# Variance Reduction: Techniques

SNU/NUKE/EHK

# Estimation Mean and Variance

- To estimate the integral

$$I = \int_0^5 \frac{1}{1+x^2}\,dx \,, \qquad (1)$$

- Define the function

$$f(x) = \begin{cases} 1/5, & 0 \le x \le 5 \\ 0, & \text{otherwise} \end{cases} \qquad (2) \qquad \text{so that} \quad \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

- Now define the random variable $V(x)$ as

$$V(x) = \frac{5}{1+x^2} \quad \text{so that} \quad \int_0^5 V(x)\cdot f(x)\,dx = \int_0^5 \frac{1}{1+x^2}\,dx \,.$$

# Estimation Mean and Variance (cont.)

- The expected value of V is then

$$E(V) = \int_{-\infty}^{\infty} V(x) f(x)\,dx = \int_0^5 \frac{1}{1+x^2}\,dx, \qquad (3)$$

which gives $E(V) = \tan^{-1}(5) - \tan^{-1}(0) \doteqdot 1.373400.$

- The mean value of $V^2$ is, on the other hand,

$$E(V^2) = \int_{-\infty}^{\infty} V^2(x) f(x)\,dx = \int_0^5 \left(\frac{5}{1+x^2}\right)^2 \frac{dx}{5} \approx 3.914270.$$

# Estimation Mean and Variance (cont.)

- The variance of the random variable V is then

$$var(V) = E(V^2) - [E(V)]^2 \doteqdot 2.028042, \text{ and}$$

the standard deviation is, therefore,

$$\sigma(V) \approx \sqrt{2.028042} \approx 1.424093.$$

Note $var(\bar{V}) = \frac{1}{n} var(V) = \frac{1}{n}\sigma^2(V)$

| ex. $var(\bar{V}) = 2.028042 \times 10^{-4}$ with $n = 10000$ |
|---|

# Improvement of Certainty in Estimate

- The purpose of MC calculation is usually to obtain an estimate of the expected value of a random variable.
- The usual measure of the certainty of the result is the standard deviation of the estimate obtained.
- The improvement in the certainty of an estimate of the mean can be made by increasing the number of samples:

$$var(\bar{S}) = \frac{1}{n} var(X) = \frac{\sigma^2(X)}{n}. \qquad (18)$$

which might cost too much time.

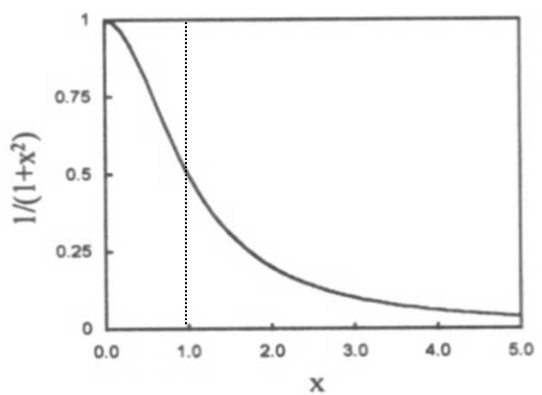
- Other methods for improvement in certainty?

# Stratified Sampling

- The range over which the independent variable is sampled is divided into strata and each stratum is sampled separately.

- Exercise
  - Estimate the integral of

$$I = \int_{x=0}^{5} \frac{1}{1+x^2} dx \qquad (1)$$



# Stratified Sampling: exercise

- Step 1. Define the pdf $f_1(x)$ for the first stratum

$$f_1(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{elsewhere} \end{cases} \quad (4)$$

and the pdf $f_2(x)$ for the second stratum

$$f_2(x) = \begin{cases} 0.25, & 1 \le x \le 5 \\ 0\ , & \text{elsewhere} \end{cases} \quad (5)$$

# Stratified Sampling: exercise (cont.)

- Step 2. Define the corresponding random variables as

$$V_1(x) = \frac{1}{1+x^2} \quad (6) \qquad \text{and} \quad V_2(x) = \frac{4}{1+x^2} \qquad (7)$$

so that

$$\int_0^1 V_1(x) f_1(x) dx = \int_0^1 \frac{1}{1+x^2} dx \, ; \text{ and } \int_1^5 V_2(x) f_2(x) dx = \int_1^5 \frac{1}{1+x^2} dx \, , \qquad (8)$$

which leads to

$$\int_0^1 V_1(x) f_1(x) dx + \int_1^5 V_2(x) f_2(x) dx = \int_0^1 \frac{1}{1+x^2} dx + \int_1^5 \frac{1}{1+x^2} dx = \int_0^5 \frac{1}{1+x^2} dx$$

$$\text{or } I = E(V_1) + E(V_2). \qquad (9)$$

# Stratified Sampling: exercise (cont.)

- Step 3. Calculate the expected value $E(V_1)$ in the first stratum and $E(V_2)$ in the second stratum:

$$E(V_1) = \int_{x=0}^{1} V_1(x) f_1(x) dx = \int_{x=0}^{1} \frac{1}{1+x^2} dx \approx 0.7853982 \qquad (10)'$$

$$E(V_2) = \int_{x=1}^{5} V_2(x) f_2(x) dx = \int_{x=1}^{5} \frac{4}{1+x^2} \frac{dx}{4} \approx 0.5880026 \qquad (10)''$$

