Case Studies in Bayesian Data Science

4: The Bootstrap as an Approximate BNP Method

David Draper

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

draper@ucsc.edu

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

27 Nov 2015

users.soe.ucsc.edu/∼draper/Reading-2015-Day-5.html

Statistics

- **Statistics** is the **study** of **uncertainty**: how to **measure it well**, and how to **make good choices** in the **face** of it.

Statistics

- **Statistics** is the **study** of **uncertainty**: how to **measure it well**, and how to **make good choices** in the **face** of it.
- **Uncertainty** is a **state** of **incomplete information** about **something of interest** to **You**.

Statistics

- **Statistics** is the **study** of **uncertainty**: how to **measure it well**, and how to **make good choices** in the **face** of it.
- **Uncertainty** is a **state** of **incomplete information** about **something of interest** to **You**.
- Call the **something of interest** to **You** θ (**think** of a **vector** in $\Re^k$).

Statistics

- **Statistics** is the **study** of **uncertainty**: how to **measure it well**, and how to **make good choices** in the **face** of it.
- **Uncertainty** is a **state** of **incomplete information** about **something of interest** to **You**.
- Call the **something of interest** to **You** θ (**think** of a **vector** in $\Re^k$).

Case Study: (**Giorgio Ballardin**) **Treatment** T: a **marketing email** in the **Business** and **Industrial category**; **Control** C: **no such email**.

Statistics

- **Statistics** is the **study** of **uncertainty**: how to **measure it well**, and how to **make good choices** in the **face** of it.

- **Uncertainty** is a **state** of **incomplete information** about **something of interest** to **You**.

- Call the **something of interest** to **You** θ (**think** of a **vector** in $\Re^k$).

Case Study: **(Giorgio Ballardin) Treatment** T: a **marketing email** in the **Business** and **Industrial category**; **Control** C: **no such email**.

Design: controlled trial (A/B test), with **256,721** ***eBay*** **representative users randomized** (**128,349** to T, **128,372** to C) and **followed** for **7 days, starting** on **14 July 2014**.

Statistics

- **Statistics** is the **study** of **uncertainty**: how to **measure it well**, and how to **make good choices** in the **face** of it.

- **Uncertainty** is a **state** of **incomplete information** about **something of interest** to **You**.

- Call the **something of interest** to **You** θ (**think** of a **vector** in $\Re^k$).

Case Study: **(Giorgio Ballardin) Treatment** T: a **marketing email** in the **Business** and **Industrial category**; **Control** C: **no such email**.

Design: controlled trial (A/B test), with **256,721** ***eBay*** **representative users randomized** (**128,349** to T, **128,372** to C) and **followed** for **7 days, starting** on **14 July 2014**.

Outcome of **interest: Gross Merchandise Bought (GMB)**, i.e., **total $ generated** by **buyers** in the **7–day period**.

Statistics

- **Statistics** is the **study** of **uncertainty**: how to **measure it well**, and how to **make good choices** in the **face** of it.
- **Uncertainty** is a **state** of **incomplete information** about **something of interest** to **You**.
- Call the **something of interest** to **You** θ (**think** of a **vector** in $\Re^k$).

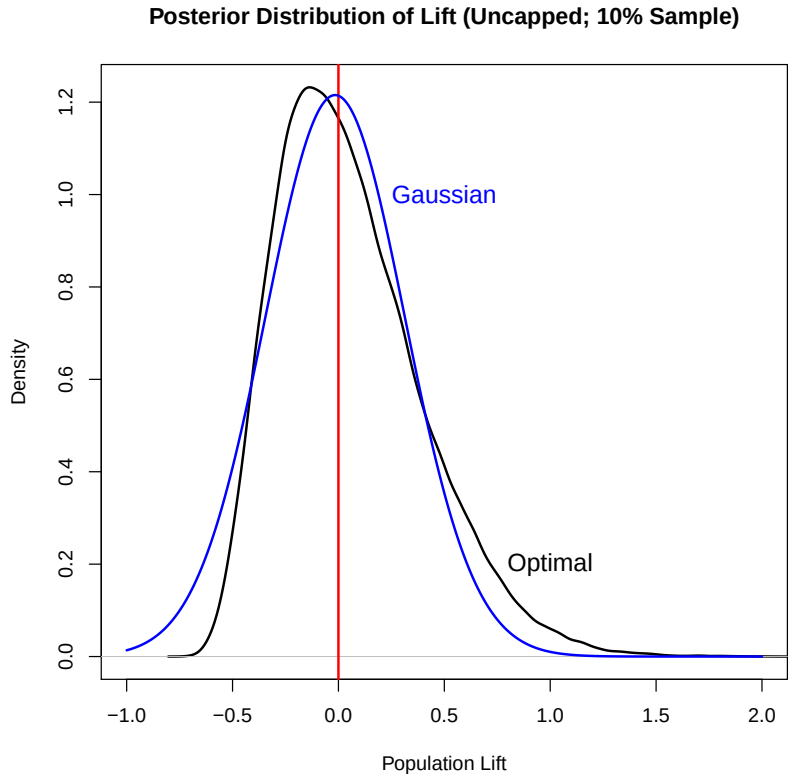
Case Study: **(Giorgio Ballardin) Treatment** T: a **marketing email** in the **Business** and **Industrial category**; **Control** C: **no such email**.

Design: controlled trial (A/B test), with **256,721** ***eBay*** **representative users randomized** (**128,349** to T, **128,372** to C) and **followed** for **7 days, starting** on **14 July 2014**.

Outcome of **interest: Gross Merchandise Bought (GMB)**, i.e., **total $ generated** by **buyers** in the **7–day period**.

