MODULE 1

PROBABILITY AND RANDOM VARIABLES
Probability Space

- A PROBABILITY SPACE is made up of a sample space, a set of events, and a probability measure.

- Central to the concept of probability space is an EXPERIMENT, the outcome of which cannot be known a priori with certainty.

- The SAMPLE SPACE $S$ is the set of all possible outcomes of the experiment.

  ▶ Ex: If the experiment is the single throwing of a fair die, then $S$ consists of the six possible outcomes:

  $\begin{array}{ccc}
  \cdot & \cdot & \cdot \\
  \cdot & \cdot & \cdot \\
  \cdot & \cdot & \cdot \\
  \cdot & \cdot & \cdot \\
  \cdot & \cdot & \cdot \\
  \cdot & \cdot & \cdot \\
  \end{array}$

  ▶ Ex: If the experiment is to measure the temperature in Norman at 2:00 in the afternoon, then $S = \mathbb{R}$, the set of real numbers.

- Note: the elements of $S$ must be disjoint. That is, for any trial of the experiment, the outcome can be one and only one element of $S$. 
• The SET OF EVENTS $\mathcal{G}$ is a set whose members are all possible subsets of $S$.

  ▶ The empty set $\emptyset$ and the set $S$ itself are both members of $\mathcal{G}$.

  ▶ Members of $\mathcal{G}$ can generally be constructed by taking unions, intersections, and complements of members of $S$.

  ▶ Formally, $\mathcal{G}$ is known as a $\sigma$-algebra, but we will not concern ourselves with the details of this.

Ex: The event “the temperature in Norman at 2:00 is in the interval $[72, 96)$ could be an event in $\mathcal{G}$ (temperature experiment).

Ex: The event NOT( (one dot) OR (two dots) ) could be an event in $\mathcal{G}$ (die experiment).

▶ Two events $A, B \in \mathcal{G}$ are mutually exclusive if $A \cap B = \emptyset$. 
• The PROBABILITY MEASURE $P$ is a function with domain is $\mathcal{S}$ and range $[0, 1]$.
  
  ▶ Intuitively, we think of the number $P(A)$ as the probability that the event $A$ occurs as the outcome of a random trial of the experiment.

• $P$ must satisfy the following three axioms:
  
  1. $P(A) \geq 0 \ \forall \ A \in \mathcal{S}$.
  2. $P(\emptyset) = 1$.
  3. If $\{A_i\}$ is a countable collection of mutually exclusive sets in $\mathcal{S}$, then
     
     $$P\left(\bigcup_i A_i\right) = \sum_i P(A_i).$$

• The following three properties are among the consequences of the three axioms:
  
  1. $P(A^C) = 1 - P(A)$.
  2. $P(\emptyset) = 0$.
  3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

• Together, $\mathcal{S}$, $\mathcal{G}$, and $P$ makeup the PROBABILITY SPACE $(\mathcal{S}, \mathcal{G}, P)$. 
Notes

- For the temperature experiment, suppose that $T$ is the measured temperature and that, on a given day, all temperatures between 75°F and 95°F are equally likely.

  - If $A$ is the event $T = 80°$ then $P(A) = 0$, since there are an infinite number of possible temperatures in the range $[75, 95]$.

  - For the event $A = \{T \in [75, 85]\}$, however, the probability is nonzero. In fact, $P(A) = 0.5$ in this case.

- For the die experiment, suppose that the die is fair, so that $P(\text{one dot}) = P(\text{two dots}) = \ldots = P(\text{six dots}) = \frac{1}{6}$.

  - Let $A = \text{odd number of dots} = \square \cup \square \cup \square$.
  - $B = \text{number of dots} < 4 = \square \cup \square \cup \square \cup \square$.
  - $P(A) = P(B) = \frac{1}{2}$
  - $P(A \cap B) = P(\square \cup \square) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)
  = \frac{1}{2} + \frac{1}{2} - \frac{1}{3} = \frac{2}{3}$
**Conditional Probability**

\[
P(B|A) = \frac{P(B \cap A)}{P(A)}
\]

**Bayes' Theorem**:

\[
P(B|A) = \frac{P(A|B)P(B)}{P(A)}
\]

**Theorem of Total Probability**:

Let \( \{A_i\} \) be exhaustive \((\forall A_i = \Omega)\) and mutually exclusive. Then

\[
P(B) = \sum_i P(B|A_i)P(A_i)
\]
COROLLARY: \[ P(A_k|B) = \frac{P(B|A_k) P(A_k)}{\sum_i P(B|A_i) P(A_i)} \]

DEF: Events A and B are called independent if \[ P(A \cap B) = P(A)P(B). \]

This implies that \[ P(A|B) = P(A), \]
\[ P(B|A) = P(B). \]

Events that are not independent are called dependent.

RANDOM VARIABLES
- A random variable (RV) is a function with domain \( S \) and range \( \mathbb{R} \).
- An RV "maps" the experimental outcomes to real numbers.

EX:

- Diagram showing mapping from \( S \) to \( \mathbb{R} \) with outcomes 1 to 6.
Once the correspondence between experimental outcomes and real numbers has been made using an RV, the original sample space can often be ignored.

This gives us a consistent way to treat a variety of experiments with a common mathematical framework.

An RV for a sample space with a finite number of experimental outcomes is called a discrete RV.

