Variance reduction

Art B. Owen

Stanford University

Adapted from “Monte Carlo theory, methods and examples”

http://statweb.stanford.edu/~owen/mc/

Variance reduction

Probability is based on a random outcome $\omega \in \Omega$
with some sets $E \subset \Omega$,
and their probabilities $\mathbb{P}(E) \equiv \mathbb{P}(\omega \in E)$

In Monte Carlo, we control ω

Suppose that

$\mu = \mathbb{E}(f_0(\boldsymbol{x}))$ for $\boldsymbol{x} \sim p_0$, and

$\mu = \mathbb{E}(f_1(\boldsymbol{x}))$ for $\boldsymbol{x} \sim p_1$

Then we can work with **either** of those.

Outline

1) Antithetic sampling
2) Stratification
3) Control variates
4) Common random variables

Efficiency

Method	Variance	Cost
Old	σ_0^2/n_0	$n_0 c_0$
New	σ_1^2/n_1	$n_1 c_1$

To get $\mathrm{Var}(\hat{\mu}) = \tau^2$ we need $n_j = \sigma_j^2/\tau^2$.

That will cost $n_j c_j$.

The **relative** efficiency of the **new** method is

$$\frac{\text{old cost}}{\text{new cost}} = \frac{c_0 \sigma_0^2/\tau^2}{c_1 \sigma_1^2/\tau^2} = \frac{\sigma_0^2}{\sigma_1^2} \times \frac{c_0}{c_1}$$

Does not depend on τ^2 or n.

Variance reduction

Addresses the first factor σ_0^2/σ_1^2.

Keep an eye on the second factor c_0/c_1.

Also increasing σ_j^2 while lowering c_j could pay

How much reduction is 'worth it'?

It depends.

A 10% improvement might not be worth the nuisance,

unless the task is taking months of CPU [e.g., graphical rendering]

Reducing cost from 1 second to 0.01 seconds

Only saves you 0.99 seconds

but might allow you to embed your algorithm inside a loop

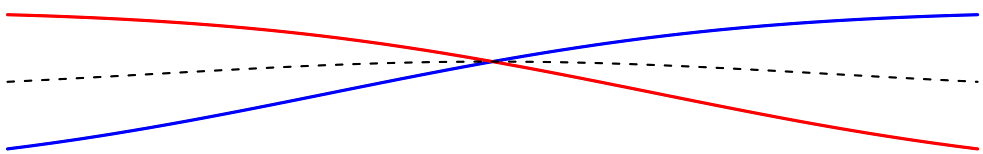
Simplicity has great value, though it is hard to quantify.

Antithetic sampling

Suppose that $f(x)$ is increasing over $0 \leqslant x \leqslant 1$.

If x_i is large then so is $f(x_i)$.

Antithetic sampling looks also at $f(1-x_i)$ to balance it out.



Antithetic estimator

$$\hat{\mu}_{\text{anti}} = \frac{1}{n/2} \sum_{i=1}^{n/2} \frac{f(\boldsymbol{x}_i) + f(\tilde{\boldsymbol{x}}_i)}{2} = \frac{1}{n} \sum_{i=1}^{n/2} \big(f(\boldsymbol{x}_i) + f(\tilde{\boldsymbol{x}}_i) \big)$$

More generally

For $\mu = \mathbb{E}(f(\boldsymbol{X}))$ for $\boldsymbol{X} \sim p$, suppose that

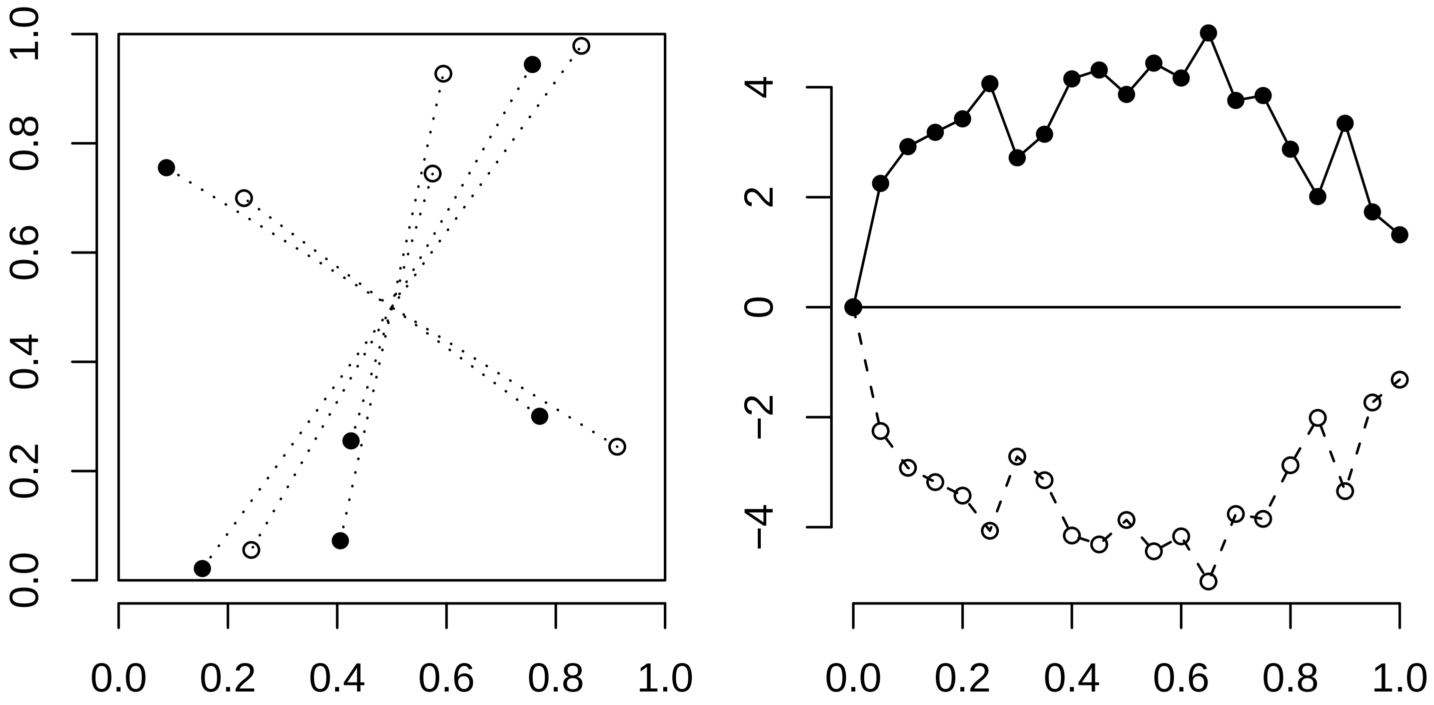
1) $\tilde{\boldsymbol{X}} \sim p$, and
2) $\tilde{\tilde{\boldsymbol{X}}} = \boldsymbol{X}$,