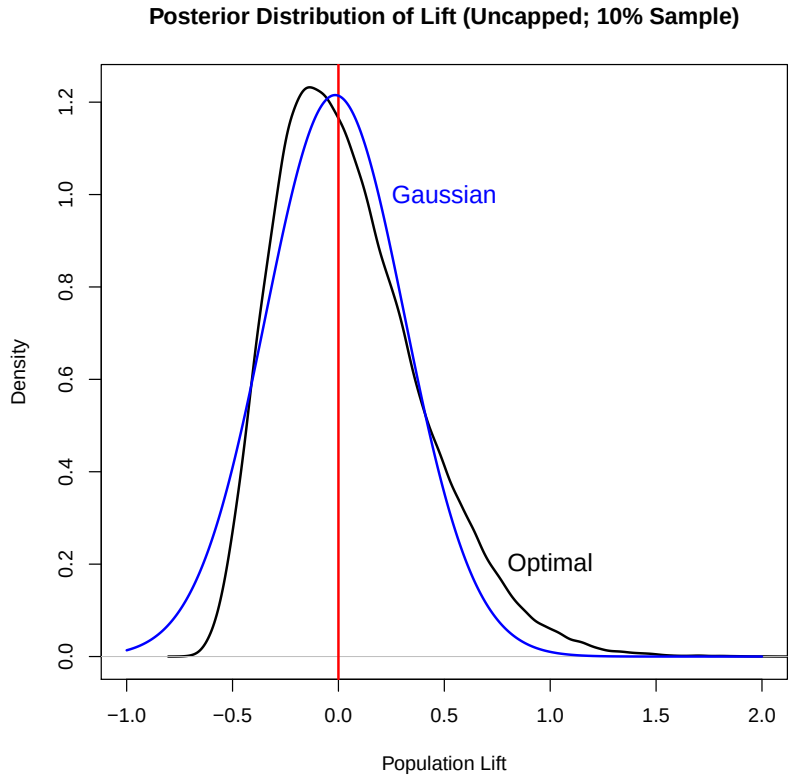
$\theta =$ **lift** $\left(\frac{\mu_T - \mu_C}{\mu_C}\right)$ that **would** be **observed** if **all** ***eBay*** **users** in the **population** $\mathcal{P}$ of **interest** (**future** ***eBay*** **users**) were to **counterfactually either receive** the **email** (**resulting population mean** $= \mu_T$) or **not receive it** (**population mean** $= \mu_C$): **here** $k = 1$.

Where This Talk is Headed



Let $\eta = P(\theta > 0 | \textbf{data, background information})$ in a **segment (subset, stratum)** of **users** that **comprises 10%** of **eBay traffic** in the **Case Study** (**12,837 observations** in **each** of T and C).

Where This Talk is Headed



Let $\eta = P(\theta > 0 | \textbf{data, background information})$ in a **segment (subset, stratum)** of **users** that **comprises 10%** of **eBay traffic** in the **Case Study** (**12,837 observations** in **each** of T and C).

Here **Gaussian analysis (current eBay best practice) produces** an **estimate** of η that's **too low** by **about 5%**, in **relation** to an **optimal analysis**; this **underestimation gets worse** with **decreasing segment sample size**.

Probability; Inference, Prediction, Decision-Making

- **Acquiring** a **relevant data set** D (**think** of a **vector** in $\Re^n$) is a **good way** to **decrease Your uncertainty** about θ.

Probability; Inference, Prediction, Decision-Making

- **Acquiring** a **relevant data set** D (**think** of a **vector** in $\Re^n$) is a **good way** to **decrease Your uncertainty** about θ.

Case Study: $D =$ the **256,721 GMB values obtained** in the **experiment, along with** a **binary vector** of **length 256,721 identifying group membership** ($\mathbf{1} = T$, $\mathbf{0} = C$).

Probability; Inference, Prediction, Decision-Making

- **Acquiring** a **relevant data set** D (**think** of a **vector** in $\Re^n$) is a **good way** to **decrease Your uncertainty** about θ.

Case Study: D = the **256,721 GMB values obtained** in the **experiment, along with** a **binary vector** of **length 256,721 identifying group membership** (**1** = T, **0** = C).

- **Probability** is the **branch** of **mathematics devoted** to **quantifying uncertainty**.

Probability; Inference, Prediction, Decision-Making

- **Acquiring** a **relevant data set** D (**think** of a **vector** in $\Re^n$) is a **good way** to **decrease Your uncertainty** about θ.

Case Study: $D =$ the **256,721 GMB values obtained** in the **experiment, along with** a **binary vector** of **length 256,721 identifying group membership** ($\mathbf{1} = T$, $\mathbf{0} = C$).

- **Probability** is the **branch** of **mathematics devoted** to **quantifying uncertainty**.

Statistics makes **vigorous use** of **probabilistic (stochastic) models** of the **world**, to

Probability; Inference, Prediction, Decision-Making

- **Acquiring** a **relevant data set** D (**think** of a **vector** in $\Re^n$) is a **good way** to **decrease Your uncertainty** about θ.

Case Study: D = the **256,721 GMB values obtained** in the **experiment, along with** a **binary vector** of **length 256,721 identifying group membership** (**1** = T, **0** = C).

- **Probability** is the **branch** of **mathematics devoted** to **quantifying uncertainty**.

Statistics makes **vigorous use** of **probabilistic (stochastic) models** of the **world**, to

— **generalize outward** from D to $\mathcal{P}$ (**inference**) [**initially** this ***seems* to be** the **main point** of the **Case Study**];

Probability; Inference, Prediction, Decision-Making

- **Acquiring** a **relevant data set** D (**think** of a **vector** in $\Re^n$) is a **good way** to **decrease Your uncertainty** about θ.

Case Study: D = the **256,721 GMB values obtained** in the **experiment, along with** a **binary vector** of **length 256,721 identifying group membership** (**1** = T, **0** = C).

- **Probability** is the **branch** of **mathematics devoted** to **quantifying uncertainty**.

Statistics makes **vigorous use** of **probabilistic (stochastic) models** of the **world**, to

— **generalize outward** from D to $\mathcal{P}$ (**inference**) [**initially** this ***seems* to be** the **main point** of the **Case Study**];

— **estimate future data values** D^* (**prediction**) [but the **Case Study** is **also about predicting future data values** for ***eBay* users**]; and

Probability; Inference, Prediction, Decision-Making

- **Acquiring** a **relevant data set** D (**think** of a **vector** in $\Re^n$) is a **good way** to **decrease Your uncertainty** about θ.

Case Study: D = the **256,721 GMB values obtained** in the **experiment, along with** a **binary vector** of **length 256,721 identifying group membership** (**1** = T, **0** = C).

- **Probability** is the **branch** of **mathematics devoted** to **quantifying uncertainty**.

Statistics makes **vigorous use** of **probabilistic (stochastic) models** of the **world**, to

— **generalize outward** from D to $\mathcal{P}$ (**inference**) [**initially** this ***seems* to be** the **main point** of the **Case Study**];

— **estimate future data values** D^* (**prediction**) [but the **Case Study** is **also about predicting future data values** for ***eBay* users**]; and

— **help people choose** the **best course of action** (**decision theory**) [and the **real point** of the **Case Study** is to ***decide* whether** the T **email is or is not** a **good thing** to **repeat** in the **future**].

Frequentist and Bayesian Probability

Here I'll **focus** on **inference** about θ; in a **future talk** I'll **show** that **treating problems** of **this kind** **decision-theoretically** can **lead** to **better eBay results**.

Frequentist and Bayesian Probability

Here I'll **focus** on **inference** about θ; in a **future talk** I'll **show** that **treating problems** of **this kind** **decision-theoretically** can **lead** to **better eBay results**.