An RV for a sample space with an infinite number of experimental outcomes is called a continuous RV.

Cumulative Distribution Function (CDF) for Continuous RV

For an RV $X$, the CDF is defined as

$$ F_X(x) = P(X \leq x) $$

The CDF is also referred to simply as the probability "distribution" function.

Properties:

1. $\lim_{x \to -\infty} F_X(x) = 0$

2. $\lim_{x \to \infty} F_X(x) = 1$

3. $F_X(x)$ is nondecreasing in $x$. 

4. $P(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$. 

PAGE 17
Probability Density Function (pdf) for Continuous RV

- The pdf, or probability "density" is defined by
  \[ f_x(x) = \frac{d}{dx} F_x(x). \]

Properties:
1. \[ F_x(x) = \int_{-\infty}^{x} f_x(\theta) d\theta \]
2. \[ P(x_1 < x \leq x_2) = F_x(x_2) - F_x(x_1) = \int_{x_1}^{x_2} f_x(\theta) d\theta \]
3. \[ \int_{-\infty}^{\infty} f_x(\theta) d\theta = 1 \]
4. \[ f_x(x) > 0. \]

Expected Value (mean) for a Continuous RV

- The mean of an RV \( X \) is written \( \bar{X} \), \( E[X] \), \( EX \), \( \langle x \rangle \), \( \mu \), or \( \mu \).
- By definition,
  \[ E[X] = \int_{-\infty}^{\infty} \theta f_x(\theta) d\theta. \]

Functions of a Continuous RV

- Suppose \( g : \mathbb{R} \to \mathbb{R} \) is a function.
- Then we can consider \( y = g(X) \), a function of the RV \( X \).
- The RV \( X \) maps an experimental outcome \( x \in \mathcal{S} \) to a number \( x \in \mathbb{R} \). The function \( g \) then maps \( x \) to a second number \( y \in \mathbb{R} \).
In symbols:

\[ X \xrightarrow{g} y \]

The expected value, or mean, of \( g(X) \) is given by

\[ E[g(X)] = \int_{-\infty}^{\infty} g(\theta)f_x(\theta)\,d\theta. \]

**Note:** Since the expectation integral is a linear operator, we have that, with \( \alpha, \beta \in \mathbb{R} \),

\[ E[\alpha g_1(X) + \beta g_2(X)] = \alpha E[g_1(X)] + \beta E[g_2(X)]. \]

**Moments of a Continuous RV**

- The "\( k^{th} \) moment" of an RV \( X \) is defined by

\[ E[X^k] = \int_{-\infty}^{\infty} \theta^k f_x(\theta)\,d\theta. \]

- The zeroth moment is always equal to 1.
- The first moment is identical to the mean.

**Central Moments of a Continuous RV**

- The "\( k^{th} \) central moment of the RV \( X \) is the expected value of the function \( g(x) = (x - \bar{x})^k \):

\[ E[(x - \bar{x})^k] = \int_{-\infty}^{\infty} (\theta - \bar{x})^k f_x(\theta)\,d\theta \]
Notes:

- The zeroth central moment is always equal to 1.
- The first central moment is always equal to zero.
- The second central moment is called the "variance" $\sigma_x^2$:

$$\sigma_x^2 = E[(x - \bar{x})^2] = \int_{-\infty}^{\infty} (\theta - \bar{x})^2 f_x(\theta) d\theta$$

$$= E[x^2] - (E[x])^2.$$

- The square root of the variance is called the "standard deviation" $\sigma_x$:

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{E[(x - \bar{x})^2]}$$

- The third central moment divided by the cube of the standard deviation is called the "skewness" of $X$:

$$\text{skewness} = \frac{E[(x - \bar{x})^3]}{\sigma_x^3}$$

- The fourth central moment divided by the square of the variance is called the "kurtosis" of $X$:

$$\text{kurtosis} = \frac{E[(x - \bar{x})^4]}{\sigma_x^4}$$
Characteristic Function of a Continuous RV

- The characteristic function of an RV $X$ is given by
  \[ \psi_X(\omega) = \int_{-\infty}^{\infty} f_X(x) e^{i\omega x} \, dx = E[e^{i\omega X}] \]

  Note:
  \[ \psi_X(-\omega) = \mathcal{F}\{f_X(x)\} \quad ; \quad f_X(x) = \mathcal{F}^{-1}\{\psi_X(-\omega)\} \]

- The pdf can be recovered from $\psi_X(\omega)$ by
  \[ f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi_X(\omega) e^{-i\omega x} \, d\omega \]

Moment Generating Function of a Continuous RV

- We can expand the characteristic function in a power series:
  \[ \psi_X(\omega) = E\left[ \sum_{k=0}^{\infty} \frac{(i\omega x)^k}{k!} \right] = \sum_{k=0}^{\infty} \frac{(i\omega)^k}{k!} E[X^k] \]

- Furthermore,
  \[ \frac{d^k}{dw^k} \psi_X(\omega) \bigg|_{w=0} = \left[ \frac{d^k}{dw^k} \int_{-\infty}^{\infty} f_X(x) e^{i\omega x} \, dx \right]_{w=0} \]
  \[ = \left[ \int_{-\infty}^{\infty} (i\omega)^k f_X(x) e^{i\omega x} \, dx \right]_{w=0} \]
  \[ = i^k \int_{-\infty}^{\infty} x^k f_X(x) \, dx \]
  \[ = i^k E[X^k] \]
The function \( \phi_X(\omega) = \gamma_X(\omega \frac{\omega}{\theta}) \) is called the "moment generating function" of the RV \( X \).