like $\tilde{x} = 1 - x$ does for $x \sim \mathbf{U}(0, 1)$.

Antithetics

$$\tilde{\boldsymbol{x}} = 1 - \boldsymbol{x}, \boldsymbol{x} \in [0,1]^d \qquad \tilde{S}(t) = -S(t), \quad 0 \leqslant t \leqslant 1$$

Some samples and antithetic counterparts



Antithetic variance

After a little algebra

$$\mathrm{Var}(\hat{\mu}_{\mathrm{anti}}) = \frac{\sigma^2}{n}(1+\rho), \quad \rho = \mathrm{Corr}(f(\boldsymbol{X}), f(\tilde{\boldsymbol{X}}))$$

Because $-1 \leqslant \rho \leqslant 1$

$$0 \leqslant \frac{\mathrm{Var}(\hat{\mu}_{\mathrm{anti}})}{\mathrm{Var}(\hat{\mu})} \leqslant 2$$

Worst case: we double σ^2.

Sometimes: lots of work to generate $\boldsymbol{x}$ and only a little for $\tilde{\boldsymbol{x}}$.

Odd and even functions

$$f(\boldsymbol{x}) = f_{\mathrm{E}}(\boldsymbol{x}) + f_{\mathrm{O}}(\boldsymbol{x})$$

$$f_{\mathrm{E}}(\boldsymbol{x}) \equiv \frac{1}{2}\bigl(f(\boldsymbol{x}) + f(\tilde{\boldsymbol{x}})\bigr) \qquad \sigma^2_{\mathrm{E}} = \mathrm{Var}(f_{\mathrm{E}}(\boldsymbol{X}))$$

$$f_{\mathrm{O}}(\boldsymbol{x}) \equiv \frac{1}{2}\bigl(f(\boldsymbol{x}) - f(\tilde{\boldsymbol{x}})\bigr) \qquad \sigma^2_{\mathrm{O}} = \mathrm{Var}(f_{\mathrm{O}}(\boldsymbol{X}))$$

After more algebra

$$\begin{pmatrix} \mathrm{Var}(\hat{\mu}) \\ \mathrm{Var}(\hat{\mu}_{\mathrm{anti}}) \end{pmatrix} = \frac{1}{n}\begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix}\begin{pmatrix} \sigma^2_{\mathrm{E}} \\ \sigma^2_{\mathrm{O}} \end{pmatrix}$$

Antithetics remove the odd component but double the even one.

We like it for odd f.

Exercise: $\rho = (\sigma^2_{\mathrm{E}} - \sigma^2_{\mathrm{O}})/(\sigma^2_{\mathrm{E}} + \sigma^2_{\mathrm{O}})$

Expected log return

We invest $\lambda_k \geqslant 0$ in stock k with $\sum_k \lambda_k = 1$.
Stock k grows by e^{X_k} per day.

Our fortune grows like $\exp(N\mu + o_p(N))$, where

$$\mu(\lambda) = \mathbb{E}\Bigl(\log\Bigl(\sum_k \lambda_k e^{X_k}\Bigr)\Bigr)$$

Example from notes

K stocks, $\lambda_k = 1/K$, $X_k \sim \mathcal{N}(0.001, 0.03^2)$
$t_{(4)}$ copula with $\Sigma = 0.3 \times \mathbf{1}\mathbf{1}^\mathsf{T} + 0.7 \times I$

Results from notes

Stocks	Period	Correlation	Reduction	Estimate	Uncertainty
20	week	-0.99957	2320.0	0.00130	6.35×10^{-6}
500	week	-0.99951	2030.0	0.00132	6.49×10^{-6}
20	year	-0.97813	45.7	0.06752	3.27×10^{-4}
500	year	-0.99512	40.2	0.06850	3.33×10^{-4}

