Rationale of Monte Carlo Approximation

SNU/NUKE/EHK

Monte Carlo Simulation

- Once the pdf's are known, the MC simulation can proceed by random sampling of parametric choices from the pdf's.
- Many simulations are then performed (multiple "histories") and the desired result is taken as an average over the number of observations.
- Along with the result, its statistical error ("variance") is informed implying the number of MC trials that is needed to achieve a certain level of reliability.

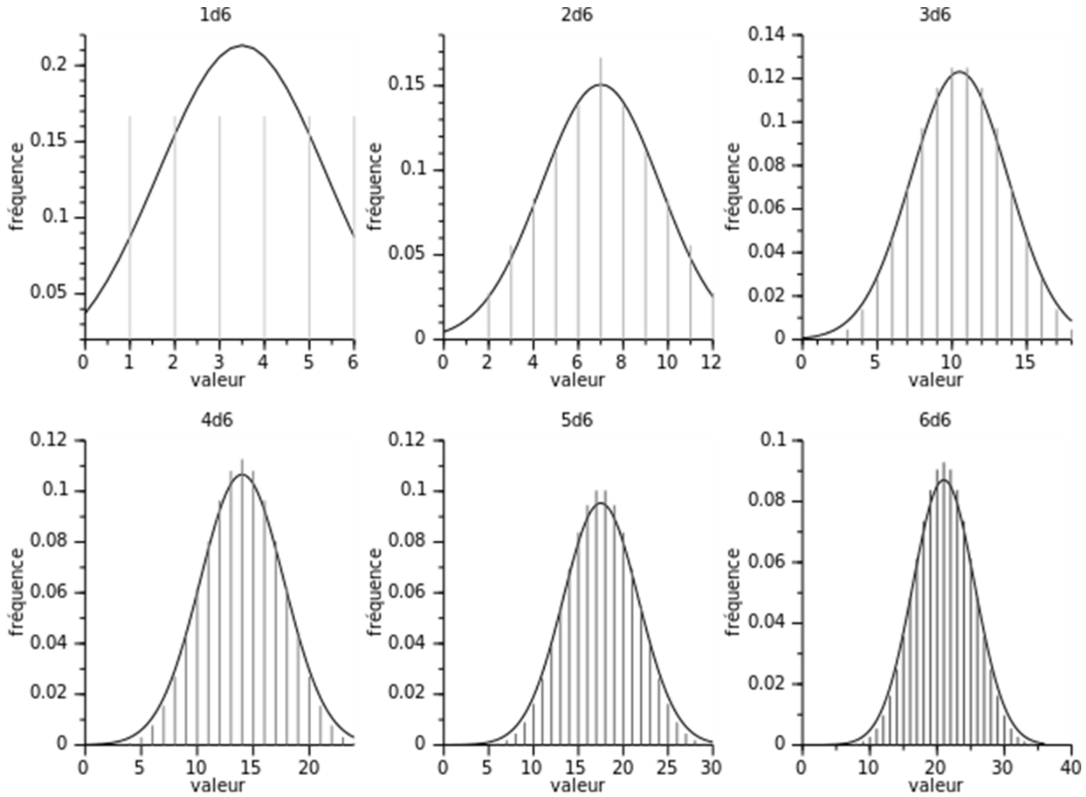
Central Limit Theorem

- Given certain conditions, the arithmetic mean (G) of a sufficiently large number of iterates ($g_i(x)$, $i=1, ..n$) of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution.

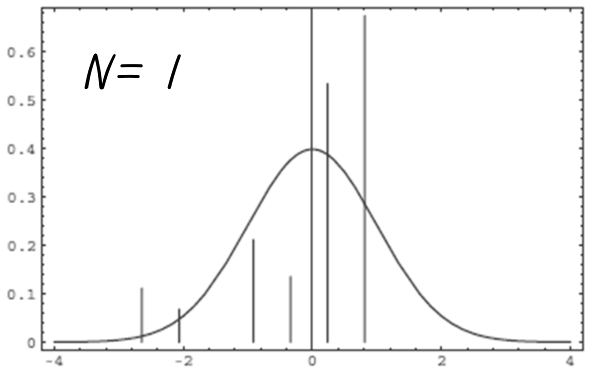
- Since real-world quantities are often the balanced sum of many unobserved random events, the central limit theorem also provides a partial explanation for the prevalence of the normal probability distribution. It also justifies the approximation of large-sample statistics to the normal distribution in controlled experiments.

