2024-1

Multivariate Data Analysis

Youjin Shin
(yj.shinn@catholic.ac.kr)

Probability Error

- **Probability error** refers to the deviation or discrepancy between the observed outcomes and the expected probabilities of these outcomes.

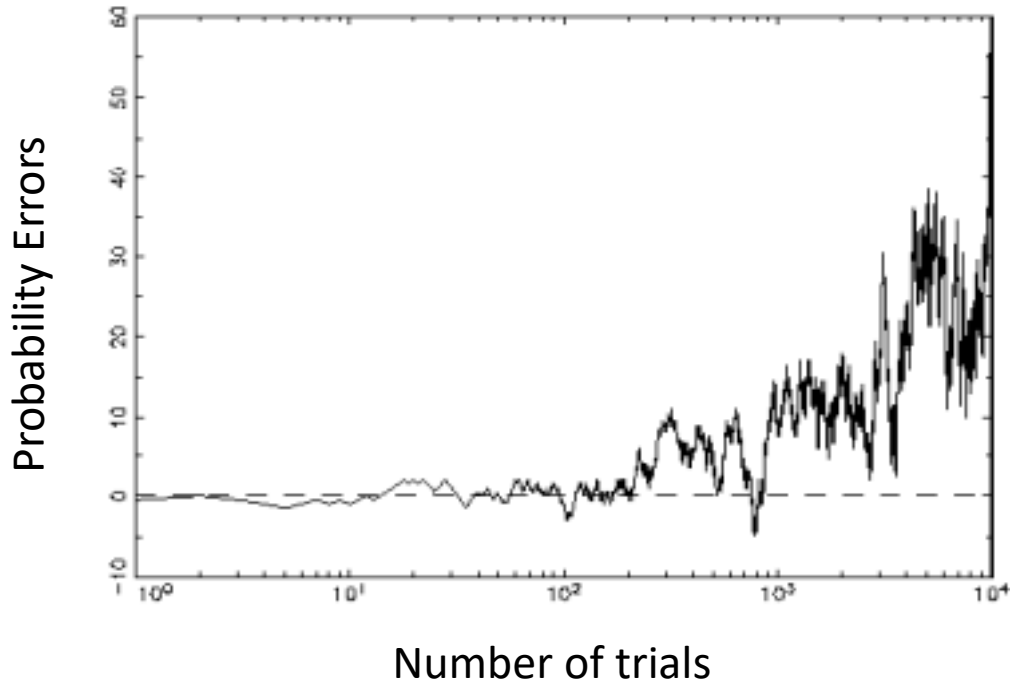
  - **Example:** Tossing a Coin
  - Imagine you toss a fair coin 10 times. You might anticipate getting heads approximately 5 times and tails 5 times, reflecting the 50% chance for each outcome (expected probabilities).

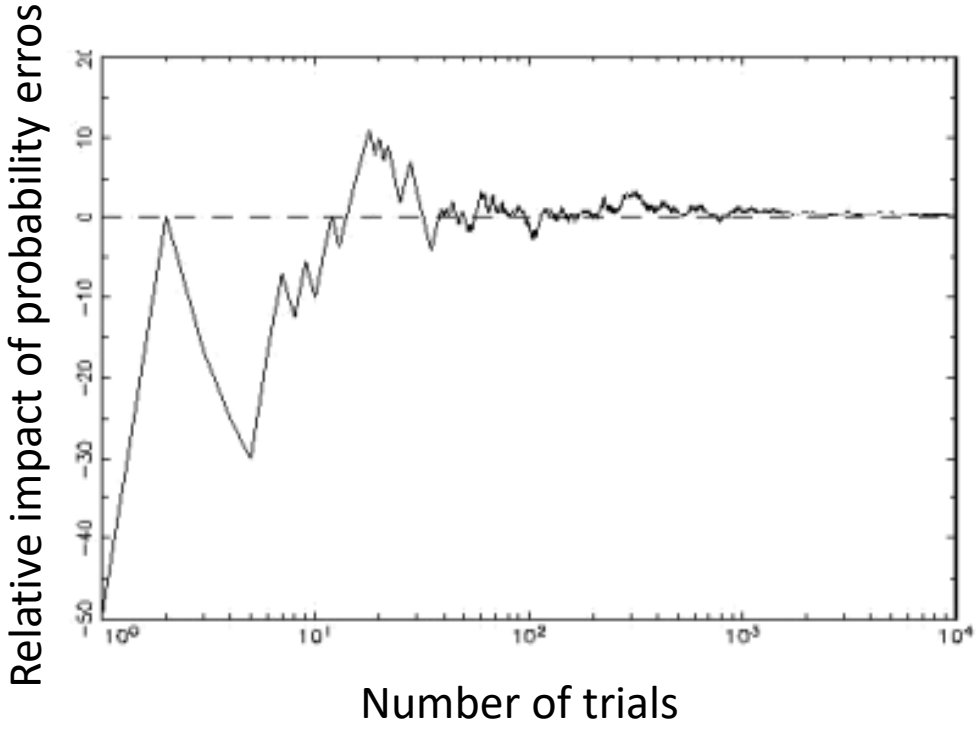
  - However, in real world, you might get with 6 heads and 4 tails.
    - Probability error to get heads : +1
    - It is the discrepancy between the expected number of heads (5) and the actual number of heads (6)
    - For tails, with 4 occurrences instead of the expected 5, the probability error would be -1.

