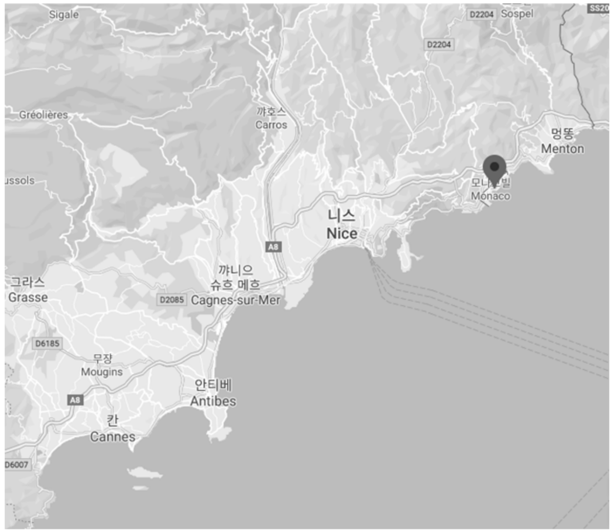
Introduction: Monte Carlo method

SNU/NUKE/EHK

Monte Carlo in Monaco



Terminology

- Monte Carlo method
  = a statistical simulation method

- Simulation
  = the imitation of the operation of a real-world process or system over time

- Statistical simulation
  ≡ any method that utilizes sequences of random numbers to perform the simulation

Statistical vs. Deterministic

- Numerical discretization method
  - is applied to ordinary or partial differential equations that describe underlying physical system.
  - discretizing those differential equations and then solving a set of algebraic equations for the unknown state of the system.

- In Monte Carlo simulation,
  - the physical process is simulated directly with no need to write down the mathematical expressions on the behavior of the system
  - the use of random sampling is essential to arrive at the solution of the physical problem.

Modeling for MC Simulation

- Modeling
  = the transition from the physical situation to a "mathematical relation"
- Modeling of the physical process by one or more probability density functions (pdf's)
  –→ an essential component of Monte Carlo simulation
  – The pdf's may have their origins in experimental data or in a theoretical model describing the physics of the process.

Monte Carlo Simulation

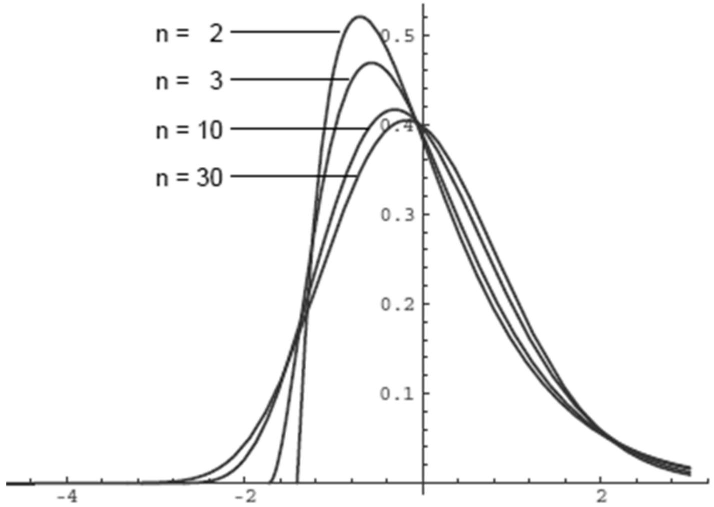
- Once the pdf's are known, the MC simulation can proceed by random sampling of parametric choices from the pdf's.
- Many simulations are then performed (multiple "histories") and the desired result is taken as an average over the number of observations.
- Along with the result, its statistical error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

