Chapter 1

Monte Carlo Methods

Goals:

1. basics of Monte Carlo methods
2. design of a Monte Carlo study

1.1 Basics of Monte Carlo methods

In a previous chapter, we developed the **crude Monte Carlo** estimator of the expectation

$$\mu \;=\; \mathrm{E}_\theta[g(X)]$$

given $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f(\underset{\sim}{x}|\theta)$ with the same distribution as X, the strong law of large numbers (SLLN) implies that

$$\hat{\mu} \;=\; \frac{1}{n}\sum_{i=1}^{n} g(X_i) \;\to\; \mu \quad \text{as} \quad n \to \infty.$$

The precision of our estimate $\hat{\mu}$ is dictated by

$$\mathrm{Var}(\hat{\mu}) \;=\; \frac{1}{n}\mathrm{Var}_\theta g(X) \;\equiv\; \frac{\sigma_g^2}{n},$$

which can be estimated via

$$\hat{\sigma}_g^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(g(X_i)-\hat{\mu})^2 \;\equiv\; \text{sample variance of } g(X_i)\text{s.}$$

Note that $\hat{\mu}$ is *unbiased*, and typically in large samples

$$\hat{\mu} \;\dot{\sim}\; \text{Normal}\left(\mu, \hat{\sigma}_g^2/n\right).$$

The precision of $\hat{\mu}$ depends on σ_g^2 and n. We will discuss several methods that aim to increase precision, besides increasing n. Note that more complex methods may increase precision for a given n, but may incur increased programming effort or computational time. Some assessment of the trade-offs between variance reduction and added labor or cost needs to be made.

1.1.1 Control variates

As before, suppose we wish to estimate (assuming x continuous)

$$\mu \;\equiv\; \mathrm{E}_\theta[g(X)] \;=\; \int g(x)f(x|\theta)\,\mathrm{d}x.$$

If we have a $g^*(x)$ that is "similar to" $g(x)$ and for which

$$\tau \;\equiv\; \mathrm{E}_\theta[g^*(X)] \;=\; \int g^*(x)f(x|\theta)\,\mathrm{d}x$$

is *known*, then writing

$$\begin{aligned}\mu \;&=\; \int\{g(x)-g^*(x)\}f(x|\theta)\,\mathrm{d}x+\tau \\ &=\; \mathrm{E}_\theta[g(x)-g^*(x)]+\tau\end{aligned}$$

we can use crude Monte Carlo to estimate $\mathrm{E}_\theta[g(x) - g^*(x)]$. That is,

$$\begin{aligned}\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n}\{g(x_i) - g^*(x_i)\} + \tau \\ &= \frac{1}{n}\sum_{i=1}^{n}\{g(x_i)\} - \frac{1}{n}\sum_{i=1}^{n}\{g^*(x_i)\} + \tau\end{aligned}$$

with

$$\begin{aligned}\mathrm{Var}[\hat{\mu}] &= \frac{1}{n}\mathrm{Var}_\theta[g(X) - g^*(X)] \\ &= \frac{1}{n}\{\mathrm{Var}_\theta[g(X)] + \mathrm{Var}_\theta[g^*(X)] - 2\mathrm{Cov}_\theta[g(X), g^*(X)]\}.\end{aligned}$$

If $g^*(X)$ mimics $g(X)$, then $\mathrm{Var}_\theta[g(X)] \doteq \mathrm{Var}_\theta[g^*(X)]$ and

$$\begin{aligned}\mathrm{Var}[\hat{\mu}] &\doteq \frac{1}{n}\{2\mathrm{Var}_\theta[g(X)] - 2\mathrm{Var}_\theta[g(X)]\mathrm{Corr}_\theta[g(X), g^*(X)]\} \\ &< \frac{1}{n}\mathrm{Var}_\theta[g(X)] \quad \text{if} \quad \mathrm{Corr}_\theta[g(X), g^*(X)] > \frac{1}{2}.\end{aligned}$$

Thus, reduction in variability relative to crude MC if $\mathrm{Corr}_\theta[g(X), g^*(X)] > \frac{1}{2}$.