  - This deviation is natural in random processes and illustrates the concept of probability error: the difference between what is expected, based on probability theory, and what actually occurs in practice.

Probability Error

- **The law of large numbers**
  - As we increase the number of coin tosses to a very large number, the absolute values of the probability errors increase.
  - The relative frequency (proportion) of heads and tails will get closer to their theoretical probability (50% each), reducing the impact of probability error in the long run.





Expected Value (Mean)

- **The expected value** of a random variable provides a measure of the center or the average outcome you would expect to see if you were to repeat the experiment an infinite number of times.
  - For a discrete random variable, the expected value is calculated using the formula:

    $E[X] = \sum_i x_i p_i$, where x_i represents the ith possible value of X, and p_i is the probability of X taking the value x_i.

  - **Example(1):** In the example of tossing a coin, expected value is calculated as blow:

    $$E[X] = (1 \times 0.5) + (0 \times 0.5) = 0.5$$

  - **Example(2):** In the example of rolling a fair six-sided dice, the possible outcomes are 1, 2, 3, 4, 5, or 6, each with an equal probability of 1/6. The expected value of the roll can be calculated as follows:

    $$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

    - This means, on average, you can expect to roll a 3.5 on a fair six-sided dice.

Expected Value (Mean)

◆ Expected Value vs. Mean

- **Expected Value** is a probabilistic concept and refers to the long-run average outcome of a random variable if an experiment is repeated many times. It's a theoretical mean of the random variable's probability distribution.

- **Mean**, in a statistical context, typically refers to the average of observed data. When discussing a data set, the mean is calculated by summing all the observations and dividing by the number of observations.

◆ For a given probability distribution or in a theoretical context, the expected value and the mean are essentially the same concept. The expected value is the mean of the distribution.

Standard Error

- **Standard Error** specifically measures the variability or uncertainty in the estimate of a sample statistic (like the sample mean) relative to the true population parameter.
  - It's most commonly used to describe the variability (error) of sample means around the population mean.
  - The standard error of the mean is calculated as the standard deviation of the population divided by the square root of the sample size as blow:

$$\text{standard deviation } \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

