Machine Learning & Pattern Recognition

SONG Xuemeng

sxmustc@gmail.com

http://xuemeng.bitcron.com/

Review of Probability

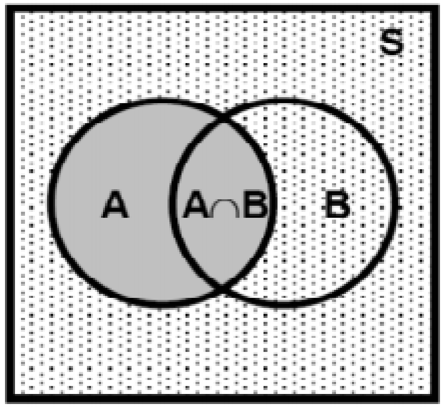
- Probability
  - Axioms and properties
  - Conditional probability
  - Law of total probability
  - Bayes theorem
- Random Variables
  - Discrete
  - Continuous
- Random Vectors
- Gaussian Random Variables

Basics of Probability

- **Definitions (informal)**
  - **Probabilities** are numbers assigned to events that indicate “how likely” it is that the event will occur when a random experiment is performed.
  - **A probability law** for a random experiment is a rule that assigns probabilities to the events in the experiment.
  - **The sample space S** of a random experiment is the set of all possible outcomes.
- **Axioms of probability**
  - Axiom 1: $0 \leq P[A]$
  - Axiom 2: $P(S) = 1$
  - Axiom 3: if $A_i \cap A_j = \emptyset$, then $\mathrm{P}[A_i \cup A_j] = \mathrm{P}[A_i] + \mathrm{P}[A_j]$

Basics of Probability

- $P[A^C] = 1 - P[A]$
- $P[A] \leq 1$
- $P[\emptyset] = 0$



- Given $\{A_1, A_2, \dots, A_N\}$, if $\{A_i \cap A_j = \emptyset, \forall i, j\} \Rightarrow P\left[\bigcup_{k=1}^{N} A_k\right] = \sum_{k=1}^{N} P[A_k]$
- $P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1 \cap A_2]$
- $P\left[\bigcup_{k=1}^{N} A_k\right] = \sum_{k=1}^{N} P[A_k] - \sum_{j<k}^{N} P\left[A_j \cap A_k\right] + (-1)^N P[A_1 \cap A_2 \cap \cdots \cap A_N]$
- If $A_1 \subset A_2$, then $P[A_1] \leq P[A_2]$

Conditional Probability

- If A and B are two events, the probability of event A when we already know that event B has occurred is defined by the relation

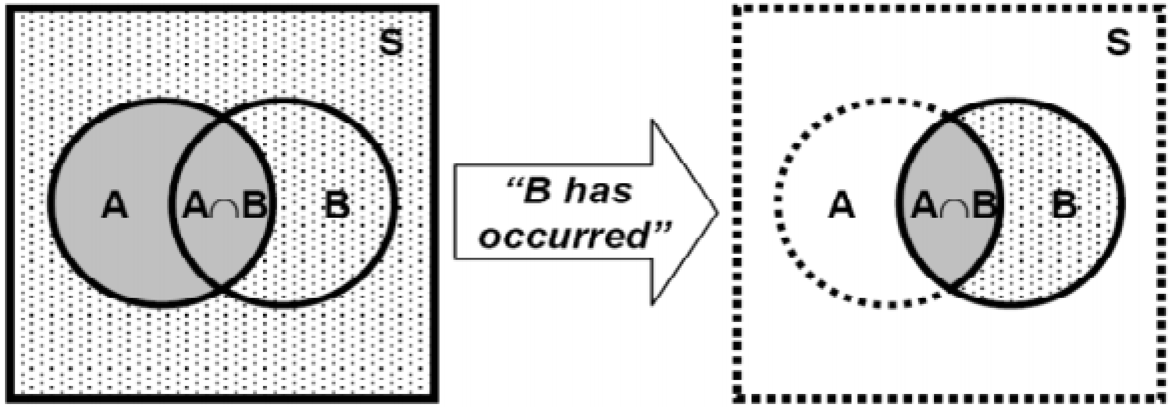
$$P[A|B] = \frac{P[A \cap B]}{P[B]} \text{ for } P[B] > 0 \qquad \text{(product rule)}$$

- This conditional probability $P[A \cap B]$ is read:
  - "The conditional probability of A conditioned on B" or simply
  - "The probability of A given B"

Conditional Probability

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

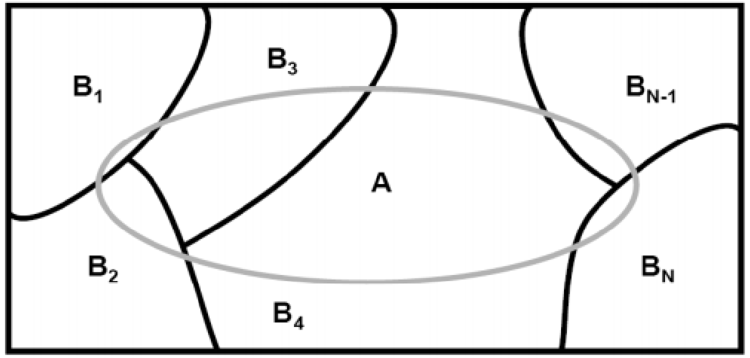
- **Interpretation**
  - The new evidence "B has occurred" has the following effects:
    - The original sample space S (the whole square) becomes B (the rightmost circle);
    - The event A becomes $A \cap B$.
  - $P[B]$ simply re-normalizes the probability of events that occur jointly with B.



Law of Total Probability

- Let $B_1, B_2, \ldots, B_N$ be mutually exclusive events whose union equals the sample space S. We refer to theses sets as a *partition* of S.
- An event A can be represented as:

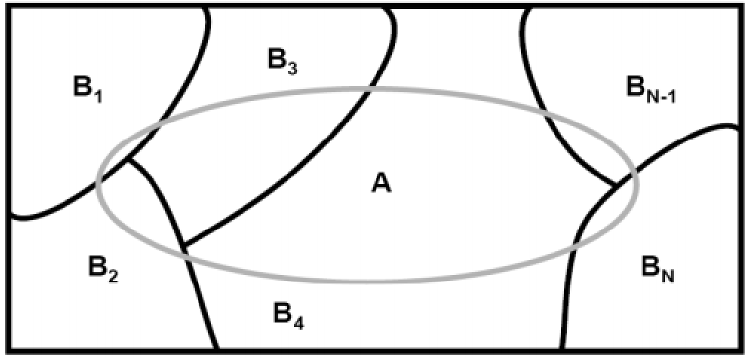
$$A = A \cap S = A \cap (B_1 \cup B_2 \cup \cdots \cup B_N)$$
$$= (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_N)$$



Law of Total Probability

- Let $B_1, B_2, \ldots, B_N$ be mutually exclusive events whose union equals the sample space S. We refer to theses sets as a *partition* of S.
- An event A can be represented as:

$$A = A \cap S = A \cap (B_1 \cup B_2 \cup \cdots \cup B_N)$$
$$= (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_N)$$



E.g., A: There is a traffic jam in Beijing.

B_1: It is a rainy day in Beijing.

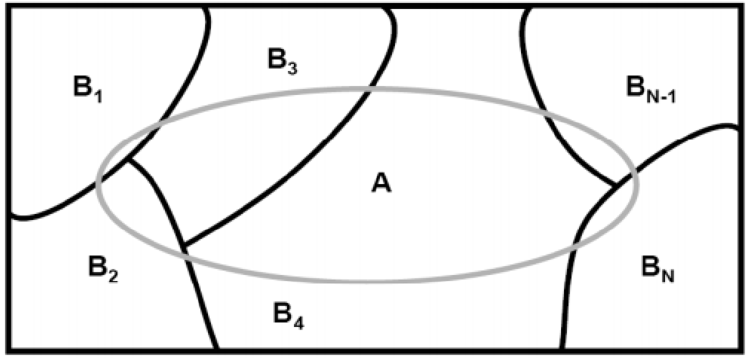
B_2: It is not a rainy day in Beijing.

$A \cap B_1$: There is a traffic jam on a rainy day in Beijing.

$A \cap B_2$: There is a traffic jam on a non-rainy day in Beijing.

Law of Total Probability

- Let $B_1, B_2, \ldots, B_N$ be mutually exclusive events whose union equals the sample space S. We refer to theses sets as a *partition* of S.
- An event A can be represented as:



$$A = A \cap S = A \cap (B_1 \cup B_2 \cup \cdots \cup B_N)$$
$$= (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_N)$$

E.g., A: A person is lying.

B_1: The person is a man.

B_2: The person is a woman.

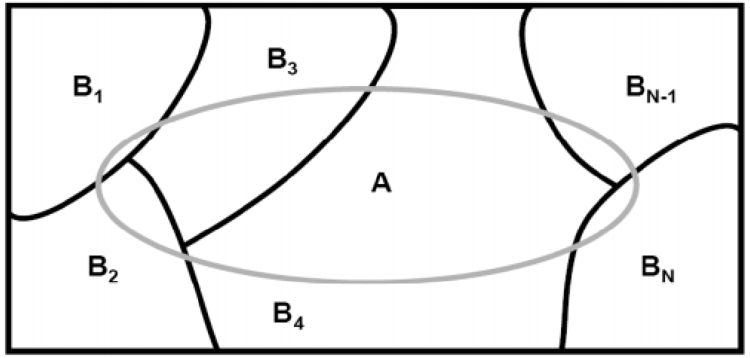
$A \cap B_1$: A man is lying.

$A \cap B_2$: A woman is lying.

Law of Total Probability

- Let $B_1, B_2, \ldots, B_N$ be mutually exclusive events whose union equals the sample space S. We refer to theses sets as a *partition* of S.
- An event A can be represented as:

$$A = A \cap S = A \cap (B_1 \cup B_2 \cup \cdots \cup B_N)$$
$$= (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_N)$$



E.g., A: The word “university” would appear in the document.

B_1: The document belongs to topic 1.

B_2: The document belongs to topic 2.

⋮

B_N: The document belongs to topic N.

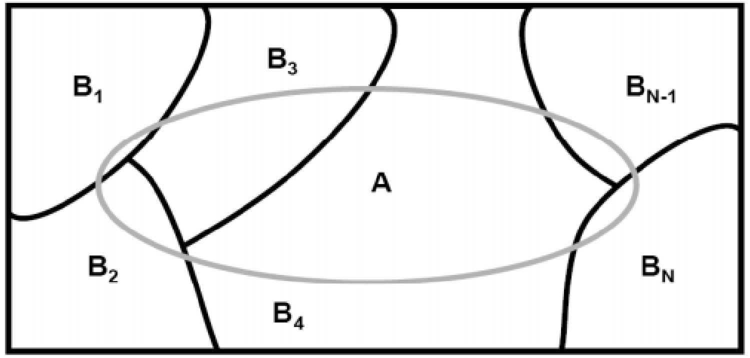
$A \cap B_i$: The word “university” would appear in a document belongs to topic i.

Assume that there are N topics in total and each document must belong to only one topic.

Law of Total Probability

- Let $B_1, B_2, \ldots, B_N$ be mutually exclusive events whose union equals the sample space S. We refer to theses sets as a *partition* of S.
- An event A can be represented as:

$$A = A \cap S = A \cap (B_1 \cup B_2 \cup \cdots \cup B_N)$$
$$= (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_N)$$



$$P[A] = P[A \cap B_1] + P[A \cap B_2] + \cdots + P[A \cap B_N]$$
$$= P[A|B_1]P[B_1] + \cdots + P[A|B_1]P[B_N] = \sum_{k=1}^{N} P[A|B_k]P[B_k]$$

Law of Total Probability

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

E.g., A: The word “university” would appear in the document.

B_1: The document belongs to topic 1.

B_2: The document belongs to topic 2.

$\vdots$

B_N: The document belongs to topic N.

Assume that there are N topics in total and each document must belong to only one topic.

$P[B_i]$: Probability that the document belongs to topic i.

$P[A|B_i]$: Probability that the word “university” would appear if the document belongs to topic i.

$P[A]$: Probability that the word “university” would appear in the document.

Bayes Theorem

- Given B_1, B_2,…, B_N, a partition of the sample space S. Suppose that event A occurs; what is the probability of event B_j?

Bayes Theorem

- Given $B_1, B_2, \dots, B_N$, a partition of the sample space S. Suppose that event A occurs; what is the probability of event B_j?

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

Bayes Theorem

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

Assume that there are N topics in total and each document must belong to only one topic.

E.g., A: The word “university” would appear in the document.

B_1: The document belongs to topic 1.

B_2: The document belongs to topic 2.

$\vdots$

B_N: The document belongs to topic N.

Bayes Theorem

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

Assume that there are N topics in total and each document must belong to only one topic.

E.g.,

A: The word "university" would appear in the document.

B_1: The document belongs to topic 1.

B_2: The document belongs to topic 2.

⋮

B_N: The document belongs to topic N.

If we know …

$P[B_i]$: Probability that the document belongs to topic i.

$P[A|B_j]$: Probability that word "university" would appear if the document belongs to topic j.

Bayes Theorem

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

Assume that there are N topics in total and each document must belong to only one topic.

E.g.,

A: The word “university” would appear in the document.

B_1: The document belongs to topic 1.

B_2: The document belongs to topic 2.

$\vdots$

B_N: The document belongs to topic N.

If we know …

$P[B_i]$: Probability that the document belongs to topic i.

$P[A|B_j]$: Probability that word “university” would appear if the document belongs to topic j.

Then we can get…

$P[B_j|A]$ = what is the meaning?

Bayes Theorem

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

Assume that there are N topics in total and each document must belong to only one topic.

E.g.,

A: The word "university" would appear in the document.

B_1: The document belongs to topic 1.

B_2: The document belongs to topic 2.

$\vdots$

B_N: The document belongs to topic N.

If we know …

$P[B_i]$: Probability that the document belongs to topic i.

$P[A|B_j]$: Probability that word "university" would appear if the document belongs to topic j.

Then we can get…

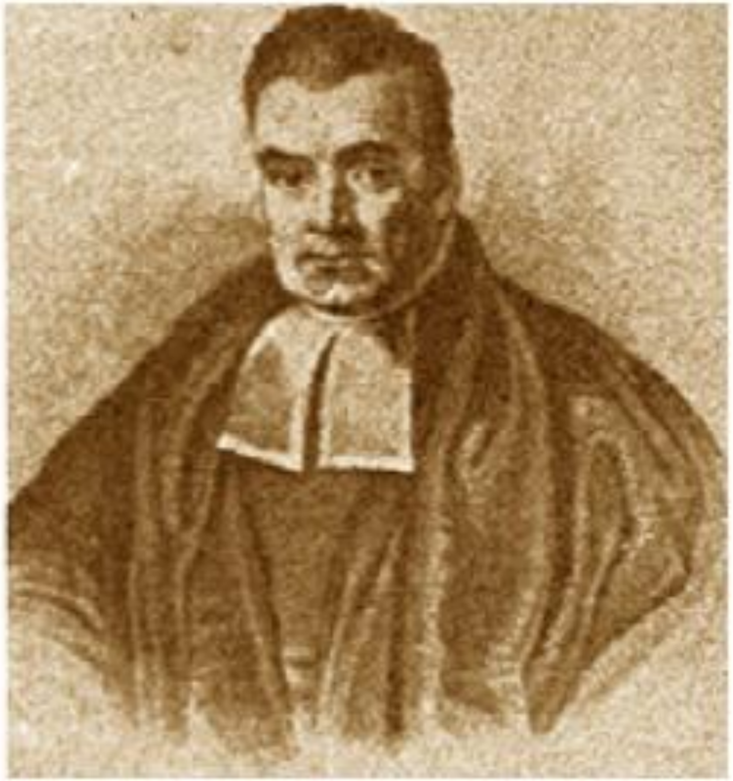
$P[B_j|A]$ = Probability that the document belongs to topic j if word "university" appears in it.

Bayes Theorem

- Given $B_1, B_2, \ldots, B_N$, a partition of the sample space S. Suppose that event A occurs; what is the probability of event B_j?

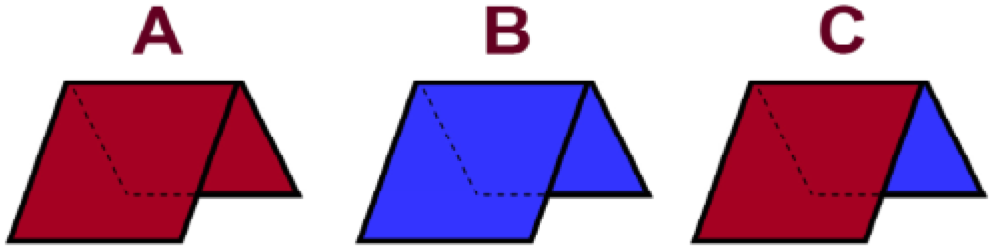
$$P[B_j|A] = \frac{P[B_j \cap A]}{P[A]} = \frac{P[A|B_j]P[B_j]}{\sum_{k=1}^{N} P[A|B_k]P[B_k]}$$

- This is known as Bayes Theorem or Bayes Rule, one of the most useful relations in probability and statistics.
  - Bayes theorem is definitely the fundamental relation in statistical pattern recognition.



Rev. Thomas Bayes (1702-1761)

Exercise



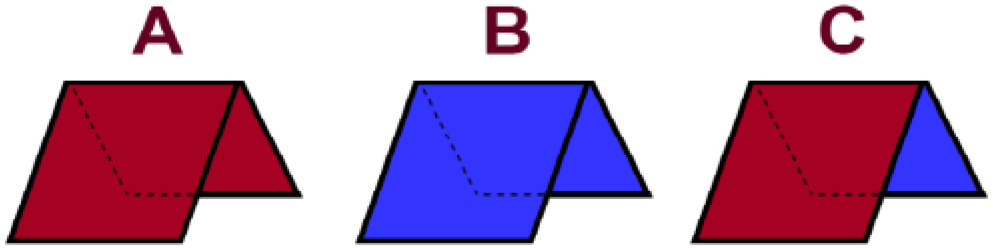
- Before I show you the color of one side of the card:

$$P(A) = P(B) = P(C) = \frac{1}{3}$$

- After I show you the color of one side of the card which turns out to be **RED**, what can you infer about the card?

Q: Is the card more or equally likely to be C?

Exercise: An Intuitive Approach



A	🟥	🟥
B	🟦	🟦
C	🟥	🟦

$$P(red \cap A) = \frac{1}{3}$$

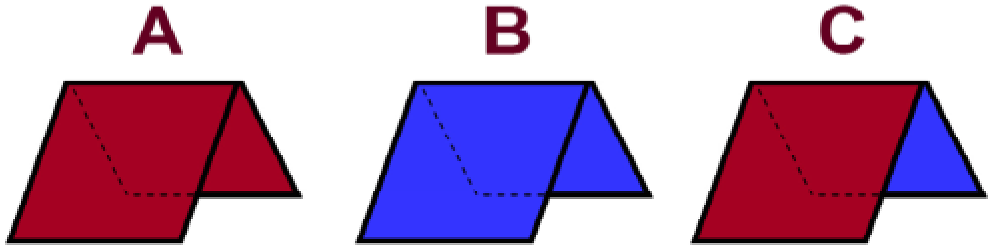
$$P(red \cap B) = 0$$

$$P(red \cap C) = \frac{1}{6}$$

$$P(red) = 1/2$$

$$P(A|red) = \frac{P(red \cap A)}{P(red)} = \frac{2}{3} \qquad P(C|red) = \frac{P(red \cap C)}{P(red)} = \frac{1}{3}$$

Exercise: Bayes Formulation



$P(red|A) = 1$ $\quad$ $P(red|B) = 0$ $\quad$ $P(red|C) = \frac{1}{2}$

$$P(red) = P(red|A)P(A) + P(red|B)P(B) + P(red|C)P(C) = \frac{1}{2}$$

$$P(A|red) = \frac{P(red|A)P(A)}{P(red)} = \frac{2}{3}$$

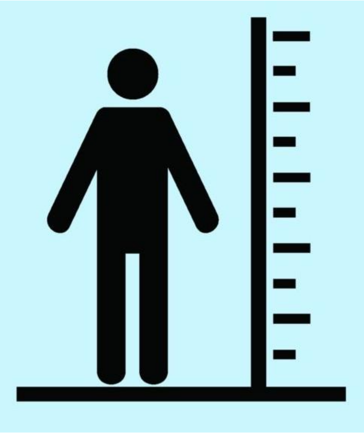
$$P(B|red) = \frac{P(red|B)P(B)}{P(red)} = 0 \qquad P(C|red) = \frac{P(red|C)P(C)}{P(red)} = \frac{1}{3}$$

Random Variables

- When we perform a random experiment we are usually interested in some measurement or numerical attribute of the outcome

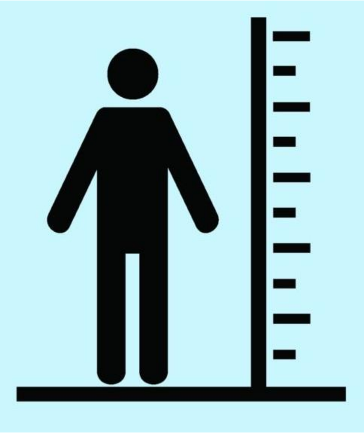
Random Variables

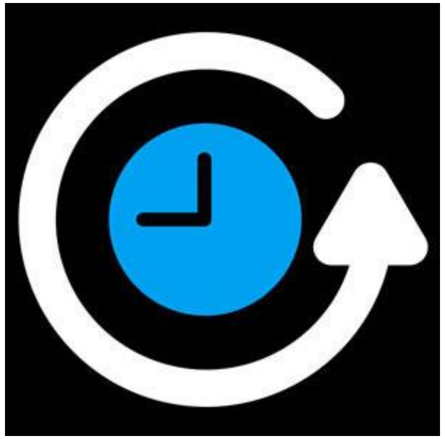
- When we perform a random experiment we are usually interested in some measurement or numerical attribute of the outcome:
  - When sampling a population ⟶ Interested in their heights



