STRATIFIED SAMPLING

Stratified Sampling Background: suppose $\theta = E[X]$. Simple MC simulation would use $\theta \approx \bar{X}$. If X simulation depends on some discrete Y with pmf $P\{Y = y_i\} = p_i$, $i = 1, \ldots, k$, and X can be simulated given $Y = y_i$,

$$E[X] = \sum_{i=1}^{k} p_i E[X|Y = y_i] \approx \sum_{i=1}^{k} p_i \bar{X}_i = \Theta.$$

Θ is the **stratified sampling** estimator of θ.

Note: y_i's subdivide sampling space into k "strata";

Analysis: If np_i samples are used to determine $\bar{X}_i$, then

$$Var(\bar{X}_i) = \frac{Var(X|Y = y_i)}{np_i}, \text{ and}$$

$$Var(\Theta) = \sum_{i=1}^{k} p_i^2 Var(\bar{X}_i) = \frac{1}{n} \sum_{i=1}^{k} p_i Var(X|Y = y_i) = \frac{1}{n} E[Var(X|Y)].$$

Now $Var(X) = E[Var(X|Y)] + Var(E[X|Y])$, and $Var(\bar{X}) = \frac{1}{n} Var(X)$, so

$$Var(\Theta) = \frac{1}{n}(Var(X) - Var(E[X|Y])) = Var(\bar{X}) - \frac{1}{n} Var(E[X|Y]).$$

Stratified sampling simulation can reduce variance significantly.

Note: if $S_i^2/(np_i)$ is the sample variance for $\bar{X}_i$ with np_i samples,

$$Var(\Theta) = \sum_{i=1}^{k} p_i^2 Var(\bar{X}_i) \approx \frac{1}{n} \sum_{i=1}^{k} p_i S_i^2.$$

STRATIFIED SAMPLING CONT.

Application: one-dimensional integration, where

$$\theta = E[h(U)] = \int_0^1 h(x)dx.$$

Strata are subintervals $[\frac{i-1}{k}, \frac{i}{k}]$, with $p_i = \frac{1}{k}$.

For subinterval i, generate $U_{ij} = \frac{i-1}{k} + \frac{1}{k}U$, $j = 1, \ldots, n_i$, $i = 1, \ldots, k$.

Example: $\frac{\pi}{4} = E[\sqrt{1-U^2}] = \int_0^1 \sqrt{1-x^2}dx \approx 0.7853982$.
Try $n = 5000$: simple MC ("raw simulation").
Matlab

```
N = 5000; U = rand(1,N); X = sqrt(1-U.^2);
disp( [mean(X) 2*std(X)/sqrt(N)])
      0.78064     0.0063975
```

STRATIFIED SAMPLING CONT.

Example continued: $\theta = E[h(U)] = \int_0^1 h(x)dx$.
With stratified sampling there are different choices for k.
a) Try $k = n$, $n_i = 1$, so use $U_{ij} = U_i = \frac{i-1}{n} + \frac{1}{n}U$, $i = 1, \ldots, n$.
Matlab

```
US = [0:N-1]/N + U/N; XS = sqrt(1-US.^2);
disp( [mean(XS) 2*std(XS)/sqrt(N)] )
       0.7854     0.0063134
```

b) Try $k = 500$, $n_i = 10$, $p_i = \frac{1}{500}$, so use
$U_{ij} = \frac{i-1}{500} + \frac{1}{500}U$, $j = 1, \ldots, 10$, $i = 1, \ldots, 500$.
then $Var(\bar{\Theta}) \approx \sum_{i=1}^{500} \frac{1}{10(500)^2} S_i^2 = \frac{1}{500} \sum_{i=1}^{500} \frac{S_i^2}{5000}$.
Matlab

```
K = 500; Ni = N/K;
for i = 1 : K
 XS = sqrt(1-((i-1+rand(1,Ni))/K).^2);
 XSB(i) = mean(XS); SS(i) = var(XS);
end, SST = mean(SS/N);
disp( [mean(XSB) 2*sqrt(SST)] )
    0.78541    3.227e-05
```

STRATIFIED SAMPLING CONT.

