Introduktion til sandsynlighedsteori og statistik

January 7, 2018

1 Gennerelt

Bog
Instruktor:
Navn: Jeanett Pelck Petersen
Mail: 201303926@post.au.dk eller jeanett_92@hotmail.dk
Kontor: F1

2 Basic Concepts

- Probabillity theory is allows us to describe random phenomena in the world around us
- When we say something is random we say that our knowlegde of the outcome is limited

2.1 Sums

Geometric series formular for any $a, x \in \mathbb{R}$:

$$a + ax + ax^2 + ax^3 + \cdots + ax^{n-1} = \sum_{k=0}^{n-1} ax^k = a\frac{1-x^n}{1-x} \tag{1}$$

For $|x| < 1$

$$a + ax + ax^2 + ax^3 + \cdots = \sum_{k=0}^{\infty} ax^k = a\frac{1}{1-x} \tag{2}$$

2.2 Sets

$$A = \{x \mid \text{x satisfies some property}\} \text{or} A = \{x \ : \ \text{x satisfies some property}\} \tag{3}$$

- *"|"* and *":"* are pronounced *"such that"*
- Two sets are equal if they have the same elements
- The **universal set** is the set of all things
  - E.g. if we roled a die the universal set would be:
    $S = \{1, 2, 3, 4, 5, 6\}$
  - Universal set is often denoted with S
- Vendiagrams is a digram showing the relationship between one or more sets and the universal set

2.2.1 Set operations

- A union of sets is written: $\bigcup_{i=1}^{n} A_i$
- The **complement** of a set A is denoted by A^c or $\hat{A}$
- Two sets are **mutually exclusive** or **disjoined** if they don't share any elements
- A collection of nonempty sets is a partition af a set A if they are disjoined and their union is A
- The **Multiplication principle** says that number of elements in $A \times B$ is $|A| \times |B|$

2.2.2 Cardinality of a set

- The **cardinary** of a set is basically the size of the set and is therefore denoted $|A|$
- If A is a finite set, the **cardinary** is simply the number of elements in the set
- Inclusion-exclusion principle: $|A \cup B| = |A| + |B| - |A \cap B|$
- There are two types of infinte sets called **countable** (e.g. $\mathbb{N}, \mathbb{Q}$ and $\mathbb{Z}$) and **uncountable** (e.g. $\mathbb{R}$)

- Sets are **countable** if the following is true

  1. If it is a finite set or
  2. It can be put down one to one with the natural numbers which is called infinite countable
     A set is called **uncountable** if it is not **countable**

- Any set on a real line [a,b],(a,b],[a,b),[a,b],(a,b],[a,b), or (a,b) where $a < b$ is uncountable
- Any subset of $\mathbb{N}, \mathbb{Q}$ and $\mathbb{Z}$ is uncountable
- Any subset of a countable set is countable.
- Any superset of an uncountable set is uncountable.
- If $A_1, A_2, \cdots$ is a list of countable sets, then the set $\bigcup_i A_i$ is also countable.
- If A and B is countable then $A \times B$ is also countable

2.3 Functions

$$f : A \to B \tag{4}$$

- The input of a function is called the domain
- The output of a function is called the codomain
- The **range** of the function is all the possible values of f(x)

2.4 Random experiments

- **Outcome** is the result of a random experiment
- The set of all possible outcomes is called the **sample space**
- When repeatting a random experiment several times each one of them is called a trial
- Our goal with random experiments is to assign a probability to certain events
- An **event** is a collection of certain outcomes

- A **probabillity** measure $P(A)$ is assigned to an event A between 0 and 1 which is how likely the event is

  Axioms of probability:
- Axiom 1: For any event, $P(A) \geq 0$
- Axiom 2: Probabillity of the sample space S is $P(S) = 1$
- Axiom 3: If $A_1, A_2, A_3 \cdots$ are disjoint events, then $P(A_1 \cup A_2 \cup A_3 \cdots) = P(A_1) + P(A_2), P(A_3) \cdots$

Rules from axioms of probability:

1. For any event A,$P(A^c) = 1 - P(A)$
2. The probability of the empty set is zero, i.e., $P(\emptyset) = 0$
3. For any event A, $P(A) \leq 1$
4. $P(A - B) = P(A) - P(A \cap B)$
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
6. If $A \subset B$ then $P(A) \leq P(B)$

Notation: - $P(A \cap B) = P(A \text{ and } B) = P(A, B)$ - $P(A \cup B) = P(A \text{ or } B)$

- To find the probality of a event there are usually two steps
  1. Use the specific information that we have about the random experiment
  2. Use the probability axioms

- A probabillity P on S is **discret** if S is countable
- If $A \subset S$ is an event, then A is also countable, and by the third axiom of probability we can write

$$P(A) = P(\bigcup_{s_j \in A} \{s_j\}) = \sum_{s_j \in A} P(s_j)$$

In a finite sample space S, where are outcomes are equally possible, the probability of any event A can be found by

$$P(A) = \frac{|A|}{|S|}$$

2.5 Conditional Probability

If A and B are two events in a sample space S, then the **conditional probability of** A **given** B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ when } P(B) > 0 \tag{5}$$

- A conditional probability is itself a probability measure, so it satisfies the probability axioms
- When B is a subset of A the probability of A is 1

For tree events A, B and C $P(C) > 0$ we have:

- $P(A^c|C) = 1 - P(A|C)$
- $P(\emptyset|C) = 0$
- $P(A|C) \leq 1$
- $P(A - B|C) = P(A|C) - P(A \cap B|C)$
- $P(A \cup B|C) = P(A|C) + P(B|C) - P(A \cap B|C)$
- $P(A|C) \leq P(B|C)$

Inclusion-exclusion principle:

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

Chain rule for conditional probability:

1. $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$
2. $P(A_1 \cap A_2 \cap \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2, A_1) \cdots P(A_n|A_{n-1}A_{n-2} \cdots A_1)$

2.5.1 Independence

Definition of independence:

$$P(A \cap B) = P(A)P(B) \wedge P(A|B) = P(A) \tag{6}$$

- If A, B events are disjoined where $P(A)$, $P(B) > 0$, then A and B is not independent

- If A and B are independent then

    1. A and B^c are independent
    2. A^c and B are independent
    3. A^c and B^c are independent

- If $A_1, A_2, \cdots, A_n$ are independent then:

$$P\big(A_1 \cup A_2 \cup \cdots \cup A_n\big) = 1 - \big(1 - P(A_1)\big)\big(1 - P(A_2)\big) \cdots \big(1 - P(A_n)\big). \tag{7}$$

2.5.2 Total probability

Law of total probability - If $B_1, B_2, B_3, \cdots$ is a partition of the sample space S, then for any event A we have

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i). \tag{8}$$

2.5.3 Bayers rule

- For any two events A and B, where $P(A) \neq 0$, we have:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{9}$$

- If $B_1, B_2, B_3, \cdots$ form the partition of the sample space S, and A is any event with $P(A) \neq 0$, we have

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)} \tag{10}$$

2.5.4 Conditional independence

Two events A and B are **conditionally independed** if given an event C where $P(C) > 0$

$$P(A \cap B|C) = P(A|C)P(B|C) \tag{11}$$

3 Combinators

- **Multiplication Principle:**
  - If we preform r experiments such that the kth experiment has n_k possible outcomes, for $k = 1, 2 \ldots, r$. The there are a total of $n_1 \cdot n_2 \cdot \cdots \cdot n_r$ possible outcomes for the sequence of r experiments

- **Sampling** from a set means choosing an element from that set.
- **Sampling with replacement** means that we put eachs objects back after each draw, in **sampling wihtout replacement** we don't
- If order matters it is called **ordered sampling** otherwise, it is called **unordered**
- If we have n elements there is $n!$ ways of ordering them

Sampling type	Posiblities
ordered sampling with replacement	n^k
ordered sampling without replacement	$P_k^n = \frac{n!}{(n-k)!}$
unordered sampling without replacement	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$
unordered sampling with replacement	$\binom{n+k-1}{k}$

3.1 Unordered sampling without replacement

- **K-permutation** of the elements in a set is choosing k elements, ordered with no repetion
- $\binom{n}{k}$ is called the **binomial coefficient**
- The total way to divide n distinct objects into two groups A and B such that group A consists of k objects and group B consists of $n-k$ objects is $\binom{n}{k}$

- Binomial theorem states that for an integer $n > 0$

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}. \tag{12}$$

- Binomial coefficient rules:
  1. We have $\sum_{k=0}^{n} \binom{n}{k} = 2^n$
  2. For $0 \leq k < n$, we have $\binom{n+1}{k+1} = \binom{n}{k+1} + \binom{n}{k}$
  3. We have $\binom{m+n}{k} = \sum_{i=0}^{k} \binom{m}{i}\binom{n}{k-i}$ (Vandermonde's identity).