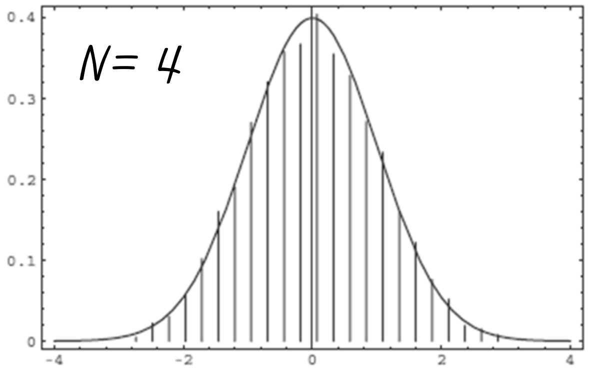
CLT: example

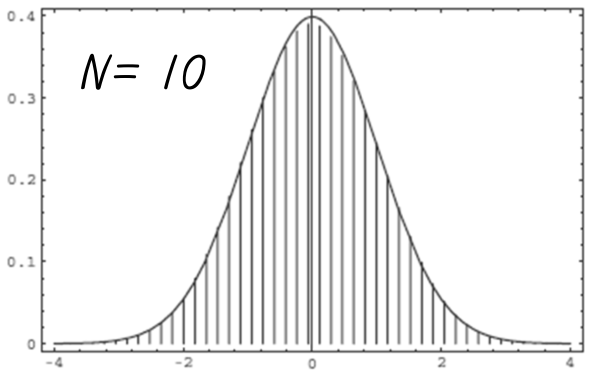
- A simple example:
  Rolling a large number of identical, unbiased dice. The distribution of the sum (or average) of the rolled numbers will be well approximated by a normal distribution.

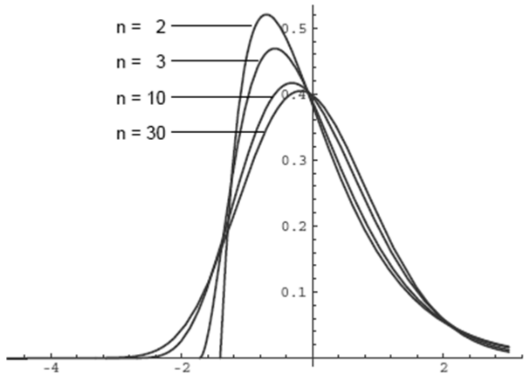


주사위를 n개 흔들 때 나오는 눈의 합 $S_n = X_1 + \ldots + X_n$의 분포가
n이 확대됨에 따라 정규 분포에 의한 근사치에 접근한 모습