Example, median Let $\underset{\sim}{X} = (X_1, \ldots, X_n)$ be a sample from some distribution with known $\mathrm{E}[x_i] = \tau$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $M =$ sample median of X_is, and suppose we wish to estimate

$$\mathrm{E}[M] \equiv \mu \qquad \text{(nonstandard notation)}$$

given R samples, each giving M_r and $\bar{x}_r$. Consider using

$$\hat{\mu} = \frac{1}{R}\sum_{r=1}^{R}\{M_r - \bar{x}_r\} + \tau$$

as the estimate. That is, use $\bar{x}$ as a control variate for estimating $\mathrm{E}[M]$.

```
# Gamma(2, 4) distribution, with E[X]=a*b and Var[X]=a*b^2.
a <- 2
b <- 4

# true median
qgamma(0.5, a, scale=b)

## [1] 6.713

# true mean of gamma distribution
tau <- a*b
tau

## [1] 8



# sample from gamma distribution
R <- 1e4 # samples
n <- 25  # sample size
x <- matrix(rgamma(R*n, a, scale=b), ncol=n) # draw R random samples in rows


# bootstrap estimate of variability of M
M <- apply(x, 1, median)
c(mean(M), var(M))

## [1] 6.806 1.638



# using mean as control variate
x.bar <- apply(x, 1, mean)
c(mean(x.bar), var(x.bar))

## [1] 8.005 1.289



# Check that the correlation between our variate of interest (median)
#   and our control variate (mean) is at least 1/2
cor(M, x.bar)

## [1] 0.7612
```

```
# This estimate of mu, the true median, has lower variance than x.bar
mu.hat <- mean(M - x.bar) + tau
c(mu.hat, var(M - x.bar))

## [1] 6.8005 0.7149
```

Example, cdf Let $\underset{\sim}{X} = \{X_1, X_2, \ldots, X_k\}$ and $T(\underset{\sim}{X}) \equiv$ some statistic. Suppose we wish to estimate

$$\mu \equiv \mu(t) = \Pr[T(X) \leq t] = \mathrm{E}[1_{(T(X)\leq t)}],$$

that is, estimate the cumulative distribution function (cdf) of $T(\underset{\sim}{X})$. In other words, the cdf of $T(\underset{\sim}{X})$ is the probability that statistic $T(\underset{\sim}{X})$ is less than t for each quantile t. The crude MC estimate is the *empirical cdf*: given $T(\underset{\sim}{X}_1), T(\underset{\sim}{X}_2), \ldots, T(\underset{\sim}{X}_n)$,

$$\begin{aligned}\hat{\mu} \equiv \hat{\mu}(t) &= \frac{1}{n}\sum_{i=1}^{n} 1_{(T(X_i)\leq t)} \\ &= \frac{\text{number of } (T(X_i) \leq t)}{n}.\end{aligned}$$

Suppose statistic $S(\underset{\sim}{X})$ mimics $T(\underset{\sim}{X})$ and the cdf of $S(\underset{\sim}{X})$

$$\tau(t) \equiv \Pr[S(X) \leq t]$$

is known. Then the control variate estimate is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\{1_{(T(X_i)\leq t)} - 1_{(S(X_i)\leq t)}\} + \tau(t)$$

$$\{1_{(T(X_i)\leq t)} - 1_{(S(X_i)\leq t)}\} = \begin{cases} 1 & T \leq t, S > t \\ 0 & T \leq t, S \leq t \text{ or } T < t, S < t \\ -1 & T > t, S \leq t \end{cases}.$$

The variance reduction could be substantial.

This idea was used in the "Princeton Robustness Study[1]", which among other things considered distributional properties of trimmed mean-like t-statistic

$$t_T \;=\; \frac{\bar{x}_T - \theta}{\mathrm{SE}[\bar{x}_T]} \qquad \text{(based on sample size, } k\text{)}.$$

If the underlying population distribution is Normal with mean θ, you can use

$$t \;=\; \frac{\bar{x} - \theta}{\mathrm{SE}[\bar{x}]} \;\sim\; t_{k-1}$$

as a control variable for estimating cdf of t_T.