- A **third ingredient** $(\theta, D, \mathcal{B})$ is **crucial** in **probability modeling**: $\mathcal{B} = \{B_1, \ldots, B_b\}$ (b a **positive finite integer**), in **which** the B_i are **propositions (true/false statements, all true) exhaustively summarizing** the **context** of the **problem** and the **nature** of the **data-gathering process**; $\mathcal{B}$ **summarizes** Your **background information**.

Frequentist and Bayesian Probability

Here I'll **focus** on **inference** about θ; in a **future talk** I'll **show** that **treating problems** of **this kind** **decision-theoretically** can **lead** to **better eBay results**.

- A **third ingredient** $(\theta, D, \mathcal{B})$ is **crucial** in **probability modeling**: $\mathcal{B} = \{B_1, \ldots, B_b\}$ (b a **positive finite integer**), in **which** the B_i are **propositions (true/false statements, all true) exhaustively summarizing** the **context** of the **problem** and the **nature** of the **data-gathering process**; $\mathcal{B}$ **summarizes** Your **background information**.

Case Study: **Here** $\mathcal{B}$ **includes propositions** such as (**users** were **assigned at random** to T or C).

Frequentist and Bayesian Probability

Here I'll **focus** on **inference** about θ; in a **future talk** I'll **show** that **treating problems** of **this kind** **decision-theoretically** can **lead** to **better eBay results**.

- A **third ingredient** $(\theta, D, \mathcal{B})$ is **crucial** in **probability modeling**: $\mathcal{B} = \{B_1, \ldots, B_b\}$ (b a **positive finite integer**), in **which** the B_i are **propositions (true/false statements, all true) exhaustively summarizing** the **context** of the **problem** and the **nature** of the **data-gathering process**; $\mathcal{B}$ **summarizes** Your **background information**.

Case Study: **Here** $\mathcal{B}$ **includes propositions** such as (**users** were **assigned at random** to T or C).

- **Two concepts** of **probability** have **proven useful** over the **last 350 years**:

Frequentist and Bayesian Probability

Here I'll **focus** on **inference** about θ; in a **future talk** I'll **show** that **treating problems** of **this kind** **decision-theoretically** can **lead** to **better eBay results**.

• A **third ingredient** $(\theta, D, \mathcal{B})$ is **crucial** in **probability modeling**: $\mathcal{B} = \{B_1, \ldots, B_b\}$ (b a **positive finite integer**), in **which** the B_i are **propositions (true/false statements, all true) exhaustively summarizing** the **context** of the **problem** and the **nature** of the **data-gathering process**; $\mathcal{B}$ **summarizes** Your **background information**.

Case Study: **Here** $\mathcal{B}$ **includes propositions** such as (**users** were **assigned at random** to T or C).

• **Two concepts** of **probability** have **proven useful** over the **last 350 years**:

— **(frequentist)** Ω **(set)** $= \{$**all possible outcomes** of a **repeatable process**$\}$; $P(A)$ (A a **set**) = **long-run limiting relative frequency** with which $A \subseteq \Omega$ **occurs** in **countably infinitely many repetitions**; and

Frequentist and Bayesian Probability

Here I'll **focus** on **inference** about θ; in a **future talk** I'll **show** that **treating problems** of **this kind** **decision-theoretically** can **lead** to **better eBay results**.

• A **third ingredient** $(\theta, D, \mathcal{B})$ is **crucial** in **probability modeling**: $\mathcal{B} = \{B_1, \ldots, B_b\}$ (b a **positive finite integer**), in **which** the B_i are **propositions (true/false statements, all true) exhaustively summarizing** the **context** of the **problem** and the **nature** of the **data-gathering process**; $\mathcal{B}$ **summarizes** Your **background information**.

Case Study: **Here** $\mathcal{B}$ **includes propositions** such as (**users** were **assigned at random** to T or C).

• **Two concepts** of **probability** have **proven useful** over the **last 350 years**:

— **(frequentist)** Ω **(set)** $= \{$**all possible outcomes** of a **repeatable process**$\}$; $P(A)$ (A a **set**) = **long-run limiting relative frequency** with which $A \subseteq \Omega$ **occurs** in **countably infinitely many repetitions**; and

— **(Bayesian)** $P(A|\mathcal{B})$ = **weight of evidence (information)** in **favor** of the **truth** of **proposition** A, **given** that **all of** $\mathcal{B}$**'s propositions** are **true**.

Bayes is Optimal; Prior and Likelihood Distributions

Fact: If Your goal is summarizing all available relevant information about the truth of a proposition A, then Bayesian probability is optimal for this task.

Case Study: Here we want to assess probabilities such as $p(\theta > 0|D\,\mathcal{B})$, the probability that the lift θ from the email campaign would be positive — given the data set D and the background information $\mathcal{B}$ — if the email were to be sent to all future eBay users.

Actually we want to assess probabilities of the form $p(\theta > c|D\,\mathcal{B})$ for all $c \in \Re$, which (since our uncertainty about θ is continuous on $\Re$) is equivalent to assessing the probability density $p(\theta|D\,\mathcal{B})$.

- Computing $p(\theta|D\,\mathcal{B})$ requires specifying two additional ingredients:

— $p(\theta|\mathcal{B})$, summarizing all available information about θ external to D (universal [and incorrect] name: prior distribution; correct name: external-information distribution); and

— $p(D|\theta\,\mathcal{B})$, which, when viewed as a function $\ell(\theta|D\,\mathcal{B})$ of θ for fixed D and density-normalized (the result is the likelihood distribution), summarizes all available information about θ internal to D.

Bayes's Theorem

Fact: The only way to combine the two information sources $p(\theta|\mathcal{B})$ and $\ell(\theta|D\,\mathcal{B})$ to produce an optimal summary of Your total information about θ is Bayes's Theorem:

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

(universal [and incorrect] name for $p(\theta|D\,\mathcal{B})$: posterior distribution; correct name: total-information distribution).

Case Study (1970s Version): Captopril, a new type of anti-hypertension drug, was developed in the mid-1970s.

• Nothing was known about captopril's effects prior to the first experiment on it (MacGregor et al., 1979; I've changed a few of the details for ease of exposition): 24 representative hypertensive people, randomized (12 to C [placebo], 12 to T [captopril]; SD = standard deviation; outcome variable = systolic blood pressure [mmHg] at the end of the trial).

group	sample size	sample mean	sample SD
C	12	185.3	17.1
T	12	166.8	14.9

Captopril Case Study

Summary: sample sizes $(n_C, n_T) = (12, 12)$; sample means $(\bar{y}_C, \bar{y}_T) = (185.3, 166.8)$; sample SDs $(s_C, s_T) = (17.1, 14.9)$.

Intuitive estimated lift $\hat{\theta} = \frac{\bar{y}_T - \bar{y}_C}{\bar{y}_C} = \frac{166.8 - 185.3}{185.3} \doteq -0.0998 = -10.0\%$.