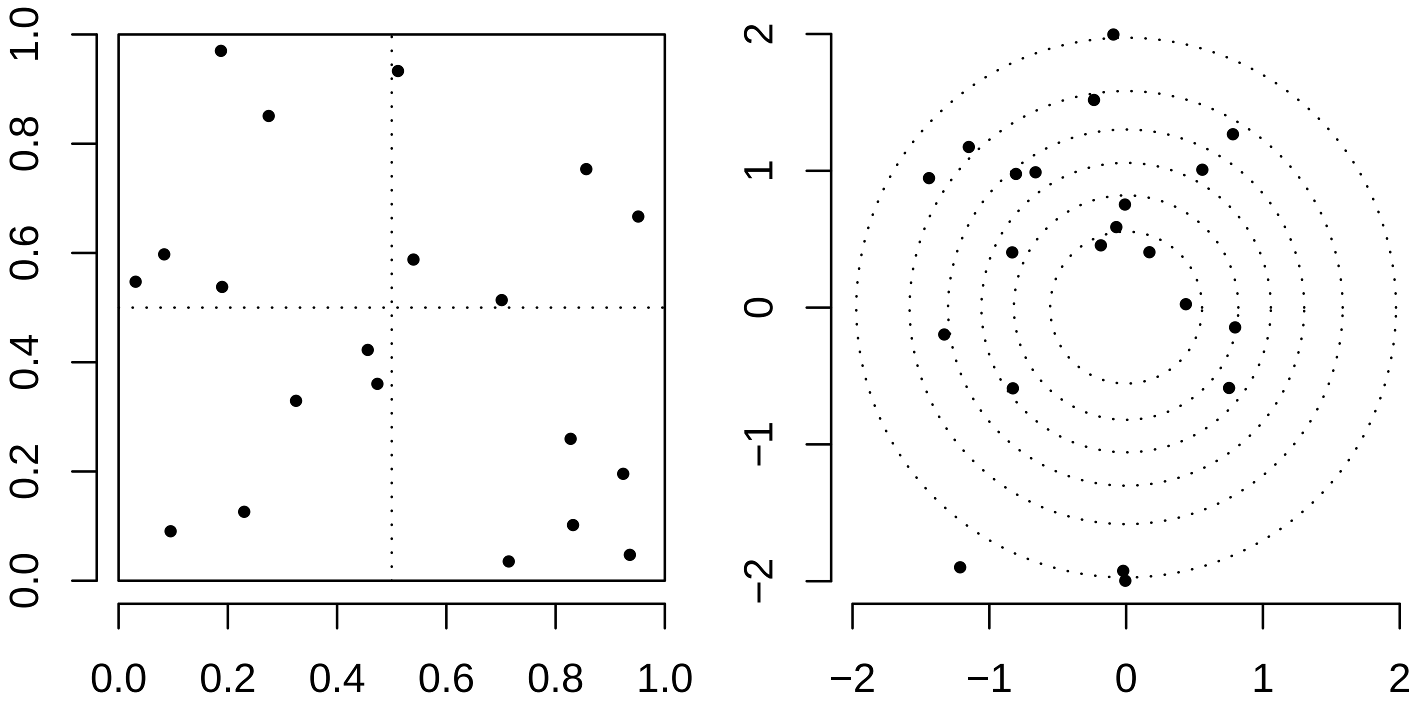
About antithetics

- The best way to see if it helps is to do it.
- Partial antithetics, flipping just some components of x also works.

Stratification

Partition $\mathcal{D} = \cup_{j=1}^{J} \mathcal{D}_j$, Sample $\boldsymbol{X}_{ij} \in \mathcal{D}_j$, $i = 1, \ldots, n_j$

Some stratified samples



5 points per subsquare or 3 points per 'ring'.

Stratification

Let $p_j(\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{D}_j)$.

Get $\boldsymbol{x}_{ij}$ from p_j

Moments

$$\hat{\mu}_{\text{strat}} = \sum_{j=1}^{J} \omega_j \times \frac{1}{n_j} \sum_{i=1}^{n_j} f(\boldsymbol{x}_{ij}), \qquad \omega_j = \mathbb{P}(\boldsymbol{X} \in \mathcal{D}_j)$$

$$\mathbb{E}(\hat{\mu}_{\text{strat}}) = \sum_{j=1}^{d} \mu_j = \mu$$

$$\text{Var}(\hat{\mu}_{\text{strat}}) = \sum_{j=1}^{d} \omega_j^2 \times \frac{\sigma_j^2}{n_j}$$

For stratum means μ_j and variances σ_j^2.

Within and between

$$f(\boldsymbol{x}) = f_W(\boldsymbol{x}) + f_B(\boldsymbol{x}) = \underbrace{\mu_{j(\boldsymbol{x})}}_{\text{within}} + \underbrace{f(\boldsymbol{x}) - \mu_{j(\boldsymbol{x})}}_{\text{between}}$$

