Chapter 4

Probability and distribution

Professor Jung Jin Lee
Soongsil University, Korea
New Uzbekistan University, Uzbekistan

Chapter 4 Probability and distribution

- 4.1 Probability
  - 4.1.1 Calculation rules of probability and conditional probability
  - 4.1.2 Bayes theorem
- 4.2 Random variable and distribution
  - 4.2.1 Binomial distribution
  - 4.2.2 Normal distribution
- 4.3 Multivariate distribution
- 4.4 Estimation of a distribution

4.1 Probability

- Similar events occur frequently or are carried out around our lives.

- A machine is producing products repeatedly at a production plant.
  => The product is either normal or defective, but it is unknown what it will be.

- We order pizza at home every Sunday.
  => It usually takes 30 minutes to be delivered to the house, but the exact time is unknown.

4.1 Probability

- What these examples have in common is as follows.
  ① The repetition of similar events.
  ② Various possible outcomes are known.
  ③ There is no telling what exactly will happen.
- An event with these three characteristics is called a statistical experiment.
  => subject to the study and application of statistics.
- **Examples of non-statistical experiments?**

4.1 Probability

- A sample space is a set of all possible outcomes from a statistical experiment.
  - usually marked by S such as S = {normal, defective}
- A subset of this sample space is referred to as an event.
  - denoted in English capitals A, B, C as A = {defective}
- When the number of elements in a sample space is finite or counted indefinitely, it is called a discrete sample space.
- When the number of elements in a sample space is infinite and uncountably innumerable, it is called a continuous sample space.

4.1 Probability

- Probability is the 'representation of the likelihood of an event occurring' between 0 and 1.
  - If an event is likely to occur, the probability is close to 1.
  - If it is unlikely to occur, the probability is close to zero.

❖ Classical definition of probability

- Assume <u>all elements in the sample space are likely to occur equally</u>.

$$P(A) = \frac{\text{Number of elements belonging to event A}}{\text{Total number of elements in sample space}}$$

$$P(A) = \frac{\text{Measurement of elements belonging to event A}}{\text{Measurement of elements in sample space}}$$

* Measurement can be length, area, volume etc.

4.1 Probability

[Example] An office worker went on a business trip to a city with two restaurants (A and B) near his lodging. He hesitated about which restaurant to go to and threw a die to count the number of points that appeared on the top. If he had odd numbers, he would go to restaurant A; if he had even numbers, he would go to restaurant B. What is the probability that restaurant A would be picked?

<Answer>

- The sample space in this statistical experiment, which counts the number of points on the top by throwing a dice, is {1, 2, 3, 4, 5, 6}.
- The number of odd events is {1, 3, 5}, so there are three elements. Therefore, the probability that restaurant A will be selected is 3/6 = 1/2.

4.1 Probability

[Example] Order pizza from home every Sunday. The time it takes for a pizza to be delivered to the house has the same possibility for any time from 10 to 30. What is the probability that a pizza delivery will be delivered between 20 and 25 minutes?

<Answer>

- The sample space in this example is all values from 10 to 30 minutes {(10,30)}.
- The events where pizza is delivered between 20 and 25 minutes is { (20,25) }.
- Therefore, the probability of this event is (25-20) / (30-10) = 0.25 by measuring the distance of the interval.

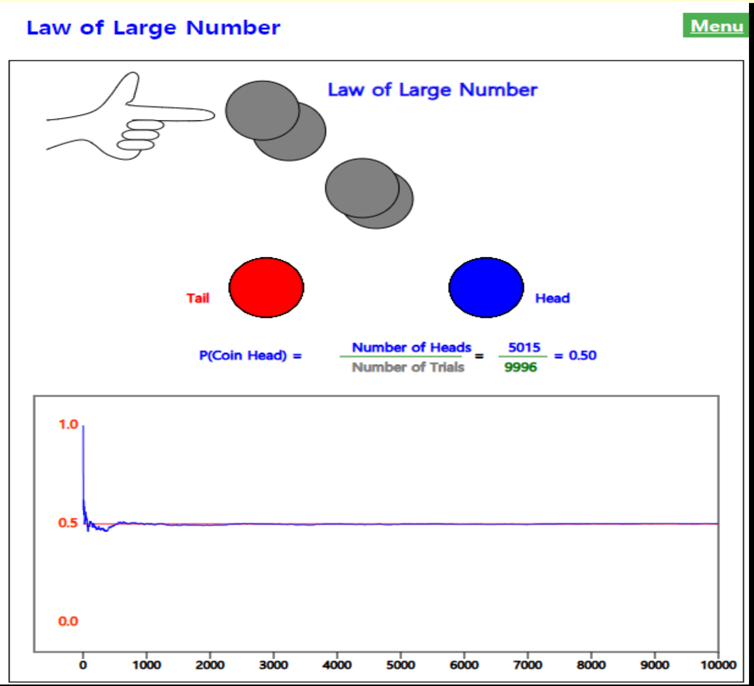
4.1 Probability

❖ **Relative frequency definition of probability**

- **The probability that event A will occur is the rate at which event A occurs when statistical experiments are conducted under the same conditions repeatedly.**