- A **Bernoulli Trial** is a random experiment that has two possible outcomes that we can label "success" or "failure" such as
  - Tossing a coin were you can define heads or tails as "failure" or "success"
  - You take a test, where the outcomes are pass or fail
- Success is denoted by p
- The probability of failure is denoted by $q = p - 1$
- An experiment where n independent Bernoulli trials is performed and we count the number of success, it is called a **binomial** experiment

- **Binomial formula:** For n independent Bernoulli trials where each trial has success probability p, the probability of k successes is given by

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

- **Multinomial formula:**
  - If $n = n1 + n2 + ... + n_r$, where all $ni \geq 0$ are integers, then the number of ways to divide n distinct objects to r distinct groups of sizes $n1, n2, ..., n_r$ is given by

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1! n_2! ... n_r!}.$$

- **Multinomial theoroem:**

$$(x_1 + x_2 + \cdots + x_r)^n = \sum_{n_1+n_2+\cdots+n_r=n} \binom{n}{n_1, n_2, \ldots, n_r} x_1^{n_1} x_2^{n_2} ... x_r^{n_r} \tag{13}$$

4 Discrete Random Variables

4.1 Random variables

- A random variable is a real-valued function that assigns a numerical value to each possible outcome of the random experiment
- Random variables are shown by capital letters such as X, Y, Z where as number values are shown by small letters such as x, y, z

- A random variable X is a function from the sample space to real numbers

$$X : S \to \mathbb{R} \tag{14}$$

- The range of a random variable X is shown by $Range(X)$ or R_X is the set of possible values of X

- X is a discrete random variable if its range is countable

4.2 Probability Mass Function (PMF)

- If X is a discrete random variable then we can list the elements in R_x:

$$R_x = \{x_1, x_2, x_3, ...\} \tag{15}$$

- The event $A = \{X = x_k\}$ is defined as the set of outcomes s in the sample space S where there corresponding value of X is equal to x_k

$$A = \{s \in S \mid X(s) = x_k\} \tag{16}$$

- The probabilities of events $\{X = x_k\}$ are formally shown by the **probability mass function (pmf)** of X

Definition 3.1
Let X be a discrete random variable with range $R_x = \{x_1, x_2, x_3, ...\}$. This function is called the probablity masse function (PMF) of X:

$$P_X(x_k) = \mathrm{P}(X = x_k), \text{ for } k = 1, 2, 3, ..., \tag{17}$$

- For discrete random variables the PMF is also called the **probability distribution**

- Properties of PMF:
  1. $0 \leq P_X(x) \leq 1$ for all x
  2. $\sum_{x \in R_X} P_X(x) = 1$
  3. For any set $A \subset R_X, P(X \in A) = \sum_{x \in A} P_X(x)$
- If it is desired that the PMF to take numbers not in R_X we define $P_X(x)$ to be

$$P_X(x) = \begin{cases} P(X = x) & \text{if } x \text{ is in } R_X \\ 0 & \text{otherwise} \end{cases}$$

4.3 Independent Random Variables

Definition 3.2
Consider two discrete random variables X and Y, X and Y is independent if

$$P\Big(X = x, Y = y\Big) = P(X = x)P(Y = y), \qquad \text{for all } x, y. \tag{18}$$

- If two random variables are independent you can write

$$P\Big(X \in A, Y \in B\Big) = P(X \in A)P(Y \in B), \qquad \text{for all sets } A \text{ and } B. \tag{19}$$

- The following is also true is two random variables are independent

$$P(Y = y \mid X = x) = P(Y = y), \text{ for all } x, y. \tag{20}$$

Definition 3.3
Consider n discrete random variables $X_1, X_2, X_3, ..., X_n$. We say that $X_1, X_2, X_3, ..., X_n$ are independent if

$$P\Big(X_1 = x_1, X_2 = x_2, ..., X_n = x_n\Big) = P(X_1 = x_1)P(X_2 = x_2)...P(X_n = x_n), \text{ for all } x_1, x_2, ..., x_n. \tag{21}$$

4.4 Special Distributions

4.4.1 Bernoulli Distribution

Definition 3.4
A random variable X is said to be a Bernoulli random variable with parameter p, shown as $X \sim Bernoulli(p)$, if its PMF is given by where $0 < p < 1$

$$P_X(x) = \begin{cases} p & \text{for } x = 1 \\ 1 - p & \text{for } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

- The bernoulli random variable is also called the **indicator** random variable.
- The indicator random variable I_A is defined for an event A as

$$I_A = \begin{cases} 1 & \text{if the event } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

- The indicator random variable for an event A has the Bernoulli distribution with parameter $p = P(A)$, so we can write

$$I_A \sim Bernoulli\big(P(A)\big).$$

4.4.2 Geometric Distribution

- A geometric distribution is can usually be thought of as repeating independent Bernoulli trials until the first success
- In a geometric distribution we usually use q as $1 - p$
- Some define the geometric district distribution as number of failures before success here it is defined as the numbers of experiments that lead to success

Definition 3.5

- A random variable X is said to be a geometric random variable with parameter p, shown as $X \sim Geometric(p)$ if its PMF is given by where $0 < p < 1$

$$P_X(k) = \begin{cases} p(1-p)^{k-1} & \text{for } k = 1, 2, 3, ... \\ 0 & \text{otherwise} \end{cases}$$

4.4.3 Binomial Distribution

- A random experiment involving a binomial distribution can be a coin that has $P(H) = p$. The coin is tossed n times and we define X to be the total number of heads observed.

Definition 3.6

- A random variable X is said to be a binomial random variable with parameters n and p, shown as $X \sim Binomial(n, p)$, if its PMF is given by where $0 < p < 1$

$$P_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{for } k = 0, 1, 2, \cdots, n \\ 0 & \text{otherwise} \end{cases}$$

Lemma 3.1

- If $X_1, X_2, ..., X_n$ are independent $Bernoulli(p)$ random variables, then the random variable X defined by $X = X_1 + X_2 + ... + X_n$ has a $Binomial(n, p)$ distribution.

4.4.4 Negative Binomial (Pascal) Distribution

- The negative binomial distribution is a generalization of the geometric distribution.
- It relates to the random experiment of repeated independent trials until observing m successes.
- $Pascal(1, p) = Geometric(p)$

Definition 3.7

- A random variable X is said to be a Pascal random variable with parameters m and p, shown as $X \sim Pascal(m, p)$, if its PMF is given by where $0 < p < 1$

$$P_X(k) = \begin{cases} \binom{k-1}{m-1} p^m (1-p)^{k-m} & \text{for } k = m, m+1, m+2, m+3, ... \\ 0 & \text{otherwise} \end{cases}$$

4.4.5 Hypergeometric Distribution

- Hypergeometric Distribution can be thought of as chosing k marbles at random out of a bag with b blue maples and r red maples and then the random variable X is the number of blue

marbles

Definition 3.8

- A random variable X is said to be a Hypergeometric random variable with parameters b,r and k, shown as $X \sim Hypergeometric(b, r, k)$, if its range is $R_X = \{\max(0, k-r), \max(0, k-r)+1, \max(0, k-r)+2, ..., \min(k, b)\}$, and its PMF is given by

$$P_X(x) = \begin{cases} \frac{\binom{b}{x}\binom{r}{k-x}}{\binom{b+r}{k}} & \text{for } x \in R_X \\ 0 & \text{otherwise} \end{cases}$$

4.4.6 Poisson Distribution

- The poisson distribution is often used counting occurences of certain events in a inteval of time and space

Definition 3.9

- A random variable X is said to be a Poisson random variable with parameter λ, shown as $X \sim Poisson(\lambda)$, if its range is $R_X = \{0, 1, 2, 3, ...\}$, and its PMF is given by

$$P_X(k) = \begin{cases} \frac{e^{-\lambda}\lambda^k}{k!} & \text{for } k \in R_X \\ 0 & \text{otherwise} \end{cases}$$

4.5 Cumulative Distribution Function

Definition 3.10
The cumulative distribution function (CDF) of random variable X is defined as

$$F_X(x) = P(X \leq x), \text{ for all } x \in \mathbb{R} \tag{22}$$

- For all $a \leq b$, we have

$$P(a < X \leq b) = F_X(b) - F_X(a) \tag{23}$$

4.6 Expectation

- The expected value is defined as the weighted value in the range
- Expected value = mean = average
- Different notations for expected value of X: $EX = E[X] = E(X) = \mu_X$

Definition 3.11

- Let X be a discrete random variable with range $R_X = \{x_1, x_2, x_3, ...\}$ (finite or countably infinite). The expected value of X, denoted by EX is defined as

$$EX = \sum_{x_k \in R_X} x_k \mathrm{P}(X = x_k) = \sum_{x_k \in R_X} x_k P_X(x_k) \tag{24}$$

Theorem 3.2

- Expectation is linear

- We have
  - $E[aX + b] = aEX + b$ for $a, b \in \mathbb{R}$
  - $E[X_1 + X_2 + \cdots + X_n] = EX_1 + EX_2 + \cdots + EX_n$, for any set of random variables $X_1, X_2, \cdots, X_n$

4.7 Functions of Random Variables

- If X is a random variable and $Y = g(X)$, then Y itself is a random variable.
  - Thus, we can talk about its PMF, CDF, and expected value.
  - The range of Y can be written as

$$R_Y = \{g(x) | x \in R_X\} \tag{25}$$

- If the PMF of X is already known the PMF of Y can be found by

$$\begin{aligned} P_Y(y) &= P(Y = y) \\ &= P(g(X) = y) \\ &= \sum_{x:g(x)=y} P_X(x) \end{aligned} \tag{26}$$