$$\sigma_B^2 = \sum_{j=1}^{J} \omega_j (\mu_j - \mu)^2$$

$$\sigma_W^2 = \sum_{j=1}^{J} \omega_j^2 \sigma_j^2$$

Proportional sampling: $n_j \propto \omega_j$

After some algebra

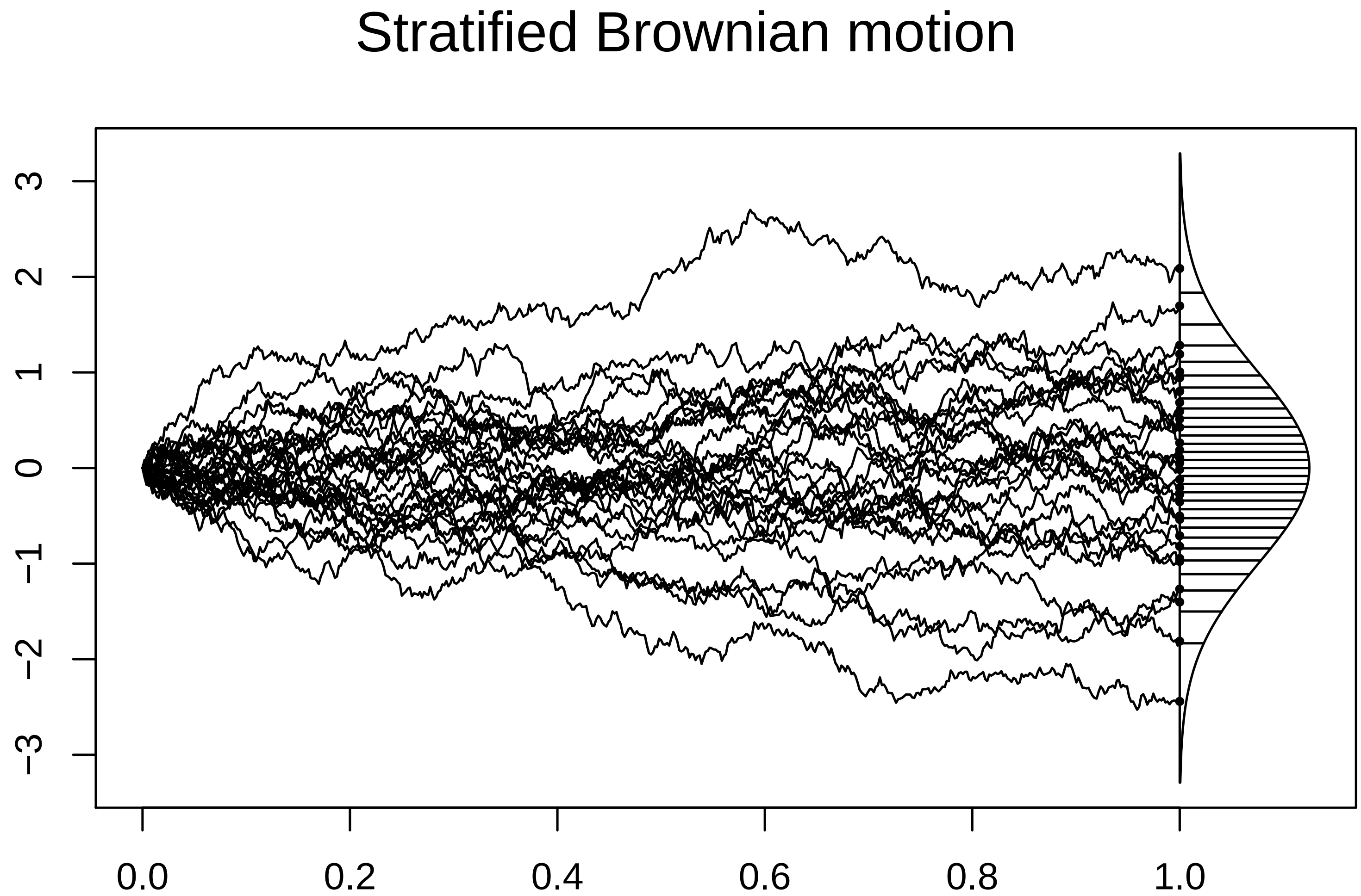
$$\begin{pmatrix} \mathrm{Var}(\hat{\mu}) \\ \mathrm{Var}(\hat{\mu}_{\mathrm{strat}}) \end{pmatrix} = \frac{1}{n} \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \sigma_W^2 \\ \sigma_B^2 \end{pmatrix}$$

Good strata give large σ_B^2.

Stratified process

Make final points representative.

Fill in conditionally.



Exercises

Post stratification: What if we sample $\boldsymbol{x}_i$ IID and then group them into strata afterwards?

What if we choose the strata after seeing the $\boldsymbol{x}_i$?

Non proportional sampling: What if n_j **not** proportional to ω_j?