Optimal Distribution of Samples : to reduce Var more, assume n_i samples for stratum i, with $\sum_{i=1}^{k} n_i = n$.

- If $\bar{X}_i \approx E[X|Y = y_i]$ using n_i samples, let

$$\bar{\Theta} = \sum_{i=1}^{k} p_i \bar{X}_i, \text{ with } Var(\bar{\Theta}) = \sum_{i=1}^{k} p_i^2 Var(\bar{X}_i).$$

- Given estimate s_i^2 for each $Var(\bar{X}_i) \approx \frac{s_i^2}{n_i}$, choose n_i's to minimize $\sum_{i=1}^{k} p_i^2 s_i^2 / n_i$, subject to $\sum_{i=1}^{k} n_i = n$;

  Lagrange multiplier solution is $n_i = n \frac{p_i s_i}{\sum_{j=1}^{k} p_j s_j}$.

- Algorithm: use small n to estimate s_i's, larger n for $\bar{\Theta}$;

$$Var(\bar{\Theta}) \approx \sum_{i=1}^{k} \frac{p_i^2}{n_i} S_i^2,$$

  is estimated using sample variances S_i^2 from strata.

STRATIFIED SAMPLING CONT.

Continued Example: $\frac{\pi}{4} = E[\sqrt{1-U^2}] = \int_0^1 \sqrt{1-x^2}dx$

c) $k = 500$, $p_i = \frac{1}{500}$, optimal $n_i = 5000 s_i / (\sum_{i=1}^{500} s_i)$.

$U_{ij} = \frac{i-1}{500} + \frac{1}{500} U$,$j = 1, \ldots, n_i$, $i = 1, \ldots, 500$,

$Var(\bar{\Theta}) \approx \frac{1}{500} \sum_{i=1}^{500} \frac{S_i^2}{500 n_i}$.

Matlab

```
ON = max( 3, round(N*sqrt(SS)/sum(sqrt(SS))) );
for i = 1 : K, XS = sqrt(1-((i-1+rand(1,ON(i)))/K).^2);
  XSB(i) = mean(XS); SS(i) = var(XS);
end,  SST = mean(SS./(500*ON));
disp( [mean(XSB) 2*sqrt(SST)] )
    0.78538   1.6446e-05
```

STRATIFIED SAMPLING CONT.

More Stratified Sampling Examples

- Estimate $P = \int_{-\infty}^{\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}} \cos(t^2) dt$.
  For simulation convert to $x \in [0,1]$ using $x = \Phi(t)$:
  so $dx = \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt$, $P = \int_0^1 \cos(\Phi^{-1}(x)^2) dx$.

  a) Simple MC Matlab

```
f = @(x)cos(norminv(x).^2); N = 5000; U = rand(1,N); X = f(U);
disp( [mean(X) 2*std(X)/sqrt(N)])   % Simple MC
      0.57205     0.016998
```

  b) $k = 500$, $n_i = 10$, $p_i = \frac{1}{500}$, so use $U_{ij} = \frac{i-1}{500} + \frac{1}{500} U$, $j = 1, \ldots, 10$, $i = 1, \ldots, 500$.
  with $Var(\bar{\Theta}) \approx \sum_{i=1}^{500} \frac{1}{10(500)^2} S_i^2 = \frac{1}{500} \sum_{i=1}^{500} \frac{S_i^2}{5000}$. Matlab

```
K = 500; Ni = N/K;
for i = 1 : K, XS = f((i-1+rand(1,Ni))/K);
 XSB(i) = mean(XS); SS(i) = var(XS);
end, SST = mean(SS/N);
disp( [mean(XSB) 2*sqrt(SST)] )    % Stratified sampling
     0.56901    0.0014701
```

STRATIFIED SAMPLING CONT.