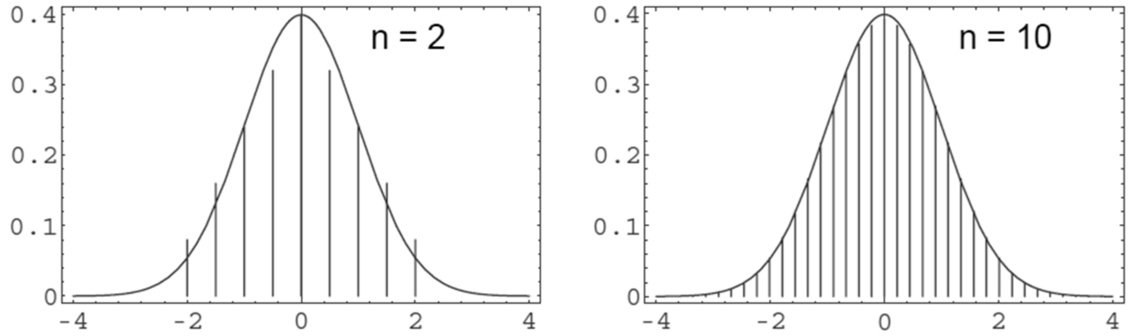
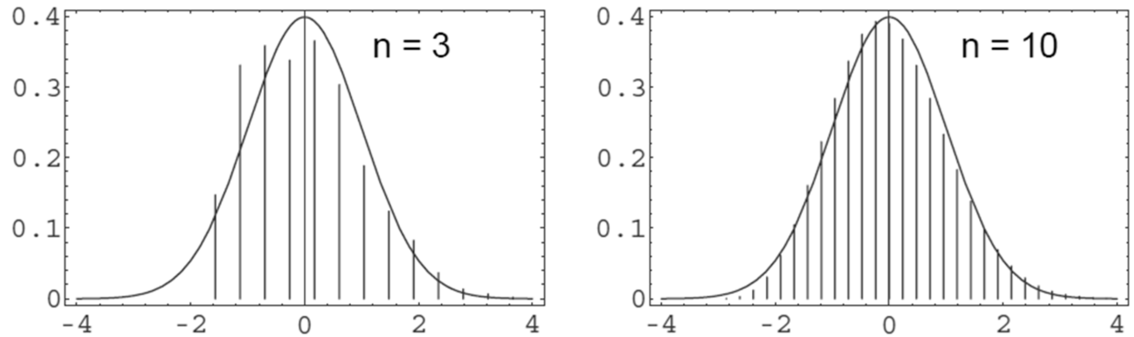
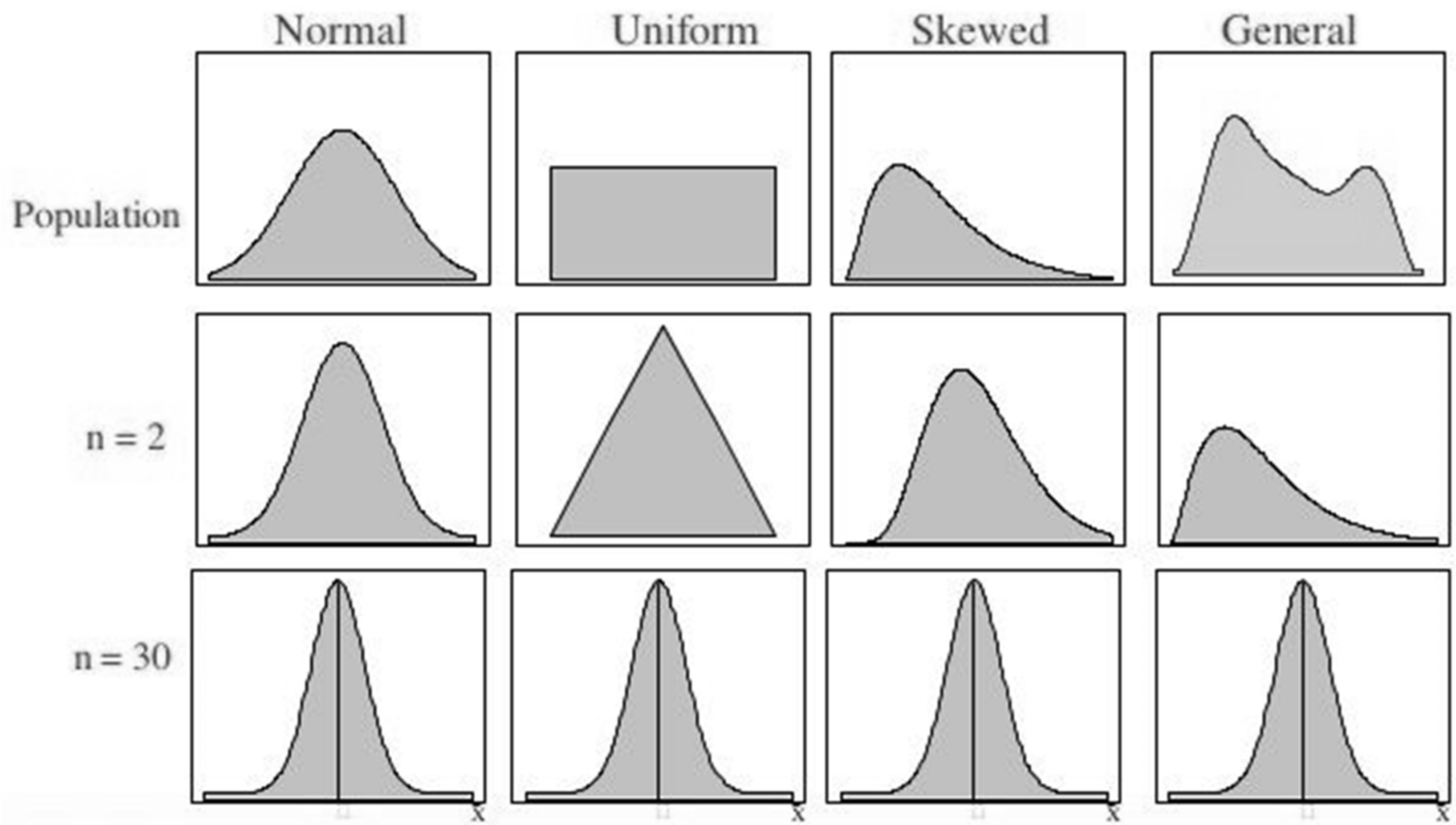


