STAT4028: Probability and Mathematical Statistics

Charles Christopher Hyland

Semester 1 2020

Abstract

Thank you for stopping by to read this. These are notes collated from lectures and tutorials as I took this course.

Contents

1.1 Sigma Algebras and Measures **1**
1.1.1 Motivation . . . 1
1.1.2 Vitali Sets . . . 1
1.1.3 Sigma Algebras . . . 2
1.1.4 Measures . . . 4
2.1.5 Algebras . . . 1
2.1.6 Caratheodory Extension Theorem . . . 1
2.1.7 Construction of Measures . . . 3
3.1.8 Uniqueness of measures . . . 1
3.1.9 Lebesgue-Stieltjes Integral . . . 2
3.1.10 Completion of Measure Spaces . . . 4
3.1.11 First Borel-Cantelli Lemma . . . 5

4.2 Random Variables **1**
4.2.1 Random Variables . . . 1
4.2.2 Distributions of random variables . . . 7

5.3 Independence **1**
5.3.1 Independence of sets of events . . . 1
5.3.2 Second Borel-Cantelli Lemma and Kolmogorov's 0-1 Law . . . 5
5.3.3 Kolmogorov's 0-1 law . . . 6

6.4 Integration **1**
6.4.1 Integration of Simple Functions . . . 1
6.4.2 Integration of non-negative measurable functions . . . 4
7.4.3 Convergence Theorems . . . 1
7.4.4 Relationship between Riemann and Lebesgue Integrals . . . 2
7.4.5 Abstract Lebesgue Integral and Integrable Functions . . . 6

8.5 Lebesgue Spaces **1**
8.5.1 $\mathcal{L}^p$-spaces . . . 1
8.5.2 Modes of convergence . . . 2
8.5.3 Inequalities in the $\mathcal{L}^p$-space . . . 8
8.5.4 Completeness of the $\mathcal{L}^p$-space . . . 10
8.5.5 L^p-spaces . . . 11

9.6 Expectation **1**
9.6.1 Expectation . . . 1
9.6.2 The Radon-Nikodym Theorem . . . 3
9.6.3 Substitution theorem for random variables . . . 5
9.6.4 Expectation of Functions of Random Variables . . . 6
9.6.5 Convex Functions . . . 11
9.6.6 Variance and Covariance . . . 12

10.7 Product Measures **1**
10.7.1 Caratheodory Extension Theorem for Product Measures . . . 1
10.7.2 Product Measure . . . 2
10.7.3 Fubini's Theorem . . . 5
10.7.4 Product Measure on n-dimensional space . . . 9
10.7.5 Random Vectors . . . 9

10.7.6 Change of variables for Random Vectors 10

11.8 Conditional Expectation **1**
11.8.1 Introduction 1
12.8.2 Properties of conditional expectation 1

13.9 Exponential Families **1**
13.9.1 Moment Generating Functions 1
13.9.2 Sufficient Statistics 2
13.9.3 Introduction 3
14.9.4 Exponential Family 1
14.9.5 The Mean-Value Parameter 3

15.10 Local Asymptotic Normality **1**
15.10.1 Taylor's Theorem 1
15.10.2 Weak Convergence 1
15.10.3 Asymptotic Statistics 2
15.10.4 Local Asymptotic Normality 3
15.10.5 Le Cam's First Lemma 5
15.10.6 LAN Property in Exponential Families 7
16.10.7 Le Cam's Third Lemma 1
16.10.8 Conditions implying LAN property 2
16.10.9 Strict Conditions for the LAN property to hold 5
16.10.10 LAN property for Location and Scale Families 7
17.10.11 L^2 differentiability implies LAN 1

18.11 Optimality of estimators **1**
18.11.1 Regular Estimators 1
18.11.2 Scores version of Le Cam's 3rd Lemma 2
18.11.3 RAJN Estimators 6
19.11.4 Asymptotically Efficient Estimators 1
19.11.5 Construction of AE Estimators 2
20.11.6 Newton-Raphson Algorithm 1
20.11.7 Identifying RAJN Estimators 2
20.11.8 Estimators based on moments and quantiles 4
21.11.9 Efficient Influence Function 1
21.11.10 Scalar Nuisance Parameter Model 2
21.11.11 General Setting Nuisance Parameter 3
21.11.12 Regularity in the presence of nuisance parameters 4
22.11.13 Estimating Smooth functions of parameters 1

23.12 Hypothesis Testing **1**
23.12.1 Asymptotically Optimal Parametric Testing 1
23.12.2 Local Alternatives 2
23.12.3 Classical Tests 3
24.12.4 Classical Tests are asymptotically equivalent 1
25.12.5 Properties of test statistics 1
25.12.5.1 Limiting Distribution Under Null Hypothesis 1
25.12.5.2 Limiting Distribution Under Nearby Alternatives 3
26.12.6 Regular Quadratic Form Test Statistics 1
27.12.7 Positive Definite Matrices 1
27.12.8 Orthogonal Decomposition Of Tests 1

28.13 Optimality of Bayesian Methods **1**
28.13.1 Introduction to Bayesian Methods . . . 1
29.13.2 Hellinger Distance . . . 1
30.13.3 Bayes Posterior Mean . . . 1
31.13.4 Expectation-Maximization Algorithm . . . 1
32.13.5 Normal Location Contamination Mixture Model . . . 1

0.14 Measure Theory Recap **1**
0.14.1 Classes of subsets . . . 1
0.14.2 Functions on Sets . . . 3
0.14.3 Extension Theorems . . . 5
0.14.4 Caratheodory Theorem . . . 5
0.14.5 Monotone Classes . . . 8
0.14.6 Lebesgue Measure . . . 8
0.14.7 Complete Measures . . . 9
0.14.8 Approximation and Regularity . . . 9
0.14.9 Measurability . . . 10
0.14.10 Simple Functions . . . 11
0.14.11 Integrating Simple Functions . . . 11
0.14.12 Integrating non-negative Functions . . . 12
0.14.13 Integrating measurable functions . . . 12
0.14.14 Measurability of functions . . . 12
0.14.15 Properties of Integrals . . . 13
0.14.16 Convergence Theorems for non-negative functions . . . 13
0.14.17 Lebesgue-Stieltjes Integral . . . 14
0.14.18 Measure on finite product of spaces . . . 15
0.14.19 Countable Product Space . . . 16
0.14.20 Fubini's Theorem . . . 16
0.14.21 Radon-Nikodym Theorem . . . 17
0.14.22 Convergence of functions . . . 17
0.14.23 Hölder and Minkowski Inequalities . . . 19
0.14.24 L^p-Spaces . . . 19

STAT4028: Probability and Mathematical Statistics

1. Non-measurable sets, Sigma algebra, Measure Spaces

1.1 Sigma Algebras and Measures

1.1.1 Motivation

We give a motivation on why we are interested in probability theory.

Let Ω be a finite or countable set. Let $\mathbb{P}$ be a function $\mathbb{P} : \Omega \to [0, 1]$. In the set up of random experiments with finite or countably many possible outcomes, we can easily deal with it with our discrete probability space $(\Omega, \mathbb{P})$.

However, the issue comes when Ω is an uncountable space. Suppose we are interested in assigning probabilities to the interval [0,1], which contains an uncountable number of points. The probability of us picking any point $\omega \in [0, 1]$ will be zero (we can always be more and more precise). However, summing over all points in [0,1] will give us 0 which does not make sense. This brings us to our first realisation that we are unable to assign a measure to all sets in Ω and hence we must be careful.

1.1.2 Vitali Sets

Let us take $\Omega = I = [0, 1]$. For now, we say a measure on I generalises the notion of length. The Lebesgue measure on I, which we denote by λ assigns to every interval its length $\lambda(a, b) = b - a$. We will require that λ is countably additive and translation invariant.

We will show that there exists sets which are not measurable.

Definition 1.1 *(Rationals). A point is in the rationals if $q \in \mathbb{Q}$ can be expressed as $q = \frac{x}{y}$ where $x, y \in \mathbb{Z}$ are integers.*

Definition 1.2 *(Coset). The coset of rationals $\mathbb{Q}$ in the reals $\mathbb{R}$ is*

$$x + \mathbb{Q} = \{x + q : q \in \mathbb{Q}\}$$

for each $x \in \mathbb{R}$.

Lemma 1.3 *Let $x, y \in \mathbb{R}$. We can group real numbers into categories of cosets*

1. *If $y - x \in \mathbb{Q}$, then $x + \mathbb{Q} = y + \mathbb{Q}$*
2. *If $y - x \notin \mathbb{Q}$, then the cosets $x + \mathbb{Q}$ and $y + \mathbb{Q}$ are **disjoint**.*

Definition 1.4 *(Vitali set). A subset $V \subseteq [0, 1]$ is called a Vitali set if V contains a single point from each coset of $\mathbb{Q}$ in $\mathbb{R}$.*

Remark 1.5 *In order to pick points as described, we require the axiom of choice.*

Definition 1.6 *(Axiom of choice). The choice function is defined on a set of sets whereby it takes an element from each set. The axiom of choice states that a choice function always exists on a set of sets.*

Theorem 1: Non-measurable sets always exist

Let $(\mathbb{R}, \mathcal{B}, \mu)$ be a measure space on $\mathbb{R}$. Then, for any measurable set $A \in \mathcal{B}$ where $\mu(A) > 0$, there exists a non-measurable subset $B \subset A$.

Remark 1.7 *To show the existence, we can look at $A \subset (0,1)$ and by the density of rationals and translation invariant property of the Lebesgue measure, shift the set to our desired set.*

Theorem 2: Vitali sets are not Lebesgue-measurable

If $V \subseteq [0,1]$ is a Vitali set, then V is not Lebesgue-measurable.

We proceed to construct an example of a Vitali set.

Claim 1.8 *There exists a set $B \subset [0,1)$ where we set $\{q_n\} = \mathbb{Q} \cap [0,1)$ and $B_n = q_n \bigoplus B$.*

1. *B_n are disjoint sets;*
2. *$\cup_n B_n = [0,1)$.*

Corollary 1.9 *We cannot assign a Lebesgue measure to the set B*

$$1 = \sum_{i=1}^{\infty} \lambda(B_i) = \sum_{i=1}^{\infty} \lambda(B) = \infty.$$

Hence, due to Vitali sets, we need to be more selective to which sets we can assign a measure if we want to generalise the notion of length.

1.1.3 Sigma Algebras

Definition 1.10 *(Sample space and points). The set of all possible outcomes is called the **sample space** Ω. A point $\omega \in \Omega$ is called a **sample point**.*

Definition 1.11 *(Events). **Events** are subsets of Ω to which we can assign a probability. An event A has **occurred** if the sample point $\omega \in \Omega$ satisfies $\omega \in A$.*

We can now define the class of sets to which we can assign a measure to.

Definition 1: σ-algebra

A class of sets $\mathcal{F}$ is a σ-algebra (σ-field) if

1. $\Omega \in \mathcal{F}$
2. If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$
3. If $A_1, A_2, ... \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_n \in \mathcal{F}$ (Closed under **countable** union)

We have the following useful properties from set theory.

Definition 2: De Morgan's Laws

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Lemma 1.12 *Let A and B be sets. Then*

$$\bigcup (A_i) \cap B = \bigcup (A_i \cap B)$$
$$\bigcap (A_i) \cap B = \bigcap (A_i \cap B)$$

Furthermore, we have that $A \cap B \subseteq A$ and $A = A \cup (A \cap B)$.

We can now describe properties of $\sigma - algebra$ that will be useful for us.

Proposition 1: Properties of σ-algebra

Let $\mathcal{F}$ be a $\sigma - algebra$. We have the following properties for $\sigma - algebra$

1. If $A_n \in \mathcal{F}$ for $n \in \mathbb{N}$ then $\cap_{i=1}^{\infty} A_n \in \mathcal{F}$ (Closed under countable intersection)
2. $\emptyset \in \mathcal{F}$
3. If $A_1, ..., A_N \in \mathcal{F}$ then $\cup_{i=1}^{N} A_n \in \mathcal{F}$ (Closed under **finite** union)
4. If $A_1, ..., A_N \in \mathcal{F}$ then $\cap_{i=1}^{N} A_n \in \mathcal{F}$ (Closed under **finite** intersection)

Definition 1.13 *(Measurable Space). The pair (Ω, F) is called a measurable space.*

Definition 1.14 *(Power set). Let Ω be a set. Then, the power set is $\mathcal{P}(\Omega) = 2^{\Omega}$ is the largest $\sigma - algebra$ of Ω.*

We now want to be able to generate the smallest $\sigma - algebra$ $\mathcal{F}$ that contains all open and closed sets in $\Omega = \mathbb{R}$. We now introduce the tools we will need in order to do so.

Proposition 1.15 *(Intersection of $\sigma-algebra$ is a $\sigma-algebra$). Let I be an arbitrary index set and suppose that for every $i \in I$, $\mathcal{A}_i$ is a $\sigma-algebra$ of subsets of Ω. Then,*

$$\mathcal{A} \coloneqq \bigcap_{i \in I} \mathcal{A}_i \subseteq \mathcal{P}(\Omega)$$

is a $\sigma-$algebra of subsets of Ω.

Definition 3: σ-algebra generated by $\mathcal{A}$

Let $\mathcal{A} \subset 2^{\Omega}$. There exists a minimal $\sigma-$algebra that contains $\mathcal{A}$ which we denote by $\sigma(\mathcal{A})$. This σ-algebra is generated by the **intersection of all σ-algebras** which contains $\mathcal{A}$.

We describe a special type of σ-algebra which we will need later onwards.

Definition 4: Borel $\sigma-$algebra

Let $\Omega = \mathbb{R}$. Define $\mathcal{B} = \sigma(\{G \subset \mathbb{R} : \text{G is open}\})$. $\mathcal{B}$ is called the **Borel** $\sigma-$algebra. A set $A \in \mathcal{B}$ is called a **Borel set.**

We can also construct the Borel $\sigma - algebra$ by looking at the $\sigma - algebra$ generated by intervals in $\mathbb{R}$.

Claim 1.16 *Let $\mathcal{B}$ be the Borel $\sigma - algebra$. Then,*

$$\mathcal{F}_1 = \sigma(\{(-\infty, x] : x \in \mathbb{R}\})$$

and

$$\mathcal{F}_2 = \sigma(\{[a, b] : a, b \in \mathbb{R}\}).$$

We have that

$$\mathcal{B} = \mathcal{F}_1 = \mathcal{F}_2.$$

Example 1.17 *(Cylinder $\sigma - algebra$). Let $\Omega = \{0,1\}^{\mathbb{N}}$ where $\omega = (\omega_1, \omega_2, ...)$ where $\omega_i \in \{0,1\}$. The **cylinder sets** are defined by $A_i\{\omega : \omega_j = 0 \quad , 1 \leq j \leq i\}$ and $A_i^c = \{\omega : \omega_j = 1 \quad , 1 \leq j \leq i\}$. A_i is known as **cylinder sets**. Then,*

$$\mathcal{F} = \sigma\{A_i : i \in \mathbb{N}\}$$

*is the **cylinder** $\sigma - algebra$.*

Lemma 1.18 *The cyliner $\sigma - algebra$ is isomorphic to the Borel $\sigma - algebra$ $\mathcal{B}$ on the interval (0,1).*

1.1.4 Measures

We can now assign a measure to sets of interest.

Definition 5: Measure

A set function $\mu : \mathcal{F} \to [0, \infty]$ is a measure on $(\Omega, \mathcal{F})$ if

1. $\mu(\emptyset) = 0$,
2. If $\{A_i\}_{i \in \mathbb{N}} \in \mathcal{F}$ are disjoint sets, then $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ (countable additivity).

Definition 1.19 *(Measure Space). The triple $(\Omega, \mathcal{F}, \mu)$ is called a measure space.*

Remark 1.20 *If $\mu(\Omega) = 1$ then μ is called a **probability measure** and denoted by $\mathbb{P}$. We then call $(\Omega, \mathcal{F}, \mathbb{P})$ a **probability space.***

Definition 1.21 *(Finite and $\sigma-$finite measures). Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. The measure μ is **finite** if $\mu(\Omega) < \infty$. The measure μ is $\sigma-$**finite** if for sets $A_n \in \mathcal{F}$ such that $\cup_{n\in\mathbb{N}} A_n = \Omega$, then $\mu(A_n) < \infty$ for all $n \in \mathbb{N}$.*

Lemma 1.22 *The Lebesgue measure λ on $\mathbb{R}$ is a $\sigma-$finite measure.*

Remark 1.23 *However, the Lebesgue measure λ is **not** finite on $\mathbb{R}$.*

Definition 1.24 *(Counting Measure). Define the measurable space $(\Omega, \mathcal{P}(\Omega))$ where $\mathcal{P}(\Omega)$ is the power set of Ω. Then, the counting measure μ is defined as*

$$\mu(A) = |A|$$

for all $A \in \mathcal{P}(\Omega)$ and $|A|$ is the cardinality of the set A.

Proposition 2: Properties of a measure

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then

1. (Finite additivity) If $A_i \in \mathcal{F}$ for $1 \leq i \leq N$ are disjoint sets, then
$$\mu(\bigcup_1^N A_i) = \sum_1^N \mu(A_i)$$
2. (Monotonicity) If $A \subset B$ then $\mu(A) \leq \mu(B)$
3. (Sub-additivity) If $A \subset \bigcup_n A_n$ then
$$\mu(A) \leq \sum_n \mu(A_n)$$
4. If $\mu(\Omega) < \infty$, then the inclusion exclusion formula holds for $A_1, ..., A_n \in \mathcal{F}$
$$\mu(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i) - \sum_{i<j} \mu(A_i \cap A_j) + \sum_{i<j<k} \mu(A_i \cap A_j \cap A_j) + ... + (-1)^{n-1}\mu(A_1 \cap ... \cap A_n)$$

Definition 1.25 *(Increasing and decreasing sequence of sets). Let $\{A_n\}_{n\in\mathbb{N}}$ be a sequence of sets such that $A_n \subset A_{n+1}$ and $\cup_{n\in\mathbb{N}} A_n = A$. We then write that $A_n \uparrow A = \lim_{n\to\infty} \cup_n A_n$. Likewise, if $A_n \supset A_{n+1}$ and $\cap_{n\in\mathbb{N}} A_n = A$, we write that $\lim_{n\to\infty} \cap_n A_n = A_n \downarrow A$.*

Proposition 3: Continuity of the measure

1. (Continuity from below) Let $A_n \in \mathcal{F}$ for $n \in \mathbb{N}$. If $A_n \uparrow A$ then

$$\lim_{n\to\infty} \mu(A_n) = \mu(\bigcup_{n=0}^{\infty} A_n).$$

2. (Continuity from above) Let $A_n \in \mathcal{F}$ for $n \in \mathbb{N}$. If $A_n \downarrow A$ and $\mu(A_0) < \infty$ then

$$\lim_{n\to\infty} \mu(A_n) = \mu(\bigcup_{n=0}^{\infty} A_n).$$

Definition 1.26 *(Discrete probability space). Let Ω be a countable set. Let $\mathcal{F} = 2^{\Omega}$. Then,*

$$P(A) = \sum_{\omega\in\Omega} p(\omega)$$

where $p(\omega) \geq 0$ and $\sum_{\omega\in\Omega} p(\omega) = 1$. If Ω is a finite set, we then let $p(\omega) = \frac{1}{|\Omega|}$.

STAT4028: Probability and Mathematical Statistics

2. Construction of Measures

We are interested in constructing measures on the measurable space $(\Omega, \mathcal{F})$. The intuition on how to do so is that we want to take a smaller collection of sets $\mathcal{A} \subset \mathcal{F}$ and build up from there. In particular, we will come to see that $\mathcal{A}$ is an **algebra**.

2.1.5 Algebras

We will denote $\mathcal{F}$ as the $\sigma - algebra$ for the rest of this section and an algebra as $\mathcal{A}$.

Definition 6: Algebra

The class $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is an **algebra** if

1. $\Omega \in \mathcal{A}$
2. If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$
3. If $A_i \in \mathcal{A}$ then $\bigcup_{i=1}^{N} \in \mathcal{A}$.

Remark 2.27 *In some text, an algebra is also known as a* ***field****.*

Example 2.28 *Define the sample space* $\Omega = (0, 1]$. *Define the algebra*

$$\mathcal{B}_0 = \{\cup_{i=1}^{n}(a_i, b_i] : 0 \leq a_i < b_i \leq 1; n \in \mathbb{N}\}$$

The class $\mathcal{B}_0$ *is an algebra but not a* $\sigma - algebra$.

Claim 2.29 *If* $\mathcal{F}$ *is a* $\sigma - algebra$, *then* $\mathcal{F}$ *is an algebra.*

2.1.6 Caratheodory Extension Theorem

Definition 2.30 *(Pre-measure). Let* $\mathcal{A}$ *be an algebra. Let* $\mu_0 : \mathcal{A} \to \overline{\mathbb{R}}^+$. *The function* μ_0 *is called a* ***pre-measure*** *if*

1. $\mu_0(\emptyset) = 0$
2. *For every countable sequence* $\{E_i\}_{i \in \mathbb{N}}$ *of pairwise disjoint sets whose union lies in* $\mathcal{A}$, *then*

$$\mu_0\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \sum_{i=1}^{\infty} \mu_0(E_i).$$

Remark 2.31 *The pre-measure is just a measure defined on the domain of an algebra.*

Pre-measures will be extremely useful in helping us define outer-measures.

Example 2.32 *Let* $\mathcal{B}_0 = \{\cup_{i=1}^n (a_i, b_i] : 0 \leq a_i < b_i \leq 1, n \in \mathbb{N}\}$. *For a set* $A \in \mathcal{B}_0$, *we define the Lebesgue pre-measure*

$$\lambda(A) = \sum_{i=1}^{n} (b_i - a_i).$$

Then, λ *is a pre-measure on* $\mathcal{B}_0$.

The λ pre-measure we defined on the algebra $\mathcal{B}_0$ is what we will want to extend to the Borel $\sigma - algebra$ and hence create the Lebesgue measure.

Theorem 3: Caratheodory's Extension Theorem

Let $\mathcal{A}$ be an algebra. Then, if μ is a $\sigma-$finite measure on the algebra $\mathcal{A}$, then μ has an **unique** extension to a measure on $\sigma(\mathcal{A})$.

Corollary 2.33 *The Lebesgue measure on* $([0,1], \mathcal{B})$ *is well defined.*

We will now show the steps on how to prove Caratheodory's extension theorem. First, we give the extension of the pre-measure we defined earlier.

Proposition 4: Outer Measure

Let Ω be the sample space and let $\mathcal{A}$ be an algebra of the space Ω. Let μ be the pre-measure we defined on $\mathcal{A}$. Then, we define that for any set $E \subseteq \Omega$, the set function

$$\mu^*(E) = \inf_{\{A_i\}_{i\geq 1}} \Big\{ \sum_{i \geq 1} \mu(A_i) : A \subseteq \bigcup_{i \geq 1} A_i \text{ and } A_i \in \mathcal{A} \Big\}.$$

Here, μ^* is known as the outer-measure.

Remark 2.34 *Here, we take the infimum over all coverings* $\{A_i\}_{i\geq 1}$ *of the set* E. *Furthermore, note that this is defined for all sets* $E \in \Omega$, *not sets that are measurable.*

The outer measure provided an upper bound on the pre-measure $\mu(E) \leq \mu^*(E)$.

Claim 2.35 *Let* μ^* *be the outer measure and* μ *be the pre-measure. If a set* E *is in the algebra* $\mathcal{A}$ *then*

$$\mu(A) = \mu^*(A).$$

We can state a series of properties of the outer-measure, which are similar to the properties of the pre-measure.

Lemma 2.36 *(Null set is zero). Let* μ^* *be the outer measure. Then*

$$\mu^*(\emptyset) = 0.$$

Lemma 2.37 *(Monotonicity of outer measure) Let* μ^* *be an outer-measure. For any two sets* $E, F \in \Omega$, *if* $E \subseteq F$ *then* $\mu(E) \leq \mu(F)$.

Lemma 2.38 *(Subadditivity of outer measure). Let μ^* be the outer measure. Then for a set $A \subset \cup_{k\geq 1} A_k$, we have that*

$$\mu^*(A) \leq \sum_{k\geq 1} \mu^*(A_k)$$

2.1.7 Construction of Measures

Now, we will define the collection of sets to which we will call measurable. Intuitively, the measurable sets will be the sets for which the upper bound $\mu^*(E)$ is tight, that is $\mu^*(E) + \mu^*(E^c) = 1$.

Definition 7: μ^*-measurable set

Let $\mu^* : \mathcal{P}(\Omega) \to \overline{\mathbb{R}}^+$ be an outer measure. We call a set $A \subseteq \Omega$ a μ^*-measurable set if

$$\mu^*(S) = \mu^*(S \cap A) + \mu^*(S \cap A^c)$$

for all sets $S \subseteq \Omega$. We define the family of measurable sets $\mathcal{M}$ where $A \in \mathcal{M}$ if it is μ^*-measurable.

Lemma 2.39 *Let $\mathcal{A}$ be the algebra on the set Ω and μ^* the outer-measure. Then, if $E \in \mathcal{A}$ then E is μ^*-measurable.*

Proposition 5: The family of measurable sets if a $\sigma - algebra$

Let Ω be our set of interest and $\mathcal{A}$ be the algebra on Ω. Let μ^* be the outer-measure. Let $\mathcal{M}$ be the family of μ^*-measurable sets. Then, we have that

1. $\mathcal{M} \supseteq \mathcal{A}$
2. $\mathcal{M}$ is a $\sigma - algebra$
3. From (1) and (2), we have that
$$\mathcal{M} \supseteq \mathcal{F}(\mathcal{A}).$$
where $\mathcal{F}(\mathcal{A})$ is the $\sigma - algebra$ generated by the algebra $\mathcal{A}$.

Proposition 6: The Outer Measure restricted to $\mathcal{M}$ is a measure

Let μ^* be the outer-measure we have defined and $\mathcal{M}$ the family of μ^*-measurable sets. Additionally, let ν be the pre-measure defined on the algebra $\mathcal{A}$. Then

$$\mu^*|_{\mathcal{M}} \colon \mathcal{M} \to \overline{\mathbb{R}}^+$$

is a **measure** μ. Furthermore, μ is an **extension** to the pre-measure ν as $\mu(A) = \nu(A)$ for all sets $A \in \mathcal{A}$.

From all this, we can now concisely state Caratheodory's extension theorem.

Theorem 4: Caratheory's Extension Theorem

Let $\mathcal{A}$ be an algebra on Ω and let $\nu : \mathcal{A} \to \overline{\mathbb{R}}^+$ be a pre-measure on $\mathcal{A}$. Let $\mathcal{F}(\mathcal{A})$ be the $\sigma - algebra$ generated by $\mathcal{A}$. Then, there exists a measure $\mu : \mathcal{F}(\mathcal{A}) \to \overline{\mathbb{R}}^+$ such that $\mu|_{\mathcal{A}} = \nu$. Moreover, if ν is $\sigma-$finite, then the extension μ is **unique** and also $\sigma-$finite.

Corollary 2.40 *Let $\mathcal{B}_0$ be the Borel algebra defined before and let λ_0 be the Lebesgue outer-measure defined on it. Then, the Lebesgue measure λ on $([0,1], \mathcal{B})$ is well-defined!*

STAT4028: Probability and Mathematical Statistics

3. Uniqueness of measures

3.1.8 Uniqueness of measures

This lecture, we work on proving uniqueness of the extension of the pre-measure in Caratheodory's theorem. In particular, we want to assert the claim.

Claim 3.41 *Assume that μ is the extension of the pre-measure ν on the algebra $\mathcal{A}$. If ν is $\sigma-$finite, then it is the* ***unique extension*** *of ν onto $\mathcal{F}(\mathcal{A})$.*

One particular issue is that we may have two probability measures that agree on a collection of sets E but they do not agree on the $\sigma - algebra$ generated $\sigma(E)$. Therefore, we need to be careful when we have that $\mu = \nu$ on the algebra $\mathcal{A}$ and make sure that they still agree on $\sigma(\mathcal{A})$. To do so, we require new structures.

Definition 8: π-system

A collection of sets P is a π-system if it is closed under intersection, that is, for all $A, B \in P$

$$A \cap B \in P.$$

Definition 9: λ-system

A collection of sets L is a $\lambda-$system (d-system) if

1. $\Omega \in L$
2. $A, B \in L$ where $A \subset B$ implies that $B \backslash A \in L$
3. $A_n \in L$ with $A_n \uparrow A$ implies that $A \in L$.

We now state a theorem which we will need in showing the uniqueness of the extension in Caratheodory's theorem.

Theorem 5: Dynkin's $\pi - \lambda$ Theorem

If $\mathcal{P}$ is a $\pi-$system and $L \supset \mathcal{P}$ is a $\lambda-$system, then $\sigma(\mathcal{P}) \subset L$.

Proposition 7: $\sigma - algebra$ and $\pi - \lambda$ System

The collection of sets $F \subset 2^{\Omega}$ is a σ-algebra if and only if F is both a π-system and a λ-system.

Lemma 3.42 *For $e \subset 2^{\Omega}$, define $\lambda(e)$ to be the intersection of all $\lambda-$systems that contains e. Then $\lambda(e)$ is a $\lambda-$system and is the smallest $\lambda-$system that contains e.*

Theorem 6: $\sigma - algebra$-generated by a $\pi-$system

Let $\mathcal{P} \subset 2^{\Omega}$ be a $\pi-$system. Then, the $\lambda-$system generated by P is agrees with the $\sigma - algebra$ generated by P

$$\lambda(\mathcal{P}) = \sigma(\mathcal{P}).$$

We can now finally show the uniqueness property for Caratheodory's theorem.

Theorem 7: Measures agree on $\sigma - algebra$ generated by a $\pi-$system

If the measures μ and ν agree on a π-system P, then they agree on the σ-algebra $\sigma(P)$.

Proof: First, define the collection of sets

$$\mathcal{L} = \{B \in \sigma(P) : \mu(B) = \nu(B)\}.$$

Now, we show that $\mathcal{L}$ is a $\lambda-$system.

1. $\mu(\Omega) = 1 = \nu(\Omega)$
2. Suppose that $A, B \in \mathcal{L}$ and $A \subset B$. Then
$$\mu(B \setminus A) = \mu(B) - \mu(A) = \nu(B) - \nu(A) = \nu(B \setminus A)$$
and therefore $B \setminus A \in \mathcal{L}$
3. If $A_n \in \mathcal{L}$ and $A_n \uparrow A$ then
$$\mu(A) = \lim_{n\to\infty} \mu(A_n) = \lim_{n\to\infty} \nu(A_n) = \nu(A)$$
and therefore $A \in \mathcal{L}$.

Then, by Dynkin's $\pi - \lambda$ theorem, we have that

$$\sigma(P) \subset \mathcal{L}$$

Now recall, that by construction of $\mathcal{L}$, for every set $B \in \mathcal{L}$, we have that $\mu(B) = \nu(B)$. Therefore, as $\sigma(P) \subset \mathcal{L}$, we have that $\mu(B) = \nu(B)$ for every set $B \in \sigma(P)$ and hence μ agrees with ν on $\sigma(P)$. ■

We have started with an algebra $\mathcal{A}$ with a pre-measure ν. We then extend the measure to the σ-algebra $\sigma(\mathcal{A})$. Hence, from the theorem above, $\mathcal{A}$ is also a π-system, so any measures that agree on $\mathcal{A}$, will also agree on $\sigma(\mathcal{A})$.

3.1.9 Lebesgue-Stieltjes Integral

We can now generalise the Lebesgue measure.

Definition 10: Borel measure

Let μ be a measure on the $\sigma - algebra$ $\mathcal{F}$. The measure μ is a Borel measure if the Borel $\sigma - algebra$ $\mathcal{B} \subseteq \mathcal{F}$. That is, any Borel set B is $\mu-$measurable.

There is a relationship between right continuous increasing functions F on $\mathbb{R}$ and a Borel measure μ_F.

Proposition 8: Relationship between increasing right continuous functions and Borel Measures

Let F be a right continuous and increasing function F on $\mathbb{R}$. Let μ_F be a Borel measure on $\mathbb{R}$.

1. Given the function F, we can construct a Borel measure μ_F;
2. Given the Borel measure μ_F, we can construct a F.

Remark 3.43 *If $F(-\infty) = 0$ and $F(\infty) = 1$, then F is the cumulative distribution function.*

Definition 11: Regular Measure

Let Ω be a topological space. Let $\mathcal{B}$ be the Borel $\sigma - algebra$, which is the smallest $\sigma - algebra$ that contains all open sets. Let $\mu : \mathcal{F} \to \overline{\mathbb{R}}^+$ and $\mathcal{F} \supseteq \mathcal{B}$.
μ is **regular** if for all sets $A \in \mathcal{F}$ and $\epsilon > 0$, there exist sets $F \subseteq A \subseteq G$ where F is closed, G is open such that

$$\mu(G \backslash F) \leq \epsilon.$$

Remark 3.44 *We can approximate any set $A \in \mathcal{F}$ from below by a closed set and from above by an open set.*

We now describe how we construct one from another.

Proposition 9: Constructing increasing right continuous functions

Let μ be a regular Borel measure on $\mathbb{R}$. Then, we define the function

$$F_0(t) = \begin{cases} \mu((0,t]) & t > 0 \\ -\mu((t,0]) & t < 0 \end{cases}.$$

Finally, we define the function $F(t) = F_0(t) + \mu((-\infty])$. F is an increasing right continuous function.

We now show how to derive a Borel measure from a increasing right continuous function.

Definition 12: Lebesgue-Stieltjes Outer Measure

Let F be a right continuous increasing function. We define the **Lebesgue-Stieltjes outer measure** $\mu_F^* : \mathcal{P}(\Omega) \to \overline{\mathbb{R}}^+$ as

$$\mu_F^*(A) = inf\Big\{ \sum_{k=0}^{\infty} (F(b_k) - F(a_k)) : A \subseteq \bigcup_{k=0}^{\infty} (a_k, b_k] \Big\}$$

for every set $A \in \mathcal{P}(\Omega)$.

Definition 13: Lebesgue-Stieltjes $\sigma - algebra$

Let μ_F^* be the Lebesgue-Stieltjes outer measure. Define the set $\mathcal{L}$ as all the sets $A \in \Omega$ such that

$$\mu_F^*(E) = \mu_F^*(E \cap A) + \mu_F^*(E \cap A^c)$$

for all sets $E \in \Omega$. Then, the collection of sets $\mathcal{L}$ is a $\sigma - algebra$.

Proposition 10: Lebesgue-Stieltjes Measure

Let F be an increasing right continuous function. Let μ_F^* be the Lebesgue-Stieltjes outer measure induced by F. Let $\mathcal{L}$ be the Lebesgue-Stieltjes $\sigma - algebra$ associated with μ_F^*. Then, the **Lebesgue-Stieltjes measure induced by F** μ_F is the outer-measure μ_F^* restricted to $\mathcal{L}$.

Theorem 8: Lebesgue-Stieltjes measure function

Let $F : \mathbb{R} \to \mathbb{R}$ such that F is non-decreasing and right continuous. Then, there exists a unique Borel measure μ on $(\mathbb{R}, \mathcal{B})$ such that

$$\mu((a, b]) = F(b) - F(a).$$

μ is the **Lebesgue-Stieltjes measure function**.

Remark 3.45 *If we set $F(x) = x$, then $\mu((a, b]) = b - a$ is the **Lebesgue measure**.*

Definition 3.46 *(Lebesgue-Stieltjes Integral). Let f be μ_F-measurable function. Then, the Lebesgue-Stieltjes integral is defined as*

$$\int_{\mathbb{R}} f d\mu_F.$$

Remark 3.47 *This is a generalisation of the Lebesgue integral which we will see later!*

3.1.10 Completion of Measure Spaces

We are now interested in including **null sets** into our $\sigma - algebra$. Why do we care about having complete measure spaces? Suppose if two functions f,g satisfy $f(x) = g(x)$ for all $x \in X \backslash N$ where N has measure zero. We would like to treat f and g as essentially the same thing. However, without completeness, it could be the case that f is measurable but g is not. Furthermore, many theorems in measure theory, for instance Fubini or Radon-Nikodym, needs completeness to make full sense.

Definition 14: Complete measure space

Let $\mathcal{F} \subseteq \mathcal{P}$ be a $\sigma - algebra$ and let $\mu : \mathcal{F} \to \overline{\mathbb{R}}^+$ be a measure. Then, $(\Omega, \mathcal{F}, \mu)$ is **complete** if for every set $A \in \mathcal{F}$ such that $\mu(A) = 0$, then for all sets $E \subseteq A$, we have that $E \in \mathcal{F}$. The sets E are known as **negligible sets**.

Observation 3.48 *By monotonicity of the measure, we have that the negligible set $\mu(E) = 0$.*

We now state how to actually complete a measure space and that this completion is unique.

Proposition 11: Completion of a $\sigma-algebra$

Let $\mathcal{F} \subseteq \mathcal{P}$ be a $\sigma-algebra$. Define $\mu : \mathcal{F} \to \overline{\mathbb{R}}^+$ be a measure. Then

$$\overline{\mathcal{F}} = \{A \cup N : A \in \mathcal{F} \text{ and } N \subseteq E \in \mathcal{F} \text{ such that } \mu(E) = 0\}.$$

Then, $\overline{\mathcal{F}}$ is a $\sigma-algebra$.

Proposition 12: Completion of a measure

Let $\mathcal{F} \subseteq \mathcal{P}$ be a $\sigma-algebra$. Define $\mu : \mathcal{F} \to \overline{\mathbb{R}}^+$ be a measure. Then, we define the measure $\overline{\mu} : \overline{\mathcal{F}} \to \overline{\mathbb{R}}^+$ to be

$$\overline{\mu}(A \cup N) = \mu(A)$$

for all sets $A \in \overline{\mathcal{F}}$. Clearly, $\overline{\mu}|_{\mathcal{F}} = \mu$.

Theorem 9: Completion of a measure space

Define the measure and $\sigma-algebra$ pair $(\mu, \mathcal{F})$. Furthermore, assume that Ω is $\sigma-$finite with respect to μ. Then, the completion $(\Omega, \overline{\mu}, \overline{\mathcal{F}})$ is unique.

Remark 3.49 *It is always possible to complete a measure and a $\sigma-algebra$.*

Proposition 13: Lebesgue $\sigma-algebra$ and Borel $\sigma-algebra$

The Lebesgue $\sigma-$algebra $\mathcal{B}_c$ is the completion of the Borel $\sigma-algebra$ $\mathcal{B}$ with respect to the Lebesgue measure λ.
That is, we can extend λ to $\mathcal{B}_c$ where $E \subset \mathcal{B}_c$ whereby there exists $A \in B$ and $B \subset N \in B$ such that $E = A \cup B$ with $\lambda(N) = 0$.

Remark 3.50 *The Lebesgue $\sigma-algebra$ is just the completion of the $\mathcal{B}$!*

Corollary 3.51 *The Lebesgue measure is a Borel measure.*

Remark 3.52 *There are sets which are Lebesgue measurable but not Borel measurable. The Borel $\sigma-$algebra is generated by open sets whereas the Lebesgue $\sigma-algebra$ is generated by open sets and null sets.*

3.1.11 First Borel-Cantelli Lemma

There is a very important duality between sets and functions. That is, with a set, we can construct a function known as an indicator function.

Definition 15: Indicator Function

Let Ω be a sample space. Then, the indicator function of a set $A \subset \Omega$ is defined as

$$1_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A. \end{cases}$$

Remark 3.53 *In this course, we will generally be working with measurable sets $A \in \mathcal{F}$ when constructing indicator functions.*

This definition gives us some nice properties.

Lemma 3.54 *Suppose that $A, B \subset \Omega$. Then*

1. *$1_A \leq 1_B$ if and only if $A \subseteq B$*
2. *$1_{A^c} = 1 - 1_A$*

We are now interested in analysing sequence of events. However, first recall some probabilistic terminology.

1. The set of all possible outcomes Ω is the **sample space**
2. A point $\omega \in \Omega$ is a **sample point**
3. **Events** are subsets of Ω to which we can assign a probability measure to. That is, they are in the $\sigma - algebra$ $\mathcal{F}$
4. We say that an **event has occurred** $A \in \mathcal{F}$ if the **outcome (sample point)** $\omega \in A$

Definition 16: Almost everywhere

For a general measure space $(\Omega, \mathcal{F}, \mu)$, we say that any statement S holds **almost everywhere** (a.e.) if

$$\mu(\{\omega \in \Omega : S(\omega) \text{ is false}\}) = 0.$$

Remark 3.55 *In a probability space, we say that the statement S holds **almost surely** (a.s.).*

We now revise the idea of limit inferiors and limit superiors from analysis.

Definition 3.56 *(Limit inferior/superior). The limit inferior of a sequence $\{x_n\}_{n\geq 1}$ is*

$$\liminf_{n\to\infty} x_n = \lim_{n\to\infty} (\inf_{k\geq n} x_k).$$

The limit superior of a sequence $\{x_n\}_{n\geq 1}$ is

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty} (\sup_{k\geq n} x_k).$$

We can apply this idea to sequences of events. We will see later that the definition of convergence for random variables depends on manipulating sequences of events, which require limits of sets.

Definition 3.57 *(Infimum and supremum of sets). Suppose that $\{A_n\}_n \subset \Omega$ is a sequence of events.*

1. *The **infimum** of the sequence of events is*

$$\inf_{k\geq n} A_k := \bigcap_{k=n}^{\infty} A_k$$

2. *The **supremum** of the sequence of events is*

$$\sup_{k\geq n} A_k := \bigcup_{k=n}^{\infty} A_k$$

Definition 17: Infinitely often/Limit superior for sets

Let $\{A_n\}_n$ be a sequence of events. We define the subset of events

$$lim\ sup\ A_n = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n = \{\omega : \forall m \text{ there exists } n = n(\omega) \geq m \text{ with } \omega \in A_n\} = (A_n \text{ i.o.})$$

Remark 3.58 *Note that if we let $B_m = \bigcup_{n\geq m} A_n$, we have that B_m is a **decreasing sequence of sets.***

Intuitively, we say that we are interested in collection of sample points ω for which that for all values of m, there exists a rank $n(\omega)$ such that the sample point ω appears in every event A_n..

Proposition 14: The limit superior of measurable sets is measurable

Let $\{A_n\}_n$ be a sequence of events on the measure space $(\Omega, \mathcal{F}, \mu)$. Then, the set

$$(A_n \text{ i.o}) \in \mathcal{F}$$

Proof: First, recall that

$$(A_n \text{ i.o.}) = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n$$

First, recall that countable union of measurable sets is measurable

$$\bigcup_{n=m}^{\infty} A_n \in \mathcal{F}$$

Next, the countable intersection of measurable sets is measurable

$$\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n \in \mathcal{F}$$

Therefore, we have shown that

$$(A_n \text{ i.o}) \in \mathcal{F}.$$

■

Proposition 3.59 *Let 1_A be the indicator function of a measurable set A_n. Then*

$$\overline{lim} 1_{A_n}(\omega) = 1_{\overline{lim} A_n}(\omega)$$

for all $\omega \in \Omega$.

Proof:(Sketch). It can be claimed that

$$\inf 1_{A_n}(\omega) = 1_{\inf A_n}(\omega) \quad \sup 1_{A_n}(\omega) = 1_{\sup A_n}(\omega)$$

Now, write out

$$\overline{lim} 1_{A_n}(\omega) = \inf \sup 1_{A_n}(\omega)$$

and apply the claim twice to get

$$1_{\inf \sup A_n}(\omega) = 1_{\overline{lim} A_n}(\omega)$$

■

We now state one of the most important lemmas of the course. The Borel-Cantelli lemma is a useful tool for proving almost sure convergence.

Theorem 10: First Borel-Cantelli lemma

Let $\{A_n\}$ be a sequence of events such that $\sum_n \mathbb{P}(A_n) < \infty$. Then,

$$\mathbb{P}((A_n \text{ i.o})) = \mathbb{P}(\lim_{n\to\infty} \sup\ A_n) = 0.$$

Proof: Let $B_m = \cup_{n\geq m} A_n$. Then, we have that

$$B_m \downarrow B = \cap_{m=1}^{\infty} B_m = \overline{\lim} A_n.$$

Therefore, we have that

$$\mathbb{P}(B_m) \to \mathbb{P}(\overline{\lim} A_n)$$

However, due to the subaddivity of the probability measure

$$\mathbb{P}(B_m) \leq \sum_{n\geq m} \mathbb{P}(A_n) \to 0$$

as $n \to \infty$ since $\sum_n \mathbb{P}(A_n) < \infty$ implies that $\mathbb{P}(A_n) \to 0$. We can therefore conclude that

$$\mathbb{P}(B_m) \to \infty$$

which in turn implies that the set $(A_n\ i.o.)$ will have measure zero. ■

We now define something that is the opposite of infinitely often.

Definition 18: Eventually/Limit infimum for sets

Define the probability space $(\Omega, \mathcal{F})$. Let $\{A_n\}_n$ be a sequence of events. We define the subset

$$\lim\ inf\ A_n = \bigcup_m \bigcap_{n\geq m} A_n = \{\omega : \omega \in A_n \text{ for all sufficiently large n}\} = (A_n \text{ eventually})$$

Remark 3.60 *A way to interpret* $(A_n$ *eventually)* *is that the point* ω *appears in all but finitely many sets.*

We give some properties of the limit infimum for sets.

Proposition 15: The limit inferior of measurable sets is measurable

Define the probability space $(\Omega, \mathcal{F})$. Let $\{A_n\}_n$ be a sequence of events on the measure space $(\Omega, \mathcal{F}, \mu)$. The set

$$(A_n \text{ eventually}) \in \mathcal{F}.$$

Proof: We have that $\bigcap_{n\geq m} A_n \in \mathcal{F}$ if $A_n \in \mathcal{F}$. Then, we can conclude that $\bigcup_m \bigcap_{n\geq m} A_n \in \mathcal{F}$. ■

Proposition 16: Relationship between limit superior and limit inferior of sets

Let $\{A_n\}_n$ be a sequence of measurable sets. Then

$$\underline{\lim}\, A_n \subset \overline{\lim}\, A_n$$

Proof:(Sketch). First, suppose that

$$\omega \in \bigcup_{m=1}^{\infty} \bigcap_{n\geq m} A_n$$

This equivalent to saying that there exists $N_0 \in \mathbb{N}$ such that $\omega \in A_n$ for all $n \geq N_0$. That is,

$$\omega \in \bigcap_{n\geq N_0} A_n$$

Now, the key insight is that this implies that

$$\omega \in \bigcup_{n\geq N_0} A_n$$

Then, we use the fact that

$$\omega \in \bigcup_{n\geq N_0} A_n \subseteq \bigcup_{n=1}^{\infty} A_n$$

Then, this means that $\omega \in \bigcup_{n=m}^{\infty} A_n$ for all $m \geq 1$. Hence, we can conclude that

$$\omega \in \bigcap_{m=1}^{\infty} \bigcup_{n\geq m} A_n = \overline{lim}\, A_n$$

■

Proposition 17: Complements between limit inferior and limit superior of sets

Define the probability space $(\Omega, \mathcal{F})$. Let $(A_n$ eventually) and $(A_n$ i.o) be the limit inferior and limit superior of a sequence of sets $\{A_n\}_n$. Then

$$(A_n^c \text{ eventually}) = (A_n \text{ i.o.})^c$$

Proof:

$$(A_n^c \text{ eventually}) = \bigcup_{m=1}^{\infty} \bigcap_{n\geq m} A_n^c = \bigcup_{m=1}^{\infty} \Big(\bigcup_{n\geq m} A_n \Big)^c = \Big(\bigcap_{m=1}^{\infty} \bigcup_{n\geq m} A_n \Big)^c = (A_n \text{ i.o.})^c$$

■

Theorem 11: Fatou's Lemma for sets

Define the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For the sets $A_n \in F$, we have that

$$\mathbb{P}(\overline{lim}\ A_n) \geq \overline{lim}\ \mathbb{P}(A_n)$$

$$\mathbb{P}(\underline{lim}\ A_n) \leq \underline{lim}\ \mathbb{P}(A_n)$$

Proof:(Sketch). First, define the decreasing sequence

$$B_m = \bigcup_{n \geq m} A_n$$

whereby we have that $B_m \downarrow \overline{lim}\ A_n$. Therefore, we have that

$$\mathbb{P}(B_m) \downarrow \mathbb{P}(\overline{lim}\ A_n)$$

Now, note that for a fixed n, $B_n \supset A_n$. Therefore, by monotonicity

$$\mathbb{P}(B_m) \geq \mathbb{P}(A_m)$$

From this, we take limit superior of both sides

$$\overline{lim}\ \mathbb{P}(B_m) \geq \overline{lim}\ \mathbb{P}(A_m)$$

Hence, we can conclude that

$$\mathbb{P}(\overline{lim}\ A_n) \geq \overline{lim}\ \mathbb{P}(A_n)$$

To prove the other result, repeat the same process except let $B_m = \bigcap_{n \geq m} A_n$. ■

Proposition 3.61 *Let 1_A be the indicator function of a measurable set A_n. Then*

$$\underline{lim} 1_{A_n}(\omega) = 1_{\underline{lim} A_n}(\omega)$$

for all $\omega \in \Omega$.

Proof: Recall that

$$\inf 1_{A_n}(\omega) = 1_{\inf A_n}(\omega) \quad \sup 1_{A_n}(\omega) = 1_{\sup A_n}(\omega)$$

Then, apply this result twice to the expression

$$\underline{lim} 1_{A_n}(\omega) = \sup_{k \geq 1} \inf_{n \geq k} 1_{A_n}(\omega).$$

■

STAT4028: Probability and Mathematical Statistics

4. Random Variables and Distributions

4.2 Random Variables

4.2.1 Random Variables

Today, we look at random variables and the distribution function of random variables.

Definition 4.62 *(Preimage of a function). Define a function $f : (\Omega_1, \mathcal{F}_1) \to (\Omega_2, \mathcal{F}_2)$. Then, the preimage of a set $B \in \mathcal{F}_2$ is given by*

$$f^{-1}(B) = \{\omega \in \Omega_1 : f(\omega) \in B\}$$

Remark 4.63 *Note that we use the phrase preimage as the function may not necessarily have an inverse but it will have a preimage.*

Now, we can define what we mean by a measurable function.

Definition 19: Measurable function

A function $f : (\Omega_1, \mathcal{F}_1) \to (\Omega_2, \mathcal{F}_2)$ is measurable with respect to $\mathcal{F}_1$ if for all measurable sets $A \in \mathcal{F}_2$, we have

$$f^{-1}(A) \in \mathcal{F}_1.$$

Example 4.64 *The constant function $f = c$ is a measurable function as defined by*

$$f^{-1}(A) = \begin{cases} \emptyset & c \notin A \\ \Omega & c \in A \end{cases}$$

for all $A \in \mathcal{F}_2$ as $\emptyset, \Omega \in \mathcal{F}_1$.

Definition 20: Random Variable

Define the function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathbb{B})$ where $\mathbb{B}$ is the Borel $\sigma-$algebra. Then X is a random variable (RV) if it is measurable with respect to $\mathcal{F}$.

Lemma 4.65 *Let Ω be a discrete probability space. Then, any function $X : \Omega \to \mathbb{R}$ is a random variable.*

We state important properties of preimages of functions that we will need.

Claim 4.66 *Let $f : (\Omega_1, \mathcal{F}_1) \to \Omega_2$. Then, we have that*

1. $f^{-1}(A^c) = \left(f^{-1}(A)\right)^c$
2. $f^{-1}(\cup_n A_n) = \cup_n f^{-1}(A_n)$

We can now show an important result whereby if a collection of sets is has a measurable preimage, then the $\sigma - algebra$ generated by the collection of sets also has a measurable preimage.

Proposition 18: Measurability of $\sigma - algebra$ generated by sets

Define the function $f : (\Omega_1, \mathcal{F}_1) \to \Omega_2$. Suppose that for any set in the collection of sets $A \in \mathcal{A} \subset 2^{\Omega_2}$, we have that

$$f^{-1}(A) \in \mathcal{F}_1.$$

Then, for any set in the $\sigma - algebra$ generated $A \in \sigma(\mathcal{A})$, we have that

$$f^{-1}(A) \in \mathcal{F}_1$$

Proof: We have a collection of sets $\mathcal{A}$ such that for any set $A \in \mathcal{A}$, we have that $f^{-1}(A) \in \mathcal{F}_1$. We want to show that any set $A \in \sigma(\mathcal{A})$, we have that $f^{-1}(A) \in \mathcal{F}_1$.

First, define the set

$$\beta = \{A \in \Omega_1 : f^{-1}(A) \in \mathcal{F}_1\}$$

Clearly, by construction, we have that

$$\mathcal{A} \subseteq \beta$$

Now, the goal is to show two things

1. β is a $\sigma-$algebra
2. $\sigma(\mathcal{A}) \subseteq \beta$

These two things will imply that every set $A \in \sigma(\mathcal{A})$, we have that $f^{-1}(A) \in \mathcal{F}_1$, the result we want to show.

First, we show that β is a $\sigma-$algebra

1. $\emptyset \in \beta$ as $f^{-1}[\emptyset] = \emptyset \in \mathcal{F}_1$
2. Let $A \in \beta$. Then, as $\mathcal{F}_1$ is a $\sigma-$algebra, $f^{-1}(A) \in \mathcal{F}_1$ implies that $f^{-1}(A)^c = f^{-1}(A^c) \in \mathcal{F}_1$. Hence, $A^c \in \beta$.
3. Suppose $A_k \in \beta$ for $k \in \mathbb{N}$. Then as $\mathcal{F}_1$ is a $\sigma-$algebra, $f^{-1}[A_k] \in \mathcal{F}_1$ we have that $\cup_{k\in\mathbb{N}} f^{-1}[A_k] = f^{-1}[\cup_{k\in\mathbb{N}} A_k] \in \mathcal{F}_1$. Therefore, $\cup_{k\in\mathbb{N}} \in \beta$.

Hence, β is a $\sigma-$algebra.

Now, we argue that $\sigma(\mathcal{A}) \subseteq \beta$. First, by definition of $\mathcal{A}$, we have that $f^{-1}[A] \in \mathcal{F}_1$ for all sets $A \in \mathcal{A}$. Hence, $\mathcal{A} \subset \beta$. Now, $\sigma(\mathcal{A})$ is the smallest $\sigma-$algebra that contains $\mathcal{A}$. As a result, $\sigma(\mathcal{A}) \subset \beta$.

Hence, we can conclude that $f^{-1}(A) \in \mathcal{F}_1$ for every set $A \in \sigma(\mathcal{A})$. ■

With this proposition, we can specify an easier way to verify whether is a function a random variable.

Theorem 12: Test for a random variable

Let $(\Omega, \mathcal{F})$ be a measurable space. Let $X : \Omega \to \mathbb{R}$ be a function. Then, the following are equivalent

1. X is a random variable;
2. $X^{-1}[(\alpha, \infty]] \in \mathcal{F}$ is measurable for all $\alpha \in \mathbb{Q}$;
3. $X^{-1}[[\alpha, \infty]] \in \mathcal{F}$ is measurable for all $\alpha \in \mathbb{Q}$;
4. $X^{-1}[[-\infty, \alpha)] \in \mathcal{F}$ is measurable for all $\alpha \in \mathbb{Q}$;
5. $X^{-1}[[-\infty, \alpha]] \in \mathcal{F}$ is measurable for all $\alpha \in \mathbb{Q}$.

We can replace $\mathbb{Q}$ by $\mathbb{R}$ in the above statement.

Proof: First, define the collection of sets $\mathcal{A} = \{[(\alpha, \infty]] : \alpha \in \mathbb{Q}\}$. Then, if $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{A}$, this means that

$$X^{-1}(B) \in \mathcal{F}$$

for all sets B in the $\sigma - algebra$ generated $\sigma(\mathcal{A})$. However, recall that by definition, the Borel $\sigma - algebra$ $\mathcal{B}$ is generated by

$$\mathcal{B} = \sigma(\mathcal{A})$$

Therefore, we have shown that every Borel set has a measurable preimage and therefore by definition, X is a random variable. ■

Definition 21: σ-algebra generated by random variable

Let $X : \Omega \to \mathbb{R}$ be a random variable. The set

$$\sigma(X) = \mathcal{G} = \{X^{-1}(B) : B \in \mathcal{B}\}$$

is called the $\sigma-$algebra generated by the random variable X where $\mathcal{B}$ is the Borel $\sigma - algebra$ and it is denoted by $\sigma(X)$.

We investigate the above definition.

Proposition 19: Smallest $\sigma - algebra$ generated by random variable

Let $X : \Omega \to \mathbb{R}$ and let

$$\sigma(X) = \mathcal{G} = \{X^{-1}(B) : B \in \mathcal{B}\}$$

where $\mathcal{B}$ is the Borel $\sigma - algebra$. Then $\sigma(X) = \mathcal{G}$ is a $\sigma-$algebra and is the smallest one for which X is a random variable.

Proof:(Sketch). It is straightforward to show that $\mathcal{G}$ is a $\sigma - algebra$. However, to show that $\mathcal{G}$ is the smallest one is not as easy. First, recall that the Borel $\sigma - algebra$ is generated by $\mathcal{B} = \sigma(\mathcal{A})$ where $\mathcal{A}$ is the semi-algebra of all open intervals. To show that $\mathcal{G}$ is the smallest $\sigma - algebra$ for which X is a random variable, we need to show that

$$\mathcal{G} = X^{-1}(\sigma(\mathcal{A})) = \sigma(X^{-1}(\mathcal{A}))$$

where the right hand term indicates that $\mathcal{G}$ is the smallest $\sigma - algebra$.

$\rightarrow$ First, as $\mathcal{A} \subseteq \sigma(\mathcal{A})$ this means that

$$X^{-1}(\mathcal{A}) \subseteq X^{-1}(\sigma(\mathcal{A}))$$

Now, recall that $X^{-1}(\sigma(\mathcal{A}))$ is a $\sigma - algebra$. By minimality, this means that

$$\sigma(X^{-1}(\mathcal{A})) \subseteq X^{-1}(\sigma(\mathcal{A}))$$

$\leftarrow$ Define the collection of sets

$$\mathcal{F}_2 = \{B \subset \mathbb{R} : X^{-1}(B) \in \sigma(X^{-1}(\mathcal{A}))\}$$

It is easy to show that $\mathcal{F}_2$ is a $\sigma - algebra$. By definition,

$$X^{-1}(\mathcal{F}_2) \subset \sigma(X^{-1}(\mathcal{A})) \tag{*}$$

Also, since $X^{-1}(\mathcal{A}) \subset \sigma(X^{-1}(\mathcal{A}))$, this means that $\mathcal{A} \subset \mathcal{F}_2$ and therefore

$$\sigma(\mathcal{A}) \subseteq \mathcal{F}_2 \tag{**}$$

as $\mathcal{F}_2$ is a $\sigma - algebra$. Combining both (*) and (**), we therefore have

$$X^{-1}(\sigma(\mathcal{A})) \subseteq X^{-1}(\mathcal{F}_2) \subset \sigma(X^{-1}(\mathcal{A}))$$

and hence we can conclude that $X^{-1}(\sigma(\mathcal{A})) \subseteq \sigma(X^{-1}(\mathcal{A}))$.

From this, we can therefore conclude that

$$\mathcal{G} = X^{-1}(\sigma(\mathcal{A})) = \sigma(X^{-1}(\mathcal{A}))$$

and hence, $\mathcal{G}$ is the smallest $\sigma - algebra$ for which X is a random variable. ■

Lemma 4.67 *Let $A \subset \Omega$ and let $\mathcal{F}$ be a σ-algebra. Then, the indicator function 1_A is a random variable if and only if $A \in \mathcal{F}$.*

Definition 22: $\sigma - algebra$ generated by collection of random variables

If $X_1, X_2, ...$ are maps where $X_i : \Omega \to \mathbb{R}$, then the $\sigma-$algebra generated by $X_1, ..., X_n$ for all $n \in \mathbb{N}$ is the smallest $\sigma-$algebra with respect to which $X_1, ..., X_n$ are random variables. We denote this by

$$\sigma(X_1, ..., X_n) = \bigcup_{i=1}^{n} \sigma(X_i)$$

Claim 4.68 *For all $n \in \mathbb{N} \cup \{\infty\}$, we have that*

$$\sigma(X_1, ..., X_n) = \sigma(\sigma(X_1), ..., \sigma(X_n)).$$

Definition 23: Borel function

Let $f : (\mathbb{R}, \mathbb{B}) \to (\mathbb{R}, \mathbb{B})$ be a measurable function with respect to the Borel $\sigma - algebra$ $\mathcal{B}$. Then, f is known as a **Borel function**.

Proposition 20: Continuous and Borel Functions

Define the function $f : (\mathbb{R}, \mathbb{B}) \to (\mathbb{R}, \mathbb{B})$.

1. If f is a continuous function, then f is a Borel-measurable function.
2. If f is a monotone function, then f is a Borel-measurable function.

Proof: Suppose that f is a continuous function. Then, construct the set

$$\mathcal{L} = \{f^{-1}[O] : O \in \mathcal{B}\}$$

where due to the fact that f is continuous, all elements of $\mathcal{L}$ will be an open set. In fact, $\mathcal{L}$ is a $\sigma - algebra$ which contains all open sets in $\mathcal{B}$. But, by definition, this is the Borel $\sigma - algebra$ and hence f is measurable with respect to the Borel $\sigma - algebra$.

Recall that a function is measurable if $f^{-1}[(-\infty, a]] \in \mathcal{F}$ for all $a \in \mathbb{Q}$. Now, WLOG, assume that f is monotone increasing. Then, notice that

$$\{f^{-1}[(-\infty, a]]\} = \{t : f(t) \leq a\}$$

The set $\{t : f(t) \leq a\}$ will be either

1. $\emptyset$
2. $\mathbb{R}$
3. $(-\infty, z]$

Clearly, $\emptyset, \mathbb{R} \in \mathcal{B}$. Now, for (3), let $a \in \mathbb{R}$ and therefore we are interested in the set $\{t : f(t) \leq a\}$. Then, if

$$x \in \{t : f(t) \leq a\}$$

we have that for all $y < x$, by monotonicity of f, $f(y) < f(x) \leq a$ and hence

$$y \in \{t : f(t) \leq a\}$$

With this in mind, define $b = \sup\{t : f(t) \leq a\}$ and therefore

$$\{t : f(t) \leq a\} = (-\infty, b]$$

which is a Borel set. Hence, the preimage of all Borel sets $(-\infty, a]$ is a Borel set $(-\infty, b]$. Hence, f is Borel measurable. ■

Remark 4.69 *All continuous functions are Borel functions, but not all Borel functions are continuous. Likewise, not all measurable functions are Borel measurable.*

Proposition 4.70 *Let $(\Omega, \mathcal{F}, P)$ be a probability space and Y, Z be metric spaces. If $f : X \to Y$ is measurable and $\phi : Y \to Z$ is continuous, then*

$$\phi \circ f : X \to Z$$

*is **measurable**.*

Proof: Let U be an open set in Z. Then $\phi^{-1}[U]$ is an open set in Y. Furthermore, $f^{-1}[\phi^{-1}[U]] \in \mathcal{A}$. ■

Remark 4.71 *If we are working on the real line $\mathbb{R}$, we in fact only require ϕ is a Borel function (measurable). Recall that every continuous function is a Borel function.*

Proposition 4.72 *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $f = (f_1, ..., f_k) : \Omega \to \mathbb{R}^k$ be a function. Then f is measurable if and only if every component function $f_k : \Omega \to \mathbb{R}$ is measurable.*

So now, with the last 2 propositions, we can construct new random variables by viewing them as compositions of measurable and continuous functions.

Theorem 13: Operations on Random Variables

Let X and Y be random variables and $f : (\mathbb{R}, \mathbb{B}) \to (\mathbb{R}, \mathbb{B})$ be a Borel function. Then the following operations yields random variables

1. X+Y
2. X · Y
3. $\frac{X}{Y}$ if Y $\neq 0$
4. f(X)
5. $|X|$, max(X,Y), min(X,Y).

Proof:(Sketch). Define the function $F(\omega) = (X(\omega), Y(\omega))$, which is a measurable function as X, Y are measurable functions respectively. Then, $\psi(x, y) = x + y$ is a continuous function. Hence,

$$\psi \circ F$$

is a Borel measurable function. ■

Remark 4.73 *Note that for max, we can express that as*

$$\max(X, Y) = \frac{1}{2}\Big(X + Y + |X - Y|\Big)$$

In the context of limits, it is convenient to allow random variables to attain the value of $\pm\infty$. In particular, we define the extended Borel $\sigma - algebra$ as $\overline{\mathcal{B}} = \sigma([-\infty, x) : x \in \mathbb{R})$. Hence a function $X : (\Omega, \mathcal{F}) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ is a random variable if it is measurable.

Theorem 14: Sequence of random variables

Let $(\Omega, \mathcal{F}, P)$ be a probability space. If $X_n \coloneqq \{X_n\}_{n \geq 1}$ are a sequence of random variables, then so is

1. $\inf X_n$
2. $\sup X_n$
3. $\liminf\ X_n$
4. $\limsup\ X_n$

Moreover, if $X_n \to X$ pointwise, then X is a random variable.

Proof:(Sketch). Recall that if $X^{-1}[[\alpha,\infty]]$ is measurable for all $\alpha \in \mathbb{R}$, then X is measurable. Suppose that $inf\ X_n \geq \alpha$ for $\alpha \in \mathbb{R}$. This implies that $X_n(\omega) \geq \alpha$ for all $n \in \mathbb{N}$. Then

$$inf\ X_n^{-1}\Big[[\alpha,\infty]\Big] = \bigcap_{n\in\mathbb{N}} X_n^{-1}\Big[[\alpha,\infty]\Big]$$

is measurable since X_n is measurable (random variable) for all $n \in \mathbb{N}$. Now, $\sup X_n = -\inf(-X_n)$ to show that $\sup X_n$ is measurable.
Recall that $\liminf\ X_n = \sup_{n\in\mathbb{N}}\Big(\inf_{k\geq n}\{X_k, X_{k+1}, ...\Big)$. Then applying the fact that the supremum and infimum of random variables is random variable, we get our result.
Finally, if $X_n \to X$ pointwise, then the $\liminf\ X_n = \limsup\ X_n = X$ and hence X is a random variable. ■

Remark 4.74 *One of the useful things is that limits preserves measurability of functions unlike continuity.*

4.2.2 Distributions of random variables

Random variables X induces a probability measures on $(\mathbb{R},\mathcal{B})$ called its **distribution**.

