U.P. Rajarshi Tandon Open
University, Prayagraj

MScSTAT – 101N/ MASTAT – 101N
Measure & Probability Theory

Course Design Committee

Dr. Ashutosh Gupta **Chairman**
Director, School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. S. Lalitha **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics
D. D. U. Gorakhpur University, Gorakhpur.

Prof. Shruti **Member-Secretary**
Professor, School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Dr. Ashok Kumar **Writer**
Department of Statistics ***(Unit 1-7)***
University of Lucknow, Lucknow

Dr. Pratyasha Tripathi **Writer**
Department of Statistics ***(Unit 8-11)***
Tilka Manjhi Bhagalpur University, Bhagalpur, Bihar

Prof. Shruti **Editor**
School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj

Prof. Shruti **Course Coordinator**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 101N/ MASTAT – 101N **MEASURE & PROBABILITY THEORY**
©UPRTOU
First Edition: ***July 2023***
ISBN: 978-81-19530-86-1

Printed and Published by Col. Vinay Kumar, Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2024.

Printed By : Chandrakala Universal Pvt. Ltd. 42/7 Jawahar Lal Neharu Road, Prayagraj.

SUSHI-052

Blocks & Units Introduction

The present SLM on ***Measure and Probability Theory*** consists of eleven units with three blocks.

The ***Block - I – Measure Theory,*** is the first block, which is divided into four units.

The ***Unit - 1 – Measure,*** is the first unit of present self-learning material describes Field, -Field, Borel Field. Measure, Measure on R^n, Properties of Measure, Outer Measure, Extension of measures, Extension Theorem, Outer Extension. Simple Functions, Integration, Non-Negative Integrable Functions, Integrable Measurable Functions.

In ***Unit – 2 – Convergence,*** the main emphasis on the Measure Space, Measurable Functions, Combinations of measurable function, point wise Convergence, Convergence in measure.

In ***Unit – 3 – Lebesgue Measure,*** we have focussed mainly on Lebesgue-Stielitjes Measure, Lebesgue-Stieltjes Integral, Riemann-Stieltjes Integration, Lebesgue Dominated Convergence Theorem, Monotone Convergence Theorem, Fatou Lemma, Fubini's Theorem.

In ***Unit – 4 – Signed Measures,*** is being introduced the Signed measures, Hahn and Jordan decomposition, Absolute Continuity, The Radon-Nikodym Theorem, Derives of Signed Measures. Product Space, Cartesian Products of two Measurable Spaces, Section, Product Measures.

The ***Block - II – Probability Measure, Distribution Function and Inequalities*** is the second block in which we have three units.

In ***Unit – 5 – Probability Measure*** is discussed with Probability space of a random experiment, probability measures, random variables as a measurable function. Field induced by a sequence of random variables.

In ***Unit – 6 – Distribution Functions*** has been introduced by discussing Decomposition of distribution functions in purely discrete, absolutely continuous and singular components.

The ***Unit –7 - Probability Inequalities*** dealt with CR-Inequality, Chebyshev's Inequality, Cauchy-Schwartz Inequality, Holder Inequality, Minkowski Inequality, Jensen Inequality, Lyapunov Inequality, Kolmogorov Inequality, Hajck-Renyki Inequality.

The ***Block - III – Convergence, Characteristics Function and Limit Theorems*** has four units.

Unit – 8 – Convergence dealt with Sequences of distribution functions, weak and complete convergence of sequence of distribution function, Different types of convergence of sequence of random variables distribution function of random vectors.

Unit –9 – Law of Large Numbers, comprises the Weak Law of Large Numbers (WLLN), Strong Law of Large Numbers (SLLN), Khinchin's Theorem, Borel Zero-One Law, Borel-Cantelli Lemmas.

In ***Unit – 10 – Characteristic Function***, we have discussed the Helly – Bray Lemma and Theorem, Weak Compactness Theorem, Kolmogorav Theorems, Characteristic Function, Inversion Theorem, Continuity Theorem, Uniqueness Theorem.

The ***Unit – 11 – Central Limit Theorems*** discussed One Dimensional Central Limit Problem: Lindeberg-Levy, Lyapunov, Lindeberg-Feller Theorems.

At the end of every block/unit the summary, self-assessment questions and further readings are given.

References

- Billingsley, P. (1995). Probability and Measure. John Wiley & Sons.
- Breiman, L. (1992). Probability. SIAM.
- Chow, Y. S., & Teicher, H. (1988). Probability Theory: Independence, Interchangeability, Martingales. Springer.
- Durrett, R. (2010). Probability: Theory and Examples. Cambridge University Press.
- Feller, W. (1968). An Introduction to Probability Theory and Its Applications, Vol. 1. John Wiley & Sons.
- Khinchin, A. (1943). Mathematical Foundations of Statistical Mechanics. Dover Publications.
- Kolmogorov, A. N. (1950). Foundations of the Theory of Probability. Chelsea Publishing Co.
- Donald L. Cohn (2010). Measure Theory. Springer Science+Business Media.
- Ross, S. (2014). A First Course in Probability. Pearson.
- Shiryaev, A. N. (1996). Probability. Springer.

Further Reading

- "An Introduction to Probability Theory and Its Applications, Vol. 1" by William Feller, John Wiley & Sons.
- “An Introduction to Measure-Theoretic Probability” by George G. Roussas, Academic Press.
- "Probability and Measure" by Patrick Billingsley, John Wiley & Sons.
- "A First Course in Probability" by Sheldon Ross, Pearson.
- "The Law of Large Numbers: Probability and Statistics and Their Applications" by M. Loève, Chelsea Publishing.
- "The Doctrine of Chances: Probabilistic Aspects of Gambling" by Stewart N. Ethier, Springer.
- "Statistics" by Robert S. Witte and John S. Witte, Wiley.

U.P. Rajarshi Tandon Open
University, Prayagraj

MScSTAT – 101N/ MASTAT – 101N

Measure & Probability Theory

Block: 1 Measure Theory

Course Design Committee

Dr. Ashutosh Gupta Director, School of Sciences U. P. Rajarshi Tandon Open University, Prayagraj	**Chairman**
Prof. Anup Chaturvedi Ex. Head, Department of Statistics University of Allahabad, Prayagraj	**Member**
Prof. S. Lalitha Ex. Head, Department of Statistics University of Allahabad, Prayagraj	**Member**
Prof. Himanshu Pandey Department of Statistics D. D. U. Gorakhpur University, Gorakhpur.	**Member**
Prof. Shruti Professor, School of Sciences U.P. Rajarshi Tandon Open University, Prayagraj	**Member-Secretary**

Course Preparation Committee

Dr. Ashok Kumar Department of Statistics University of Lucknow, Lucknow	**Writer**
Prof. Shruti School of Sciences U.P. Rajarshi Tandon Open University, Prayagraj	**Editor**
Prof. Shruti School of Sciences, U. P. Rajarshi Tandon Open University, Prayagraj	**Course Coordinator**

MScSTAT – 101N/ MASTAT – 101N **MEASURE & PROBABILITY THEORY**
©UPRTOU
First Edition: ***July 2023***
ISBN: 978-81-19530-86-1

Printed and Published by Col. Vinay Kumar, Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2024.

Printed By : Chandrakala Universal Pvt. Ltd. 42/7 Jawahar Lal Neharu Road, Prayagraj.

Block & Unit Introduction

The ***Block - I – Measure Theory,*** is the first block, which is divided into four units.

The ***Unit - 1 – Measure,*** is the first unit of present self-learning material describes Field, -Field, Borel Field. Measure, Measure on R^n, Properties of Measure, Outer Measure, Extension of measures, Extension Theorem, Outer Extension. Simple Functions, Integration, Non-Negative Integrable Functions, Integrable Measurable Functions.

In ***Unit – 2 – Convergence,*** the main emphasis on the Measure Space, Measurable Functions, Combinations of measurable function, point wise Convergence, Convergence in measure.

In ***Unit – 3 – Lebesgue Measure,*** we have focussed mainly on Lebesgue-Stielitjes Measure, Lebesgue-Stieltjes Integral, Riemann-Stieltjes Integration, Lebesgue Dominated Convergence Theorem, Monotone Convergence Theorem, Fatou Lemma, Fubini's Theorem.

In ***Unit – 4 – Signed Measures,*** is being introduced the Signed measures, Hahn and Jordan decomposition, Absolute Continuity, The Radon-Nikodym Theorem, Derives of Signed Measures. Product Space, Cartesian Products of two Measurable Spaces, Section, Product Measures.

At the end of every unit the summary, self-assessment questions and further readings are given.

UNIT: 1 MEASURE

Structure

1.1 Introduction

The introduction of the important classes of sets in an abstract space, which are those of a field, a σ-field, including the Borel σ-field, and a monotone class. They are illustrated by concrete examples, and their relationships are studied.

The concept of a measure is defined, and some of its basic properties are established. We then proceed with the introduction of an outer measure, study its relationship to the underlying measure, and determine the class of sets measurable with respect to the outer measure. These results are used as a basis toward obtaining an extension of a given measure from a field to the σ-field generated by this field.

1.2 Objectives

By the end of this unit, the learner should be able to:

- Understand the basic concept of measure theory.
- Distinguish between field and sigma field.
- Describe the properties of measure and their extension.
- Identify various types of measure.
- Understand the concept of measurable function.

1.3 Concept of Set Theory

Set: Set is a collection of well-defined and distinct objects and it is denoted by capital letters of English alphabets. E.g. – $A, B, C, \dots X, Y, Z$

Thus, the $A = \{1, 2, 3\}$ is a set but $B = \{1, 1, 3\}$ is not because 1 appears twice in the second collection.

The set of Natural numbers $N = \{1,2,3, \dots\}$

The set of vowel letters $V = \{a, e, i, o, u\}$

The set $A = \{a, b, c\}$

The set of integers $I = \{\dots, -3, -2, -1,0, 1, 2, 3, \dots\}$

$X = \{x: x\varepsilon z^{+}, z^{+} < 20\}$

$N = \{1,2,3, \dots \dots \dots .19\}$

$\mathcal{Y} = \{set\ of\ complex\ no.\}$

1.3.1 Different Types of Sets

Singleton Set: If there is only one object or element in a set, then it is called a singleton set. E.g., the set $A = \{a\}$ is a singleton set with only one element a.

Finite Set: If the elements in a set are finite in number, then it is called a finite set. Thus $S = \{1, 3, 9, 27\}$ is a finite set.

Infinite Set: If the elements in a set are infinite in number, it is called an infinite set. Thus, $S = \{x : x = 3n,$ where n is a whole number$\}$

= $\{1, 3, 9, 27, 81 \ldots\ldots.\}$ is an infinite set.

Empty Set: A set containing no elements is called an empty set or void set or null set and it is denoted by $\{\phi\}$ or ϕ.

Set of Sets: A set itself may sometimes be an element of another set, i.e., the objects or elements of a set may be sets themselves, then the latter set is called the set of sets. E.g., the set of all lines in a plane, since a line itself is a set of points.

Equal Sets: We say two sets S and T are equal if they have exactly same elements i.e., each element of set S is equal to each element of T and indicate this by writing $S = T$.

Proper subset: A set T of S will be called a proper subset if $T < S$ and we write this fact by the notation $T \subset S$, which implies that every element of T is in S, and S contains at least one element which does not belongs to T.

Note: The empty set $\emptyset$ is a proper subset of every set except itself. If a set contains n elements, then 2^n subset can be obtained.

Power Set: A set formed by all the subsets of set S including the set itself and the empty or null set as its elements is called the power set of S. The power set is denoted by $P(S)$.

E.g., if the set $A = \{1, 2, 3\}$ is then the power set of the set is $\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}$.If the number of elements in $S = n$, then the number of elements in $P(S) = 2^n$.

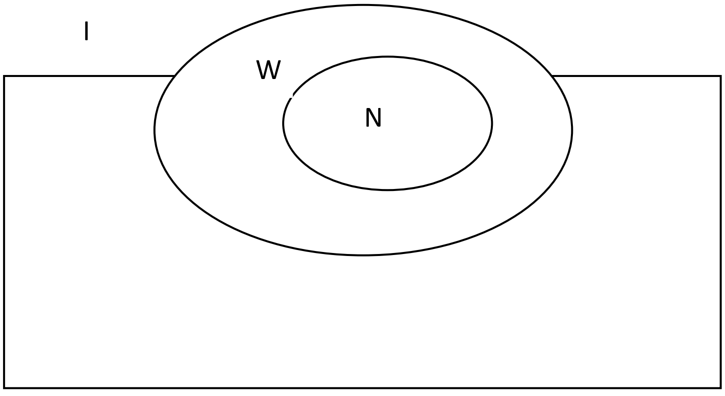
Countable and Uncountable Sets: A set S whose elements can be put in one to one corresponding with a set I of positive integers is called a countable or enumerable set, otherwise it is called uncountable or non-enumerable.

Cardinality: Number of elements in a set is called cardinality. The set S cardinality is denoted by $|S|$.
E.g., for the set $S = \{2, 4, 6\}$, the cardinality is $|S| = 3$.

Universal Set: A set which contains all objects including itself without repetition of elements. It is denoted by U.
E.g., U = {set of starts in the sky}, U = {Set of people on the earth},
U = {Set of natural numbers}

Complement Set: The complement set is a set in which has all the elements of universal set except the elements given in the set. For set S, the compliment set is denoted by S' or $\bar{S}$.
E.g., Let $U = \{1, 2, 3, 4, 5, 6, 7\}$ and $S = \{2, 4, 6\}$, then the complement set of S is $\bar{S} = \{1, 3, 5, 7\}$.

Venn Diagram: Venn Diagrams are the diagrams which represent the logical relationship between sets.
E.g., the set of natural numbers (N) is a subset of set of whole numbers (W) which is a subset of integers (I). We can represent this relationship through Venn diagram in the following way.



1.3.2 Operations on Sets

The operations on sets are performed on two or more sets to obtain a combination of elements as per the operation performed on them. In a set theory, there are three major types of operations performed on sets, such as:

(i) **Union of Sets:** If two sets A and B are given, then the union of A and B is equal to the set that contains all the elements present in set A and set B and it is denoted by $A \cup B$.
E.g., let set $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5, 6\}$,
then $A \cup B = \{1, 2, 3, 4, 3, 4, 5, 6\} = \{1, 2, 3, 4, 5, 6\}$.

(ii) **Intersection of Sets:** If two sets A and B are given, then the intersection of A and B is equal to the set that contains the common elements present in set A and set B and it is denoted by $A \cap B$.
E.g., let set $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5, 6\}$, then $A \cap B = \{3, 4\}$.

(iii) **Difference of Sets:** The difference of two sets A and B is defined as set of elements which belong to A but not to B and is denoted by $A - B$.
E.g., let set $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5, 6\}$, then $A - B = \{1, 2\}$.

Index and Indexed Set: Let A_r be a non-empty set, for each r in a set ∇, where $\nabla = \{1, 2, 3, \dots, r, \dots\}$. Here, $A_1, A_2, A_3, \dots, A_r, \dots$ called indexed set and $\Delta = \{1, 2, 3, \dots\}$ is known as index set.

Notation:

$\{A_r : r \in \Delta\}\ or\ \{A_r\}_{r \in \Delta}$

$A_1 = \{1, 2, 3\}$

$A_2 = \{4, 5, 6\}$

$A_3 = \{2, 3, 4, 5\}$

$\Delta = \{1, 2, 3\}\ or\ (chain)$

Hereditary Property: A non-empty family $A, \{A_r\}$ of sets is said to be hereditary, if $A_r \subset A_s, A_s \subset A \Rightarrow A_r \subset A$

Therefore, $a + b = b + a$

$a + (b + c) = (a + b) + c$

$a\,(b+c) \;=\; a.b \;+\; a.c.$

$a \;+\; (b.c) \;=\; (a+b)\,.\,(a+c)$

Theorems:

(A) Commutative Law:

(i) $A \cup B = B \cup A$

(ii) $A \cap B = B \cap A$

(B) Associative Law:

(i) $(A \cup B) \cup C = A \cup (B \cup C)$

(ii) $(A \cap B) \cap C = A \cap (A \cap C)$

(C) Distribution law:

(i) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

(ii) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

(iii) $A - (B \cup C) = (A - B) \cup (A - C)$

(iv) $A - (B \cap C) = (A - B) \cup (A - C)$

(D) De. Morgan Law:

(i) $(A \cup B)' = A' \cap B'$

(ii) $(A \cap B)' = A' \cup B'$

1.4 Ring and σ- Ring

A non-empty subset of Ω is called a ring of sets, if ant $A, B \in \underline{C}$ (family of C):

(i) $A - B \in \underline{C}$

(ii) $A \cup B \in \underline{C}$

It is closed under the formation of unions or differences

$$A \cap B = A - (A - B), \qquad A\Delta B = (A - B) \cup (B - A)$$

where $A \,\Delta\, B$ stands for the symmetric difference of A and B implies

$A \cap B \in \underline{C},\; A\Delta B \;\in \underline{C},$ if $\underline{C}$ is a ring of set. Also, $\emptyset \in\; \underline{C}, \emptyset = A - A.$

A non-empty family $\underline{C}$ of subset of Ω is called a σ- *ring* of sets if it is closed under the formation of differences and countable Unions that is if,

$A, B \in \underline{C} = A - B \in \underline{C}$ and, $A;\ \in \underline{C} = \bigcup_{i=1}^{\infty} A_i \in \underline{C}$

if $\bigcap_{i=1}^{\infty} A_i = A - A_i$, where $A = \bigcap_{i=1}^{\infty} A_i$

then, this equality says that σ- ring is also closed under the formation of countable intersections.

1.5 Fields, σ- Field and Boral σ- Field

A *collection of subsets* $C$of a set Ω is termed as *class of subsets* of Ω. Let C be the class of subsets of Ω, then

(i) **Union:** C is said to be closed under the union, if for any sets $A, B \in C, A \cup B \in C$.

(ii) **Intersection:** C is said to be closed under the intersection, if for any sets $A, B \in C$, $A \cap B \in C$.

(iii) **Complement:** X is said to be closed under the complement, if for any set $A \in C$, $\bar{A} \in C$.

(iv) **Finite Union and Countable Union:** C is said to be closed under the finite union if for any sets $A_1, A_2, \dots, A_n \in C$, $\cup_{i=1}^{n} A_i \in C$. Further if $n \to \infty$ and if the condition $\cup_{i=1}^{n} A_i \in C$ is satisfied then, C is said to closed under countable unions.

(v) **Finite Intersection and Countable Intersection:** C is said to be closed under the finite intersection if for any sets $A_1, A_2, \dots, A_n \in C$, $\cap_{i=1}^{n} A_i \in C$. Further if $n \to \infty$ and if the condition $\cap_{i=1}^{n} A_i \in C$ is satisfied then, C is said to closed under countable unions.

1.5.1 Field

The class of subsets $C$of non empty set Ω is called a field on Ω, if the following conditions satisfying:

(i) $\phi \in \mathrm{C}, \Omega \in \mathrm{C}$.

(ii) For any set $A \in C$, $\bar{A} \in C$ i.e., it is closed under complement.

(iii) For any sets $A_1, A_2, \dots, A_n \in C$, $\cup_{i=1}^{n} A_i \in C$ i.e., it is closed under finite unions.

1.5.2 $\sigma -$ Field or $\sigma -$Algebra

A class of subsets C of a non-empty set Ω, then C is called a σ-field on Ω, if the following conditions satisfies:

(i) $\phi \in \mathrm{C},\ \Omega \in C$

(ii) For any set $A \in C$, then $\bar{A} \in C$ i.e., it is closed under complement.

(iii) For any sets $A_1, A_2, \dots A_n \dots, \in C$, then $\cup_{i=1}^{\infty} A_i \in C$ i.e., it is closed under infinite countable unions.

Remark 1. Any finite/countable/arbitrary intersection of σ- fields is also a σ field.

Remark 2. The smallest σ- field is $C_s = \{\phi, \Omega\}$ and the largest σ- field is $C_l = P(\Omega)$, the set of all subset of C.

Remark 3. The smallest σ-field containing C is called a σ- field generated by C or minimal σ-field containing C.

Remark 4. Every σ- field is a field.

Proof. The conditions (i) and (ii) s are same in field and σ- field.

For (iii), if $A_1 = A$, $A_2 = B$, $A_n = \phi$, for $n \geq 3$, then each $A_n \in \mathrm{C}$, and C is a σ- field. Therefore, $\cup_{n=1}^{\infty} \in C$,

then $A_1 \cup A_2 \cup \dots U A_n \in C$

$\Rightarrow A \cup B \cup \phi \cup \dots \cup \phi \in C$

$\Rightarrow A \cup B \in C$

Hence, every σ- field is a field.

1.5.3 Borel σ- Field

Let $\Omega = \mathcal{R}$, a real line and C is a collection of open intervals for real line, and is denoted as $C = \{(a, b, a < b, a, b \in \mathcal{R}\}$, then the smallest σ- field on $\mathcal{R}$ containing C is called a Borel σ-Field on real line $\mathcal{R}$, and it is denoted by $\mathcal{B}$. The sets in $\mathcal{B}$ are called Borel sets.

Borel σ-Field on $\mathcal{R}^n$:

Borel σ- field $\boldsymbol{\mathcal{B}_n}$ of the subsets of $\mathcal{R}^n$, the n-dimensional Euclidean space with points $(x_1, x_2, \dots x_n)$ is the smallest σ- field containing n-dimensional open rectangle

$B = \{(x_1, x_2, \dots x_n) | a_i < x_i < b_i; i = 1,2, \dots, n\}$

where $a_i, b_i \in \mathcal{R}, i = 1,2,\dots,n$ and a_i's and b_i's are arbitrary.

1.6 Measurable Space and Measurable Sets

The pair (Ω, C) is called a measurable, where Ω is any non empty set and C is σ- field on Ω. The sets in C are called measurable sets.

1.7 Measure

Let the pair (Ω, C) be a measurable space. A measure on C is a function μ defined as $\mu: C \to \mathcal{R}^* = \mathcal{R} \cup \{\pm\infty\}$ such that

(i) $\mu(\phi) = 0$.

(ii) $\mu(A) \geq 0 \;\forall\, A \in C$ (μ is nonnegative).

(iii) If $A_1, A_2, \dots, A_n \in C$ is a collection of pairwise disjoint sets i.e., $A_n \cap A_m = \phi$ for $n \neq m$, then $\mu(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$ (μ is countable additivity or σ-additive).

A triplet $(\Omega, \mathrm{C}, \mu)$ is called a measure space. If μ assumes only the values a and ∞, then, the measure space $(\Omega, \mathrm{C}, \mu)$ is called degenerate.

Remark 1. If addition to (i), (ii) and (iii), μ also satisfies that $\mu(C) = 1$, then μ is called a **probability measure** on measurable space (Ω, C) and it is denoted by P. A triplet (Ω, C, P) is called a probability space.

Example: Let Ω is non empty set and $C = \{\Omega, \phi)\}$, also, $\mu_1(\phi) = 0$, $\mu(\Omega) = 1$ and $\mu_2(\phi) = 0, \mu_2(\Omega) = \infty$, check whether $\mu_1$and μ_2 are measure or probability measure?

Solution: For μ_1

(i) $\mu_1(\phi) = 0$

(ii) $\mu_1(\Omega) = 1 > 0$, so, $\mu_1(A) \geq 0$ for each $A \in C$.

(iii) If $A_n, n = 1,2,\dots$, $A_n \in C$ is the collection of disjoint sets then, $\mu_1(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu_1(A_n)$.

Let A_1 be Ω and $A_2, A_3, \dots$, be ϕ, then

$$\mu_1(A_1 \cup A_2 \cup \dots A_n \cup \dots) = \mu_1(\Omega \cup \phi \cup \dots \cup \phi \dots)$$

$$= \mu_1(\Omega) = 1$$

Also, $\sum_{n=1}^{\infty} \mu_1(A_1) + \mu_1(A_2) + \cdots + \mu_1(A_n) + \cdots .. = \mu_1(\Omega) + \mu_1(\phi) + \cdots + \cdots$

$$= 1 + 0 + 0 \ldots + 0 + \cdots = 1$$ Hence,

$\mu_1(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu_1(A_n)$.

(iv) $\mu_1(\Omega) = 1$

Hence, μ_1 is the probability measure on (Ω, C).

For μ_2,

(i) $\mu_2(\phi) = 0$

(ii) $\mu_2(\Omega) = 1 > 0$, so, $\mu_2(A) \geq 0$ for each $A \in C$.

(iii) If $A_n;\ n = 1,2, \ldots,\ A_n \in C$ is the collection of disjoint sets then, $\mu_2(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu_2(A_n)$.

Let A_1 be Ω and $A_2, A_3, \ldots$, be ϕ, then

$$\mu_2(A_1 \cup A_2 \cup \ldots A_n \cup \ldots) = \mu_2(\Omega \cup \phi \cup \ldots \cup \phi \ldots) = \infty$$

Also, $\sum_{n=1}^{\infty} \mu_2(A_1) + \mu_2(A_2) + \cdots + \mu_2(A_n) + \cdots .. = \mu_2(\Omega) + \mu_2(\phi) + \cdots + \cdots$

$$= \infty + 0 + 0 \ldots + 0 + \cdots = \infty$$

Hence, $\mu_2(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu_2(A_n)$.

But it does not follow the (iv) property i.e., $\mu_2(\Omega) \neq 1$, therefore, μ_2 is the measure only but not probability measure.

1.7.1 Properties of Measure

Properties 1: Let (Ω, C, μ) be measure space. Then μ is finitely additive, i.e., if $A_1, A_2, \ldots, A_n$ is a sequence of disjoint sets then, $\mu(\cup_{m=1}^{n} A_m) = \sum_{m=1}^{n} \mu(A_m)$.

Proof: Given that μ is ameasure on $\Omega, C)$, μ is countably additive that is there is a sequence $A_1, A_2, \ldots, A_n, A_{n+1}, \ldots$ of disjoint set such that

$$\mu(A_1, A_2, \ldots, A_n, A_{n+1}, \ldots) = \sum_{n=1}^{\infty} \mu(A_n) \quad (1)$$

Now, choose $A_{n+1} = A_{n+2} = \cdots = \phi$ i.e., $A_m = \phi, m > n$.

Hence from eq. (1),

$$\mu(\cup_{m=1}^{n} A_m) = \sum_{m=1}^{n} \mu(A_m)$$

So, μ is finite additivity.

Properties 2: Let $A, B \in C$, such that $A \subseteq B$, the $\mu(A) \leq \mu(B)$, it is called monotonicity property.

Proof. Define $B = A \cup (B - A)$

Taking μ of both sides, then we have

$$\mu(B) = \mu\big(A \cup (B - A)\big)$$
$$= \mu(A) + \mu(B - A),$$

because $A\ and\ B - A,$ are disjoint sets and μ is finite additive.

Then, $\mu(B) \geq \mu(A)$ as $\mu(B - A) \geq 0,$ by definition

Hence $\mu(A) \leq \mu(B)$.

Note: If μ is a probability measure then $\mu(B - A) = \mu(B) - \mu(A)$, because finite terms can be subtracted but infinite cannot.

Properties 3: Let $(\Omega, \mathrm{C}, \mu)$ ne measure space. For any sets $A, B \in C$, then $\mu(A \cup B) \leq \mu(A) + \mu(B)$ and if A and B are disjoint, then $\mu(A \cup B) = \mu(A) + \mu(B)$.

Proof. Define

$$A \cup B = A \cup (B - A)$$

Take μ on both sides, we have

$$\mu(A \cup B) = \mu\big(A \cup (B - A)\big)$$

Now, A and $B - A$ are the disjoint sets and μ is finitely additive, then

$$\mu(A \cup B) = \mu(A) + \mu(B - A) \quad (2)$$

Since, $B - A \subseteq B$

Take μ on both sides and using Property 2, we have

$$\mu(A - B) \leq \mu(B) \quad (3)$$

From (2) and (3), we have

$$\mu(A \cup B) \leq \mu(A) + \mu(B)$$

Also, if A and B are disjoint, then

$$\mu(A \cup B) = \mu(A) + \mu(B)$$

Property 4: For $A_1, A_2, \dots, A_n \in C$, then $\mu(\cup_{m=1}^{n} A_m) \leq \sum_{m=1}^{n} \mu(A_m)$ for any n.

Proof: Now consider, for $n = 1$,

$\mu(A_1) = \mu(A_1)$, which is true.

For $n = 2$, we have

$$\mu\left(A_1 \cup_{A_2}\right) \leq \mu(A_1) + \mu(A_2) \qquad \text{(Form Property 3)}$$

Now, for $n = k$, suppose it is true that

$$\mu\left(\cup_{m=1}^{k} A_m\right) \leq \sum_{m=1}^{k} \mu(A_m)$$

For $n = k + 1$, suppose $A_{k+1} \in C$, we have

$$\mu\left(\cup_{m=1}^{k+1} A_m\right) \leq \mu\left(\cup_{m=1}^{k} A_m \cup A_{k+1}\right)$$

Then by Property (3), we have

$$\mu\left(\cup_{m=1}^{k+1} A_m\right) \leq \mu\left(\cup_{m=1}^{k} A_m\right) + \mu(A_{k+1})$$

Hence, by the property of finite additive

$$\mu\left(\cup_{m=1}^{k+1} A_m\right) \leq \mu\left(\cup_{m=1}^{k+1} A_m\right)$$

$$\mu\left(\cup_{m=1}^{k+1} A_m\right) \leq \mu\left(\cup_{m=1}^{k} A_m\right) + \mu(A_{k+1})$$

Hence proved by the principal of mathematical induction for positive integer values of n.

Property 5: Let (Ω, C, μ) be measure space, $A, B \in C$, then $\mu(A \cup B) + \mu(A \cap B) = \mu(A) + \mu(B)$.

Proof: Define

$$A \cup B = A \cup (B - A)$$

Taking μ measure on both sides, we have

$$\mu(A \cup B) = \mu\big(A \cup (B - A)\big)$$

$$\Rightarrow \ \mu(A \cup B) = \mu(A) + \mu(B - A) \qquad (4)$$

as A and $B - A$ sets are disjoint sets and μ is finitely additive.

We can define set B as

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

Taking μ measure on both sides, we have

$$\mu(B) = \mu(A \cap B) + \mu(B - A) \qquad (5)$$

as $A \cap B)$ and $(B - A)$ are disjoint sets.

From (4) and (5), we have

$$\mu(A \cup B) + \mu(A \cap B) = \mu(A) + \mu(B)$$

Hence proved.

Definition: Let $\Omega, C, \mu)$ be a measure space, then

(i) **Finite Measure:** A measure μ is said to be a finite measure if $\mu(\Omega) < \infty$.
Aliter: A set $A \in C$ is said to be a set of finite measure if $\mu(A) < \infty$.

(ii) ***σ*- Finite Measure:** The measure μ is called a σ- finite measure if there is a sequence of sets $A_n \in \Omega$, such that $\Omega = \cup_{n=1}^{\infty} (A_n)$ and $\mu(A_n) < \infty, \ \ \forall\, n.$
Aliter: A set $A \in C$ is of σ- finite measure if A is a union of a number of measurable sets and such that in the union has a finite measure that is for $A = \cup_{i=1}^{\infty} A_1;\ \mu(A_i) < \infty \ \forall\, i.$

Remarks

1. A counting measure is not a finite measure.
2. If μ is a σ- finite, then every measurable set is of σ- finite measure, because μ is a σ- finite, hence for each $\mu(A_n) < \infty \ \forall\, n.$
3. ***μ* Null Set:** A set $A \in C$ is null if $A = \phi$, then $\mu(A) = 0$ is called μ null set.

4. **Complete Measure:** A measure space (Ω, C, μ) or a measure μ is called complete measure if C contains all μ null subsets i.e., if $B \in C$, $\mu(B) = 0$ and $A \subseteq B$, then $A \in C$ or i.e., every subset if a measurable μ- null set should be measurable.
5. Every measure is not a complete measure but it can be completed.

1.7.2 Extension of Measures

Let $P(\Omega)$ be the class of all subsets of Ω and let C, C' be two subclasses of $P(\Omega)$. Let ϕ, ϕ' also be two set functions defined on C, C' respectively and taking values in $\mathcal{R}$. The we call φ' is an *extension* of φ and φ is a restriction of φ', if $C \subset C'$ and $\varphi = \varphi'$ on C.

Definition: Let (Ω, C, μ) be a measure space. Let $A \in C$, we can again let E and F equal to A, then a measure μ' is called an extension of measure μ if,

(i) $\bar{\mu}(\phi) \geq 0$

(ii) $\mu'(\cup_n A_n) = \mu(\cup_n En) = \sum_n\big(\mu(En)\big) = \sum_n \mu'(A_n)$

where $\{A_n\}$ be a sequence of disjoints sets in C and for each$n$ again choose sets E_n and F_n in C such that $E_n \subseteq A_n \subseteq F_n$ and $\mu(F_n - E_n) = 0$.

1.7.3 Extension Theorem

Theorem: *Any measure* μ′ *on a semiring S is uniquely extended to a measure* μ *on the generated ring R*(*S*). *If the initial measure was* σ*-additive, then the extension is* σ*-additive as well.*

Proof: Let (Ω, C, μ)be a measure space. If an extension μ' of μ exists, then it shall satisfy

$$\mu(A) = \sum_{m=1}^{n} \mu'(A)$$

where $A_m \in C$.

We need to show two statements for this definition.

1. Consistency, i.e., independence of the value from a presentation of $A_m \in R(S)$ as $C = \cup_{m=1}^{n} A_m$, where $A_m \in C$.

For two different presentation $C = \cup_{l=1}^{n} A_l$ and $C = \cup_{m=1}^{n} B_m$.

Define $E_{lm} = A_l \cap B_m$ which will be pair-wise disjoint.

By the additivity of μ', we have

$$\mu'(A_l) = \sum_m \mu'(E_{lm}) \text{ and } \mu'(B_m) = \sum_l \mu'(E_{lm})$$

Then,

$$\sum_l \mu'(A_l) = \sum_l \sum_m \mu'\,(E_{lm}) = \sum_m \sum_j \mu'(E_{lm}) = \sum_m \mu'(B_m)$$

2. ***Additivity:*** For $C = \cup_{m=1}^{n} A_m$, where $A_m \in R(S)$.

We can present $A_m = \cup_{l=1}^{n(m)} E_{lm}$, $E_{lm} \in S$.