- Law of the unconscious statistician (LOTUS) for discrete random variables:

$$E[g(X)] = \sum_{x_k \in R_X} g(x_k) P_X(x_k)$$

4.8 Variance

- The variance of a random variable X, with mean $EX = \mu_X$, is defined as

$$\text{Var}(X) = E\big[(X - \mu_X)^2\big].$$

- The **standard deviation** of a random variable X is defined as

$$\text{SD}(X) = \sigma_X = \sqrt{\text{Var}(X)}$$

- Computational formula for the variance:

$$\text{Var}(X) = E\big[X^2\big] - \big[EX\big]^2$$

- **Theorem 3.3:** For a random variable X and real numbers a and b,

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

- **Theorem 3.4:** If $X_1, X_2, \cdots, X_n$ are independent random variables and $X = X_1 + X_2 + \cdots + X_n$,then

$$\text{Var}(X) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n)$$

5 Continuous Random Variables

- A random variable X with CDF $F_X(x)$ is said to be continuous if $F_X(x)$ is a continuous function for all $x \in \mathbb{R}$

Definition 4.2

- Consider a continuous random variable X with an absolutely continuous CDF $F_X(x)$. The function $f_X(x)$ is defined by:

$$f_X(x) = \frac{dF_X(x)}{dx} = F'_X(x), \qquad \text{if } F_X(x) \text{ is differentiable at } x$$

is called the probability density function (PDF) of X

- Since the PDF is the derivative of the CDF we can find the CDF in the following way

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du.$$

- Consider a continuous random variable X with a PDF $f_X(x)$

1. $f\_X(x) \geq 0$ for all $x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f_X(u)du = 1$
3. $P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(u)du$
4. For at set A: $P(X \in A) = \int_A f_X(u)du$

- The **range** of a continuous random variable is defined as the set of real numbers x for which the PDF is larger than zero

$$R_X = \{x | f_X(x) > 0\}$$

5.1 Expected value and Variance

- The expected value of a continuous random variable is defined as

$$EX = \int_{-\infty}^{\infty} x f_X(x)dx$$

- Law of the unconscious statistician (LOTUS) for continuous random variables:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)dx$$

- For a continuous random variable we can write variance as

$$Var(X) = E\big[(X-\mu_X)^2\big] = \int_{-\infty}^{\infty}(x-\mu_X)^2 f_X(x)dx \tag{27}$$

$$= EX^2 - (EX)^2 = \int_{-\infty}^{\infty} x^2 f_X(x)dx - \mu_X^2 \tag{28}$$

5.2 Functions

Theorem 4.1 - If X is a continuous random variable and $g : \mathbb{R} \to \mathbb{R}$ is a strictly monotonic differentiable function. Let $Y = g(x)$. Then the PDF of Y is given by

$$f_Y(y) = \begin{cases} f_x(g^{-1}(x)) \cdot |g^{-1'}(y)| & \\ 0 & \text{if } g(x) = y \text{ does not have a solution} \end{cases}$$

Theorem 4.2 - If we have a random variable X with domain R_X and let $Y = g(X)$. If we can partition R_X into a finite number of intervals where $g(x)$ is strictly monotone and differentiable on each partition. Then the PDF for Y is given by

$$f_Y(y) = \sum_{i=1}^{n} \frac{f_X(x_i)}{|g'(x_i)|} = \sum_{i=1}^{n} f_X(x_i).\left|\frac{dx_i}{dy}\right|$$

where $x_1, x_2, \ldots, x_n$ are real solutions to $g(x) = y$

5.3 Special Distributions

5.3.1 Uniform Distribution

- A continuous random variable X is said to have a uniform distribution over the interval $[a, b]$, shown as $X \sim Uniform(a, b)$, if its PDF is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x < a \text{ or } x > b \end{cases}$$

- Its mean is given by

$$EX = \frac{a+b}{2}$$

- Its variance is given by

$$Var(X) = \frac{(b-a)^2}{12}$$

5.3.2 Exponential Distribution

- A continuous variable X is said to have an exponential distribution with parameter $\lambda > 0$, shown as $X \sim Exponential(\lambda)$ if its PDF is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- If $X \sim Exponential(\lambda)$ then $EX = \frac{1}{\lambda}$ and $Var(X) = \frac{1}{\lambda^2}$
- If X is exponential with a parameter $\lambda > 0$, then X is a **memoryless** random variable

$$P(X > x + a \mid X > a) = P(X > x), \quad \text{for } a, x \geq 0.$$

5.3.3 Normal (Gaussian) Distribution

- A continuous random variable Z is said to be a standard normal (standard Gaussian) random variable, shown as $Z \sim N(0,1)$ is its PDF is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}, \qquad \text{for all } z \in \mathbb{R}.$$

- If $Z \sim N(0,1)$, then $EZ = 0$ and $Var(Z) = 1$
- The CDF of the normal distribution is denoted Φ function:

$$\Phi(x) = P(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{u^2}{2}\right\} du.$$

- Some properties of the Φ function
  1. $\lim_{x\to\infty} \Phi(x) = 1, \quad \lim_{x\to-\infty} \Phi(x) = 0$
  2. $\Phi(0) = \frac{1}{2}$
  3. $\Phi(-x) = 1 - \Phi(x)$ for all $x \in \mathbb{R}$

- If Z is a standard normal random variable and $X = \sigma Z + \mu$, then X is a random normal variable with mean μ and variance σ^2

$$X \sim N(\mu, \sigma)$$

- If X is a normal random variable with mean μ and variance σ^2 i.e. $X \sim N(\mu, \sigma^2)$ then

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

$$F_X(x) = P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

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

- $X \sim N(\mu_X, \sigma_X^2)$ and $Y = aX + b$ where $a, b \in \mathbb{R}$ then $Y \sim N(\mu_Y, \sigma_Y^2)$ where

$$\mu_Y = a\mu_X + b, \quad \sigma_Y^2 = a^2\sigma_X^2.$$

5.3.4 Gamma Distribution

- A continuous random variable is said to have a gamma distribution with parameters $a > 0$ and $\lambda > 0$, shown as $X \sim Gamma(\alpha, \lambda)$ if its PDF is given by

$$f_X(x) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- A gamma distribution with parameter one is the an exponential distribution $Gamma(1, \lambda) = Exponential(\lambda)$

- If $X \sim Gamma(\alpha, \lambda)$ then
$$EX = \frac{\alpha}{\lambda}, \qquad Var(X) = \frac{\alpha}{\lambda^2}.$$
- For any positive real number a:
  1. $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$
  2. $\int_0^\infty x^{\alpha-1}e^{-\lambda x}dx = \frac{\Gamma(\alpha)}{\lambda^\alpha}$, for $\lambda > 0$;
  3. $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$;
  4. $\Gamma(n) = (n-1)!$, for $n = 1, 2, 3, \cdots$;
  5. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

5.4 Mixed random variables

- There is some random variables which is neither continuous or discrete they are called **Mixed random variables**
  - They are a mixture of both
- The CDF of a mixed random variable can be written as the sum of a continuous and a staircase function
$$F_Y(y) = C(y) + D(y)$$
- If differentiate the continuous part of the CDF we get $c(y)$ which is not a valid CDF because it does not summarize to one
$$c(y) = \frac{dC(y)}{dy}, \text{ wherever } C(y) \text{ is differentiable.}$$
- Let $\{y_1, y_2, y_3, ...\}$ be a set of jump points of $D(y)$ for which $P(X = y_k) > 0$ We then have
$$\int_{-\infty}^{\infty} c(y)dy + \sum_{y_k} P(Y = y_k) = 1.$$
- The expected value can be obtained as
$$EY = \int_{-\infty}^{\infty} yc(y)dy + \sum_{y_k} y_k P(Y = y_k)$$

5.4.1 The delta function

Definition 4.3
We define the delta function $\delta(x)$ as a function with the following properties:

1. $\delta(x) = \begin{cases} \infty & x = 0 \\ 0 & \text{otherwise} \end{cases}$
2. $\delta(x) = \frac{d}{dx}u(x)$ where $u(x)$ is a unit step
3. $\int_{-\epsilon}^{\epsilon} \delta(x)dx = 1$ for any $\epsilon > 0$

4. For any $\epsilon > 0$ and any function $g(x)$ that is continuous over $(x_0 - \epsilon, x_0 + \epsilon)$ we have

$$\int_{-\infty}^{\infty} g(x)\delta(x - x_0)dx = \int_{x_0-\epsilon}^{x_0+\epsilon} g(x)\delta(x - x_0)dx = g(x_0) \tag{29}$$

- For a discrete random variable X with range $RX = \{x_1, x_2, x_3, \dots\}$ and PMF $PX(x_k)$, we define the (generalized) probability density function (PDF) as

$$f_X(x) = \sum_{x_k \in R_X} P_X(x_k)\delta(x - x_k).$$

- The (generalized) PDF of a mixed random variable can be written in the form

$$f_X(x) = \sum_k a_k \delta(x - x_k) + g(x),$$

- where $ak = P(X = xk)$, and $g(x) \geq 0$ does not contain any delta functions. Furthermore, we have

$$\int_{-\infty}^{\infty} f_X(x)dx = \sum_k a_k + \int_{-\infty}^{\infty} g(x)dx = 1$$

6 Joint Distributions

6.1 For two discrete variables

6.1.1 Joint PMF

- The **joint probability mass function** of two discrete random variable X and Y is defined as

$$P_{XY}(x, y) = P(X = x, Y = y)$$

- The joint range for X and Y can be defined as

$$R_{XY} = \{(x, y) \mid P_{XY}(x, y) > 0\}$$

- For two discrete random variables we have

$$\sum_{(x_i,y_j)\in R_{XY}} P_{XY}(x_i, y_j) = 1$$

- To find $P((X, Y) \in A$ for any set $A \subset \mathbb{R}$ we have

$$P\big((X, Y) \in A\big) = \sum_{(x_i,y_j)\in(A\cap R_{XY})} P_{XY}(x_i, y_j)$$

- The marginal PMFs of X and Y

$$\begin{aligned} P_X(x) &= \sum_{y_j \in R_Y} P_{XY}(x, y_j), \qquad \text{for any } x \in R_X \\ P_Y(y) &= \sum_{x_i \in R_X} P_{XY}(x_i, y), \qquad \text{for any } y \in R_Y \end{aligned} \tag{30}$$