The moments of \( X \) can be obtained from the moment generating function by

\[
E[X^k] = \left[ \frac{d^k}{d\omega^k} \phi_X(\omega) \right]_{\omega=0}
\]

---

**Gaussian or "Normal" Random Variable**

- The Gaussian distribution is used often because it approximately describes a large class of naturally occurring random phenomena.

- The terms "Gaussian" and "Normal" are used interchangeably.

- A Gaussian variable is described by two parameters:
  - The mean \( \mu \).
  - The variance \( \sigma^2 \).

- The pdf of a Gaussian variable \( X \) is given by

\[
f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right]
\]
The cdf of a Gaussian variable is given by

\[ F_x(x) = \int_{-\infty}^{x} f_x(x) \, dx \]

\[ = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right] \, dx \]

\[ = \int_{-\infty}^{(x-\mu)/\sigma} \frac{1}{\sqrt{2\pi} \sigma} \exp \left[ -\frac{1}{2} \theta^2 \right] \, d\theta \]

\[ = \Phi \left( \frac{x-\mu}{\sigma} \right), \]

where \( \Phi(x) \) is the cdf of a Gaussian variable with zero mean and unit variance.

The function \( \Phi(x) \) cannot be expressed in terms of elementary functions. To evaluate it, we must use numerical integration or tables.

Be very careful when using tables. Sometimes they are normalized in strange ways or specified in terms of the closely related "error function" \( \Phi_e(x) \).

The shorthand notation \( X \sim N(\mu, \sigma^2) \) is often used to indicate that the RV \( X \) is normally distributed with mean \( \mu \) and variance \( \sigma^2 \).

The characteristic function for a \( N(\mu, \sigma^2) \) variable is given by

\[ \phi_x(\omega) = \exp \left[ j\mu \omega - \frac{1}{2} \sigma^2 \omega^2 \right] \]
Discrete Random Variables

- Let \((\mathcal{S}, \mathcal{G}, P)\) be a probability space.
- Suppose that the number of elements in \(\mathcal{S}\) is finite. The RV \(X\) maps the outcomes to a finite set of numbers \([x_1, x_2, \ldots, x_n]\)
  where \(x_1 < x_2 < \ldots < x_n\).
- Then \(X\) is a discrete RV.
- Let \(P(x_k) = p_k\).
- Then \(\sum_{k=1}^{n} p_k = 1\).
- The cdf is given by

\[
F_X(x) = \begin{cases} 
0 & , x < x_1 \\
\sum_{k=1}^{i} p_k & , x_1 \leq x < x_{i+1} \\
1 & , x > x_n 
\end{cases}
\]

\[
F_X(x) \quad \begin{array}{c}
p_1 \{ \} \\
p_2 \{ \} \\
p_3 \{ \} \\
p_n \{ 1 \}
\end{array}
\]

\[
\begin{array}{c}
x_1 \\
x_2 \\
x_3 \\
x_{n-1} \\
x_n
\end{array}
\]
Taking the derivative of the cdf, we obtain a sum of weighted Dirac deltas for the pdf:

\[ f_X(x) = \frac{d}{dx} F_X(x) = \sum_{k=1}^{n} p_k \delta(x-x_k) \]

\[ f_X(x) \]

- The mean is given by

\[ E[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \sum_{k=1}^{n} p_k x_k \]

- The variance is given by

\[ \sigma^2 = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x) dx = \sum_{k=1}^{n} (x_k - E[X])^2 p_k \]

- The characteristic function is given by

\[ \phi_X(\omega) = \sum_{k=1}^{n} p_k e^{i\omega x_k} \]
The expected value of the function $g(x)$ is given by

$$E[g(x)] = \int_{-\infty}^{\infty} g(x) f_X(x) \, dx = \sum_{k=1}^{n} g(x_k) p_k$$

**Mixed Random Variables**

- Suppose $X$ is an RV that takes uncountably many values $x \in \mathbb{R}$, but that a finite set of these values $\{x_1, x_2, \ldots, x_n\}$ each occur with nonzero probability.

- Then $X$ is a \textit{“mixed”} RV. The cdf and pdf have characteristics of both continuous and discrete RVs.

- Let $P(x_k) = p_k$.

- Note that $P(x) = 0$ for $x \neq x_k$, $k = 1, 2, \ldots, n$.

- In this case, the cdf $F_X(x)$ has a step discontinuity of height $p_k$ at $x_k$:

- The pdf has a Dirac delta of weight $p_k$ at $x_k$:
Conditional Density

- The conditional distribution for the RV $X$ given that event $A$ has occurred is given by

$$F_{X|A}(x) = P(X \leq x | A)$$

- The corresponding conditional density is given by

$$f_{x|A}(x) = \frac{d}{dx} F_{x|A}(x)$$

- The corresponding event-on-density conditional probability is given by

$$P(A | X=x) = \lim_{\Delta x \to 0} P(A | x \leq X \leq x+\Delta x)$$

- The Bayes' formula relating the conditional density and conditional probability is

$$f_{x|A}(x) = \frac{P(A | X=x) f_x(x)}{P(A)}$$
Multiple Random Variables

- "Bivariate" refers to a situation involving two RVs.
- "Multivariate" refers to a situation involving two or more RVs.