Thus $C = \cup_{m=1}^{n} \cup_{l=1}^{n(m)} C_{lm}$ and

$$\mu(A) = \sum_{m=1}^{n} \sum_{l=1}^{n(m)} \mu'(E_{lm}) = \sum_{m=1}^{n} \mu(A_m)$$

Finally, show the σ-additivity.

For a set $A = \cup_{m=1}^{\infty} A_m$, where A and $A_m \in R(S)$, find presentations $A = \cup_{l=1}^{n} B_l$, $B_l \in$ and $A_m = \cup_{v=1}^{u(m)} B_{vm}$, $B_{vm} \in S$.

Define $E_{lmv} = B_l \cap B_{lm} \in S$, then $B_j = \cup_{m=1}^{\infty} \cup_{v=1}^{u(m)} E_{lmv}$ and $A_m = \cup_{l=1}^{n} \cup_{(v=1)}^{u(m)} E_{lmv}$

Then, from σ- additivity of μ'

$$\mu(A) = \sum_{l=1}^{n} \mu'(B_l) = \sum_{l=1}^{n} \sum_{m=1}^{\infty} \sum_{v=1}^{u(m)} \mu'(E_{lmv}) = \sum_{m=1}^{\infty} \sum_{l=1}^{n} \sum_{v=1}^{u(m)} \mu'(E_{lmv}) = \sum_{m=1}^{\infty} \mu(A_m)$$

where we changed the summation order in series with non-negative terms.

1.8 Outer Measure

Let Ω be non empty set, μ^* be an extended real valued function defined on $P(\Omega)$, the μ^* is called an outer measure if

(i) $\mu^*(\phi) = 0,\ \ \mu^*(A) \geq 0$ for $A \in P(\Omega)$

(ii) For, $A, B \in P(\Omega)$, $A \subseteq B$, then $\mu^*(A) \leq \mu^*(B)$ i.e., μ^* has the monotonicity property.

(iii) If $< A_n >$ is $(P\Omega)$, then $\mu^*(\cup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \mu^*(A_n)$ i.e., μ^* is a countably sub-additive.

Remark: Every measure is an outer measure but an outer measure may or may not be measure.

$\boldsymbol{\mu}^*$ Measurability: Let μ^* ne outer measure on $P(\Omega)$ and, also, let $E \subseteq \Omega$, $E^c \subseteq \Omega$, then set E is said to be μ^* measurable, if $A \subseteq \Omega$, then $\mu^*(A) = \mu^*(A \cap E) + \mu^*(A \cap E^c)$. This is also called Caratheodory Measurability or an extension of measure.

1.9 Measurable Function

Let (Ω, C) be a measurable space and let f be a finite real valued function of Ω, then f is said to be a measurable function if for each real α such that

$\{x \in \Omega | f(x) > \alpha\} \in C$ or

$\{x \in \Omega | f(x) \geq \alpha\} \in C$ or

$\{x \in \Omega | f(x) < \alpha\} \in C$ or

$\{x \in \Omega | f(x) \leq \alpha\} \in C$

Note: All these sets are equivalent set.

Indicator Function: Let Ω be any non empty set, and also let $A \subseteq \Omega$, then indicator function of A is denoted by I_A and defined as

$$I_A = \begin{cases} 1\,, & \Omega \in A \\ 0\,, & \Omega \notin A \end{cases}$$

Simple Function: Let f be a real valued defined on Ω, then f is called simple function if it takes only finite many distinct values.

Note 1: Any function f is a simple function if it can be expressed as a linear combination of indictor functions of disjoint sets.

Proof: Let f be a real valued function.

Let $\alpha_1, \alpha_2, \dots \alpha_n$ be n distinct values taken by f.

Defined, $A_i = \{x \in \mathcal{R} | f(x) = \alpha_i\}, \quad i = 1, ,2, \dots, n$ and $A_i's$ are disjoint sets (also $\alpha_i's$ are distinct), therefore,

$A_i \cap A_j = \phi$ for $i \neq j$ and $\cup_{i=1}^{n} A_i = \mathcal{R}$

$$T_{A_i}(x) = \begin{cases} 1 \ , & x \in A_i \\ 0 \ , & x \notin A_i \end{cases}$$

i.e., $f(x) = \alpha_i$

Hence, $f(x) = \alpha_1 I_{A_1}(x) + \alpha_2 I_{A_2}(x) + \cdots + \alpha_n I_{A_n}(x)$, because if $f(x) = \alpha_1$, then $I_{A_1}(x) = 1$ all other indicator functions are zero.

Thereby, $f(x) = \sum_{i=1}^{n} \alpha_i \, I_{A_i}(x)$

Hence proved that any function of it can be expressed as a finite linear combination of $I_A(x)$ of disjoint sets.

Note 2: The sum, difference, and product of two simple functions is again simple function. The limits of simple function if they exist are measurable.

Note 3: If f is C measurable function, then $A_i = \{x_i \in \Omega | \ \mathrm{f(x)} = \alpha_\mathrm{i}\} \in C$ and $A_i = \{x \in \Omega | f(x) \geq \alpha_i\} - \{x \in \Omega | f(x) > \alpha\}$

Elementary Function: A function taking infinite number of distinct values is called an elementary function and defined as

$$f(x) = \sum_{i=1}^{\infty} \alpha_i \, I_{A_i}(x)$$

Step Function: A measurable function f with finite range is called *step function*.

1.10 Integrals

Let (Ω, C, μ) be s measure space and let $E \in C$ be measurable set and for a measurable function f on E, the integral of f is given by

$$\mu(f) = \int_E f d\mu = \int_E f(x)\mu(x)$$

When $(\Omega, \mathrm{C}) = (\mathcal{R}, \mathcal{B})$ (that is real line and borel σ- field) and μ is measure, then the integral of f is given by

$$\mu(f) = \int_{\mathcal{R}} f(x) \, dx = \sum_{i=1}^{n} \alpha_i \mu(A_i)$$

where $A_i \in C \; \forall \; i$.

Remarks 1: Let f and g be two measurable functions on (Ω, C, μ) and let $E \in C$, then the following holds:

(i) If $f \leq g$ on E, then $\int_E f \; d\mu \leq \int_E g \; d\mu$

(ii) If $A \subset B$ and $f \geq 0$, then $\int_A f \; d\mu \leq \int_B f d\mu$

(iii) If c is a constant, then $\int_E c \; f \; d\mu = c \int_E f \; d\mu$

(iv) $f \cong 0$ on E, then $\int_E f \; d\mu = 0$ even if $\mu(E) = \infty$

(v) If $\mu(E) = 0$, then $\int_E f \; d\mu = 0$if $f(x) = \infty$ on E.

(vi) If $f \geq 0$, then $\int_E f \; d\mu = \int_\Omega g f_E f d\mu$

2. Suppose f and g are two simple functions, then

$\int_\Omega (f + g) d\mu = \int_\Omega f \; d\mu + \int_\Omega g \; d\mu$

1.11 Non-Negative Integrable Function

Let (Ω, C, μ) be a measure space and f be a non-negative extended real valued measurable function on Ω, then the integral of f is

$$\mu(f) = \int_\Omega f \; d\mu = sup \int_\Omega \tau \; d\mu$$

Is known as non-negative integrable function, where τ ranges over all the simple functions τ for which $0 \leq \tau \leq f$ on Ω.

1.12 Integrable Measurable Function

Let (Ω, C, μ) be a measure space. A measurable function f on Ω is integrable over Ω with respect to μ if $|f|$ is integrable over Ω. For such function, we define the integral f over Ω with respect to μ as

$\int_\Omega f \; d\mu = \int_\Omega f^+ \; d\mu - \int_\Omega f^- \; d\mu$

For a measurable subset E on Ω, f is integrable over E if f_{Ω_E} is integrable over $\Omega$and we define the integral of f over E as

$\int_E f \; d\mu = \int_\Omega f_{\Omega_E} d\mu$

Remark: Let (Ω, C, μ) be a measure space and let f and g be two integrable function over Ω, then

(i) **Linearity:** For $a, b \in \mathcal{R}$, $af + bg$ is integrable over Ω and

$$\int_{\Omega} (af + bg) d\mu = \int_{\Omega} f \, d\mu + b \int_{\Omega} g \, d\mu$$

(ii) **Monotonicity:** If $f \leq g$, then

`$\int_{\Omega} f \, d\mu \leq \int_{\Omega} g \, d\mu$

(iii) **Additivity over Domains:** If A and B are disjoint measurable sets, then

$\int_{\mathrm{A \cup B}} f \, d\mu = \int_{\mathrm{A}} f \, d\mu + \int_{\mathrm{B}} f \, d\mu$

Important: The set of integrable functions is a subset of the set of measurable functions. We cannot define the integral of a non-negative function f unless f is measurable, and we cannot define the integral of a measurable function f (possibly negative) unless f is integrable.

1.13 Self-Assessment Exercises

1 Let the class A consist of the single set A and the class B consist of the single set B. What are $A \cup B$ and $A \cap B$?

2 What are the rings, fields, σ-rings and σ-fields generated by the following classes of sets?

(a) $\epsilon = \{E\}$, the class consisting of one fixed set E only

(b) ϵ is the class of all subsets of a fixed set E.

(c) ϵ is the class of all sets containing exactly two points.

3 Let X be an uncountably infinite set and ϵ_1 the class of sets which are either countable or have countable complements. Is ϵ_1 a ring? A field? A σ-ring? Let ϵ_2 be the class of all countable subsets of X. Is ϵ_2 a ring? A field? A σ- ring?

4 Let ϵ be any nonempty class of sets and let P be the class of all possible finite intersections of the form $E_1 \cap E_2 \cap \ldots \cap E_n$; $n = 1, 2, \ldots$, where $E1 \in \epsilon$ and for each $j = 2, \ldots, n$, either $E_j \in \epsilon \; or \; E_j^c \in \epsilon$. Then show that P is a semiring, $P \supset \epsilon$, and $R(P) = R(\epsilon)$.

5 Let μ be a measure defined on a ring R. Show that the class of sets $E \in R$ with μ(E) finite, forms a ring.

6 Let ϵ be a class of sets and μ be a measure on $R(\epsilon)$ such that $\mu(\epsilon) < \infty$ for all $E \in \epsilon$. Show that μ is a finite measure on $R(\epsilon)$.

7 Let X be any space with two or more points. Write $\mu(\phi) = 0$, and $\mu(E) = 1$ for $E \neq \phi$. Is μ an outer measure, a measure?

8 If μ^* is an outer measure and E, F are two sets, E being μ^*-measurable, show that

$$\mu^*(E) + \mu^*(F) = \mu^*(E \cup F) + \mu^*(E \cap F).$$

9 Let x_0 be a fixed point of space X. Is $\mu^*(E) = \chi_E(x_0)$ an outer measure?

10 If $|f|$ is a measurable function on (Ω, C), is f measurable?

11 Give a measurable space (Ω, C) and finite measures μ and ν on it that satisfy $\mu(\Omega) = \nu(\Omega)$ but are such that $\{A \in C : \mu(A) = \nu(A)\}$ is not a σ-algebra.

12 Is it always the case that if all the A_n are in ϵ then $\mu(\cup_n A_n) \leq \sum_n \mu(A_n)$.

13 By answering these questions, learners will be able to gauge their understanding of the unit's key concepts and their ability to apply them in various contexts.

1.14 Summary

In this unit on measure in probability and statistics, we delved into topics, including rings and sigma-rings, fields, and sigma-fields, which play a pivotal role in defining measurable spaces and measurable sets. The central theme of measure theory revolves around the notion of a measure on R^n, and we explored the key properties of measures and the concept of outer measure. The extension of measures is a crucial concept, and the Extension Theorem serves as a fundamental tool in this context. The concepts of simple functions that provide a foundation for integration were explored, as well as the non-negative integrable and integrable measurable functions, which are integral components of the broader study of Lebesgue integration and analysis. These topics collectively form the cornerstone of measure theory, offering powerful tools for mathematical analysis and understanding complex sets and functions.

1.15 References

1. Billingsley, P. (1995). Probability and Measure. John Wiley & Sons.

2. Durrett, R. (2019). Probability: Theory and Examples. Cambridge University Press.

3. Cohn, D.L. (2013). Measure Theory. Springer New York

4. Klenke, A. (2014). Probability Theory: A Comprehensive Course. Springer.

5. Loève, M. (1977). Probability Theory I. Springer-Verlag.

6. Shiryaev, A. N. (1996). Probability. Springer.

7. Roussasan, G.G. (2014). Introduction to Measure-Theoretic Probability. Academic Press.

1.16 Further Reading

- "Probability and Measure" by Patrick Billingsley, John Wiley & Sons.
- "A Probability Path" by Sidney I. Resnick, Birkhäuser.
- "Convergence of Probability Measures" by Patrick Billingsley, John Wiley & Sons.
- "Weak Convergence and Empirical Processes: With Applications to Statistics" by A. W. van der Vaart and Jon A. Wellner, Springer.
- "An Introduction to Probability Theory and Its Applications (Vol. 1)" by William Feller, John Wiley & Sons.
- "Probability: Theory and Examples" by Rick Durrett, Cambridge University Press.
- "Real Analysis and Probability" by R. M. Dudley, Cambridge University Press.
- "Foundations of Modern Probability" by Olav Kallenberg, Springer.
- “Introduction to Measure-Theoretic Probability” by Roussasan, G.G, Academic Press.
- "Probability with Martingales" by David Williams, Cambridge University Press.

UNIT: 2 CONVERGENCE

Structure

2.1 Introduction

In this chapter we deal with some functions, namely measurable function, Lebesgue measurable functions, equivalent function, and characteristics function which are naturally linked with the notion of measurable sets. One can ask why we need such a function, the best answer to this question is that we ultimately intend to define an integration process modeled on Riemann integration which should be stronger than Riemann integration. Since functions are the objects which we integrate, we need to define a special class of functions which we will like to integrate.

There is a concept called convergence in measure for sequences of measurable functions $f_n \to f$ that is especially useful in the theory of probability. In that context, it is useful to know that the probability of a random variable f_n differing from the random variable f by more than ϵ

is very small. Firstly, we will dive deep in the measure space and measurable function. A very important function i.e., Lebesgue function will also be discussed in this chapter. In this chapter we also define Lebesgue measurable functions and prove its basic properties. We also define a special type of measurable function called simple function and mainly show that any measurable function can be expressed as the limit of a sequence of simple functions.

We will consider sequences of real functions and show that with respect to Lebesgue measure, the pointwise convergence can be generalized into the notion of convergence almost everywhere which means usual pointwise convergence. We also define the convergence in measure for sequences of measurable functions. We will dive deep into understanding how these sequences stabilize or converge. In doing so, we will encounter different 'flavors' of convergence, each with its distinct characteristics and implications.

2.2 Objectives

By the end of this unit, the learner should be able to:

- Understand the basic concept of convergence.
- Concept of the measurable function and Lebesgue measurable function and also distinguish between them.
- Describe the concept the characteristics function.
- Identify various types of convergence in measure.
- Understand the combination of measurable function.

2.3 Measure Space and Measurable Function

Measure spaces are the fundamental objects of measure theory, the branch of mathematics that studies generalized concepts of volume. It contains the underlying set, the subset of this set that can be used for measurement (σ-algebra), and the method used for measurement (measurement). One important example of measurement spaces is the probability space.

Most of the theory of measurable functions and integration does not depend on the specific features of the measure space on which the functions are defined. Measurable functions in measure theory are analogous to continuous functions in topology. A continuous function pulls back open sets to open sets, while a measurable function pulls back measurable sets to measurable sets.

2.3.1 Measure Space

Let Ω be a non empty set, C be a σ-filed on Ω, and μ is a measure, then the triplet $(\Omega, \mathrm{C}, \mu)$ is called a measure space.

2.3.2 Measurable Function

Let (Ω, C) be a measurable space and let f be a finite real valued function of Ω, then f is said to be a measurable function if for each real α such that

$\{x \in \Omega | f(x) > \alpha\} \in C$ or

$\{x \in \Omega | f(x) \geq \alpha\} \in C$ or

$\{x \in \Omega | f(x) < \alpha\} \in C$ or

$\{x \in \Omega | f(x) \leq \alpha\} \in C$

2.3.3 Lebesgue Measurable Function

We change (Ω, C) to $(\mathcal{R}, \mathcal{B})$. Let $(\mathcal{R}, \mathcal{B})$ be Lebesgue measurable function and let f be a finite real valued function of $\mathcal{R}$, then f is said to be a measurable function if, for each real α

$\{x \in \mathcal{R} | f(x) > \alpha\} \in \mathcal{B}$ or

$\{x \in \mathcal{R} | f(x) \geq \alpha\} \in \mathcal{B}$ or

$\{x \in \mathcal{R} | f(x) < \alpha\} \in \mathcal{B}$ or

$\{x \in \mathcal{R} | f(x) \leq \alpha\} \in \mathcal{B}$

where $\mathcal{B}$ is the Borel σ- field of Lebesgue measurable space.

Example 1: Let $\Omega = \{1,2,3,4\}$ and let σ- field on Ω is $C = \{\phi, \Omega\}$. Define $f(x) = 1 \ \forall \ x \in \Omega$. Check the measurability of function f.

Solution: For a real α,

$$\{x \in \Omega | f(x) > \alpha\} = \begin{cases} \Omega, & \alpha < 1 \\ \phi, & \alpha \geq 1 \end{cases}$$

Since $C = \{\phi, \Omega\}$, then $(\Omega, \phi) \in C$, hence $\{x \in \Omega | f(x) > \alpha\} \in C$, then f is the measurable function.

Example 2: Let $f(x) = a,\ \Omega = \{1,2,3,4\}$ and $C = \{\phi, \Omega\}$. Check whether the function f is measurable?

Solution: For real α

$$\{x \in \Omega | f(x) > \alpha\} = \begin{cases} \Omega, & \alpha < a \\ \phi, & \alpha \geq a \end{cases}$$

Since $C = \{\phi, \Omega\}$, then $(\Omega, \phi) \in C$, hence $\{x \in \Omega | f(x) > \alpha\} \in C$, then f is the measurable function i.e., every constant function f is the measurable function.

Definition: - Let μ be a measure function on a σ- ring of subsets of Ω.

(i) **Completeness of Measure Function:** The measure function μ is said to be complete, if $B \in C$ such that, $\mu(B) = 0,\ A \subset B = A \in C$ i.e., if all the subsets of measure 0 are measurable in this case (Ω, C, μ) is called a *complete measure space.*

(ii) Any set $A \in C$ is said to have finite measure if $\mu(A) < \infty$.

(iii) The measure of any set $A \in C$ is said to be σ- finite, if there exist a sequence $\langle A_n \rangle \in R$ such that $n \ni N$ such that

a. $A \subset \bigcup_{n=1}^{\infty} A_n$

b. $\mu\,(A_n) < \infty \quad \forall\, n \in N$

(iv) The measure function n is called finite or σ-finite according as the measure of every set is finite or σ-finite.

(v) The measure function μ is called totally finite or totally σ- finite if,

a. $x \in C$ i.e., C is algebra sets.

b. $\mu(x)$ is finite or σ- finite.

2.3.4 Equivalent Function

Two functions f and g defined on the same set E are said to be equivalent If $\mu[E(f \neq g)] = 0$ i.e., if there exist $A, B, \subset E$ such that $E = A \cup B$, on A and $f = g$ on A, $\mu(B) = 0,\ \mu(A) \neq 0$.

2.3.5 Characteristic Functions

Let A be a subset of set E. The characteristic function K_A of A is defined as,

$$K_A(x) = \begin{cases} 1 & if\ x\ \in A \\ 0, & if\ x\ \in E - A. \end{cases}$$

The function $K_A(x)$ is measurable if and only if A is measurable.

Note: Existence of non-measurable set implies that the existence of non-measurable function.

2.4 Combination of Measurable Function

Theorem: Let (Ω, C) be a measurable space. The functions $f: \Omega \to R$ and $\mathrm{g}: \Omega \to R$ are two measurable functions, then the functions

(a) $f + c$ (b) cf (c) $f + g$ (d) $f - g$

(e) f^2 (f) fg (g) $|f|$ are measurable functions, where c is a real valued constant.

Proof: Let f be a measurable function defined over the measurable se t E.

(a) For any real α such that

$\{x \in \Omega | f(x) + c > \alpha\} = \{x \in \Omega | f(x) > \alpha - c\}$ (If c is any constant then, $(\alpha - c)$ is a real value)

$$= \{x \in \Omega | f(x) > \alpha'\} \in C$$

Therefore, f is the measurable function.

(b) For any real α such that

$\{x \in \Omega | f(x) + c > \alpha\} \in C$

Case (i): when $c > 0$,

$\{x \in \Omega | f(x) > \alpha/c\} \in C = \{x \in \Omega | f(x) > \alpha'\} \in C$ (because f is the measurable function with respect to C and α/c is also a real number)

Case (ii): When $c < 0$

$\{x \in \Omega | f(x) < \alpha/c\} \in C = \{x \in \Omega | f(x) < \alpha'\} \in C$ (because f is the measurable function with respect to C and α/c is also a real number)

Case (iii): When $c = 0$, then

$$h(x) = cf(x) = 0 \text{ , if } c = 0 | x \in \Omega$$

For any real α, $\{x \in \Omega |\ \mathrm{h(x)} > \alpha\} = \begin{cases} \Omega \ , \alpha < 0 \\ \phi \ , \ \alpha \geq 0 \end{cases} \ \in C$

Therefore, h is the measurable function with respect to C, hence f will be a measurable function.

(c) First, we prove that $A = \{x \in \Omega | f(x) > g(x)\} \in C$

where f and g are the measurable functions.

Let $x \in A, f(x) > g(x)$

$\Rightarrow$ there exist a rational number r such that

$$f(x) > r > g(x)$$

Let $\{r_1, r_2, \dots r_n\}$ be a sequence of all such rational numbers,

Therefore, $A = \cup_{n=1}^{\infty} \{x \in \Omega | \mathrm{f(x)} > \mathrm{r_n} > \mathrm{g(x)}\}$

$= \cup_{n=1}^{\infty} [\{x \in \Omega | \mathrm{f(x)} > \mathrm{r_n}\} \cap \{x \in \Omega | \mathrm{g(x)} < \mathrm{r_n}\}]$

$= \cup_{n=1}^{\infty} [\{x \in \Omega | \mathrm{x} > \mathrm{f}^{-1}(r_n)\} \cap \{x \in \Omega | \mathrm{x} < \mathrm{g}^{-1}(\mathrm{r_n})\}]$ (if f is measurable, then inverse images $\in C$)

$= \cup_{n=1}^{\infty} (A_n \cap B_n)$

$\Rightarrow A_n \in C$ ($\because f$ is a measurable function)

$B_n \in C$ ($\therefore g$ is a measurable function).

Since, C is a σ- field, hence intersection also $\in C$, thereby

$A_n \cap B_n \in C$

$\Rightarrow \cup_{n=1}^{\infty} (A_n \cap B_n) \in C$

$\Rightarrow \{x \in \Omega | f(x) + g(x) > \alpha\} = \{x \in \Omega | f(x) > \alpha - g(x)\}$

Let $\alpha - g(x) = h(x)$.

Since g is a measurable function,

$\Rightarrow -g$ is also a measurable function

($\because cg$ is measurable function for $c = -1$)

$\Rightarrow \alpha - g(x)$ is the measurable function

($\because c + g(x)$ is measurable function with the result $c = \alpha$)

Since, f is the measurable function and $\alpha - g(x)$ is also measurable function. Hence,

$$\{x \in \Omega | f(x) > \alpha - g(x)\} \in C$$

$$\therefore \{x \in \Omega | f(x) > g(x)\} \in C$$

$\Rightarrow f + g$ is the measurable function.

(d) The proof will follow by changing $h(x) = a + g(x)$ in (c).

(e) For any real α, consider

$$\{x \in \Omega | f^2(x) > \alpha\} = \begin{cases} \phi, & \alpha \leq 0 \\ x \in \Omega | -\sqrt{\alpha} < f(x) < \sqrt{\alpha}, & \alpha > 0 \end{cases} \in C$$

$\Rightarrow \left\{x \in \Omega | f(x) > -\sqrt{\alpha}\right\} \cap \left\{x \in \Omega | f(x) < \sqrt{\alpha}\right\} \in C$

$\Rightarrow A \cap B$

$\because A, B \in C$, and f is measurable function. $A \cap B \in C$, f is measurable function.

$\therefore -\sqrt{\alpha}$ and $\sqrt{\alpha}$ are the real numbers

Hence, $\{x \in \Omega | f^2(x) > \alpha\} \in C$

Therefore, $f^2(x)$ is a measurable function.

(f) Since,

$$4fg = (f + g)^2 - (f - g)^2$$

$$\Rightarrow fg = \frac{1}{4}[(f + g)^2 - (f - g)^2]$$

We know that $(f + g)$ and $(f - g)$ are measurable functions, and cf is also a measurable function, So

$$fg = \frac{1}{4}[(f + g)^2 - (f - g)^2] \in C$$

Hence, fg is a measurable function.

(g) For any real α, consider

$$\{x \in \Omega || f(x)| > \alpha\} = \begin{cases} \phi, & \alpha \leq 0 \\ x \in \Omega | -\alpha < f(x) < \alpha, & \alpha > 0 \end{cases} \in C$$

$$\Rightarrow \{x \in \Omega | f(x) > -\alpha\} \cap \{x \in \Omega | f(x) < \alpha\} \in C$$

$\Rightarrow A \cap B$

$\because A, B \in C$, and f is measurable function. $A \cap B \in C$, f is measurable function.

$\therefore -\alpha$ and α are the real numbers

$\Rightarrow \{x \in \Omega | |f(x)| > \alpha\} \in C$

Hence, $|f(x)|$ is a measurable function.

Theorem: Let f be a measurable function and $f \neq 0$, then $1/f$ also a measurable function.

Proof: For any real α,

$\{x \in \Omega | 1/f(x) < \alpha\}$

$$= \begin{cases} \{x \in \Omega | f(x) < 0\}, & \alpha = 0 \\ \{x \in \Omega | f(x) < 0\} \cup \{x \in \Omega | f(x) < 1/\alpha\}, & \alpha > 0 \\ \{x \in \Omega | 1/\alpha < f(x) < 0\}, & \alpha < 0 \end{cases}$$

Since f is measurable function and C is a σ- field ($\because$ closed under unions and intersections).

Hence all set $\in C$,

$\Rightarrow \{x \in \Omega | 1/f(x) < \alpha\} \in C$

Hence, $1/f$ is a measurable function.

Almost Everywhere Property: Suppose then that (Ω, C, μ) is a fixed measure space. Suppose that some property holds at all points of $A \in C$, where $\mu(A^c) = 0$, then this property is said to hold almost everywhere (abbreviated "a.e." or "a.e. (μ)"). For example, if f is a function on Ω, the statement $f \geq 0$ a.e. means that there is a set $A \in C, \mu(A^c) = 0$, such that $f(x) \geq 0$ for all $x \in A$.

Note that the set where $f(x) < 0$ is to be a subset of the set A^c. The precise set where the property does not hold is not necessarily measurable unless, of course, μ is a complete measure.

Thus, as defined above, to say that a property holds i.e., means that it holds at all points of A, where A is a measurable set with $\mu(A^c) = 0$. For an example, to say that a function f is defined a.e. on Ω means that f is defined for all $x \in A$, where $A \in C, \mu(A^c) = 0$. To say that two functions f and g are equal a.e. on Ω means that $f(x) = g(x)\ \forall\, x \in A\ (\in C)$, where $\mu(A^c) = 0$, and so on.

2.5 Convergence in Measure

Let $(\Omega, \mathrm{C}, \mu)$ ne a measure space. Consider a sequence $\{f_n\}$ of measurable functions defined on $E \in \Omega$ and taking values in $\mathcal{R}$. If f be a measurable function on E for which f and each f_n are finite a.e. on E. The sequence $\{f_n\}$ *converges in measure* on E to f provided for each $\eta > 0$

$$\lim_{n\to\infty} \mu(\{x \in E||f_n(x) - f(x)| > \eta\}) = 0$$

Note 1: When we say "$\{f_n\}$ converges in measure on E to f" it is implied that f and each f_n are measurable and finite a.e. on E.

Note 2: Assume E has finite measure. Let $\{f_n\}$ be a sequence of measurable functions on E that converges pointwise a.e. on E to f, and f and each f_n are finite a.e. on E, then $\{f_n\} \to f$ in measure on E.

2.6 Point wise Convergence

Let $(\Omega, \mathrm{C}, \mu)$ ne a measure space. Consider a sequence $\{f_n\}$ of measurable functions defined on $E \in \Omega$, if there exist a measurable function f on E such that

$$\lim_{n\to\infty} f_n(x) = f(x), \quad \forall\, x \in E$$

Then, we say that the sequence $\{f_n\}$ converge point wise to f on E.

Convergence Almost Every Where: Let $\{f_n\}$ be a sequence of measurable function defined on E, there exist a measurable function f over E and a set A such that

(i) $\mu(A) = 0$

(ii) $\lim_{n\to\infty} f_n(x) = f(x), \quad \forall\, x \in E - A.$

(i.e., f_n converges point wise to f and $E - A$),

then we say that the sequence f_n converges to f almost everywhere.

Uniform Convergence: Let A $\{f_n\}$ be a measurable function defined over a measurable set E is said to be converge uniformly almost everywhere to a measurable function f, if there exist $E \subset \Omega$ as $n \to \infty$ if and only if

$$\lim_{n\to\infty}\left(\underbrace{sup}_{x\in E}|f_n(x)-f(x)|\right)=0.$$

2.7 Self-Assessment Questions

1 What is the fundamental idea behind convergence in measure theory?

2 Define the term "measurable function".

3 Let X be the real line (R), and Ω the σ-field consisting of $X, \phi, (-\infty, 0], (0, \infty)$. What functions defined on X are Ω measurable?

4 Explain the role of a sequence of measurable functions in measure theory.

5 Distinguish between the almost everywhere convergence and uniform convergence of a sequence of measurable functions.

6 What is "point-wise convergence"?

7 When do we say a sequence of measurable function converges in measure?

8 How is the measurable function of a sequence different from that of a simple function?

9 Why is the concept of convergence in measure important?

10 Given a sequence of measurable functions where you know they converge almost surely to a function $f(,)$, what can you infer about their uniform convergence?

True or False Questions

11 By answering these questions, learners will be able to gauge their understanding of the unit's key concepts and their ability to apply them in various contexts.

2.8 Summary

This unit covered the fundamental concepts of measure theory, such as measure space, where measurable functions play a pivotal role. Among these functions, the Lebesgue measurable function plays a crucial foundation for analysis. We delved into equivalent and characteristic functions, both essential in representing and understanding measurable sets and their properties. Also, the concept of combinations of measurable functions that allow for the manipulation of data and the examination of intricate relationships is studied. The concept of convergence in measure, which provides a valuable tool for studying the behaviour of sequences of functions, with point-wise convergence offering insights into the individual behaviour of functions within these sequences, was studied. These topics provided a deeper understanding of the measurable functions and properties of Lebesgue integration and mathematical analysis.

2.9 References

- Billingsley, P. (1995). Probability and Measure. John Wiley & Sons.
- Grimmett, G., & Stirzaker, D. (2001). Probability and Random Processes. Oxford University Press.
- Klenke, A. (2014). Probability Theory: A Comprehensive Course. Springer.
- Loève, M. (1977). Probability Theory I. Springer-Verlag.
- Shiryaev, A. N. (1996). Probability. Springer.
- Cohn, D.L. (2013). Measure Theory. Springer New York
- Roussasan, G.G. (2014). Introduction to Measure-Theoretic Probability. Academic Press.
- Leadbetter, R., Cambanis, S. & Pipiras, V. (2014). A Basic Course in Measure and Probability Theory for Applications. Cambridge University Press.

2.10 Further Reading

- "Probability and Measure" by Patrick Billingsley, John Wiley & Sons.
- "A Probability Path" by Sidney I. Resnick, Birkhäuser.
- "Convergence of Probability Measures" by Patrick Billingsley, John Wiley & Sons.
- "An Introduction to Probability Theory and Its Applications (Vol. 1)" by William Feller, John Wiley & Sons.
- "Probability: Theory and Examples" by Rick Durrett, Cambridge University Press.
- "Real Analysis and Probability" by R. M. Dudley, Cambridge University Press.
- "Foundations of Modern Probability" by Olav Kallenberg, Springer.
- “Basic Course in Measure and Probability Theory for Applications” by
- "Probability with Martingales" by David Williams, Cambridge University Press. Leadbetter, R., Cambanis, S. & Pipiras, V. Cambridge University Press.

UNIT: 3 LEBESGUE MEASURE

Structure

3.1 Introduction

When we dive deeper into math, we will find out that the intuitive knowledge gradually becomes insufficient, e.g. when we encounter more tricky sets and begin to think about what can be integrated, the common sense of "area" is not so applicable anymore. Then "measure" comes into play. The Lebesgue measure, named after French mathematician Henri Lebesgue, is the standard way of assigning a measure to a geometric object. In this chapter we introduce the Lebesgue measure and integral as intuitive and understandable as possible.

The Riemann integral is the 'ordinary' integral encountered in everyday mathematics, and can be dealt with (in some situations) by the well-known methods of integration. Sometimes, however, this definition of the integral proves inadequate for applications elsewhere in mathematics. The generalization we look upon, the Riemann-Stieltjes integral, integrates a given function with respect to another. The benefit of the concept is that, among other things, it includes Riemann integrals and finite sums as special cases.

An important generalization is the Lebesgue–Stieltjes integral, which generalizes the Riemann–Stieltjes integral in a way analogous to how the Lebesgue integral generalizes the Riemann integral. The concept of Lebesgue dominated convergence and monotone convergence theorem are also discussed.

3.2 Objectives

By the end of this unit, the learner should be able to:

- Understand the basic concept of Lebesgue measure.
- Describe the concept of Borel measurable function.
- Understand the construction of the Lebesgue integral
- Distinguish between Lebesgue-Stielitijes integral and Riemann integral.
- Describe the concept of Lebesgue dominated convergence and monotone convergence.
- Identify various types of convergence for measurable function

3.3 Borel Measurable Function

Let $\mathcal{B}$ be the Borel field of subset $\mathcal{R}$. C be a σ- field of subset of Ω. If $X^{-1}(B) \in \mathcal{B}$ for all Borel sets $B \in \mathcal{B}$, then X is said to be a measurable function. If $\Omega = \mathcal{R}$, then X is said to be Borel measurable function.