Example, Multinomial Suppose

$$\underset{\sim}{X} = \{X_1, X_2, \ldots, X_k\} \;\sim\; \text{Multinomial}(m, \underset{\sim}{\theta}),$$

where $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$. Two standard statistics for testing $H_0 : \theta_1 = \theta_{01}, \ldots, \theta_k = \theta_{0k}$ are the Pearson statistic

$$P \;=\; \sum_{i=1}^{k} \frac{(x_i - m\theta_{0i})^2}{m\theta_{0i}}$$

and the likelihood ratio statistic

$$G^2 \;=\; 2\sum_{i=1}^{k} x_i \log_e \left(\frac{x_i}{m\theta_{0i}}\right).$$

[1] John W. Tukey (1973). *The Estimators of the Princeton Robustness Study.* Princeton University, Department of Statistics.

Note that $0 \log_e(0) \equiv 0$. In large samples, both P and $G^2 \dot{\sim} \chi^2_{k-1}$ when H_0 is true. One way to study the closeness of χ^2_{k-1} approximation is through the moments: how close do the moments of P and G^2 match those of the χ^2_{k-1} distribution? The moments of P are tractable, but the moments of G^2 are not. This suggests using P as a control variate for estimating moments of G^2. For example, suppose we wish to estimate

$$\mathrm{E}[G^2] \;=\; \mu.$$

We know

$$\mathrm{E}[P] \;=\; \mathrm{E}[\chi^2_{k-1}] \;=\; k-1.$$

Thus, given n multinomial samples, estimate μ via

$$\hat{\mu} \;=\; \frac{1}{R}\sum_{r=1}^{R}\{G_r^2 - P_r\} + (k-1),$$

where G_r^2 and P_r are the values of G^2 and P from the rth sample.

1.1.2 Antithetic variates (AV)

Suppose we have two estimators $\hat{\mu}_1$ and $\hat{\mu}_2$ of μ and each has variance σ^2/n when based on a sample of size n. If the correlation ρ between these estimators is negative[2], then the estimator

$$\hat{\mu}_{\mathrm{AV}} \;=\; \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$$

[2]Antithetic means "directly opposed or contrasted; mutually incompatible".

has

$$\begin{aligned}
\mathrm{Var}[\hat{\mu}_{\mathrm{AV}}] &= \frac{1}{4}\{\mathrm{Var}[\hat{\mu}_1] + \mathrm{Var}[\hat{\mu}_2] + 2\mathrm{Cov}[\hat{\mu}_1, \hat{\mu}_2]\} \\
&= \frac{1}{4}\{\mathrm{Var}[\hat{\mu}_1] + \mathrm{Var}[\hat{\mu}_2] + 2\rho\sqrt{\mathrm{Var}[\hat{\mu}_1]\mathrm{Var}[\hat{\mu}_2]}\} \\
&= \frac{1}{4n}\{\sigma^2 + \sigma^2 + 2\rho\sigma^2\} \\
&= \frac{1}{2n}\sigma^2(1+\rho) \\
&< \frac{\sigma^2}{2n}
\end{aligned}$$

where the last term is the variance of either $\hat{\mu}_1$ or $\hat{\mu}_2$ based on a sample of size $2n$. That is, averaging the two estimators based on the same sample of size n (necessary to make estimators correlated) is better than doubling the sample size using either estimator individually.

Put another way, two negatively correlated estimators can be combined to provide a more precise estimator than either estimate individually, even when the combined estimator is based on half the number of samples.

The AV method is often difficult to implement since you need to find negatively correlated estimators. This can often be done in situations with certain symmetry constraints.

Example, AV Suppose $X \sim \mathrm{Normal}(0,1)$ and we wish to estimate

$$\mu = \mathrm{E}[h(X)] \qquad \text{where} \qquad h(X) = \frac{X}{2^X - 1}.$$

Since $-X \sim \mathrm{Normal}(0,1)$, the distribution of $h(X)$ and $h(-X)$ are identical and thus $\mathrm{E}[h(-X)] = \mu$. Based on a sample of $n = 10000$, we find the AV sample is much more precise than that of either individual estimate based on $n = 20000$ samples.

```r
# define h(x)
f.h <- function(x) {
  h <- x / (2^x - 1)
  return(h)
}

# sample from normal distribution
R <- 1e4 # samples
x <- rnorm(R) # draw R random samples
x2 <- rnorm(R) # double the samples for later comparison

# calculate h(x) and h(-x)
h.x <- f.h(x)
h.negx <- f.h(-x)

# these are negatively correlated, so the AV approach is profitable
cor(h.x, h.negx)
```

```
## [1] -0.9527
```

```r
# estimate
combine.h.x <- (h.x + h.negx) / 2
mu.hat.AS <- mean(combine.h.x)
mu.hat.AS
```

```
## [1] 1.499
```

```r
# sd of AV estimate
sd(combine.h.x)
```

```
## [1] 0.07764
```

```r
# sd of individual estimate based on 2*R samples
h.x2 <- f.h(x2)
h.negx2 <- f.h(-x2)

sd(c(h.x, h.x2))
```

```
## [1] 0.5086
```

```r
sd(c(h.negx, h.negx2))
```

```
## [1] 0.507
```