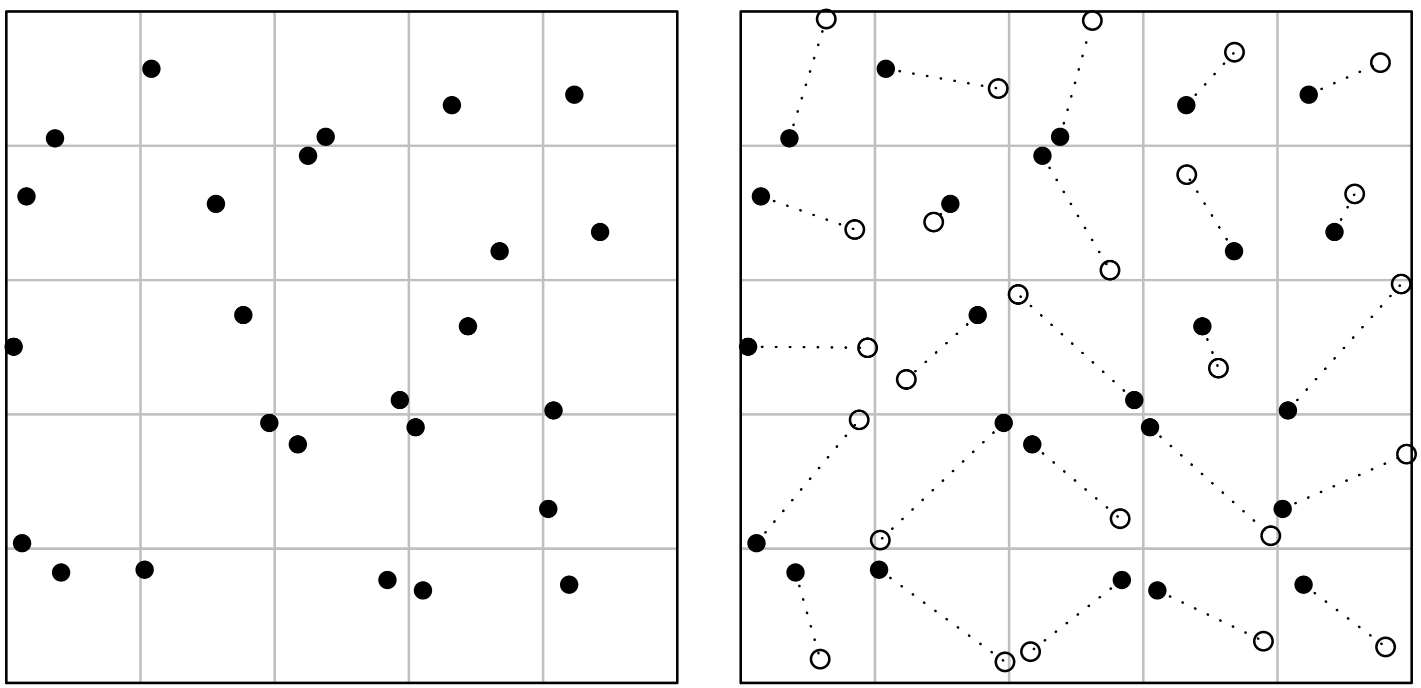
d dimensional stratification

Can get $\mathrm{Var}(\hat{\mu}_{\mathrm{strat}}) = O(n^{-1-2/d})$,

Or $O(n^{-1-4/d})$ with antithetics,

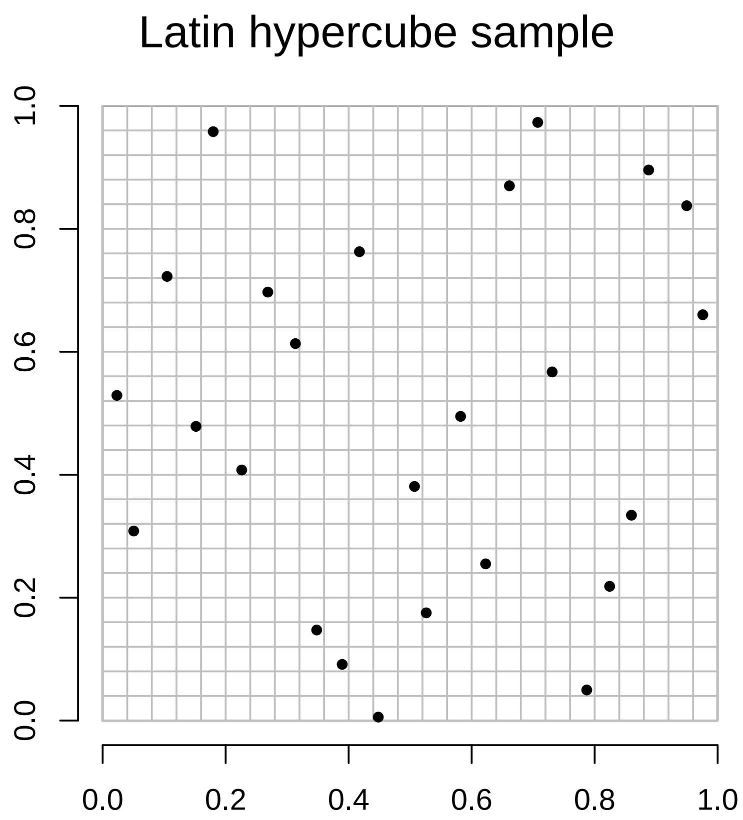
and some smoothness.

Grid based stratification



Original Antithetic

Latin hypercube sampling



- Stratify each dimension
- $x_{ij} = (\pi_j(i) - U_{ij})/n$
- π_j permutes $1, 2, \ldots, n$
- $U_{ij} \sim \mathbf{U}(0, 1)$
- Can have $d > n$

LHS ctd

For any $f \in L^2[0,1]^d$

$$\mathrm{Var}(\hat\mu_{\mathsf{LHS}}) \leqslant \frac{\sigma^2}{n-1}$$

So it is never much worse than plain MC.

ANOVA of $[0,1]^d$

Hoeffding (1948), Sobol' (1967)

$$f(\boldsymbol{x}) = \mu + f_1(x_1) + \cdots + f_d(x_d) + f_{1,2}(x_1,x_2) + \text{et cetera}$$

LHS gets the additive part at $o_p(n^{-1/2})$
the rest at $O_p(n^{-1/2})$
Stein (1987)

Orthogonal array sampling

We can balance bivariate margins too.

Ulitimate balance from quasi-Monte Carlo.

Control variates

We **want** $\mu = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$

and for $f \approx h$

we **know** $\theta = \int h(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$

Difference estimator

$$\hat{\mu}_{\mathrm{diff}} = \theta + \frac{1}{n}\sum_{i=1}^{n}\bigl(f(\boldsymbol{x}_i) - h(\boldsymbol{x}_i)\bigr) \equiv \theta + \hat{\mu} - \hat{\theta}$$

Ratio estimator

$$\hat{\mu} = \theta \times \frac{\hat{\mu}}{\hat{\theta}}$$

Product estimator

$$\hat{\mu} = \frac{\hat{\mu} \times \hat{\theta}}{\theta}$$

These can all help but there's something better.

Regression estimator

$$\hat{\mu}_\beta = \frac{1}{n}\sum_{i=1}^{n}\big(f(\boldsymbol{x}_i) - \beta h(\boldsymbol{x}_i)\big) + \beta\theta$$

$$\mathbb{E}(\hat{\mu}_\beta) = \mu, \quad \text{for any } \beta$$

The best β

$$\mathrm{Var}(\hat{\mu}_\beta) = \frac{1}{n}\Big(\mathrm{Var}(f(\boldsymbol{X})) - 2\beta\mathrm{Cov}(f(\boldsymbol{X}), h(\boldsymbol{X})) + \beta^2\mathrm{Var}(h(\boldsymbol{X}))\Big)$$

So it is a least squares problem. Optimal β yields

$$\mathrm{Var}(\hat{\mu}_{\beta_{\mathrm{opt}}}) = \frac{1}{n}\sigma^2(1-\rho^2)$$

$\rho \equiv \mathrm{Corr}(f(\boldsymbol{X}), h(\boldsymbol{X}))$

Via regression

Given $\int p(\boldsymbol{x})h_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \theta_j$ for $j = 1, \dots, J$

$$\hat{\mu}_\beta = \frac{1}{n}\sum_{i=1}^{n}\bigl(f(\boldsymbol{x}_i) - \beta^{\mathsf{T}}\boldsymbol{h}(\boldsymbol{x}_i)\bigr) + \beta^{\mathsf{T}}\theta$$

$$\hat{\beta} = \text{by least squares}$$

Short cut

Regress $Y_i \equiv f(\boldsymbol{x}_i)$ on $X_{ij} \equiv h_j(\boldsymbol{x}_i) - \theta_j$

Then $\hat{\mu}_{\hat{\beta}}$ is the **intercept**. You also get a standard error.