3.4 Lebesgue-Stieltjes Measure

Consider a generalization of one-dimensional Lebesgue measure on $\mathcal{R}$. These measures are obtained from an increasing, right-continuous function $F : \mathcal{R} \to \mathcal{R}$ (i.e., $F(x+0) = F(x)$) and assign to a half-open interval $(a, b]$, the measure

$$\mu_F\ ((a,b]) = F(b) - F(a), \quad -\infty < a < b < \infty$$

The measure μ_F is called the Lebesgue-Stieltjes measure on $\mathcal{B}$ corresponding to the function F.

The use of half-open intervals is significant here because a Lebesgue-Stieltjes measure may assign nonzero measure to a single point. Thus, unlike Lebesgue measure, we need not have $\mu_F\ ([a,b]) = \mu_F\ ((a,b])$. Half-open intervals are also convenient because the complement of a half-open interval is a finite union of (possibly infinite) half-open intervals of the same type. Thus, the collection of finite unions of half-open intervals forms an algebra or field.

Remark 1: A nondecreasing (right-continuous) function F has at most countably many discontinuities. Equivalently the corresponding Lebesgue–Stieltjes measure μ_F has at most countably many atoms.

Proof: If $F : \mathcal{R} \to \mathcal{R}$ is monotonic increasing (i.e., $x_1 < x_2$ implies $F(x_1) < F(X_2)$, then the left-limit $F(x_-)$ and the right-limit $F(x_+)$ exist at every $x \in \mathcal{R}$ and $F(x_-) < F(x) < F(x_+)$. Hence F has at most countable many discontinuities, and there are jump discontinuities, i.e., $F(x_-) < F(x_+)$.

3.5 Lebesgue-Stieltjes Integral

Let (Ω, C, μ) be a measure space and f be the simple function of such that

$f(x) = \sum_{i=1}^{n} \alpha_i I_{A_i}(x)$,

where $A_i = \{x | f(x) = \alpha_i\}$, $i = 1,2, \dots, n$ and $A_i^{\prime}$s are pairwise disjoint.

Then, the function f is Lebesgue integral if $\mu(A_i) < \infty$ (finite) for every index i for which $\alpha_i^{\prime} s \neq 0$ and

$\int_x f(x)\, d\mu(x)$ or $\int f d\mu = \sum_{i=1}^{n} \alpha_i \mu(A_i)$

and

$$\int_E f\, d\mu = \sum_{i=1}^{n} \alpha_i \mu(A_i \cap E)\,;\ E \subseteq X$$

(exists or not), and call it the Lebesgue-Stieltjes integral of f with respect to μ on $(a,b]$.

3.6 Riemann Integral

Suppose that f is a function on $\mathcal{R}$ that is defined and bounded on the interval $I = [a, b]$ and $\mathcal{P} = \mathcal{P}[a, b]$ is the set of all partitions of $[a, b]$. Then the *upper Riemann integral* and the *lower Riemann integral* are defined by

$$\int_a^{\bar{b}} f(x)dx = \inf_{P\in\mathcal{P}} U(P, f) \qquad \text{and} \int_{\bar{a}}^{b} f(x)dx = \underset{P\in\mathcal{P}}{sup}\, U(P, f),$$

respectively. If $\int_a^{\bar{b}} f(x)dx = \int_{\bar{a}}^{b} f(x)dx$, then f is called Riemann integrable or just integrable on $I$and the common value of the integral is denoted by $\int_a^b f(x)dx$.

3.7 Riemann-Stieltjes Integral

Suppose that f is a function on $\mathcal{R}$ that is defined and bounded on the interval $I = [a, b]$ and $\mathcal{P} = \mathcal{P}[a, b]$ is the set of all partitions of $[a, b]$, and α is function that is defined and monotonically increasing on I. Then, the *upper Riemann-Stieltjes integral* and the *lower Riemann-Stieltjes integral* are defined by

$$\int_a^{\bar{b}} f(x)d\alpha(x) = \inf_{P\in\mathcal{P}} U(P, f, \alpha) \qquad \text{and} \int_{\bar{a}}^{b} f(x)d\alpha(x) = \underset{P\in\mathcal{P}}{sup}\, U(P, f, \alpha),$$

respectively. If $\int_a^{\bar{b}} f(x)d\alpha(x) = \int_{\bar{a}}^{b} f(x)d\alpha(x)$, then f is called Riemann integrable or just integrable on $I$and the common value of the integral is denoted by $\int_a^b f(x)\, d\alpha(x)$ or $\int_a^b f\, d\alpha$.

Theorem: If f is a function that is continuous on the interval $I = [a, b]$, then f is Riemann-Stieltjes integrable on $[a, b]$.

Proof: Let α be monotonically increasing on I and f be continuous on I. Suppose that $\epsilon > 0$ is given. Then there exists an $\eta > 0$ such that $[\alpha(b) - \alpha(a)]\eta < \epsilon$. By the Uniform Continuity theorem, f is uniformly continuous in $a, b]$ fromwhic it follows that there exists a $\delta > 0$ such that

$$[u, v \in I \wedge |u - v| < \delta \Rightarrow |f(u) - f(v)| < \epsilon] \quad \text{for all } u, v$$

Let $\mathcal{P} = \{x_0 = a, x_1, \dots, x_{n-1}, x_n = b\}$be a partition of $[a, b]$ for which $\mathcal{P} < \delta$, for each j, $j = 1,2, \dots, n$ set $M_j = \sup_{x_{j-1}\le x\le x_j} f(x)$ and $m_j = \inf_{x_{j-1}\le x\le x_j} f(x)$. Then $M_j - m_j \le \eta$ and

$$U(\mathcal{P},f,\alpha)-L(\mathcal{P},f,\alpha)=\sum_{j=1}^{n}\left(M_j-m_j\right)\triangle\alpha_j\leq\sum_{j=1}^{n}\triangle\alpha_j=\eta[\alpha(b)-\alpha(a)]<\epsilon$$

Since $\epsilon>0$ was arbitrary, we have that there exists $\mathcal{P}$ such that

$(\mathcal{P}\wedge U(\mathcal{P},f,\alpha)-L(\mathcal{P},f,\alpha)<\epsilon)$ for all ϵ and $\epsilon>0$.

In view of the integrability criterion, $f\in\mathcal{R}(\alpha)$. Because α was arbitrary, we conclude that f is Riemann-Stieltjes integrable (with respect to any monotonically increasing function on $[a,b]$).

3.8 Lebesgue Dominated Convergence Theorem

Statement: Suppose $f_n:R\rightarrow[-\infty,\infty]$ are (Lebesgue) measurable functions such that the pointwise $limit\, f(x)=\lim_{n\rightarrow\infty}f_n(x)$ exists. Assume there is an integrable $g:R\rightarrow[0,\infty]$ with $|f_n(x)|\leq g(x)$ for each $x\in R$. Then f is integrable as is f_n for each n, and

$$\lim_{n\rightarrow\infty}\int_{\mathcal{R}}f_nd\mu=\int_{\mathcal{R}}\lim_{n\rightarrow\infty}f_n\,d\mu=\int_{\mathcal{R}}f\,d\mu$$

Proof: Since $|f_n(x)|\leq g(x)$ and g is integrable $\int_R|f_n|d\mu\leq\int_R g\,d\mu<\infty$. So f_n is integrable. We know f is measurable (as a pointwise limit of measurable functions) and then,

similarly, $|f(x)|=\lim_{n\rightarrow\infty}|f_n(x)|\leq g(x)$ implies that f is integrable too.

The proof does not work properly if $g(x)=\infty$ for some x.

We know that $g(x)<\infty$ almost everywhere.

So, we can take $E=\{x\in R: g(x)=\infty\}$ and multiply g and each of the functions f_n and f by $1-\chi E$ to make sure all the functions have finite values. As we are changing them all only on the set E of measure 0, this change does not affect the integrals or the conclusions. We assume then all have finite values.

Let $h_n=g-f_n$, so that $h_n\geq 0$.

By Fatou's lemma

$$\lim_{n\rightarrow\infty}inf\int_R(g-f_n)d\mu\geq\int_R\lim_{n\rightarrow\infty}\inf(g-f_n)d\mu=\int_R(g-f)d\mu$$

and that gives

$$\underset{n\to\infty}{lim}\inf\left(\int_R g d\mu - \int_R f_n d\mu\right) = \int_R g\, d\mu - \lim_{n\to\infty}\sup\int_R f_n d\mu \geq \int_R g d\mu - \int_R f d\mu$$

or $\lim\limits_{n\to\infty}\sup\int_R f_n d\mu \leq \int_R f d\mu$ (1)

Repeat this Fatou's lemma argument with $g + f_n$ rather than $g - f_n$.

We get

$$\lim_{n\to\infty}\inf\int_R (g + f_n) d\mu \geq \int_R \lim_{n\to\infty}\inf(g + f_n) d\mu = \int_R (g + f) d\mu$$

and that gives

$$\lim_{n\to\infty}\inf\left(\int_R g d\mu + \int_R f_n d\mu\right) = \int_R g d\mu + \lim_{n\to\infty}\inf\int_R f_n d\mu \geq \int_R g d\mu + \int_R f d\mu$$

or $\lim\limits_{n\to\infty}\inf\int_R f_n d\mu \geq \int_R f d\mu$ (2)

Combining (1) and (2), we get

$$\int_R f d\mu \leq \lim_{n\to\infty}\inf\int_R f_n d\mu \leq \lim_{n\to\infty}\sup\int_R f_n d\mu \leq \int_R f d\mu$$

which force

$$\int_R f d\mu = \lim_{n\to\infty}\inf\int_R f_n d\mu = \lim_{n\to\infty}\sup\int_R f_n d\mu$$

And that gives the result because if

$\lim\limits_{n\to\infty}\sup a_n = \lim\limits_{n\to\infty}\inf a_n$ (for a sequence $(a_n)_{n=1}^{\infty}$,

it implies that

$\lim\limits_{n\to\infty} a_n$ exists and $\lim\limits_{n\to\infty} a_n = \lim\limits_{n\to\infty}\sup a_n = \lim\limits_{n\to\infty}\inf a_n$.

3.9 Lebesgue Monotonic Convergence Theorem (or) Beppa- Levi's Theorem

Let $\{f_n\}$ be a non-decreasing sequence of integrable function defind over a measurable set E. Let $\lim\limits_{n\to\infty} f_n$ be integrable over E, then

$$\lim_{n\to\infty}\int_E f_n\,(x) dx = \int_E \lim_{n\to\infty} f_n\,(x)\, dx$$

Proof: Let be a non-decreasing sequence of integrable function defined over a measurable set E. Let $\lim_{n\to\infty} f_n$ be integrable over E

To prove that

$$\lim_{n\to\infty} \int_E f_n\,(x)dx = \int_E \lim_{n\to\infty} f_n\,(x)dx$$

Since f_n is non-decreasing sequence and hence

$f_1 \le f_2 \le f_3 \le \cdots \le f_n\,, \forall\, n$

$\Rightarrow f_1 \le f_n, \quad \forall\ n$

$\Rightarrow f_1 - f_n \ \le 0$

$\Rightarrow f_n \ \ge 0\,, \quad \forall\ n$

where $y_n = f_n - f_1$

more over f_n is a sequence of integrable function $\{y_n\}$ is a sequence of integral

Finally, $\{y_n\}$ is a sequence of non-negative integrable function. On applying this to the Lebesgue bounded convergence theorem.

Note: To give the Lebesgue bounded convergence theorem. (Statement)

$$\lim_{n\to\infty} \int_E \ y_n\,dx = \int_E \lim_{n\to\infty} y_n\,dx$$

or $\lim_{n\to\infty} \int_E (f_n - f_1)\,dx = \int_E \lim_{n\to\infty} \big((f_n - f_1)\big)\,dx$

or $\lim_{n\to\infty} \int_E f_n^{(x)}\,dx = \int_E \lim_{n\to\infty} f_n(x)\ d_x$

Proved.

3.10 Fatou’s Lemma’s

Let $\{f_n\}$ be sequence of non-negative integrable functions defined over a measurable set E such that $\lim_{n\to\infty} \inf f_n = f$ almost everywhere on E and $\lim_{n\to\infty} \inf \int f_n(x) < \infty$, then

$$\int_E f(x)\,d_x \le \lim_{n\to\infty} inf \int_E f_n(x)\ dx$$

Proof: Let be a sequence of non-negative integrable function defined over a measurable set E such that,

(i) $\lim_{n\to\infty} inf\ f_n = f$ almost everywhere on E.

(ii) $\lim_{n\to\infty} inf\ \int f_n(x)dx < \infty$

Then, to prove that

$$\int_E f(x)\,dx \leq \lim_{n\to\infty} inf\ \int f_n(x)\,dx$$

We define $g_k(x) = \inf\ \{f_n(x)\}$

$\inf\ \{f_n(x)\}$ such that $n \geq k$

Then $g_n(x) \leq f_n(x),\ \ \forall\ n$ and therefore

$$\int_E g_n(x)\ dx \leq \int_E f_n(x)\ dx$$

Consequently

$$\lim_{n\to\infty} \int_E g_n(x)\ dx \leq \lim_{n\to\infty} inf\ \int_E f_n(x)\ dx \tag{3}$$

Since is an increasing sequence of non-negative integrable function and hence Lebesgue monotonic convergence theorem

$$\begin{aligned}\lim_{n\to\infty} \int_E g_n(x)\ dx\ &= \int_E \lim_{n\to\infty} g_n(x)\ dx\\ &= \int_E \lim_{n\to\infty} \inf\ f_n(x)dx\\ &= \int_E f(x)\,dx\end{aligned}$$

From (3) we have

$$= \int_E f(x)\,dx \leq \lim_{n\to\infty} inf.\ \int_E f_n(x)\,dx$$

Proved.

3.11 Fubini's Theorem

Fubini's theorem is a powerful tool that provides conditions for interchanging the order of integration in a double integral. Given that sums are essentially special cases of integrals (with respect to discrete measures), it also gives conditions for interchanging the order of summations, or the order of a summation and an integration.

Fubini's theorem holds under two different sets of conditions: (a) nonnegative functions g (b) functions g whose absolute value has a finite integral. We state the two versions separately, because of some subtle differences.

Theorem 1: Let $g\colon \Omega_1 \times \Omega_2 \to \mathcal{R}$ be a nonnegative measurable function. Let $\mathcal{P} = \mathcal{P}_1 \times \mathcal{P}_2$ be a product measure. Then,

(a) $\int_{\Omega_2} g(\omega_1, \omega_2)\ d\mathcal{P}_2$ is a measurable function of ω_1

(b) $\int_{\Omega_1} g(\omega_1, \omega_2)\ d\mathcal{P}_1$ is a measurable function of ω_2

(c) We have

$$\int_{\Omega_1}\left[\int_{\Omega_2} g(\omega_1, \omega_2)\, d\mathcal{P}_2\right] d\mathcal{P}_1 = \int_{\Omega_2}\left[\int_{\Omega_1} g(\omega_1, \omega_2)\, d\mathcal{P}_1\right] d\mathcal{P}_2$$

$$= \int_{\Omega_1\times\Omega_2} g(\omega_1, \omega_2)\ d\mathcal{P}$$

Note that some of the integrals above may be infinite, but this is not a problem; since everything is nonnegative, expressions of the form $\infty - \infty$ do not arise.

Proof: For simple functions $g = \sum_{i=1}^{n} a_i 1_{E_i}, E_i \in C_1 \times C_2$ statement

(a) Since measurability of $\omega_1 \to \mathcal{P}_2(E_{\omega_1})$. For a general g consider a sequence of simple functions

$$g_r(\omega_1, \omega_2) \nearrow g(\omega_1, \omega_2) \qquad \forall \omega_1, \omega_2$$

as $r \to \infty$. Then we have shown that

$$f_r(\omega_1) = \int_2 g_r(\omega_1, \omega_2)\ d\mathcal{P}_2$$

are C_1 measurable and monotonically increasing $f_r \nearrow f$. By the monotonic convergence theorem

$$f(\omega_1) \triangleq \lim_{r\to\infty} \int_2 g_r(\omega_1, \omega_2)\ d\mathcal{P}_2 \tag{4}$$

$$= \int_2 \lim_{r\to\infty} g_r(\omega_1, \omega_2)\ d\mathcal{P}_2 \tag{5}$$

$$= \int_2 g(\omega_1, \omega_2)\ d\mathcal{P}_2. \tag{6}$$

Since f is a limit of measurable $f_r's - f$ must be measurable. By (6) the integral over $\mathcal{P}_2$ is also C_1 measurable. This establishes (a) and (b) by symmetry. Now

$$\mathcal{P}(E) = \int \mathcal{P}_2(E_{\omega_1})\mathcal{P}_1(d\omega_1) \tag{7}$$

$$= \int \mathcal{P}_1(E_{\omega_2})\mathcal{P}_2(d\omega_2). \tag{8}$$

Finally (c), for a simple function g is just (7)-(8), while for a general function g we just need to integrate (7) interchanging $\int$ and $\lim$ by the monotonic convergence theorem at will. Recall now that a function is said to be integrable if it is measurable and the integral of its absolute value is finite.

Theorem 2. Let $g: \Omega_1 \times \Omega_2 \to \mathcal{R}$ be a measurable function such that

$$\int_{\Omega_1 \times \Omega_2} |g(\omega_1, \omega_2)| d\mathcal{P} < \infty,$$

where $\mathcal{P} = \mathcal{P}_1 \times \mathcal{P}_2$.

(a) For almost all $\omega_1 \in \Omega_1, g(\omega_1, \omega_2)$ is an integrable function of ω_2.

(b) For almost all $\omega_2 \in \Omega_2, g(\omega_1, \omega_2)$ is an integrable function of ω_1.

(c) There exists an integrable function $h: \Omega_1 \to \mathcal{R}$ such that $\int_{\Omega_2} g(\omega_1, \omega_2) d\mathcal{P}_2 = h(\omega_1)$, a.s. (i.e., except for a set of ω_1 of zero $\mathcal{P}_1$-measure for which $\int_{\Omega_2} g(\omega_1, \omega_2)\, d\mathcal{P}_2$ is undefined or infinite).

(d) There exists an integrable function $h: \Omega_2 \to \mathcal{R}$ such that $\int_{\Omega_1} g(\omega_1, \omega_2) d\mathcal{P}_1 = h(\omega_2)$, a.s. (i.e., except for a set of ω_2 of zero $\mathcal{P}_2$-measure for which $\int_{\Omega_1} g(\omega_1, \omega_2)\, d\mathcal{P}_1$ is undefined or infinite).

(e) We have

$$\int_{\Omega_1} \left[\int_{\Omega_2} g(\omega_1, \omega_2) d\mathcal{P}_2\right] d\mathcal{P}_1 = \int_{\Omega_2} \left[\int_{\Omega_1} g(\omega_1, \omega_2) d\mathcal{P}_1\right] d\mathcal{P}_2$$
$$= \int_{\Omega_1 \times \Omega_2} g(\omega_1, \omega_2) d\mathcal{P}.$$

Proof. By now converting from a non-negative case to integrable case should be familiar. Theorem 2 is no exception: Given a function g, decompose it into its positive and negative parts, apply Theorem 1 to each part, and in the process make sure that you do not encounter expressions of the form $\infty - \infty$.

3.12 Self-Assessment Questions

1 What is the fundamental idea behind Boral measurable function?

2 Define the term " Lebesgue-Stielitijes Measure” in the context of measure theory.

3 Explain what it means for a Riemann integral and Riemann-Stielitijes Integral.

4 State and prove the Lebesgue dominated convergence theorem.

5 Give the applications of the dominated convergence theorem.

6 Let $g, f_n , n = 1,2, \ldots$ be the integrable functions on a measure space (Ω, C, μ) such that $\left|f_{n(x)} \leq g(x)\right|$ almost everywhere for each n, Show that

$$\int \left(\lim_{n\rightarrow} sup\ f_n\right) d\mu \geq \lim_{n\rightarrow\infty} sup \int f_n d\mu$$

7 Let μ be Lebesgue measure on the real line. Let

$$f_n(x) = \begin{cases} -n^2, & 0 < x < 1/n \\ \ \ 0, & \text{otherwise} \end{cases}$$

3.13 Summary

This unit comprehensively explored the concept of measure theory and integration by delving into Borel measurable functions, which serve as the foundation for understanding the measurability of real-valued functions. The Lebesgue-Stieltjes measure was introduced, which enables the incorporation of various real-valued functions as measures. The concepts of Lebesgue-Stieltjes integral and Riemann integral were covered, highlighting their strengths and limitations. The unit offered insights into the Riemann-Stieltjes integration, which bridges the gap between Riemann and Lebesgue-Stieltjes integrals, offering flexible integration methods. Other crucial theorems such as the Lebesgue Dominated Convergence Theorem, Monotone Convergence Theorem, Fatou Lemma, and Fubini's Theorem, each contributing to a deeper understanding of limit processes, convergence, and inequalities in integration, were also studied.

3.14 References

- Billingsley, P. (1995). Probability and Measure. John Wiley & Sons.

- Durrett, R. (2019). Probability: Theory and Examples. Cambridge University Press.
- Grimmett, G., & Stirzaker, D. (2001). Probability and Random Processes. Oxford University Press.
- Klenke, A. (2014). Probability Theory: A Comprehensive Course. Springer.
- Leadbetter, R., Cambanis, S. & Pipiras, V. (2014). A Basic Course in Measure and Probability Theory for Applications. Cambridge University Press.
- Loève, M. (1977). Probability Theory I. Springer-Verlag.
- Shiryaev, A. N. (1996). Probability. Springer.
- Roussasan, G.G. (2014). Introduction to Measure-Theoretic Probability. Academic Press.

3.15 Further Reading

- "Probability and Measure" by Patrick Billingsley, John Wiley & Sons.
- "A Probability Path" by Sidney I. Resnick, Birkhäuser.
- "Convergence of Probability Measures" by Patrick Billingsley, John Wiley & Sons.
- "An Introduction to Probability Theory and Its Applications (Vol. 1)" by William Feller, John Wiley & Sons.
- “An Introduction to Measure-Theoretic Probability” by George G. Roussas, Academic Press.
- "Real Analysis and Probability" by R. M. Dudley, Cambridge University Press.
- "Foundations of Modern Probability" by Olav Kallenberg, Springer.
- "Limit Theorems for Stochastic Processes" by Jean Jacod and Albert N. Shiryaev, Springer.
- "Probability with Martingales" by David Williams, Cambridge University Press.
- "Probability with Martingales" by David Williams, Cambridge University Press. Leadbetter, R., Cambanis, S. & Pipiras, V. Cambridge University Press.

UNIT: 4 SIGNED MEASURES

Structure

4.1 Introduction

A signed measure on a measurable space is a set function which has all the properties of a measure, except that of non-negativity. It is shown that signed measures are essentially got by taking the difference of two measures. The notion of absolute continuity is introduces and the famous Radon-Nikodym theorem is proved for σ-finite signed measures. The notion of singularity, of one measure with respect to another, is studied.

Lebesgue measure on R generalizes the notion of the length of an interval. In this chapter, we see how two-dimensional Lebesgue measure on R^2 generalizes the notion of the area of a rectangle. More generally, we construct new measures that are the products of two measures.

Cartesian product is the product of any two sets, but this product is actually ordered i.e, the resultant set contains all possible and ordered pairs such that the first element of the pair belongs to the first set and the second element belongs to the second set. In addition to this, many real-life objects can be represented by using cartesian products such as a deck of cards, chess boards, computer images, etc. Most of the digital images displayed by computers are represented as pixels which are graphical representations of cartesian products.

4.2 Objectives

By the end of this unit, the learner should be able to:

- Understand the basic concept of signed measure.
- Distinguish between measure and signed measure.
- Describe the concept of product of measure space.
- Describe the concept of Cartesian product.
- Understand the difference between Cartesian product and product of measure space.

4.3 Signed Measure

Suppose (Ω, C) be a measurable space. A function $\mu: C \rightarrow \mathbb{R}$ is called signed measure on (Ω, C), if function ν has the following properties:

(i) Either one of the following is true

- $\mu(C) < \infty \ \forall A \in C$
- $\mu(C) > -\infty \ \forall \ A \in C$

(ii) $\mu(\phi) = 0$

(iii) For any sequence $\{A_n\}_{n=1}^{\infty} \subset C$ of pairwise disjoint sets, then

$$\mu(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$$

Note 1: Here we adopt the convention that if one term in the right-hand side of (i) is equal to ±∞, then the entire sum is equal to ±∞. It is important to use condition (i), which avoids situations when one term is ∞ and another term is $-\infty$.

Note 2: A set $E \in M$ is said to be positive (negative, null) for the signed measure μ if $F \in M, F \subset E \Rightarrow \mu(F) \geq 0\ (\leq 0, = 0)$

Remark 1: Assume that (Ω, C) is a measurable space and that μ_1, μ_2 are measures on (Ω, C). If at least one of the μ_i's is finite, then

$$\mu = \mu_1 - \mu_2$$

is a signed measure on (Ω, C).

Remark 2: Let (Ω, C, μ) be a measure space and let f be an integrable function. Then if $f = f^+ - f^-$, we have that

$$\lambda(E) = \int_E f\, d\mu = \int_E f^+ d\mu - \int_E f^- d\mu$$

is a signed measure on(Ω, C).

Remark 3: Let μ ne signed measure on (Ω, C), then if

(i) $\{E_n\}$ is an increasing sequence, $E_n \in C, E_n \subset E_{n+1}$, then $\mu(\cup_n E_n) = \lim_{n\to\infty} \mu(E_n)$

(ii) $\{E_n\}$ is an decreasing sequence, $E_n \in C, E_n \supset E_{n+1}$, and $\mu(E_1)$ is finite, then $\mu(\cap_n E_n) = \lim_{n\to\infty} \mu(E_n)$

4.4 Jordan and Hahn Decompositions

Theorem: Hahn Decomposition: Let μ be a signed measure on (Ω, C). Then there exist a positive set P and a negative set N for μ such that

$$P \cup N = \Omega \text{ and } P \cap N = \phi.$$

This partition is unique, in the sense that if P' and N' are another such pair of sets, then $P\Delta P' = N\Delta N'$ is null.

Proof: Without loss of generality, we assume that μ does not take value $+\infty$.

Let $c = \sup\{\mu(E): A\ is\ \mu - positive\}$. Then there exists a sequence of sets $E_n \in C$ such that $\mu(E_n) \to c$.

We consider the fields $\mathcal{A}_n$ generated by the sets $E_1, E_2, \dots E_n$.

Let now $F_n \in \mathcal{A}_n$ be such that

$$\mu(F_n) = \max_{F \in \mathcal{A}_n} \mu(F)$$

then $\mu(E_n) \leq \mu(F_n) \leq c$

For any $D \in \mathcal{A}_n$ such that

$$D \cap F_n = \phi$$

We know that $\mu(D) \leq 0$ and for any$G \in \mathcal{A}_n$ such that $G \subset F_n$, we have

$$\mu(G) \geq 0$$

Let now, $B_m = \cap_{m=1}^{\infty} F_m$

Then, $\mu(B_m) = \lim\limits_{n \to \infty} \mu(\cap_{m=1}^{n} F_m)$

We have $D_{m,n} = \cap_{m=1}^{n-1} F_m \backslash F_n \in \mathcal{A}_n$, and $\mu\left(D_{m,n}\right) \leq 0$. Since $D_{m,n} \cap F_n = \phi$

Thus, $\mu(\cap_{m=1}^{n} F_m) \geq \mu(\cap_{m=1}^{n-1}) \geq \mu(F_n)$ and $\mu(B_m) \geq \mu(F_m)$.

Now consider the set $P = \cup_m B_m$, then $\mu(P) = c$

If $E \subset P$, then $\mu(P \backslash E) = \mu(P) - \mu(F)$ and thus, $\mu(E) \geq 0$ and if $F \subset P^c$,

Then $\mu(F \cup P) = \mu(F) + \mu(P)$ and $\mu(F) \leq 0$. It means that P is a positive set and its complement $N = \Omega \backslash P$ is negative.

The decomposition given by the theorem is called *Hahn decomposition.* It is usually not unique, but if $X = P \cup N = P' \cup N'$, where $P \cap N = P' \cap N' = \phi$ and P, P' are positive sets and N, N' are negative sets, then $\mu(N \Delta N') = \mu(P \Delta P') = 0$.

Definition: Two measures μ and ν on (Ω, C) are said to be mutually singular if there are disjoint sets $A, B \in C$ with $\Omega = A \cup B$ and $\mu(A) = 0$ while $\nu(B) = 0$. In this case, we write $\mu \perp \nu$.

Theorem: Jordan Decomposition: Let μ be a signed measure on (Ω, C). Then there are two unique positive measures μ^+ and μ^- on (Ω, C) such that

$$\mu^+ \perp \mu^- \text{ and } \mu = \mu^+ - \mu^-.$$

Furthermore, if λ and ν are two any positive measures with

$$\mu = \lambda - \nu$$

then for each $E \in C$, we have

$$\lambda(E) \geq$$

Proof: Let $\{\mu^+, \mu^-\}$ be a Jordan decomposition of μ arising from the Hahn decomposition $\{P, N\}$. Then, we have

$$\mu = \mu^+ - \mu^-$$

Moreover, since

$$\mu^+(N) = \mu(P \cap N) = \mu^-(P) = 0$$

We have that $\mu^+ \perp \mu^-$

Let λ and ν be any two positive measures with

$$\mu = \lambda - \nu$$

and let $E \in A$. Then

$$\begin{aligned} \mu^+(E) &= \mu(P \cap E) \\ &= \lambda(P \cap E) - \nu(P \cap E) \\ &\leq \lambda(P \cap E) \\ &\leq \lambda(E) \end{aligned}$$

A similar argument shows that $\nu(E) \geq \mu^-(E)$.

Finally assume that $\mu^+ \perp \mu^-$

Let $\{A, B\}$ be a partition of Ω so that $\lambda(B) = 0$ and $\nu(A) = 0$.

For each $E \in C$, we have

$$\mu(E \cap A) = \lambda(E \cap A) - \nu(E \cap A) = \lambda(E \cap A) \geq 0$$

That is A is positive.

Similarly, B is negative, so $\{A, B\}$ is a Hahn decomposition. It follows that for an $E \in C$

$$\mu^+(E) = \mu(E \cap P) = \mu(E \cap A) = \lambda(E \cap A) = \lambda(E)$$

$$\mu^{-}(E) = -\mu(E \cap N) = -\mu(E \cap B) = \nu(E \cap B) = \nu(E)$$

Definition: The decomposition of signed measure ν on measure space (Ω, C) into the difference of two (nonnegative) measures given in the Jordan Decomposition Theorem is called the *Jordan decomposition of* μ.

Definition (Absolute Continuity): Let μ and ν be the measures on (Ω, C). The measure μ is said to be absolutely continuous with respect to ν, if every null set of ν is also a null set of ν and it is denoted by $\mu \ll \nu$.

Aliter: The measures μ and ν on (Ω, C). Then, $\mu \ll \nu$, if and only if

$$\forall\, E \in C{:}\, \nu(E) = 0 \Rightarrow \mu(E) = 0.$$

Definition (Mutually Singular): Let μ and ν be the measures on (Ω, C). The measures μ and ν are mutually singular, if there exists $E \in C$ such that E is null for μ and E^c is null for ν and it is denoted by $\mu \perp \nu$.

Lemma: Let (Ω, C, μ) be a measure space and let ν be a measure or signed measure defined on C, then $\nu \ll \mu$ if and only if $\nu_+ \ll \nu$ and $\nu_- \ll \nu$. Also, $\nu \ll \mu$ if and only if $|\nu| \ll \mu$.

Proof: $C = P \cup N$, where P is the positive and N is the negative set determined by Hahn Decomposition on Ω, so that $P = \Omega \backslash N$. Then $\nu_+(E) = \nu(E \cap P)$ and $\nu_-(E) = -\nu(E \cap N)$ for all $E \in C$.

If $\nu_+ \ll \nu$ and $\nu_- \ll \nu$, then $\nu(E) = \nu_+(E) - \nu_-(E) = 0$ for all $E \in C$satisfying $\mu(E) = 0$ and therefore, $\nu \ll \mu$.

Conversely, suppose that $\nu \ll \mu$.

Let $E \in C$ satisfy $\mu(E) = 0$. Then

$$\mu(E \cap P) = 0 \text{ and } \mu(E \cap N) = 0, \text{ and}$$

Therefore,

$$\nu_+(E) = \nu(E \cap P) = 0 \text{ and } \nu_-(E) = -\nu(E \cap N) = 0$$

Thus, $\nu_+ \ll \nu$ and $\nu_- \ll \nu$ if and only if $\nu \ll \mu$.

Now,

$0 \leq \nu_+(E) \leq |\nu|(E), 0 \leq v_-(E) \leq |\nu|(E)$ for all $E \in C$.

and $\quad |\nu|(E) = \nu_+(E) + \nu_-(E)$ for all $E \in C$

Therefore, $|\nu|(E) = 0$ if and only if $\nu_+(E) = 0$ and $\nu_-(E) = 0$

It follows that $|\nu| \ll \mu$ if and only if $\nu_+ \ll \nu$ as required.

4.5 Radon-Nikodym Theorem

Theorem: Let μ and ν be two finite measures on a measurable space (Ω, C) (μ is positive measure and ν is signed measure). If μ is absolutely continuous with respect to ν i.e., $\mu \ll \nu$, there is a function h that is integrable with respect to ν such that for all $E \in C$

$$\mu(E) = \int_E h\, d\mu$$

and moreover, h is unique upto almost everywhere equivalence.

Proof: Let μ and ν be two finite measures on C.

Define the positive finite Borel measure $z = \mu + \nu$

Let H denote $L^2(\Omega, C, z)$. For all $f \in H$, by the fact that $z \geq \nu$ and then the Cauchy-Schwarz inequality,

$$\int_X |f| d\nu \leq \int_X 1|f| dz \leq \left(\int_X 1 dz\right)^{1/2} \left(\int_X |f|^2 dz\right)^{1/2} = \left(\nu(X)\right)^{1/2} \left(\int_X |f|^2 dz\right)^{1/2} \quad (1)$$

Thus, for all $f \in H, f \in L^1(\Omega, C, \nu)$, and we may define a linear function L on H by

$$L(f) = \int_X f d\nu$$

It follows from (1) that for all $f \in H$,

$$|L(f)| \leq \int_X |f| d\nu \leq \left(v(X)\right)^{\frac{1}{2}} \parallel f \parallel_H$$

Therefore, L is bounded, and by the Riesz Representation Theorem, there exists a unique functiong $g \in H$ such that

$$\int_X f d\nu = \int_X fg\, dz \quad (2)$$

Hence, for any $E \in C$, $\nu(E) \geq \int_E g\, dz \geq 0$, and this means that

$$0 \leq g(x) \leq 1$$

Almost everywhere with respect to z.

