Monte Carlo Methods

Simon Jackman

Stanford University
http://jackman.stanford.edu/BASS

February 3, 2012

The Monte Carlo principle

anything we want to know about a random variable θ can be learned by sampling many times from $p(\theta)$, the density of θ.

The Monte Carlo principle

anything we want to know about a random variable θ can be learned by sampling many times from p(θ), the density of θ.

- Example: compute $E(\theta)$, $\theta \in \Theta \subseteq \mathbb{R}$.
- Analytically: $E(\theta) = \int_{\Theta} \theta p(\theta) d\theta$.

The Monte Carlo principle

anything we want to know about a random variable θ can be learned by sampling many times from $p(\theta)$, the density of θ.

- Example: compute $E(\theta), \theta \in \Theta \subseteq \mathbb{R}$.
- Analytically: $E(\theta) = \int_{\Theta} \theta p(\theta) d\theta$. This math might be “too hard”.

The Monte Carlo principle

anything we want to know about a random variable θ can be learned by sampling many times from $p(\theta)$, the density of θ.

- Example: compute $E(\theta)$, $\theta \in \Theta \subseteq \mathbb{R}$.
- Analytically: $E(\theta) = \int_\Theta \theta p(\theta) d\theta$. This math might be “too hard”.
- Monte Carlo estimate:
  - sample θ from $p(\theta)$; call these $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}$; T large.
  - compute the following *Monte Carlo estimate* of $E(\theta)$:

$$\widehat{E(\theta)}_T = \sum_{t=1}^{T} \theta^{(t)}/T$$

The Monte Carlo principle

anything we want to know about a random variable θ can be learned by sampling many times from $p(\theta)$, the density of θ.

- Example: compute $E(\theta)$, $\theta \in \Theta \subseteq \mathbb{R}$.
- Analytically: $E(\theta) = \int_\Theta \theta p(\theta) d\theta$. This math might be "too hard".
- Monte Carlo estimate:
  - sample θ from $p(\theta)$; call these $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}$; T large.
  - compute the following *Monte Carlo estimate* of $E(\theta)$:

$$\widehat{E(\theta)}_T = \sum_{t=1}^{T} \theta^{(t)}/T$$

- accuracy of estimate $\widehat{E(\theta)}_T$ improves as $T \to \infty$.

The Monte Carlo principle

> *anything we want to know about a random variable θ can be learned by sampling many times from p(θ), the density of θ.*

- Example: compute $E(\theta), \theta \in \Theta \subseteq \mathbb{R}$.
- Analytically: $E(\theta) = \int_\Theta \theta p(\theta) d\theta$. This math might be "too hard".
- Monte Carlo estimate:
  - sample θ from $p(\theta)$; call these $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}$; T large.
  - compute the following *Monte Carlo estimate* of $E(\theta)$:

$$\widehat{E(\theta)}_T = \sum_{t=1}^{T} \theta^{(t)}/T$$

- accuracy of estimate $\widehat{E(\theta)}_T$ improves as $T \to \infty$.
- all this generalizes to (a) vectors $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)'$; (b) estimates of functionals of $\boldsymbol{\theta}$, $h(\boldsymbol{\theta})$.

Brief history of Monte Carlo methods

- Buffon's needle, estimating π; see `simpi` in my `pscl` package for `R`
- Lord Kelvin
- Enrico Fermi
- Metropolis and Ulam; Manhattan Project, Los Alamos

Brief history of Monte Carlo methods

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 247 SEPTEMBER 1949 *Volume 44*

THE MONTE CARLO METHOD

NICHOLAS METROPOLIS AND S. ULAM
Los Alamos Laboratory

We shall present here the motivation and a general description of a method dealing with a class of problems in mathematical physics. The method is, essentially, a statistical approach to the study of differential equations, or more generally, of integro-differential equations that occur in various branches of the natural sciences.

Simulation Consistency

Theorem (Simulation Consistency, Independent Draws, Part 1)

Suppose $\{\boldsymbol{\theta}^{(t)}\}$ *is a sequence of independent draws from the density* $f(\boldsymbol{\theta})$*, with* $\boldsymbol{\theta} \in \mathbb{R}^k$ *and* $h : \mathbb{R}^k \to \mathbb{R}$*. Then*

$$\bar{h}^{(T)} = T^{-1} \sum_{t=1}^{T} h(\boldsymbol{\theta}^{(t)}) \stackrel{a.s.}{\to} E[h(\boldsymbol{\theta})]$$

Proof.