Definition 24: Distribution function

Let $X : (\Omega,\mathcal{F},\mathbb{P}) \to (\mathbb{R},\mathbb{B})$ be a random variable, where $\mathbb{P}$ is a probability measure on $(\Omega,\mathcal{F})$. Then, the random variable X induces a probability measure on $(\mathbb{R},\mathbb{B})$ called its **distribution**

$$\mu(A) = \mathbb{P}(\{\omega : X(\omega) \in A\}) = \mathbb{P}(X \in A)$$

for all Borel sets $A \in \mathbb{B}$.

Remark 4.75 *Note that the distribution pulls the Borel set $A \subset \mathbb{R}$ back to $(\Omega,\mathcal{F},\mathbb{P})$ and we apply the probability measure $\mathbb{P}$ to it.*

Claim 4.76 *The triple $(\mathbb{R},\mathbb{B},\mu)$ is a probability space.*

Proof: First, recall that for a Borel set $A \subset \mathbb{B}$, we have that $\mu(A) = \mathbb{P}(\{\omega : X(\omega) \in A\})$. Now, let $\{A_i\}_{i\in I}$ be disjoint Borel sets. Now, since A_i are disjoint, this means that if $X \in \cup_i A_i$, then it is only in 1 A_j. Furthermore, as ω cannot be mapped to multiple A_i, then $\{\omega : X(\omega) \in A_i\} \subset \mathcal{F}$ are disjoint. So, we have

$$\mu(\cup_i A_i) = \mathbb{P}(\{\omega : X(\omega) \in \cup_i A_i\}) = \mathbb{P}(\cup_i\{\omega : X(\omega) \in A_i\}) = \sum_i \mathbb{P}(\{\omega : X(\omega) \in A_i\}) = \sum_i \mu(A_i).$$

■

However, we normally describe the distribution of a random variable through its distribution function, also known as the cumulative distribution function.

Definition 25: Distribution Function

Let $X : (\Omega,\mathcal{F},\mathbb{P}) \to (\mathbb{R},\mathbb{B},\mu)$ be a random variable. The distribution **distribution function** F_X of the random variable X is

$$F_X(x) = \mu((-\infty,x]) = \mathbb{P}(X \leq x) = \mathbb{P}(\{\omega : X(\omega) \leq x\})$$

for $x \in \mathbb{R}$.

We now state properties associated to the CDF.

Theorem 15: Properties of CDF

Let F be the Cumulative Distribution Function for a random variable X. Then, F satisfies the following properties:

1. F is non-decreasing;
2. $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$
3. F(x) is right continuous
4. $F(X_-) = \lim_{t\to x^-} F(x) = P(X \leq x)$
5. $P(X = x) = F(x) - F(x_-)$.

Proof: (1). Let $x \leq y$, then $\{X \leq x\} \subseteq \{X \leq y\}$. Then, by monotonicty of the probability measure

$$F(x) = \mathbb{P}(X \leq x) \leq \mathbb{P}(X \leq y) = F(y)$$

(2). As $x \to \infty$, then $\{\omega : X(\omega) \leq x\} \to \Omega$. Therefore $\mathbb{P}(\Omega) = 1$. Likewise, $\{\omega : X(\omega) \leq x\} \to \emptyset$ as $x \to -\infty$.
(3). We want to show that

$$\lim_{y\to x^+} F(y) = F(x).$$

So then, define the sequence $(x, y_n] \downarrow \emptyset$, which means that

$$(x, y_1] \supseteq (x, y_2] \supseteq ...$$

Then, we have that

$$\bigcap_n (x, y_n] = (x, x] = \emptyset.$$

Now, applying the CDF to this

$$F(y_n) - F(x) = \mathbb{P}((x, y_n]) \to 0$$

and therefore we have that

$$F(y_n) \downarrow F(x).$$

■

Theorem 16: Every CDF is associated to a random variable

If a function F satisfies

1. F is non-decreasing;
2. $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$
3. F(x) is right continuous

then it is the CDF of some random variable.

Remark 4.77 *Recall that if a function F satisfies (1) and (3), then there exists a unique measure on $(\mathbb{R}, \mathbb{B})$ with $\mu[a, b] = F(b) - F(a)$ which we called the Lebesgue-Stieltjes measure. Let $(\Omega, \mathcal{F}, P) = (\mathbb{R}, \mathbb{B}, \mu)$ and define $X(\omega) = \omega$. Then, P is a probability measure and $F_X = F$*

Definition 26: Equal in distribution

Let X and Y be random variables defined on the same measurable spaces $(\Omega, \mathcal{F})$. We say that X and Y are equal in distribution if they induce the same distribution μ on $(\mathbb{R}, \mathbb{B})$. That is

$$P(X \leq x) = F_X(x) = \mu((-\infty, x]) = F_Y(x) = P(Y \leq x)$$

for all $x \in \mathbb{R}$.

STAT4028: Probability and Mathematical Statistics

5. Independence

5.3 Independence

5.3.1 Independence of sets of events

We first recall definitions from elementary probability theory. Remember, that we can interpret $\sigma-$algebras as description of information. What we will do is to generalise the notions of independence of events to the notion of independence of information.

Definition 27: Independence of a finite number of events

Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. The events $A_1, A_2, ..., A_n$ are independent if

$$\mathbb{P}(\bigcap_{i\in I} A_i) = \prod_{i\in I} \mathbb{P}(A_i)$$

for all finite $I \subset \{1, ..., n\}$.

Remark 5.78 *Note that for the criterion of independence of a finite number of events, we require*

$$\sum_{k=2}^{n} \binom{n}{k} = 2^n - n - 1$$

equations to hold in order for the collection of events to be independent.

These notions of independence of sets hold for their complements too.

Claim 5.79 *If $A_1, ..., A_n$ are independent events, then so are $A_1^c, A_2, ..., A_n$.*

Proof: We have that

$$\mathbb{P}(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} \mathbb{P}(A_i)$$

So now, let A_1^c be our first event. Then, we have that

$$\mathbb{P}(A_1^c \cap \Big(\bigcap_{i=2}^{n} A_i\Big)) = \mathbb{P}(\bigcap_{i=2}^{n} A_i \setminus (\cap_{i=2}^{n} A_i \cap A_1)) = \mathbb{P}(\cap_{i=2}^{n} A_i) - \mathbb{P}(\cap_{i=2}^{n} A_i \cap A_1)$$

$$= \prod_{i=2}^{n} \mathbb{P}(A_i) - \prod_{i=1}^{n} \mathbb{P}(A_i) = \prod_{i=2}^{n} \mathbb{P}(A_i)\Big[1 - \mathbb{P}(A_1)\Big] = \prod_{i=2}^{n} \mathbb{P}(A_i)\mathbb{P}(A_1^c)$$

■

Claim 5.80 *If $A_1, ..., A_n$ are independent if and only if $B_1, ..., B_n$ are independent, where for each i, $B_i = A_i$ or $B_i = A_i^c$.*

Proof: Use induction to repeatedly apply the previous claim. ■

Proposition 21: Equivalent formulation of independence of events

Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. The events $A_1, A_2, ..., A_n$ are independent if

$$P(\bigcap_{i=1}^{n} B_i) = \prod_{i=1}^{n} P(B_i)$$

where for each i = 1,...,n, B_i equals either A_i or Ω.

We can extend the notion of independence of events to independence for **sets of events**. However, recall that on our probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we have 1 main $\mathcal{F}$ and we can construct sub-$\sigma - algebra$ from it.

Definition 28: Sub-σ-algebra

Let $(\Omega, \mathcal{F}, P)$ be a probability space. A set $g \subset 2^{\Omega}$ is a sub$-\sigma-$algebra of $\mathcal{F}$ if g is a σ-algebra and $g \subset \mathcal{F}$.

Definition 29: Independence of sub-$\sigma-$algebra

Let $(\Omega, \mathcal{F}, P)$ be a probability space. The sub-$\sigma-$algebras $g_1, ..., g_n$ are independent if whenever $A_i \in g_i$ for i=1,...,n, we have that

$$P(\bigcap_{1}^{n} A_i) = \prod_{1}^{n} P(A_i).$$

Recall earlier definitions of random variables.

Definition 5.81 *(Independence of random variables). Random variables $X_1, ..., X_n$ are independent if for any Borel set $B_i \in \mathcal{B}$*

$$P(X_1 \in B_1, ..., X_n \in B_n) = \prod_{i=1}^{n} P(X_i \in B_i) = \prod_{1}^{k} P(\{\omega : X_i(\omega) \in B_i\}).$$

Proposition 5.82 *(Sub-σ − algebra-generated by indicator function). Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Let 1_A be the indicator function for the measurable set $A \in \mathcal{F}$. Then, the sub-σ − algebra generated by the indicator function is*

$$\sigma(1_A) = \{\emptyset, \Omega, A, A^c\}.$$

If we had a collection $\{\mathcal{F}_i\}_{i \in I}$ of sub-$\sigma-$algebras, we can think that each $\mathcal{F}_i$ represents some information. We now make teh connection between independent events and random variables.

Proposition 22: Relating independent events to independent random variables

For $A_1, ..., A_n \in \mathcal{F}$, the following are equivalent:

1. A_i are independent events;
2. The indicator functions 1_{A_i} are independent random variables;
3. The sub-σ-algebras generated by the indicator function 1_{A_i}, given by $g_i = \sigma(A_i) = \{\emptyset, \Omega, A_i, A_i^c\}$ are independent for $i \in I$.

Proof: $(1 \to 2)$. First, recall that

$$\{1_A = a\} = \{\omega : 1_A(\omega) = a\} = \begin{cases} A & \text{if } a = 1 \\ A^c & \text{if } a = 0 \\ \emptyset & \text{if } a \neq 0, 1 \end{cases}$$

Therefore, we can say that

$$\mathbb{P}(1_A = 1) = \mathbb{P}(A) \quad \mathbb{P}(1_A = 0) = \mathbb{P}(A^c)$$

WLOG, let us look at n = 2. Then,

$$\mathbb{P}(1_{A_1} = 1, 1_{A_2} = 1) = \mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) = \mathbb{P}(1_{A_1} = 1) \cdot \mathbb{P}(1_{A_2} = 1)$$

since A_1, A_2 are independent by assumption. Likewise, we can do something similar for

$$\mathbb{P}(1_{A_1} = 0, 1_{A_2} = 1) = \mathbb{P}(A_1^c \cap A_2) = \mathbb{P}(A_1^c) \cdot \mathbb{P}(A_2) = \mathbb{P}(1_{A_1} = 0) \cdot \mathbb{P}(1_{A_2} = 1)$$

We can repeat this for all other cases to show that 1_{A_1} and 1_{A_2} are independent random variables. Furthermore, we can show by induction that this holds for $n \geq 2$ independent events.

$(2 \to 3)$. We have that the sub-$\sigma - algebra$ generated by the indicator function 1_{A_i} is

$$\sigma(1_{A_i}) = \{\emptyset, \Omega, A_i, A_i^c\}.$$

Define $B_i = \{\emptyset, \Omega, A_i, A_i^c\}$ where $B_i \in g_i$. Then, as 1_{A_i} are independent, then A_i are independent. From earlier results, we then have that B_i are independent. Therefore, we have that

$$\mathbb{P}(\bigcap_{i=1}^{n} B_i) = \prod_{i=1}^{n} \mathbb{P}(B_i)$$

for $B_i \in g_i$ and therefore the sub-$\sigma - algebra$ g_i are independent.

$(3 \to 1)$. For each g_i, define $B_i = A_i$. Then, as g_i is independent, then

$$\mathbb{P}(\bigcap_{i=1}^{n} A_i) = \mathbb{P}(\bigcap_{i=1}^{n} B_i) = \prod_{i=1}^{n} \mathbb{P}(B_i) = \prod_{i=1}^{n} \mathbb{P}(A_i)$$

and therefore A_i are independent events. ■

We can come up with a new notion of independence for random variables.

Theorem 17: Independence of random variables

The collection of random variables $\{X_i\}_{1\leq i\leq n}$ are independent if and only if the collection of sub-$\sigma-$algebras generated by random variables

$$\{\sigma(X_i)\}_{1\leq i\leq n}$$

are independent.

Proof:(Sketch). As $X_1, ..., X_n$ are independent, then for any Borel set $B_i \in \mathcal{B}$, we have that

$$P(X_1 \in B_1, ..., X_n \in B_n) = \prod_{i=1}^{n} P(X_i \in B_i) = \prod_{1}^{k} P(\{\omega : X_i(\omega) \in B_i\}).$$

As $\{X \in B_j\} = \sigma(X)$ for all Borel sets $B_j \in \mathcal{B}$, it then follows that $\{\sigma(X_i)\}_{1\leq i\leq n}$ are independent. ■

Theorem 18: Indepedent random variables and $\sigma - algebra$

If $\mathcal{G}_1, ..., \mathcal{G}_n$ are independent $\sigma-$algebras and if $X_i \in \mathcal{G}_i$ (i.e. $\sigma(X_i) \subset \mathcal{G}_i$) then $X_1, ..., X_n$ are **independent** random variables.

Proof: As $\sigma(X_i) \subset \mathcal{G}_i$ and $\mathcal{G}_i$ are independent, then $\sigma(X_i)$ are independent. Therefore, from the previous result, X_i are also independent. ■

We can now extend the notion of independence to sets of events $\mathcal{A}_1, ..., \mathcal{A}_n$.

Definition 30: Independence for sets of events

Define the probabiliy space $(\Omega, \mathcal{F}, \mathbb{P})$. Define sets of events $\mathcal{A}_1, ..., \mathcal{A}_n$, where for each i, we have that $\mathcal{A}_i \subset \mathcal{F}$ and $\Omega \in \mathcal{F}_i$. Then, $\mathcal{A}_i$ are independent if for any choice of $A_i \in \mathcal{A}_i$ for i = 1,...,n, we have that

$$\mathbb{P}(\cap_{i=1}^{n} A_i) = \prod_{i=1}^{n} \mathbb{P}(A_i).$$

Proposition 5.83 *The sets of events $\mathcal{A}_1, ..., \mathcal{A}_n$ are independent if and only if whenever $A_i \in \mathcal{A}$ for i = 1,...,n are independent events.*

Proof: This follows by definition. ■

We can now describe a result which allows for us to test for independence in an efficient manner.

Theorem 5.84 *(Test for independence). Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Suppose that for i = 1,...,n we have that the collection of sets, $\mathcal{A}_i \subset \mathcal{F}$ which satisfies*

1. $\Omega \in \mathcal{A}_i$
2. *$\mathcal{A}_i$ is a $\pi-$system*
3. *$\mathcal{A}_1, ..., \mathcal{A}_n$ are independent.*

Then, $\sigma(\mathcal{A}_1), \mathcal{A}_2, ..., \mathcal{A}_n$ are independent.

Proof:(Sketch). First, for n = 2, fix an event $A_2 \in \mathcal{A}_2$. Then, define the set of all independent events

$$\mathcal{L} = \{A \in \mathcal{F} : \mathbb{P}(AA_2) = \mathbb{P}(A)\mathbb{P}(A_2)\}$$

Then, we can show that $\mathcal{L}$ is a $\lambda-$system. Furthermore, it is clear that $\mathcal{A}_1 \subset \mathcal{L}$. Then, by Dynkin's $\pi - \lambda$ theorem, we have that

$$\sigma(\mathcal{A}_1) \subset \mathcal{L}$$

using the fact that $\mathcal{A}_1$ is a $\pi-$system. Therefore, $\sigma(\mathcal{A}_1), \mathcal{A}_2$ are independent. We can then extend this argument by induction to show that $\sigma(\mathcal{A}_1), \mathcal{A}_2, ..., \mathcal{A}_n$ are independent. ■

The following corollary is what we really care about.

Theorem 19: Criterion for independence

Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Let $\mathcal{A}_i$ be a collection of classes of events satisfying:

1. $\Omega \in \mathcal{A}_i$
2. $\mathcal{A}_i$ is a $\pi-$system
3. $\mathcal{A}_1, ..., \mathcal{A}_n$ are independent.

Then, we have that

$$\sigma(\mathcal{A}_1), \sigma(\mathcal{A}_2), ..., \sigma(\mathcal{A}_n)$$

are independent classes of events.

A typical example of the application of our independence critera is the independence of random variables using their cumulative distribution functions.

Theorem 20: Independence of random variables through CDF

The random variables $X_1, ..., X_n$ are independent if and only if their CDFs

$$P(\bigcap_1^n \{X_i \leq x_i\}) = \prod_1^n P(X_i \leq x_i)$$

for all $x_i \in \overline{\mathbb{R}}$.

Proof: (Sketch). First, recall the fact that $\{(-\infty, x] : x \in \mathbb{R}\}$ is a $\pi-$system on $\mathbb{R}$. Then, $\mathcal{A}_i \subset \mathcal{F}$ and $\sigma(\mathcal{A}_i) = \sigma(X_i)$. It then follows by the criterion for independence that $X_1, ..., X_n$ are independent. ■

5.3.2 Second Borel-Cantelli Lemma and Kolmogorov's 0-1 Law

Recall that the first Borel-Cantelli lemma states that whenever the probabilities of the events A_n decay fast enough, then it is (almost surely) impossible for the events to occur infinitely often.

The second Borel-Cantelli lemma states that if the probabilities of the events A_n do not decay fast, and additionally if the events are independent, then the events must (almost surely) occur infinitely often.

Theorem 21: The second Borel-Cantelli Lemma

Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Let $A_n \in \mathcal{F}$ be a sequence of **independent** events such that $\sum_{k=1}^{\infty} P(A_k) = \infty$. Then,

$$\mathbb{P}(A_n \text{ i.o.}) = 1.$$

Proof: Define $p_n = \mathbb{P}(A_n)$. Then, we have that

$$\mathbb{P}(\bigcap_{n=m}^{N} A_n^c) = \prod_{n=m}^{N} \Big(1 - \mathbb{P}(A_n)\Big) \leq \prod_{n=m}^{N} e^{-\mathbb{P}(A_n)} = e^{-\sum_{n=m}^{N} \mathbb{P}(A_n)} \to 0$$

as $n \to \infty$. Therefore, we have that

$$\mathbb{P}(\bigcap_{n=m}^{\infty} A_n^c) = 0.$$

Then, recall that $(A_n \text{ i.o.})^c = (A_n^c \text{ eventually})$ and therefore

$$\mathbb{P}((A_n \text{ i.o.})^c) = \mathbb{P}(\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n^c) = 0$$

and therefore, we can conclude that

$$\mathbb{P}((A_n \text{ i.o.})) = 1.$$

■

Remark 5.85 *This is known as Borel zero-one law in some places.*

5.3.3 Kolmogorov's 0-1 law

We now work towards Kolmogorov's 0-1 law which states that under the independence assumption, any event which is not sensitive to finitely many individual values, has probability either zero or one.

Let $\{X_i\}_{i \geq 1}$ be a sequence of random variables. We can define

$$\sigma(X_1, X_2, ...) = \sigma(\{X_i\}_{i \in \mathbb{N}})$$

and interpret this as the information contained in the entire sequence.

We note that

$$\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, ...)$$

can be interpreted as the information contained in the sequence without its first n members. That is, $\mathcal{T}_n$ can be interpreted as the information in the sequence which is not affected by the first n values.

Definition 31: Tail $\sigma-$algebra

Let $X_1, X_2, ...$ be random variables and define

$$\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, ...)$$

for $n \in \mathbb{N}$. Then, define

$$\mathcal{T} = \bigcap_{n=0}^{\infty} \mathcal{T}_n.$$

We have that $\mathcal{T}$ is called the **tail $\sigma-$algebra** of the sequence $\{X_n\}$. The events $A \in \mathcal{T}$ are called **tail events**.

Recall that the intersection of $\sigma-$algebras is a $\sigma-$algebra. As $\mathcal{T}$ contains less information than any $\mathcal{T}_k$ for $k \in \mathbb{N}$, it describes the information which is not affected by any finite number of changes in the sequence.

Lemma 5.86 *Let $\mathcal{T}$ be the tail $\sigma-$algebra. We have that*

$$\mathcal{T} \subset ... \subset \mathcal{T}_{k+1} \subset \mathcal{T}_k \subset ... \subset \mathcal{T}_2 \subset \mathcal{T}_1.$$

That is, we have that the tail $\sigma-$algebra $\mathcal{T} \subset \sigma\Big(\{X_i\}_{i \in I}\Big)$.

Lemma 5.87 *We have that*

$$\mathcal{T} = \bigcap_{n=m}^{\infty} \mathcal{T}_n$$

for any $m \in \mathbb{N}$.

We give some examples of tail events and random variables.

Claim 5.88 *Show that the following are tail events in $\mathcal{T}$*

1. $\{\omega : \lim_n X_n(\omega)$ *exists*$\}$
2. $\{\omega : \sum_{i=1}^{\infty} X_n(\omega)$ *converges*$\}$
3. $\{\omega : \lim_n \frac{S_n(\omega)}{n}$ *exists*$\}$
4. $\{\omega : \overline{\lim} \frac{S_n(\omega)}{n} > 0\}$

where $S_n = \sum_{i=1}^{n} X_i$.

Proof: (1)
(2) First, recall that $\sum_{n=1}^{\infty} X_n(\omega)$ converges if and only if $\sum_{n=m}^{\infty} X_n(\omega)$ converges. Therefore

$$(\sum_{n}^{\infty} X_n \text{ converges}) = (\sum_{n=m+1}^{\infty} X_n \text{ converges}) \in \mathcal{T}_m$$

whereby this holds for all m and after intersecting over m. ■

We now have an important result regarding the probabilities of tail events.

Theorem 22: Kolmogorov's 0-1 Law

Suppose that $X_1, X_2, ...$ is a sequence of independent random variables. Then for any tail event $A \in \mathcal{T}$, we have that either

$$P(A) = 0 \quad \text{or} \quad P(A) = 1.$$

We now look at some of the implications of Kolmogorov's 0-1 Law.

Theorem 5.89 *Suppose that $X_1, X_2, ...$ is a sequence of independent random variables. Then, any $\mathbb{R}-$valued random variable Y that is measurable with respect to the tail σ-algebra $\mathcal{T}$ is* ***almost surely constant****, that is, there exists some $c \in \mathbb{R}$, such that*

$$P(Y = c) = 1.$$

Proof:(Sketch). Use Kolmogorov's 0-1 law where Y is a $\mathcal{T}-$measurable function. Then, define $c \coloneqq inf\{x \in \mathbb{R} : P(Y \leq y) = 1\}$. Then, by the right-continuity of CDFs, we can get the desire result where $P(Y = c) = P(Y \leq c) - P(Y < c)$. ■

Corollary 5.90 *We have the following 3 corollaries.*

1. *If $A_1, A_2, ...$ are independent events, then $P(A_n \text{ i.o.}) \in \{0, 1\}$.*
2. *Define $S_n = \sum_1^n X_k$. Then $P(limS_n \text{ exists}) \in \{0, 1\}$.*
3. *$P(lim\frac{S_n}{n} = \mu) \in \{0, 1\}$.*

STAT4028: Probability and Mathematical Statistics

6. Lebesgue Integration

6.4 Integration

6.4.1 Integration of Simple Functions

We are now interested in constructing integrals over probability spaces.

Definition 32: Simple functions

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. The function f is a **simple non-negative function** $f : \Omega \to \overline{\mathbb{R}}$ if

$$f(\omega) = \sum_{i=1}^{n} \alpha_i 1_{A_i}(\omega)$$

where A_i are **disjoint sets** with $A_i \in \mathcal{F}$ and $\alpha_i \in \mathbb{R}^+$.

Remark 6.91 *The representation of f is not unique since we do not assume that α_i is unique.*

Remark 6.92 *If A_i are not disjoint, we can set $B_k = A_k \cap (A_1 \cup A_2 \cup ... \cup A_{k-1})^c$. Then, $A_k = \bigcup_{j=0}^{n} A_k \cap B_j$ is a disjoint union. Hence, $1_{A_k} = \sum_{j=0}^{n} 1_{A_k \cap B_j}$. Therefore, our simple function can be expressed as*

$$\sum_{k=0}^{n} \alpha_k 1_{A_k} = \sum_{k=0}^{n} \sum_{j=0}^{n} \alpha_k 1_{A_k \cap B_j}.$$

Lemma 6.93 *A simple function $f : (\Omega, \mathcal{F}) \to (\mathbb{R}^+, \overline{\mathcal{B}})$ is a **measurable** function.*

Proof: Recall that f is measurable if and only if $f^{-1}((-\infty, a]) \in \mathcal{F}$ for all $a \in \mathbb{R}$. Then

$$f^{-1}((-\infty, a]) = \bigcup \{A_i : i \in \{1, 2, ..., n\} \text{ such that } \alpha_i \leq a\}$$

As each A_i is measurable, then the union is measurable and hence f is measurable. ■

Definition 33: Space of simple non-negative functions

We denote $\mathcal{S}^+$ as the set of **simple non-negative functions**.

Remark 6.94 *The space of simple functions form a **vector space**.*

Proposition 23: Properties of Simple Functions

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $f, g \in \mathcal{S}^+$. Then, the following are also in $\mathcal{S}^+$:

1. $\alpha f + \beta g$ for all $\alpha, \beta > 0$;
2. $f \cdot g$;
3. $\frac{f}{g}$ if $g(x) \neq 0$ for all $x \in X$;
4. $min(f, g)$ and $max(f, g)$.

Proof:(Sketch). Recall that

$$max(f,g) = \frac{f+g+|f-g|}{2} \quad min(f,g) = \frac{f+g-|f-g|}{2}$$

Then show that the sum of simple functions is a simple function, the scalar multiple of a simple function is a simple function, and the modulus of a simple function is a simple function. ■

Definition 34: Lebesgue Integral of Simple Functions

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $f = \sum_{i=1}^{n} \alpha_i 1_{A_i}$ be a simple function. The Lebesgue integral of f with respect to μ is defined as

$$\int_\Omega f(\omega) d\mu(d\omega) = \int_\Omega f d\mu = \sum_{i=1}^{n} \alpha_i \mu(A_i).$$

where we define $0 \cdot \infty = \infty \cdot 0 = 0$.

Example 6.95 *(Lebesgue and Riemann integral agree on indicator function). Let $(\Omega, \mathcal{F}, \mu) = (\mathbb{R}, \mathbb{B}, \lambda)$ where λ is the Lebesgue measure. Then, define $f = \alpha 1_{(a,b)}$ where $\alpha > 0$ and $-\infty < a < b < \infty$. Then the Lebesgue integral is the same as the Riemann integral.*

$$\textit{(Lebesgue)} \int f d\mu = \alpha\lambda((a,b)) = \alpha(b-a) = \int_{\mathbb{R}} f(x)dx \textit{(Riemann)}$$

We have a potential issue as if a simple function has 2 different representation, the integrals may not be the same.

Proposition 24: Integral of simple functions is well-defined

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $f : \Omega \to \overline{\mathbb{R}}^+$ be a positive simple function. Then, for any two standard representations for f, that is

$$f = \sum_{i=0}^{n} a_i 1_{A_i} = \sum_{j=0}^{m} b_j 1_{B_j}$$

for $E_i, F_j \in \mathcal{F}$. It holds that

$$\sum_{i=0}^{n} a_i \mu(A_i) = \sum_{j=0}^{m} b_j \mu(B_j).$$

That is, the integral $\int_\Omega f d\mu$ is well-defined.

Proof:(Sketch). A simple function f may have two different representations

$$f = \sum_{i=0}^{n} a_i 1_{A_i} = \sum_{j=0}^{m} b_j 1_{B_j}$$

However, the integral may give different results, which is undesirable

$$\sum_{i=0}^{n} a_i \mu(A_i) \neq \sum_{j=0}^{m} b_j \mu(B_j).$$

However, first recall we can partition the sample space Ω up

$$\Omega = \cup_i \cup_j E_{ij} = \cup_i \cup_j (A_i \cap B_j)$$

and that

$$A_i = \bigcup_{j=0}^{m} (A_i \cap B_j) \quad B_j = \bigcup_{i=0}^{n} (B_j \cap A_i)$$

Therefore, we have that

$$\int f d\mu = \sum_{i=0}^{n} a_i \mu(A_i) = \sum_{i=0}^{n} a_i \mu(\bigcup_{j=0}^{m} (A_i \cap B_j)) = \sum_{i=0}^{n} \sum_{j=0}^{m} a_i \mu(A_i \cap B_j)$$

$$= \sum_{j=0}^{m} b_j 1_{B_j} = \int f d\mu$$

■

Proposition 25: Elementary Properties of Integrals

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $f, g \in \mathcal{S}^+$. Then, we have that

1. $\int (\alpha f + \beta g) d\mu = \alpha \int f d\mu + \beta \int g d\mu$;
2. If $f \leq g$, then $\int f d\mu \leq \int g d\mu$;

Proof:(Sketch). Let $f = \sum_{i=1}^{n} \alpha_i 1_{A_i}$ and $g = \sum_{j=1}^{m} \beta_j 1_{B_j}$. We then let $A_0 = \left(\cup_1^n A_i\right)^c$ and $B_0 = \left(\cup_1^n B_j\right)^c$. We also set $\alpha_0 = \beta_0 = 0$. Then, we have

$$\begin{cases} f = \sum_{i=0}^{n} \alpha_i 1_{A_i} \\ g = \sum_{j=0}^{m} \beta_j 1_{B_j} \end{cases}$$

Now, $\{A_i\}_{i=0}^n$ and $\{B_j\}_{j=0}^m$ are disjoint partitions of Ω. Then $(A_i \cap B_j) \in \mathcal{F}$ and is also a partition of Ω.

Hence

$$f + g = \sum_{i=0}^{n} \sum_{j=0}^{m} (\alpha_i \beta_j) 1_{A_i \cap B_j}.$$

Then recalling that $A_i = \cup_{j=0}^{m} (A_i \cap B_j)$, we can get the result that

$$\int (f+g) d\mu = \sum_{i=0}^{n} \alpha_i \mu(A_i) + \sum_{j=0}^{m} \beta_j \mu(B_j) = \int f d\mu + \int g d\mu$$

Additionally

$$\int \alpha f d\mu = \sum_{i=0}^{n} \alpha \alpha_i \mu(A_i) = \alpha \sum_{i=0}^{n} \alpha_i \mu(A_i)$$

Finally, as $f \leq g$, this means that $g - f \geq 0$. Then

$$\int g d\mu = \int f d\mu + \int (g-f) d\mu \geq \int f d\mu.$$

■

We now define what it means for integrals to not see sets of measure zero.

Proposition 26: Integrals do not see sets of measure zero

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $f \in \mathcal{S}^+$ Let N be a set of measure zero such that $f(\omega) = 0$ for all $\omega \in N$. Then

$$\int_{\Omega} f d\mu = \int_{X \setminus N} f d\mu.$$

Proof: First, note that $f = 1_{\Omega \setminus N} f + 1_N f$

$$\int_{\Omega} f d\mu = \int_{\Omega} 1_{\Omega \setminus N} f + 1_N f d\mu = \int_{\Omega} 1_{\Omega \setminus N} f d\mu + \int_{\Omega} 1_N f d\mu$$

Now, we note that

$$\int_{\Omega} 1_N f d\mu = \sum_{i=0}^{n} 1_N \alpha_i \mu(A_i \cap N) = \sum_{i=0}^{n} \alpha_i \mu(A_i \cap N) = 0$$

because $\mu(A_i \cap N) \leq \mu(N) = 0$. ■

6.4.2 Integration of non-negative measurable functions

Definition 35: Set of all non-negative measurable functions

We define $\mathcal{F}^+$ as the set of all $f : \Omega \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ that are measurable and non-negative.

Claim 6.96 *The set of all non-negative measurable functions $\mathcal{F}^+$ is a vector space.*

We now state that every measurable function can be approximated by simple measurable functions. This will be very important for us when we want to take Lebesgue integrals of non-negative functions.

Proposition 27: Simple Approximation Theorem

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $f : \Omega \to [0, \infty]$ be a measurable function. Then, there exists a sequence of simple measurable functions $\phi : \Omega \to [0, \infty)$ such that

$$0 \leq \phi_n(\omega) \leq \phi_{n+1}(\omega) \leq f(\omega)$$

for all $n \in \mathbb{N}$ and all $\omega \in \Omega$.
Moreover, $\phi_n(\omega) \to f(\omega)$ as $n \to \infty$ for all $\omega \in \Omega$. That is, $\phi_n \to f$ pointwise.

With the Simple approximation theorem, we can now define Lebesgue integrals of non-negative functions.

Definition 36: Lebesgue Integral of non-negative functions

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. The **Lebesgue integral** of $f \in \mathcal{F}^+$ with respect to μ is defined as

$$\int_\Omega f d\mu = \sup \left\{ \int_\Omega h d\mu : h \leq f \text{ and } h \in \mathcal{S}^+ \right\}$$

where $\mathcal{F}^+$ is the space of non-negative measurable functions and $\mathcal{S}^+$ is the space of simple measurable functions.

Claim 6.97 *If $f \in \mathcal{S}^+$, then the definition of the Lebesgue integral of simple functions agrees with the Lebesgue integral of non-negative functions.*

Proposition 28: Properties of the Lebesgue integral of $\mathcal{F}^+$ functions

Let $f, g \in \mathcal{F}^+$ be measurable non-negative functions and $\alpha, \beta \in \geq 0$. Then

1. $f \leq g$ implies that $\int f d\mu \leq \int g d\mu$
2. $\int (\alpha f + \beta g) d\mu = \alpha \int f d\mu + \beta \int g d\mu$.

Proof:(Sketch). First, using simple measurable functions such that $0 \leq \phi \leq f$ and $0 \leq \psi \leq g$, then by definition of supremum

$$\int f d\mu \leq \int g d\mu$$

For the second result, again express the functions in terms of their simple functions and use previously shown properties of Lebesgue integral of simple functions. ■

Proposition 6.98 *(Integral of functions equal almost everywhere). Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $f, g \in \mathcal{F}^+$. Let N be a set of measure zero where $f(\omega) = g(\omega)$ for all $\Omega \backslash N$. Then $f = g$* ***almost everywhere****. Furthermore,*

$$\int_\Omega f d\mu = \int_\Omega g d\mu.$$

Proof: Using the fact that the Lebesgue integrals does not see sets of measure zero, we have that

$$\int_\Omega f d\mu = \int_\Omega 1_{\Omega \setminus N} f d\mu = \int_\Omega 1_{\Omega \setminus N} g d\mu = \int_\Omega g d\mu.$$

■

Proposition 29: A function is zero almost everywhere

Let $f \in \mathcal{F}^+$ be a non-negative measurable function. Then, $\int_\Omega f d\mu = 0$ if and only if $f = 0$ almost everywhere.

Remark 6.99 *As a result, if the integral of a non-negative function f is zero, we **cannot** conclude that the function f is the zero function.*

Proposition 30: Functions are equal almost everywhere

Let $f, g \in \mathcal{F}^+$ be non-negative measurable functions. Assume that $f = g$ almost everywhere. Then

$$\int f d\mu = \int g d\mu.$$

We are now interested in figuring out how to actually compute the integral $\int f d\mu$. We show an example of when we cannot switch the order of limits and integration.

Example 6.100 *Let $(\Omega, \mathcal{F}, \mu) = ([0,1], \mathcal{B}, \lambda)$. Let $f_n = n 1_{0,1/n}$. Then*

$$\int f_n d\mu = n[\frac{1}{n} - 0] = 1$$

but this does not converge to

$$\int f d\mu = 0$$

As a result of this example, we need to be careful to find a sequence of functions $f_n \uparrow f$ and therefore $\int f_n d\mu = \int f d\mu$. We state the condition for when we are allowed to interchange the limit and integral.

Theorem 23: Weak Monotone Convergence Theorem

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Suppose that $\{f_n\}$ is a sequence of simple non-negative functions $f_n \in \mathcal{S}^+$. Furthermore, assume that $f_n \uparrow f$ pointwise. Then $f \in \mathcal{F}^+$ is a non-negative measurable function. Furthermore, we have that

$$\lim_{n\to\infty} \int_\Omega f_n d\mu = \int_\Omega \lim_{n\to\infty} f_n d\mu = \int_\Omega f d\mu.$$

Remark 6.101 *The "weak" part comes from the fact that $f_n \in \mathcal{S}^+$. We can later make this $f_n \in \mathcal{F}^+$.*

Claim 6.102 *The lemma for the weak monotone convergence theorem holds for $f = 1_{A_i}$ where $A_i \in \mathcal{F}$.*

Claim 6.103 *The lemma for the weak monotone convergence theorem holds for $f \in \mathcal{S}^+$.*

STAT4028: Probability and Mathematical Statistics

7. Convergence Theorems and Integrability

7.4.3 Convergence Theorems

First, we recall the definition of pointwise convergence.

Definition 7.104 *(Pointwise Convergent). Let $\{f_n\}_{n\geq 1}$ be a sequence of functions on the domain X. Then, $\{f_n\}_{n\geq 1}$ is said to be **Pointwise Convergent** to the function f if $\lim_{n\to\infty} f_n(x) = f(x)$ if for all $x \in X$. That is, for all $\epsilon > 0$, there exists a $N = N(\epsilon, x) \in \mathbb{N}$ such that if $n \geq N$, then*

$$|f_n(x) - f(x)| < \epsilon.$$

Also, recall the following sufficient condition for convergence of a sequence.

Claim 7.105 *(Sufficient conditions for convergence of a sequence). If the sequence $\{a_n\}_{n\geq 1}$ is bounded and monotonic, then the sequence $\{a_n\}_{n\geq 1}$ is convergent.*

We state some important theorems. The MCT is one of the key results of Lebesgue integrals.

Theorem 24: Monotone Convergence Theorem

Let $(X, \mathcal{A}, \mu)$ be a measure space. For every $n \in \mathbb{N}$, let $f_n \in \mathcal{F}^+$ and $0 \leq f_n(x) \leq f_{n+1}(x)$ for almost every $x \in X$. Then, there exists a measurable function $f \in \mathcal{F}^+$ such that $f_n(x) \to f(x)$ as $n \to \infty$ for almost every $x \in X$ and

$$\lim_{n\to\infty} \int_X f_n d\mu = \int_X \lim_{n\to\infty} f_n d\mu = \int_X f d\mu.$$

Remark 7.106 *In lectures, we did not extend the MCT to holding almost everywhere when also considering null sets, but we do did this in the tutorials.*

We can also state a theorem that applies to **arbitrary** sequences of non-negative functions by using the limit inferior.

Theorem 25: Fatou's Lemma

Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. Let $f_n \in \mathcal{F}^+$ be non-negative measurable functions. Then

$$\int_\Omega \lim_{n\to\infty} \inf f_n d\mu \leq \lim_{n\to\infty} \inf \int_\Omega f d\mu.$$

We can state something similar using the limit superior.

Theorem 26: Reverse Fatou's Lemma

Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. Let $f_n \in \mathcal{F}^+$ be non-negative measurable functions. Furthermore, assume that $f_n \leq g$ where $g \in \mathcal{F}^+$ and $\int g d\mu < \infty$. Then

$$\int_\Omega \overline{\lim} f_n d\mu \geq \overline{\lim} \int_\Omega f d\mu.$$

Remark 7.107 *Note that we require a dominating function in the reverse-Fatou lemma.*

7.4.4 Relationship between Riemann and Lebesgue Integrals

First, we recall the Riemann integral. A Riemann integral is defined on a bounded real-valued function $f : [a, b] \to \mathbb{R}$ and the domain of f is a **compact** interval.

Definition 7.108 *(Finite partition). A finite partition of [a,b] is defined as* $\mathcal{P} = \{a = x_0 < x_1 < ... < x_n = b\}$. *Let* $m_k = \inf_{x_{k-1} \leq x \leq x_k} f(x)$ *and* $M_k = \sup_{x_{k-1} \leq x \leq x_k} f(x)$ *for* $1 \leq k \leq n$.

Definition 7.109 *(Lower and upper Riemann sums). We define the upper and lower Riemann sums as*

$$\underline{I}(f, \mathcal{P}) = \sum_{k=1}^{n} m_k \Delta_k$$

$$\overline{I}(f, \mathcal{P}) = \sum_{k=1}^{n} M_k \Delta_k$$

where $\Delta_k = x_k - x_{k-1}$ *for* $1 \leq k \leq n$.

Definition 7.110 *(Lower and upper Riemann Integrals). Finally, we define the lower and upper Riemann integrals as*

$$\underline{I} = \sup_{\mathcal{P}} \underline{I}(f, \mathcal{P})$$

$$\overline{I} = \inf_{\mathcal{P}} \overline{I}(f, \mathcal{P}).$$

Definition 7.111 *(Riemann Integrable). A function f is said to be Riemann integrable if*

$$I(f) = \underline{I}(f) = \overline{I}(f).$$

We recall some things from calculus.

Proposition 7.112 *Every continuous function* $f \in C[a, b]$ *is Riemann integrable.*

We will now show that we can evaluate Lebesgue integrals using Riemann integrals. We will set the domain as $[0, 1]$ but the following argument holds for any domain where $[a, b]$ for $a, b \geq 0$.

Proposition 31: Simple functions approximating Riemann integral

Suppose that $f : [0,1] \to \mathbb{R}^+$ is a properly Riemann integrable function with $\int_0^1 f(x)dx$. Then define the measure space $([0,1], \mathcal{B}, \lambda)$ where $\mathcal{B}$ is the Borel $\sigma - algebra$ and λ is the Lebesgue measure. Then, there exists sequences $L_n, U_n \in \mathcal{S}^+$ of simple non-negative functions such that

1. $L_n \uparrow L \leq f$ and $U_n \downarrow U \geq f$;
2. $\int_{[0,1]} L_n d\lambda \uparrow \int_0^1 f(x)dx$ and $\int_{[0,1]} U_n d\lambda \downarrow \int_0^1 f(x)dx$.

Proof:(Sketch). We take

$$L_n = \sum_{k=0}^{2^n-1} \inf_{x \in I_k} f(x) 1_{I_k}$$

$$U_n = \sum_{k=0}^{2^n-1} \sup_{x \in I_k} f(x) 1_{I_k}$$

where the interval is $I_k = (\frac{k}{2^n}, \frac{k+1}{2^n})$. Both $L_n, U_n \in \mathcal{S}^+$. Furthermore,

$$\int_{[0,1]} L_n d\lambda = \sum_{k=0}^{2^n-1} \inf_{x \in I_k} f(x) \mu(1_{I_k}) = \sum_{k=0}^{2^n-1} \inf_{x \in I_k} f(x) \frac{1}{2^n} \to \int_{[0,1]} f(x)dx$$

by the definition of the Riemann integral. Likewise, a similar argument holds for U_n. ■

Claim 7.113 *The functions $U, L \in \mathcal{B}^+$. That is, they are positive functions measurable with respect to the Borel $\sigma - algebra$.*

Proof: The simple functions U_n, L_n are non-negative Borel-measurable functions and hence their pointwise limits are non-negative Borel-measurable functions. ■

We have that $L \leq U$ and by monotonicty of integrals

$$\int L_n d\lambda \leq \int L d\lambda \leq \int U d\lambda \leq \int U_n d\lambda.$$

Furthermore, we have shown that $\int_{[0,1]} L_n d\lambda \uparrow \int_0^1 f(x)dx$ and $\int_{[0,1]} U_n d\lambda \downarrow \int_0^1 f(x)dx$.

Lemma 7.114 *By the squeeze law, we have that*

$$\int_{[0,1]} L d\lambda = \int_{[0,1]} U d\lambda.$$

Corollary 7.115 *The functions*

$$L = U$$

almost everywhere.

Proof: We have that

$$\int_{[0,1]} L d\lambda = \int_{[0,1]} U d\lambda.$$

$$\int_{[0,1]} Ld\lambda - \int_{[0,1]} (U-L)d\lambda = \int_{[0,1]} Ud\lambda.$$

$$\int_{[0,1]} (U-L)d\lambda = 0$$

which means that $U - L = 0$ almost everywhere. Hence, this implies that $L = U$ almost everywhere. ■

So, as we defined $L \leq f \leq U$, from the previous lemma, we have shown that on $\{\omega : L(\omega) = U(\omega)\}$, we have that $L = f = U$. Furthermore, the set $\{\omega : L(\omega) \neq U(\omega)\}$ has measure zero.

However, we are unable to simply take the Lebesgue integral of f as we do not know if it is measurable with respect to the Lebesgue $\sigma - algebra$. We only know it is Riemann integrable. Therefore, we need to show that it is.

First, define the function

$$\tilde{f} = f \circ 1_{L=U} = L \circ 1_{L=U}$$

since $L \leq f \leq U$ means that $f = L$ when $L = U$.

Claim 7.116 *The function $\tilde{f} = L \circ 1_{L=U} \in \mathcal{B}^+$ is a non-negative function that is measurable with respect to the Borel $\sigma - algebra$. Furthermore, $\tilde{f} = L$ almost everywhere.*

Proof: We had that $L \in \mathcal{B}^+$ and the set $\{L = U\}$ was defined to be in the Borel $\sigma - algebra$. ■

From the previous claim, we have that

$$\int_{[0,1]} f(x)dx = \int_{[0,1]} Ld\lambda = \int_{[0,1]} \tilde{f}d\lambda.$$

We are now really close to showing that

$$\int_{[0,1]} f(x)dx = \int_{[0,1]} fd\lambda.$$

Proposition 32: Expressing Riemann integrable function as Borel integrable function

Let f be the Riemann integrable function and $\tilde{f} = f \circ 1_{L=U} \in \mathcal{B}^+$. Then $f = \tilde{f}$ almost everywhere. Furthermore,

$$f = \tilde{f} + f \circ 1_{L \neq U}.$$

Recall that the Lebesgue $\sigma - algebra$ $\mathcal{L}$ is the completion of the Borel $\sigma - algebra$ $\mathcal{B}$ with respect to the Lebesgue measure λ.

Lemma 7.117 *Let $([0,1], \mathcal{B}, \lambda)$ be a measure space. We can complete the measure space with respect to the Lebesgue measure to be $([0,1], \mathcal{L}, \lambda)$ where $\mathcal{L}$ is the Lebesgue $\sigma - algebra$.*

Theorem 27: Riemann integrable function is Lebesgue measurable

Let f be a Riemann integrable function on [0,1]. Then, the function $f : ([0,1], \mathcal{L}, \lambda) \to (\overline{\mathbb{R}}^+, \overline{\mathcal{B}})$ is measurable. That is, f is Lebesgue-measurable.

Proof: Take a Borel set $B \in \overline{\mathcal{B}}$. Then

$$f^{-1}(B) = f^{-1}(B) \cap \{\{L = U\} \cup \{L \neq U\}\}$$
$$= (f^{-1}(B) \cap \{L = U\}) \cup (f^{-1}(B) \cap \{L \neq U\})$$

Recall that $f = \tilde{f} + f \circ 1_{L \neq U}$ where $\tilde{f}$ is only defined on $\{L = U\}$

$$= (\tilde{f}^{-1}(B) \cap \{L = U\}) \cup (f^{-1}(B) \cap \{L \neq U\})$$

Now, we know that for $(\tilde{f}^{-1}(B) \cap \{L = U\})$, $\tilde{f}$ was shown to be measurable with respect to $\mathcal{B}^+ \subset \mathcal{L}$.

For the second term, we have that $f^{-1}(B) \cap \{L \neq U\} \subseteq \{L \neq U\}$. However, this was shown to be a set of measure zero in $\mathcal{B}^+$. Therefore, $f^{-1}(B) \cap \{L \neq U\} \subseteq \{L \neq U\} \in \mathcal{B}^+ \subseteq \mathcal{L}$.

Therefore, f is measurable with respect to the Borel $\sigma - algebra$. However, as $\mathcal{B} \subseteq \mathcal{L}$, this means that f is measurable with respect to the Lebesgue $\sigma - algebra$. ■

Theorem 28: Riemann and Lebesgue integrals

Let f be a Riemann integrable function over [0,1]. Then f is Lebesgue integrable with respect to $([0,1], \mathcal{L}, \lambda)$. Furthermore

$$\int_{[0,1]} f d\lambda = \int_0^1 f(x)dx.$$

Remark 7.118 *This is extremely useful as we can now evaluate Lebesgue integrals using Riemann integrals and using the fundamental theorem of calculus.*

Now, we can generalise what we have just shown.

Proposition 7.119 *Every function f that is Borel-measurable and finite is Lebesgue integrable.*

Theorem 29: Lebesgue's Characterisation of Riemann integrable function

Let f be a bounded real-valued function defined on a compact interval [a,b]. Then f is Riemann integrable if and only if the set of all discontinuity points of f is a set of Lebesgue measure zero.

Theorem 30: Every Riemann integrable function is Lebesgue integrable

Every function $f \in C[a,b]$ has their Riemann integral and Lebesgue integral agreeing.

Hence, every function that is Riemann integrable is Lebesgue integrable. Therefore, we can use tools from calculus when we wish to compute Lebesgue integrals of continuous functions.

Proposition 33: Riemann integral and Lebesgue integral for non-negative functions

Let $f(x)$ be a non-negative Riemann integrable function on an interval $(a,b]$. Then, f is also Lebesgue integrable (with the Lebesgue measure λ) on $(a,b]$ with their integrals agreeing

$$\int_a^b f(x)dx = \int_{(a,b]} f d\lambda$$

7.4.5 Abstract Lebesgue Integral and Integrable Functions

Thus far, we have only looked at integrals of non-negative integrable functions. In this section, we are now interested in the integrals of functions $f : \Omega \to \mathbb{R}$. Recall that the integral of a function f is the area of the function that is above 0 less the area of the function below 0. Hence, we require $|f|$ to be a finite integral in order for us to take the integral of f. This motivates our definitions of integrable functions.

Definition 37: Decomposition of extended real-valued function

If $f : \Omega \to \overline{\mathbb{R}}$ is a extended real-valued function, we define the positive and negative parts $f^+, f^- : \Omega \to [0, \infty]$ of the function f by

$$f^+ = \max\{f, 0\}, \quad f^- = \max\{-f, 0\} = -(\min\{f, 0\})$$

Therefore, we can write the function f as

$$f = f^+ - f^-$$

Lemma 7.120 *Let $f : (\Omega, \mathcal{F}) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ be a measurable function. Then, the functions $f^-, f^+ \in \mathcal{F}^+$ are non-negative measurable functions.*

Proof: We can write

$$f^+ = \frac{|f|+f}{2} \quad f^- = \frac{|f|-f}{2}$$

where we have shown that each component was a non-negative measurable function. ■

Lemma 7.121 *f is measurable if and only if f^+ **and** f^- are measurable.*

Before the next few definitions, we explain there is a difference between the **existence** of integral and **integrability** in Lebesgue integration. Integrability refers to the integral being finite. However, we can still define an integral $\int f d\mu$ but it **needs not** be integrable. Indeed, we have said from before that $\int g d\mu$ for a nonnegative measurable g, we allowed it to take the value ∞.

With that in mind, we now define how to take the integral of a measurable function.

Definition 38: Abstract Lebesgue Integral

Let $f : (\Omega, \mathcal{F}) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ be a measurable function. We define

$$\int_\Omega f d\mu \coloneqq \int_\Omega f^+ d\mu - \int_\Omega f^- d\mu$$

provided that **at least one** of the integrals $\int_\Omega f^+ d\mu, \int_\Omega f^- d\mu$ is finite.

Remark 7.122 *Alternatively, we can say that a function f is well-defined if*

$$min\{\int_\Omega f^+ d\mu, \int_\Omega f^- d\mu\} < \infty.$$

*The integral $\int f d\mu$ **does not exist** if and only if $\int f^+ d\mu = \infty$ and $\int f^- d\mu = \infty$ since $\infty - \infty$ is not defined.*

Remark 7.123 *We use the convention that* $\infty - c = \infty$ *and* $c - \infty = \infty$.

We now draw out the distinction for when a function is integrable.

Definition 39: Integrable function

Define the measure space $(\Omega, \mathcal{F}, \mu)$. A measurable function f is **integrable** if

$$\max\Big(\int f^+ d\mu, \int f^- d\mu\Big) < \infty.$$

Proposition 34: Equivalent formulation of integrability

Define the measure space $(\Omega, \mathcal{F}, \mu)$. Let f be a measurable function. As $f = f^+ - f^-$, this implies that $|f| = f^+ + f^-$. Then, we can say that f is a $\mu-$integrable function if

$$\int_\Omega |f| d\mu < \infty.$$

Proof: As $|f| = f^+ + f^-$, we have that

$$\int |f| d\mu = \int f^+ d\mu + \int f^- d\mu.$$

The above equation is finite if and only if $\max\{\int f^+ d\mu, \int f^- d\mu\} < \infty$. Hence, f is integrable. ■

Definition 40: $\mathcal{L}^1 - Space$

Define the measure space $(\Omega, \mathcal{F}, \mu)$. We define the $\mathcal{L}^1$-space as the space of functions that are integrable with respect to the measure μ. That is

$$\mathcal{L}^1(\Omega, \mathcal{F}, \mu) = \{f : \int_\Omega |f| d\mu < \infty.\}$$

We can write $\mathcal{L}^1(\Omega, \mathcal{F}, \mu) = \mathcal{L}^1(d\mu)$.

Proposition 35: Condition for integrability of a function

Define the measure space $(\Omega, \mathcal{F}, \mu)$. Let f be a measurable function. Then $f \in \mathcal{L}^1(d\mu)$ if and only if $|f| \in \mathcal{L}^1(d\mu)$.

We now state some properties of integrating in $\mathcal{L}^1$.

Lemma 7.124 *(Integrating over subset). Define the measure space* $(\Omega, \mathcal{F}, \mu)$. *For any* $\mu-$*measurable function set* A *and* $f \in \mathcal{L}^1(d\mu)$, *we define integrating over a set* $A \subset \Omega$ *as*

$$\int_A f d\mu \coloneqq \int_X 1_A f d\mu$$

where 1_A *is the indicator function.*

Proposition 36: $\mathcal{L}^1$ is a vector space

Define the measure space $(\Omega, \mathcal{F}, \mu)$. Let $f, g \in \mathcal{L}^1(d\mu)$ and $\alpha, \beta \in \mathbb{R}$. Then, we have that $\alpha f + \beta g \in \mathcal{L}^1(d\mu)$. That is, $\mathcal{L}^1$ is a vector space.

Proof: First, we use the fact that

$$\Big|\alpha f + \beta g\Big| \leq |\alpha||f| + |\beta||g|$$

Then, we note that integration is monotone with respect to $\mathcal{F}^+$. Therefore, the right hand side is in $\mathcal{L}^1$ and therefore the left hand side is in $\mathcal{L}^1$. ■

Proposition 37: Properties of Lebesgue integrals

Define the measure space $(\Omega, \mathcal{F}, \mu)$. Let $f, g \in \mathcal{L}^1(d\mu)$. Then we have that

1. For all $\alpha, \beta \in \mathbb{R}$
$$\int_\Omega (\alpha f + \beta g) d\mu = \alpha \int_\Omega f d\mu + \beta \int_\Omega g d\mu.$$
2. If $A, B \in \mathcal{F}$ are disjoint, then
$$\int_{A \cup B} f d\mu = \int_A f d\mu + \int_B f d\mu.$$
3. We have the **triangle inequality**
$$\Big|\int_\Omega f d\mu\Big| \leq \int_\Omega |f| d\mu.$$

Remark 7.125 *Every function that is Riemann integrable is Lebesgue integrable. However, the converse does not hold.*

STAT4028: Probability and Mathematical Statistics

8. $\mathcal{L}^p$-spaces

8.5 Lebesgue Spaces

8.5.1 $\mathcal{L}^p$-spaces

We now extend upon $\mathcal{L}^1$-spaces.

Definition 41: $\mathcal{L}^p$-norm

Define the measure space $(\Omega, \mathcal{F}, \mu)$. Let $1 \leq p < \infty$ and $f : \Omega \to \mathbb{R}$. We call

$$||f||^p := \Big(\int_\Omega |f|^p d\mu \Big)^{1/p}$$

the $\mathcal{L}^p-$norm of f.

Remark 8.126 *Alternatively, we can say $f \in \mathcal{L}^p$ if $|f|^p \in \mathcal{L}^1$. That is, $\int |f|^p d\mu < \infty$.*

Definition 42: $\mathcal{L}^p$-space

Define the measure space $(\Omega, \mathcal{F}, \mu)$. We define the space

$$\mathcal{L}^p(\Omega, \mathcal{F}, \mu) = \{f : \Omega \to \mathbb{R} : \text{f is measurable and } ||f||_p < \infty\}.$$

We call $\mathcal{L}^p(d\mu)$ the Lebesgue space.

Proposition 38: $\mathcal{L}^p(d\mu)$ is a vector space

The space $\mathcal{L}^p(\Omega, \mathcal{F}, \mu)$ is a vector space.

Proof: First, suppose that $f, g \in \mathcal{L}^p$. Then

$$|f+g|^p \leq \Big(|f|+|g|\Big)^p \leq \Big(2 \cdot \max\{f, g\}\Big)^p \leq 2^p \Big(\max\{f^p, g^p\}\Big) \leq 2^p \Big(|f|^p + |g|^p\Big)$$

Then, on the right hand side

$$\int_\Omega 2^p \Big(|f|^p + |g|^p\Big) d\mu < \infty$$

and therefore $|f+g|^p \in \mathcal{L}^1$. This means that $f + g \in \mathcal{L}^p$.

Now, assume $\alpha \in \mathbb{R}$ and $f \in \mathcal{L}^p$.

$$\int_\Omega \alpha |f|^p d\mu = \alpha \int_\Omega |f|^p d\mu < \infty$$

and therefore $\alpha f \in \mathcal{L}^p$. ■

Claim 8.127 *Define the measure space $(\Omega, \mathcal{F}, \mu)$. If $f, g \in \mathcal{F}$, $f \leq g$ and $\int f^- d\mu < \infty$, then $\int g d\mu$ exists and*

$$\int f d\mu \leq \int g d\mu.$$

We can actually strengthen conditions if two functions are equal almsot everywhere.

Claim 8.128 *Define the measure space $(\Omega, \mathcal{F}, \mu)$. Let $f, g \in \mathcal{F}$ be measurable functions and $f = g$ a.e. Then $\int f d\mu$ if well-defined if and only if $\int g d\mu$ is well-defined then*

$$\int f d\mu = \int g d\mu.$$

We revisit some theorems we have previously stated with stronger conditions of it holding a.e.

Corollary 8.129 *The MCT, Fatou's lemma, and the reversed Fatou lemma hold if the conditions hold a.e:*

1. *(Fatou's Lemma) If f_n are measurable functions with $f_n \geq 0$ a.e. then $\int \underline{\lim} f_n d\mu \leq \underline{\lim} \int f_n d\mu$*
2. *(MCT) If in addition $f_n \to f$ monotonically pointwise a.e. then $\int f_n d\mu \to \int f d\mu$ as $n \to \infty$.*
3. *(Reversed Fatou) If f, f_n are measurable functions with $0 \leq f_n \leq g$ a.e and $g \in \mathcal{L}^1$ then $\int \overline{\lim} f_n d\mu \geq \overline{\lim} \int f_n d\mu$.*

Claim 8.130 *(Strengthened reverse Fatou Lemma). Define the measure space $(\Omega, \mathcal{F}, \mu)$. Suppose f_n are measurable with $f_n \leq g$ a.e and $g \in L^1$ then*

$$\int \overline{\lim} f_n d\mu \geq \overline{\lim} \int f_n d\mu.$$

8.5.2 Modes of convergence

First, recall that the definition of pointwise convergence that for a sequence of functions $\{f_n\}_{n \geq 1}$ that for all $x \in X$ and for all $\epsilon > 0$, there exists a $N = N(\epsilon, x) \in \mathbb{N}$ such that if $n \geq N$, then

$$|f_n(x) - f(x)| < \epsilon.$$

Now, for almost surely convergent, we relax this assumption by not requiring it not to converge for all $\omega \in \Omega$ but the ones that don't, belong to a set of measure zero. That is, let X_n be a sequence of random variables on the same probability space $(\Omega, \mathcal{F}, P)$. Then

$$P(\omega \in \Omega : lim_{n \to \infty} X_n(\omega) = X(\omega)) = 1.$$

That is, the set of points ω that it does not converge to has a probability of zero.

Definition 43: Convergence almost everywhere

Suppose $\{f_n\}_{n \geq 1}$ is a sequence of measurable functions $(\Omega, \mathcal{F}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathbb{B}})$. Then f_n converges in f **almost everywhere** (**almost surely** if μ is a probability measure) if $\lim f_n(\omega) = f(\omega)$ pointwise on a set $E \in \mathcal{F}$ and $\mu(E^c) = 0$.

Definition 44: Convergence in $\mathcal{L}^p$

Suppose $\{f_n\}_{n\geq 1}$ is a sequence of measurable functions $(\Omega, \mathcal{F}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathbb{B}})$. Suppose that $f \in \mathcal{F}$ is a measurable function. We say that $f_n \to f$ in $\mathcal{L}^p$ if

$$||f_n - f||^p = \int_\Omega |f_n - f|^p d\mu \to 0$$

as $n \to \infty$.

Remark 8.131 *This is also known as convergence in the p-th mean as we will soon see that expectation of a random variable is just the Lebesgue integral. Alternatively, we say that X_n converges to X in $p-th$ moment if*

$$\lim_{n\to\infty} E\Big[|X_n - X|^p\Big] = 0.$$

When $p = 2$, we call this ***convergence in quadratic mean****.*

Proposition 39: Convergence in p-th mean is integrable

Let $\{f_n\} \in \mathcal{L}^p(X)$. If $f_n \to f$ in $\mathcal{L}^p$, that is $\int |f_n - f|^p d\mu \to 0$, then $f \in \mathcal{L}^p$.

Proof: First, we have that $\lim_{n\to\infty} \int |f_n - f|^p d\mu = 0$. Then, for every $\epsilon > 0$, there exists $N_\epsilon > 0$ such that

$$\int |f_n - f|^p d\mu \leq \frac{\epsilon}{2^{p-1}}$$

for all $n \geq N_\epsilon$. Then, by adding $f_n - f_n$ and the triangle inequality

$$\int |f|^p d\mu \leq 2^{p-1} \int |f - f_n|^p d\mu + 2^{p-1} \int |f_n|^p d\mu = 2^{p-1} \frac{\epsilon}{2^{p-1}} + 2^{p-1} \int |f_n|^p d\mu < \infty$$

Therefore, $\int |f|^p d\mu < \infty$ and hence $f \in \mathcal{L}^p$. ■

Without additional assumptions, there is no relationship between convergence almost everywhere and converge in $\mathcal{L}^p$.

Corollary 8.132 *Suppose $\{f_n\}_{n\geq 1}$ is a sequence of measurable functions $(\Omega, \mathcal{F}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathbb{B}})$. If $f_n \to f$ almost everywhere, it* ***does not*** *mean that $f_n \to f$ in $\mathcal{L}^1$.*

Proof: Define the measure space $((0,1), \mathcal{B}, \lambda)$. Define the sequence of functions $f_n = n1_{(0,\frac{1}{n})}$. Then $f_n \to 0$ a.e. but $\int |f_n - f| d\mu = \int n1_{(0,\frac{1}{n})} d\mu = 1$. ■

We now introduce a new limit theorem which is extremely useful.

Theorem 31: Dominated Convergence Theorem

Let $\{f_n\} \in \mathcal{F}$ be a sequence of measurable functions on a measure space $(\Omega, \mathcal{F}, \mu)$. Suppose that the sequence $f_n \to f$ pointwise almost everywhere. Furthermore, suppose that $|f_n(\omega)| \leq g(\omega)$ for all $\omega \in \Omega$ a.e for $g \in \mathcal{L}^1(\Omega)$. Then

1. $f_n, f \in \mathcal{L}^1$ for all $n \in \mathbb{N}$
2. $\int_\Omega |f_n - f| d\mu = 0$ as $n \to \infty$. That is, $f_n \to f$ in $\mathcal{L}^1$.

Proof: It is clear that $f_n, f \in \mathcal{L}^1$ as both are bounded by $g \in \mathcal{L}^1$. Now, let

$$\phi_n = |f - f_n| \in \mathcal{F}^+$$

Then,

$$\phi_n \to 0 \text{ a.e.}$$

Furthermore, $\phi_n \in \mathcal{L}^1$ as

$$\phi_n(\omega) \leq |f(\omega)| + |f_n(\omega)| \leq 2g(\omega)$$

for a.e. $\omega \in \Omega$.

Then, for a.e. $\omega \in \Omega$

$$\overline{lim} \int \phi_n d\mu \leq \int \overline{lim} \phi_n d\mu = 0$$

where the first inequality comes from reverse Fatou's lemma and the final equality comes from the fact that $\phi_n \to 0$ a.e. Therefore, we can conclude that $f_n \to f$ in $\mathcal{L}^1$. ■

Remark 8.133 *Two big takeaways is that the sequence of functions and its limit f_n, f are integrable. Additionally, we can strengthen the convergence from pointwise to $\mathcal{L}^1$ convergence.*

We state a very important result from the dominated convergence theorem.

Proposition 40: Interchange limit and integrals due to DCT

Let $\{f_n\} \in \mathcal{F}$ be a sequence of measurable functions on a measure space $(\Omega, \mathcal{F}, \mu)$. If $f_n \to f$ in $\mathcal{L}^1$, then

$$\lim_{n\to\infty} \int f_n d\mu = \int \lim_{n\to\infty} f_n d\mu = \int f d\mu.$$

That is, if $f_n \to f$ in $\mathcal{L}^1$, then $\int f_n d\mu \to \int f d\mu$.

Proof: We have that $\int |f_n - f| d\mu \to 0$. Then, by the triangle inequality, we have that

$$\left| \int_\Omega f d\mu - \int_\Omega f_n d\mu \right| \leq \int_\Omega |f - f_n| d\mu < \epsilon$$

for all $n \geq N_\epsilon$. Therefore, we can conclude that

$$\int f_n d\mu \to \int f d\mu.$$

■

We now show a case where the converse of the above state **does not hold**. That is, being able to interchange the limit and integrals $\int f_n d\mu \to \int f d\mu$ does not necessarily imply $f_n \to f$ converges in $\mathcal{L}^1$.

First, we recall that the Monotone Convergence Theorem requires an increasing sequence of non-negative measurable functions $f_n \uparrow f$.