Figure 9.7: Distribution of standardized sums.



CLT: example



Expected Values in MC Simulation

- In many cases, the true mean is not known and the purpose of the Monte Carlo simulation is <u>to estimate the true mean</u>.
- The estimates of MC simulation are denoted by

$$\hat{x} \text{ and } \hat{g}(x),$$

hoping that

$$\hat{x} \text{ and } \hat{g}(x) \text{ are good approximations to } \overline{x} \text{ and } \overline{g}(x).$$

Sums of Random Variables

- Draw N samples $x_1, x_2, \ldots, x_N$ from $f(x)$ and define the following linear combination

$$G = \sum_{n=1}^{N} c_n g_n(x_n) \qquad (1)$$

where c_n's are real constants and $g_n(x)$ are real-valued functions <u>independent from each other</u>.

- The mean and variance of G are

$$E(G) = \bar{G} = E\left[\sum_{n=1}^{N} c_n g_n(x_n)\right] = \sum_{n=1}^{N} c_n E[g_n(x_n)] = \sum_{n=1}^{N} c_n \overline{g_n}(x_n) \qquad (2)$$

$$var[G] = var\left[\sum_{n=1}^{N} c_n g_n(x_n)\right] = \sum_{n=1}^{N} c_n^{\,2}\, var[g_n(x_n)] \qquad (3)$$

Sums of Random Variables (cont.)

- Consider the special case where $g_n(x) = g(x)$ and $c_n = 1/N$:

$$G = \frac{1}{N}\sum_{n=1}^{N} g(x_n) \qquad (4)$$

- The expectation value for G is

$$\bar{G} = E\left[\frac{1}{N}\sum_{n=1}^{N} g(x_n)\right] = \frac{1}{N}\sum_{n=1}^{N} E[g(x_n)] = \frac{1}{N}\sum_{n=1}^{N} \bar{g}(x_n) = \bar{g}(x) \quad (5)$$

---→ The expectation value for the average (not the average itself) of N observations of the random variable $g(x)$ is simply the expectation value for $g(x)$.

– The simple average is an unbiased estimator for the mean.

Sums of Random Variables (cont.)

- Also, the variance is given by

$$var[G] = var\left[\frac{1}{N}\sum_{n=1}^{N} g_n(x_n)\right] = \left(\frac{1}{N}\right)^2 \sum_{n=1}^{N} var[g(x_n)]$$

