

Lecture 1 / Week 1

Classes of Events

Definition A **random experiment** is an experiment where the outcome cannot be predicted in advance.

Example We can observe the price of an asset at t_0 , price in $[0, T]$

Definition The set of all the possible outcomes is called **sample space** Ω .

Example The price at time t_0 : $\Omega = [0, \infty]$
Log return : $\Omega = (-\infty, \infty) = \mathbb{R}$ (the set of all real numbers),
Note: $[-\infty, \infty] =$ extended real line

Example Observe a price between time 0 and time T . If the price moves continuously $\implies \Omega = \{w = w(t) \mid w : [0, T] \rightarrow [0, \infty], w \text{ is continuous}\}$

Insert here Figure 1

Definition Let Ω be a sample space. An **event** is a subset of Ω . In other words it is the set of possible outcomes of the experiment.

Example Price at t_0 ; Event $A = [0, a)$. So the event says, the price at t_0 is lower than a . Price in $[0, T]$.

An event could also be A : the price at $\frac{T}{2}$ is lower than a .
Formally, $A = \{w = w(t) : w(\frac{T}{2}) < a\}$

Insert here Figure 2

Definition $\cup_{i \in I} A_i =$ The union of A_i 's (at least one of the A_i 's)

Definition $\cap_{i \in I} A_i =$ The intersection of A_i 's (all of the A_i 's)

Definition $A^C =$ The complement of A (not A).

Example In the previous example, $A^C =$ the price at $\frac{T}{2} \geq a$. Formally,
 $A^C = \{w = w(t) : w(\frac{T}{2}) \geq a\}$

Insert here Figure 3

Definition A **class** of events is a set of events with certain properties.

Definition Let Ω be the sample space and \mathcal{F} a class of events in Ω . \mathcal{F} is a **sigma-algebra** (σ -algebra) if and only if it has the following properties:

1. $\Omega \in \mathcal{F}$. (So in words, the entire set should belong to the class)
2. If $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$. (If an event belongs to the class, then its complement should also belong to the class.)
3. If $A_1, A_2, \dots \in \mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$. (If a sequence of events belong to the class, then their countable union should also belong to the class. (Note: \cup^{∞} :countable union, \cup^n where $n \in \mathbb{N}$: finite union)

Insert here Figure 4

Remark \mathcal{F} is closed under complements and unions.

Claim A σ -algebra is closed under intersections.

Proof Suppose we have a sequence of events $A_1, A_2, \dots \in \mathcal{F}$, then (by property 2) $A_1^C, A_2^C, \dots \in \mathcal{F}$. Then, by property 3, we know that $\cup_{i=1}^{\infty} A_i^C \in \mathcal{F}$. Using the De Morgan's Law, $\cup_{i=1}^{\infty} A_i^C = (\cap_{i=1}^{\infty} A_i)^C \in \mathcal{F}$. Let's call $(\cap_{i=1}^{\infty} A_i)^C = B \in \mathcal{F}$, but then $B^C = \cap_{i=1}^{\infty} A_i \in \mathcal{F}$. *QED*.

Claim \mathcal{F} is closed under unions iff (if and only if) \mathcal{F} is closed under intersections. (Given property 1 and property 2)

Proof ' \Rightarrow ': This implication we have just proved.

' \Leftarrow ': Suppose we have a sequence of events $A_1, A_2, \dots \in \mathcal{F}$, then (by property 2) $A_1^C, A_2^C, \dots \in \mathcal{F}$. By hypothesis, we know $\cap_{i=1}^{\infty} A_i^C \in \mathcal{F}$. Using the De Morgan's Law, $\cap_{i=1}^{\infty} A_i^C = (\cup_{i=1}^{\infty} A_i)^C \in \mathcal{F}$. Then by property 2, $(\cup_{i=1}^{\infty} A_i) \in \mathcal{F}$. *QED*.

Remark We know that \mathcal{F} is closed under countable unions (intersections). Can we also say that it is closed under finite unions (intersections)? $A_1, A_2, \dots, A_n \in \mathcal{F} \Rightarrow (?) \cup_{i=1}^n A_i \in \mathcal{F}$. Yes, indeed we can!

Proof By property 1, we know that $\Omega \in \mathcal{F}$. Since $\Omega^C = \emptyset$, but then $\emptyset \in \mathcal{F}$. So, we can take the infinite sequence $A_1, A_2, \dots, A_n, \emptyset, \emptyset, \emptyset, \dots$. Then, $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$ implies $\cup_{i=1}^n A_i \in \mathcal{F}$. *QED*.

Definition A class of events that is closed under complements and finite unions denoted as \mathcal{A} where $\Omega \in \mathcal{A}$ is called an **algebra (field)** of events.