# Stratified Sampling: exercise (cont.)

- Step 4. Calculate the variances $V_1$ for the first stratum and $V_2$ for the second stratum:

$$E(V_1^2) = \int_{x=0}^{1} \left( \frac{1}{1+x^2} \right)^2 dx \approx 0.6426991 \qquad (11)'$$

$$E(V_2^2) = \int_{x=1}^{5} \left( \frac{4}{1+x^2} \right)^2 \frac{dx}{4} \approx 0.5606200 \qquad (11)''$$

Hence,

$$var(V_1) = E(V_1^2) - [E(V_1)]^2 \approx 0.0258488 \qquad (12)'$$

$$var(V_2) = E(V_2^2) - [E(V_2)]^2 \approx 0.2148729 \qquad (12)''$$

# Stratified Sampling: exercise (cont.)

- Step 5. The value of the original integral is

$$I = \int_0^5 \frac{1}{1+x^2} dx = E(V_1) + E(V_2) \approx 1.373400.$$

- Step 6. Compare the mean and the variance

$$E(V_1) + E(V_2) \fallingdotseq 1.373400 \quad \text{as (3)}$$

$$\begin{aligned} var[E(V_1)+E(V_2)] \\ &= var[E(V_1)] + var[E(V_2)] \\ &= var(V_1)/n_1 + var(V_2)/n_2 \text{ and} \end{aligned}$$

the standard deviation is $\sigma(\bar{V}_1 + \bar{V}_2) = \sqrt{var(\bar{V}_1 + \bar{V}_2)}$ .

- ex. $var[E(V_1)+E(V_2)] \fallingdotseq 4.8144 \times 10^{-5}$ with $n_1$=5000 and $n_2$=5000

ex. $var(\bar{V}) = 2.028042 \times 10^{-4}$ with $n = 10000$

# Demonstration of the Stratification Property: nomenclature

- Conditional probability $P\{X=x \mid B\}$
  $\equiv$ the probability of $X=x$ given condition $B$.
- $E(V \mid A) \equiv$ the expected value of random variable $V(x)$ over the set $A$ of $x$'s.
- "~" indicates the subtraction of one set from another.

# Demonstration of the Stratification Property

- <u>Theorem</u>

  Given a variable function $V(x)$ for $x$ in a set $R$ of real numbers such that $E(V \mid A)$ in a subset $A$ and $E(V \mid R)$ and $var(V \mid R)$ in the set $R$ exist and are finite, if $E(V \mid A)$ does not equal $E(V \mid R)$ then the variance of the estimate of $E(V \mid R)$ using the strata defined by $A$ and $R{\sim}A$ is less than the variance of the estimate without stratification.

# Demonstration of the Stratification Property : Proof

- Let $f(x)$ be the pdf of $x$ for $V(x)$ and let

$$a = \int_A f(x)dx. \quad \text{Then} \quad 1 - a = \int_{R\sim A} f(x)dx.$$

- Define

$$f_1(x) = \begin{cases} \dfrac{f(x)}{a}, & \text{for } x \text{ in } A \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad f_2(x) = \begin{cases} \dfrac{f(x)}{1-a}, & \text{for } x \text{ not in } A \\ 0, & \text{otherwise} \end{cases}$$

# Demonstration of the Stratification Property : Proof (cont.)

- Then

  $E(V \mid A) = \bar{V}_1 = \int_A V(x) f_1(x) dx = \frac{1}{a} \int_A V(x) f(x) dx$ and

  $E(V \mid R\sim A) = \bar{V}_2 = \int_{R\sim A} V(x) f_2(x) dx = \frac{1}{1-a} \int_{R\sim A} V(x) f(x) dx$

  whereas $E(V \mid R) = \bar{V} = \int_R V(x) f(x) dx$.

  And thus,

  $E(V \mid R) = a \cdot E(V \mid A) + (1-a) \cdot E(V \mid R\sim A)$ or $\bar{V} = a\bar{V}_1 + (1-a)\bar{V}_2$.