Joint Distribution:

- The joint cdf of the n RVs $X_1, X_2, \ldots, X_n$ is given by

$$F_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = P(X_1 \leq x_1 \cap X_2 \leq x_2 \cap \cdots \cap X_n \leq x_n)$$

- Properties:

$$\lim_{x_1, x_2, \ldots, x_n \to -\infty} F_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = 0$$

$$\lim_{x_1, \ldots, x_n \to \infty} F_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = 1$$

Joint Density:

- The joint pdf of the n RVs $X_1, X_2, \ldots, X_n$ is given by

$$f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F_{X_1, \ldots, X_n}(x_1, \ldots, x_n)$$
- The joint CDF can be recovered from the joint PDF by integration:

\[ F_{x_1, \ldots, x_n}(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f_{x_1, \ldots, x_n}(\theta_1, \ldots, \theta_n) \, d\theta_1 \cdots d\theta_n \]

- The joint PDF integrates to 1:

\[ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{x_1, \ldots, x_n}(\theta_1, \ldots, \theta_n) \, d\theta_1 \cdots d\theta_n = 1 \]

- The marginal density of the variable \( X_k \) may be obtained from the joint PDF by "integrating out" the other variables:

\[ f_{x_k}(x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{x_1, \ldots, x_n}(x_1, \ldots, x_n) \, dx_1 \cdots dx_{k-1} \, dx_{k+1} \cdots dx_n \]

\[ \text{integrals} \]

(\( dx_k \) missing)

- The marginal CDF for \( X_k \) is then obtained by

\[ F_{x_k}(x_k) = \int_{-\infty}^{x_k} f_{x_k}(\theta) \, d\theta \]
The variables $X_1, \ldots, X_n$ are called mutually independent if

$$F_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = \prod_{k=1}^{n} F_{X_k}(x_k)$$

This is equivalent to

$$f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = \prod_{k=1}^{n} f_{X_k}(x_k)$$

i.e., they are independent if the joint cdf is the product of the marginal cdfs and the joint pdf is the product of the marginal pdfs.

Note: For discrete RVs, all of the above multivariate formulas involving integrals reduce to sums.

**Functions of Multiple RVs**

- Suppose $g: \mathbb{R}^n \rightarrow \mathbb{R}$ is a function.
- Let $X_1, \ldots, X_n$ be $n$ RVs with joint cdf $F_{X_1, \ldots, X_n}(x_1, \ldots, x_n)$ and joint pdf $f_{X_1, \ldots, X_n}(x_1, \ldots, x_n)$.
- Then $g(X_1, \ldots, X_n)$ is a function of the $n$ RVs that takes scalar values in $\mathbb{R}$. 
- The expected value of $g(x_1, \ldots, x_n)$ is

$$E[g(x_1, \ldots, x_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f_{x_1, \ldots, x_n}(x_1, \ldots, x_n) \, dx_1 \cdots dx_n$$

- If the RVs are mutually independent and $g$ is separable so that $g(x_1, \ldots, x_n) = \prod_{k=1}^{n} g_k(x_k)$, then

$$E[g(x_1, \ldots, x_n)] = \prod_{k=1}^{n} E[g_k(x_k)]$$

**EX:** Suppose $g(x_1, \ldots, x_n) = \sum_{k=1}^{n} \alpha_k x_k$ and $x_1, \ldots, x_n$ are mutually independent. Then

$$E[g(x_1, \ldots, x_n)] = E\left[ \sum_{k=1}^{n} \alpha_k x_k \right]$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{k=1}^{n} \alpha_k x_k f_{x_1, \ldots, x_n}(x_1, \ldots, x_n) \, dx_1 \cdots dx_n$$

$$= \sum_{k=1}^{n} \alpha_k \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_k f_{x_1, \ldots, x_n}(x_1, \ldots, x_n) \, dx_1 \cdots dx_n$$

$$= \sum_{k=1}^{n} \alpha_k \int_{-\infty}^{\infty} x_k f_{x_k}(x_k) \, dx_k \prod_{l=1}^{n} \int_{-\infty}^{\infty} f_{x_l}(x_l) \, dx_l$$

$$= \sum_{k=1}^{n} \alpha_k \int_{-\infty}^{\infty} x_k f_{x_k}(x_k) \, dx_k$$

$$= \sum_{k=1}^{n} \alpha_k E[X_k]$$.
- We can also consider a vector-valued function $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$.