We estimate that captopril causes a 10% reduction in systolic blood pressure (sounds like a big win), but how much uncertainty is associated with this estimate, in generalizing inferentially from the patients in the experiment to $\mathcal{P} = \{\text{all hypertensive patients}\}$?

We need to finish the model specification to answer this question.

- $p(\theta|\mathcal{B})$ — the "prior" distribution for θ (given $\mathcal{B}$):

Since nothing was known about captopril prior to this experiment, the external-information distribution should contain essentially no information.

In other words, from an entropy point of view it should be close to uniform, so take $p(\theta|\mathcal{B}) \dot{\propto} 1$ (this is a diffuse or flat prior).

Captopril Case Study (continued)

- $p(D|\theta\,\mathcal{B})$ — the "sampling" distribution for D given θ and $\mathcal{B}$:

Off-the-shelf specification for this is as follows — let $\{y_{iC}\}_{i=1}^{n_C}$ and $\{y_{jT}\}_{j=1}^{n_T}$ be the C and T outcome values, respectively; then

$$\begin{aligned}(y_{iC}|\mu_C\,\sigma_C^2\,\mathcal{B}\,\mathcal{G}) &\overset{\text{IID}}{\sim} N(\mu_C, \sigma_C^2) \\ (y_{jT}|\mu_T\,\sigma_T^2\,\mathcal{B}\,\mathcal{G}) &\overset{\text{IID}}{\sim} N(\mu_T, \sigma_T^2)\,, \end{aligned} \tag{2}$$

in which $\mathcal{G}$ = assumption of Gaussian sampling distributions in C and T.

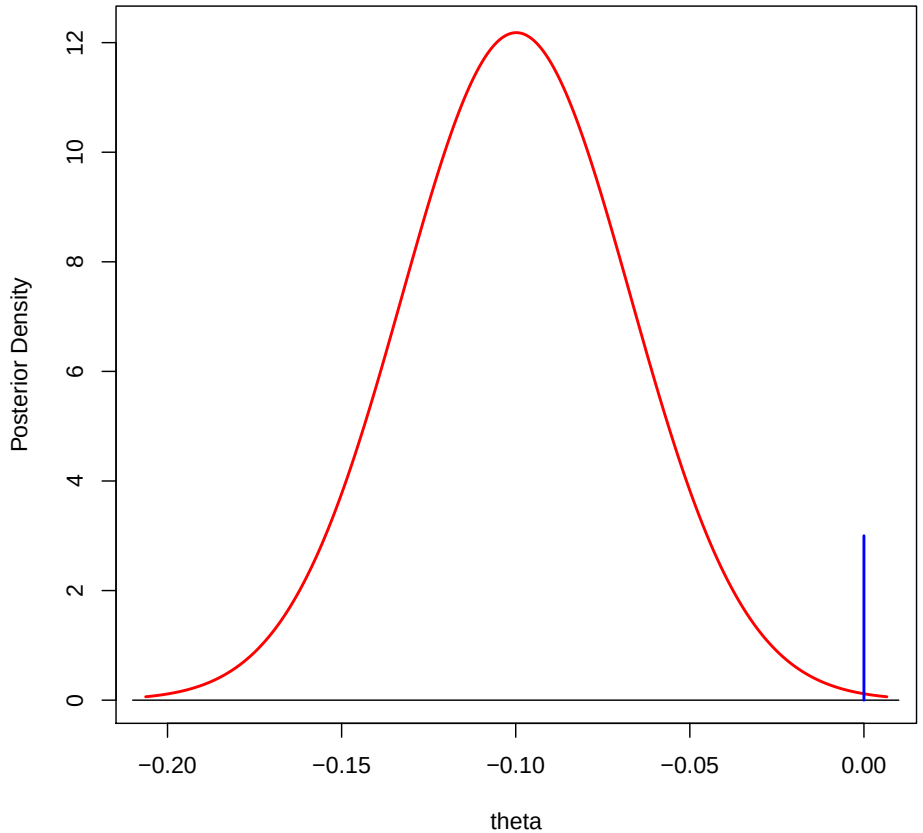
Fact: With this sampling distribution, the induced likelihood distribution for θ is

$$\ell(\theta|D\,\mathcal{B}\,\mathcal{G}) \doteq \text{Normal with mean } \hat{\theta} \text{ and SD } \sqrt{\frac{\bar{y}_T^2\,s_C^2}{\bar{y}_C^4\,n_C} + \frac{s_T^2}{\bar{y}_C^2\,n_T}}\,, \tag{3}$$

and, with the prior distribution $p(\theta|\mathcal{B}) \dot{\propto} 1$, the resulting posterior distribution is

$$(\theta|D\,\mathcal{B}\,\mathcal{G}) \doteq N\left(\hat{\theta}, \sqrt{\frac{\bar{y}_T^2\,s_C^2}{\bar{y}_C^4\,n_C} + \frac{s_T^2}{\bar{y}_C^2\,n_T}}\right) \doteq N(-0.0998, 0.0334^2)\,. \tag{4}$$

Captopril Case Study (continued)



The signal-to-noise ratio here is $\frac{|\text{posterior mean of } \theta|}{\text{posterior SD of } \theta} \doteq \frac{0.0998}{0.0334} \doteq 2.99$, and the posterior probability $p(\theta < 0 | D\,\mathcal{B}\,\mathcal{G})$ that captopril would be beneficial, on average, if administered to the population of {all hypertensive patients similar to those in this study} — given the data set D, the background information $\mathcal{B}$, and the Gaussian sampling-distribution assumption $\mathcal{G}$ — is about 0.999.

Optimal Bayesian Model Specification

Of course we don't want $p(\theta < 0 | D\,\mathcal{B}\,\mathcal{G})$, because $\mathcal{G}$ is not part of the known-to-be-true background information $\mathcal{B}$; we want $p(\theta < 0 | D\,\mathcal{B})$.

Definition (Draper, 2014): Given $(\theta, D, \mathcal{B})$ from

$\mathcal{C} =$ (problem context, data-gathering protocol),

a Bayesian model specification $[p(\theta|\mathcal{B}), p(D|\theta\,\mathcal{B})]$ is optimal if it includes only assumptions rendered true by the structure of $\mathcal{C}$.

Fact: One way to achieve optimal Bayesian model specification is via Bayesian non-parametric (BNP) methods, which place prior distributions on cumulative distribution functions (CDFs).