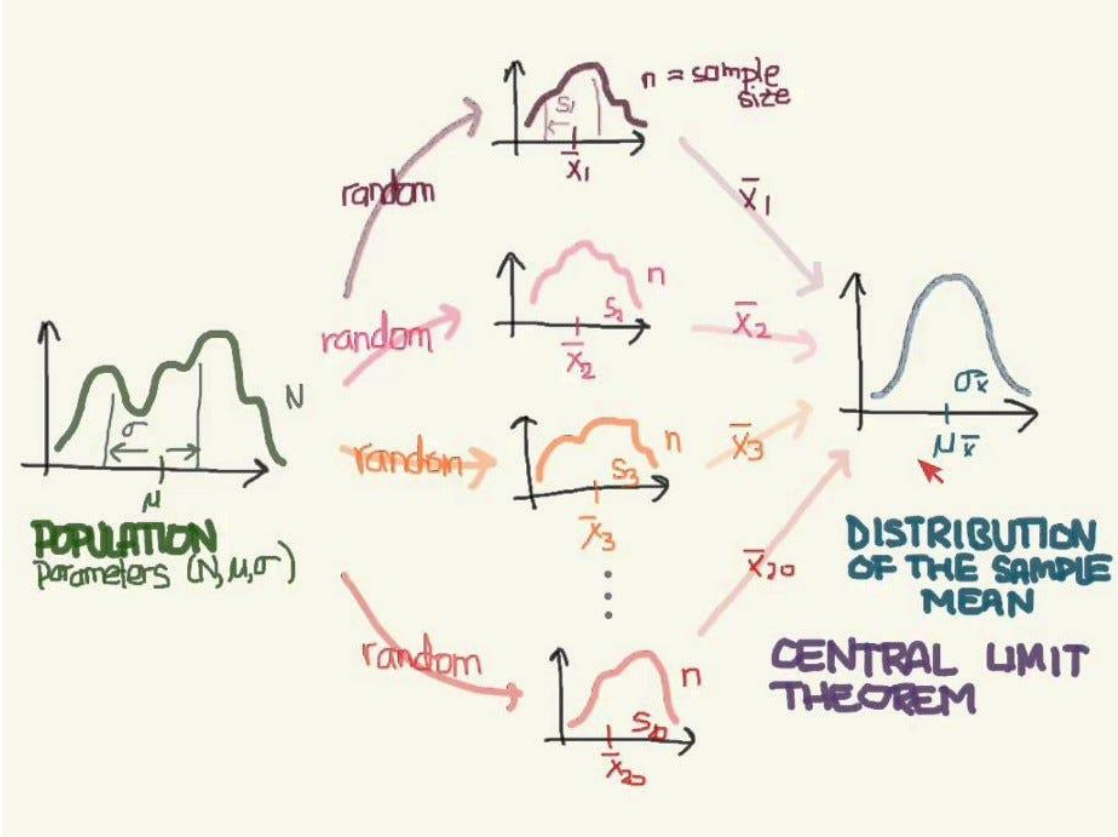
$$\text{variance} = \sigma^2$$

$$\text{standard error } (\sigma_{\bar{x}}) = \frac{\sigma}{\sqrt{n}}$$

where:

$\bar{x}$ = the sample's mean

n = the sample size



Standard Error

- **Standard Error** specifically measures the variability or uncertainty in the estimate of a sample statistic (like the sample mean) relative to the true population parameter.
  - It's most commonly used to describe the variability (error) of sample means around the population mean.
  - The standard error of the mean is calculated as the standard deviation of the population divided by the square root of the sample size as blow:

$$\text{standard deviation } \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$\text{variance} = \sigma^2$$