- Since $E(V \mid A) \neq E(V \mid R)$, defining $\delta = E(V \mid A) - E(V \mid R)$ gives

  $E(V \mid R\sim A) = E(V \mid R) - \frac{a}{1-a}\delta$.

  Therefore, if $\delta > 0$, $E(V \mid A) > E(V \mid R)$ then $E(V \mid R\sim A) < E(V \mid R)$ and vice versa.

# Demonstration of the Stratification Property : Proof (cont.)

- The variance of V is given by

$$var(V) = \int_R [V(x) - E(V)]^2 f(x)dx$$

$$= \int_A [V(x) - E(V)]^2 f(x)dx + \int_{R-A} [V(x) - E(V)]^2 f(x)dx \, .$$

- The variance of the estimate of the mean of V or $\bar{V}$ , based on n samples is then [var(V)]/n, or

$$var(\bar{V}) = \frac{1}{n}\int_A [V(x) - E(V)]^2 f(x)dx + \frac{1}{n}\int_{R-A} [V(x) - E(V)]^2 f(x)dx \quad (13)$$

# Demonstration of the Stratification Property : Proof (cont.)

- In an unbiased sampling scheme with total n samples for these two strata, nP(A) or na samples are used for stratum A and nP(R~A) or n(1−a) are used for stratum R~A. Therefore, the variance of $[a\bar{V}_1 + (1-a)\bar{V}_2]$ is

$$var[a\bar{V}_1 + (1-a)\bar{V}_2]$$

$$= a^2 \cdot var(\bar{V}_1) + (1-a)^2 \cdot var(\bar{V}_2)$$

$$= \frac{a^2}{na} \cdot var(V_1) + \frac{(1-a)2}{n(1-a)} \cdot var(V_2)$$

$$= \frac{a}{n}\int_A [V(x) - \bar{V}_1]^2 f_1(x)dx + \frac{(1-a)}{n}\int_{R\sim A}[V(x) - \bar{V}_2]^2 f_2(x)dx$$

$$= \frac{1}{n}\int_A [V(x) - \bar{V}_1]^2 f(x)dx + \frac{1}{n}\int_{R\sim A}[V(x) - \bar{V}_2]^2 f(x)dx \quad (14)$$

# Demonstration of the Stratification Property : Proof (cont.)

- By comparing (13) and (14), $var[\bar{V}] = var[a\bar{V}_1 + (1-a)\bar{V}_2]$ if $\bar{V} = \bar{V}_1 = \bar{V}_2$.

$$var[\bar{V}] = \frac{1}{n}\int_A [V(x) - \bar{V}]^2 f(x)dx + \frac{1}{n}\int_{R\sim A}[V(x) - \bar{V}]^2 f(x)dx \quad (13)$$

$$var[a\bar{V}_1 + (1-a)\bar{V}_2]$$
$$= \frac{1}{n}\int_A [V(x) - \bar{V}_1]^2 f(x)dx + \frac{1}{n}\int_{R\sim A}[V(x) - \bar{V}_2]^2 f(x)dx \quad (14)$$

# Demonstration of the Stratification Property : Proof (cont.)

- When $E(V \mid A) \neq E(V \mid R)$, which is the assumption of the theorem, consider a function $h(v)$ defined by

$$h(v) = \int_A [V(x) - v]^2 f(x)dx \; .$$

- Taking the derivative of $h(v)$ with respect to $v$ gives

$$\begin{aligned}
\frac{dh(v)}{dv} &= -2\int_A [V(x) - v]f(x)dx \\
&= -2\int_A V(x)f(x)dx + 2\int_A v \cdot f(x)dx \\
&= -2a\overline{V}_1 + 2av = 2a(v - \overline{V}_1)
\end{aligned}$$

and $dh(v)/dv = 0$ implies $v = \bar{V}_1$ .

# Demonstration of the Stratification Property : Proof (cont.)

- Note that the second derivative of $h(v)$ with respect to $v$ is $2a > 0$. Then, $v = \bar{V}_1$ as the minimum value for $h(v)$, which means that if $E(V \mid A) \neq E(V \mid R)$ then

$$\int_A [V(x) - \bar{V}_1]^2 f(x)dx < \int_A [V(x) - \bar{V}]^2 f(x)dx.$$

Similarly, if $E(V \mid R{\sim}A) \neq E(V \mid R)$ then

$$\int_{R\sim A} [V(x) - \bar{V}_2]^2 f(x)dx < \int_{R\sim A} [V(x) - \bar{V}]^2 f(x)dx.$$

- Therefore, $var[a\bar{V}_1 + (1-a)\bar{V}_2] < var[\bar{V}]$.