The AV approach combines two estimates of the same parameter as best we can, that is, by averaging them. A real gain comes about if the estimates have negative correlation.

In general, if we have estimates $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_\ell$ of μ with covariance matrix

$$\mathbf{\Sigma} \;=\; [\mathrm{Cov}(\hat{\mu}_i, \hat{\mu}_j)],$$

then we can use generalized LS to get the optimal estimate, that is, set

$$\hat{\mu}^* \;=\; \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_\ell \end{bmatrix} \;=\; \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \mu + \underset{\sim}{\varepsilon} \;=\; \underset{\sim}{1}\mu + \underset{\sim}{\varepsilon}, \qquad \mathrm{Cov}[\varepsilon] = \mathbf{\Sigma},$$

then the best estimate is

$$\hat{\mu} \;=\; (\underset{\sim}{1}^\top \mathbf{\Sigma}^{-1} \underset{\sim}{1})^{-1} \underset{\sim}{1}^\top \mathbf{\Sigma}^{-1} \hat{\mu}^*.$$

Remarks

- Typically estimate $\mathbf{\Sigma}$ with $\hat{\mathbf{\Sigma}}$ and plug that into $\hat{\mu}$,
- with two estimates with equal variance, the estimate is always the average, and
- depending on $\mathbf{\Sigma}$, could potentially reduce n and get same precision as using individual estimator $\hat{\mu}_j$.

1.1.3 Importance sampling (IS)

As before, we wish to estimate

$$\mu \;\equiv\; \mathrm{E}_\theta[g(X)] \;=\; \int g(x) f(x|\theta)\, \mathrm{d}x.$$

Assume θ is fixed and let $f(x) \equiv f(x|\theta)$. the crude MC estimate $\hat{\mu}$ is unbiased with

$$\begin{aligned} \mathrm{Var}(\hat{\mu}) &= \frac{1}{n}\mathrm{Var}_\theta[g(X)] \\ &= \frac{1}{n}(\mathrm{E}_\theta[g^2(X)] - \mu^2) \\ &= \frac{1}{n}\left(\int g^2(x) f(x|\theta)\,\mathrm{d}x - \mu^2\right). \end{aligned}$$

Importance sampling seeks to reduce $\mathrm{Var}(\hat{\mu})$ as follows. Note that for any other density $h(x)$

$$\begin{aligned} \mu &= \int g(x) f(x)\,\mathrm{d}x \\ &= \int g(x)\frac{f(x)}{h(x)} h(x)\,\mathrm{d}x \\ &= \int g(x) w(x) h(x)\,\mathrm{d}x \\ &= \mathrm{E}_h[g(x)w(x)], \end{aligned}$$

which is the expectation with respect to $h(x)$. Thus, drawing a sample of size n, $X_1, X_2, \ldots, X_n$, *from* $h(\underset{\sim}{x}|\theta)$, we can use the MC estimate

$$\hat{\mu}_{\mathrm{IS}} = \frac{1}{n}\sum_{i=1}^{n} g(x_i) w(x_i)$$

as an unbiased estimator of μ with

$$\begin{aligned} \mathrm{Var}(\hat{\mu}_{\mathrm{IS}}) &= \frac{1}{n}\mathrm{Var}_h[g(X)w(X)] \\ &= \frac{1}{n}(\mathrm{E}_h[g^2(X)w^2(X)] - \mu^2). \end{aligned}$$

Note that the expected value of the weight function

$$\begin{aligned} \mathrm{E}_h[w(x)] &= \int \frac{f(x)}{h(x)} h(x)\,\mathrm{d}x \\ &= \int f(x)\,\mathrm{d}x \\ &= 1, \end{aligned}$$

that is, the average weight is 1.

