

# Ch02\_Basic Probability Theory

---

**Prof. Cheolsoo Park**



# Basic Statistics

---

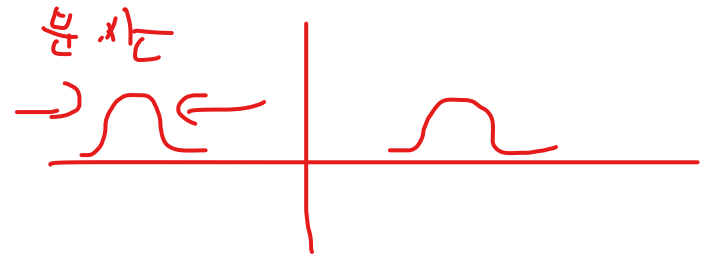
- Statistics is a study about an objective decision-making process through the data, converted to information
- In pattern recognition, various statistics method is utilized, and analysis based on prior knowledge is necessary
- Under the uncertainty, various probability theory is applied to classify unknown data into several categories



# Basic Statistics

- Terminology of statistics

- A population
  - Entire data for the analysis
- Samples
  - Some part of data among all
- Sample distribution
  - Distributions of sample data



- Statistical parameters

- Mean (1<sup>st</sup> order)
  - A point of center for gravity
- Variance (2<sup>nd</sup> order)
  - How much scattered numerical data is

- Standard deviation

- Square root of variation

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Eq. 1 numerical  
expression of mean

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Eq. 2 numerical  
expression of variance

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Eq. 3 numerical  
expression of STD



# Basic Statistics

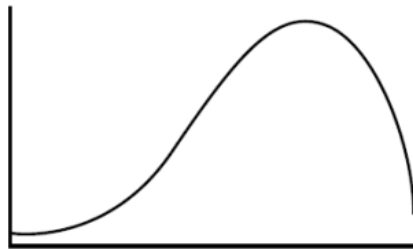
- Skewness (3<sup>rd</sup> order)

- A measure of the asymmetry of the data probability distribution
- Negative skew : the left tail is longer; the mass of the distribution is concentrated on the right of the figure
- Positive skew : the right tail is longer; the mass of the distribution is concentrated on the left of the figure
- If the distribution is symmetry, skewness is 0

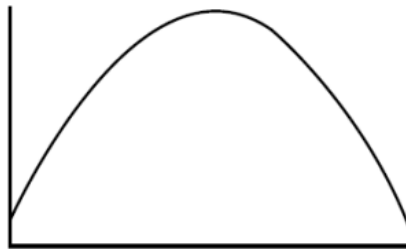
$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left( \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}}$$

where  $\bar{x}$  is the [sample mean](#),  
 $s$  is the [sample standard deviation](#),  
and the numerator  $m_3$  is the sample third central [moment](#).

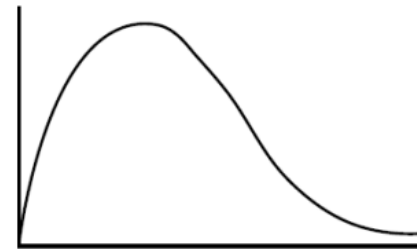
Eq. 6 skewness



(a) Negative Skew



(b) Skewness = 0



(c) Positive Skew

# Basic Statistics

- Kurtosis (4<sup>th</sup> order)
  - Kurtosis is a descriptor of the shape of a probability distribution

$$Kurtosis = \frac{\mu_4}{\sigma^4} = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{\left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right]^2}$$

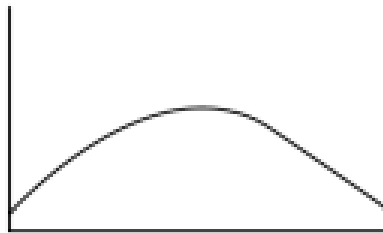
where  $\mu_4$  is the fourth moment about the mean  
and  $\sigma$  is the standard deviation.