Continued Example: $P = \int_0^1 \cos(\Phi^{-1}(x)^2)dx$.

c) $k = 500$, $p_i = \frac{1}{500}$, optimal $n_i = 5000 s_i / (\sum_{i=1}^{500} s_i)$.
$U_{ij} = \frac{i-1}{500} + \frac{1}{500} U$,$j = 1, \ldots, n_i$, $i = 1, \ldots, 500$,
with $Var(\bar{\Theta}) \approx \frac{1}{500} \sum_{i=1}^{500} \frac{S_i^2}{500 n_i}$. Matlab

```
ON = max( 3, round(N*sqrt(SS)/sum(sqrt(SS))) )
766 283 92 30 57 83 86 76 61 50 36 35 24
  ...
 3   3  3  3  3  3  3  3  3   3  3  3  3
  ...
 31 43 44 61 77 79 86 98 67 43 99 239 998

for i = 1 : K
 XS = f((i-1+rand(1,ON(i)))/K);
 XSB(i) = mean(XS); SS(i) = var(XS);
end,  SST = mean(SS./(500*ON));
disp( [mean(XSB) 2*sqrt(SST)] ) % Optimal stratified sampling
      0.56901    0.0002475
```

STRATIFIED SAMPLING CONT.

- Video Poker Example: 5 cards dealt, 1-5 discards are replaced.
  Let $C = \binom{52}{5}^{-1}$, and assume payoff table is

Hand	Payoff	Probability
RF: Royal Flush	800	4C
SF: Straight Flush	50	36C
4K: Four-of-a-kind	25	624C
FH: Full House	8	3744C
FL: Flush	5	5108C
ST: Straight	4	10200C
3K: Three-of-a-kind	3	54912C
2P: Two Pair	2	123552C
HP: High Pair(> 10)	1	337920C
LP: Low Pair($<= 10$)	0	760320C
OT: Other	0	1-$\sum p_i \approx .5010527$.

Textbook Strategy: no replacement cards dealt if initial deal $> 3K$;
otherwise keep pairs and triplets but replace other cards.

Raw simulation: estimate $\theta = E[X]$ using many runs, where each run
a) uses random permutation of [1:52] (p.51) to shuffle cards;
b) picks top 5 cards, discarding and replacing using text strategy;
c) and computes X from table.

STRATIFIED SAMPLING CONT.

Video Poker Example with stratification: if payoff is RV X,

$$\begin{aligned}E[X] = {} & E[X|RF]P\{RF\}+E[X|SF]P\{SF\}+E[X|4K]P\{4K\}+E[X|FH]P\{FH\} \\ & +E[X|FL]P\{FL\}+E[X|ST]P\{ST\}+E[X|3K]P\{3K\}+E[X|2P]P\{2P\} \\ & +E[X|HP]P\{HP\}+E[X|LP]P\{LP\}+E[X|OT]P\{OT\} \\ = {} & 0.044976+4.241443(.021128)+E[X|2P]P\{2P\} \\ & +E[X|HP]P\{HP\}+E[X|LP]P\{LP\}+E[X|OT]P\{OT\}.\end{aligned}$$

$E[X|2P], E[X|HP], E[X|LP]$, can also be computed analytically.

$E[X|2P] \approx 2.5106,\ E[X|HP] \approx 1.5264,\ E[X|LP] \approx .81351.$

Stratified sampling: for each run (using .044976 $\neq$ text value .051290).

a) use random permutation of [1:52] (p.51) to shuffle cards;

b) pick top 5 cards, if hand $\geq LP$, start new run;

c) pick next 5 cards; compute X_{OT} and X using

$$\begin{aligned}\hat{\theta} = {} & 0.044976+4.241443097(.021128451)+E[X|2P]P\{2P\} \\ & +E[X|HP]P\{HP\}+E[X|LP]P\{LP\}+X_{OT}P\{OT\}\end{aligned}$$

Compute $E[\hat{\theta}]$ using many runs.

Notice: using $P\{OT\} \approx .5$, $Var(\hat{\theta}) \approx (.5)^2 Var(X_{OT})$.