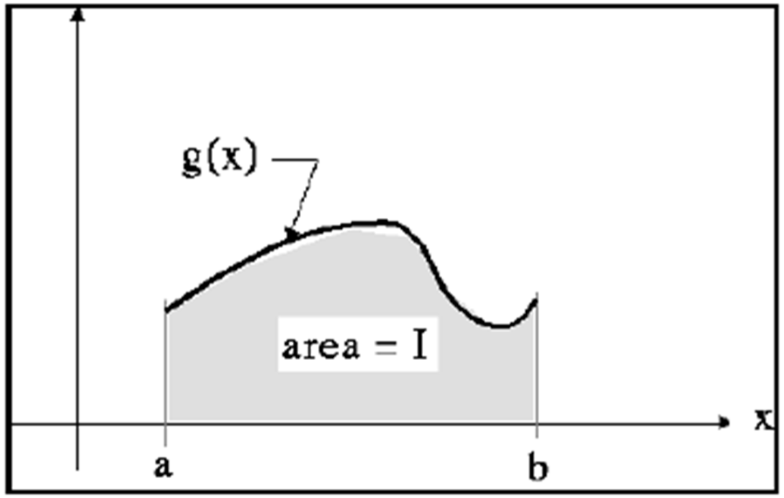
$$= \left(\frac{1}{N}\right)^2 N \cdot var[g(x_n)] = \left(\frac{1}{N}\right) var[g(x_n)] \qquad (6)$$

---→ The variance in <u>the average value of N samples of g(x)</u> is smaller than the variance in the original random variable g(x) by a factor of N.

Monte Carlo Integration

- To evaluate the following definite integral,

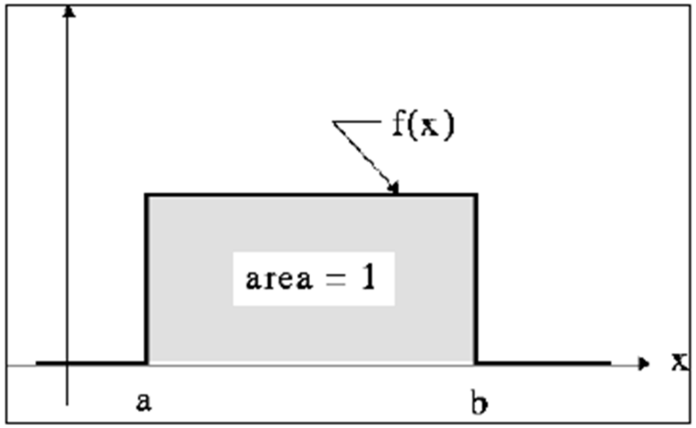
$$I = \int_a^b g(x)dx \qquad (7)$$



Monte Carlo Integration (cont.)

- To manipulate the definite integral into a form that can be solved by MC, we define the following function on [a,b].

$$f(x) = \begin{cases} 1/(b-a), & a \le x \le b \\ 0, & \text{otherwise} \end{cases} \tag{8}$$



  - $f(x)$ is a uniform pdf on the interval [a, b].

Monte Carlo Integration (cont.)

- Insert Eq. (8) into Eq. (7) to obtain the following expression

$$I = (b-a)\int_a^b g(x)f(x)dx. \qquad (9)$$

- Given that $f(x)$ is a pdf of x, the <u>integral</u> on the RHS in (9) is simply the expectation value for $g(x)$:

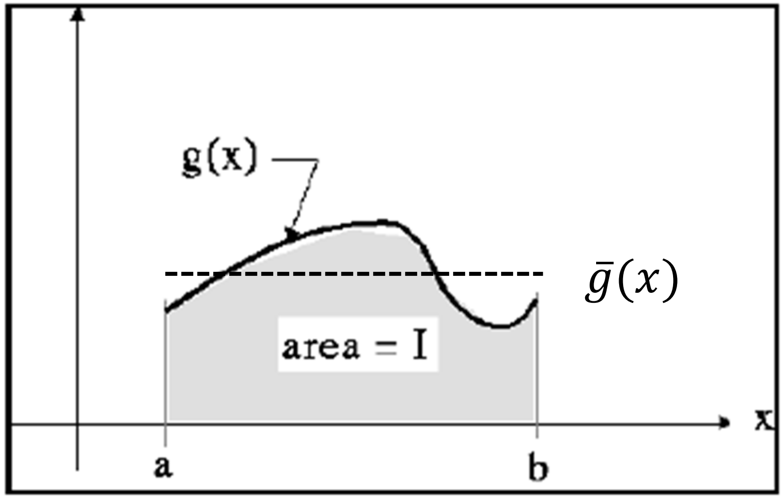
$$I = (b-a)\,\overline{g}(x) \qquad (10)$$

Monte Carlo Integration

- To evaluate the following definite integral,

$$I = \int_a^b g(x)dx \qquad (7)$$

$$I = (b-a)\ \bar{g}(x) \qquad (10)$$



Monte Carlo Integration (cont.)