Since the average weight is one, some weights may be very large ($\gg 1$). IS tends to work well when $w(x)$ is large only when $g(x)$ is *small*. This requires the choice of $h(x)$ to be made carefully!

Remarks

1. IS is a crude MC, so we can estimate $\mathrm{Var}(\hat{\mu}_{\mathrm{IS}})$ via

$$\mathrm{Var}(\hat{\mu}_{\mathrm{IS}}) = \frac{\sigma^2_{\mathrm{IS}}}{n}$$

where

$$\sigma^2_{\mathrm{IS}} = \frac{1}{n-1} \sum_{i=1}^{n} \{g(x_i)w(x_i) - \hat{\mu}_{\mathrm{IS}}\}^2,$$

which is the sample variance of the $g(x_i)w(x_i)$s.

2. Another IS estimate is obtained by writing

$$\begin{aligned}\mu &= \frac{\int g(x)f(x)\,\mathrm{d}x}{\int f(x)\,\mathrm{d}x} \\ &= \frac{\int g(x)\frac{f(x)}{h(x)}h(x)\,\mathrm{d}x}{\int \frac{f(x)}{h(x)}h(x)\,\mathrm{d}x} \\ &= \frac{\int g(x)w(x)h(x)\,\mathrm{d}x}{\int w(x)h(x)\,\mathrm{d}x} \\ &= \frac{\mathrm{E}_h[g(x)w(x)]}{\mathrm{E}_h[w(x)]}.\end{aligned} \tag{1.1}$$

This also makes sense because $\mathrm{E}_h[w(x)] = 1$.

Given $X_1, X_2, \ldots, X_n$ from $h(\underset{\sim}{x}|\theta)$, estimate μ via

$$\begin{aligned}\hat{\mu} &= \frac{\frac{1}{n}\sum_{i=1}^n g(x_i)w(x_i)}{\frac{1}{n}\sum_{i=1}^n w(x_i)} \\ &= \frac{1}{n}\sum_{i=1}^n g(x_i)w^*(x_i)\end{aligned}$$

where

$$w^*(x_i) = \frac{w(x_i)}{\frac{1}{n}\sum_{\ell=1}^n w(x_\ell)}$$

are the normalized weights.

This approach is important because we can think of $f(x)$ in (1.1) not as a density but as a kernel of a density. That is, the actual density is

$$cf(x_i) = \frac{f(x_i)}{\int f(x)\,\mathrm{d}x}.$$

That is, we don't need to know the normalization constant, which makes this a useful strategy in Bayesian calculations.

3. Sometimes IS is used because sampling from $h(x)$ is easier than sampling from $f(x)$.

Example of IS, Beta Suppose $X \sim \text{Beta}(\alpha, \beta)$ with density

$$f(x) \;=\; \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1$$

and we wish to compute the moment generating function (mgf) of X,

$$\text{M}_X(t) \;=\; \text{E}[e^{tx}] \;=\; \int_0^1 e^{tx} f(x)\,\text{d}x,$$

for which there is no closed-form solution.

Define $h(x) = 1$ for $0 < x < 1$, and 0 otherwise, that is $h(x)$ is Uniform$(0, 1)$. Then,

$$\begin{aligned} \text{M}_X(t) \;&=\; \int_0^1 e^{tx} \frac{f(x)}{1} h(x)\,\text{d}x \\ &=\; \int_0^1 e^{tx} w(x) h(x)\,\text{d}x \\ &=\; \text{E}_h[e^{tx} f(x)], \end{aligned}$$

where the expectation is taken with respect to $X \sim \text{Uniform}(0, 1)$.

If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(0, 1)$, the IS estimate is

$$\hat{\mu}_{\text{IS}} \;=\; \frac{1}{n} \sum_{i=1}^{n} e^{tx} f(x).$$

We can do crude MC by sampling $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Beta}(\alpha, \beta)$ and computing

$$\hat{\mu} \;=\; \frac{1}{n} \sum_{i=1}^{n} e^{tx}.$$

How well does this work?

1.2 Some basics on designing an MC study

Principals of experimental design apply to designing an MC study. For a given parameter μ (or set of parameters) that we wish to estimate, we need to assess

- the sample size needed to obtain a specified precision (1/variance), and
- whether the crude MC can be improved upon.