Now let $A = \{x: g(x) > 0\}$, or what is the same

$$A^c = \{x: g(x) = 0\}$$

Taking$f = 1_{A^c}$ in (2), we see that

$$\mu(A^c) = 0$$

Therefore, if we define a measure $\mu^{(s)}$ by

$\mu^{(s)}(E) = \mu(A^c \cap E)$ for all $E \in C$

$\mu^{(s)}(A) = 0$

Since, $\mu^{(s)}(A) = 0$ and $v(A^c) = 0$, then $\mu^{(s)}$ and v are mutually singular.

Next define $\mu^{(ac)}$ by

$$\mu^{(ac)} = \mu - \mu^{(s)}$$

or, what is the same

$\mu^{(ac)}(E) = \mu(E \cap A)$ for all $E \in C$

It remains to find h, which we shall show is given by $h = (1 - g)/g$ on A. To see this, use $z = \mu + v$ to rewrite (2) as

$\int_X f(1 - g)dv = \int_X fg\, d\mu$ for all $f \in H$ (3)

Now, let E be any measurable subset of A, and for each positive number N define

$$f_N = 1_E \min\{g^{-1}, N\}$$

Since, $g > 0$ on E, g^{-1} is defined and finite

$$1_E g^{-1} = \lim_{N\to} f_N \qquad (4)$$

almost everywhere.

Moreover, since f_N is bounded, it belongs to H, hence from (3)

$\int_X f_N(1 - g)dv = \int_X f_N g\, d\mu$

By (4) and the Lebesgue monotone convergence theorem,

$$\int_E \frac{1-g}{g} dv = \lim_{N\to} \int_X f_N(1 - g)dv$$

$$= \lim_{N\to} \int_X f_N g d\mu = \mu(E).$$

Taking $E = A$,

$$\int_E \frac{1-g}{g} dv = \mu(A) \leq \mu(\Omega) < \infty$$

Hence the non-negative measurable function h defined by

$$h(x) = \begin{cases} 0 \ ; x \in A^c \\ \frac{1-g(x)}{g(x)} \ ; x \in A \end{cases}$$

is integrable with respect to ν and for all measurable sets E,

$$\mu^{(ac)} = \mu(E \cap A) = \int_E h \, d\nu$$

It follows immediately that if $\nu(E) = 0$, then $\mu^{(ac)}(E) = 0$, so that $\mu^{(ac)}$ is indeed absolutely continuous with respect to ν.

Also, by definition, there exist sets $A_j \in C$ such that

$$\nu(A_j) = 0 \text{ and } \lambda_j(A_j^c) = 0 \text{ for } j = 1, 2$$

where λ_j is another measure such that $\lambda_j \perp \nu$ for $j = 1, 2$.

Let $b - A_1 \cup A_2$. Then, $\nu(B) \leq \nu(A_1) + \nu(A_2) = 0$

Hence, $\nu(B) = 0$.

Consequently, $\rho_j(B) = 0$,

Where ρ_j is a measure such that $\rho_j \ll \nu$ and $\mu = \lambda_j + \rho_j$ for $j = 1, 2$.

Now also let,

$$B^c = A_1^c + A_2^c \subset A_j^c \text{ for } j = 1, 2.$$

Hence, $\lambda_j(B^c) = 0$ for $j = 1, 2$.

Now, for any $E \in C$, and each $j = 1, 2$

$$\rho_j = \rho_j(E \cap B) + \rho_j(E \cap B^c) = \rho_j(E \cap B^c)$$

$$= \rho_j(E \cap B^c) + \lambda_j(E \cap B^c) = \mu(E \cap B)$$

Where, we have used, successively, the fact that

$$\rho_j(E \cap B) = 0$$

$$\lambda_j(E \cap B^c) = 0$$

and $\quad \mu = \lambda_j + \rho_j$

Thus, $\rho_j(E) = \mu(E \cap B)$ for $j = 1, 2$

which shows that $\rho_1 = \rho_2$.

Finally, $\lambda_j = \mu - \rho_j$, so $\lambda_1 = \lambda_2$.

4.6 Products of Measure Spaces

Definition: Let X and Y are set. A rectangle in $X \times Y$ is a set of the form $A \times B$, where $A \subset X$ and $B \subset Y$.

4.6.1 Product Space

Given two measurable spaces A and B, the product space $A \times B$ is the cartesian product of the sets A and B, endowed with the σ-algebra generated by the sets of the form $E \times F$, where E is measurable in A and F is measurable in B.

4.6.2 Product of two σ- fields

Definition: Let (Ω_1, C_1) and (Ω_2, C_2) be two measurable spaces. The product σ-field $C \otimes S$ on $\Omega_1 \times \Omega_2$ is defined as the σ-field generated by the collection of all sets of the from

$\{A_1 \times A_2 : A_1 \in C_1, A_2 \in C_2\}$

The sets in this collection are called measurable rectangles.

Note: $C_1 \otimes C_2 \neq C_1 \times C_2$ because $C_1 \times C_2$ may not be closed on A^c or $A_1 \cup A_2$.

4.6.3 Product Measure

Let (Ω_1, C_1, μ_1) and $(\Omega_2, \mathrm{C}_2, \mu_2)$ be σ- finite measurable spaces. Then there is a unique measure π on σ-field $C_1 \times C_2$ such that

$\pi(A \times B) = \mu_1(A)\mu_2(B)$

Hold for each $A \in C_1$ and $B \in C_2$. Furthermore, the measure under π of an arbitrary set $E \in A \times B$ is given by

$$\pi(E) = \int_{\Omega_1} \mu_2\left(E_{\Omega_1}\right)\mu_1(dx) = \int_{\Omega_1} \mu_1\left(E_{\Omega_2}\right)\mu_1(dy)$$

The measure π is called the product of μ_1 and μ_2.

4.7 Cartesian Product of Two Sets

The Cartesian product of two sets A and B is

$$A \times B = \{(a, b) : a \in A, b \in B\}.$$

In a similar way we define the Cartesian product of $n \in N$ sets. The repeated Cartesian product of the same set, denoted as

$$A^d = A \times \cdots \times A, \quad d \in N,$$

is the set of d-tuples or d-dimensional vectors whose components are elements in A.

4.7.1 Cartesian Product of Two Measurable Spaces

Let (Ω_1, C_1, μ_1) and (Ω_2, C_2, μ_2) be two measure spaces. We consider the Cartesian product $\Omega_1 \times \Omega_2$ and for $A \subset \Omega_1$ and $B \subset \Omega_2$, we call $A \times B$ a *rectangle*. If $A \in C_1$ and $B \in C_2$, we call $A \times B$ a *measurable rectangle* provided $\mu_1(A) < \infty$ and $\mu_2(B) < \infty$.

Note: The obvious choice for the measure of $A \times B$ is $\mu_1(A) \cdot \mu_2(B)$.

4.8 Self-Assessment Questions

1 Let (Ω, C, μ) be a measure space and v_1 and v_2 be signed measure on C. Prove the following:

(i) If $v_1 \perp \mu$ and $v_2 \perp \mu$, then $v_1 + v_2 \perp \mu$.

(ii) If $v_1 \ll \mu$ and $v_2 \ll \mu$, then $v_1 + v_2 \ll \mu$

(iii)If $v_1 \ll \mu$, then $|v_1 \ll \mu|$, and conversely

(iv)If $v_1 \ll \mu$ and $v_2 \perp \mu$, then $v_1 \perp v_2$

(v) If $v_1 \ll$ and $v_1 \perp \mu$, then $v_1 \equiv 0$

2 Let m be Lebesgue measure on the interval $[0, +\infty]$, and let μ be a finite unsigned measure. Then

(i) Show that μ is a continuous measure if and only if the function $x \mapsto \mu([0, x])$ is continuous.

(ii) Show that μ is an absolutely continuous measure with respect to m if and only if the function $x \mapsto \mu([0,x])$ is absolutely continuous.

3 Given two sets $C = \{1,2,6\}$ and $D = \{8,3\}$. Find the cartesian product $C \times D$.

4 Find the cartesian product of C^2, where, $C = \{1,2\}$.

5 If $C = \emptyset$ and $D = \{1, 4, 6, -1, 7\}$, then find the number of elements in the cartesian product $C \times D$.

True or False Questions

6 For the two non-empty sets A and B, the cartesian product A × B = B × A.

4.9 Summary

This chapter delved into the intricate realm of measure theory, focusing on the concept of signed measures, which allows for the integration of functions with both positive and negative values. The Jordan decomposition theorem for measures was studied, which is a corollary of the Hahn decomposition theorem and is useful for the Lebesgue decomposition theorem, along with the Radon-Nikodym Theorem, which enables the representation of one measure with respect to another and offers a deeper understanding of measure transformations. The study extended to the derivatives of signed measures, a fundamental tool for understanding how measures change with respect to each other. Topics like product spaces and the Cartesian Product of two Sets were also explored, highlighting the concepts of product of two σ Fields, Product Measure and Cartesian Product of two Measurable Spaces.

4.10 References

- Billingsley, P. (1995). Probability and Measure. John Wiley & Sons.
- Durrett, R. (2019). Probability: Theory and Examples. Cambridge University Press.
- Grimmett, G., & Stirzaker, D. (2001). Probability and Random Processes. Oxford University Press.
- Klenke, A. (2014). Probability Theory: A Comprehensive Course. Springer.
- Loève, M. (1977). Probability Theory I. Springer-Verlag.

- Leadbetter, R., Cambanis, S. & Pipiras, V. (2014). A Basic Course in Measure and Probability Theory for Applications. Cambridge University Press.
- Roussasan, G.G. (2014). Introduction to Measure-Theoretic Probability. Academic Press.

4.11 Further Reading

- "Probability and Measure" by Patrick Billingsley, John Wiley & Sons.
- "A Probability Path" by Sidney I. Resnick, Birkhäuser.
- "Convergence of Probability Measures" by Patrick Billingsley, John Wiley & Sons.
- "Weak Convergence and Empirical Processes: With Applications to Statistics" by A. W. van der Vaart and Jon A. Wellner, Springer.
- "An Introduction to Probability Theory and Its Applications (Vol. 1)" by William Feller, John Wiley & Sons.
- "Probability: Theory and Examples" by Rick Durrett, Cambridge University Press.
- "Real Analysis and Probability" by R. M. Dudley, Cambridge University Press.
- "Foundations of Modern Probability" by Olav Kallenberg, Springer.
- "Limit Theorems for Stochastic Processes" by Jean Jacod and Albert N. Shiryaev, Springer.
- "Probability with Martingales" by David Williams, Cambridge University Press.
- “Introduction to Measure-Theoretic Probability” by Roussasan, G.G. Academic Press.

U.P. Rajarshi Tandon Open
University, Prayagraj

MScSTAT – 101N/ MASTAT – 101N

Measure & Probability Theory

Block: 2 Probability Measure, Distribution Function and Inequalities

Course Design Committee

Course Preparation Committee

Dr. Ashok Kumar **Writer**
Department of Statistics
University of Lucknow, Lucknow

Prof. Shruti **Editor**
School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj

Prof. Shruti **Course Coordinator**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 101N/ MASTAT – 101N **MEASURE & PROBABILITY THEORY**
©UPRTOU
First Edition: ***July 2023***
ISBN: 978-81-19530-86-1

Printed and Published by Col. Vinay Kumar, Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2024.

Printed By : Chandrakala Universal Pvt. Ltd. 42/7 Jawahar Lal Neharu Road, Prayagraj.

Block & Units Introduction

The ***Block - II – Probability Measure, Distribution Function and Inequalities*** is the second block in which we have three units.

In ***Unit – 5 – Probability Measure*** is discussed with Probability space of a random experiment, probability measures, random variables as a measurable function. Field induced by a sequence of random variables.

In ***Unit – 6 – Distribution Functions*** has been introduced by discussing Decomposition of distribution functions in purely discrete, absolutely continuous and singular components.

The ***Unit –****7* ***- Probability Inequalities*** dealt with CR-Inequality, Chebyshev's Inequality, Cauchy-Schwartz Inequality, Holder Inequality, Minkowski Inequality, Jensen Inequality, Lyapunov Inequality, Kolmogorov Inequality, Hajck-Renyki Inequality.

At the end of every unit the summary, self-assessment questions and further readings are given.

UNIT: 5 PROBABILITY MEASURES

Structure

5.1 Introduction

Probability theory has become increasingly important in multiple parts of science. Modern probability theory makes major use of measure theory. As we will see, a probability measure is simply a measure such that the measure of the whole space equals 1. However, probability theory is not simply the special case of measure theory where the whole space has measure 1. The questions that probability theory investigates differ from the questions natural to measure theory.

Even when concepts in probability theory have the same meaning as well-known concepts in measure theory, the terminology and notation can be quite different. Thus, one goal of this chapter is to introduce the vocabulary of probability theory. This difference in vocabulary between probability theory and measure theory occurred because the two subjects had different historical developments, only coming together in the first half of the twentieth century.

5.2 Objectives

By the end of this unit, the learner should be able to:

- Understand the basic concept of probability measure.

- Distinguish between the probability space and sample space
- Understand the concept of random experiment.
- Describe the concept of random variables.
- Distinguish between random experiment and random variable.
- Identify the relationship between random variable and measurable function.

5.3 Basic Definitions

Random Experiment: An experiment with more than one out comes which can be repeated any number of times under more or less similar condition but the outcomes of which very irregularly from repetition to repetition is called a random experiment. That is for a random experiment there are more than one outcome and the outcome very from repetition to repetition.

Example 1: Tossing a coin. The outcome of a trial can be either tail or head showing up.

Sample space: A sample space is the set of all possible simple outcomes (or sample points) of a random experiment.

Example: For the experiment of roll of a die, the sample space is

$S = \{1, 2, 3, 4, 5, 6\}$.

Probability Space: A probability space is a triplet $(\Omega, \mathcal{F}, P)$ such as:

(i) A sample space, $\mathit{\Omega}$, which is the set of all possible outcomes of a random experiment.

(ii) The σ-field or σ- algebra $\mathcal{F}$ is a set of subsets of Ω such that

(a) $\phi, \Omega \in F$, (b) if $A \in F$, then $A^c \in F$, (c) if $A_n \in F$ for $n = 1,2\ldots$, then $\cup A_n \in F$.

In words, F is closed under complementation and under countable unions, and contains the empty set. Elements of F are called measurable sets.

(iii) The ***probability measure*** is any function $P: F \to [0,1]$ is such that if $A_n \in F$ and are pairwise disjoint, then $P(\cup A_n) = \sum P(A_n)$ (countable additivity) and such that $P(\Omega) = 1$. $P(A)$ is called the *probability* of A.

Example: Suppose $n \in Z^+$ and Ω is a set containing exactly n elements. Let C denote the collection of all subsets of Ω. Then

$$\frac{\text{counting measure on } \Omega}{n}$$

is a probability measure on (Ω, F).

Remark 1: Let m denote Lebesgue measure on the interval $[0, 1]$. Then m is a probability measure on $([0, 1], B)$, where B denotes the σ-algebra of Borel subsets of $[0, 1]$.

Definition: Suppose (Ω, F, P) is a probability space. An event A is said to happen *almost surely* if the probability of A is 1, or equivalently if $P(\Omega \backslash A) = 0$.

5.4 Random Variable and Measurable Function

Definition: Any measurable function X on Ω is called a random variable.

Random variable is a measurable function from the measure space (Ω, F) to $(R, \mathcal{B})$. That is, a function $X: \Omega \to R$ such that the preimage of every set in $\mathcal{B}$ is in F. Say X is F-measurable.

Aliter: A random variable is a function $X: \Omega \to R$. It is said to be measurable with respect to F (or we say that X is a random variable w.r.t F) if for every Borel set $B \in \mathcal{B}(R)$

$$X^{-1}(B) := \{\omega \in \Omega : X(\omega) \in B\} \in F.$$

That is X is measurable with respect to F if all possible inverse of X can be found in F.

For any event X and any $x \in R$, the probability $P(X \leq x)$ is defined, which is referred to as the probability that the random variable X is bounded by x.

For example, if A is any event, then the indicator function I_A on Ω is a random variable. Fix a random variable X on Ω.

Recall that, for any Borel set $A \subset R$, the set

$$\{X \in A\} = X^{-1}(A)$$

is measurable. Hence, the set $\{X \in A\}$ is an event and we can consider the probability $P(X \in A)$ that X is in A.

Set for any Borel set A ⊂ R,

$$P_X\,(A) = P(X \in A)$$

Then we obtain a real-valued functional $P_X(A)$ on the σ-algebra $\mathcal{B}(R)$ of all Borel sets in R.

Lemma: For any random variable X, P_X is a probability measure on $\mathcal{B}(R)$. Conversely, given any probability measure μ on $\mathcal{B}(R)$, there exists a probability space and a random variable X on it such that $P_X = \mu$.

Proof: We can write $P_X(A) = P(X^{-1}(A))$. Since X^{-1} preserves all set-theoretic operations, by this formula the probability measure P on F induces a probability measure on $\mathcal{B}$.

For example, check additivity:

$$P_X(A \sqcup B) \;=\; P(X^{-1}(A \sqcup B)) \;=\; P(X^{-1}(A) \sqcup X^{-1}(B))$$
$$= P(X^{-1}(A)) \;+\; P(X^{-1}(B))$$

The σ-additivity is proved in the same way. Note also that

$$P_X(R) = P\big(X^{-1}(R)\big) = P(\Omega) = 1$$

For the converse statement, consider the probability space

$$(\Omega, F, P) = (R, \mathcal{B}, \mu)$$

and the random variable on it

$$X(x) \;=\; x.$$

Then

$$P_X(A) = P(X \in A) = P(x{:}\, X(x) \in A) = \mu(A)$$

Hence, each random variable X induces a probability measure P_X on real Borel sets.

The measure P_X is called the distribution of X.

In probabilistic terminology, $P_X(A)$ is the probability that the value of the random variable X occurs to be in A.

Any probability measure on $\mathcal{B}$ is also called a *distribution*. As we have seen, any distribution is the distribution of some random variable.

Consider examples distributions on R. There is a large class of distributions possessing a density, which can be described as follows. Let f be a non-negative Lebesgue integrable function on R such that

$$\int_R f \, d\lambda = 1$$

where $\lambda = \lambda_1$ is the one-dimensional Lebesgue measure.

Define the measure μ on $\mathcal{B}(R)$ by

$$\mu(A) = \int_A f \, d\lambda$$

Since, μ is a measure and $\mu(R) = 1$, μ is a $probability\ measure$. The Function f is called the density of μ or the *density function* of μ.

Definition: Let (Ω, F) be a measurable space and $\{X_i, i \in I\}$ be a sequence of random variables on (Ω, F). The σ-field generated by $X_i, i \in I$, denoted as $\sigma(X_i, i \in I)$, is the smallest σ-field G on Ω such that all the random variables X_i are G-measurable.

Example 1: Let (Ω, F) be a measurable space. If X_0 is a constant random variable (i.e., $X_0(\omega) = c \in R, \forall \omega \in \Omega$),

then $\sigma(X_0) = \{\emptyset, \Omega\}$.

Example 2: Let $\Omega = \{1,2,3,4,5,6\}$.

The following are σ $-$fields on Ω:

$$F_1 = \{\emptyset, \{1\}, \{2,3,4,5,6\}, \Omega\}.$$

$$F_2 = \{\emptyset, \{1,3,5\}, \{2,4,6\}, \Omega\}$$

$$F_3 = \{\emptyset, \{1,2,3\}, \{4,5,6\}, \Omega\}.$$

and $F = \mathcal{P}(\Omega)$. Then,

$$\sigma(X_1) = P(\Omega) \text{ and}$$

$$\sigma(X_2) = F_2.$$

5.5 Independent Events and Independent Random Variables

The notion of independent events, which we now define, is one of the key concepts that distinguishes probability theory from measure theory.

Suppose (Ω, F, P) is a probability space.

Then, two events A and B are called independent if

$$P(A \cap B) = P(A) \cdot P(B)$$

More generally, a family of events $\{A_k\}_{k\in I}$ is called independent if

$$P(A_{k1} \cap \cdot \cdot \cdot \cap A_{kn}) = P(A_{k1}) \cdot \cdot \cdot \ P(A_{kn})$$

whenever $k_1, \ldots, k_n$ are distinct elements of I.

Example: *Independent events: coin tossing*

Suppose $\Omega = \{H, T\}^3$, where H and T are symbols that you can think of as denoting "heads" and "tails". Thus, elements of Ω are 3-tuples of the form $\omega = (\omega_1, \omega_2, \omega_3)$, where each ωj is H or T. Let F be the collection of all subsets of Ω, and let $P = (\text{counting measure on } \Omega)/8$, as we expect from a fair coin toss. Let $A = \{\omega \in \Omega : \omega_1 = \omega_2 = H\}$ and $B = \{\omega \in \Omega : \omega_3 = H\}$. Then A contains two elements and thus $P(A) = \frac{1}{4}$, corresponding to probability $\frac{1}{4}$ that the first two coin tosses are all heads. Also, B contains four elements and thus $P(B) = \frac{1}{2}$, corresponding to probability $\frac{1}{2}$ that the third coin toss is heads.

Now $P(A \cap B) = \frac{1}{8} = P(A) \cdot P(B)$,

where the first equality holds because $A \cap B$ consists of only the one element (H, H, H) and the second equality holds because $P(A) = 1/4$, thus the above equations shows that A and B are independent.

If the events are independent then the random variable corresponding to them also independent random variables.

5.6 Self-Assessment Questions

1 What is the fundamental idea behind the probability theory?

2 What is the fundamental idea behind the random variable in probability theory?

3 Suppose (Ω, F, P) is a probability space and $A \in F$. Prove that A and $\Omega \setminus A$ are independent if and only if $P(A) = 0$ or $P(A) = 1$.

4 Define the term “random variable”.

5 Suppose P is Lebesgue measure on $[0, 1]$. Give an example of two disjoint Borel subsets sets A and B of $[0, 1]$ such that $P(A) = P(B) = \frac{1}{2}$, $[0, \frac{1}{2}]$ and A are independent, and $[0, \frac{1}{2}]$ and B are independent.

6 Suppose (Ω, F, P) is a probability space and $A, B \in F$. Prove that the following are equivalent:

(i) A and B are independent events.

(ii) A and $\Omega \setminus B$ are independent events.

(iii) $\Omega \setminus A$ and B are independent events.

(iv) $\Omega \setminus A$ and $\Omega \setminus B$ are independent events.

7 Give an example of a probability space (Ω, F, P) and events A, B_1, B_2 such that A and B_1 are independent, A and B_2 are independent, but A and $B_1 \cup B_2$ are not independent.

8 By answering these questions, learners will be able to gauge their understanding of the unit's key concepts and their ability to apply them in various contexts.

5.7 Summary

This unit highlights the basic concept of a probability measure, essential for quantifying uncertainty and randomness, and distinguishes between probability and sample space. The concept of a random experiment and random variables were introduced, providing a structured framework for studying uncertain events and their outcomes. The crucial distinction between random experiments, which encompass the broader scenario, and random variables, which focus on specific quantities or characteristics within that context, was also defined. The relationship between random variables and measurable functions was explored, along with examples for both of these concepts.

5.8 References

- Billingsley, P. (1995). Probability and Measure. John Wiley & Sons.
- Durrett, R. (2019). Probability: Theory and Examples. Cambridge University Press.
- Grimmett, G., & Stirzaker, D. (2001). Probability and Random Processes. Oxford University Press.
- Klenke, A. (2014). Probability Theory: A Comprehensive Course. Springer.
- Lehmann, E. L., & Casella, G. (2006). Theory of Point Estimation. Springer.
- Loève, M. (1977). Probability Theory I. Springer-Verlag.
- Shiryaev, A. N. (1996). Probability. Springer.

5.9 Further Reading

- "Probability and Measure" by Patrick Billingsley, John Wiley & Sons.
- "A Probability Path" by Sidney I. Resnick, Birkhäuser.
- "Convergence of Probability Measures" by Patrick Billingsley, John Wiley & Sons.
- "Weak Convergence and Empirical Processes: With Applications to Statistics" by A. W. van der Vaart and Jon A. Wellner, Springer.
- "An Introduction to Probability Theory and Its Applications (Vol. 1)" by William Feller, John Wiley & Sons.
- "Probability: Theory and Examples" by Rick Durrett, Cambridge University Press.
- "Real Analysis and Probability" by R. M. Dudley, Cambridge University Press.
- "Foundations of Modern Probability" by Olav Kallenberg, Springer.
- "Limit Theorems for Stochastic Processes" by Jean Jacod and Albert N. Shiryaev, Springer.
- "Probability with Martingales" by David Williams, Cambridge University Press.
- “Introduction to Measure-Theoretic Probability” by Roussasan, G.G, Academic Press.

UNIT: 6 DISTRIBUTION FUNCTIONS

Structure

6.1 Introduction

There is a very interesting thing about probability: everything seems to be so obvious, but when we investigate it a bit more into depth, it suddenly turns out that we don't actually understand it. The probability distribution which is usually encountered in our early stage of learning probability is the uniform distribution. Uniform means all the event has the same probability of happening.

6.2 Objectives

By the end of this unit, the learner should be able to:

- Understand the basic concept of probability distribution.
- Understand the distribution function of random vectors.
- Distinguish between probability distribution and distribution function.
- Describe the decomposition of distribution function.

- Distinguish between density function and distribution function.

6.3 Probability Distribution

Definition: Suppose (Ω, F, P) is a probability space and X is a random variable. The probability distribution of X is the probability measure P_X defined on (R, B) by

$$P_X(B) = P(X \in B) = P\big(\, X^{-1}(B)\big)$$

6.4 Distribution Function

Definition: Suppose a random variable $X: (\Omega, \mathcal{B}, P) \to R \cup \{-\infty, \infty\}$, the function $F: R \to [0,1]$ defined by

$$F(x) = P(\{\omega \in \Omega: X(\omega) \leq x\})$$

The function F is called *distribution function* or cumulative distribution function, if it satisfies the following properties:

(i) $\lim_{t \to -\infty} F(t) = 0$

Generally written as $F(-\infty) = 0$

(ii) $\lim_{t \to \infty} F(t) = 1$

Generally written as $F(\infty) = 1$

(iii) F is an non decreasing function i.e., if $s < t$, then $F(s) \leq F(t)$

(iv) F is right continuous i.e., for all t $\lim_{s \searrow t} F(s) = F(t)$

Example: Consider the random variable X defined on the probability space

$$X: ([0,1], \mathcal{B}, \mu) \to R,\ X(t) = t$$

such that

$$\mu(\{\omega \in [0,1]: X(\omega) \leq t\,\}) = \mu([0, t]) = t$$

The random variable X has the uniform distribution on $[0,1]$.

The cumulative distribution of uniform distribution on $[0,1]$ is defined as

$$F(t) = \begin{cases} 0, & t < 0 \\ t, & t \in [0,1] \\ 1, & t > 1 \end{cases}$$

So, F is strictly increasing function and hence one-to-one on $[0,1]$. Thus, $F|_{[0,1]}$ has an inverse function $\left(F|_{[0,1]}\right)^{-1}: [0,1] \to [0,1]$. In fact, $X = \left(F|_{[0,1]}\right)^{-1}$.

6.5 Decomposition of Distribution Function

Theorem: Every nondecreasing and right continuous function F on R has a decomposition

$$F(x) = F_1(x) + F_2(x) + F_3(x), \quad x \in R$$

where F_1, F_2, F_3 are nondecreasing and right continuous function, and F_1 is purely discreate, $F_2$is absolutely continuous, and F_3 is singular. Each of F_1, F_2, F_3 is unique up to an additive constant. F has at most countably many discontinuities, arising solely from possible jumps in the discreate component F_1.

Proof: Let F be a nondecreasing right continuous function defined on R and μ_F its corresponding Lebesgue-Stieltjes measure on $\mathcal{B}$. Let $\mu_F = \nu_1 + \nu_2 + \nu_3$ be the decomposition of μ_F into its three components.

If $\mu_F \ll m$, F is said to be absolutely continuous with density function

$$f = d\mu_F/dm.$$

Since $\mu_F\{(a,b]\} < \infty$ for all $-\infty < a < b < \infty$, it follows from the Radon-Nikodym Theorem that

$$(F(b) - F(b) = \mu_F\{(a,b]\} = \int_{(a,b]} f(t)dt = \int_a^b f(t)dt$$

Then, for each a and for all x,

$$F(x) = F(a) + \int_a^x f(t)dt$$

where we write $\int_a^x f(t)dt = -\int_x^a f(t)dt$ when $x < a$

If F is continuous and $\mu_F \perp m$, F is said to be (continuous) singular.

Recall that F is continuous if and only if $\mu_F(\{x\}) = 0$ for all $x \in R$. Thus, F is singular means that $\mu_F \perp m$ and $\mu_F(\{x\}) = 0$, for all $x \in R$.

If μ_F is atomic (discreate), then F is called discreate.

Then $\mu_F\ (C^c)) = 0$ for some countable set $C = \{x_n\}_{n=1}^{\infty}$ and for $-\infty < a < b < \infty$,

$$F(b) - F(a) = \mu_F\{(a,b]\} = \mu_F\{(a,b] \cap C\} = \sum_{a<x_n<b} \mu_F(\{x_n\})$$

Thus if,

$$p_n = \mu_F(\{x_n\}),$$

then $F(x) = F(a) + \sum_{a<x_n<b} p_n$ for all $x \geq a$.

Therefore, by noting that $F(x) - F(0) = F_1(x) + F_2(x) + F_3(x)$ and by adding the constant $F(0)$ any of the F_i's. Since each ν_i $(i = 1,2,3)$ is unique and each $F_i$is unique up to an additive constant by above Theorem.

Since F has at most countably many (jump) discontinuities. This also follows from the above decomposition since each absolutely continuous and singular components of a Lebesgue–Stieltjes measure have no atoms. Hence the only atoms arise from the discrete component.

If we assume that F is a distribution function, then the *Decomposition of Distribution* can be done by this theorem.

Therefore, the decomposition of distribution function is purely discreate, absolutely continuous and singular.

6.6 Density Function

Suppose X is a random variable on some probability space. If there exists $h \in L^1(R)$ (where $L^1\ (R)$ will always refer to the vector space of real-valued F-measurable functions on Ω such that $\int_\Omega |\ f\ | dP < \infty$) such that

$$F(x) = \int_{-\infty}^{x} h d\mu$$

for all $x \in R$, then h is called the density function of X.

Example: Let $\alpha > 0$ and

$$h(x) = \begin{cases} 0 & , \ x < 0 \\ \alpha e^{-\alpha x} & , x \geq 0 \end{cases}$$

Then this density function h is called the exponential density on $[0, \infty)$.

For the corresponding random variable $X(x) = x$ for $x \in R$, the distribution function $F(x)$ is given by

$$F(x) = \begin{cases} 0 & , \ x < 0 \\ 1 - e^{-\alpha x} & \ x \geq 0 \end{cases}$$

6.7 Absolutely Continuous and Singular Components

The concept of absolute continuity for a real-valued function of a real variable is particularly important when studying the various forms of the Fundamental Theorem of Calculus for the Lebesgue integral.

Definition: Let $f: [a, b] \to R$. Then the following definitions regarding f

(i) *Continuous:* f is continuous at $x_0 \in [a, b]$ if and if only if each $\epsilon > 0$ there exists $\delta > 0$ such that $x \in [a, b]$ and $|x - x_0| < \delta$ implies that $|f(x) - f(x_0)| < \epsilon$.

(ii) *Uniformly continuous:* f is uniformly continuous at $[a, b]$ if and if only if each $\epsilon > 0$ there exists $\delta > 0$ such that $x, y \in [a, b]$ and $|x - y| < \delta$ implies that $|f(x) - f(y)| < \epsilon$.

(iii)*Absolutely continuous:* f is absolutely continuous at $[a, b]$ if and if only if each $\epsilon > 0$ there exists $\delta > 0$ such that for each $n \in N$

$$a \leq x_1 < y_1 \leq x_2 < y_2 \leq \cdots \leq x_n < y_n \leq b \text{ with } \sum_{i=1}^{n}(y_i - x_i) < \delta$$

implies that

$$\sum_{i=1}^{n}|f(y_i) - f(x_i)| < \epsilon.$$

Definition: *(Singular Function).* A function $f: [a, b] \to C$ or $f: R \to C$ is singular if f is differentiable at almost every point in its domain and $f\,' = 0$ almost everywhere, where f' is the derivative of f.

6.8 Self-Assessment Questions

1 What is the fundamental concept behind the theory of distribution function in probability?

2 Show that the product of two absolutely continuous functions on a closed finite interval $[a, b]$ is absolutely continuous.

3 Let $f(x) = \begin{cases} x^n \sin\frac{2\pi}{x} & \text{if } x \in (0,1] \\ 0 & \text{otherwise} \end{cases}$

where $n \in N$. Prove the following conclusions.

a. f is continuous at each point of $[0,1]$.

b. f is uniformly continuous on $[0,1]$.

c. f is not absolutely continuous on $[0,1]$ if n = 1 but f is absolutely continuous provided n > 1.

4 Suppose (Ω, F, P) is a probability space and X is a random variable. Prove that the following are equivalent:

(a) F is a continuous function on R.

(b) F is a uniformly continuous function on R.

(c) $P(X = x) = 0$ for every $x \in R$.

5 How is the distribution function of a random vector different from that of a single random variable?

6 Let X be a random variable with distribution function

$$F(x) = \begin{cases} 0 & , \ x < 0 \\ \frac{1}{2} & , \ x = 0 \\ \frac{1}{2} + \frac{x}{2} & , 0 < x < 1 \\ 1 & , \ 1 \leq x \end{cases}$$

Write the decomposition form of $F(x)$.

7 By answering these questions, learners will be able to gauge their understanding of the unit's key concepts and their ability to apply them in various contexts.

6.9 Summary

The unit delved into the fundamental concepts of probability theory, starting with the core notion of probability distributions, which are pivotal in characterizing the likelihood of various outcomes in a random experiment. Distribution functions, often referred to as cumulative distribution functions, were also studied to have a comprehensive view of how probabilities accumulate for different values. The core concept of decomposition of distribution functions in purely discrete, absolutely continuous and singular components was covered in the unit, illuminating the various ways in which probabilities are distributed. Through this study, we gained insights into the probability distributions, distribution functions and decomposition of distribution functions.