- Now, we draw samples x_n from the pdf $f(x)$, evaluate $g(x_n)$, and form G, the average of $g(x)$:

$$G = \frac{1}{N}\sum_{n=1}^{N} g(x_n) \qquad (11)$$

Monte Carlo Integration (cont.)

- Recall

$$\bar{G} = E\left[\frac{1}{N}\sum_{n=1}^{N} g(x_n)\right] = \frac{1}{N}\sum_{n=1}^{N} E[g(x_n)] = \frac{1}{N}\sum_{n=1}^{N} \bar{g}(x_n) = \bar{g}(x) \quad (5)$$

- Hence,

$$I = (b-a)\cdot\bar{g} = (b-a)\cdot\bar{G} \approx (b-a)\cdot G = (b-a)\cdot\left(\frac{1}{N}\sum_{n=1}^{N} g(x_n)\right) \qquad (12)$$

where the interval [a, b] is finite.

Monte Carlo Integration (cont.)

- Recall

$$var[G] = \left(\frac{1}{N}\right) var[g(x_n)], \qquad (6)$$

which relates the variance in the integral G to the true variance in g(x).

- We might expect the error (σ) in the estimate of I to decrease by the factor $N^{-1/2}$.

Rejection Technique

- *An alternative method of sampling based on the fact that the integral of pdf $f(x)$ from $-\infty$ to x or $F(x)$ is the area under the curve of $f(x)$ over this interval.*

- *The "rejection technique" is one very simple technique for selecting points uniformly within the area under the curve and thereby calculating the area.*