A sample size calculation requires some knowledge of uncertainty, possibly based on a "pilot study". To make things concrete, suppose we have a statistic

$$T(\underset{\sim}{X}) \;=\; T(X_1, X_2, \ldots, X_n)$$

for which we wish to estimate the CDF

$$\Pr[T(X) \leq t] \;=\; \mathrm{E}[1_{(T(X)\leq t)}].$$

More generally, we would consider estimating $\Pr[T(X) \in C]$ for some set C. If we, for the moment, assume that t is fixed, then all we are doing is estimating the probability

$$p \;=\; \Pr[T(X) \leq t].$$

For crude MC, we generate n copies $X_1, X_2, \ldots, X_n$ from the same distribution as X, and compute

$$\begin{aligned} \hat{p} &= \frac{1}{n}\sum_{i=1}^{n} 1_{(T(X_i)\leq t)} \\ &= \frac{\text{number of } \{T(X_i) \leq t\}}{n} \\ &= \{\text{sample proportion } \leq t\}. \end{aligned}$$

We know

$$\begin{aligned} \text{Var}[\hat{p}] &= \frac{1}{n}\text{Var}[1_{(T(X_i)\leq t)}] \\ &= \frac{1}{n}p(1-p) \end{aligned}$$

which can be estimated via

$$\text{Var}[\hat{p}] = \frac{1}{n}\hat{p}(1-\hat{p})$$

or (a close approximation)

$$\text{Var}[\hat{p}] \doteq \frac{1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n}\{1_{(T(X_i)\leq t)} - \hat{p}\}^2\right).$$

Thus, our general results can be applied to this setting.

Given this *method*, how do you choose n? One approach is based on the margin-of-error (MOE). We note that an approximate 95% CI for p based on $\hat{p}$ is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

which implies that p is within (approximately) $2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ of $\hat{p}$ in 95% of samples. That is, the error on $\hat{p}$ as an estimate of p is within

$$\text{MOE} \;=\; 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

in 95% of samples. Since $p(1-p) \leq 0.25$,

$$\text{MOE} \;\leq\; 2\sqrt{\frac{0.25}{n}} \;=\; \frac{1}{\sqrt{n}}.$$

If we pre-specify a desired MOE, then choosing

$$\frac{1}{\sqrt{n}} \;=\; \text{MOE} \qquad \text{implies} \qquad n \;=\; \frac{1}{\text{MOE}^2}$$

gives the desired result. For a MOE of 0.01, we need $n = \frac{1}{0.01^2} = 100^2 = 10000$. For a MOE of 0.05, we need $n = \frac{1}{0.05^2} = 20^2 = 400$. In general, decreasing the MOE by a factor of two requires quadrupling n.

Note that this is a worst case scenario. If you know $p \doteq 0.1$, then

$$\text{MOE} \;\doteq\; 2\sqrt{\frac{0.1 \times 0.9}{n}} \;=\; \frac{2(0.3)}{\sqrt{n}} \;=\; \frac{0.6}{\sqrt{n}}$$

or

$$n \;\doteq\; \frac{0.6^2}{\text{MOE}^2} \;=\; \frac{0.36}{\text{MOE}^2},$$

which reduces the necessary sample size by a factor of approximately 3 relative to using $p = 0.5$.

Remark If p is "really small", that is, a tail probability, you probably wish a MOE of no greater than 0.01 or less!

1.3 Using the same stream of random numbers

This can have an effect of "pairing".

Suppose again we have a univariate random variable, X, (though the following holds for multivariate, as well) and now we wish to estimate both

$$\begin{aligned} p_T &= \Pr[T(X) \leq t] \\ p_S &= \Pr[S(X) \leq t] \end{aligned}$$

for two different statistics $T(X)$ and $S(X)$ and a fixed t.

One approach would be to use *separate* random samples of size n and crude MC

$$\begin{aligned} \hat{p}_T &= \frac{1}{n}\sum_{i=1}^{n} 1_{(T(X_i)\leq t)} = \frac{\text{number of } \{T(X) \leq t\}}{n} \\ \hat{p}_S &= \frac{1}{n}\sum_{i=1}^{n} 1_{(S(X_i^*)\leq t)} = \frac{\text{number of } \{S(X_i^*) \leq t\}}{n}. \end{aligned}$$

This gives

$$\begin{aligned} \text{Var}[\hat{p}_T] &= \frac{p_T(1-p_T)}{n} \\ \text{Var}[\hat{p}_S] &= \frac{p_S(1-p_S)}{n} \end{aligned}$$

and, since samples are independent

$$\text{Var}[\hat{p}_T - \hat{p}_S] = \frac{1}{n}\{p_T(1-p_T) + p_S(1-p_S)\}.$$

This is a two independent proportions problem.