- In this case, $g(x_1, \ldots, x_n)$ is the vector

$$g(x_1, \ldots, x_n) = [y_1 \ldots y_m]^T = \begin{bmatrix} g_1(x_1, \ldots, x_n) \\
g_2(x_1, \ldots, x_n) \\
\vdots \\
g_m(x_1, \ldots, x_n) \end{bmatrix}$$

- To find the expected value of $g(x_1, \ldots, x_n)$, we take the expectation on an element-by-element basis, and $E[g(x_1, \ldots, x_n)]$ is the vector

$$E[g(x_1, \ldots, x_n)] = E \begin{bmatrix} g_1(x_1, \ldots, x_n) \\
g_2(x_1, \ldots, x_n) \\
\vdots \\
g_m(x_1, \ldots, x_n) \end{bmatrix}$$

$$= \begin{bmatrix} E[g_1(x_1, \ldots, x_n)] \\
E[g_2(x_1, \ldots, x_n)] \\
\vdots \\
E[g_m(x_1, \ldots, x_n)] \end{bmatrix}$$
Joint Moments

- The joint moment of order \( k_1, \ldots, k_n \) of the \( n \) RVS \( X_1, \ldots, X_n \) is given by

\[
E \left[ x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n} \right] = E \left[ \prod_{l=1}^{n} x_l^{k_l} \right] \\
= \int_{\mathbb{R}^n} \prod_{l=1}^{n} x_l^{k_l} f_{x_1, \ldots, x_n}(x_1, \ldots, x_n) \, dx_1 \cdots dx_n
\]

Joint Central Moments

- The joint central moment of order \( k_1, \ldots, k_n \) of the \( n \) RVS \( X_1, \ldots, X_n \) is given by

\[
E \left[ \prod_{l=1}^{n} (x_l - \overline{x}_l)^{k_l} \right] \\
= \int_{\mathbb{R}^n} \prod_{l=1}^{n} (x_l - \overline{x}_l)^{k_l} f_{x_1, \ldots, x_n}(x_1, \ldots, x_n) \, dx_1 \cdots dx_n
\]
Covariance and Correlation for Two RVs

- Let $X_1$ and $X_2$ be two RVs with joint pdf $f_{X_1, X_2}(x_1, x_2)$.

- The covariance of $X_1$ and $X_2$ is the joint central moment
  \[ \text{Cov}(X_1, X_2) = \iint_{\mathbb{R}^2} (x_1 - \overline{x}_1)(x_2 - \overline{x}_2) f_{X_1, X_2}(x_1, x_2) \, dx_1 \, dx_2 \]

- Using the Cauchy–Schwarz inequality, it is easy to show that
  \[ |\text{Cov}(X_1, X_2)| \leq \sigma_{X_1} \sigma_{X_2} \]

- The correlation coefficient between $X_1$ and $X_2$ is given by
  \[ \rho_{X_1, X_2} = \frac{\text{Cov}(X_1, X_2)}{\sigma_{X_1} \sigma_{X_2}} \]

- Note that $|\rho_{X_1, X_2}| \leq 1$.

  \[ \Rightarrow \] if $\rho = 1$, then $X_1 = \alpha X_2$ with $\alpha > 0$. (Perfectly correlated)

  \[ \Rightarrow \] if $\rho = -1$, then $X_1 = \alpha X_2$ with $\alpha < 0$. In this case, $X_1$ and $X_2$ are said to be anti-correlated.

  \[ \Rightarrow \] if $\rho = 0$, then $X_1$ and $X_2$ are said to be uncorrelated.
Notes:

- In general, \( \text{Cov}(X_1, X_2) = E[X_1X_2] - E[X_1]E[X_2] \)

- If \( X_1 \) and \( X_2 \) are independent, then

\[
E[X_1X_2] = \int_{\mathbb{R}^2} x_1x_2 f_{x_1,x_2}(x_1,x_2) \, dx_1 \, dx_2
\]

\[
= \int_{\mathbb{R}^2} x_1x_2 f_{x_1}(x_1)f_{x_2}(x_2) \, dx_1 \, dx_2
\]

\[
= \int_{\mathbb{R}} x_1 f_{x_1}(x_1) \, dx_1 \int_{\mathbb{R}} x_2 f_{x_2}(x_2) \, dx_2
\]

\[
= \overline{x_1} \overline{x_2}
\]

\[\Rightarrow\] So independence implies \( \text{Cov}(X_1, X_2) = 0 \)

\[\Rightarrow\] This implies \( \rho_{x_1, x_2} = 0 \)

\[\Rightarrow\] So independence implies uncorrelated.

\[\rightarrow\star\text{ The converse is not true.}\]

**DEF:** if \( E[X_1X_2] = 0 \), then \( X_1 \) and \( X_2 \) are called **orthogonal**.

- If \( X_1 \) and \( X_2 \) are orthogonal and if \( \overline{x_1} = 0 \) or \( \overline{x_2} = 0 \), then \( X_1 \) and \( X_2 \) are **uncorrelated**.
Covariance for Random Vectors

- Suppose \( \tilde{X} \) and \( \tilde{Y} \) are random vectors (vectors of RVs) with

\[
\tilde{X} = [x_1, \ldots, x_n]^T \quad \text{and} \quad \tilde{Y} = [y_1, \ldots, y_m]^T.
\]

- The covariance of \( \tilde{X} \) and \( \tilde{Y} \) is then a covariance matrix defined by

\[
\text{Cov}(\tilde{X}, \tilde{Y}) = \begin{bmatrix}
\text{Cov}(x_1, y_1) & \text{Cov}(x_1, y_2) & \cdots & \text{Cov}(x_1, y_m) \\
\text{Cov}(x_2, y_1) & \text{Cov}(x_2, y_2) & \cdots & \text{Cov}(x_2, y_m) \\
\vdots & \vdots & \ddots & \vdots \\
\text{Cov}(x_n, y_1) & \text{Cov}(x_n, y_2) & \cdots & \text{Cov}(x_n, y_m)
\end{bmatrix}
\]

\[
= \mathbb{E}[(\tilde{X} - \mathbb{E}[\tilde{X}])(\tilde{Y} - \mathbb{E}[\tilde{Y}])^T].
\]

