on $l \geq 2$ **distinct values** $v = (v_1, \ldots, v_l)$, and let $s = (s_1, \ldots, s_l)$ be the **vector** of **counts** ($s_1 = \#(y_i = v_1)$ and so on).

If I judge the y_i to be part of an **infinitely exchangeable sequence**, then to be **coherent** my joint predictive distribution $p(y_1, \ldots, y_n)$ must have the **hierarchical** form

$$\begin{array}{rcl} \theta & \sim & p(\theta) \\ (s|\theta) & \sim & \text{Multinomial}(n, \theta), \end{array} \tag{5}$$

where $\theta = (\theta_1, \ldots, \theta_l)$ and θ_j is the **limiting relative frequency** of v_j values in the infinite sequence.

Bayesian Qual/Quant Inference

The **likelihood** for (vector) θ in this case has the form

$$l(\theta|y) = c \prod_{j=1}^{l} \theta_j^{s_j}, \tag{6}$$

from which it's evident that the **conjugate prior** for the **Multinomial likelihood** is of the form

$$p(\theta) = c \prod_{j=1}^{l} \theta_j^{\alpha_j - 1}, \tag{7}$$

for some $\alpha = (\alpha_1, \ldots, \alpha_l)$ with $\alpha_j > 0$ for $j = 1, \ldots, l$; this is the **Dirichlet**(α) distribution, a **multivariate generalization** of the Beta family.

Here the **conjugate updating rule** is

$$\left\{ \begin{array}{c} \theta \sim \text{Dirichlet}(\alpha) \\ (s|\theta) \sim \text{Multinomial}(n, \theta) \end{array} \right\} \rightarrow (\theta|y) \sim \text{Dirichlet}(\alpha + s), \tag{8}$$

where $s = (s_1, \ldots, s_l)$ and s_j is the **number of** v_j **values** $(j = 1, \ldots, l)$ in the data set $y = (y_1, \ldots, y_n)$.

Furthermore, by **direct analogy** with the $l = 2$ case,

(a) the **Dirichlet**(α) prior acts like a **(prior) data set** with α_j v_j values $(j = 1, \ldots, l)$ and a **prior sample size** of $\sum_{j=1}^{l} \alpha_j$, and

(b) to achieve a relatively **diffuse (low-information-content)** prior for θ (if that's what **context** suggests I should aim for) I should try to choose all of the α_j **not far from 0**.

Bayesian Qual/Quant Inference

To **summarize**:

(A) if the **data vector** $y = (y_1, \ldots, y_n)$ takes on l **distinct** values $v = (v_1, \ldots, v_l)$ (**real numbers or not**) and I judge (my uncertainty about) the infinite sequence $(y_1, y_2, \ldots)$ to be **exchangeable**, then (by a **representation theorem** of de Finetti) **coherence** compels me (i) to **think about** the quantities $\theta = (\theta_1, \ldots, \theta_l)$, where θ_j is the **limiting relative frequency** of the v_j values in the infinite sequence, and (ii) to **adopt** the Multinomial model

$$\begin{aligned} \theta & \sim p(\theta) \\ p(y_i|\theta) & = c \prod_{j=1}^{l} \theta_j^{s_j}, \end{aligned} \tag{9}$$

where s_j is the **number** of y_i values equal to v_j;

(B) if context suggests a **diffuse** prior for θ a convenient (**conjugate**) choice is **Dirichlet**(α) with $\alpha = (\alpha_1, \ldots, \alpha_l)$ and all of the α_j **positive but close to 0**; and

(C) with a **Dirichlet**(α) prior for θ the **posterior** is **Dirichlet**(α'), where $s = (s_1, \ldots, s_l)$ and $\alpha' = (\alpha + s)$.

Note, remarkably, that the v_j values themselves **make no appearance** in the model; this modeling approach is **natural** with **categorical** outcomes but can also be used when the v_j are **real numbers**.

For example, for **real-valued** y_i, if (as in the **IHGA case study** in Part 1) interest focuses on the **(underlying population) mean** in the infinite sequence $(y_1, y_2, \ldots)$, this is $\mu_y = \sum_{j=1}^{l} \theta_j \, v_j$, which is just a **linear function** of the θ_j with **known coefficients** v_j.