Fact: Without loss of any generality, an optimal Bayesian model specification for $\{y_{iC}\}_{i=1}^{n_C}$ and $\{y_{jT}\}_{j=1}^{n_T}$ in the current Case Study involves Dirichlet-process (DP) priors, as follows:

$$\begin{aligned} (F_C|\mathcal{B}) &\sim DP(\alpha, F_{0C}) \\ (y_{iC}|F_C\,\mathcal{B}) &\overset{\text{IID}}{\sim} F_C \end{aligned} \tag{5}$$

and similarly for $\{y_{jT}\}_{j=1}^{n_T}$, where F_C is the CDF of the outcome values in the population of (patients, eBay users) similar to those in the experiment.

Fact: With no information about F_C external to D, the optimal BNP analysis is based on the DP posterior

$$(F_C|D\,\mathcal{B}) \sim DP\left(n_C, \hat{F}_{n_C C}\right) , \tag{6}$$

where $\hat{F}_{n_C C}$ is the empirical CDF based on $\{y_{iC}\}_{i=1}^{n_C}$.

Definition: Given a real-valued data set $y = (y_1, \ldots, y_n)$, the (frequentist) bootstrap distribution of the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ may be approximated by

(a) choosing a sample of values y_i^* at random with replacement from the y vactor and computing $\bar{y}^* = \frac{1}{n}\sum_{i=1}^{n} y_i^*$, and

(b) repeating (a) M times (for large positive integer $M \geq$ 100,000) and making a histogram or kernel density trace of the values $(\bar{y}_1^*, \ldots, \bar{y}_M^*)$.

Fact (Draper 2014): The posterior distribution $p(\mu_C|D\,\mathcal{B})$ induced by $DP\left(n_C, \hat{F}_{n_C C}\right)$ distribution may be sampled from accurately and quickly by (frequentist) bootstrapping the sample mean and interpreting the resulting distribution as a good approximation to $p(\mu_C|D\,\mathcal{B})$.

Summary of Conclusions

• captopril: bnp analysis coincides with gaussian-assumption analysis, because clt has kicked in even with only 12 obs per group, because skewness and kurtosis values in C and T are both so close to 0

• gold-standard analysis at ebay up til about the middle of 2012: hope that captopril gaussian-assumption analysis is 'close to optimal'; no proof that this hope is justified

• gold-standard analysis at ebay for about a year, from jul 2012 through jun 2013: complicated parametric frequentist analysis using "torso-tail" distribution (this is demonstrably sub-optimal)

• current gold-standard analysis at ebay: back to gaussian-assumption analysis, with "capping", still based on hope rather than proof

fact: gmb has hideously non-gaussian skewness and kurtosis values

fact: but the gaussian-assumption analysis is still approximately optimal, provided that the C and T sample sizes n.C and n.T are large enough for the clt to save us

eBay Case Study Details

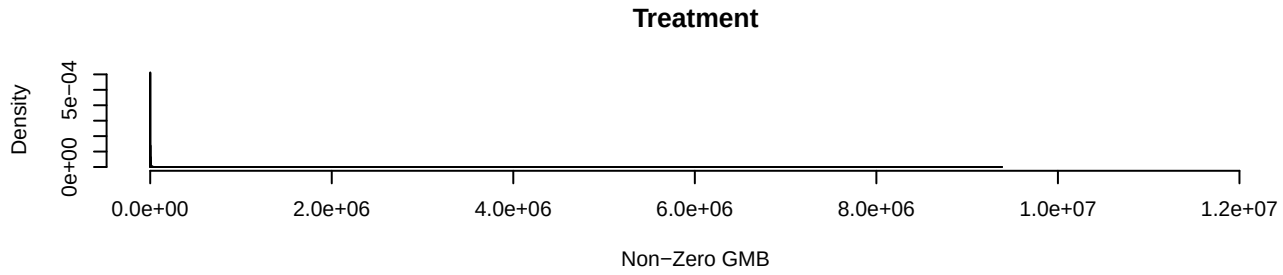
group	number of zero values	number of nonzero values	total number of values	proportion of zero values	nonzero mean	nonzero SD	total mean	total SD
treatment	90,006	38,343	128,349	0.7013	3,618.0	60,476	1080.9	33,096
control	89,863	38,509	128,372	0.7000	3,387.5	66,554	1016.2	36,485

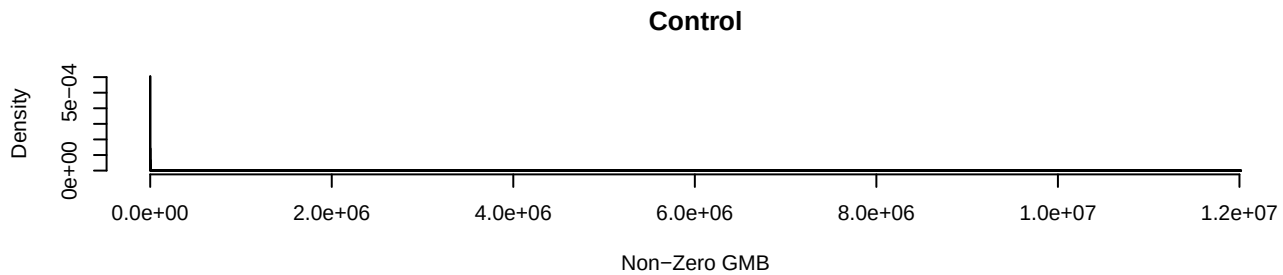
group	all values skewness	all values kurtosis	non-zero values skewness	non-zero values kurtosis	all values noise-to-signal ratio	non-zero values noise-to-signal ratio
treatment	205.9	52,887.9	112.8	15,861.1	30.62	16.72
control	289.1	92,750.5	158.7	27,902.6	35.90	19.65

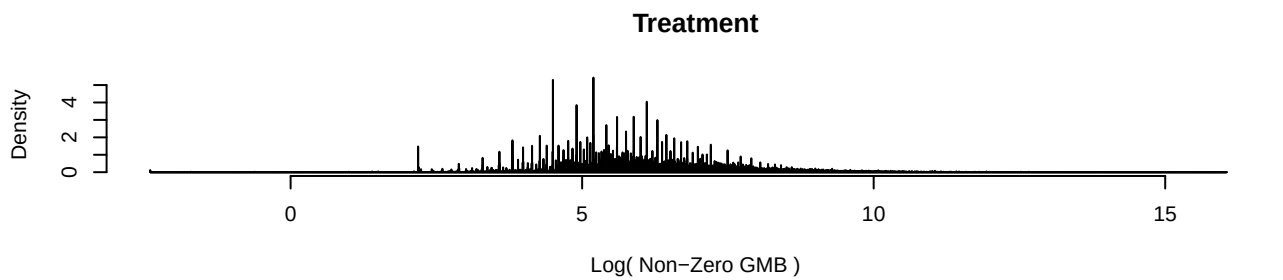
	nonzero values min	nonzero values max
treatment	0.09	9,381,532
control	0.09	12,018,199