6.1.2 Joint CDF

- The **joint cumulative distribution function** of two random variables X and Y is defined as:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y).$$

- Marginal CDFs of X and Y:

$$\begin{aligned} F_X(x) &= F_{XY}(x, \infty) = \lim_{y \to \infty} F_{XY}(x, y), \qquad \text{for any } x, \\ F_Y(y) &= F_{XY}(\infty, y) = \lim_{x \to \infty} F_{XY}(x, y), \qquad \text{for any } y \end{aligned} \tag{31}$$

- For a joint CDF the following must be true

$$\begin{aligned} &F_{XY}(\infty, \infty) = 1, \\ &F_{XY}(-\infty, y) = 0, \qquad \text{for any } y, \\ &F_{XY}(x, -\infty) = 0, \qquad \text{for any } x. \end{aligned}$$

- **Lemma 5.1:** For two random variables X and Y, and real numbers $x1 \leq x2$, $y1 \leq y2$, we have

$$\begin{aligned} P(x_1 < X \leq x_2,\ \ y_1 < Y \leq y_2) = \\ F_{XY}(x_2, y_2) - F_{XY}(x_1, y_2) - F_{XY}(x_2, y_1) + F_{XY}(x_1, y_1). \end{aligned}$$

6.1.3 Conditioning and Independence

- For a discrete variable X and a event A, the **conditional PMF** of X given A is defined as

$$\begin{aligned} P_{X|A}(x_i) &= P(X = x_i | A) \\ &= \frac{P(X = x_i \text{ and } A)}{P(A)}, \quad \text{for any } x_i \in R_X. \end{aligned}$$

- The **conditional CDF** of X given A is defined as

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

- For discrete variables X and Y the **conditional PMF** of X given Y and vice versa is defined as

$$\begin{aligned} P_{X|Y}(x_i | y_j) &= \frac{P_{XY}(x_i, y_j)}{P_Y(y_j)}, \\ P_{Y|X}(y_j | x_i) &= \frac{P_{XY}(x_i, y_j)}{P_X(x_i)} \end{aligned}$$

for any $x_i \in R_X$ and $y_j \in R_Y$.

X

- Two discrete random variables X and Y are independent if

$$P_{XY}(x,y) = P_X(x)P_Y(y), \quad \text{for all } x,y.$$

Or

$$F_{XY}(x,y) = F_X(x)F_Y(y), \quad \text{for all } x,y.$$

- Conditional expectation of X

$$E[X|A] = \sum_{x_i \in R_X} x_i P_{X|A}(x_i),$$
$$E[X|Y = y_j] = \sum_{x_i \in R_X} x_i P_{X|Y}(x_i|y_j)$$

- Law of total probability:

$$P(X \in A) = \sum_{y_j \in R_Y} P(X \in A|Y = y_j)P_Y(y_j), \quad \text{for any set } A.$$

- Law of total expectation

1. If $B_1, B_2, B_3, \ldots$ is a partition of the sample space S

$$EX = \sum_i E[X|B_i]P(B_i)$$

2. For random variable X and discrete variable Y

$$EX = \sum_{y_j \in R_Y} E[X|Y = y_j]P_Y(y_j)$$

6.1.4 Functions of Two Random Variables

- For two random variables X and Y and $Z = g(X,Y)$ where $g : \mathbb{R}^2 \to \mathbb{R}$

$$\begin{aligned} P_Z(z) &= P(g(X,Y) = z) \\ &= \sum_{(x_i,y_j) \in A_z} P_{XY}(x_i,y_j), \text{ where } A_z = \{(x_i,y_j) \in R_{XY} : g(x_i,y_j) = z\}. \end{aligned}$$

- Law of the unconscious statistician (LOTUS) for two discrete random variables:

$$E[g(X,Y)] = \sum_{(x_i,y_j) \in R_{XY}} g(x_i,y_j)P_{XY}(x_i,y_j)$$

6.1.5 Conditional Expectation

- Let X and Y be two random variables and g and h be two functions then the following is true
$$E[g(X)h(Y)|X] = g(X)E[h(Y)|X]$$
- Law of Iterated Expectations
$$E[X] = E[E[X|Y]]$$
- If X and Y are independent variables then
  1. $E[X|Y] = EX$
  2. $E[g(X)|Y] = E[g(X)]$
  3. $E[XY] = EXEY$
  4. $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$

- Law of Total Variance
$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y])$$

6.2 For two continuous random variables

6.2.1 Joint PDF

- Two random variable X and Y are **jointly continuous** if there exists a nonnegative function such that we have $f_{XY} : \mathbb{R}^2 \to \mathbb{R}$, such that for any set $A \in \mathbb{R}^2$ we have:
$$P\big((X,Y) \in A\big) = \iint_A f_{XY}(x,y)dxdy$$
- The function $f_{XY}(x,y)$ is called the **joint probability density function (PDF)** of X and Y
- The range of (X,Y) in a joint PDF is
$$R_{XY} = \{(x,y) \mid f_{X,Y}(x,y) > 0\}.$$
- The following must be true for a joint PDF
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1$$
- Marginal PDFs
$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy, \quad \text{for all } x,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx, \quad \text{for all } y.$$

6.2.2 Joint CDF

- The **joint cumulative function** of two random variables X and Y is defined as:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y).$$

- The joint CDF satisfies the following properties

1. $F_X(x) = F_{XY}(x, \infty)$ for any x (marginal CDF of X)
2. $F_Y(y) = F_{XY}(\infty, y)$ for any y (marginal CDF of Y)
3. $F_{XY}(\infty, \infty) = 1$
4. $F_{XY}(-\infty, y) = F_{XY}(x, -\infty) = 0$
5. $P(x_1 < X \leq x_2,\ y_1 < Y \leq y_2) = F_{XY}(x_2, y_2) - F_{XY}(x_1, y_2) - F_{XY}(x_2, y_1) + F_{XY}(x_1, y_1)$
6. If X and Y are independent then $F_{XY}(x, y) = F_X(x)F_Y(y)$

6.2.3 Conditioning and independence

- If X is a continuous random variable, and A is an event that $a < X < b$ (where possibly $a = -\infty$ or $b = \infty$, then

$$F_{X|A}(x) = \begin{cases} 1 & x > b \\ \frac{F_X(x) - F_X(a)}{F_X(b) - F_X(a)} & a \leq x < b \\ 0 & x < a \end{cases}$$

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{P(A)} & a \leq x < b \\ 0 & \text{otherwise} \end{cases}$$

- In general for a random variable X and an event A we have the following

$$E[X|A] = \int_{-\infty}^{\infty} x f_{X|A}(x) dx,$$
$$E[g(X)|A] = \int_{-\infty}^{\infty} g(x) f_{X|A}(x) dx,$$
$$\text{Var}(X|A) = E[X^2|A] - (E[X|A])^2$$

- For two jointly continuous random variables X and Y, we have:

1. The conditional PDF of X given $Y = y$:

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

2. The conditional probability that $X \in A$ given $Y = y$

$$P(X \in A|Y = y) = \int_A f_{X|Y}(x|y) dx$$

3. The conditional CDF of X given $Y = y$

$$F_{X|Y}(x|y) = P(X \leq x|Y = y) = \int_{-\infty}^{x} f_{X|Y}(x|y)dx$$

- For two jointly continuous random variables X and Y, we have:

1. Expected value of X given $Y = y$

$$E[g(X)|Y = y] = \int_{-\infty}^{\infty} g(x)f_{X|Y}(x|y)dx$$

2. Conditional LOTUS

$$E[g(X)|Y = y] = \int_{-\infty}^{\infty} g(x)f_{X|Y}(x|y)dx$$

3. Conditional variance of X given $Y = y$

$$Var(X|Y = y) = E[X^2|Y = y] - (E[X|Y = y])^2$$

- Two continuous random variables X and Y are independent if and only if

$$f_{XY}(x, y) = f_X(x)f_Y(y), \quad \text{for all } x, y$$

$$F_{XY}(x, y) = F_X(x)F_Y(y), \quad \text{for all } x, y$$

- If two continuous random variables X and Y are independent then the following is true

$$E[XY] = EXEY,$$
$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

- Law of Total Probability

$$P(A) = \int_{-\infty}^{\infty} P(A|X = x)f_X(x)\ dx$$

- Law of Total Expectation

$$E[Y] = \int_{-\infty}^{\infty} E[Y|X = x]f_X(x)\ dx = E[E[X|Y]]$$

- Law of Total Variance

$$\text{Var}(Y) = E[\text{Var}(Y|X)] + \text{Var}(E[Y|X])$$

6.2.4 Functions of two random variables

- LOTUS for two continuous random variables

$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)\ dxdy$$

- **Theorem 5.1** Let X and Y be two jointly continuous random variables. Let $(Z,W) = g(X,Y) = (g_1(X,Y), g_2(X,Y))$, where $g : \mathbb{R}^2 \to \mathbb{R}^2$ is a continuous one-to-one (invertible) function with continuous partial derivatives. Let $h = g^{-1}$, i.e., $(X,Y) = h(Z,W) = (h_1(Z,W), h_2(Z,W))$. Then Z and W are jointly continuous and their joint PDF, $f_{ZW}(z,w)$, for $(z,w) \in R_{ZW}$ is given by

$$f_{ZW}(z,w) = f_{XY}(h_1(z,w), h_2(z,w))|J|,$$

  where J is the Jacobian of h defined by

$$J = \det \begin{bmatrix} \frac{\partial h_1}{\partial z} & \frac{\partial h_1}{\partial w} \\ \frac{\partial h_2}{\partial z} & \frac{\partial h_2}{\partial w} \end{bmatrix} = \frac{\partial h_1}{\partial z}.\frac{\partial h_2}{\partial w} - \frac{\partial h_2}{\partial z}\frac{\partial h_1}{\partial w}.$$