Claim 8.134 *(The Monotone Convergence Theorem does not imply $\mathcal{L}^1$ convergence). Let $\{f_n\} \in \mathcal{F}^+$ be a sequence of measurable **non-negative** functions on a measure space $(\Omega, \mathcal{F}, \mu)$. Furthermore, assume that $f_n \uparrow f$. Therefore, we have that*

$$\int f_n d\mu \uparrow \int f d\mu.$$

*However, **it does not follow** that $f_n \to f$ in $\mathcal{L}^1$.*

However, if we can also specify a dominating function, then we can then conclude that being able to interchange limits and integrals $\int f_n d\mu \to \int f d\mu$ **does indeed** imply $f_n \to f$ converges in $\mathcal{L}^1$.

Proposition 41: Dominated convergence theorem and convergence in $\mathcal{L}^1$

Let $\{f_n\} \in \mathcal{F}^+$ be a sequence of measurable **non-negative** functions on a measure space $(X, \mathcal{F}, \mu)$. Assume that $f_n \uparrow f$ is a montone sequence which converges pointwise. Furthermore, assume that $f \in \mathcal{L}^1$. Then, by the Monotone Convergence Theorem

$$\int f_n d\mu \uparrow \int f d\mu.$$

Furthermore, by the dominated convergence theorem, where f is the dominating function, we have that

$$||f - f_n|| = \int_\Omega |f - f_n| d\mu \to 0.$$

That is, $f \to f_n$ in $\mathcal{L}^1$.

We can state a condition in order for convergence almost everywhere to imply convergence in $\mathcal{L}^1$.

Theorem 32: Scheffe's Lemma

Suppose that $f_n, f \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu)$ and that $f_n \to f$ a.e. Then $f_n \to f$ in $\mathcal{L}^1$ if and only if

$$\int |f_n| d\mu \to \int |f| d\mu$$

Remark 8.135 *Scheffe's lemma implies that if $f_n \to f$ in $\mathcal{L}^1$, it **does not mean** that $f_n \to f$ almost everywhere if $\int |f_n| d\mu \to \int |f| d\mu$.*

We now state a weaker type of convergence.

Definition 45: Convergence in measure

Let $\{f_n\}_{n \geq 1}$ be a sequence of measurable functions in $(\Omega, \mathcal{F}, \mu)$. The sequence f_n **converges to f in measure** if for all $\epsilon > 0$

$$\mu\Big(\omega : |f_n - f| > \epsilon\Big) \to 0$$

as $n \to \infty$.

Remark 8.136 *If μ is a probability measure, then this is known as **convergence in probability**. We write that a sequence of random variables $\{X_n\}$ converges in probability towards the random variable X if for all $\epsilon > 0$*

$$\lim_{n \to \infty} P(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}) = \lim_{n \to \infty} P(|X_n - X| > \epsilon) = 0.$$

We can see that this is a weaker type of convergence as it is implied by other forms of convergence.

Theorem 33: Convergence almost everywhere implies convergence in measure

Let $\{f_n\}_{n\geq 1}$ be a sequence of measurable functions in $(\Omega, \mathcal{F}, \mu)$. If $f_n \to f$ almost everywhere and $\mu(\Omega) < \infty$, then $f_n \to f$ in measure.

Proof: We will make sure of the dominating convergence theorem where the constant function 1 will be our dominating function and hence why we require $\mu(\Omega) < \infty$. Fix $\epsilon > 0$ and let $A_n = \{\omega : |f_n - f| > \epsilon\}$ and define $g_n = 1_{A_n}$. Since $f_n \to f$ a.e. then $g_n \to 0$ a.e. Furthermore, as the function is bounded $g_n \leq 1$ whereby $\int_\Omega 1 d\mu = \mu(\Omega) < \infty$, this implies that the constant function $1 \in \mathcal{L}^1$. Then, as a result of the dominating convergence theorem, $g_n \in \mathcal{L}^1$ and $g_n \to 0$ in $\mathcal{L}^1$. Therefore, we have that

$$\mu(A_n) = \int_\Omega g_n d\mu \to \int 0 d\mu = 0$$

Therefore, $f_n \to f$ in measure. ■

Remark 8.137 *In probability spaces, as $\mathbb{P}(\Omega) = 1$, this means that **convergence almost surely always implies convergence in probability**.*

We elaborate with an example on why we need a finite measure.

Example 8.138 *Let $f_n = 1_{(n,\infty)}$ defined on the measure space $([0,\infty), \mathcal{B}, \lambda)$. Clearly, $f_n \to 0$ almost everywhere. However, $\lambda(\{\omega : |f_n(\omega) - 0| > \frac{1}{2}\}) = \infty$. Therefore, f_n does not convergence to 0 in measure.*

The converse of convergence in measure does not imply convergence almost everywhere.

Corollary 8.139 *Let f_n be measurable functions on $(\Omega, \mathcal{F}, \mu)$. Suppose that $f_n \to f$ in measure. This **does not mean** that $f_n \to f$ almost everywhere.*

We now state an important inequality in probability theory.

Proposition 42: Markov's Inequality

Define a measure space $(\Omega, \mathcal{F}, \mu)$. Let $f \in \mathcal{F}^+$ be a non-negative measurable function and $\alpha \in \mathbb{R}^+$. Then, **Markov's inequality** states that

$$\mu\Big(\{\omega \in \Omega : f(\omega) \geq \alpha\}\Big) \leq \frac{1}{\alpha} \int_\Omega f d\mu.$$

If $(\Omega, \mathcal{F}, \mu) = (\Omega, \mathcal{F}, \mathbb{P})$, and we defined $X : (\Omega, \mathcal{F}) \to (\overline{\mathbb{R}}^+, \overline{\mathcal{B}})$ to be a random variable, then Markov's inequality is

$$\mathbb{P}\Big(X \geq \alpha\Big) \leq \frac{1}{\alpha}\mathbb{E}[X].$$

Proof: As $f \in \mathcal{F}^+$, we have that

$$f \geq f \cdot 1_{\{f \geq \alpha\}} \geq \alpha 1_{\{f \geq \alpha\}} \geq 0$$

Then, taking integrals

$$\int f d\mu \geq \int \alpha 1_{f \geq \alpha} d\mu = \alpha \mu(\{\omega : f(\omega) \geq \alpha\})$$

■

We can use Markov's inequality to help us prove another convergence relationship.

Proposition 43: Convergence in $\mathcal{L}^1$ implies convergence in measure

Let f_n be measurable functions on $(\Omega, \mathcal{F}, \mu)$. Suppose that $f_n \to f$ in $\mathcal{L}^1$. Then, $f_n \to f$ in measure.

Proof: Fix $\epsilon > 0$. Then, by Markov's inequality,

$$\mu(|f_n - f| > \epsilon) \leq \frac{1}{\epsilon} \int |f_n - f| d\mu \to 0$$

as $n \to \infty$ since $f_n \to f$ in $\mathcal{L}^1$. ■

Remark 8.140 *This actually holds for $f_n \to f$ in $\mathcal{L}^p$ for any $p > 0$. In probability language, this says that* ***convergence in mean*** *implies* ***convergence in probability.***

Whilst we may not have that convergence in measure imply convergence almost everywhere, we have something close to it.

Claim 8.141 *Let f_n be measurable functions on $(\Omega, \mathcal{F}, \mu)$. Suppose that $f_n \to f$ in measure. Then, there exists a sub-sequence $\{n_k\}$ such that $f_{n_k} \to f$ almost everywhere as $k \to \infty$.*

Corollary 8.142 *The claim also applies if $f_n \to f$ in $\mathcal{L}^1$.*

However, if we add in an extra assumption, we can show that **convergence in measure implies convergence in $\mathcal{L}^1$**. The dominated convergence theorem provides **sufficient conditions** under which almost everywhere convergence of a sequence of functions implies convergence in the $\mathcal{L}^1$ norm.

Proposition 44: Convergence in measure with dominating function implies convergence in $\mathcal{L}^1$

Let f_n be measurable functions on $(\Omega, \mathcal{F}, \mu)$. Suppose that $f_n \to f$ in measure. Furthermore, assume that there exists an integrable dominating function $g \in \mathcal{L}^1$ such that $|f_n| \leq g$ almost everywhere. Then $f_n \to f$ in $\mathcal{L}^1$.

Proof: We prove by contradiction. Assume that $\int |f_n - f| d\mu \not\to 0$. Then, this means that for all $\epsilon > 0$, there exists a subsequence $\{n_k\}$ such that for all k

$$\int |f_{n_k} - f| d\mu > \epsilon \tag{*}$$

However, by assumption $f_{n_k} \to f$ in measure. By our previous result, there exists a further subsequence $\{m_k\} \subset \{n_k\}$ such that

$$f_{m_k} \to f \text{ a.e.}$$

Then, recall that $f_{m_k} \leq g \in \mathcal{L}^1$. Therefore, by the dominating convergence theorem

$$f_{m_k} \to f \text{ in } \mathcal{L}^1$$

This contradicts the original claim (*) and therefore we can conclude that $f_n \to f$ in $\mathcal{L}^1$. ■

Remark 8.143 *Therefore, if we ever have convergence in measure and a dominating function, then we can strengthen this to convergence in* $\mathcal{L}^1$.

We now state the weakest form of convergence.

Definition 46: Weak Convergence

A sequence of probability measures μ_n converges weakly to a probability measure μ if for all $x \in \mathbb{R}$, with $\mu(\{x\}) = 0$, we have that

$$\mu_n((-\infty, x]) \to \mu((-\infty, x])$$

We can in fact relate this form of convergence to the distribution function F.

Definition 47: Converges in Distribution

Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables. We say that the sequence X_n converges in distribution to a random variable X if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for every number $x \in \mathbb{R}$ at which F is continuous. Here, F_n and F are the cumulative distribution functions of the random variables X_n and X respectively.

Proposition 45: Relationship between weak convergence and convergence in distribution

Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables. Then, if the random variables X_n converges in distribution,

$$F_{X_n} \to F_X$$

then, $X_n \to X$ weakly.

Remark 8.144 *The reason it is called* ***weak*** *convergence is that the* $\{X_n\}$ *and* X ***does not need*** *to be defined on the same probability space as in previous forms of convergence.*

The following claim shows the relationship between this new form of convergence to those previously seen.

Claim 8.145 *Suppose that* $X_n \to X$ *in probability. Then* $X_n \to X$ *weakly.*

8.5.3 Inequalities in the $\mathcal{L}^p$-space

Earlier, we alluded to the fact that $||f||_p$ is known as the $\mathcal{L}^p-$norm. We shall prove some theorems using it and also show that it is in fact a **pseudonorm**.

Lemma 8.146 *(Young's Inequality). Let* $p, q > 1$ *be such that*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then,

$$st \leq \frac{1}{p}s^p + \frac{1}{q}t^q$$

for all $s, t \geq 0$.

Using Young's inequalities, we can derive inequalities for $\mathcal{L}^p-$norms.

Proposition 46: Hölder's Inequality

Let f,g be measurable functions and $p, q > \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$||f \cdot g|| = \int |f \cdot g| d\mu \leq \Big(\int |f|^p d\mu \Big)^{1/p} \Big(\int |g|^q d\mu \Big)^{1/q} = ||f||_p ||g||_q$$

for all $f \in \mathcal{L}^p(\Omega)$ and $g \in \mathcal{L}^q(\Omega)$.

Hölder's inequality is actually a generalisation of the Cauchy-Schwarz inequality.

Theorem 8.147 *(Cauchy-Schwarz Inequality). Let* $p = q = 2$. *Then, let* f, g *be measurable functions* $f, g \in \mathcal{L}^2(X)$. *Then*

$$||fg||_2 = \Big| \int_\Omega f g d\mu \Big|^2 \leq \int_\Omega |f^2| d\mu \int_\Omega |g^2| d\mu = ||f||_2^2 ||g||_2^2$$

Proposition 47: Minkowski's Inequality

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Fix $p \in [1, \infty)$. Then, for measurable functions $f, g \in \mathcal{L}^p(\Omega)$, then

$$||f + g||_p = \Big(\int |f + g|^p d\mu \Big)^{1/p} \leq \Big(\int |f|^p d\mu \Big)^{1/p} + \Big(\int |g|^p d\mu \Big)^{1/p} = ||f||_p + ||g||_p.$$

Remark 8.148 *This is the triangle inequality for the* $\mathcal{L}^p$ *norm* $||.||_p$.

We can now summarise the properties of why $\mathcal{L}^p-$norm is in fact a pseudonorm.

Proposition 8.149 *(Properties of the* $\mathcal{L}^p$*-pseudonorm). Let* $1 \leq p < \infty$. *Then, for* $f, g \in \mathcal{L}^p(X)$ *and* $\alpha \in \mathbb{R}$, *we have*

1. $||f||_p \geq 0$ *with equality if and only if* $f = 0$ *almost everywhere.*
2. $||\alpha f||_p = |\alpha| \, ||f||_p$
3. $||f + g||_p \leq ||f||_p + ||g||_p$ *(Minkowski's inequality).*

Remark 8.150 *The reason that* $\mathcal{L}^p$*-norm is not a proper norm since* $||f||_p = 0$ ***does not*** *necessarily imply that* $f = 0$ *is the zero function. This is important as in a normed vector space* V, *there is a* ***unique vector*** $v \in V$ *with* $||v|| = 0$, *known as the* ***zero vector****.*

We will see later on how can we turn the pseudnorm $||.||_p$ in a proper norm.

8.5.4 Completeness of the $\mathcal{L}^p$-space

We now look at sequences of functions in the $\mathcal{L}^p-$space.

Definition 8.151 *(Cauchy sequence). Let $\{f_n\}, f \in \mathcal{L}^p$. We say that $\{f_n\}_{n\geq 1}$ is a **Cauchy sequence** in $\mathcal{L}^p$ if for every $\epsilon > 0$, there exists a $n_0 \in \mathbb{N}$ such that*

$$||f_n - f_m||_p < \epsilon$$

for all $n, m > n_0$.

Remark 8.152 *Recall that*

$$||f_n - f_m||_p = \left(\int |f_n - f_m|^p d\mu\right)^{\frac{1}{p}}.$$

Lemma 8.153 *Every convergent sequence is a Cauchy sequence.*

We want to show that $\mathcal{L}^p$ is complete. That is, every Cauchy sequence converges. The idea to show this is that if a Cauchy sequence has a convergent subsequence, then it converges. Note that for general sequences, this is not true.

First, recall that in $\mathcal{L}^1$, we have shown that if $f_n \to f$ in L^1, then there exists a sub-sequence $\{n_k\}$ such that $f_{n_k} \to f$ a.e. We extend this idea to $\mathcal{L}^p(X)$ space.

We use the next two lemmas to show completeness of the $\mathcal{L}^p(X)$ space.

Lemma 8.154 *Suppose that $\{f_n\}$ is a Cauchy sequence in $\mathcal{L}^p(X)$. If $\{f_{n_k}\}$ is a convergent subsequence, with $f_{n_k} \to f$ in $\mathcal{L}^p(X)$, then $f_n \to f$ in $\mathcal{L}^p(X)$.*

Lemma 8.155 *Let $\{g_k\}$ be a sequence of functions in $\mathcal{L}^p(X)$ such that*

$$\sum_{k=1}^{\infty} ||g_k||_p < \infty.$$

Then, there exists a function $f \in \mathcal{L}^p(X)$ such that $f = \sum_{k=0}^{\infty} g_k$ converges pointwise almost everywhere and in $\mathcal{L}^p(X)$.

We can now state the main theorem of this section.

Theorem 34: Completeness of $\mathcal{L}^p$

Let $\{f_n\}_{n\geq 1}$ be a Cauchy sequence in $\mathcal{L}^p$. Then, there exists a function $f \in \mathcal{L}^p$ such that $f_n \to f$ in $\mathcal{L}^p$. Moreover, $\{f_n\}_{n\geq 1}$ has a subsequence that converges almost everywhere.

8.5.5 L^p-spaces

We have seen in the previous section that $\mathcal{L}^p-$norm is not a proper norm since $||f||_p= 0$ does not imply that $f = 0$ is the zero function. We can can only conclude that $f = 0$ almost everywhere. What we do is that we consider functions to be equal if they are equal almost everywhere. Hence, we can define an equivalence relation where $f \sim g$ if f = g almost everywhere.

Definition 48: Equivalence class of functions equal almost everywhere

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let f be a measurable function. We define the equivalence class of f by

$$[f] = \{g : \Omega \to \mathbb{R} : \text{g is measurable and } f = g \text{ almost everywhere}\}.$$

From this, we can introduce a family of vector spaces of these equivalence classes.

Definition 49: L^p-space

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. For $1 \leq p < \infty$, we define

$$L^p(\Omega, \mathcal{F}, \mu) \coloneqq \{[f] : f \in \mathcal{L}^p(\Omega, \mathcal{F}, \mu)\}.$$

We set the L^p-norm as

$$\Big|\Big|[f]\Big|\Big|_p \coloneqq ||f||_p$$

Similarly, we set

$$[f] + [g] \coloneqq [f + g] \quad \text{and} \quad \alpha[f] \coloneqq [\alpha f]$$

for all $[f], [g] \in L^p(\Omega)$ and $\alpha \in \mathbb{R}$.

Remark 8.156 *In practice, we write $f \in L^p(\Omega)$ and call it an L^p-function rather than [f]. In many situations, we choose a representative in the equivalence class from $\mathcal{L}^p(\Omega)$.*

Remark 8.157 *The norm $||.||_p$ on $L^p(\Omega)$ is a proper norm now.*

Claim 8.158 *(The $L^p(\Omega)$ class is the quotient space). Define the measure space $(\Omega, \mathcal{F}, \mu)$. Define $\mathcal{L}^p(\Omega)$ and $L^p(\Omega)$ to be as before. Then, define $\mathcal{N}(\Omega) = \{g : \Omega \to \mathbb{R} | g$ is measurable and $g = 0$ almost everywhere$\}$. Then, $\mathcal{N}(\Omega)$ is a vector subspace of $\mathcal{L}^p(\Omega)$ and*

$$L(\Omega) = \frac{\mathcal{L}^p(\Omega)}{\mathcal{N}(\Omega)}$$

is the quotient space.

Lemma 8.159 *Every inner product induces a norm. However, only a norm which satisfies the parallelogram law can define an inner product.*

Theorem 35: L^p-norm is a Banach space

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then, for $1 \leq p < \infty$, define the vector space $L^p(\Omega)$. The space $L^p(\Omega)$ is a complete normed space with respect to the norm $||.||_p$.

Remark 8.160 *A complete normed vector space is called a* ***Banach space****.*

Furthermore, recall that a complete inner product space is called a **Hilbert space**. There is only **one** L^p-space that is a Hilbert space.

Proposition 48: L^2-space is a Hilbert space

Let $L^2(\Omega)$ be a Banach space with respect to the $||.||_2$ norm. We have that $L^2(\Omega)$ is a complete inner product space with the inner product

$$\sqrt{\langle f, g \rangle} = \int_\Omega fg \, d\mu.$$

STAT4028: Probability and Mathematical Statistics

9. Expectation and Radon-Nikodym Theorem

9.6 Expectation

9.6.1 Expectation

Definition 50: Expectation

Define the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Define the random variable $X : (\Omega, \mathcal{F}) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$. The expectation of the random variable X is

$$E(X) = \int_\Omega X d\mathbb{P}$$

provided the integral is well-defined. That is, the expectation is the Lebesgue integral with respect to the probability measure $\mathbb{P}$.

Remark 9.161 *Recall that $E(X) = \int_\Omega X d\mathbb{P}$ is well-defined if and only if $\min(\int_\Omega X^+ d\mathbb{P}, \int_\Omega X^- d\mathbb{P}) < \infty$. That is, either $E(X^+) < \infty$ and $E(X^-) \leq \infty$ **or** $E(X^+) \leq \infty$ and $E(X^-) < \infty$.*

Lemma 9.162 *X does not admit an expectation if and only if $E(X^+) = \infty$ **and** $E(X^-) = \infty$.*

Recall that for a random variable X defined on $(\Omega, \mathcal{F}, \mathbb{P})$, we define the probability measure induced by X μ_X on $(\overline{\mathbb{R}}, \overline{\mathbb{B}})$ as

$$\mu_X(B) = \mathbb{P}(X \in B)$$

for all $B \in \overline{\mathbb{B}}$.

Definition 51: Discrete Random Variable

A random variable X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ is **discrete** if there exists a distinct sequence $\{x_k\}_k \subset \overline{\mathbb{R}}$ such that

$$\sum_k \mathbb{P}(X = x_k) = \sum_k \mu_X(\{x_k\}) = 1.$$

We denote by $\mathbb{P}_X(x) = \mu_X(\{x\})$ the **probability mass function** of the discrete random variable X. The set $\{x_k : \mathbb{P}_X(x_k) > 0\}$ is called the **support** of the probability mass function.

Remark 9.163 *In this construction, it is clear that X will take on one of the values $\{x_k\}$, that is, $\mathbb{P}(X \in \{x_k\}) = 1$. However, in general, $\mathbb{P}(X(\Omega) \neq \{x_k\})$.*

Proposition 49: Expectation of discrete random variable is well-defined

Define a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let X be a discrete random variable with support $\{x_k\}$. Then, $E(X)$ is well-defined if and only if

$$\min\left\{\sum_{k:x_k>0} x_k \mathbb{P}_X(x_k), -\sum_{k:x_k<0} x_k \mathbb{P}_X(x_k)\right\} < \infty.$$

We now have a new definition for the expectation of a random variable.

Definition 52: Expectation of a discrete random variable

Let X be a discrete random variable. Let the expectation E(X) be well-defined. Then, the expectation of the discrete random variable X is

$$E(X) = \sum_k x_k \mathbb{P}_X(x_k).$$

We also state the condition for when the expectation is not only defined but integrable.

Proposition 50: Integrability of discrete random variable

Let X be a discrete random variable. Furthermore, assume that E(X) is defined. Then, $X \in L^1$ if and only if $\sum_k |x_k| \mathbb{P}_X(x_k) < \infty$.

Remark 9.164 *If $X \in L^1$, this implies that $E(X) < \infty$ where we require E(X) to be defined in the first place.*

We now work our way up to the Law of the unconscious statistician.

Lemma 9.165 *Let X be a discrete random variable. Let $h : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ be a measurable function. Then, $Y = h(X)$ is a random variable. Furthermore, Y is a **discrete** random variable.*

Hence, we can show that we can compute the expectation of Y using the distribution function induced by the random variable X.

Theorem 36: Law of the unconscious statistician

Let X be a discrete random variable. Let $Y = h(X)$ be a discrete random variable where $h : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ is a measurable function. Then, if $E(X)$ is well-defined, then $E(Y)$ is well-defined. Furthermore, we can write

$$E(Y) = \sum_k h(x_k) \mathbb{P}_X(x_k)$$

where $E(Y) < \infty$ if $X \in L^1$.

9.6.2 The Radon-Nikodym Theorem

We have established distributions and cumulative distribution functions. We will now work our way towards defining density functions.

Definition 9.166 *(Integrals over subsets). Let $A \in \mathcal{F}$. Let f be a measurable function. We then define the integral of f over A as*

$$\int_A f d\mu = \int_\Omega f 1_A d\mu$$

provided the right hand side exists.

Definition 53: Measures with a density

Let $(X, \mathcal{A}, \mu)$ be a measure space and $g : X \to [0, \infty)$ be a measurable function. For $A \in \mathcal{A}$, we define

$$\nu(A) \coloneqq \int_A g d\mu.$$

for all $A \in \mathcal{A}$. We have that ν is a measure. We call g the **density** of ν with respect to μ.

Proof: We prove the 2 conditions needed for a measure.

$$\nu(\emptyset) = \int_\emptyset g d\mu = \int_X 1_\emptyset g d\mu = 0.$$

Let $A_k \in \mathcal{A}, k \in \mathbb{N}$ be disjoint sets. Then

$$\nu(\bigcup_{k=0}^{\infty} A_k) = \int_{\bigcup_{k=0}^{\infty} A_k} g d\mu = \int_X 1_{\bigcup_{k=0}^{\infty} A_k} g d\mu = \int_X \sum_{k=0}^{\infty} 1_{A_k} g d\mu = \sum_{k=0}^{\infty} \int_X 1_{A_k} g d\mu = \sum_{k=0}^{\infty} \nu(A_k)$$

where the second last equality arises from the monotone convergence theorem applied to power series. Hence, ν is a measure. ■

We are now interested in the question that if μ and ν are arbitrary measures both defined on the $\sigma-$algebra $\mathcal{A}$, does ν have a density with respect to μ. That is, does there exist a measurable function $g : X \to [0, \infty)$ such that

$$\nu(A) = \int_A g d\mu$$

for all $A \in \mathcal{A}$.

The general answer is **no**. However, there is a condition needed for this to always hold.

Definition 54: Absolutely Continuous

Let μ, ν be measures defined on the same $\sigma-$algebra of subsets of X. We call ν **absolutely continuous** with respect to μ if for every $A \in \mathcal{A}$, $\mu(A) = 0$ implies that $\nu(A) = 0$. If that is the case, we write

$$\nu << \mu.$$

Remark 9.167 *This can also be thought of as ν being **dominated** by the measure μ.*

Hence, if we have two measures that are absolutely continuous, then we have a density function.

Theorem 37: Radon-Nikodym Theorem

Define the measure space $(\Omega, \mathcal{F}, \mu)$. Suppose μ and ν are $\sigma-$finite measures defined on the same $\sigma-$algebra $\mathcal{F}$. Moreover, assume that $\nu << \mu$. Then, ν has a density function $f : \Omega \to [0, \infty)$ with respect to μ such that

$$\nu(A) = \int_A f d\mu$$

for all $A \in \mathcal{F}$.
Moreover, the density is **essentially unique**, that is, any two density function are equal $\mu-$almost everywhere.

Remark 9.168 *We can write*

$$\int_\Omega f d\nu = \int_\Omega f \frac{d\nu}{d\mu} d\mu$$

if we let $\frac{d\nu}{d\mu} = f$ *where* f *is the density function from the Radon-Nikodym theorem. In this case,* f *is also known as the* ***Radon-Nikodym derivative*** *of* ν *with respect to* μ *which exists if* $\nu << \mu$.

We now relate all these measure theory concepts to probability theory. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be our probability space. Let X be our random variable to $(\mathbb{R}, \mathbb{B})$. Recall that the **distribution** measure of the random variable X is the probability measure μ_X on $\mathbb{B}$ defined by

$$\mu_X(B) = \mathbb{P}[X^{-1}(B)]$$

for all $B \in \mathbb{B}$. That is, μ_X is the **push-forward** of the measure $\mathbb{P}$ by the map X.

Finally, we can also work with the real-valued F_X defined by

$$F_X(x) = \mu_X((-\infty, x]) = \mathbb{P}[X \leq x]$$

for $x \in \mathbb{R}$. We call F_x the **(cumulative) distribution function** of the random variable X. We can now relate the Radon-Nikodym theorem to all of this.

Definition 55: Distribution function is Absolutely Continuous to the Lebesgue Measure

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let X be the random variable $(\Omega, \mathcal{F}) \to (\mathbb{R}, \mathbb{B})$. The distribution measure $\mu_X = \mathbb{P}[X^{-1}(B)]$, is **absolutely continuous** with respect to the Lebesgue measure λ on $\mathbb{B}(\mathbb{R})$ if $\mu_X(\mathbb{R}) = 1$ and there exists a measurable function $f : (\mathbb{R}, \mathcal{B}) \to ([0, \infty], \mathcal{B})$ such that

$$\mu_X(B) = \int_B f d\lambda$$

for all Borel sets $B \in \mathbb{B}$. We call the function f the **(probability) density function** of μ_X.

Remark 9.169 *In other words, the Radon-Nikodym derivative* $\frac{d\mu_X}{d\lambda}$ *is the probability density function (pdf) of X.*

Lemma 9.170 *There is a one-to-one correspondence between distribution measures and distribution functions.*

Remark 9.171 *Many random variables can induce the same distribution measure and cumulative distribution function. However, the converse does not hold.*

We now show how we can get continuous random variables. First, we need a different characterisation of absolutely continuous.

Theorem 38: Characterisation of absolutely continuous cumulative distribution function

Let μ be a finite Borel measure on $\mathbb{R}$ and $F(t) \coloneqq \mu((-\infty, t])$ be the cumulative distribution function associated with μ. Denote by λ, the Lebesgue measure on $\mathbb{R}$. Then, $\mu << \lambda$ if and only if for every $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\sum_{k=1}^{\infty} |F(b_k) - F(a_k)| < \epsilon$$

for every countable collection (a_k, b_k) of disjoint intervals such that

$$\sum_{k=1}^{\infty} |b_k - a_k| < \delta.$$

Hence, we can state where continuous random variables come from.

Proposition 51: Cumulative Distribution Function is Continuous

Let F be a cumulative distribution function associated to the Borel measure μ. Then, if F is absolutely continuous, then F is a continuous function.

Remark 9.172 *This gives us that if our random variable X has a cumulative distribution function that is absolutely continuous, then the cumulative distribution function is continuous. This is what is meant by a continuous random variable.*

9.6.3 Substitution theorem for random variables

In practice, we don't want (and can't) integrate the random variable X over the sample space Ω. We would would rather work with the distribution of X. In particular, we want to express integrals over Ω with respect to $\mathbb{P}$ into integrals over $\mathbb{R}$ with respect to the distribution function μ_X. First, we recall a few technical results.

Lemma 9.173 *Let $X : \Omega \to \mathbb{R}$ be a random variable on the probability space $(\Omega, \mathcal{F}, P)$ and $h : \mathbb{R} \to \mathbb{R}$ be Borel measurable. Then $h \circ X : \Omega \to \mathbb{R}$ is measurable.*

Claim 9.174 *Let X be a random variable on the probability space $(\Omega, \mathcal{F}, P)$. Then, the set function*

$$\mu_X(B) = P(X^{-1}(B)) = P(\{\omega : X(\omega) \in B\})$$

for all Borel sets $B \in \mathcal{B}$ is a measure.

We now state a proposition we will use alot later. We will state it for the case of a simple function but it can easily be extended to non-negative and measurable function.

Proposition 52: Change of variables formula

Let $X : \Omega \to \mathbb{R}$ be a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $h : \mathbb{R} \to [0, \infty)$ be a Borel non-negative simple function. Then

$$\int_\Omega h \circ X d\mathbb{P} = \int_{\mathbb{R}} h d\mu_X.$$

Proof: As h is a simple function, let $h = \sum_{k=1}^n a_k 1_{A_k}$. Now

$$(1_{A_k} \circ X)(\omega) = 1_{A_k}(X(\omega))$$

where $X(\omega) \in A_k$ if and only if $\omega \in X^{-1}(A_k)$. Hence, we can write

$$h \circ X = \sum_{k=1}^n a_k 1_{A_k} \circ X = \sum_{k=1}^n a_k 1_{X^{-1}(A_k)}.$$

Now, we can integrate to get our desired result

$$\int_\Omega h \circ X d\mathbb{P} = \int_\Omega \sum_{k=1}^n a_k 1_{A_k} \circ X d\mathbb{P} = \int_\Omega \sum_{k=1}^n a_k 1_{X^{-1}(A_k)} d\mathbb{P}$$

$$= \sum_{k=1}^n \int_\Omega a_k 1_{X^{-1}(A_k)} d\mathbb{P} = \sum_{k=1}^n a_k \mathbb{P}(X^{-1}(A_k))$$

where recall that by definition $\mu_X(A_k) = \mathbb{P}(X^{-1}(A_k))$, so therefore we have that

$$= \sum_{k=1}^n a_k \mu_X(A_k) = \int_{\mathbb{R}} h d\mu_X$$

where we used the fact that h is a simple function $h = \sum_{k=1}^n a_k 1_{A_k}$. ■

Notice that the change of variables integrates over the domain of the functions!

We can extend the change of variables formula to h being non-negative and measurable using the usual machinary in measure theory.

9.6.4 Expectation of Functions of Random Variables

We first define the change of variables theorem for **non-negative** functions of random variables.

Theorem 39: Change of variables formula for non-negative random variables

Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$. Let μ_X be the distribution probability measure on $(\overline{\mathbb{R}}, \overline{\mathcal{B}})$ induced by X. Now, suppose that $h : (\overline{\mathbb{R}}, \overline{\mathcal{B}}) \to (\overline{\mathbb{R}}^+, \overline{\mathcal{B}}^+)$ is a measurable and non-negative function. Let $Y = h(X)$. Then

1. $Y \in \mathcal{F}^+$ is a measurable non-negative random variable;
2. $E(Y) = \int_\Omega Y dP = \int_\Omega h \circ X dP = \int_{\mathbb{R}} h d\mu_X$.

Proof: $Y \in \mathcal{F}^+$ is clear as $X \in \mathcal{F}^+$ and h is a measurable non-negative function.

We can apply the change of variables formula to $E(Y) = \int_\Omega Y dP$ to get the desired result where

$$E(Y) = \int_{\mathbb{R}} h d\mu_X.$$

■

Before we extend the change of variables theorem to hold for all measurable functions $h \in \mathcal{F}$, we require another condition.

Proposition 9.175 *Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$. Suppose that the function $h : \mathbb{R} \to \mathbb{R}$ is Borel measurable. Then, $h \circ X \in L^1(\Omega, P)$ if and only if $h \in L^1(\mathbb{R}, \mu_X)$ where μ_X is the distribution measure induced by the random variable X.*

Now, we extend the above theorem to measurable functions h.

Theorem 40: Change of variables formula for L^1 functions

Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$. Let μ_X be the distribution probability measure on $(\overline{\mathbb{R}}, \overline{\mathcal{B}})$ induced by X. Now, suppose that $h : (\overline{\mathbb{R}}, \overline{\mathcal{B}}) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ is a measurable function. Furthermore, assume that $h \in L^1(\mathbb{R}, \mu_X)$. Then we have that

$$E(Y) = \int_\Omega Y dP = \int_\Omega h \circ X dP = \int_{\mathbb{R}} h d\mu_X.$$

Proof:(Sketch). This theorem follows from the change of variables formula for non-negative random variables as we can write $h = h^+ - h^-$. ■

Remark 9.176 *The change of variables formula for non-negative random variables and L^1 functions hold for **any** random variable.*

Recall that if μ_X is absolutely continuous with respect to the Lebesgue measure λ, then there exists a density function f. We first state the following claim to help us prove the subsequent theorem.

Proposition 53: Radon-Nikodym for distribution and Lebesgue measure

Suppose μ_X is a probability measure on $(\overline{\mathbb{R}}, \overline{\mathbb{B}})$ that is absolutely continuous with respect to the Lebesgue measure λ and the Radon-Nikodym derivative is the density function f. Then, for any measurable function $h : (\overline{\mathbb{R}}, \overline{\mathbb{B}}) \to (\overline{\mathbb{R}}, \overline{\mathbb{B}})$, we have that

$$\int_{\overline{\mathbb{R}}} h d\mu_X = \int_{\mathbb{R}} h d\mu_X = \int_{\mathbb{R}} h \cdot f d\lambda.$$

Proof: First, recall that $\mu(\mathbb{R}) = 1$. This gives us

$$\int_{\overline{\mathbb{R}}} h d\mu_X = \int_{\mathbb{R}} h d\mu_X + \int_{\{-\infty,\infty\}} h d\mu_X$$

Now, we look at the second term

$$\int_{\{-\infty,\infty\}} h d\mu_X \leq \int_{\{-\infty,\infty\}} \infty d\mu_X = \infty \mu_X(\{-\infty, \infty\}) = \infty \cdot 0 = 0.$$

Therefore, we have that

$$\int_{\overline{\mathbb{R}}} h d\mu_X = \int_{\mathbb{R}} h d\mu_X$$

We now want to show

$$\int_{\mathbb{R}} h d\mu_X = \int_{\mathbb{R}} h \cdot$$

First, let $h = 1_A$ be an indicator function for some $A \in \overline{\mathcal{B}}$. Then, we have that

$$\int_{\mathbb{R}} h d\mu_X = \int_{\mathbb{R}} 1_A d\mu_X = \mu_X(A) = \int_A f d\lambda = \int_{\mathbb{R}} 1_A f d\lambda = \int_{\mathbb{R}} h \cdot f d\lambda$$

We can then extend this to show simple functions by using the linearity of the integral. Then, for non-negative measurable functions, we can use the monotone convergence theorem. Finally, for measurable functions, we can split it up into positive and negative parts. ■

Theorem 41: Expectation with density functions

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let X be a random variable. Let $h : (\overline{\mathbb{R}}, \overline{\mathbb{B}}) \to (\overline{\mathbb{R}}, \overline{\mathbb{B}})$ be a measurable function. Define $Y = h(X)$. If X has a density function f, then μ_X is absolutely continuous with respect to the Lebesgue measure λ. Then we can define the expectation of Y using the Lebesgue integral

$$E(Y) = \int_\Omega h \circ X dP = \int_{\mathbb{R}} h d\mu_X = \int_{\mathbb{R}} h \cdot f d\lambda$$

with the left hand side defined if and only if the right hand side is defined.

We are now interested in expressing expectations in terms of the Riemann integral rather than the Lebesgue integral.

Theorem 9.177 *(Expectation of non-negative functions using Riemann integrals). Define the probability space $(\Omega, \mathcal{F}, P)$. Let X be a random variable that has a density function f and let $h : \mathbb{R} \to \mathbb{R}^+$ be a measurable function such that $g(x) = h(x)f(x)$ is a Riemann integrable function on any finite interval. Denote $\int_{-\infty}^{\infty} dx$ as the Riemann integral and $\int d\lambda$ as the Lebesgue integral. Then*

$$E(Y) = \int_{\mathbb{R}} h \cdot f d\lambda = \int_{-\infty}^{\infty} h(x) f(x) dx.$$

Proof:(Sketch). First write $E(Y) = \int_{\mathbb{R}} h \circ f d\lambda$ using the expectation with density functions theorem. Then, recall that a function that is Riemann integrable on a finite interval is equal to the Lebesgue integral of the function. ■

Corollary 9.178 *Define the probability space $(\Omega, \mathcal{F}, P)$. Let X be a random variable that has a density function f and let $h : \mathbb{R} \to \mathbb{R}^+$ be a measurable function such that $g(x) = h(x)f(x)$ is a Riemann integrable function on any finite interval. Then, $X \in L^1(\Omega)$ if and only if the following Riemann integral is finite*

$$\int_{-\infty}^{\infty} |x| f(x) dx < \infty.$$

We now generalise the above theorem to hold for any measurable function h.

Theorem 42: Expectation of measurable functions using Riemann integrals

Define the probability space $(\Omega, \mathcal{F}, P)$. Let X be a random variable that has a density function f and let $h : \mathbb{R} \to \mathbb{R}$ be a measurable function such that $g(x) = h(x)f(x)$ is a Riemann integrable function on any finite interval. Denote $\int_{-\infty}^{\infty} dx$ as the Riemann integral and $\int d\lambda$ as the Lebesgue integral. Then

$$E(Y) = \int_{\mathbb{R}} h \circ f d\lambda = \int_{-\infty}^{\infty} h(x)f(x)dx.$$

Corollary 9.179 *Define the probability space $(\Omega, \mathcal{F}, P)$. Let X be a random variable that has a density function f and let $h : \mathbb{R} \to \mathbb{R}$ be a measurable function such that $g(x) = h(x)f(x)$ is a Riemann integrable function on any finite interval. Then, $X \in L^1(\Omega)$ if and only if the following Riemann integral is finite*

$$\min\left\{ \int_{-\infty}^{\infty} h^+(x)f(x)dx, \int_{-\infty}^{\infty} h^-(x)f(x)dx \right\} < \infty.$$

In this case, then

$$E(Y) = \int_{-\infty}^{\infty} h(x)f(x)dx.$$

Corollary 9.180 *Define the probability space $(\Omega, \mathcal{F}, P)$. Let X be a random variable that has a density function f and let $h(x) = x$ be the identity function. Suppose that $g(x) = xf(x)$ is Riemann integrable on any finite interval. Then $E(X)$ is well-defined if and only if*

$$\int_{-\infty}^{0} xf(x)dx > -\infty \quad or \quad \int_{0}^{\infty} xf(x)dx < \infty.$$

If $E(X)$ is well-defined, then

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

We can now state an important result that we only need to know the distribution function of the random variable X rather than X itself in order to compute expectation and variance.

Proposition 54: Expectation and variance of a random variable

Define the probability space $(\Omega, \mathcal{F}, P)$. Let X be a random variable with distribution measure μ_X. Suppose that X has density function f. Furthermore, suppose that $E(X)$ is well-defined. Then

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

and

$$Var(X) = \int_{-\infty}^{\infty} x^2 f(x)dx - E[X]^2.$$

We now state some theorem regarding expectation of random variables.

Definition 9.181 *(Bounded random variable). Let X be a random variable. X is said to be bounded if*

$$X \leq C < \infty$$

almost surely.

Recall that earlier we did not establish a direct relationship between convergence almost surely and convergence in L^1. We can now do so if we impose an additional assumption.

Theorem 43: Bounded Convergence Theorem

Let X_n be a sequence of random variables defined on the probability space $(\Omega, \mathcal{F}, P)$. If X_n are bounded and $X_n \to X$ almost surely, then

$$X_n \to X \text{ in } L^1$$

Proof: We have that as X_n are bounded random variables, then $X_n \leq C$ for all n. Therefore

$$\lim_{n\to\infty} \int_\Omega |X_n - X| d\mathbb{P} = \lim_{n\to\infty} \int_{\Omega\setminus N} |X_n - X| d\mathbb{P} = \int_{\Omega\setminus N} \lim_{n\to\infty} |X_n - X| d\mathbb{P} = 0$$

where N is a set of measure zero where $X_n \not\to X$. The interchange of integral and limit arises from the dominating convergence theorem, whereby we then use the fact that $X_n \to X$ a.s. on $\Omega \setminus N$. ■

Remark 9.182 *Recall that in a probability space, convergence in L^1 is also known as **convergence in the mean.** Therefore, if our random variables are bounded and they converge almost surely, then they also converge in mean. The bounded convergence theorem is a weakened form of the dominating convergence theorem whereby the dominating function is now a constant.*

We now state a series of claims regarding the expectation of sequences of random variables.

Claim 9.183 *Let X be a sequence of random variables defined on $(\Omega, \mathcal{F}, P)$. Then, we have the following.*

1. *(Finite additivity) If $X_i \in L^1(\Omega)$, then $E(\sum_{i=1}^n X_i) = \sum_{i=1}^n E(X_i)$.*
2. *(Countable additivity) If $X_n \geq 0$ almost surely, then $E(\sum_{i=1}^\infty X_n) = \sum_{i=1}^\infty E(X_i)$.*
3. *(Finite expectation implies bounded almost surely) If X is a random variable with $E(X) < \infty$, then $P(X < \infty) = 1$.*
4. *If $X_n \geq 0$ almost surely and $\sum_{i=1}^\infty E(X_n) < \infty$ then $\sum_{i=1}^\infty X_i < \infty$ almost surely. Furthermore, $X_n \to 0$ almost surely as $n \to \infty$.*

Proof:(Sketch). (1) Finite additivity can be proven by induction where for n = 2

$$E(X_1 + X_2) = \int_\Omega X_1 + X_2 d\mathbb{P} = \int_\Omega X_1 d\mathbb{P} + \int_\Omega X_2 d\mathbb{P} = E(X_1) + E(X_2) = \sum_{i=1}^2 E(X_i).$$

(2) We can extend finite addivity to countable additivity by defining $Y_n = \sum_{i=1}^n X_i$ and noting that $Y_n \leq Y_{n+1}$. We can then use finite addivitiy and the monotone convergence theorem to get our desired result

$$\lim_{n\to\infty} \sum_{i=1}^n E(X_i) = \lim_{n\to\infty} E(\sum_{i=1}^n X_i) = \lim_{n\to\infty} \int_\Omega Y_n d\mathbb{P} = \int_\Omega \lim_{n\to\infty} Y_n d\mathbb{P} = E(\sum_{i=1}^\infty X_i).$$

(3) We show this by proof by contradiction. Suppose that $\mathbb{P}(X < \infty) \neq 1$. Then $\mathbb{P}(X = \infty) = \epsilon$. Then, define the set S to be the set where $X = \infty$. Then, we have that

$$E(X) = \int_\Omega X d\mathbb{P} = \int_{\Omega\setminus S} X d\mathbb{P} + \int_S X d\mathbb{P} = \infty.$$

(4) First, by countable additivity, we have that

$$E(\sum_{i=1}^{\infty} X_n) = \sum_{i=1}^{\infty} E(X_i) < \infty$$

Then, by (3), a random variable with finite expectation implies that it is bounded almost surely

$$\mathbb{P}(\sum_{i=1}^{\infty} X_i < \infty) = 1 \tag{*}$$

Define the set

$$S = \{\omega : \sum_{i=1}^{\infty} X_i(\omega) < \infty\}$$

and by (*), this implies that $\mathbb{P}(S) = 1$. Now, for a fixed $\omega \in \Omega$, we have that

$$\lim_{n\to\infty} \sum_{i=1}^{n} X_n(\omega) \leq C$$

where we define C to be the supremum bound for all X_i. We can therefore apply the bounded convergence theorem

$$\lim_{n\to\infty} X_n(\omega) = \lim_{n\to\infty} \sum_{i=1}^{n} X_n(\omega) - \lim_{n\to\infty} \sum_{i=1}^{n-} X_n(\omega) = C - C = 0.$$

Therefore, we can conclude that for every $X_n \to 0$ almost surely. ■

We now state a theorem that is useful for us when computing moments.

Theorem 44: L^p-spaces are nested

Suppose X is a measurable function defined on the measure space $(\Omega, \mathcal{F}, \mu)$ with $\mu(\Omega) < \infty$. Then

1. If $X \in L^p(\Omega)$ for $p \geq 1$, then
$$X \in L^r(\Omega)$$
for all $r \in [1, p]$.
2. Furthermore, if $\mu = P$, then
$$||X||_r \leq ||X||_p.$$

Remark 9.184 *We will soon see that if a random variable $X \in L^2(\Omega)$, then that means $X \in L^1(\Omega)$ as well. This will be helpful for computing momentss.*

9.6.5 Convex Functions

We go on a quick detour on convexity.

Definition 56: Convex Function

A function $\phi : I \to \mathbb{R}$ is convex on $I \subset \mathbb{R}$ where I is some open interval if for all $x, y \in I$ and $t \in [0, 1]$, we have that

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y).$$

Lemma 9.185 *The p-norms are convex functions.*

Theorem 45: Jensen's Inequality

Let $\phi : I \to \mathbb{R}$ be a convex function on an open interval $I \subset \mathbb{R}$ and let X be a random variable in L^1 with $P(X \in I) = 1$. Then, $E[\phi(X)]$ is well defined and

$$E[\phi(X)] \geq \phi(E[X]).$$

Corollary 9.186 *Let $\phi(x) = x^2$. Then, we have that*

$$[E(X)]^2 \leq E(X^2)$$

9.6.6 Variance and Covariance

We move on to defining variance and covariance.

Lemma 9.187 *Suppose that $X \in L^1(\Omega)$. Then, $\mu = E(X) < \infty$. Then, $E(X - \mu)^2$ is well-defined.*

We now want to see if $E(X - \mu)^2$ is integrable.

Lemma 9.188 *Let X be a random variable such that $X \in L^1(\Omega)$. We have that $(X - \mu)^2 \in L^1$ if and only if $X \in L^2$.*

Proof: We have that $(X - \mu)^2 = X^2 - 2\mu X + \mu^2$. From this, it is clear that $(X - \mu)^2 \in L^1$ if and only if $X^2 \in L^1$ ■

Definition 57: Variance

Let X be a random variable such that $X \in L^2(\Omega)$. Then, the **variance** of X is defined by

$$Var(X) = E((X - \mu)^2) < \infty.$$

We state a useful theorem involving the variance of a random variable.

Theorem 46: Chebychev's Inequality

For $X \in L^1$ with $\mu = E[X]$, we have that

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{Var(X)}{\epsilon^2}.$$

We now define the notion to relate how 2 random variables relate to each other.

Definition 58: Covariance

Define X, Y to be random variable such that $X, Y \in L^1(\Omega)$. Furthermore, assume that their product $X \cdot Y \in L^1(\Omega)$. We define the covariance of X and Y to be

$$Cov(X,Y) = E[X - E(X)][Y - E(Y)] = E[XY] - E(X) \cdot E(Y)$$

We now extend the notion of variance of a sum of random variables.

Proposition 55: Variance of sum of random variables

Let $X_1, ..., X_n$ be random variables such that $X_1, ..., X_n \in L^2(\Omega)$. Then the product $X_i X_j \in L^1(\Omega)$ for all i, j. Furthermore,

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + \sum_{i,j} Cov(X_i, X_j).$$

Remark 9.189 *Generally, if $X, Y \in L^1(\Omega)$, it does not follow that $X.Y \in L^1(\Omega)$. It does follow if $X, Y \in L^2(\Omega)$.*

Theorem 9.190 *(Cauchy-Schwarz). Let $X_i, X_j \in L^1(\Omega)$. Then*

$$||X_i X_j||_1 \leq ||X_i||_2 \cdot ||X_j||_2$$

Theorem 47: Independence of random variables implies expectations is independent

Let X and Y be random variables such that X,Y are independent and $X, Y \in L^1(\Omega)$. Then $X \cdot Y \in L^1(\Omega)$ and

$$E[XY] = E[X]E[Y].$$

Corollary 9.191 *If X, Y are independent L^1 random variables, then*

1. *$Cov(X, Y) = 0$*
2. *If X, Y are also in $L^2(\Omega)$, then $Var(X + Y) = Var(X) + Var(Y)$.*

Claim 9.192 *If X_i are independent random variables and $\phi_i : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ are measurable functions, then $Y_i = \phi_i(X_i)$ are independent random variables.*

Theorem 48: Version of Strong Law of Large Numbers

Suppose that X_i are independent and identically distributed random variables with finite fourth moment $||X_i||_4 \leq C < \infty$. Let $S_n = \sum_{i=1}^{n} X_i$. Then,

$$\frac{S_n}{n} \to \mu = E(X_i)$$

almost surely as $N \to \infty$.

STAT4028: Probability and Mathematical Statistics

10. Product Measures

10.7 Product Measures

10.7.1 Caratheodory Extension Theorem for Product Measures

We now want to move into working with multiple random variables. This section only looks at the case of having two measurable spaces $(\Omega_j, \mathcal{F}_j)$ for j=1,2. There are two ways to define measures on product spaces. We can use Caratheodory's extension theorem on a semi-algebra of rectangles or we can use the monotone class theorem as seen in lectures. We will illustrate both methods. First, we present a series of steps to derive the product measure using Caratheodory's extension theorem.

Definition 59: Rectangles

Define the measure spaces $(\Omega_i, \mathcal{F}_i, \mu_i)$ for i = 1,2. Define the product space $\Omega_1 \times \Omega_2 = \{(x, y) : x \in \Omega_1, y \in \Omega_2\}$. Then, define $\mathcal{F}_1 \times \mathcal{F}_2$ as the class of rectangles $E_1 \times E_2$ where $E_j \in \mathcal{F}_j$.

Lemma 10.193 *The class of rectangles $\mathcal{F}_1 \times \mathcal{F}_2$ is a semi-algebra.*

Definition 60: x and y section of a rectangle

Define the measure spaces $(\Omega_i, \mathcal{F}_i, \mu_i)$ for i = 1,2. Let $\mathcal{F}_1 \times \mathcal{F}_2$ be the class of rectangles. Then, take a rectangle $A \in \mathcal{F}_1 \times \mathcal{F}_2$. Then, the x-section A_x and the y-section A^y of the rectangle A is defined as

$$\begin{cases} A_x = \{y \in \Omega_2 : (x, y) \in A\} & \text{for } x \in \Omega_1 \\ A^y = \{x \in \Omega_1 : (x, y) \in A\} & \text{for } y \in \Omega_2. \end{cases}$$

Furthermore, $A_x \subseteq \Omega_2$ and $A^y \subseteq \Omega_1$.

We can define measurable sets from the x and y section.

Proposition 56: The x and y section of rectangles are measurable

Let $\mathcal{F}_1 \times \mathcal{F}_2$ be the class of rectangles. Let $A \in \mathcal{F}_1 \times \mathcal{F}_2$ be a rectangle. Denote A_x as the x-section of the rectangle and A^y the y-section of the rectangle. Then

$$\begin{cases} A_x \in \mathcal{F}_2 & \text{for all } x \in \mathcal{F}_2 \\ A^y \in \mathcal{F}_1 & \text{for all } x \in \mathcal{F}_1 \end{cases}$$

We can take the semi-algebra of rectangles, define a pre-measure on it, and from that, use Caratheodory's extension theorem to define the product measure.

Proposition 57: Unique extension to construct product measure

Define the measure spaces $(\Omega_i, \mathcal{F}_i, \mu_i)$ for i = 1,2. Let $\mathcal{F}_1 \times \mathcal{F}_2$ be the class of rectangles. Define the pre-measure on $\mathcal{F}_1 \times \mathcal{F}_2$ as

$$\mu(E_1 \times E_2) = \mu_1(E_1)\mu_2(E_2).$$

Then, $\mu(.)$ is $\sigma-$additive. Furthermore, if μ_i are $\sigma-$finite, then by Caratheodory's extension theorem, we can extend μ uniquely to the $\sigma - algebra$ generated by the class of rectangles $\sigma(\mathcal{F}_1 \times \mathcal{F}_2)$.

10.7.2 Product Measure

We now illustrate the construction of the product measure using the Monotone class theorem.

Definition 10.194 *(Cartesian Product of Sample Spaces) Define the measurable spaces $(\Omega_j, \mathcal{F}_j)$ for j=1,2. Let $\Omega = \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_i \in \Omega_i\}$.*

Claim 10.195 *Define the measurable spaces $(\Omega_j, \mathcal{F}_j)$ for j=1,2. Let $I = \{A_1 \times A_2 : A_i \in \mathcal{F}_i\}$. Then I is a $\pi-$system.*

Proof: Let $A_1 \times A_2 = A \in I$ and $B_1 \times B_2 \in I$ where $A_i \in \mathcal{F}_i$ and $B_i \in \mathcal{F}_i$. Then

$$A \cap B = (A_1 \times A_2) \cap (B_1 \times B_2) = (A_1 \cap B_1) \times (A_2 \cap B_2)$$

We have that $(A_j \cap B_j) \in \mathcal{F}_j$ as $A_j, B_j \in \mathcal{F}_j$ and $\mathcal{F}_j$ is a $\sigma - algebra$. Therefore, by definition of I,

$$A \cap B \in I$$

and therefore we can conclude that I is a $\pi-$system. ■

Definition 61: Product $\sigma - algebra$

Define the measurable spaces $(\Omega_j, \mathcal{F}_j)$ for j=1,2. Let $I = \{A_1 \times A_2 : A_i \in \mathcal{F}_i\}$. Then, $\mathcal{F} = \sigma(I)$ is the **product** $\sigma - algebra$ which we denote by $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$.

Definition 62: Co-ordinate map

Let Ω be the Cartesian product of Ω_1, Ω_2 and $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$. The i^{th} coordinate map $p_i : \Omega \to \Omega_i$ is defined as

$$p_i(\omega_1, \omega_2) = \omega_i.$$

Claim 10.196 *Let p_i be the coordinate map. Then, the map p_i is measurable with respect to the product $\sigma - algebra$ $\mathcal{F}$.*

Proof: First, recall that $\Omega = \Omega_1 \times \Omega_2$. The i-th coordinate map p_i is defined as

$$p_i : \Omega \to \Omega_i$$

where $p_i(\omega_1, \omega_2) \to \omega_i$. Then, to show that p_i is measurable with respect to $\mathcal{F}$, we need to show that for each $A \in \mathcal{F}_i$, we have that $p_i^{-1}(A) \in \mathcal{F}$. It is clear that

$$p_1^{-1}(A_1) = A_1 \times \Omega_2 \quad \text{for all } A_1 \in \mathcal{F}_1$$

$$p_2^{-1}(A_1) = \Omega_1 \times A_2 \quad \text{for all } A_2 \in \mathcal{F}_2$$

As $A_1 \in \mathcal{F}_1$ and $\Omega_2 \in \mathcal{F}_2$, we can conclude that $A_1 \times \Omega_2 \in I$ and therefore $A_1 \times \Omega_2 \in \sigma(I) = \mathcal{F}$. A similar argument holds to argue that $\Omega_1 \times A_2 \in \sigma(I)$. Therefore, p_i is measurable with respect to $\mathcal{F}$. ■

Claim 10.197 *(Equivalent way of constructing the product $\sigma-$algebra). Let p_i be the i^{th} co-ordinate map for i = 1,2. Then, an alternative method to construct the product $\sigma-$algebra would be*

$$\mathcal{F} = \sigma(p_1, p_2).$$

Equivalently, we can construct the product $\sigma-$algebra through

$$\mathcal{F} = \sigma(B_1 \times \Omega_2, \Omega_1 \times B_2 : B_i \in \mathcal{F}_i).$$

Proof: We have already shown that

$$p_1^{-1}(A_1) = A_1 \times \Omega_2 \quad p_2^{-1}(A_2) = \Omega_1 \times A_2$$

for all $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$. Therefore, we have that

$$\sigma(p_1, p_2) = \sigma(A_1 \times \Omega_2, \Omega_1 \times A_2 : A_i \in \mathcal{F}_i)$$

Additionally, we know that

$$(A_1 \times \Omega_2) \cap (\Omega_1 \times A_2) = A_1 \times A_2$$

and therefore, we can conclude that

$$\sigma(p_1, p_2) = \sigma(A_1 \times \Omega_2, \Omega_1 \times A_2 : A_i \in \mathcal{F}_i) = \sigma(A_1 \times A_2 : A_i \in \mathcal{F}_i) = \sigma(I) = \mathcal{F}.$$

■

We now show that a function from the product measurable space is measurable when fixing one of its parameters. However, we first need a theorem that we will use repeatedly.

Theorem 49: The Monotone-Class Theorem

Let $\mathcal{H}$ be a class of bounded functions from $\Omega \to \mathbb{R}$ such that

1. $\mathcal{H}$ is a vector space over $\mathbb{R}$
2. $1_\Omega \in \mathcal{H}$
3. If $f_n \in \mathcal{H}$ satisfies that $0 \leq f_n \leq C < \infty$ and $f_n \uparrow$, then $\lim f_n \in \mathcal{H}$.

Then, if $1_A \in \mathcal{H}$ for all sets A in a $\pi-$system I, then $\mathcal{H}$ contains every bounded $\sigma(I)-$measurable function on Ω.

Remark 10.198 *Condition 2 states that the constant function is in $\mathcal{H}$. Condition 3 is known as closed under bounded convergence.*

Before we can prove the monotone-class theorem, we need a lemma.

Lemma 10.199 *Let $\mathcal{H}$ be the class of bounded functions described in the montone-class theorem. Define the collection*

$$\mathcal{L} = \{A \subset \Omega : 1_A \in \mathcal{H}\}$$

Then, $\mathcal{L}$ is a $\lambda-$system.

Proof: $\Omega \in \mathcal{L}$ by assumption of $\mathcal{H}$. Now, suppose that $A, B \in \mathcal{L}$ with $A \subset B$. Then,

$$1_{B\setminus A} = 1_B - 1_A \in \mathcal{H}$$

as $1_A, 1_B \in \mathcal{H}$ by assumption and $\mathcal{H}$ is a vector space. Finally, suppose that $A_n \in \mathcal{L}$ with $A_n \uparrow A$. Then $1_{A_n} \in \mathcal{H}$ for all $n \in \mathbb{N}$. Then

$$1_{A_n} \uparrow 1_A \in \mathcal{H}$$

due to the closed under bounded convergence property of $\mathcal{H}$. Therefore, $1_A \in \mathcal{H}$ and we can conclude that $A \in \mathcal{L}$. Therefore, $\mathcal{L}$ is a $\lambda-$system. ■

We can prove the montone class theorem.

Proof:(Montone-Class theorem). Let $\mathcal{L} = \{A \subset \Omega : 1_A \in \mathcal{H}\}$. By assumption, I has the property that $1_A \in \mathcal{H}$ for all sets $A \in I$. Therefore

$$I \subset \mathcal{L}$$

and by Dynkin's $\pi - \lambda$ theorem, we have that $\sigma(I) \subset \mathcal{L}$. Now, recall that all bounded simple functions is a linear combination of indicator functions. Therefore, we denote $\delta_b(\sigma(I))$ to be the sets with simple function associated to the sets in $\sigma(I)$. We have that

$$\delta_b(\sigma(I)) \subset \mathcal{L}$$

Now, if $f \in \sigma(I)$ and $0 \leq f \leq M < \infty$, then defining $f_n = \alpha^{(n)} \circ f$, we have that $f_n \in \mathcal{H}$ as $\mathcal{H}$ is a vector space. Furthermore, $0 \leq f_n \leq M$ and hence hence $f_n \uparrow f$ implies that $f \in \mathcal{H}$.

Now, suppose that $f \in \sigma(I)$ is any measurable function $|f| \leq M < \infty$. We can write $f = f^+ - f^-$. By our previous result, $f^+, f^- \in \mathcal{H}$ and therefore $f \in \mathcal{H}$ again due to the fact that $\mathcal{H}$ is a vector space. ■

We now show that functions from the product sample space Ω are measurable with respect to their arguments.

Proposition 58: Measurable functions of the measurable product space

Let $(\Omega, \mathcal{F})$ be a measurable space where Ω is the product sample space and $\mathcal{F}$ is the product $\sigma-$*algebra*. Define the function $f : (\Omega, \mathcal{F}) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ to be a measurable function. Then

1. For all $\omega_1 \in \Omega_1$, we have that $f(\omega_1, .) : (\Omega_2, \mathcal{F}_2) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ is measurable with respect to $\mathcal{F}_2$;
2. For all $\omega_2 \in \Omega_2$, we have that $f(., \omega_2) : (\Omega_1, \mathcal{F}_1) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ is measurable with respect to $\mathcal{F}_1$.

Remark 10.200 *That is, for 1, we can fix a point $\omega_1 \in \Omega_1$. Then, varying $\omega_2 \in \Omega_2$, we have that the function is measurable with respect to $\mathcal{F}_2$.*

Proof: Define the collection of functions

$$\mathcal{H} = \{f : \Omega \to \mathbb{R} \text{ such that f is bounded and satisfies conditions (1) and (2)}\}$$

We can show that $\mathcal{H}$ satisfies the conditions for the monotone class theorem for the set of real-valued bounded functions. Now, recall the $\pi-$system

$$I = \{A_1 \times A_2 : A_i \in \mathcal{F}_i\}$$

We have that $1_A \subset \mathcal{H}$ for all sets $A \in I$ where we define

$$1_A = 1_{A_1 \times A_2}(\omega_1, \omega_2) = 1_{A_1}(\omega_1)1_{A_2}(\omega_2)$$

Therefore, by the monotone class theorem, we have that $\mathcal{H}$ contains every bounded $\sigma(I)-$measurable function. Now, recal lthat by definition, the product $\sigma - algebra = \sigma(I) = \sigma(p_1, p_2)$ where p_i is the i-th projection map, which is measurable with respect to $\sigma - algebra$. Therefore, by (*)

$$\sigma(I) = \{\text{bounded and } \mathcal{F} - \text{measurable functions}\} \subset \mathcal{H}$$

However, by definition of $\mathcal{H}$ satisfying the conditions (1) and (2), we can conclude that all functions in $\mathcal{H}$ are bounded and satisfy conditions (1) and (2). ■

It is worth noting that our construction of the product $\sigma - algebra$ works for a countable product of measure spaces but problematic if we had an uncountable product of measure spaces as $\sigma - algebra$ are not necessarily closed by uncountable intersections, which was an operation we used frequently in our construction of the product $\sigma - algebra$.

10.7.3 Fubini's Theorem

We could define the product measure by taking rectangles $A_1 \times A_2$ where $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$ and using Caratheodory's theorem to extend the measure $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$. However, Fubini's theorem lets us define the product measure in a different manner.

Recall that we said that $f(\omega_1, .)$ is measurable with respect to $\mathcal{F}_2$ for all $\omega_1 \in \Omega_1$. That means we can integrate $f(\omega_1, .)$ with respect to μ_2. Likewise, as $f(\cdot, \omega_2)$ is measurable with respect to $\mathcal{F}_1$, we can integrate $f(\cdot, \omega_2)$ with respect to μ_1.

Definition 63: Integral functions

Assume that μ_i are finite measures on $(\Omega_i, \mathcal{F}_i)$ for i=1, 2. Define the function $f : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ to be bounded and measurable. Then $f(\omega_1, .)$ and $f(., \omega_2)$ are bounded and measurable. We define the integral functions

$$\begin{cases} I_1(\omega_1; f) = \int_{\Omega_2} f(\omega_1, .)d\mu_2 \\ \\ I_2(\omega_2; f) = \int_{\Omega_1} f(., \omega_2)d\mu_1 \end{cases}$$

which are well defined functions from $\Omega_i \to \mathbb{R}$. That is, $I_1 : \Omega_1 \to \mathbb{R}$ and $I_2 : \Omega_2 \to \mathbb{R}$.

Remark 10.201 *As the function f is bounded and measurable, this means that its integral is well-defined.*

We now have integral functions $I_i : \Omega_i \to \mathbb{R}$. We now want to again integrate these integrals, but before that, we need to show that I_1 is measurable with respect to $\mathcal{F}_1$ and I_2 is measurable with respect to $\mathcal{F}_2$.

Proposition 59: First form of Fubini's theorem

Define the measure spaces $(\Omega_i, \mathcal{F}_i, \mu_i)$ for $i = 1, 2$. Define the function $f : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ to be bounded and measurable. We then have the following.

1. $I_1(\cdot; f) : (\Omega_1, \mathcal{F}_1) \to (\mathbb{R}, \mathcal{B})$ is bounded and measurable with respect to $\mathcal{F}_1$.
2. $I_2(\cdot; f) : (\Omega_2, \mathcal{F}_2) \to (\mathbb{R}, \mathcal{B})$ is bounded and measurable with respect to $\mathcal{F}_2$.
3. We have Fubini's theorem
$$\int_{\Omega_1} I_1(\cdot; f) d\mu_1 = \int_{\Omega_2} I_2(\cdot; f) d\mu_2.$$

Proof: First, we define the class of all bounded functions

$$\mathcal{H} = \{f : \Omega \to \mathbb{R} \text{ such that f is bounded and satisfy conditions 1, 2, 3}\}$$

We can show that $\mathcal{H}$ satisfy all the requirements for the monotone class theorem from before. Furthermore, $\mathcal{H}$ contains 1_A for every set $A \in I = \{A_1 \times A_2 : A_i \in \mathcal{F}\}$ where we define the indicator function

$$1_A(\omega_1, \omega_2) = 1_{A_1}(\omega_1) 1_{A_2}(\omega_2)$$

Now, recall that $\sigma(I)$ is the set of all $\sigma-$measurable function. Then, by the monotone class theorem

$$\mathcal{H} \supset \text{ every bounded } \mathcal{F} - \text{measurable function}$$

By construction,

$$\mathcal{H} \subset \text{ every bounded } \mathcal{F} - \text{measurable function}$$

Therefore, we can conclude that for any function in the class $f \in \mathcal{H}$, it satisfies conditions 1, 2, 3. ■

Remark 10.202 *Note that we now further integrate I_1 (which itself is an integral with respect to μ_2) with respect to μ_1.*

We now can define the product measure using Fubini's theorem.