6.10 References

- Billingsley, P. (1995). Probability and Measure. John Wiley & Sons.
- Durrett, R. (2019). Probability: Theory and Examples. Cambridge University Press.
- Grimmett, G., & Stirzaker, D. (2001). Probability and Random Processes. Oxford University Press.
- Klenke, A. (2014). Probability Theory: A Comprehensive Course. Springer.
- Lehmann, E. L., & Casella, G. (2006). Theory of Point Estimation. Springer.
- Loève, M. (1977). Probability Theory I. Springer-Verlag.
- Roussasan, G.G. (2014). Introduction to Measure-Theoretic Probability. Academic Press
- Shiryaev, A. N. (1996). Probability. Springer.

6.11 Further Reading

- "Probability and Measure" by Patrick Billingsley, John Wiley & Sons.
- "A Probability Path" by Sidney I. Resnick, Birkhäuser.
- "Convergence of Probability Measures" by Patrick Billingsley, John Wiley & Sons.
- "Weak Convergence and Empirical Processes: With Applications to Statistics" by A. W. van der Vaart and Jon A. Wellner, Springer.

- "An Introduction to Probability Theory and Its Applications (Vol. 1)" by William Feller, John Wiley & Sons.
- "Probability: Theory and Examples" by Rick Durrett, Cambridge University Press.
- "Real Analysis and Probability" by R. M. Dudley, Cambridge University Press.
- "Foundations of Modern Probability" by Olav Kallenberg, Springer.
- "Limit Theorems for Stochastic Processes" by Jean Jacod and Albert N. Shiryaev, Springer.
- "Probability with Martingales" by David Williams, Cambridge University Press.

UNIT: 7 PROBABILITY INEQUALITIES

Structure

7.1 Introduction

Probability inequalities are mathematical expressions that provide bounds or limits on the probabilities of certain events or random variables. These inequalities are fundamental tools in probability theory and statistics and are used to make statements about the likelihood of events occurring in various situations. The Cramér-Rao Inequality is a key idea in estimation theory because it limits the range of any unbiased estimator that is used to figure out a parameter. It is useful for judging how well estimators work because it sets a lower bound on the variance. On the other hand, Chebyshev's Inequality is a basic probability tool that tells how likely a random variable will be more than a certain number of standard deviations away from its mean. It shows how data likes to cluster around the mean. The Cauchy-Schwartz Inequality can be used in many ways to figure out the inner product of two random variables or vectors. Holder Inequality adds to what Cauchy-Schwartz can do by setting limits on the integrals of products of functions, which is very important in many areas, like probability theory and functional analysis. The Minkowski Inequality is an important part of measure theory, functional analysis, and probability. It shows that the norms of sums of random variables are related to the norms of their sums. This makes it a great tool for limiting the tails of probability distributions. Jensen's Inequality is a very important part of both convex analysis and probability theory and gives us a way to look at predictions by creating an inequality for the expected value of a random variable that is convex. This inequality can be used in many different situations. Also, the Lyapunov Inequality is used a lot in probability and statistics, which enables us to figure out the moments of a random variable by looking at the higher moments. The Kolmogorov Inequality is a key tool for figuring out the tail probabilities of the maximum of independent random variables and is also an important tool for understanding the extreme value theory and rare events. Lastly, the Hajek-Renyi Inequality is an important part of probability theory, especially considering big differences. It limits the chances of events deviating significantly from their expected values. This is useful for figuring out how likely extreme results will happen in different situations.

7.2 Objectives

By the end of this unit, the learner should be able to:

- Understand the basic concept of inequalities in probability theory.
- Describe the importance of Chebycheve’s Inequality.
- Describe the role of C-R Inequality in the probability theory.
- Understand the concept of convex function.
- Understand the different inequalities with their applications.

7.3 Chebychev’s Inequality

Statement: It provides an upper bound to the probability that the absolute deviation of a random variable from its mean will exceed a given threshold.

If X is a random variable with mean μ and variance σ^2, then for $k > 0$,

$$P\{|X-\mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

or

$$P\{|X-\mu| < k\} \geq 1-\frac{\sigma^2}{k^2}$$

Proof: Consider

$$\sigma^2 = Var(X)$$

$$= \int_{-\infty}^{\infty} (t-\mu)^2 f_X(t)dt$$

$$\geq \int_{-\infty}^{\mu-k} (t-\mu)^2 f_X(t)dt + \int_{\mu+k}^{\infty} (t-\mu)^2 f_X(t)dt,$$

Where the last line is by restricting the region over which we integrate a positive function. Then this is

$$\geq \int_{-\infty}^{\mu-k} k^2 f_X(t)dt + \int_{\mu+k}^{\infty} k^2 f_X(t)dt,$$

Since $t \leq \mu - k \Rightarrow k \leq |t-k| \Rightarrow k^2 \leq (t-\mu)^2$.

But we rearrange and use the definition of density function to get

$$= k^2\left(\int_{-\infty}^{\mu-k} f_X(t)dt + \int_{\mu+k}^{\infty} f_X(t)dt\right)$$

$$= k^2 P(X \leq \mu - k \; or \; X \geq \mu + k)$$

$$= k^2 P(|X-\mu| \geq k).$$

Thus,

$\sigma^2 \geq k^2 P(|X-\mu| \geq k)$,

And dividing through by k^2 gives

$P\{|X-\mu| \geq k\} \leq \frac{\sigma^2}{k^2}$.

Example: Suppose that we extract an individual at random from a population whose members have an average income of \$40,000, with a standard deviation of \$20,000. What is the probability of extracting an individual whose income is either less than \$10,000 or greater than \$70,000? In the absence of more information about the distribution of income, we cannot compute this probability exactly. However, we can use Chebyshev's inequality to compute an upper bound to it.

Solution: If X denotes income, then X is less than \$10,000 or greater than \$70,000 if and only if

$$|X - \mu| \geq k$$

where $\mu = 40{,}000$ and $k = 30{,}000$.

The probability that this happens is:

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2} = \frac{400{,}000{,}000}{900{,}000{,}000} = 4/9$$

Therefore, the probability of extracting an individual outside the income range \$10,000-\$70,000 is less than $4/9$.

7.4 Cauchy-Schwartz Inequality

Statement: If X and Y are random variables taking real values then

$$[E(XY)]^2 \leq E(X^2)E(Y^2). \qquad (1)$$

Proof: let us consider a real valued function of the real variable t, defined by

$$Z(t) = E(X + tY)^2$$

which is always non-negative, since $(X + tY)^2 \geq 0$, for all real X, Y and t.

Thus $\quad Z(t) = E(X + tY)^2 \geq 0 \;\; \forall\, t.$

$$\Rightarrow Z(t) = E[X^2 + 2tXY + t^2Y^2]$$

$$= E(X^2) + 2tE(XY) + t^2E(Y^2) \geq 0, \text{ for all } t.$$

Obviously, $Z(t)$ is a quadratic expression in $'t'$ where t is constant, i.e., not random. Clearly $E(X + tY)^2 \geq 0$ for all real values of t. Now recall that for $a, b, c,$ the polynomial $at^2 + bt + c$ remains non-negative as t changes if and only if $a \geq 0$ and the discriminant $b^2 - 4ac \leq 0$. So

$$b^2 - 4ac = 4E(XY)^2 - 4E(X^2)E(Y^2)$$

So, $\quad 4E(XY)^2 - 4E(X^2)E(Y^2) \leq 0$

$$\Rightarrow E(XY)^2 - E(X^2)E(Y^2) \leq 0$$

$$\Rightarrow E(XY)^2 \le E(X^2)E(Y^2)$$

Remark 1: The sign of equality holds in (1) if and only if

$$E(X+tY)^2 = 0 \;\forall\, t \quad \Rightarrow \quad P\{(X+tY)^2 = 0\} = 1.$$

Remark 2: If the random variable X takes the real values $x_1, x_2, \dots, x_n$ and the r.v. Y takes the real values $y_1, y_2, \dots, y_n$ then Cauchy-Schwartz inequality implies:

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right)^2 \le \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right)$$

$$\Rightarrow \quad (\textstyle\sum_{i=1}^{n} x_i y_i)^2 \le (\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2)$$

The sign of equality holds if and only if $\frac{x_i}{y_i} = constant = k \quad \forall\, i = 1, 2, \dots, n.$

Remark 3: Replacing X by $|X - E(X)| = |X - \mu_x|$and taking $Y = 1$ in (*), we get

$$[E|X-\mu_x|]^2 \le E|X-\mu_x|^2 . E(1)$$

$\Rightarrow$ (Mean Deviation about mean)$^2 \le Variance(X)$

$\Rightarrow$ M.D.$\le$ S.D.

7.5 Holder's Inequality

Lemma: Let $a, b > 0$ and $p, q > 1$ satisfy $p^{-1} + q^{-1} = 1$. Then $p^{-1}a^p + q^{-1}b^q \ge ab$ with equality if and only if $a^p = b^q$.

Proof: for fix $b > 0$, let $g(a; b) = p^{-1}a^p + q^{-1}b^q - ab$.

We require that $g(a;\ b) \ge 0$ for all a.

Differentiating w.r.t. a for fixed b yields

$$\frac{\delta g}{\delta a} = a^{p-1} - b = 0$$

$$\Rightarrow a^{p-1} = b$$

so that $g(a;\ b)$ is minimized (the second derivative is strictly positive at all a) when $a^{p-1} = b$, and at this value of a, the function takes the value

$$p^{-1}a^p + q^{-1}(a^{p-1})^q - a^p = p^{-1}a^p + q^{-1}a^p - a^p$$

(using $p^{-1} + q^{-1} = 1$ which implies that $p + q = pq$, so $(a^{p-1})^q = a^{pq-q} = a^p$)

Now $\frac{\delta^2 g}{\delta a^2} = (p-1)a^{(p-2)} > 0$

As the second derivative is strictly positive at all a, the minimum is attained at the unique value of a where $a^{p-1} = b$, where, raising both sides to power q yields $a^p = b^q$.

Theorem: (Holder's Inequality): Suppose X and Y be two random variables and $p, q > 1$ such that $p^{-1} + q^{-1} = 1$, then

$$|E(XY)| \le E|XY| \le \{E|X|^p\}^{1/p}\,\{E|Y|^q\}^{1/q}.$$

Proof: we know that

$$-|XY| \le XY \le |XY|$$

$$\Rightarrow \quad E\{-|XY|\} \le E(XY) \le E|XY|$$

$$\Rightarrow \quad -E\{|XY|\} \le E(XY) \le E|XY|$$

(since if $X \ge 0$ then $E(X) \ge 0$ and if $X \ge Y$ then $E(X) \ge E(Y)$)

Therefore, $|E(XY)| \le E|XY|$.

Now to prove,

$$E|XY| \le \{E|X|^p\}^{1/p}\,\{E|Y|^q\}^{1/q},$$

In Lemma set

$$a = \frac{|X|}{\{E|X|^p\}^{1/p}} \quad \text{and} \quad b = \frac{|Y|}{\{E|Y|^q\}^{1/q}},$$

We get

$$p^{-1}\frac{|X|^p}{E|X|^p} + q^{-1}\frac{|Y|^q}{E|Y|^q} \ge \frac{|XY|}{\{E|X|^p\}^{1/p}\{E|Y|^q\}^{1/q}}$$

Taking expectations on both sides, we get

$$p^{-1}\frac{E|X|^p}{E|X|^p} + q^{-1}\frac{E|Y|^q}{E|Y|^q} \ge \frac{E|XY|}{\{E|X|^p\}^{1/p}\{E|Y|^q\}^{1/q}}$$

$$\Rightarrow p^{-1} + q^{-1} \ge \frac{E|XY|}{\{E|X|^p\}^{1/p}\{E|Y|^q\}^{1/q}}$$

$$\Rightarrow E|XY| \le \{E|X|^p\}^{1/p}\{E|Y|^q\}^{1/q}.$$

Remark: if $p = q = 2$, then $E|XY| \leq \sqrt{E|X|^2 E|Y|^2}$ which *is* ***Chauchy-Schwartz inequality***.

7.6 Minkowski Inequality

Statement: Let X and Y be two random variables and $1 < p < \infty$. Then

$$\{E|X+Y|^p\}^{1/p} \leq \{E|X|^p\}^{\frac{1}{p}} + \{E|Y|^p\}^{\frac{1}{p}}.$$

Proof: By the triangle inequality $|X+Y| \leq |X| + |Y|$

Multiply both sides by $|X+Y|^{p-1}$, we get

$$|X+Y||X+Y|^{p-1} \leq |X||X+Y|^{p-1} + |Y||X+Y|^{p-1}$$

$$\Rightarrow \; |X+Y|^p \leq |X||X+Y|^{p-1} + |Y||X+Y|^{p-1}$$

Taking expectation on both sides, we get

$$E|X+Y|^p \leq E\{|X||X+Y|^{p-1}\} + E\{|Y||X+Y|^{p-1}\}$$

Now using Holder's inequality on right hand side, we have

$$E|X+Y|^p \leq \{E|X|^p\}^{\frac{1}{p}}\left\{E|X+Y|^{(p-1)q}\right\}^{1/q} + \{E|Y|^p\}^{\frac{1}{p}}\left\{E|X+Y|^{(p-1)q}\right\}^{1/q}$$

$$\Rightarrow \frac{E|X+Y|^p}{\left\{E|X+Y|^{(p-1)q}\right\}^{\frac{1}{q}}} \leq \left[\{E|X|^p\}^{\frac{1}{p}} + \{E|Y|^p\}^{\frac{1}{p}}\right]$$

$$\Rightarrow \frac{E|X+Y|^p}{\{E|X+Y|^p\}^{1-\frac{1}{p}}} \leq \left[\{E|X|^p\}^{\frac{1}{p}} + \{E|Y|^p\}^{\frac{1}{p}}\right]$$

(since, $p^{-1} + q^{-1} = 1 \; \Rightarrow (p-1)q = p$))

$$\Rightarrow E|X+Y|^p \leq \left[\{E|X|^p\}^{\frac{1}{p}} + \{E|Y|^p\}^{\frac{1}{p}}\right]$$

Continuous Convex Function: A continuous function $g(x)$ on the interval I is convex if for every x_1 and $x_2, \frac{x_1+x_2}{2} \in I$, we have

$$g\left(\frac{x_1+x_2}{2}\right) \leq \frac{1}{2}g(x_1) + \frac{1}{2}g(x_2)$$

7.7 Jensen's Inequality

Statement: If g is a continuous and convex function on the interval I, and X is a random variable whose values are in I with probability 1, then

$$E[g(X)] \geq g[E(X)],$$

provided the expectations exists.

Proof: Let $l(X)$ be tangent to $g(X)$. Let $g(X) = a + bX$ for some a and b.

Since $g(X)$ is convex, then

$$g(X) \geq l(X) = a + bX$$

$$\Longrightarrow \quad g(X) \geq a + bX$$

$$\Longrightarrow E[g(X)] \geq a + bE(X)$$

$$= l\big(E(x)\big) = g\big(E(x)\big)$$

Hence, $E[g(X)] \geq g\big(E(x)\big)$.

Corollary to Jensen's Inequality: If $g$is a continuous and concave function on the interval I, and X is a random variable whose values are in I with probability 1, then

$$E[g(X)] \leq g[E(X)],$$

provided the expectations exists.

For example,

- $g(x) = x^2$is convex function then $E(X^2) \geq \{E(X)\}^2$.
- $g(x) = \log x$ is a concave function then $E\{\log(X)\} \leq \log\{E(X)\}$.

7.8 Lyapunov Inequality

Statement: Suppose X and Y be two random variables and $p, q > 1$ such that $p^{-1} + q^{-1} = 1$, then

$$\{E|X|^r\}^{1/r} \leq \{E|X|^s\}^{1/s} \text{ for } 1 < r < s < \infty.$$

Proof: Put $Y = 1$ with probability 1 in Holder's inequality. Then for $1 < p < \infty$,

$$E|X| \leq \{E|X|^p\}^{\frac{1}{p}}.$$

Let $1 < r < p$. Then

$$E|X|^r \leq \{E|X|^{pr}\}^{\frac{1}{p}}$$

and letting $s = pr > r$ yields

$$E|X|^r \leq \{E|X|^s\}^{\frac{r}{s}}$$

So that

$$\{E|X|^r\}^{1/r} \leq \{E|X|^s\}^{\frac{r}{s}} \text{ for } 1 < r < s < \infty.$$

7.9 Kolmogorov Inequality

Statement: Suppose $X_1, X_2, \dots, X_n$ are independent with $E(X_i) = 0$ and $Var(X_i) < \infty$. $S_j = X_1 + \cdots + X_j$. Then,

$$P\left(\max_{1\leq j\leq n}|S_j| \geq \epsilon\right) = \frac{var(S_n)}{\epsilon^2}.$$

Proof: Let $T = \min\{j \leq n: |S_j| \geq \epsilon\}$, with minimum of empty set being ∞. Then, $\{T \leq j\}$ or $\{T = j\}$ only depends on $X_1, \dots, X_j$ and as a result

$$\{T \geq j\} = \{T \leq j-1\}^c = \{S_i \leq \epsilon, 1 \leq i \leq j-1\}$$

Only depends on $X_1, \dots, X_{j-1}$ and therefore is independent of $X_j, X_{j+1}, \dots$.

Write,

$$P\left(\max_{1\leq j\leq n}|S_j| \geq \epsilon\right) = P(T \leq n) \leq \epsilon^{-2}E\left(|S_r|^2 1_{\{T\leq n\}}\right) \leq \epsilon^{-2}E(|S_{T\wedge n}|^2)$$

$$= \epsilon^{-2}E\left(\left|\sum_{j=1}^{T\wedge n} X_j\right|^2\right) = \epsilon^{-2}E\left(\left|\sum_{j=1}^{n} X_j 1_{\{T\geq j\}}\right|^2\right)$$

$$= \epsilon^{-2}\left\{E\left(\sum_{j=1}^{n} X_j^2 1_{\{T\geq j\}}\right) + 2\sum_{1\leq i\leq j\leq n}^{n} E(X_j X_i 1_{\{T\geq j\}} 1_{\{T\geq i\}}\right\}$$

$$= \epsilon^{-2}\left\{\sum_{j=1}^{n} E\left(X_j^2\right)P(T \geq j) + 2\sum_{1\leq i\leq j\leq n}^{n} E(X_j)E(X_i 1_{\{T\geq j\}} 1_{\{T\geq i\}})\right\}$$

$$= \epsilon^{-2} \sum_{j=1}^{n} E\left(X_j^2\right) P(T \ge j) + 0$$

$$\le \frac{var(S_n)}{\epsilon^{-2}}.$$

7.10 Hajck-Renyki Inequality

Statement: Let $\{X_n, n \ge 1\}$ be an associated sequence of random variables with $Var\left(X_j\right) = \sigma_j^2$ and $\{b_n, n \ge 1\}$ be a positive non-decreasing sequence of real numbers. Then, for any $\epsilon > 0$,

$$P\left\{\max_{1\le k\le n}\left|\frac{1}{b_n}\sum_{i=1}^{k}[X - E(X_i)]\right| \ge \epsilon\right\} \le 4\epsilon^{-2}\left\{\sum_{j=1}^{n}\frac{Var(X_j)}{b_j^2} + \sum_{1\le j\ne k\le n}\frac{Cov(X_j,X_k)}{b_j b_k}\right\}$$

Proof: Let $Y_j = b_j^{-1}[X - E(X_i)]$. It is clear that $\{Y_n, n \ge 1\}$ is a zero mean associated sequence.

Let $S_n = \sum_{j=1}^{n}[X - E(X_i)]\,, n \ge 1.$ Let $b_0 = 0$.

Note that $S_k = \sum_{j=1}^{k} b_j Y_j = \sum_{j=1}^{k}\left(\sum_{i=1}^{j}(b_i - b_{i-1})\right)Y_j$

$$= \sum_{i=1}^{k}(b_i - b_{i-1})\left(\sum_{j=i}^{k} Y_j\right)$$

Since $b_k^{-1}\sum_{(i=1)}^{k}(b_i - b_{i-1}) = 1$, it follows that

$$\left[\left|\frac{S_k}{b_k}\right| \ge \varepsilon\right] \subset \left[\max_{1\le i\le k}\left|\sum_{j=i}^{k} Y_j\right| \ge \varepsilon\right]$$

and hence

$$\left[\max_{1\le k\le n}\left|\frac{S_k}{b_k}\right| \ge \varepsilon\right] \subset \left[\max_{1\le k\le n}\max_{1\le i\le k}\left|\sum_{j=i}^{k} Y_j\right| \ge \varepsilon\right]$$

$$= \left[\max_{1\le i\le k\le n}\left|\sum_{j=1}^{k} Y_j - \sum_{j=1}^{i} Y_j\right| \ge \varepsilon\right] \subset \left[\max_{1\le i\le n}\left|\sum_{j=1}^{i} Y_j\right| \ge \frac{\varepsilon}{2}\right].$$

Therefore,

$$P\left(\max_{1\le k\le n}\left|\frac{S_k}{b_k}\right| \ge \varepsilon\right) \le P\left(\max_{1\le i\le n}\left|\sum_{j=1}^{i} Y_j\right| \ge \frac{\varepsilon}{2}\right).$$

Applying the Chebyschev's inequality, we get that

$$P\left(\max_{1\le k\le n}\left|\frac{S_k}{b_k}\right|\ge\varepsilon\right)\le 4\varepsilon^{-2}E\left(\max_{1\le i\le n}\left|\sum_{j=1}^{i}Y_j\right|^2\right).$$

We now apply the Kolmogorov-type inequality, for the expression on right-hand side of the above inequality, valid for partial sums of associated random variables $\{Y_j, 1\le j\le n\}$ with mean zero. Hence, we have

$$P\left(\max_{1\le k\le n}\left|\frac{S_k}{b_k}\right|\ge\varepsilon\right)\le 4\varepsilon^{-2}E\left[\sum_{j=1}^{n}Y_j\right]^2$$

$$= 4\varepsilon^{-2}Var\left[\sum_{j=1}^{n}Y_j\right]$$

$$= 4\varepsilon^{-2}\left\{\sum_{j=1}^{n}Var(Y_j)+\sum_{1\le j\ne k\le n}Cov(Y_j,Y_k)\right\}$$

$$= 4\varepsilon^{-2}\left\{\sum_{j=1}^{n}\frac{Var(X_j)}{b_j^2}+\sum_{1\le j\ne k\le n}\frac{Cov(X_j,X_k)}{b_jb_k}\right\}.$$

From the non-decreasing positive property of the sequence $\{b_n, n\ge 1\}$, it follows that

$$P\left(\max_{1\le k\le n}\left|\frac{1}{b_n}\sum_{i=1}^{k}[X_i-E(X_i)]\right|\ge\varepsilon\right)\le 4\varepsilon^{-2}\left\{\sum_{j=1}^{n}\frac{Var(X_j)}{b_j^2}+\sum_{1\le j\ne k\le n}\frac{Cov(X_j,X_k)}{b_jb_k}\right\}$$

Proving the Hajck-Renyki Inequality.

7.11 C-R Inequality

Likelihood Function: Let $X_1, X_2, \dots, X_n$ have a joint density function $f(X_1, X_2, \dots, X_n|\theta)$, where θ is the parameter. Given $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ is observed, the function of θ defined by

$$L(\theta) = L(\theta|x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n|\theta)$$

is called the *likelihood function*.

Regularity conditions for Cramer-Rao Inequality

(i) The parameter space Θ is a non degenerate open interval on the real line $R(-\infty,\infty)$.

(ii) For almost all $x = (x_1, x_2 \dots, x_n)$ and for all $\theta \in \Theta$,

$$\frac{\partial}{\partial\theta}L(x,\theta)$$

exists and is independent θ.

(iii)The range of integration is independent of the parameter θ, so that $f(x,\theta)$ is differentiable under integral sign.

If range is not independent of θ and f is zero at the extremes of the range i.e., $f(a,\theta) = 0 = f(b,\theta)$, then

$$\frac{\partial}{\partial\theta}\int_a^b f\,dx = \int_a^b \frac{\partial f}{\partial\theta}dx - f(a,\theta)\frac{\partial a}{\partial\theta} + f(b,\theta)\frac{\partial b}{\partial\theta}$$

$\Rightarrow \frac{\partial}{\partial\theta}\int_a^b f\,dx = \int_a^b \frac{\partial f}{\partial\theta}dx$,since $f(a,\theta) = 0 = f(b,\theta)$.

(iv)The conditions of uniform convergence of integrals are satisfied so that differentiation under the integral sign is valid.

(v) $I(\theta) = E\left[\left\{\frac{\partial}{\partial\theta}\log L(x,\theta)\right\}^2\right]$, exists and is positive for all $\theta \in \Theta$.

Statement: If t is an unbiased estimator for $\gamma(\theta)$, a function of parameter θ, then

$$Var(t) \geq \frac{\left\{\frac{\partial}{\partial\theta}\gamma(\theta)\right\}}{E\left(\frac{\partial}{\partial\theta}\log L\right)^2} = \frac{\{\gamma'(\theta)\}^2}{I(\theta)}$$

where $I(\theta)$ is the information on θ, supplied by the sample.

In other words, Cramer-Rao inequality provides a lower bound $\frac{\{\gamma'(\theta)\}^2}{I(\theta)}$, to the variance of an unbiased estimator of $\gamma(\theta)$.

Proof: Let X be a random variable having the p.d.f. $f(x,\theta)$ and let L be the likelihood function of the random sample $(x_1, x_2, \dots x_n)$ from this population. Then

$L = L(x,\theta) = \prod_{i=1}^n f(x_i,\theta)$

Since L is the joint p.d.f. of $(x_1, x_2, \dots x_n)$, then

$\int L(x,\theta)dx = 1$

where $\int dx = \int \int \ldots \int dx_1 dx_2 \ldots dx_n$

Differentiating with respect to θ and using regularity conditions given above, we get

$$\int \frac{\partial}{\partial\theta} L\, dx = 0$$

$$\Rightarrow \int \left(\frac{\partial}{\partial\theta} \log L\right) dx = 0$$

$$\Rightarrow E\left(\frac{\partial}{\partial\theta} \log L\right) = 0 \tag{2}$$

Now, let $t = t(x_1, x_2, \ldots, x_n)$ be an unbaised estimator of $\gamma(\theta)$ such that

$$E(t) = \gamma(\theta)$$

$$\Rightarrow \int t\, L\, dx = \gamma(\theta)$$

Differentiating with respect to θ, we get

$$\int t \frac{\partial L}{\partial\theta} dx = \gamma'(\theta)$$

$$\Rightarrow \int \left(\frac{\partial}{\partial\theta} \log L\right) L\, dx = \gamma'(\theta)$$

$$\Rightarrow E\left(t\ \frac{\partial}{\partial\theta} \log L\right) = \gamma'(\theta)$$

$$Cov\left(t, \frac{\partial}{\partial\theta} \log L\right) = E\left(t\ \frac{\partial}{\partial\theta} \log L\right) - E(t) E\left(\frac{\partial}{\partial\theta} \log L\right) = \gamma'(\theta)$$

We have

$$[r(X, Y)]^2 \leq 1$$

$$\Rightarrow \left[Cov\left(t, \frac{\partial}{\partial\theta} \log L\right)\right]^2 \leq Var(X) Var(Y)$$

$$\therefore \left[Cov\left(t, \frac{\partial}{\partial\theta}\log L\right)\right]^2 \leq Var\ (t) Var\left(\frac{\partial}{\partial\theta}\log L\right)$$

$$\Rightarrow [\gamma'(\theta)]^2 \leq Var\ (t) \left[E\left\{\frac{\partial}{\partial\theta}\log L\right\}^2 - \left\{E\left(\frac{\partial}{\partial\theta}\log L\right)\right\}^2\right]$$

$$\Rightarrow [\gamma'(\theta) \leq Var(t)]E\left\{\frac{\partial}{\partial\theta}\log L\right\}^2$$

$$\Rightarrow Var(t) \geq \frac{[\gamma'(\theta]^2}{E\left\{\frac{\partial}{\partial\theta}\log L\right\}^2}$$

Which is Cramer-Rao inequality.

Remark 1: If t is an unbiased estimator of parameter θ i.e., $\gamma(\theta) = \theta$ then $\gamma'(\theta) = 1$,

$$Var(t) \geq \frac{1}{E\left(\frac{\partial}{\partial\theta}\log L\right)^2} = \frac{1}{I(\theta)}$$

where $I(\theta) = E\left\{\left(\frac{\partial}{\partial\theta}\log L\right)^2\right\}$ is called by R.A. Fisher as the amount of information on θ supplied by the sample $(x_1, x_2, \dots, x_n)$ and its reciprocal $1/I(\theta)$, as the information limit to the variance of estimator $t = t(x_1, x_2, \dots, x_n)$.

Remark 2: An estimator T is said to be unbiased estimator of $\gamma(\theta)$ if $E(T) = \gamma(\theta)$.

7.12 Self-Assessment Questions

1 What is the importance of various inequalities in probability theory?

2 Suppose a fair coin is tossed 50 times. The bound on the probability that the number of heads will be greater than 35 or less than 15 can be found using Chebyshev's Inequality.

3 Let X be a random variable such that $E(X) = 0$ and $P(-3 < X < 2) = \frac{1}{2}$. Find the bound of its variance.

4 Let X be a random variable such that $E(X) = 1$ and $Var(X) = 1$. Find a lower bound to the probability $P \geq 3)$.

5 Define the term "convex function".

6 Explain what is means of unbiased estimator.

7 By answering these questions, learners will be able to gauge their understanding of the unit's key concepts and their ability to apply them in various contexts.

7.13 Summary

The unit “Probability Inequalities” covered a diverse array of inequalities such as CR-Inequality, Chebyshev's Inequality, Cauchy-Schwartz Inequality, Holder Inequality, Minkowski Inequality, Jensen Inequality, Lyapunov Inequality, Kolmogorov Inequality, and Hajck-Renyki Inequality, which serve as valuable tools for understanding the distribution and properties of random variables.

Studying these inequalities provides the concepts of bounds, relationships, and constraints, contributing to a deeper understanding of probability distributions, moments, and statistical analysis. Studying these inequalities is indispensable in various contexts, as they allow us to derive meaningful conclusions and make informed decisions in probability and statistics.

7.14 References

- Billingsley, P. (1995). Probability and Measure. John Wiley & Sons.
- Durrett, R. (2019). Probability: Theory and Examples. Cambridge University Press.
- Grimmett, G., & Stirzaker, D. (2001). Probability and Random Processes. Oxford University Press.
- Klenke, A. (2014). Probability Theory: A Comprehensive Course. Springer.
- Lehmann, E. L., & Casella, G. (2006). Theory of Point Estimation. Springer.
- Roussasan, G.G. (2014). Introduction to Measure-Theoretic Probability. Academic Press
- Loève, M. (1977). Probability Theory I. Springer-Verlag.
- Shiryaev, A. N. (1996). Probability. Springer.

7.15 Further Reading

- "Probability and Measure" by Patrick Billingsley, John Wiley & Sons.
- "A Probability Path" by Sidney I. Resnick, Birkhäuser.
- "Convergence of Probability Measures" by Patrick Billingsley, John Wiley & Sons.
- "Weak Convergence and Empirical Processes: With Applications to Statistics" by A. W. van der Vaart and Jon A. Wellner, Springer.
- "An Introduction to Probability Theory and Its Applications (Vol. 1)" by William Feller, John Wiley & Sons.
- "Probability: Theory and Examples" by Rick Durrett, Cambridge University Press.
- "Real Analysis and Probability" by R. M. Dudley, Cambridge University Press.
- "Foundations of Modern Probability" by Olav Kallenberg, Springer.
- "Limit Theorems for Stochastic Processes" by Jean Jacod and Albert N. Shiryaev, Springer.
- "Probability with Martingales" by David Williams, Cambridge University Press.

U.P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 101N/ MASTAT – 101N
Measure & Probability Theory

Block: 3 Convergence, Characteristic Function and Limit Theorems

Course Design Committee

Dr. Ashutosh Gupta **Chairman**
Director, School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. S. Lalitha **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics
D. D. U. Gorakhpur University, Gorakhpur.

Prof. Shruti **Member-Secretary**
Professor, School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Dr. Pratyasha Tripathi **Writer**
Department of Statistics
Tilka Manjhi Bhagalpur University, Bhagalpur, Bihar

Prof. Shruti **Editor**
School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj

Prof. Shruti **Course Coordinator**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 101N/ MASTAT – 101N **MEASURE & PROBABILITY THEORY**
©UPRTOU
First Edition: ***July 2023***
ISBN: 978-81-19530-86-1

Printed and Published by Col. Vinay Kumar, Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2024.

Printed By : Chandrakala Universal Pvt. Ltd. 42/7 Jawahar Lal Neharu Road, Prayagraj.

Block & Units Introduction

The ***Block - III – Convergence, Characteristics Function and Limit Theorems*** has four units.

Unit – 8 – Convergence dealt with Sequences of distribution functions, weak and complete convergence of sequence of distribution function, Different types of convergence of sequence of random variables distribution function of random vectors.

Unit –9 – Law of Large Numbers, comprises the Weak Law of Large Numbers (WLLN), Strong Law of Large Numbers (SLLN), Khinchin's Theorem, Borel Zero-One Law, Borel-Cantelli Lemmas.

In ***Unit – 10 – Characteristic Function***, we have discussed the Helly – Bray Lemma and Theorem, Weak Compactness Theorem, Kolmogorav Theorems, Characteristic Function, Inversion Theorem, Continuity Theorem, Uniqueness Theorem.

The ***Unit – 11 – Central Limit Theorems*** discussed One Dimensional Central Limit Problem: Lindeberg-Levy, Lyapunov, Lindeberg-Feller Theorems.

At the end of every unit the summary, self-assessment questions and further readings are given.

UNIT:8 CONVERGENCE

Structure

8.1 Introduction

In the fascinating world of probability and statistics, the idea of convergence stands as one of the cornerstone concepts. At its core, convergence deals with understanding how certain mathematical entities, primarily sequences, approach a specific value or behavior as they advance. But why is such a concept vital?

Imagine making repeated measurements in an experiment. As you accumulate data, you'd expect your conclusions or results to stabilize or 'converge' to a certain truth. Similarly, in probability, as we gather more data or consider more trials, we expect our outcomes to settle down to a particular behavior. This 'settling down' is encapsulated by the notion of convergence.