- Let X and Y be two jointy continous random variables and $Z = X + Y$, then

$$f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(w, z-w)dw = \int_{-\infty}^{\infty} f_{XY}(z-w, w)dw.$$

- If X and Y are also independed then

$$\begin{aligned} f_Z(z) &= f_X(z) * f_Y(z) \\ &= \int_{-\infty}^{\infty} f_X(w) f_Y(z-w)dw = \int_{-\infty}^{\infty} f_Y(w) f_X(z-w)dw. \end{aligned}$$

- **Therem 5.2:** If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independed then

$$X + Y \ \sim \ N\Big(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2\Big).$$

6.3 Covariance and Correlation

- The covariance between X and Y is defined as

$$\mathrm{Cov}(X,Y) = E\big[(X - EX)(Y - EY)\big] = E[XY] - (EX)(EY)$$

- The covariance has the following properties

1. $\text{Cov}(X, X) = \text{Var}(X)$
2. If X and Y are independed then $\text{Cov}(X, Y) = 0$
3. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
4. $\text{Cov}(aX, Y) = a\text{Cov}(X, Y)$
5. $\text{Cov}(X + c, Y) = \text{Cov}(X, Y)$
6. $\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$
7. More generally

$$\text{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \text{Cov}(X_i, Y_j).$$

- If $\text{Cov}(X, Y) = 0$ X and Y are not necessarily independed
- Generally for $a, b \in \mathbb{R}$

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y)$$

- The **correlation coefficient** of two random variables X and Y is defined as the covariance of the standardized versions of X and Y
  - The sandardized versions of X and Y is defined as

$$U = \frac{X - EX}{\sigma_X}, \quad V = \frac{Y - EY}{\sigma_Y}$$

- The **correlation coefficient** denoted by ρ_{XY} or $\rho(X, Y)$ is obtained by normalizing the covariance

$$\rho_{XY} = \rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var(X) Var(Y)}}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- Properties of the **correlation coefficient**
  1. $-1 \leq \rho(X, Y) \leq 1$
  2. If $\rho(X, Y) = 1$ then $Y = aX + b$, where $a > 0$
  3. If $\rho(X, Y) = -1$ then $Y = aX + b$, where $a < 0$
  4. $\rho(aX + b, cY + d) = \rho(X, Y)$ for $a, c > 0$
- **Definition 5.2:** Consider two random variable X and Y:
  - If $\rho(X, Y) = 0$ we say that X and Y are **uncorrelated**
  - If $\rho(X, Y) > 0$ we say that X and Y are **positivly** correlated
  - If $\rho(X, Y) < 0$ we say that X and Y are **negativly** correlated
- In general the following is true about two random variables X and Y

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

- If X and Y are uncorrelated then

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$

- More generally if $X_1, X_2, ..., X_n$ are pairwise uncorrelated i.e. $\rho(X_i, X_j) = 0$ for $i \neq j$, then

$$\text{Var}(X_1 + X_2 + ... + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + ... + \text{Var}(X_n).$$

- If X and Y are independed then they are uncorrelated

7 Multiple random variables

- All the concepts used for two random variables can be extended to more variables

- Let $X_1, X_2, \ldots, X_n$ be n discrete variables then the joint PMF of $X_1, X_2, \ldots, X_n$ is defined as

$$P_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n).$$

- For n jointly continuous random variables $X_1, X_2, \ldots, X_n$ the joint PDF is defined to be the function $f_{X_1X_2...X_n}(x_1, x_2, ..., x_n)$ such that the probability of any set $A \subset \mathbb{R}^n$ is given by the integral of the PDF over the set A. In particular, for a set $A \subset \mathbb{R}^n$, we can write

$$P\Big((X_1, X_2, \cdots, X_n) \in A\Big) = \int \cdots \int_A \cdots \int f_{X_1X_2\cdots X_n}(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n.$$

- The marginal PDF of X_i can be obtained by integrating all other X_j's

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1X_2...X_n}(x_1, x_2, ..., x_n) dx_2 \cdots dx_n$$

- The join CDF of n random variables $X_1, X_2, \ldots, X_n$ is defined as

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

7.1 Independence

- Random variables $X_1, X_2, \ldots, X_n$ are independent if for all $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n).$$

- If $X_1, X_2, \ldots, X_n$ are discrete, then they are independent if for all $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$, we have

$$P_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = P_{X_1}(x_1)P_{X_2}(x_2) \cdots P_{X_n}(x_n).$$

- If $X_1, X_2, \ldots, X_n$ are continuous, then they are independent if for all $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$, we have p

$$f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \cdots f_{X_n}(x_n)$$

- If random variables $X_1, X_2, \ldots, X_n$ are independent, then we have

$$E[X_1X_2 \cdots X_n] = E[X_1]E[X_2] \cdots E[X_n].$$

Definition 6.1

- Random variables are said to be **independent and identically distributed (i.i.d.)** if they are independent and have the same marginal distributions:

$$F_{X_1}(x) = F_{X_2}(x) = ... = F_{X_n}(x), \text{ for all } x \in \mathbb{R}$$

- Then the following must also be true

$$E[X_1X_2 \cdots X_n] = E[X_1]^n$$

8 Statistical Inference 1: Classical methods

- **Statistical Inference** is a collection of methods that deal with drawing conclusions from data that are prone to random variation

8.1 Random sampling

- The collection of random variables $X_1, X_2, X_3, \ldots, X_n$ is said to be a random sample of size n if they are independent and identically distributed i.e.
  1. $X_1, X_2, X_3, \ldots, X_n$ are independent random variables and
  2. they have the same distribution i.e.

$$F_{X_1}(x) = F_{X_2}(x) = ... = F_{X_n}(x), \qquad \text{for all } x \in \mathbb{R}.$$

- The **point estimator** is a function of the random sample $\hat{\Theta} = h(X_1, X_2, \cdots, X_n)$ that is used to estimate an unknown quantity

- Properties of random sampling
  1. the X_i's are independent
  2. $F_{X_1}(x) = F_{X_2}(x) = ... = F_{X_n}(x) = F_X(x)$
  3. $EX_i = EX = \mu < \infty$
  4. $0 < \text{Var}(X_i) = \text{Var}(X) = \sigma^2 < \infty$

- The **sample mean** is defined as

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}.$$

- Properties of the sample mean
  1. $\overline{EX} = \mu$
  2. $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$
  3. Weak Law of Large Numbers (WLLN):

$$\lim_{n\to\infty} P(|\overline{X} - \mu| \geq \epsilon) = 0.$$

  4. Central limit Theorem: The random variable

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + X_2 + ... + X_n - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to the standard normal random variable as n goes to infinity that is

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x), \qquad \text{for all } x \in \mathbb{R}$$

where $\Phi(x)$ is the standard normal CDF

- If we let $X_1, X_2, \ldots X_n$ be a random sample from a continuous distribution function with CDF $F_X(x)$ then if we order X_i's from the smallest to the largest and denote the resulting sequence of random variables as

$$X_{(1)}, X_{(2)}, \cdots, X_{(n)}.$$

- $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ is called the **order statistics** of the random sample $X_1, X_2, \ldots X_n$

- **Theorem 8.1:** Let $X_1, X_2, \ldots X_n$ be a random sample from a continuous distribution with CDF $F_X(x)$ and PDF $f_X(x)$. Let $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ be the order statistics of the random sample $X_1, X_2, \ldots X_n$. Then the CDF and PDF of $X_{(i)}$ are given by

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} f_X(x) \big[F_X(x)\big]^{i-1} \big[1 - F_X(x)\big]^{n-i}, \tag{32}$$

$$F_{X_{(i)}}(x) = \sum_{k=i}^{n} \binom{n}{k} \big[F_X(x)\big]^k \big[1 - F_X(x)\big]^{n-k}. \tag{33}$$

Also the join PDF of $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ is given by

$$f_{X_{(1)},\cdots,X_{(n)}}(x_1, x_2, \cdots, x_n) = \begin{cases} n! f_X(x_1) f_X(x_2) \cdots f_X(x_n) & \text{for } x_1 \leq x_2 \leq x_2 \cdots \leq x_n \\ 0 & \text{otherwise} \end{cases}$$

8.2 Point Estimation

- If θ is the unknown parameter to be estimated, where θ is a fixed (non-random) quantity, we estimate θ by defining a point estimator $\hat{\Theta}$ that is a function of the random sample i.e.