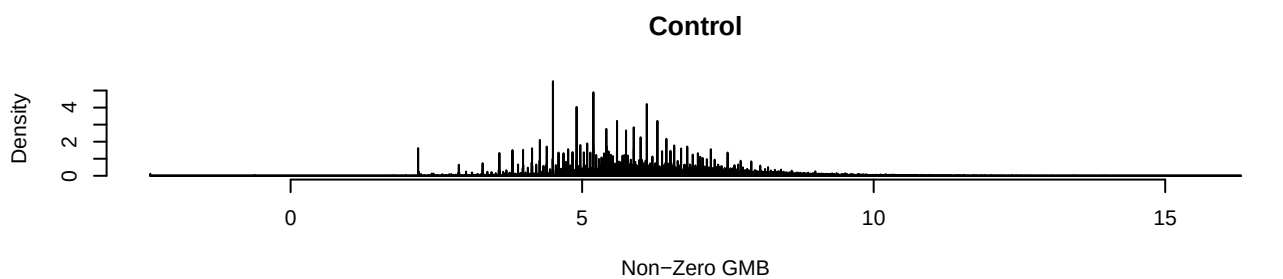
lift estimate +0.0636 = + 6.36%
sd of lift estimate 0.1400 = +14.00%
p(theta > 0 | data, background information): gaussian 0.675 optimal 0.696

eBay Case Study Details (continued)