❖ **Law of Large Numbers**

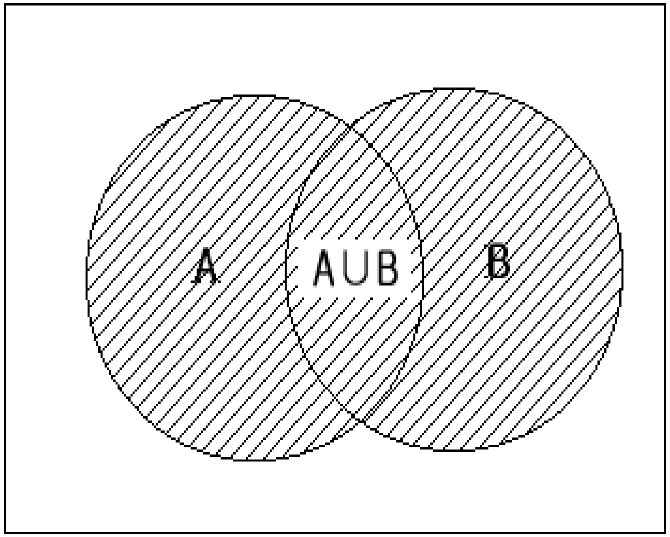
If a coin is thrown many times, the probability of {Head} event converges to half



4.1 Probability

❖ Addition rule of probability

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



If $A \cap B = \emptyset$, then the rule becomes as follows.

$$\mathbf{P(A \cup B) = P(A) + P(B)}$$

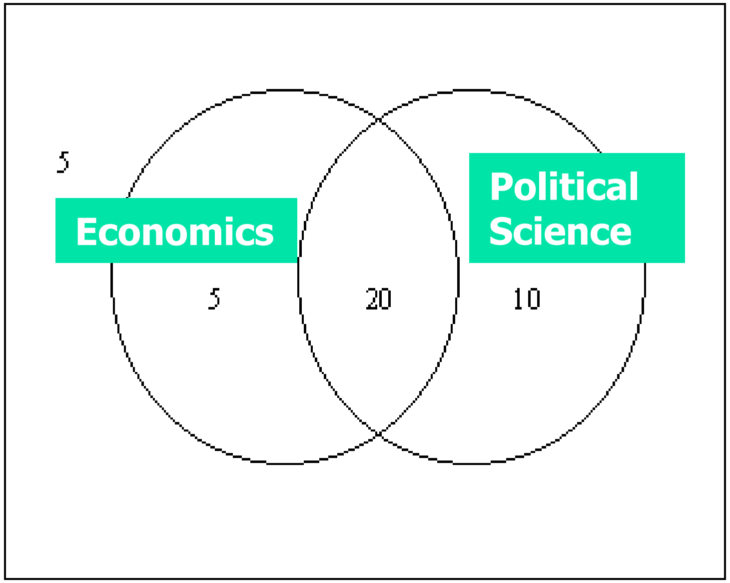
Events A and B are called mutually exclusive.

4.1 Probability

[Example] Out of 40 sophomores in the statistics department this semester, 25 students are taking economics(A), 30 students are taking political science(B), and 20 students are taking both subjects. When I meet one of the sophomores, what is the probability of this student taking either economics or political science?

<Answer>

- Since 25 students take economics and 20 students take both courses, 25 - 20 = 5 students take only economics.
- Since 30 students take political science, 30 - 20 = 10 students take political science
- The number of students taking economics or politics is 5 + 10 + 20 = 35.
- The probability of students taking economics or politics is 35 / 40.



4.1 Probability

❖ Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{if } P(B) \neq 0$$

❖ Multiplication Rule of Probability

$$P(A \cap B) = P(A)\, P(B|A)$$

If $P(B|A) = P(B)$, then $P(A \cap B) = P(A)\, P(B)$

** A and B are independent events

4.1 Probability

[Example] Of the 30 sophomores in the Department of Statistics, there are 10 males and 20 females; one of the males is from the province, and five females are from the province.

1) When selecting a student, what is the probability that he is from a province?
2) When I selected a student, she was a female. What is the probability that this student is from a province?
3) When I selected a student, he was from a province. What's the probability of this student being a male?
4) When selecting a student, what is the probability that he is male and from Baku?

4.1 Probability

<Answer>

	Baku(S)	Province(C)	Total
Male(M)	9	1	10
Female(F)	15	5	20
Total	24	6	30

1) $P(C) = 6/30$.
2) The probability that a student is from a province among girls is 5/20. It is expressed as $P(C|F)$ and is called conditional probability.
3) The probability of a male from a province origin is $P(M|C) = 1/6$.
4) The probability is $P(M \cap S)$ is 9/30. Alternatively,

$$P(M \cap S) = P(M)\,P(S|M) = (10/30) \times (9/10) = 9/30$$

It shows that $P(S|M)$ can be calculated by dividing $P(M \cap S)$ by $P(M)$.