$$\hat{\Theta} = h(X_1, X_2, \cdots, X_n).$$

- If $\theta = EX$ we may choose the $\hat{\Theta}$ to be the sample mean

$$\hat{\Theta} = \overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

8.2.1 Evaluating estimators

- Let $\hat{\Theta} = h(X_1, X_2, \cdots, X_n)$ be a point estimator for θ. The **bias** of the point estimator $\hat{\Theta}$ is defined by

$$B(\hat{\Theta}) = E[\hat{\Theta}] - \theta.$$

- In general we want to have a bias close to 0
- Let $\hat{\Theta} = h(X_1, X_2, \cdots, X_n)$ be a point estimator for a parameter θ. We say that $\hat{\Theta}$ is an **unbiased** of estimator of θ if

$$B(\hat{\Theta}) = 0, \qquad \text{for all possible values of } \theta.$$

- The **mean squared error** (MSE) of a point estimator $\hat{\theta}$ is defined as

$$MSE(\hat{\Theta}) = E\big[(\hat{\Theta} - \theta)^2\big].$$

- The MSE is a measure of the distance between $\hat{\Theta}$ and θ
- A smaller MSE is generally indicative a better estimator
- The MSE can generally be written as

$$MSE(\hat{\Theta}) = \text{Var}(\hat{\Theta}) + B(\hat{\Theta})^2.$$

- We say that an estimator is **consistent** if as the sample size n get larger $\hat{\Theta}$ converges to the real value of θ
- Let $\hat{\Theta}\_1, \hat{\Theta}\_2, \ldots \hat{\Theta}\_n \ldots$, be a sequence of point estimators for θ. We say that $\hat{\Theta}_n$ is a **consistent** estimator of θ if

$$\lim_{n\to\infty} P\big(|\hat{\Theta}_n - \theta| \geq \epsilon\big) = 0, \text{ for all } \epsilon > 0.$$

Theorem 8.2

- Let $\hat{\Theta}\_1, \hat{\Theta}\_2, \ldots$ be a sequence of point estimators of θ if

$$\lim_{n\to\infty} MSE(\hat{\Theta}_n) = 0,$$

then $\hat{\Theta}_n$ is a consistent estimator of θ

8.2.2 For mean and variance

- Let $X_1, X_2, X_3 \ldots, X_n$ be a random sample with mean $EX_i = \mu < \infty$ and variance $0 < \text{Var}(X_i) = \sigma^2 < \infty$. Then the **sample variance** of this random sample is defined as

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{k=1}^{n} X_k^2 - n\overline{X}^2\right).$$

- The sample variance is an unbiased estimator of σ^2. The **sample standard deviation** is defined as

$$S = \sqrt{S^2}$$

and is commonly used as an estimator for σ. Nevertheless, S is a biased estimator for σ

8.2.3 Maximum likelihood estimation

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with a parameter θ. Suppose that we have observed $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$

  1. If X_i's are discrete, then the **likelihood function** is defined as

$$L(x_1, x_2, \cdots, x_n; \theta) = P_{X_1X_2\cdots X_n}(x_1, x_2, \cdots, x_n; \theta).$$

  2. If X_i's are jointly continuous, then the likelihood function is defined as

$$L(x_1, x_2, \cdots, x_n; \theta) = f_{X_1X_2\cdots X_n}(x_1, x_2, \cdots, x_n; \theta).$$

- In some problems it is easier to work with the **log likelihood function** given by

$$\ln L(x_1, x_2, \cdots, x_n; \theta).$$

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with a parameter θ. Given that we have observed $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ a maximum likelihood estimate of θ, shown by $\hat{\theta}_{ML}$ is the value of θ that maximizes the likelihood function

$$L(x_1, x_2, \cdots, x_n; \theta)$$

- A maximum estimator (MLE) of a parameter θ shown by $\hat{\Theta}_{ML}$ is a random variable
  - $\hat{\Theta}_{ML}(X_1, X_2, \cdots, X_n)$ whose value when $X_1, X_2, \ldots, X_n$ is given by $\hat{\theta}_{ML}$

- **Asymptotic properties of MLEs:** Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with a parameter θ. Let $\hat{\Theta}_{ML}$ denote the maximum likelihood estimator (MLE) of θ. Then under some mild regularity conditions:

  1. $\hat{\Theta}_{ML}$ is asymptotically consistent i.e.

  $$\lim_{n\to\infty} P(|\hat{\Theta}_{ML} - \theta| > \epsilon) = 0$$

  2. $\hat{\Theta}_{ML}$ is asymptotically unbiased i.e.

  $$\lim_{n\to\infty} E[\hat{\Theta}_{ML}] = \theta.$$

  3. As n becomes large $\hat{\Theta}_{ML}$ is approximately a random variable. More precisely the random variable

  $$\frac{\hat{\Theta}_{ML} - \theta}{\sqrt{\text{Var}(\hat{\Theta}_{ML})}}$$

  converges in distribution to $N(0,1)$

8.3 Interval estimation

- Let $X_1, X_2, \ldots X_n$ be a random sample from a distribution with parameter θ that is to be estimated, the goal is to find two estimators for θ

  1. the low estimator $\hat{\Theta}_l = \hat{\Theta}_l(X_1, X_2, \cdots, X_n)$
  2. the high estimator $\hat{\Theta}_h = \hat{\Theta}_h(X_1, X_2, \cdots, X_n)$

- The interval estimator is given by the interval $[\hat{\Theta}_l, \hat{\Theta}_h]$

- Let $X_1, X_2, \ldots X_n$ be a random sample from a distribution with parameter θ that is to be estimated. An **interval estimator** with **confidence level** $1-\alpha$ consists of two estimators $\hat{\Theta}_l(X_1, X_2, \cdots, X_n)$ and $\hat{\Theta}_h(X_1, X_2, \cdots, X_n)$ such that

  $$P\left(\hat{\Theta}_l \leq \theta \text{ and } \hat{\Theta}_h \geq \theta\right) \geq 1-\alpha,$$

  for every possible value of θ. Equivalently, we say that $[\hat{\Theta}_l, \hat{\Theta}_h]$ is a $(1-\alpha)100\%$ **confidence interval** for θ

8.3.1 Find interval estimators

Pivotal Quantity
Let $X_1, X_2, \ldots, X_n$ be an random sample from a distribution with parameter θ that is to be estimated. The random variable Q is said to be a pivot or a pivotal quantity, if it has the following properties

1. It is a function of the observed data $X_1, X_2, \ldots, X_n$ and the unknown parameter θ, but it does not depend on any other unknown parameters:

   $$Q = Q(X_1, X_2, \cdots, X_n, \theta).$$

2. The probability distribution of Q does not depend on θ or any other unknown parameters

Interval estimators:

- Assumptions: A random sample $X_1, X_2, \ldots, X_n$ is given from a distribution with known variance $Var(X_i) = \sigma^2 < \infty$; n is large
- Parameter to be Estimated: $\theta = EX_i$
- Confidence Interval: $\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$ is approximately a $(1-\alpha)100\%$ confidence interval for θ

- Assumptions: A random sample $X_1, X_2, \ldots, X_n$ is given from a $Bernoulli(\theta)$; n is large
- Parameter to be Estimated: θ
- Confidence Interval: $\left[\overline{X} - \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}\right]$ is approximately $(1-\alpha)100\%$ confidence interval for θ. This is a conservative confidence interval as it is obtained using an upper bound for σ

- Assumptions: A random sample $X_1, X_2, \ldots, X_n$ is given from a $Bernoulli(\theta)$; n is large
- Parameter to be Estimated: θ
- Confidence Interval: $\left[\overline{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}, \overline{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right]$ is approximately a $(1-\alpha)100\%$ confidence interval for θ

- Assumptions: A random sample $X_1, X_2, \ldots, X_n$ is given from a distribution with unknown variance $Var(X_i) = \sigma^2 < \infty$; n is large
- Parameter to be Estimated: $\theta = EX_i$
- Confidence Interval: If S is the sample standard deviation

$$S = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2} = \sqrt{\frac{1}{n-1}\left(\sum_{k=1}^{n}X_k^2 - n\overline{X}^2\right)}$$

then the interval

$$\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$$

is approximately a $(1-\alpha)100\%$ confidence interval for θ

8.4 Confidence intervals for normal samples

8.4.1 The Chi-Squared Distribution

- **Definition 8.1:** If $Z_1, Z_2, \ldots, Z_n$ are independent standard normal random variables, the random variable Y defined as
$$Y = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$
is said to have a chi-squared distribution with n degrees of freedom shown by
$$Y \sim \chi^2(n)$$

- Properties of the chi-squared distribution:

  1. The chi-squared distribution in a special case of the gamma distribution. More specifically
  $$Y \sim Gamma\left(\frac{n}{2}, \frac{1}{2}\right).$$
  Thus
  $$f_Y(y) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, \quad \text{for } y > 0.$$
  2. $EY = n$, $Var(Y) = 2n$
  3. For any $P \in [0, 1]$ and $n \in \mathbb{N}$, we define $\chi^2_{p,n}$ as the real value for which
  $$P(Y > \chi^2_{p,n}) = p,$$
  where $U \sim \chi^2(n)$

- **Theorem 8.3:** Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(\mu, \sigma)$ random variables. Also let S^2 be the standard variance for this random sample. Then the random variable Y is defined as
$$Y = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$$
has a chi-squared distribution with $n-1$ degrees of freedom, i.e., $Y \sim \chi^2(n-1)$. Moreover, $\hat{X}$ and S^2 are independent random variables.