- \( \text{Cov}(\tilde{X}, \tilde{X}) \) is called the covariance matrix of the random vector \( \tilde{X} \), denoted \( \text{Cov}(\tilde{X}) \).
Note:

\[
\text{Cov}(\mathbf{x}) = \begin{bmatrix}
\sigma_{x_1}^2 & \rho_{x_1,x_2} \sigma_{x_1} \sigma_{x_2} & \cdots & \rho_{x_1,x_n} \sigma_{x_1} \sigma_{x_n} \\
\rho_{x_2,x_1} \sigma_{x_2} \sigma_{x_1} & \sigma_{x_2}^2 & \cdots & \rho_{x_2,x_n} \sigma_{x_2} \sigma_{x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{x_n,x_1} \sigma_{x_n} \sigma_{x_1} & \rho_{x_n,x_2} \sigma_{x_n} \sigma_{x_2} & \cdots & \sigma_{x_n}^2
\end{bmatrix}
\]

- The covariance matrix of a random vector is always positive semidefinite; for any vector \( \mathbf{a} \),

\[
\mathbf{a}^T \text{Cov}(\mathbf{x}) \mathbf{a} \geq 0.
\]

\( \rightarrow \) The eigenvalues of \( \text{Cov}(\mathbf{x}) \) are all nonnegative.

Note: The joint characteristic function and moment generating function for a random vector are easily generated using the multidimensional Fourier transform.
Conditional Density for Two RVs

- Let X and Y be two RVs with joint density $f_{X,Y}(x,y)$ and marginal densities $f_X(x)$ and $f_Y(y)$.

- Then

$$f_{X|Y=y}(x) = \lim_{\Delta y \to 0} \frac{f_{X|\{y \leq Y \leq y+\Delta y\}}(x)}{f_Y(y)}$$

- This has meaning only if $f_Y(y) \neq 0$.

Sum of Two Independent RVs

- Let X and Y be two independent RVs with marginal densities $f_X(x)$ and $f_Y(y)$.

- Let Z be a third RV defined by $Z = X + Y$.

- Then

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-\theta)f_Y(\theta)\,d\theta$$

$$= \int_{-\infty}^{\infty} f_X(\theta)f_Y(z-\theta)\,d\theta$$

$$= f_X(z) * f_Y(z). \quad \text{(Convolution)}$$
In the book, this result is obtained using a standard argument.

Here, we will take a different approach based on the characteristic function.

Since $X$ and $Y$ are independent, we have that

$$
\Phi_z(\omega) = E\left[ e^{i\omega Z} \right] \\
= E\left[ e^{i\omega (X+Y)} \right] \\
= E\left[ e^{i\omega X} e^{i\omega Y} \right] \\
= E\left[ e^{i\omega X} \right] E\left[ e^{i\omega Y} \right] \quad \text{(because independent)} \\
= \Phi_X(\omega) \Phi_Y(\omega)
$$

So $\Phi_z(-\omega) = \Phi_X(-\omega) \Phi_Y(-\omega)$

Taking the inverse Fourier Transform:

$$
f_z(z) = \mathcal{F}^{-1}\left\{ \Phi_z(-\omega) \right\} \\
= \mathcal{F}^{-1}\left\{ \Phi_X(-\omega) \Phi_Y(-\omega) \right\} \\
= f_x(z) * f_y(z)
$$
Transformations of an RV

- Let $X$ be an RV and $g: \mathbb{R} \rightarrow \mathbb{R}$ be a function.
- Then we can define a new RV $Y$ by the transformation
  \[ Y = g(X) \]

- If $g$ is a "one-to-one" function (a bijection), then it is invertible and $g^{-1}$ is also a function.
  \[ X = g^{-1}(Y) = \frac{Y-3}{5} \]

- If $g$ is one-to-one and $Y = g(X)$, then the density of $Y$ is given by
  \[ f_Y(y) = \left| \frac{d}{dy} g^{-1}(y) \right| f_X\left( g^{-1}(y) \right) \]

  ⇒ A derivation and examples are given in the book.

  ⇒ If $g$ is not one-to-one, it is generally necessary to resort to fundamental probability concepts to find $f_Y(y)$.

  ⇒ An example multivariate transformation is also given in the book.
Jointly Gaussian Variables

Let $X_1, X_2, \ldots, X_n$ be $n$ RVs.
Let $\mathbf{X}$ be the random vector
\[ \mathbf{X} = [x_1, x_2, \ldots, x_n]^T \]
Let $\mathbf{m}_X$ be the expected value of $\mathbf{X}$:
\[ \mathbf{m}_X = \mathbf{E}[\mathbf{X}] = [\mathbf{E}[X_1], \mathbf{E}[X_2], \ldots, \mathbf{E}[X_n]]^T \]
\[ = [m_1, m_2, \ldots, m_n]^T \]
Let $\mathbf{C}_X$ be the covariance matrix of $\mathbf{X}$:
\[ \mathbf{C}_X = \text{cov}(\mathbf{X}, \mathbf{X}) = \mathbf{E} \left\{ (\mathbf{X} - \mathbf{m}_X)(\mathbf{X} - \mathbf{m}_X)^T \right\} \]
Let $|\mathbf{C}_X| = \det \mathbf{C}_X$