$$P(S|M) = \frac{P(M \cap S)}{P(M)} = \frac{9/30}{10/30} = \frac{9}{10}$$

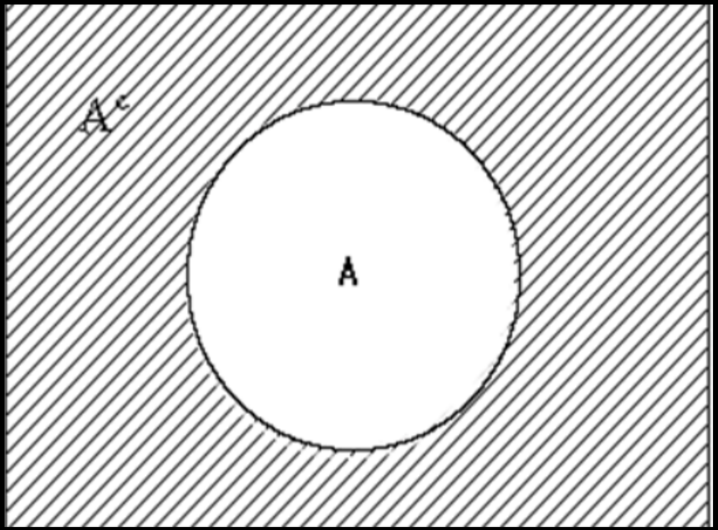
$$P(M \cap S) = P(S)\,P(M|S) = (24/30) \times (9/24)$$

4.1 Probability

❖ **Probability of a complementary event**

If A^C denotes a complementary event of A, then P(A^C) can be calculated as follows.

$$P(A^C) = 1 - P(A)$$



4.2 Random variable and distribution

- Two coins are thrown repeatedly.

  S = {'Tail-Tail', 'Tail-Head', 'Head-Tail' and 'Head-Head'}

  The probability of each element of the sample space is 1/4.
- We are interested in counting the number of heads or tails. If X is 'number of heads', possible value of X are 0, 1, or 2.
- A function corresponding to a real number between [0,1] for each element of the sample space is a random variable.

Sample Space	**X=Number of {Head}**
'Tail-Tail'	**0**
'Head-Tail'	**1**
'Tail-Head'	**1**
'Head-Head'	**2**

- If the possible values of a random variable are finite or countably infinite, it is a discrete random variable.
- If the possible values of a random variable are uncountably infinite, it is a continuous random variable.

4.2 Random variable and distribution

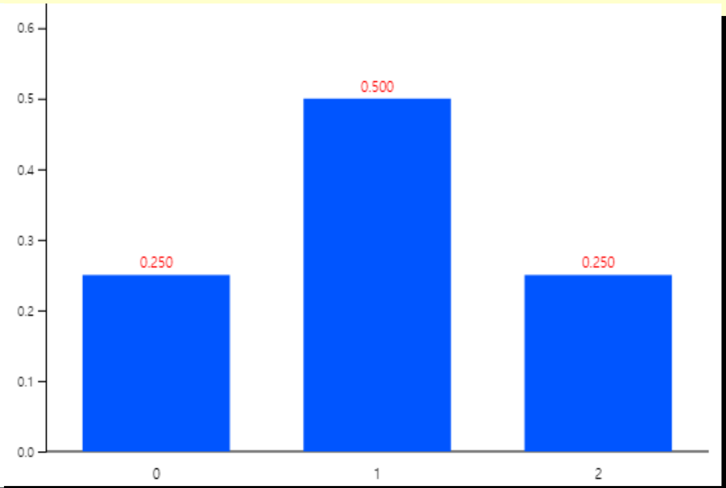
- The probability that the random variable X will be zero is 1/4 because it is the probability of an event {Tail-Tail}, the probability that X being 1 is 2/4 because P({Tail-Head, Head-Tail}) is 2/4, and the probability that X being 2 is 1/4 because P({Head-Head}) is 1/4.
- The summarized probability for the value of the random variable X is called the probability distribution function of X denoted as f(x).

1) Table style

X = x	P(X=x)
0	1/4
1	2/4
2	1/4
계	1

2) Function style

$$f(x) = 1/4,\ x=0$$
$$= 2/4,\ x=1$$
$$= 1/4,\ x=2$$



4.2 Random variable and distribution

- The mean and variance of a random variable X are measures of the central tendency and dispersion.
- The mean of X, E(X) called an expectation of X, and the variance of X, V(X), are defined as follows.

$$E(X) = \mu = \sum_{i=1}^{n} x_i P(\mathrm{X} = x_i)$$

$$V(X) = \sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 P(\mathrm{X} = x_i) = \sum_{i=1}^{n} x_i^2 P(\mathrm{X} = x_i) - \mu^2$$

4.2 Random variable and distribution

[Example] Find the expectation and variance of the random variable X = the number of heads when tossing a coin twice.

<Answer>

$$E(X) = \mu = \sum_{i=1}^{n} x_i \mathrm{P}(\mathrm{X} = x_i) = 0 \times \frac{1}{4} + 1 \times \frac{2}{4} + 2 \times \frac{1}{4} = 1$$

$$V(X) = \sigma^2 = \sum_{i=1}^{n} x_i^2 \mathrm{P}(\mathrm{X} = x_i) - \mu^2 = 0^2 \times \frac{1}{4} + 1^2 \times \frac{2}{4} + 2^2 \times \frac{1}{4} - 1^2 = \frac{1}{2}$$

4.2 Random variable and distribution

❖ Expectation and variance of aX + b

$$E(aX+b) = a\,E(X) + b$$
$$V(aX+b) = a^2 V(X)$$

❖ Standardized random variable

If X has mean μ and standard deviation σ,

$$Z = \frac{X - \mu}{\sigma}$$

is called a standardized random variable with

$E(Z) = 0$, $V(Z) = 1$

4.2 Random variable and distribution

[Example] The mean of a mid-term score on statistics was 60 points, and the variance was 100. To adjust the score, the professor is thinking of the alternatives. Find the mean and variance of each alternative.