8.4.2 The t Distribution

- **Definition 8.2:** Let $Z \sim N(0, 1$, and $Y \sim \chi^2(n)$, where $n \in \mathbb{N}$. Also assume that Z and Y are independent. The random variable T defined as
$$T = \frac{Z}{\sqrt{Y/n}}$$
is said to have a t-distribution with n degrees of freedom shown by
$$T \sim T(n)$$

- Properties:

1. The t-distribution has a bell-shaped PDF centered at 0, but its PDF is more spread out than the normal PDF
2. $ET = 0$, for $n > 0$. But ET, is undefined for $n = 1$.
3. $Var(T) = \frac{n}{n-2}$ for $n < 2$. But variance is undefined for $n = 1, 2$
4. As n becomes large, the t density approaches the standard normal PDF. More formally, we can write
$$T(n) \xrightarrow{d} N(0,1).$$
5. For any $p \in [0,1]$ and $n \in$, we define $t_{p,n}$ as the real value for which
$$P(T > t_{p,n}) = p.$$

- Since the t-distribution has a symmetric PDF, we have
$$t_{1-p,n} = -t_{p,n}.$$

- **Theorem 8.4:** Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(\mu, \sigma)$ random variables. Also let S^2 be the standard variance for this random sample. Then let the random variable T defined as
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$
has a t-distribution with $n-1$ degrees of freedom, i.e. $T \sim T(n-1)$

8.4.3 More Interval estimators

- Assumptions: A random sample $X_1, X_2, \ldots, X_n$ is given from a $N(\mu, \sigma^2)$ distribution where $Var(X_i) = \sigma^2$ is known
- Parameter to be estimated: $\mu = EX_i$
- Confidence interval $\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$ is a $(1-\alpha)100\%$ confidence interval for mu

- Assumptions: A random sample $X_1, X_2, \ldots, X_n$ is given from a $N(\mu, \sigma^2)$ distribution where $Var(X_i) = \sigma^2$ are unknown
- Parameter to be estimated: $\mu = EX_i$
- Confidence interval $\left[\overline{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right]$ is a $(1-\alpha)100\%$ confidence interval for mu

- Assumptions: A random sample $X_1, X_2, \ldots, X_n$ is given from a $N(\mu, \sigma^2)$ distribution where $\mu = EX_i$ and $Var(X_i) = \sigma^2$ are unknown
- Parameter to be estimated: $Var(X_i) = \sigma^2$
- Confidence interval $\left[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right]$ is a $(1-\alpha)100\%$ confidence interval for mu

8.5 Hypothesis testing

- Let S be the set of possible values for θ, we can partition S into two disjoint sets S_0 and S_1. Then let H_0 be the hypothesis that $\theta \in S_0$ and let H_1 be the hypothesis that $\theta \in S_1$
  - H_0 (the **null** hypothesis): $\theta \in S_0$
  - H_1 (the **alternative** hypothesis): $\theta \in S_1$

- **Definition 8.3:** Let $X_1, \ldots, X_n$ be a random sample of interest. A **statistic** is a real valued function of data. For example the sample mean defined as

$$W(X_1, X_2, \cdots, X_n) = \frac{X_1 + X_2 + ... + X_n}{n},$$

  is a statistic. A **test statistic** is a statistic based on which we build our test

- To decide whether to choose H_0 or H_1 we
  - choose a test statistic $W(X_1, X_2, \cdots, X_n)$
  - define a set $A \subset \mathbb{R}$ as the possible values of W for which we would accept H_0 called the **acceptance region**
  - the set $R = \mathbb{R} - A$ is said to be the **rejection region**

- **Type 1 error** is defined as the event that we reject H_0 when H_0 is true.
  - The probability of a type 1 error is

$$P(\text{type I error} \mid \theta) = P(\text{Reject } H_0 \mid \theta) \tag{34}$$
$$= P(W \in R \mid \theta), \quad \text{for } \theta \in S_0. \tag{35}$$

    If the probability of a type 1 error satisfies

$$P(\text{type I error}) \leq \alpha, \quad \text{for all } \theta \in S_0,$$

    then we say that the test has **significance level** α or simply the test is a α test

- **Type 2 error** is defined as the event that we accept H_0 when H_0 is false.
  - The probability is a function of θ and is shown by β:

$$\beta(\theta) = P(\text{Accept } H_0 \mid \theta), \quad \text{for } \theta \in S_1.$$

8.5.1 Hypothesis test for mean

- To decide between the following hypothesis

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

the null hypothesis is a simple hypothesis and the alternative is a two-sided, this hypothesis test is called a two sided hypothesis test

Table 8.2: Two-sided hypothesis testing for the mean: $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$.

Case	Test Statistic	Acceptance Region
$X_i \sim N(\mu, \sigma^2)$, σ known	$W = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\lvert W \rvert \leq z_{\frac{\alpha}{2}}$
n large, X_i non-normal	$W = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\lvert W \rvert \leq z_{\frac{\alpha}{2}}$
$X_i \sim N(\mu, \sigma^2)$, σ unknown	$W = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\lvert W \rvert \leq t_{\frac{\alpha}{2}, n-1}$

- If the hypothesis is defined as follows

$$\mu \leq \mu_0 \vee \mu \geq mu_0$$

$$\mu > \mu_0 \vee \mu < mu_0$$

the null hypothesis and the alternative are one-sided this is called a one-sided hypothesis test:

Table 8.3: One-sided hypothesis testing for the mean: $H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$.

Case	Test Statistic	Acceptance Region
$X_i \sim N(\mu, \sigma^2)$, σ known	$W = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$W \leq z_\alpha$
n large, X_i non-normal	$W = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$W \leq z_\alpha$
$X_i \sim N(\mu, \sigma^2)$, σ unknown	$W = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$W \leq t_{\alpha, n-1}$

Table 8.4: One-sided hypothesis testing for the mean: $H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$.

Case	Test Statistic	Acceptance Region
$X_i \sim N(\mu, \sigma^2)$, σ known	$W = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$	$W \geq -z_\alpha$
n large, X_i non-normal	$W = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$	$W \geq -z_\alpha$
$X_i \sim N(\mu, \sigma^2)$, σ unknown	$W = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$	$W \geq -t_{\alpha,n-1}$

8.5.2 P-values

- **P-value** is the lowest significance level α that results in rejecting the null hypothesis
- Consider a hypothesis test for choosing between H_0 and H_1. Let W be the test statistic, and w_1 be the observed value of W

  1. Assume H_0 is true
  2. The P-value is $P(\text{type I error})$ then the test threshold c is chosen to be $c = w_1$

8.5.3 Likelihood ratio test

- Let $X_1, \ldots, X_n$ be a random sample from a distribution with parameter θ. Suppose we have observed $X_1 = x_1, \ldots X_n = x_n$. To decide between two simple hypothesis

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta = \theta_1$$

  we define

$$\lambda(x_1, x_2, \cdots, x_n) = \frac{L(x_1, x_2, \cdots, x_n; \theta_0)}{L(x_1, x_2, \cdots, x_n; \theta_1)}.$$

  To perform a **likelihood ratio test (LRT)**, we choose a constant c. We reject H_0 if $\lambda < c$ and accept it if $\lambda > c$. The value of c can be chosen based on the desired a

- Let $X_1, \ldots, X_n$ be a random sample from a distribution with parameter θ. Suppose we have observed $X_1 = x_1, \ldots X_n = x_n$. Define

$$\lambda(x_1, x_2, \cdots, x_n) = \frac{\sup\{L(x_1, x_2, \cdots, x_n; \theta) : \theta \in S_0\}}{\sup\{L(x_1, x_2, \cdots, x_n; \theta) : \theta \in S\}}.$$

  To perform a **likelihood ration test (LRT)**, we choose a constant c in $[0, 1]$. We reject H_0 if $\lambda < c$ and accept it if $\lambda > c$. The value of c can be chosen based on the desired a

9 Discrete-Time Markov Chains

- Consider the random process $\{X_n, n = 1, 2, \cdots\}$, where $R_{X_i} = S \subset \{0, 1, 2, \cdots\}$. We say this process is a Markov chain if

$$P(X_{m+1} = j | X_m = i, X_{m-1} = i_{m-1}, \cdots, X_0 = i_0) = P(X_{m+1} = j | X_m = i),$$

  for all $m, j, i, i_0, i_1, \cdots, i_{m-1}$. If the number of states is finite. e.g. $S = \{0, 1, 2, \cdots, r\}$ we call it a **finite** Markov chain
- If $X_n = j$, we say that the process is in state j
- The numbers $P(X_{m+1} = j | X_m = i)$ are called the **transition probabilities**
  - It is assumed that they do not depend on time. That is $P(X_{m+1} = j | X_m = i$ do not depend on m
- The following is defined

$$p_{ij} = P(X_{m+1} = j | X_m = i).$$

  we have in particular

$$\begin{aligned} p_{ij} &= P(X_1 = j | X_0 = i) && (36)\\ &= P(X_2 = j | X_1 = i) && (37)\\ &= P(X_3 = j | X_2 = i) = \cdots. && (38) \end{aligned}$$

  if the process is in state i, it will a transition to state j with probability p_{ij}.