Random Variables

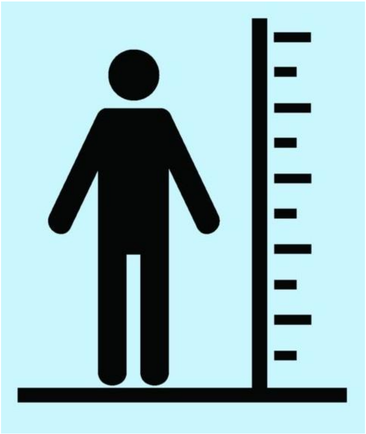
- When we perform a random experiment we are usually interested in some measurement or numerical attribute of the outcome:
  - When sampling a population ⟶ Interested in their heights
  - When rating the performance of two computers ⟶ Interested in the execution time of a benchmark

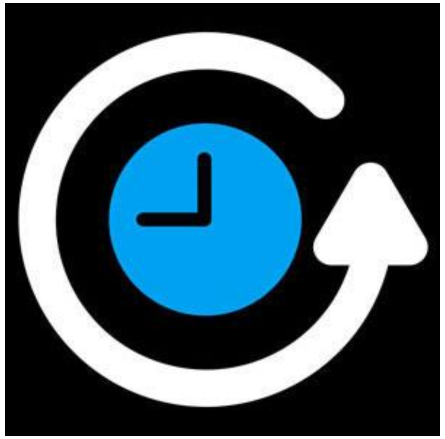


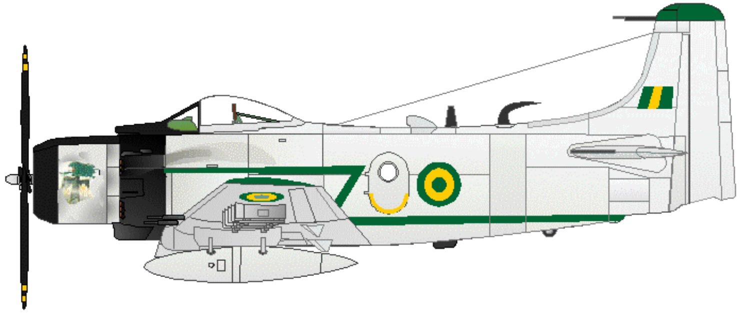


Random Variables

- When we perform a random experiment we are usually interested in some measurement or numerical attribute of the outcome:
  - When sampling a population ⟶ Interested in their heights
  - When rating the performance of two computers ⟶ Interested in the execution time of a benchmark
  - When recognizing an intruder aircraft ⟶ Interested in the parameters that characterize its shape

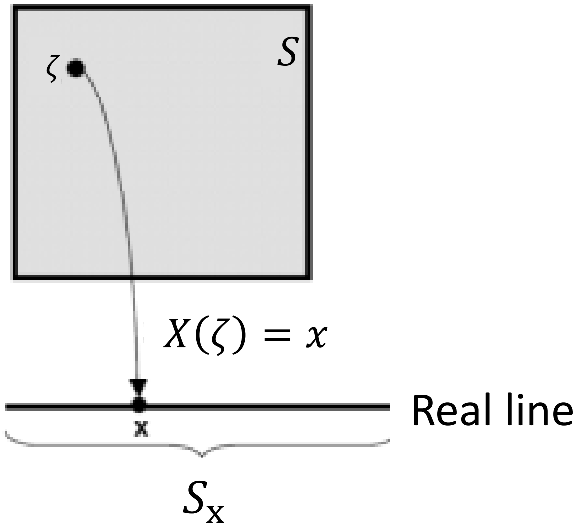






Random Variables

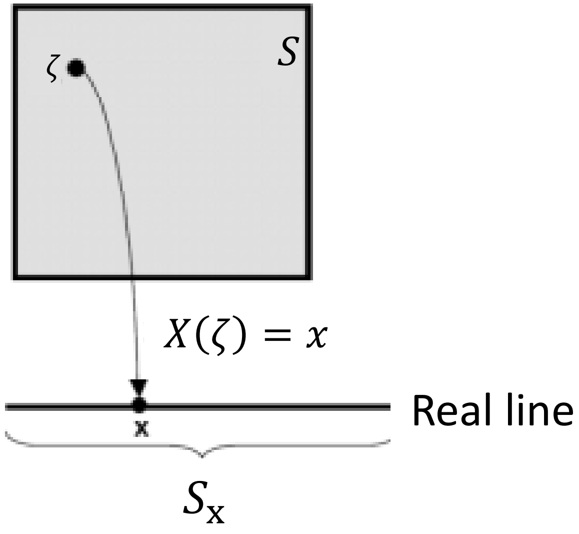
- A random variable X is a function that assigns a real number $X(\zeta)$ to each outcome ζ in the sample space of a random experiment.
  - This function $X(\zeta)$ is performing a mapping from all the possible elements in the sample space onto the real line (real numbers).



Random Variables

- A random variable X is a function that assigns a real number $X(\zeta)$ to each outcome ζ in the sample space of a random experiment.
  - $X(\cdot)$ is performing a mapping from all the possible elements in the sample space onto the real line (real numbers).

- The function X is fixed and deterministic
  - E.g, the rule "count the number of heads in three coin tosses".
  - The randomness the observed values is due to the underlying randomness of the argument ζ (the outcome of the experiment) of the function X



Two Types of Random Variables

- **Discrete Random Variable**
  - Has countable number of values
  - E.g., the resulting number of rolling a dice (any number from 1,2,3,4,5,6)
  - Probability distribution is defined by probability mass function (pmf) 概率质量函数

Two Types of Random Variables

- **Discrete Random Variable**
  - Has countable number of values
  - E.g., the resulting number of rolling a dice (any number from 1,2,3,4,5,6)
  - Probability distribution is defined by probability mass function (pmf) 概率质量函数

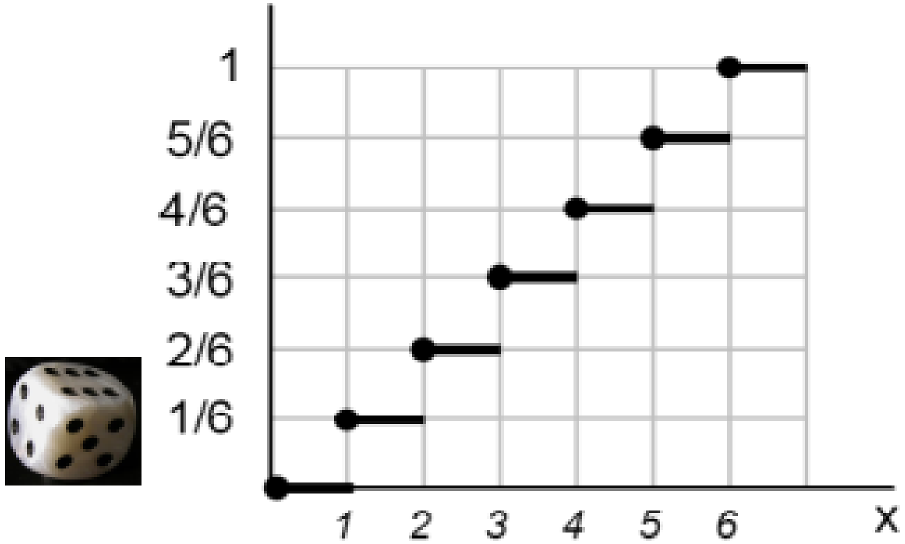
- **Continuous Random Variable**
  - Has values that are continuous
  - E.g., the weight of an individual (any real number within the range of human weight)
  - Probability distribution is defined by probability density function (pdf) 概率密度函数

Cumulative Distribution Function 累积分布函数

- The cumulative distribution function $F_X(\mathrm{x})$ of a random variable X is defined as the probability of the event $\{X \leq \mathrm{x}\}$

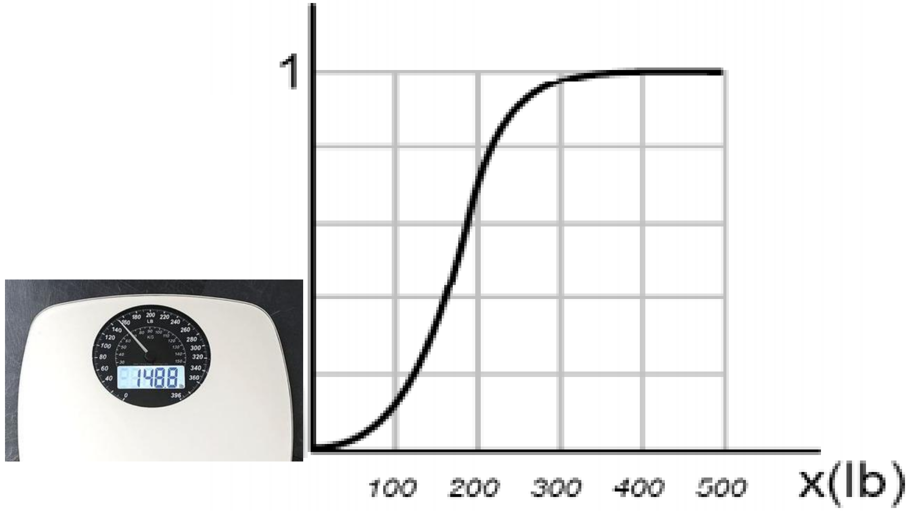
$$F_X(\mathrm{x}) = P[X \leq \mathrm{x}] \text{ for } -\infty < \mathrm{x} < +\infty$$

CDF for discrete RV



cdf for rolling a dice

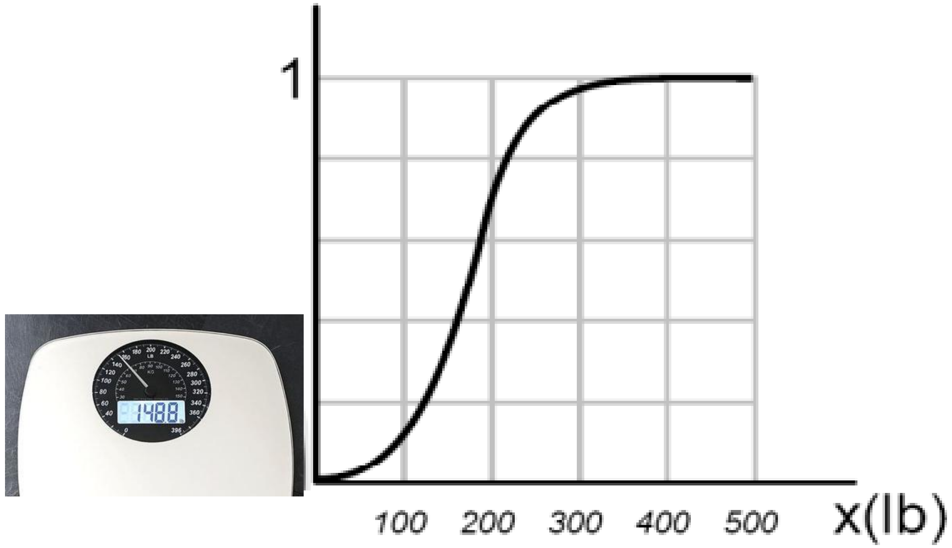
CDF for continuous RV



cdf for a person's weight

Properties of CDF

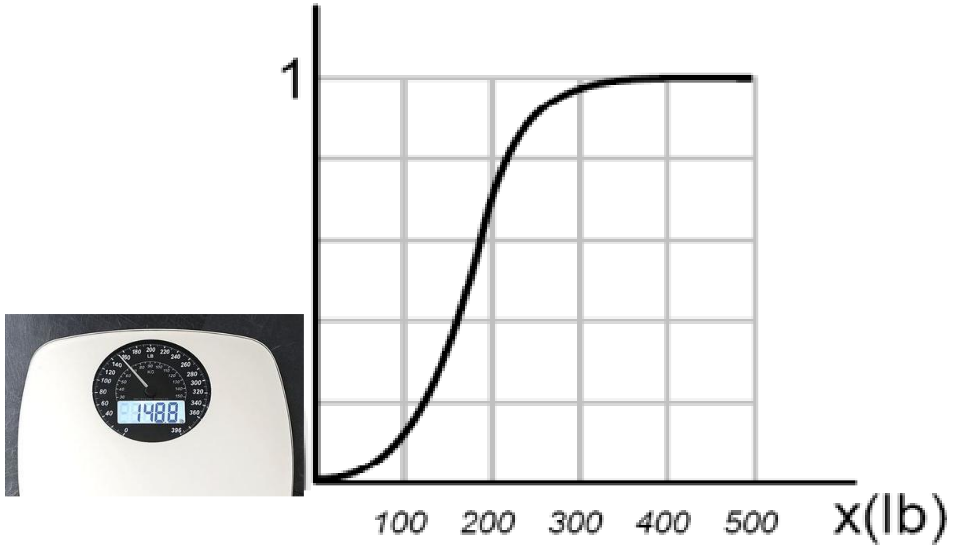
- $0 \le F_X(x) \le 1$
- $\lim_{x \to \infty} F_X(x) = 1, \quad \lim_{x \to -\infty} F_X(x) = 0$
- $F_X(a) \le F_X(b)$ if $a \le b$
- $F_X(b) = \lim_{h \to 0} F_X(b + h) = F_X(b^+)$



cdf for a person's weight

Properties of CDF

- $0 \le F_X(x) \le 1$
- $\lim_{x \to \infty} F_X(x) = 1, \quad \lim_{x \to -\infty} F_X(x) = 0$
- $F_X(a) \le F_X(b)$ if $a \le b$
- $F_X(b) = \lim_{h \to 0} F_X(b+h) = F_X(b^+)$

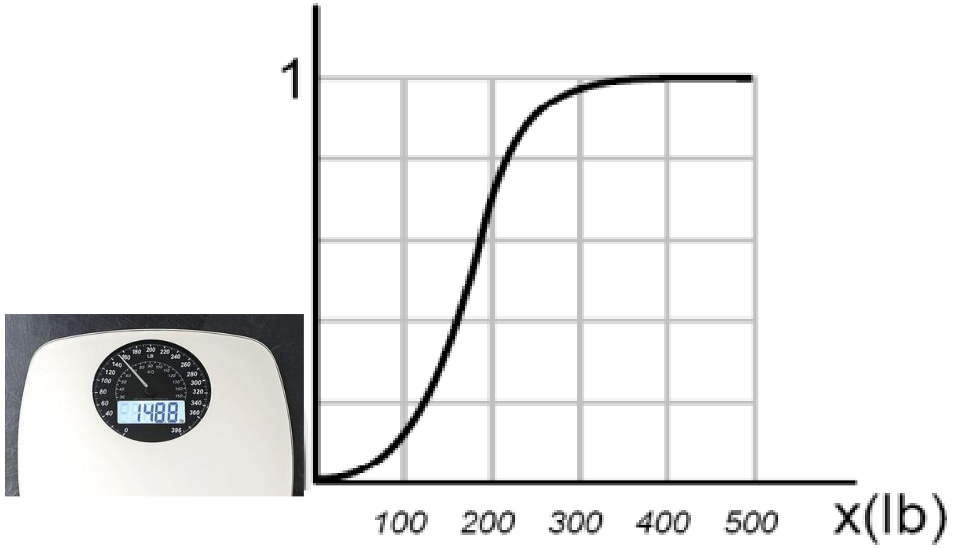


cdf for a person's weight

$$\boxed{P(a < X \le b) = F(b) - F(a)}$$

Properties of CDF

- $0 \leq F_X(x) \leq 1$
- $\lim_{x \to \infty} F_X(x) = 1, \qquad \lim_{x \to -\infty} F_X(x) = 0$
- $F_X(a) \leq F_X(b)$ if $a \leq b$
- $F_X(b) = \lim_{h \to 0} F_X(b+h) = F_X(b^+)$



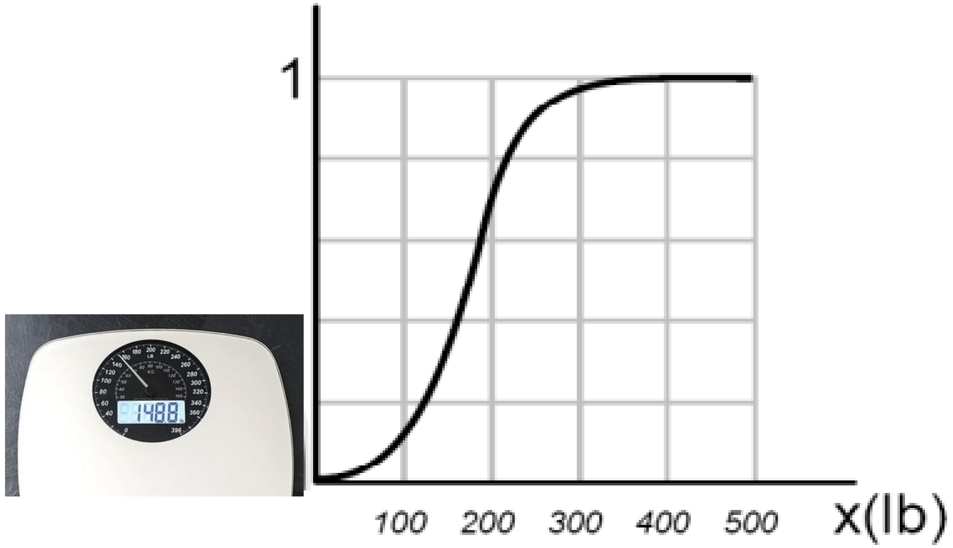
cdf for a person's weight

$$\boxed{P(a < X \leq b) = F(b) - F(a)}$$

P(a person's weight between 100 and 200) =?

Properties of CDF

- $0 \leq F_X(x) \leq 1$
- $\lim_{x\to\infty} F_X(x) = 1, \quad \lim_{x\to-\infty} F_X(x) = 0$
- $F_X(a) \leq F_X(b)$ if $a \leq b$
- $F_X(b) = \lim_{h\to 0} F_X(b+h) = F_X(b^+)$



cdf for a person's weight

$$\boxed{P(a < X \leq b) = F(b) - F(a)}$$

$$P(\text{a person's weight between 100 and 200}) = F(200) - F(100)$$

Discrete RV: Probability Mass Function

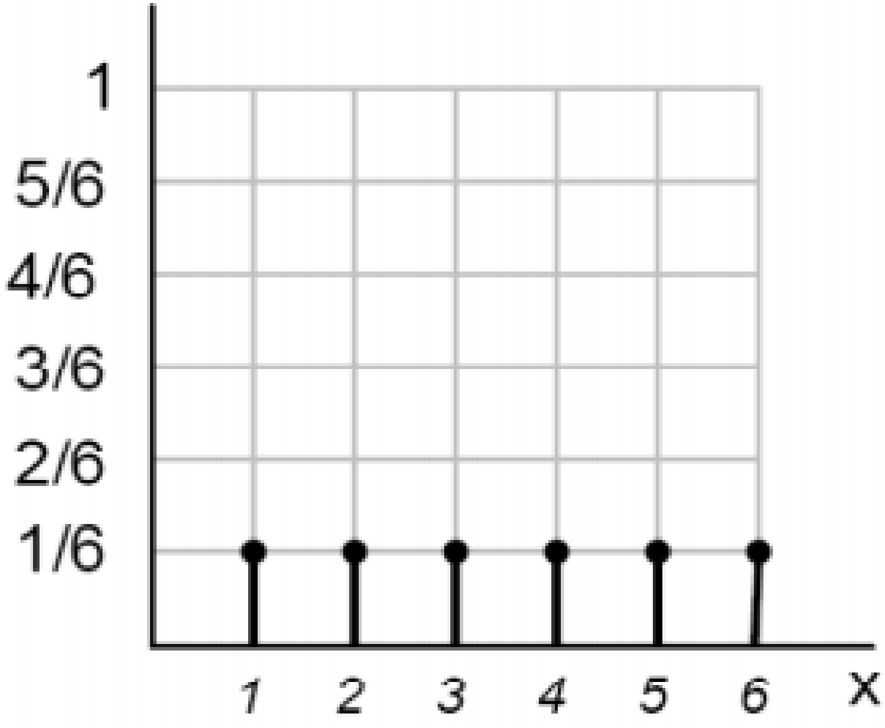
- Given a discrete RV X, the probability mass function is defined as

$$P(a) = P(X = a)$$

- Satisfies all axioms of probability

- CDF satisfies

$$\mathrm{F_X(a)} = P(X \leq a) = \sum_{k \leq a} P(X = k)$$



pmf for rolling a (fair) dice

Continuous RV: Probability Density Function

- Probability density function is the derivative of CDF,

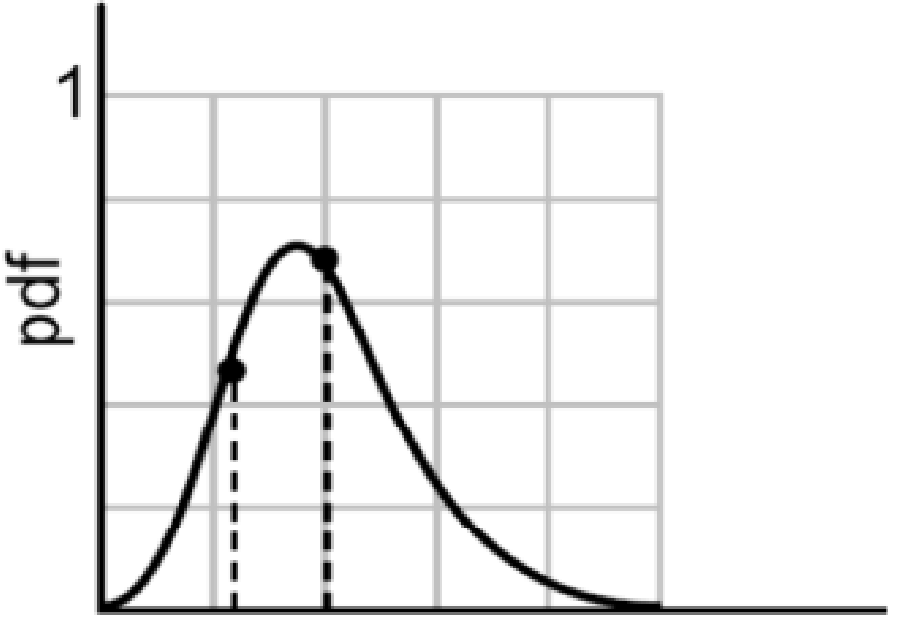
$$\mathrm{f_X(x)} = \frac{\mathrm{dF_X(x)}}{\mathrm{dx}}$$

- CDF satisfies

$$\mathrm{F_X}(a) = P(X \leq a) = \int_{-\infty}^{a} \mathrm{f_X(x)}dx$$

$$P(a < X \leq b) = \int_{a}^{b} \mathrm{f_X(x)}dx$$

General usage



pdf for a person's weight

Statistical Characterization of RVs

- The **cdf** or the **pdf** are **SUFFICIENT** to characterize a random variable.
- A random variable can be also **PARTIALLY** characterized by other measures:

Statistical Characterization of RVs

- The **cdf** or the **pdf** are **SUFFICIENT** to characterize a random variable.
- A random variable can be also **PARTIALLY** characterized by other measures:

Expectation $\mathrm{E[X]} = \mu = \int_{-\infty}^{+\infty} \mathrm{x f_x(x) dx}$

Variance $\mathrm{VAR[X]} = \mathrm{E[(X - E[X])^2]} = \mathrm{E}[X^2] - (E[X])^2 \int_{-\infty}^{+\infty} (\mathrm{x} - \mu)^2 \mathrm{f_x(x) dx}$

Standard deviation $\mathrm{STD[X]} = \sqrt{\mathrm{VAR[X]}}$

Statistical Characterization of RVs

- For two random variables X and Y,

Covariance $COV[X,Y] = E[\{X - E[X]\}\{Y - E[Y]\}] = E[XY] - E[X]E[Y]$

The extent to which X and Y vary together.