Remark If a class of events is a σ -algebra, then it is an algebra, but reverse is not generally true. Formally, $\sigma\text{-ALG} \Rightarrow \text{ALG}$, but $\text{ALG} \not\Rightarrow \sigma\text{-ALG}$. If Ω is finite, an algebra is also a σ -algebra.

Example $\Omega = (0, 1]$

$\mathcal{A} = \{\text{finite unions of intervals of the type } (a, b] \text{ with } 0 \leq a \leq b \leq 1\}$

Exercise \mathcal{A} is an algebra.

Proof We have to check whether it satisfies the 3 properties of algebra. Property 1: Taking $a=0$ and $b=1$, we show that $\Omega \in \mathcal{A}$. Property 3: $\cup_{i=1}^n (a_i, b_i] \in \mathcal{F}$. By induction, we can see that the finite union belongs to the algebra. When $n=1$, it holds by definition. If we take $(a_1, b_1] \cup (a_2, b_2]$. We get a union of the same form. If this holds for n and it can be shown that it also holds for $n+1$. Property 2: The complements of the sets $(a, b]$, have the form $(a, b]^C = (0, a] \cup (b, 1]$. Since they are union of elements of \mathcal{A} , they belong to the algebra.

Remark \mathcal{A} is not a σ -algebra.

Proof Take sets of the form: $\cap_2^\infty (\frac{1}{2} - \frac{1}{n}, 1] \Rightarrow (0, 1] \cap (\frac{1}{2} - \frac{1}{3}, 1] \cap \dots = [\frac{1}{2}, 1] \notin \mathcal{A}, (= \{x \in \mathbb{R} : \frac{1}{2} - \frac{1}{n} < x \leq 1 \text{ for every } n\})$, since the countable intersection does not belong to \mathcal{A} , the class is not a σ -algebra. *QED*.

Definition Let Ω be a sample space. Suppose \mathcal{C} is a class of events.

$\sigma(\mathcal{C}) = \cap_{\mathcal{G} \text{ is } \sigma\text{-ALG and } \mathcal{G} \supset \mathcal{C}} \mathcal{G}$, $\sigma(\mathcal{C})$ is called σ -algebra **generated by** \mathcal{C} .

- 1) $\sigma(\mathcal{C})$ is a σ -algebra.
- 2) It is the smallest σ -algebra containing \mathcal{C} .

Proof Property (1): Every intersection of σ -algebra is a σ -algebra. Suppose $\{\mathcal{F}_\theta\}_{\theta \in \Theta}$, $\cap_{\theta \in \Theta} \mathcal{F}_\theta = \mathcal{F}$.

- 1) $\Omega \in \mathcal{F}_\theta \forall \theta \Rightarrow \Omega \in \cap_{\theta \in \Theta} \mathcal{F}_\theta$
- 2) $A \in \mathcal{F} = \cap_{\theta \in \Theta} \mathcal{F}_\theta \Rightarrow A \in \mathcal{F}_\theta \forall \theta \Rightarrow A^C \in \mathcal{F}_\theta \forall \theta \Rightarrow A^C \in \cap_{\theta \in \Theta} \mathcal{F}_\theta = \mathcal{F}$
- 3) If $A_1, A_2, \dots \in \mathcal{F} = \cap_{\theta \in \Theta} \mathcal{F}_\theta \Rightarrow A_1, A_2, \dots \in \mathcal{F}_\theta \forall \theta \Rightarrow \cup_{i=1}^\infty A_i \in \mathcal{F}_\theta \forall \theta \Rightarrow \cup_{i=1}^\infty A_i \in \cap_{\theta \in \Theta} \mathcal{F}_\theta = \mathcal{F}$. *QED*.

Property (2): If \mathcal{F} is σ -ALG $\supseteq \mathcal{C} \Rightarrow \mathcal{F} \supseteq \sigma(\mathcal{C}) \Rightarrow \mathcal{F}$ is in the intersection, so $\sigma(\mathcal{C})$ must be the smallest σ -ALG. *QED*

Definition Let $\Omega = \mathbb{R}$, $\mathcal{C} = \{(a, b) : -\infty < a < b < \infty\}$ a general class (Note that it is not σ -algebra). The **Borel** σ -algebra on $\mathbb{R} = \sigma(\mathcal{C}) \Rightarrow B(\mathbb{R})$. The **Borel** σ -algebra on \mathbb{R} is the σ -algebra generated by \mathcal{C} . Note that it contains the singletons $\{a\}$.

$$\begin{aligned} \{a\} &= \cap_{n=1}^\infty (a - \frac{1}{n}, a + \frac{1}{n}) \in B(\mathbb{R}). \\ (a, b) &= (a, b) \cup \{b\} \in B(\mathbb{R}). \text{similarly, } [a, b), [a, b] \in B(\mathbb{R}) \\ (a, b] &\in B(\mathbb{R}), [a, b] \in B(\mathbb{R}), \end{aligned}$$