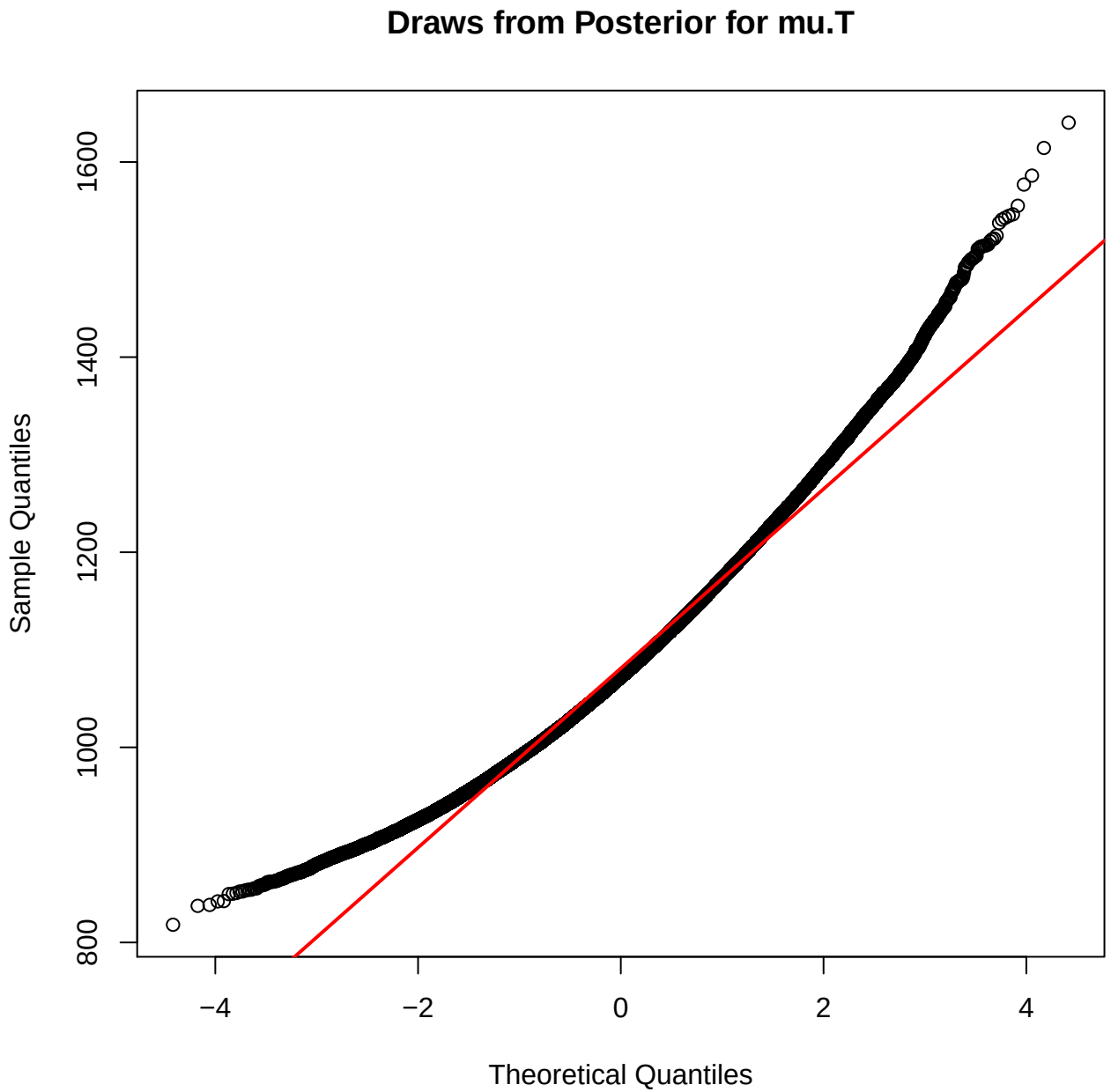




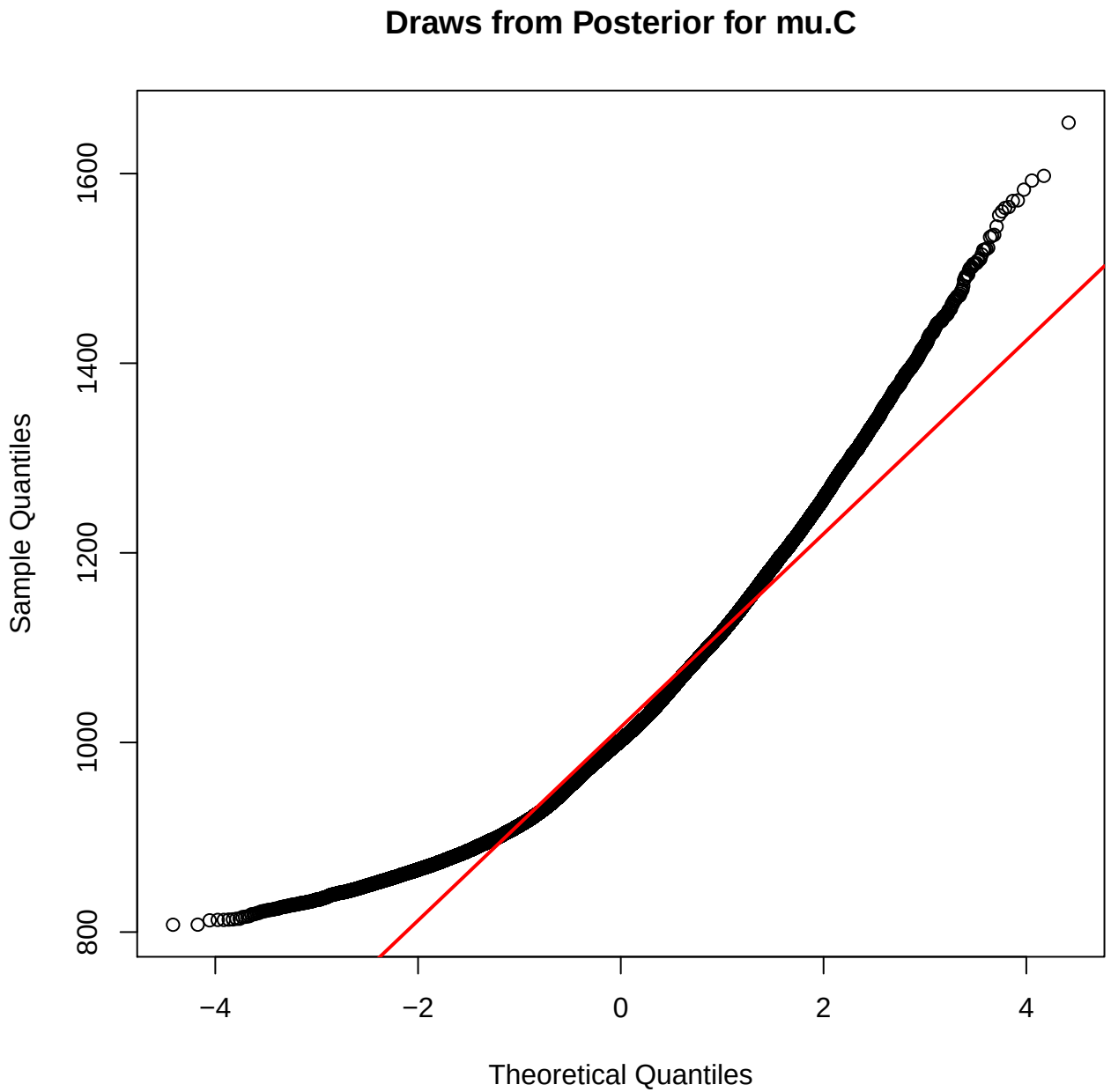




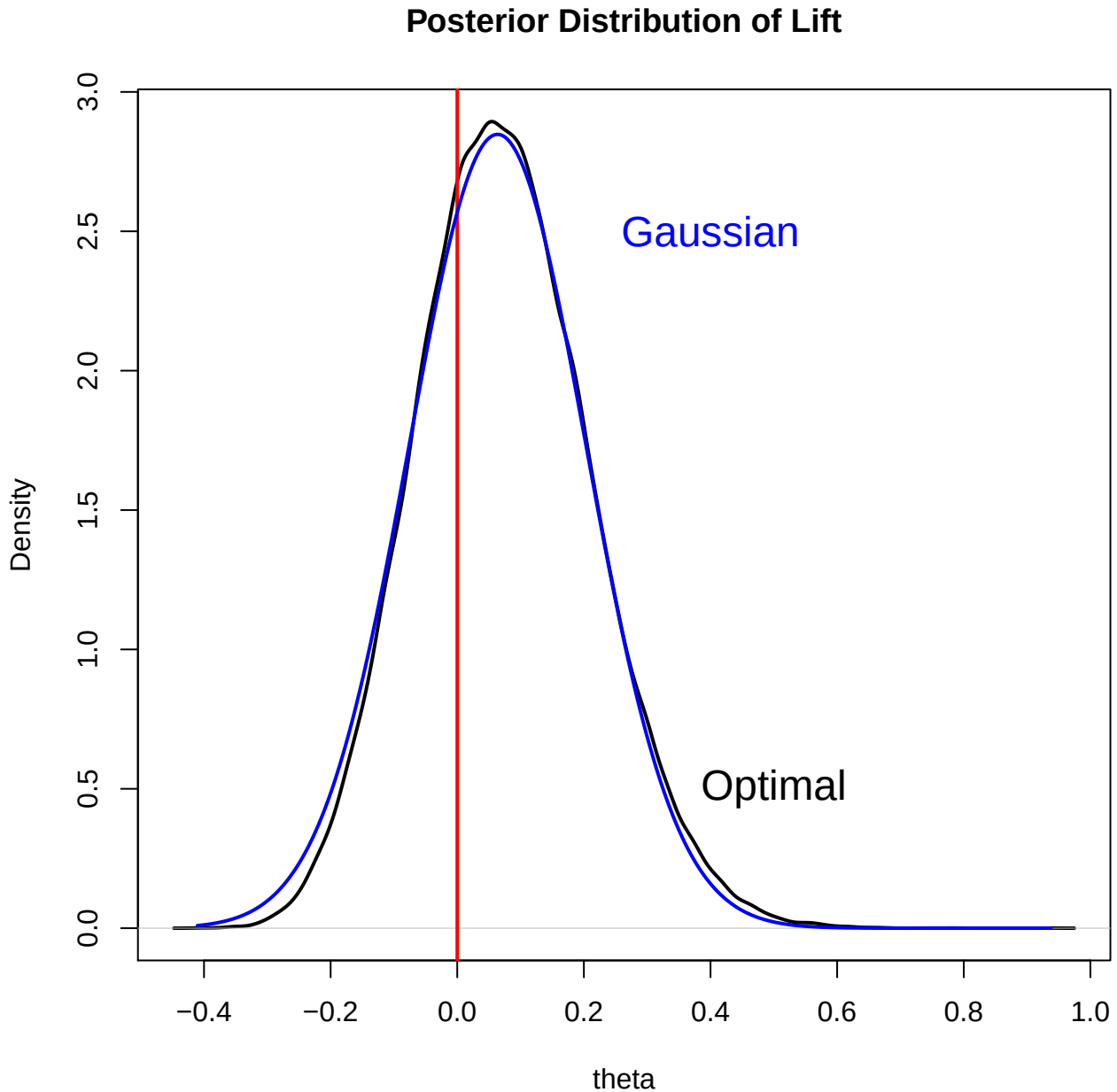
eBay Case Study Details (continued)



eBay Case Study Details (continued)



eBay Case Study Details (continued)



R Code For Parallel Bootstrapping

```
library( doParallel )
n.processors <- makeCluster( 24 )
registerDoParallel( n.processors )

parallel.mean.bootstrap <- function( y, M, n, p.hat.0 ) {
  foreach( i = 1:M, .inorder = F, .multicombine = T,
    .combine = 'c' ) %dopar% {
    sum( sample( y, n - rbinom( 1, n, p.hat.0 ),
      replace = T ) ) / n
  }
}

seed <- 1
set.seed( seed )
M.b <- 100000

system.time(
  mu.T.star.uncapped.1 <-
    parallel.mean.bootstrap( nonzero.T.values.uncapped, M.b, n.T,
    p.hat.0.T )
)
```