$|COV[X,Y]| \leq \sqrt{VAR[X]VAR[Y]}$ Cauchy–Schwarz inequality.

Variance $VAR[X+Y] = VAR[X] + VAR[Y] - COV[X,Y]$

If X and Y are independent, $VAR[X+Y] = VAR[X] + VAR[Y]$

Interpretation of The Correlation Coefficient ρ

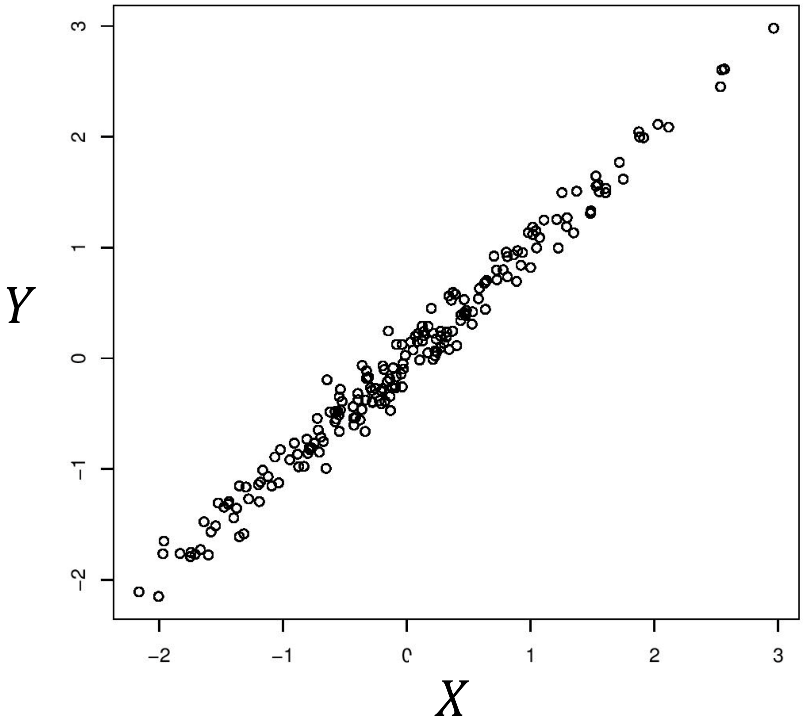
- Correlation coefficient ρ (normalized covariance)

$$\rho(X,Y) = \frac{COV[X,Y]}{\sqrt{VAR[X]VAR[Y]}}$$

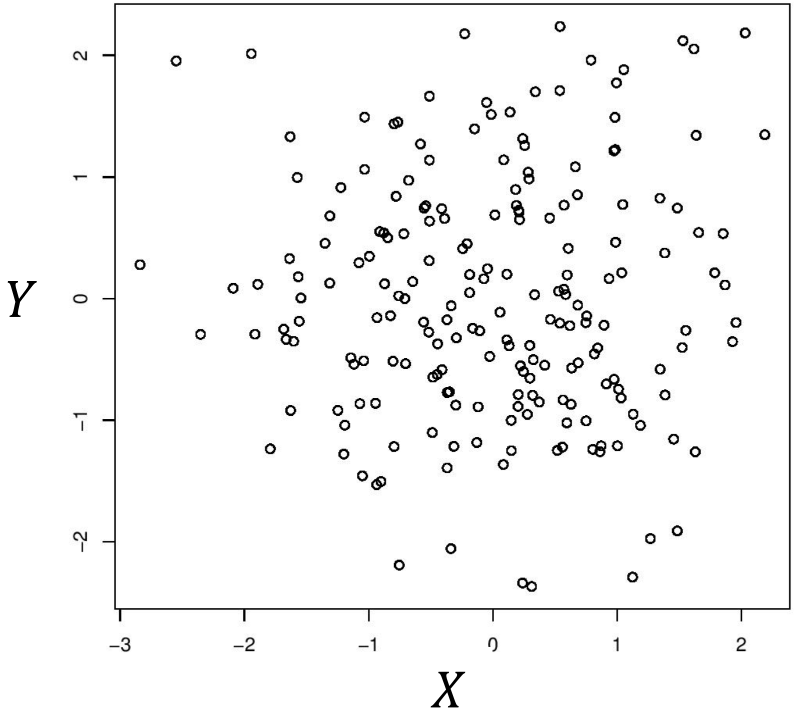
- $\rho(X,Y)$ measures the strength and direction of the linear relationship between X and Y.
- If X and Y have non-zero variance, then $\rho(X,Y) \in [-1,1]$.
- Y is a linearly increasing function of X if and only if $\rho(X,Y) = 1$
- Y is a linearly decreasing function of X if and only if $\rho(X,Y) = -1$
- X and Y are uncorrelated, if and only if $\rho(X,Y) = 0$

Interpretation of The Correlation Coefficient ρ

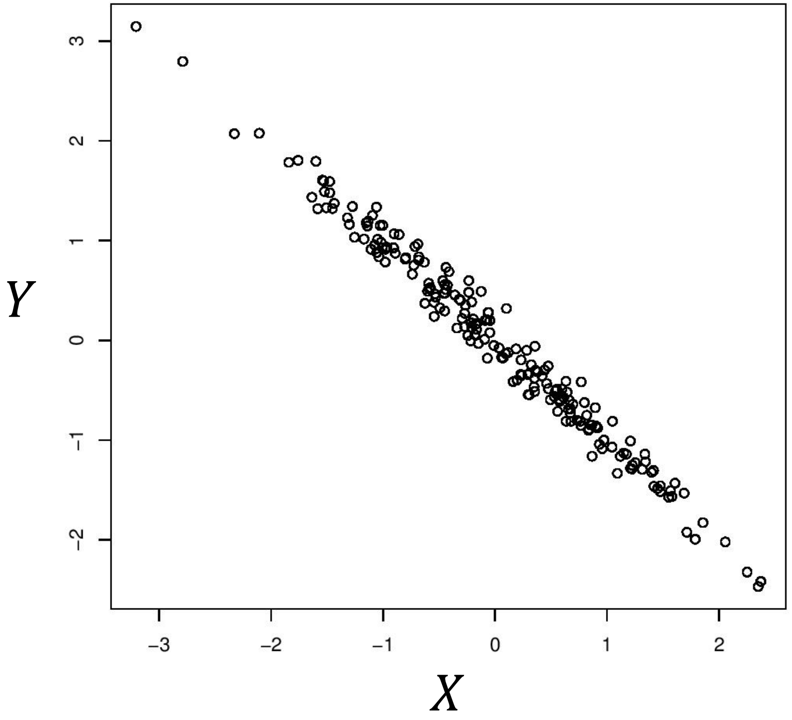
- Y is a linearly increasing function of X if and only if $\rho(X, Y) = 1$
- Y is a linearly decreasing function of X if and only if $\rho(X, Y) = -1$
- X and Y are uncorrelated, if and only if $\rho(X, Y) = 0$



$\rho(X, Y) = 0.99$



$\rho(X, Y) = 0$



$\rho(X, Y) = -0.99$

Interpretation of The Correlation Coefficient ρ

- Correlation coefficient ρ (normalized covariance)

$$\rho(X,Y) = \frac{COV[X,Y]}{\sqrt{VAR[X]VAR[Y]}}$$

- $\rho(X,Y)$ measures the strength and direction of the linear relationship between X and Y.
- If X and Y have non-zero variance, then $\rho(X,Y) \in [-1,1]$.
- Y is a linearly increasing function of X if and only if $\rho(X,Y) = 1$
- Y is a linearly decreasing function of X if and only if $\rho(X,Y) = -1$
- X and Y are uncorrelated, if and only if $\rho(X,Y) = 0$

Can you prove that for any two RV's X and Y, if $\rho(X,Y) = 0$, then there must be no linear dependence between them (i.e., "uncorrelated"=="linearly independent")?

Random Vectors

- A function that assigns a **vector** of **real numbers** to each outcome ζ in the sample space S. (An **extension** of RV’s.)

Random Vectors

- A function that assigns a **vector** of **real numbers** to each outcome ζ in the sample space S. (An **extension** of RV's.)

- The notions of cdf and pdf are replaced by "joint cdf" and "joint pdf".
- Given random vector, $\vec{\mathbf{X}} = [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{N}]^\mathrm{T}$, we define,

Random Vectors

- A function that assigns a **vector** of **real numbers** to each outcome ζ in the sample space S. (An **extension** of RV's.)

- The notions of cdf and pdf are replaced by "joint cdf" and "joint pdf".
- Given random vector, $\vec{\mathbf{X}} = [x_1, x_2, \ldots, x_N]^T$, we define,

Joint cdf $\quad F_{\vec{\mathbf{X}}}(\vec{\mathbf{X}}) = P_{\vec{\mathbf{X}}}[\{X_1 \le x_1\} \cap \{X_2 \le x_2\} \cap \cdots \cap \{X_N \le x_N\}]$

Joint pdf $\quad f_{\vec{\mathbf{X}}}(\vec{\mathbf{X}}) = \frac{\partial^N F_{\vec{\mathbf{X}}}(\vec{\mathbf{X}})}{\partial X_1 \partial X_2 \ldots \partial X_N}$

Random Vectors

- Marginal pdf: the pdf of a subset of all the random vector dimensions
  - Can be obtained by integrating out the variables that are not interest.

E.g., for a two-dimensional random vector $\vec{\mathbf{X}} = [\mathrm{x}_1, \mathrm{x}_2]^{\mathrm{T}}$, where we have the joint pdf $f_{x_1 x_2}(x_1 x_2)$, then the marginal pdf of x_1,

$$\mathrm{f}_{\mathrm{x}_1}(\mathrm{x}_1) = \int_{\mathrm{x}_2=-\infty}^{\mathrm{x}_2=+\infty} \mathrm{f}_{\mathrm{x}_1\mathrm{x}_2}(\mathrm{x}_1\mathrm{x}_2)\mathrm{d}\mathrm{x}_2$$

Statistical Characterization of Random Vectors

- A random vector can be fully characterized by its joint cdf or joint pdf
- Alternatively, we can partially describe a random vector with measures as follows.

Mean vector $E[\boldsymbol{X}] = [E[X_1], E[X_2], \dots, E[X_N]]^T = [\mu_1 \mu_2 \dots \mu_N] = \boldsymbol{\mu}$

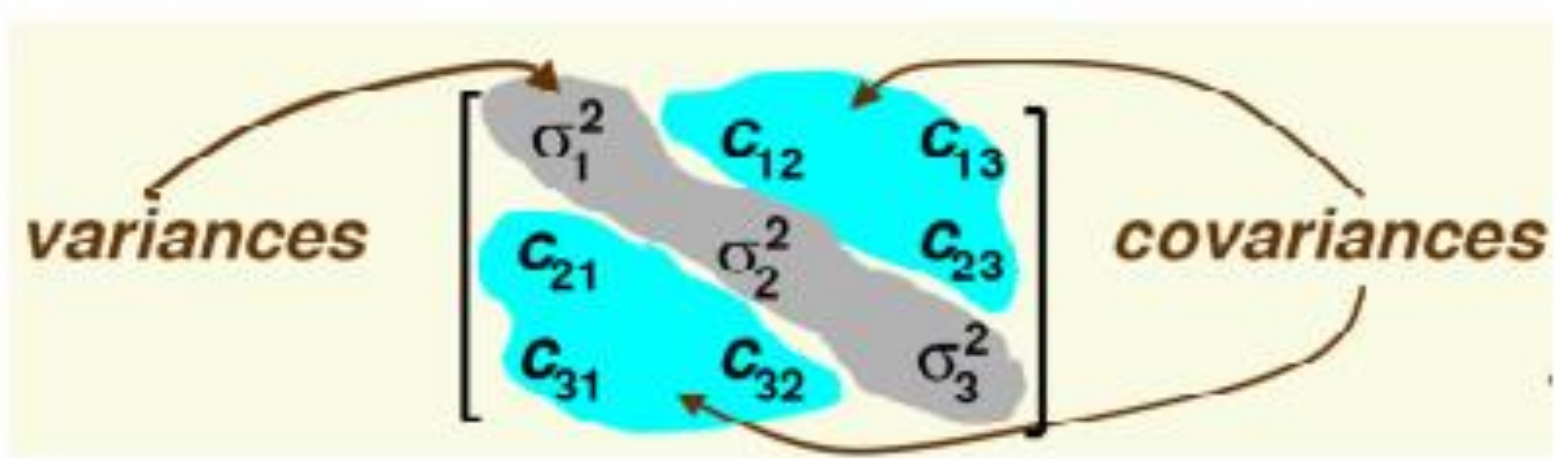
Covariance matrix

$$COV[\boldsymbol{X}] = \boldsymbol{\Sigma} = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\boldsymbol{T}}]$$

$$= \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] \dots E[(X_1 - \mu_1)(X_N - \mu_N)] \\ \ddots \\ E[(X_N - \mu_N)(X_1 - \mu_1)] \dots E[(X_N - \mu_N)(X_N - \mu_N)] \end{bmatrix} = \begin{bmatrix} {\sigma_1}^2 \dots c_{1N} \\ \dots \\ c_{N1} \dots {\sigma_N}^2 \end{bmatrix}$$

Statistical Characterization of Random Vectors

- A rar
- Alter ollows.



Mean v

Covariance matrix

$$COV[\boldsymbol{X}] = \boldsymbol{\Sigma} = E[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^{\boldsymbol{T}}]$$

$$= \begin{bmatrix} E[(X_1-\mu_1)(X_1-\mu_1)] \ldots E[(X_1-\mu_1)(X_N-\mu_N)] \\ \ddots \\ E[(X_N-\mu_N)(X_1-\mu_1)] \ldots E[(X_N-\mu_N)(X_N-\mu_N)] \end{bmatrix} = \begin{bmatrix} {\sigma_1}^2 \ldots c_{1N} \\ \ldots \\ c_{N1} \ldots {\sigma_N}^2 \end{bmatrix}$$

Covariance Matrix

- The covariance matrix indicates the tendency of each pair of dimensions (features) in a random vector to vary together, i.e., to co-vary.

- **Important Properties**
  - If x_i and x_k tend to increase together, then $c_{ik} > 0$
  - If x_i tends to decrease when x_k increases, then $c_{ik} < 0$
  - If x_i and x_k are uncorrelated, then $c_{ik} = 0$
  - $|c_{ik}| \leq \sigma_i \sigma_k$, where σ_i is the standard deviation of x_i
  - $c_{ii} = {\sigma_i}^2 = \mathrm{VAR}(x_i)$
  - Symmetric: $c_{ji} = c_{ij}$

Covariance Matrix

- The covariance matrix indicates the tendency of each pair of dimensions (features) in a random vector to vary together, i.e., to co-vary.

- **Important Properties**
  - If $\mathrm{x_i}$ and $\mathrm{x_k}$ tend to increase together, then $\mathrm{c_{ik}} > 0$
  - If $\mathrm{x_i}$ tends to decrease when $\mathrm{x_k}$ increases, then $\mathrm{c_{ik}} < 0$
  - If $\mathrm{x_i}$ and $\mathrm{x_k}$ are uncorrelated, then $\mathrm{c_{ik}} = 0$
  - $|\mathrm{c_{ik}}| \leq \sigma_\mathrm{i}\sigma_\mathrm{k}$, where σ_i is the standard deviation of $\mathrm{x_i}$
  - $\mathrm{c_{ii}} = \sigma_\mathrm{i}{}^2 = \mathrm{VAR(x_i)}$
  - Symmetric: $\mathrm{c_{ji}} = \mathrm{c_{ij}}$
  - Positive semi-definite:
    - Eigenvalues are nonnegative
    - Determinant is nonnegative, $|C| \geq 0$

Covariance Matrix: Quiz

- You are given the heights and weights of a certain set of individuals in unknown units. Which one of the following four matrices is the most likely to be the sampled covariance matrix?

$$\text{(a)}\begin{bmatrix} 1.232 & 0.867 \\ -0.867 & 2.791 \end{bmatrix} \qquad \text{(b)}\begin{bmatrix} 1.232 & -0.867 \\ -0.867 & 2.791 \end{bmatrix}$$

$$\text{(c)}\begin{bmatrix} 1.232 & 0.867 \\ 0.867 & 2.791 \end{bmatrix} \qquad \text{(d)}\begin{bmatrix} 1.232 & 3.307 \\ 3.307 & 2.791 \end{bmatrix}$$

Covariance Matrix: Quiz

- You are given the heights and weights of a certain set of individuals in unknown units. Which one of the following four matrices is the most likely to be the sampled covariance matrix?

$$(a)\begin{bmatrix} 1.232 & 0.867 \\ -0.867 & 2.791 \end{bmatrix} \qquad (b)\begin{bmatrix} 1.232 & -0.867 \\ -0.867 & 2.791 \end{bmatrix}$$

$$\boxed{(c)\begin{bmatrix} 1.232 & 0.867 \\ 0.867 & 2.791 \end{bmatrix}} \qquad (d)\begin{bmatrix} 1.232 & 3.307 \\ 3.307 & 2.791 \end{bmatrix}$$

Covariance Matrix

- **Uncorrelation VS. Independence**
  - Two random variables $\mathrm{x_i}$ and $\mathrm{x_j}$ are uncorrelated (linearly independent) if $\mathrm{E[x_i x_k] = E[x_i]E[x_k]}$, i.e., $\rho(\mathrm{x_i}, \mathrm{x}_k) = 0$
  - Two random variables $\mathrm{x_i}$ and $\mathrm{x_j}$ are independent if $\mathrm{P(x_i \cap x_k) = P(x}_i\mathrm{)P(x_k)}$.
    - The joint pdf factorizes into the product of the factors (marginal), one involving only $\mathrm{x_i}$ and one involving only $\mathrm{x_k}$.

Covariance Matrix

- **Uncorrelation VS. Independence**
  - Two random variables $\mathrm{x_i}$ and $\mathrm{x_j}$ are uncorrelated (linearly independent) if $\mathrm{E[x_i x_k] = E[x_i]E[x_k]}$, i.e., $\rho(\mathrm{x_i}, \mathrm{x}_k) = 0$
  - Two random variables $\mathrm{x_i}$ and $\mathrm{x_j}$ are independent if $\mathrm{P(x_i \cap x_k) = P(x}_i\mathrm{)P(x_k)}$.
    - The joint pdf factorizes into the product of the factors (marginal), one involving only $\mathrm{x_i}$ and one involving only $\mathrm{x_k}$.

- One is based on probability while the other one based on expectation.
- Two variables that are independent have zero covariance (uncorrelated).
- Two variables that have $\rho(\mathrm{x_i}, \mathrm{x}_k) \neq 0$ are dependent.
- For two variables $\rho(\mathrm{x_i}, \mathrm{x}_k) = 0$, there must be no linear dependence between them.

Covariance Matrix

- **Uncorrelation VS. Independence**
  - Two random variables $\mathrm{x_i}$ and $\mathrm{x_j}$ are uncorrelated (linearly independent) if $\mathrm{E[x_i x_k] = E[x_i]E[x_k]}$, i.e., $\rho(\mathrm{x_i}, \mathrm{x}_k) = 0$
  - Two random variables $\mathrm{x_i}$ and $\mathrm{x_j}$ are independent if $\mathrm{P(x_i \cap x_k) = P(x}_i\mathrm{)P(x_k)}$.
    - The joint pdf factorizes into the product of the factors (marginal), one involving only $\mathrm{x_i}$ and one involving only $\mathrm{x_k}$.

- Independence is a stronger requirement than $\rho(\mathrm{x_i}, \mathrm{x}_k) = 0$, as independence also excludes nonlinear relationship.
- It is possible for two variables $\mathrm{x_i}$ and x_k are dependent with $\rho(\mathrm{x_i}, \mathrm{x}_k) = 0$.