All finite unions of intervals $\in B(\mathbb{R})$

$$(-\infty, a] = \cup_{n=1}^\infty (a - n, a] \in B(\mathbb{R})$$

$$\mathcal{C}' = \{(a, b] : -\infty < a < b < \infty\} \quad \sigma(\mathcal{C}') = B(\mathbb{R})$$

$$\mathcal{C}'' = \{(\infty, a] : -\infty < a < \infty\} \quad \sigma(\mathcal{C}'') = B(\mathbb{R}) \quad (\text{Proof p.14})$$

$$\begin{aligned} \mathbb{R}^K &= \{(x_1, x_2, \dots, x_K) : x_i \in \mathbb{R}\} \\ \mathcal{C} &= \{(a_1, b_1) \times (a_2, b_2) \times \dots \times (a_K, b_K) : -\infty < a_i < b_i < \infty\} \\ &\quad (x_1, x_2, \dots, x_K) : x_i \in (a_i, b_i) \quad \forall i. \end{aligned}$$

Insert figure here 5

Remark $B(\mathbb{R}^K) = \sigma(\mathcal{C})$ contains singletons, finite and countable sets, open sets, closed sets...

Definition Let Ω be a sample space. Fix a σ -algebra \mathcal{F} . \mathcal{F} contains all the relevant events $A_1, A_2, \dots, \in \mathcal{F}, \cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

A **probability measure** is a function $P: \mathcal{F} \rightarrow \mathbb{R}$ that satisfies

- 1) $P(A) \geq 0$
- 2) $P(\Omega) = 1$
- 3) If A_1, A_2, \dots are disjoint events then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ where $A_i \cap A_j = \emptyset, i \neq j$. (σ -additive)

Example Observe a price from time 0 to T . (Note that the sample space Ω is a set of functions). \mathcal{F} is a σ -algebra of all events.

$$\mathcal{C} = \{\text{price at time } t \text{ is lower than } a \text{ (} p_t < a \text{)}, \forall a \in \mathbb{R}, \forall t \in [0, T]\}, \mathcal{F} = \sigma(\mathcal{C}).$$

Insert here Figure 6

Definition Events until time t_0 : $\sigma\{\text{price at time } t \text{ is smaller than } a, \forall a \in \mathbb{R}, \forall t \leq t_0\} = \mathcal{F}_{t_0}$. Note that $\mathcal{F}_{t_0} \subseteq \mathcal{F}$. \mathcal{F} is the **universe σ -algebra** and \mathcal{F}_t is the **sub σ -algebra** of \mathcal{F} .

Definition $\mathcal{F}_0 = \{\emptyset, \Omega\}$ is the smallest σ -algebra of subsets of Ω .

Definition $\mathcal{P}(\Omega) = \{\text{All the subsets of } \Omega\}$ is the biggest σ -algebra. (**Part σ -algebra**).

When $\Omega = \mathbb{R}$, we can choose $\mathcal{P}(\mathbb{R})$, but in general $B(\mathbb{R})$ is chosen instead, since it is smaller. It can be proved that $B(\mathbb{R}) \subseteq \mathcal{P}(\mathbb{R})$. If we choose $\mathcal{P}(\mathbb{R})$, then we cannot define a sensible probability measure on it. The only probability on $\mathcal{P}(\mathbb{R})$ is the discrete probability, where only the points have positive probability.

Insert here Figure 7

Remark We have already shown that $\cap_{\theta \in \Theta} \mathcal{F}_\theta$ is a σ -algebra. Yet, $\cup_{\theta \in \Theta} \mathcal{F}_\theta$ is not a σ -algebra. But still we can obtain a σ -algebra generated by $\cup_{\theta \in \Theta} \mathcal{F}_\theta \Rightarrow \sigma(\cup_{\theta \in \Theta} \mathcal{F}_\theta) = \vee_{\theta \in \Theta} \mathcal{F}_\theta$. This σ -algebra is generated in the same way we have seen before (Definition 10); take a class \mathcal{C} , then $\sigma(\mathcal{C}) = \cap_{\mathcal{G}}$ is σ -ALG and $\mathcal{G} \supset \mathcal{C}$, we only have to set $\mathcal{C} = \cup_{\theta \in \Theta} \mathcal{F}_\theta$.

Example The following example helps us to see that $\cup_{\theta \in \Theta} \mathcal{F}_\theta$ is not a σ -algebra.