Eq. 7 kurtosis

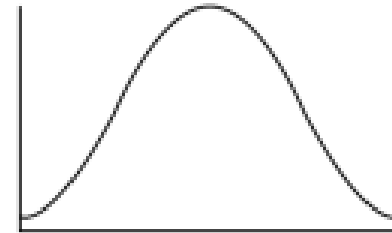
$$Kurtosis = \begin{cases} 3 \text{ less } (< 3) : & \text{platykurtic} \\ 3 (= 3) & : \text{normal} \\ 3 \text{ more } (> 3) : & \text{leptokurtic} \end{cases}$$



(a)  $K = 3$  less



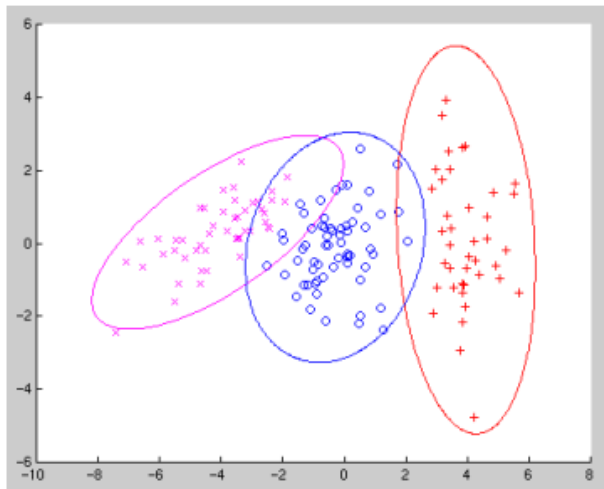
(b)  $K = 3$



(c)  $K = 3$  more

# Covariance and Correlation

- Covariance
  - How much related two random variables are with changing together
  - Zero covariance means no correlation between two



Example of 3 random variables

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \\ C(x, y) &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n-1} \left( \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y} - \sum_{i=1}^n \bar{x} y_i + \sum_{i=1}^n \bar{x} \bar{y} \right) \\ &= \frac{1}{n-1} \left( \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y} - \sum_{i=1}^n \bar{x} y_i + n \bar{x} \bar{y} \right)\end{aligned}$$

Eq. 4 covariance of bivariate

# Covariance and Correlation

- Correlation

- Normalised covariance by the standard deviation of two variables

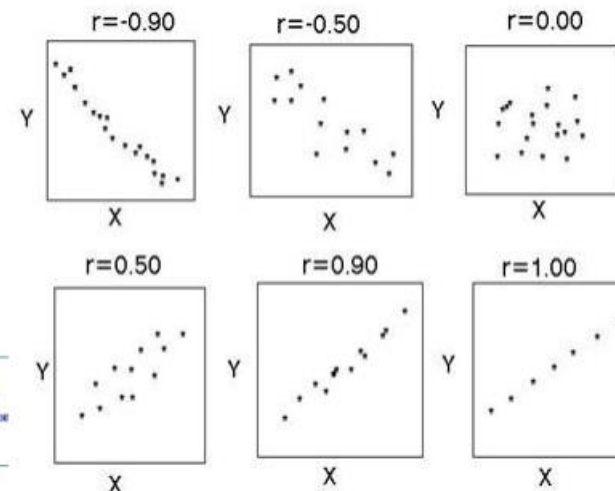
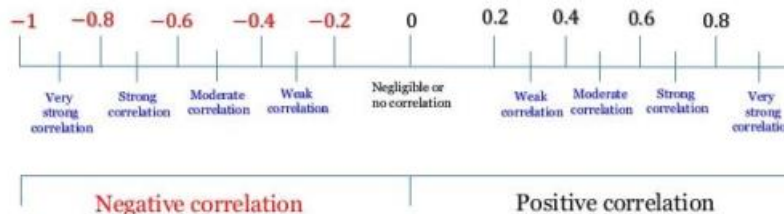
- Varied from -1 to 1

$$-1 \leq \rho_{xy} \leq 1$$

$$\rho = \frac{C(x, y)}{\sigma_x \sigma_y} = \frac{C(x, y)}{\sqrt{V_{(x)}} \sqrt{V_{(y)}}} \quad * \sigma_x, \sigma_y \text{ is STD of variable } x, y$$

## Correlation Coefficient Interpretation Guideline

The correlation coefficient (r) ranges from -1 (a perfect negative correlation) to 1 (a perfect positive correlation). In short,  $-1 \leq r \leq 1$ .



linearity for 'r'

# Basic Statistics

$x_1$ : hours spending in library per a week

$x_2$ : Test score of machine learning exam

$x_3$ : days of class absence

$x_1$	$x_2$	$x_3$
35	11	0.5
5	12	0
0	5	10
20	9	4.5
15	8	5

①  $x_1$ ,  $x_2$  correlation coefficient

$$\rho_{12} = \frac{(35-15)(11-9) + (5-15)(12-9) + (0-15)(5-9)}{3\sqrt{187.5}\sqrt{7.5}} = 0.4667$$