Estimated β

Our $\hat{\beta}$ is random, not fixed.

It's usually ok: $\hat{\beta} - \beta_{\text{opt}} = O_p(n^{-1/2})$.

For $J \ll n$.

Control variates

Maybe h has closed form and f is a 'tweak'

The h_j can be polynomials.

The h_j can be densities p_j.

Don't forget the additional cost of computing h_j .

Multiple everything

1) Multiple regression for control variates
2) Latin hypercube sampling is multiple stratification
3) Multiple importance sampling (coming later)
4) There is also multiple antithetic sampling

Moment matching

We get $\boldsymbol{x}_i$ but we know $\theta \equiv \mathbb{E}(\boldsymbol{X})$.

Adjust them: $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i + \theta - \bar{\boldsymbol{x}}$.

Or we know $\Sigma \equiv \mathbb{E}((\boldsymbol{X} - \theta)(\boldsymbol{X} - \theta)^{\mathsf{T}})$.

Rescale them

Boyle et al (1997) show it is like control variates with perhaps sub-optimal β.

Reweighting

Use $\sum_i w_i f(\boldsymbol{x}_i)$ where

$$\sum_{i=1}^{n} w_i \boldsymbol{h}(\boldsymbol{x}_i) = \theta, \quad \text{and} \quad \sum_{i=1}^{n} w_i = 1 \qquad (*)$$

The regression estimator already does this

but it can have $w_i < 0$

If we want positive weights

we can use empirical likelihood

maximize $\prod_i w_i$ subject to $(*)$

Conditioning

Sometimes we can integrate out part of the problem.

$$\int_0^1 \int_0^1 e^{g(x)y}\,\mathrm{d}y\,\mathrm{d}x = \int_0^1 h(x)\,\mathrm{d}x, \quad h(x) = (e^{g(x)} - 1)/g(x)$$

For $h(\boldsymbol{x}) = \mathbb{E}(f(\boldsymbol{x}, \boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{x})$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{y}_i) \quad \text{vs} \quad \hat{\mu}_{\mathrm{cond}} = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}_i)$$

$$\mathrm{Var}(\hat{\mu}_{\mathrm{cond}}) = \frac{1}{n}\mathrm{Var}\big(f(\boldsymbol{X}, \boldsymbol{Y}) \mid \boldsymbol{X}\big) \leqslant \frac{1}{n}\mathrm{Var}\big(f(\boldsymbol{X}, \boldsymbol{Y})\big) = \mathrm{Var}(\hat{\mu})$$

But check

whether h costs more than f.

Rao-Blackwell theorem

In statistical theory:

> If we can find **any** unbiased estimate $\hat{\theta}$ of θ,
>
> and a complete sufficient statistics S

Then $\mathbb{E}(\hat{\theta} \mid S)$ is a minimum variance unbiased estimate of θ.

Rao-Blackwellization

In Monte Carlo conditioning is sometimes called Rao-Blackwellization.

There is usually no sufficient statistic.

Example: roulette Wilson (1965)

Number	Wheel 1	Wheel 2
00	2127	1288
1	2082	1234
36	2221	1251
24	2192	1164 **w**
3	2008	1438 **b**
15	2035	1264
17	2044	1326
32	2133	1302
20	1912 **w**	1227
7	1999	1192
11	1974	1278
18	2191	1392
31	2192	1306
19	2284 **b**	1330
8	2136	1266
12	2110	1224
. . .	. . .	. . .
10	2121	1320
27	2158	1336
Avg	2100	1279.16

Hole 19

Hole 19 is the best on wheel 1. Seems to pay $2284/2100$ times average.

That would be a long term win.

What is $\mathbb{P}(\text{19 is best})$?

If counts C_j are $\mathrm{Mult}(N, \boldsymbol{p})$ and prior $\boldsymbol{p} \sim \mathrm{Dir}(1, \dots, 1)$

then $\boldsymbol{p} \mid \text{counts} \sim \mathrm{Dir}(\cdots, 1 + C_j, \cdots)$

$$\mathbb{P}\Big(p_{19} = \max_{1 \leqslant j \leqslant 38} p_j\Big)$$

Dirichlet via normalized Gamma

Recall $p_j \overset{\mathrm{d}}{=} \frac{X_j}{\sum_k X_k} \quad X_j \sim \mathrm{Gam}(1 + C_j)$

$$\mathbb{P}(p_{19} \text{ best} \mid X_{19} = x_{19}) = \prod_{k \neq 19} \mathbb{P}(X_k \leqslant x_{19}) \equiv h(x_{19})$$

So we sample $X_{19} \sim \mathrm{Gam}(C_{19} + 1)$ and average $h(X_{19})$.