9.1 State Transition Matrix and Diagram

- The transitions probabilities is often listed in a matrix
  - is called the **state transition matrix** or **transition probability matrix** and is usually shown by P
  - Assuming the states are $1, 2, \cdots, r$, then the state transition matrix is shown by

$$P = \begin{bmatrix} p_{11} & p_{12} & ... & p_{1r} \\ p_{21} & p_{22} & ... & p_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ p_{r1} & p_{r2} & ... & p_{rr} \end{bmatrix}.$$

    for $p_{ij} \geq 0$, and for all i we have

$$\sum_{k=1}^{r} p_{ik} = \sum_{k=1}^{r} P(X_{m+1} = k | X_m = i) = 1.$$

- A Markov chain is usually shown by a **state transition diagram**
  - If there is no arrow from state i to state j, then $p_{ij} = 0$

9.2 Probability distributions

- Given a Markov chain $\{X_n, n = 0, 1, 2, ...\}$ where $X_n \in S = \{1, 2, \cdots r\}$. The following is defined
$$\pi^{(n)} = \begin{bmatrix} P(X_n = 1) & P(X_n = 2) & \cdots & P(X_n = r) \end{bmatrix},$$
- The following is then true
$$\pi^{(n+1)} = \pi^{(n)} P, \text{ for } n = 0, 1, 2, \cdots;$$
$$\pi^{(n)} = \pi^{(0)} P^n, \text{ for } n = 0, 1, 2, \cdots.$$
- Given a Markov chain $\{X_n, n = 0, 1, 2, ...\}$ the n-step probabilities $p_{ij}^{(n)}$ is defined as
$$p_{ij}^{(n)} = P(X_n = j | X_0 = i), \text{ for } n = 0, 1, 2, \cdots,$$
and the n-step transition matrix, $P^{(n)}$, as
$$P^{(n)} = \begin{bmatrix} p_{11}^{(n)} & p_{12}^{(n)} & ... & p_{1r}^{(n)} \\ p_{21}^{(n)} & p_{22}^{(n)} & ... & p_{2r}^{(n)} \\ \vdots & \vdots & \vdots & \vdots \\ p_{r1}^{(n)} & p_{r2}^{(n)} & ... & p_{rr}^{(n)} \end{bmatrix}.$$
- The Chapman-Kolmogorov equation can be written as
$$\begin{aligned} p_{ij}^{(m+n)} &= P(X_{m+n} = j | X_0 = i) \\ &= \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n)}. \end{aligned}$$
- The n step matrix is given by
$$P^{(n)} = P^n, \text{ for } n = 1, 2, 3, \cdots.$$

9.3 Classification of States

- A state j is **accessible** from state i, written as $i \to j$, if $p_{ij}^{(n)} > 0$ for some n. It is assumed that every state is accessible from itself since $p_{ii}^{(0)} = 1$
- Two state i and j are said to **communicate**, written as $i \leftrightarrow j$ if they are accessible from each other, in other words
$$i \leftrightarrow j \text{ means } i \to j \text{ and } j \to i.$$
- Communication is an **equivalence relation**, which means
  - every state communicates with itself $i \leftrightarrow i$;
  - if $i \leftrightarrow j$ then $j \leftrightarrow i$;

- if $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$

- Markov chains can be partitioned into communicating classes such that only member of the same class communicate with each other
- A Markov chain is said to be **irreducible** if all the states communicate with each other

- For any state i we define

$$f_{ii} = P(X_n = i, \text{ for some } n \geq 1 | X_0 = i).$$

  State i is **recurrent** if $f_{ii} = 1$, and is **transient** if $f_{ii} < 1$
- If two states are in the same class they are either both recurrent or transient
  - Therefore we can say that a class is transient if all the states in it are transient and the same is the case for recurrent
- Consider a discrete-time Markov chain. Let V be the total number of visits to state i
  1. If i is a recurrent state, then
$$P(V = \infty | X_0 = i) = 1.$$
  2. If i is a transient state, then
$$V | X_0 = i \ \sim \ Geometric(1 - f_{ii}).$$

- The **period** of a state i is the largest integer d satisfying the following property: $p_{ii}^{(n)} = 0$ whenever n is not divisible by d. The period of i is shown as $d(i)$. If $p_{ii}^{(n)} = 0$ for all $n > 0$, we let $d(i) = \infty$
  - If $d(i) < 1$, we say that state i is **periodic**
  - If $d(i) = 1$, we say that state i is **aperiodic**
- The states in the same communicating class have the same period
  - a class therefor can be periodic or aperiodic

$$\text{If } i \leftrightarrow j, \text{ then } d(i) = d(j)$$

- Consider a finite irreducible Markov chain X_n
  1. If there is a self-transition in the chain ($p_{ii} > 0$ for some i), then the chain is aperiodic.
  2. Suppose that you can go from state i to state i in l steps i.e. $p_{ii}^{(l)} > 0$. Also suppose that $p_{ii}^{(m)} > 0$ for some m. If $gcd(l, m) = 1$, then state i is aperiodic.
  3. The chain is aperiodic if and only if there exists a positive integer n such that all the elements of the matrix P^n are strictly positive, i.e.

$$p_{ij}^{(n)} > 0, \ \text{ for all } i, j \in S.$$

9.4 Using the Law of Total Probability with Recursion

9.4.1 Absorption Probabilities

- Consider a finite Markov chain $\{X_n, n = 0, 1, 2, \cdots\}$ with state space $S = \{0, 1, 2, \cdots r\}$. Suppose that all states are either absorbing or transient. Let $l \in S$ be an absorbing state. Define
$$a_i = P(\text{absorption in } l | X_0 = i), \quad \text{for all } i \in S.$$
By the above definition, we have $a_l = 1$, and $a_j = 0$ if j is another absorbing state. To find unknown values of a_i's we can use the following equations
$$a_i = \sum_k a_k p_{ik}, \quad \text{for } i \in S.$$

9.4.2 Mean Hitting Times

- Consider a finite Markov chain $\{X_n, n = 0, 1, 2, \cdots\}$ with state space $S = \{0, 1, 2, \cdots r\}$. Let $A \subset S$ be a set of states. Let T be the first time the chain visits a state in A. For all $i \in S$ define
$$t_i = E[T | X_0 = i].$$
- By the above definition we have, we have $t_j = 0$, for all $j \in A$. To find the unknown values of t_i's we can use the following equations
$$t_i = 1 + \sum_k t_k p_{ik}, \quad \text{for } i \in S - A.$$

9.4.3 Mean Return Times

- Consider a finite Markov chain $\{X_n, n = 0, 1, 2, \cdots\}$ with state space $S = \{0, 1, 2, \cdots r\}$. Let $l \in S$ be a state. Let r_l be the **mean return time** to state l. Then
$$r_l = 1 + \sum_k t_k p_{lk},$$
where t_k is the expected time until the chain hits state l given $X_0 = k$. Specifically
$$\begin{aligned} t_l &= 0, \\ t_k &= 1 + \sum_j t_j p_{kj}, \quad \text{for } k \neq l. \end{aligned}$$

9.5 Limiting and stationary distributions

- The probability distribution $\pi = [\pi_0, \pi_1, \pi_2, \cdots]$ is called the **limiting distribution** of the Markov chain X_n if
$$\pi_j = \lim_{n \to \infty} P(X_n = j | X_0 = i)$$
for all $i, j \in S$ we have
$$\sum_{j \in S} \pi_j = 1.$$

- Consider a finite Markov chain $\{X_n, n = 0, 1, 2, \cdots\}$ where $X_n \in S = \{1, 2, \cdots, r\}$. Assume that the chain is irreducible and aperiodic. Then

1. The set of equations

$$\pi = \pi P,$$
$$\sum_{j \in S} \pi_j = 1$$

has a unique solution
2. The unique solution to the above equations is the limiting distribution of the Markov chain i.e.

$$\pi_j = \lim_{n \to \infty} P(X_n = j | X_0 = i),$$

for all $i, j \in S$
3. We have

$$r_j = \frac{1}{\pi_j}, \quad \text{for all } j \in S,$$

where r_j is the mean return time to state j

9.6 Countably Infinite Markov Chains

- Let i be a recurrent state. Assuming $A_0 = i$, let R_i be the number for transitions needed to return to i, i.e.

$$R_i = \min\{n \geq 1 : X_n = i\}.$$

If $r_i = E[R_i | X_0 = i] < \infty$, then i is said to be **positive recurrent**. If $E[R_i | X_0 = i] = \infty$, then i is said to be **null recurrent**

Theorem 11.2

- Consider an infinite Markov chain $\{X_n, n = 0, 1, 2, ...\}$ where $X_n \in S = \{0, 1, 2, \cdots\}$. Assume that the chain is irreducible and and aperiodic. The one of the following cases can occur:

1. All states are transient and

$$\lim_{n \to \infty} P(X_n = j | X_0 = i) = 0, \text{ for all } i, j.$$

2. All states are null recurrent

$$\lim_{n \to \infty} P(X_n = j | X_0 = i) = 0, \text{ for all } i, j.$$

3. All states are positive recurrent. In this case, there exists a limiting distribution, $\pi = [\pi_0, \pi_1, \cdots]$, where

$$\pi_j = \lim_{n \to \infty} P(X_n = j | X_0 = i) > 0,$$

for all $i, j \in S$. The limiting distribution in unique solution to the equations

$$\pi_j = \sum_{k=0}^{\infty} \pi_k P_{kj}, \quad \text{for } j = 0, 1, 2, \cdots,$$
$$\sum_{j=0}^{\infty} \pi_j = 1.$$

We also have

$$r_j = \frac{1}{\pi_j}, \quad \text{for all } j = 0, 1, 2, \cdots,$$

where r_j is the mean return time to state j