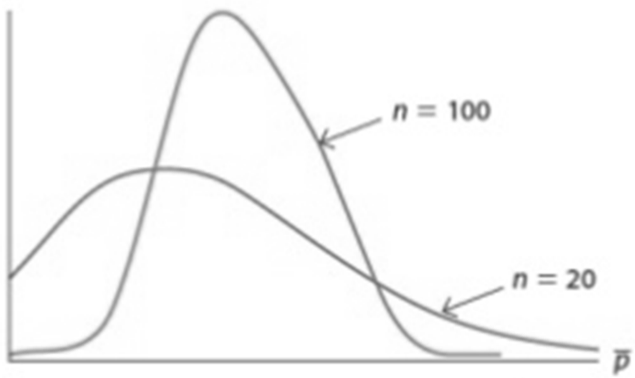
Central Limit Theorem

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

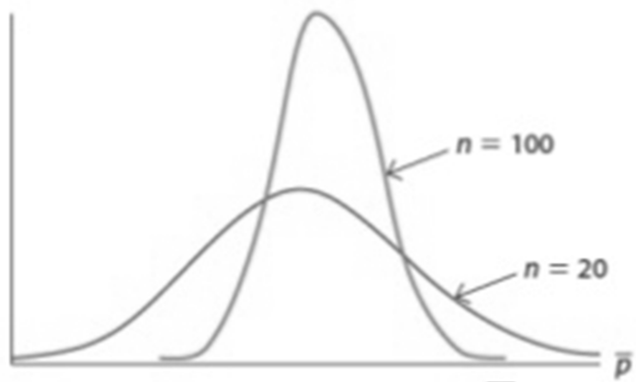
- Since real-world quantities are often the balanced sum of many unobserved random events, the central limit theorem also provides a partial explanation for the prevalence of the normal probability distribution. It also justifies the approximation of large-sample statistics to the normal distribution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

Central Limit Theorem (cont.)





Sampling distribution of $\bar{P}$ when the population proportion is $p = 0.10$.



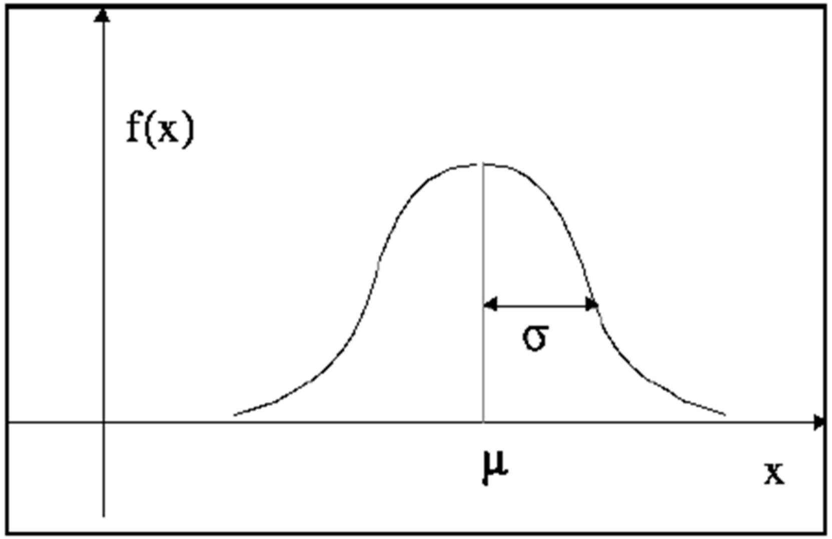
Sampling distribution of $\bar{P}$ when the population proportion is $p = 0.30$.

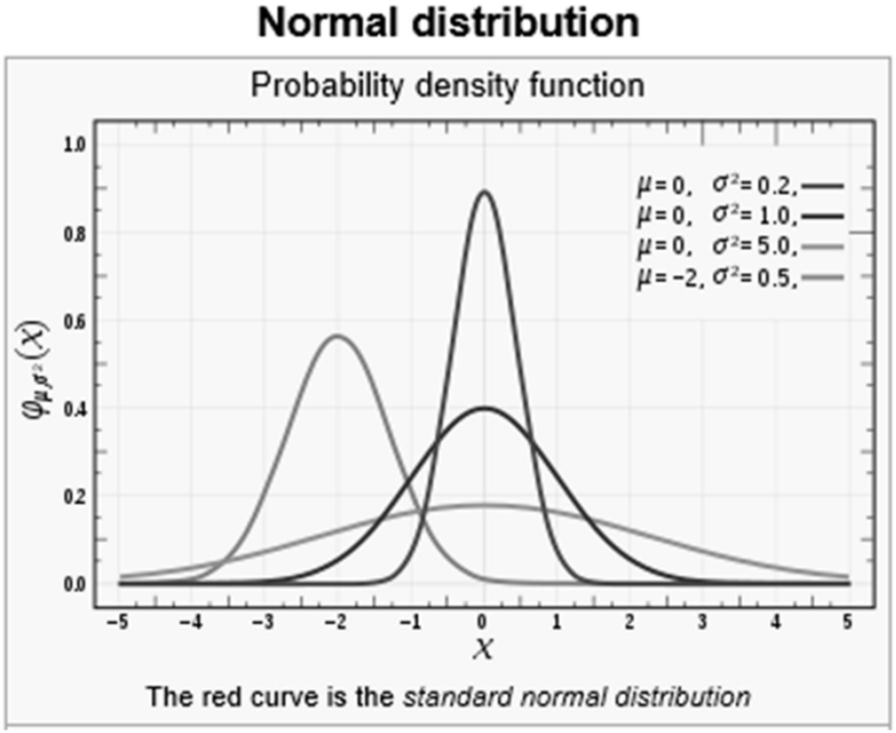
- *Gaussian (Normal) distribution*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad , -\infty < x < \infty$$