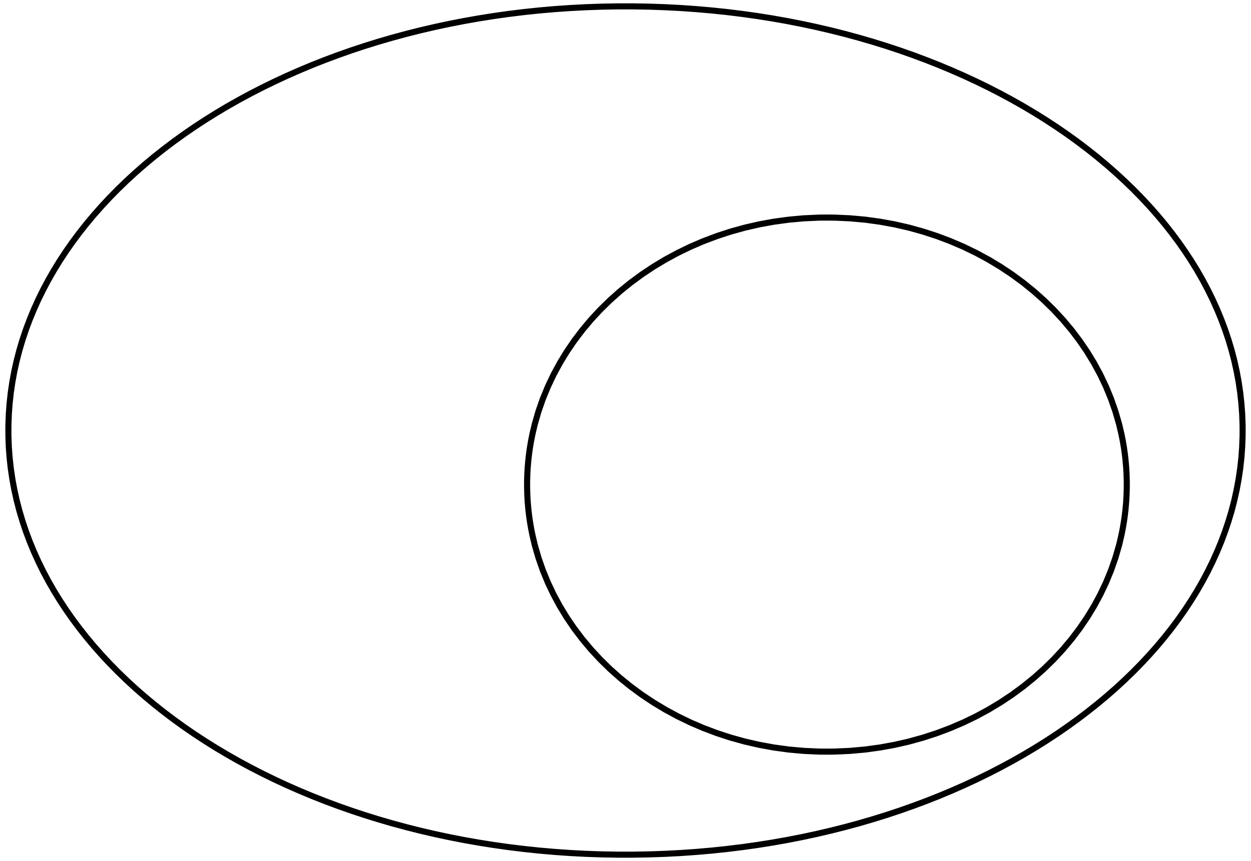
Covariance Matrix

- **Uncorrelation VS. Independence**
  - Two random variables $\mathrm{x_i}$ and $\mathrm{x_j}$ are uncorrelated (linearly independent) if $\mathrm{E[x_i x_k] = E[x_i]E[x_k]}$, i.e., $\rho(\mathrm{x_i}, \mathrm{x}_k) = 0$
  - Two random variables $\mathrm{x_i}$ and $\mathrm{x_j}$ are independent if $\mathrm{P(x_i \cap x_k) = P(x}_i\mathrm{)P(x_k)}$.
    - The joint pdf factorizes into the product of the factors (marginal), one involving only $\mathrm{x_i}$ and one involving only $\mathrm{x_k}$.

- Independence is a stronger requirement than $\rho(\mathrm{x_i}, \mathrm{x}_k) = 0$, as independence also excludes nonlinear relationship.
- It is possible for two variables $\mathrm{x_i}$ and x_k are dependent with $\rho(\mathrm{x_i}, \mathrm{x}_k) = 0$.
- E.g., suppose $Y = X^2$. Clearly, X and Y are not independent, as Y is completely determined by X. However, $COV(X, Y) = 0$.

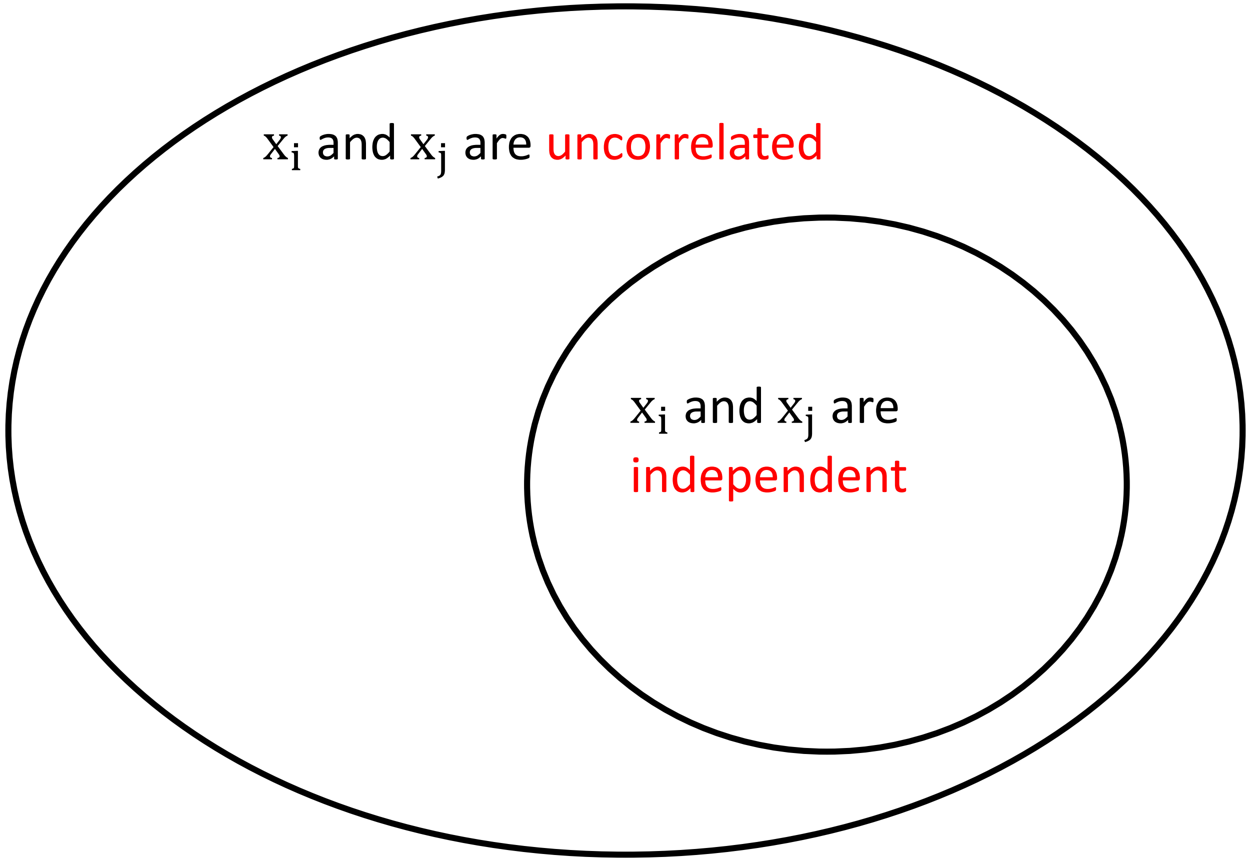
Covariance Matrix

- **Uncorrelation VS. Independence**
  - Uncorrelated (linearly independent): $\mathrm{E}[\mathrm{x_i x_k}] = \mathrm{E}[\mathrm{x_i}]\mathrm{E}[\mathrm{x_k}]$
  - Independent : $\mathrm{P}[\mathrm{x_i} \cap \mathrm{x_k}] = \mathrm{P}[\mathrm{x_i}]\mathrm{P}[\mathrm{x_k}]$.



Covariance Matrix

- **Uncorrelation VS. Independence**
  - Uncorrelated (linearly independent): $\mathrm{E}[x_i x_k] = \mathrm{E}[x_i]\mathrm{E}[x_k]$
  - Independent : $\mathrm{P}[x_i \cap x_k] = \mathrm{P}[x_i]\mathrm{P}[x_k]$.


x_i and x_j are uncorrelated
x_i and x_j are independent

The Normal or Gaussian Distribution of a RV

Deutsche Mark

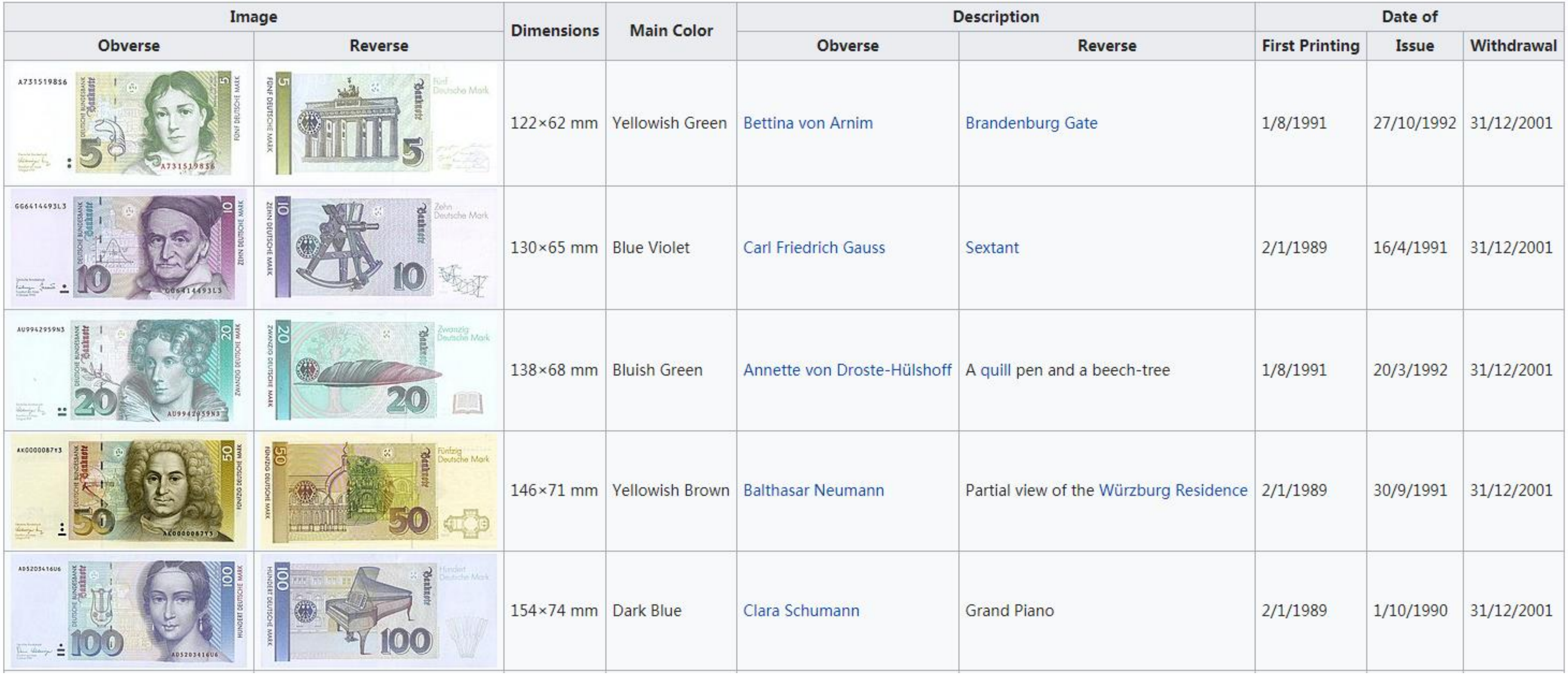
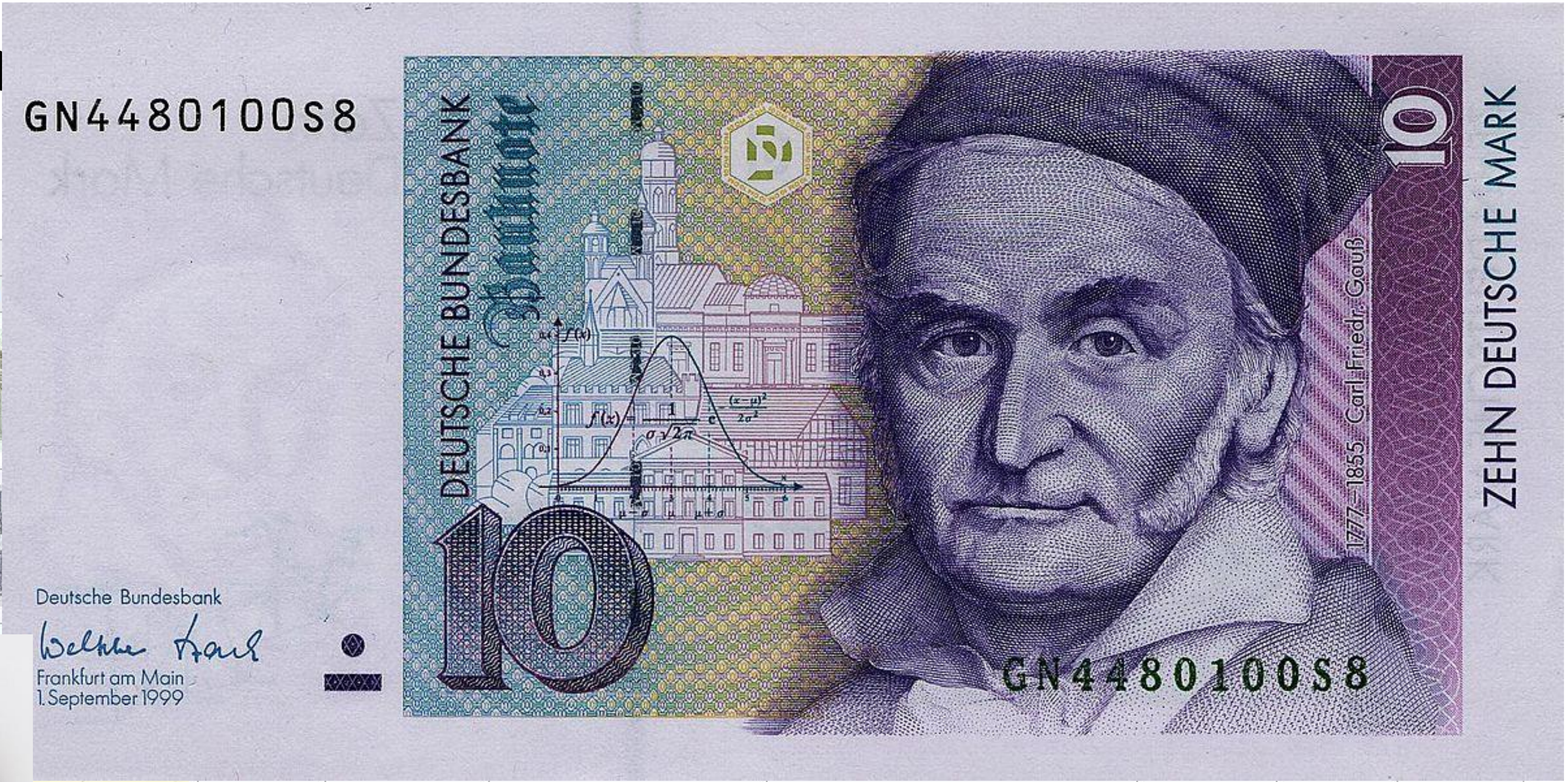


Image		Dimensions	Main Color	Description		Date of		
Obverse	Reverse			Obverse	Reverse	First Printing	Issue	Withdrawal
		122×62 mm	Yellowish Green	Bettina von Arnim	Brandenburg Gate	1/8/1991	27/10/1992	31/12/2001
		130×65 mm	Blue Violet	Carl Friedrich Gauss	Sextant	2/1/1989	16/4/1991	31/12/2001
		138×68 mm	Bluish Green	Annette von Droste-Hülshoff	A quill pen and a beech-tree	1/8/1991	20/3/1992	31/12/2001
		146×71 mm	Yellowish Brown	Balthasar Neumann	Partial view of the Würzburg Residence	2/1/1989	30/9/1991	31/12/2001
		154×74 mm	Dark Blue	Clara Schumann	Grand Piano	2/1/1989	1/10/1990	31/12/2001

The Norma

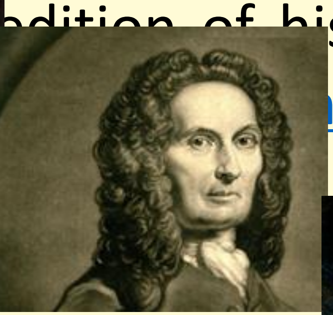


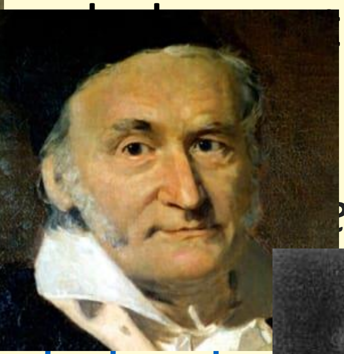
	146×71 mm	Yellowish Brown	Balthasar Neumann	Partial view of the Würzburg Residence	2/1/1989	30/9/1991	31/12/2001
	154×74 mm	Dark Blue	Clara Schumann	Grand Piano	2/1/1989	1/10/1990	31/12/2001

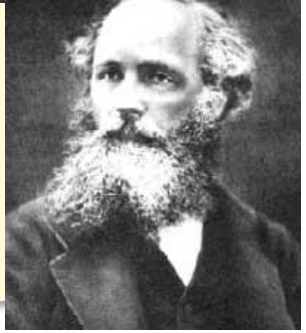
Brief History

- In 1738, de Moivre published in the second edition of his "*The Doctrine of Chances*" the study of the coefficients in the binomial expansion of $(a + b)^n$.
- In 1774, Laplace first posed the problem of aggregating several observations… and first calculated the value of the integral $\int e^{-t2}\,dt = \sqrt{\pi}$ in 1782…
- In 1809 Gauss published his monograph "*Theoria motus corporum coelestium in sectionibus conicis solem ambientium*" where he introduces several important statistical concepts, such as the method of least squares, the method of maximum likelihood, and the *normal distribution.*
- In 1809 an American mathematician Adrain published two derivations of the normal probability law, simultaneously and independently from Gauss.
- In the middle of the 19th century Maxwell demonstrated that the normal distribution is not just a convenient mathematical tool, but may also occur in natural phenomena: "The number of particles whose velocity….

Brief History







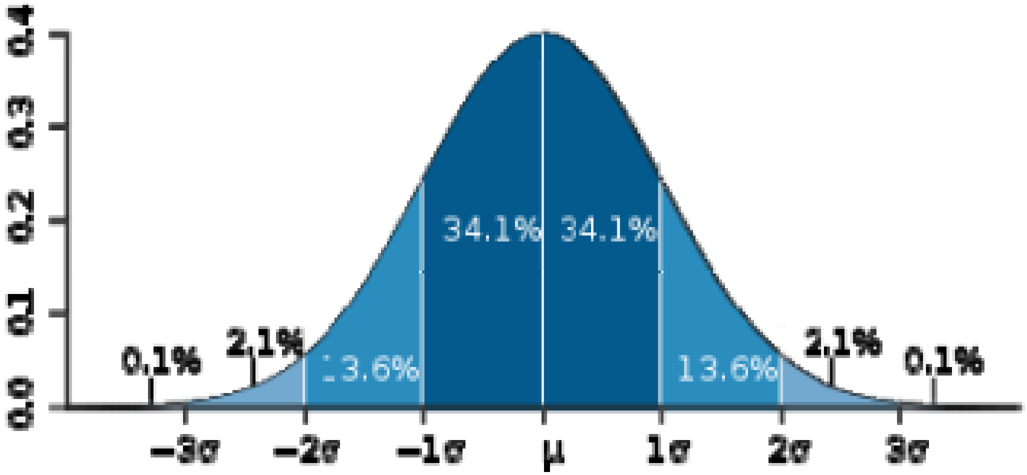
- In 1738, de Moivre published in t edition of his "*The Doctrine of Chances*" the study of the coefficients in the bi nsion of ($a + b$)^n.
- In 1774, Laplace first posed the problem of agg ions… and first calculated the value of the integral ∫ *e*
- In 1809 Gauss published his monograph "*Theoria motus co estium in sectionibus conicis solem ambientium*" where he i important statistical concepts, such as the method of least the method of maximum likelihood, and the *normal distribution.*
- In 1809 an American mathematician Adrain published two derivatio normal probability law, simultaneously and independently from Gauss
- In the middle of the 19th century Maxwell demonstrated that the normal distribution is not just a convenient mathematical tool, but may also occur in natural phenomena: "The number of particles whose velocity….