	$x_1$	$x_2$	$x_3$
평균	15	9	4
분산	187.5	7.5	16.37

②  $x_2$ ,  $x_3$  correlation coefficient

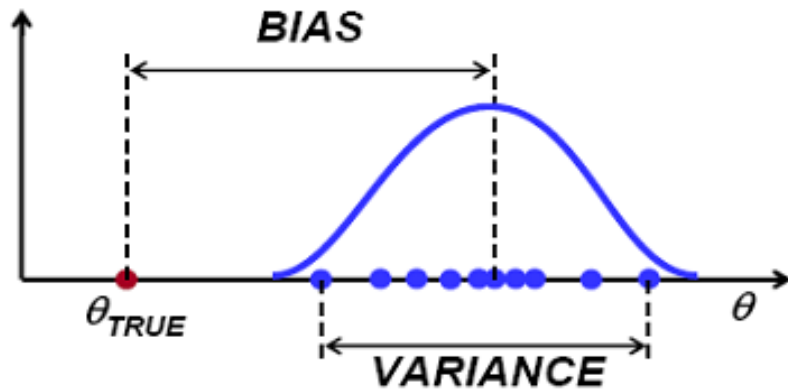
$$\rho_{23} = \frac{(11-9)(0.5-4) + (12-9)(0-4) + (5-9)(10-4) + (8-9)(5-4)}{4\sqrt{7.5}\sqrt{16.37}} = -0.9926$$





# Bias and Variance

- Bias and Variance



Ex) designing optimal size of mobile phone  
corresponding to users' preference



(a) High bias, Low variance



(b) Low bias, High variance



(c) High bias, High variance



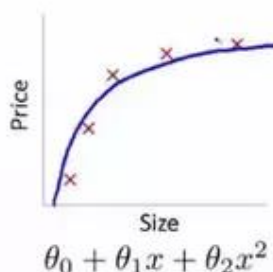
(d) Low bias, Low variance

# Bias and Variance

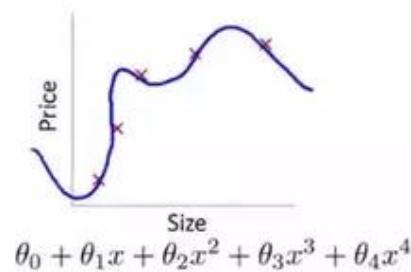
- $\text{Error}(X) = \text{noise}(X) + \text{bias}(X) + \text{variance}(X)$ 
  - noise : irreducible error due to its intrinsic property of data
  - bias and variance : reducible error depending on the model
- Bias and Variance Tradeoff
  - The bias is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (**underfitting**).
  - The variance is an error from sensitivity to small fluctuations in the training set. High variance can cause **overfitting**: modeling the random noise in the training data, rather than the intended outputs.



High bias  
(underfit)



“Just right”



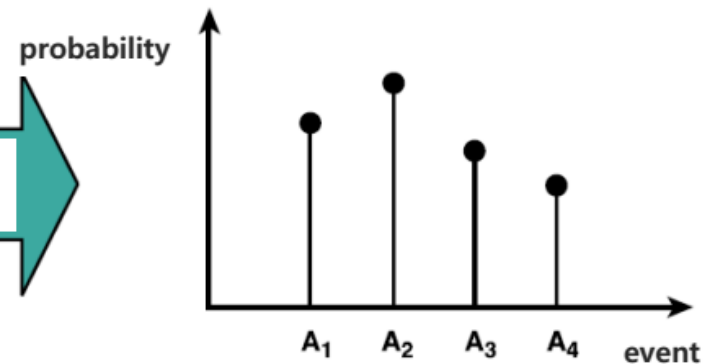
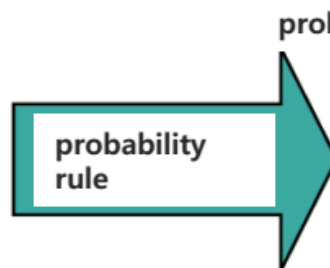
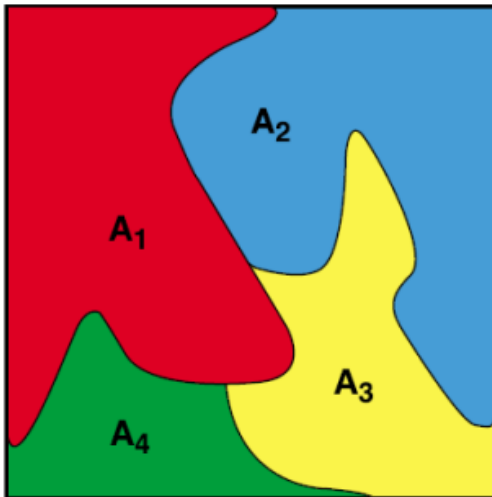
High variance  
(overfit)