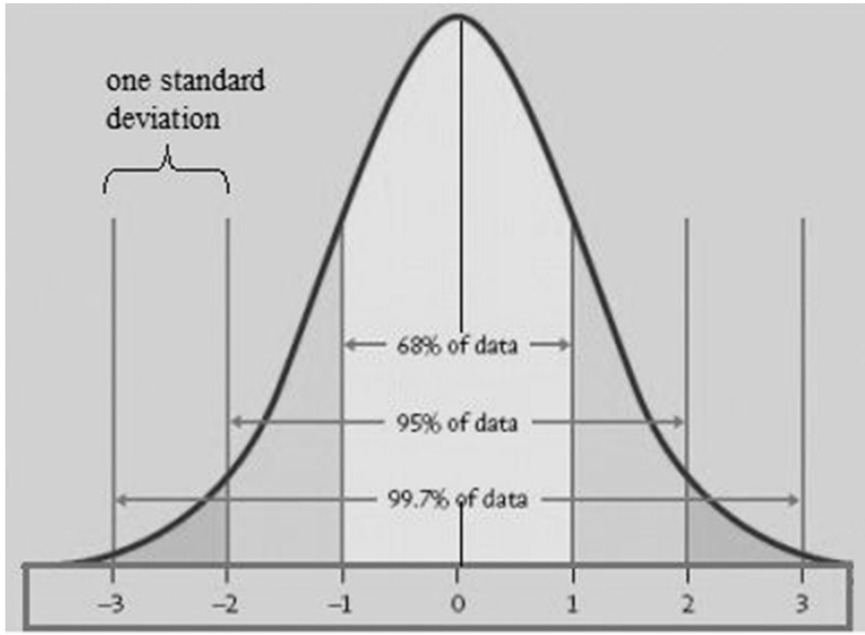
which is a two-parameter (μ and σ) distribution.

– μ and σ^2 are the mean and variance of the distribution.





parametric properties in normal distribution



Moment

- The nth moment of a real-valued function f(x) of a real variable value x <u>about a value c</u> is

$$\mu_n = \int_{-\infty}^{\infty} (x-c)^n f(x)dx$$

- If f(x) is a probability density function of x, then

$$\mu_n \equiv \int_{-\infty}^{\infty} (x-c)^n f(x)dx = E\left[(x-c)^n\right]$$

- For discrete variable values x_i and p_i,

$$\mu_n = \sum_i p_i (x_i - c)^n = E\left[(x-c)^n\right]$$

1st Moment:

$$\mu_1 = \int_{-\infty}^{\infty} (x-c)^1 f(x)\,dx$$

- If f(x) is a probability density function of x, then

$$\mu_1 \equiv \int_{-\infty}^{\infty} (x-c)^1 f(x)\,dx = E\left[(x-c)^1\right]$$

- For discrete variable values x_i and p_i,

$$\mu_1 = \sum_i p_i\,(x_i - c)^1 = E\left[(x-c)^1\right]$$

→ Mean of the variable values in f(x)