Bayesian Qual/Quant Inference

This fact makes it possible to draw an **analogy** with the **distribution-free** methods that are at the heart of **frequentist non-parametric** inference: when your **outcome variable** takes on a **finite number** of **real** values v_j, **exchangeability** compels a **Multinomial likelihood** on the **underlying frequencies** with which the v_j occur; you are not required to build a **parametric model** (e.g., normal, lognormal, ...) on the y_i values themselves.

In this sense, therefore, model (14)—particularly with the **conjugate Dirichlet** prior—can serve as a kind of **low-technology Bayesian non-parametric** modeling: this is the basis of the **Bayesian bootstrap** (Rubin 1981).

Moreover, if you're **in a hurry** and you're already familiar with `WinBUGS` you can readily carry out **inference** about quantities like μ_y above in that environment, but there's **no need to do MCMC** here: **ordinary Monte Carlo** (MC) sampling from the **Dirichlet**(α') posterior distribution is perfectly **straightforward**, e.g., in `R`, based on the following **fact**:

To generate a **random draw** $\theta = (\theta_1, \ldots, \theta_l)$ from the **Dirichlet**(α') distribution, with $\alpha' = (\alpha'_1, \ldots, \alpha'_l)$, **independently draw**

$$g_j \stackrel{\text{indep}}{\sim} \Gamma(\alpha'_j, \beta), \quad j = 1, \ldots, l \tag{10}$$

(where $\Gamma(a, b)$ is the **Gamma distribution** with parameters a and b) and compute

$$\theta_j = \frac{g_j}{\sum_{m=1}^{l} g_j}. \tag{11}$$

Any $\beta > 0$ will do in this calculation; $\beta = 1$ is a **good choice** that leads to **fast random number generation**.

Bayesian Qual/Quant Inference

The **downloadable version** of `R` doesn't have a **built-in function** for making **Dirichlet draws**, but it's easy to write one:

```
rdirichlet = function( n.sim, alpha ) {

  l = length( alpha )

  theta = matrix( 0, n.sim, l )

  for ( j in 1:l ) {

    theta[ , j ] = rgamma( n.sim, alpha[ j ], 1 )

  }

  theta = theta / apply( theta, 1, sum )

  return( theta )

}
```

The **Dirichlet**(α) distribution has the following **moments**: if $\theta \sim \text{Dirichlet}(\alpha)$ then

$$E(\theta_j) = \frac{\alpha_j}{\alpha_0}, \; V(\theta_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}, \; C(\theta_j, \theta_{j'}) = -\frac{\alpha_j \alpha_{j'}}{\alpha_0^2(\alpha_0 + 1)},$$

where $\alpha_0 = \sum_{j=1}^{l} \alpha_j$ (note the **negative correlation** between components of θ).

This can be used to **test** the function above:

Bayesian Qual/Quant Inference

```
> alpha = c( 5.0, 1.0, 2.0 )

> alpha.0 = sum( alpha )

> test = rdirichlet( 100000, alpha )    # 15 seconds at 550 Unix MHz

> apply( test, 2, mean )

[1] 0.6258544 0.1247550 0.2493905

> alpha / alpha.0

[1] 0.625 0.125 0.250

> apply( test, 2, var )

[1] 0.02603293 0.01216358 0.02071587

> alpha * ( alpha.0 - alpha ) / ( alpha.0^2 * ( alpha.0 + 1 ) )

[1] 0.02604167 0.01215278 0.02083333

> cov( test )

             [,1]         [,2]         [,3]
[1,]  0.026032929 -0.008740319 -0.017292610
[2,] -0.008740319  0.012163577 -0.003423259
[3,] -0.017292610 -0.003423259  0.020715869

> - outer( alpha, alpha, "*" ) / ( alpha.0^2 * ( alpha.0 + 1 ) )

             [,1]         [,2]         [,3]
[1,] -0.043402778 -0.008680556 -0.017361111
[2,] -0.008680556 -0.001736111 -0.003472222     # ignore diagonals
[3,] -0.017361111 -0.003472222 -0.006944444
```

Bayesian Qual/Quant Inference

Example: re-analysis of **IHGA data** from Part 1; recall **policy** and **clinical interest** focused on $\eta = \frac{\mu_E}{\mu_C}$.