Definition 64: Product Measure

Let $(\Omega_i, \mathcal{F}_i, \mu_i)$ be measure spaces and μ_i is finite for $i = 1, 2$. We define the **product measure** $\Omega : \mathcal{F} \to \mathbb{R}^+$ as

$$\mu(A) = \int_{\Omega_1} I_1(\cdot, 1_A) d\mu_1 = \int_{\Omega_2} I_2(\cdot, 1_A) d\mu_2$$

for $A \in \mathcal{F}$.

Remark 10.203 *More explicitly,*

$$\begin{cases} \int_{\Omega_1} I_1(\cdot, 1_A) d\mu_1 = \int_{\Omega_1} \left(\int_{\Omega_2} 1_A(\omega_1, \cdot) d\mu_2 \right) d\mu_1 \\ \int_{\Omega_2} I_2(1_A, \cdot) d\mu_2 = \int_{\Omega_2} \left(\int_{\Omega_1} 1_A(\cdot, \omega_2,) d\mu_1 \right) d\mu_2 \end{cases}$$

Proposition 60: Uniqueness of the product measure

Let μ be the product measure defined above. We then have the following claims.

1. μ is a measure on the measure product space $(\Omega, \mathcal{F})$.
2. $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ for all $A_i \in \mathcal{F}_i$.
3. If ν is another measure on $(\Omega, \mathcal{F})$ with
$$\nu(A_1 \times A_2) = \mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$$
for all $A_i \in \mathcal{F}_i$. Then $\mu = \nu$.

That is, the product measure μ is a unique extension of μ_1, μ_2.

Proof: For (1), recall that

$$\mu(A) = \int_{\Omega_1} I_1(\cdot; 1_A) d\mu_1 = \int_{\Omega_1} \int_{\Omega_2} 1_A(\omega_1, \omega_2) d\mu_2 d\mu_1$$

Let $A = \emptyset$. Then

$$\mu(\emptyset) = \int_{\Omega_1} I_1(\cdot; 1_\emptyset) d\mu_1 = \int_{\Omega_1} \int_{\Omega_2} 1_\emptyset(\omega_1, \omega_2) d\mu_2 d\mu_1 = \int_{\Omega_1} 0 d\mu_1 = 0$$

Therefore $\mu(\emptyset) = 0$. Now, suppose that $A_1, A_2, ... \in \mathcal{F}$ are disjoint sets. Then, note that for a fixed $\omega_1 \in \Omega_1$ we have that

$$I_1(\omega_1; 1_{\cup A_k}) = \int_{\Omega_2} 1_{\cup A_k}(\omega_1, \omega_2) d\mu_2 = \int_{\Omega_2} \sum_{k=1}^{\infty} 1_{A_k}(\omega_1, \omega_2) d\mu_2 = \int_{\Omega_2} \lim_{n\to\infty} \sum_{k=1}^{n} 1_{A_k}(\omega_1, \omega_2) d\mu_2$$

and then by monotone convergence theorem

$$= \lim_{n\to\infty} \sum_{k=1}^{n} \int_{\Omega_2} 1_{A_k}(\omega_1, \omega_2) d\mu_2 = \sum_{k=1}^{\infty} I_1(\omega_1; 1_{A_k}) \tag{*}$$

Therefore, we have that

$$\mu(\cup_k A_k) = \int_{\Omega_1} I_1(\omega_1; 1_{\cup_{A_k}}) d\mu_1 = \int_{\Omega_1} \sum_{k=1}^{\infty} I_1(\omega_1; 1_{A_k}) d\mu_1 = \int_{\Omega_1} \lim_{n\to\infty} \sum_{k=1}^{n} \int_{\Omega_2} 1_{A_k}(\omega_1, \omega_2) d\mu_2 d\mu_1$$

$$= \lim_{n\to\infty} \sum_{k=1}^{n} \int_{\Omega_1} \int_{\Omega_2} 1_{A_k}(\omega_1, \omega_2) d\mu_2 d\mu_1$$

$$= \sum_{k=1}^{\infty} \int_{\Omega_1} I_1(\omega_1; 1_{A_k}) d\mu_1 = \sum_{k=1}^{\infty} \mu(A_k)$$

by definition.

Now for claim 2, recall that $1_{A_1 \times A_2}(\omega_1, \omega_2) = 1_{A_1}(\omega_1) 1_{A_2}(\omega_2)$. Then, we have that

$$I_1(\omega_1; 1_{\cup A}) = \int_{\Omega_2} 1_A(\omega_1, \omega_2) d\mu_2 = 1_{A_1}(\omega_1) \int_{\Omega_2} 1_{A_2}(\omega_2) d\mu_2$$

Therefore, we have that

$$\mu(A) = \mu(A_1 \times A_2) = \int_{\Omega_1} I_1(\omega_1; A_1 \times A_2) d\mu_1 = \int_{\Omega_1} 1_{A_1}(\omega_1) \int_{\Omega_2} 1_{A_2}(\omega_2) d\mu_2 d\mu_1$$

$$= \int_{\Omega_1} 1_{A_1}(\omega_1) d\mu_1 \int_{\Omega_2} 1_{A_2}(\omega_2) d\mu_2 = \mu_1(A_1)\mu_2(A_2)$$

For (3), first recall that $\mathcal{F} = \sigma(I) = \sigma(A_1 \times A_2 : A_i \in \mathcal{F}_i)$. Furthermore, recall that if two probability measures μ, ν agree on a $\pi-$system I, then they agree on the $\sigma - algebra$ generated by that $\pi-$system $\sigma(I)$. Therefore, as I was shown to be a $\pi-$system where $\nu(A_1 \times A_2) = \mu(A_1 \times A_2)$ agree for all $A_1 \times A_2 \in I$, then they also agree on $\sigma(I) = \mathcal{F}$. ■

We now a product measure using the Lebesgue measure.

Definition 65: Lebesgue's Measure on $\mathbb{R}^2$

There exists a unique measure λ_2 on $(\mathbb{R}^2, \mathcal{B}^2)$ such that for any intervals $I_j = (a_j, b_j]$, we have that

$$\lambda_2(I_1 \times I_2) = \lambda(I_1)\lambda(I_2) = (b_1 - a_1)(b_2 - a_2).$$

Here, $\mathcal{B}^2 = \sigma(\mathcal{B} \times \mathcal{B}) = \mathcal{B}(\mathbb{R}^2)$.

We now state two important theorems. Here, we relax the condition for f to be bounded to now being non-negative or integrable.

Theorem 50: Tonelli's Theorem

Let $(\Omega_j, \mathcal{F}_j, \mu_j)$ be measure spaces for j=1,2 where Ω_j is $\sigma-$finite with respect to μ_j. Define μ to be the product measure on the product measure space $(\Omega, \mathcal{F})$. Take a non-negative measurable function $f : \Omega \to \overline{\mathbb{R}}^+$. Then

1. $I_1(.; f) : (\Omega_1, \mathcal{F}_1) \to (\mathbb{R}, \mathcal{B})$ is bounded and measurable with respect to $\mathcal{F}_1$.
2. $I_2(.; f) : (\Omega_2, \mathcal{F}_2) \to (\mathbb{R}, \mathcal{B})$ is bounded and measurable with respect to $\mathcal{F}_2$.
3. $$\int_{\Omega_1} I_1(.; f) d\mu_1 = \int f d\mu = \int_{\Omega_2} I_2(.; f) d\mu_2.$$

Theorem 51: Fubini's Theorem

Let $(\Omega_j, \mathcal{F}_j, \mu_j)$ be measure spaces for j=1,2 where Ω_j is $\sigma-$finite with respect to μ_j. Define μ to be the product measure on the product measure space $(\Omega, \mathcal{F})$. Take an integrable and measurable function $f : \Omega \to \mathbb{R}$ where $f \in L^1(\Omega)$. Then

1. $I_1(.; f) : (\Omega_1, \mathcal{F}_1) \to (\mathbb{R}, \mathcal{B}) \in L^1(\Omega_1)$.
2. $I_2(.; f) : (\Omega_2, \mathcal{F}_2) \to (\mathbb{R}, \mathcal{B}) \in L^1(\Omega_2)$.
3. $$\int_{\Omega_1} I_1(.; f) d\mu_1 = \int f d\mu = \int_{\Omega_2} I_2(.; f) d\mu_2.$$

10.7.4 Product Measure on n-dimensional space

We can actually what we have seen to if we had $(\Omega_i, \mathcal{F}_i, \mu_i)$ for i = 1,...,n that are $\sigma-$finite measure spaces. We can inductively define a product $\sigma-algebra$ by $\mathcal{F} = \mathcal{F}_1 \times ... \times \mathcal{F}_n$ and a product measure by $\mu = \mu_1 \times ... \times \mu_n$.

Proposition 61: Tonelli's Theorem on n-dimensional space

Define the product space $(\Omega, \mathcal{F})$ of n measurable spaces $(\Omega_i, \mathcal{F}_i, \mu_i)$. Then, let $f : (\Omega, \mathcal{F}) \to (\mathbb{R}^+, \mathcal{B})$ be a nonnegative and measurable function. Then, for any permutation $\pi \in S_n$

$$\int f d\mu = \int_{\Omega_{\pi(1)}} ... \int_{\Omega_{\pi(n)}} f d\mu_{\pi(1)} ... d\mu_{\pi(n)}.$$

Remark 10.204 *The same holds for Fubini's theorem if* $f \in L^1(\Omega, \mathcal{F}, \mu)$.

Proposition 62: Lebesgue measure on $\mathbb{R}^n$

There exists a unique Lebesgue measure λ_n on $(\mathbb{R}^n, \mathcal{B}^n)$ such that for any intervals $I_j = (a_j, b_j]$

$$\lambda_n(I_1 \times ... \times I_n) = \prod_{j=1}^{n} \lambda(I_j) = \prod_{j=1}^{n} \Big(b_j - aj\Big)$$

10.7.5 Random Vectors

We now want to work in dimensions of size n.

Definition 66: Random Vector

A measurable function $\mathbf{X} = (X_1, ..., X_n) : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}^n)$ is called a **random vector.**

Claim 10.205 ***X*** *is a random vector if and only if* X_i *is a random variable for i = 1,...,n.*

Definition 67: Probability measure induced by random vector

Define the product measure space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, the random vector $\mathbf{X} = (X_1, ..., X_n)$ induces a probability measure on $(\overline{\mathbb{R}}^n, \overline{\mathcal{B}}^n)$ by

$$\mu_{\mathbf{X}}(B) = \mathbb{P}(\mathbf{X} \in B)$$

for all Borel set $B \in \overline{\mathcal{B}}^n$. $\mu_{\mathbf{X}}$ is called the probability measure induced by the random vector X.

Definition 68: Density function of random vector

Define the product measure space $(\Omega, \mathcal{F}, \mathbb{P})$. Define the random vector $\mathbf{X} = (X_1, ..., X_n)$ and its induced probability measure $\mu_{\mathbf{X}}$. Then, if $\mu_{\mathbf{X}}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$ λ_n if $\mu_{\mathbf{X}}(\mathbb{R}^n) = 1$ and there exists a non-negative and measurable function $f : \mathbb{R}^n \to [0, \infty]$ such that

$$\mu_{\mathbf{X}}(B) = \int_B f d\lambda_n$$

for all Borel sets $B \in \mathcal{B}^n$. Then, f is called the density of $\mu_{\mathbf{X}}$.

Claim 10.206 *If $\boldsymbol{X} = (X_1, ..., X_n)$ has a density, then so does each X_i for i = 1,...,n. Furthermore,*

$$f_{X_i}(x) = \int ... \int f(\boldsymbol{x}) d\lambda(x_1)...d\lambda(x_{i-1})d\lambda(x_{i+1})...d\lambda(x_n).$$

Claim 10.207 *If the random variables $X_1, ..., X_n$ are independent and if each X_i has density f_i, then $\boldsymbol{X} = (X_1, ..., X_n)$ has a density*

$$f(x_1, ..., x_n) = \prod_{i=1}^{n} f_i(x_i).$$

10.7.6 Change of variables for Random Vectors

We now have a multi-dimensional analogue of the change of variables theorem.

Theorem 52: Change of variables formula for non-negative randon vector

Let $\mathbf{X}$ be a random vector defined on a probability $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mu_{\mathbf{X}}$ be its induced measure on $(\overline{\mathbb{R}}^+, \overline{\mathcal{B}}^n)$. Suppose that $h : (\overline{\mathbb{R}}, \overline{\mathcal{B}}^n) \to (\overline{\mathbb{R}^+}, \overline{\mathcal{B}}^+)$ is measurable and let $\mathbf{Y} = h(\mathbf{X})$. Then $\mathbf{Y}$ is a non-negative random variable and

$$\int \mathbf{Y} d\mathbb{P} = E(\mathbf{Y}) = E(h(\mathbf{X})) = \int_{\mathbb{R}^n} h d\mu_{\mathbf{X}}.$$

Theorem 53: Change of variables formula for measurable randon vector

Let $\mathbf{X}$ be a random vector defined on a probability $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mu_{\mathbf{X}}$ be its induced measure on $(\overline{\mathbb{R}}^+, \overline{\mathcal{B}}^n)$. Suppose that $h : (\overline{\mathbb{R}}, \overline{\mathcal{B}}^n) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ is measurable and let $\mathbf{Y} = h(\mathbf{X})$. Then $\mathbf{Y}$ is a random variable.

1. $E(\mathbf{Y}) = \int \mathbf{Y} d\mathbb{P}$ is defined if and only if $\int_{\mathbb{R}^n} h d\mu_{\mathbf{X}}$ is defined and then $E(\mathbf{Y}) = \int_{\mathbb{R}^n} h d\mu_{\mathbf{X}}$.
2. $\mathbf{Y} = h(\mathbf{X}) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ if and only if $h \in L^1(\overline{\mathbb{R}}^n, \overline{\mathcal{B}}^n, \mu_{\mathbf{X}}$. Then

$$\int \mathbf{Y} d\mathbb{P} = E(\mathbf{Y}) = E(h(\mathbf{X})) = \int_{\mathbb{R}^n} h d\mu_{\mathbf{X}}.$$

Theorem 10.208 *If the random vector* $\boldsymbol{X}$ *has a density* f*, then*

$$E(\boldsymbol{Y}) = E(h(\boldsymbol{X})) = \int_{\mathbb{R}^n} h \circ f d\lambda_n$$

with the LHS defined if and only if the RHS is. Furthermore, if $h \circ f$ *is Riemann integrable, we can compute* $E(\boldsymbol{Y})$ *using iterated* ***Riemann integrals***

$$\int_{\mathbb{R}^n} h \circ f d\lambda_n = \int_{\mathbb{R}} ... \int_{\mathbb{R}} h \circ f d\lambda(x_1)...d\lambda(x_n).$$

STAT4028: Probability and Mathematical Statistics

11. Conditional Expectation

11.8 Conditional Expectation

11.8.1 Introduction

Definition 69: Conditional Expectation

Let $(\Omega, \mathcal{F}, P)$ be a probabiliy space and $X : \Omega \to \mathbb{R}$ be a random variable. Let $X \in L^1(\Omega, \mathcal{F}, P)$. Moreover, let $\mathcal{F}_0 \subseteq \mathcal{F}$ be a $\sigma - algebra$.
Then, the random variable X_0 is called the conditional expectation of X given $\mathcal{F}_0$ if

1. X_0 is measurable with respect to the $\sigma - algebra$ $\mathcal{F}_0$;
2. We have that $\int_A X_0 dP = \int_A X dP$ for all $A \in \mathcal{F}_0$.

If both conditions are satisfied, we write

$$E[X|\mathcal{F}_0] \coloneqq X_0$$

and say that X_0 is a **version of** $E[X|\mathcal{F}_0]$.

We can show that the conditional expectation exists and is unique. First, we require a lemma.

Lemma 11.209 *Let X_0 be a random variable that satisfies the 2 conditions to be a version of the conditional expectation $E[X|\mathcal{F}_0]$. Then, $X_0 \in L^1(\Omega, \mathcal{F}, P)$.*

Proposition 63: Existence and uniqueness of conditional expectation

For every random variable $X \in L^1(\Omega, \mathcal{F}, P)$ and every $\sigma-algebra$ $\mathcal{F}_0 \subseteq \mathcal{F}$, the conditional expectation $E[X|\mathcal{F}_0]$ exists and is unique.

STAT4028: Probability and Mathematical Statistics

12. Conditional Expectation

12.8.2 Properties of conditional expectation

Proposition 64: Properties of conditional expectation

Define the probability space $(\Omega, \mathcal{F}, P)$. Define two random variables X and Y. Assume that $X, Y \in L^1(\Omega, \mathcal{F}, P)$. Then we have the following.

1. $E(aX + bY|\mathcal{F}) = aE(X|\mathcal{F}) + E(Y|\mathcal{F})$;
2. If $X \leq Y$ then $E(X|\mathcal{F}) \leq E(Y|\mathcal{F})$;
3. If $X_n \geq 0$ and $X_n \uparrow X$ with $E[X] < \infty$, then

$$E[X_n|\mathcal{F}] \uparrow E[X|\mathcal{F}].$$

Theorem 12.210 *Let ϕ be a convex function and let $E[|X|], E[|\phi(X)|] < \infty$. Then, we have that*

$$\phi(E[X|\mathcal{F}]) \leq E[\phi(X)|\mathcal{F}].$$

Theorem 12.211 *If $X \in \mathcal{F}$ and $E[|Y|], E[|XY|] < \infty$ then*

$$E[XY|\mathcal{F}] = XE[Y|\mathcal{F}].$$

STAT4028: Probability and Mathematical Statistics

13. Cramér-Rao Lower Bound

13.9 Exponential Families

We revisit ideas on the Cramér-Rao lower bound and examine the assumptions behind them. We additionally revisit ideas from previous courses.

13.9.1 Moment Generating Functions

Let X be a random variable. Recall that the k-th moment of X is given by $E[X^k]$.

Definition 70: Moment Generating Function

Let X be a random variable with CDF F_X. The Moment Generating Function (MGF) M_X of X is given by

$$M_X(t) = E[e^{tX}]$$

provided that the expectation exists for t in a neighbourhood of 0.

Hence, if X is a continuous random variable, we can write

$$M_X(t) = \int_{supp(X)} e^{tx} f_X(x) dx$$

where $supp(X)$ is the support of the random variable X.

Theorem 54: Moments of random variable

If the random variable X has the MGF M_X, then

$$E[X^n] = \frac{d^n}{dt^n} M_X(t)|_{t=0}.$$

That is, the n-th moment of X is given by the n-th derivative of the MGF of X evaluated at t = 0.

Proof: We assume that we can interchange differentiation with respect to t and the integral.

$$\begin{aligned}
\frac{d}{dt} M_X(t) &= \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \\
&= \int_{-\infty}^{\infty} \frac{d}{dt} e^{tx} f_X(x) dx \\
&= \int_{-\infty}^{\infty} \left(x e^{tx} \right) f_X(x) dx \\
&= E[X e^{tX}].
\end{aligned}$$

Then, evaluating at t = 0

$$\frac{d}{dt}M_X(t)|_{t=0}= E[Xe^{0X}] = E[X].$$

We can follow a similar argument for the n-th moment. ■

Definition 71: Cumulant Generating Function

Let X be a random variable with MGF M_X. Then, the cumulant generating function $K_X(t)$ of X is given by

$$K_X(t) = log(M_X(t)) = logE[e^{tX}].$$

13.9.2 Sufficient Statistics

Definition 13.212 *(Statistic). A statistic $T(\underset{\sim}{X})$ is a function of the data vector $\underset{\sim}{X}$.*

Definition 72: Sufficient Statistic

A statistic $T(\underset{\sim}{X})$ is a **sufficient statistic** for θ if the conditional distribution of the sample $\underset{\sim}{X}$ given the value $T(\underset{\sim}{X})$ does not depend on θ.

Theorem 13.213 *(Sufficiency principle). If $T(\underset{\sim}{X})$ is a sufficient statistic for the parameter θ, then any inference on θ should depend on the sample $\underset{\sim}{X}$ only through the sufficient statistic $T(\underset{\sim}{X})$.*

We can say that 2 samples $\underset{\sim}{x}, \tilde{y}$ are equal if $T(\underset{\sim}{x}) = T(\tilde{y})$ even if $\underset{\sim}{x} \neq \tilde{y}$.

We can view that the function $T(.)$ is a partition of the sample space $\mathcal{X}$ whereby

$$\mathcal{T} = \{t : t = T(\underset{\sim}{x}) \text{ for a } \underset{\sim}{x} \in \mathcal{X}\}.$$

Hence, we can partition the sample space

$$\mathcal{X} = \bigcup_{t\in\mathcal{T}} A_t$$

whereby $A_t = \{\underset{\sim}{x} \in \mathcal{X} : T(\underset{\sim}{x}) = t\}$.

Theorem 13.214 *If $p(\underset{\sim}{x}|\theta)$ is the joint PMF/PDF of $\underset{\sim}{X}$ and $q(t|\theta)$ is the PMF/PDF of $T(\underset{\sim}{X})$, then $T(\underset{\sim}{X})$ is a sufficient statistic for θ if for every $x \in \mathcal{X}$ the ratio $\frac{p(\underset{\sim}{x}|\theta)}{q(T(\underset{\sim}{x})|\theta)}$ is constant as a function of θ.*

Theorem 55: Factorisation Theorem

Let $f(\underset{\sim}{x}|\theta)$ be the joint PMF/PDF of the sample. Then, $T(\underset{\sim}{X})$ is a sufficient statistic for θ if and only if there exists functions $g(t|\theta)$ and $h(\underset{\sim}{x})$ such that for all sample points $x \in \mathcal{X}$ and all parameter points $\theta \in \Theta$,

$$f(\underset{\sim}{x}|\theta) = g(T(\underset{\sim}{x})|\theta)h(\underset{\sim}{x}).$$

13.9.3 Introduction

Suppose that $\{p_\theta(.) : \theta \in \Theta\}$ is a 1-parameter family of densities with respect to a $\sigma-$finite measure ν on $\mathbb{R}^d$, where d is the dimension of the data.

Lemma 13.215 *Given a density p_θ and measure ν, we can define a probability measure*

$$\mathbb{P}_\theta(B) = \int_B p_\theta(\underset{\sim}{x}) d\nu(\underset{\sim}{x})$$

where B is a Borel set and ν is either the Lebesgue or counting measure.

Definition 73: Score function

Let $p_\theta(\underset{\sim}{x})$ be the density function. Then, the score function is defined as

$$\ell_\theta^\circ(\underset{\sim}{x}) = \frac{\partial}{\partial\theta} log\ p_\theta(\underset{\sim}{x}).$$

Definition 13.216 *(Support). The support $\mathcal{X}_\theta$ is the smallest set whose complement has 0 probability* $\mathbb{P}_\theta(\mathcal{X}_\theta^c) = 0$.

Definition 74: Assumption 1: Support of distribution does not depend on the parameter

The support $\mathcal{X}_\theta = \mathcal{X}$ does not depend on the parameter θ.

Remark 13.217 *This rules out the uniform $U[0, \theta]$ distribution.*

From assumption 1, we have that the integral of the density over the support is 1. That is

$$\mathbb{P}_\theta(\mathcal{X}) = \int ... \int_{\mathcal{X}} p_\theta(\underset{\sim}{x}) d\nu(\underset{\sim}{x}) = 1 \tag{13.1}$$

From this, we can state our next assumption.

Definition 75: Assumption 2: Density is twice differentiable

The density function p_θ is twice differentiable with respect to θ everywhere inside the integral sign. That is,

$$p_\theta^{'} = \frac{\partial p_\theta}{\partial\theta}$$

$$p_\theta^{''} = \frac{\partial^2 p_\theta}{\partial\theta^2}$$

exists for all $x \in \mathcal{X}$.

Using assumption (1) and assumption (2), we have that the integral of the derivatives of the density function is zero

$$\int ... \int_{\mathcal{X}} p_\theta^{'}(\underset{\sim}{x}) d\nu(\underset{\sim}{x}) = 0 \tag{13.2}$$

$$\int ... \int_{\mathcal{X}} p_{\theta}^{''}(\underset{\sim}{x}) d\nu(\underset{\sim}{x}) = 0 \tag{13.3}$$

for all $\theta \in \Theta$.

Since the density function is non-negative $p_{\theta}(\underset{\sim}{x}) > 0$ for all $x \in \mathcal{X}$, this means we can divide and multiply by the density in equation (13.2).

$$\int ... \int_{\mathcal{X}} \frac{p_{\theta}^{'}(\underset{\sim}{x})}{p_{\theta}(\underset{\sim}{x})} p_{\theta}(\underset{\sim}{x}) d\nu(\underset{\sim}{x}) = \int ... \int_{\theta} \ell_{\theta}^{\circ}(\underset{\sim}{x}) p_{\theta}(\underset{\sim}{x}) d\nu(\underset{\sim}{x}) = \mathbb{E}_{\theta}[\ell_{\theta}^{\circ}(\underset{\sim}{x})] = 0. \tag{13.1}$$

This leads to an interesting result.

Proposition 65: Expectation of the score function is zero

Under assumption (1) and (2), the expectation of the score function $\ell_{\theta}^{\circ}(\underset{\sim}{x})$ is 0

$$\mathbb{E}_{\theta}[\ell_{\theta}^{\circ}(\underset{\sim}{x})] = 0.$$

We can show that the second derivative of the score function as

$$\ell_{\theta}^{\circ\circ} = \frac{\partial^2}{\partial \theta^2} log\ p_{\theta} = \frac{p_{\theta}^{''}}{p_{\theta}} - \ell_{\theta}^{\circ}.$$

The variance of the score function is therefore given by

$$Var_{\theta}[\ell_{\theta}^{\circ}(\underset{\sim}{x})] = -\int ... \int_{\mathcal{X}} \ell_{\theta}^{\circ\circ} p_{\theta} d\nu(\underset{\sim}{x}).$$

Definition 13.218 *(Unbiased estimator). The function $\hat{\theta}(\underset{\sim}{x})$ is an unbiased estimator of the parameter θ if*

$$E_{\theta}[\hat{\theta}(\underset{\sim}{x})] = \int ... \int_{\mathcal{X}} \hat{\theta}(\underset{\sim}{x}) p_{\theta}(\underset{\sim}{x}) d\nu(\underset{\sim}{x}) = \theta$$

for all $\theta \in \Theta$.

We can take the derivative of the expression for an unbiased estimator to derive our next assumption.

Definition 76: Assumption 3: Differentiate expression for unbiased estimators inside integral sign

Define the function $\hat{\theta}(\underset{\sim}{x})$ to be an unbiased estimator of the parameter θ denoted by

$$E_{\theta}[\hat{\theta}(\underset{\sim}{x})] = \int ... \int_{\mathcal{X}} \hat{\theta}(\underset{\sim}{x}) p_{\theta}(\underset{\sim}{x}) d\nu(\underset{\sim}{x}) = \theta$$

We then assume that we are able to differentiate inside this expression to get

$$\int ... \int_{\mathcal{X}} \hat{\theta}(\underset{\sim}{x}) p_{\theta}^{'}(\underset{\sim}{x}) d\nu(\underset{\sim}{x}) = 1$$

Regarding assumptions 2 and 3, we note that this is in fact an application of Leibnitz's rule.

Proposition 66: Leibnitz's Rule

Suppose that $f(x,\theta)$ is a function with parameter θ. Suppose that there exists a dominating function $g \in L^1$. Then

$$\frac{d}{d\theta}\int_a^b f(x,\theta)dx = \int_a^b \frac{\partial}{\partial\theta} f(x,\theta)dx$$

Leibnitz's rule essentially asks when are we able to interchange limits and integration as the derivative is a form of a limit

$$\frac{\partial}{\partial\theta} f(x,\theta) = \lim_{\delta\to 0}\frac{f(x,\theta+\delta)-f(x,\theta)}{\delta}.$$

In particular, we are able to do so under the assumptions of the dominating convergence theorem.

We state an important result of assumption 3.

Proposition 67: The covariance of unbiased estimator and score function is 1

Assume that assumptions (1)-(3) holds. Let $\hat{\theta}(.)$ be an unbiased estimator of θ. Furthermore, let $\ell^{\circ}_{\theta}(\underset{\sim}{x})$ be the score function. Then, the covariance of the unbiased estimator and score function is

$$Cov_{\theta}[\hat{\theta}(\underset{\sim}{x}), \ell^{\circ}_{\theta}(\underset{\sim}{x})] = 1.$$

Proof: We have that

$$1 = \int ... \int_{\mathcal{X}} \hat{\theta}(\underset{\sim}{x}) p'_{\theta}(\underset{\sim}{x}) d\nu(\underset{\sim}{x})$$

$$= \int ... \int_{\mathcal{X}} \hat{\theta}(\underset{\sim}{x}) \ell^{\circ}_{\theta}(\underset{\sim}{x}) p_{\theta}(\underset{\sim}{x}) d\nu(\underset{\sim}{x})$$

due to the fact that the expectation of the score function is zero and by the definition of covariance

$$= Cov_{\theta}[\hat{\theta}(\underset{\sim}{x}), \ell^{\circ}_{\theta}(\underset{\sim}{x})].$$

■

Hence, if we have an unbiased estimator, its covariance with the score function is 1. We can use this result to determine the variance of our estimator.

Definition 13.219 *(Correlation Inequality). The correlation inequality between an unbiased estimator $\hat{\theta}$ and score function ℓ°_{θ} is*

$$Cov_{\theta}[\hat{\theta}(\underset{\sim}{x}), \ell^{\circ}_{\theta}(\underset{\sim}{x})]^2 \leq Var_{\theta}(\hat{\theta}(\underset{\sim}{x})) Var_{\theta}(\ell^{\circ}_{\theta}(\underset{\sim}{x}))$$

with equality if and only if $\hat{\theta}(\underset{\sim}{x})$ and $\ell^{\circ}_{\theta}(\underset{\sim}{x})$ are lienarly related so that

$$\ell^{\circ}_{\theta}(\underset{\sim}{x}) = a_{\theta} + b_{\theta}\hat{\theta}(\underset{\sim}{x}).$$

Definition 77: Cramer-Rao Lower Bound

Suppose that $\hat{\theta}$ is an unbiased estimator of θ. Furthermore, let $\ell_\theta^\circ(\underset{\sim}{x})$ be the score function. Then, the Cramer-Rao lower bound for the estimator $\hat{\theta}$ is

$$Var_\theta(\hat{\theta}(\underset{\sim}{x})) \geq \frac{1}{Var_\theta(\ell_\theta^\circ(\underset{\sim}{x}))}.$$

An estimator is minimum variance if the score function is a linear function of the estimator.

Proposition 68: Condition for minimum variance estimator

Suppose that $\hat{\theta}$ is an unbiased estimator of θ. Furthermore, let $\ell_\theta^\circ(\underset{\sim}{x})$ be the score function. Then, the estimator $\hat{\theta}$ attains the Cramer-Rao lower bound if the score function is of the form

$$\ell_\theta^\circ(\underset{\sim}{x}) = \frac{\hat{\theta}(\underset{\sim}{x}) - \theta}{Var_\theta(\hat{\theta}(\underset{\sim}{x}))}.$$

STAT4028: Probability and Mathematical Statistics

14. Exponential Families

14.9.4 Exponential Family

The exponential family is a widely unified family of distributions. There are many famous distributions which are examples of the exponential family when we restrict the space to the real line $\mathbb{R}$. We can see that there are properties underlying all these different distributions when we view it from an exponential family point of view.

The 2 important elements required to define an exponential family is a $\sigma-$finite measure and sufficient statistic. We first recall the definition of the exponential family.

Definition 78: Exponential family density

We define the probability density function of the exponential family

$$q_{\underset{\sim}{\eta}}(\underset{\sim}{x}) = e^{\underset{\sim}{\theta}(\underset{\sim}{\eta})^T \underset{\sim}{s}(\underset{\sim}{x}) - c(\underset{\sim}{\eta}) - H(\underset{\sim}{x})}$$

for $\underset{\sim}{x} \in \mathbb{R}^n$. In particular, we have that $\underset{\sim}{\eta} \in \mathcal{H} \subseteq \mathbb{R}^d$, $\underset{\sim}{\theta}(\underset{\sim}{\eta}) : \mathbb{R}^d \to \mathbb{R}^d$, $H : \mathbb{R}^n \to \mathbb{R}$, $c : \mathbb{R}^d \to \mathbb{R}$. Finally, $\underset{\sim}{s}(.) : \mathbb{R}^n \to \mathbb{R}^d$ is the sufficient statistic.

However, it may be difficult to write things out in terms of $\underset{\sim}{\theta}(\underset{\sim}{\eta})$. Therefore, if $\underset{\sim}{\theta}(\underset{\sim}{\eta})$ is a homeomorphism, we can then just suppress the $\underset{\sim}{\eta}$ and work with only $\underset{\sim}{\theta}$ instead. We can then just convert things back into $\underset{\sim}{\theta}(\underset{\sim}{\eta})$ when we are done. This will be known as the natural parameterisation.

Definition 79: Natural Parameter Space

Suppose that $\nu(.)$ is a $\sigma-$finite measure on $\mathbb{R}^n$ and $s : \mathbb{R}^n \to \mathbb{R}^d$ is a d-dimensional sufficient statistic. We define the **natural parameter space** as

$$\mathcal{N} = \{\underset{\sim}{\theta} \in \mathbb{R}^d : \int_{\mathbb{R}^n} e^{\underset{\sim}{\theta}^T \underset{\sim}{s}(\underset{\sim}{x})} d\nu(\underset{\sim}{x}) < \infty\}.$$

We see that the natural parameter space depends on the measure ν and sufficient statistic $s(.)$.

Definition 14.220 *(MGF and CGF of the sufficient statistic). Suppose that $\nu(.)$ is a $\sigma-$finite measure on $\mathbb{R}^n$ and $s : \mathbb{R}^n \to \mathbb{R}^d$ is a d-dimensional sufficient statistic. Let $\mathcal{N}$ be the natural parameter space. For $\theta \in \mathcal{N}$, we define the moment generating function of the sufficient statistic*

$$m(\underset{\sim}{\theta}) = \int_{\mathbb{R}^n} e^{\underset{\sim}{\theta}^T \underset{\sim}{s}(\underset{\sim}{x})} d\nu(\underset{\sim}{x})$$

where $m(\underset{\sim}{\theta}) < \infty$ for all $\underset{\sim}{\theta} \in \mathcal{N}$.

Furthermore, we define the cumulant generating function of the sufficient statistic as

$$K(\theta) = log\ m(\theta).$$

Now, we can express the exponential family in the canonical (natural) parameterisation.

Definition 80: Exponential family density with natural parameterisation

For all $\underset{\sim}{\theta} \in \mathcal{N}$, we define

$$p_{\underset{\sim}{\theta}}(\underset{\sim}{x}) = e^{\underset{\sim}{\theta}^T \underset{\sim}{s}(\underset{\sim}{x}) - K(\underset{\sim}{\theta})}$$

where $p_{\underset{\sim}{\theta}}$ is a probability density and $K(\underset{\sim}{\theta})$ is the cumulant generating function of the sufficient statistic. Then, any Borel set is assigned mass

$$\mathbb{P}_{\underset{\sim}{\theta}}(B) = \int_B e^{\underset{\sim}{\theta}^T \underset{\sim}{s}(\underset{\sim}{x}) - K(\underset{\sim}{\theta})} d\nu(\underset{\sim}{x})$$

where $\mathbb{P}_{\underset{\sim}{\theta}}$ is the probability measure. Here, $s : \mathbb{R}^n \to \mathbb{R}^d$ is the sufficient statistic, $\underset{\sim}{x} \in \mathbb{R}^n$, $\mu \in \mathcal{N} \subseteq \mathbb{R}^d$. The family $\{p_{\underset{\sim}{\theta}} : \underset{\sim}{\theta} \in \mathcal{N}\}$ is the family of exponential family densities with a natural parameterisation.

Theorem 14.221 *Suppose the sequence* $\{a_n : n = 0, 1, 2, ...\}$ *is such that for some* $r > 0$

$$\lim_{k\to\infty} \sum_{n=0}^{k} a_n x^n = \ell(x)$$

exists and is finite for all $|x| < r$. *Then* r *is known as the* ***radius of convergence****. Then, we can differentiate it term by term any number of times*

$$\frac{d^k}{dx^k} \ell(x)|_{x=0} = a_k \cdot k!$$

for k = 1,2,3...

For certain exponential families, we can differentiate with respect to θ any number of times inside the integral.

Theorem 56: Power series theorem for the MGF of the sufficient statistic

Let $\mathcal{N} = \{\theta : \int e^{\underset{\sim}{\theta}^T \underset{\sim}{s}(\underset{\sim}{x})} d\nu(\underset{\sim}{x} < \infty\}$ be the natural parameter space. Let $\mathcal{N}$ be an open interval and $\theta_0 \in \mathcal{N}$ be an interior point. Then, with the moment generating function of the sufficient statistic $m(\theta) = \int_{\mathcal{X}} e^{\underset{\sim}{\theta}^T \underset{\sim}{s}(\underset{\sim}{x})} d\nu(\underset{\sim}{x})$, we can differentiate $m(\theta)$ any number of times inside the integral sign for all $\theta \in \mathcal{N}$

$$\frac{d^k m(\theta)}{d\theta^k}|_{\theta=\theta_0} = m_k(\theta_0) = \int_{\mathbb{R}} [\underset{\sim}{s}(\underset{\sim}{x})]^k e^{\underset{\sim}{\theta}^T \underset{\sim}{s}(\underset{\sim}{x})} d\nu(\underset{\sim}{x}).$$

Remark 14.222 *If* $\mathcal{N}$ *is an open interval, then every point* $\theta \in \mathcal{N}$ *is an interior point.*

This is a significant result because this means that assumption 2 and 3 for CRLB holds!

Corollary 14.223 *Assumptions 2 and 3 for the CRLB holds for exponential families for all* $\theta \in \mathcal{N}$.

14.9.5 The Mean-Value Parameter

In this last section, we show that we can further parameterise the exponential family when its in its natural parameterisation. The reason we are interested is that for exponential families, the CRLB is only obtained when we have a mean value parameterisation. That is, the MLE is a function of the mean of the sufficient statistic and the mean of the sufficient statistic is an unbiased estimator of the mean value parameter. Therefore, we have that we can now get an unbiased estimator of the parameter and the relationship of the score function and the estimator will be such that the CRLB will be obtained.

Reall that the cumulant generating function is defined as $K(\theta) = log\ M(\theta)$.

Proposition 69: First and second derivative of the cumulant generating function

Let $s(\underset{\sim}{x})$ be the sufficient statistic and let $K(\theta)$ be the cumulant generating function of the sufficient statistic. Then

$$\frac{\partial}{\partial\theta_i}K(\theta) = E_\theta(s_i(\underset{\sim}{x})).$$

Furthermore, we have that

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}K(\theta) = Cov_\theta(s_i(\underset{\sim}{x}), s_j(\underset{\sim}{x})).$$

Combining these two facts, we have that

$$\begin{cases} \mathbb{E}_\theta(\underset{\sim}{s}(\underset{\sim}{x})) = K^{'}(\theta) \\ \\ Var_\theta(\underset{\sim}{s}(\underset{\sim}{x})) = K^{''}(\theta) \end{cases}$$

With this, we can now define our new parameterisation.

Definition 81: Mean value parameter

We define a smooth change of parameter from θ to μ as

$$\theta \to \mu(\theta) = \mathbb{E}_\theta(\underset{\sim}{s}(\underset{\sim}{X})).$$

That is, the expected value of the sufficient statistic is the mean-value parameter.

Theorem 57: Existence of mean-value map inverse

The inverse map from the mean-value parameter μ to θ is well-defined if the covariance matrix of the sufficient statistic $s(\underset{\sim}{X})$ is non-singular.

Remark 14.224 *The Jacobian of the mean-value change of parameters is the covariance matrix of the sufficient statistic.*

Definition 14.225 *(Mean-Value Parameter Space). The mean-value parameter space $\mathcal{M}$ is the image under*

the mean value transformation of the natural parameter space. That is

$$\mathcal{M} = \{\mu \in \mathbb{R} : \int_{\mathcal{X}} e^{\mu s(\underset{\sim}{x})} d\nu(\underset{\sim}{X}) < \infty\}.$$

Definition 14.226 *(Convex Hull). The convex hull of a set A is the smallest convex set that contains the set A.*

Remark 14.227 *The mean-value parameter space $\mathcal{M}$ is the convex hull of the support of the original base measure $\nu(.)$.*

Proposition 14.228 *The support $\mathcal{X}$ of the natural parameter space $\mathcal{N}$ agrees with the support of the mean-value parameter space $\mathcal{M}$.*

We recall a result from vector calculus which we will need to write out the mean-value parameterisation of the exponential family.

Definition 14.229 *Let $\alpha : I_1 \to \mathbb{R}$ and $\beta : I_2 \to \mathbb{R}$ be parameterised smooth curves. Then, β is a reparameterisation of α if there exists a smooth function ψ such that $\psi : I_2 \to I_1$ with a smooth inverse*

$$\beta : \alpha \circ \psi.$$

Definition 82: Density of mean value parameterisation

The family of densities for the exponential family under the mean value parameterisation $\{q_\mu(.) : \mu \in \mathcal{M}\}$ is given by

$$q_\mu(\underset{\sim}{x}) = p_\theta(\mu)(\underset{\sim}{x}) = e^{\theta(\mu)s(\underset{\sim}{x}) - K(\theta(\mu))}.$$

Claim 14.230 *The support of the mean-value parameter space $\mathcal{M}$ is the same support as the natural parameter space $\mathcal{N}$.*

Proposition 14.231 *The mean value parameterisation satisfies the 3 regularity conditions.*

Proposition 70: The sufficient statistic is MVUE

Under the mean-value parameterisation, the sufficient statistic $\underset{\sim}{s}(\underset{\sim}{x})$ is a minimum variance unbiased estimator of the mean-value parameter $\mu = \theta(\mu)$.

Proof: First, the sufficient statistic $s(\underset{\sim}{x})$ is an unbiased estimator for the mean value parameter μ. The score function has the form

$$\ell^\circ_\mu = \frac{s(\underset{\sim}{x}) - \mu}{Var_\mu[s(\underset{\sim}{x})]}$$

which is the form we need for the unbiased estimator $s(\underset{\sim}{x})$ to attain the Cramér-Rao lower bound. ■

STAT4028: Probability and Mathematical Statistics

15. Local Asymptotic Normality

15.10 Local Asymptotic Normality

In this lecture, we introduce the notion of local asymptotic normality. That is, the distribution of the log likelihood ratio is asymptotically normal in a local sense. Furthermore, the LAN property guarantees us the existence of a score vector and information matrix at our point of analysis. Finally, this will allow us to analyse other sequences that are nearby.

15.10.1 Taylor's Theorem

First, we recall somethings from calculus. We can approximate differentiable functions using a Taylor polynomial and from there, derive the Taylor series.

Definition 15.232 *(Taylor Polynomial). Let f be a function that is differentiable n times on an open interval containing $x = c$. Then, the* ***Taylor polynomial of f at c*** *is*

$$P_n(x) = f(c) + \frac{f^{(1)}(c)}{1!}(x-c) + \frac{f^{(2)}(c)}{2!}(x-c)^2 + ... + \frac{f^{(n)}(c)}{n!}(x-c)^n.$$

Theorem 58: Taylor's Theorem

Suppose that f is n + 1 tiems differentiable on an interval containing x = c. Let $P_n(x) = f(c) + \frac{f^{(1)}(c)}{1!}(x-c) + ... + \frac{f^{(n)}(c)}{n!}(x-c)^n$ be the n^{th} order Taylor polynomial of f at x = c. Then,

$$f(x) = P_n(x) + E_n(x)$$

where $E_n(x)$ is the error term of $P_n(x)$ from f(x) and for ϵ between c and x, the **Lagrange remainder** form of the error E_n is given by the formula

$$E_n(x) = \frac{f^{(n+1)}(\epsilon)}{(n+1)!}(x-c)^{n+1}.$$

Corollary 15.233 *Suppose that derivatives of all orders exists on an interval containing x and c. Then, Taylor's theorem holds and if for any x we have that $\lim_{n\to\infty} E_n(x) = 0$ then*

$$\lim_{n\to\infty} P_n(x) = f(x).$$

15.10.2 Weak Convergence

So far, we have seen 3 modes of convergence (a.s convergence, convergence in probability and $\mathcal{L}^p-$convergence). We will now introduce another one called **weak convergence** or **convergence in distribution**. This is different to the other 3 forms of convergence as the convergence only depends on their distribution.

Definition 83: Weak Convergence

Let $(\Omega, \mathcal{B})$ be a measurable space where (Ω, d) is a metric space and $\mathcal{B}$ is the Borel $\sigma - algebra$ on Ω. Let $\{\mu_n\}_{n\in\mathbb{N}}$ be a sequence of probability measures on $(\Omega, \mathcal{B})$. We say that μ_n **converges weakly** to a probability measure μ on $(\Omega, \mathcal{B})$ and write $\mu_n \xrightarrow{w} \mu$ if

$$\int f d\mu_n \to \int f d\mu$$

for all $f \in C_b(\Omega)$ where $C_b(\Omega)$ denotes the set of all continuous and bounded functions $f : \Omega \to \mathbb{R}$.

From this, convergence in distribution is simply random variables weakly converging.

Definition 84: Convergence in distribution

A sequence $\{X_n\}_{n\in\mathbb{N}}$ of random variables is said to **converge in distribution** to the random variable X if $\mu_{X_n} \xrightarrow{w} \mu_X$ where μ_X is the distribution measure induced by the random variable X. We then write this as $X_n \xrightarrow{d} X$.

In other words, a characterisation of convergence in distribution is that the expected value of bounded continuous functionals converges. We give other characterisation of convergence in distribution.

Lemma 15.234 *(Portmanteau Lemma). For any random variables X_n and X, the following statements are equivalent.*

1. $X_n \xrightarrow{d} X$
2. $P(X_n \leq x) \to P(X \leq x)$
3. $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ *for all bounded, Lipschitz functions f*

Proposition 15.235 *(Uniqueness of weak convergence). Suppose that $\{\mu_n\}_{n\in\mathbb{N}}$ is a sequence of probability measures on $(\Omega, \mathcal{B})$ such that $\mu_n \xrightarrow{w} \mu$ and $\mu_n \xrightarrow{w} \mu'$. Then $\mu = \mu'$.*

15.10.3 Asymptotic Statistics

Why do we care about asymptotic statistics? In most statistical models in a frequentist setting, exact inference is not feasible and hence asymptotic results can provide convenient approximations. Furthermore, we can gain valuable theoretical insights into complex models and complex problems. We can also define new measures of optimality in an asymptotic setting.

Definition 15.236 *(Asymptotic distribution). Let us define the sequence of non-random constants $\{a_n\}, \{b_n\}$. Let $\{\hat{\theta}_n\}$ be a sequence of estimators. Then, suppose that*

$$b_n(\hat{\theta}_n - a_n) \xrightarrow{d} G$$

where G is a distribution. Then, $\{\hat{\theta}_n\}$ is said to have the asymptotic distribution G.

We describe some other useful results we will need. The continuous-mapping theorem states that if a sequence of random vectors X_n converges to X and g is continuous, then $g(X_n)$ converges to g. This holds for multiple forms of convergence.

Theorem 15.237 *(Continuous Mapping Theorem). Let g be a continuous function at every point of a set C such that* $\mathbb{P}(X \in C) = 1$. *Then*

1. *If* $X_n \xrightarrow{d} X$ *then* $g(X_n) \xrightarrow{d} g(X)$
2. *If* $X_n \xrightarrow{p} X$ *then* $g(X_n) \xrightarrow{p} g(X)$
3. *If* $X_n \xrightarrow{a.s.} X$ *then* $g(X_n) \xrightarrow{a.s.} g(X)$

Definition 15.238 *(Bounded in probability). Let* $\{X_n\}$ *be a sequence of random vectors. Then, the sequence is bounded in probability if for every* $\epsilon > 0$, *there exists a constant M such that*

$$\sup_n \mathbb{P}(|X_n| > M) < \epsilon$$

Lemma 15.239 *Every weakly converging sequence* X_n *is bounded in probability.*

In the i.i.d setting

$$\underset{\sim}{S_n} = n^{-1/2} \sum_{i=1}^{n} \underset{\sim}{\ell_\theta^\circ}(\underset{\sim}{X_i})$$

Then, to compute the information

$$\mathbb{E}[\underset{\sim}{S_n}\underset{\sim}{S_n}^T] = \frac{1}{n}\Big[\sum_{i=1}^{n} \underset{\sim}{\ell_\theta^\circ}(\underset{\sim}{X_i}) \cdot \underset{\sim}{\ell_\theta^\circ}(\underset{\sim}{X_i})^T\Big]$$

15.10.4 Local Asymptotic Normality

We are now interested in **sequences of parametric models** for the data $\underset{\sim}{x}$. That is, we will define a sequence of probability models $\{p_{n\theta}(.)\}$ whereby we will have a different sample space $\mathcal{X}_n$ and measure $\nu_n(.)$ for each n. The LAN property as we will soon see captures a particular aspect of regularity in regular parametric models.

Definition 85: Log Likelihood Ratio

Suppose that for each $n = 1, 2, ...$ and each $\underset{\sim}{\theta} \in \Theta \subseteq \mathbb{R}^d$, we have a probability distribution $\mathbb{P}_{n\underset{\sim}{\theta}}$ on a sample space $\mathcal{X}_n$ whose density function with respect to the measure $\nu_n(.)$ is $p_{n\underset{\sim}{\theta}}(.)$. Fix an interior point $\underset{\sim}{\theta_0} \in \Theta$. Define the sequence of supports $A_n = A_n(\underset{\sim}{\theta_0}) = \{\underset{\sim}{x} : \mathcal{X}_n : p_{n\underset{\sim}{\theta}}(\underset{\sim}{x}) > 0\}$. Then, for any other $\underset{\sim}{\theta_1} \in \Theta$, we define the **log likelihood ratio** as

$$L_n(\underset{\sim}{x}; \underset{\sim}{\theta_1} | \underset{\sim}{\theta_0}) = \begin{cases} \frac{log\ p_{n\underset{\sim}{\theta}_1}(\underset{\sim}{x})}{log\ p_{n\underset{\sim}{\theta_0}}(\underset{\sim}{x})} & \text{for } \underset{\sim}{x} \in A_n \\ \\ 0 \quad \text{for } \underset{\sim}{x} \in A_n^c \end{cases}$$

where we define L_n to be zero outside of the support A_n.

Remark 15.240 *We can give a different interpretation of the log likelihood ratio. Suppose that $\mathbb{P}_{n\underset{\sim}{\theta_1}}$ is absolutely continuous with respect to $\mathbb{P}_{n\underset{\sim}{\theta_0}}$. Then, the log likelihood ratio is the logarithm of the Radon-Nikodym derivative of these two probability measures.*

We now want to do "local Pitman analysis" of the log likelihood ratio L_n. We want to restrict our analysis to a local area where $\underset{\sim}{\theta_n} = \underset{\sim}{\theta_0} + n^{-\frac{1}{2}}\underset{\sim}{h}$. We can think of this as we are taking a step $\underset{\sim}{h}$ away from the original point $\underset{\sim}{\theta_0}$. Then, we can define the LAN property as a local property around $\underset{\sim}{\theta_0}$.

Definition 86: Local Asymptotic Normality

Let $L_n(\underset{\sim}{x}; \underset{\sim}{\theta_1} | \underset{\sim}{\theta_0})$ be the log likelihood of $\underset{\sim}{\theta_1}$ against the true value $\underset{\sim}{\theta_0}$. We say that the **local asymptotic normality** property holds at $\underset{\sim}{\theta_0}$ if there exists a symmetric positive definite matrix $\underset{\sim}{J} = \underset{\sim}{J}(\underset{\sim}{\theta_0})$ and a random vector $\underset{\sim}{S_n} = \underset{\sim}{S_n}(\underset{\sim}{X_n}; \underset{\sim}{\theta_0})$ such that

$$\underset{\sim}{S_n}(\underset{\sim}{X_n}; \underset{\sim}{\theta_0}) \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{J})$$

where $\underset{\sim}{0}$ is the zero mean vector and covariance matrix $\underset{\sim}{J}$ such that when $\underset{\sim}{X_n} \sim \mathbb{P}_{n\underset{\sim}{\theta_0}}$, for any d-dimensional vector $\underset{\sim}{h} \in \mathbb{R}^d$, the log likelihood ratio satisfies

$$L_n(\underset{\sim}{x}; \underset{\sim}{\theta_0} + n^{-\frac{1}{2}}\underset{\sim}{h} | \underset{\sim}{\theta_0}) = \underset{\sim}{h}^T \underset{\sim}{S_n} - \frac{1}{2}\underset{\sim}{h}^T \underset{\sim\sim}{J}\underset{\sim}{h} + R_n$$

where the remainder $R_n = R_n(\underset{\sim}{\theta_0}; \underset{\sim}{h}) \xrightarrow{p} 0$.

Definition 15.241 *(Vector of scores and information matrix). The random d-dimensional vector $\underset{\sim}{S_n}$ is the vector of scores and $\underset{\sim}{J}$ is the information matrix.*

Remark 15.242 *The information matrix $\underset{\sim}{J}$ is usually the covariance matrix of the first derivatives of the*

log likelihood. Therefore, the information matrix can be interpreted to be the asymptotic variance of the scores.

Remark 15.243 *We will later see conditions which imply the LAN property but it is worth noting that if the conditions which imply LAN holds, then we can show that* $\underset{\sim}{S}_n(\underset{\sim}{X}_n;\underset{\sim}{\theta}_0) \xrightarrow{d} \mathcal{N}(\underset{\sim}{0},\underset{\sim}{J})$ *using the central limit theorem.*

There are a few things we can remark. First, the score function converges to the multivariate normal distribution through the central limit theorem. This leads to the following result.

Proposition 71: Log likelihood Ratio is asymptotically normal under LAN

The log likelihood ratio is asymptotically normal at $\underset{\sim}{\theta}_0$ if the LAN property holds at $\underset{\sim}{\theta}_0$. That is, if $\underset{\sim}{X}_n \sim \mathbb{P}_{n\underset{\sim}{\theta}_0}$, then

$$L_n \xrightarrow{d} \mathcal{N}\Big(-\frac{\underset{\sim}{h}^T \underset{\sim}{J} \underset{\sim}{h}}{2}, \underset{\sim}{h}^T \underset{\sim}{J} \underset{\sim}{h}\Big)$$

Hence, essentially, the LAN property states that the log likelihood ratio is a linear combination of the scores vector.

Proposition 72: Density of a nearby sequence

Under the support, the density of a nearby sequence is given by the density under $\underset{\sim}{\theta}_0$ times the log likelihood ratio.

$$\mathbb{P}_{n(\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h})} = e^{L_n}\mathbb{P}_{n\underset{\sim}{\theta}_0}$$

Proof: First, recall that the log likelihood ratio is given by

$$L_n = log\Big\{\frac{\mathbb{P}_{n(\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h})}}{\mathbb{P}_{n\underset{\sim}{\theta}_0}}\Big\}$$

We can arrange this to get our desired result. ■

15.10.5 Le Cam's First Lemma

We are now interested in the consequences of the local asymptotic normality property. We note that the LAN conditions only happens when the true distribution is $\mathbb{P}_{\underset{\sim}{\theta}_0}$ under the sequence $\{\mathbb{P}_n(\underset{\sim}{\theta}_0)\}$. However, Le Cam showed that it also implies important properties for sequences that are nearby. In particular, we want to look at the nearby sequence $\{\mathbb{P}_n(\underset{\sim}{\theta}_0 + n^{-\frac{1}{2}}\underset{\sim}{h})\}$.

For a fixed n, the $\mathbb{P}_{n\underset{\sim}{\theta}}$ need not all have the same support unlike the exponential family.

Definition 87: Contiguity

Denote $(\Omega_n, \mathcal{F}_n)$ be a sequence of measurable spaces, each equipped with two measures P_n and Q_n. We say that Q_n is contiguous with respect to P_n if for every sequence A_n of measurable sets, $P_n(A_n) \to 0$ implies that $Q_n(A_n) \to 0$.

Remark 15.244 *We can say that two sequences of probability measures are contiguous if asymptotically, they share the same support. The notion of contiguity extends the concept of absolute continuity to the sequences of measures.*

Le Cam's first lemma states that sequences that are "nearby" are contiguous with one another. That is, the supports agree in an asymptotic sense.

Theorem 59: Le Cam's First Lemma

Suppose that for each $n = 1, 2, ...$ and each $\underset{\sim}{\theta} \in \Theta \subseteq \mathbb{R}^d$, we have a probability distribution $\mathbb{P}_{n\underset{\sim}{\theta}}$ on a sample space $\mathcal{X}_n$ whose density function with respect to the measure $\nu_n(.)$ is $p_{n\underset{\sim}{\theta}}(.)$. Fix an interior point $\theta_0 \in \Theta$. Define the sequence of supports $A_n = A_n(\underset{\sim}{\theta}_0) = \{\underset{\sim}{x} : \mathcal{X}_n : p_{n\underset{\sim}{\theta}}(\underset{\sim}{x}) > 0\}$. Suppose that the LAN property holds at $\underset{\sim}{\theta}_0$. Then for a nearby sequence, the probability of a nearby alternative

$$\mathbb{P}_{n(\underset{\sim}{\theta}_0 + n^{-\frac{1}{2}}\underset{\sim}{h})}(A_n) \to 1.$$

Proof: The goal is to show that for any $\epsilon > 0$, we have that $\mathbb{P}_{n(\underset{\sim}{\theta}_0 + n^{-\frac{1}{2}}\underset{\sim}{h})}(A_n) \geq 1 - \epsilon$ for the sequence of supports A_n. First, the LAN property implies that for $\underset{\sim}{X}_n \sim \mathbb{P}_{n\underset{\sim}{\theta}_0}$, we have that the log likelihood ratio

$$L_n \xrightarrow{d} \mathcal{N}\Big(\frac{-\delta^2}{2}, \delta^2\Big)$$

where $\delta^2 = \underset{\sim}{h}^T J \underset{\sim}{h}$.

We let F_n be the CDF of the log likelihood ratio L_n under the true distribution $\mathbb{P}_{n\underset{\sim}{\theta}_0}$. Let F, f be the CDF and PDF of the limiting distribution of the log likelihood ratio $\mathcal{N}(\frac{-\delta^2}{2}, \delta^2)$. Additionally, let G, g be the CDF and PDf of the limiting distribution of $\mathcal{N}(\frac{+\delta^2}{2}, \delta^2)$.

Now, we can use the fact that for all v,

$$g(v) = e^v f(v).$$

Fix $\epsilon > 0$ and choose M_ϵ such that the upper quantile $G(M_\epsilon) > 1 - \epsilon$. We are interested in the probability under the nearby parameter of the support

$$\mathbb{P}_{n(\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h})}(A_n) = \int_{A_n} d\mathbb{P}_{n(\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h})}$$

$$= \int_{A_n} e^{L_n} d\mathbb{P}_{n\underset{\sim}{\theta}_0} = \int_{A_n} e^{L_n} p_{n\underset{\sim}{\theta}_0}(\underset{\sim}{x}) d\nu_n(\underset{\sim}{x}) = \int_{\mathcal{X}_n} e^{L_n} d\mathbb{P}_{n\underset{\sim}{\theta}_0} = \int_{-\infty}^{\infty} e^v dF_n(v)$$

A characterisation of convergence in distribution is convergence of bounded continuous function

$$\geq \int_{-\infty}^{\infty} e^v 1\{v \leq M_\epsilon\} dF_n(v) \to \int_{-\infty}^{\infty} e^v 1\{v \leq M_\epsilon\} dF(v)$$

since $F_n \xrightarrow{d} F$ where $e^v 1\{v \leq M_\epsilon\}$ is a bounded continuous function. Hence, we have that

$$= \int_{-\infty}^{\infty} 1\{v \leq M_\epsilon\} dG(v) = G(M_\epsilon) > 1 - \epsilon.$$

As ϵ was arbitrary, this proves the theorem as

$$\mathbb{P}_{n(\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h})}(A_n) \geq 1 - \epsilon$$

for arbitrary ϵ. ■

Remark 15.245 *Le Cam's first lemma is significant as the support under the true parameter and a nearby sequence may be different for a fixed n, but in the limit, they are the same.*

15.10.6 LAN Property in Exponential Families

We now give an example for the exponential family. Let $\{\mathbb{P}_{\underset{\sim}{\theta}} : \underset{\sim}{\theta} \in \mathcal{N}\}$ be an exponential distribution on $\mathbb{R}^k$ where $\mathcal{N} = \{\int e^{\underset{\sim}{\theta} s(x)} < \infty\}$ is the natural parameter space. Therefore, the density function of $\mathbb{P}_{\underset{\sim}{\theta}}$ with respect to the $\sigma-$finite measure $\nu(.)$ is of the form

$$p_{\underset{\sim}{\theta}}(\underset{\sim}{x}) = e^{\underset{\sim}{\theta}^T \underset{\sim}{s}(\underset{\sim}{x}) - K(\underset{\sim}{\theta})}.$$

Suppose that $\underset{\sim}{X}_1, ..., \underset{\sim}{X}_n$ are i.i.d with common distribution $\mathbb{P}_{\underset{\sim}{\theta}_0}$ for some true unknown $\underset{\sim}{\theta}_0 \in \mathcal{N}$. Then, fix $\underset{\sim}{h} \in \mathbb{R}^d$ where $d << k$ and consider the log likelihood ratio

$$L(h) = \log \prod_{i=1}^{n} \Big\{ \frac{p_{\underset{\sim}{\theta}_0 + n^{-\frac{1}{2}}\underset{\sim}{h}}(\underset{\sim}{X}_i)}{p_{\underset{\sim}{\theta}_0}(\underset{\sim}{X}_i)} \Big\} = \underset{\sim}{h}^T n^{-1/2} \sum_{i=1}^{n} \underset{\sim}{s}(\underset{\sim}{x}_i) - n[K(\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h}) - K(\underset{\sim}{\theta}_0)] \qquad (*)$$

We can interpret K to be the cumulant generating function if $\nu(.)$ is a probability measure. Furthermore, it is infinitely smooth from the theorem that allows us to infinitely differentiate the moment generating function of the sufficient statistic. We can then therefore apply a 2nd order Taylor series expansion on K.

$$K(\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h}) = K(\underset{\sim}{\theta}_0) + n^{-1/2} h^T \underset{\sim}{K}^{'}(\underset{\sim}{\theta}_0) + n^{-1/2} \underset{\sim}{h}^T \underset{\sim}{K}^{''}(\underset{\sim}{\theta}^*_n) \underset{\sim}{h}^T$$

for some $\underset{\sim}{\theta}^*_n$ between $\underset{\sim}{\theta}_0$ and $\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h}$. Furthermore, $K^{'}, K^{''}$ are the vector of 1st order and matrix of 2nd order partial derivatives.

By element-wise continuity of $\underset{\sim}{K}^{''}(.)$, we have that

$$\underset{\sim}{h}^T \underset{\sim}{K}^{''}(\underset{\sim}{\theta}^*_n) \underset{\sim}{h}^T \to \underset{\sim}{h}^T \underset{\sim}{K}^{''}(\underset{\sim}{\theta}_0) \underset{\sim}{h}^T$$

as $\underset{\sim}{\theta}^*_n \to \underset{\sim}{\theta}_0$ in a shrinking neighbourhood.

We can now take this second order Taylor series expansion of the cumulant generating function and plug it back into our log likelihood ratio in (*).

Proposition 73: LAN Property of exponential family

Suppose all the conditions we elaborated in this section holds. Therefore, for an exponential family, we have a log likelihood ratio

$$L_n(\underset{\sim}{h}) = \underset{\sim}{h}^T \underset{\sim}{S}_n - \frac{1}{2}\underset{\sim}{h}^T \underset{\sim\sim}{J}\underset{\sim}{h} + R_n(\underset{\sim}{h})$$

where the score function vector

$$\underset{\sim}{S}_n = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \left[\underset{\sim}{s}(x_i) - \underset{\sim}{K}^{'}(\underset{\sim}{\theta}_0\right] \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{J}).$$

Furthermore, $\underset{\sim}{J} = \underset{\sim}{K}^{''}(\underset{\sim}{\theta}_0)$ is the 2nd order partial derivatives matrix, which was the covariance matrix of the sufficient statistic. Finally, the remainder term $R_n(\underset{\sim}{h}) \xrightarrow{p} 0$.

Remark 15.246 *Recall that $\underset{\sim}{K}^{'}(\underset{\sim}{\theta}_0)$ is the mean-value of the sufficient statistic. Therefore, the difference between the score and $\underset{\sim}{K}^{'}(\underset{\sim}{\theta}_0)$ has mean zero. Then by the central limit theorem, converges to a multivariate normal.*

We will now show that this rescaled, localised log likelihood ratio is similar to that for a multivariate Gaussian normal based on a single observation.

Definition 15.247 *(Multivariate Normal with shifted location of one observation). Suppose the vector $\underset{\sim}{S} \sim \mathcal{N}(\underset{\sim\sim}{J}\underset{\sim}{h}, \underset{\sim}{J})$ for a known $\underset{\sim}{J}$ but unknown $\underset{\sim}{h}$. This is a shifted multivariate normal distribution. Then, the density for one observation of $\underset{\sim}{S}$ is*

$$f_{\underset{\sim}{h}}(s) = \frac{1}{(2\pi)^{d/2}} \frac{1}{(det\ J)^{1/2}} e^{-\frac{1}{2}\left\{(s-\underset{\sim\sim}{J}\underset{\sim}{h})^T \underset{\sim}{J}^{-1}(s-\underset{\sim\sim}{J}\underset{\sim}{h})\right\}}$$

Proposition 74: Log likelihood ratio for multivariate normal

The log likelihood ratio for the multivariate normal distribution between values $\underset{\sim}{h}$ and $\underset{\sim}{0}$ is given by

$$log\{\frac{f_{\underset{\sim}{h}}(\underset{\sim}{s})}{f_{\underset{\sim}{0}}(\underset{\sim}{s})}\} = \underset{\sim}{h}^T \underset{\sim}{s} - \frac{1}{2}\underset{\sim}{h}^T \underset{\sim\sim}{J}\underset{\sim}{h}.$$

Hence, the log likelihood ratio of the multivariate normal with shifted location of one observation is the same as the log likelihood of the exponential family near the true value!