Some results:	$E[X]$	Variance	Error	N
Simple MC sampling:	0.8509	2.79	0.0334	10000
Stratified sampling:	0.85774	0.15	0.0246	1000

STRATIFIED SAMPLING CONT.

Multdimensional Integrals: consider the 2-dimensional

$$\theta = \int_0^1 \int_0^1 g(x,y)dxdy$$

- Simple stratified sampling.
  If subdivision of $[0,1] \times [0,1]$ is same for x and y, strata are squares $s_{ij} = [\frac{i-1}{k}, \frac{i}{k}] \times [\frac{j-1}{k}, \frac{j}{k}]$, for $i = 1:k$, $j = 1:k$, so

$$\theta = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} E[X|(x,y) \in s_{ij}], \quad p_{ij} = \frac{1}{k^2},$$

$$\bar{\Theta} = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} \bar{X}_{ij}, \text{ with } \bar{X}_{ij} \approx E[X|(x,y) \in s_{ij}],$$

$$\bar{X}_{ij} = \frac{1}{n_{ij}} \sum_{l=1}^{n_{ij}} g\Big(\frac{i-1}{k} + \frac{U_l}{k}, \frac{j-1}{k} + \frac{V_l}{k}\Big), \quad Var(\bar{\Theta}) \approx \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{p_{ij}^2}{n_{ij}} S_{ij}^2.$$

  Note: if $n_{ij} = n/k^2$, $Var(\bar{\Theta}) \approx \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{S_{ij}^2}{n}$.

STRATIFIED SAMPLING CONT.

Two-dimensional example: $\theta = \int_0^1 \int_0^1 e^{(x+y)^2} dxdy$. Matlab

```
g = @(x)exp(sum(x).^2); N = 10000; X = g(rand(2,N));
disp([mean(X) 2*std(X)/sqrt(N)]) % Simple MC
       4.8975      0.11808
K = 20; Nij = N/K^2;
for i = 1 : K                    % Stratified
  for j = 1 : K, XS = g([i-1+rand(1,Nij);j-1+rand(1,Nij)]/K);
    XSb(i,j) = mean(XS); SS(i,j) = var(XS);
  end
end, SST = mean(mean(SS/N));
disp([mean(mean(XSb)) 2*sqrt(SST) ])
       4.8967     0.010679
```

Note: for $n-$dimensional integrals the growth in number of terms k^n in the $\bar{\Theta}$ sum, $\bar{\Theta} = \sum_{i_1=1}^{k} \cdots \sum_{i_n=1}^{k} \frac{\bar{X}_{i_1,\ldots,i_n}}{k^n}$, results in an infeasible method for large n; however, the terms in the sum could be sampled with $\bar{\Theta} \approx \sum_{l=1}^{m} \frac{\bar{X}_{\mathbf{I}}}{m}$, where $\mathbf{I} = (I_1, I_2, \ldots, I_n)$, with $I_j \sim Uniform(1:k)$.

Other strategies, e.g.

with nonuniform n_{ij}'s, different k's for each variable, have been studied.

STRATIFIED SAMPLING CONT.

- Transformation Method for Monotone Functions

$$\theta = \int_0^1 \cdots \int_0^1 g(x_1, x_2, \ldots, x_n) dx_1 \cdots dx_n.$$

If $x_1 = e^{-y_n y_1}, x_2 = e^{-y_n(y_2 - y_1)}, \ldots \; x_n = e^{-y_n(1-y_{n-1})}$,

$$\theta = \int_0^\infty y_n^{n-1} e^{-y_n} \int_0^1 \int_0^{y_{n-1}} \int_0^{y_{n-2}} \cdots \int_0^{y_2} g(x_1(\mathbf{y}), \ldots, x_n(\mathbf{y})) dy_1 \cdots \; dy_n.$$

Notes: $0 < y_1 < \cdots < y_{n-1} < 1$; and $y_n^{n-1} e^{-y_n}/n!$ is a $Gamma(n, 1)$ pdf.