The claim is a restatement of the strong law of large numbers (e.g., Proposition B.10; BASS). □

Simulation Consistency

Theorem (Simulation Consistency, Independent Draws, Part 2)

Further, suppose that for $p \in (0, 1)$, there is a unique q_p such that $Pr[h(\boldsymbol{\theta}) \leq q_p] \geq p$ and $Pr[h(\boldsymbol{\theta}) \geq q_p] \geq 1 - p$ are both true. Consider $q_p^{(T)} \in \mathbb{R}$ such that

$$T^{-1} \sum_{t=1}^{T} \mathcal{I}(-\infty < h(\boldsymbol{\theta}^{(t)}) < q_p^{(T)}) \geq p$$

where $p \in (0, 1)$ and $\mathcal{I}(\cdot)$ is a binary indicator function, equal to 1 if its argument is true, and zero otherwise. Then $q_p^{(T)} \stackrel{a.s.}{\rightarrow} q_p$.

Proof.

Geweke (2005, Theorem 4.1.1); Rao (1973, 423); van der Vaart (1998, 305). □

Learning about a uniform random variable

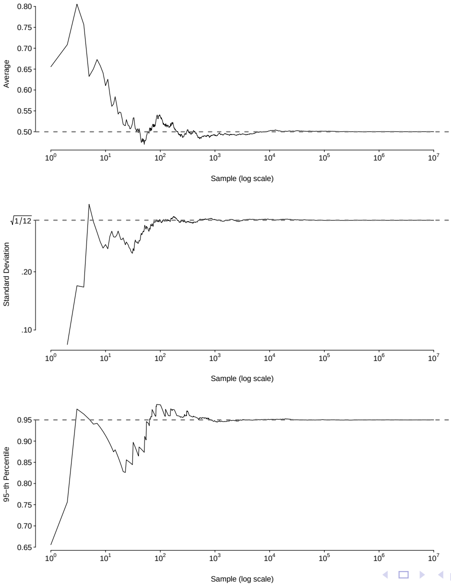
- Suppose $\theta \sim \text{Unif}(0, 1)$
- But we have forgotten that $E(\theta) = .5$, $\text{sd}(\theta) = \sqrt{1/12}$, $\Pr(\theta \leq q) = q$, $q \in (0, 1)$ etc.
- Monte Carlo methods? Use `rnorm` in `R`.

1. the average of the sampled values, $\bar{\theta}^{(T)} = T^{-1} \sum_{t=1}^{T} \theta_t$, will be very close to $E(\theta) = .5$. Importantly, $\bar{\theta}^{(T)}$ gets arbitrarily close to .5 as $T \rightarrow \infty$.
2. likewise, the standard deviation of the sampled values

$$\text{sd}(\theta)^{(T)} = \left[(T - 1)^{-1} \sum_{t=1}^{T} (\theta_t - \bar{\theta}^{(T)})^2 \right]^{1/2}$$

   will get arbitrarily close to $\sqrt{1/12}$ as $T \rightarrow \infty$.
3. the p-th quantile of the sampled values, $q_p^{(T)}$, $p \in (0, 1)$; $q_p^{(T)} \overset{a.s.}{\rightarrow} p$.

Learning about a uniform random variable (Figure 3.1)



Monte carlo integration/marginalization (method of composition)

- $\boldsymbol{\theta} = (\theta_1, \theta_2)$, with $\theta_j \in \Theta_j \subseteq \mathbb{R}, j = 1, 2$.
- Posterior density: $p(\boldsymbol{\theta}|\mathbf{y})$.
- But interest centers on the marginal posterior density of θ_1,

$$p(\theta_1|\mathbf{y}) = \int_{\Theta_2} p(\theta_1, \theta_2|\mathbf{y}) d\theta_2 = \int_{\Theta_2} p(\theta_1|\theta_2, \mathbf{y}) p(\theta_2|\mathbf{y}) d\theta_2.$$

- Method of composition:

  1: **for** $t = 1$ to T **do**
  2: sample $\theta_2^{(t)}$ from $p(\theta_2|\mathbf{y})$
  3: sample $\theta_1^{(t)}$ from $p(\theta_1|\theta_2^{(t)}, \mathbf{y})$.
  4: **end for**

- $\theta_1^{(t)} \sim p(\theta_1|\mathbf{y})$, as desired.