Exercises

Find this value

Find $\mathbb{P}(19 \text{ is second best})$

Find $\mathbb{P}(19 \text{ pays})$

Empirical Bayes

Common variates

Now $f \approx g$, and we want $\Delta = \mathbb{E}(f(\boldsymbol{X}) - g(\boldsymbol{X}))$. So use

$$\hat{\Delta} = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i)$$

Intuitively better than

$$\hat{\Delta} = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i) - \sum_{i=n+1}^{2n} g(\boldsymbol{x}_i)$$

Who would even do that?

Really better so long as $\mathrm{Corr}(f(\boldsymbol{X}), g(\boldsymbol{X})) > 0$.

Especially if cost $\boldsymbol{x} \sim p$ is large.

Coupling

Same f, different p:

Now $\Delta = \mathbb{E}(f(\boldsymbol{X}) \mid \boldsymbol{X} \sim p) - \mathbb{E}(f(\boldsymbol{X}) \mid \boldsymbol{X} \sim q)$

$$\hat{\Delta} = \frac{1}{n}\sum_{i=1}^{n} f(\psi_p(\boldsymbol{u}_i)) - f(\psi_q(\boldsymbol{u}_i))$$

Here $\psi_p(\boldsymbol{U}) \sim p$ and $\psi_q(\boldsymbol{U}) \sim q$

Parametric p

$$\boldsymbol{X} = \psi_\theta(\boldsymbol{U}) \sim p(\cdot;\theta) \qquad \theta \in \Theta$$

Called the “reparametrization trick” in machine learning.

It supports differentation wrt θ.

A space of fs

From a parametric function

$$\mu(\theta) = \int h(\boldsymbol{x}, \theta) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \quad \theta \in \Theta \subset \mathbb{R}^d$$

$$\hat{\mu}(\theta_j) = \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}_i, \theta_j), \quad j = 1, \dots, J$$

Double loop over i and j.

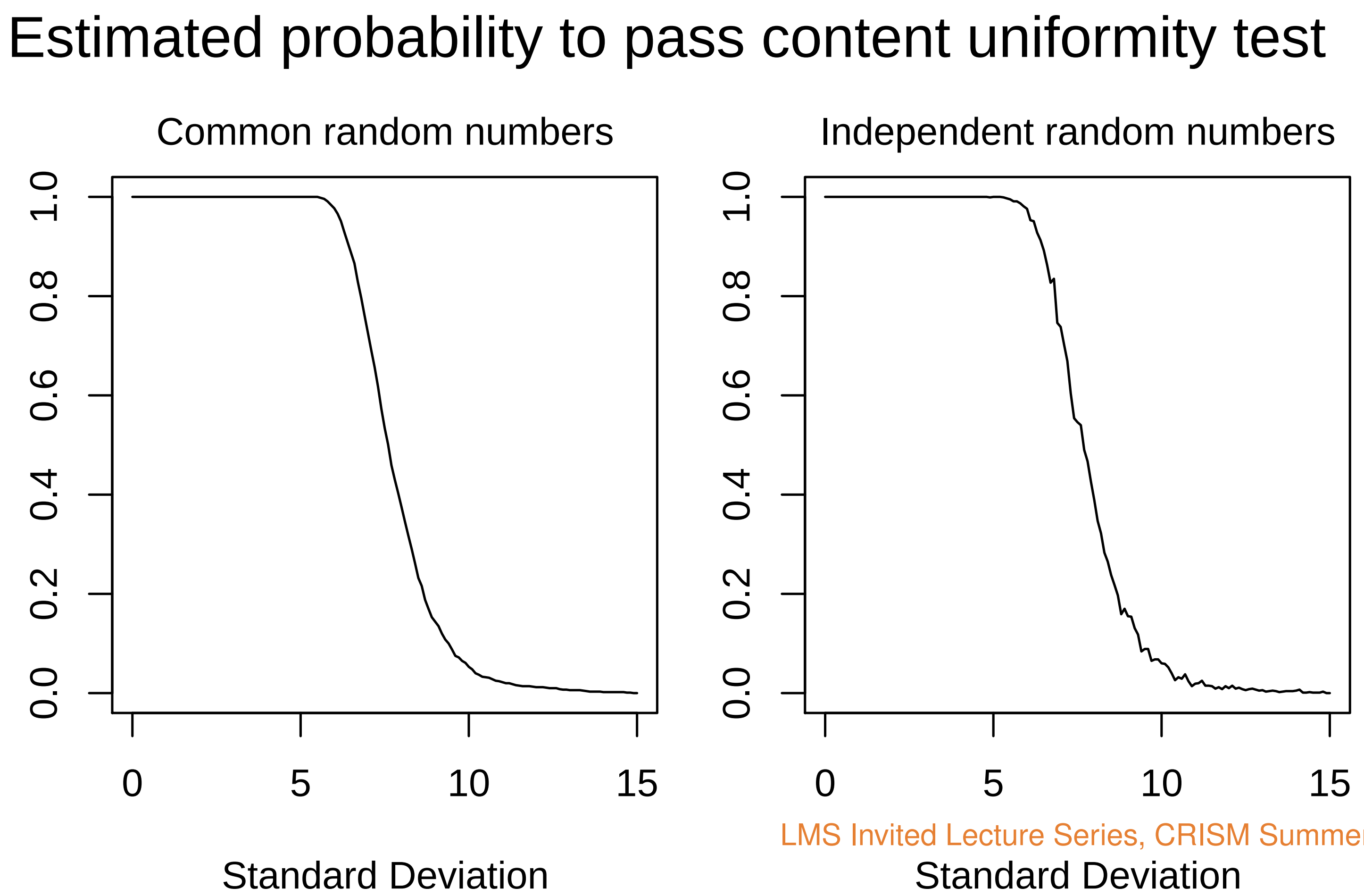
If j is the outer loop, reset your random seed!