(based on suggestions from **Chris Severs**)

Summary of Conclusions (continued)

(ebay, not captopril) case study: n.C and n.T are just barely big enough for gaussian-assumption analysis to be decent

fact: when clt has not yet kicked in, gaussian-assumption analysis will be conservative in the right tail (positive lift) and liberal in the left tail

conservative in the right tail means that the gaussian-assumption analysis might say $p(\theta > 0|D\,\mathcal{B}\,\mathcal{G}) = 0.88$ when really the optimal analysis concludes that $p(\theta > 0|D\,\mathcal{B}) = 0.97$

this conservatism can be noticeable if n.C and n.T are quite small and the outcome variable is quite skewed and kurtotic

i'll take a position later on optimal ebay analysis vis à vis capping

Summary of A/B Testing Analysis Algorithms

Design: Identify $n = (n_C + n_T)$ eBay users representative of

$\mathcal{P} = \{$all future eBay users relevant to this experiment$\}$

(You have to specify relevant).

Randomize n_C of these users to C (current best eBay environment without the T intervention) and n_T to T (identical to C but with the T intervention).

(This is a completely-randomized experiment; better designs exist, but that's another talk.)

Data summaries: sample means $(\bar{y}_C, \bar{y}_T)$, sample SDs (s_C, s_T) for an outcome y such as GMB.

Inferential target: population lift $\theta = \frac{\mu_T - \mu_C}{\mu_C}$, in which μ_C (μ_T) is the population mean of y under the C (T) condition.

Algorithm (Gaussian approximation): (extremely fast, but may underestimate the posterior probability that the T intervention is beneficial, especially in segments with small sample sizes)

Gaussian Approximation Algorithm

$$\hat{\theta} = \frac{\bar{y}_T - \bar{y}_C}{\bar{y}_C}, \qquad \widehat{SD}(\hat{\theta}) = \sqrt{\frac{\bar{y}_T^2 \, s_C^2}{\bar{y}_C^4 \, n_C} + \frac{s_T^2}{\bar{y}_C^2 \, n_T}}$$

$$p(\theta > 0 | D \, \mathcal{B} \, \mathcal{G}) \quad \doteq \quad 1 - \Phi \left[\frac{-\hat{\theta}}{\widehat{SD}(\hat{\theta})} \right], \tag{7}$$

in which $\Phi(\cdot)$ is the standard normal CDF.

inferential suggestion (not yet a proper decision algorithm): consider launching the T if $p(\theta > 0 | D \, \mathcal{B} \, \mathcal{G}) > c$, where conventional (not necessarily in any sense optimal) values of c include 0.9, 0.95, and 0.99

this logic may be applied not only to the entire data set but also to smaller segments defined by covariates (features) (e.g., separately for male and female eBay users)

arriving at many such inferential suggestions —

$\{$entire data set, segment 1, segment 2, $\ldots$, segment $S\}$

— (for large S) creates a multiplicity problem that's best solved with Bayesian decision theory (another talk)

Gaussian Approximation Algorithm (continued)

R code to implement this approximate algorithm:

```
lift.estimate <- ( y.bar.T - y.bar.C ) / y.bar.C

SD.lift.estimate <- sqrt( ( y.bar.T^2 * s.C^2 ) /
  ( y.bar.C^4 * n.C ) + s.T^2 / ( y.bar.C^2 * n.T ) )

gaussian.posterior.probability.of.improvement <-
  1 - pnorm( ( 0 - lift.estimate ) / SE.lift.estimate )
```

even with (n_C, n_T) each on the order of 10–100 million, this code takes less than 1 second to run on a laptop with one decent core and decent RAM

approximate validity of Gaussian algorithm depends on (n_C, n_T) and the sample skewness and kurtosis values in each of C and T

$(*)$ unfavorable conditions for this algorithm: $\{$small sample size, large skewness, large kurtosis$\}$ in either or both groups

in a future white paper (published to the experimentation wiki) i'll quantify $(*)$

Optimal Analysis Algorithm

Algorithm (optimal analysis): (accurate assessment of the posterior probability that the T intervention is beneficial, but may be slow; however, the bootstrap is embarrassingly parallelizable)

to make a valid draw μ_T^* from the posterior distribution $p(\mu_T | y^T \mathcal{B})$ induced by the $DP\left(n, \hat{F}_T\right)$ posterior on F_T,

(a) choose a random sample $(y_1^{T*}, \ldots, y_{n_T}^{T*})$ of size n_T with replacement from the data vector y^T, and

(b) compute $\mu_T^* = \frac{1}{n_T} \sum_{\ell=1}^{n_T} y_\ell^{T*}$;

now repeat this M_b times (for large M_b) and use a histogram or kernel density trace of the resulting μ_T^* draws to approximate $p(\mu_T | y^T \mathcal{B})$.

this reasoning obviously applies in parallel to obtain the corresponding posterior $p(\mu_C | y^C \mathcal{B})$ for the control-group population mean, and then to simulate from $p(\theta | y \mathcal{B})$, where $y = \left(y^C, y^T\right)$, You just

(a) bind the columns $(\mu_{C1}^*, \ldots, \mu_{CM_b}^*)$ and $(\mu_{T1}^*, \ldots, \mu_{TM_b}^*)$ together to make a matrix with M_b rows and 2 columns,

Optimal Analysis Algorithm

(b) calculate $\theta^*_m = \frac{\mu^*_{Tm} - \mu^*_{Cm}}{\mu^*_{Cm}}$ in row $m = 1, \ldots, M_b$ of this matrix, and

(c) use a histogram or kernel density trace of the resulting M_b θ^* draws to approximate $p(\theta | D\,\mathcal{B})$.

	Elapsed Time (Sec) With		Bootstrap Distribution of mu.T.star			
Mb	8 Threads	24 Threads	Mean	SD	Skewness	Kurtosis
10,000	104.82	65.67	9.1279	0.036707	0.070319	-0.095457
			9.1276	0.037139	0.053797	0.017913
100,000	1049.81	694.97	9.1278	0.037074	0.041394	0.00094482
			9.1276	0.037086	0.048562	0.0087070

	Elapsed Time (Sec) With		Bootstrap Distribution of mu.C.star			
Mb	8 Threads	24 Threads	Mean	SD	Skewness	Kurtosis
10,000	114.64	---	9.2031	0.042402	0.046275	0.019909
100,000	1076.14	---	9.2031	0.042352	0.086158	0.058135