In the context of random variables, which can be thought of as unpredictable quantities driven by chance, understanding convergence is akin to predicting how these quantities will behave as the number of trials or observations increases. This predictability can be crucial, especially when making decisions based on these variables.

In this unit, our journey begins with sequences of distribution functions, which capture the probabilistic behavior of sequences of random variables. We will dive deep into understanding how these sequences stabilize or converge. In doing so, we will encounter different 'flavors' of convergence, each with its distinct characteristics and implications. From the weak convergence, which concerns the convergence of distributions, to the almost sure convergence, which delves into almost certain outcomes, our exploration will be both broad and deep.

Furthermore, as we dive deeper, we'll encounter the realm of random vectors. Unlike a single random variable, which can be visualized as a singular uncertain quantity, random vectors represent multiple such quantities simultaneously. How do these vectors behave in tandem? What does convergence mean in this multi-dimensional setting?

So, as we embark on this exploration of convergence, we are not just studying an abstract mathematical concept. We are gearing up to understand the very behavior and essence of randomness and uncertainty, and how they manifest themselves in the face of increasing trials or observations. Buckle up for a deep dive into the heart of probability theory!

8.2 Objectives

By the end of this unit, the learner should be able to:

- Understand the basic concept of convergence in probability theory.
- Distinguish between sequence of distribution functions and sequence of random variables.
- Describe the difference between weak and complete convergence.

- Identify various types of convergence for sequences of random variables.
- Understand the distribution function of random vectors.

8.3 Sequence of Distribution Functions

In probability and statistics, distribution functions (or cumulative distribution functions, CDFs) give the probability that a random variable is less than or equal to a particular value. When we talk about a sequence of distribution functions, we are considering a series of such functions, each one corresponding to a different random variable or distribution.

Main Concept: Think of each distribution function in the sequence as representing the probability distribution of a different group of people's heights. As you move through the sequence, you might be looking at groups from different towns, countries, or even different time periods. The sequence can tell you how the overall pattern (or distribution) of heights changes from one group to the next.

1. Convergence of Distribution Functions: If, for every value (like a specific height), the probability of being less than or equal to that value gets closer and closer for each group in the sequence, then the sequence of distribution functions is said to converge.

Example: Suppose you're studying the heights of high school graduates over several decades. If over time, the probability of finding a student taller than 6 feet increases and gets closer to a specific value (say 20%), then the sequence of distribution functions representing the heights of students over the decades is converging at the 6 feet mark.

2. Pointwise Convergence: If a sequence of distribution functions converges for a particular value (like a specific height), then it is said to have pointwise convergence at that height.

Example: Using the above high school graduates' example, if the sequence is converging only at 6 feet but not necessarily at other heights, then it has pointwise convergence at 6 feet.

3. Uniform Convergence: If a sequence of distribution functions converges for all values (all possible heights) simultaneously, it's said to have uniform convergence.

Example: If for our high school graduates, the sequence is converging not just at 6 feet but at 5 feet, 5.5 feet, 6.5 feet, and so on (for every possible height), then the sequence has uniform convergence.

Definition:

A sequence of distribution functions refers to an ordered set of cumulative distribution functions (CDFs). Let us say $\{F_n\}$ is a sequence of distribution functions. For each n, $F_n(x)$gives the probability that the random variable associated with that distribution function is less than or equal to x. That is, for each n,

$$F_n(x) = P(X_n \leq x)$$

Where $\{X_n\}$ is the sequence of random variables associated with $\{F_n\}$.

Properties:

Since each F_n is a distribution function, it will have the standard properties of any *cdf*

- $0 \leq F_n \leq 1$
- F_n is non-decreasing.
- $\log_{x \to -\infty} F_n(x) = 0$, and $\log_{x \to \infty} F_n(x) = 1$
- F_n is right continuous, meaning for any point x, $\log_{y \to x} F_n(y) = F_n(x)$.

8.3.1 Supporting Lemmas and Results

Monotonicity: Distribution functions are non-decreasing. This means that as you move to higher values (like taller heights), the probability can't decrease. The probability of someone being less than 5 feet is obviously lower than the probability of someone being less than 6 feet.

Right Continuity: If you consider any specific height, and then think about heights just a tiny bit taller, the probability won't suddenly jump up. It might stay the same or increase slightly, but it won't make abrupt jumps.

Boundedness: Distribution functions are always between 0 and 1, inclusive. This means that the probability of a person having a height less than or equal to any value will always be between 0% and 100%.

In practice, these ideas about sequences of distribution functions can be very useful. For example, statisticians might use them to study how certain characteristics of populations change over time or across different contexts. Understanding whether a sequence of distribution functions converges, and in what way, can provide insights into underlying trends or patterns in the data.

8.3.2 Convergence

One of the central interests regarding a sequence of distribution functions is whether it converges to another distribution function. There are different modes of convergence, like weak convergence and complete convergence, which we discussed previously.

Convergence basically means that elements (or values) in a sequence get closer and closer to a particular value as we progress further in the sequence.

1. ***Simple Numerical Sequence:***

Consider the sequence: $\frac{1}{2}, \frac{3}{4}, \frac{5}{6}, \ldots$

As you can see, each term is just a bit less than 1 and gets closer and closer to 1 as we keep progressing. So, we say that this sequence converges to 1.

Simply it can understand as, it is like trying to reach a destination, and with each step, you are covering half the remaining distance. You are getting very close, but technically, you will never quite reach it. However, for all practical purposes, you are converging to that destination.

2. ***Real-world Example - Zooming In On a Picture*****:**

Imagine you are looking at a digital picture of a vast landscape on your computer, and you decide to zoom in on a specific tree. Each time you zoom in, the tree takes up a larger and larger portion of your screen. If you could zoom in infinitely, the entire screen would essentially be filled by that tree. So, as the number of zooms increases, the proportion of the screen filled by the tree converges to 100%.

3. ***Functional Convergence - Cooking on a Stove*****:**

Imagine you are heating a pot of water on a stove. At the start, the temperature of the water might be at room temperature, say 20°C. As time progresses and you continue heating, the temperature rises: 30°C, 40°C, 80°C, 90°C, and so on. If you are aiming to boil the water, it will tend to reach but never exceed 100°C (at standard atmospheric pressure). In this case, the temperature of the water is converging to 100°C.

4. ***Geometric Example - Drawing a Triangle Inside a Circle:***

Imagine a circle, and inside that circle, you draw a triangle (3 sides). Now, increase the number of sides: draw a square (4 sides), then a pentagon (5 sides), and so on. As you increase the number of sides indefinitely, the shape you draw starts looking more and more like the circle itself. In this sense, the perimeter of these shapes is converging to the circumference of the circle.

5. ***Convergence in Technology - Data Streaming*****:**

Imagine you are streaming a video online. At first, due to slower internet, the video might start in low resolution. But as the buffer builds up and the internet catches up, the video quality improves, moving to 480p, then 720p, and then to 1080p. Here, the video quality is converging to its highest available resolution as time progresses and the conditions (internet speed) improve.

In all these examples, the idea is the same: As you progress (whether through a sequence, over time, by zooming in, etc.), values or conditions tend to get closer and closer to a specific target or value, which we refer to as convergence.

8.3.3 Applications

Studying sequences of distribution functions has various applications:

Limit Theorems:

The convergence of sequences of distribution functions plays a role in probability's limit theorems, like the Central Limit Theorem or the Law of Large Numbers.

Asymptotic Analysis:

In statistics, studying the behavior of estimators as the sample size grows (asymptotic behavior) often involves looking at sequences of distribution functions.

Stochastic Processes**:**

When dealing with processes indexed by time or another parameter, sequences of distribution functions help describe the probabilistic behavior of the process.

Joint Distribution Functions:

When working with random vectors or multiple random variables simultaneously, sequences can be described in terms of joint distribution functions. For a bivariate case, $F_n(x, y)$ would represent the probability that the first random variable in the nth distribution is less than or equal to x and the second is less than or equal to y.

Examples: Consider a sequence of Bernoulli random variables X_n with parameter p_n, where p_n is a sequence of numbers in [0,1]. The distribution function $F_n(x)$ is simple: it jumps from 0 to $1-p_n$ at 0 and from $1-p$ n to 1 at 1. As n changes, if the sequence p_n changes (say it converges to some p), the sequence of distribution functions F_n will reflect that change.

In essence, a sequence of distribution functions provides a rigorous framework to study the probabilistic behavior of a sequence of random variables. Understanding this sequence's properties and behavior, especially its convergence, is pivotal in both theoretical and applied probability and statistics.

8.4 Weak and Complete Convergence of Sequence of Distribution Functions

The convergence of sequences of distribution functions is crucial in understanding the behaviour of random sequences and in inferential statistics. When discussing the convergence of distribution functions, two main types of convergence are typically highlighted: weak convergence and complete convergence. Let us delve into these concepts:

8.4.1 Weak Convergence (Convergence in Distribution)

In simple words, weak convergence (also known as convergence in distribution) refers to a sequence of random variables getting closer and closer in terms of their distributional behavior, even if they don't necessarily get close in terms of their actual values. Essentially, if you were to sketch the histograms or plots of these random variables, they would start looking more and more alike as you go further in the sequence, converging to the same shape or distribution in the limit.

Definition:

A sequence of distribution functions F_n is said to converge weakly (or in distribution) to a distribution function F if, for all points x at which F is continuous,

$$\log_{n\to\infty} F_n(x) = F(x)$$

This form of convergence is also represented as

$$F_n \overset{w}{\to} F \;\; or \;\; F_n \overset{D}{\to} F$$

Intuition:

The weak convergence can be visualized as the graph of F_n approaching the graph of F at all continuous points of F. It implies that the distributions represented by F_n become closer and closer to the distribution represented by F as n grows larger.

Significance:

This type of convergence is foundational for many central results in probability and statistics, like the Central Limit Theorem.

Examples:

Central Limit Theorem (CLT): This is a classic example of convergence in distribution. Imagine you're rolling a fair six-sided die. The average of a few rolls will not necessarily resemble any specific pattern. But if you were to roll the dice, say, 1000 times, and average the results, and then do this whole process repeatedly, the averages would start to cluster around a particular value (3.5 for a fair die). Furthermore, the way these averages spread around 3.5 would follow a normal (bell-shaped) distribution. The CLT says that as the number of rolls (or observations) increases, the distribution of the average tends to a normal distribution, regardless of the original distribution. So, the sequence of averages converges in distribution to a normal distribution.

Law of Large Numbers: Imagine you are flipping a fair coin. You are interested in the proportion of heads you get. At first, after a few flips, you might get sequences like H, T, H, T (50% heads) or H, H, T, H (75% heads). But if you flip the coin a very large number of times, the proportion of heads you observe will get closer and closer to 0.5. In terms of convergence in distribution, as the number of coin flips grows, the distribution of the proportion of heads you get converges to a single point at 0.5.

These examples illustrate how various sequences of random variables, under specific conditions, can have their behavior (in terms of probabilities) converge to that of another random variable or a specific point.

8.4.2 Complete Convergence

In simple words, weak convergence (also known as convergence in distribution) refers to a sequence of random variables getting closer and closer in terms of their distributional behavior, even if they don't necessarily get close in terms of their actual values. Essentially, if you were to sketch the histograms or plots of these random variables, they would start looking more and more alike as you go further in the sequence, converging to the same shape or distribution in the limit.

In straightforward terms, complete convergence is a concept that deals with how quickly and consistently a sequence of random variables approaches a limit.

A sequence of random variables is said to converge completely to a limit if, when you sum up the probabilities that the random variables deviate from the limit by more than some small amount, that total sum is finite. In other words, not only should the sequence converge, but the deviations from the limit should become rare enough, fast enough, to make this sum finite.

In a more intuitive sense, complete convergence is stronger than just convergence in probability but not as strong as almost sure convergence. It ensures that the random variables in the sequence get closer to the limit quickly enough for our mathematical comfort.

Definition: A sequence of distribution function F_n is said to converge completely to a distribution F, if for every $\epsilon > 0$

$$\log_{n\to\infty} P\left(sup_x |F_n(x) - F(x)| > \epsilon\right) = 0$$

Intuition: Unlike weak convergence, which requires the convergence to occur only at the continuous points of F, complete convergence mandates the convergence of F_n to F to be uniform over all real numbers. In other words, the graphs of F_n and F should become arbitrarily close over the entire real line as n increases.

Significance: Complete convergence is a stronger form of convergence than weak convergence. If a sequence of distribution functions converges completely to F, it also converges weakly to F, but the converse is not always true.

8.4.3 Supporting Results

Kolmogorov's Three-Series Theorem: A necessary and sufficient condition for a series of random variables to converge almost surely is given by their variances and covariances. If the individual and cross terms of these series converge, then the series itself will converge almost surely. This theorem provides a way to ensure that the complete convergence condition holds.

Examples*:*

Converging Sum of Random Variables: Consider a sequence of random variables Z_n where each Z_n is independently and identically distributed, and assume $E(Z_n^2) = 1$ (i.e., the expectation of Z_n^2 is 1). Now, consider a new sequence
$Y_n = \frac{Z_n}{2^n}$. The sum of Y_n, i.e.,
$S = \sum_{n=1}^{\infty} Y_n$, converges completely. Here, the decreasing factor of $\frac{1}{2^n}$ ensures that the series of probabilities decays quickly enough for complete convergence.

Borel-Cantelli Lemma: This is a foundational result that can be applied to understand complete convergence. Let us take an example. Consider flipping a fair coin. Let An be the event that we get heads on the nth flip. Now, let us consider a modified scenario. Instead of flipping the coin indefinitely, we will flip it 2^n times on the nth day. The probability of getting heads on all 2^n flips on the nth day is $\left(\frac{1}{2}\right)^{2^n}$. The sum of these probabilities over all days is: $\sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^{2^n}$.
This sum is finite. According to the Borel-Cantelli Lemma, the event that we get heads on all 2^n flips on the nth day will happen only finitely many times. This is an example where probabilities converge fast enough to ensure complete convergence in terms of events happening.

Complete convergence helps in scenarios where we do not just want to know that something will happen eventually, but we also want to ensure that the "not happening" becomes rarer at a specific rate.

Comparison:

While both forms of convergence describe how a sequence of distribution functions approach another, the requirements for complete convergence are more stringent. Weak convergence is concerned mainly with the behavior at continuous points of the limit function. In contrast, complete convergence requires a sort of "uniform closeness" between the converging functions and the limit function across the entire real line.

In many statistical applications, especially when dealing with large samples, weak convergence (or convergence in distribution) becomes the focus, primarily because of its connections to limit theorems and the asymptotic behavior of statistics. Complete convergence, while fundamental, is less commonly encountered in standard statistical practice.

8.5 Different Types of Convergence of Sequence of Random Variable

The concept of convergence plays a foundational role in this understanding. As we delve deeper into the subject, it becomes evident that there are multiple ways in which a sequence of random variables can approach a limit. Each mode of convergence captures a unique aspect of this limiting behavior, dictated by specific criteria and with its own set of implications. In this section, we shall explore various modes of convergence, including almost sure convergence, convergence in probability, weak convergence (or convergence in distribution), and convergence in mean.

8.5.1 Almost Sure Convergence or Convergence almost Everywhere

Definition:

A sequence of random variable $X_1, X_2, \dots$ is said to converge almost surely to a random variable X, if the probability that X_n converges to X as n approaches infinity is 1. This is denoted as:

$$X_n \xrightarrow{a.s.} X$$

Intuition:

For a fixed outcome in the sample space, the sequence $X_1(\omega), X_1(\omega), ...$is a real sequence, and this sequence converges to $X(\omega)$. This holds true for all outcomes ω except possibly for a set with probability zero.

8.5.2 Convergence in Probability

Definition:

A sequence of random variable $X_1, X_2, ...$ converges in probability to X if for every $\epsilon > 0$

$$\log_{n\to\infty} P[|X_n - X| > \epsilon] = 0$$

It is denoted as

$$X_n \overset{P}{\to} X$$

Intuition:

as n grows, the probability that X_n differs from X by more than a small amount ϵ becomes increasingly smaller.

8.5.2.1 Supporting Results

Characterization via Subsequences:

If X_n converges to X in probability, then for every subsequence X_{nk}, there exists a further subsequence X_{nk_l} such that X_{nk_l} converges to X almost surely.

Boundedness in Probability:

If X_n converges to X in probability and $|X_n$ |≤M almost surely for all n, then $|X|$≤M almost surely.

Continuous Mapping Theorem:

If X_n converges to X in probability and g is a continuous function, then $g(X_n)$ converges to $g(X)$ in probability.

Sum of Converging Sequences:

If X_n converges to X in probability and Y_n converges to Y in probability, then Xn+Yn converges to X+Y in probability.

Relationship with Other Convergences:

Convergence almost surely implies convergence in probability. Convergence in $\mathrm{L^p}$ (for p>0) implies convergence in probability

Examples:

Let X_n be a random variable such that X_n=$\begin{cases} n & with\,probability\;\frac{1}{n} \\ 0 & with\,probability\;1-\frac{1}{n} \end{cases}$

Does X_n converges to 0 in probability?

Solution: $P(|X_n - 0| > \epsilon) = P(X_n > \epsilon) = \begin{cases} \frac{1}{n} & if\ \epsilon < n \\ 0 & if\ \epsilon \geq n \end{cases}$

$for\ any\ fixed\ \epsilon > 0, as\ n\ \rightarrow \infty, P(X_n > \epsilon) = \frac{1}{n},$ which will tend to zero as n tends to infinity. Thus X_n converges to 0 in probability.

Example: let $Y_n = \frac{1}{n}$ for all natural numbers n. Does Y_n converge to 0 in probability?

Solution: Since Y_n is a deterministic sequence and nota sequence of random variables in the usual sense, Y_n converges to 0 almost surely, and hence also in probability.

8.5.3 Convergence in Mean (Convergence in $\mathrm{L^p}$)

In simpler terms, convergence in L^p means that the average size of the p-th power of the differences between the sequence and its limit gets closer and closer to zero. This mode of convergence is stronger than convergence in probability but weaker than almost sure convergence. It is particularly useful when considering the behavior of the moments (like variances) of random variables.

Definition:

A sequence of random variable $X_1, X_2, \ldots .\,convergence$ in mean of order p (L^p) to X if:

$$\log_{n\rightarrow\infty} E[|X_n - X|^p] = 0; for\ some\ p \geq 1$$

Intuition:

the expected value of the p^{th} power of the difference between X_n and X goes to zero as n approaches infinity. The most common case is $p = 2$ which is convergence in mean square.

8.5.3.1 Supporting Results

1. ***Cauchy Criterion for Convergence in L^p***

A sequence of random variable $X_1, X_2, \dots .convergence$ in L^p if and only if for every $\epsilon > 0, there\ exists\ N(\epsilon)$:

$$E[|X_n - X|^p] < \boldsymbol{\epsilon}$$

2. ***Relationship with Convergence in Probability***

If X_n converges in distribution to X in L^p for some in $p > 0$, then X_n also converges to X in probability.

3. ***Bounded Convergence in L^p:***

If $|X_n| \leq \boldsymbol{M}$ almost surely for all n, and X_n converges to X almost surely then X_n converges to X in L^p for in $1 \leq p < \infty$.

Example: Convergence in L^2

Consider $X_n = \begin{cases} n & with\ probability\ \frac{1}{n} \\ 0 & with\ probability\ 1 - \frac{1}{n} \end{cases}$

Check if X_n converges to 0 in mean squared.

Solution: compute $E[|X_n - 0|^2]$

$$E[X_n^2] = n^2 \frac{1}{n} + 0.\left(1 - \frac{1}{n}\right) = n$$

Here

$E[X_n^2] = n$, *which does not converge to 0 as n tends to infinity. Thus* X_n *does not* Convergence in L^2.

Example: Convergence in L^1

Consider $X_n = \begin{cases} 1 & with\ probability\ \frac{1}{n} \\ 0 & otherwise \end{cases}$

Check if X_n converges to 0 in $\mathbf{L^1}$**.**

Solution: compute $E[X_n - 0] = \frac{1}{n}$

which converge to 0 as n tends to infinity. Thus X_n Convergences in L^1.

"Convergence in mean" describes the behavior of random variables' expected values as they tend to a limit. It provides a strong form of convergence (especially for p>1) but is less commonly encountered in basic statistics than convergence in probability or almost sure convergence.

8.5.4 Convergence in Distribution or Weak Convergence

Definition:

A sequence of random variable $X_1, X_2, \ldots$ converges in distribution to X if the cumulative distribution function X_n denoted as , $F_n(x)$, converges to the cdf of X, $F(x)$ at all points x where $F(x)$ is continuous. This is denoted as:

$$X_n \overset{D}{\to} X$$

Intuition:

The distribution or the probabilistic behavior of $X_n$becomes closer to that of X as n increases. Note that this does not necessarily mean the random variables themselves are getting close.

Each type of convergence has its significance. For instance, convergence in probability is commonly used in the context of estimators in statistics, while convergence in distribution is the basis for the central limit theorem.

It is essential to understand that these modes of convergence are not equivalent. For instance, almost sure convergence implies convergence in probability, but the converse is not necessarily true.

8.5.4.1 Supporting Results and Theorems

1. ***Skorohod's Representation Theorem:***

This result asserts that if X_n converges in distribution to X, then there exist probability spaces and random variables Y_n and Y on these spaces such that:

a). The distribution of Y_n and Y are the same as those of X_n and X respectively.

b). Y_n converges almost surely to Y.

2. ***Portmanteau Lemma***:

A sequence of random variables X_n converges in distribution to X if and only if one of the following holds:

a). For every bounded continuous function g:

$$\log_{n\to\infty} E[g(X_n)] = E[g(X)]$$

b). for every open set U:

$$\log \inf_{n\to\infty} P(X_n \in U) = P(X \in U)$$

c). for every closed set G:

$$\log \sup_{n\to\infty} P(X_n \in G) = P(X \in G)$$

3. Continuous Mapping Theorem:

If X_n converges in distribution to X and g is a function continuous at points where $F_X(x)$, is continuous then $g(X_n)$ converges in distribution to $g(X)$,

4. Slutsky's Theorem:

If X_n converges in distribution to X and Y_n converges in probability to c, then:

a). $X_n + Y_n$ converges in distribution to $X + c$.

b). $X_n . Y_n$ converges in distribution to $X . c$.

5. Cramer-Wold Thorem:

For random vectors X_n converges in distribution to X if and only if for every t in Euclidean space. $t^{'}X_n$ converges in distribution to $t^{'}X$.

Proofs for these theorems can be quite involved, especially in a short format. They typically involve intricate uses of measure theory, tightness of measures, and properties of distribution functions. For a complete exposition of these proofs, I recommend referring to comprehensive probability theory books such as "Probability and Measure" by Patrick Billingsley or "Weak Convergence and Empirical Processes" by A. W. van der Vaart and Jon A. Wellner. These texts offer step-by-step detailed proofs for the above theorems and many more concepts related to weak convergence.

Example: Let X_n be a sequence of random variables such that X_n has the probability density function (pdf)

$$\begin{cases} n^2x & 0 \le x \le \frac{1}{n} \\ 0 & otherwise \end{cases}$$

Determine if X_n converges in distribution, and if so, to which random variable.

Solution:

Firstly, note that $f_n(x)$ is a valid pdf because:

$$\int_{-\infty}^{\infty} f_n(x)dx = \int_0^{\frac{1}{n}} n^2x\,dx = n^2\left(\frac{1}{2n^2} - 0\right) = 1$$

Given the pdf we can find the cumulative distribution function $F_n(x)$:

$$\begin{cases} 0 & x < 0 \\ \int_0^x n^2t\,dt & 0 \le x \le \frac{1}{n} \\ 1 & x > \frac{1}{n} \end{cases}$$

For $0 \le x \le \frac{1}{n}$,

$$F_n(x) = n^2\frac{x^2}{2}$$

So

$$\begin{matrix} 0 & x < 0 \\ n^2\frac{x^2}{2} & 0 \le x \le \frac{1}{n} \\ 1 & x > \frac{1}{n} \end{matrix}$$

Now let us find the limit:

$$\log_{n\to\infty} F_n(x)$$

For $x < 0, the\ limit\ is\ 0$,

For $x = 0, the\ limit\ is\ 0$, but for For $x > 0$, since $x > \frac{1}{n}$ for large n, the limit is 1. Therefore X_n converges in distribution to X that is 0 with probability 1 (a degenerate random variable at 0).

8.6 Distribution Function of Random Vectors

The distribution function (or cumulative distribution function, CDF) of a random vector is a generalization of the distribution function of a single random variable to multiple dimensions. It captures the joint behavior of multiple random variables simultaneously.

Generation of the distribution function of a single random variable to multiple dimensions. It captures the joint behavior of multiple random variables simultaneously.

For a random vector $X = (X_1, X_2, \dots., X_k)$ where $X_i's$ are random variables, the joint distribution function or joint *cdf* is defined as:

$$F_X(x_1, x_2, \dots., x_k) = P(X_1 \leq x_1, X_2 \leq x_2, \dots., X_k \leq x_k)$$

This function gives the probability that the random variable X_1 takes a value less than or equal to x_1 , the random variable X_2 takes a value less than or equal to x_2 and so on, simultaneously.

This joint distribution function characterizes the entire probabilistic behavior of the random vector X and can be used to deduce various properties, such as joint probabilities and marginal distributions, of the random variables involved.

Example 1: Two-dimensional random vector

Let X=(X,Y) be a random vector representing the height (X) and weight (Y) of individuals in a certain population. Suppose the joint CDF, F_X(x,y), gives the probability that a randomly chosen individual has height less than or equal to x cm and weight less than or equal to y kg.
If we want to know the probability that a person's height is less than 170 cm and weight is less than 65 kg, we evaluate: F_X (170,65).

Example 2: **Transformations of Random Variables**

Let's consider a random vector W=(X,Y) where X represents the age of a car and Y represents its price. If cars depreciate over time, the joint CDF might reflect that older cars (larger x values) tend to have lower prices.

So, if F_W (3,10,000) is 0.8, it means there's an 80% chance that a randomly selected car will be 3 years old or younger and priced at $10,000 or less.

Example 3: Consider the random vector X= (X1, X2), where X1 and X2 represent the outcomes of rolling two six-sided dice, respectively.

The joint CDF, F_X (x1, x2), gives the probability that the first die shows a number less than or equal to x1 and the second die shows a number less than or equal to x2.

Let us say we want to find the probability that the first die shows 3 or less and the second die shows 4 or less. We evaluate: F_X (3,4)

Considering each die has equal probability for each side: F_X (3,4) = 3/6 ×4/6 = 1/3

Thus, there is a 1/3 chance that the first die will show 3 or fewer and the second die will show 4 or fewer.

8.7 Self-Assessment Exercises

1 What is the fundamental idea behind convergence in probability theory?

2 Define the term "distribution function" in the context of random variables.

3 Explain what it means for a sequence of distribution functions Fn(x) to converge to F(x).

4 How does understanding the sequence of distribution functions aid in grasping the behavior of a sequence of random variables?

5 Distinguish between weak convergence and complete convergence of a sequence of distribution functions.

6 What is "almost sure convergence"? How does it differ from "convergence in probability"?

7 Describe "convergence in Lp norm". For what values of p is this definition relevant?

8 When do we say a sequence of random variables converges in distribution?

9 How is the distribution function of a random vector different from that of a single random variable?

10 Define the joint distribution function for a random vector X=(X1,X2,...,Xk).

11 Why is the concept of convergence crucial when considering the Central Limit Theorem or the Law of Large Numbers?

12 Given a sequence of random variables where you know they converge almost surely to a random variable X, what can you infer about their convergence in probability?

True or False Questions

13 Weak convergence of a sequence of distribution functions implies complete convergence.

14 Almost sure convergence of a sequence of random variables implies convergence in distribution.

Scenario-Based Question

15 Consider a casino game where the outcome of each game is a random variable. If a player keeps playing indefinitely, under what type of convergence would you expect the player's average winnings (or losses) to stabilize?

16 By answering these questions, learners will be able to gauge their understanding of the unit's key concepts and their ability to apply them in various contexts.

8.8 Summary

In this unit on convergence in the realm of probability and statistics, we delved into understanding how sequences, especially those of random variables, approach specific behaviors or values as they progress. Starting with the basic sequence of distribution functions, we explored how the probabilistic attributes of these sequences stabilize. The unit differentiated between various modes of convergence, namely weak and complete convergence for distribution functions, and almost sure convergence, convergence in probability, convergence in L^p norm, and convergence in distribution for sequences of random variables. This knowledge was further expanded to the multi-dimensional setting of random vectors, elucidating the intricate nature of their joint distribution functions. Through this comprehensive exploration, we gained insights into the fundamental behavior and essence of randomness and uncertainty as they manifest in increasing observations or trials.

8.9 References

1. Billingsley, P. (1995). Probability and Measure. John Wiley & Sons.

2. Durrett, R. (2019). Probability: Theory and Examples. Cambridge University Press.

3.. Grimmett, G., & Stirzaker, D. (2001). Probability and Random Processes. Oxford University Press.

4. Klenke, A. (2014). Probability Theory: A Comprehensive Course. Springer.

5. Lehmann, E. L., & Casella, G. (2006). Theory of Point Estimation. Springer.

6. Loève, M. (1977). Probability Theory I. Springer-Verlag.

7. Shiryaev, A. N. (1996). Probability. Springer.

8.10 Further Reading

- "Probability and Measure" by Patrick Billingsley, John Wiley & Sons.
- "A Probability Path" by Sidney I. Resnick, Birkhäuser.
- "Convergence of Probability Measures" by Patrick Billingsley, John Wiley & Sons.
- "Weak Convergence and Empirical Processes: With Applications to Statistics" by A. W. van der Vaart and Jon A. Wellner, Springer.
- "An Introduction to Probability Theory and Its Applications (Vol. 1)" by William Feller, John Wiley & Sons.
- "Probability: Theory and Examples" by Rick Durrett, Cambridge University Press.
- "Real Analysis and Probability" by R. M. Dudley, Cambridge University Press.
- "Foundations of Modern Probability" by Olav Kallenberg, Springer.
- "Limit Theorems for Stochastic Processes" by Jean Jacod and Albert N. Shiryaev, Springer.
- "Probability with Martingales" by David Williams, Cambridge University Press.

UNIT:9 LAW OF LARGE NUMBERS

Structure

9.1 Introduction

The Law of Large Numbers (LLN) is a pivotal theorem in probability theory, gracefully bridging the realms of theory and application, elucidating the deterministic pathways emerging amidst probabilistic chaos. Originating from the formative work of Jacob Bernoulli and later refined by numerous statisticians and mathematicians, the LLN provides profound insights into the long-term behavior of series of random experiments or processes. This unit unveils the layers of the LLN and explores its numerous manifestations and related principles. Beginning with an exploration of the essential concepts and theoretical foundations of the Weak and Strong Laws of Large Numbers, we plunge into the fascinating world where averages of random variables showcase predictable patterns as the sample size burgeons. We'll unfold Khinchin's theorem, offering a nuanced perspective on conditions for the weak law to hold for independent and identically distributed random variables. Then, navigating through the theoretical landscapes shaped by the Borel-zero-one law and Borel-Cantelli lemma, we shall explore the probability spaces where seemingly random events exhibit stark determinism upon infinite repetitions. Through this journey, we will not only decipher the logical beauty embedded in probabilistic theories but also comprehend their profound implications on practical statistical inferences and real-world phenomena, enabling us to gaze into the apparently chaotic realms of randomness with a lens of predictable stability.

9.2 Objectives

By the end of the Unit, learner will be able to:

- Understand the basic concept of the Law of Large Numbers.
- Differentiate between the Weak and Strong Law of Large Numbers.
- Grasp the essence of Khinchin's theorem.
- Understand the Borel-zero-one law and its implications.
- Familiarize oneself with the Borel-Cantelli lemma.

9.3 Weak Law of Large Numbers (WLLN)

This law states that the sample average converges in probability towards the expected value. Given a sequence of random variables with the same expected value, the weak LLN provides conditions under which the sample averages approach this expected value.

Statement:

The weak Law of Large Numbers states that as the size of a sample drawn from a population increase, the sample mean (or average) will get closer to the population mean. Formally, if $X_1, X_2, ...$ are independent and identically distributed (*i.i.d*) random variables with expected value $E[X_i] = \mu \, and \;\; variance \, Var[X_i] = \sigma^2 < \infty, then \, for \, any \, \epsilon > 0$;

$$P\left(\left|\frac{X_1, X_2, \dots, X_n}{n} - \mu\right| \geq \epsilon\right) \rightarrow 0$$

As $n \rightarrow \infty$.

Proof:

For simplicity, we will use Chebyshev's inequality to prove the WLLN.

Given that $X_1, X_2, \ldots$ are independent and identically distributed (*i.i.d*) random variables with expected value $E[X_i] = \mu \; and \;\; variance \; Var[X_i] = \sigma^2 < \infty$, let $S_n = X_1 + X_2 + \cdots + X_n$

The expected value of S_n is

$$E[S_n] = n\mu$$

and

$$V[S_n] = n\sigma^2$$

because the variables are independent. Now consider the sample mean $\overline{X_n} = \frac{S_n}{n}$

we are interested in

$$P(|\overline{X_n} - \mu| \geq \epsilon)$$

Using Chebyshev's inequality

$$P(|\overline{X_n} - \mu| \geq \epsilon) \leq \frac{Var(\overline{X_n})}{\epsilon^2}$$

Where $Var(\overline{X_n}) = \frac{\sigma^2}{n}$

Thus, the inequality will become

$$P(|\overline{X_n} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

As $n \to 0$,

$P(|\overline{X_n} - \mu| \geq \epsilon) \to 0.$

Example: Consider flipping a fair coin. Let us represent heads by 1 and tails by 0. If we were to flip the coin infinitely many times, the Weak Law of Large Numbers tells us that the average value will converge to the true mean of the distribution, which is 0.5 (since it's a fair coin).

So, if you were to flip the coin 10 times, you might get an average of 0.6 or 0.4 or some other value. But if you flip it 1,000 times, the average will be much closer to 0.5. And as you continue, say 10,000 or 100,000 times, the average will get even closer to 0.5, thanks to the WLLN.

Example: Let $\{X_n\}$ be a sequence of i.i.d. random variables, where each X_n represents the outcome of a fair die roll. Let us demonstrate that the sample mean converges to the expected value using WLLN.