Method of Composition to sample from a t density

$$
\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) &= \int_0^\infty p(\boldsymbol{\beta}|\sigma^2,\mathbf{y},\mathbf{X})p(\sigma^2|\mathbf{y},\mathbf{X})d\sigma^2 \\
p(\sigma^2|\mathbf{y},\mathbf{X}) &\equiv \text{inverse-Gamma}(\nu/2, \nu s^2/2) \\
p(\boldsymbol{\beta}|\sigma^2,\mathbf{y},\mathbf{X}) &\equiv N(\mathbf{b}, \sigma^2\mathbf{B})
\end{aligned}
$$

1: **for** $t = 1$ to T **do**
2: sample $\sigma^{2\,(t)}$ from $p(\sigma^2) \equiv \text{inverse-Gamma}(\nu/2, \nu s^2/2)$
3: sample $\boldsymbol{\beta}^{(t)}$ from $p(\boldsymbol{\beta}|\sigma^{2\,(t)}) \equiv N(\mathbf{b}, \sigma^{2\,(t)}\mathbf{B})$
4: **end for**

i.e., $\boldsymbol{\beta}^{(t)} \sim p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) \equiv$ student - t.

Monte carlo inference, functions of parameters

- 2-by-2 table:

	x_i		
y_i	0	1	
0	n_0 - r_0	n_1 - r_1	
1	r_0	r_1	
	n_0	n_1	n

- Two success probabilities: $0 \leq \theta_0, \theta_1 \leq 1$; $\Pr(y_i = 1 | x_i = j) = \theta_j$, $j \in \{0, 1\}$.
- MLEs: $\hat{\theta}_j = r_j / n_j, j \in \{0, 1\}$.
- Bayesian analysis: independent, conjugate Beta priors over each θ_j, $\theta_j \sim \text{Beta}(\alpha_j, \beta_j)$. posterior densities are independent Beta densities $\theta_j | r_j, n_j \sim \text{Beta}(\alpha_j + r_j, \beta_j + n_j - r_j)$.

Inference for Difference of Two Binomial Proportions

- $q = \theta_1 - \theta_0$; difference of two binomial proportions
- $p(\theta_j | r_j, n_j) \equiv$ Beta.
- But what is $p(q | r_0, r_1, n_0, n_1)$?
- Until recently (Pham-Gia and Turkkan 1993), the density of the difference of two Betas was unknown (unavailable in closed form)
- Characterize this density by Monte Carlo methods.

Algorithm:

1. sample $\theta_0^{(t)}$ from Beta$(\alpha_0 + r_0, \beta_0 + n_0 - r_0)$
2. sample $\theta_1^{(t)}$ from Beta$(\alpha_1 + r_1, \beta_1 + n_1 - r_1)$
3. compute $q^{(t)} = \theta_1^{(t)} - \theta_0^{(t)}$

Repeat many times, $t = 1, \ldots, T$; sampled $q^{(t)}$ are a sample from the posterior density of q; summarize numerically, make histogram, etc.

Example 3.2; War and revolution in Latin America

Sekhon (2005); Geddes (1990); Skocpol (1979):

	Revolution	No Revolution
Defeated & Invaded/Lost Territory	1	7
Not Defeated within 20 years	2	74