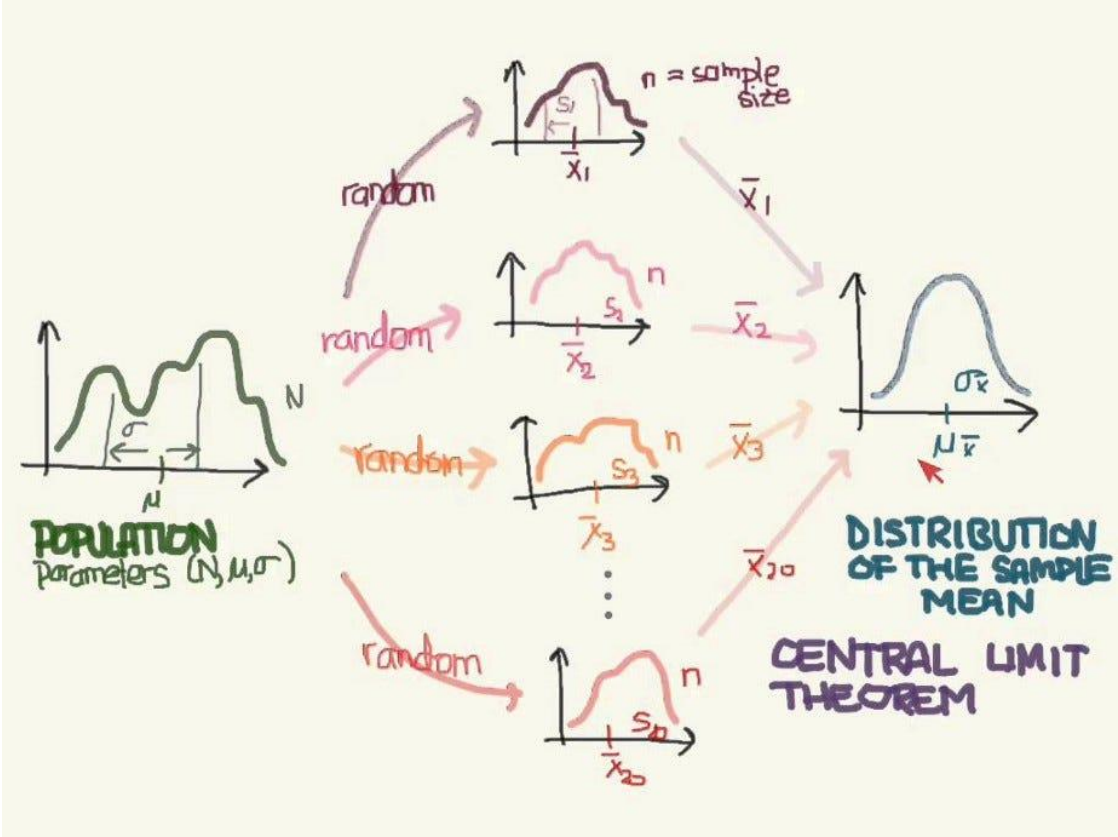
$$\text{standard error } (\sigma_{\bar{x}}) = \frac{\sigma}{\sqrt{n}}$$

(the standard error of the sample mean (SEM))

where:

$\bar{x}$ = the sample's mean

n = the sample size



Standard Error

Derivation [edit]

The standard error on the mean may be derived from the variance of a sum of independent random variables,[6] given the definition of variance and some properties thereof. If $x_1, x_2, \ldots, x_n$ is a sample of n independent observations from a population with mean $\bar{x}$ and standard deviation σ, then we can define the total

$$T = (x_1 + x_2 + \cdots + x_n)$$

which due to the Bienaymé formula, will have variance

$$\operatorname{Var}(T) = \big(\operatorname{Var}(x_1) + \operatorname{Var}(x_2) + \cdots + \operatorname{Var}(x_n)\big) = n\sigma^2.$$

where we've approximated the standard deviations, i.e., the uncertainties, of the measurements themselves with the best value for the standard deviation of the population. The mean of these measurements $\bar{x}$ is given by

$$\bar{x} = T/n.$$

The variance of the mean is then

$$\operatorname{Var}(\bar{x}) = \operatorname{Var}\left(\frac{T}{n}\right) = \frac{1}{n^2}\operatorname{Var}(T) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

The standard error is, by definition, the standard deviation of $\bar{x}$ which is the square root of the variance:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

For correlated random variables the sample variance needs to be computed according to the Markov chain central limit theorem.

Standard Error

- **Example** : In the example of rolling a fair six-sided dice, suppose we roll the dice 36 times
- First, calculate the standard deviation of a single dice roll's outcome:

$$\sigma = \sqrt{\frac{(1-3.5)^2+(2-3.5)^2+(3-3.5)^2+(4-3.5)^2+(5-3.5)^2+(6-3.5)^2}{6}} \approx 1.7078$$

- The standard error of the sample mean (SEM) can be calculated using the formula:

$$SEM = \frac{\sigma}{\sqrt{n}}$$

- For our dive rolling, with n=36, the standard error would be:

$$SEM = \frac{1.7078}{\sqrt{36}} = \frac{1.7078}{6} \approx 0.2846$$

- This standard error tells us about the variability we might expect in the average outcome of 36 rolls from the true mean (3.5) of the dice's outcomes. If we repeated this experiment of 36 rolls many times, calculating the mean each time, we'd expect the distribution of those means to have a standard deviation (i.e., standard error) of approximately 0.2846.