2nd Moment:

$$\mu_2 = \int_{-\infty}^{\infty} (x-c)^2 f(x)dx$$

- *If $f(x)$ is a probability density function of x, then*

$$\mu_2 \equiv \int_{-\infty}^{\infty} (x-c)^2 f(x)dx = E\left[(x-c)^2\right]$$

- *For discrete variable values x_i and p_i,*

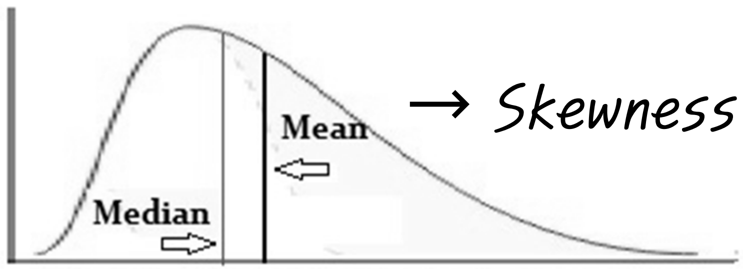
$$\mu_2 = \sum_i p_i (x_i - c)^2 = E\left[(x-c)^2\right]$$

→ *Variance of the variable values in $f(x)$*

3rd Moment:

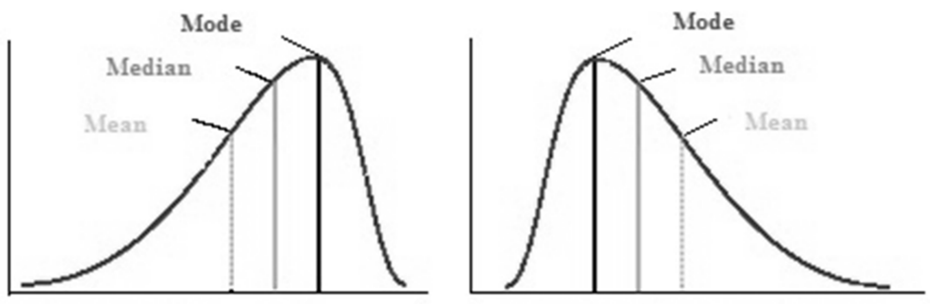
$$\boxed{\mu_3} \equiv \int_{-\infty}^{\infty} (x-c)^3 f(x)\,dx = E\left[(x-c)^3\right]$$

$$\boxed{\mu_3} = \sum_i p_i\,(x_i - c)^3 = E\left[(x-c)^3\right]$$



Right skewed distribution: Mean is to the right

→ *Skewness* $\underbrace{(\text{왜도, 歪度})}_{=\ \text{lack of symmetry}}$ *of the distribution* $f(x)$



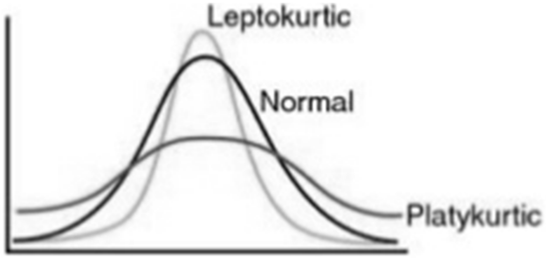
Left-Skewed (Negative Skewness) Right-Skewed (Positive Skewness)

- *skewness* with left tails < $\underbrace{0 \text{ of normal distribution}}_{\text{with median = mean}}$ < *skewness* with right tails

4^{th} Moment:

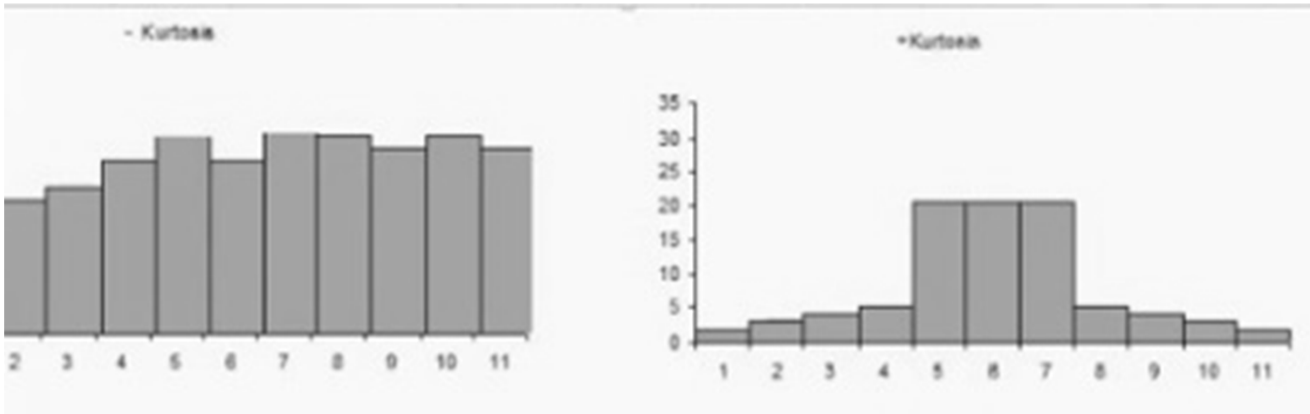
$$\boxed{\mu_4} \equiv \int_{-\infty}^{\infty} (x-c)^{4} f(x)dx = E\left[(x-c)^{4}\right]$$