	Number of Hospitalizations										
Group	0	1	2	3	4	5	6	7	n	Mean	SD
Control	138	77	46	12	8	4	0	2	287	0.944	1.24
Experimental	147	83	37	13	3	1	1	0	285	0.768	1.01

In this **two-independent-samples** setting I can apply de Finetti's representation theorem **twice, in parallel**, on the C and E data.

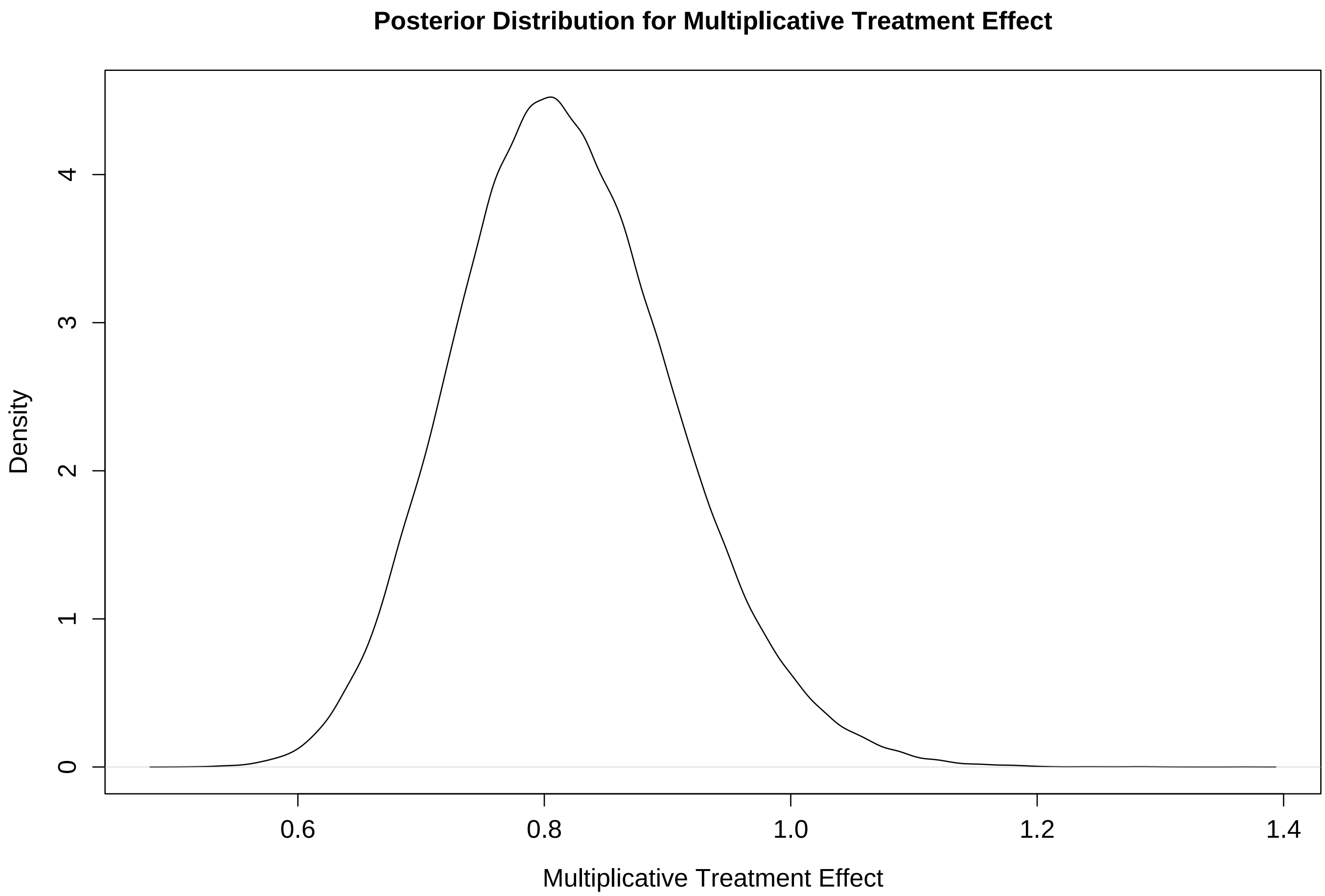
I don't know much about the **underlying frequencies** of $0, 1, \ldots, 7$ hospitalizations under C and E **external** to the data, so I'll use a **Dirichlet**$(\epsilon, \ldots, \epsilon)$ **prior** for both θ_C and θ_E with $\epsilon = 0.001$, leading to a **Dirichlet**$(138.001, \ldots, 2.001)$ **posterior** for θ_C and a **Dirichlet**$(147.001, \ldots, 0.001)$ **posterior** for θ_E (other small positive choices of ϵ yield **similar results**).

```
> alpha.C = c( 138.001, 77.001, 46.001, 12.001, 8.001, 4.001, 0.001,
    2.001 )

> alpha.E = c( 147.001, 83.001, 37.001, 13.001, 3.001, 1.001, 1.001,
    0.001 )

> theta.C = rdirichlet( 100000, alpha.C )   # 17 sec at 550 Unix MHz

> theta.E = rdirichlet( 100000, alpha.E )   # also 17 sec

> print( post.mean.theta.C = apply( theta.C, 2, mean ) )

[1] 4.808015e-01 2.683458e-01 1.603179e-01 4.176976e-02 2.784911e-02
[6] 1.395287e-02 3.180905e-06 6.959859e-03

> print( post.SD.theta.C <- apply( theta.C, 2, sd ) )

[1] 0.0294142963 0.0261001259 0.0216552661 0.0117925465 0.0096747630
[6] 0.0069121507 0.0001017203 0.0048757485
```