Bernoulli Distribution

- In everyday life, outcomes are frequently categorized into pairs
  - increase/decrease
  - success/failure
  - pass/fail
  - alive/dead

- Such binary outcomes are referred to as **Bernoulli trials**.

- **The Bernoulli distribution** is one of the simplest discrete distributions, based on a Bernoulli trial where the outcomes are binary, represented as 0 (failure) or 1 (success).
  - This distribution is characterized by a single parameter p, which represents the probability of success. The probability of failure is represented as $1 - p$

Binomial Distribution

◆ **Binomial distribution** models the number of successes in a given number of independent Bernoulli trials. For example, tossing a coin 10 times and modeling the number of heads obtained can be represented using a binomial distribution.

- ➢ This distribution is defined by two parameters, n and p,
  - where n is the number of trials,
  - and p is the probability of success in each trial.
- ➢ The outcomes can range from 0 to n, where each value represents the number of successes in n trials.

◆ While the Bernoulli distribution models a single binary trial, the binomial distribution aggregates the outcomes of multiple Bernoulli trials. Thus, the binomial distribution can be seen as the sum of multiple Bernoulli trials.

Binomial Distribution

- The binomial distribution in which the random variable X follows parameters n and p is expressed as follows:

  X~B(n,p)

- *When X~B(n,p),* The probability mass function (PMF) of *X* is given by:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where:

- $P(X = k)$ is the probability of getting exactly k successes.
- $\binom{n}{k}$ is the binomial coefficient, calculated as $\frac{n!}{k!(n-k)!}$, which gives the number of ways k successes can occur in n trials.
- p is the probability of success on a single trial.
- $1 - p$ is the probability of failure on a single trial.
- n is the number of trials.
- k is the number of successes (ranging from 0 to n).

Binomial Distribution

◆ (EX) Toss a coin

- ➢ When you toss a coin 10 times, the number of times heads can appear ranges from 0 to 10, with the respective probabilities being 0.001, 0.01, 0.044, 0.117, 0.205, 0.246, 0.205, 0.117, 0.044, 0.01, and 0.001.

$$k=0: \frac{10!}{0!\cdot 10!}\left(\frac{1}{2}\right)^{0}\cdot\left(1-\frac{1}{2}\right)^{10}=0.0010$$

$$k=1: \frac{10!}{1!\cdot 9!}\left(\frac{1}{2}\right)^{1}\cdot\left(1-\frac{1}{2}\right)^{9}=0.0098$$

$$k=2: \frac{10!}{2!\cdot 8!}\left(\frac{1}{2}\right)^{2}\cdot\left(1-\frac{1}{2}\right)^{8}=0.0439$$

$$\vdots$$

$$k=10: \frac{10!}{10!\cdot 0!}\left(\frac{1}{2}\right)^{10}\cdot\left(1-\frac{1}{2}\right)^{0}=0.0010$$

$$P(X=k)=\binom{n}{k}p^{k}(1-p)^{n-k}$$

where:

- $P(X=k)$ is the probability of getting exactly k successes.
- $\binom{n}{k}$ is the binomial coefficient, calculated as $\frac{n!}{k!(n-k)!}$, which gives the number of ways k successes can occur in n trials.
- p is the probability of success on a single trial.
- $1-p$ is the probability of failure on a single trial.
- n is the number of trials.
- k is the number of successes (ranging from 0 to n).