$$\boxed{\mu_4} = \sum_i p_i (x_i - c)^{4} = E\left[(x-c)^{4}\right]$$



flatness

→ *Kurtosis* (첨도, 尖度) *of the distribution* $f(x)$



light tails (little data in tails) heavy tails (lots of data in tails)

- *kurtosis* with light tails < 3 of normal distribution < *kurtosis* with heavy tails

1^{st} moment: measure of the central location in f(x)

$$\mu_1 \equiv E[x] = \bar{X} = \int_{-\infty}^{\infty} x \cdot f(x)dx \text{ , or } = \sum_{i=1}^{n} p_i \cdot x : mean\ of\ x\ from\ f(x)$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i : sample\ mean$$

2^{nd} moment: measure of the dispersion

$$\mu_2 \equiv E[(x_i - \bar{X})^2] = \sigma^2 = \int_{-\infty}^{\infty} (x - \bar{X})^2 \cdot f(x)dx\,, or\ = \sum_{i=1}^{n} p_i \cdot (x_i - \bar{X})^2$$

$: variance\ of\ x\ in\ f(x)$

$s^2 \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2$ $: sample\ variance$

$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$ $: corrected\ sample\ variance$

3^{rd} moment: measure of asymmetry

$$\mu_3 \equiv E[(x_i - \bar{X})^3] = \int_{-\infty}^{\infty} (x - \bar{X})^3 \cdot f(x)dx\,, or = \sum_{i=1}^{n} p_i \cdot (x_i - \bar{X})^3$$

$$skewness \equiv E\left[\left(\frac{x_i - \bar{X}}{\sigma}\right)^3\right] = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(x_i - \bar{X})}{\sigma}\right]^3 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{{\mu_2}^{3/2}} \;: assymetry$$

$$\cong \frac{\frac{1}{N-1}\ \sum_{i=1}^{N}(x_i - \bar{x})^3}{\left[\frac{1}{N-1}\ \sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{3/2}}$$