1) Add 20 points to each student's score.
2) Each student's score is multiplied by 1.4.
3) Multiply each student's score by 1.2 and add 10 points.

<Answer>

- X is the mid-term score, E(X) = 60 and V(X) = 100.

1) The mean and variance of X + 20 are as follows.

$$E(X + 20) = E(X) + 20 = 60 + 20$$
$$V(X + 20) = V(X) = 100$$

2) The mean and variance of 1.4X are as follows.

$$E(1.4X) = 1.4\ E(X) = 1.4 \times 60 = 84$$
$$V(1.4X) = 1.4^2\ V(X) = 1.96 \times 100 = 196$$

3) The mean and variance of 1.2X + 10 are as follows.e 1.4X.

$$E(1.2X + 10) = 1.2\ E(X) + 10 = 1.2 \times 60 + 10 = 82$$
$$V(1.2X + 10) = 1.2^2\ V(X) = 1.44 \times 100 = 144$$

4.2 Random variable and distribution

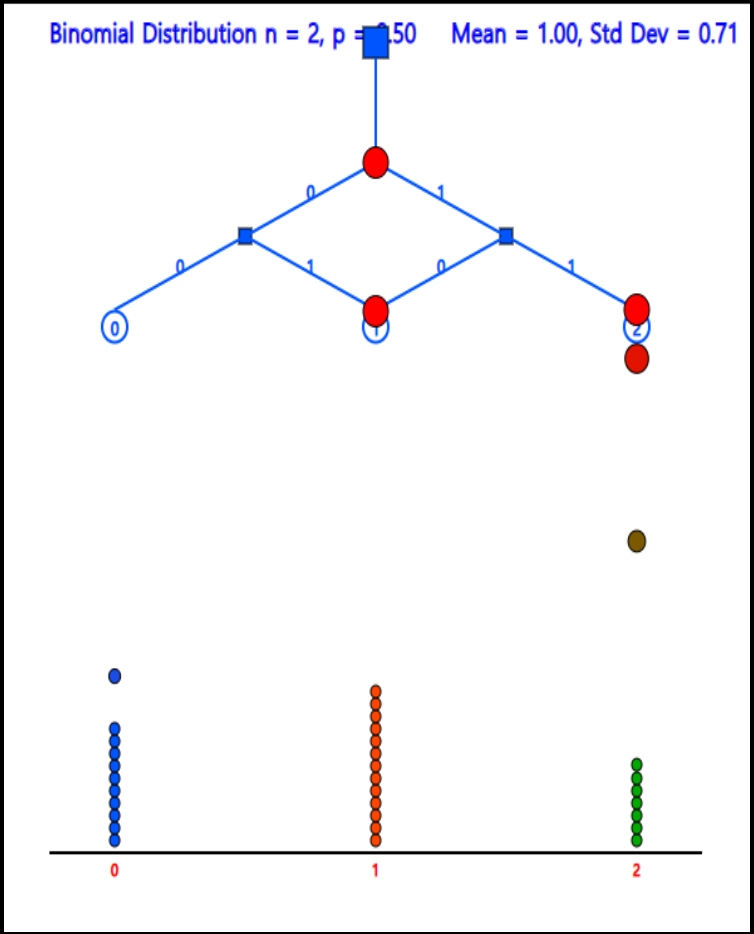
❖ Binomial Distribution

- Similar examples, such as tossing a coin several times and counting the number of heads, are observed around us.
  - Products are inspected and classified as defective or good.
  - Ask one voter about the pros and cons of a particular candidate.
- These experiments have two possible outcomes, such as {defective, normal}, {pros, cons}. However, the probability of outcomes in each experiment is different.
- These experiments are specifically called the Bernoulli trial, and one outcome in the two is referred to as 'success' and the other as 'failure'.
- Bernoulli trials are repeated, and the number of ‘success’ is counted.
  - Throw a coin five times and examine the number of heads.
  - Inspect 100 products and count the number of defective products.
  - Count the number of voters in favor of a candidate among 50 voters.

4.2 Random variable and distribution

❖ Binomial Distribution

- The 'counting of success' when performing this independently repeated Bernoulli trial with the same probability of success is called a binary random variable, and its distribution is called a binomial distribution.