STRATIFIED SAMPLING CONT.

$$\theta = \int_0^1 \cdots \int_0^1 g(\mathbf{x})d\mathbf{x} = \int_0^\infty y_n^{n-1} e^{-y_n} \int_0^1 \int_0^{y_{n-1}} \int_0^{y_{n-2}} \cdots \int_0^{y_2} g(\mathbf{x}(\mathbf{y}))d\mathbf{y},$$

with $x_1 = e^{-y_n y_1}, x_2 = e^{-y_n(y_2-y_1)}, \ldots\ x_n = e^{-y_n(1-y_{n-1})}$.

For simulation, first generate $Y_n \sim Gamma(n,1)$, then generate

$U_1, \ldots, U_{n-1} \sim Uni(0,1)$, and set $Y_i = U_{(i)}$ (sorted), for $i = 1, \ldots, n-1$.

If $Y_n \sim Gamma^{-1}(n,1,U_n)$ (inversion), stratification can be used with U_n.

Example 2-d continued; let $x_1 = e^{-y_2 y_1}, x_2 = e^{-y_2(1-y_1)}$, then

$$\theta = \int_0^1 \int_0^1 e^{(x_1+x_2)^2} dx_1 dx_2 = \int_0^\infty y_2 e^{-y_2} \int_0^1 e^{(x_1(y_1,y_2)+x_2(y_1,y_2))^2} dy_1\ \ dy_2,$$

Matlab test

```
N = 10000; U=rand(2,N); Y(1,:)=U(1,:); Y(2,:)=gaminv(U(2,:),2,1);
X(1,:) = exp(-Y(1,:).*Y(2,:)); X(2,:) = exp(-Y(2,:).*(1-Y(1,:)));
Z = g(X); disp([mean(Z) 2*std(Z)/sqrt(N)])
    4.9077 0.11989    % Raw MC
K = 1000; Ni = N/K;
for i = 1:K, U = rand(2,Ni), Y(2,:) = gaminv((i-1+U(2,:))/K,2,1);
  X(1,:) = exp(-U(1,:).*Y(2,:)); X(2,:) = exp(-Y(2,:).*(1-U(1,:)))
  XS = g(X); XSB(i) = mean(XS); SS(i) = var(XS);
end, SST = mean(SS/N); disp([mean(XSB) 2*sqrt(SST)])
    4.9014   0.010342 % Stratified Sampling
```

STRATIFIED SAMPLING CONT.

Compound RVs: assume iid RVs X_i and RV N:

let $S_n(\mathbf{X}) = \sum_{i=1}^{n} \alpha_i X_i$; estimate $P\{S_N(\mathbf{X}) > c\}$.

E.g. X_i is insurance claim i, N is # claims by time T.

In many cases, distribution for N is known, e.g. Poisson(λ), with $p_i = e^{-\lambda}\frac{(\lambda)^i}{i!}$.

"Raw simulation": for K runs, generate N and N X_is;

count proportion with $\sum_{i=1}^{N} X_i > c$.

Stratified simulation: given pmf for N, let $g_n(\mathbf{x}) = I(S_n(\mathbf{x}) > c)$.

$$\theta = \sum_{n=0}^{m} E[g_n(\mathbf{X})]p_n + E[g_N(\mathbf{X})|N > m](1 - \sum_{n=0}^{m} p_n)$$

. Simulation run: given an m

a) compute $m' = N|N > m$ from conditional pmf;

b) generate $X_1, \ldots, X_{m'}$ and set

$$\Theta = \sum_{n=0}^{m} g_n(\mathbf{X})p_n + g_{m'}(\mathbf{X})(1 - \sum_{n=0}^{m} p_n).$$

Then $\bar{\Theta} \approx P\{S_N(\mathbf{X}) > c\}$.

E.g. Poisson(λ), has conditional pmf with $p_{m+i} = e^{-\lambda}\frac{(\lambda)^{m+i}}{\alpha(m+i)!}$,

and $\alpha = 1 - \sum_{n=0}^{m} p_n$.