4^{th} moment: measure of peakedness

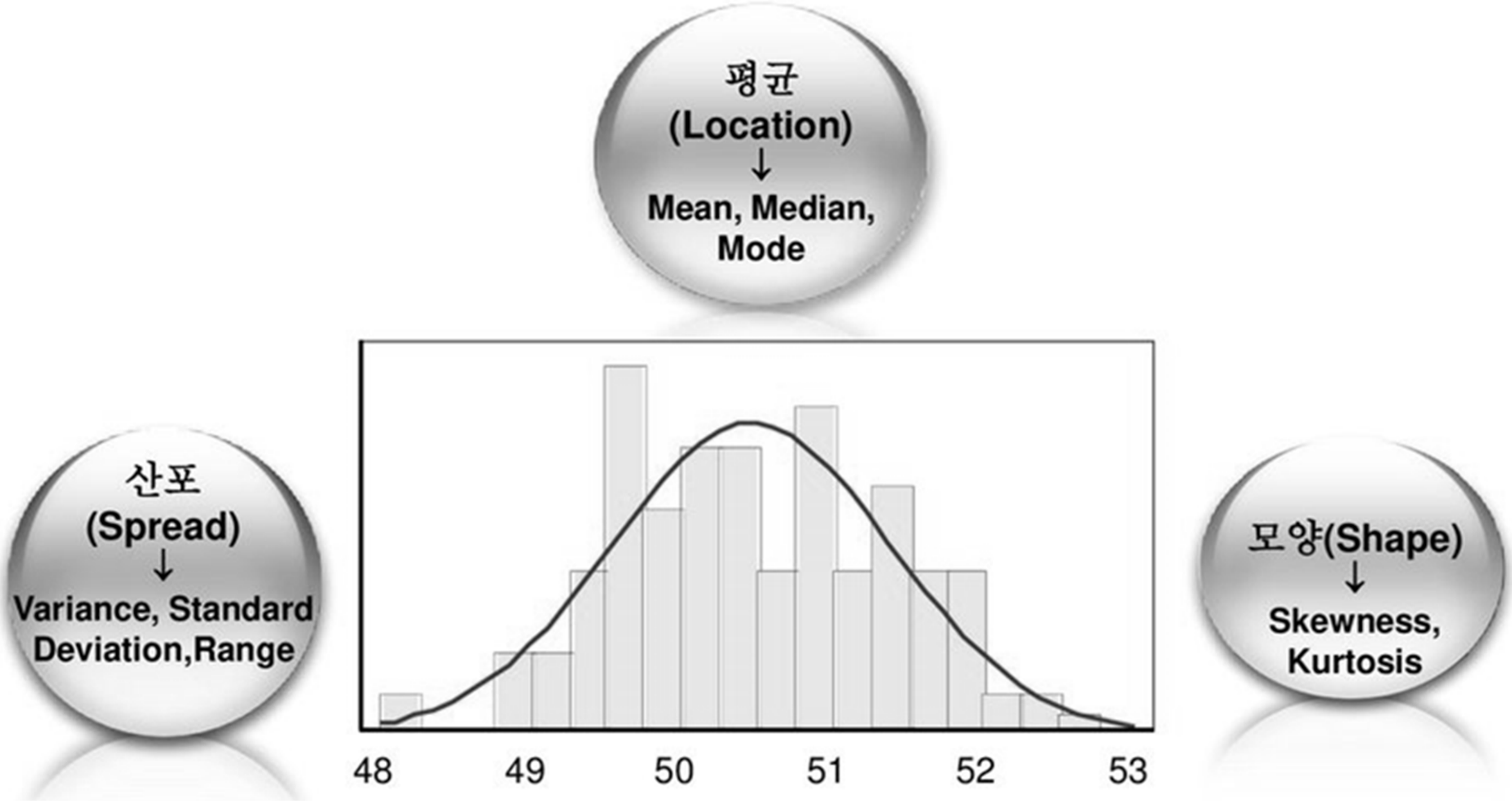
$$\mu_4 \equiv \int_{-\infty}^{\infty} (x-\bar{X})^2 \cdot f(x)dx = E[(x_i-\bar{X})^2] \quad or \quad \mu_2 = \sum_{i=1}^{n} p_i \cdot (x_i-\bar{X})^2$$

$$kurtosis \equiv E\left[\left(\frac{x_i-\bar{X}}{\sigma}\right)^4\right] = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(x_i-\bar{X})}{\sigma}\right]^4 = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{{\mu_2}^2}:$$

$$peakedness\ or\ attendance\ of\ outliers$$

$$\cong \frac{\frac{1}{N-1}\ \sum_{i=1}^{N}(x_i-\bar{x})^4}{\left[\frac{1}{N-1}\ \sum_{i=1}^{N}(x_i-\bar{x})^2\right]^2}$$

- *Specification of a variable in distribution*



Terminology in MC

- *the mathematical process*

  -→ an experiment in MC

  – The experiment has a number of possible outcomes, to which we assign probabilities.

- *the collection of possible outcomes*

  -→ the sample space S of the experiment

- *one realization of the experiment*

  -→ a trial results in an outcome s in the sample space S.

- *The consequence of the outcomes of multiple experiments*

  -→ the occurrence of a specific event E_k

Ex. computational experiment

- An experiment consists of one roll of a normal die and observation the top face of the die
- Every realization (i.e., each trial) results in one of six faces being the top face. The outcomes s_i are those six faces, and the sample space S consists of these six outcomes.
- An event can be defined in terms of the possible outcomes. The possible outcomes are:
  - E_1 : top face is an even number;
  - E_2 : top face is larger than 4;
  - E_3 : top face is equal to 2; etc.
- Disjoint events that can not happen at the same time. (Ex. E_2 and E_3)