The multivariate normal density is given by
\[ f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{C}_X|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m}_X)^T \mathbf{C}_X^{-1} (\mathbf{x} - \mathbf{m}_X) \right\} \]

Note:
\[ f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(x_1, x_2, \ldots, x_n): \mathbb{R}^n \rightarrow [0, \infty) \]
In the bivariate case,

\[
\begin{pmatrix}
\hat{X} \\
\hat{Y}
\end{pmatrix} = 
\begin{pmatrix}
X_1 \\
X_2
\end{pmatrix}, \quad
\begin{pmatrix}
\hat{X} \\
\hat{Y}
\end{pmatrix} = 
\begin{pmatrix}
X_1 \\
X_2
\end{pmatrix}, \quad
\begin{pmatrix}
m_1 \\
m_2
\end{pmatrix} = 
\begin{pmatrix}
\mu_1 \\
\mu_2
\end{pmatrix}
\]

\[
C_{\hat{X}} = 
\begin{pmatrix}
\sigma_1^2 & \rho \sigma_1 \sigma_2 \\
\rho \sigma_1 \sigma_2 & \sigma_2^2
\end{pmatrix}
\]

where \(\rho = \text{corr} (X_1, X_2)\)

\[
|C_{\hat{X}}| = \sigma_1 \sigma_2 (1 - \rho^2)
\]

\[
C_{\hat{X}}^{-1} = 
\begin{pmatrix}
\frac{1}{(1 - \rho^2) \sigma_1^2} & -\frac{-\rho}{(1 - \rho^2) \sigma_1 \sigma_2} \\
-\frac{\rho}{(1 - \rho^2) \sigma_1 \sigma_2} & \frac{1}{(1 - \rho^2) \sigma_2^2}
\end{pmatrix}
\]

So

\[
f_{\hat{X}} (\hat{X}) = f_{X_1, X_2} (X_1, X_2)
\]

\[
= \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho^2}} \exp \left\{ \frac{-1}{2(1 - \rho^2)} \left[ \frac{(X_1 - m_1)^2}{\sigma_1^2} - \frac{2\rho (X_1 - m_1)(X_2 - m_2)}{\sigma_1 \sigma_2} + \frac{(X_2 - m_2)^2}{\sigma_2^2} \right] \right\}
\]
Suppose $X_1$ and $X_2$ are uncorrelated.

Note: this is a weaker assumption than independent.

Then $\rho = 0$ and $\Sigma_x = \text{diag}(\sigma_1^2, \sigma_2^2)$.

So

$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{2\pi \sigma_1 \sigma_2} \exp \left\{ -\frac{1}{2} \left[ \frac{(x_1 - m_1)^2}{\sigma_1^2} + \frac{(x_2 - m_2)^2}{\sigma_2^2} \right] \right\}$$

$$= \frac{1}{\sqrt{2\pi \sigma_1^2}} e^{-\frac{(x_1 - m_1)^2}{2\sigma_1^2}} \cdot \frac{1}{\sqrt{2\pi \sigma_2^2}} e^{-\frac{(x_2 - m_2)^2}{2\sigma_2^2}}$$

the product of the marginal densities.

(For $X_1, X_2$ uncorrelated).

$\Rightarrow$ Thus, for Gaussian variables,

uncorrelated $\iff$ independent
Linear Transformation of Jointly Gaussian Variables

- Let $\tilde{X} = [x_1, x_2, \ldots, x_n]^T$ be a vector of $n$ jointly Gaussian RVs.

- The multivariate density of $\tilde{X}$ is

$$f_{\tilde{X}}(\tilde{x}) = (2\pi)^{-\frac{n}{2}} |C_{\tilde{x}}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} (\tilde{x} - \tilde{m}_{\tilde{x}})^T C_{\tilde{x}}^{-1} (\tilde{x} - \tilde{m}_{\tilde{x}}) \right\}$$

- Define a random vector $\tilde{Y} = [Y_1, Y_2, \ldots, Y_n]^T$ that is linearly related to $\tilde{X}$ by

$$\tilde{Y} = A \tilde{X} + \tilde{b}$$

where $A$ is an $n \times n$ invertible matrix of constants and $\tilde{b}$ is an $n \times 1$ vector of constants.

- We will denote the inverse transformation by

$$\tilde{X}(\tilde{Y}) = A^{-1} \tilde{Y} - A^{-1} \tilde{b}.$$ 

- The multivariate density of $\tilde{Y}$ is then given by

$$f_{\tilde{Y}}(\tilde{y}) = f_{\tilde{X}}(\tilde{X}(\tilde{y})) \left| J(\frac{\tilde{X}}{\tilde{Y}}) \right|$$

where $\left| J(\frac{\tilde{X}}{\tilde{Y}}) \right|$ is the magnitude of the Jacobian

$$J(\frac{\tilde{X}}{\tilde{Y}}) = \det(A^{-1})$$

see book, page 54.
Thus, the density of \( \tilde{Y} \) is given by

\[
f_{\tilde{Y}} (\tilde{y}) = \frac{| \text{det} A^{-1} |}{(2\pi)^{\frac{d}{2}} | C_\tilde{X} |^{\frac{1}{2}}} \times \exp \left\{ -\frac{1}{2} (A^{-1} \tilde{y} - A^{-1} \tilde{b} - \tilde{m}_\tilde{X})^T C_\tilde{X}^{-1} (A^{-1} \tilde{y} - A^{-1} \tilde{b} - \tilde{m}_\tilde{X}) \right\}
\]