The Normal or Gaussian Distribution of a RV

- Probability density function:

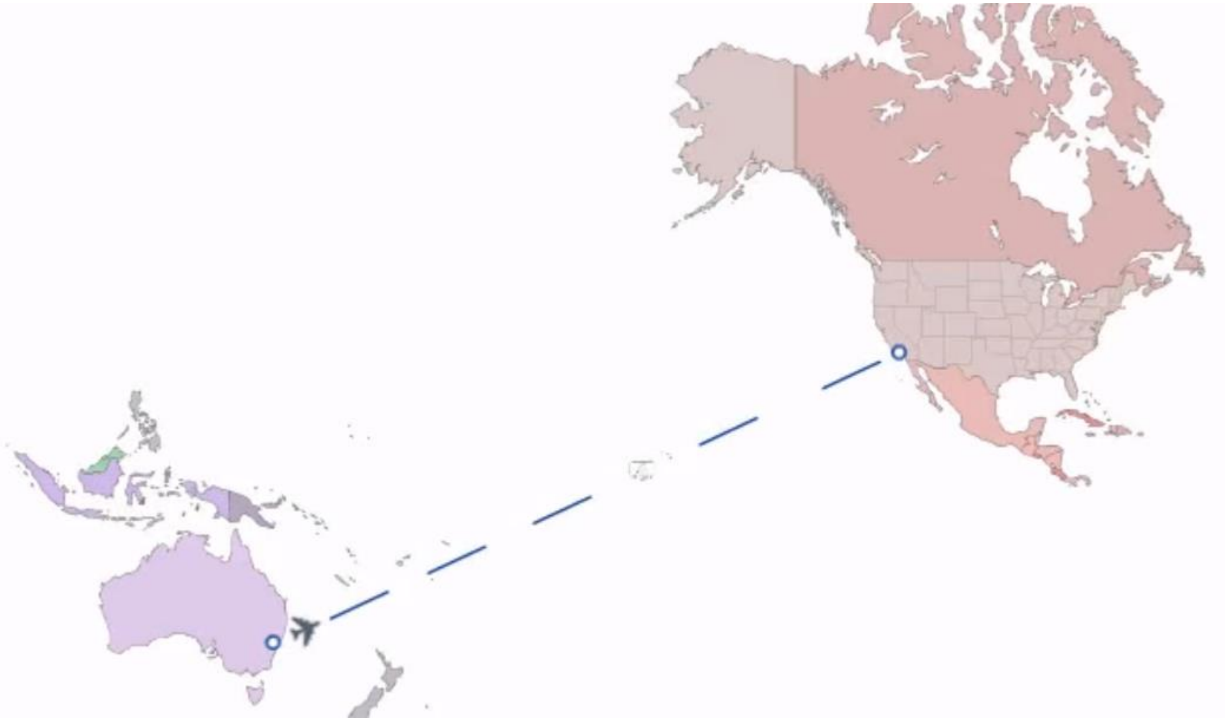
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2\right]$$

- μ = mean (or expected value) of x
- σ^2 = expected squared deviation or variance

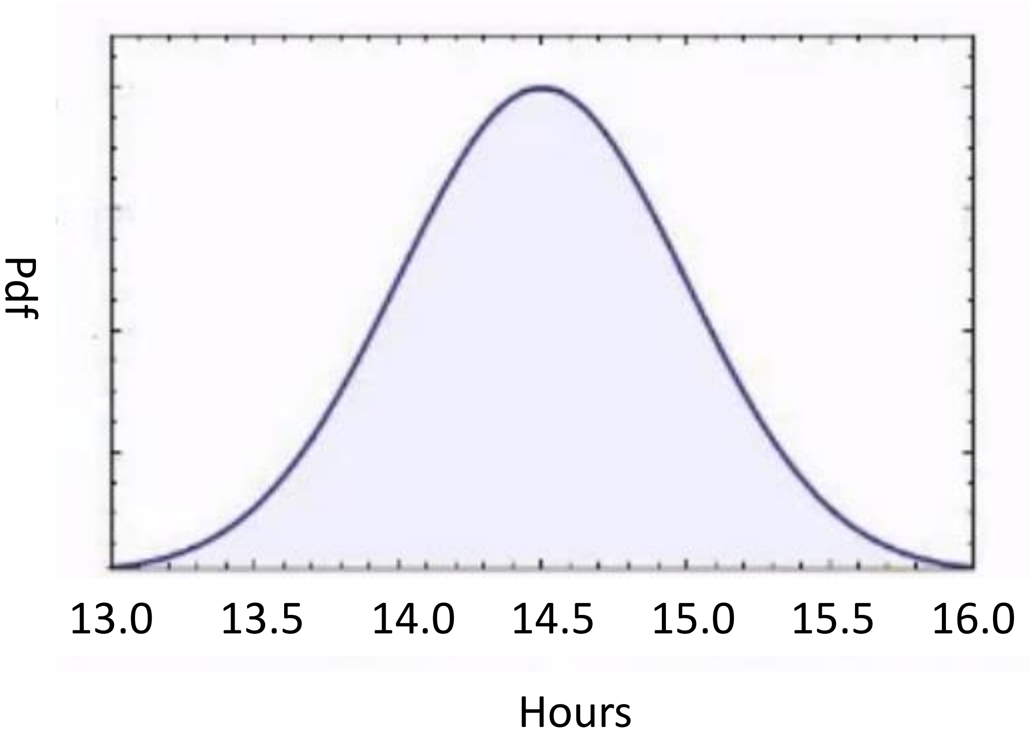


The Normal or Gaussian Distribution of a RV

- How long does the flight from Sydney to Los Angeles take?

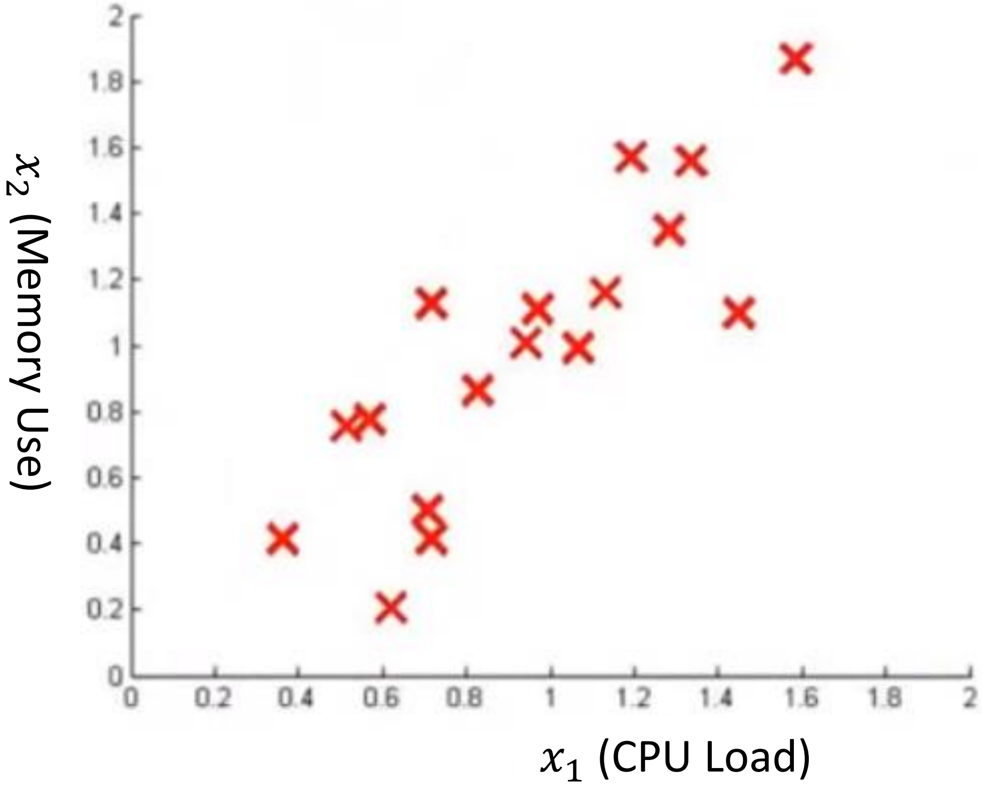


- $\mu = 14.5$ hours
- $\sigma = 0.5$ hours



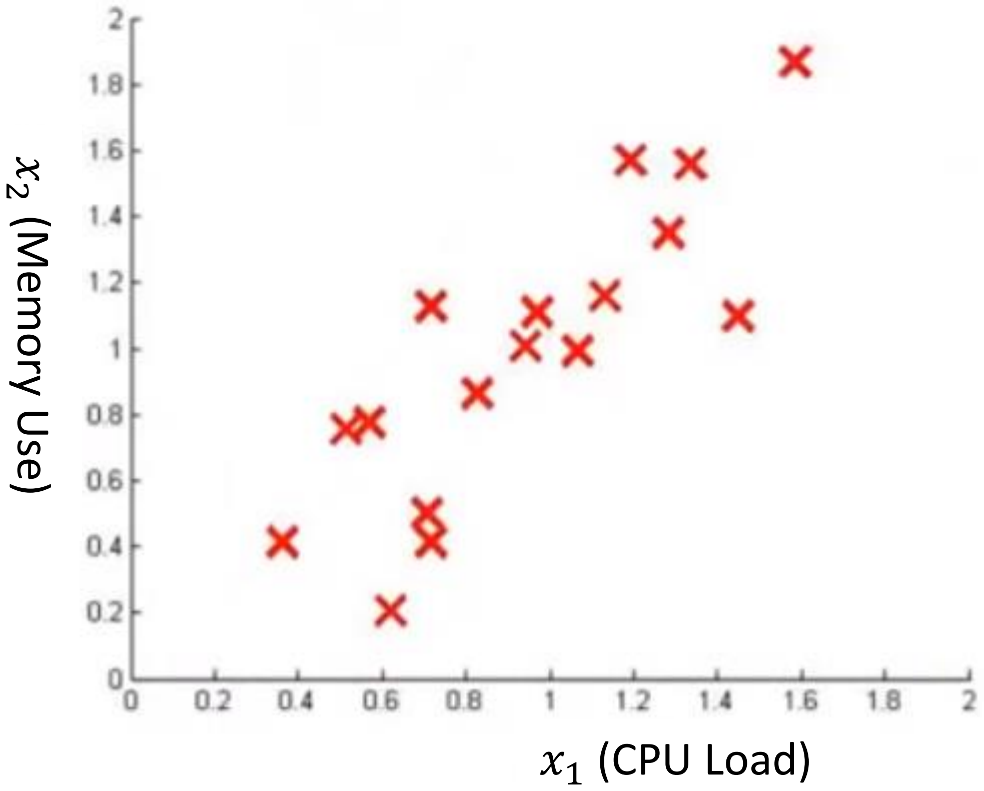
Multivariate Gaussian

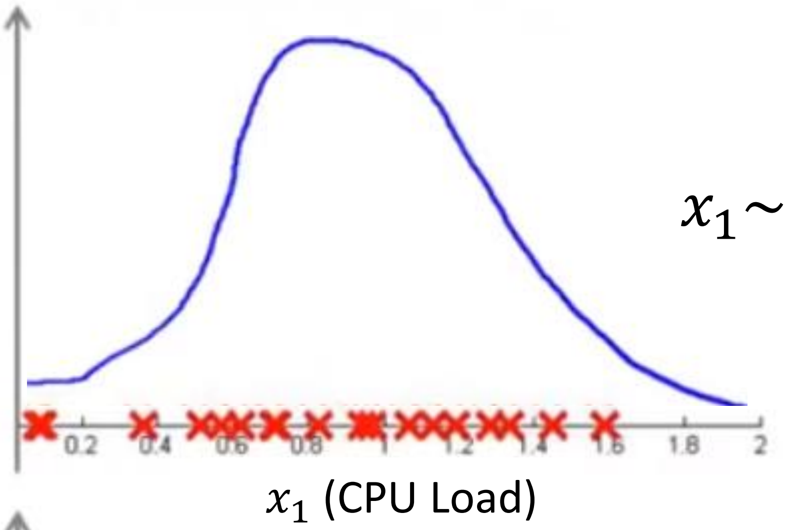
- Motivation example: monitoring machines in a data center.
- If we model the variables x_1 and x_2 separately.



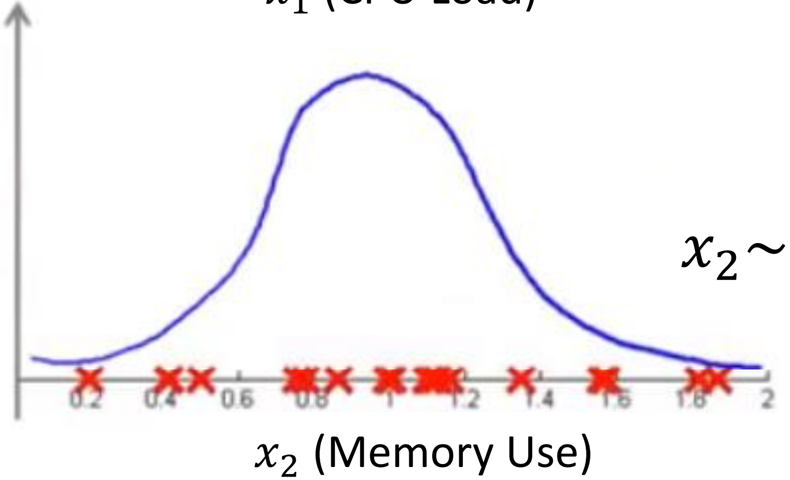
Multivariate Gaussian

- Motivation example: monitoring machines in a data center.
- If we model the variables x_1 and x_2 separately.





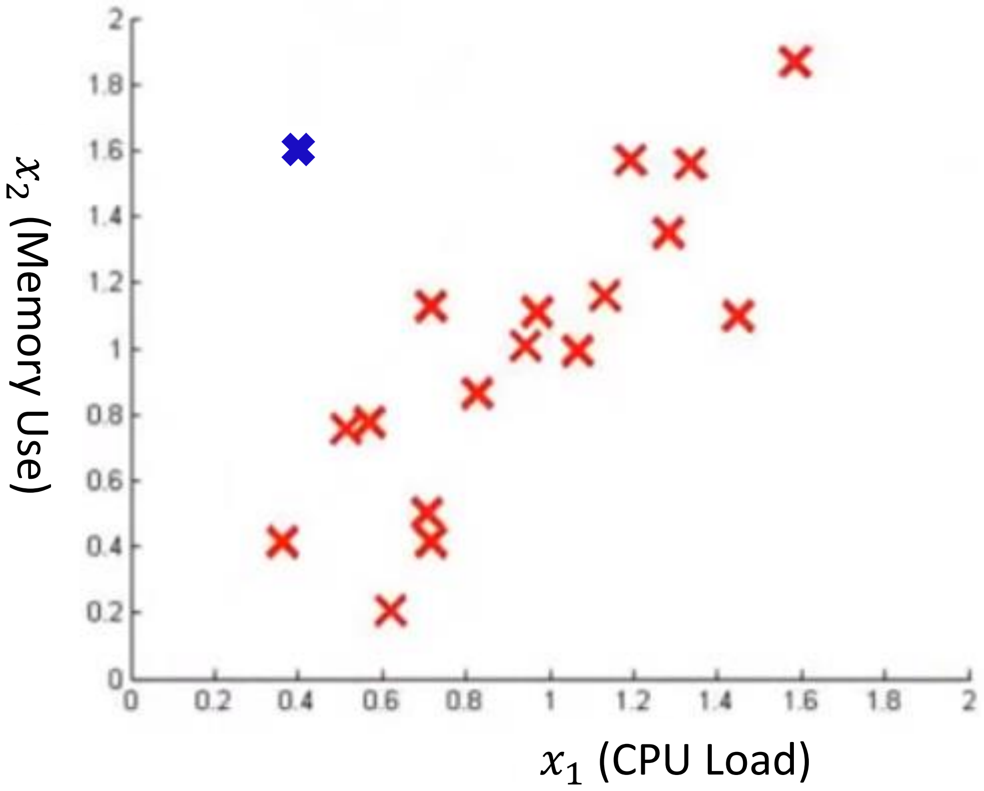
$x_1 \sim p(x_1; \mu_1, \sigma_1)$

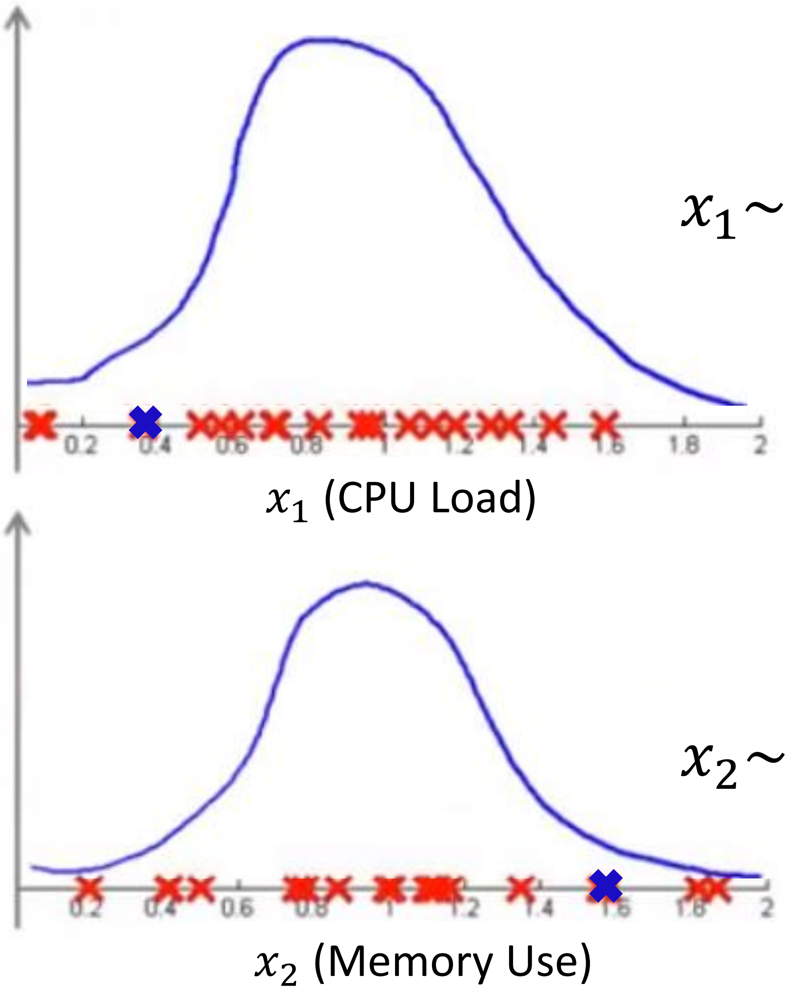


$x_2 \sim p(x_2; \mu_2, \sigma_2)$

Multivariate Gaussian

- Motivation example: monitoring machines in a data center.
- If we model the variables x_1 and x_2 separately.



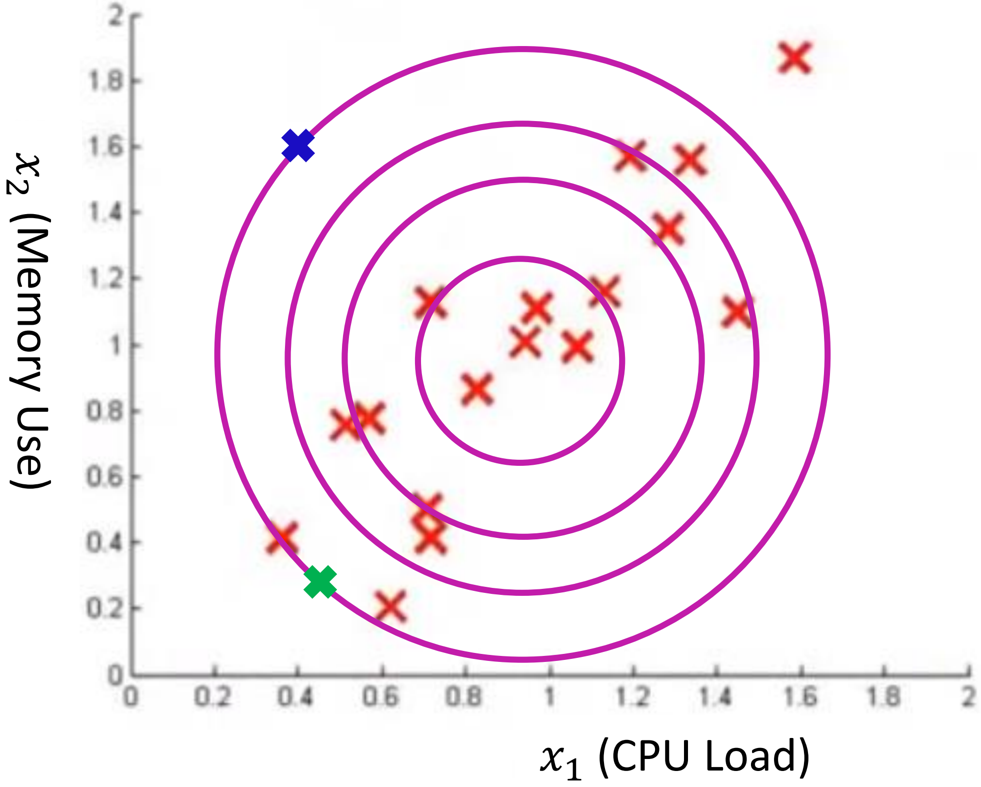


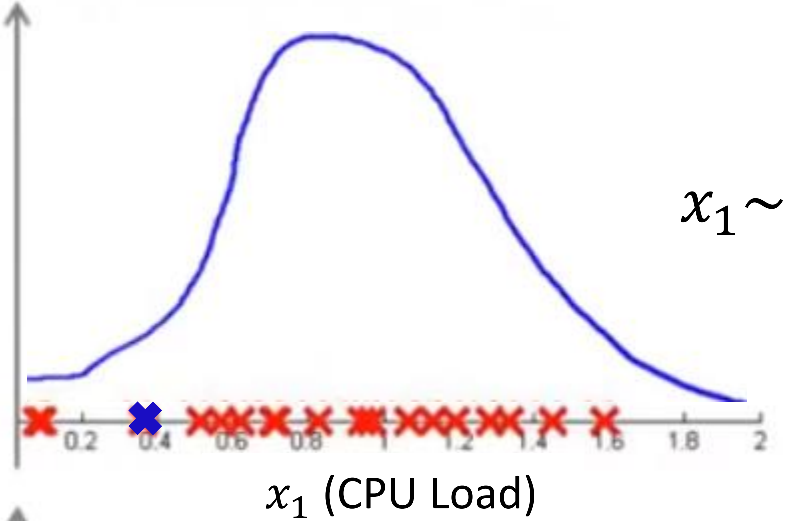
$$x_1 \sim p(x_1; \mu_1, \sigma_1)$$

$$x_2 \sim p(x_2; \mu_2, \sigma_2)$$

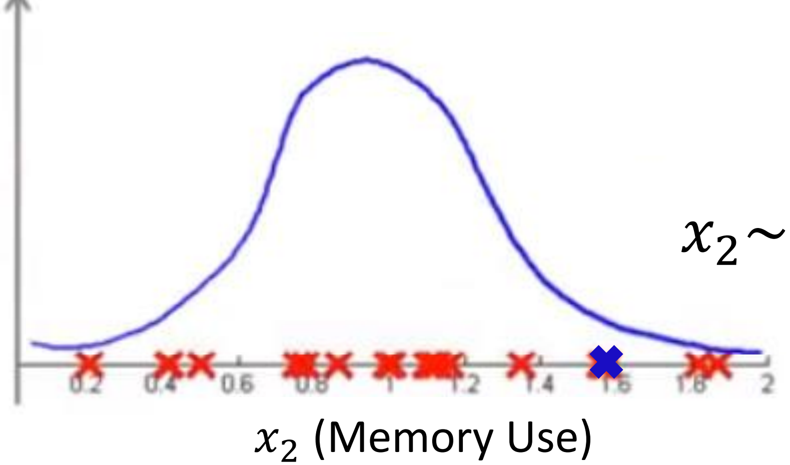
Multivariate Gaussian

- Motivation example: monitoring machines in a data center.
- If we model the variables x_1 and x_2 separately.