4.2 Random variable and distribution

❖ Binomial Distribution

☞ Binomial Distribution

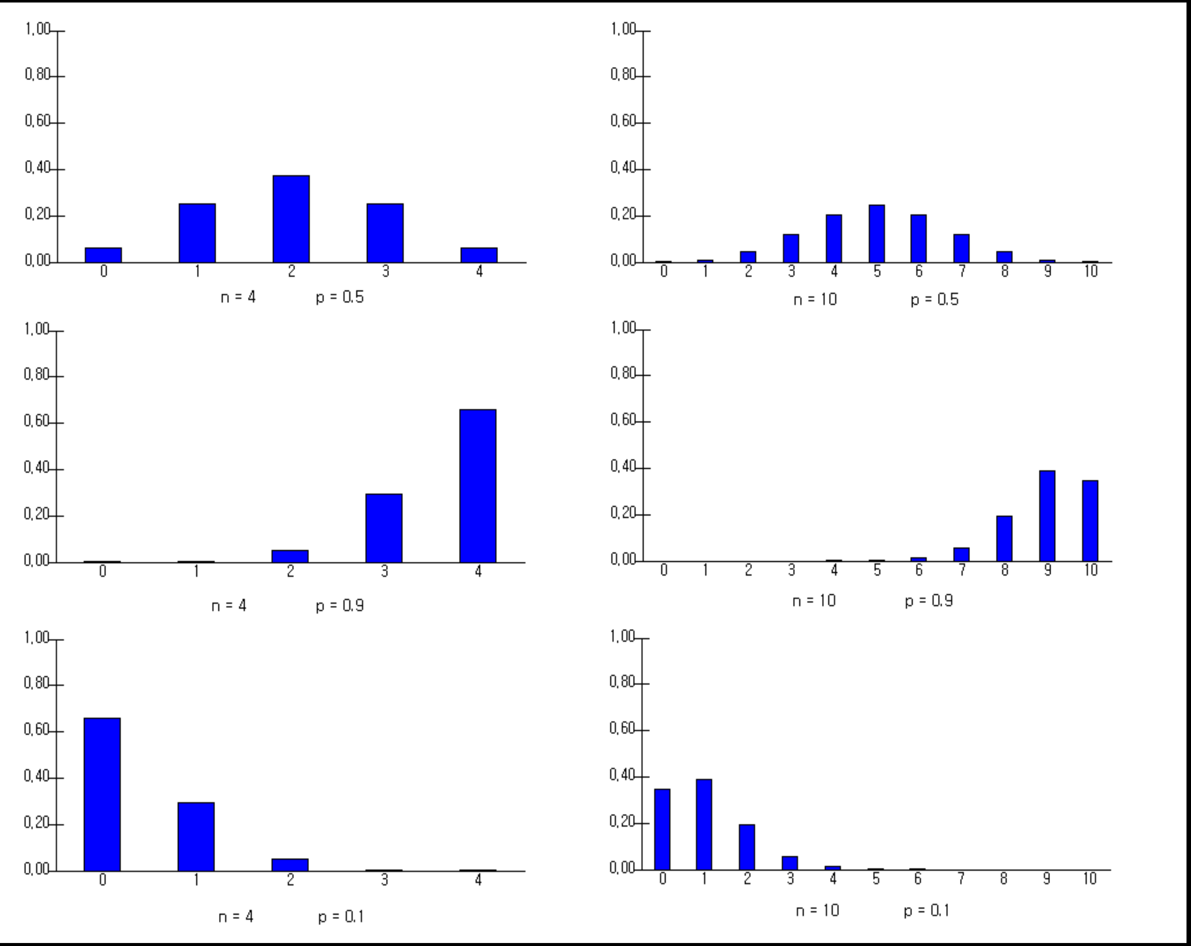
If the probability of success is p in a Bernoulli trial and the trial is repeated n times independently, the probability distribution function that the random variable X = 'the number of success' is x is as follows. It is called a binomial distribution and denoted as $B(n,p)$.

$$f(x) = {}_nC_x \, p^x \, (1-p)^{n-x} \quad , \qquad x = 0,1,2,\ldots,n$$

The expectation and variance of the binomial distribution are E(X) = np, V(X) = $np(1-p)$.

4.2 Random variable and distribution

❖ Binomial Distribution



4.2 Random variable and distribution

❖ Binomial Distribution

[Example] Past experience shows that a salesperson from an insurance company has a 20% chance of meeting a customer and insuring that person. The salesperson is scheduled to meet 10 customers this morning. Calculate the following probabilities.

1) What is the probability that three customers will get insurance?
2) What is the probability that two or more customers will get insurance?
3) How many people on average would sign up?
   And its standard deviation?

4.2 Random variable and distribution

❖ Binomial Distribution

- This is a Binomial distribution when n = 10, p = 0.2.

1) The probability that three customers will get insurance is as follows.

$$P(X=3) = {}_{10}C_3(0.2)^3(1-0.2)^{10-3} = 0.2013$$

2) The probability that two or more customers will get insurance may use the complement event as follows.

$$\begin{aligned} P(X \geq 2) &= 1 - P(X=0) - P(X=1) \\ &= 1 - {}_{10}C_0(0.2)^0(1-0.2)^{10} - {}_{10}C_1(0.2)^1(1-0.2)^{10-1} \\ &= 1 - 0.1074 - 0.2684 = 0.6242 \end{aligned}$$

3) Expectation and standard deviation are as follows..