Content uniformity trials

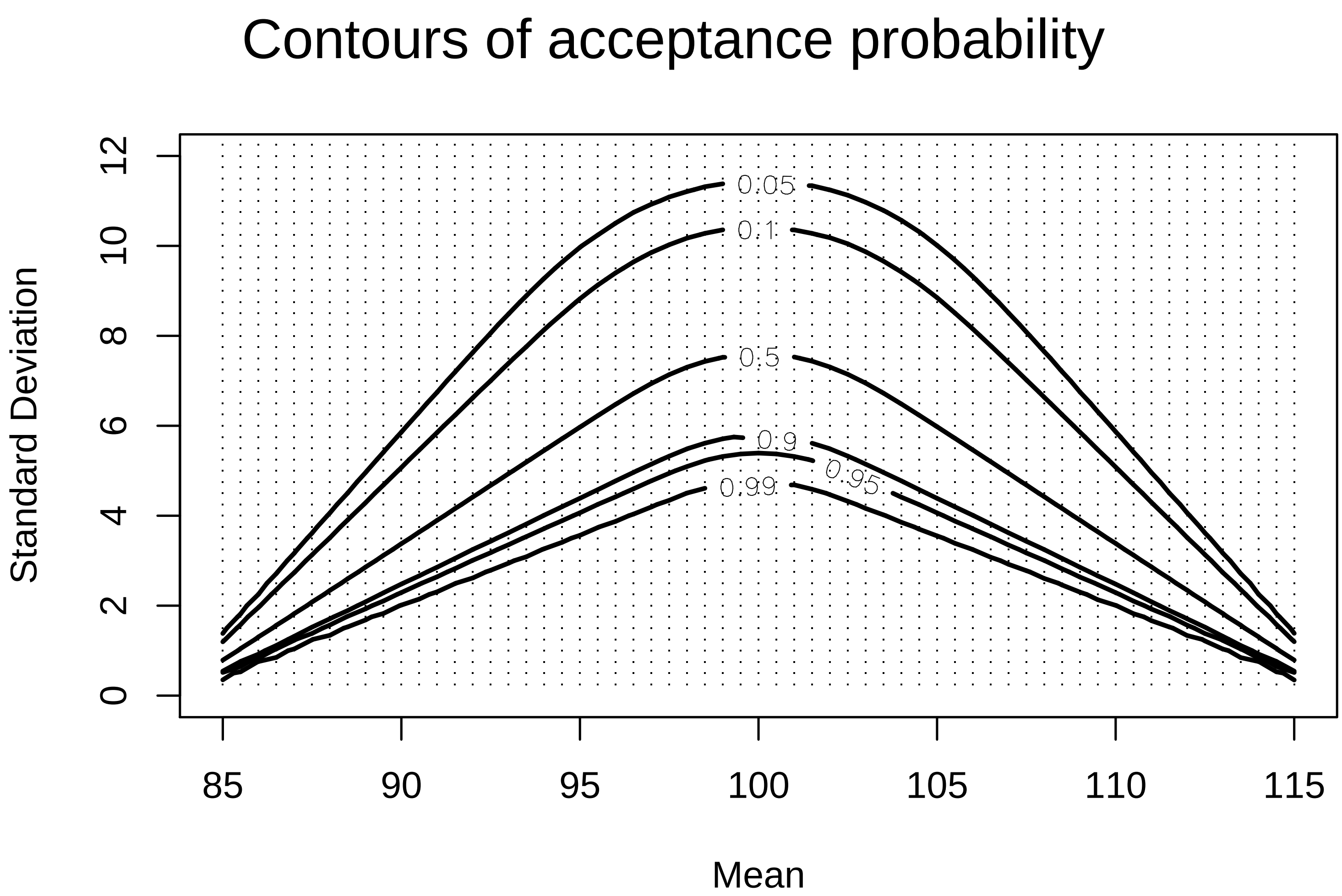
Will a batch of medications meet their specified doses?

Complicated multistage sampling rule from regulator.

Target potency 100. Suppose $X \sim \mathcal{N}(100, \sigma^2)$.



Vary μ and σ



Order statistics

Product fails when r out of k components have failed.

Component times $X_j \overset{\text{iid}}{\sim} F$

Mean failure time

Sample $X_{ij} \overset{\text{iid}}{\sim} F, \quad i = 1, \ldots, n, \quad j = 1, \ldots, k$

Sort $X_{i(1)} \leqslant X_{i(2)} \leqslant \cdots \leqslant X_{i(k)}$

Average the $X_{i(r)}$

Via inversion

If $u_1, \ldots, u_k \overset{\text{iid}}{\sim} \mathbf{U}(0,1)$

then $u_{(r)} \sim \text{Beta}(r, k-r+1)$

Generate $v_i \overset{\text{iid}}{\sim} \text{Beta}(r, k-r+1)$

Average $F^{-1}(v_i)$.

Control variates plus

Plus antithetics

Antithetic sampling for f with a control variate h.

It helps if f_E is correlated with h_E

Correlation from the ‘odd parts’ does no good.

Plus stratification

It helps if f and h are correlated ‘within strata’.

Plus LHS

It helps if the “nonadditive parts” of f and h are correlated.

You can’t subtract the same source of variance twice.

Thanks

- Lecturers: Nicolas Chopin, Mark Huber, Jeffrey Rosenthal
- Guest speakers: Michael Giles, Gareth Roberts
- The London Mathematical Society: Elizabeth Fisher, Iain Stewart
- CRISM & The University of Warwick, Statistics
- Sponsors: Amazon, Google
- Partners: ISBA, MCQMC, BAYSM
- Poster: Talissa Gasser, Hidamari Design
- NSF: DMS-1407397 & DMS-1521145
- Planners: Murray Pollock, Christian Robert, Gareth Roberts
- Support: Paula Matthews, Murray Pollock, Shahin Tavakoli