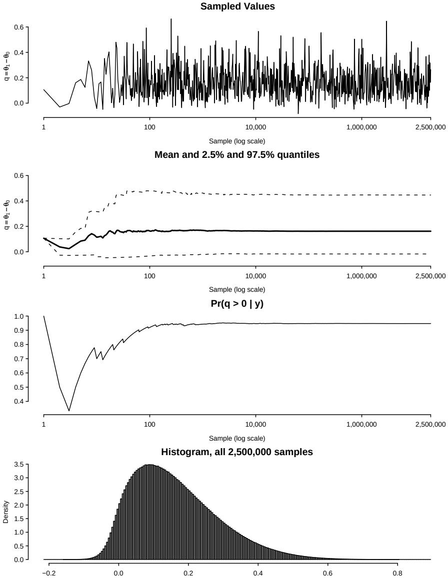
- Each observation spans 20 years for each Latin American country
- The sole observation in the top left of the cross-tabulation is Bolivia: it suffered a military defeat in 1935, and a social revolution in 1952.
- The two observations in the lower left of the table are Mexico (revolution in 1910) and Nicaragua (revolution in 1979).
- The MLEs are $\hat{\theta}_0 = 2/(2+74) = .026$ and $\hat{\theta}_1 = 1/(7+1) = .125$, suggesting that revolutions are much more likely conditional on military defeat than conditional on not having experienced a military defeat.
- Uniform priors $\theta_j \sim \text{Beta}(1,1), j = 0, 1$; posterior densities $\theta_0 | r_0, n_0 \sim \text{Beta}(3, 75)$ and $\theta_1 | r_1, n_1 \sim \text{Beta}(2, 8)$.

Example 3.2; War and revolution in Latin America

We seek the posterior density of $q = \theta_1 - \theta_0$; we use Monte Carlo methods:

for $t = 1$ to T **do**
  sample $\theta_1^{(t)}$ from $p(\theta_1|\mathbf{y}) \equiv \text{Beta}(2, 8)$
  sample $\theta_0^{(t)}$ from $p(\theta_0|\mathbf{y}) \equiv \text{Beta}(3, 75)$
  $q^{(t)} \leftarrow \theta_1^{(t)} - \theta_0^{(t)}$
end for

Example 3.2; War and revolution in Latin America



Example 3.2; War and revolution in Latin America

R code is trivial:

```
nsims <- 1e6
theta1 <- rbeta(nsims,2,8)
theta0 <- rbeta(nsims,3,75)
q <- theta1 - theta0
summary(q)
mean(q>0)
```

In JAGS

```
model{
        ## model for the data
        for(i in 1:2){
           r[i] ~ dbin(theta[i],n[i])
        }

        ## priors
        for(i in 1:2){
           theta[i] ~ dbeta(1,1)
        }

        ## quantity of interest
        q <- theta[2] - theta[1]
}
```

Sampling algorithms

Suppose $\theta \sim p$? How to sample from p?

1. inverse-CDF method
2. importance sampling
3. rejection sampling
4. slice sampler

Inverse-CDF method

- $\theta \sim p$, $\theta \in \Theta \subseteq \mathbb{R}$.
- $F(q) = \Pr(\theta \leq q) = \int_{-\infty}^{q} p(\theta)d\theta$; n.b., $F : \Theta \mapsto (0, 1)$.
- Suppose F^{-1} exists, is computable; $F^{-1} : (0, 1) \mapsto \Theta$
- Inverse-CDF algorithm:

  1: **for** $t = 1$ to T **do**
  2:     sample $p^{(t)} \sim \text{Unif}(0, 1)$
  3:     $\theta^{(t)} \leftarrow F^{-1}(p^{(t)})$
  4: **end for**

- Reasonably rare that an inverse-CDF exists. E.g., not available in closed form for the normal, but good approximations exist; e.g., Wichura (1988), used in the `pnorm` function in R).

Importance Sampling

- we can evaluate the target density at any given point in its support; i.e., we can compute $p(\theta) \, \forall \, \theta \in \Theta$.
- no algorithm for direct sampling from $p(\theta)$.
- we *can* sample from a density $s(\theta)$, where $s(\theta)$ has the property that $p(\theta) > 0 \Rightarrow s(\theta) > 0, \forall \, \theta \in \Theta$.
- Exploit the following identity: consider some $h(\theta)$, then

$$E[h(\theta)] = \int_\Theta h(\theta)p(\theta)d\theta = \int_\Theta h(\theta)p(\theta)/s(\theta)s(\theta)d\theta.$$

- Importance sampling algorithm:

1: **for** $t = 1$ to T **do**
2: sample $\theta^{(t)} \sim s(\theta)$.
3: $w^{(t)} \leftarrow p(\theta^{(t)})/s(\theta^{(t)})$
4: **end for**
5: $\bar{h}^{(T)} \leftarrow T^{-1} \sum_{t=1}^{T} h(\theta^{(t)}) w^{(t)}$

Accept-Reject Sampling (von Neumann 1951)