Hence, we have a big takeaway.

Theorem 60: LAN Property always hold for exponential families

Let $\{\mathbb{P}_{\underset{\sim}{\theta}} : \underset{\sim}{\theta} \in \mathcal{N}\}$ be a family of exponential distributions where $\mathcal{N}$ is the natural parameter space. Then, the LAN property holds for all $\underset{\sim}{\theta}$ that is an interior point of $\mathcal{N}$.

Corollary 15.248 *The LAN property still holds after exponential families after a smooth change of parameters*

$$\eta \to \theta(\eta)$$

where the smooth change is three times differentiable.

STAT4028: Probability and Mathematical Statistics

16. Le Cam's Third Lemma

16.10.7 Le Cam's Third Lemma

Recall from last lecture that Le Cam's first lemma tells us that supports of nearby alternatives are equal asymptotically. Furthermore, LAN tells us what happens under the true value $\underset{\sim}{\theta}_0$. In this lecture, we will see that Le Cam's third lemma tells us the joint distribution of statistics and nearbly alternatives are asymptotically normal. Furthermore, we will also introduce the concept of L_2-differentability which is a condition that will imply the LAN property.

Le Cam's third lemma tells us that the joint distribution of a nearby alternative $\mathbb{P}_{n(\underset{\sim}{\theta}_0+n^{-\frac{1}{2}}\underset{\sim}{h})}$ of both the log likelihood ratio L_n and statistic Y_n is asymptotically normal.

Theorem 61: Le Cam's Third Lemma

Suppose that the LAN property holds at $\underset{\sim}{\theta}_0$. Furthermore, assume that Y_n is a statistic such that under the value at which LAN holds, $\mathbb{P}_{n\underset{\sim}{\theta}_0}$, we have the following limiting behaviour

$$\begin{pmatrix} Y_n \\ L_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} 0 \\ -\frac{\delta^2}{2} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \beta \\ \beta & \delta^2 \end{pmatrix} \Big) \tag{*}$$

where Y_n has limiting mean 0 and variance σ^2, $\delta^2 = \underset{\sim}{h}^T J \underset{\sim}{h}$, and β is the covariance between Y_n and L_n. Then, under a nearby alternative sequence $\mathbb{P}_{n(\underset{\sim}{\theta}_0+n^{-1/2}\underset{\sim}{h})}$

$$\begin{pmatrix} Y_n \\ L_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} \beta \\ +\frac{\delta^2}{2} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \beta \\ \beta & \delta^2 \end{pmatrix} \Big) \tag{**}$$

Proof: Let F_n denote the joint CDF of $(Y_n, L_n)^T$ under $\mathbb{P}_{n\underset{\sim}{\theta}_0}$. Let F and f denote the CDF and PDF of the bivariate normal distribution (*). Let G and g denote the CDF and PDF of the bivariate normal distribution (**). Let G_n denote the CDF of $(Y_n, L_n)^T$ under the nearby sequence $\mathbb{P}_{n(\underset{\sim}{\theta}_0+n^{-1/2}\underset{\sim}{h})}$.

Then, the joint CDF of $G_n(y, l)$ is

$$G_n(y,l) = \mathbb{P}_{n(\underset{\sim}{\theta}_0+n^{-1/2}\underset{\sim}{h})}\Big(\{Y_n \le y, L_n \le l\} \cap A_n\Big) + \mathbb{P}_{n(\underset{\sim}{\theta}_0+n^{-1/2}\underset{\sim}{h})}\Big(\{Y_n \le y, L_n \le l\} \cap A_n^c\Big)$$

where $A_n = \{\underset{\sim}{x} : \mathbb{P}_{n\underset{\sim}{\theta}_0}(\underset{\sim}{x}) > 0\}$ is the support of $\mathbb{P}_{n\underset{\sim}{\theta}_0}$. However, by Le Cam's first lemma, we have that

$$\mathbb{P}_{n(\underset{\sim}{\theta}_0+n^{-1/2}\underset{\sim}{h})}\Big(\{Y_n \le y, L_n \le l\} \cap A_n^c\Big) \xrightarrow{p} 0.$$

Therefore, we restrict our attention to the first term. We now want to show that $G_n \to G$.

$$\mathbb{P}_{n(\underset{\sim}{\theta}_0+n^{-1/2}\underset{\sim}{h})}\Big(\{Y_n \le y, L_n \le l\} \cap A_n\Big) = \int_{A_n} 1\{Y_n \le y, L_n \le l\} d\mathbb{P}_{n(\underset{\sim}{\theta}_0+n^{-1/2}\underset{\sim}{h})}$$

$$= \int_{A_n} e^{L_n} 1\{Y_n \leq y, L_n \leq l\} d\mathbb{P}_{n\underset{\sim}{\theta}_0} = \int_{\mathcal{X}_n} e^{L_n} 1\{Y_n \leq y, L_n \leq l\} d\mathbb{P}_{n\underset{\sim}{\theta}_0}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^v 1\{u \leq y, v \leq l\} dF_n(u,v) \rightarrow \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^v 1\{u \leq y, v \leq l\} dF(u,v)$$

since $F_n \xrightarrow{d} F$.

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1\{u \leq y, v \leq l\} dG(u,v) = G(y,l).$$

Therefore, we have shown that $G_n(y,l) \rightarrow G(y,l)$. ■

More succinctly, if a statistic Y_n is asymptotically jointly normal with the log likelihood ratio, then for a nearby alternative, Y_n is still asymptotically jointly normal with the log likelihood ratio with the limiting mean of Y_n now the covariance under $\underset{\sim}{\theta}_0$. The nearby limiting sequence only has an effect on altering the limiting means.

16.10.8 Conditions implying LAN property

We are now interested in investigating the conditions for which the LAN property will hold. In particular, classical conditions involving the third order partial derivatives so that we can compute third order Taylor expansions of the log likelihood of $\underset{\sim}{\theta}$ are actually too stringent. That is, we can actually relax these classical conditions and the LAN property will still hold. We will soon see that weak conditions involving a first derivative of the square root of the density are all that is needed. We will later see that if a family of densities is $L_2(\nu)$-differentiable at $\underset{\sim}{\theta}_0$ then the LAN property holds at $\underset{\sim}{\theta}_0$.

Definition 16.249 *(L_2-norm and space). Let ν be a measure. We define the $L_2(\nu)$ class as the class of functions*

$$L_2(\nu) = \{g : \int g^2 d\nu < \infty\}.$$

Furthermore, for any $g \in L_2(\nu)$, we define the $L_2(\nu)$-norm of g as

$$||g|| = \sqrt{\int g^2 d\nu}.$$

Proposition 16.250 *The $L_2(\nu)$ space is a Hilbert space.*

Proposition 75: Existence of partial derivatives of log density

Suppose we have a parametric family of distributions $\{\mathbb{P}_{\underset{\sim}{\theta}} : \underset{\sim}{\theta} \in \Theta\}$ for $\Theta \subseteq \mathbb{R}^d$ and that each posseses a density $p_{\underset{\sim}{\theta}}(.)$ with respect to a $\sigma-$finite measure $\nu(.)$. If the density $p_{\underset{\sim}{\theta}}(\underset{\sim}{x}) > 0$ and the vector of partial derivatives

$$p^{\circ}_{\underset{\sim}{\theta}}(\underset{\sim}{x}) = \Big(\frac{\partial}{\partial\theta_1} p_{\underset{\sim}{\theta}}(\underset{\sim}{x}), ..., \frac{\partial}{\partial\theta_d} p_{\underset{\sim}{\theta}}(\underset{\sim}{x})\Big)^T$$

exists, then the vector of partial derivatives of the log-density $\ell_{\underset{\sim}{\theta}}(\underset{\sim}{x}) = log\ p_{\underset{\sim}{\theta}}(\underset{\sim}{x})$ also exist. This is given by

$$\ell^{\circ}_{\underset{\sim}{\theta}}(\underset{\sim}{x}) = \frac{p^{\circ}_{\underset{\sim}{\theta}}(\underset{\sim}{x})}{p_{\underset{\sim}{\theta}}(\underset{\sim}{x})}$$

Corollary 16.251 *Under the same conditions, the square root of the density* $\sqrt{p_{\underset{\sim}{\theta}}(\underset{\sim}{x})}$ *has a vector of partial derivatives*

$$\sqrt{p_{\underset{\sim}{\theta}}}^{\circ} = \frac{1}{2}\frac{p^{\circ}_{\underset{\sim}{\theta}}(\underset{\sim}{x})}{\sqrt{p_{\underset{\sim}{\theta}}(\underset{\sim}{x})}} = \frac{1}{2}\ell^{\circ}_{\underset{\sim}{\theta}}(\underset{\sim}{x})\sqrt{p_{\underset{\sim}{\theta}}(\underset{\sim}{x})}$$

To motivate the definition of L_2-derivative, first recall the definition of the Fréchet derivative.

Definition 16.252 *(Fréchet derivative). Let V and W be normed vector spaces and $U \subset V$ be an open subset of V. A function $f : U \to W$ is called Fréchet differentiable at $x \in U$ if there exists a bounded linear operator $A : V \to W$ such that*

$$\lim_{||h||\to 0} \frac{||f(x+h) - f(x) - Ah||_W}{||h||_V} = 0.$$

Definition 88: L_2-derivative

For any $\underset{\sim}{h} \in \mathbb{R}^d$, for a value $\underset{\sim}{x}$ where $p_{\underset{\sim}{\theta}}(\underset{\sim}{x}) > 0$ and $p^{\circ}_{\underset{\sim}{\theta}}(\underset{\sim}{x})$ exists, the L_2-derivative at $\underset{\sim}{x}$ is

$$\frac{\sqrt{p_{\underset{\sim}{\theta}+\underset{\sim}{h}}(\underset{\sim}{x})} - \sqrt{p_{\underset{\sim}{\theta}}(\underset{\sim}{x})} - \frac{1}{2}\underset{\sim}{h}^T \ell^{\circ}_{\underset{\sim}{\theta}}(\underset{\sim}{x})}{|\underset{\sim}{h}|} \to 0$$

as $||h|\to 0$ where $|\underset{\sim}{h}|\sqrt{h_1^2 + ... + h_d^2}$ is the Euclidean norm.

Proposition 16.253 *(Square root of densities is in L_2). The square of densities $\sqrt{p_{\underset{\sim}{\theta}}}$ is square integrable with respect to $\nu(.)$ and hence $\sqrt{p_{\underset{\sim}{\theta}}} \in L_2(\nu)$.*

We can now give a different characterisation of L_2-differentability based on the L_2-norm.

Proposition 76: Condition for L_2-differentiability

The family of densities $\{p_{\underset{\sim}{\theta}}(.) : \underset{\sim}{\theta} \subseteq \Theta\}$ is said to be $L_2(\nu)-$differentiable at $\underset{\sim}{\theta_0}$ if there exists a vector of functions $\underset{\sim}{\underset{\sim}{\ell^{\circ}_{\theta}}}$ such that

$$\int \frac{\left[\sqrt{p_{\underset{\sim}{\theta}+h}} - \sqrt{p_{\underset{\sim}{\theta}}} - \frac{1}{2}h^T \underset{\sim}{\underset{\sim}{\ell^{\circ}_{\theta}}} \sqrt{p_{\underset{\sim}{\theta}}}\right]^2}{|h|} d\nu \to 0$$

as $|h| \to 0$.

Remark 16.254 *The product* $\underset{\sim}{\underset{\sim}{\ell^{\circ}_{\theta}}} \sqrt{p_{\underset{\sim}{\theta}}}$ *is half the* L_2*-derivative of the square root of the density* $\sqrt{p_{\underset{\sim}{\theta}}}$.

Here, we have that it is converging to zero in the $L_2(\nu)$ norm instead of pointwise in $\underset{\sim}{x}$ in our original definition of $L_2(\nu)$-differentability. We will soon see that $L_2(\nu)-$differentiability is sufficient for LAN to hold.

16.10.9 Strict Conditions for the LAN property to hold

We now want to look at what are stricter conditions for the LAN property to hold. They may be stricter but they are also much easier to verify.

Definition 16.255 *(Cube). Define the parameter vector* $\theta_0 = (\theta_{01}, ..., \theta_{0d})^T$. *Then, the* $\epsilon-$*neighbourhood (cube) of* θ_0 *is given by*

$$C(\theta_0, \epsilon) = \{\theta = (\theta_1, ..., \theta_d)^T \in \Theta : \max_a |\theta_a - \theta_{0a}| < \epsilon\}$$

We now state the "restrictive" condition for LAN property.

Proposition 77: Restrictive Conditions for LAN Property

Suppose $\underset{\sim}{X}_1, ..., \underset{\sim}{X}_n$ are iid random vectors with common density $p_{\theta_0}(.)$ where θ_0 is an interior point of the parameter space $\Theta \subset \mathbb{R}^d$ of the parametric family of densities $\{p_\theta(.) : \theta \in \Theta\}$. Define the log density to be $\ell_\theta(\underset{\sim}{x}) = log\ p_\theta(\underset{\sim}{x})$. The 3 conditions needed for the LAN property to hold are as follows.

1. For some $\epsilon > 0$ and all $\theta \in C(\theta_0, \epsilon)$, the third-order partial derivatives with respect to elements of θ exists for (p_{θ_0}-almost) all $\underset{\sim}{x}$ and satisfy

$$\Big|\frac{\partial^3 \ell_\theta(\underset{\sim}{x})}{\partial\theta_a \partial\theta_b \partial\theta_c}\Big| \leq A(\underset{\sim}{x})$$

   for a function $A(.)$ satisfying $\mathbb{E}[A(\underset{\sim}{X}_1)] < \infty$ for all $a, b, c = 1, ..., d$;

2. The expectation of the log density is zero

$$\mathbb{E}[\ell^\circ_{\theta_0}(\underset{\sim}{X}_1)] = \underset{\sim}{0};$$

3. The information matrix

$$\underset{\sim}{J} = \mathbb{E}[\ell^\circ_{\theta_0}(\underset{\sim}{X}_1)\ell^\circ_{\theta_0}(\underset{\sim}{X}_1)^T] = -\mathbb{E}[\ell^{\circ\circ}_{\theta_0}(\underset{\sim}{X}_1)]$$

   exists and is positive definite.

Then, the LAN condition holds at θ_0 with score function $\underset{\sim}{S}_n = n^{-1/2}\sum_{i=1}^n \ell^\circ_{\theta_0}(\underset{\sim}{X}_i)$ and information matrix $\underset{\sim}{J}$ given above.

Proof:(Sketch). As the log density has 3rd-order derivatives, we can take Taylor expansions on the log likelihood under nearby alternatives $\sum_{i=1}^n \ell_{\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h}}(\underset{\sim}{X}_i)$. We then show that the third order term in this expansion is a remainder term R_n, which is bounded by $A(\underset{\sim}{x})$ and therefore we can show that $R_n \xrightarrow{p} 0$. Then, the second order term can be shown to converge to the information matrix $\underset{\sim}{J}$ and the first order term is the score. Then, we can write out the log likelihood ratio

$$L_n(\underset{\sim}{X}, \underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h}; \underset{\sim}{\theta}_0) = \sum_{i=1}^n \ell_{\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h}}(\underset{\sim}{X}_i) - \sum_{i=1}^n \ell_{\underset{\sim}{\theta}_0}(\underset{\sim}{X}_i) = \underset{\sim}{h}^T \underset{\sim}{S}_n - \frac{1}{2}\underset{\sim}{h}^T \underset{\sim}{J}\underset{\sim}{h} + R_n$$

and hence the LAN property holds. ■

Corollary 16.256 *If the restrictive assumptions for LAN holds, we can also compute the information matrix by taking the negative of the expectation of the Hessian of the score vector.*

We now want to show that under these restrictive assumptions, that the score vector converges to a normal distribution. First, recall the multivariate central limit theorem.

Proposition 16.257 *Suppose that* $\underset{\sim}{X} = (x_1, x_2, ..., x_k)^T \in \mathbb{R}^k$ *is a random vector with covariance* Σ. *Assume that* $\mathbb{E}[x_i] < \infty$. *Then, if* $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ...$ *is a sequence of i.i.d copies of* $\underset{\sim}{X}$ *then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\underset{\sim}{X}_i - \mathbb{E}[\underset{\sim}{X}_i] \right) \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{J})$$

Proposition 78: Score function converges to a normal distribution

Suppose that the restrictive conditions for the LAN property hold. Then

$$\underset{\sim}{S}_n \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{J})$$

Proof: By the second assumption, we have that

$$\mathbb{E}[\underset{\sim}{S}_n] = \mathbb{E}[n^{-1/2} \sum_{i=1}^{n} \ell^{\circ}_{\theta_0}(\underset{\sim}{X}_i)] = \underset{\sim}{0}$$

By the third assumption, we have that $\underset{\sim}{J}$ is the covariance matrix of $\underset{\sim}{S}_n$. Then, by the multivariate central limit theorem, we have that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\ell^{\circ}_{\theta_0}(\underset{\sim}{X}_i) - \mathbb{E}[\ell^{\circ}_{\theta_0}(\underset{\sim}{X}_i)] \right) = \underset{\sim}{S}_n - 0 \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{J})$$

■

We can actually relax the first assumption in LAN to the following. That is, instead of having a fixed cube, we can now have a shrinking cube, where the speed in which it is shrinking is slower than $n^{-1/2}$. Additionally, instead of it holding for all $\underset{\sim}{x}$, it only holds for all $\underset{\sim}{x}$ in a set B_n which is approaching one. That is, the probability of not being in B_n quickly approaches $\underset{\sim}{0}$. Thirdly, the third derivative is bounded by $A_n(\underset{\sim}{x})$ where the function A_n now depends on n and this only applies to $\underset{\sim}{x}$ inside the set B_n. Finally, as A_n is no longer fixed, we require the limit superior of A_n is finite.

Proposition 79: Relaxation of assumption 1 in restrictive conditions for LAN

For some $\epsilon_n \to 0$ more slowly than $n^{-1/2}$ and all $\theta \in C(\theta_0, \epsilon_n)$, the third-order partial derivatives exists with respect to the elements of θ for all $\underset{\sim}{x}$ in a set B_n with $P_{\theta_0}(B_n^c) \to 0$ faster than n^{-1}. Furthermore, the third-order partial derivatives satisfy

$$1\{\underset{\sim}{x} \in B_n\}\Big|\frac{\partial^3 \ell_\theta(\underset{\sim}{x})}{\partial\theta_a\partial\theta_b\partial\theta_c}\Big| \leq A_n(\underset{\sim}{x})$$

for functions $\{A_n(.)\}$ satisfying $lim\sup_{n\to\infty}\mathbb{E}[A_n(\underset{\sim}{X}_1) < \infty$ for all $a, b, c = 1, ..., d$.

16.10.10 LAN property for Location and Scale Families

We now want to apply LAN to location and scale families. We can now describle location and scale families based on the following theorem.

Theorem 16.258 *Let f(x) be any probability density function and let μ and $\sigma > 0$ be any given constants. Then*

$$g(x|\mu, \sigma) = \frac{1}{\sigma}f(\frac{x-\mu}{\sigma})$$

is a probability density function.

Definition 16.259 *(Location family). Let f(x) be a probability density function. Then, the family of probability density functions $\{f(x-\mu) : \mu \in \mathbb{R}\}$ is called the location family with standard probability density function f(x) and μ is called the location parameter for the family.*

Proposition 16.260 *(Properties on standard probability density function). Suppose f(x) is a probability density function on the real line and for some $-\infty \leq a < b \leq \infty$, we have that*

1. *$f(x) > 0$ for all $a < x < b$ and $\int_a^b f(x)dx = 1$*
2. *f(.) is twice continuously differentiable at all $a < x < b$*
3. *$\lim_{x\to a} f(x) = \lim_{x\to b} f(x) = \lim_{x\to a} f'(x) = \lim_{x\to b} f'(x) = 0$*
4. *$\int_a^b \frac{f'(x)^2}{f(x)}dx < \infty$*

Then, define the location family $p_\theta(x) = f(x-\theta)$. We then have that conditions (2) and (3) for the restrictive conditions for LAN property hold. That is,

$$\mathbb{E}[\ell^\circ_{\theta_0}(\underset{\sim}{X}_1)] = \underset{\sim}{0};$$

and

$$\underset{\sim}{J} = \mathbb{E}[\ell^\circ_{\theta_0}(\underset{\sim}{X}_1)\ell^\circ_{\theta_0}(\underset{\sim}{X}_1)^T] = -\mathbb{E}[\ell^{\circ\circ}_{\theta_0}(\underset{\sim}{X}_1)]$$

exists and is positive definite.

Proposition 80: LAN property for Location models

Any location models whose standard probability density satisfies the above conditions, will satisfy conditions 2 and 3 in the restrictive conditions for LAN property. That is, a location model will have an information matrix and expectation of the log density is 0.

Remark 16.261 *This means that to show LAN property holds for a location model, we only need to prove condition 1!*

Definition 16.262 *(Scale family). Let f(x) be a probability density function. Then, the family of probability density functions* $\{\frac{1}{\sigma}f(\frac{x}{\sigma}) : \sigma \in \mathbb{R}\}$ *is called the scale family with standard probability density function f(x) and* σ *is called the scale parameter for the family.*

Definition 16.263 *(Location-Scale family). Let f(x) be a probability density function. Then, the family of probability density functions* $\{\frac{1}{\sigma}f(\frac{x-\mu}{\sigma}) : \mu, \sigma \in \mathbb{R}\}$ *is called the location-scale family with standard probability density function f(x),* μ *is called the location parameter and* σ *is called the scale parameter for the family.*

Proposition 81: LAN property for Location-Scale models

Any location-scale models whose standard probability density satisfies the above conditions, will satisfy conditions 2 and 3 in the restrictive conditions for LAN property. That is, a location-scale model will have an information matrix and expectation of the log density is 0.

STAT4028: Probability and Mathematical Statistics

17. L^2 differentiability implies LAN

17.10.11 L^2 differentiability implies LAN

In this lecture, we want to show that we no longer require second or third derivatives to exists in order for the LAN property to hold. We only require L_2-differentiability.

We now state a collection of lemmas which can be used to help prove the main theorem of this section.

Proposition 17.264 *(Continuity of the inner product). Suppose we have sequences of functions $\{f_n\}$ and $\{g_n\}$ whereby $f_n \to f$ and $g_n \to g$ in $L_2(\nu)$. That is, $\int (f_n - f)^2 d\nu \to 0$. Then*

$$\int f_n g_n d\nu \to \int f g d\nu.$$

Lemma 17.265 *(Convergence in probability to a constant). Let $X_1, X_2, ...$ be a sequence of random variables where $\mathbb{E}[X_n] \to \mu$ and $Var[X_n] \to 0$. Then*

$$X_n \xrightarrow{p} \mu.$$

Lemma 17.266 *We may write*

$$2log(1 + \frac{x}{2}) = x - \frac{x^2}{4} + x^2 R(x)$$

where $R(x) \to 0$ as $x \to 0$.

Lemma 17.267 *Suppose we have a sequence of random variables $\{A_n\}$ and $\{B_n\}$ that have the mean square $\lim_{n\to\infty} \sup E[A_n^2] < \infty$ and $E((A_n - B_n)^2) \to 0$. Then, we may write $A_n^2 = B_n^2 + C_n$ with*

$$E(|C_n|) \to 0.$$

Lemma 17.268 *For a non-negative random variable X with $E(X) < \infty$, then for any $C_n \to \infty$, we have that*

$$E(X.1\{\underset{\sim}{X} > C_n\}) \to 0.$$

Lemma 17.269 *For identically disitributed random variables $\underset{\sim}{X}_1, ...\underset{\sim}{X}_n$, for any real C,*

$$\mathbb{P}(\max_{i=1,...,n} \underset{\sim}{X}_i > C) \leq n\mathbb{P}(\underset{\sim}{X}_1 > C).$$

We now state the important theorem.

Theorem 62: L_2-differentiability implies LAN

Let $\{p_{\underset{\sim}{\theta}}(.) : \underset{\sim}{\theta} \in \Theta\}$ for $\Theta \subseteq \mathbb{R}^d$ be a family of densities with respect to a $\sigma-$finite measure $\nu(.)$ and suppose that the map

$$\underset{\sim}{\theta} \to \sqrt{p_{\underset{\sim}{\theta}}}$$

is $L_2(\nu)$-differentiable at $\underset{\sim}{\theta_0}$ and that $\underset{\sim}{X}_1, ..., \underset{\sim}{X}_n$ are i.i.d random vectors with common density $p_{\underset{\sim}{\theta_0}}(.)$. Then, for any fixed vector $\underset{\sim}{h} \in \mathbb{R}^d$, one may write

$$\int \Big[\sqrt{n}\Big(\sqrt{p_{\underset{\sim}{\theta}+n^{-1/2}\underset{\sim}{h}}} - \sqrt{p_{\underset{\sim}{\theta}}}\Big) - \frac{1}{2}\underset{\sim}{h}^T \underset{\sim}{\ell}^{\circ}_{\underset{\sim}{\theta_0}} \sqrt{p_{\underset{\sim}{\theta_0}}}\Big]^2 d\nu \to 0.$$

Then, under these conditions and that there exists a $\underset{\sim}{\ell}^{\circ}_{\underset{\sim}{\theta_0}}$ which satisfies L_2-differentability, we have that

1. $E(\underset{\sim}{\ell}^{\circ}_{\underset{\sim}{\theta_0}}(\underset{\sim}{X}_i)) = \underset{\sim}{0}$
2. The matrix $\underset{\sim}{J} = \underset{\sim}{J}(\underset{\sim}{\theta_0} = E[\underset{\sim}{\ell}^{\circ}_{\underset{\sim}{\theta_0}}(\underset{\sim}{X}_i).\underset{\sim}{\ell}^{\circ}_{\underset{\sim}{\theta_0}}(\underset{\sim}{X}_i)^T]$ exists
3. The LAN property holds at $\underset{\sim}{\theta_0}$ with score vector

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underset{\sim}{\ell}^{\circ}_{\underset{\sim}{\theta_0}}(\underset{\sim}{X}_i)$$

and information matrix $\underset{\sim}{J}$.

Remark 17.270 *Note that the information matrix is not necessarily the expectation of the Hessian matrix of the score vector.*

Hence, in order for the LAN property to hold, we only require that the map of the square root of the family of densities is L_2-differentiable.

STAT4028: Probability and Mathematical Statistics

18. RAJN Estimators

18.11 Optimality of estimators

In this lecture, we describe what regular and AJN estimators are. We also show additional assumptions we require on top of LAN in order to derive AJN estimators. This restrictive assumption of an estimator being AJN with the scores is the main assumption that will allow us to derive much of theory in the next few sections. This leads us to have a new formulation of Le Cam's third lemma. We then see that RAJN estimators are analogous to unbiased estimators in the Cramér-Rao setting. Finally, we show that we can always decompose a RAJN estimator to its prediction component and extra estimation error.

18.11.1 Regular Estimators

Definition 18.271 *(Stochastic convergence). Let Y_n be a random sequence. The random sequence is $o_p(1)$ if $Y_n \xrightarrow{P} 0$ as $n \to \infty$. The random sequence is $O_p(1)$ if for every $\epsilon > 0$, there exists a constant M such that $lim_{n\to\infty} P(|Y_n| > M) < \epsilon$ as $n \to \infty$.*

In our local asymptotic framework, we are interested in finding something analogous to unbiased estimators.

Definition 18.272 *(Translation-invariant property). An estimator $\hat{\theta}_n$ has a translation property if the estimation error $\hat{\theta}_n - \theta$ has a distribution **not** depending on θ.*

Remark 18.273 *If an estimator has the translation invariance property then the estimation error is the same no matter what value of θ is.*

Definition 89: Regular Estimator

An estimator $\hat{\underset{\sim}{\theta}}_n$ is regular if it locally asymptotically behaves like a translation-invariant location parameter estimator. That is if

$$\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{\Sigma}(\underset{\sim}{\theta}_0))$$

under $\underset{\sim}{\theta}_0$, then for any $\underset{\sim}{h} \in \mathbb{R}^d$, writing $\underset{\sim}{\theta}_n(\underset{\sim}{h}) = \underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h}$, the estimation error for any nearby sequence has the same limiting distribution regardless of $\underset{\sim}{h}$

$$\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_n(\underset{\sim}{h})) \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{\Sigma}(\underset{\sim}{\theta}_0)).$$

That is, the asymptotic distribution of the estimation error possibly depends on $\underset{\sim}{\theta}_0$ but does not depend on the local deviation $\underset{\sim}{h}$.

Remark 18.274 *As we move the nearby alternative $\underset{\sim}{\theta}_n$ around near the true parameter $\underset{\sim}{\theta}_0$, the distribution of the estimator error moves with $\underset{\sim}{\theta}_n$ but only in its location. The shape and spread does not change.*

Corollary 18.275 *Moment and quantile estimators are typically regular estimators.*

Establishing regular estimators allows us to rule out pathological estimators such as Hodge's estimator.

18.11.2 Scores version of Le Cam's 3rd Lemma

We are now interested in restricting our attention to statistics $\underset{\sim}{Y}_n$ such that

$$\underset{\sim}{Y}_n \xrightarrow{d} \mathcal{N}(0, \Sigma_Y).$$

From this, we can apply Le Cam's 3rd lemma to this after making the extra assumption that $\underset{\sim}{Y}_n$ is asymptotically jointly normal with the log likelihood ratio L_n.

Proposition 18.276 *(Alternative limiting distribution of statistic). Suppose we have a model where the LAN property holds at $\underset{\sim}{\theta}_0$ for some score vector $\underset{\sim}{S}_n$ and information matrix $\underset{\sim}{J}$. Suppose a statistic vector $\underset{\sim}{Y}_n$ is such that under the true parameter $\underset{\sim}{\theta}_0$ for every vector $\underset{\sim}{h} \in \mathbb{R}^d$*

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ L_n(\underset{\sim}{h}) \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} 0 \\ -\frac{1}{2}\underset{\sim}{h}^T \underset{\sim}{J}\underset{\sim}{h} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \sigma_{YL}(\underset{\sim}{h}) \\ \sigma_{YL}^T(\underset{\sim}{h}) & \underset{\sim}{h}^T \underset{\sim}{J}\underset{\sim}{h} \end{pmatrix} \Big) \tag{*}$$

where L_n is the log likelihood ratio. Now, applying Le Cam's third lemma, we get that under the nearbly alternative sequence $\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h}$, the limiting joint distribution is

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ L_n(\underset{\sim}{h}) \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} \sigma_{YL}(\underset{\sim}{h}) \\ +\frac{1}{2}\underset{\sim}{h}^T \underset{\sim}{J}\underset{\sim}{h} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \sigma_{YL}(\underset{\sim}{h}) \\ \sigma_{YL}^T(\underset{\sim}{h}) & \underset{\sim}{h}^T \underset{\sim}{J}\underset{\sim}{h} \end{pmatrix} \Big)$$

Remark 18.277 *Here $\sigma_{YL}(\underset{\sim}{h})$ is the vector whose j-th element is the covariance β in Le Cam's third lemma. This is a more restrictive assumption than LAN as we now require a $\underset{\sim}{Y}_n$ which is asymptotically jointly normal with L_n.*

Remark 18.278 *In fact, under random sampling, statistics of the form*

$$\underset{\sim}{Y}_n = n^{-1/2} \sum k(z_i) + o_p(1)$$

where k(Z) has mean zero and finite variance will satisfy both the condition of being asymptotically normal and having a joint limiting distribution with L_n of the form of (*). In particular, sample moments, quantiles, and unbiased estimators satisfy this.

From Le Cam's third lemma, we were able to derive the limiting local-alternative distribution of the statistic $\underset{\sim}{Y}_n$ extremely easy just from knowing its limiting distribution under the true parameter of interst. Then,

recalling that under $\underset{\sim}{\theta}_0$, due to the LAN property, we have that the log likelihood ratio can be expressed as

$$L_n(\underset{\sim}{h}) = \underset{\sim}{h}^T \underset{\sim}{S}_n - \frac{1}{2}\underset{\sim}{h}^T \underset{\sim}{\underset{\sim}{J}} \underset{\sim}{h} + o_p(1).$$

That is, the likelihood ratio is a linear combination of the scores. We can therefore replace $L_n(\underset{\sim}{h})$ with the first 2 terms on the right hand side.

Proposition 18.279 *(Limiting joint distribution of likelihood ratio expressed with score vector). The likelihood ratio is a linear combination of the scores, which implies that the limiting joint distribution of the vector*

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ L_n(\underset{\sim}{h}) + \frac{1}{2}\underset{\sim}{h}^T \underset{\sim}{\underset{\sim}{J}} \underset{\sim}{h} \end{pmatrix}$$

is the same as

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ \underset{\sim}{h}^T \underset{\sim}{S}_n \end{pmatrix}$$

We can therefore get a different formulation of the LAN property seen in (*) at the beginning of this section in terms of the scores vector that under $\underset{\sim}{\theta}_0$

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ \underset{\sim}{h}^T \underset{\sim}{S}_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} \underset{\sim}{0} \\ \underset{\sim}{0} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \sigma_{YL}(\underset{\sim}{h}) \\ \sigma_{YL}^T(\underset{\sim}{h}) & \underset{\sim}{h}^T \underset{\sim}{\underset{\sim}{J}} \underset{\sim}{h} \end{pmatrix} \Big) \tag{**}$$

since under the LAN assumption, the score vector $\underset{\sim}{S}_n \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{\underset{\sim}{J}})$.

We now make a quick detour. Sometimes, it is better to take transforms of random vectors in order to show convergence.

Definition 90: Characteristic Function

Let $\underset{\sim}{X} \in \mathbb{R}^d$ be a random vector. Then, the characteristic function is defined by

$$\psi(\underset{\sim}{t}) = \mathbb{E}[e^{i \underset{\sim}{t}^T \underset{\sim}{X}}]$$

for $\underset{\sim}{t} \in \mathbb{R}^d$.

Lemma 18.280 *The Characteristic function always exists for a random vector.*

Lemma 18.281 *The characteristic function $\psi(\underset{\sim}{t})$ is continuous and bounded for every $\underset{\sim}{t} \in \mathbb{R}^d$.*

By the Portmanteau lemma, we can then show that convergence in characteristic function is equivalent to convergence in distribution.

Theorem 63: Lévy's Continuity Theorem

If $\underset{\sim}{X}, \underset{\sim}{X}_1, \underset{\sim}{X}_2, ...$ are random vectors with respective characteristic functions $\psi, \psi_1, \psi_2, ...$, then $\underset{\sim}{X}_n \xrightarrow{d} \underset{\sim}{X}$ if and only if

$$\psi_n(\underset{\sim}{t}) \to \psi(\underset{\sim}{t})$$

for all $\underset{\sim}{t} \in \mathbb{R}^d$.

We can use Lévy's continuity theorem to show another important result.

Proposition 82: Cramér-Wold Theorem

et $\underset{\sim}{X} \in \mathbb{R}^d$ and $\underset{\sim}{X}_1, ..., \underset{\sim}{X}_n \in \mathbb{R}^d$ be random variables. Furthermore, let $\underset{\sim}{t} \in \mathbb{R}^d$. Then

$$\underset{\sim}{X}_n \xrightarrow{d} \underset{\sim}{X}$$

if and only if

$$\underset{\sim}{t}^T \underset{\sim}{X}_n \xrightarrow{d} \underset{\sim}{t}^T \underset{\sim}{X}$$

for all $\underset{\sim}{t} \in \mathbb{R}^d$.

So now let us revisit

$$\begin{pmatrix} Y_n \\ \underset{\sim}{h}^T \underset{\sim}{S}_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} \underset{\sim}{0} \\ \underset{\sim}{0} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \sigma_{YL}(\underset{\sim}{h}) \\ \sigma_{YL}^T(\underset{\sim}{h}) & \underset{\sim}{h}^T J \underset{\sim}{h} \end{pmatrix} \Big) \tag{**}$$

Lemma 18.282 *The covariance matrix $\sigma_{YL}(\underset{\sim}{h})$ between the statistic $\underset{\sim}{Y}$ and log likelihood ratio is a linear function of $\underset{\sim}{h}$.*

Proof: Suppose that $\begin{pmatrix} Y \\ \underset{\sim}{S} \end{pmatrix}$ is a multivariate normal distribution such that $\begin{pmatrix} Y \\ \underset{\sim}{h}^T \underset{\sim}{S} \end{pmatrix}$ has the exact multivariate distribution of equation (**) via the Cramer-Wold theorem. Then, we have that

$$\sigma_{YL}(\underset{\sim}{h}) = \mathbb{E}[\underset{\sim}{Y} \underset{\sim}{S}^T \underset{\sim}{h}] = \Sigma_{YS} \underset{\sim}{h}$$

where $\Sigma_{YS} = \mathbb{E}[\underset{\sim}{Y} \underset{\sim}{S}^T]$ is the cross-covariance. ■

As a result of the Cramér-Wold theorem, the assumptions (*) and (**) are equivalent to the Augmented LAN property, which is where we now impose an additional assumption on the random vector $\underset{\sim}{Y}_n$.

Theorem 64: Augmented LAN (LAN#)

Suppose that the LAN property holds at $\underset{\sim}{\theta}_0$ for some score $\underset{\sim}{S}_n$ and information matrix $\underset{\sim}{J}$. Furthermore, assume that for some positive semi-definite matrix $\underset{\sim}{\Sigma}_Y$, the statistic vector $\underset{\sim}{Y}_n$ is jointly asymptotically normal with the scores

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ \underset{\sim}{S}_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\underset{\sim}{0}, \begin{pmatrix} \underset{\sim}{\Sigma}_Y & \underset{\sim}{\Sigma}_{YS} \\ \underset{\sim}{\Sigma}_{SY} & \underset{\sim}{J} \end{pmatrix}\Big)$$

under $\underset{\sim}{\theta}_0$.

Proof: We have shown that the under $\underset{\sim}{\theta}_0$, the limiting joint distribution of the statistic $\underset{\sim}{Y}_n$ and the shifted score $\underset{\sim}{h}^T \underset{\sim}{S}_n$ is given by

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ \underset{\sim}{h}^T \underset{\sim}{S}_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} \underset{\sim}{0} \end{pmatrix}, \begin{pmatrix} \underset{\sim}{\Sigma}_Y & \underset{\sim}{\Sigma}_{YS}\underset{\sim}{h} \\ \underset{\sim}{\Sigma}_{SY}\underset{\sim}{h} & \underset{\sim}{h}^T \underset{\sim}{J} \underset{\sim}{h} \end{pmatrix}\Big)$$

Let us denote the random vector of the target distribution

$$\begin{pmatrix} \underset{\sim}{Y} \\ \underset{\sim}{S} \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\underset{\sim}{0}, \begin{pmatrix} \underset{\sim}{\Sigma}_Y & \underset{\sim}{\Sigma}_{YS} \\ \underset{\sim}{\Sigma}_{SY} & \underset{\sim}{J} \end{pmatrix}\Big)$$

Suppose that $\underset{\sim}{Y}$ has dimension c and $\underset{\sim}{S}$ has dimension d. Then, we have that for any vector $\underset{\sim}{g} \in \mathbb{R}^c$

$$\begin{pmatrix} \underset{\sim}{g} & 1 \end{pmatrix} \begin{pmatrix} \underset{\sim}{Y}_n \\ \underset{\sim}{h}^T \underset{\sim}{S}_n \end{pmatrix} = \underset{\sim}{g}^T \underset{\sim}{Y}_n + \underset{\sim}{h}^T \underset{\sim}{S}_n \xrightarrow{d} \underset{\sim}{g}^T \underset{\sim}{Y} + \underset{\sim}{h}^T \underset{\sim}{S}$$

However, this is equivalent to

$$\underset{\sim}{t}^T \begin{pmatrix} \underset{\sim}{Y}_n \\ \underset{\sim}{h}^T \underset{\sim}{S}_n \end{pmatrix} \xrightarrow{d} \underset{\sim}{t}^T \begin{pmatrix} \underset{\sim}{Y} \\ \underset{\sim}{h}^T \underset{\sim}{S} \end{pmatrix}$$

for all $\underset{\sim}{t} \in \mathbb{R}^{c+d}$. Then, by the Cramér-Wold theorem, where $\underset{\sim}{t}^T \underset{\sim}{X}_n \xrightarrow{d} \underset{\sim}{t}^T \underset{\sim}{X}$ implies that $\underset{\sim}{X}_n \xrightarrow{d} \underset{\sim}{X}$, we have that

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ \underset{\sim}{S}_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\underset{\sim}{0}, \begin{pmatrix} \underset{\sim}{\Sigma}_Y & \underset{\sim}{\Sigma}_{YS} \\ \underset{\sim}{\Sigma}_{SY} & \underset{\sim}{J} \end{pmatrix}\Big)$$

■

Remark 18.283 *LAN# is an assumption about the random statistic vector Y. LAN# is the LAN property with the added condition that the statistic is also asymptotically jointly normal with the scores vector. The original form of Le Cam's third lemma described that the statistic is asymptotically jointly normal with the likelihood ratio. Restricting our analysis to statistics that satisfy LAN# throughout the rest of this course will allow us to easily develop alot of new theory!*

We have now imposed another condition on top of LAN for the statistic $\underset{\sim}{Y}_n$. It is worthwhile to note that the covariance matrix $\underset{\sim}{\Sigma}_Y$ does **not** need to have full-rank.

Using the augmented LAN assumption, we can deduce a new version of scores version of Le Cam's third lemma where we repeat the process we have just done for the nearby sequence $\underset{\sim}{\theta}_0 + n^{1/2}\underset{\sim}{h}$.

Proposition 83: Scores Version of Le Cam's Third Lemma

Suppose that LAN# holds for the statistic $\underset{\sim}{Y}_n$. Then, under a nearby alternative $\underset{\sim}{\theta}_0 + n^{1/2}\underset{\sim}{h}$, we have that

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ \underset{\sim}{S}_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} \underset{\sim}{\Sigma}_{YS}\underset{\sim}{h} \\ \underset{\sim}{J}\underset{\sim}{h} \end{pmatrix}, \begin{pmatrix} \underset{\sim}{\Sigma}_Y & \underset{\sim}{\Sigma}_{YS} \\ \underset{\sim}{\Sigma}_{SY} & \underset{\sim}{J} \end{pmatrix} \Big)$$

That is, the asymptotic mean of $\underset{\sim}{Y}_n$ is now a linear combination of $\underset{\sim}{h}$ given by $\underset{\sim}{\Sigma}_{YS}\underset{\sim}{h}$.

Proof: First, recall that from (**), we have that

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ \underset{\sim}{h}^T\underset{\sim}{S}_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} \underset{\sim}{0} \\ \underset{\sim}{0} \end{pmatrix}, \begin{pmatrix} \underset{\sim}{\Sigma}_Y & \underset{\sim}{\sigma}_{YL}(\underset{\sim}{h}) \\ \underset{\sim}{\sigma}_{YL}^T(\underset{\sim}{h}) & \underset{\sim}{h}^T\underset{\sim}{J}\underset{\sim}{h} \end{pmatrix} \Big) \tag{**}$$

under $\underset{\sim}{\theta}_0$. ■

Remark 18.284 *We have rewritten Le Cam's third lemma in terms of the score function and as a result, we get more information because now, we see that the limiting mean of the statistic $\underset{\sim}{Y}_n$ is a linear combination of $\underset{\sim}{\Sigma}_{SY}$.*

18.11.3 RAJN Estimators

Definition 91: Asymptotically Jointly Normal Estimator

We say that an estimator $\underset{\sim}{\hat{\theta}}_n = \underset{\sim}{\hat{\theta}}_n(\underset{\sim}{X}_1, ..., \underset{\sim}{X}_n)$ is asymptotically jointly normal (AJN) with the scores $\underset{\sim}{S}_i$ at $\underset{\sim}{\theta}_0$ if the estimation error

$$\underset{\sim}{Y}_n = \sqrt{n}(\underset{\sim}{\hat{\theta}}_n - \underset{\sim}{\theta}_0)$$

satisfies LAN#. That is, under a nearby alternative $\underset{\sim}{\theta}_0 + n^{1/2}\underset{\sim}{h}$, we have that

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ \underset{\sim}{S}_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} \underset{\sim}{\Sigma}_{YS}\underset{\sim}{h} \\ \underset{\sim}{J}\underset{\sim}{h} \end{pmatrix}, \begin{pmatrix} \underset{\sim}{\Sigma}_Y & \underset{\sim}{\Sigma}_{YS} \\ \underset{\sim}{\Sigma}_{SY} & \underset{\sim}{J} \end{pmatrix} \Big)$$

Remark 18.285 *This essentially says that when the estimator is rescaled, it is asymptotically jointly normal with the scores vector.*

We now want to combine the 2 properties of regularity and AJN to derive a new class of estimators.

Definition 92: RAJN Estimator

A RAJN estimator is an estimator that is both regular and asymptotically jointly normal with the scores.

Theorem 65: RAJN Estimators are analogous to unbiased estimators

An estimator $\hat{\underset{\sim}{\theta}}_n$ is a RAJN estimator if and only if its rescaled estimation error has a limiting covariance matrix with the score vector equal to the identity matrix. That is, $\underset{\sim}{\Sigma}_{\underset{\sim}{Y}\underset{\sim}{S}} = \mathbb{I}$ where $\underset{\sim}{Y}_n = \sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_n) + \underset{\sim}{h}$.

Proof: Suppose $\hat{\underset{\sim}{\theta}}_n$ is a RAJN estimator. Then, due to the fact that is AJN, then under the nearby sequence $\underset{\sim}{\theta}_n = \underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h}$, we have that the estimation error is asymptotically normal

$$\sqrt{n}(\hat{\theta}_n - \underset{\sim}{\theta}_n) = \underset{\sim}{Y}_n - \underset{\sim}{h} \xrightarrow{d} \mathcal{N}(\underset{\sim}{\Sigma}_{\underset{\sim}{Y}\underset{\sim}{S}}\underset{\sim}{h} - \underset{\sim}{h}, \underset{\sim}{\Sigma}_{\underset{\sim}{Y}}).$$

However, as $\hat{\underset{\sim}{\theta}}_n$ is a regular estimator, we require that the limiting distribution to be free of $\underset{\sim}{h}$ under the alternative $\underset{\sim}{\theta}_n$. This only holds if the cross-covariance of the estimation error and score vector

$$\underset{\sim}{\Sigma}_{\underset{\sim}{Y}\underset{\sim}{S}} = \mathbb{I}$$

where $\mathbb{I}$ is the d-by-d identity matrix. ■

Remark 18.286 *This property is directly analogous to the Cramer-Rao result that under regularity conditions where any unbiased estimator has covariance with the score function equal to one.*

Now, recall the correlation inequality.

Definition 18.287 *(Correlation Inequality). The correlation inequality between an unbiased estimator $\hat{\theta}$ and score function ℓ_θ° is*

$$Cov_\theta[\hat{\theta}(\underset{\sim}{x}), \ell_\theta^\circ(\underset{\sim}{x})]^2 \leq Var_\theta(\hat{\theta}(\underset{\sim}{x}))Var_\theta(\ell_\theta^\circ(\underset{\sim}{x}))$$

with equality if and only if $\hat{\theta}(\underset{\sim}{x})$ and $\ell_\theta^\circ(\underset{\sim}{x})$ are linearly related so that

$$\ell_\theta^\circ(\underset{\sim}{x}) = a_\theta + b_\theta\hat{\theta}(\underset{\sim}{x}).$$

Hence, we can now get an analogous asymptotic multivariate generalisations of the correlation inequality.

Theorem 66: Orthogonal Decomposition Theorem

Any estimator which is RAJN at $\underset{\sim}{\theta}_0$ has the representation

$$\hat{\underset{\sim}{\theta}}_n = \underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{J}^{-1}\underset{\sim}{S}_n + n^{-1/2}\underset{\sim}{Z}_n$$

under $\underset{\sim}{\theta}_0$. Alternatively, we can say that the estimation error for any RAJN estimator can be decomposed into a common component and an added on noise

$$\sqrt{n}[\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0] = \underset{\sim}{J}^{-1}\underset{\sim}{S}_n + \underset{\sim}{Z}_n$$

where the extra estimation error $\underset{\sim}{Z}_n$ is asymptotically independent of the scores vector

$$\begin{pmatrix} \underset{\sim}{Z}_n \\ \underset{\sim}{S}_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\underset{\sim}{0}, \begin{pmatrix} \Sigma_Y - \underset{\sim}{J}^{-1} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{I} \end{pmatrix}\Big)$$

Proof: This follows from the Scores version of Le Cam's 3rd lemma after defining the statistic

$$\underset{\sim}{Y}_n = \sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0)$$

and

$$\underset{\sim}{Z}_n = \underset{\sim}{Y}_n - \underset{\sim}{J}^{-1}\underset{\sim}{S}_n$$

where the cross-covariance $\underset{\sim}{\Sigma}_{YS} = \mathbb{I}$ due to the RAJN estimator.

■

Remark 18.288 *This result arises due to the estimator being both regular and AJN with the scores vector.*

The significance of the fact that the extra estimation noise is asymptotically independent of the scores is that we can then write

$$AVar\Big(\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0)\Big) = AVar(\underset{\sim}{J}^{-1}\underset{\sim}{S}_n) + AVar(\underset{\sim}{Z}_n)$$

So if we want to minimise the variance of the estimation error, we need $AVar(\underset{\sim}{Z}_n) = 0$.

In fact, we can see the orthogonal decomposition theorem as an asymptotic version of the Cramér-Rao lower bound

$$\underset{\sim}{Y}_n = n^{1/2}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_n) \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{J}^{-1} + \Sigma_Z)$$

where the best we can do is if $\Sigma_Z = \underset{\sim}{0}$.

From this, the best we can do is to not have the $n^{-1/2}\underset{\sim}{Z}_n$ term in our estimation error whilst every estimator has $\underset{\sim}{J}^{-1}\underset{\sim}{S}_n$ component. Hence, the best we can do to minimise the estimation error is to remove this term. This will lead to a new definition of optimality which we will investigate in the next section.

STAT4028: Probability and Mathematical Statistics

19. Asymptotically Efficient Estimators

19.11.4 Asymptotically Efficient Estimators

In this lecture, we now define what we mean by efficient estimators. We then introduce the regular scores assumption which assumes that the LAN property holds in an area and that the scores vectors are "differentiable". From this, we can then expand the class of estimators that are efficient. With this expansion of efficient estimators, we will look at how to construct efficient estimator starting from an arbitrary $\sqrt{n}-$consistent estimators. This will be of a form of a 1-step adjustment to turn our estimator into an asymptotically efficient estimator. That is, (near) roots of score equations can provide efficient estimates.

We have described the class of RAJN estimators. We now want to know which of these RAJN estimators are the most optimal. This is analogous to what we have seen before whereby RAJN estimators are analogous to unbiased estimators. Furthermore, the estimators that are asymptotically efficient are analogous to unbiased estimators that have minimum variance through attaining the Cramér-Rao lower bound.

Definition 93: Asymptotically Efficient Estimators

Let $\underset{\sim}{\hat{\theta}_n}$ be a RAJN estimator. Then $\underset{\sim}{\hat{\theta}_n}$ is asymptotically efficient (AE) at $\underset{\sim}{\theta_0}$ if it can be written as

$$\sqrt{n}(\underset{\sim}{\hat{\theta}_n} - \underset{\sim}{\theta_0}) = \underset{\sim}{J}^{-1}\underset{\sim}{S}_n + o_p(1).$$

Remark 19.289 *Note that this follows from the orthogonal decomposition thereom where the extra noise error Z_n is now $o_p(1)$*

$$\sqrt{n}[\underset{\sim}{\hat{\theta}_n} - \underset{\sim}{\theta_0}] = \underset{\sim}{J}^{-1}\underset{\sim}{S}_n + \underset{\sim}{Z}_n$$

Theorem 67: Sufficient condition for RAJN estimator

Suppose the LAN property holds at $\underset{\sim}{\theta_0}$ with score vector $\underset{\sim}{S}_n$ and information matrix $\underset{\sim}{J}$. Then, if we can write an estimator $\underset{\sim}{\hat{\theta}_0}$ in this form

$$\sqrt{n}(\underset{\sim}{\hat{\theta}_0} - \underset{\sim}{\theta_0}) = \underset{\sim}{J}^{-1}\underset{\sim}{S}_n + o_p(1)$$

then $\underset{\sim}{\hat{\theta}_0}$ is a RAJN (regular, asymptotically jointly normal with the scores) estimator at $\underset{\sim}{\theta_0}$.

Proof: First, denote the estimation error $\underset{\sim}{Y}_n = \sqrt{n}(\underset{\sim}{\hat{\theta}_0} - \underset{\sim}{\theta_0})$. Then, due to the LAN property, the joint behaviour with the scores is

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ \underset{\sim}{S}_n \end{pmatrix} = \begin{pmatrix} \underset{\sim}{J}^{-1}\underset{\sim}{S}_n + o_p(1) \\ \underset{\sim}{S}_n \end{pmatrix} = \begin{pmatrix} \underset{\sim}{J}^{-1}\underset{\sim}{S}_n \\ \underset{\sim}{S}_n \end{pmatrix} + o_p(1) \xrightarrow{d} \mathcal{N}\Big(\underset{\sim}{0}, \begin{bmatrix} \underset{\sim}{J}^{-1} & \underset{\sim}{I} \\ \underset{\sim}{I} & \underset{\sim}{J} \end{bmatrix}\Big)$$

Hence, we have that $\hat{\underset{\sim}{\theta}}_n$ is AJN with the scores. Then, under a nearby sequence $\underset{\sim}{\theta}_n = \underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h}$, we have by the Scores version of Le Cam's third lemma

$$\underset{\sim}{Y}_n \xrightarrow{d} \mathcal{N}\left(\underset{\sim}{\underset{\sim}{I}}\underset{\sim}{h}, \underset{\sim}{J}^{-1}\right) = \mathcal{N}\left(\underset{\sim}{h}, \underset{\sim}{J}^{-1}\right)$$

However, the estimation error $\underset{\sim}{Y}_n = \sqrt{n}(\hat{\underset{\sim}{\theta}}_0 - \underset{\sim}{\theta}_0)$ is under $\underset{\sim}{\theta}_0$. Therefore, the estimation error under $\underset{\sim}{\theta}_n$ is

$$\sqrt{n}(\hat{\underset{\sim}{\theta}}_0 - \underset{\sim}{\theta}_n) = \underset{\sim}{Y}_n - \underset{\sim}{h} \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{J}^{-1})$$

Hence, $\hat{\underset{\sim}{\theta}}_n$ is also a regular estimator. ■

Remark 19.290 *It is AJN as it is asymptotically linear combination of the scores, which means it is AJN with the scores. It is regular because under nearby sequences $\underset{\sim}{S}_n$ enjoys a shift of $\underset{\sim}{h}$.*

19.11.5 Construction of AE Estimators

We now have a characterisation of optimal estimators for RAJN estimators. However, we can relax the conditions for the estimators if we impose additional assumptions on the model. That is, previously, the LAN property only held at a given point, whereby we will now extend the assumption that **the LAN property holds in an area near the point** $\underset{\sim}{\theta}_0$.

Proposition 19.291 *(Neighbourhood LAN property). We shall now assume that LAN holds for all $\underset{\sim}{\theta}$ in a neighbourhood of $\underset{\sim}{\theta}_0$ with scores $\underset{\sim}{S}_n(\underset{\sim}{\theta})$ and information matrix $\underset{\sim}{J} = \underset{\sim}{J}(\underset{\sim}{\theta})$ which now depends on $\underset{\sim}{\theta}$.*

We now have that the scores $\underset{\sim}{S}_n(\underset{\sim}{\theta})$ and information $\underset{\sim}{J}(\underset{\sim}{\theta})$ is a function depending on $\underset{\sim}{\theta}$. We shall now impose a new assumption on the scores vector in order to construct efficient estimators.

Definition 94: Regular Scores Assumption

Assume that the LAN property holds in a neighbourhood of $\underset{\sim}{\theta}_0$. Then, under $\underset{\sim}{\theta}_0$ as the parameter value, we assume that we can write the scores vector as a function

$$\underset{\sim}{S}_n(\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h}) = \underset{\sim}{S}_n(\underset{\sim}{\theta}_0) - \underset{\sim}{\underset{\sim}{J}}(\underset{\sim}{\theta}_0)\underset{\sim}{h} + R_n(\underset{\sim}{\theta}_0, \underset{\sim}{h})$$

where for any finite $0 < M < \infty$, under $\underset{\sim}{\theta}_0$, we have that

$$\sup_{|h| \leq M} R_n(\underset{\sim}{\theta}_0, \underset{\sim}{h}) \xrightarrow{p} 0.$$

We can therefore express the score vector as

$$\underset{\sim}{S}_n(\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h}) = \underset{\sim}{S}_n(\underset{\sim}{\theta}_0) - \underset{\sim}{\underset{\sim}{J}}(\underset{\sim}{\theta}_0)\underset{\sim}{h} + o_p(1)$$

uniformly in bounded h.

Remark 19.292 *One way to interpret the regular scores assumption is that it is a smoothness assumption on the scores vector such that the scores vector is differentiabile in a probability sense.*

As we will see many times, the uniformly in bounded h assumption is needed for when we are plugging in a random $\underset{\sim}{h}$. Recall that the ordinary LAN assumption only holds for a fixed $\underset{\sim}{h}$. If we strengthen it to uniformly bounded $\underset{\sim}{h}$, then we can plug in a random variable instead of a fixed $\underset{\sim}{h}$ and the analogous property still holds. This is tied in to the idea of neighbourhood LAN property as when we start constructing estimators and test statistics, the score vector will be evaluated at a random value of the parameter. When we need to evaluate the scores at a non-fixed parameter, then the LAN assumption alone is not enough. This regular scores assumption allows us to define $\underset{\sim}{S}_n$ in a neighbourhood of $\underset{\sim}{\theta}_0$. Therefore, the regular score can be thought of as an expansion of the score vector in the neighbourhood of $\underset{\sim}{\theta}_0$.

Corollary 19.293 *Under the regular scores assumption, the scores vector has derivatives with respect to* $\underset{\sim}{\theta}$ *"in probability" and these are the elements of the information matrix* $-\underset{\sim}{J}(\underset{\sim}{\theta}_0)$.

Now, under the regular scores assumption, we are actually now able to broaden the class of estimators from the class of RAJN estimators that are asymptotically efficient.

Definition 19.294 *(Bounded in probability). An estimator* $\underset{\sim}{\tilde{\theta}}_n$ *is said to be bounded in probability if for all* $\epsilon > 0$*, there exists* $0 < M_\epsilon < \infty$ *such that*

$$\lim_{n\to\infty} \sup \mathbb{P}_{\underset{\sim}{\theta}_0}\left(|\sqrt{n}(\underset{\sim}{\tilde{\theta}}_n - \underset{\sim}{\theta}_0)| > M_\epsilon\right) \leq \epsilon.$$

Remark 19.295 *Having an estimator bounded in probability means that the mass of a distribution does not flow off to* ∞.

We define a broader class of estimators.

Definition 95: $\sqrt{n}-$consistent estimators

An estimator $\underset{\sim}{\tilde{\theta}}_n$ is said to be $\sqrt{n}-$consistent at $\underset{\sim}{\theta}_0$ if

$$\sqrt{n}(\underset{\sim}{\tilde{\theta}}_n - \underset{\sim}{\theta}_0) = O_p(1)$$

where $O_p(1)$ is bounded in probability.

Remark 19.296 *Recall that consistent estimators are those that converges to the parameter.* $\sqrt{n}-$*consistent specifies the rate at which it converges to the parameter.*

Proposition 84: Sufficient conditions for $\sqrt{n}-$consistent estimators

All regular estimators and all AJN estimators are $\sqrt{n}-$consistent.

However, a $\sqrt{n}$-consistent estimator need not be regular or AJN.

Theorem 68: Expression of the score function for $\sqrt{n}-$consistent estimators

Suppose the regular scores assumption holds at $\underset{\sim}{\theta}_0$. Then, any $\sqrt{n}-$consistent estimator $\underset{\sim}{\tilde{\theta}}_n$ satisfies

$$S_n(\underset{\sim}{\tilde{\theta}}_n) = S_n(\underset{\sim}{\theta}_0) - J(\underset{\sim}{\theta}_0)\sqrt{n}(\underset{\sim}{\tilde{\theta}}_n - \underset{\sim}{\theta}_0) + o_p(1)$$

under $\underset{\sim}{\theta}_0$.

Remark 19.297 *Recall the regular scores assumption*

$$\underset{\sim}{S}_n(\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h}) = \underset{\sim}{S}_n(\underset{\sim}{\theta}_0) - \underset{\sim}{J}(\underset{\sim}{\theta}_0)\underset{\sim}{h} + o_p(1)$$

We replace $\underset{\sim}{h}$ *with* $\sqrt{n}(\underset{\sim}{\tilde{\theta}}_n - \underset{\sim}{\theta}_0)$ *to get*

$$\underset{\sim}{S}_n(\underset{\sim}{\theta}_0 + n^{-1/2}\sqrt{n}(\underset{\sim}{\tilde{\theta}}_n - \underset{\sim}{\theta}_0)) = \underset{\sim}{S}_n(\underset{\sim}{\theta}_0) - \underset{\sim}{J}(\underset{\sim}{\theta}_0)\sqrt{n}(\underset{\sim}{\tilde{\theta}}_n - \underset{\sim}{\theta}_0) + o_p(1)$$

$$S_n(\underset{\sim}{\tilde{\theta}}_n) = S_n(\underset{\sim}{\theta}_0) - J(\underset{\sim}{\theta}_0)\sqrt{n}(\underset{\sim}{\tilde{\theta}}_n - \underset{\sim}{\theta}_0) + o_p(1)$$

We come to an important corollary.

Proposition 85: $\sqrt{n}-$consistent estimators are RAJN and A.E. under certain conditions

Suppose that the regular scores assumption holds, the estimator $\underset{\sim}{\tilde{\theta}}_n$ is $\sqrt{n}-$consistent, and the scores evaluated at that estimator $\underset{\sim}{\tilde{\theta}}_n$ has the property that

$$\underset{\sim}{S}_n(\underset{\sim}{\tilde{\theta}}_n) \xrightarrow{p} 0.$$

Then, $\underset{\sim}{\tilde{\theta}}_n$ is a RAJN estimator and A.E.

Proof: By the theorem for the expression of the score function for $\sqrt{n}-$estimators, we have that under $\underset{\sim}{\theta}_0$

$$S_n(\underset{\sim}{\tilde{\theta}}_n) = S_n(\underset{\sim}{\theta}_0) - J(\underset{\sim}{\theta}_0)\sqrt{n}(\underset{\sim}{\tilde{\theta}}_n - \underset{\sim}{\theta}_0) + o_p(1)$$

$$\sqrt{n}(\underset{\sim}{\tilde{\theta}}_n - \underset{\sim}{\theta}_0) = J(\underset{\sim}{\theta}_0)^{-1}\left[S_n(\underset{\sim}{\theta}_0) - S_n(\underset{\sim}{\tilde{\theta}}_n)\right] + o_p(1)$$

Then, as $\underset{\sim}{S}_n(\underset{\sim}{\tilde{\theta}}_n) \xrightarrow{p} 0$

$$\sqrt{n}(\underset{\sim}{\tilde{\theta}}_n - \underset{\sim}{\theta}_0) = J(\underset{\sim}{\theta}_0)^{-1}S_n(\underset{\sim}{\theta}_0) + o_p(1)$$

where recall this expression by definition is the meaning for an estimator to be asymptotically efficient and a sufficient condition for an estimator to be a RAJN estimator. ■

Proposition 86: All roots to score equations are asymptotically efficient

Under the regular scores assumption, any solution $\hat{\underset{\sim}{\theta}}_n$ to the score equation

$$\underset{\sim}{S}_n(\hat{\underset{\sim}{\theta}}_n) = \underset{\sim}{0}$$

is asymptotically efficient.

Remark 19.298 *Maximum likelihood estimators are not always optimal. However, if the regular scores assumption holds, which implies that LAN holds, then MLE is optimal. That is, the scores function evaluated at the MLE estimate is 0. Therefore, this shows that the MLE estimate is asymptotically efficient.*

Now, we have the question of what to do if the regular scores assumption holds, we have a $\sqrt{n}-$consistent estimator $\tilde{\underset{\sim}{\theta}}_n$ **but** the score equation evaluated at $\tilde{\underset{\sim}{\theta}}_n$ is not zero, i.e. $\underset{\sim}{S}_n(\tilde{\underset{\sim}{\theta}}_n) \neq \underset{\sim}{0}$.

What we can do is, under certain conditions, take our $\sqrt{n}-$consistent estimator $\tilde{\underset{\sim}{\theta}}_n$ and transform it such that the score equation will be zero and therefore the estimator will be RAJN and asymptotically efficient.

We first describe the condition needed to do this.

Proposition 87: Continuous Information Condition

The LAN property holds in a neighbourhood of $\underset{\sim}{\theta}_0$ with information matrix $\underset{\sim}{J}(\underset{\sim}{\theta})$ being continuous at $\underset{\sim}{\theta}_0$. That is, $\underset{\sim}{J}^{-1}(\underset{\sim}{\theta}_0)$ exists.

Remark 19.299 *Recall that $X_n \xrightarrow{p} \mu$ and if $g(.)$ is a continuous function, then $g(X_n) \xrightarrow{p} g(\mu)$. Hence, if we have that $\tilde{\underset{\sim}{\theta}}_n \xrightarrow{p} \underset{\sim}{\theta}_0$ and if $J(.)$ is continuous at $\underset{\sim}{\theta}_0$ then we have that*

$$\underset{\sim}{J}(\tilde{\underset{\sim}{\theta}}_n) \xrightarrow{p} \underset{\sim}{J}(\underset{\sim}{\theta}_0)$$

This will allow for us to justify our estimates of the information matrix.

We come to the main theorem of the lecture on how to construct asymptotically efficient estimators from $\sqrt{n}-$consistent estimators.