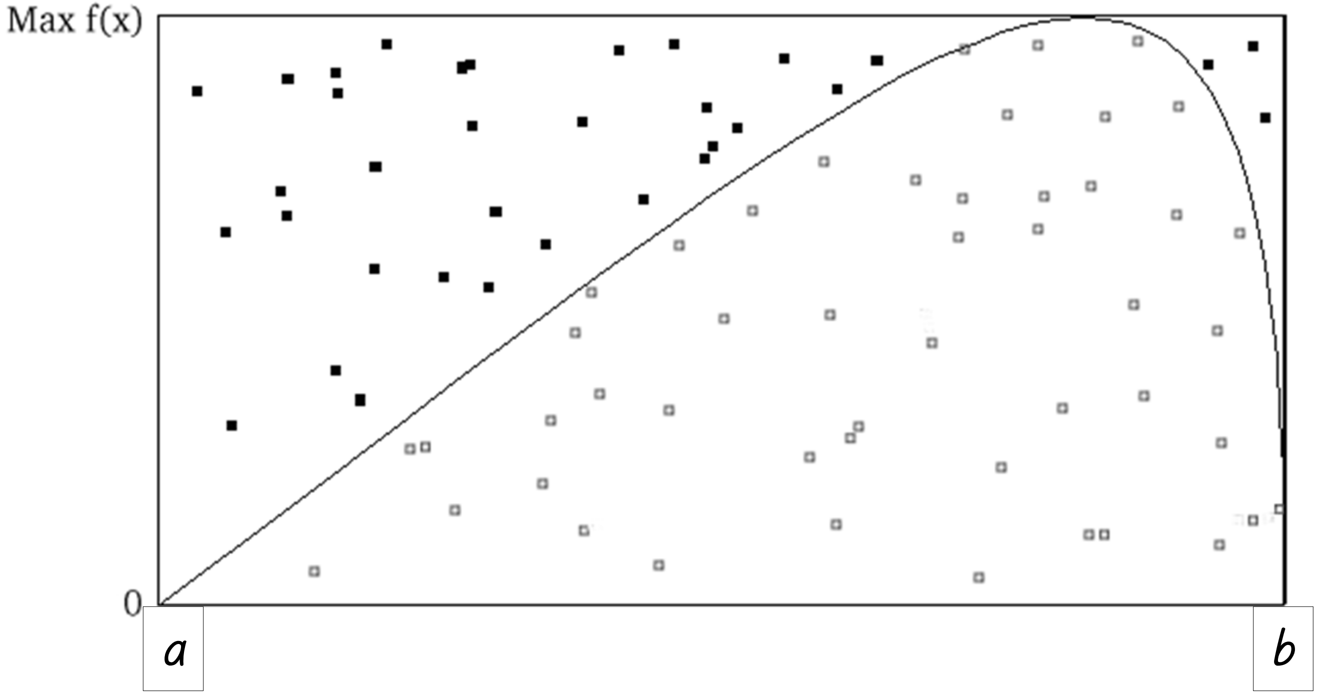
- *Ex. calculation of pi*

Rejection Technique (cont.)

- Consider a pdf $f(x)$ that has a maximum value less than or equal to some number in the range $a \leq x \leq b$ and is zero outside the range.

$$\begin{cases} f(x) \leq M, & a \leq x \leq b \\ f(x) = 0, & \text{otherwise} \end{cases}$$

- The area under the curve defined by $f(x)$ is enclosed within the rectangle bounded by $0 \leq y \leq M$ and $a \leq x \leq b$.


Max f(x)
0
a
b

Rejection Technique (cont.)

- Calculation of the area under the curve $f(x)$ by sampling random numbers $\xi_i (I = 1, 2, \cdots)$
  - Step 1. select a point (x, y) by uniform chance within the rectangle:

$$x = a + (b-a)\xi_1 \quad \text{and} \quad y = M\xi_2.$$

  - Step 2. examine whether the point (x, y) falls under the curve $f(x)$:

    If it does, accept the point. Otherwise, reject it!

  - Step 3. $\text{Area} = (1 - \text{fraction of rejection}) \cdot [(b-a)M]$

- For the rejection technique to be efficient, ensure that the fraction of points that are rejected is small.

Estimation of Means and Variances

- Given a discrete set of n samples of the random variable X, which consists of the numbers $S=\{x_1, x_2, \dots x_n\}$, the mean value of S or the sample mean is

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n} x_i \, , \qquad (13)$$

which is the best estimate for the mean of X one can make.

- The variance of the sample S is

$$var(S) = \sigma_I^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{S})^2 \, , \qquad (14)$$

which is a valid estimate for the variance σ^2 of the variable X.

Estimation of Means and Variances (cont.)

- If the true mean is known, the variance is given by

$$var(S) = \sigma_2^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 . \qquad (15)$$

- For small value of n, a correction is necessary:

$$var(S) = \sigma_3^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \overline{S})^2 . \qquad (16)$$

- σ_1^2 is the variance of the sample S whereas σ_2^2 and σ_3^2 are the unbiased estimates of the variance of X.

Degrees of freedom in statistics
$\equiv$ the number of observations that are free to vary.
= the number of independent scores that go into the estimate
minus the number of parameters used as intermediate steps.

Estimation of Means and Variances (cont.)

- In practice, one uses a large value of n with the true mean of X unknown.
- For n greater than ~30, the difference between (14) and (15) is sufficiently small that it can be ignored.

  ---→ One normally uses Eq. (14) to estimate the variance of random variables.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{S})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^{\,2} - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2. \qquad (17)$$

Estimation of Means and Variances (cont.)

$$var[G] = var\left[\frac{1}{N}\sum_{n=1}^{N} g_n(x_n)\right] = \left(\frac{1}{N}\right)^2 \sum_{n=1}^{N} var[g(x_n)]$$

$$= \left(\frac{1}{N}\right)^2 N \cdot var[g(x_n)] = \left(\frac{1}{N}\right) var[g(x_n)] \qquad (6)$$

- From (6),

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

- The standard deviation of the estimate of the mean is then

$$StdDev(\bar{S}) = \sqrt{\frac{1}{n} var(X)} = \frac{1}{\sqrt{n}} StdDev(X) = \frac{\sigma}{\sqrt{n}}. \qquad (19)$$

- Fractional Standard Deviation (fsd) ≡ StdDev/Mean
  - A result with fsd greater than about 0.1 should be questionable!