If the goal is to estimate p_T and p_S but also to estimate $p_T - p_S$ accurately, then we should identify a way to make $\hat{p}_T$ and $\hat{p}_S$ positively correlated (similar to the control variate idea) since

$$\begin{aligned}\mathrm{Var}[\hat{p}_T - \hat{p}_S] &= \mathrm{Var}[\hat{p}_T] + \mathrm{Var}[\hat{p}_S] - 2\mathrm{Cov}[\hat{p}_T, \hat{p}_S] \\ &= \frac{1}{n}\{p_T(1-p_T) + p_S(1-p_S)\} - 2\mathrm{Cov}[\hat{p}_T, \hat{p}_S].\end{aligned}$$

If $T(X)$ and $S(X)$ are similar, just using the same stream of random numbers is often sufficient (and more efficient!)

With the same sample $X_1, X_2, \ldots, X_n$, calculate

$$\begin{aligned}\hat{p}_T &= \frac{1}{n}\sum_{i=1}^{n} 1_{(T(X_i)\le t)} = \frac{\text{number of } \{T(X) \le t\}}{n}, \\ \hat{p}_S &= \frac{1}{n}\sum_{i=1}^{n} 1_{(S(X_i^*)\le t)} = \frac{\text{number of } \{S(X_i^*) \le t\}}{n},\end{aligned}$$

and

$$\hat{p}_T - \hat{p}_S = \frac{1}{n}\sum_{i=1}^{n}\{1_{(T(X_i)\le t)} - 1_{(S(X_i)\le t)}\} = \frac{1}{n}\sum_{i=1}^{n}\Delta_t\{T(X_i), S(X_i)\},$$

where

$$\begin{aligned}\Delta_t\{T(X_i), S(X_i)\} &= \{1_{(T(X_i)\le t)} - 1_{(S(X_i)\le t)}\} \\ &= \begin{cases} 1 & T \le t, S > t \\ 0 & T \le t, S \le t \text{ or } T < t, S < t \\ -1 & T > t, S \le t \end{cases}.\end{aligned}$$

Let the joint distribution of the indicators be given by the following 2-by-2 contingency table

	S		
T	$S \le t$	$S > t$	
$T \le t$	p_{TS}	$p_{T\bar{S}}$	p_T
$T > t$	$p_{\bar{T}S}$	$p_{\bar{T}\bar{S}}$	$p_{\bar{T}}$
	p_S	$p_{\bar{S}}$	1

where $p_{TS} = \Pr[T \le t, S \le t]$, $p_{T\bar{S}} = \Pr[T \le t, S > t]$, etc. Then

$$\begin{aligned} \mathrm{E}[\Delta_t\{T(X_i), S(X_i)\}] &= 1p_{T\bar{S}} - 1p_{\bar{T}S} \\ &= (p_{TS} + p_{T\bar{S}}) - (p_{\bar{T}S} + p_{TS}) \\ &= p_T - p_S, \end{aligned}$$

that is,

$$\mathrm{E}[\hat{p}_T - \hat{p}_S] = p_T - p_S$$

and

$$\begin{aligned} \mathrm{Var}[\Delta_t\{T(X_i), S(X_i)\}] &= \mathrm{E}[\Delta_t^2] - (\mathrm{E}[\Delta_t])^2 \\ &= 1p_{T\bar{S}} - 1p_{\bar{T}S} - (p_T - p_S)^2 \\ &= p_{T\bar{S}} - p_{\bar{T}S} - ((p_{TS} + p_{T\bar{S}}) - (p_{\bar{T}S} + p_{TS}))^2 \\ &\vdots \qquad \text{(a little work)} \\ &= p_{T\bar{S}}(1 - p_{T\bar{S}}) + p_{\bar{T}S}(1 - p_{\bar{T}S}) + 2p_{T\bar{S}}p_{\bar{T}S}. \end{aligned}$$