Theorem 69: A.E Construction Theorem

Suppose that the regular scores assumption and the continuous information condition both hold at $\underset{\sim}{\theta}_0$. Then, if $\tilde{\underset{\sim}{\theta}}_n$ is a $\sqrt{n}-$consistent estimator, the new estimator

$$\hat{\underset{\sim}{\theta}}_n = \tilde{\underset{\sim}{\theta}}_n + n^{-1/2}\underset{\sim}{J}^{-1}(\tilde{\underset{\sim}{\theta}}_n)\underset{\sim}{S}_n(\tilde{\underset{\sim}{\theta}}_n)$$

is asymptotically efficient.

Remark 19.300 *As the estimator is asymptotically efficient, it is also a RAJN estimator.*

Proof: Assume that the regular scores assumption holds. We have that every asymptotically efficient estimator $\hat{\underset{\sim}{\theta}}_n$ satisfies

$$\hat{\underset{\sim}{\theta}}_n = \underset{\sim}{\theta}_0 + n^{-1/2}\Big(\underset{\sim}{J}^{-1}(\underset{\sim}{\theta}_0)\underset{\sim}{S}_n(\underset{\sim}{\theta}_0) + o_p(1)\Big)$$

As $\tilde{\underset{\sim}{\theta}}_n$ is a $\sqrt{n}-$consistent estimator, we can express any $\sqrt{n}-$consistent estimator as

$$\tilde{\underset{\sim}{\theta}}_n = \underset{\sim}{\theta}_0 + n^{-1/2}\Big(\underset{\sim}{J}(\underset{\sim}{\theta}_0)^{-1}[\underset{\sim}{S}_n(\underset{\sim}{\theta}_0) - \underset{\sim}{S}_n(\tilde{\underset{\sim}{\theta}}_n)]\Big)$$

Plugging this into our expression for the asymptotically efficient estimator expression, we get that

$$\hat{\underset{\sim}{\theta}}_n = \tilde{\underset{\sim}{\theta}}_n + n^{-1/2}\Big(\underset{\sim}{J}^{-1}(\underset{\sim}{\theta}_0)\underset{\sim}{S}_n(\tilde{\underset{\sim}{\theta}}_n) + o_p(1)\Big)$$

Then, under the continuous information condition, we can replace $\underset{\sim}{\theta}_0$ by $\tilde{\underset{\sim}{\theta}}_n$ and therefore get that

$$\hat{\underset{\sim}{\theta}}_n = \tilde{\underset{\sim}{\theta}}_n + n^{-1/2}\underset{\sim}{J}^{-1}(\tilde{\underset{\sim}{\theta}}_n)\underset{\sim}{S}_n(\tilde{\underset{\sim}{\theta}}_n)$$

■

That is, if we have any $\sqrt{n}-$consistent estimator that is not optimal, we can turn it into an optimal estimator through a 1-step transformation process.

STAT4028: Probability and Mathematical Statistics

20. Newton-Raphson Algorithm

20.11.6 Newton-Raphson Algorithm

In this lecture, we descibe the Newton-Raphson algorithm as a method to optimise an intractible log likelihood. We then show that a one-step update using the Newton-Raphson algorithm under certain conditions yields an asymptotically efficient estimator of the parameter. We then describe the process for identifying whether an estimator is RAJN by checking whether is it an asymptotically linear estimator. Finally, we look at estimators based on moments and quantiles and see such methods are asymptotically linear and hence RAJN estimators.

We are now interested in making the link between maximum likelihood estimators and our current framework of LAN. In particular, this relates to the Newton-Raphson method.

Definition 20.301 *Let $p_{\underset{\sim}{\theta}}$ be a density function. Then, the log likelihood is defined to be*

$$\Lambda_n(\underset{\sim}{\theta}) = \sum_{i=1}^{n} log\ p_{\underset{\sim}{\theta}}(\underset{\sim}{X}_i)$$

Unfortunately, for many models in practice, the maximiser of Λ_n is not available in closed form. Therefore, we need to maximize Λ_n numerically using an optimisation algorithm. The Newton-Raphson algorithm is one way to do this.

Proposition 88: Newton-Raphson Method

Let $\Lambda_n(\underset{\sim}{\theta})$ be the log likelihood of the density of interest. The Newton-Raphson method looks for a solution to the score equations

$$\underset{\sim}{\Lambda_n}'(\underset{\sim}{\theta}) = \underset{\sim}{0}$$

where $\underset{\sim}{\Lambda_n}'(\underset{\sim}{\theta})$ is a vector of partial derivatives with respect to $\underset{\sim}{\theta}$.
Then, the method requires an initial starting value $\underset{\sim}{\theta}_0$. Then, our update rule from $\underset{\sim}{\theta}_i$ to $\underset{\sim}{\theta}_{i+1}$ is given by

$$\underset{\sim}{\theta}_{i+1} = \underset{\sim}{\theta}_i - \underset{\sim}{\Lambda_n}''(\underset{\sim}{\theta}_i)^{-1}\underset{\sim}{\Lambda_n}'(\underset{\sim}{\theta}_i)$$

where $\underset{\sim}{\Lambda_n}''$ is the matrix of second order partial derivatives with respect to $\underset{\sim}{\theta}$.

Therefore, in the Newton-Raphson method, we have a score function $\underset{\sim}{\Lambda_n}'(\underset{\sim}{\theta})$ and we seek to find the root of the score function. In the 1D case, we update the rule based on the tangent line at the current point we are at and see where does the tangent line intersect the x-axis.

Now, it is interesting to note that the Newton-Raphson method looks similar to our 1-step update construction of asymptotically efficient estimator except that instead of the inverse of the information matrix, we have the second derivative of the log likelihood.

However, in nice smooth models, the second derivatives of the log likelihood is close to the inverse of the information matrix. This gives us the connection between the Newton-Raphson method and 1-step update rule to construct asymptotically efficient estimators.

Theorem 70: Newton-Raphson One-Step Theorem

Assume that for all $\underset{\sim}{\theta}$ in a neighbourhood of $\underset{\sim}{\theta}_0$

1. $\Lambda_n(.)$ is twice-differentiable;
2. The LAN property holds with the scores
$$\underset{\sim}{S}_n(\underset{\sim}{\theta}) = n^{-1/2}\underset{\sim}{\Lambda}_n{}'(\underset{\sim}{\theta})$$
3. The information matrix $\underset{\sim}{J}(\underset{\sim}{\theta})$ is such that
$$n^{-1}\underset{\sim}{\Lambda}_n{}''(\underset{\sim}{\theta}_0 + n^{-1/2}\underset{\sim}{h}) = -\underset{\sim}{J}(\underset{\sim}{\theta}_0) + o_p(1)$$
uniformly bounded in $\underset{\sim}{h}$.

Then, the regular score assumption holds. Furthermore, the one-step Newton-Raphson update to any RAJN estimator $\underset{\sim}{\tilde{\theta}}_n$ given by the Newton-Raphson update
$$\underset{\sim}{\hat{\theta}}_n = \underset{\sim}{\tilde{\theta}}_n - \underset{\sim}{\Lambda}_n{}''(\underset{\sim}{\tilde{\theta}}_n)^{-1}\underset{\sim}{\Lambda}_n{}'(\underset{\sim}{\tilde{\theta}}_n)$$
is asymptotically efficient.

Remark 20.302 *What we have shown is that we can replace the* $\underset{\sim}{\Lambda}_n{}''$ *with* $-\underset{\sim}{J}(\underset{\sim}{\theta}_0)$*, which gives us the 1-step update for asymptotically efficient estimator formula.*

Remark 20.303 *It can be quite tricky to verify these assumptions. However, if we are able to show the restrictive conditions for LAN property to hold, which is alot easier to show, then all the assumptions required for the Newton-Raphson One-step theorem holds too.*

Note that the Newton-Raphson one step theorem **does not say** that we only need one step to be at the maximum. Instead, it says that for large samples, taking one step gives us an estimator that is asymptotically efficient, that is, asymptotically, it has the smallest variance. Therefore, we don't have to be at the maximum to be asymptotically efficient estimator, we just need to be near the maximum.

20.11.7 Identifying RAJN Estimators

We have introduced the class of regular and asymptotically joint normal with the scores estimators. Now, we are interested in how can we verify that a given estimator is regular or AJN?

We will now introduce a more restrictive class of estimators known as asymptotically linear estimators.

Definition 96: Asymptotically Linear Estimators

Suppose $\underset{\sim}{X}_1, ..., \underset{\sim}{X}_n$ are i.i.d random vectors with common distribution $\mathbb{P}_{\underset{\sim}{\theta}_0}$ for some family of densities $\{p_{\underset{\sim}{\theta}} : \underset{\sim}{\theta} \in \underset{\sim}{\Theta}\}$ for $\underset{\sim}{\Theta} \subseteq \mathbb{R}^d$. An estimator $\tilde{\underset{\sim}{\theta}}_n(\underset{\sim}{X}_1, ..., \underset{\sim}{X}_n)$ is said to be **asymptotically linear (A.L)** if there exists a vector of functions $\underset{\sim}{g}(\underset{\sim}{x}; \underset{\sim}{\theta})$ such that under $\underset{\sim}{\theta}_0$

1. The function has mean zero $\mathbb{E}[\underset{\sim}{g}(\underset{\sim}{x}; \underset{\sim}{\theta}_0)] = 0$;
2. The function has finite variance $\mathbb{E}[\underset{\sim}{g}(\underset{\sim}{x}; \underset{\sim}{\theta}_0)^2] < \infty$;
3. We can express the estimation error of the estimator $\tilde{\underset{\sim}{\theta}}_n$ as a sum of the function

$$\underset{\sim}{Y}_n = \sqrt{n}(\tilde{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underset{\sim}{g}(\underset{\sim}{x}_i; \underset{\sim}{\theta}_0) + o_p(1)$$

where $\underset{\sim}{g}(\underset{\sim}{x}; \underset{\sim}{\theta}_0)$ is called the **influence function** of the estimator.

Remark 20.304 *In condition 3, we can interpret the asymptotic behaviour of the estimator is dominated by the sum of the influence function.*

Now, suppose that $\tilde{\underset{\sim}{\theta}}_n$ is an asymptotically linear estimator. Then, if the LAN property holds and the score vector is of the form

$$\underset{\sim}{S}_n(\underset{\sim}{\theta}_0) = n^{-1/2} \sum_{i=1}^{n} \underset{\sim}{\ell}_{\underset{\sim}{\theta}_0}{}^{\circ}(\underset{\sim}{X}_i)$$

then, we have that

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ \underset{\sim}{S}_n \end{pmatrix} = n^{-1/2} \sum_{i=1}^{n} \begin{pmatrix} \underset{\sim}{g}(\underset{\sim}{x}; \underset{\sim}{\theta}_0) \\ \underset{\sim}{\ell}^{\circ}_{\underset{\sim}{\theta}_0}(\underset{\sim}{x}_i; \underset{\sim}{\theta}_0) \end{pmatrix} + o_p(1)$$

is an asymptotically normal random-vector using the multivariate central limit theorem. That is, $\tilde{\underset{\sim}{\theta}}_n$ is a RAJN estimator. Furthermore, it is important to note that a statistic that is asymptotically linear satisfies Scores version of Le Cam's third lemma.

From this, we have the big takeaway.

Proposition 89: A.L. estimators are RAJN

An estimator that is asymptotically linear is a RAJN estimator.

Proposition 20.305 *Method of moment estimators are RAJN estimators.*

20.11.8 Estimators based on moments and quantiles

Many estimators can be written as smooth functions of sample quantiles or sample averages. We can use the delta method in probability and other techniques to show that estimators based on averages, under smoothness conditions, are RAJN by showing that the estimator is asymptotically linear.

We can use similar techniques that estimators based on quantiles are also regular. However, it can be difficult to show that estimators based on quantiles are AJN. Therefore, we will introduce tools to help us show that quantiles are asymptotically linear.

These tools are revolve around the osallations of the uniform empirical CDF.

Definition 20.306 *(Empirical CDF). Suppose that $U_1, ..., U_n$ are i.i.d $U(0,1)$ random variables. Define the empirical CDF as*

$$G_n(u) = \frac{1}{n}\sum_{i=1}^{n} 1\{U_i \leq u\}.$$

Definition 97: Empirical Process

Let $G_n(u)$ be the empirical CDF. Then, the **empirical process** is defined as

$$H_n(u) = \sqrt{n}(G_n(u) - u).$$

Remark 20.307 *The empirical process is a centered version of the empirical CDF.*

The empirical process will be our influence function to show that quantile estimators are asymptotically linear.

The empirical CDF $G_n(u)$ can be interpreted as a step function with jumps of height $\frac{1}{n}$ which gets smaller and smoother as $n \to \infty$.

Lemma 20.308 *Let G_n be the empirical CDF. Then, for a fixed u, we have that*

$$n.G_n(u) = \sum_{i=1}^{n} 1\{U_i \leq u\} \sim Bin(n, u).$$

Corollary 20.309 *Let G_n be the empirical CDF and H_n the corresponding empirical process. Then*

$$\mathbb{E}[H_n(u)] = 0$$

$$Var[H_n(u)] = u(1-u)$$

not depending on n.

As the empirical CDF gets smoother as we increase n, we are now interested in measuring the smoothness of function. We define a measure of how to measure the smoothness of a function.

Definition 98: Modulus of continuity (osallation)

The modulus of continuity is a measure of smoothness for a given empirical process H_n defined as

$$\omega_n(\Delta) = \sup_{|u-v|=\Delta} |H_n(u) - H_n(v)|$$

for a fixed window of width Δ.

Remark 20.310 *The intuition behind the modulus of continuity is that if we have a fixed window and we move it over the function, what is the biggest difference that we see? In particular, if our function is smooth. the modulus of continuity should tend to zero as the window width* $\Delta \to 0$.

However, it may be tricky to analyse our empirical CDF as it is a random function and it has steps. Therefore, we let both $n \to \infty$ but shrink the window size $\Delta \to 0$ at a certain rate.

Proposition 20.311 *For any* $0 < C < \infty$,

$$\omega_n(Cn^{-1/2}) \xrightarrow{p} 0.$$

That is, the modulus of continuity of the sequence $Cn^{-1/2}$ *tends to probability in zero if we let the* Δ_n *shrink to zero at the rate of* $n^{-1/2}$.

Corollary 20.312 *Define the remainder of 2 nearby values of the empirical CDFs*

$$R_n(h;u) = \sqrt{n}\Big[G_n(u + n^{-1/2}h) - G_n(u)\Big] - h.$$

For any fixed u, we have that

$$\sup_{|h|\leq M} |R_n(h;u)| \xrightarrow{p} 0.$$

That is,

$$\sqrt{n}(G_n(u + n^{-1/2}h) - G_n(u)) = h + O_p(1)$$

uniformly in bounded h.

We now state that we can approximate an order statistic by the value of the empirical process.

Theorem 71: Rescaled order statistic

Let $U_{(A_n)}$ be an order statistic. Then

$$\sqrt{n}\Big[U_{(A_n)} - p\Big] = H_n(p) + O_p(1).$$

We can now state that the order statistic is an asymptotically linear estimator.

Corollary 20.313 *If* $\sqrt{n}(\frac{a_n}{n+1} - p) \to 0$, *then the order statistic* $U_{(a_n)}$ *is an A.L. estimator of p with influence function*

$$g(u) = \begin{cases} p - 1 & u \leq p \\ p & u > p \end{cases}$$

Remark 20.314 *If we do a transformation to another distribution, the order statistic from any other distribution is an A.L. estimator of that population's p-th quantile using the delta method.*

Now, we will state the Delta method. In particular, the Delta method consists of using a Taylor expansion to approximate the random vector $f(X_n)$ with a polynomial.

Proposition 20.315 *(Delta method). If we have that* $\sqrt{n}(X_n - c) \xrightarrow{d} Y$ *then we have that*

$$\sqrt{n}(f(X_n) - f(c)) \xrightarrow{d} f^{'}(c)Y.$$

Proposition 90: Delta Method in probability

Let $X_n \xrightarrow{p} C$ and $\lim_{x \to C} g(x) = \ell$. Then, we have that

$$g(X_n) \xrightarrow{p} \ell.$$

STAT4028: Probability and Mathematical Statistics

21. Efficient Influence Function

21.11.9 Efficient Influence Function

In this lecture, we analyse asymptotically linear estimators through their influence functions. From this, we can define what an efficient influence function is. We then introduce parameter estimation when we have nuisance parameters involved in the model. We then revisit previous results on Scores Version of Le Cam's third lemma involving nuisance parameters. From this, we can see that RAJN estimators are asymptotically independent of the nuisance parameters. Additionally, we can also see that asymptotically efficient estimators involving nuisance parameters can be constructed through a 1-step rule.

We have shown that an estimator that is asymptotically linear is a way to show that an estimator is a RAJN estimator. In particular, an estimator is asymptotically linear if it is dominated by the sum of influence functions. Hence, understanding how the influence functions behave gives insights into the asymptotically linear estimators.

First, let us assume that the LAN property holds at $\underset{\sim}{\theta}_0$ with the scores function $\underset{\sim}{S}_n(\underset{\sim}{\theta}_0) = n^{-1/2}\sum_{i=1}^n \underset{\sim}{\ell}^\circ_{\underset{\sim}{\theta}_0}$ and information matrix $\underset{\sim}{J}(\underset{\sim}{\theta}_0)$. Furthermore, assume that we have an estimator $\hat{\underset{\sim}{\theta}}_n$ that satisfies

$$\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) = \underset{\sim}{J}^{-1}(\underset{\sim}{\theta}_0)\underset{\sim}{S}_n(\underset{\sim}{\theta}) + o_p(1).$$

Then, we can conclude that $\hat{\underset{\sim}{\theta}}_n$ is asymptotically efficient estimator, and also a RAJN estimator at $\underset{\sim}{\theta}_0$. Furthermore, it is also asymptotically linear estimator.

Definition 99: Efficient Influence Function

We define the efficient influence function

$$g = \underset{\sim}{J}^{-1}\underset{\sim}{\ell}^\circ_{\underset{\sim}{\theta}_0}$$

for estimating $\underset{\sim}{\theta}$ at $\underset{\sim}{\theta}_0$.

Proof: Recall condition 3 needed for the influence function $\underset{\sim}{g}(\cdot, \underset{\sim}{\theta})$ that

$$\underset{\sim}{Y}_n = \sqrt{n}(\tilde{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \underset{\sim}{g}(\underset{\sim}{X}_i; \underset{\sim}{\theta}_0) + o_p(1)$$

However, if $\hat{\underset{\sim}{\theta}}_n$ is asymptotically efficient

$$\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) = \underset{\sim}{J}^{-1}(\underset{\sim}{\theta}_0)\underset{\sim}{S}_n(\underset{\sim}{\theta}) + o_p(1)$$

we can conclude that

$$\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) = \underset{\sim}{J}^{-1}(\underset{\sim}{\theta}_0)\underset{\sim}{S}_n(\underset{\sim}{\theta}) + o_p(1) = \underset{\sim}{J}^{-1}n^{-1/2}\sum_{i=1}^n \underset{\sim}{\ell}^\circ_{\underset{\sim}{\theta}_0} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \underset{\sim}{g}(\underset{\sim}{X}_i; \underset{\sim}{\theta}_0) + o_p(1)$$

■

Once we know the efficient influence function, we can figure out whether is an estimator asymptotically efficient.

Theorem 72: Test for asymptotically efficient estimator

Suppose the efficient influence function is given by $\underset{\sim}{\tilde{\ell}}_{\underset{\sim}{\theta_0}}(\underset{\sim}{X})$. Then suppose we can express our estimator of interest $\underset{\sim}{\hat{\theta}_n}$ in the form

$$\sqrt{n}\left(\underset{\sim}{\hat{\theta}_n} - \underset{\sim}{\theta_0}\right) = n^{-1/2}\sum_{i=1}^{n} \underset{\sim}{\tilde{\ell}}_{\underset{\sim}{\theta_0}}(\underset{\sim}{X_i}) + o_p(1).$$

Then, the estimator $\underset{\sim}{\hat{\theta}_n}$ is asymptotically efficient.

21.11.10 Scalar Nuisance Parameter Model

Up until now, we have been interested in estimating the full parameter of the model. What if we are only interested in estimating a subvector? That is, we now have nuisance parameters in our model which we do not care about but must account for.

We now describe the basic setup involving scalars for the parameter of interest and nuisance parameter. Consider a model $\{\mathbb{P}_{\underset{\sim}{\gamma}}\}$ where the full parameter vector $\underset{\sim}{\gamma} = \binom{\theta}{\eta}$ has **dimension 2**. Furthermore, assume that the LAN property holds at $\underset{\sim}{\gamma_0} = \binom{\theta_0}{\eta_0}$ with scores vector

$$S_n = S = \begin{pmatrix} S_\theta \\ S_\eta \end{pmatrix} = n^{-1/2}\sum_{i=1}^{n} \begin{pmatrix} \ell^\circ_\theta(\underset{\sim}{X_i}) \\ \ell^\circ_\eta(\underset{\sim}{X_i}) \end{pmatrix}$$

and the information matrix

$$J = \begin{pmatrix} J_{\theta\theta} & J_{\theta\eta} \\ J_{\eta\theta} & J_{\eta\eta} \end{pmatrix}$$

Proposition 21.316 *(Efficient Influence Function for scalar parameters). The efficient influence function for estimating $\underset{\sim}{\gamma}$ at $\underset{\sim}{\gamma_0}$ is*

$$J^{-1}\underset{\sim}{\ell}^\circ_{\underset{\sim}{\gamma}}.$$

Recall that explicitly, the inverse of the information matrix is given by

$$J^{-1} = \frac{1}{J_{\theta\theta}J_{\eta\eta} - J_{\theta\eta}^2}\begin{bmatrix} J_{\eta\eta} & -J_{\theta\eta} \\ -J_{\eta\theta} & -J_{\theta\theta} \end{bmatrix}$$

We will now explicitly write out the efficient influence function. We have that the efficient influence function written out is

$$J^{-1}\underset{\sim}{\ell}^\circ_{\underset{\sim}{\gamma}} = \begin{bmatrix} \frac{J_{\eta\eta}\ell^\circ_\theta - J_{\theta\eta}\ell^\circ_\eta}{J_{\theta\theta}J_{\eta\eta} - J_{\theta\eta}^2} & \frac{J_{\theta\theta}\ell^\circ_\eta - J_{\theta\eta}\ell^\circ_\theta}{J_{\theta\theta}J_{\eta\eta} - J_{\theta\eta}^2} \end{bmatrix}^T$$

However, we are only interested in estimating the parameter of interest θ. Therefore, we only care about the first component of the efficient influence function

$$\frac{J_{\eta\eta}\ell_\theta^\circ - J_{\theta\eta}\ell_\eta^\circ}{J_{\theta\theta}J_{\eta\eta} - J_{\theta\eta}^2}$$

Now, multiplying the efficient influence function for $\underset{\sim}{\theta}$ by $J_{\eta\eta}^{-1}$, we get the following proposition.

Proposition 91: Efficient influence function for scalar parameter of interest

The efficient influence function for estimating the parameter of interest θ is given by

$$\frac{\ell_{\theta_0}^{\circ *}}{J_{\theta\theta}^*} = \frac{\ell_\theta^\circ - J_{\theta\eta}J_{\eta\eta}^{-1}\ell_\eta^\circ}{J_{\theta\theta} - J_{\theta\eta}J_{\eta\eta}^{-1}J_{\eta\theta}}$$

Remark 21.317 *Looking at the efficient influence function, if the nuisance parameter η was known, then we would use ℓ_θ° and $J_{\theta\eta}$. However, if η is unknown, we have a loss of information and therefore we replace $J_{\theta\theta}$ with $J_{\theta\theta}^*$, which is now a reduction in the information. Now, as the asymptotic variance is the inverse of the information, this means that the asymptotical variance increases if η is unknown.*

We can then define important quantities from the efficient influence function.

Definition 100: Scalar Effective Score and scalar effective information

Let $\frac{\ell_{\theta_0}^{\circ *}}{J_{\theta\theta}^*}$ be the efficient influence function. We define the effective score as

$$S_\theta^* = n^{-1/2}\sum_{i=1}^n \ell_\theta^{\circ *}(\underset{\sim}{X}_i).$$

We call $J_{\theta\theta}^*$ the effective information.

Remark 21.318 *The effective information can be thought of as the asymptotic variance of the effective score.*

21.11.11 General Setting Nuisance Parameter

Now, in contrast to our last section, we are in the general case where the dimension $d = d_\theta + d_\eta \geq 2$. We now have the same quantities defined in the last section but in terms of matrices.

Definition 101: Efficient influence function for parameter of interest

The efficient influence function for estimating the parameter of interest $\underset{\sim}{\theta}$ is given by

$$(\underset{\sim\theta\theta}{J}^*)^{-1}\underset{\sim}{\ell}_{\theta_0}^{\circ *}$$

Definition 102: Effective Score and effective information matrix

Let $(\underset{\sim}{J^*}_{\theta\theta})^{-1}\underset{\sim}{\ell_{\theta_0}}^{\circ *}$ be the efficient influence function. We define the effective score as

$$\underset{\sim}{S_\theta}^* = n^{-1/2}\sum_{i=1}^{n}\underset{\sim}{\ell}^{\circ *}_{\underset{\sim}{\theta}}(\underset{\sim}{X_i}).$$

We call $\underset{\sim}{J^*}_{\theta\theta}$ the effective information matrix.

Hence, everything follows from the last section by using the block-matrix inverse formulae.

Proposition 21.319 *(Block-matrix inverse formulae). Suppose that $\underset{\sim}{J}$ is an invertible square matrix*

$$\underset{\sim}{J} = \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix}$$

where J_{11}, J_{22} are invertible block matrices. Let

$$\begin{cases} J_{11.2} = J_{11} - J_{12}J_{22}^{-1}J_{21} \\ J_{22.1} = J_{22} - J_{21}J_{11}^{-1}J_{12} \end{cases}$$

Then, the inverse $\underset{\sim}{J}^{-1}$ is given by

$$\underset{\sim}{J}^{-1} = \begin{bmatrix} J_{11.2}^{-1} & -J_{11.2}^{-1}J_{12}J_{22}^{-1} \\ -J_{22.1}^{-1}J_{21}J_{11}^{-1} & J_{22.1}^{-1} \end{bmatrix}$$

Proof: It is simple to verify that

$$\underset{\sim}{J}^{-1}\underset{\sim}{J} = \mathbb{I}$$

where $\mathbb{I}$ is the identity matrix. ■

Proposition 92: Effective Information

The inverse of the effective information is given by

$$(\underset{\sim}{J^*}_{\theta\theta})^{-1} = \left(J_{\theta\theta} - J_{\theta\eta}J_{\eta\eta}^{-1}J_{\eta\theta}\right)^{-1}$$

where $J_{\theta\theta}, J_{\theta\eta}, J_{\eta\theta}$ are terms from the information matrix $\underset{\sim}{J}$.

21.11.12 Regularity in the presence of nuisance parameters

We are now interested in defining something analogous to the decomposition theorem when estimating $\underset{\sim}{\theta}$ in the presence of nuisance parameters $\underset{\sim}{\eta}$. Before that, we first need to define regularity in the presence of nuisance parameters $\underset{\sim}{\eta}$..

Definition 103: Regular Estimators in the presence of nuisance parameters

A regular estimator is one where the estimation error has the same limiting distribution under the nearby sequence

$$\begin{pmatrix} \theta_0 \\ \underset{\sim}{\eta_0} \end{pmatrix} + n^{-1/2} \begin{pmatrix} h_\theta \\ \underset{\sim}{h_\eta} \end{pmatrix}$$

as it has under the fixed center

$$\begin{pmatrix} \theta_0 \\ \underset{\sim}{\eta_0} \end{pmatrix}$$

Now, after this definition of regularity, we can then apply Scores version of Le Cam's third lemma to get the following proposition.

Proposition 93: RAJN Estimators are independent of nuisance parameters

Suppse that the LAN property holds at the true parameter value $\underset{\sim}{\gamma_0} = \binom{\theta_0}{\underset{\sim}{\eta_0}}$. Suppose that the estimator $\underset{\sim}{\tilde{\theta}_n}$ is a RAJN estimator. That is, the estimator $\underset{\sim}{\tilde{\theta}_n}$ is regular at $\underset{\sim}{\theta_0}$ and that the rescaled estimation error $Y_n = \sqrt{n}(\underset{\sim}{\tilde{\theta}_n} - \underset{\sim}{\theta_0})$ is such that

$$\begin{pmatrix} Y_n \\ \underset{\sim}{S_n} \end{pmatrix} = \begin{pmatrix} Y_n \\ \underset{\sim}{S_\theta} \\ \underset{\sim}{S_\eta} \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} 0 \\ \underset{\sim}{0} \\ \underset{\sim}{0} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \Sigma_{Y\theta} & \Sigma_{Y\eta} \\ \Sigma_{\theta Y} & J_{\theta\theta} & J_{\theta\eta} \\ \Sigma_{\eta Y} & J_{\eta\theta} & J_{\eta\eta} \end{pmatrix} \Big) = \mathcal{N}\Big(\begin{pmatrix} 0 \\ \underset{\sim}{0} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \Sigma_{YS} \\ \Sigma_{SY} & J \end{pmatrix} \Big)$$

which means that the estimator $\underset{\sim}{\tilde{\theta}_n}$ is asymptotically jointly normal with the scores vector.
Then, the cross covariance of the estimation error with the scores vector for the parameters of interest $\underset{\sim}{\theta}$ is given by

$$\Sigma_{Y\theta} = \underset{\sim}{I}$$

and the cross covariance of the estimation error with the scores vector for the nuisance parameters $\underset{\sim}{\eta}$ is given by

$$\Sigma_{Y\eta} = \underset{\sim}{0}$$

Proof: By the Scores version of Le Cam's third lemma, we have that under the nearby alternative

$$\begin{pmatrix} \theta_0 \\ \underset{\sim}{\eta_0} \end{pmatrix} + n^{-1/2} \begin{pmatrix} h_\theta \\ \underset{\sim}{h_\eta} \end{pmatrix}$$

that the estimation error under the true value $Y_n = \sqrt{n}(\underset{\sim}{\tilde{\theta}_n} - \underset{\sim}{\theta_0})$ has the distribution

$$\underset{\sim}{Y_n} \xrightarrow{d} \mathcal{N}\Big(\Sigma_{YS}\underset{\sim}{h}, \Sigma_Y\Big) = \mathcal{N}\Big(\Sigma_{Y\theta}\underset{\sim}{h_\theta} + \Sigma_{Y\eta}\underset{\sim}{h_\eta}, \Sigma_Y\Big)$$

where we decomposed the mean into the parameter of interest and the nuisance parameter. Then, we have

that the estimation error under the nearby sequence $\underset{\sim}{\gamma}_n$

$$\sqrt{n}(\underset{\sim}{\tilde{\theta}}_n - \underset{\sim}{\theta}_n) = \underset{\sim}{Y}_n - \underset{\sim}{h}_\theta \xrightarrow{d} \mathcal{N}\Big([\Sigma_{Y\theta} - \underset{\sim}{I}]\underset{\sim}{h}_\theta + \Sigma_{Y\eta}\underset{\sim}{h}_\eta, \Sigma_Y\Big)$$

However, by assumption, the estimator $\underset{\sim}{\tilde{\theta}}_n$ was assumed to be RAJN. Therefore, in order for the limiting distribution to be free of $\underset{\sim}{h}$, we therefore require that

$$\Sigma_{Y\theta} = \underset{\sim}{I}$$

$$\Sigma_{Y\eta} = \underset{\sim}{0}$$

■

Remark 21.320 *Any RAJN estimator is asymptotically independent of the nuisance scores* $\underset{\sim}{S}_\eta$. *That is, the estimator is unaffected by changing* η *whilst keeping* θ *fixed.*

We now revisit some concepts we have seen already.

Proposition 21.321 *Any asymptotically efficient estimator* $\underset{\sim}{\hat{\theta}}_n$ *can be written in the form*

$$\sqrt{n}(\underset{\sim}{\hat{\theta}}_n - \underset{\sim}{\theta}_0) = (J^*_{\theta\theta})^{-1}\underset{\sim}{S}_{n\theta}{}^* + o_p(1)$$

Additionally, we can also see that the effective scores $S^*_{n\theta}$ are asymptotically independent of the nuisance scores $\underset{\sim}{S}_\eta$. That is, the effective scores are the precise linear combinations of the scores $\underset{\sim}{S}_\theta$ and $\underset{\sim}{S}_\eta$ which is asymptotically uncorrelated with the score of the nuisance parameter $\underset{\sim}{S}_\eta$.

Proposition 94: Effective score is the linear combination of the scores

The effective score S^*_θ is given by the linear combination of scores of the parameter of interest $\underset{\sim}{S}_\theta$ and of the nuisance parameter $\underset{\sim}{S}_\eta$

$$S^*_\theta = \underset{\sim}{S}_\theta - \underset{\sim}{J}_{\theta\eta}\underset{\sim}{J}_{\eta\eta}{}^{-1}\underset{\sim}{S}_\eta$$

Remark 21.322 *We can interpret the effective score as the residuals from regressing the scores of the parameter of interest on the nuisance parameter scores. Furthermore, we can interpret the effective information as the asymptotic variance of the effective score.*

Remark 21.323 *When the co-information matrix* $J_{\theta\eta} = 0$, *we have that the effective scores is identical to the ordinary scores. That is,* η *does not matter and the performance of estimates and tests for* θ *will be the same regardless if* η *is known or not. This is described as that we can* ***adapt*** *for* η.

Corollary 21.324 *(Limiting distribution of the effective score) Assuming LAN, we have that*

$$S_\theta^* \xrightarrow{d} \mathcal{N}(J_{\theta\theta}^* h_\theta, J_{\theta\theta}^*)$$

Proposition 21.325 *(Effective score is asymptotically uncorrelated). The asymptotic covariance of the effective score and score of the nuisance parameter is zero*

$$\mathbb{E}[\underset{\sim}{S_\theta}^* \underset{\sim}{S_\eta}^T] \to \underset{\sim}{0}.$$

where $J_{\theta\theta}^*$ is the effective information.

The final remark we make is that asymptotically efficient estimators can be constructed as before by updating a $\sqrt{n}-$consistent estimator of the whole vector $\underset{\sim}{\gamma}/$ In fact, only the $\underset{\sim}{\theta}$ component of the estimators are updated. Hence, we have the same process of constructing asymptotically efficient estimators.

STAT4028: Probability and Mathematical Statistics

22. Estimating Smooth functions of parameters

22.11.13 Estimating Smooth functions of parameters

In this lecture, we now generalise the previous lectures on estimation of parameters. In particular, we want to estimate smooth function of parameters. We define regular estimators, asymptotically efficient estimators, and orthogonal decomposition theorem for smooth function of parameters.

We have seen two forms of estimations of the parameter. The first is where are interested in estimating the full parameter vector $\underset{\sim}{\theta}$. The second case is when we have a nuisance parameter involved $\underset{\sim}{\eta}$.

Now, we are interested in a generalisation. Suppose we have a parametetric model $\{P_{\underset{\sim}{\gamma}}\}$, where we are now estimating a smooth function $\underset{\sim}{\theta}(\underset{\sim}{\gamma})$ of dimension d_θ of the d-dimensional parameter $\underset{\sim}{\gamma}$.

We now give the definition for a regular estimator of a smooth function of the parameters.

Definition 104: Regular Estimator

A regular estimator $\tilde{\underset{\sim}{\theta}_n}$ at $\underset{\sim}{\gamma}_0$ is such that for any nearby sequence $\underset{\sim}{\gamma}_n = \underset{\sim}{\gamma}_0 + n^{-1/2}\underset{\sim}{h}$, the estimation error satisfies

$$\sqrt{n}(\tilde{\underset{\sim}{\theta}_n} - \underset{\sim}{\theta}(\underset{\sim}{\gamma}_n)) \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{\Sigma}(\underset{\sim}{\gamma}_0))$$

Remark 22.326 *As always, a regular estimator is one where the distribution of the estimation error does not depend on $\underset{\sim}{h}$ for any nearby sequence. The only thing that changes is the covariance matrix, which only depends on the centered true value.*

We now repeat a procedure that we have seen when estimating nuisance parameters.

Suppose that the LAN property holds at the trule value $\underset{\sim}{\gamma}_0$ with the scores vector $\underset{\sim}{S}_n$ and information matrix $\underset{\sim}{J}$. Suppose further, that the rescaled estimation error at $\underset{\sim}{\gamma}_0$

$$\underset{\sim}{Y}_n = \sqrt{n}(\tilde{\underset{\sim}{\theta}_n} - \underset{\sim}{\theta}(\underset{\sim}{\gamma}_0))$$

satisfies the LAN# property. That is, the rescaled estimation error is asymptotically joint normal with the scores vector

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ \underset{\sim}{S}_n \end{pmatrix} = \mathcal{N}\Big(\begin{pmatrix} \underset{\sim}{0} \\ \underset{\sim}{0} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \Sigma_{YS} \\ \Sigma_{SY} & J \end{pmatrix} \Big)$$

under the true value $\underset{\sim}{\gamma}_0$. Then, under the nearby sequence $\underset{\sim}{\gamma}_n = \underset{\sim}{\gamma}_0 + n^{-1/2}\underset{\sim}{h}$, we can apply Scores version of Le Cam's third lemma to get

$$\begin{pmatrix} \underset{\sim}{Y}_n \\ \underset{\sim}{S}_n \end{pmatrix} = \mathcal{N}\Big(\begin{pmatrix} \Sigma_{YS}\underset{\sim}{h} \\ \underset{\sim}{J}\underset{\sim}{h} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \Sigma_{YS} \\ \Sigma_{SY} & J \end{pmatrix} \Big)$$

Note that what we just did was for the estimation error at the true value $\underset{\sim}{\gamma}_0$. We now write the general estimation error for a nearby sequence

$$\sqrt{n}(\tilde{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}(\underset{\sim}{\gamma}_n)) = \sqrt{n}[\tilde{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}(\underset{\sim}{\gamma}_0) - \underset{\sim}{\theta}^\circ(\underset{\sim}{\gamma}_n{}^*)n^{-1/2}\underset{\sim}{h}] = \underset{\sim}{Y}_n - \underset{\sim}{\theta}_0^\circ \underset{\sim}{h} + o(1)$$

where we applied a 1-term Taylor expansion and that $\underset{\sim}{\theta}^\circ$ is the d_θ by d matrix of partial derivatives (the Jacobian matrix) and $\underset{\sim}{\theta}_0^\circ = \underset{\sim}{\theta}^\circ(\underset{\sim}{\gamma}_0)$ is the Jacobian evaluated at the true parameter of interest.

Then, by Le Cam's third lemma, we see that the general estimation error under $\underset{\sim}{\gamma}_n$ is

$$\sqrt{n}[\tilde{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}(\underset{\sim}{\gamma}_n)] \xrightarrow{d} \mathcal{N}(\Sigma_{YS}\underset{\sim}{h} - \underset{\sim 0}{\theta}^\circ \underset{\sim}{h}, \Sigma_Y)$$

However, as the estimator was assumed to be regular, we require this distribution to be free of $\underset{\sim}{h}$, which only occurs if the cross-covariance matrix in the mean is zero, which is when $\Sigma_{YS} = \underset{\sim 0}{\theta}^\circ$.

Proposition 95: Covariance of general estimation error and score

Suppose that $\tilde{\underset{\sim}{\theta}}$ is a RAJN estimator. Then, under a nearby sequence, we have that the cross-covariance of the general estimation error Y and the score vector is the Jacobian of the transformation of the parameters evaluated at the true value

$$\Sigma_{YS} = \underset{\sim 0}{\theta}^\circ = \underset{\sim}{\theta}^\circ(\underset{\sim}{\gamma}_0)$$

Remark 22.327 *Note that this generalises our earlier 2 results when deducing the cross-covariance matrix.*

In the case where we are estimating the whole vector $\underset{\sim}{\theta}(\underset{\sim}{\gamma}) = \underset{\sim}{\gamma}$*, the cross-covariance matrix was* $\Sigma_{YS} = \underset{\sim}{I}$.

In the case where we were only estimating the parameters of interest and ignoring the nuisance parameters, that is, $\underset{\sim}{\theta}(\underset{\sim}{\gamma})$ *was the first* d_θ *elements of* $\underset{\sim}{\gamma}$*, the cross covariance was given by*

$$\Sigma_{YS} = (\Sigma_{Y\theta}, \Sigma_{Y\eta}) = (\underset{\sim}{I}, \underset{\sim}{0})$$

where $\Sigma_{Y\theta}, \Sigma_{Y\eta}$ *was the cross-covariance with the parameter of interest and nuisance parameter respectively.*

Hence, all our earlier results on the cross-covariance actually depends on the Jacobian of the transformation.

This will now lead to generalised definitions of asymptotically efficient estimators and orthogonal decomposition theorem.

Proposition 96: Orthogonal Decomposition Theorem

Let $\underset{\sim}{\tilde{\theta}_n}$ be a RAJN estimator of $\underset{\sim}{\theta}(\underset{\sim}{\gamma})$. Then, define

$$\underset{\sim}{Z_n} = \sqrt{n}(\underset{\sim}{\tilde{\theta}_n} - \underset{\sim}{\theta}(\underset{\sim}{\gamma_0})) - \underset{\sim}{\theta}^{\circ}\underset{\sim}{J}^{-1}\underset{\sim}{S} = \underset{\sim}{Y_n} - \underset{\sim}{\theta}^{\circ}\underset{\sim}{J}^{-1}\underset{\sim}{S}$$

where $\underset{\sim}{Y_n} = \underset{\sim}{\tilde{\theta}_n} - \underset{\sim}{\theta}(\underset{\sim}{\gamma_0})$ is the estimation error. Then, the limiting distribution under the null $\underset{\sim}{\gamma_0}$ is given by

$$\begin{bmatrix} \underset{\sim}{Z_n} \\ \underset{\sim}{S_n} \end{bmatrix} \xrightarrow{d} \mathcal{N}\Big(\underset{\sim}{0}, \begin{bmatrix} \Sigma_Y - \underset{\sim}{\theta}^{\circ}\underset{\sim}{J}^{-1}\underset{\sim}{\theta}^{\circ T} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{J} \end{bmatrix}\Big)$$

Therefore, any estimator is made up of two parts

$$\underset{\sim}{Y_n} = \underset{\sim}{\theta}^{\circ}\underset{\sim}{J}^{-1}\underset{\sim}{S} + \underset{\sim}{Z_n}$$

Then, the best that we can do is for $\underset{\sim}{Z_n} = \underset{\sim}{0}$ which occurs when the covariance matrix

$$\Sigma_Y = \underset{\sim}{\theta}^{\circ}\underset{\sim}{J}^{-1}\underset{\sim}{\theta}^{\circ T}$$

Remark 22.328 *Here, $\underset{\sim}{Z_n}$ is AJN with the scores as the estimator was assumed to be AJN.*

Hence, if we choose an estimator $\underset{\sim}{\tilde{\theta}_n}$ which has a covariance matrix $\Sigma_Y = \underset{\sim}{\theta}^{\circ}\underset{\sim}{J}^{-1}\underset{\sim}{\theta}^{\circ T}$, that is the best that we can do. This brings us to our definition of asymptotically efficeint estimators.

Definition 105: Asymptotically Efficient Estimator

An estimator $\underset{\sim}{\hat{\theta}_n}$ is asymptotically efficient if we can write the estimation error

$$\sqrt{n}(\underset{\sim}{\hat{\theta}_n} - \underset{\sim}{\theta}(\underset{\sim}{\gamma_0})) = \underset{\sim}{\theta}^{\circ}\underset{\sim}{J}^{-1}\underset{\sim}{S} + o_p(1)$$

under the true value $\underset{\sim}{\gamma_0}$.

Remark 22.329 *Note that this form is what we stated in the orthogonal decomposition theorem except that the extra estimation error $\underset{\sim}{Z_n}$ now converges to 0 in probability.*

Now recall, that we can view asymptotically efficient estimators from the perspective of it being an asymptotically linear function and therefore it has an associated efficient influence function.

Definition 106: Efficient Influence Function, Effective Score and information

Let $\underset{\sim}{\hat{\theta}_n}$ be an asymptotically linear estimator. Then, the efficient influence function is

$$\underset{\sim}{\theta}^{\circ}\underset{\sim}{J}^{-1}\underset{\sim}{\underset{\sim}{\ell}^{\circ}}_{\gamma}$$

The asymptotic covariance matrix of $\underset{\sim}{\theta}^{\circ}\underset{\sim}{J}^{-1}\underset{\sim}{S}$ is given by

$$\underset{\sim}{\theta}^{\circ}\underset{\sim}{J}^{-1}\underset{\sim}{\theta}^{\circ T}$$

The effective score is given by

$$S_n^* = J^* \underset{\sim}{\theta}^{\circ}\underset{\sim}{J}^{-1}\underset{\sim}{S}$$

where the effective information is given by

$$J^* = (\underset{\sim}{\theta}^{\circ}\underset{\sim}{J}^{-1}\underset{\sim}{\theta}^{\circ T})^{-1}$$

Remark 22.330 *Recall that the effective scores is the score associated to the parameters of interest. These 2 forms of the effective score and effective information generalises from yesterday.*

We are now interested in constructing asymptotically efficeint estimators. We can in fact construct them by applying smooth transformations of efficient estimators of the original parameter.

Proposition 97: Smooth function of efficient estimators are efficient

Suppose that $\underset{\sim}{\hat{\gamma}_n}$ is an asymptotically efficient estimator for the full parameter vector $\underset{\sim}{\gamma}$. Then, by the delta method in probability, the new estimator defined by taking a smooth function $\underset{\sim}{\theta}$ of our estimator, denoted by $\underset{\sim}{\tilde{\theta}_n} = \underset{\sim}{\theta}(\underset{\sim}{\hat{\gamma}_n})$ is also an asymptotically efficient estimator of $\underset{\sim}{\theta}(\underset{\sim}{\gamma})$ at $\underset{\sim}{\gamma_0}$.

Hence, taking a smooth function of an efficient estimator gives you an efficient estimator of a smooth function.

However, we can now incorporate the ideas of one-step estimators whereby we do not necessarily need to start off with an asymptotically efficient estimator of the whole vector. In fact, just a $\sqrt{n}-$consistent estimator is enough.

Proposition 98: Constructing efficient estimators through one step update

Suppose that we have the following conditions holding at $\underset{\sim}{\gamma_0}$

1. LAN property
2. Regular scores
3. Continuous information

Furthermore, suppose that

1. The Jacobian $\underset{\sim}{\theta}^{\circ}(\underset{\sim}{\gamma})$ is continuous at $\underset{\sim}{\gamma_0}$
2. $\underset{\sim}{\tilde{\gamma}_n}$ is a $\sqrt{n}-$consistent estimator of $\underset{\sim}{\gamma}$ at $\underset{\sim}{\gamma_0}$

Then, the one-step estimator given by

$$\underset{\sim}{\hat{\theta}_n} = \underset{\sim}{\theta}(\underset{\sim}{\tilde{\gamma}_n}) + n^{-1/2}\, \underset{\sim}{\theta}^{\circ}\, (\underset{\sim}{\tilde{\gamma}}) \underset{\sim}{J}^{-1}(\underset{\sim}{\tilde{\gamma}_n}) \underset{\sim}{S}_n(\underset{\sim}{\tilde{\gamma}_n})$$

is an asymptotically efficient estimator for $\underset{\sim}{\theta}(\underset{\sim}{\gamma})$ at $\underset{\sim}{\gamma_0}$.

Remark 22.331 *We can think of the one step update as that we take a smooth function of a $\sqrt{n}-$consistent estimator and then we apply a correction term to it. This is a generalisation of the one-step estimators we have seen so far.*

In fact, there is an even nicer way of expressing one-step estimators.

Proposition 99: Concise form of one-step estimators

All one-step estimators can be written as

$$\underset{\sim}{\hat{\theta}_n} = \underset{\sim}{\theta}(\underset{\sim}{\tilde{\gamma}_n}) + \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{\tilde{\ell}_{\theta}}(\underset{\sim}{X_i}; \underset{\sim}{\tilde{\gamma}_n})$$

where $\underset{\sim}{\tilde{\ell}_{\theta}}(.; \underset{\sim}{\tilde{\gamma}_n}) = \underset{\sim}{\theta}^{\circ} \underset{\sim}{J}^{-1}(\underset{\sim}{\gamma}) \underset{\sim}{\ell_{\gamma}}^{\circ}$ is the efficient influence function for estimating $\underset{\sim}{\theta}(\underset{\sim}{\gamma})$ at $\underset{\sim}{\gamma}$.

STAT4028: Probability and Mathematical Statistics

23. Asymptotically Optimal Parametric Testing

23.12 Hypothesis Testing

23.12.1 Asymptotically Optimal Parametric Testing

In this lecture, we describe the set up of hypothesis testing. We then describe what it means to look at local alternatives of a test. Finally, we introduce 4 classical tests.

Suppose we model data as values taken by i.i.d random vectors $\underset{\sim}{X}_1, ..., \underset{\sim}{X}_n$ with common density in a parametric family $\{p_{\underset{\sim}{\gamma}}\}$ indexed by a d-dimensional parameter vector $\underset{\sim}{\gamma}$.

Now, a hypothesis test is when we are trying to determine whether the observed data constitutes evidence **against** a null hypothesis H_0, which states that a certain function $\underset{\sim}{\theta}(\underset{\sim}{\gamma})$ of the parameters is equal to a reference value $\underset{\sim}{\theta}^\circ$.

In particular, the function of the parameters $\underset{\sim}{\gamma}$, will be the special case where $\underset{\sim}{\theta}(\underset{\sim}{\gamma})$ is the first d_θ components of $\underset{\sim}{\gamma}$ where $d_\theta \leq d$. Then, if we define a nuisance parameter mapping

$$\underset{\sim}{\gamma} \to \underset{\sim}{\eta}(\underset{\sim}{\gamma})$$

we can therefore define a function of the parameter vector to be

$$\underset{\sim}{\gamma} \to \begin{pmatrix} \underset{\sim}{\theta}(\underset{\sim}{\gamma}) \\ \underset{\sim}{\eta}(\underset{\sim}{\gamma}) \end{pmatrix}$$

where this mapping is smoothly invertible. Therefore, we define

$$\underset{\sim}{\gamma} = \begin{pmatrix} \underset{\sim}{\theta} \\ \underset{\sim}{\eta} \end{pmatrix}$$

with $d_{\underset{\sim}{\theta}} + d_{\underset{\sim}{\eta}} = d$ where $\underset{\sim}{\theta}$ is the parameter of interest and $\underset{\sim}{\eta}$ is the nuisance parameter.

We now assume that the LAN property holds in at least a neighbourhood of the true value $\underset{\sim}{\gamma}^\circ = \begin{pmatrix} \underset{\sim}{\theta}^\circ \\ \underset{\sim}{\eta}^\circ \end{pmatrix}$.

This will allow us to define the scores vector

$$\underset{\sim}{S}_n = \underset{\sim}{S}_n(\underset{\sim}{\theta}, \underset{\sim}{\eta}) = \begin{pmatrix} S_\theta \\ S_\eta \end{pmatrix}$$

and the information matrix

$$\underset{\sim}{J} = \underset{\sim}{J}(\underset{\sim}{\theta}, \underset{\sim}{\eta}) = \begin{bmatrix} \underset{\sim}{J}_{\theta\theta} & \underset{\sim}{J}_{\theta\eta} \\ \underset{\sim}{J}_{\eta\theta} & \underset{\sim}{J}_{\eta\eta} \end{bmatrix}$$

Proposition 100: Block form of inverse information matrix

Suppose that the information matrix $\underset{\sim}{J}$ is partitioned according to

$$\underset{\sim}{J} = \underset{\sim}{J}(\theta, \eta) = \begin{bmatrix} \underset{\sim}{J}_{\theta\theta} & \underset{\sim}{J}_{\theta\eta} \\ \underset{\sim}{J}_{\eta\theta} & \underset{\sim}{J}_{\eta\eta} \end{bmatrix}$$

Then, we may write the corresponding inverse information matrix $\underset{\sim}{J}^{-1}$ as

$$\underset{\sim}{J}^{-1} = \begin{bmatrix} \underset{\sim}{J}^{\theta\theta} & \underset{\sim}{J}^{\theta\eta} \\ \underset{\sim}{J}^{\eta\theta} & \underset{\sim}{J}^{\eta\eta} \end{bmatrix}$$

where

$$\underset{\sim}{J}^{\theta\theta} = \left(\underset{\sim}{J}_{\theta\theta} - \underset{\sim}{J}_{\theta\eta} \underset{\sim}{J}^{-1}_{\eta\eta} \underset{\sim}{J}_{\eta\theta} \right)^{-1}$$

$$\underset{\sim}{J}^{\eta\eta} = \left(\underset{\sim}{J}_{\eta\eta} - \underset{\sim}{J}_{\eta\theta} \underset{\sim}{J}^{-1}_{\theta\theta} \underset{\sim}{J}_{\theta\eta} \right)^{-1}$$

$$\underset{\sim}{J}^{\theta\eta} = -\underset{\sim}{J}^{\theta\theta} \underset{\sim}{J}_{\theta\eta} \underset{\sim}{J}^{-1}_{\eta\eta}$$

$$\underset{\sim}{J}^{\eta\theta} = -\underset{\sim}{J}^{\eta\eta} \underset{\sim}{J}_{\eta\theta} \underset{\sim}{J}^{-1}_{\theta\theta}$$

Now, as our mapping is just the first d_θ components of the parameter $\underset{\sim}{\gamma}$, the Jacobian of our transformation will be the identity matrix. Then, recall that for an asymptotically efficient estimator, its estimation error has the form

$$\underset{\sim}{J}^{-1} \underset{\sim}{S} = J^{\theta\theta} S_\theta + J^{\theta\eta} S_\eta = (J^*)^{-1} S^*_\theta$$

where again, we define the effective score to be

$$S^*_\theta = S_\theta - J_{\theta\eta} J^{-1}_{\eta\eta} S_\eta$$

which is the linear combination of the scores S_θ, S_η that is uncorrelated with S_η. Additionally, we define the effective information

$$J^*_\theta = J_{\theta\theta} - J_{\theta\eta} J^{-1}_{\eta\eta} J_{\eta\theta}$$

Additionally, recall that the effective information $J^*_\theta = (J^{\theta\theta})^{-1}$.

23.12.2 Local Alternatives

We will now develop theory for analysing and comparing the power of tests.

Definition 23.332 *(Power of a test). The power of a binary hypothesis test is the probability that the test rejects the null hypothesis H_0 when a specific alternative hypothesis H_1 is true. That is, the power is given by*

$$\mathbb{P}(\textit{reject } H_0 : H_1 \textit{ is true})$$

So the size of a test is the probability of rejecting a true hypothesis. If the size of a test is bounded by α for all values in the null hypothesis, we say that the level of the test is α.

The results which we will soon show are asymptotic. That is, as $n \to \infty$, the power of our tests will go to 1.

We shall study cases the true value $\underset{\sim}{\theta} = \underset{\sim n}{\theta} \neq \underset{\sim 0}{\theta}$ but $\underset{\sim n}{\theta} \to \underset{\sim 0}{\theta}$ as $n \to \infty$ at just the right speed.

Definition 107: Local Alternative

The local alternative is a pertubation of both the parameter of interest and nuisance parameter by a local shift.

$$\underset{\sim}{\gamma_n}^{\circ} = \underset{\sim}{\gamma}^{\circ} + n^{-1/2}\underset{\sim}{h} = \begin{pmatrix} \theta^{\circ} \\ \underset{\sim}{\eta}^{\circ} \end{pmatrix} + n^{-1/2} \begin{pmatrix} h_{\theta} \\ \underset{\sim}{h_{\eta}} \end{pmatrix}$$

We shall write

$$\underset{\sim}{\hat{\gamma}_n} = \begin{pmatrix} \hat{\theta}_n \\ \underset{\sim}{\hat{\eta}_n} \end{pmatrix}$$

for an asymptotically efficient estimator of the full parameter vector. Likewise, we write

$$\underset{\sim}{\tilde{\gamma}_n} = \begin{pmatrix} \tilde{\theta}_n \\ \underset{\sim}{\tilde{\eta}_n} \end{pmatrix}$$

for a $\sqrt{n}-$consistent estimator of the full parameter vector. We now introduce a new concept.

Definition 108: Constrained Estimators

Hold the parameter of interest fixed at the null value $\underset{\sim}{\theta} = \underset{\sim}{\theta}^{\circ}$. Then, we respectively define the constrained asymptotically efficient estimator and $\sqrt{n}$-consistent estimator of the nuisance parameter $\underset{\sim}{\eta}$ as

$$\underset{\sim}{\hat{\gamma}_n}^{\circ} = \begin{pmatrix} \theta^{\circ} \\ \underset{\sim}{\hat{\eta}_n}^{\circ} \end{pmatrix}$$

$$\underset{\sim}{\tilde{\gamma}_n} = \begin{pmatrix} \theta^{\circ} \\ \underset{\sim}{\tilde{\eta}_n}^{\circ} \end{pmatrix}$$

That is, $\underset{\sim}{\hat{\eta}_n}^{\circ}$ and $\underset{\sim}{\tilde{\eta}_n}^{\circ}$ are (respectively) asymptotically efficient and $\sqrt{n}-$consistent estimators of the nuisance parameter under the extra assumption that the null hypothesis H_0 is true.

23.12.3 Classical Tests

We will now introduce classical tests we have seen before.

Definition 109: Likelihood Ratio Test

First, denote the log likelihood function at $\underset{\sim}{\gamma}$ as

$$\Lambda_n(\underset{\sim}{\gamma}) = \sum_{i=1}^{n} log\ p_{\underset{\sim}{\gamma}}(\underset{\sim}{x_i}).$$

Then, the likelihood ratio test statistic is given by

$$LR = 2[\Lambda_n(\hat{\underset{\sim}{\gamma}}_n) - \Lambda_n(\hat{\underset{\sim}{\gamma}}_n{}^{\circ})]$$

where $\hat{\underset{\sim}{\gamma}}_n$ and $\hat{\underset{\sim}{\gamma}}_n{}^{\circ}$ are (respectively) unconstrained and constrained asymptotically efficient estimators.

Remark 23.333 *To interpret this, $\hat{\underset{\sim}{\gamma}}_n{}^{\circ}$ can be thought of as MLE under the constrained assumption of H_0.*

The subsequent tests require a consistent estimate under the null hypothesis H_0 of the effective information $\underset{\sim\theta}{J^*}$ or its inverse $\underset{\sim}{J^{\theta\theta}}$, which we will shall denote the estimations (respectively) as $\hat{J}^*_\theta$ and $\hat{J}^{\theta\theta}$.

Proposition 23.334 *(Estimates of the effective information and its inverse). Let $\underset{\sim\theta}{J^*}$ be the effective information. Denote $\hat{J}^*_\theta$ as the estimator of the effective information. Two particular forms of the effective information estimator are*

$$\hat{J}^*_\theta = \underset{\sim\theta}{J^*}(\tilde{\underset{\sim}{\gamma}}_n)$$

$$\hat{J}^*_\theta = -\Lambda_n^{''}(\tilde{\underset{\sim}{\gamma}}_n)$$

Let $\underset{\sim}{J^{\theta\theta}}$ be the inverse of the effective information. Denote $\hat{J}^{\theta\theta}$ as the estimator of the inverse of the effective information. Two particular forms of the inverse of the effective information estimators are

$$\hat{J}^{\theta\theta} = \underset{\sim}{J^{\theta\theta}}(\tilde{\underset{\sim}{\gamma}}_n)$$

$$\hat{J}^{\theta\theta} = -\Lambda_n^{''}(\tilde{\underset{\sim}{\gamma}}_n)^{-1}$$

under suitable regularity conditions.

Definition 110: Rao-Score Test

The Rao Score test has the test statistic

$$R = S_\theta(\hat{\underset{\sim}{\gamma}}_n{}^{\circ})^T \hat{J}^{\theta\theta} S_\theta(\hat{\underset{\sim}{\gamma}}_n{}^{\circ})$$

where $\hat{\underset{\sim}{\gamma}}_n{}^{\circ}$ is a constrained asymptotically efficient estimator of the nuisance parameter with the parameter of interest fixed at the hypothesised value $\underset{\sim}{\theta} = \underset{\sim}{\theta}^{\circ}$.

Remark 23.335 *In Econometrics, this is known as the Lagrange multiplier test.*

Definition 111: Wald Test

The Wald test has the test statistic

$$W = n(\underset{\sim}{\hat{\theta}_n} - \underset{\sim}{\theta_n})^T \underset{\sim}{\hat{J}}^{*}_{\theta} (\underset{\sim}{\hat{\theta}_n} - \underset{\sim}{\theta_n})$$

where $\underset{\sim}{\hat{\theta}_n}$ is an asymptotically efficient estimator of the parameter of interest.

Definition 112: Neyman-Rao Test

The Neyman-Rao test has the test statistic

$$NR = S^*_{\theta}(\underset{\sim}{\tilde{\gamma}_n}^{\circ})^T \hat{J}^{\theta\theta} S^*_{\theta}(\underset{\sim}{\tilde{\gamma}_n}^{\circ})$$

where $\underset{\sim}{\tilde{\gamma}_n}^{\circ}$ is a constrained $\sqrt{n}$-consistent estimator of the nuisance parameter with the parameter of interest fixed at the hypothesised value $\underset{\sim}{\theta} = \underset{\sim}{\theta}^{\circ}$.

Remark 23.336 *The Neyman-Rao test evaluates the effective scores of a $\sqrt{n}-$consistent estimator of the nuisance parameter whereas the Rao-Score test evaluates the ordinary score of an asymptotically efficient estimator of the nuisance parameter.*

Proposition 23.337 *The Rao-Score test and the Neyman-Rao test are equivalent when the regular scores of the asymptotically efficient estimator is equal to the effective score of the $\sqrt{n}-$estimator. This occurs when the information matrix is a diagonal matrix.*

Remark 23.338 *When there is no nuisance parameter η, then the effective score is the same as the ordinary score. Therefore, the Rao test is equivalent to the Neyman-Rao test.*

STAT4028: Probability and Mathematical Statistics

24. Classical Tests are equivalent

24.12.4 Classical Tests are asymptotically equivalent

We shall show in this lecture that under certain conditions, all 4 classical test statistics are asymptotically equivalent. First, we relax the assumption of regular scores. Then, we introduce the quadratic form random variable with which all our tests converge to. We then state a new strengthened LAN property.

We will first require a proposition to show us that evaluating scores of asymptotically efficient estimators and $\sqrt{n}-$consistent estimators are equivalent to evaluating effective scores.

Proposition 101: Expression for the effective score at true values

Suppose that for a parameteric model, the LAN property holds in a neighbourhood of the true values $\underset{\sim}{\gamma}^\circ = \begin{pmatrix} \underset{\sim}{\theta}^\circ \\ \underset{\sim}{\eta}^\circ \end{pmatrix}$. Furthermore, assume that the regular scores assumption and the continuous information matrix assumption holds. Let $\underset{\sim}{\hat{\eta}_n}^\circ$ and $\underset{\sim}{\tilde{\eta}_n}^\circ$ be the constrained asymptotically efficient estimator and constrained $\sqrt{n}-$consistent estimator of the nuisance parameter when the parameter of interest is fixed at the hypothesised value $\underset{\sim}{\theta} = \underset{\sim}{\theta}^\circ$. Then

1. $\underset{\sim\theta}{S}(\underset{\sim}{\theta}^\circ; \underset{\sim}{\hat{\eta}_n}^\circ) = \underset{\sim\theta}{S^*}(\underset{\sim}{\theta}^\circ; \underset{\sim}{\eta_n}^\circ) + o_p(1)$
2. $\underset{\sim\theta}{S^*}(\underset{\sim}{\theta}^\circ; \underset{\sim}{\tilde{\eta}_n}^\circ) = \underset{\sim\theta}{S^*}(\underset{\sim}{\theta}^\circ; \underset{\sim}{\eta_n}^\circ) + o_p(1)$

where $\underset{\sim}{S_\theta}^*$ is the effective score for $\underset{\sim}{\theta}$.

Proof:(Sketch). By the Regular scores assumption and the fact that any $\sqrt{n}-$consistent estimator $\underset{\sim}{\tilde{\theta}_n}$ satisfies under $\underset{\sim}{\theta}_0$

$$S_n(\underset{\sim}{\tilde{\theta}_n}) = S_n(\underset{\sim}{\theta}_0) - J(\underset{\sim}{\theta}_0)\sqrt{n}(\underset{\sim}{\tilde{\theta}_n} - \underset{\sim}{\theta}_0) + o_p(1) \qquad (*)$$

we can rewrite the score $\underset{\sim}{S}(\underset{\sim}{\theta}^\circ; \underset{\sim}{\hat{\eta}_n}^\circ)$ using (*). ■

Remark 24.339 *This is useful for us as we can use it to show that the Rao-Score test, which uses the ordinary score, is equal to the Neyman-Rao test, which uses the effective score are equal under certain conditions.*

However, the regular scores assumption in the previous proposition can in fact be weakened to the following assumption. This is because we only need local deviations of $\underset{\sim}{h}$ in the nuisance parameter $\underset{\sim}{\eta}$ only.

Definition 113: Null-Regular Scores Assumption

Define the null values $\underset{\sim}{\gamma}^\circ = \begin{bmatrix} \theta^\circ \\ \underset{\sim}{\eta}^\circ \end{bmatrix}$ and $\underset{\sim}{h} = \begin{bmatrix} h \\ \underset{\sim}{h}_\eta \end{bmatrix}$. Then, define the remainder $R_n(\underset{\sim}{\gamma}^\circ; \underset{\sim}{h})$ via

$$S_n(\underset{\sim}{\gamma}^\circ + n^{-1/2}\underset{\sim}{h}) = S_n(\underset{\sim}{\gamma}^\circ) - J(\underset{\sim}{\gamma}^\circ)\underset{\sim}{h} + R_n(\underset{\sim}{\gamma}^\circ; \underset{\sim}{h})$$

Then, for any finite $0 < M < \infty$

$$\sup_{|\underset{\sim}{h}_\eta| \leq M} |R_n(\underset{\sim}{\gamma}^\circ; \underset{\sim}{h})| \xrightarrow{p} 0$$

where $\underset{\sim}{h}_\theta = \underset{\sim}{\theta}^\circ$ is fixed.