$E(X) = np = 10 \times 0.2 = 2$

$V(X) = np(1-p) = 10 \times 0.2 \times 0.8 = 1.6$

Standard deviation = $\sqrt{1.6}$ = 1.265

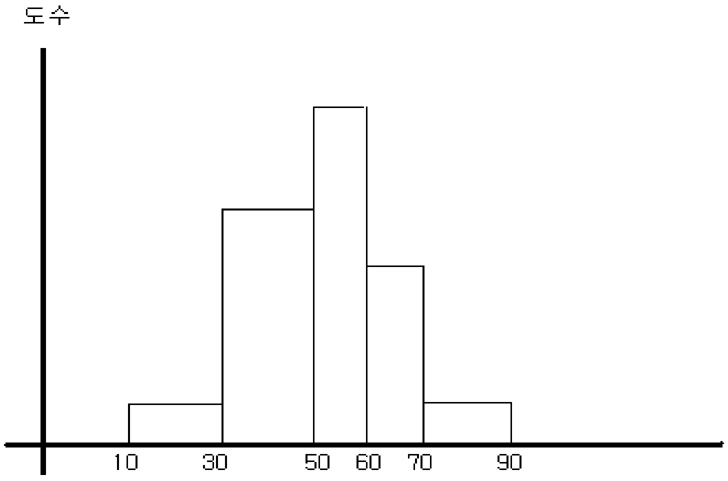
4.2 Random variable and distribution

❖ Normal Distribution

- Consider a statistical experiment that measures how long it takes for an office worker to get to work from home.
  => Commuting time usually takes about 30 minutes
- Define a random variable X as the 'time to work place'.
  => Infinite numbers of possible values for X
  => It is called a continuous random variable.
- The probability of an interval is of interest.
  =>What is the probability of X between 25 and 35 minutes?

4.2 Random variable and distribution

❖ Normal Distribution

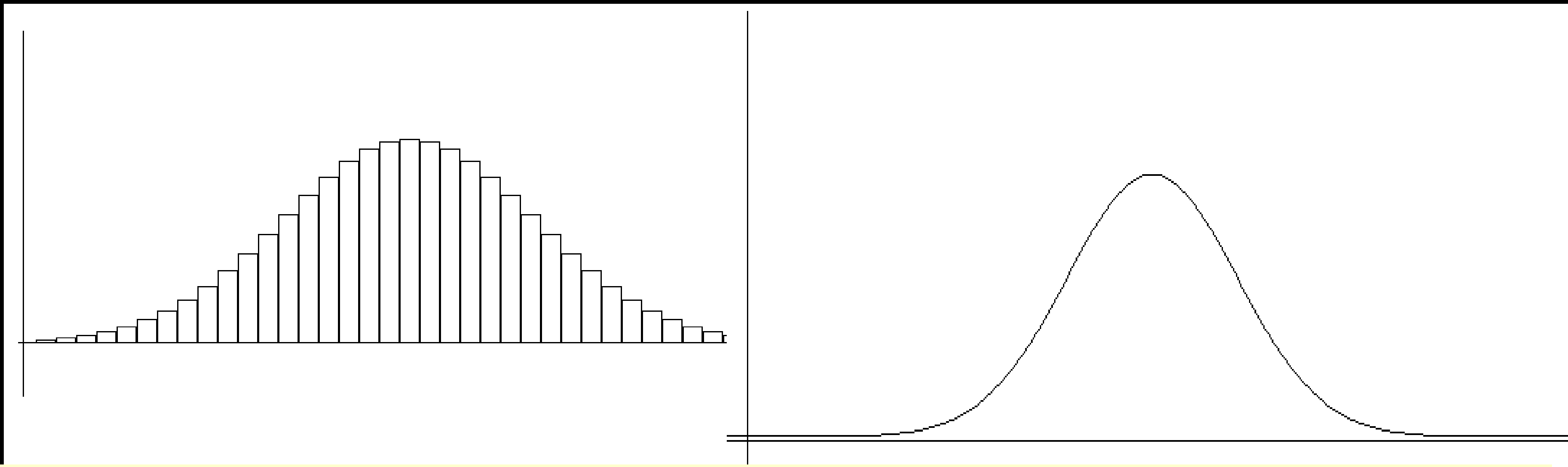


Interval ($a \le X < b$)	Frequency	Probability
$10 \le X < 30$	5	5/100
$30 \le X < 50$	30	30/100
$50 \le X < 60$	40	40/100
$60 \le X < 70$	20	20/100
$70 \le X < 90$	5	5/100
Total	100	1

- $P(30 \le X < 60) = 30/100 + 40/100 = 70/100$

4.2 Random variable and distribution

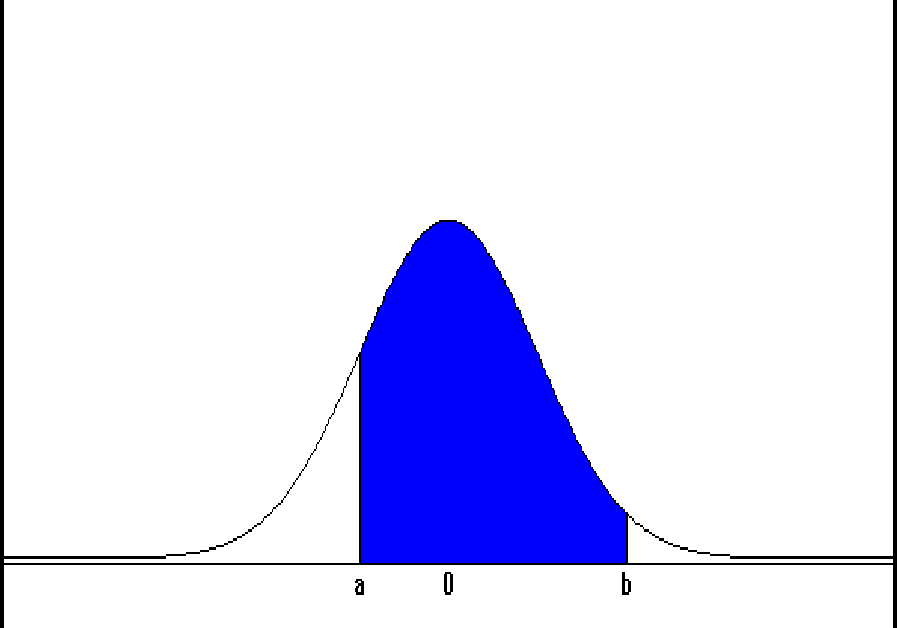
❖ Normal Distribution



- If you increase the number of data and close to zero the width of the interval, this histogram will be approximated to the continuous.
- It is called a probability distribution function of a continuous random variable.