# Optimal Choice of the Number of Samples

- Let $w$ denote the variance of the mean of the stratified sample, i.e.,

$$w = var[a\overline{V_1} + (1-a)\overline{V_2}].$$

made with total $n$ samples, with $m$ samples for the first stratum and $n-m$ samples for the second stratum.

This gives

$$w(m) = \frac{a^2}{m} var(V_1) + \frac{(1-a)^2}{n-m} var(V_2)\ .$$

- By setting $dw(m)/dm = 0$, the optimum value $m$ is

$$m = \frac{n}{1+\sqrt{k}} \quad \text{where} \quad k = \frac{(1-a)^2 var(V_2)}{a^2 \cdot var(V_1)}\ .$$

# Biased Sampling Scheme

- In the general case of biased sampling, the samples are picked from a modified pdf.
- Consider the random variable $V(x)$ with pdf $f(x)$. The expected value of $V$ is then given by

$$E(V) = \int_R V(x) f(x)\, dx. \qquad (1)$$

- Given a different pdf $g(x)$, such that $g(x) > 0$ everywhere that $V(x)f(x) > 0$, one has

$$E(V) = \int_R \frac{V(x) f(x)}{g(x)} \cdot g(x)\, dx. \qquad (2)$$

# Biased Sampling Scheme (cont.)

- The random variable V' associated with the pdf g(x) is

$$V'(x) = V(x)\cdot f(x)/g(x) \quad \text{and} \quad E(V') = E(V).$$

- The proper selection of g(x) can result in the variance being significantly reduced.
  - Select g(x) such that

$$\underline{V'(x)} = V(x)\cdot f(x)/g(x) = c \equiv \text{Constant.} \qquad (3)$$

Then,

$$E(V) = \underline{E(V')} = \int_R \frac{V(x)f(x)}{g(x)}\cdot g(x)dx = \int_R c\cdot g(x)dx = c.$$

and

$$var(V') = \int_R \left[(V'(x) - E(V')\right]^2 g(x)dx = 0.$$

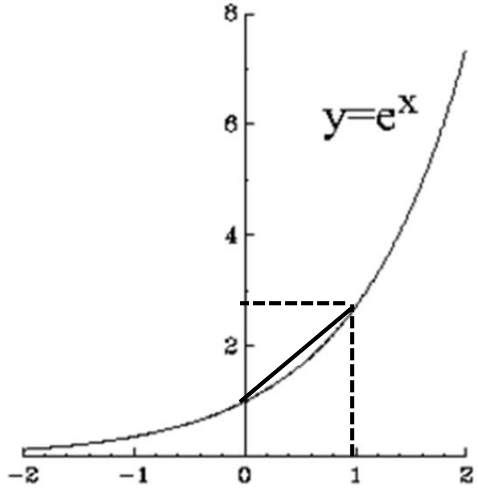
# Biased Sampling Scheme (cont.)

- Therefore, in theory, it is possible to choose a modified pdf such that the sampling process gives an answer with zero variance if one can find $g(x)$ that has the same shape as $V(x)f(x)$ to satisfy (3).

# Biased Sampling Scheme: example

- Consider the variable $V(x) = e^x$ and the pdf

$$f(x) = \begin{cases} 1, & \text{for } 0 \le x \le 1 \\ 0, & \text{elsewhere} \end{cases} . \qquad (4)$$




- Then

$$E(V) = \int_0^1 V(x) f(x)\,dx = \int_0^1 e^x\,dx \approx 1.71828 .$$

- The MC estimates with $10^6$ samples without biasing are
  - the integral estimate, $\bar{V}(x)$ = 1.71825757
  - standard deviation of random variable values $V(x_i)$'s = 0.49168896
  - standard deviation of the integral estimate $\bar{V}(x)$ = <u>0.00049169</u>

(5)

# Biased Sampling Scheme: example (cont.)

- Select a pdf for biasing
  - Considering a straight line from (0,1) to (1,2.718) is somewhat the same shape as V(x)f(x), define a modified pdf as

$$g(x) = \frac{1+1.718x}{1.859}. \qquad (6)$$

- The MC results by the biased sampling are
  - the integral value = 1.71823566
  - standard deviation of random variable = 0.06507303
  - standard deviation of the integral estimate = 0.00006507

(7)

--→ The result obtained by biased sampling is more reliable.

# Biased Sampling/ weighting factor

- A new random variable V' is defined for a modified pdf g(x) as

$$V'(x) = V(x) \cdot f(x)/g(x) \quad \text{to give} \quad E(V') = E(V).$$

- The original random variable V(x) is weighted by

$$V'(x)/V(x) = f(x)/g(x) \equiv \text{weighting factor}$$

to compensate the modification of pdf from f(x) to g(x).

# Homework #3