Remark 24.340 *This is a weakened version of the Regular scores assumption as we now hold* $\underset{\sim}{\theta} = \underset{\sim}{\theta}^\circ$. *Recall that in regular scores, we required that*

$$\sup_{|\underset{\sim}{h}| \leq M} |R_n(\underset{\sim}{\gamma}^\circ; \underset{\sim}{h})| \xrightarrow{p} 0$$

but now we only require it to hold for the nuisance parameter $\underset{\sim}{h}_\eta$ *component.*

We now define a random variable that will be extremely important for us.

Definition 114: Quadratic Form

We define the quadratic form

$$Q_n = (\underset{\sim}{S}^*_\theta)^T (\underset{\sim}{J}^*_\theta)^{-1} \underset{\sim}{S}^*_\theta = (\underset{\sim}{S}^*_\theta)^T \underset{\sim}{J}^{\theta\theta} \underset{\sim}{S}^*_\theta$$

where $\underset{\sim}{S}^*_\theta$ is the effective score and $\underset{\sim}{J}^*_\theta$ is the effective information. That is $\underset{\sim}{J}^{\theta\theta} = \left(\underset{\sim}{J}_{\theta\theta} - \underset{\sim}{J}_{\theta\eta} \underset{\sim}{J}^{-1}_{\eta\eta} \underset{\sim}{J}_{\eta\theta} \right)^{-1}$

Now, we state the proposition that our classical test statistics are asymptotically equivalent to this quadratic form random variable.

Theorem 73: Classical Test Equivalence Part 1

Suppose that the LAN property holds in the neighbourhood of the true value $\underset{\sim}{\gamma}^\circ = \begin{bmatrix} \theta^\circ \\ \underset{\sim}{\eta}^\circ \end{bmatrix}$ and both the null-regular scores and continuous information also hold at $\underset{\sim}{\gamma}^\circ$. Then

1. The Rao-Score Statistic $R = Q_n + o_p(1)$
2. The Wald-Score Statistic $W = Q_n + o_p(1)$
3. The Neyman-Rao Statistic $NR = Q_n + o_p(1)$

Proof: First, recall the Rao Score test which has the test statistic

$$R = S_\theta(\hat{\underset{\sim}{\gamma}_n}^\circ)^T \hat{J}^{\theta\theta} S_\theta(\hat{\underset{\sim}{\gamma}_n}^\circ)$$

where $\hat{\underset{\sim}{\gamma}_n}^\circ$ is a constrained asymptotically efficient estimator of the nuisance parameter with the parameter of interest fixed at the hypothesised value $\underset{\sim}{\theta} = \underset{\sim}{\theta}^\circ$. Then, by first part of the theorem on the expression of the effective score

$$\underset{\sim\theta}{S}(\underset{\sim}{\theta}^\circ; \hat{\underset{\sim}{\eta}_n}^\circ) = \underset{\sim\theta}{S^*}(\underset{\sim}{\theta}^\circ; \underset{\sim}{\eta_n}^\circ) + o_p(1)$$

and by continuity of information, we have that the Rao Score test

$$R = S_\theta(\hat{\underset{\sim}{\gamma}_n}^\circ)^T \hat{J}^{\theta\theta} S_\theta(\hat{\underset{\sim}{\gamma}_n}^\circ) = \underset{\sim\theta}{S^*}(\underset{\sim}{\gamma_n}^\circ)^T \underset{\sim}{\hat{J}}^{\theta\theta} \underset{\sim\theta}{S^*}(\underset{\sim}{\gamma_n}^\circ) = Q_n \tag{1}$$

Now, recall that the Neyman-Rao test has the test statistic

$$NR = S^*_\theta(\tilde{\underset{\sim}{\gamma}_n}^\circ)^T \hat{J}^{\theta\theta} S^*_\theta(\tilde{\underset{\sim}{\gamma}_n}^\circ)$$

where $\tilde{\underset{\sim}{\gamma}_n}^\circ$ is a constrained $\sqrt{n}$-consistent estimator of the nuisance parameter with the parameter of interest fixed at the hypothesised value $\underset{\sim}{\theta} = \underset{\sim}{\theta}^\circ$. Now, recall the second part of the theorem on the expression of the effective score

$$\underset{\sim\theta}{S^*}(\underset{\sim}{\theta}^\circ; \tilde{\underset{\sim}{\eta}_n}^\circ) = \underset{\sim\theta}{S^*}(\underset{\sim}{\theta}^\circ; \underset{\sim}{\eta_n}^\circ) + o_p(1)$$

and by continuity of the information assumption, we have that the Neyman-Rao test

$$NR = S^*_\theta(\tilde{\underset{\sim}{\gamma}_n}^\circ)^T \hat{J}^{\theta\theta} S^*_\theta(\tilde{\underset{\sim}{\gamma}_n}^\circ) = S^*_\theta(\underset{\sim}{\gamma_n}^\circ)^T \hat{J}^{\theta\theta} S^*_\theta(\underset{\sim}{\gamma_n}^\circ) = Q_n \tag{2}$$

Finally, recall that the Wald test has the test statistic

$$W = n(\hat{\underset{\sim}{\theta}_n} - \underset{\sim}{\theta}^\circ)^T \underset{\sim\theta}{\hat{J}}^* (\hat{\underset{\sim}{\theta}_n} - \underset{\sim}{\theta}^\circ)$$

where $\hat{\underset{\sim}{\theta}}_n$ is an asymptotically efficient estimator of the parameter of interest. Now, recall that an asymptotically efficient estimator $\hat{\underset{\sim}{\theta}}_n$ can have its estimation error written as

$$\sqrt{n}\Big[\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}^\circ\Big] = (\underset{\sim}{J}_\theta^*)^{-1}\underset{\sim}{S_\theta}^* + o_p(1)$$

Hence, we can rewrite the Wald test statistic

$$W = n(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}^\circ)^T \hat{\underset{\sim}{J}}_\theta^{\,*}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}^\circ) = (\underset{\sim}{S_\theta}^*)^T((\underset{\sim}{J}_\theta^*)^{-1})^T\underset{\sim}{J}_\theta^*((\underset{\sim}{J}_\theta^*)^{-1})\underset{\sim}{S_\theta}^* = \underset{\sim}{S}_\theta^*(\underset{\sim}{\gamma_n}^\circ)^T\hat{J}^{\theta\theta}\underset{\sim}{S}_\theta^*(\underset{\sim}{\gamma_n}^\circ) = Q_n \quad (3)$$

■

Now, in order to handle the likelihood ratio test, we need to strengthen the LAN assumption where we replace the pertubation $\underset{\sim}{h}$ with a random version $\sqrt{n}[\hat{\underset{\sim}{\gamma}}^n - \underset{\sim}{\gamma}^\circ]$. That is, recall that in the LAN assumption, we have that for each fixed $\underset{\sim}{h}$, the remainder term goes to zero in probability.

Definition 115: Uniform LAN Condition

Define the log likelihood ratio of two nearby values

$$L_n(\underset{\sim}{\gamma}^\circ + n^{-1/2}\underset{\sim}{h}; \underset{\sim}{\gamma}^\circ) = \sum_{i=1}^n log\ p_{\underset{\sim}{\gamma}^\circ + n^{-1/2}\underset{\sim}{h}}(\underset{\sim}{X}_i) - \sum_{i=1}^n log\ p_{\underset{\sim}{\gamma}^\circ}(\underset{\sim}{X}_i)$$

We say that the Uniform LAN propety (ULAN) holds at $\underset{\sim}{\gamma}^\circ$ if there exists a score vector $\underset{\sim}{S}_n$ and a symmetric positive definite information matrix $\underset{\sim}{J}$ such that the remainder defined by

$$L_n(\underset{\sim}{\gamma}^\circ + n^{-1/2}\underset{\sim}{h}; \underset{\sim}{\gamma}^\circ) = \underset{\sim}{h}^T\underset{\sim}{S}_n - \frac{1}{2}\underset{\sim}{h}\underset{\sim}{J}^T\underset{\sim}{h} + R_n(\underset{\sim}{\gamma}^\circ; \underset{\sim}{h})$$

satisfies a uniformly bounded $\underset{\sim}{h}$ where for any $0 < M < \infty$

$$\sup_{|h|\leq M} |R_n(\underset{\sim}{\gamma}^\circ; \underset{\sim}{h})| \xrightarrow{p} 0.$$

In other words, LAN holds uniformly in bounded $\underset{\sim}{h}$.

Remark 24.341 *First, note that in the normal LAN, we had that the remainder goes to zero for a fixed h. Now, it goes to zero for a range of bounded h, that is, we can now plug in a random* $\underset{\sim}{h}$ *given by*

$$\underset{\sim}{h} = \sqrt{n}[\hat{\underset{\sim}{\gamma}}_n - \underset{\sim}{\gamma}^\circ].$$

Proposition 102: Conditions for ULAN

The restrictive conditions for the LAN property implies that the ULAN property holds.

Proof: Under the restrictive assumptions for LAN, we can do a third order Taylor expansion of the log likelihood ratio. From that, we can show that the linear term converges in distribution to a normal distribution, the quadratic term converges in probability to $-\underset{\sim}{J}$ and the cubic term is the remainder that goes to zero. ■

We can now write the expression of log likelihood ratios under ULAN, which we will need in the proof of the limiting distribution of the likelihood ratio statistic.

Lemma 24.342 *(Expression of likelihood-ratio under ULAN). Suppose that the ULAN property holds at* $\underset{\sim}{\gamma}^{\circ} = \begin{pmatrix} \theta^{\circ} \\ \underset{\sim}{\eta}^{\circ} \end{pmatrix}$. *Denote* $L_n(.)$ *to be the log likelihood ratio.*

1. *If* $\underset{\sim n}{\hat{\gamma}}$ *is an unconstrained asymptotically efficient estimator for the whole vector, then*

$$L_n(\underset{\sim n}{\hat{\gamma}}; \underset{\sim}{\gamma}^{\circ}) = \frac{1}{2} \underset{\sim}{S}^T(\underset{\sim}{\gamma}^{\circ})^T \underset{\sim}{J}(\underset{\sim}{\gamma}^{\circ})^{-1} \underset{\sim}{S}(\underset{\sim}{\gamma}^{\circ}) + o_p(1)$$

   where $\underset{\sim}{S}$ *is the score vector.*

2. *If* $\underset{\sim n}{\hat{\gamma}}^{\circ} = \begin{pmatrix} \gamma^{\circ} \\ \underset{\sim}{\hat{\eta}}^{\circ} \end{pmatrix}_{\sim n}$ *is a constrained asymptotically efficient estimator of the nuisance parameter* η, *then*

$$L_n(\underset{\sim n}{\hat{\gamma}}^{\circ}; \underset{\sim}{\gamma}^{\circ}) = \frac{1}{2} \underset{\sim}{S_{\eta}}^T(\underset{\sim}{\gamma}^{\circ})^T \underset{\sim}{J_{\eta\eta}}(\underset{\sim}{\gamma}^{\circ})^{-1} \underset{\sim}{S_{\eta}}(\underset{\sim}{\gamma}^{\circ}) + o_p(1)$$

Remark 24.343 *Note that the second case is a special case of the first case whereby we are now only estimating the nuisance parameter.*

Theorem 74: Classical Test Equivalence Part 2

Suppose that the ULAN property holds at $\underset{\sim}{\gamma}^{\circ}$, we then have that the likelihood-ratio statistic

$$LR = Q_n^2 + o_p(1).$$

Hence, we see that all 4 classical tests are asymptotically equivalent as they all converge to the same quadratic form random variable Q_n.

STAT4028: Probability and Mathematical Statistics

25. Properties of test statistics

25.12.5 Properties of test statistics

In this lecture, we look at local alternatives for hypothesis testing. We first introduce the central and non-central Chi-squared distributions. Then, we show that the limiting behaviour of classical test statistics under the null alternative is a χ^2-distribution under H_0 for any value of the nuisance parameter due to asymptotic similarity and a non-central χ^2-distribution under the alternative hypothesis.

First, recall that we have shown that all classical tests are asymptotically equivalent to the quadratic form

$$Q_n = (\underset{\sim}{S_\theta}^*)^T \underset{\sim}{J}^{\theta\theta} (\underset{\sim}{S_\theta}^*).$$

Therefore, we can deduce various properties of the classical tests by analysing the properties of Q_n.

25.12.5.1 Limiting Distribution Under Null Hypothesis

Proposition 103: Limiting distribution of Q_n under null hypothesis

If the parameter of interest has dimension d_θ, then, when H_0 is true, we have that the limiting distribution of Q_n is

$$Q_n \xrightarrow{d} \chi^2_{d_\theta}.$$

We will now prove why is this proposition true. First, recall a fact from linear algebra.

Definition 25.344 *(Eigendecomposition of a matrix). Let $\underset{\sim}{\Sigma}$ be a matrix of full rank. We can then express it in terms of its eigendecomposition*

$$\underset{\sim}{\Sigma} = \underset{\sim}{U}\underset{\sim}{D}\underset{\sim}{U}^T$$

where $\underset{\sim}{U}$ is an orthogonal matrix which contains eigenvectors and $\underset{\sim}{D}$ is a diagonal matrix containing eigenvalues, whose diagonal elements are all positive.

Lemma 25.345 *If $\underset{\sim}{U}$ is an orthogonal matrix, then*

$$\underset{\sim}{U}\underset{\sim}{U}^T = \mathbb{I}.$$

We can now state the proposition that will help us prove that the quadratic form has a χ^2-distribution.

Proposition 25.346 *Suppose that we have the multivariate normal*

$$\underset{\sim}{X} \sim \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{\Sigma})$$

where $\underset{\sim}{X}$ has dimension d and the covariance matrix $\underset{\sim}{\Sigma}$ has full rank d, then

$$\underset{\sim}{X}^T \underset{\sim}{\Sigma}^{-1} \underset{\sim}{X} \sim \chi^2_d.$$

Proof: First, due to the eigendecomposition of a full rank vector, we have that we can express the covariance matrix as

$$\underset{\sim}{\Sigma}^{-1/2} = \underset{\sim}{U}\underset{\sim}{D}^{-1/2}\underset{\sim}{U}$$

where $\underset{\sim}{D}^{-1/2}$ is the inverse square root of the diagonal entries of $\underset{\sim}{D}$. Then, we can do a change of variables where for $\underset{\sim}{X} \sim \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{\Sigma})$, we can write the rescaled random variable

$$\underset{\sim}{Z} = \underset{\sim}{\Sigma}^{-1/2}\underset{\sim}{X}.$$

This new random variable has mean zero

$$\mathbb{E}[\underset{\sim}{Z}] = \underset{\sim}{\Sigma}^{-1/2}\mathbb{E}[\underset{\sim}{X}] = \underset{\sim}{0}$$

and its variance is given by

$$\begin{aligned}\mathbb{E}[\underset{\sim}{Z}\underset{\sim}{Z}] &= \underset{\sim}{\Sigma}^{-1/2}\mathbb{E}[\underset{\sim}{X}\underset{\sim}{X}]\underset{\sim}{\Sigma}^{-1/2} \\ &= \underset{\sim}{\Sigma}^{-1/2}\underset{\sim}{\Sigma}\underset{\sim}{\Sigma}^{-1/2} \\ &= \underset{\sim}{U}\underset{\sim}{D}^{-1/2}\underset{\sim}{U}^T\underset{\sim}{U}\underset{\sim}{D}\underset{\sim}{U}^T\underset{\sim}{U}\underset{\sim}{D}^{-1/2}\underset{\sim}{U}^T \\ &= \underset{\sim}{U}\underset{\sim}{D}^{-1/2}\underset{\sim}{D}\underset{\sim}{D}^{-1/2}\underset{\sim}{U}^T = \underset{\sim}{U}\underset{\sim}{U}^T = \mathbb{I}\end{aligned}$$

Therefore, $\underset{\sim}{Z}$ consists of d i.i.d $\mathcal{N}(\underset{\sim}{0}, \mathbb{I})$ random variables and therefore, the sum of squares of $\underset{\sim}{Z}$ is a χ^2-distribution. Therefore, we have that

$$\begin{aligned}\underset{\sim}{Z}^T\underset{\sim}{Z} &= \underset{\sim}{X}^T\underset{\sim}{\Sigma}^{-1/2}\underset{\sim}{\Sigma}^{-1/2}\underset{\sim}{X} \\ &= \underset{\sim}{X}^T\underset{\sim}{\Sigma}^{-1}\underset{\sim}{X} \sim \chi^2_d.\end{aligned}$$

■

We can therefore apply the above proposition to our quadratic form!

Corollary 25.347 *Recall that the effective score* $(\underset{\sim}{S_\theta}^*)$ *is an asymptotically normal distribution with the effective information matrix as its variance and the inverse covariance matrix is* $\underset{\sim}{J}^{\theta\theta}$. *Therefore, we have that the limiting distribution is*

$$(\underset{\sim}{S_\theta}^*)^T\underset{\sim}{J}^{\theta\theta}(\underset{\sim}{S_\theta}^*) \sim \chi^2_{d_\theta}$$

Definition 116: Asymptotic Similarity

Suppose that the null hypothesis is true $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0$. Then, asymptotic similarity is when the limiting distribution of the test statistic is the same for any nuisance parameter $\underset{\sim}{\eta}$.

Corollary 25.348 *The Quadratic form* Q_n *has the property of asymptotic similarity.*

25.12.5.2 Limiting Distribution Under Nearby Alternatives

Recall that Le Cam's third lemma tells us what happens to the limiting distribution when we perturb the true value. We now want to investigate the limiting distribution of the quadratic form under nearby alternatives. First of all, recall that Le Cam's third lemma gives the limiting distribution under local alternatives of the form

$$\underset{\sim}{\gamma_n}^\circ = \underset{\sim}{\gamma}^\circ + n^{-1/2}\underset{\sim}{h} = \begin{pmatrix} \theta^\circ \\ \sim \\ \eta^\circ \\ \sim \end{pmatrix} + n^{-1/2} \begin{pmatrix} h_\theta \\ \sim \\ h_\eta \\ \sim \end{pmatrix}$$

So, first of all, recall that the score vector

$$\underset{\sim}{S} = \begin{pmatrix} S_\theta \\ \sim \\ S_\eta \\ \sim \end{pmatrix}$$

has the limiting distribution under the true value $\underset{\sim}{\gamma}^\circ$

$$\underset{\sim}{S} \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{J})$$

where $\underset{\sim}{J}$ is the information matrix and under the nearby sequence $\underset{\sim}{\gamma_n}^\circ$

$$\underset{\sim}{S} \xrightarrow{d} \mathcal{N}(\underset{\sim}{J}\underset{\sim}{h}, \underset{\sim}{J})$$

where

$$\underset{\sim}{h} = \begin{pmatrix} h_\theta \\ \sim \\ h_\eta \\ \sim \end{pmatrix}$$

Lemma 25.349 *Let $\underset{\sim}{S}$ be the full score vector. Then, for any matrix $\underset{\sim}{M}$ with $d = d_\theta + d_\eta$ columns, we have that*

$$\underset{\sim}{M}\underset{\sim}{S} \xrightarrow{d} \begin{cases} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{M}\underset{\sim}{J}\underset{\sim}{M}^T) & \text{under } \underset{\sim}{\gamma}^\circ \\ \\ \mathcal{N}(\underset{\sim}{M}\underset{\sim}{J}\underset{\sim}{h}, \underset{\sim}{M}\underset{\sim}{J}\underset{\sim}{M}^T) & \text{under } \underset{\sim}{\gamma_n}^\circ \end{cases}$$

We are now interested in the case of the limiting distribution of the effective score, which can be written as $\underset{\sim}{M}\underset{\sim}{S}$ where $\underset{\sim}{M} = (\underset{\sim}{I}, -\underset{\sim}{J_{\theta\eta}}\underset{\sim}{J_{\eta\eta}^{-1}})$.

Proposition 104: Limiting distribution of effective score under alternative

The limiting distribution of the effective score under the alternative $\underset{\sim}{\gamma_n}^\circ$ is given by

$$\underset{\sim}{S_\theta}^* \xrightarrow{d} \mathcal{N}(\underset{\sim}{J_\theta}^*\underset{\sim}{h_\theta}, \underset{\sim}{J_\theta}^*)$$

where $\underset{\sim}{J_\theta}^*$ is the effective information.

Note that this distribution is free of $\underset{\sim}{h_\eta}$, which is an important property.

Definition 117: Regularity

If the limiting distribution is free of the nuisance parameter $\underset{\sim}{h_\eta}$, then it is known as the property of regularity.

In hypothesis testing, we specify a value of $\underset{\sim}{\theta}$ for our null hypothesis. Therefore, if we change $\underset{\sim}{\theta}$, we expect the test statistic to change and depend on $\underset{\sim}{h_\theta}$. However, we do not want to detect a change in $\underset{\sim}{\eta}$ and in fact, to ignore any changes from $\underset{\sim}{h_\eta}$. This is what regularity guarantees for us.

Corollary 25.350 *For $\underset{\sim}{h_\theta} = \underset{\sim}{0}$, the limiting distribution is $\chi^2_{d_\theta}$ regardless of the value of $\underset{\sim}{h_\eta}$.*

Remark 25.351 *This can be thought of as a local version of asymptotic similarity.*

Therefore, we can conclude from this section that if we have a regular model, we always will have a limiting distribution of

$$\chi^2_{d_\theta}$$

regardless of the nuisance parameter $\underset{\sim}{h_\eta}$.

Now, we are interested in what happens when we shift $\underset{\sim}{h_\theta}$. First, we introduce a distribution which we will need later.

Definition 118: Non-central Chi-squared distribution

Let $X_1, ..., X_d$ be independent random variables with $X_i \sim \mathcal{N}(\mu_i, 1)$. Then, the random variable

$$S = \sum_{i=1}^{d} X_i^2 \sim \chi^2_d(\delta)$$

where $\delta = \sum_{i=1}^{d} \mu_i^2$ is said to have a non-central Chi-squared distribution with d degrees of freedom and **noncentrality parameter** δ.

Now, recall that under the local alternative $\underset{\sim}{\gamma_n}^\circ$, we have that the effective score

$$\underset{\sim}{S_\theta}^* \xrightarrow{d} \mathcal{N}(\underset{\sim}{J_\theta}^* \underset{\sim}{h_\theta}, \underset{\sim}{J_\theta}^*)$$

where $\underset{\sim}{J_\theta}^*$ is the effective information. We are now interested in what is the limiting distribution under the alternative for the quadratic form

$$Q_n = (\underset{\sim}{S_\theta}^*)^T (\underset{\sim}{J_\theta}^*)^{-1} (\underset{\sim}{S_\theta}^*) = (\underset{\sim}{S_\theta}^*)^T (\underset{\sim}{J}^{\theta\theta})(\underset{\sim}{S_\theta}^*)$$

First, let us define the random variable

$$\underset{\sim}{Z} = (\underset{\sim}{J}^{\theta\theta})^{1/2} \underset{\sim}{S_\theta}^* = (\underset{\sim\theta}{J^*})^{-1/2} \underset{\sim}{S_\theta}^*$$

Then, we can show that

$$\mathbb{E}[\underset{\sim}{Z}] = (\underset{\sim\theta}{J^*})^{-1/2} \underset{\sim}{J_\theta}^* \underset{\sim}{h_\theta}$$

and the limiting covariance matrix is

$$\mathbb{E}[\underset{\sim}{Z}\underset{\sim}{Z}^T] = \mathbb{I}.$$

Therefore, $\underset{\sim}{Z}$ consists of asymptotically normal random variables with variance $\mathbb{I}$. Therefore, the sum of squares $\underset{\sim}{Z}^T\underset{\sim}{Z}$ is a χ^2-distribution. Finally, we can show that

$$\underset{\sim}{Z}^T\underset{\sim}{Z} = Q_n$$

and hence is equal to our quadratic form.

Proposition 105: Limiting distribution of Q_n under alternative hypothesis

If the parameter of interest has dimension d_θ, then, under a local alternative, we have that the limiting distribution of Q_n is

$$Q_n \xrightarrow{d} \chi^2_{d_\theta}(\delta_\theta).$$

where $\delta_\theta = \underset{\sim\theta}{h}^T \underset{\sim\theta\sim\theta}{J^*h}$ is the non-centrality parameter.

This non-centrality parameter $\delta_\theta = \underset{\sim\theta}{h}^T \underset{\sim\theta\sim\theta}{J^*h}$ is the biggest that the non-centrality can be. We will investigate this further in the next lecture.

Under null hypothesis, we have Chi-squared regardless of the nuisance parameter due to asymptotic similarity.

STAT4028: Probability and Mathematical Statistics

26. Regular Quadratic Form Test Statistics

26.12.6 Regular Quadratic Form Test Statistics

In this lecture, we first define the equivalence of RAJN estimators in the context of testing, which are regular quadratic form test statistics. That is, within this class of quadratic form tests, which are based on test statistics constructed from statistics that are AJN with the scores vector, we derive the efficiency of all classical tests within this class. We then derive the limiting distribution of regular quadratic form test statistics, which is a non-central chi-squared distribution. Finally, we show that if we have a test statistic with $K > d_\theta$ variables, we can get a more powerful test by transforming our test statistic to only use d_θ variables.

First, we define the equivalence of a regular estimator in the context of testing.

Definition 119: Regular Test Statistic

An asymptotically similar test statistic T_n is regular at $\underset{\sim}{\gamma}_0$ if under the nearby sequence

$$\underset{\sim}{\gamma}_n^\circ = \begin{pmatrix} \theta^\circ \\ \underset{\sim}{} \\ \eta^\circ \\ \underset{\sim}{} \end{pmatrix} + n^{-1/2} \begin{pmatrix} h \\ \underset{\sim}{\theta} \\ h \\ \underset{\sim}{\eta} \end{pmatrix}$$

the limiting distribution satisfies the 2 conditions

1. The limiting distribution possibly depends on $\underset{\sim}{\gamma}_0$, $\underset{\sim}{h}_\theta$ but not $\underset{\sim}{h}_\eta$
2. For any $\underset{\sim}{h}_\theta \neq \underset{\sim}{0}$, this differs from that under the null hypothesis $\underset{\sim}{\gamma}_0$

Remark 26.352 *Notice that for regularity, we require that the distribution changes if we shift in the $\underset{\sim}{\theta}$ direction.*

The reason that we require the second condition for a regular test statistic is to assure local power in every direction of $\underset{\sim}{h}_\theta$. That is, for tests with smooth power functions, this is akin to requiring partial derivatives of the power function with respect to θ to be non-zero and partial derivatives with respect to η to be zero.

We now define the equivalence of an AJN estimator in the context of testing. In fact, we will restrict our test statistics to be of the following class.

Definition 120: Quadratic Form Test Statistic

A test statistic T_n for testing H_0 is a quadratic form test statistic if under the null hypothesis $\underset{\sim}{\gamma}_0 = \begin{pmatrix} \theta \\ \underset{\sim}{0} \\ \eta \\ \underset{\sim}{0} \end{pmatrix}$ can be written as

$$T_n = \underset{\sim}{Y_n}^T \underset{\sim}{\Sigma}_Y^{-1} \underset{\sim}{Y_n} + o_p(1)$$

for some K-vector $\underset{\sim}{Y_n}$ $(K \geq d_\theta)$ satisfying LAN# with a non-singular asymptotic covariance matrix Σ_Y.

Remark 26.353 *Recall that under LAN#, we have that $\underset{\sim}{Y_n}$ is AJN with the scores vector. The reason we require $\underset{\sim}{Y_n}$ to satisfy LAN# is to ensure asymptotic similarity.*

We now state the limiting distribution of a regular quadratic form test statistic.

Proposition 106: Non-Centrality of a regular quadratic form test statistic

Suppose that T_n is a regular quadratic form test statistic.
Now, as T_n is a quadratic form test statistic, then under the local alternative $\underset{\sim}{\gamma_n}^\circ = \underset{\sim}{\gamma}^\circ + n^{-1/2}\underset{\sim}{h}$, the limiting distribution is

$$T_n \xrightarrow{d} \chi^2_k(\delta)$$

with the non-centrality parameter being $\delta = \underset{\sim}{h}^T \underset{\sim}{\Sigma}_{SY} \underset{\sim}{\Sigma}_Y^{-1} \underset{\sim}{\Sigma}_{YS} \underset{\sim}{h}$ and is asymptotically similar.
Now, as T_n is a regular statistic, then the cross covariance with the scores of interest Σ_{YS_θ} is of rank d_θ and the cross covariance with the nuisance parameter is 0 $\Sigma_{YS_\eta} = \underset{\sim}{0}$ in which case, the non-centrality parameter simplifies as it now only depends on θ

$$\delta = \underset{\sim}{h_\theta}^T \Sigma_{S^*_\theta Y} \Sigma_Y^{-1} \Sigma_{YS^*_\theta} \underset{\sim}{h_\theta}$$

as $\Sigma_{YS_\theta} = \Sigma_{YS^*_\theta}$.

Remark 26.354 *What this proposition state is that because T_n is a quadratic form test statistic, then we have a non-central χ^2-distribution under the local alternative. Furthermore, because T_n is regular, then the cross-covariance with the score of the nuisance parameter is zero and therefore the non-centrality parameter simplifies.*

We will later show that the non-centrality parameter of the classical test statistics under the alternative hypothesis can be no bigger than $\delta = \underset{\sim}{h_\theta}^T \Sigma_{S^*_\theta Y} \Sigma_Y^{-1} \Sigma_{YS^*_\theta} \underset{\sim}{h_\theta}$.

First, we recall another fact from linear algebra.

Lemma 26.355 *Suppose that the quadratic form test statistic $T_n = \underset{\sim}{Y_n}^T \underset{\sim}{\Sigma_Y}^{-1} \underset{\sim}{Y_n}$ for $\underset{\sim}{Y_n}$ satisfying LAN# with asymptotic covariance matrix $\underset{\sim}{\Sigma_Y}$. Then*

$$T_n = \underset{\sim}{Z_n}^T \underset{\sim}{Z_n}$$

where $\underset{\sim}{Z_n} = \underset{\sim}{\Sigma_Y}^{-1/2}\underset{\sim}{Y_n}$ *where now* $\underset{\sim}{Z_n}$ *satisfies LAN# with limiting covariance matrix* $\mathbb{I}$.

Our next proposition shows that if our statistic has dimension $K > d_\theta$, we should restrict it to transform it to only look at $K = d_\theta$ in order to gain power for our test.

Proposition 107: Power in a test is in first d_θ components

Suppose that the test statistic

$$T_n = \underset{\sim}{Y_n}^T \underset{\sim}{Y_n}$$

is a regular quadratic form test statistic for a K-dimensional vector $\underset{\sim}{Y_n}$ where $K > d_\theta$. Then, there exists an orthogonal K x K matrix (A, B) where A has d_θ columns such that we can construct 2 new vectors

$$\begin{pmatrix} U_n \\ V_n \end{pmatrix} = \begin{pmatrix} A^T \\ B^T \end{pmatrix} Y_n.$$

Then, under a nearby sequence $\underset{\sim}{\gamma_n}^\circ = \underset{\sim}{\gamma}^\circ + n^{-1/2} \begin{pmatrix} h_\theta \\ \underset{\sim}{h_\eta} \end{pmatrix}$, we can rewrite the original statistic

$$T_n = \underset{\sim}{Y_n}^T \underset{\sim}{Y_n} = \underset{\sim}{U_n}^T \underset{\sim}{U_n} + \underset{\sim}{V_n}^T \underset{\sim}{V_n}$$

where we have that

$$\underset{\sim}{U_n}^T \underset{\sim}{U_n} \xrightarrow{d} \chi^2_{d_\theta}(\delta)$$

$$\underset{\sim}{V_n}^T \underset{\sim}{V_n} \xrightarrow{d} \chi^2_{K-d_\theta}(0)$$

where $\chi^2_{d_\theta}(\delta)$ is a non-central chi-squared distribution and $\chi^2_{K-d_\theta}(0)$ is a central chi-squared distribution.
Furthermore, $\underset{\sim}{U_n}^T \underset{\sim}{U_n}$ and $\underset{\sim}{V_n}^T \underset{\sim}{V_n}$ are asymptotically independent, which implies that

$$\underset{\sim}{U_n}^T \underset{\sim}{U_n} + \underset{\sim}{V_n}^T \underset{\sim}{V_n} \sim \chi^2_K(\delta)$$

with non-centrality parameter $\delta = \underset{\sim}{h_\theta}^T \Sigma_{S^*_\theta Y} \Sigma_Y^{-1} \Sigma_{Y S^*_\theta} \underset{\sim}{h_\theta}$.

Remark 26.356 *Any regular quadratic form test statistic in K variables ($K > d_\theta$) may be characterised as the sum of d_θ variables that carries all the signal and another $(K - d_\theta)$ which carries only added noise. Therefore, using only* $\underset{\sim}{U_n}^T \underset{\sim}{U_n}$ *instead of* $\underset{\sim}{Y_n}^T \underset{\sim}{Y_n}$ *increases the power of our test as the non-centrality parameter is the signal. That is, the non-centrality is the signal, the characteristic that makes it differ from its null hypothesis behaviour.*

Corollary 26.357 *We will restrict attention to regular quadratic form statistics in d_θ variables.*

Remark 26.358 *If we have more than d_θ variables, we can reduce it using the orthogonal matrix and therefore get a more powerful test.*

STAT4028: Probability and Mathematical Statistics

27. Orthogonal Decomposition Of Tests

In this lecture, will do an analogue of the orthogonal decomposition theorem whereby we will show that in the class of regular quadratic form test statistics, we can't get better power than the classical statistics.

27.12.7 Positive Definite Matrices

We first recall some things from linear algebra.

Definition 121: Positive Definite Matrix

A d-by-d matrix M is positive definite if for all $\underset{\sim}{h} \in \mathbb{R}^d$ where $\underset{\sim}{h} \neq \underset{\sim}{0}$, we have that

$$\underset{\sim}{h}^T \underset{\sim}{M} \underset{\sim}{h} > 0.$$

We then write that $\underset{\sim}{M} > 0$.

Remark 27.359 *We say that M is positive semi-definite if for all* $\underset{\sim}{h} \in \mathbb{R}^d$ *where* $\underset{\sim}{h} \neq \underset{\sim}{0}$ *if*

$$\underset{\sim}{h}^T \underset{\sim}{M} \underset{\sim}{h} \geq 0$$

Lemma 27.360 *If* $\underset{\sim}{M} > 0$*, then* $\underset{\sim}{M}$ *is invertible.*

Now, note that if

$$\underset{\sim}{M} - \underset{\sim}{N} \geq 0$$

we can write $\underset{\sim}{M} \geq \underset{\sim}{N}$.

Claim 27.361 *If* $\underset{\sim}{M} \geq \underset{\sim}{N} > \underset{\sim}{0}$ *and* $\underset{\sim}{M} > \underset{\sim}{0}$*. Then*

$$\underset{\sim}{N}^{-1} \geq \underset{\sim}{M}^{-1} > 0.$$

27.12.8 Orthogonal Decomposition Of Tests

Recall that we can restrict our attention to d_θ variables for regular quadratic form test statistics. We now will develop an analogue for orthogonal decomposition in the context of hypothesis testing.

Suppose that T_n is a regular quadratic form statistic in $\underset{\sim}{Y_n}$ of length d_θ such that it has the LAN# property, which means that

$$\begin{pmatrix} Y_n \\ S_\theta \\ S_\eta \end{pmatrix} \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \begin{pmatrix} \Sigma_Y & \Sigma_{YS_\theta} & \underset{\sim}{0} \\ \Sigma_{S_\theta Y} & J_{\theta\theta} & J_{\theta\eta} \\ \underset{\sim}{0} & J_{\eta\theta} & J_{\eta\eta}) \end{pmatrix})$$

where Σ_{YS_θ} is of rank d_θ (full-rank) and is invertible. We can do an implicit transformation and assume that $\Sigma_{YS_\theta} = \underset{\sim}{I}$.

Lemma 27.362 *We can always express a regular quadratic test statistic T_n with a cross-covariance with the score*

$$\Sigma_{YS_\theta} = \underset{\sim}{I}.$$

We will illustrate on why is this the case. First, define the new variable

$$\underset{\sim}{Z_n} = \Sigma_{YS_\theta}^{-1} \underset{\sim}{Y_n}.$$

Now, we can express the test statistic

$$T_n = \underset{\sim}{Y_n}^T \underset{\sim}{\Sigma_{Y_n}}^{-1} \underset{\sim}{Y_n} + o_p(1) = \underset{\sim}{Z_n}^T \underset{\sim}{\Sigma_{YS_\theta}} \underset{\sim}{\Sigma_Y}^{-1} \underset{\sim}{\Sigma_{S_\theta Y}} \underset{\sim}{Z_n}$$

From this, $\underset{\sim}{Z_n}$ also satisfies the LAN# property whereby

$$\begin{pmatrix} \underset{\sim}{Z_n} \\ \underset{\sim}{S_\theta} \\ S_\eta \end{pmatrix} \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \begin{pmatrix} \Sigma_Y & \underset{\sim}{I} & \underset{\sim}{0} \\ \underset{\sim}{I} & J_{\theta\theta} & J_{\theta\eta} \\ \underset{\sim}{0} & J_{\eta\theta} & J_{\eta\eta}) \end{pmatrix}$$

Finally, we can now define the remainder as

$$\underset{\sim}{R_n} = \underset{\sim}{Z_n} - (J_\theta^*)^{-1} S_\theta^* \tag{*}$$

where $(J_\theta^*)^{-1}$ is the inverse of the effective information and S_θ^* is the effective score.

Lemma 27.363 *The asymptotic covariance of the effective score and regular of θ is*

$$\mathbb{E}[(S_\theta^*)S_\theta^T] = J_\theta^*$$

The asymptotic covariance of the effective score of θ and the score of the nuisance parameter is

$$\mathbb{E}[(S_\theta^*)S_\eta^T] = \underset{\sim}{0}$$

Additionally, we have the remainder $\underset{\sim}{R_n}$ also satisfies the LAN# property

$$\begin{pmatrix} \underset{\sim}{R_n} \\ \underset{\sim}{S_\theta} \\ S_\eta \end{pmatrix} \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \begin{pmatrix} \Sigma_R & \Sigma_{RS_\theta} & \underset{\sim}{0} \\ \Sigma_{S_\theta R} & J_{\theta\theta} & J_{\theta\eta} \\ \underset{\sim}{0} & J_{\eta\theta} & J_{\eta\eta}) \end{pmatrix}$$

where we have that $\Sigma_{RS_\theta} = \underset{\sim}{0}$. That is, this remainder $\underset{\sim}{R_n}$ is uncorrelated with S_θ.

From equation (*), we re-arrange the equation to get

$$\underset{\sim}{Z_n} = (J_\theta^*)^{-1} S_\theta^* + \underset{\sim}{R_n}$$

Lemma 27.364 *We can interpret our test statistic $\underset{\sim}{Z}_n$ to be the decomposition of a linear combination of the effective score and an asymptotically independent remainder $\underset{\sim}{R}_n$.*

Remark 27.365 *This is significant as every regular quadratic form test statistic can be decomposed into the optimal variables associated with the classical test statistics and some noise.*

Since we shown that $\Sigma_{RS_\theta} = \underset{\sim}{0}$, we get the following proposition.

Proposition 27.366 *Define the random variable $\underset{\sim}{Z}_n = (J_\theta^*)^{-1} S_\theta^* + \underset{\sim}{R}_n$. Then, we have that the covariance matrix of $\underset{\sim}{Z}_n$ satisfies the following relationship*

$$\Sigma_Z = (J_\theta^*)^{-1} + \Sigma_R \geq (J_\theta^*)^{-1}.$$

where J_θ^* is the effective information matrix.

Remark 27.367 *Recall that if X and Y are uncorrelated, then $Var(X+Y) = Var(X) + Var(Y)$.*

Then, using the relationship $\Sigma_Z \geq (J_\theta^*)^{-1}$, we have that

$$J_\theta^* - \Sigma_Z^{-1} \geq \underset{\sim}{0}.$$

Then, by definition of positive semi-definite, for any $\underset{\sim}{h}_\theta \in \mathbb{R}^d$, we have that

$$\underset{\sim}{h}_\theta^T [J_\theta^* - \Sigma_Z^{-1}] \underset{\sim}{h}_\theta \geq \underset{\sim}{0}.$$

We have a following result from this.

Theorem 27.368 *The non-centrality of the regular quadratic form test statistic T_n is given by*

$$\underset{\sim}{h}_\theta^T \Sigma_Z^{-1} \underset{\sim}{h}_\theta \leq \underset{\sim}{h}_\theta^T J_\theta^* \underset{\sim}{h}_\theta$$

with equality only if

$$\Sigma_Z = (J_\theta^*)^{-1} + o_p(1).$$

If this is the case, we can then express the test statistic

$$T_n = (J_\theta^*)^{-1} S_\theta^* + o_p(1) = Q_n + o_p(1)$$

where Q_n is the quadratic form associated with the classical test statistics.

Remark 27.369 *Recall that the power of a test only depends on the non-centrality of the parameter of the non-central χ^2-distribution. That is, the bigger the non-centrality parameter, the more power in the test.*

Now, regular quadratic form test statistics are the analogue to RAJN estimators and tests that can be expressed in the quadratic form of the classical test statistics is analogous to asymptotically efficient estimators.

Proposition 108: Orthogonal Decomposition of Tests

For any d_θ-dimensional regular quadratic form statistic T_n with an associated vector $\underset{\sim}{Y}_n$, we can first express it as $\underset{\sim}{Z}_n = \Sigma_{YS_\theta}^{-1}\underset{\sim}{Y}_n$. Then, we can decompose our regular quadratic form test statistic

$$\underset{\sim}{Z}_n = (J_\theta^*)^{-1}S_\theta^* + \underset{\sim}{R}_n$$

Then, for under any $\underset{\sim}{\gamma_n}^\circ = \begin{pmatrix} \theta_0 \\ \underset{\sim}{\eta_0} \end{pmatrix} + n^{-1/2}\begin{pmatrix} h_\theta \\ \underset{\sim}{h_\eta} \end{pmatrix}$, the noncentrality parameter is

$$\delta_\theta = \underset{\sim}{h_\theta}^T \underset{\sim}{J_\theta}^* \underset{\sim}{h_\theta}$$

with equality only if

$$T_n = Q_n + o_p(1).$$

Remark 27.370 *The power of our test associated with $\underset{\sim}{Z}_n$ is the non-centrality parameter $\underset{\sim}{h_\theta}^T\Sigma_Z^{-1}\underset{\sim}{h_\theta}$, where this achieves the maximal power*

$$\underset{\sim}{h_\theta}^T\Sigma_Z^{-1}\underset{\sim}{h_\theta} = \underset{\sim}{h_\theta}^T J_\theta^* \underset{\sim}{h_\theta}$$

when the test statistic can be expressed as the quadratic form of the classical test statistics. Therefore, the noncentrality parameter of our original test statistic now have the same noncentrality parameter of the classical test statistic.

Corollary 27.371 *The four classical test statistics has the maximal power, i.e, the highest non-centrality parameter term.*

In fact, the Neyman-Pearson lemma can be applied to verify that there are in fact the most powerful test statistics.

Definition 122: Asymptotically Efficient Test

A quadratic form test statistic T_n is asymptotically efficient if it is regular and has noncentrality parameter $\delta_\theta = \underset{\sim}{h_\theta}^T \underset{\sim}{J_\theta}^* \underset{\sim}{h_\theta}$.

The limiting (local) power is determined by the non-centrality. Thus, the classical tests provides various options for constructing regular quadratic form test statistics, that are optimal according to the decomposition theorem. That is, we have 4 different ways of constructing optimal tests.

The Rao Score statistic R requires a constrained asymptotically efficient estimator of the nuisance parameter $\underset{\sim}{\eta}$ and a consistent estimator of the inverse of the effective information $(J_\theta^*)^{-1}$.

The Wald statistic W requires an asymptotically efficient estimator of the parameter of interest $\underset{\sim}{\theta}$ and a consistent estimator of the effective information (J_θ^*).

The Neyman-Rao statistic NR requires a $\sqrt{n}-$consistent estimator of the nuisance parameter $\underset{\sim}{\eta}$ and consistent estimators of the various sublocks of the information matrix $J_{\theta\eta}, J_{\eta\eta}^{-1}, J^{\theta\theta}$.

The Likelihood-Ratio statistic needs both an asymptotically efficient estimator $\underset{\sim}{\hat{\gamma}_n}$ of the full parameter and a constrained asymptotically efficient estimator

$$\underset{\sim}{\hat{\gamma}^\circ} = \begin{pmatrix} \theta^\circ \\ \underset{\sim}{\hat{\eta}_n}^\circ \end{pmatrix}$$

In different situations, we use different tests depending on whether can we obtain an asymptotically efficient estimator.

STAT4028: Probability and Mathematical Statistics

28. Optimality of Bayesian Methods

28.13 Optimality of Bayesian Methods

28.13.1 Introduction to Bayesian Methods

In this lecture, we revisit some concepts from Bayesian decision theory and introduce the set up.

Definition 28.372 *(Loss function). Let $\underset{\sim}{X}$ be the data and $\tilde{\underset{\sim}{\theta}}(\underset{\sim}{X})$ be our estimator. We define the loss function to be*

$$D(\tilde{\underset{\sim}{\theta}}(\underset{\sim}{X})$$

.

Definition 28.373 *(Risk). Let $D(\tilde{\underset{\sim}{\theta}}(\underset{\sim}{X})$ be the loss function associated to the estimator $\tilde{\underset{\sim}{\theta}}(\underset{\sim}{X})$. Then, the Bayes risk is the expectation of the loss function*

$$R(\underset{\sim}{\theta}|\tilde{\underset{\sim}{\theta}}(\underset{\sim}{X})) = \mathbb{E}_{\underset{\sim}{\theta}}[D(\tilde{\underset{\sim}{\theta}}(\underset{\sim}{X})|\underset{\sim}{\theta}].$$

Definition 123: Conditional Bayes Risk

Let $w(.) \geq 0$ be a weight function defined on the parameter space. The conditional Bayes risk is defined as

$$B_w(\tilde{\underset{\sim}{\theta}}(\underset{\sim}{X})) = \int_{\Theta} w(\underset{\sim}{\theta}) R(\underset{\sim}{\theta}|\tilde{\underset{\sim}{\theta}}(\underset{\sim}{X}))d\underset{\sim}{\theta}$$

Definition 124: Bayesian Method

Given a prior of weight function $w(.) \geq 0$ defined on the parameter space. A Bayesian method $\hat{\underset{\sim}{\theta}}_B$ is one which minimises the conditional Bayes Risk

$$\int ... \int D(\tilde{\underset{\sim}{\theta}}(\underset{\sim}{X}); \underset{\sim}{\theta}) q_n(\underset{\sim}{\theta}|\underset{\sim}{X}) d\underset{\sim}{\theta}$$

where

$$q_n(\underset{\sim}{\theta}|\underset{\sim}{X}) = \frac{w(\underset{\sim}{\theta}) p(\underset{\sim}{X}; \underset{\sim}{\theta})}{\int ... \int w(\underset{\sim}{\theta}) p(\underset{\sim}{X}; \underset{\sim}{\theta})} d\underset{\sim}{\theta}$$

is the posterior density.

Remark 28.374 *Note that we do not require that the weight function $w(.)$ is in L^1 but we do require that the posterior density $q_n(\underset{\sim}{\theta}|\underset{\sim}{X}) \in L^1$.*

Proposition 109: Bayes estimator is the posterior mean

When the loss function is the squared error loss $(\underset{\sim}{\theta}_1 - \underset{\sim}{\theta}_2)^T \underset{\sim}{M} (\underset{\sim}{\theta}_1 - \underset{\sim}{\theta}_2)$ for a positive definite matrix $\underset{\sim}{M}$, then the Bayes estimator is the posterior mean

$$\hat{\underset{\sim}{\theta}}_B = \int \underset{\sim}{\theta} q_n(\underset{\sim}{\theta}|\underset{\sim}{X}) d\underset{\sim}{\theta}.$$

We will show in the next sections that if we strengthen the ULAN condition, we can show that the posterior mean $\hat{\underset{\sim}{\theta}}_B$ is an asymptotically efficient estimator.

STAT4028: Probability and Mathematical Statistics

29. Hellinger Distance

29.13.2 Hellinger Distance

In this lecture, we will introduce the Hellinger distance. From this, we state propositions which bounds the likelihood ratio as we move away from the true value. We can then introduce the Bayesian ULAN assumption, which we require in order to take integrals on the parameter space.

Definition 125: Hellinger Distance

Suppose that $\underset{\sim}{X}_1, ..., \underset{\sim}{X}_n$ are i.i.d with density function p(.) that is absolutely continuous with respect to a measure $\underset{\sim}{\nu}(.)$, and let $q(.)$ be another density function absolutely continuous with respect to $\nu(.)$. Then, the Hellinger distance is defined as

$$H(p,q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2 d\nu}.$$

Remark 29.375 *The Hellinger distance is a proper metric.*

Proposition 29.376 *(Large deviation on likelihood ratio). Suppose $\underset{\sim}{X}_1, ..., \underset{\sim}{X}_n$ are i.i.d with density $p(.)$ with respect to a measure $\nu(.)$ and let $q(.)$ be another density such that the $L_2(\underset{\sim}{\nu})-$distance*

$$\sqrt{\int [\sqrt{p} - \sqrt{q}]^2 d\nu} = \epsilon > 0.$$

Then, the likelihood ratio

$$\mathbb{P}\Big(\prod_{i=1}^{n} \frac{q(\underset{\sim}{X}_i)}{p(\underset{\sim}{X}_i)} \geq e^{-\frac{n\epsilon^2}{3}} \Big) \leq e^{-\frac{n\epsilon^2}{3}}.$$

Remark 29.377 *This proposition states that if we look at the likelihood ratio away from the true density, then it is exponentially small with an exponentially high probability.*

From the above proposition, if we have a fixed density q, where $q \neq p$, then it becomes trivially small really quickly as $n \to \infty$. Therefore, we require the density q to depend on n. We now want to look at densities that are essentially like a n^{-1} distance away from the true density p(.) as anything further than that goes to zero. Therefore, we can also state an uniform version of the large deviation proposition.

Proposition 110: Uniform bound of large deviations of likelihood ratio

Suppose that a family of probability densities

$$\mathcal{P} = \{p_{\underset{\sim}{\theta}}\}$$

is indexed by a d-dimensional parameter $\underset{\sim}{\theta}$ is such that $\sqrt{p_{\underset{\sim}{\theta}}}$ is $L_2(\nu)-$differentiable at $\underset{\sim}{\theta_0}$. Then, there exists positive universal constants $A, C_1, C_2, C_3 = C_3(\mathcal{P})$ such that for sufficiently large n, if $\epsilon > C_3 n^{-1/2}$ where $C_3 n^{-1/2}$ is a threshold value

$$\mathbb{P}\Big\{ \sup_{\substack{q \in \mathcal{P} \\ \sqrt{\int [\sqrt{q} - \sqrt{p_{\underset{\sim}{\theta_0}}}]^2} \geq \epsilon}} \prod_{i=1}^{n} \frac{q(\underset{\sim}{X_i})}{p_{\underset{\sim}{\theta_0}}(\underset{\sim}{X_i})} \geq e^{-C_1 n \epsilon^2} \Big\} \leq A e^{-C_2 n \epsilon^2}$$

Remark 29.378 *Note that this is the same as our previous proposition on the large deviation of the likelihood ratio. However, we are now expanding from a fixed density q to now that the likelihood ratio is small for any density q outside a ball of ϵ using the Hellinger distance.*

The reason we have this proposition on the uniform bound of large deviations of the likelihood ratio is that in our Bayesian methods, we need to integrate over the entire parameter space, and therefore we need to control the likelihood-ratio for when we are far away from the true value. This proposition will help us disregard anything that is "far away" from the true value.

However, it may be difficult in working with Hellinger distance. Therefore, the following lemma shows how we can convert between Hellinger distance and distance in the parameter space.

Lemma 29.379 *Suppose that a family of probability densities*

$$\mathcal{P} = \{p_{\underset{\sim}{\theta}}\}$$

is indexed by a d-dimensional parameter $\underset{\sim}{\theta}$ is such that $\sqrt{p_{\underset{\sim}{\theta}}}$ is $L_2(\nu)-$differentiable at $\underset{\sim}{\theta_0}$. Furthermore, assume that for a vector $\underset{\sim}{\ell^\circ}$ of functions

$$\frac{\int \left[\sqrt{p_{\underset{\sim}{\theta_0} + \underset{\sim}{h}}} - \sqrt{p_{\underset{\sim}{\theta_0}}} - \frac{1}{2} \underset{\sim}{h}^T \underset{\sim}{\ell^\circ} \sqrt{p_{\underset{\sim}{\theta_0}}} \right]^2}{|h|^2} \leq g(|h|)$$

for a function $g(x) \downarrow 0$ as $x \to 0$. Furthermore, assume that the covariance of the score functions

$$\underset{\sim}{J} = \int \underset{\sim}{\ell^\circ} \underset{\sim}{\ell^{\circ T}} p_{\underset{\sim}{\theta_0}} d\nu$$

is positive definite. Then, there exists constants $0 < \delta < \infty$ and $0 < K_1 < K_2 < \infty$ such that for $|h| \leq \delta$,

$$K_1 |h|^2 \leq \int \left[\sqrt{p_{\underset{\sim}{\theta_0} + \underset{\sim}{h}}} - \sqrt{p_{\underset{\sim}{\theta_0}}} \right]^2 \leq K_2 |h|^2.$$

Remark 29.380 *This lemma tells us that we can convert between the Hellinger distance and the usual Euclidean distance in the parameter space as long as the distances are small enough. That is, if we have a ball in Hellinger distance bounded by a small h, then we can approximate it with a ball in Euclidean distance.*

First, we state a lemma which we will need for the Bayesian ULAN condition.

Lemma 29.381 *Suppose that* $J_n \xrightarrow{p} J$ *and* $S_n \xrightarrow{d} \mathcal{N}(0, J)$. *Then, with probability* $\to 1$, *the ellipsoid*

$$E_1 = \{\underset{\sim}{h} : (\underset{\sim}{h} - J_n^{-1} S_n) J_n (\underset{\sim}{h} - J_n^{-1} S_n) \leq n^{\epsilon}\}$$

is included in the ball $\{\underset{\sim}{h} : |\underset{\sim}{h}| \leq n^{\epsilon}\}$.

Remark 29.382 *This is an ellipsoid centered at* $J_n^{-1} S_n$ *with size* n^{ϵ}.

We can now strengthen the ULAN condition for it to work in our Bayesian set up where we now still require the remainder to go to 0 but at a slower rate.

Theorem 75: Bayesian ULAN

Let us denote the likelihood-ratio

$$e^{L_n(\underset{\sim}{\theta}_0 + n^{-1/2} \underset{\sim}{h}; \underset{\sim}{\theta}_0)} = \prod_{i=1}^{n} \frac{p_{\underset{\sim}{\theta}_0 + n^{-1/2} \underset{\sim}{h}}(\underset{\sim}{X}_i)}{p_{\underset{\sim}{\theta}_0}(\underset{\sim}{X}_i)}.$$

Then, assume that

1. A positive definite symmetric matrix $\underset{\sim}{J}$
2. A score vector $\underset{\sim}{S}_n \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{J})$ under $\underset{\sim}{\theta}_0$
3. $\underset{\sim}{J}_n \xrightarrow{p} \underset{\sim}{J}$ under $\underset{\sim}{\theta}_0$
4. There exists $0 < \epsilon < \frac{1}{2}$ such that the remainder R_n given by

$$L_n(\underset{\sim}{\theta}_0 + n^{-1/2} \underset{\sim}{h}; \underset{\sim}{\theta}_0) = \underset{\sim}{h}^T \underset{\sim}{S}_n - \frac{1}{2} \underset{\sim}{h}^T \underset{\sim}{J}_n \underset{\sim}{h} + R_n(\underset{\sim}{h})$$

which satisfies that

$$\sup_{|h| \leq n^{\epsilon}} |R_n(\underset{\sim}{h})| \xrightarrow{p} 0.$$

Remark 29.383 *So note that condition 4 of the Bayesian rate is no longer for bounded* h *but for* h *that goes to infinity at a slow rate.*

STAT4028: Probability and Mathematical Statistics

30. Bayes Posterior Mean

30.13.3 Bayes Posterior Mean

Theorem 76: Bayesian Posterior Mean Theorem

Assume that $\underset{\sim}{X}_1, ..., \underset{\sim}{X}_n$ are i.i.d with common density p_{θ_0}. Furthermore, assume that the family $\{p_\theta\}$ is $L_2(\nu)-$differentiable at θ_0 and that the Bayesian ULAN conditions hold at θ_0. Finally, assume that the prior $w(\theta)$ is integrable, $\int ... \int |\theta| w(\theta) d\theta < \infty$ and $w(.)$ is continuous at θ_0.
Then, if $\hat{\theta}_B$ is the corresponding posterior mean, then we have that

$$\sqrt{n}(\hat{\theta}_B - \theta_0) = \underset{\sim}{J_n}^{-1} \underset{\sim}{S_n} + o_p(1)$$

STAT4028: Probability and Mathematical Statistics

31. Expectation-Maximization Algorithm

31.13.4 Expectation-Maximization Algorithm

The EM algorithm is useful for finding out the maximum likelihood estimates or the maximum a posteriori (MAP) estimates of parameters in a latent variable model. Many existing approaches to this kind of problem can be unified as a special case of the EM algorithm.

We first recall some concepts we will need.

Definition 31.384 *(Convex Function). A convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is such that for all $\underset{\sim}{x}, \underset{\sim}{y} \in \mathbb{R}^d$ and all $0 \leq w \leq 1$ where*

$$f((1-w)\underset{\sim}{x} + w\underset{\sim}{y}) \leq (1-w)f(\underset{\sim}{x}) + wf(\underset{\sim}{y}).$$

We now state that every convex function has tangent lines to points of the function.

Proposition 31.385 *(Tangent line). A property of convex function as that at every $\underset{\sim}{x}_0$ where $f(\underset{\sim}{x}_0) < \infty$, there exists a linear function*

$$g(\underset{\sim}{x}) = f(\underset{\sim}{x}_0) + \underset{\sim}{v}^T(\underset{\sim}{x} - \underset{\sim}{x}_0)$$

which satisfies

1. $g(\underset{\sim}{x}_0) = f(\underset{\sim}{x}_0)$
2. $g(\underset{\sim}{x}) \leq f(\underset{\sim}{x})$ *for all* $\underset{\sim}{x} \in \mathbb{R}^d$

Remark 31.386 *Such a linear function g(.) need not be unique.*

Corollary 31.387 *Suppose that the convex function f(.) is differentiable. Then, we can take the gradient vector $\underset{\sim}{f}^\circ$ to construct the linear function*

$$g(\underset{\sim}{x}) = f(\underset{\sim}{x}_0) + \underset{\sim}{f}^{\circ T}(\underset{\sim}{x} - \underset{\sim}{x}_0)$$

Proposition 111: Jensen's Inequality

Suppose $\underset{\sim}{X}$ is a d-dimensional random vector such that $\underset{\sim}{\mu} = \mathbb{E}[\underset{\sim}{X}]$ exists. If f(.) is a convex function, then

$$\mathbb{E}[f(\underset{\sim}{X})] \geq f(\mathbb{E}[\underset{\sim}{X}]) = f(\underset{\sim}{\mu})$$

Corollary 31.388 *Using Jensen's inequality, we can show that*

$$\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$$

for a random variable $X \in L^2(\Omega)$.

We now introduce a measure of the difference between two density functions.

Definition 126: Kullback-Leibler Divergence

Suppose that $p(.)$ is the density with respect to a $\sigma-$finite measure $\nu(.)$ of a probability distribution over $\mathbb{R}^d$. Let $\mathcal{X}$ denote the support of the density $p(.)$ and suppose $q(.)$ is another probability density with respect to the measure $\nu(.)$ with the same support $\mathcal{X}$. Then, the Kullback-Leibler divergence is given by

$$\int_{\S} log\Big\{\frac{p(\underset{\sim}{x})}{q(\underset{\sim}{x})}\Big\}p(\underset{\sim}{x})d\nu(\underset{\sim}{x})$$

which equals 0 if $p = q$ almost everywhere.

Remark 31.389 *The KL-divergence is not symmetric nor does it necessarily satisfy the triangle-inequality. Therefore, it is not a proper metric.*

The KL-divergence from q(.) to the reference density p(.) measures how different q(.) is to p(.).

We can now describe the EM algorithm. First, we describe the setup.

Suppose we have 2 measurable spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$. Let $\nu(.)$ and $\lambda(.)$ be $\sigma-$finite measures on $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ respectively.

Let $p(\underset{\sim}{x}, \underset{\sim}{y}|\underset{\sim}{\theta})$ be a parametric family of joint probability densities with respect to the product measure, which assigns to the product set $A \times B$, the measure $\nu(A)\lambda(B)$, where $A \in \mathcal{A}$ and $B \in \mathcal{B}$.

We model $\underset{\sim}{y}$ as the observed values of $\underset{\sim}{Y}$ and let $\underset{\sim}{X}$ be the unobserved variables. $\underset{\sim}{X}$ is also known as **latent variables**.

Definition 127: Marginal density of observed data

The marginal density of the observed data $\underset{\sim}{Y}$ is

$$f(\underset{\sim}{Y}|\underset{\sim}{\theta}) = \int p(\underset{\sim}{X}, \underset{\sim}{Y}|\underset{\sim}{\theta})d\nu(\underset{\sim}{X})$$

Definition 128: Conditional density of latent variable

The conditional density of the latent variable on the observed data, $\underset{\sim}{X}$ given $\underset{\sim}{Y} = \underset{\sim}{y}$ is

$$q(\underset{\sim}{X}|\underset{\sim}{y}, \underset{\sim}{\theta}) = \frac{p(\underset{\sim}{X}, \underset{\sim}{Y}|\underset{\sim}{\theta})}{f(\underset{\sim}{Y}|\underset{\sim}{\theta})}$$

which assigns to the set $A \in \mathcal{A}$ the probability

$$\int_A q(\underset{\sim}{X}|\underset{\sim}{y}, \underset{\sim}{\theta}) d\nu(\underset{\sim}{X}).$$

The goal of the EM algorithm is that there parameters $\underset{\sim}{\theta}$ which we wish to infer.

Suppose we wish to compute the maximum likelihood estimate

$$\underset{\sim}{\hat{\theta}} = \max_{\underset{\sim}{\theta}} f(\underset{\sim}{y}|\underset{\sim}{\theta})$$

However, this is an issue with this.

Proposition 31.390 *Computing the maximum likelihood estimate for the marginal density of the observed data*

$$\underset{\sim}{\hat{\theta}} = \max_{\underset{\sim}{\theta}} f(\underset{\sim}{y}|\underset{\sim}{\theta})$$

is NP-hard.

However, maximizing the complete data likelihood is relatively easy to do. That is, it is hard to maximize the marginal density of the observed data $f(\underset{\sim}{y}|\underset{\sim}{\theta})$ but it is alot easier to maximize the complete data density $p(\underset{\sim}{X}, \underset{\sim}{Y}|\underset{\sim}{\theta})$, however, we do not know what $\underset{\sim}{X}$ is which is what the EM algorithm tries to solve!

Definition 129: EM Log Likelihood

The conditional expectation of the complete data log likelihood is known as the EM log likelihood

$$\ell_{EM}(\underset{\sim}{\theta}|\underset{\sim}{\theta}_0) = \int log\ p(\underset{\sim}{x}, \underset{\sim}{y}|\underset{\sim}{\theta}) q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}_0) d\nu(\underset{\sim}{x})$$

where $\underset{\sim}{\theta}_0$ is a guess. That is, it is taking the conditional expectation of the log likelihood of the joint density under the conditional density of the latent variable on the observed data. Equivalently, we can express it as

$$\mathbb{E}_{q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}_0)}[log\ p(\underset{\sim}{X}, \underset{\sim}{y}|\underset{\sim}{\theta})|\underset{\sim}{Y} = \underset{\sim}{y}]$$

We can now state an importat theorem that guarantees why does the EM algorithm work.

Theorem 77: Fundamental EM-Algorithm Lemma

Suppose that $\underset{\sim}{\theta}_1$ is such that for the EM log likelihood, we have that

$$\ell_{EM}(\underset{\sim}{\theta}_1|\underset{\sim}{\theta}_0) \geq \ell_{EM}(\underset{\sim}{\theta}_0|\underset{\sim}{\theta}_0).$$

Then, the marginal density of the observed data has the relationship

$$f(\underset{\sim}{y}|\underset{\sim}{\theta}_1) \geq f(\underset{\sim}{y}|\underset{\sim}{\theta}_0)$$

Proof: First, note that we can express the log complete data density as

$$log\ p(\underset{\sim}{x}, \underset{\sim}{y}|\underset{\sim}{\theta}) = log\ f(\underset{\sim}{y}|\underset{\sim}{\theta}) + log\ q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta})$$

Then, we have that

$$\ell_{EM}(\underset{\sim}{\theta}|\underset{\sim}{\theta}_0) = \int \Big[log\ p(\underset{\sim}{x}, \underset{\sim}{y}|\underset{\sim}{\theta})q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}_0)\Big]d\nu(\underset{\sim}{x}) = \int \Big[(log\ f(\underset{\sim}{y}|\underset{\sim}{\theta}) + log\ q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}))q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}_0)\Big]d\nu(\underset{\sim}{x})$$

$$= log\ f(\underset{\sim}{y}|\underset{\sim}{\theta}) + \int log\ q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta})q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}_0)d\nu(\underset{\sim}{x})$$

Therefore, we can write the hypothesis of the theorem using these expressions

$$\ell_{EM}(\underset{\sim}{\theta}_1|\underset{\sim}{\theta}_0) \geq \ell_{EM}(\underset{\sim}{\theta}_0|\underset{\sim}{\theta}_0)$$

$$log\ f(\underset{\sim}{y}|\underset{\sim}{\theta}_1) + \int log\ q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}_1)q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}_0)d\nu(\underset{\sim}{x}) \geq log\ f(\underset{\sim}{y}|\underset{\sim}{\theta}_0) + \int log\ q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}_0)q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}_0)d\nu(\underset{\sim}{x})$$

$$log\ f(\underset{\sim}{y}|\underset{\sim}{\theta}_1) - log\ f(\underset{\sim}{y}|\underset{\sim}{\theta}_0) \geq \int log\left\{\frac{q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}_0)}{q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}_1)}\right\} \cdot q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}_0)d\nu(\underset{\sim}{x}) \geq 0$$

where the last inequality arises from the definition of the KL-divergence theorem. ■

Remark 31.391 *What the fundamental EM-algorithm lemma states is that we can increase the observed data likelihood by replacing* $\underset{\sim}{\theta}_0$ *with* $\underset{\sim}{\theta}_1$.