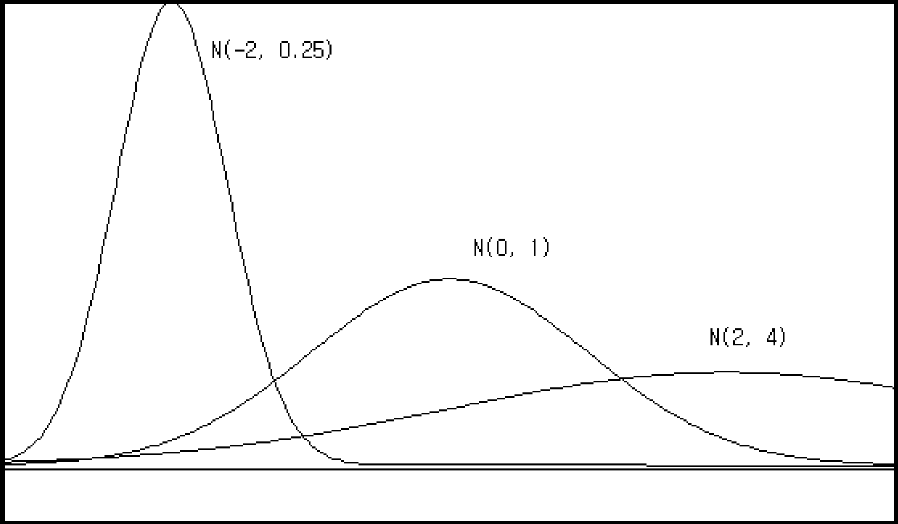
4.2 Random variable and distribution

❖ Normal Distribution



$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \qquad -\infty < x < \infty$$

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$



$$P(a < X < b) = \int_a^b f(x)dx$$

4.2 Random variable and distribution

❖ Normal Distribution

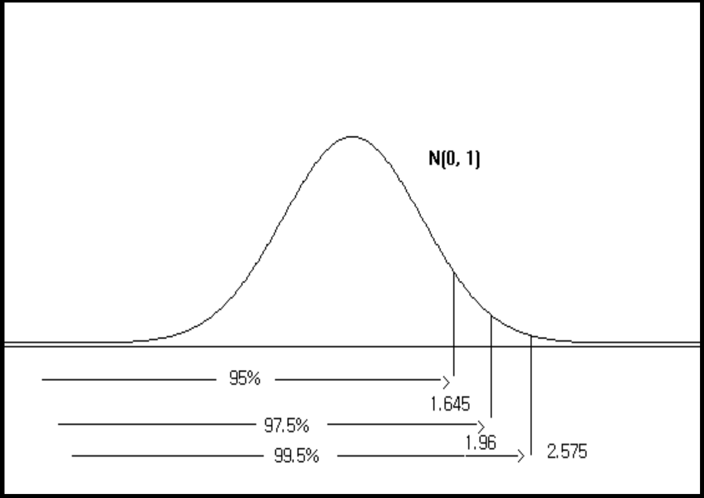
1) It is a continuous function in the shape of a bell.
2) It is symmetrical with respect to the mean μ.
3) There are an infinite number of normal distributions according to the value of μ and σ.
4) The probability of interval $[\mu - \sigma, \mu + \sigma]$ is 0.68, the probability of interval $[\mu - 2\sigma, \mu + 2\sigma]$ is 0.95, and the probability of interval $[\mu - 3\sigma, \mu + 3\sigma]$ is 0.997.
   => Most of values located in $\mu \pm 3\sigma$, few values outside

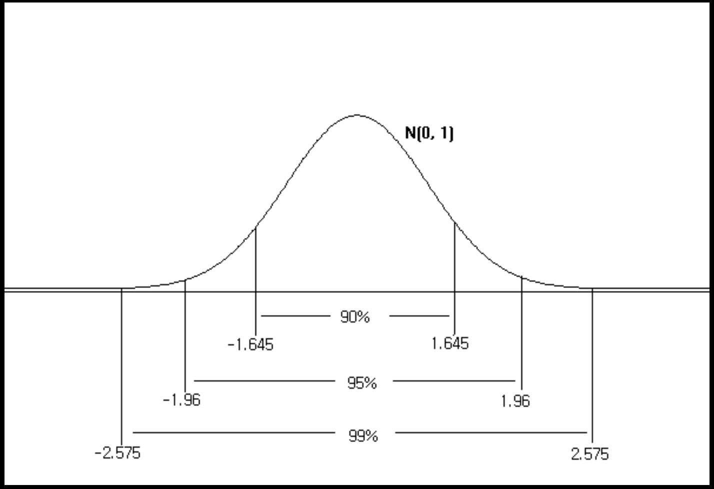
4.2 Random variable and distribution

❖ Normal Distribution

☞ When X is a normal random variable with a mean μ and variance σ^2, $(X-\mu)/\sigma$ follows the standard normal distribution. Therefore, the probability P(a< X <b) of the interval $[a,b]$ of X is as follows.