E(X) and Var(X) of X~B(n,p)

- When X is a binomial random variable that follows a binomial distribution, let's calculate $E(X_1)$ and $Var(X_1)$ of this X.
  - First, we calculate mean and variance of Bernoulli random variable (X_1)

$$P(X_1 = x) = p^x(1-p)^{(1-x)}, \qquad x = 0, 1$$

$$E(X_1) = \sum_x xP(X_1 = x) \quad = 0 \times (1-p) + 1 \times p = p$$

$$Var(X_1) = E(X_1 - E(X_1))^2 = \sum_x (x-p)^2 P(X_1 = x) \quad = (0-p)^2 (1-p) + (1-p)^2 p = p(1-p)$$

  - Next, we calculate mean and variance of Bernoulli random variables $(X_1, X_2, \cdots, X_n)$, where the success probability of variables is equal to P.

$$E(X) = E(X_1 + X_2 + \cdots + X_n) = nE(X_1) = np$$

$$Var\ (X) = Var(X_1 + X_2 + \cdots + X_n) = nVar(X_1) = np(1-p)$$

E(X) and Var(X) of X~B(n,p)

- QUIZ: You are planning to toss a fair coin 100 times. Probability of getting a head (p) and Probability of getting a tail ($1-p$) are both ½. Answer the following questions:

1. What is the probability of getting heads in this single trial?
    1. A) 0
    2. B) 0.25
    3. C) 0.5
    4. D) 1
2. Calculate the expected value ($E[X]$) for this single trial.
    1. A) 0
    2. B) 0.25
    3. C) 0.5
    4. D) 1
3. Determine the variance ($Var(X)$) for the number of heads in this single trial.
    1. A) 0.25
    2. B) 0.5
    3. C) 0.75
    4. D) 1

E(X) and Var(X) of X~B(n,p)

- QUIZ: You are planning to toss a fair coin 100 times. Probability of getting a head (p) and Probability of getting a tail ($1-p$) are both ½. Answer the following questions:

1. What is the probability of getting heads in this single trial?
    1. A) 0
    2. B) 0.25
    3. C) 0.5
    4. D) 1

$$P(X_1 = 1) = p^x(1-p)^{(1-x)} = 1/2^1(1-1/2)^{(1-1)} = 1/2$$

$$P(X_1 = 0) = p^x(1-p)^{(1-x)} = 1/2^0(1-1/2)^{(1-0)} = 1/2$$

2. Calculate the expected value ($E[X]$) for this single trial.
    1. A) 0
    2. B) 0.25
    3. C) 0.5
    4. D) 1

$$E(X_1) = \sum_x xP(X_1 = x) \quad = 0 \cdot P(tail) + 1 \cdot P(head)$$

$$= 0 \times (1-p) + 1 \times p = p$$

$$= 0 \times (1-1/2) + 1 \times 1/2 = 1/2$$

3. Determine the variance ($Var(X)$) for the number of heads in this single trial.
    1. A) 0.25
    2. B) 0.5
    3. C) 0.75
    4. D) 1

$$Var(X_1) = E(X_1 - E(X_1))^2 = \sum_x (x-p)^2 P(X_1 = x)$$

$$= (0-p)^2(1-p) + (1-p)^2 p = p(1-p) \quad = \quad 1/2(1-1/2) = 1/4$$

E(X) and Var(X) of X~B(n,p)

- QUIZ: You are planning to toss a fair coin 100 times. Probability of getting a head (p) and Probability of getting a tail ($1-p$) are both ½. Answer the following questions:

4. What is the expected number of times the coin will land on heads?
    1. A) 25
    2. B) 50
    3. C) 75
    4. D) 100
5. What is the variance of the number of times the coin will land on heads?
    1. A) 12.5
    2. B) 25
    3. C) 50
    4. D) 100

E(X) and Var(X) of X~B(n,p)

◆ QUIZ: You are planning to toss a fair coin 100 times. Probability of getting a head (p) and Probability of getting a tail ($1-p$) are both ½. Answer the following questions:

4. What is the expected number of times the coin will land on heads?
   1. A) 25
   2. B) 50
   3. C) 75
   4. D) 100

$$E(X) = E(X_1 + X_2 + \cdots + X_n) = nE(X_1) = np = 100 \cdot 1/2 = 50$$

5. What is the variance of the number of times the coin will land on heads?
   1. A) 12.5
   2. B) 25
   3. C) 50
   4. D) 100

$$Var\ (X) = Var(X_1 + X_2 + \cdots + X_n) = nVar(X_1) = np(1-p) = 100 \cdot 1/2 \cdot 1/2 = 25$$


Bye Bye