Theorem 78: EM Algorithm

The EM-algorithm starts at a starting value $\underset{\sim}{\theta}_0$. Then, the update rule for the parameter $\underset{\sim}{\theta}$ is given by

$$\underset{\sim}{\theta}_{j+1} = \max_{\underset{\sim}{\theta}} \ell_{EM}(\underset{\sim}{\theta}|\underset{\sim}{\theta}_j) = \max_{\underset{\sim}{\theta}} \Big\{ \int log\ p(\underset{\sim}{x}, \underset{\sim}{y}|\underset{\sim}{\theta})q(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}_j)d\nu(\underset{\sim}{x}) \Big\}$$

Then, the resultant sequence $\underset{\sim}{\theta}_0, \underset{\sim}{\theta}_1, \underset{\sim}{\theta}_2, ...$ is guaranteed to not decrease the observed data likelihood

$$f(\underset{\sim}{y}|\underset{\sim}{\theta}_0) \leq f(\underset{\sim}{y}|\underset{\sim}{\theta}_1) \leq f(\underset{\sim}{y}|\underset{\sim}{\theta}_2) \leq ...$$

Corollary 31.392 *If a global maximum likelihood estimator exists θ^*, then θ^* is a fixed point of the algorithm.*

Remark 31.393 *It is not true in general that the algorithm converges to the global MLE. Furthermore, the rate of convergence may be very slow.*

STAT4028: Probability and Mathematical Statistics

32. Expectation-Maximization Algorithm Examples

32.13.5 Normal Location Contamination Mixture Model

Let $X_i \sim B(1,\eta)$ and $Y_i = X_i\theta + Z_i$ where $Z_i \sim \mathcal{N}(0,1)$.

Lemma 32.394 *The conditional distribution of Y is*

$$f_{Y|X}(y|x) = \Phi(y)^{1-x}\Phi(y-\theta)^x$$

Proof: Recall that X is either 0 or 1. Therefore, we have that

$$Y_i|X_i = 0 \sim \mathcal{N}(0,1) \quad \text{and} \quad Y_i|X_i = 1 \sim \mathcal{N}(\theta,1)$$

That means

$$\mathbb{P}(Y \le y|X=0) = \Phi(y) \quad \text{and} \quad \mathbb{P}(Y \le y|X=1) = \Phi(y-\theta)$$

Therefore, we can conclude for the conditional distribution

$$\begin{aligned} f_{Y|X}(y|x) &- 1\{x=0\}\Phi(y) + 1\{x=1\}\Phi(y-\theta) \\ &= \Phi(y)^{1-x}\cdot\Phi(y-\theta)^x \end{aligned}$$

■

Proposition 32.395 *The joint density of $\underset{\sim}{X} = (x_1,...,x_n)^T$ and $\underset{\sim}{Y} = (y_1,...,y_n)^T$ is given by*

$$p(\underset{\sim}{x},\underset{\sim}{y}|\theta,\eta) = \prod_{i=1}^{n}\Big[(1-\eta)\Phi(y_i)\Big]^{1-x_i}\Big[\eta\Phi(y_i-\theta)\Big]^{x_i}$$

Proof: Recall that $\mathbb{P}(X=0) = 1-\eta$ and $\mathbb{P}(X=1) = \eta$. Therefore, using the expression for the conditional distribution of Y, we have that

$$p(\underset{\sim}{x},\underset{\sim}{y}|\theta,\eta) = p(\underset{\sim}{x}|\theta,\eta)p(\underset{\sim}{y}|\underset{\sim}{x},\theta,\eta) = \prod_{i=1}^{n}\Big[(1-\eta)\Phi(y_i)\Big]^{1-x_i}\Big[\eta\Phi(y_i-\theta)\Big]^{x_i}$$

■

First, derive the log-likelihood of the complete data joint density

$$log\ p(\underset{\sim}{x},\underset{\sim}{y}|\theta,\eta) = \sum_{i=1}^{n}(1-X_i)log\ \eta + \sum_{i=1}^{n}X_i\ log\ \eta + \theta\sum_{i=1}^{n}x_i\cdot y_i - \frac{\theta^2}{2}\sum_{i=1}^{n}x_i + nlog\frac{1}{\sqrt{2\pi}} - \frac{\sum_{i=1}^{n}y_i^2}{2}$$

Then, we can compute the estimators of the parameters by maximizing this log-likelihood

$$\hat{\eta} = \frac{1}{n}\sum_{i=1}^{n}x_i \quad \hat{\theta} = \frac{\sum_{i=1}^{n}x_i\cdot y_i}{\sum_{i=1}^{n}x_i}$$

We can then compute the marginal density of the observed data $\underset{\sim}{y}$ as

$$f(\underset{\sim}{y}|\eta,\theta) = \int p(\underset{\sim}{x},\underset{\sim}{y}|\eta,\theta)d\nu(\underset{\sim}{x}) = \prod_{i=1}^{n}\Big\{(1-\eta)\Phi(y_i)+\eta\Phi(y_i-\theta)\Big\}$$

We now remark that the observed data log likelihood is extremely difficult to maximize

$$log\ f(\underset{\sim}{y}|\eta,\theta) = \sum_{i=1}^{n} log\ \Big\{(1-\eta)\Phi(y_i)-\eta\Phi(y_i-\theta)\Big\}$$

Therefore, we require the EM algorithm.

(1) First, we need initial estimates of the parameters θ,η. We can use the Method-Of-Moment estimators as our initial guess

$$\hat{\theta} = \frac{\overline{Y}}{\eta} \quad \hat{\eta} = \frac{\overline{Y}^2}{\overline{Y}^2 + V - 1}$$

where V is the sample variance of the observed data $\underset{\sim}{Y}$.

(2) We now compute the conditional density of the latent variable $\underset{\sim}{X}$ given the observed data $\underset{\sim}{Y} = \underset{\sim}{y}$

$$q(\underset{\sim}{x}|\underset{\sim}{y},\theta,\eta) = \frac{p(\underset{\sim}{x},\underset{\sim}{y}|\theta,\eta)}{f(\underset{\sim}{y}|\theta,\eta)} = \prod_{i=1}^{n}\Big\{\Big(\frac{(1-\eta)\Phi(y_i)}{(1-\eta)\Phi(y_i)+\eta\Phi(y_i-\theta)}\Big)^{1-x_i}\Big(\frac{\eta\Phi(y_i-\theta)}{(1-\eta)\Phi(y_i)+\eta\Phi(y_i-\theta)}\Big)^{x_i}\Big\}$$

where we can see that

$$X_i|Y_i = y_i \sim B(1,\gamma_i)$$

where

$$\gamma_i = \frac{\eta\Phi(y_i-\theta)}{(1-\eta)\Phi(y_i)+\eta\Phi(y_i-\theta)}$$

Now, we note that the complete data log likelihood $p(\underset{\sim}{x},\underset{\sim}{y}|\theta,\eta)$ is linear in X_i. Therefore, we have that

$$\mathbb{E}[g(\underset{\sim}{X})|\underset{\sim}{Y} = \underset{\sim}{y}]$$

but $g(\underset{\sim}{X})$ is of the form $\sum_{i=1}^{n} h(X_i)$ as the data is i.i.d and $X_i|Y_i$ are independent

$$= \sum_{i=1}^{n}\mathbb{E}[h(X_i)|Y_i = y_i]$$

and as $h(x_i) = a + bX_i$ as seen in the complete data log likelihood, we have that

$$= a + b\sum_{i=1}^{n}\mathbb{E}[X_i|Y_i = y_i] = a + b\sum_{i=1}^{n}\gamma(y_i)$$

as $X_i|Y_i = y_i \sim B(1,\gamma(y_i))$. Therefore, going back to our log likelihood of the complete data,

$$log\ p(\underset{\sim}{x},\underset{\sim}{y}|\theta,\eta) = \sum_{i=1}^{n}(1-X_i)log\ \eta + \sum_{i=1}^{n}X_i\ log\ \eta + \theta\sum_{i=1}^{n}x_i\cdot y_i - \frac{\theta^2}{2}\sum_{i=1}^{n}x_i + nlog\frac{1}{\sqrt{2\pi}} - \frac{\sum_{i=1}^{n}y_i^2}{2}$$

we replace x_i by the expression

$$\gamma_i^\circ = \mathbb{E}_{\theta_0,\eta_0}[X_i|Y_i = y_i]\frac{\eta_0\Phi(y_i-\theta_0)}{(1-\eta_0)\Phi(y_i)+\eta_0\Phi(y_i-\theta_0)}$$

to arrive at

$$log_{\theta,\eta}(\theta,\eta|\theta_0,\eta_0) = log\eta\sum_{i=1}^{n}\gamma_i^\circ + log(1-\eta)(n-\sum_{i=1}^{n}\gamma_i^\circ) + n\ log\frac{1}{\sqrt{2\pi}} - \frac{n\sum_{i=1}^{n}y_i^2}{2} + \theta\sum_{i=1}^{n}\gamma_i^\circ y_i - \frac{\theta^2}{2}\sum_{i=1}^{n}\gamma_i^\circ$$

Maximizing this gives us

$$\eta_1 = \frac{1}{n}\sum_{i=1}^{n}\gamma_i^\circ \quad \theta_1 = \frac{\sum_{i=1}^{n}\gamma_i^\circ y_i}{\sum_{i=1}^{n}\gamma_i^\circ}$$

We summarise the steps taken.

Let $\underset{\sim}{x}$ be the latent variable and let $\underset{\sim}{y}$ be the observed variable.

1. Derive the complete data log-likelihood $log\ p(\underset{\sim}{x},\underset{\sim}{y}|\theta,\eta)$
2. Derive the estimators of the parameters $\hat{\theta},\hat{\eta}$
3. Derive the marginal density of the observed variable $f(\underset{\sim}{y}|\theta,\eta)$
4. Determine your starting values of your parameters θ_0,η_0 (e.g. using M.O.M estimators)
5. Determine the conditional distribution of the latent variable given the observed variable

$$q(\underset{\sim}{x}|\underset{\sim}{y},\theta,\eta) = \frac{p(\underset{\sim}{x},\underset{\sim}{y}|\theta,\eta)}{f(\underset{\sim}{y}|\theta,\eta)}$$

6. Carry out the "expectation step"

$$\gamma_i^\circ = \mathbb{E}_{\theta_0,\eta_0}\left[\underset{\sim}{X_i}|\underset{\sim}{Y_i} = \underset{\sim}{y_i}\right]$$

7. Maximize the complete data log-likelihood by replacing the latent variable with γ_i° to get

$$log_{\theta\eta}p(\theta,\eta|\theta_0,\eta_0)$$

to find the estimates of the parameters $\hat{\theta},\hat{\eta}$

STAT4028: Probability and Mathematical Statistics

0. Prerequisite

0.14 Measure Theory Recap

0.14.1 Classes of subsets

For this whole chapter, we denote Ω as the space.

Definition 0.396 *(Powerset). We define as the set of all subsets of Ω as the **powerset** $\mathcal{P} = 2^{\Omega}$.*

Definition 0.397 *(Semi-algebra). The class $\mathcal{S} \subseteq \mathcal{P}$ is a **semi-algebra** if*

1. $\Omega \in \mathcal{P}$
2. *If $A, B \in \mathcal{S}$, then $A \cap B \in \mathcal{S}$*
3. *If $A \in \mathcal{S}$, then there exists a finite disjoint union of sets $E_1, ...E_n$ where $E_i \in \mathcal{S}$ for $1 \leq i \leq n$ such that $A^c = \bigcup\!\!\!\cdot_{i=1}^{n} E_i$*

Remark 0.398 *In some text, a semi-algebra is also known as a **semi-ring**.*

Example 0.399 *(Intervals on $\mathbb{R}$ is a semi-algebra). Let $\Omega = \mathbb{R}$. We then define the class of sets to include*

1. $\{(a, b] : a < b; a, b \in \mathbb{R}\} \in \mathcal{S}$
2. $\{(-\infty, b] : b \in \mathbb{R}\} \in \mathcal{S}$
3. $\{(a, \infty] : a \in \mathbb{R}\} \in \mathcal{S}$
4. $\emptyset, \mathbb{R} \in \mathcal{S}$

Then, $\mathcal{S}$ is a semi-algebra.

Definition 0.400 *(Algebra). The class $\mathcal{A} \subseteq \mathcal{P}$ is an **algebra** if*

1. $\Omega \in \mathcal{A}$
2. *If $A, B \in \mathcal{A}$ then $A \cap B \in \mathcal{A}$*
3. *If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$.*

Remark 0.401 *In some text, an algebra is also known as a **ring**.*

Claim 0.402 *If $\mathcal{A}$ is an algebra, then $\mathcal{A}$ is a semi-algebra.*

Proposition 0.403 *If $E, F \in \mathcal{A}$ then $E \cup F \in \mathcal{A}$.*

We now move on to our main interest in measure theory.

Definition 0.404 *(σ-algebra). Let $\mathcal{F} \subseteq \mathcal{P}$. $\mathcal{F}$ is a $\sigma-$algebra if*

1. $\Omega \in \mathcal{F}$
2. *If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$*
3. *If $A_j \in \mathcal{F}$ then $\cup_{j \geq 1} A_j \in \mathcal{F}$.*

Claim 0.405 *If $\mathcal{F}$ is a σ-algebra, then $\mathcal{F}$ is an algebra.*

Proposition 0.406 *Let $\mathcal{A}_i \subseteq \mathcal{P}$ where $i \in I$ is an index set and $\mathcal{A}_i$ is an algebra. Then*

$$\mathcal{A} = \bigcap_{i \in I} \mathcal{A}_i$$

is an algebra.

Remark 0.407 *This proposition also holds for the intersection of $\sigma-$algebras $\mathcal{F}_i$ is a $\sigma-$algebra.*

Definition 0.408 *(The algebra generated by a set). Let $\mathcal{C} \subseteq \mathcal{P}$ be a collection of sets. Then, the algebra generated by $\mathcal{C}$ is defined as*

$$\mathcal{A}(\mathcal{C}) = \bigcap_{i \in I} \mathcal{A}_i$$

where $\mathcal{A}_i$ are algebras which contains $\mathcal{C}$.

Proposition 0.409 *(Properties of algebras generated by a set). Let $\mathcal{C} \subset \mathcal{P}$ be a collection of sets. Let $\mathcal{A}$ be the algebra generated by $\mathcal{C}$. Then*

1. *$\mathcal{A}$ contains $\mathcal{C}$*
2. *For any other algebra $\beta \supseteq \mathcal{C}$, then $\beta \supseteq \mathcal{A}$. That is, $\mathcal{A}$ is the smallest algebra that contains $\mathcal{C}$.*

Remark 0.410 *These ideas also apply to the $\sigma -$ algebra $\mathcal{F}$ generated by a set $\mathcal{C} \subseteq \mathcal{P}$.*

We now see that the algebra $\mathcal{A}(\mathcal{S})$ generated by a semi-algebra $\mathcal{S}$ is **simple** in the sense that any set $E \subset \mathcal{A}(\mathcal{S})$ can be expressed as a finite union of sets in the semi-algebra $\mathcal{S}$. However, this property does not hold for the $\sigma - algebra$ generated by the semi algebra $\mathcal{F}(\mathcal{S})$.

Proposition 0.411 *(The Algebra generated by semi-algebra is simple). Let $\mathcal{S} \subseteq \mathcal{P}$ be a semi-algebra. Let $\mathcal{A}(\mathcal{S})$ be the algebra generated by $\mathcal{S}$. Then, a set $A \in \mathcal{A}(\mathcal{S})$ if and only if there exists a finite disjoint collection of sets $E_1, ..., E_n$ where $E_i \in \mathcal{S}$ for $1 \leq i \leq n$ where*

$$A = \bigcupdot_{i=1}^{n} E_i.$$

Proof:(Sketch). We can take the set

$$\beta = \{\bigcupdot_{j=1}^{n} F_j : F_j \in \mathcal{S}\}$$

where $\beta \subseteq \mathcal{P}$. We can show that

1. β is an algebra;
2. $\beta \supset \mathcal{S}$.

Hence, this implies that

$$\beta \supset \mathcal{A}(\mathcal{S}).$$

Hence, we can take any set $E \in \mathcal{A}(\mathcal{S})$ and the above implies that $E \in \beta$ and hence can be written as disjoint union of sets in $\mathcal{S}$. ■

This is significant as any element in $\mathcal{A}(\mathcal{S})$ can be written as a finite disjoint union of elements in the semi-algebra $\mathcal{S}$.

0.14.2 Functions on Sets

Definition 0.412 *(Additive set function). Let $\mathcal{C} \subseteq \mathcal{P}$ be a class of sets. Furthermore, assume that $\emptyset \in \mathcal{C}$. Define the set function*

$$\mu : \mathcal{C} \to \overline{\mathbb{R}}^{+}.$$

mu is ***additive*** *if*

1. $\mu(\emptyset) = 0$;
2. *Let $E_1, ..., E_n \in \mathcal{C}$ be disjoint sets where $E = \bigcupdot_{j=1}^{n} E_j$ and $E \in \mathcal{C}$. Then*

$$\mu(E) = \sum_{j=1}^{n} E_j.$$

Definition 0.413 *(σ-additive set function). Let $\mathcal{C} \subseteq \mathcal{P}$ be a class of sets. Furthermore, assume that $\emptyset \in \mathcal{C}$. Define the set function*

$$\mu : \mathcal{C} \to \overline{\mathbb{R}}^{+}.$$

mu is ***σ-additive*** *if*

1. $\mu(\emptyset) = 0$;
2. *Let $\{E_i\}_{i \geq 1} \in \mathcal{C}$ be disjoint sets where $E = \bigcupdot_{j \geq 1} E_j$ and $E \in \mathcal{C}$. Then*

$$\mu(E) = \sum_{j=1}^{\infty} E_j.$$

The second property is what is referred to as $\sigma-$additivity.

Remark 0.414 *$\sigma-$additivivty is the same as additivity except that the second condition holds for a* ***countably infinite*** *number of disjoint sets. Some text calls additivity,* ***finite-additivity*** *and $\sigma-$additivity,* ***countably additive****.*

Observation 0.415 *We can relax the first condition required for addivity. If there exists $A \in \mathcal{C}$ such that $\mu(A) < \infty$, then we can express $A = A \cup \emptyset$. Then*

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

by addivity. Then, as $\mu(A) < \infty$ by assumption, we have that $\mu(\emptyset) = 0$. Hence, as long as we have a set A such that $\mu(A) < \infty$, then condition 1 follows from condition 2 for a (σ) additive set function.

Lemma 0.416 *Let μ be a $\sigma-$additive set function. If $E \subseteq F$, then*

$$\mu(E) \leq \mu(F).$$

Proposition 0.417 *If μ is a $\sigma-$additive set function, then μ is an additive set function.*

Example 0.418 *(An additive set function that is not σ-additive). Let $\Omega = (0, 1)$. Let $\mathcal{C} = \{(a, b] : 0 \leq a < b < 1\}$. Define the set function $\mu : \mathcal{C} \to \overline{\mathbb{R}}^+$ as*

$$\mu((a, b]) = \begin{cases} +\infty & \text{if } a = 0 \\ b - a & \text{if } a > 0 \end{cases}$$

Let $\{E_n\}_{ngeq1}$ be a sequence of sets. We will write $E_n \uparrow E$ if $E_n \subseteq E_{n+1}$ and $\bigcup_{n\geq 1} E_n = E$.

We will write $E_n \downarrow E$ if $E_n \supseteq E_{n+1}$ and $\bigcap_{n\geq 1} E_n = E$.

Definition 0.419 *(Continuous from above). Let $\mathcal{C} \subseteq \mathcal{P}$ and $\mu : \mathcal{C} \to \overline{\mathbb{R}}^+$. Take $E \in \mathcal{C}$. Then, μ is* ***continuous from below*** *at E if for all sequences $\{E_j\}_{j\geq 1}$ such that $E_j \in \mathcal{C}$ and $E_n \uparrow E$, we have that*

$$\mu(E_n) \to \mu(E)$$

as $n \to \infty$.

Definition 0.420 *(Continuous from below). Let $\mathcal{C} \subseteq \mathcal{P}$ and $\mu : \mathcal{C} \to \overline{\mathbb{R}}^+$. Take $E \in \mathcal{C}$. Then, μ is* ***continuous from below*** *at E if for all sequences $\{E_j\}_{j\geq 1}$ such that $E_j \in \mathcal{C}$, $E_n \downarrow E$, and $\mu(E_0) < \infty$, we have that*

$$\mu(E_n) \to \mu(E)$$

as $n \to \infty$.

The set function μ is continuous if it is continuous from above and below.

We now state a proposition to help us determine if a function is $\sigma-$additive.

Proposition 0.421 *(Tests for a function being $\sigma-$additive). Let $\mathcal{A} \subseteq \mathcal{P}$ be an algebra. Define the* ***additive*** *set function $\mu : \mathcal{A} \to \overline{\mathbb{R}}^+$. Then, we have the following:*

1. *μ is $\sigma-$additive if μ is continuous at E for all $E \in \mathcal{A}$;*
2. *If μ is continuous from below, then μ is $\sigma-$additive;*
3. *If μ is continuous from above at $\emptyset$ and μ is finite, then μ is $\sigma-$additive.*

Remark 0.422 *We generally use (2) and (3) to show an additive set function is $\sigma-$additive.*

0.14.3 Extension Theorems

We now look at extending an additive set function on a semi-algebra to an algebra with this extension being unique.

Theorem 0.423 *(Simple extension theorem). Let $\mathcal{S} \subseteq \mathcal{P}$ be a semi-algebra. Define $\mu : \mathcal{S} \to \overline{\mathbb{R}}^+$ to be an additive set function. Let $\mathcal{A}(\mathcal{S})$ be the algebra generated by the semi-algebra. Then, there exists a set function $\nu : \mathcal{A}(\mathcal{S}) \to \overline{\mathbb{R}}^+$ such that*

1. *ν is additive;*
2. *$\nu(A) = \mu(A)$ for all $A \in \mathcal{S}$ (ν is an extension of μ)*
3. *ν is an **unique** extension of μ onto $\mathcal{A}(\mathcal{S})$.*

Observation 0.424 *To define the extension, recall that the collection $\mathcal{A}(\mathcal{S})$ is simple, that is, for every set $E \in \mathcal{A}(\mathcal{S})$, there exists a finite number of disjoint sets $E_j \in \mathcal{S}$ such that $E = \bigcup\!\!\!\cdot_{j=1}^{n} E_j$. Hence, we can define our extension as*

$$\nu(E) = \sum_{j=1}^{n} \mu(E_j).$$

for every set $E \in \mathcal{A}(\mathcal{S})$.

Corollary 0.425 *If $\mu : \mathcal{S} \to \overline{\mathbb{R}}^+$ is $\sigma-$**additive**, then there exists a unique $\sigma-$additive extension $\nu : \mathcal{A}(\mathcal{S}) \to \overline{\mathbb{R}}^+$.*

0.14.4 Caratheodory Theorem

Now, we are interested in taking a $\sigma-$additive set function $\nu : \mathcal{A}(\mathcal{S}) \to \overline{\mathbb{R}}^+$ and extending it to a $\sigma-$additive function $\pi : \mathcal{F}(\mathcal{A}) \to \overline{\mathbb{R}}^+$ where $\mathcal{F}(\mathcal{A})$ is the $\sigma - algebra$ generated by the algebra $\mathcal{A}$. However, it is not as straight forward as what we did when we extended it onto the algebra $\mathcal{A}(\mathcal{S})$ as we cannot express every set $E \in \mathcal{F}(\mathcal{A})$ as a finite disjoint union of sets in $\mathcal{A}$ or $\mathcal{S}$. Hence, we require a more elaborate construction.

We outline the steps of what we will need to do.

1. Define an outer-measure $\pi^* : \mathcal{P} \to \overline{\mathbb{R}}^+$.
2. Define the collection $\mathcal{M} \subseteq \mathcal{P}$ where $\mathcal{M}$ is a $\mathcal{F}$ and that $\mathcal{M} \supseteq \mathcal{F}(\mathcal{A})$.
3. Show that $\pi^*|_{\mathcal{M}}$ is $\sigma-$additive.
4. Show that $\pi^*|_{\mathcal{A}} = \nu$
5. Show that $\pi^*|_{\mathcal{M}}$ is an unique extension (under certain conditions).

Note that the $\sigma - algebra$ generated by semi-algebras and algebras are the same, hence the difference does not matter and we can repeat all of the above with $\sigma - algebra$ generated by semi-algebras $\mathcal{S}$, which actually gives us a more general definition.

Before we proceed, we tie up some definitions.

Definition 0.426 *(Pre-measure). Let $\mathcal{A}$ be an algebra (ring). Let $\mu_0 : \mathcal{A} \to \overline{\mathbb{R}}^+$. The function μ_0 is called a **pre-measure** if*

1. $\mu_0(\emptyset) = 0$
2. *For every countable sequence $\{E_i\}_{i\in\mathbb{N}}$ of pairwise disjoint sets whose union lies in $\mathcal{A}$, then*

$$\mu_0\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \sum_{i=1}^{\infty} \mu_0(E_i).$$

Remark 0.427 *This is what we have called a $\sigma-$additive function this whole time! Additionally, we can drop the first condition and only require $\sigma-$addivity if there exists a set $A \in \mathcal{A}$ such that $\mu_0(A) < \infty$. This is known as a pre-measure as it is defined on an algebra $\mathcal{A}$ rather than a $\sigma-$algebra $\mathcal{F}$.*

Pre-measures will be extremely useful in helping us define outer-measures.

Definition 0.428 *(Outer-measure). Let Ω be the set and $\mathcal{P}$ be the powerset of Ω. Let $\mu : \mathcal{P} \to \overline{\mathbb{R}}^+$ be a set function. μ is an **outer-measure** if*

1. $\mu(\emptyset) = 0$;
2. *(Monotonicity) For any two sets $E, F \in \Omega$, if $E \subseteq F$ then $\mu(E) \leq \mu(F)$;*
3. *(σ-subaddivitiy) For **any** sequence $\{E_j\}_{j\geq 1}$ of subsets of Ω, then*

$$\mu\Big(\bigcup_{j=1}^{\infty} E_j\Big) \leq \sum_{j=1}^{\infty} \mu\Big(E_j\Big).$$

Remark 0.429 *Note that condition (3) require $\sigma-$subaddivitiy and not $\sigma-$additivity.*

Proposition 0.430 *Let Ω be the set and let $\mathcal{A}$ be an algebra of the space Ω. Let ν be the pre-measure we defined on $\mathcal{A}$. Then, we define that for any set $A \subseteq \Omega$, the set function*

$$\pi^*(A) = \inf_{\{E_i\}_{i\geq 1}} \Big\{ \sum_{i\geq 1} \nu(E_i) : A \subseteq \bigcup_{i\geq 1} E_i \text{ and } E_i \in \mathcal{A} \Big\}.$$

Then, π^* is an outer-measure.

Remark 0.431 *Here, we take the infimum over all coverings $\{E_i\}_{i\geq 1}$ of the set A.*

Now,we will define the collection of sets to which we will call measurable.

Definition 0.432 *(μ^*-measurable set). Let $\mu^* : \mathcal{P}(\Omega) \to \overline{\mathbb{R}}^+$ be an outer measure. We call a set $A \subseteq \Omega$ a μ^*-measurable set if*

$$\mu^*(S) = \mu^*(S \cap A) + \mu^*(S \cap A^c)$$

for all sets $S \subseteq \Omega$.

Definition 0.433 *(family of measurable sets). Let $\pi^* : \mathcal{P} \to \overline{\mathbb{R}}^+$ be the outer-measure we defined earlier using the pre-measure ν. We denote the family of measurable sets $\mathcal{M}$ where $A \in \mathcal{M}$ if*

$$\pi^*(E) = \pi^*(E \cap A) + \pi^*(E \cap A^c)$$

for all sets $E \subseteq \Omega$.

Proposition 0.434 *Let Ω be our set of interest and $\mathcal{A}$ be the algebra on Ω. Let π^* be the outer-measure we have defined. $\mathcal{M}$ the family of π^*-measurable sets. Then, we have that*

1. *$\mathcal{M} \supseteq \mathcal{A}$*
2. *$\mathcal{M}$ is a $\sigma - algebra$*
3. *From (1) and (2), we have that*
$$\mathcal{M} \supseteq \mathcal{F}(\mathcal{A}).$$

Proposition 0.435 *Let π^* be the outer-measure we have defined and $\mathcal{M}$ the family of π^*-measurable sets. Additionally, let ν be the pre-measure defined on the algebra $\mathcal{A}$. Then*

$$\pi^*|_{\mathcal{M}} \colon \mathcal{M} \to \overline{\mathbb{R}}^+$$

is $\sigma-$additive and hence a measure π. Furthermore, π is an extension to ν as $pi(A) = \nu(A)$ for all sets $A \in \mathcal{A}$.

We need one more definition before we show uniqueness.

Definition 0.436 *(σ-finite). Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then, μ is $\sigma-$finite if there are sets $\{E_j\}_{j\geq 1}$ such that $E_j \uparrow \Omega$ such that*

$$\mu(E_j) < \infty$$

for all $j \in \mathbb{N}$.

Claim 0.437 *Assume that π is the extension of the pre-measure ν on $\mathcal{A}$. If ν is $\sigma-$finite, then it is the **unique extension** of ν onto $\mathcal{F}(\mathcal{A})$.*

From all this, we can now concisely state Caratheodory's extension theorem.

Theorem 79: Caratheory's Extension Theorem

Let $\mathcal{A}$ be an algebra on Ω and let $\nu : \mathcal{A} \to \overline{\mathbb{R}}^+$ be a pre-measure on $\mathcal{A}$. Let $\mathcal{F}(\mathcal{A})$ be the $\sigma - algebra$ generated by $\mathcal{A}$. Then, there exists a measure $\mu : \mathcal{F}(\mathcal{A}) \to \overline{\mathbb{R}}^+$ such that $\mu|_{\mathcal{A}} = \nu$. Moreover, if ν is $\sigma-$finite, then the extension μ is unique and also $\sigma-$finite.

We give an application.

Proposition 0.438 *Any pre-measure on an algebra containing all intervals of $\mathbb{R}$ can be extended to the Borel algebra of $\mathbb{R}$.*

0.14.5 Monotone Classes

Definition 0.439 *(Monotone Classes of sets). Let Ω be our set. Let $g \subseteq \mathcal{P}$. g is a monotone class if*

1. *If $A_j \in g$ for all $j \geq 1$, then $A_j \uparrow A$ implies that $A \in g$;*
2. *If $B_j \in g$ for all $j \geq 1$, then $B_j \downarrow B$ implies that $B \in g$.*

Claim 0.440 *Suppose we had a family of monotone classes g_i for $i \in I$ and $g_i \in \mathcal{P}$. Then*

$$\bigcap_{i \in I} g_i$$

is a monotone class.

Proposition 0.441 *Every $\sigma - algebra$ $\mathcal{F}$ is a monotone class.*

Definition 0.442 *(Smallest monotone class). Let $C \subseteq \mathcal{P}$. Then, the smallest monotone class contaning C is*

$$g(C) = \bigcap_{i \in I} g_i$$

where $C \subseteq g_i$ for all $i \in I$.

We now turn to the theorem of interest.

Theorem 80: Monotone class and $\sigma - algebra$

Let us define $\mathcal{A} \subseteq \mathcal{P}$ as the algebra on Ω. Denote $\mathcal{M}(\mathcal{A})$ as the monotone class generated by $\mathcal{A}$. Denote $\mathcal{F}(\mathcal{A})$ as the $\sigma - algebra$ generated by $\mathcal{A}$. Then, we have that

$$\mathcal{M}(\mathcal{A}) = \mathcal{F}(\mathcal{A}).$$

We can use this theorem to show that if a collection of sets is a monotone class, then it is also a $\sigma - algebra$.

0.14.6 Lebesgue Measure

We can use Caratheory's theorem to construct the Lebesgue measure, which is the extension of the measure defined on the algebra of $\mathbb{R}$ intervals.

Definition 0.443 *(Lebesgue algebra). Let us define the semi-algebra $\mathcal{S} = \{\emptyset, \mathbb{R}, \{(a,b] : a, b \in \mathbb{R}\}, \{(a,\infty) : a \in \mathbb{R}\}, \{(-\infty, b] : b \in \mathbb{R}\}\}$. The Lebesgue algebra is the algebra generated $\mathcal{A}(\mathcal{S})$.*

We now define a measure on this algebra and then extend it to get the Lebesgue measure.

Claim 0.444 *Let us define the set function $\mu : \mathcal{A} \to \overline{\mathbb{R}}^+$ where*

$$\begin{cases} \mu(\emptyset) = 0 \\ \mu(\mathbb{R}) = \infty \\ \mu((a,b]) = b - a \\ \mu((a,\infty)) = \infty \\ \mu((-\infty, b]) = \infty \end{cases}$$

Then, μ is a pre-measure.

Claim 0.445 *The real line $\mathbb{R}$ is $\sigma-$finite with respect to μ.*

Proof: Let $E_n = (-n, n]$ and $\mathbb{R} = \bigcup_{n\in\mathbb{N}} E_n$. Then, $\mu(E_n) = -2n < \infty$. ■

From the fact that μ is a pre-measure and $\mathbb{R}$ is $\sigma-$finite, then by Caratheodory's theorem, we can find a unique extension λ on the $\sigma-algebra\ \mathcal{F}(\mathcal{A})$ where restricting λ to $\mathcal{A}$ gives us back μ. This λ is the **Lebesgue measure**.

0.14.7 Complete Measures

Many theorems in measure theory, for instance Fubini or Radon-Nikodym, needs completeness to make full sense. The notion of almost everywhere wouldn't make sense if it holds for a set but not subsets of that set.

Definition 0.446 *(Complete measure). Let $\mathcal{F} \subseteq \mathcal{P}$ be a $\sigma-algebra$ and let $\mu : \mathcal{F} \to \overline{\mathbb{R}}^+$ be a measure. Then, $(\mathcal{F}, \mu)$ is **complete** if for every set $A \in \mathcal{F}$ such that $\mu(A) = 0$, then for all sets $E \subseteq A$, we have that $E \in \mathcal{F}$. The sets E are known as **negligible sets**.*

Observation 0.447 *By monotonicity of the measure, we have that $\mu(E) = 0$.*

Claim 0.448 *Let $\mathcal{F} \subseteq \mathcal{P}$ be a $\sigma-algebra$. Define $\mu : \mathcal{F} \to \overline{\mathbb{R}}^+$ be a measure. Then*

$$\overline{\mathcal{F}} = \{A \cup N : A \in \mathcal{F} \text{ and } N \subseteq E \in \mathcal{F} \text{ such that } \mu(E) = 0\}.$$

Then, $\overline{\mathcal{F}}$ is a $\sigma-algebra$.

Clearly, $\overline{\mathcal{F}} \supset \mathcal{F}$.

Definition 0.449 *Let $\mathcal{F} \subseteq \mathcal{P}$ be a $\sigma-algebra$. Define $\mu : \mathcal{F} \to \overline{\mathbb{R}}^+$ be a measure. Then, we define the measure $\overline{\mu} : \overline{\mathcal{F}} \to \overline{\mathbb{R}}^+$ to be*

$$\overline{\mu}(A \cup N) = \mu(A)$$

for all sets $A \in \overline{\mathcal{F}}$.

Clearly, $\overline{\mu}|_{\mathcal{F}} = \mu$.

Proposition 0.450 *Define the measure and $\sigma-$algebra pair $(\mu, \mathcal{F})$. Furthermore, assume that Ω is $\sigma-$finite with respect to μ. Then, the completion $(\overline{\mu}, \overline{\mathcal{F}})$ is unique.*

Hence, it is always possible to complete a measure and $\sigma-algebra$.

0.14.8 Approximation and Regularity

Let us define π^* as the outer-measure and let $A \in \mathcal{M}$ be a measurable set such that $\pi^*(A) < \infty$. We are interested on whether can we approximate $\pi^*(A)$ with a set $E \in \mathcal{F}(\mathcal{A})$. That is, does there exist a set $E \in \mathcal{F}(\mathcal{A})$ such that $A \subseteq E$ and

$$\pi^*(A) = \pi^*(E).$$

Can we approximate any measurable set from above by a set in $\mathcal{F}(\mathcal{A})$?.

First, we show that we can approximate any set in $\mathcal{F}$ by a set in the $\mathcal{A}$, which we can write as the finite disjoint union of sets in the semialgebra $\mathcal{S}$. However, we pay a price in not having an exact approximation.

Theorem 0.451 *Let $\mathcal{A} \subseteq \mathcal{P}$ be an algebra and $\mathcal{F}(\mathcal{A})$ be a $\sigma-$algebra. Define the measure $\mu : \mathcal{F} \to \overline{\mathbb{R}}^+$. Let us define the set $A \in \mathcal{F}$ and assume $\mu(A) < \infty$. Take $\epsilon > 0$. Then, there exists a set $E \in \mathcal{A}$ such that*

$$\mu(E \backslash A) + \mu(A \backslash E) < \epsilon.$$

Corollary 0.452 *If Ω is $\sigma-$finite with respect to μ, then we can extend $(\mathcal{F}(\mathcal{A}), \mu)$ to the completion $(\overline{\mathcal{F}}, \overline{\mu})$. Sp tjem of $A \in \overline{\mathcal{F}}$ amd $\epsilon > 0$, then there exists a set $E \in \mathcal{A}$ such that*

$$\overline{\mu}(E \backslash A) + \overline{\mu}(A \backslash E) < \epsilon.$$

Definition 130: Regular Measure

Let Ω be a topological space. Let $\mathcal{B}$ be the Borel $\sigma-$*algebra*, which is the smalelst $\sigma-$*algebra* which contains all open sets. Let $\mu : \mathcal{F} \to \overline{\mathbb{R}}^+$ and $\mathcal{F} \supseteq \mathcal{B}$.
μ is **regular** if for all sets $A \in \mathcal{F}$ and $\epsilon > 0$, there exist sets $F \subseteq A \subseteq G$ where F is closed, G is open such that

$$\mu(G \backslash F) \leq \epsilon.$$

Remark 0.453 *We can approximate any set $A \in \mathcal{F}$ from below by a closed set and from above by an open set.*

Lemma 0.454 *Assume $\mathcal{B} \subseteq \mathcal{F}$. If μ is regular then $\mathcal{F} \subseteq \overline{\mathcal{B}}_{\overline{\mu}}$ where $\overline{\mathcal{B}}_{\overline{\mu}}$ where $\overline{\mathcal{B}}$ is the completion of $\mathcal{B}$ with respect to μ.*

Theorem 0.455 *(Regular Lebesgue Measure). The Lebesgue measure is regular. Let μ be the Lebesgue measure where*

$$\mu : \mathcal{L} \to \overline{\mathbb{R}}^+$$

where $\mathcal{L}$ is the Lebesgue $\sigma-$algebra (the extension of semi-algebra of real line intervals).

We now denote F_σ as the countable union of closed sets and G_δ as the countable intersection of open sets.

We can now state that we can approximate any Lebesgue measurable set by a countable collection of closed and open sets.

Theorem 0.456 *For any Lebesgue measurable set $A \in \mathcal{L}$, there exists $R \in F_\sigma$ and $S \in G_\delta$ such that $R \subseteq A \subseteq S$ such that*

$$\mu(S \backslash R) = 0.$$

0.14.9 Measurability

Definition 0.457 *(Measurable functions). Define a measure space $(\Omega, \mathcal{F}, \mu)$. Let $f : \Omega \to \overline{\mathbb{R}}^+$. Define $\overline{\mathcal{B}}$ to be the extended Borel $\sigma-$algebra. Then, a function f is measurable if $f^{-1}(A) \in \mathcal{F}$ for all $A \in \overline{\mathcal{B}}$.*

Remark 0.458 *This is important when we integrate as we will need to take the measure $\mu^{-1}(f^{-1}(A))$.*

It is extremely tedious to check that every set in $\overline{\mathcal{B}}$ has a measurable preimage, hence the following theorem helps us narrow the number of sets we need to check.

Theorem 0.459 *(Measurability of a function). Define a measure space $(\Omega, \mathcal{F}, \mu)$. Let $f : \Omega \to \overline{\mathbb{R}}^+$. Then, the function f is measurable if and only if one of the following holds*

1. *$f^{-1}((-\infty, x]) \in \mathcal{F}$ for all $x \in \mathbb{R}$;*
2. *$f^{-1}((-\infty, x)) \in \mathcal{F}$ for all $x \in \mathbb{R}$;*
3. *$f^{-1}((x, \infty)) \in \mathcal{F}$ for all $x \in \mathbb{R}$;*
4. *$f^{-1}([x, \infty)) \in \mathcal{F}$ for all $x \in \mathbb{R}$.*

Hence, we can simply check whether $\{\omega \in \Omega : f(\omega) \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$.

0.14.10 Simple Functions

Lemma 0.460 *A simple function 1_A is measurable if and only if A is measurable.*

Definition 0.461 *(Simple function). Define a measure space $(\Omega, \mathcal{F}, \mu)$. Then f is a simple function if we can write it in the form*

$$f = \sum_{j=1}^{n} c_j 1_{E_j}$$

where $c_j \in \mathbb{R}$, $E_j \in \mathcal{F}$, $E_j \cap E_k = \emptyset$ for $j \neq k$ and $\cup_{j=1}^{n} E_j = \Omega$.

0.14.11 Integrating Simple Functions

Definition 0.462 *(Integral of a simple function). Define a measure space $(\Omega, \mathcal{F}, \mu)$. Let f be a simple function. Then, we define the integral of f with respect to μ as*

$$I(f) = \sum_{j=1}^{n} c_j \mu(E_j).$$

We state an important theorem that we can approximate any non-negative function by a sequence of monotonic simple functions.

Theorem 0.463 *(Simple approximation theorem). Let $f : \Omega \to \overline{\mathbb{R}}^+$ be a measurable function. Then, there exists a sequence of simple function $\{f_n\}_{n \geq 1}$ such that $f_n \geq 0$ and $f_n \uparrow f$.*

0.14.12 Integrating non-negative Functions

Definition 0.464 *(Integral of non-negative function). Let $f : \Omega \to \overline{\mathbb{R}}^+$ be a measurable function. Then, the integral of f is defined as*

$$I(f) = \lim_{n\to\infty} I(f_n).$$

Lemma 0.465 *The integral of a non-negative function $f : \Omega \to \overline{\mathbb{R}}^+$ does not depend on the sequence of non-negative simple functions chosen to approximate f.*

0.14.13 Integrating measurable functions

We are now interested in measurable functions $f : \Omega \to \overline{\mathbb{R}}^+$. Then, we can decompose f into its positive and negative part. That is, we can define

$$\begin{cases} f^+ = \max(f, 0) \\ f^- = \max(-f, 0). \end{cases}$$

We have that f^+, f^- are measurable and we can express $f = f^+ - f^-$.

Remark 0.466 *Note that if either f^+ or f^- is infinity, then the other term has to be zero.*

Definition 0.467 *(Integrable function). We say that a function f is integrable if* $\min\{f^+, f^-\} < \infty$.

Definition 0.468 *(Integrating measurable functions). Define the measurable function $f : \Omega \to \overline{\mathbb{R}}^+$. Then, define f^+, f^- as before. We define the respective integrals $I(f^+)$ and $I(f^-)$. Now, if we have that* $\min\{f^+, f^-\} < \infty$*, then we define the integral of f as*

$$I(f) = I(f^+) - I(f^-).$$

0.14.14 Measurability of functions

Proposition 0.469 *(Properties of measurable functions). Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then define $f, g : \Omega \to \overline{\mathbb{R}}^+$ to be measurable functions. Let $\alpha \in \mathbb{R}$. The following functions are measurable.*

1. $\alpha f \in \mathcal{F}$
2. $\alpha + f \in \mathcal{F}$
3. $f + g \in \mathcal{F}$
4. $f^2 \in \mathcal{F}$
5. $\frac{1}{f} \in \mathcal{F}$
6. $f^+, f^-, |f| \in \mathcal{F}$
7. $fg \in \mathcal{F}$

Proposition 0.470 *Let $\{f_n\}_{n\geq 1}$ be a sequence of functions $f_n : \Omega \to \overline{\mathbb{R}}^+$ and $f_n \in \mathcal{F}$. We have that the following functions are measurable.*

1. $\sup_n f_n \in \mathcal{F}$
2. $\inf_n f_n \in \mathcal{F}$
3. $\lim_{n\to\infty} \sup f_n \in \mathcal{F}$
4. $\lim_{n\to\infty} \inf f_n \in \mathcal{F}$
5. *If* $f_n \to f$, *then* $f \in \mathcal{F}$.

Proposition 0.471 *Let* Ω *be a topological space. Define the measure space* $(\Omega, \mathcal{F}, \mu)$ *and let* $\mathcal{F} \supseteq \mathcal{B}$ *where* $\mathcal{B}$ *is the Borel* $\sigma-$*algebra. Then, define the function* $f : \Omega \to \overline{\mathbb{R}}^+$. *Then, if the function* f *is continuous, it is also measurable.*

0.14.15 Properties of Integrals

Proposition 0.472 *Let* $(\Omega, \mathcal{F}, \mu)$ *be a measure space and let* f, g *be integrable with respect to* $\mathcal{F}$. *Then,*

$$\int \Big(f + g\Big) d\mu = \int f d\mu + \int g d\mu.$$

Proposition 0.473 *Let* $(\Omega, \mathcal{F}, \mu)$ *be a measure space and let* f *be integrable with respect to* $\mathcal{F}$. *Then*

$$\Big| \int f d\mu \Big| \leq \int |f| d\mu.$$

Proposition 0.474 *Let* $(\Omega, \mathcal{F}, \mu)$ *be a measure space and* $c \in \mathbb{R}$. *Let* f *be integrable with respect to* $\mathcal{F}$. *Then*

$$\int cf d\mu = c \int f d\mu.$$

Proposition 0.475 *(Monotonicty of integral). If* $f \geq 0$ *then* $\int f d\mu \geq 0$. *If* $f \geq g$, *then* $\int f d\mu \geq \int g d\mu$.

Proposition 0.476 *If* $f = g$ *a.e., then we have that*

$$\int f d\mu = \int g d\mu.$$

Proposition 0.477 *If* $h : \Omega \to \overline{\mathbb{R}}^+$ *and* $h \in \mathcal{F}$. *Let* $|h| \leq f$. *If* f *is integrable, then* h *is integrable.*

0.14.16 Convergence Theorems for non-negative functions

Theorem 0.478 *(Monotone Convergence Theorem). Let* $(\Omega, \mathcal{F}, \mu)$ *be a measure space. Let* $f_n \geq 0$ *be a sequence of non-negative functions where* $f_n \in \mathcal{F}$ *for all* $n \geq 1$. *Suppose that* $f_n \uparrow f$ *pointwise. Then*

$$\int f_n d\mu \uparrow \int f d\mu.$$

Remark 0.479 *The MCT still holds if we replace* $f_n \geq 0$ *by* $f_n \geq g$ *where* g *is integrable as we can then define* $g_n = f_n - g \geq 0$.

Definition 0.480 *(Uniform integrability). Let $f \geq 0$ on a measure space $(\Omega, \mathcal{F}, \mu)$. Define $\mu_f(A) = \int_A f d\mu$. Then, the set A is uniformly integrable if for all $\epsilon > 0$, there exists a $\delta > 0$ such that $\mu(A) < \delta$ implies that $\mu_f(A) < \epsilon$.*

Definition 0.481 *(Fatou's Lemma). Let $f_n \geq 0$. Then*

$$\int \underline{lim} f_n d\mu \leq \underline{\lim} \int f_n d\mu.$$

Remark 0.482 *We can repalce $f_n \geq 0$ by $f_n \geq g$ where g is integrable.*

Definition 0.483 *(Reversed Fatou's Lemma). Let $f_n \geq 0$. Then*

$$\int \overline{lim} f_n d\mu \geq \overline{\lim} \int f_n d\mu.$$

Remark 0.484 *We can repalce $f_n \geq 0$ by $f_n \leq g$ where g is integrable.*

Using the two Fatou's lemma, we define the dominated convergence theorem.

Theorem 0.485 *(Dominated Convergence Theorem). Let $f_n \to f$ where $|f_n| \leq g$ where g is integrable. Then f is integrable and*

$$\int f_n d\mu \to \int f d\mu.$$

0.14.17 Lebesgue-Stieltjes Integral

There is a relationship between right continuous increasing functions F on $\mathbb{R}$ and Borel measure μ_F.

If we start off with a regular Borel measure μ on $\mathbb{R}$, we can define a right continuous increasing function by

$$F_0(t) = \begin{cases} \mu((0,t]) & t > 0 \\ -\mu((t,0]) & t < 0 \end{cases}.$$

Then, we derive the **distribution function** by $F(t) = F_0(t) + \mu((-\infty])$.

Now, let us start off with a right continuous increasing function F. First, define the **Lebesgue-Stieltjes outer measure** by

$$\mu_F^*(A) = inf\{\sum_{k=0}^{\infty}(F(b_k) - F(a_k)) : A \subseteq \bigcup_{k=0}^{\infty}(a_k, b_k]\}.$$

Then, create a $\sigma - algebra$ by collecting all sets $E \in X$ such that $\mu_F^*(E) = \mu_F^*(E \cap A) + \mu_F^*(E \cap A^c)$ for all $A \in X$. Then, if we restrict μ_F^*, we get the **Lebesgue-Stieltjes measure induced by F** denoted by μ_F. We have that μ_F is a Borel measure (every Borel set is inside the $\sigma - algebra$).

Theorem 0.486 *Every regular Borel measure μ on $\mathbb{R}$ is associated to an increasing right continuous function $F : \mathbb{R} \to \mathbb{R}$.*

Definition 0.487 *(Lebesgue-Stieltjes Integral). Let f be μ_F-measurable. Then, the Lebesgue-Stieltjes integral is defined as*

$$\int_{\mathbb{R}} f d\mu_F.$$

Remark 0.488 *If we set $F(x) = x$, then we will recover the Lebesgue measure and resultingly, the Lebesgue integral.*

0.14.18 Measure on finite product of spaces

We define the set up on the measure spaces $(\Omega_j, \mathcal{F}_j, \mu_j)$ for j=1,2. We can define $\Omega = \Omega_1 \times \Omega_2 = \{(x, y) : x \in \Omega_1, y \in \Omega_2\}$. Our aim is to now define a $\sigma - algebra$ on Ω and a measure μ on Ω.

Definition 0.489 *(Rectangle). Define the measure spaces $(\Omega_j, \mathcal{F}_j, \mu_j)$ for j=1,2. Then, consider the sets $E_j \in \mathcal{F}_j$. Define the set*

$$E_1 \times E_2 = \{(x, y) : x \in E_1, y \in E_2\}.$$

The set $E_1 \times E_2$ is called a rectangle.

Definition 0.490 *(Class of rectangles). Define the measure spaces $(\Omega_j, \mathcal{F}_j, \mu_j)$ for j=1,2. We define the **class of rectangles** $\mathcal{F}_1 \times \mathcal{F}_2$ as*

$$\mathcal{F}_1 \times \mathcal{F}_2 = \{E_1 \times E_2 : E_1 \in \mathcal{F}_1, E_2 \in \mathcal{F}_2\}.$$

Proposition 0.491 *The family of rectangles $\mathcal{F}_1 \times \mathcal{F}_2$ is a semi-algebra.*

Definition 0.492 *($\sigma - algebra$ of rectangles). Let $\mathcal{F}_1 \times \mathcal{F}_2$ be the family of rectangles. We define the $\sigma - algebra$ of rectangles as*

$$\mathcal{F} = \mathcal{F}_1 * \mathcal{F}_2 = \sigma(\mathcal{F}_1 \times \mathcal{F}_2).$$

We now want to define a pre-measure on the semi-algebra $\mathcal{F}_1 \times \mathcal{F}_2$ and then by Caratheodory's theorem, uniquely extend this to a measure on $\mathcal{F}$.

Proposition 0.493 *Define $\mathcal{F}_1 \times \mathcal{F}_2$ as the semi-algebra of rectangles. Then, define the set function $\mu : \mathcal{F}_1 \times \mathcal{F}_2 \to \overline{\mathbb{R}}^+$ as*

$$\mu(E_1 \times E_2) = \mu_1(E_1)\mu_2(E_2)$$

where we define $0.\infty = 0$.

Proposition 0.494 *The set function $\mu : \mathcal{F}_1 \times \mathcal{F}_2 \to \overline{\mathbb{R}}^+$ is a pre-measure. That is, it is σ-additive.*

Lemma 0.495 *Take the set $A \in \mathcal{F}$. For any set $A \subseteq \Omega$, we define for $x \in \Omega_1$ and for $y \in \Omega_2$, we define*

$$\begin{cases} A_x = \{y \in \Omega_2 : (x, y) \in A\} \\ A^y = \{x \in \Omega_1 : (x, y) \in A\}. \end{cases}$$

Clearly, $A_x \subseteq \Omega_2$ and $A^y \subseteq \Omega_1$. Then, for all $x \in \Omega_1$, we have that $A_x \in \mathcal{F}_2$ and for all $y \in \Omega_2$, we have that $A^y \in \mathcal{F}_1$.

Claim 0.496 *Assume that Ω_1, Ω_2 is $\sigma-$finite with respect to μ_1, μ_2 respectively. Then, Ω is $\sigma-$finite with respect to μ.*

Proposition 0.497 *Let μ be the pre-measure defined on the semi-algebra $\mathcal{F}_1 \times \mathcal{F}_2$. Let $\mathcal{F}$ be the $\sigma - algebra$ generated by the semi-algebra $\mathcal{F}_1 \times \mathcal{F}_2$. Let Ω be σ-finite with respect to μ. Then, by Caratheodory's theorem, we can **uniquely** extend the pre-measure μ onto the $\sigma - algebra$ $\mathcal{F}$. This extended measure is known as the **product measure**.*

0.14.19 Countable Product Space

0.14.20 Fubini's Theorem

Fubini's theorem is a theorem which gives us the sufficient conditions for when we can exchange the order of integrals.

We describe the setup. Let $(\Omega_i, \mathcal{F}_i, \mu_i)$ for i=1,2. We also have Ω_i to be σ-finite with respect to μ_i. Then, we define the product measure $\mu = \mu_1 * \mu_2$ and the product $\sigma - algebra$ as $\mathcal{F} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$. We now have a function $f : \Omega_1 \times \Omega_2 \to \overline{\mathbb{R}}$. We are interested for when does the following hold:

$$\int_{\Omega_1} \Big[\int_{\Omega_2} f_x(y) d\mu_2(y) \Big] d\mu_1(x) = \int f d\mu.$$

This equation has 2 implicit things we need to show. First, $f_x(y)$ is $\mathcal{F}_2-$measurable so that we are able to integrate it with respect to μ_2. Then, we need to show $\int_{\Omega_2} f_x(y) d\mu_2(y)$ is measurable with respect to $\mathcal{F}_1$ for any fixed x so that we can integrate it with respect to μ_1.

Claim 0.498 *Let $f : \Omega \to \overline{\mathbb{R}}$ be a function that is measurable with respect to $\mathcal{F} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$. Then, for all $x \in \Omega$, we define $f_x : \Omega_2 \to \overline{\mathbb{R}}$ as*

$$y \to (x, y).$$

Then, we have that f_x is measurable with respect to $\mathcal{F}_2$.

Recall that if a set $E \in \mathcal{F} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$, we defined the x-section of E and the y-section of E as:

$$\begin{cases} E_x = \{y \in \Omega_2 : (x, y) \in E\} \\ E^y = \{x \in \Omega_1 : (x, y) \in E\}. \end{cases}$$

We first state a lemma to help us prove an important theorem for Fubini's theorem.

Lemma 0.499 *Let $(\Omega_j, \mathcal{F}_j, \mu_j)$ be measure spaces for j=1,2 where Ω_j is $\sigma-$finite with respect to μ_j. Take a set $E \in \mathcal{F} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$. Then, we have the following*

1. *$E_x \in \mathcal{F}_2$;*
2. *$\mu_2(E_x)$ is well-defined;*
3. *$E^y \in \mathcal{F}_1$;*
4. *$\mu_1(E^y)$ is well-defined.*

Theorem 0.500 *Let $(\Omega_j, \mathcal{F}_j, \mu_j)$ be measure spaces for j=1,2 where Ω_j is $\sigma-$finite with respect to μ_j. Take a set $E \in \mathcal{F} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$. Then, we have the following.*

1. *$x \to \mu_2(E_x)$ is $\mathcal{F}_1-$measurable;*
2. *$y \to \mu_1(E^y)$ is $\mathcal{F}_2-$measurable;*
3. *$\int_{\Omega_1} \mu_2(E_x) d\mu_1(x) = \mu(E) = \int_{\Omega_2} \mu_1(E^y) d\mu_2(y)$.*

Theorem 0.501 *(Tonelli's Theorem). Let $(\Omega_j, \mathcal{F}_j, \mu_j)$ be measure spaces for j=1,2 where Ω_j is $\sigma-$finite with respect to μ_j. Take a function $f : \Omega_1 \times \Omega_2 \to \overline{\mathbb{R}}^+$. Then*

$$\int_{\Omega_1} \Big[\int_{\Omega_2} f_x(y) d\mu_2 \Big] d\mu_1 = \int f d\mu = \int_{\Omega_2} \Big[\int_{\Omega_1} f^y(x) d\mu_1 \Big] d\mu_2$$

Remark 0.502 *Here, we are saying that $\int_{\Omega_2} f_x(y) d\mu_2$ is measurable with respect to $\mathcal{F}_1$ and hence we can integrate it with μ_1.*

Theorem 0.503 *(Fubini's Theorem). Let $(\Omega_j, \mathcal{F}_j, \mu_j)$ be measure spaces for j=1,2 where Ω_j is $\sigma-$finite with respect to μ_j. Take a function that is measurable with respect to $\mathcal{F}$. Furthermore, assume that **f is integrable**. Then*

1. *$\int_{\Omega_2} f_x(y) d\mu_2$ is measurable with respect to $\mathcal{F}_1$;*
2. *$\int_{\Omega_1} f^y(x) d\mu_1$ is measurable with respect to $\mathcal{F}_2$;*
3. $$\int_{\Omega_1} \Big[\int_{\Omega_2} f_x(y) d\mu_2 \Big] d\mu_1 = \int f d\mu = \int_{\Omega_2} \Big[\int_{\Omega_1} f^y(x) d\mu_1 \Big] d\mu_2$$

0.14.21 Radon-Nikodym Theorem

Definition 0.504 *(Absolute continuity). Let μ, ν be measures on the space $(\Omega, \mathcal{F})$. μ is absolutely continuous with respect to ν if $\nu(A) = 0$ implies that $\mu(A) = 0$. We denote this by $\mu << \nu$.*

Claim 0.505 *(Measure with a density). Let $f \geq 0$ be a function on $(\Omega, \mathcal{F}, \mu)$. Then, we can define set function*

$$\mu_f(A) = \int_A f d\mu = \int_\Omega 1_A f d\mu.$$

where $A \in \mathcal{F}$. Then, μ_f is a measure.

The following gives us the converse, that is, if 2 measures are absolutely continuous, then there exists a non-negative function relating the two measures.

Theorem 0.506 *(Radon-Nikodym Theorem). Let μ, ν be measures on the space $(\Omega, \mathcal{F})$. Let $\mu << \nu$. Then, there exists a non-negative function $g(= \frac{d\mu}{d\nu})$ such that*

$$\mu(B) = \int_B g d\nu$$

for all $B \in \mathcal{F}$.

0.14.22 Convergence of functions

Lemma 0.507 *Let $(\Omega, \mathcal{F}, \mu)$ is a measure space. Let Ω be $\sigma-$finite with respect to μ. Assume that $\mathcal{F}$ is $\sigma-$complete. Let $f : \Omega \to \overline{\mathbb{R}}$ and $g : \Omega \to \overline{\mathbb{R}}$. If f is $\mathcal{F}$-measurable and $g = f$ almost surely, then g is $\mathcal{F}-$measurable.*

Definition 0.508 *(Pointwise Convergence). Let $f, f_n : E \to \overline{\mathbb{R}}$. Then f_n converges to f pointwise if for all $x \in E$, $f_n(x) \to f(x)$. We write this as $f_n \xrightarrow{p} f$.*

Definition 0.509 *(Almost Sure Convergence). Let $f, f_n : \Omega \to \overline{\mathbb{R}}$. Then, f_n converges to f almost surely (or almost everywhere) if there exists a set $E \in \mathcal{F}$ such that $\mu(E^c) = 0$ and f_n converges to f pointwise on E, that is, for all $x \in E$, $f_n(x) \to f(x)$.*

We can think of a.s. convergence as pointwise convergence except for a null set E.

Proposition 0.510 *(Properties of a.s. convergence). Let $f_n \to f$ almost surely. Then if $f_n \sim g_n$ and $f \sim g$ then $g_n \to g$ almost surely.*

Proposition 0.511 *(Further properties of a.s. convergence). Let $f_n \to f$ almost surely and $f_n \to g$ almost surely. Then $f = g$ almost surely.*

We recall another form of convergence from analysis.

Definition 0.512 *(Uniform Convergence). Let $f, f_n : E \to \overline{\mathbb{R}}$. Then f_n converges to f uniformly if for all $\epsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that*

$$\sup_{x \in E} |f_n(x) - f(x)| \leq \epsilon$$

for all $x \in E$ and $n \geq n_0$.

Definition 0.513 *(Uniform Convergence Almost Everywhere). Let $f, f_n : \Omega \to \overline{\mathbb{R}}$. Then f_n converges to f **uniformly almost everywhere** if there exists a set $E \in \mathcal{F}$ such that $\mu(E^c) = 0$ and f_n converges to f uniformly on E.*

Definition 0.514 *(Almost Uniform Convergence). Let $f, f_n : \Omega \to \overline{\mathbb{R}}$. Then f_n converges to f **almost uniformly** if for all $\epsilon > 0$, there exists a set $E_\epsilon \in \mathcal{F}$ such that $\mu(E_\epsilon^c) \leq \epsilon$ and f_n converges to f uniformly on E_ϵ.*

Proposition 0.515 *(Relationships between forms of convergence). Let $f, f_n : \Omega \to \overline{\mathbb{R}}$. Then, if $f_n \to f$ uniformly almost everywhere, then f_n converges to f almost everywhere **AND** also converges almost uniformly.*

Theorem 0.516 *(Ergoroff). Let $\mu(\Omega) < \infty$. Then $f_n \to f$ almost surely if and only if $f_n \to f$ almost uniformly.*

Remark 0.517 *In a probability space, this always holds!*

Definition 0.518 *(Convergence in measure). Let $f_n, f : \Omega \to \overline{\mathbb{R}}$. Then, f_n converges to f in measure if for all $\epsilon > 0$, we have that*

$$\mu\Big(\{x : |f_n(x) - f(x)| > \epsilon\}\Big) \to 0$$

as $n \to \infty$.

Proposition 0.519 *Assume that f_n converges to f in measure and f_n converges to g in measure. Then $f = g$ almost surely.*

Proposition 0.520 *If $f_n \to f$ in measure and $g_n \sim f_n$ and $g \sim f$, then $g_n \to g$ in measure.*

Lemma 0.521 *Assume that f_n converges to f in measure. Then, there exists a subsequence $\{n_k\}$ such that $f_{n_k} \to f$ almost surely.*

Remark 0.522 *Note that this only holds for a subsequence, not for the whole sequence.*

Theorem 0.523 *Assume that $\mu(\Omega) < \infty$. Then if $f_n \to f$ almost surely, then $f_n \to f$ in measure.*

Remark 0.524 *In probability theory, this means that if a sequence of random variables converges almost surely, then the sequence of random variables converges in probability.*

0.14.23 Hölder and Minkowski Inequalities

Definition 0.525 *(L^p-norm). Let $1 \leq p < \infty$. We define the L^p-norm as*

$$||f||_p = \Big(\int |f(x)|^p \mu(dx) \Big)^{\frac{1}{p}}.$$

Theorem 0.526 *(Hölder's Inequality). Let $1 < p < \infty$. Define q to be such that $\frac{1}{p} + \frac{1}{q} = 1$. Assume that $||f||_p < \infty$ and $||g||_q < \infty$. Then, we have that*

$$\int |fg| d\mu \leq ||f||_p ||g||_q.$$

Theorem 0.527 *(Minkowski Inequality). Let $1 \leq p < \infty$. Let f,g be functions such that $||f||_p < \infty$ and $||g||_p < \infty$. Then, we have that*

$$||f + g||_p \leq ||f||_p + ||g||_p.$$

Lemma 0.528 *The Minkowski inequality is a **pseudonorm**.*

0.14.24 L^p-Spaces

Definition 0.529 *($\mathcal{L}^p-$space). We define the Lebesgue space as*

$$\mathcal{L}^p = \{f : \Omega \to \overline{\mathbb{R}} : ||f||_p < \infty\}.$$

Theorem 0.530 *($\mathcal{L}_p$ is a linear space). Let $f, g \in \mathcal{L}_p$. Then*

$$f + g \in \mathcal{L}_p.$$

Definition 0.531 *(Metric in $\mathcal{L}^p$). We can define a metric on the $\mathcal{L}^p-$space as*

$$d(f, g) = ||f - g||_p.$$

Definition 0.532 *(Convergence in L^p). Define a sequence of functions $\{f_n\}_{n \geq 1}$ where $f_n \in \mathcal{L}^p$. Then, f_n converges to f in $\mathcal{L}^p$ if and only if*

$$||f_n - f||_p \to 0.$$

Definition 0.533 *(Cauchy Sequence). Let* $\{f_n\}_{n\geq 1}$ *be a sequence of functions where* $f_n \in \mathcal{L}^p$*. Then,* f_n *is a Cauchy sequence if for all* $\epsilon > 0$*, there exists a* $M_0 \in \mathbb{N}$ *such that* $||f_n - f_m||_p \leq \epsilon$ *for all* $n, m \geq M_0$*.*

Theorem 0.534 *The space* $\mathcal{L}^p$ *is complete. That is, suppose there exists a Cauchy sequence of functions* $\{f_n\}_{n\geq 1}$ *in* $\mathcal{L}^p$*, then there exists a function* $f \in \mathcal{L}^p$ *such that* $f_n \to f$ *in* $\mathcal{L}^p$*.*

Definition 0.535 *(Uniformly Integrable). A sequence of functions* $\{f_n\}_{n\geq 1}$ *is uniformly integrable if*

$$\sup_n \int_{|f_n|>A} |f_n|^p d\mu \to 0$$

as $A \to \infty$*.*