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$





4.2 Random variable and distribution

❖ Normal Distribution

Example 4.2.7 If mid-term scores (X) of a statistics course follow a normal distribution with the average of 70 points and the standard deviation of 10, calculate the following probabilities. Check the calculated values using 『eStatU』.

1) $P(X < 94.3)$
2) $P(X > 57.7)$
3) $P(57.7 < X < 94.3)$

Answer

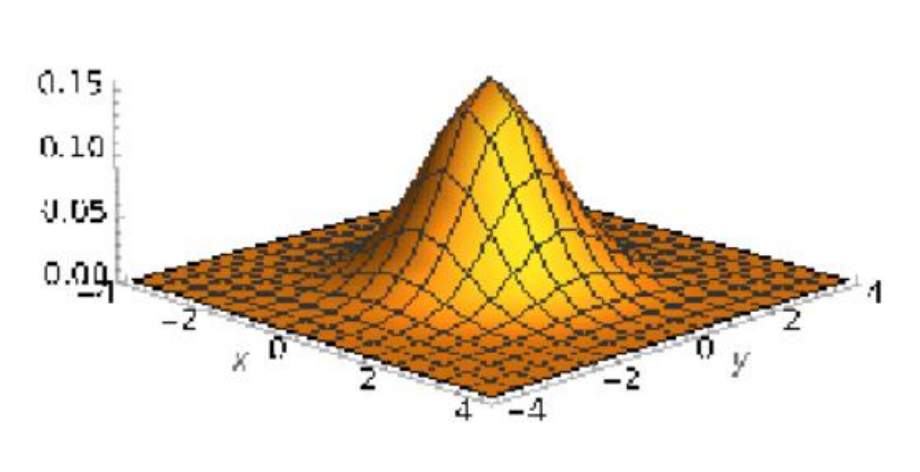
Using the transformation to the standardized normal random variable, probability calculations are as follows:

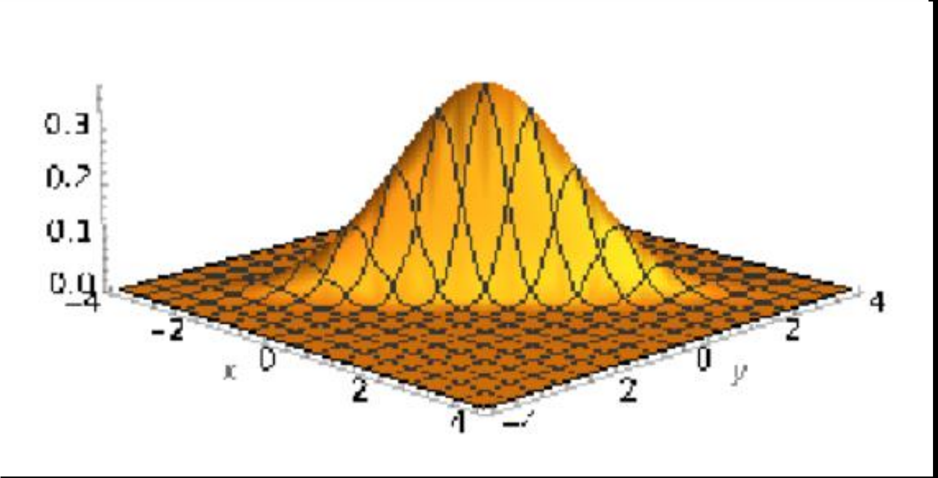
1) $P(X < 94.3) = P(\frac{X-70}{10} < \frac{94.3-70}{10}) = P(Z < 2.43) = 0.9925$
2) $P(X > 57.7) = P(\frac{X-70}{10} > \frac{57.7-70}{10}) = P(Z > -1.23 = 0.8907$
3) $P(57.7 < X < 94.3) = P(\frac{57.7-70}{10} < \frac{X-70}{10} < \frac{94.3-70}{10}) = P(-1.23 < Z < 2.43) = 0.8832$

4.3 Multivariate normal distribution

❖ Bivariate normal distribution

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} exp\left\{ \frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right] \right\}$$





❖ Multivariate normal distribution

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{m}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$$

4.4 Estimation of a distribution

❖ Maximum likelihood estimation

- Likelihood distribution

$$f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n : \boldsymbol{\theta}) = f(\boldsymbol{x}_1 : \boldsymbol{\theta}) f(\boldsymbol{x}_2 : \boldsymbol{\theta}) \cdots f(\boldsymbol{x}_n : \boldsymbol{\theta})$$

$$\frac{\partial f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n : \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

- If f(x) is a multivariate normal distribution

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{m}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\}$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$$

$$\hat{\boldsymbol{\Sigma}} = \frac{n-1}{n} \boldsymbol{S}$$

Summary

- Probability:
  - Classical and relative frequency definition
  - Addition and multiplication rules
  - Bayes theorem
- Random variable and probability distribution:
  - Random variable is a function from a sample space to real number
  - Binomial distribution, normal distribution
  - Multiivariate normal distribution
- Estimation of a distribution:
  - Maximum likelihood estimation

Thank you !!!