Thus,

$$\begin{aligned} \mathrm{Var}[\hat{p}_T - \hat{p}_S] &= \frac{1}{n}\mathrm{Var}[\Delta_t\{T(X_i), S(X_i)\}] \\ &= \frac{p_{T\bar{S}}(1 - p_{T\bar{S}}) + p_{\bar{T}S}(1 - p_{\bar{T}S}) + 2p_{T\bar{S}}p_{\bar{T}S}}{n}. \end{aligned}$$

Remarks

1. This is just a paired proportion problem, where, if we let n_{TS} = number of $(T(X_i) \le t, S(X_i) \le t)$, $n_{T\bar{S}}$ = number of $(T(X_i) \le t, S(X_i) > t)$, etc., then the 2-by-2 table of counts

	S		
T	$S \le t$	$S > t$	
$T \le t$	n_{TS}	$n_{T\bar{S}}$	n_T
$T > t$	$n_{\bar{T}S}$	$n_{\bar{T}\bar{S}}$	$n_{\bar{T}}$
	n_S	$n_{\bar{S}}$	n

leads to estimates of cell and marginal probabilities, for example

	S		
T	$S \le t$	$S > t$	
$T \le t$	$p_{TS} = n_{TS}/n$	$p_{T\bar{S}} = n_{T\bar{S}}/n$	$p_T = n_T/n$
$T > t$	$p_{\bar{T}S} = n_{\bar{T}S}/n$	$p_{\bar{T}\bar{S}} = n_{\bar{T}\bar{S}}/n$	$p_{\bar{T}} = n_{\bar{T}}/n$
	$p_S = n_S/n$	$p_{\bar{S}} = n_{\bar{S}}/n$	$1 = n/n$

Then

$$\begin{aligned} \hat{p}_T - \hat{p}_S &= \frac{(n_{TS} + n_{T\bar{S}}) - (n_{TS} + n_{\bar{T}S})}{n} \\ &= \hat{p}_{T\bar{S}} - \hat{p}_{\bar{T}S}. \end{aligned}$$

That is, the estimate of $p_T - p_S$ is based only on cases that *disagree*.

This is unbiased for $p_T - p_S$ with

$$\begin{aligned} \mathrm{Var}[\hat{p}_T - \hat{p}_S] &= \mathrm{Var}[\hat{p}_{T\bar{S}} - \hat{p}_{\bar{T}S}] \\ &= \mathrm{Var}[\hat{p}_{T\bar{S}}] + \mathrm{Var}[\hat{p}_{\bar{T}S}] - \mathrm{Cov}[\hat{p}_{T\bar{S}}, \hat{p}_{\bar{T}S}] \\ &= \frac{p_{T\bar{S}}(1 - p_{T\bar{S}}) + p_{\bar{T}S}(1 - p_{\bar{T}S}) + 2p_{T\bar{S}}p_{\bar{T}S}}{n}. \end{aligned}$$

If $T(X)$ and $S(X)$ mimic each other, expect the number or proportion of disagreements to be low, or $p_{T\bar{S}} \doteq 0$ and $p_{\bar{T}S} \doteq 0$ leading to very small $\mathrm{Var}[\hat{p}_T - \hat{p}_S]$ based on using the same sample of X_is.

2. From earlier results,

$$\begin{aligned}\mathrm{Var}[\hat{p}_T - \hat{p}_S] &= \frac{1}{n}\mathrm{Var}[\Delta_t\{T(X_i), S(X_i)\}] \\ &= \frac{p_{T\bar{S}}(1-p_{T\bar{S}}) + p_{\bar{T}S}(1-p_{\bar{T}S}) + 2p_{T\bar{S}}p_{\bar{T}S}}{n}.\end{aligned}$$

We can estimate this in two ways:

1. plug-in estimates of $p_{T\bar{S}}$ and $p_{\bar{T}S}$ from the contingency table, or
2. Compute the sample variance of $\Delta_t\{T(X_i), S(X_i)\}$ which is easy to do if you have one column with entries $1_{(T(X_i)\leq t)}$ and another with $1_{(S(X_i)\leq t)}$. Then you simply take the difference in the columns and calculate the sample variance of the differences.

1.4 Other methods

A variety of other MC techniques exist, such as

- conditioning swindles,
- Rao-Blackwellization, and
- stratified sampling.