Bayesian Qual/Quant Inference

```
> print( post.mean.theta.E <- apply( theta.E, 2, mean ) )

[1] 5.156872e-01 2.913022e-01 1.298337e-01 4.560130e-02 1.054681e-02
[6] 3.518699e-03 3.506762e-03 3.356346e-06

> print( post.SD.theta.E <- apply( theta.E, 2, sd ) )

[1] 0.029593047 0.026915644 0.019859213 0.012302252 0.006027157
[6] 0.003501568 0.003487824 0.000111565

> mean.effect.C <- theta.C %*% ( 0:7 )

> mean.effect.E <- theta.E %*% ( 0:7 )

> mult.effect <- mean.effect.E / effect.C

> print( post.mean.mult.effect <- mean( mult.effect ) )

[1] 0.8189195

> print( post.SD.mult.effect <- sd( mult.effect ) )

[1] 0.08998323

> quantile( mult.effect, probs = c( 0.0, 0.025, 0.5, 0.975, 1.0 ) )

       0%      2.5%       50%     97.5%      100%
0.5037150 0.6571343 0.8138080 1.0093222 1.3868332

> postscript( "mult.effect.ps" )

> plot( density( mult.effect, n = 2048 ), type = 'l', cex.lab = 1.25,
    xlab = 'Multiplicative Treatment Effect', cex.axis = 1.25,
    main = 'Posterior Distribution for Multiplicative Treatment Effect',
    cex.main = 1.25 )

> dev.off( )
```

Bayesian Qual/Quant Inference



Model	Posterior Mean	Posterior SD	Central 95% Interval
REPR	0.830	0.0921	$(0.665, 1.02)$
Dir-Mult	0.819	0.0900	$(0.657, 1.01)$

In this example the **low-tech BNP**, **Dirichlet-Multinomial**, **exchangeability-plus-diffuse-prior-information** model has **reproduced** the **parametric REPR results** almost exactly and without a **complicated search through model space** for a **"good"** model.

$\boxed{\textbf{NB}}$ This **approach** is an **application** of the **Bayesian bootstrap** (Rubin 1981), which (for **complete validity**) includes the **assumption** that the **observed** y_i **values form** a **complete set** of $\{$**all possible values the outcome** y **could take on**$\}$.