# Basic Probability

- Probability
  - Probability is the measure of the likelihood that an event will occur through statistic phenomenon
  - Probability is quantified as a number between 0 and 1
  - Probability rule
    - Probability rule is assigned probability of each events in random trial
    - Sample space of random trial, 'S' is all sets

sample space



# Basic Probability

---

- Numerical Probability
  - When the possibility is equal that each event of sample, 'S', will occur, it is called numerical probability of event A about  $n(A)/n(S)$
  - Think about the case of 'dice'
- Statistic Probability
  - Not only the same probability of natural phenomenon or society phenomenon is rare, but also probability of each event has uncertainty
  - In this case, experiment is performed several times and event probability can be assumed through relative frequency
  - If event occur n of r, relative frequency can be expressed  $n/r$



# Basic Probability

---

- Event and exclusive events
  - Interesting event can be occurred by individual or multiplicative (complex) factors
    - ➔ this complex results are called event
  - If event A and B can not be occurred simultaneously, they are called exclusive events



# Basic Probability

---

- Sample space and probability space
  - Sample space of an experiment or random trials is the set of all possible outcomes or results of the experiment
  - Elements of sample space is called 'sample point'
  - Partial set of sample space is named as 'event', and only one sample point event is named as 'fundamental event'
  - Probability space is set of all result that can be occurred



# Basic Probability

- Theorem of probability

theorem  $0 \leq P[A_i]$

theorem  $P[S] = 1$

theorem **If**  $A_i \cap A_j = \emptyset$  ,  $P[A_i \cup A_j] = P[A_i] + P[A_j]$

- Characteristics of probability

| 1:  $P[A^c] = 1 - P[A]$

| 2:  $P[A] \leq 1$

| 3:  $P[\emptyset] = 0$

| 4: Given  $\{A_1, A_2, \dots, A_N\}$  , if  $\{A_i \cap A_j = \emptyset \ \forall i, j\}$ , then  $P[\bigcup_{k=1}^N A_k] = \sum_{k=1}^N P[A_k]$

| 5:  $P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1 \cap A_2]$

| 6:  $P[\bigcup_{k=1}^N A_k] = \sum_{k=1}^N P[A_k] - \sum_{j < k} P[A_k \cap A_j] + \dots + (-1)^{N+1} P[A_1 \cap A_2 \cap \dots \cap A_N]$

| 7: If  $A_1 \subset A_2$  ,  $P[A_1] \leq P[A_2]$



- Conditional Probability

- $$P[A|B] = \frac{P[A \cap B]}{P[B]} \quad \text{if } P[B] > 0$$



# Basic Probability

- Joint Probability

- Event A and B are occurred simultaneously, called ‘product rule’
- If the event A and B are independent each other,  $P(A|B)=P(A)$  and  $P(A) \times P(B) = P(A \cap B)$

example of joint probability

Event A  
- even number case of dice  
- 2, 4, 6

Event B  
- multiple of 3  
- 3, 6

Event A	X	Event B
---------	---	---------

$$= \frac{3}{6} \times \frac{2}{6} = \frac{1}{6}$$

Ex)  $P(A)$  = lunch for today,  $P(B)$  = lunch for tomorrow

$P(A \cap B) = P(A) \times P(B)$  or  $P(A) \times P(B|A)$  depending on your behavior

Ex) Restaurant event R:

$P(R1)$  and  $P(R2)$  are the chance for restaurant 1 and 2

$P(Tg)$  and  $P(Tb)$  are good and bad outcomes in a toilet

Using  $P(Tg|R1)$ ,  $P(Tb|R1)$ ,  $P(Tg|R2)$  and  $P(Tb|R2)$ ,

calculate  $P(Tg, R1)$ ,  $P(Tb, R1)$ ,  $P(Tg, R2)$  and  $P(Tb, R2)$

\* Possible for  $P(Tg|R1) = P(Tg|R2) = P(Tg)$ , when they are independent



# Basic Probability

- Chain Rule

- The chain permits the calculation of any member of the joint distribution of a set of random variables using only conditional probabilities
- The chain rule produces this product of conditional probabilities:

$$P(A_1, A_2, A_3, A_4, \dots, A_n) \\ = P(A_1 | A_2, A_3, A_4, \dots, A_n) \times P(A_2 | A_3, A_4, \dots, A_n) \times \dots \times P(A_{n-1} | A_n) \times P(A_n)$$

Ex) Urn 1 has 1 black ball and 2 white balls. Urn 2 has 1 black ball and 3 white balls. We will choose the urn 1 and 2 with the same probability.