Solution: each X_n takes values from 1 to 6 with equal probability of 1/6, thus

$$E[X_n] = \frac{1}{6}.1 + \frac{1}{6}.2 + \cdots + \frac{1}{6}.6 = 3.5$$

The variance is

$$Var[X_n] = E[X_n^2] - (E[X_n])^2$$

$E[X_n^2]$=$\frac{1}{6}.1^2 + \frac{1}{6}.2^2 + \cdots + \frac{1}{6}.6^2$ =15.1667

Thus, $Var[X_n] = 15.1667 - (3.5)^2 = 2.9167$

Now consider n rolls of the die and their sample mean

$$\overline{X_n} = \frac{X_1, X_2, \dots, X_n}{n}$$

Using Chebyshev's inequality:

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{Var[X_i]}{n\epsilon^2} = \frac{2.9167}{n\epsilon^2}$$

As n tends to infinity the right-hand side goes to 0 for any fixed ϵ, confirming the WLLN.

The WLLN basically tells us that as we take more and more samples from a distribution (with a finite mean), the average of these samples gets closer and closer to the true mean of the distribution. In other words, with a sufficiently large sample size, the sample mean becomes a reliable estimator of the population mean.

Illustrative Example:

Suppose you are flipping a fair coin, where "Heads" is represented by 1 and "Tails" by 0. The expected value (mean) of a single flip is μ=0.5.

Now, if you flip the coin, say, 5 times, you might get a sequence like: 1, 0, 0, 1, 1. The sample average in this case is 0.6. This is somewhat close to the true mean of 0.5, but not exactly.

However, if you flip the coin 10,000 times, the proportion of heads you observe will likely be much closer to 0.5. For instance, you might observe 5012 heads, which gives a sample average of 0.5012.

The WLLN assures that as the number of coin flips (or general samples from any distribution with a finite mean) increases, the sample mean converges in probability to the true mean.

9.3.1 Applications

The Weak Law of Large Numbers underpins much of classical statistics and is crucial for ensuring that sample averages (and other sample statistics) provide consistent estimates of population parameters. It's frequently invoked in scenarios where we want to estimate population quantities using sample data, such as in polling, quality control, and many other areas of research and industry.

9.4 Strong Law of Large Numbers (SLLN)

Unlike the weak LLN, the strong LLN talks about almost sure convergence. That is, the sample average not only converges in probability but also with a probability of one towards the expected value.

The Strong Law of Large Numbers provides a more robust statement than the Weak Law. It states that the sample averages converge almost surely to the expected value, i.e., the probability that the sample average converges to the expected value is one.

Formally, let $X_1, X_2, \ldots$ are independent and identically distributed (*i.i.d*) random variables with expected value $E[X_i] = \mu$. Let $\bar{X}_n = \frac{X_1+X_2+\cdots+X_n}{n}$, then

$$P(\log_{n\to\infty} \bar{X}_n = \mu) = 1$$

Proof:

To prove SLLN, one of the common methods uses the Borel-Cantelli lemma for the purpose of this proof, we shall assume $E(X_i^2) < \infty$ and use the fact that $Var[X_i] = E(X_i^2) - \mu^2$. First observe that for any positive ϵ:

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{Var[X_i]}{n\epsilon^2}$$

This comes from Chebyshev's -Inequality, now summing over all n:

$$\sum_{\boldsymbol{n=1}}^{\infty} P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{Var[X_i]}{\epsilon^2} \sum_{n=1}^{\infty} \frac{1}{n}$$

The series on the right is a harmonic series and diverges. However, since each term in the series on the left represents a probability of a particular event, the sum represents a count of the number of times the sample average deviates from the mean by at least ϵ.

Using Borel-Cantelli lemma, if $\sum_{\boldsymbol{n=1}}^{\infty} P(A_n) \infty$**,** where A_n are independent events. then $P(A_n i.o.) = 1$, where $i.o.$ stands for infinitely often. But this is a contradiction since the series above is an upper bound on the number of deviations of size ϵ, implying that the sample mean cannot deviate from the true mean by ϵ infinitely often.

Therefore,

$$P(\log_{n\to\infty} \bar{X}_n = \mu) = 1$$

Which completes the proof of the SLLN.

Remark: *The difference between the Weak and Strong Law of Large Numbers lies in the mode of convergence. While WLLN speaks about convergence in probability, SLLN asserts almost sure convergence. The Strong Law guarantees that the sample mean will converge to the true mean for almost every sample path, with exceptions being possible but occurring with probability zero.*

9.5 Khinchin's Theorem

Khinchin's theorem (or the Law of the Iterated Logarithm) is a central result in probability theory and is particularly important in the context of the theory of random walks and limit theorems. The theorem gives conditions under which the "oscillations" of the partial sums of independent and identically distributed (i.i.d.) random variables are of a specific magnitude. Named after the Russian mathematician Aleksandr Khinchin, this theorem provides insights into the "extreme" fluctuations of the cumulative sum of i.i.d. random variables, especially when normalized appropriately. Khinchin's theorem provides conditions under which the weak law of large numbers holds for independent and identically distributed random variables.

Statement:

Let (A_n) be a sequence of independently and identically distributed random variables, each with finite expected value $E[X_i] = \mu$. Let $\bar{X}_n = \frac{X_1+X_2+\cdots+X_n}{n}$ be the sample mean of the first n variables. Then

$$\log_{n\to\infty} P\left(\left|\overline{X_n} - \mu\right| \geq \epsilon\right) = 0\ ;\ for\ every\ \epsilon > 0$$

In other words, the sample mean $\overline{X_n}$ converges in probability to the expected value μ.

Proof:

Using the properties of variance of $\overline{X_n}$, and the fact that the$X_i's$ are independent:

$$Var[\bar{X}_n] = Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

Given the i.i.d assumption, $Var(X_i) = Var\left(X_j\right); for\ all\ i\ and$ j. if we call it common variance σ^2. Then

$$Var[\bar{X}_n] = \frac{1}{n^2} n\sigma^2 = \sigma^2/n$$

we can apply Chebyshev's Inequality

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{Var[\bar{X}_n]}{\epsilon^2}$$

Using the above results we can write it as :

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

Thus as $n \to \infty$,

$$\log_{n\to\infty} P\left(\left|\overline{X_n} - \mu\right| \geq \epsilon\right) = 0\ ;\ for\ every\ \epsilon > 0$$

This concludes the proof. The sample mean $\overline{X_n}$ of IID random variables with a finite expected value converges in probability to the true expected value μ.

9.5.1 Chebyshev's Inequality

It is the supporting lemma above theorem.

Statement:

For any random variable X with mean μ and variance σ^2, and for any k>0, the Chebyshev inequality is given by:

$$P(|X-\mu| > k\sigma) \leq \frac{1}{k^2}$$

This inequality provides an upper bound on the probability that the random variable deviates from its mean by more than a specified amount k.

Proof:

Consider the non-negative random variable $Z = (X-\mu)^2$, expected value of Z is:

$E(Z)$ =E$(X-\mu)^2 = \sigma^2$

Which is nothing but variance of X

Now we can say that,

$P(|X-\mu| > k\sigma) = P((X-\mu)^2 > k^2\sigma^2)$,

Using Markov's inequality which states that for any non-negative random variable Y, and a>0 ; $P(Y \geq a) \leq \frac{E[Y]}{a}$

Here we can say that $Y = (X-\mu)^2$ $and\ a = k^2\sigma^2$.

Putting these values in the Markov's inequality, we get

$$P((X-\mu)^2 \geq k^2\sigma^2) \leq \frac{E[(X-\mu)^2]}{k^2\sigma^2}$$

Or,

$$P((X-\mu)^2 \geq k^2\sigma^2) \leq \frac{\boldsymbol{\sigma^2}}{k^2\sigma^2}$$

Or,

$$P((X-\mu)^2 \geq k^2\sigma^2) \leq \frac{\mathbf{1}}{k^2}$$

Hence the result.

Chebyshev's inequality is a powerful tool because it applies to any random variable with a finite mean and variance, regardless of the distribution. It provides a general bound, which in many cases might not be tight, but it's a guarantee regardless of the specific distributional form.

Example: Suppose we have a sequence of i.i.d. random variables, where each X_n represents the number of heads observed in n coin flips of a fair coin. We want to show that the sample mean of the number of heads observed converges in probability to 0.5 using Khinchin's theorem.

Solution: Each X_n is a Bernoulli random variable that takes the value 1 with probability 0.5 (if heads) and 0 with probability 0.5 (if tails). Hence the expected value is:

$$E[X_n] = 0.5 \times 1 + 0.5 \times 0 = 0.5$$

The variance is :

$$V[X_n] = E\left[X_n^2\right] - \left(E[X_n]\right)^2 = 0.25$$

For the sample mean, we use Chebyshev's inequality:

$$P(|\bar{X}_n - 0.5| > \epsilon) \leq \frac{0.25}{n\epsilon^2}$$

Here as n become large, the right- hand side tends to 0 for any fixed ϵ. by Kinchin's theorem, this confirms that the sample mean $\bar{X}_n$ of the number of heads observed in n coin flips converges in probability to 0.5.

Intuitive Explanation:

The theorem provides a boundary for the "oscillations" or "fluctuations" of the partial sums Sn. While the Central Limit Theorem tells us that the normalized sums converge in distribution to a normal distribution, Khinchin's theorem gives us a sense of the magnitude of the maximum excursions or fluctuations of these sums from their expected value.

In simpler words, even though the average behavior of the random variables may conform to the Central Limit Theorem, their wild fluctuations are bounded by the $\sqrt{2n \log logn}$. term, a slower-growing function than $\sqrt{n}$, which is used in the Central Limit Theorem.

9.5.2 Applications

Khinchin's theorem finds applications in the study of random walks and in understanding the extreme behavior of stochastic processes. For example, when considering a simple random walk (like flipping a coin and moving left or right on the number line), this theorem describes the "worst-case" or most extreme fluctuations of the walk over time.

Khinchin's theorem provides a version of the Law of Large Numbers for IID random variables. It assures us that the sample mean will get arbitrarily close to the true mean as the sample size increases. This result is foundational in statistics and probability theory.

9.6 Borel-Zero-One Law

This law is a fundamental concept in probability, stating that for a sequence of independent events, the probability of infinitely many of them occurring is either zero or one.

Statement:

Let (X_n) be a sequence of independent random variables, and let $\mathcal{F}_n\, be\; \sigma - algebra$ generated by $X_1, X_2, \ldots, X_n$. Let T be the tail $\sigma - algebra$ defined as

$$T = \bigcap_{n=1}^{\infty} \mathcal{F}_n$$

Then any event in $A\; in\; T$ has either $P(A) = 0$ or $P(A) = 1$.

Proof:

First, observe that for each n, the the tail $\sigma - algebra\; T$ is independent of the $-algebra$ $\mathcal{F}_n$. this is because $T\; is\; a\; subset\; of\; every\, \mathcal{F}_n\, for\; k \geq n$, and the independence of the $X_i's$ensures the independence of $T\; and\, \mathcal{F}_n$.

Let $A\;\; be\; any\; event\; in\; T$. Then $A \cap B\; is\; in\; T$ for any event $B\; in\, \mathcal{F}_n$ because T is a $\sigma - algebra$. Since $A\; and\; B$ are independent, we have
$P(A \cap B) = P(A).P(B)$.

However, since A belongs to the tail $\sigma - algebra$. It is unaffected by removing the effect of finitely many random variables. Thus, $A = (A \cap B)$, therefore

$$P(A) = P(A)P(B)$$

Now either $P(A) = 0$ or $P(A) = 1$,

If none of them is true, then the above equation yields a contradiction, so this case is impossible. Now if $P(A) = 0,$ then the equation holds true and $P(A) = 1$ then also the equation holds true. Thus, for any event A in the tail $\sigma - algebra\; T$, its probability is either 0 or 1 which proves the Borel's zero one law.

In simple terms, for events that only concern the "tail behavior" of a sequence of independent random variables (i.e., events that are not affected by a finite number of the variables in the sequence), the probability of such events occurring is either 0% or 100%.

Intuitive Explanation:

The Borel-Zero-One Law sheds light on the "all or nothing" nature of certain types of events in infinite sequences. If the occurrence of an event does not depend on any fixed number of initial terms of the sequence but only on the behavior of the tail (the terms far out in the sequence), then this event is either "almost sure" to happen or "almost sure" not to happen.

Example:

One of the classic examples involves the tossing of a fair coin infinitely many times. Let's consider the event A where the coin shows up heads infinitely often.

Now, no matter how the first few tosses (or any finite number of initial tosses) turn out, it does not determine whether heads will show up infinitely often. This event's outcome depends on the entire infinite sequence's behavior.

According to the Borel-Zero-One Law, the probability of this event is either 0 or 1. In this specific case, the probability is actually 1, meaning that when flipping a fair coin infinitely many times, one will almost surely see heads an infinite number of times.

9.6.1 Applications

While the Borel-Zero-One Law might seem abstract, it has practical implications. It aids in understanding the long-term behavior of sequences of independent trials. The law plays a foundational role in ergodic theory, random walks, and other domains of probability and statistics.

In essence, the Borel-Zero-One Law underscores the deterministic nature of certain probabilistic events when looking at them over infinite sequences. Even in a world of randomness and chance, some events become certain (or impossible) given enough time and trials. The Borel's Zero-One Law does not specify whether a tail event has a probability of 0 or 1. It just asserts that it is one of those two values. The power of this theorem lies in the fact that it gives us a nontrivial conclusion about a broad class of events without requiring specific knowledge of the underlying distribution of the random variables.

9.7 Borel-Cantelli Lemma

There are two parts to the Borel-Cantelli Lemma:

Lemma part-1:

If the sum of the probabilities of a sequence of events is finite, then the probability that infinitely many of them occur is zero. i.e. Given a sequence (A_n), if $\sum_{n=1}^{\infty} P(A_n) < \infty$, then the probability that infinitely many of A_n occurs is zero.

Proof:

The proof uses the fact that for any non-negative series, if its sum is finite, then the series terms must go to zero,

Consider the sequence of partial sums:

$$B_m = \bigcup_{n=m}^{\infty} A_n$$

Thus, $P(B_m)$ is the probability that at least one of the events $A_m, A_{m+1}, \ldots$ occurs. Now, using subadditivity:

$$P(B_m) \leq \sum_{n=m}^{\infty} P(A_n)$$

Since the series $\sum_{n=m}^{\infty} P(A_n)$ is convergent, its tail $\sum_{n=m}^{\infty} P(A_n)$ must go to zero as $m \to \infty$. Therefore $P(B_m) \to 0$. Thus, the probability that any of the events $A_m, A_{m+1}, \ldots$ occurs goes to 0. Meaning, the probability that infinitely many occur is 0.

Lemma part-2:

For independent events, if the sum of their probabilities is infinite, then the probability that infinitely many of them occur is one i.e. if the events (A_n) are independent and$\sum_{n=1}^{\infty} P(A_n) = \infty$ then the probability that infinitely many A_n occur is 1.

Proof:

Let B_n be the event that A_k does not occur for all $k \geq n$. We want to show that $P(B_n) = 0$ for all *n.*

$$P(B_n) = P\left(\bigcap_{k=n}^{\infty} A_k^c\right)$$

$$P(B_n) = \prod_{k=n}^{\infty} (1 - P(A_k))$$

Now consider the product: $\prod_{k=1}^{n}(1 - P(A_k))$, taking logarithm of this product, we get $\sum_{k=1}^{n} log(1 - P(A_k))$. Using inequality

$\log(1 - x) \leq -x \; for \; 0 \leq x \leq 1$, we get

$$\sum_{k=n}^{n} log(1 - P(A_k)) \leq -\sum_{k=1}^{n} P(A_k) \;\; for \; 0 \leq x \leq 1$$

Since $\sum_{k=1}^{n} P(A_k) = \infty$, the right-hand side goes to negative infinity as n tends to infinity and so does the left hand side. This means the logarithm of the product goes to negative infinity, implying that the product itself goes to 0. Hence $P(B_n) = 0 \; for \; all \; n$.

***Note:** Borel's Zero-One Law and Borel-Cantelli Lemmas are distinct but both provide deep insights into the behavior of infinite sequences of events.*

Example: consider tossing a fair coin. Let A_n be an event that you get heads on nth toss. Since the coin is fair, $P(A_n) = \frac{1}{2}$. Let us create a new sequence of events, B_n, where B_n is the event that you get heads on the $2^n th$ toss. Then $P(B_n) = \frac{1}{2^n}$.

The sum of the probabilities is:

$$\sum_{\boldsymbol{n=1}}^{\infty} P(B_n) = \sum_{\boldsymbol{n=1}}^{\infty} \frac{1}{2^n} < \infty$$

By Borel-Cantelli Lemma, the event B_n will occur only infinitely many times almost surely.

Example: consider throwing a dart at the interval [0,1] uniformly. Let A_n be an event that the dart lands in the interval $\left[0, \frac{1}{n}\right]$. Then $P(A_n) = \frac{1}{n}$

The sum of probabilities is:

$$\sum_{\boldsymbol{n=1}}^{\infty} P(A_n) = \sum_{\boldsymbol{n=1}}^{\infty} \frac{1}{n} = \infty$$

By Borel Cantelli Lemma, given that the events are independent, the dart will land in the interval $\left[0, \frac{1}{n}\right]$ infinitely often, almost surely. these examples demonstrate how the Borel-Cantelli Lemma can be applied in different contexts, using both first and second parts of lemma.

9.8 Summary

In this unit titled "Law of Large Numbers," we delve into the foundational principles governing the outcomes of repetitive experiments. Beginning with an introduction to the essence of the Law of Large Numbers, the unit differentiates between its two main forms: the Weak and Strong Law. While the Weak Law states that the sample average converges in probability towards the expected value, the Strong Law emphasizes almost sure convergence. Khinchin's theorem provides the conditions for the Weak Law to hold for independent, identically distributed variables. The unit further introduces the Borel-zero-one law, which posits that for a sequence of independent events, their infinite occurrence probability is strictly binary - either zero or one. Building on this, the Borel-Cantelli Lemma, divided into two parts, dictates the conditions under which sequences of events occur with a probability of zero or one. The unit, rich with theoretical insights, also encourages learners to delve deeper into these concepts through self-assessment and additional readings.

9.9 Self-Assessment Exercises

1. Differentiate between the weak and strong laws of large numbers.
2. What are the implications of Khinchin's theorem?
3. Under what conditions does the Borel-Cantelli lemma apply?
4. What is the primary distinction between the Weak and Strong Law of Large Numbers?
5. State, in your own words, the main premise of the Law of Large Numbers.
6. Under what conditions does the sample average converge towards the expected value according to the Weak Law of Large Numbers?
7. Explain the term "almost sure convergence" in the context of the Strong Law of Large Numbers.
8. Summarize the significance of Khinchin's theorem in relation to the Weak Law of Large Numbers.
9. Describe the Borel-zero-one law. How does it relate to sequences of independent events?
10. What are the two key parts of the Borel-Cantelli Lemma?
11. Under the Borel-Cantelli Lemma, when does the probability of infinitely many events occurring become zero?

12. If a sequence of events is independent and the sum of their probabilities is infinite, what can be inferred about their occurrence based on the Borel-Cantelli Lemma?
13. Which of the following books would provide further reading on the Law of Large Numbers: a) Convergence of Probability Measures by Patrick Billingsley, b) The Art of Probability by Rick Durrett, or c) A First Look at Rigorous Probability Theory by Jeffrey S. Rosenthal?

9.10 References

- Billingsley, P. (1995). Probability and Measure. John Wiley & Sons.
- Breiman, L. (1992). Probability. SIAM.
- Chow, Y. S., & Teicher, H. (1988). Probability Theory: Independence, Interchangeability, Martingales. Springer.
- Durrett, R. (2010). Probability: Theory and Examples. Cambridge University Press.
- Feller, W. (1968). An Introduction to Probability Theory and Its Applications, Vol. 1. John Wiley & Sons.
- Grimmett, G., & Stirzaker, D. (2001). Probability and Random Processes. Oxford University Press.
- Khinchin, A. (1943). Mathematical Foundations of Statistical Mechanics. Dover Publications.
- Kolmogorov, A. N. (1950). Foundations of the Theory of Probability. Chelsea Publishing Co.
- Ross, S. (2014). A First Course in Probability. Pearson.
- Shiryaev, A. N. (1996). Probability. Springer.

9.11 Further Reading

Here is the list of recommended books on the Law of Large Numbers, along with their respective publishers:

- "An Introduction to Probability Theory and Its Applications, Vol. 1" by William Feller, John Wiley & Sons.

- "Probability and Measure" by Patrick Billingsley, John Wiley & Sons.
- "A First Course in Probability" by Sheldon Ross, Pearson.
- "The Law of Large Numbers: Probability and Statistics and Their Applications" by M. Loève, Chelsea Publishing.
- "The Doctrine of Chances: Probabilistic Aspects of Gambling" by Stewart N. Ethier, Springer.

These books, along with their publishers, should give you a good starting point in your exploration of the Law of Large Numbers and its various implications and applications in probability and statistics.

UNIT: 10 CHARACTERISTICS FUNCTIONS

Structure

10.1 Introduction

The realm of probability theory is vast, and its applications are seen across various fields, such as statistics, finance, and engineering. Characteristic functions provide an essential tool in this field, offering a bridge between the probability and frequency domain. In this unit, we delve deep into understanding the foundation of characteristic functions and associated theorems.

Probability theory, an intricate tapestry of mathematical concepts and principles, enables us to understand and model uncertainty in various domains. From predicting stock market fluctuations to modeling the diffusion of particles in a liquid, its applications are vast and profound. Among the numerous tools and concepts in probability theory, characteristic functions stand out as particularly powerful. They provide a unique approach to understanding random variables and their distributions, primarily by mapping them from the real line into a complex plane. Such a transformation, akin to the Fourier Transform in signal processing, allows for many problems in probability to be handled more elegantly and often more simply.

As we journey through this unit, we will encounter the foundational Helly-Bray Lemma, highlighting the intimate connection between probability measures and characteristic functions. The Kolmogorov Theorem, on the other hand, provides a perspective on the convergence of random variables. This is especially crucial in the context of the Law of Large Numbers and the Central Limit Theorem, cornerstones of statistical theory.

Characteristic functions aren't just theoretical constructs; they have real-world implications. By understanding how to invert these functions through the Inversion Theorem, we can navigate from the frequency to the probability domain, unlocking a plethora of applications. Furthermore, the Continuity and Uniqueness Theorems add depth to our understanding, emphasizing the robustness of the links between distributions and their characteristic functions.

Prepare yourself for an engaging exploration, as this unit dives deep into the world of characteristic functions, unveiling their mysteries, and elucidating their pivotal role in probability theory. Whether you're a seasoned statistician or a curious learner, the insights from this unit will undoubtedly enrich your appreciation for the elegance and power of mathematical reasoning in deciphering the unpredictable nature of randomness.

10.2 Objectives

By the end of this unit, learners should be able to:

- Understand the foundational concept behind Helly-Bray Lemma.
- Grasp the significance and applications of the Kolmogorov Theorem.
- Describe and interpret the characteristic function.
- Apply the Inversion Theorem in probability scenarios.
- Understand the basis and implications of the Continuity and Uniqueness Theorems.
- Evaluate problems using the knowledge from this unit.

10.3 Helly-Bray Lemma

The Helly-Bray Lemma states that if two probability measures have the same characteristic functions, then they also have the same distribution functions. It sets the foundation for understanding the relationship between distributions and their characteristic functions.

Statement:

If two probability measures P and Q on R have the same characteristic functions, then they are identical.

Proof:

Let μ and ν be the probability measures corresponding to the random variables X and Y, such that their characteristic functions $\emptyset_{\mu}(t)\ and\ \emptyset_{v}(t)$ are identical.

Approach using the Lévy Continuity Theorem:

It's known by the Lévy Continuity Theorem that a sequence of characteristic functions converges to another characteristic function if and only if their corresponding sequence of probability measures converges weakly to the measure associated with the limiting characteristic function.

Identity in the Distribution:

If two characteristic functions are identical, it can be seen as a special case of convergence where the sequence is constant. That is, if for every t, $\emptyset_{\mu}(t) = \emptyset_{v}(t)$, then one can imagine a sequence of measures, all identical to v, converging weakly to μ. But weak convergence to a unique limit is a property of probability measures.

Since the characteristic functions are identical, and given the properties of weak convergence, it must be that μ and ν are identical.

This is a proof in broad strokes. To fill in the details would require a more extensive exposition of the weak- topology on the space of measures, the concept of weak convergence, and properties of the Fourier transform on the space of tempered distributions.

Illustrative Examples:

Bird Migration Patterns:

Suppose you are studying the migration patterns of a certain species of bird over the years. Every year, the birds seem to be favoring a new region slightly different from the last year. If these yearly patterns converge weakly to a particular pattern, then the Helly-Bray Lemma ensures that any continuous measurement you make on these patterns (like average distance traveled, assuming it's bounded) will also converge over the years.

Taste Preferences in Coffee Shops:

Imagine you own a chain of coffee shops, and over time, you notice that the preferences of customers for different types of coffee are changing. Every year, more people prefer lattes over cappuccinos, for instance. If you model this yearly preference change as a sequence of probability distributions that converges weakly to a certain distribution, then any average measure you compute on these distributions (like the average sweetness preferred, given it's bounded and continuous) will also converge.

Distribution of Books in Libraries:

Consider various public libraries and their collections of books across genres. Over decades, suppose the proportion of genres in new libraries is approaching a specific distribution (like 30% fiction, 20% science, 50% history). If this trend represents weak convergence to a particular distribution, then any continuous bounded measure (like average page count of books, if it has an upper limit) will converge across libraries as well.

Color Patterns in Art:

Let us say there is a sequence of art movements where the color palette used by artists is evolving. First, artists favored blues and greens; then, they moved to reds and yellows, and so on. If the distribution of colors used in artworks across these movements converges weakly to a certain distribution, the Helly-Bray Lemma ensures that any bounded continuous measure we compute on these color patterns (like average brightness or hue, given they're bounded) will also converge across the movements.

The real power of the lemma is in its application to sequences of probability measures and how they interact with bounded continuous functions.

10.4 Kolmogorov Theorem

This theorem deals with the convergence of random variables. It states that for a sequence of independent and identically distributed random variables, their sum, when normalized, converges in distribution to the standard normal distribution under certain conditions.

Statement:

Let (X_n) be a sequence of random variables and let X be another random variable. If the following two conditions hold:

The sequence (X_n) is such that, for every ϵ >0, there exists a compact set K_ϵ such that

$$\mathrm{P}(X_n \in K_\epsilon) > 1 - \epsilon \text{ for all } n.$$

The characteristic functions $\emptyset_{X_n}(t)$ converge pointwise to $\emptyset_X(t)$for every t ∈R. Then, the sequence of distribution functions F_{X_n} converges weakly to F_X.

Proof: **Convergence in Distribution:**

By the definition of weak convergence, for any bounded and continuous function g, we have:

$$\int g(x)dF_{X_n}(x) \rightarrow \int g(x)dF_X(x)$$

Characteristic Functions and Fourier Transform:

Given any function g with compact support and is differentiable, the Fourier inversion formula tells us that:

$$g(x) = \frac{1}{2\pi}\int e^{-itx}\hat{g}\,(t)dt$$

Where $\hat{g}$ is the Fourier transform of g.

Due to the tightness condition, there exists a compact set K such that X_n and X are both majorly contained in K. This allows us to multiply g by an indicator function I_K which equals 1 on K and 0 elsewhere, without significantly altering the integrals. The boundedness of g and the compactness of K ensure the convergence of the integral.

Using Pointwise Convergence of Characteristic Functions:

Expanding the integral of g against the distribution of X_n using the Fourier inversion formula and using the properties of the Fourier transform, we can relate this to the characteristic functions of X_n and X. The pointwise convergence of the characteristic functions guarantees the convergence of these integrals.

Using the properties of weak convergence, we deduce that if the integrals converge for a dense set of functions g, they converge for all bounded, continuous functions. Hence, the distributions F_{X_n} converge weakly to F_X.

The theorem guarantees that, under certain conditions, there is a probability space on which these random variables can be defined such that they have the given finite-dimensional distributions.

To phrase it simply, Kolmogorov's existence theorem tells us that if we know how all the finite collections of random variables behave (in terms of their joint distributions), then there exists a consistent probabilistic structure where all these random variables can "live" together.

10.4.1 Kolmogorov's Consistency Theorem (or Existence Theorem)

Given a collection of consistent finite-dimensional distributions, there exists a probability space and a collection of random variables on this space such that the finite-dimensional distributions of these random variables are precisely the given distributions.

To understand this, consider the following illustrative examples:

Weather Modeling:

Imagine you are trying to simulate the weather for an entire year, day by day. You have models that can simulate:

The weather for any individual day.

The joint weather for any pair of days.

The joint weather for any trio of days.

And so on...

If all these models are consistent (e.g., the two-day model agrees with the one-day model for each of its days), then Kolmogorov's theorem assures you that there exists a model that can simulate the weather for the entire year in a way that is consistent with all of your smaller models.

Stock Market Movements:

Suppose you have statistical models that describe how a set of stocks move:

On any individual day.

Jointly over any pair of days.

Jointly over any trio of days.

And so on...

Again, if all these models are consistent with each other, then there is a "global" model that can simulate the stock market movements over a long period in a way that is consistent with all the smaller models.

The theorem is of fundamental importance in probability theory, especially when constructing complex stochastic processes from simpler pieces. While the examples provided offer an intuitive grasp, the actual theorem is deeply mathematical and requires a comprehensive understanding of measure theory to fully appreciate.

10.5 Characteristic Function

A characteristic function provides a complex-valued function that characterizes the probability distribution of a random variable. It is particularly useful in deriving distributional properties and the sums of independent random variables.

The characteristic function of random variable X is defined as expectation of e^{itX}, where i is imaginary unit and t is a real number. Formally, for a random variable X with probability density function $f(x)$, the characteristic function $\emptyset_X(t)$ is given as

$$\emptyset_X(t) = E[e^{itX}]$$

For discrete random variables

$$\emptyset_X(t) = E\left[e^{itX}\right] = \sum e^{itX} P(X = x)$$

And, for continuous random variable

$$\emptyset_X(t) = E\left[e^{itX}\right] = \int e^{itX} f(x) dx$$

The characteristic function always exists and provides an alternative way to describe the distribution of a random variable. One of its primary advantages is that the characteristic function of the sum of independent random variables is the product of their individual characteristic functions.

Examples:

1. **Bernoulli Distribution:**

   Consider a Bernoulli random variable X taking values 0 and 1 with probability $p \; and \; (1-p)$ respectively. The characteristic function is:

$$\emptyset_X(t) = (1-p) + pe^{it}$$

2. **Uniform distribution on [0,1]:**

   If X is uniformly distributed over [0,1] then its probability density function is $f(x) = 1 \; for \; x \in [0,1]$, the characteristic function is:

$$\emptyset_X(t) = \int_0^1 e^{itX}.1.dx = \frac{e^{it}-1}{it}$$

3. **Standard Normal Distribution:**

   For a standard normal random variable Z with mean zero and variance 1, the characteristic function is

$$\emptyset_X(t) = e^{-t^2/_2}$$

The characteristic function is a powerful tool in probability theory and statistics. It is especially useful in problems of summing independent random variables, as the characteristic function of the sum is simply the product of the characteristic functions.

The characteristic function always exists for any random variable and provides a unique representation of its probability distribution. Two random variables with the same distribution will have the same characteristic function.

Uniqueness: The characteristic function uniquely determines the distribution of a random variable. If two random variables have the same characteristic function, they have the same distribution.

Inversion Formula: Given the characteristic function, it's possible (under certain conditions) to recover the distribution of the random variable.

Moments: The moments of the random variable (like mean, variance, etc.) can be derived from its characteristic function.

Illustrative Examples:

Musical Analogy: Imagine the distribution of a random variable as a musical tune. The characteristic function is like the set of frequencies that, when combined, produce that tune. Just as you can recreate a tune by combining its constituent frequencies, you can recreate a probability distribution using its characteristic function.

Light Analogy: Think of the distribution of a random variable as a pattern of light and shadow on a wall. If this pattern is produced by shining light through a complex stencil, then the characteristic function is akin to the stencil itself. It's a tool that encodes all the information about the light and shadow pattern (i.e., the distribution).

Simplified Example:

Let us discuss the case of a coin flip without going into heavy math. Suppose you have a fair coin, and you assign the value +1 for heads and -1 for tails. The distribution of this random assignment (either +1 or -1) has a particular shape or behavior. The characteristic function for this distribution captures that behavior in a complex exponential format. So, if you know the characteristic function of this coin flip scenario, you can deduce features about the original distribution, like the probability of getting heads or tails.

In more technical applications, the characteristic function plays a vital role in simplifying the math, especially when dealing with sums of independent random variables. The product of their individual characteristic functions gives the characteristic function of their sum, which can be a very handy property.

10.6 Inversion Theorem

The inversion theorem, in the context of characteristic functions, provides a method to recover the distribution of a random variable from its characteristic function. It's a cornerstone in the theory of characteristic functions. The Inversion Theorem is pivotal in probability theory. It offers a way to retrieve the probability distribution of a random variable from its characteristic function.

Statement:

Given a characteristic function $\emptyset_X(t)$ of a random variable X with cumulative distribution function $F(x), if\ F(x) is\ continuous\ at\ x_0$, then

$$F(x_0) = \frac{1}{2\pi} \log_{T\to\infty} \int_{-T}^{T} \frac{e^{itx_0}\emptyset_X(t)}{it} dt$$

Proof:

For ant real number x using the definition of the characteristic function:

$$\emptyset_X(t) = \int e^{itx}\, dF(x)$$

Multiplying both sides by $\frac{e^{-itx_0}}{it}$ and integrating it with respect to t from -T to T, we get

$$\int_{-T}^{T} \frac{e^{-itx_0}\emptyset_X(t)}{it} dt = \int_{-T}^{T} \frac{e^{-itx_0}}{it} dt \int e^{itx}\, dF(x)$$

$$= \int \left(\int_{-T}^{T} \frac{e^{-it(x_0-x)}}{it} dt \right) dF(x)$$

The function $\frac{e^{-itx_0}}{it}$ is problematic at t=0. However, one can show that:

$$\left| \int_{-T}^{T} \frac{e^{-it(x_0-x)}}{it} dt \right| \le \pi;\ \forall\, x$$

Using standard techniques from complex analysis.

One of the key insights is that. As T goes to infinity, the inner integral in the convolution closely resembles the indicator function of the interval $\left[-\frac{1}{2}, \frac{1}{2}\right]$. Specifically one can show

$$\log_{T\to\infty} \int_{-T}^{T} \frac{e^{-it(x_0-x)}}{it} dt = \pi I_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x_0 - x)$$

Where I_A is the indicator function of set A.