$x_1 \sim p(x_1; \mu_1, \sigma_1)$

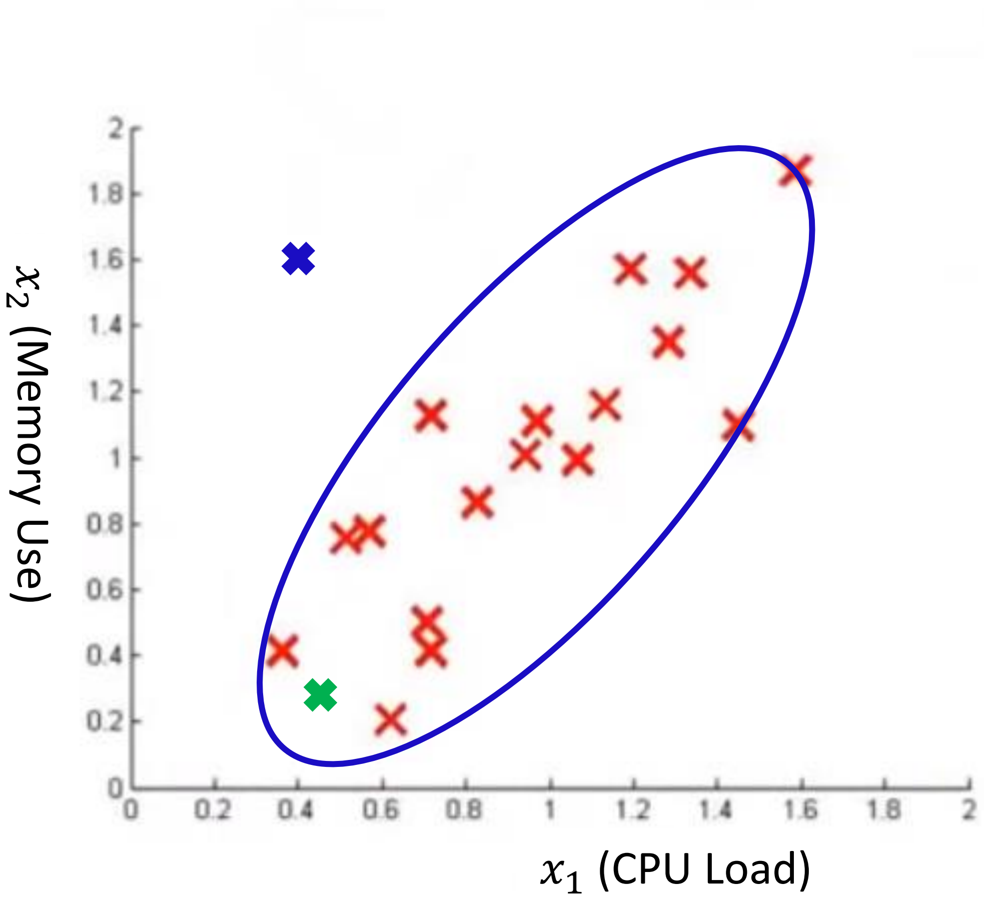


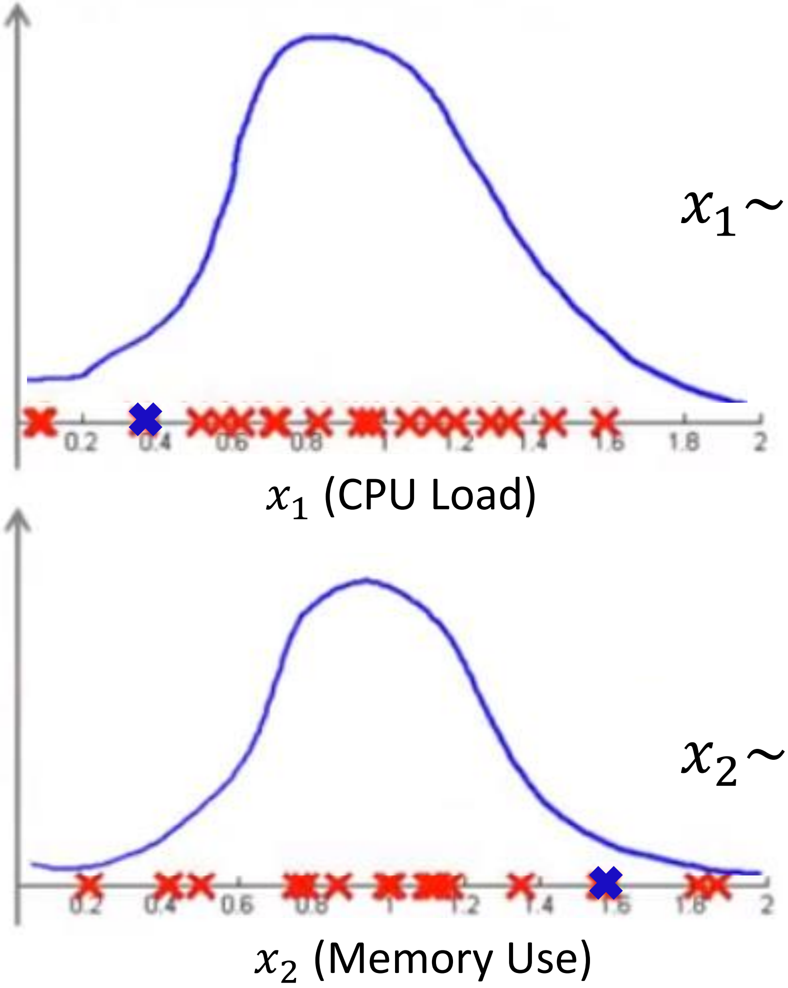
$x_2 \sim p(x_2; \mu_2, \sigma_2)$

Does the p(✖)=p(✖)?

Multivariate Gaussian

- Motivation example: monitoring machines in a data center.
- If we model the variables x_1 and x_2 separately.

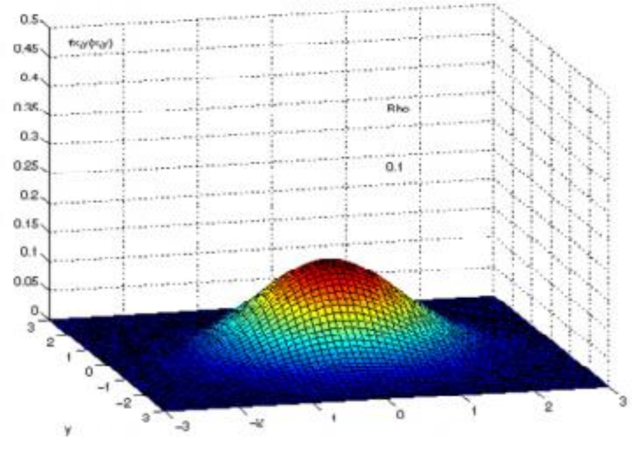




$$x_1 \sim p(x_1; \mu_1, \sigma_1)$$

$$x_2 \sim p(x_2; \mu_2, \sigma_2)$$

Multivariate Gaussian

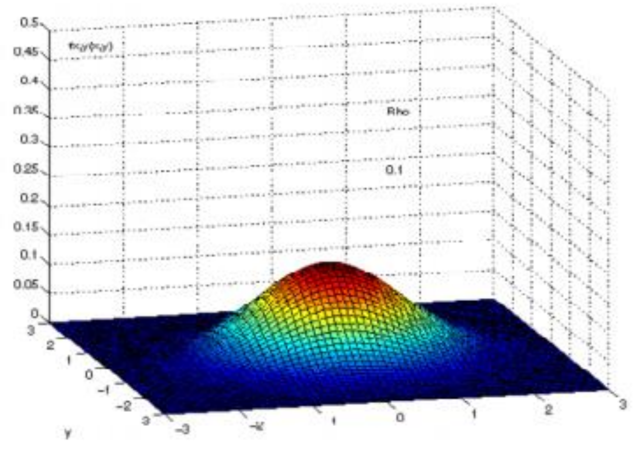


- Probability density function:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{\boldsymbol{-1}}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

- **Mean vector**: $\boldsymbol{\mu}$ **Covariance matrix:** $\boldsymbol{\Sigma}$
- Mahalanobis distance: $\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{\boldsymbol{-1}}(\boldsymbol{x}-\boldsymbol{\mu})}$

Multivariate Gaussian



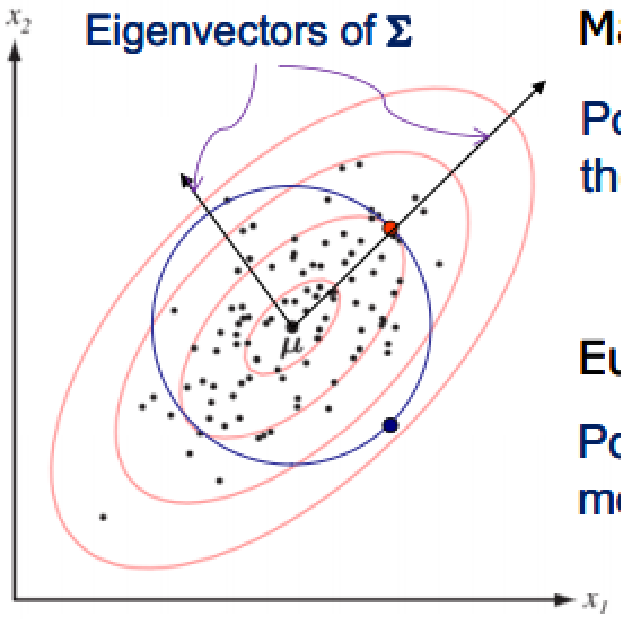
- Probability density function:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

- **Mean vector**: $\boldsymbol{\mu}$ **Covariance matrix:** $\boldsymbol{\Sigma}$
- Mahalanobis distance: $\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\boldsymbol{x}-\boldsymbol{\mu})}$

✓ Represents the distance of the test point $\boldsymbol{x}$ from the mean $\boldsymbol{\mu}$.

✓ If $\boldsymbol{\Sigma} = \boldsymbol{I}$, Mahalanobis distance ↔ Euclidean distance.



Mahalanobis Distance: $\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\boldsymbol{x}-\boldsymbol{\mu})}$

Points of equal Mahalanobis distance to the mean lie on an ellipse.

Euclidean Distance: $\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\boldsymbol{T}}(\boldsymbol{x}-\boldsymbol{\mu})}$

Points of equal Euclidean distance to the mean lie on a circle.

Independent Gaussian Models

- Special Case: Assume that x_1 and x_2 are independent.

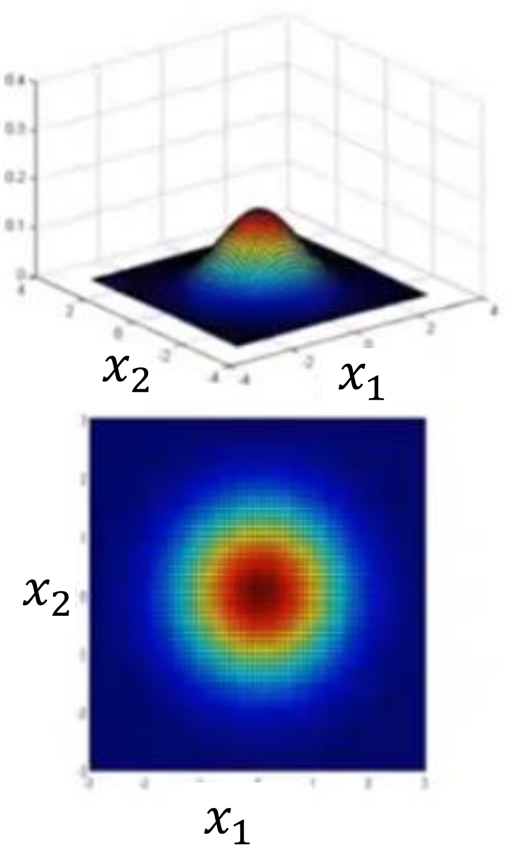
$$p(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} exp\left[-\frac{1}{2}(\frac{x_1-\mu_1}{\sigma_1})^2\right] \qquad p(x_2) = \frac{1}{\sqrt{2\pi}\sigma_2} exp\left[-\frac{1}{2}(\frac{x_2-\mu_2}{\sigma_2})^2\right]$$

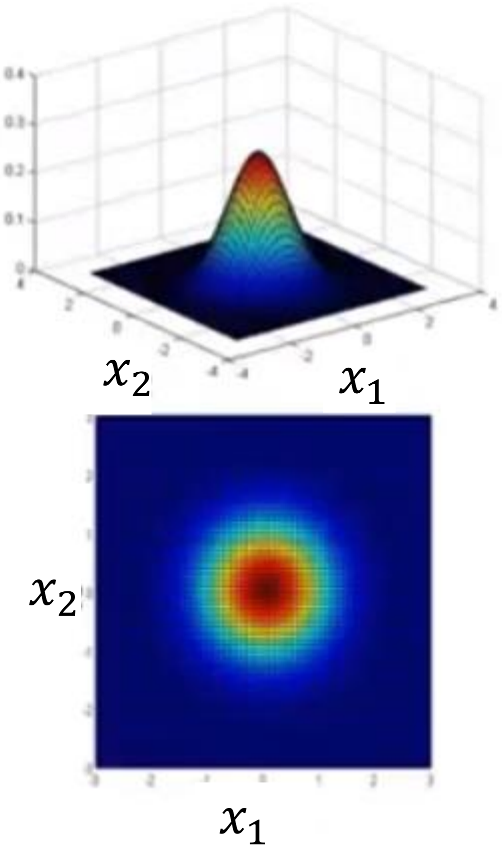
$$p(x_1)p(x_2) = \frac{1}{2\pi\sigma_1\sigma_2} exp\left[\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

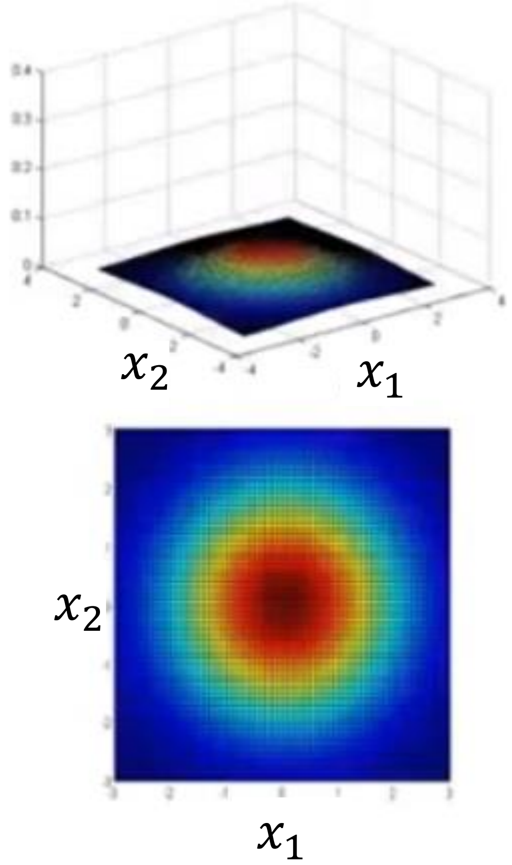
$\boldsymbol{x} = [x_1 \;\; x_2]$ $\qquad$ $\boldsymbol{\mu} = [\mu_1 \;\; \mu_2]$ $\qquad$ $\boldsymbol{\Sigma} = diag(\sigma_1^2, \sigma_2^2)$

Multivariate Gaussian Examples

$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$ $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