Note: \( E[\tilde{Y}] = E[A \tilde{X} + \tilde{b}] = A \tilde{m}_\tilde{X} + \tilde{b} = \tilde{m}_\tilde{Y} \)

Note: \( | \text{det} A^{-1} | = \frac{1}{| \text{det} A |} = \frac{1}{\sqrt{\text{det} A \text{det} A^T}} \)

Then the density of \( \tilde{Y} \) may be written as

\[
f_{\tilde{Y}} (\tilde{y}) = (2\pi)^{-\frac{d}{2}} | A C_\tilde{X} A^T |^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\tilde{y} - \tilde{m}_\tilde{Y})^T (A C_\tilde{X} A^T)^{-1} (\tilde{y} - \tilde{m}_\tilde{Y}) \right\}
\]

\( \Rightarrow \tilde{Y} \) is a multivariate normal vector with mean

\( \tilde{m}_\tilde{Y} = A \tilde{m}_\tilde{X} + \tilde{b} \)

and covariance

\( C_\tilde{Y} = A C_\tilde{X} A^T \).

\( \Rightarrow \) Given a jointly normal vector \( \tilde{X} \), it is of interest to find a linear transformation \( \tilde{Y} = A \tilde{X} + \tilde{b} \) such that the jointly Gaussian variables \( Y_1, Y_2, \ldots, Y_n \) are mutually uncorrelated, and therefore mutually independent.
We say that such a transformation ("decouples") the variables $X_1, X_2, \ldots, X_n$.

If $Y_1, Y_2, \ldots, Y_n$ are mutually uncorrelated, then the covariance matrix $C_\Phi$ is diagonal:

$$C_\Phi = \text{diag}(\sigma_{Y_1}^2, \sigma_{Y_2}^2, \ldots, \sigma_{Y_n}^2).$$

Thus, a "decoupling" transformation (if it exists) diagonalizes the covariance matrix.

Note: a matrix $A$ is positive definite if, $\forall \vec{b} \neq \vec{0}$,

$$\vec{b}^\top A \vec{b} > 0,$$

and positive semidefinite if, $\forall \vec{b} \neq \vec{0}$,

$$\vec{b}^\top A \vec{b} \geq 0.$$

**FACT:** Any covariance matrix is positive semidefinite.

**FACT:** If the magnitudes of all the correlation coefficients between the variables $X_1, X_2, \ldots, X_n$ are strictly less than unity, then $C_\Phi$ is positive definite.

**FACT:** If $C_\Phi$ is positive definite, then a "decoupling" or "diagonalizing" transformation exists.
Facts About Jointly Gaussian Variables

Note: The list at the bottom of page 56 of the book needs two minor corrections to point (2).

1. The density of a multivariate normal vector $\mathbf{X}$ is completely specified by the mean vector and covariance matrix.

2. $C_\mathbf{X}$ is positive semidefinite. All of the correlation coefficients are less than or equal to one.

3. For jointly normal variables, uncorrelated $\leftrightarrow$ independent

4. A linear transformation of a multivariate normal vector always yields another multivariate normal vector. A decoupling transformation exists if $C_\mathbf{X}$ is positive definite.

5. The marginal densities for $X_1, X_2, \ldots, X_n$ are all univariate normal densities.
Types of Convergence

- Suppose we have a sequence of RVs $Y_1, Y_2, \ldots$

- For example, let $X_1, X_2, X_3, \ldots$ be an infinite collection of iid RVs. We can define a sequence of RVs $\{Y_n\}$ by

$$
Y_1 = X_1,
Y_2 = \frac{1}{2} (X_1 + X_2),
\vdots
Y_n = \frac{1}{n} (X_1 + X_2 + \cdots + X_n).
$$

- The sequence $Y_n$ is said to converge to $\Theta$ in mean if

$$
\lim_{n \to \infty} E[(Y_n - \Theta)^2] = 0.
$$

Note: $\Theta$ could be deterministic or stochastic.

- The sequence $Y_n$ is said to converge to $\Theta$ in probability if, $\forall \varepsilon > 0$,

$$
\lim_{n \to \infty} P(|Y_n - \Theta| > \varepsilon) = 0
$$

$\Rightarrow$ If $Y_n$ converges in mean, then it converges in probability.

$\Rightarrow$ The converse is not true.
Note: if $x_1, x_2, \ldots$ is an infinite collection of iid RVs with $E[X_k] = m_x$ and $\text{Var}(X_k) = \sigma_x^2$,

and if $Y_1 = X_1$

\[ Y_n = \frac{1}{n} \sum_{k=1}^{n} X_k, \]

then

\[ E[Y_n] = E\left[\frac{1}{n} \sum_{k=1}^{n} X_k\right] \]

\[ = \frac{1}{n} \sum_{k=1}^{n} E[X_k] \]

\[ = \frac{1}{n} \cdot n \cdot m_x = m_x. \]

$\Rightarrow$ In this case, we say that $Y_n$ is an "unbiased estimator" of $m_x$.

Note: $\text{Var}(Y_n) = \frac{\sigma_x^2}{n}$ (Show this)