- $\theta \sim p(\theta)$, but can't sample from this density
- can sample from a *majorizing function* $g(\theta)$, that is, where $g(\theta) > p(\theta), \forall\, \theta$.
- can trivially find a majorizing function by rescaling a *proposal* density: $g(\theta) = cm(\theta)$:

1: **for** $t = 1$ to T **do**
2:     sample $z \sim m(\theta)$
3:     sample $u \sim \text{Unif}(0, 1)$
4:     $r \leftarrow p(z)/cm(z)$
5:     **if** $u \leq r$ **then**
6:         $\theta^{(t)} \leftarrow z$ {"accept"}
7:     **else**
8:         go to 2 {"reject"}
9:     **end if**
10: **end for**

Accept-Reject Sampling

- The target density p need only be known up to a factor of proportionality. All that matters is that we can sample from a function that majorizes the target density, and we can control that through the scaling constant, c.
- Useful in Bayesian analysis, where it is often the case that posterior densities are only known up to an (unknown) proportionality constant.
- An accept-reject algorithm produces potentially many draws that are rejected, and hence the algorithm can be computationally inefficient.
- More efficient if the majorizing function g closely approximates the target density, p; e.g., $r = p/g \approx 1$.

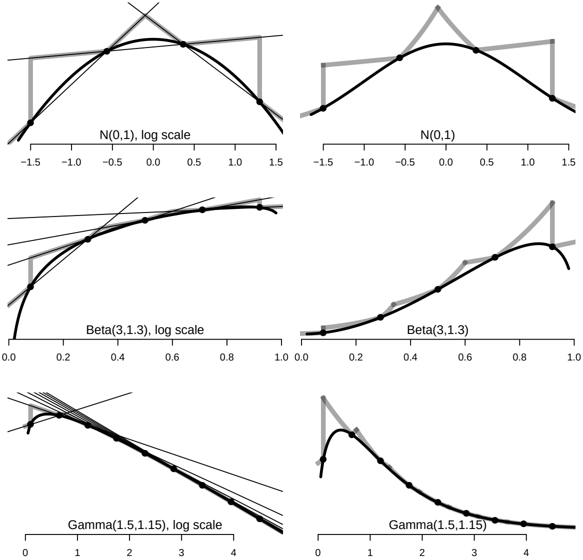
Adaptive Rejection Sampling

- hard to find a good proposal density
- but if target density $p(\theta)$ is log-concave and continuously differentiable, then use adaptive rejection sampling
- build a proposal density as a set of piecewise exponential densities bracketing the target density
- A density $p(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^k$, is log-concave if the determinant of

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 \log p}{\partial\theta_1 \partial\theta_1} & \frac{\partial^2 \log p}{\partial\theta_1 \partial\theta_2} & \cdots & \frac{\partial^2 \log p}{\partial\theta_1 \partial\theta_k} \\ \frac{\partial^2 \log p}{\partial\theta_2 \partial\theta_1} & \frac{\partial^2 \log p}{\partial\theta_2 \partial\theta_2} & \cdots & \frac{\partial^2 \log p}{\partial\theta_2 \partial\theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \log p}{\partial\theta_k \partial\theta_1} & \frac{\partial^2 \log p}{\partial\theta_k \partial\theta_2} & \cdots & \frac{\partial^2 \log p}{\partial\theta_k \partial\theta_k} \end{pmatrix}$$

is non-positive.

Adaptive Rejection Sampling



Adaptive Rejection Sampling

- algorithm is *adaptive*: successfully sampled points are added to the set of evaluation points.
- initialization/adaptation phase consists of getting a good set of evaluation points
- ARS was a critical step in developing a general purpose computer program for simulation-based Bayesian statistical analysis; e.g., Gilks and Wild (1992).
- Extension to non-log-concave densities (e.g., Gilks, Best and Tan 1995)