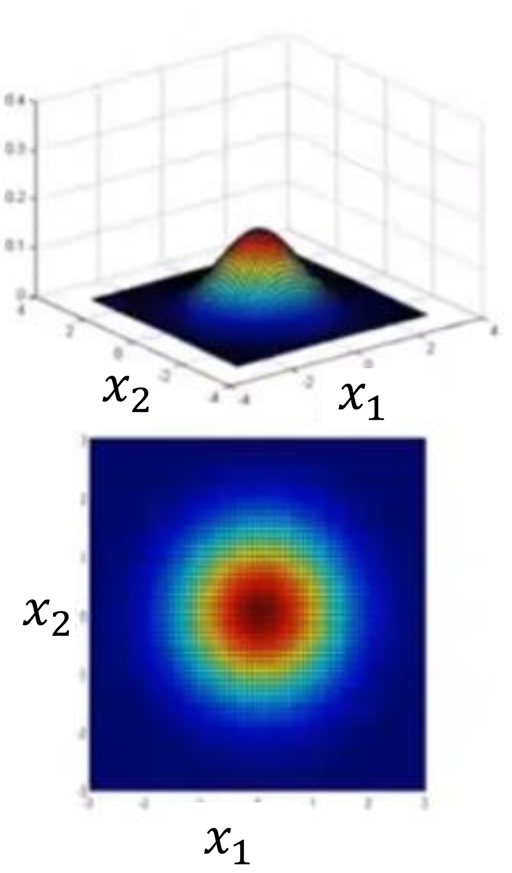


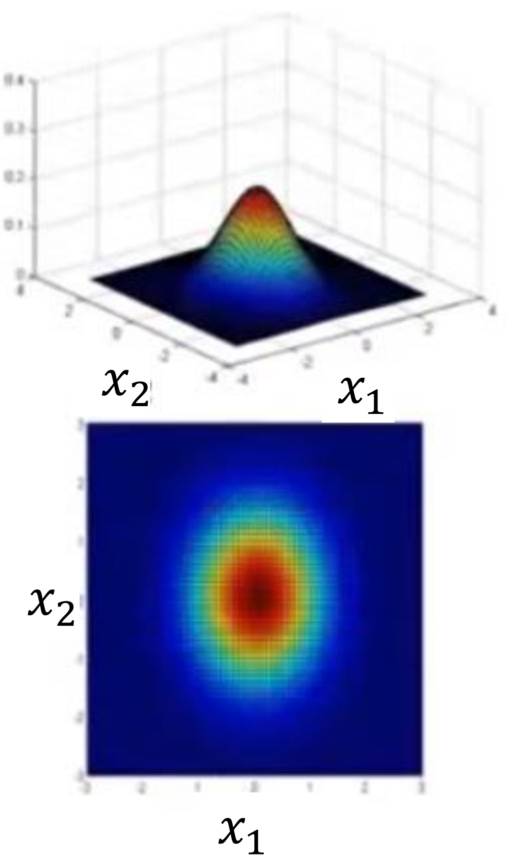


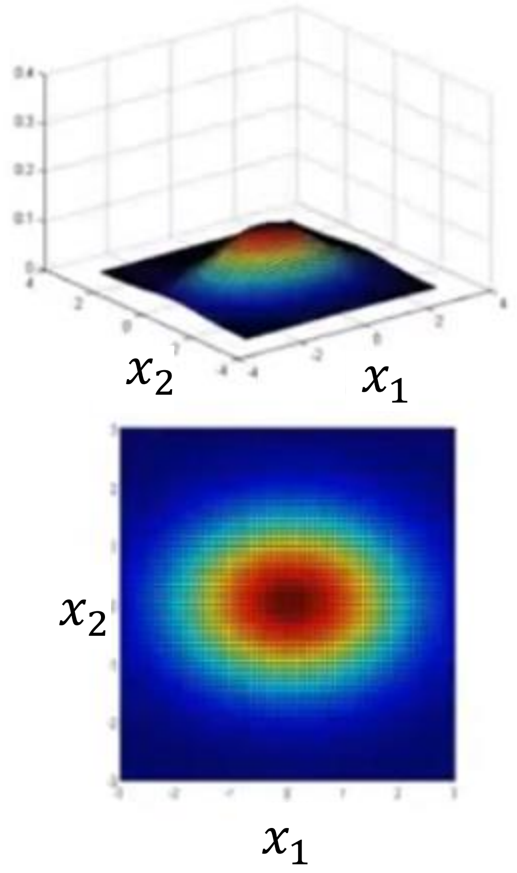


Multivariate Gaussian Examples

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

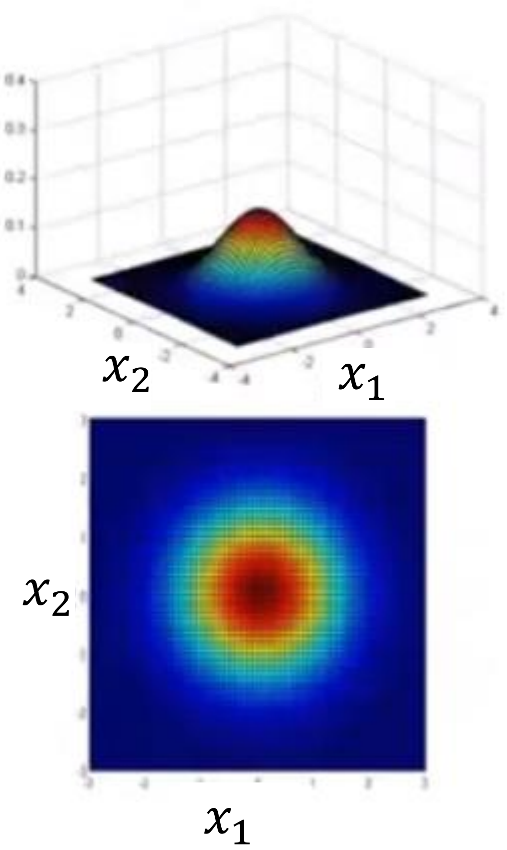


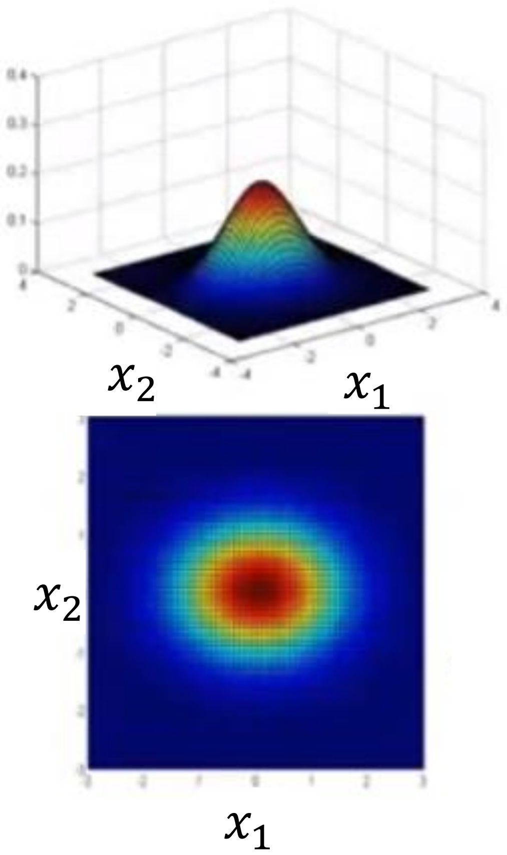


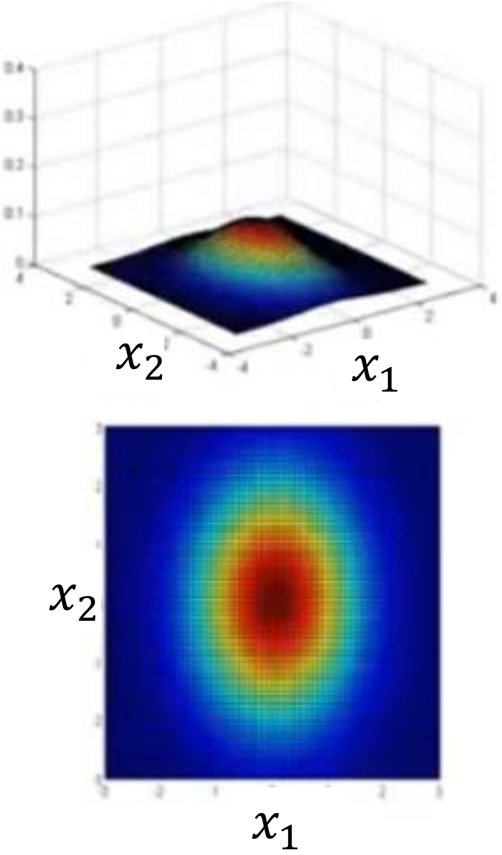


Multivariate Gaussian Examples

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 0.6 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

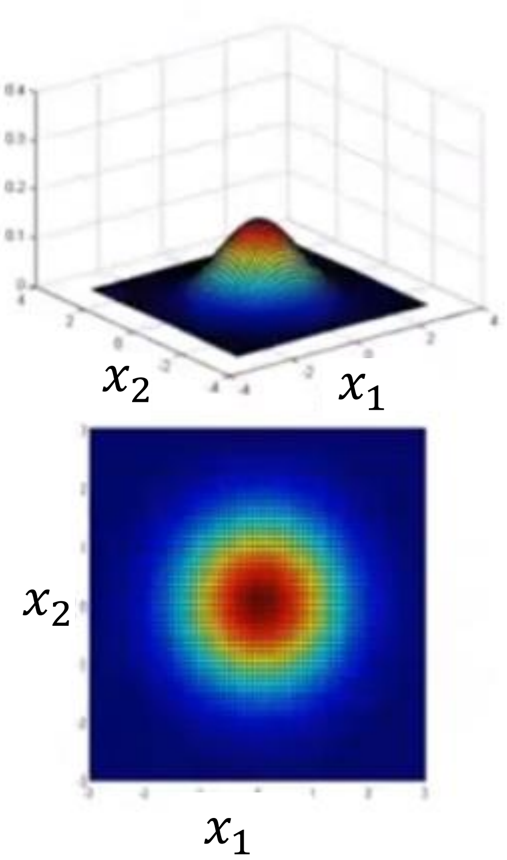






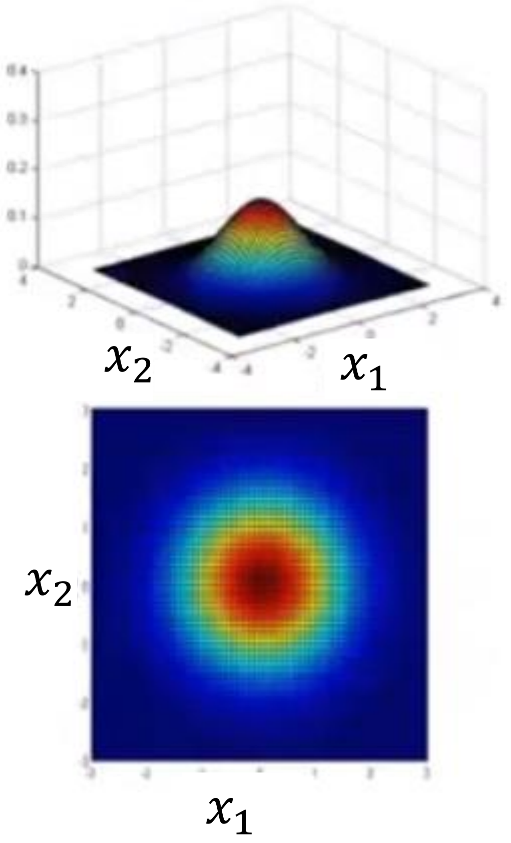
Multivariate Gaussian Examples

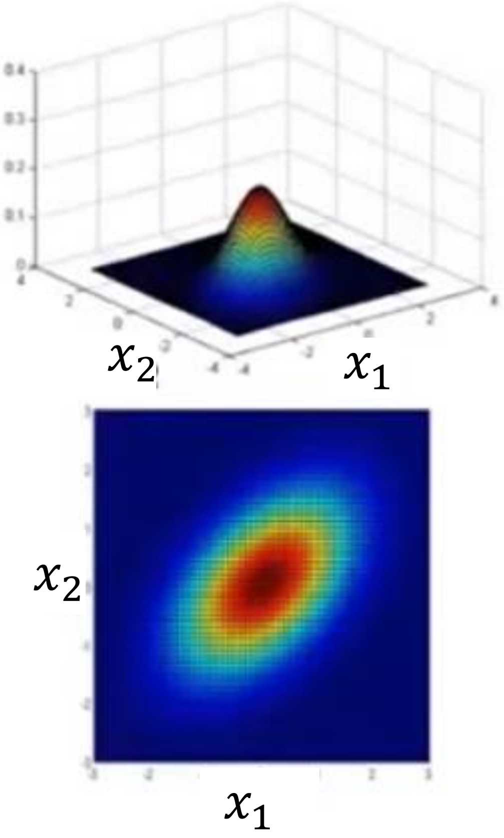
$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

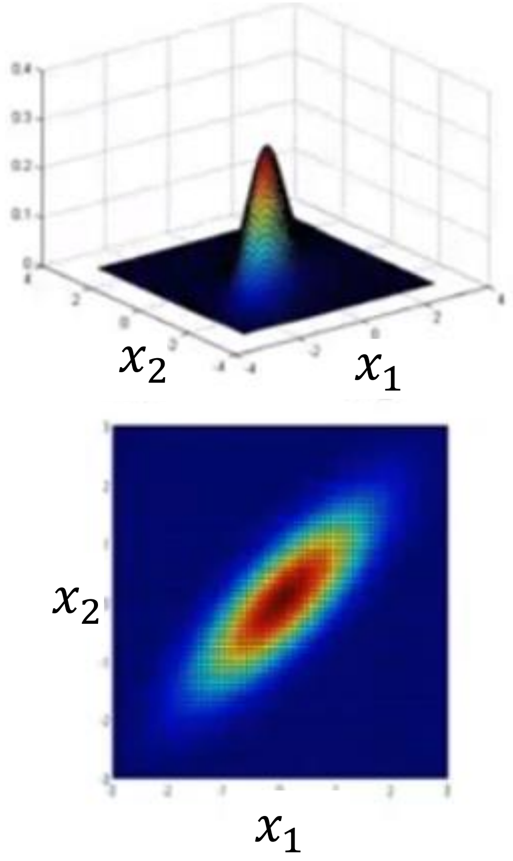


Multivariate Gaussian Examples

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$







Affine Transformation of Multivariate Gaussian

Theorem: If $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$ is an affine transformation of $\boldsymbol{X} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{A} \in \mathbb{R}^{M \times N}$, $\boldsymbol{b} \in \mathbb{R}^{M}$, then $\boldsymbol{Y} \sim \boldsymbol{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\boldsymbol{T}})$.

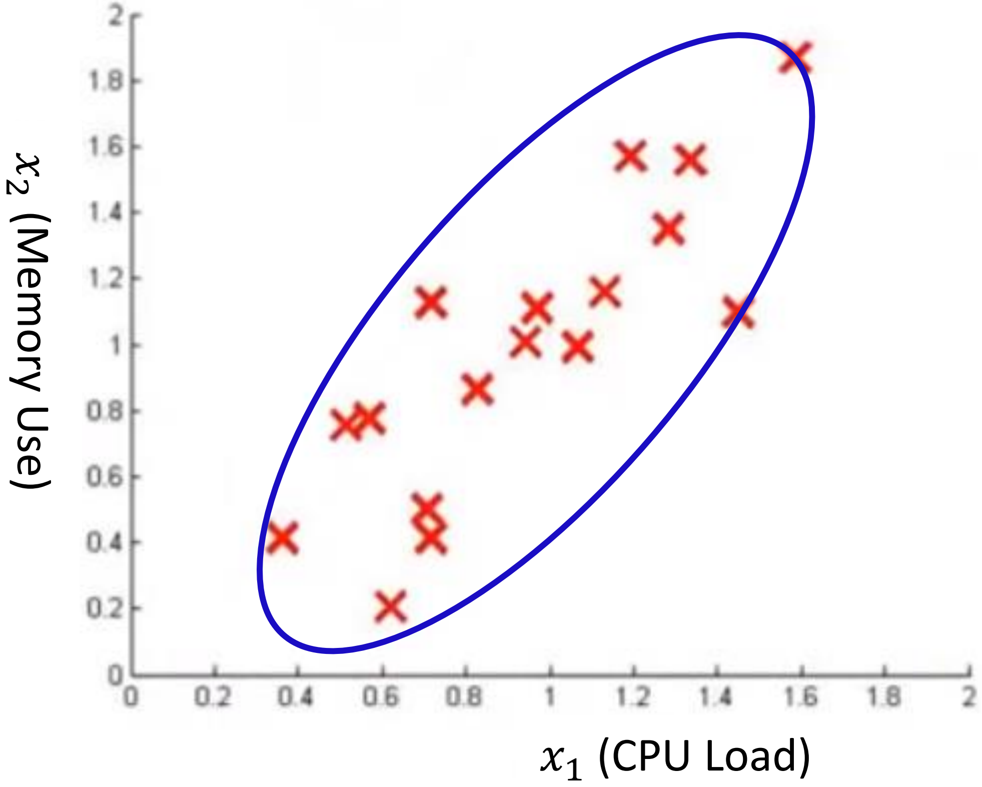
We would not prove this. **JUST REMEMBER.**

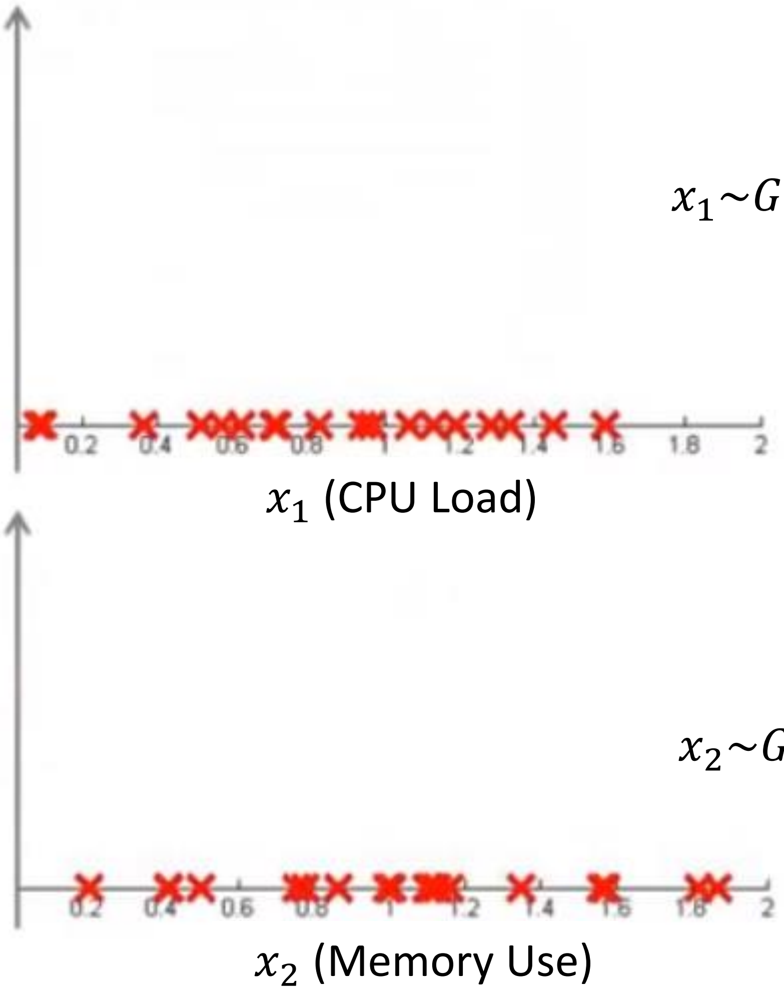
If $\boldsymbol{X} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{X} \in \mathbb{R}^{N}$, then

- **Q1**: What would the marginal pdf of multivariate Gaussian like?
  - E.g., $(X_1, X_2, X_4)^{\boldsymbol{T}} \sim ?$
- **Q2**: What would the conditional pdf of multivariate Gaussian like?
  - E.g.,$(X_1 | X_2 = x_2) \sim ?$

Marginal Pdf of the Multivariate Gaussian

Marginal pdf of the multivariate Gaussian is also Gaussian.
E.g., If $X = [x_1, x_2]$~Gaussian, then x_1~ Gaussian and x_2~ Gaussian.





$x_1 \sim Gaussian?$

$x_2 \sim Gaussian?$

Marginal Pdf of the Multivariate Gaussian

Theorem: If $\boldsymbol{Y} = \boldsymbol{AX} + \boldsymbol{b}$ is an affine transformation of $\boldsymbol{X} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{A} \in \mathbb{R}^{M \times N}$, $\boldsymbol{b} \in \mathbb{R}^{M}$, then $\boldsymbol{Y} \sim \boldsymbol{N}(\boldsymbol{A\mu} + \boldsymbol{b}, \boldsymbol{A\Sigma A}^{\boldsymbol{T}})$.

Given $\boldsymbol{X} \in \mathbb{R}^{N}$, let us see the marginal pdf of $(X_1, X_2, X_4)^{\boldsymbol{T}}$ (a subset of the X_i's).

Use the following $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{3 \times N}$$

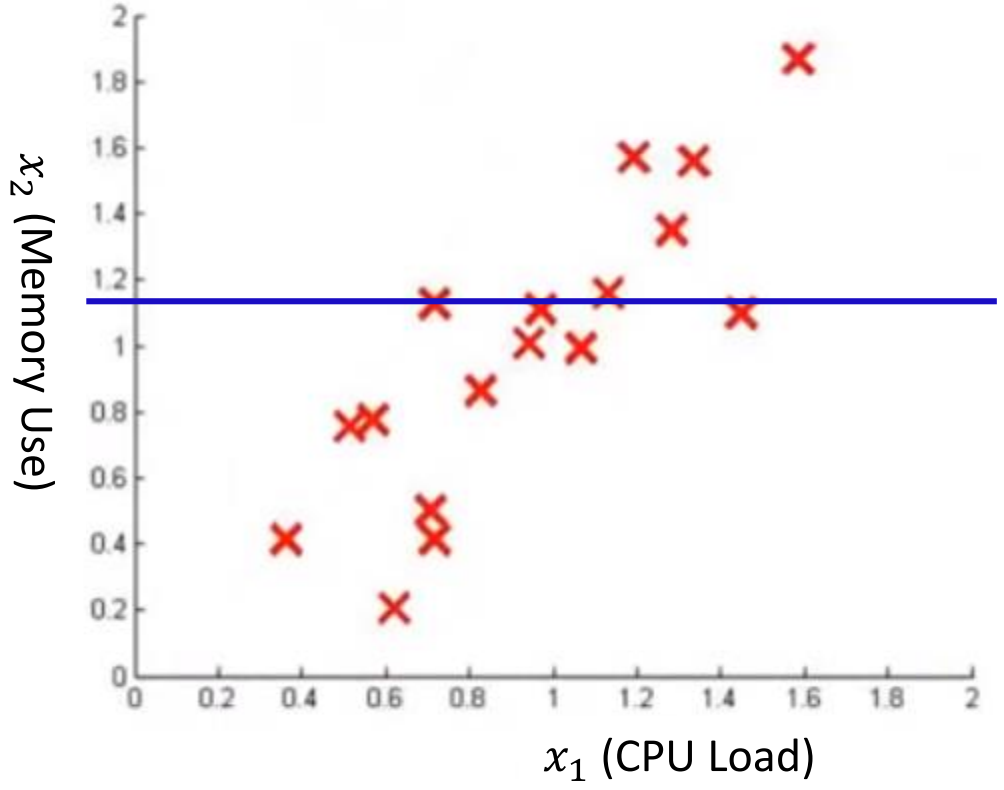
which extracts the desired elements directly!!!

Applying the above **Theorem**, we can say…

If $\boldsymbol{X} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any subset of the X_i's has a marginal distribution that is also multivariate normal.

Conditional Pdf of the Multivariate Gaussian

Conditional pdf of the multivariate Gaussian is also Gaussian.
E.g., If $X = [X_1, X_2]$~Gaussian, then.,$(X_1|X_2 = x_2)$~ Gaussian



Conditional Pdf of the Multivariate Gaussian

Theorem: Let $\boldsymbol{X} \in \mathbb{R}^N$, $\boldsymbol{X} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We do the partition as follows.

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}, \text{ where } \boldsymbol{X}_1 \in \mathbb{R}^q \text{ and } \boldsymbol{X}_2 \in \mathbb{R}^{N-q}.$$

Accordingly,

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

Then we have $(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{a}) \sim N(\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\Sigma}})$, where

$$\overline{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}_2), \qquad \overline{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Conditional Pdf of the Multivariate Gaussian

Theorem: Let $\boldsymbol{X} \in \mathbb{R}^N$, $\boldsymbol{X} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We do the partition as follows.

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}, \text{ where } \boldsymbol{X}_1 \in \mathbb{R}^q \text{ and } \boldsymbol{X}_2 \in \mathbb{R}^{N-q}.$$

Accordingly,

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

Then we have $(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{a}) \sim N(\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\Sigma}})$, where

$$\overline{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}_2), \qquad \overline{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

DIFFICULT

JUST REMEMBER.