Given the approximating property above, we can substitute it back into our original equation to get

$$\int \pi I_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x_0 - x) f F(x) = \pi \left[F\left(x_0 + \frac{1}{2}\right) - F(x_0 - \frac{1}{2}) \right]$$

Because $F(x)$ is continuous at x0, the value $F\left(x_0 + \frac{1}{2}\right) and\ F(x_0 - \frac{1}{2})$ both converges to $F(x_0)$ as the interval shrinks to zero.

Thus, combining all the above equations, we get the desired result:

$$F(x_0) = \frac{1}{2\pi}\log_{T\to\infty}\int_{-T}^{T}\frac{e^{itx_0}\emptyset_X(t)}{it}dt$$

The specific mathematical details involve complex analysis and integrals, but for a qualitative understanding, consider the following analogy:

Analogy:

Imagine you are given a puzzle, but instead of the typical jigsaw pieces, you are provided with descriptions of how each piece connects to its neighbors. The Inversion Theorem is like a guideline that allows you to use these descriptions to reconstruct the original image of the puzzle.

The Inversion Theorem is crucial for several reasons:

Uniqueness: It reinforces the fact that the characteristic function uniquely determines the distribution. If two random variables have different distributions, they will have different characteristic functions.

Practical Computation: In some scenarios, it is easier to work with characteristic functions (especially when dealing with sums or products of independent random variables). Once you have done the necessary operations in the "characteristic function domain", you can use the Inversion Theorem to return to the "probability distribution domain".

Theoretical Foundation: It serves as a foundational result in probability theory and statistics, establishing a deep link between the probabilistic properties of a random variable and its characteristic function.

Simplified Example:

Suppose you have been studying a random phenomenon (like the height of individuals in a population) and have derived its characteristic function. Now, you wish to know the probability that the height falls within a certain range. Instead of directly dealing with the raw data, you can utilize the characteristic function and the Inversion Theorem to deduce this probability.

Remember, while the above explanations aim to provide an intuitive grasp, the Inversion Theorem is rooted in complex mathematical formulations. But at its heart, it is a bridge between the world of characteristic functions and the probability distributions they represent.

10.7 Continuity Theorem

This theorem provides conditions under which convergence in distribution implies convergence of characteristic functions. It's instrumental in ensuring that transformations in the frequency domain have valid representations in the probability domain.

Statement:

Let $X_n \, and \, X$ be random variables with characteristic functions $\emptyset_{X_n}(t)$ and $\emptyset_X(t)$ respectively. If $\emptyset_{X_n}(t)$ converges to $\emptyset_X(t)$ pointwise as $n \to \infty \, for \, every \, t$ and if $\emptyset_X$ is continuous at $t = 0$, then the distribution of $X_n \, converges \, weakly \, to \, the \, distribution \, of \, X.$

Proof:

For any bounded, continuous function $g: \mathbb{R} \to \mathbb{R}$, we have

$$\int g(x) dF_{X_n}(x) \to \int g(x) dF_X(x) \, ; as \, n \to \infty$$

Where $F_{X_n} \, and \, F_X$ are cumulative distribution function of $X_n \, and \, X$ respectively.

Recall that the set of bounded, continuous functions is dense in the space of bounded, measurable functions with respect to the L^1 norm. So, any such function can be approximated arbitrarily closely by bounded, continuous functions.

Given that $\emptyset_{X_n}(t)$ converges to $\emptyset_X(t)$ pointwise as $n \to \infty$ and the continuous property of $\emptyset_X(t)$ at t=0, it can be shown, through the properties of the Fourier transformation, that for every bounded, continuous function g,

$$\int g(x) dF_{X_n}(x) \to \int g(x) dF_X(x) \, ; as \, n \to \infty$$

Also, by the definition of weak convergence which states that, for every bounded, continuous function g,

$$\int g(x) dF_{X_n}(x) \to \int g(x) dF_X(x) \, ; as \, n \to \infty$$

Given that this holds from the properties of characteristic functions, as shown in previous step, the distribution F_{X_n} converge weakly to F_X.

The proof, in essence, is about using the characteristics of the Fourier transform and its relationship to the underlying distribution.

Importance of the theorem:

Tool for Proving Convergence: Many problems in probability and statistics involve showing that a sequence of random variables (or their distributions) converges to something. The Continuity Theorem provides a mechanism for this. If we can prove that the characteristic functions converge, it can lead us to conclusions about the convergence of the random variables themselves.

Connection between Pointwise Convergence and Distribution Convergence: It offers a bridge between the pointwise convergence of characteristic functions and the convergence in distribution of random variables.

Illustrative Example:

Imagine you are watching a series of movies from a particular director. Each movie has its own "mood" or "tone," which we can think of as its distribution. Now, if someone told you that the "essence" or "core theme" (analogous to the characteristic function) of each movie is getting more and more like that of a classic film, the Continuity Theorem ensures that the mood or tone of the director's movies is also getting closer and closer to that of the classic film, under the right conditions.

This is, of course, a very abstract way to think about it. The actual theorem operates in the realm of complex functions and probability distributions. However, the essence is that by observing how the characteristic functions behave, we can deduce how the actual random processes or variables behave.

10.8 Uniqueness Theorem

The Uniqueness Theorem strengthens the link between characteristic functions and probability distributions. It essentially states that a probability distribution is uniquely determined by its characteristic function.

Statement:

If X and Y are random variables with characteristic function $\emptyset_X(t)$ and $\emptyset_Y(t)$. If $\emptyset_X(t) = \emptyset_Y(t)$; for all t in an interval containing 0, then X and Y have the same distribution.

Proof:

The main idea behind the proof is to use the Inversion Theorem to show that the two random variables have the same distribution function.

Given that $\emptyset_X(t) = \emptyset_Y(t)$ for all t , we can consider the difference between the two characteristic functions:

$$\Delta(t) = \emptyset_X(t) - \emptyset_Y(t) = 0; for\ all\ t.$$

Using the Inversion Theorem, if $F_X(x)$ and $F_Y(x)$ are cumulative distribution functions of X and Y respectively, then for any $a < b$

$$F_X(b) - F_X(a) = \frac{1}{2\pi}\log_{T\to\infty}\int_{-T}^{T}\frac{e^{-itb} - e^{-ita}}{it}\emptyset_X(t)dt$$

$$F_Y(b) - F_Y(a) = \frac{1}{2\pi}\log_{T\to\infty}\int_{-T}^{T}\frac{e^{-itb} - e^{-ita}}{it}\emptyset_Y(t)dt$$

Since, given that $\emptyset_X(t) = \emptyset_Y(t)$ for all t and using the expressions from the Inversion Theorem, we can deduce that for all $a < b$

$F_X(b) - F_X(a)$ =$F_Y(b) - F_Y(a)$

This means the two cumulative distribution functions are equal, and thus , the distributions of X and Y are the same.

The proof is essentially a direct consequence of the Inversion Theorem. It shows the powerful connection between the characteristic function of a random variable and its distribution. If two random variables have the same characteristic function over an interval containing 0, they must have the same distribution.

This theorem emphasizes that there is a one-to-one correspondence between the probability distribution of a random variable and its characteristic function.

Importance:

One-to-One Correspondence: It confirms that the characteristic function contains all the information about the distribution of a random variable. If two random variables have the same characteristic function, they are essentially "the same" in terms of their distributions.

Theoretical Clarity: The theorem clarifies the relationship between distribution functions and characteristic functions, two fundamental constructs in probability theory.

Practical Implications: In some contexts, it's easier to work with characteristic functions. Knowing that the characteristic function uniquely determines the distribution means that if you can manipulate or understand the characteristic function, you have a complete understanding of the distribution.

Illustrative Example:

Imagine every person has a unique fingerprint that represents their identity. If two people have the same fingerprint, then they are the same individual (ignoring the idea of identical twins for this analogy). Similarly, the characteristic function is like the "fingerprint" of a probability distribution. If two distributions have the same "fingerprint" (characteristic function), then they are the same distribution.

In a more practical scenario, suppose you are studying two seemingly different phenomena, but upon analysis, you find their characteristic functions to be the same. The Uniqueness Theorem tells you that the underlying probability distributions governing these phenomena are identical.

In essence, the Uniqueness Theorem reinforces the idea that the characteristic function captures all the information about a random variable's distribution. If you know the characteristic function, you know the distribution, and vice versa.

10.9 Summary

In the unit "Characteristic Functions," we delved into the world of probability distributions and their corresponding characteristic functions, which provide a Fourier transform representation of a random variable. Starting with an introduction to the underlying concepts, the unit unveiled key theorems and their implications: the Helly-Bray Lemma, which speaks to the convergence of characteristic functions; the Kolmogorov Theorem, establishing conditions for convergence; and the pivotal Inversion and Continuity Theorems, which bridge the relationship between characteristic functions and the probability distributions they represent. The unit also highlighted the Uniqueness Theorem, asserting that a random variable's distribution is uniquely determined by its characteristic function. Throughout, the material emphasized the crucial role of characteristic functions in understanding and analyzing the properties and behavior of random variables and their distributions.

10.10 Self-Assessment Exercises

1. Define the Helly-Bray Lemma in your own words.
2. Explain the importance of the Kolmogorov Theorem.
3. What is the relationship between the characteristic function and the distribution of a random variable?
4. How does the Inversion Theorem aid in deriving the probability distribution from a characteristic function?
5. What implications does the Continuity Theorem have on the convergence of random variables?
6. Explain the significance of the Uniqueness Theorem.
7. What is the definition of a characteristic function for a random variable?
8. State the Helly-Bray Lemma. How does it relate to the convergence of characteristic functions?
9. Briefly describe the essence of the Kolmogorov Theorem in the context of characteristic functions.
10. If X and Y are independent random variables with characteristic functions ϕX (t) and ϕY (t), respectively, what is the characteristic function of X+Y?
11. State the Inversion Theorem. How can it be used to recover a probability distribution from its characteristic function?
12. Describe the main premise of the Continuity Theorem. How does it relate the convergence of characteristic functions to the convergence of distributions?
13. What does the Uniqueness Theorem tell us about the relationship between a random variable's distribution and its characteristic function?

14. What are some of the practical applications of characteristic functions in probability and statistics?
15. Given the random variable X which follows a standard normal distribution, what is the form of its characteristic function?
16. Why is knowledge of complex analysis important when working with characteristic functions?

10.11 Reference

1. Billingsley, P. (1995). Probability and Measure (3rd ed.). John Wiley & Sons.
2. Feller, W. (1971). An Introduction to Probability Theory and Its Applications, Vol. II (2nd ed.). John Wiley & Sons.
3. Gnedenko, B. V., & Kolmogorov, A. N. (1968). Limit Distributions for Sums of Independent Random Variables. Addison-Wesley.
4. Lévy, P. (1937). Théorie de l'addition des variables aléatoires. Gauthier-Villars.
5. Lukacs, E. (1970). Characteristic Functions (2nd ed.). Griffin.
6. Parzen, E. (1960). Modern Probability Theory and Its Applications. John Wiley & Sons.
7. Rao, C. R. (1973). Linear Statistical Inference and Its Applications (2nd ed.). John Wiley & Sons.

10.12 Further Reading

- Billingsley, P. (1995). Probability and Measure. Wiley Series in Probability and Mathematical Statistics.
- Feller, W. (1971). An Introduction to Probability Theory and Its Applications, Vol. 2. Wiley.
- Lévy, P. (1937). Théorie de l'Addition des Variables Aléatoires. Paris: Gauthier-Villars.

UNIT: 11 CENTRAL LIMIT THEOREM

Structure

11.1 Introduction

The Central Limit Theorem (CLT) is a cornerstone of probability theory and statistics. It provides a bridge that connects the sometimes-vast differences of individual distributions to a common result: the Gaussian or normal distribution. This unit explores the details, assumptions, and several forms of the CLT, all of which are foundational for understanding statistical inferences.

In the vast realm of probability and statistics, certain principles and theorems lay the foundational understanding necessary for deeper exploration and application. Among them, the Central Limit Theorem (CLT) shines as one of the most pivotal. Whether you're a student beginning your journey in the subject, a professional leveraging data, or someone with a passing interest in the mathematical phenomena behind everyday events, the CLT has implications that touch upon a myriad of areas.

The essence of the CLT lies in its profound assertion: no matter the original distribution of a set of variables, their sum (or average) will always tend towards a normal distribution as their number grows. It's akin to the common thread that binds the fabric of randomness, bringing order and predictability in the face of inherent variability.

This unit will guide through the nuances of the CLT, its various forms, and its underlying principles. We will embark on a journey from its one-dimensional version to the more generalized Lindeberg-Levy, Lyapunov, and Lindeberg-Feller theorems. By the end of this unit, the learner not only appreciate the beauty and power of the CLT but also be equipped to apply it in real-world scenarios.

11.2 Objectives

By the end of this chapter, learner should be able to:

- Understand the significance of the Central Limit Theorem.
- Describe the one-dimensional version of the CLT.
- Comprehend the Lindeberg-Levy, Lyapunov, and Lindeberg-Feller Theorems.
- Apply the CLT in practical statistical scenarios.

11.3 One-Dimensional Central Limit Problem

Definition:

The one-dimensional central limit problem refers to the convergence in distribution of the sum of independent random variables to the normal distribution. In simpler terms, no matter what individual distributions our random variables come from, as we add more and more of them together, their normalized sum will look more and more like a standard normal distribution.

Formal Presentation: Let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed (*i.i.d.*) random variables, each with an expected value
E[X_i]=μ and a variance Var(X_i)=σ2. The sum of these variables is:

$$S_n = \sum_{i=1}^{n} X_1$$

Now, consider the normalized version: $Z_n = \frac{S_n - n\mu}{\sigma_n}$

As n approaches infinity (i.e., as we consider more and more random variables), the distribution of Z_n tends towards a standard normal distribution (with mean 0 and variance 1).

Example:

Tossing a Coin Imagine we are tossing a fair coin. Let's represent a heads by the number 1 and tails by 0. If we toss the coin once, the average value is E[X]=0.5, and the variance is Var(X)=0.25.

Now, consider tossing the coin n times and calculating the sum of the results. The expected sum is n ×0.5 and the variance is n ×0.25.

Let's define our normalized sum for n coin tosses as:

$$Z_n = \frac{S_n - 0.5n}{0.5\, n}$$

As n gets larger, according to the one-dimensional CLT, Z_n will tend towards a standard normal distribution.

Visualization:

For a small n, say 5 or 10, the distribution of the sum (or average) of the coin tosses won't look very normal. It will be binomial, in fact. But as you increase n to 30, 50, or even 100, the distribution starts looking bell-shaped or 'normal'. By the time you reach n =1000, the distribution of the average of the coin tosses will be almost indistinguishable from a normal curve.

If you have a series of random processes (like the daily average temperature in a city over several years) and the characteristic functions of these processes are getting closer and closer to some other function, then the original processes (in this case, the temperature patterns) are also approaching a specific pattern or behavior.

The importance of the CLT lies in its implications:

Universality: Regardless of the original distribution of Xi (as long as it has finite mean and variance), the sum or average of a large number of these variables will have a distribution that is approximately normal. This is why the normal distribution appears in so many natural phenomena.

Simplifies Analyses: Many statistical methods are based on the assumption of normality. The CLT provides a justification for using these methods in practice when dealing with large samples.

Illustrative Example:

Suppose you are rolling a fair six-sided die. The outcomes are 1,2,…,6 with equal probabilities. If you roll the die once, the distribution of outcomes is uniform. However, if you roll the die, say, 50 times and compute the average result each time, the distribution of those averages will start to resemble a bell-shaped curve (a normal distribution), even though the individual rolls follow a uniform distribution. The more times you roll and average, the closer the distribution of averages gets to a normal distribution. This is the essence of the Central Limit Theorem in action.

The one-dimensional central limit theorem provides a powerful tool in statistics. Many real-world phenomena can be approximated as the sum of many small, independent effects. Even if learner do not know the original distribution of these effects, we know that their sum will tend towards a normal distribution, which is a well-understood and widely studied distribution. This allows statisticians to make inferences and predictions with confidence.

11.4 Lindeberg-Levy Theorem

The Lindeberg-Levy Theorem is a specific case of the Central Limit Theorem (CLT) tailored for sequences of independent and identically distributed (*i.i.d.*) random variables. It provides conditions under which the sum of these *i.i.d.* variables, when properly normalized, converges in distribution to the standard normal distribution.

Statement:

Let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed (*i.i.d.*) random variables, each with an expected value
E[X_i]=μ and a variance Var(X_i)=σ2. The sum of these variables is:

$$S_n = \sum_{i=1}^{n} X_1$$

Now, consider the normalized version: $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$

As n approaches infinity, Z_n converges in distribution to a standard normal distribution (with mean 0 and variance 1).

The theorem essentially says that if learner takes a large number of variables from the same distribution (with a finite mean and variance) and add them up, then the average will follow a normal distribution. This behavior holds true regardless of the shape or type of their common distribution, if it has a well-defined mean and variance.

Proof:

The proof of Lindeberg Levy theorem utilizes characteristic functions.

Given that $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed (*i.i.d.*) random variables, each with an expected value

E[X_i]=μ and a variance Var(X_i)=σ2. The characteristic function of X_1 is:

$$\emptyset(t) = E\left[e^{itX_1}\right]$$

The characteristic function for is

$$\emptyset_{z_n}(t) = E\left[e^{itz_n}\right]$$

Using properties of expected values and independence, this can be expanded as

$$\emptyset_{z_n}(t) = E\left[e^{it\frac{S_n - n\mu}{\sigma\sqrt{n}}}\right]$$

$$\emptyset_{z_n}(t) = E\left[e^{it\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}}\right]$$

$$\emptyset_{z_n}(t) = \prod_{i=1}^{n} E\left[e^{it\frac{X_i - n\mu}{\sigma\sqrt{n}}}\right]$$

$$\emptyset_{z_n}(t) = \left(\emptyset\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n$$

By Taylor series expansion around 0, given as:

$\emptyset(t) = 1 + it\mu - \frac{1}{2}t^2\sigma^2 + o(t).$

The above expression can write as:

$\log_{n\to\infty}\left(\emptyset\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n = \exp\left(-\frac{1}{2}t^2\right).$

The right-hand side is the characteristic function of a standard normal distribution. Since the convergence of characteristic functions implies convergence in distribution, Zn converges in distribution to the standard normal as n approaches infinity.

Why is the Lindeberg-Levy Theorem important?

Special Case: It's a special case of the more general CLT, which applies to i.i.d. variables with a common finite variance. This theorem is often taught in introductory courses because it covers many practical scenarios and has simpler conditions than some more general versions.

Foundational: Like other versions of the CLT, it underpins many methods in statistics. If you are conducting a study and collect a large sample of i.i.d. data with a known average and variance, you can make inferences about the population using techniques that assume normality, thanks to the Lindeberg-Levy theorem.

Illustrative Example:

Consider measuring the time it takes for different people to complete a specific task, like solving a puzzle. Each person's time can be considered a random variable. Even if individual times are not normally distributed, if you take a sample of a large number of people and compute the average time, the Lindeberg-Levy theorem tells you that this average will be approximately normally distributed (given that the variance is finite). So, you could then use this to make inferences about the average time it would take for anyone randomly selected from the population to complete the puzzle.

In essence, the Lindeberg-Levy theorem is one of the foundational stones of inferential statistics, allowing researchers in various fields to draw conclusions from sample data.

The One-Dimensional Central Limit Problem specifically focuses on the convergence of the distribution of sums (or averages) of i.i.d. variables to the normal distribution in one-dimensional space. There are multi-dimensional generalizations and various other versions of the CLT that delve deeper into different aspects of convergence and types of random variables, but the one-dimensional case is the most commonly cited and taught.

11.5 Lyapunov Theorem

The Lyapunov Central Limit Theorem is a more general form of the Central Limit Theorem (CLT) that provides conditions under which the sum of independent, but not necessarily identically distributed, random variables converge in distribution to a normal distribution. It's particularly useful when the variances of the individual random variables differ or when we want to establish normal convergence without assuming identical distributions.

Statement:

Consider $X_1, X_2, X_3, \ldots$ be a sequence of independent distributed random variables, each with an expected value

E[X_i]=μ_i and a variance Var(X_i)= $\sigma_i^2 > 0$

$$S_n = \sum_{i=1}^{n} X_i$$

If there exists $\delta > 0$

$$\log_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} E\left|X_i - \mu_i\right|^{2+\delta} = 0$$

Where $s_n^2 = Var(S_n) = \sum_{i=1}^{n} \sigma^2$

Then the normalized sum,

$$Z_n = \frac{S_n - E(S_n)}{S_n}$$

Converges to standard normal distribution as n tends to infinity.

Proof:

The proof of Lyapunov's theorem is more involved and relies on the properties of characteristic functions, much like the proof of the Lindeberg-Levy theorem, but it also introduces the use of Lyapunov's inequality.

Characteristic Functions:

Start by defining the characteristic function of Zn and express it in terms of the characteristic functions of the individual random variables. i.e.

Let $\emptyset_{z_n}(t)$ be the characteristic function of Zn. Using properties of characteristic function and the independence of Xi's

$$\emptyset_{z_n}(t) = E\left[e^{it\frac{S_n - n\mu}{\sigma\sqrt{n}}}\right]$$

$$\emptyset_{z_n}(t) = E\left[e^{it\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}}\right]$$

$$\emptyset_{z_n}(t) = \prod_{i=1}^{n} E\left[e^{it\frac{X_i - n\mu}{\sigma\sqrt{n}}}\right]$$

$$\emptyset_{z_n}(t) = \prod_{i=1}^{n} \emptyset_i\left({}^{t}/_{s_n}\right)$$

Where $\emptyset_i$ is the characteristic function of $X_i - \mu_i$.

Taylor Expansion:

Use a Taylor series expansion of the characteristic function of each random variable Xi about 0.

$$\emptyset_i(t) = 1 + itE[X_i - \mu_i] - \tfrac{1}{2}t^2\left(E[X_i - \mu_i]\right)^2 + \int_{-\infty}^{\infty}\left(e^{itx} - 1 - itx\right)e^{-x^2/2}\, dF_i(x)).$$

Lyapunov's Inequality:

Using Lyapunov's condition, we can manage the integral term. Recall the condition:

$$\log_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} E\left|X_i - \mu_i\right|^{2+\delta} = 0$$

With this condition, for any fixed t, we can show that:

$$\log_{n\to\infty} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \left(e^{\frac{itx}{S_n}} - 1 - \frac{itx}{S_n}\right) e^{-x^2/2} dF_i(x) = 0$$

This step involves using the condition and the Dominated Convergence Theorem. By managing the higher order terms with Lyapunov's condition and taking the limit as n→∞, show that the characteristic function of Zn converges to that of the standard normal distribution.

Given the above steps, and using the properties of the logarithm and exponential functions, we find that:

$$\log_{n\to\infty} \left(\emptyset\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n = \exp\left(-\frac{1}{2}t^2\right)$$

Convergence in Distribution:

As previously stated, the convergence of characteristic functions implies convergence in distribution. Thus, Zn converges in distribution to a standard normal random variable as n approaches infinity.

In simpler terms, the Lyapunov condition checks if the higher moments (beyond variance) of the random variables grow at a rate that is "manageable" compared to the sum of their variances.

If this condition is satisfied, then even without identical distributions, the sum of the random variables will still converge to a normal distribution when standardized.

Why is the Lyapunov Theorem important?

Generality: This theorem covers a broader set of scenarios than the standard CLT versions because it doesn't assume identically distributed random variables.

Varied Applications: In real-world situations, it's common to encounter sums of random variables from different distributions (e.g., in risk management, where diverse risk factors may impact a total outcome).

Illustrative Example:

Imagine a company analyzing the total delay in shipping products to customers. Different products might be shipped from different warehouses, each with its own distribution of delay times. While the delay for a single product type might not be normally distributed, the total delay for a large order (with products from multiple warehouses) could be approximated as normally distributed if the Lyapunov condition is met.

Overall, the Lyapunov theorem provides a robust tool for researchers and practitioners when dealing with sums of random variables from varied sources or distributions. It generalizes the idea that under appropriate conditions, the behavior of large sums of random variables can be understood using the familiar bell-shaped curve of the normal distribution.

11.6 Lindeberg-Feller Theorem

The Lindeberg-Feller Central Limit Theorem is a generalized version of the Central Limit Theorem (CLT) that addresses the sum of independent but not necessarily identically distributed random variables. It provides a set of conditions under which the sum of these variables will converge in distribution to a normal distribution. The theorem is often simply referred to as the "Lindeberg Condition".

Statement:

Let $X_1, X_2, ...$be a sequence of independent random variables, each with expected value $E(X_i) = \mu_i$ and variance $Var(X_i) = \sigma_i{}^2$, define

$$S_n = \sum_{i=1}^{n} X_i$$

$$s_n{}^2 = \sum_{i=1}^{n} \sigma_i{}^2$$

Suppose the Lindeberg condition holds:

For every $\epsilon > 0$

$$\log_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{n} E\left[(X_i - \mu_i)^2 . 1(|X_i - \mu_i| \epsilon s_n)\right] = 0$$

Then normalized sum:

$$Z_n = \frac{S_n - E(S_n)}{s_n}$$

Converges in distribution to the standard normal distribution as n approaches to infinity.

Proof:

For the normalized sum Z_n, its characteristic function is:

$$\emptyset_{Z_n}(t) = E\left[e^{it\frac{S_n - n\mu}{\sigma\sqrt{n}}}\right]$$

$$\emptyset_{Z_n}(t) = E\left[e^{it\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}}\right]$$

$$\emptyset_{Z_n}(t) = \prod_{i=1}^{n} E\left[e^{it\frac{X_i - n\mu}{\sigma\sqrt{n}}}\right]$$

$$\emptyset_{Z_n}(t) = \prod_{i=1}^{n} \emptyset_i\left({}^{t}/_{s_n}\right)$$

Where $\emptyset_i$ is the characteristic function of $X_i - \mu_i$.

a Taylor series expansion of the characteristic function of each random variable Xi about 0.

$$\emptyset_i(t) = 1 + itE[X_i - \mu_i] - \frac{1}{2}t^2\left(E[X_i - \mu_i]\right)^2 + o(t^2)$$

The main challenge is to control the error term, $o(t^2)$. Using the Lindeberg condition, it is possible to show that for ant fixed t

$$\log_{n\to\infty}\sum_{i=1}^{n} E\left[\left(X_i-\mu_i\right)^2 .1\left(\left|X_i-\mu_i\right|\epsilon s_n\right)\right]=0$$

This essentially ensures that the impact of large deviations from the mean becomes negligible as n grows, facilitating the application of the Central Limit Theorem.

Combining the above results, we can show that:

$$\log_{n\to\infty} \emptyset\left(\frac{t}{\sigma\sqrt{n}}\right)=\exp\left(-\frac{1}{2}t^2\right)$$

This is a characteristic function of standard normal distribution.

By the continuity theorem in the theory of characteristic functions, if the characteristic functions converge, the distributions themselves converge. Hence, Zn converges in distribution to a standard normal random variable as n goes to infinity.

Significance of the theorem

Flexibility: Unlike some other forms of the CLT, the Lindeberg-Feller theorem doesn't require the random variables to be identically distributed. This makes it applicable to more diverse real-world scenarios.

Tail Behavior: The Lindeberg condition specifically inspects the behavior of the tails of the distributions. This is insightful in many applications, particularly in risk management and finance where tail behavior can be crucial.

Illustrative Example:

Consider a factory producing items, where each item can have a defect with varying probabilities, and the nature/severity of the defect can differ across product types (different distributions). While each individual product type might have its own distinct distribution of defects, if the factory produces a large mix of these products and if the conditions of the Lindeberg-Feller theorem are satisfied, the proportion of defects in large batches can be approximated using a normal distribution.

In sum, the Lindeberg-Feller theorem provides a robust framework for understanding the distribution of the sum of non-identically distributed random variables, especially as the number of variables becomes large.

The Lindeberg-Feller Central Limit Theorem is a powerful generalization of the classical Central Limit Theorem. It handles the case where the random variables are not identically distributed, as long as they satisfy the Lindeberg condition. This condition ensures that no single random variable or small group of them has a disproportionately large impact on the sum, allowing for the emergence of the normal distribution in the limit.

Let us go through some solved examples to understand the Central Limit Theorem (CLT) better.

Example 1: Dice Rolling

Suppose you roll a fair six-sided die. The expected value (mean) E[X] is 3.5 and the variance Var(X) is 2.92. If you roll the die n times and take the average each time, the CLT states that the distribution of those averages will tend to be normal as n increases, with a mean of 3.5 and variance 2.92/n.

Let us calculate the average of rolling the dice 30 times. If we repeat this experiment a large number of times and plot the averages, the distribution of these averages would approach a normal distribution. To get the standard deviation of the average of 30 rolls, you'd take the square root of 2.92/30, which is about 0.311.

If you were to ask the probability that the average of those 30 rolls is between 3 and 4, you'd use a normal distribution table (or software) with mean 3.5 and standard deviation 0.311 to find that probability.

Example 2: Sample Average of a Uniform Distribution

Suppose the time (in hours) it takes for a package to be delivered follows a uniform distribution between 2 and 6 hours. That means any time between 2 to 6 hours is equally likely. The mean E[X] is 4 hours and the variance Var(X) is $\frac{4^2}{12}\ or\ \frac{4}{3}$.

Now, let us say you order 50 packages. What is the probability that the average delivery time of these 50 packages is less than 3.9 hours?

The standard deviation of the average of 50 delivery times would be the square root of (4/3)/50, which is about 0.163.

You would then convert 3.9 hours to a z-score:

$$z = \frac{3.9 - 4}{0.163} = -0.613$$

You would then use a z-table (or software) to find the probability that a standard normal variable is less than -0.613, which would give you the desired probability.

Example 3: Binomial Approximation

Suppose 20% of a town's population are smokers. If you randomly survey 100 people, what is the probability that more than 25 of them are smokers?

Here, we are dealing with a binomial distribution with n=100 and p=0.2. The mean is np=20, and the variance is np (1−p)=16.

However, with the CLT, we can approximate this binomial with a normal distribution, especially since our n is large.

The standard deviation is the square root of 16, which is 4.

To find the probability that more than 25 people are smokers, we'd first convert 25 to a z-score:

$$z = \frac{25 - 20}{4} 1.25$$

Then, using a z-table (or software), you'd find the probability that a standard normal variable is greater than 1.25 to get the desired probability.

Note: These examples are simplifications and, in practice, certain nuances should be considered. For instance, the third example (binomial approximation) is a common application of the CLT, but one should ensure that both

np and n(1−p) are sufficiently large for the normal approximation to be valid.

11.7 Summary

In this unit, the Central Limit Theorem (CLT), a foundational concept in probability and statistics, was dissected. Beginning with an introduction, we explored the idea that the normalized sum (or average) of a large number of independent, identically distributed random variables tends to follow a normal distribution, irrespective of the original distribution of the variables. Delving deeper, the one-dimensional central limit problem highlighted this phenomenon with practical examples. The unit then introduced three theorems that generalize the CLT to different conditions:

the Lindeberg-Levy theorem, which applies when variables are identically distributed; the Lyapunov theorem, which requires only a condition on moments and works even when the distributions are not identical; and the Lindeberg-Feller theorem, another generalization to non-identical distributions but based on a different condition. Each theorem was meticulously stated, proven, and elucidated with examples. The unit closed with self-assessment questions and suggestions for further reading, reinforcing the importance and widespread applicability of the CLT in statistical theory and practice.

11.8 Self-Assessment Exercises

1. What is the primary implication of the Central Limit Theorem (CLT) concerning the sum of a large number of random variables?
2. Suppose you are given a dataset with observations from a clearly non-normal distribution. If you were to take numerous samples from this dataset and average them, what would the distribution of these averages approach, according to the CLT?
3. Under what conditions does the Lindeberg-Levy theorem state that the sum of independent random variables will be approximately normally distributed?
4. Contrast the conditions under which the Lindeberg-Levy theorem and the Lyapunov theorem are applied.
5. For the Lyapunov theorem, explain in your own words what the Lyapunov condition checks about the random variables.
6. Describe the Lindeberg condition that is central to the Lindeberg-Feller theorem. Why is this condition important?
7. Given the sequence of random variables $X_1, X_2, \ldots$, each with mean μ and variance σ2, if $S_n = \sum_{i=1}^{n} X_1$, what would be the normalized form of Sn used in the context of the CLT?
8. True or False: The Central Limit Theorem is only valid for random variables that come from a normal distribution.
9. Why is the Central Limit Theorem considered as cornerstone in the field of statistics and probability?
10. If the random variables in question are not identically distributed, which of the theorems (Lindeberg-Levy, Lyapunov, or Lindeberg-Feller) might you consider applying, and why?

These questions span a range of difficulty levels, from basic recall and understanding to more complex application and analysis, allowing for a comprehensive self-assessment of the unit's content.

11.9 References

1. Billingsley, P. (1995). Probability and Measure. John Wiley & Sons.
2. Durrett, R. (2019). Probability: Theory and Examples. Cambridge University Press.
3. Feller, W. (1971). An Introduction to Probability Theory and Its Applications, Vol. II. John Wiley & Sons.
4. Gnedenko, B. V., & Kolmogorov, A. N. (1968). Limit Distributions for Sums of Independent Random Variables. Addison-Wesley.
5. Klenke, A. (2014). Probability Theory: A Comprehensive Course. Springer.
6. Lehmann, E. L., & Romano, J. P. (2005). Testing Statistical Hypotheses. Springer.
7. Shiryaev, A. N. (1996). Probability. Springer.

11.10 Further Reading

- "A First Course in Probability" by Sheldon Ross, Pearson
- "Probability Theory: A Comprehensive Course" by Achim Klenke, Springer
- "Probability and Measure" by Patrick Billingsley, John Wiley & Sons
- "The Emergence of Probability" by Ian Hacking, Cambridge University Press
- "The Essence of Multivariate Thinking: Basic Themes and Methods" by Lisa L. Harlow, Routledge
- "Statistics" by Robert S. Witte and John S. Witte, Wiley
- "Statistical Learning with Sparsity: The Lasso and Generalizations" by Trevor Hastie, Robert Tibshirani, and Martin Wainwright, CRC Press
- "The Drunkard's Walk: How Randomness Rules Our Lives" by Leonard Mlodinow, Pantheon

Note

Note

Note