Slice Sampling §5.2.7

- $\theta \sim p(\theta), \theta \in \Theta \subseteq \mathbb{R}$, restricting ourselves to the one-dimensional case for the time being.
- This is equivalent to sampling the pair (θ, U) uniformly from the set $\mathcal{J} = \{(\theta, u) : 0 < u < p(\theta)\}$.
- i.e., let $\tilde{\Theta} = \Theta \times [0, m]$, where $p(\theta) \leq m \,\forall\, \theta \in \Theta$.
- Now pick a random point $(\theta^*, U^*) \in \tilde{\Theta}$.
- if If $0 < U^* < p(\theta^*)$, then accept the draw.
- sampling over the u dimension is equivalent to marginalizing u out of $f(\theta, u)$, i.e.,

$$p(\theta) = \int_0^{p(\theta)} f(u) du = \int_0^{p(\theta)} du, \; 0 < u < p(\theta),$$

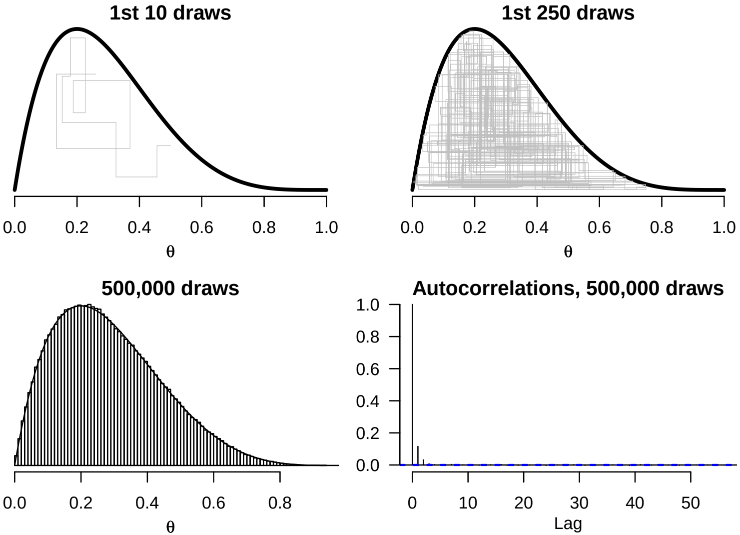
since $f(u)$ here is a constant.

Slice Sampling §5.2.7

- sequentially sample from $g(U|\theta)$ and $g(\theta|U)$
- Given $(\theta^{(t\text{-}1)}, U^{(t\text{-}1)})$:
  1: sample $U^{(t)} \sim \text{Unif}(0, p(\theta^{(t\text{-}1)}))$
  2: sample $\theta^{(t)} \sim \text{Unif}(\mathcal{A}^{(t)})$ where $\mathcal{A}^{(t)} = \{\theta : p(\theta) \geq U^{(t)}\}$.
- special case of the Gibbs sampler

Slice sampling from a Beta density, Example 5.12

$p(\theta) \equiv \text{Beta}(2, 5), \theta \in \Theta \equiv [0, 1]$



References

Geddes, Barbara. 1990. "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics." *Political Analysis* 2:131--150.

Geweke, John. 2005. *Contemporary Bayesian Econometrics and Statistics*. Hoboken, New Jersey: Wiley.

Gilks, W. R., N. G. Best and K.K.C. Tan. 1995. "Adaptive rejection Metropolis sampling withing Gibbs sampling." *Applied Statistics* 44:455--472.

Gilks, W. R. and P. Wild. 1992. "Adaptive rejection sampling for Gibbs sampling." *Applied Statistics* 41:337--348.

Pham-Gia, Thu and Noyan Turkkan. 1993. "Bayesian analysis of the difference of two proportions." *Communications in Statistics --- Theory and Methods* 22:1755--1771.

Rao, C. Radhakrishna. 1973. *Linear Statistical Inference and Its Applications*. Second ed. New York: Wiley.

Sekhon, Jasjeet S. 2005. "Making Inference from 2 × 2 Tables: The Inadequacy of the Fisher Exact Test for Observational Data and a Bayesian Alternative." Typescript. Survey Research Center, University of California, Berkeley.

Skocpol, Theda. 1979. *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge: Cambridge University Press.

van der Vaart, A. W. 1998. *Asymptotic Statistics*. Cambridge, United Kingdom: Cambridge University Press.

von Neumann, J. 1951. "Various techniques used in connection with random digits." *National Bureau of Standards Applied Mathematics Series* 12:36--38.

Wichura, Michael J. 1988. "Algorithm AS 241: The Percentage Points of the Normal Distribution." *Applied Statistics* 37:477--484. http://links.jstor.org/sici?sici=0035-9254