Geometry of the Gaussian

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\boldsymbol{1}}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

Write the eigen-decomposition for $\boldsymbol{\Sigma} = \boldsymbol{V\Lambda V^T}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \uparrow & \uparrow & \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots \\ \downarrow & \downarrow & \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \end{bmatrix} \begin{bmatrix} \leftarrow \boldsymbol{v}_{\boldsymbol{1}}^{\boldsymbol{T}} \rightarrow \\ \leftarrow \boldsymbol{v}_{\boldsymbol{2}}^{\boldsymbol{T}} \rightarrow \\ \vdots \end{bmatrix}$$

$\boldsymbol{V}$ is orthonormal (i.e., $\boldsymbol{VV^T} = \boldsymbol{I}$)

Then we do the following transformation $\boldsymbol{y} = \boldsymbol{V^T x}$

Geometry of the Gaussian

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\boldsymbol{1}}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

Write the eigen-decomposition for $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\boldsymbol{T}}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \uparrow & \uparrow & \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots \\ \downarrow & \downarrow & \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \end{bmatrix} \begin{bmatrix} \leftarrow \boldsymbol{v}_{\boldsymbol{1}}^{\boldsymbol{T}} \rightarrow \\ \leftarrow \boldsymbol{v}_{\boldsymbol{2}}^{\boldsymbol{T}} \rightarrow \\ \vdots \end{bmatrix}$$

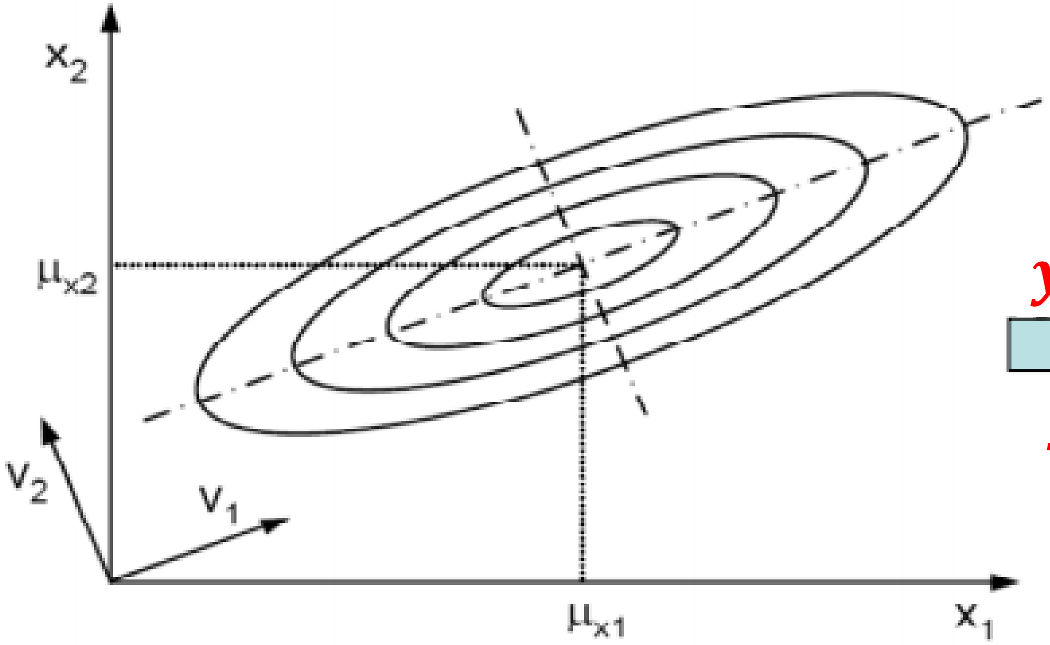
$\boldsymbol{V}$ is orthonormal (i.e., $\boldsymbol{V}\boldsymbol{V}^{\boldsymbol{T}} = \boldsymbol{I}$)

Then we do the following transformation $\boldsymbol{y} = \boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{x}$ Then $p(\boldsymbol{y}) = ?$

Geometry of the Gaussian

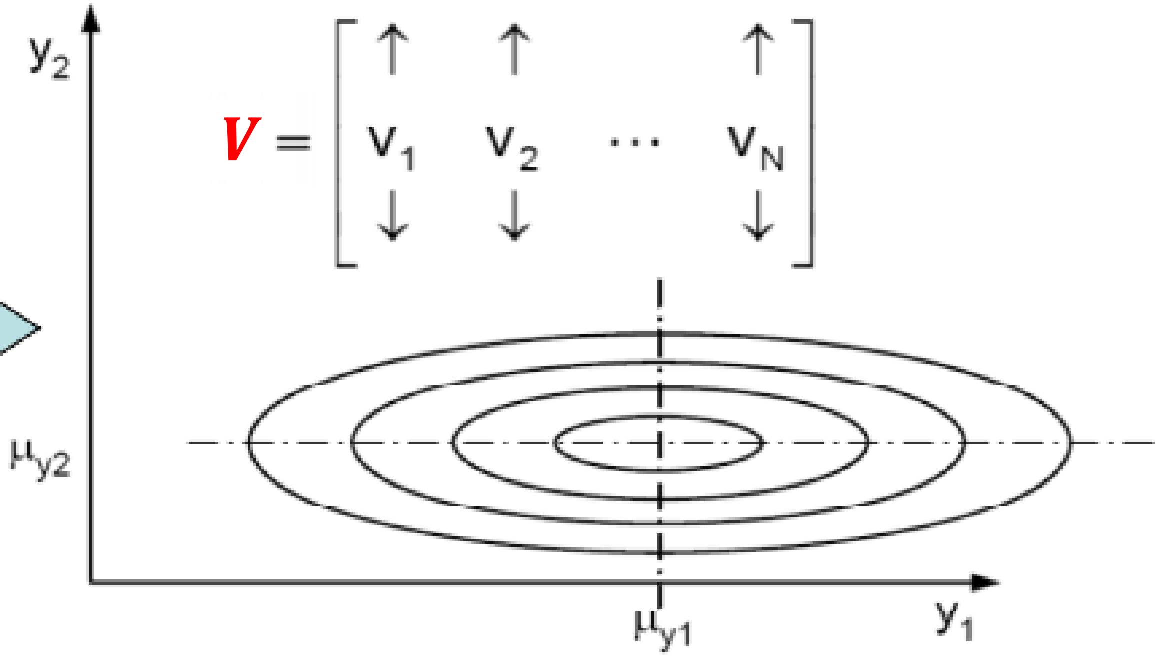
$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\, exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\boldsymbol{1}}(\boldsymbol{x}-\boldsymbol{\mu})\right] \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \uparrow & \uparrow & \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots \\ \downarrow & \downarrow & \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \end{bmatrix} \begin{bmatrix} \leftarrow \boldsymbol{v}_1^T \rightarrow \\ \leftarrow \boldsymbol{v}_2^T \rightarrow \\ \vdots \end{bmatrix}$$

$$p(\boldsymbol{y}) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\lambda_i}}\, exp\left[-\frac{\left(y_i-\mu_{y_i}\right)^2}{2\lambda_i}\right]$$



$$\boldsymbol{y} = \boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{x}$$

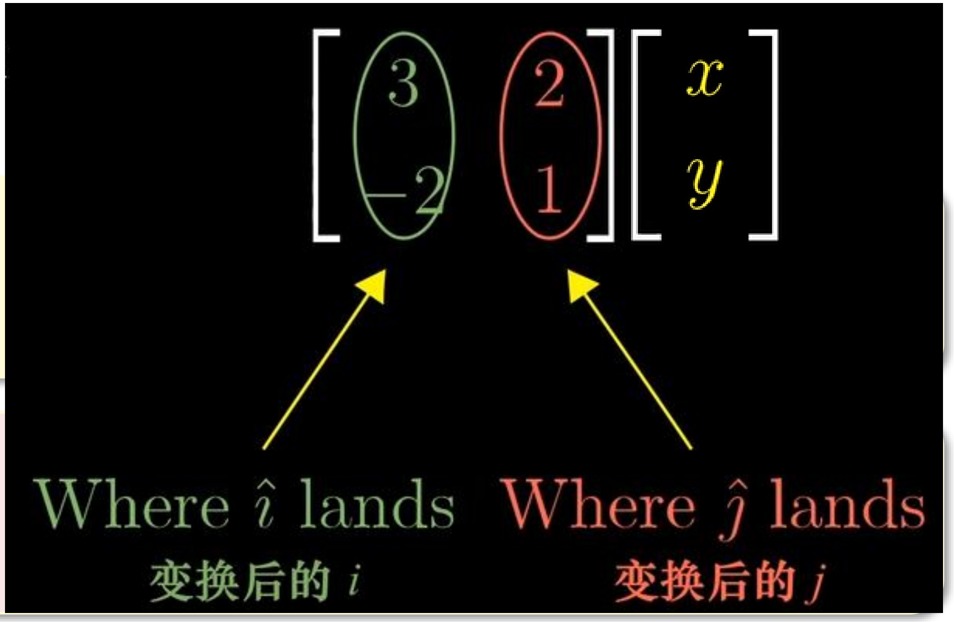
$$\boldsymbol{x} = \boldsymbol{V}\boldsymbol{y}$$



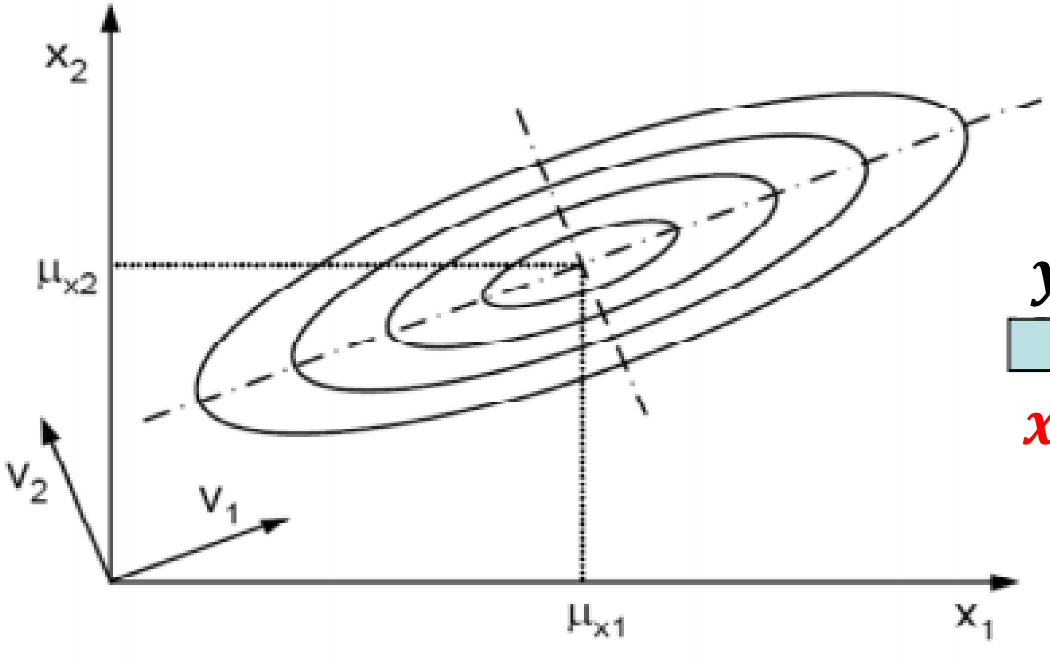
$$\boldsymbol{V} = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ v_1 & v_2 & \cdots & v_N \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

Geometry of the Gaussian

- Remember: matrix ↔ linear transformation.
- $\boldsymbol{y}$: before the transformation of $\boldsymbol{V}$. ($\boldsymbol{x}$: after)

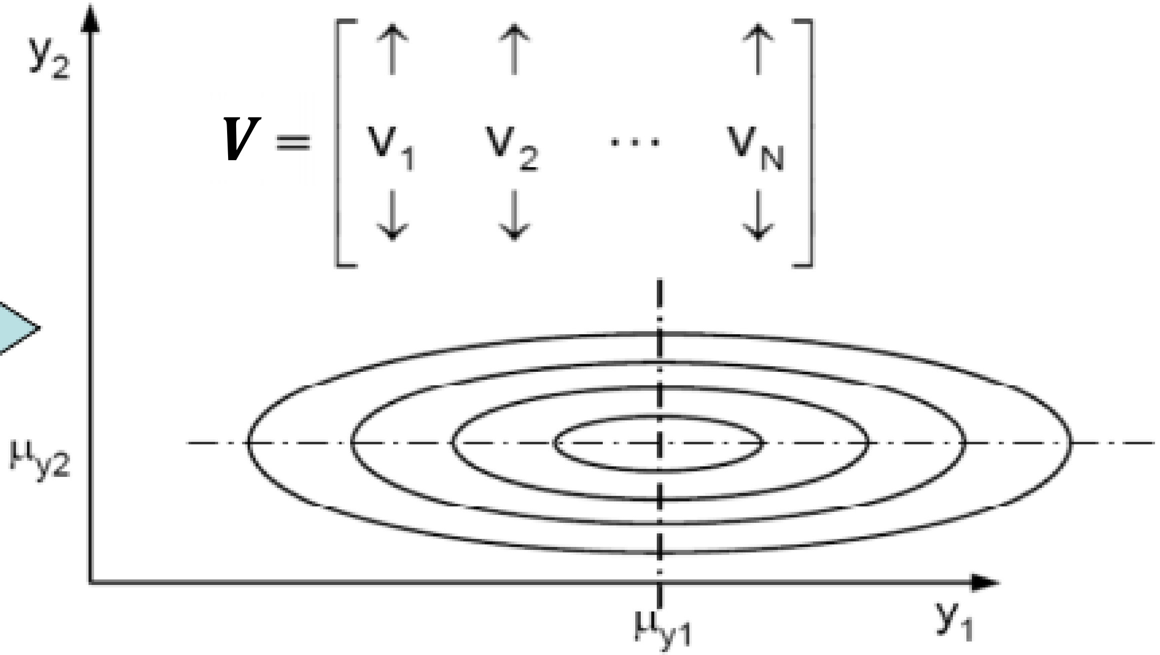

$\begin{bmatrix} 3 & 2 \\ -2 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$
Where $\hat{\imath}$ lands
变换后的 i
Where $\hat{\jmath}$ lands
变换后的 j

- ➢ Eigenvectors of $\boldsymbol{\Sigma}$ are the principle directions.
- ➢ Eigenvalues are the variances.



$$\boldsymbol{y} = \boldsymbol{V}^T\boldsymbol{x}$$

$$\boldsymbol{x} = \boldsymbol{V}\boldsymbol{y}$$



$$\boldsymbol{V} = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ v_1 & v_2 & \cdots & v_N \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

The Central Limit Theorem

- If $(X_1, X_2, \ldots, X_n)$ are independent and identically distributed (i.e., iid) continuous variables

- Define $Z = f(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$

- As $n \rightarrow$ infinity, $p(Z) \rightarrow$ Gaussian with mean $E[X_i]$ and variance $Var[X_i]/n$

- This explains the ubiquity (everywhere) of the normal probability distribution.

The Central Limit Theorem

- Flip the coin

$p(X = 1) = p; \;\; p(X = 0) = 1 - p$ **Bernoulli distribution**

$p(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$ **Binomial distribution**

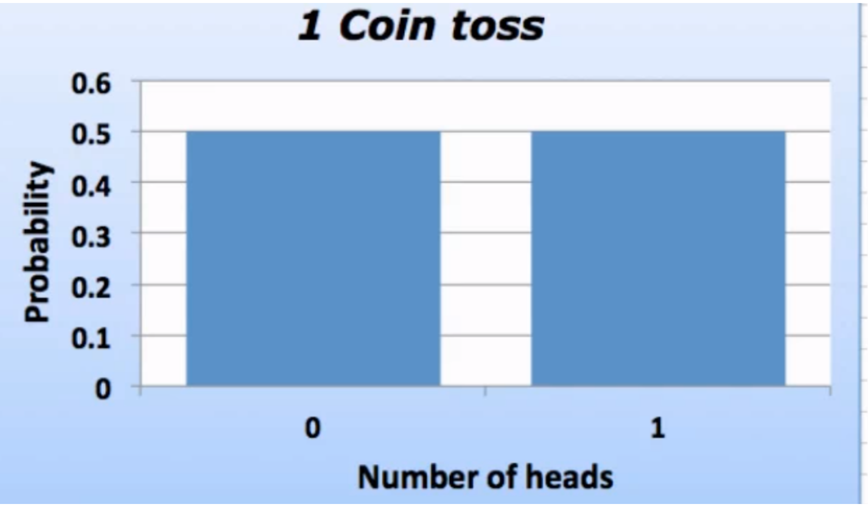
Z: the average number of heads.

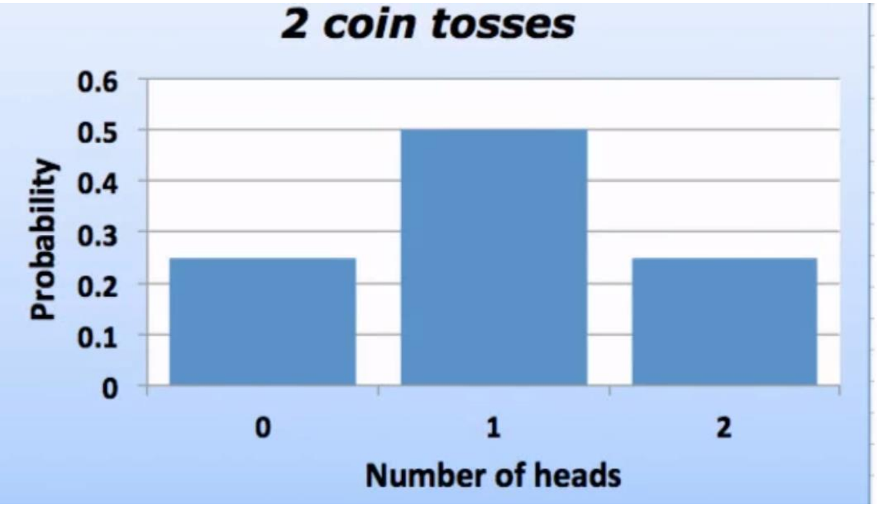
The Central Limit Theorem

$p(X=1)=p;\ \ p(X=0)=1-p$ **Bernoulli distribution**

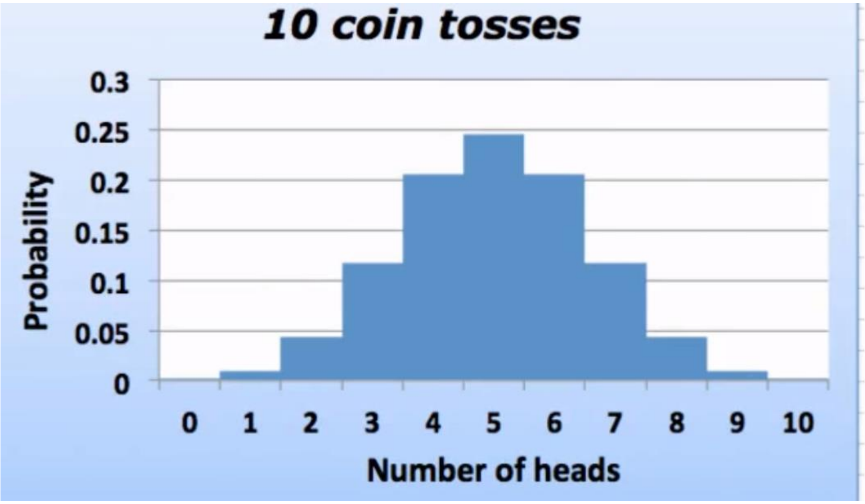
$p(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ **Binomial distribution**

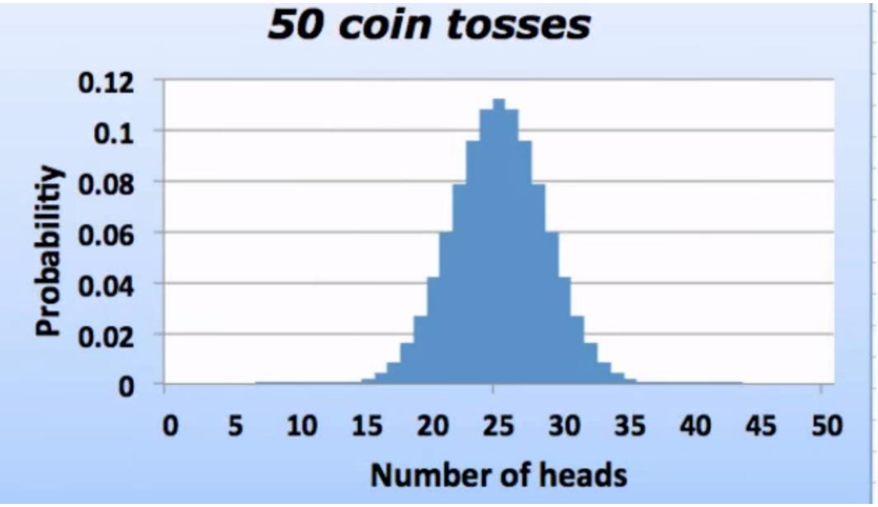
- Flip the coin





Z: the average (sum) number of heads.



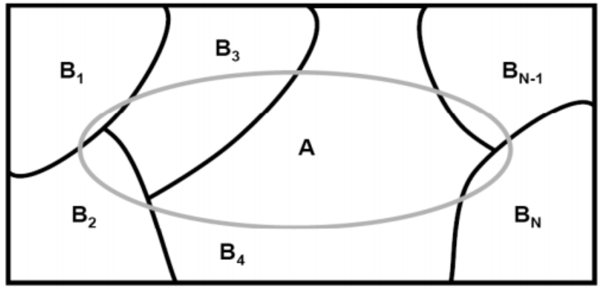


Why Gaussian

- **Analytical tractability**
  - $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are sufficient to uniquely characterize the distribution.
  - If (Gaussian) x_i's are mutually uncorrelated, then they are independent.
  - The marginal and conditional densities are also Gaussian.
  - Any linear transformation of any N jointly Gaussian RV's results in N RV's also Gaussian (affine transformation Theorem)

- **Ubiquity-Frequently observed**
  - Central limit theorem (Many distributions we wish to model are truly close to being normal distributions.

Summary



- **Bayesian Rule**
  - $P[B_j|A] = \frac{P[B_j \cap A]}{P[A]} = \frac{P[A|B_j]P[B_j]}{\sum_{k=1}^{N} P[A|B_k]P[B_k]}$
- **Covariance Matrix**
  - $\boldsymbol{COV}[\boldsymbol{X}] = \boldsymbol{\Sigma} = \boldsymbol{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^{\boldsymbol{T}}]$
  - Symmetric and Positive semi-definite
- **Uncorrelation VS. Independence**
  - Uncorrelated (linearly independent): $\mathrm{E}[\mathrm{x_i x_k}] = \mathrm{E}[\mathrm{x_i}]\mathrm{E}[\mathrm{x_k}]$
  - Independent : $\mathrm{P}[\mathrm{x_i} \cap \mathrm{x_k}] = \mathrm{P}[\mathrm{x_i}]\mathrm{P}[\mathrm{x_k}]$.
- **Multivariate Gaussian**
  - $\boldsymbol{\mu}$ = mean vector, $\boldsymbol{\Sigma}$ = covariance matrix
  - Geometry of the Gaussian
    - ✓ Eigenvectors of $\boldsymbol{\Sigma}$ are the principle directions.
    - ✓ Eigenvalues are the variances.
- **The Central Limit Theorem**