What's the probability of choosing the first urn and a white ball from it.



# Basic Probability

- Chain Rule

- The chain permits the calculation of any member of the joint distribution of a set of random variables using only conditional probabilities
- The chain rule produces this product of conditional probabilities:

$$P(A_1, A_2, A_3, A_4, \dots, A_n) \\ = P(A_1 | A_2, A_3, A_4, \dots, A_n) \times P(A_2 | A_3, A_4, \dots, A_n) \times \dots \times P(A_{n-1} | A_n) \times P(A_n)$$

Ex) Urn 1 has 1 black ball and 2 white balls. Urn 2 has 1 black ball and 3 white balls. We will choose the urn 1 and 2 with the same chance.

What's the probability of choosing the first urn and a white ball from it.

$$\Rightarrow P(B | 1) = 1/3, P(W | 1) = 2/3, P(B | 2) = 1/4, P(W | 2) = 3/4$$

$$P(1, W) = P(1) \times P(W | 1) = 1/2 \times 2/3$$



# Basic Probability

- Marginal Probability
  - The marginal distribution of a subset of a collection of random variables is the probability distribution of the variables contained in the subset
  - It gives the probabilities of various values of the variables in the subset without reference to the values of the other variables

	Y \ X	Mon	Tue	Wed	Thurs	$p_Y(Y) \downarrow$
		$x_1$	$x_2$	$x_3$	$x_4$	
Football	$y_1$	$\frac{4}{32}$	$\frac{2}{32}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{8}{32}$
Basketball	$y_2$	$\frac{2}{32}$	$\frac{4}{32}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{8}{32}$
Baseball	$y_3$	$\frac{2}{32}$	$\frac{2}{32}$	$\frac{2}{32}$	$\frac{2}{32}$	$\frac{8}{32}$
Swim	$y_4$	$\frac{8}{32}$	0	0	0	$\frac{8}{32}$
	$p_X(X) \rightarrow$	$\frac{16}{32}$	$\frac{8}{32}$	$\frac{4}{32}$	$\frac{4}{32}$	$\frac{32}{32}$

Joint and marginal distributions of a pair of discrete, random variables  $X, Y$  having nonzero mutual information  $I(X; Y)$ . The values of the joint distribution are in the  $4 \times 4$  square, and the values of the marginal distributions are along the right and bottom margins.



# Bayes' Theorem

- Bayes' theorem



$$P[B_j | A] = \frac{P[A \cap B_j]}{P[A]} = \frac{P[A | B_j] \cdot P[B_j]}{\sum_{k=1}^N P[A | B_k] \cdot P[B_k]}$$

$$P[\omega_j | \mathbf{x}] = \frac{P[\mathbf{x} | \omega_j] \cdot P[\omega_j]}{\sum_{k=1}^N P[\mathbf{x} | \omega_k] \cdot P[\omega_k]} = \frac{P[\mathbf{x} | \omega_j] \cdot P[\omega_j]}{P[\mathbf{x}]}$$

※  $\omega_j$ :  $j$ -th class

$\mathbf{x}$ : Feature Vector

- $P[\omega_j]$  : the probabilities (prior probability)
- $P[\omega_j | \mathbf{x}]$  : a conditional probability, is the probability of observing event A given that B is true.
- $P[\mathbf{x} | \omega_j]$  : the probability of observing event B given that A is true.
- $P[\mathbf{x}]$  : prior probability of X.  $P(B) = \sum P(B|A)P(A)$



# 21

INSPIRED BY A TRUE STORY.

© 2007 Sony Pictures Digital Inc. All rights reserved.

THIS FILM IS  
NOT YET RATED

COLUMBIA  
PICTURES

SONY  
PICTURES  
DIGITAL



# Bayes' Theorem

- Example of Bayes' theorem

$P(S)$  : probability of single,  $P(M)$  : probability of married

$P(C1)$  : probability of area 1,  $P(C2)$  : probability of area2

Goal :  $P(S|C1)$  and  $P(S|C2)$

$P(S)$  and  $P(M)$  : from the population statistics

$P(C1|S)$ ,  $P(C2|S)$ ,  $P(C1|M)$  and  $P(C2|M)$  : survey from people

$$\begin{aligned} \Rightarrow P(S|C1) &= \frac{P(S)P(C1|S)}{P(C1)} = \frac{P(S)P(C1|S)}{\sum_{X=S,M} P(C,X)} \\ &= \frac{P(S)P(C1|S)}{P(S)P(C1|S) + P(M)P(C1|M)} \end{aligned}$$