Take the following σ -algebra: $\mathcal{F}_\theta = \{\emptyset, \mathbb{R}, (-\infty, \theta], (\theta, \infty)\}$, then $\cup_{\theta \in \Theta} \mathcal{F}_\theta = \{\emptyset, \mathbb{R}, (-\infty, \theta], (\theta, \infty) \mid \theta \in \mathbb{R}\}$

If we take $\cup_{n=1}^{\infty} (-\infty, -\frac{1}{n}] = (-\infty, 0) \notin \cup_{\theta \in \Theta} \mathcal{F}_\theta$, even though $(-\infty, -\frac{1}{n}] \in \cup_{\theta \in \Theta} \mathcal{F}_\theta$

Remark If we have two classes such that $\mathcal{C}_1 = \{(0, 1), (2, 3), (5, 6)\}$ and $\mathcal{C}_2 = \{(-1, 0), (7, 8)\}$, then $\mathcal{C}_1 \cup \mathcal{C}_2 = \{(0, 1), (2, 3), (5, 6), (-1, 0), (7, 8)\}$.

Insert here Figure 8

Lecture 2 / Week 1

Random Variables

Definition (General Definition of Measurability): Suppose we have Ω, Ω' sample spaces and $\mathcal{F}_\Omega, \mathcal{F}_{\Omega'}$ σ -algebras. Then we define the following mapping:

$$g: \Omega \rightarrow \Omega'$$

Insert here Figure 9

Inverse Image of B (a subset of Ω') through g: The set $\{\omega \in \Omega : g(\omega) \in B\} = g^{-1}(B)$. Note that the inverse image is not necessarily a inverse function. It is just the inverse of the direct image.

Definition g is $\mathcal{F} \setminus \mathcal{F}'$ measurable if for every $B \in \mathcal{F}'$, $g^{-1}(B) \in \mathcal{F}$.

Remark An interesting case is if $\Omega' = \mathbb{R}$.

Definition Suppose we have a sample space Ω , σ -algebra \mathcal{F} , and probability measure P , then (Ω, \mathcal{F}, P) is a **probability space**.

Definition A **random variable** is a function

$$\mathcal{X}: \Omega \rightarrow \mathbb{R}$$

which is $\mathcal{F} \setminus \mathcal{B}(\mathbb{R})$ measurable.

We want to be able to define the probability of $P(\mathcal{X} \in B)$, i.e the probability of events like $(\mathcal{X} \in B) = \{\omega \in \Omega : \mathcal{X}(\omega) \in B\} = \mathcal{X}^{-1}(B)$. Note that B is a Borel set of the Borel σ -algebra $\mathcal{B}(\mathbb{R})$. For example, such a set could be $B=(0,1)$ and then $(\mathcal{X} \in (0,1)) = (0 < \mathcal{X} < 1)$. Also note that $(\mathcal{X} \in B) = \mathcal{X}^{-1}(B) \in \mathcal{F}$.

Insert here Figure 10

Definition Let (Ω, \mathcal{F}, P) be a probability space and $\mathcal{G} \subseteq \mathcal{F}$, where \mathcal{G} is a sub σ -algebra of \mathcal{F} . Then \mathcal{X} is **\mathcal{G} -measurable** if for every $B \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}\mathcal{X}^{-1}(B) &\in \mathcal{G} \\ (\mathcal{X} \in B) &\in \mathcal{G}\end{aligned}$$

Theorem Suppose we a sample space Ω and a σ -algebra \mathcal{F} , of subsets of Ω . Let \mathcal{X} be a function s.t $\mathcal{X}: \Omega \rightarrow \mathbb{R}$. Then \mathcal{X} is $\mathcal{F} \setminus \mathcal{B}(\mathbb{R})$ measurable if and only if $\forall y \in \mathbb{R}$,

$$A_y = \{\omega \in \Omega : \mathcal{X}(\omega) \leq y\} \in \mathcal{F}$$

In general, we can try to take every Borel set B , compute $\mathcal{X}^{-1}(B)$ and check if $\mathcal{X}^{-1}(B) \in \mathcal{F}$. But this would be very cumbersome, instead the theorem tells us that we can just take the inverse images of particular class of sets such as $\mathcal{X}^{-1}(-\infty, y] = \{\omega \in \Omega : \mathcal{X}(\omega) \leq y\} = \{\omega : \mathcal{X}(\omega) \in (-\infty, y]\}$ and check whether they belong to \mathcal{F} .

Proof Take the class $D = \{B \in \mathcal{B}(\mathbb{R}) : \mathcal{X}^{-1}(B) \in \mathcal{F}\}$. Note that it is a set of Borel sets s.t the inverse image belongs to the σ -algebra. It is easy to verify that D is a σ -algebra. (Exercise!) Notice that $(-\infty, y] \in D$ and $\{(-\infty, y] : y \in \mathbb{R}\} \subseteq D$. We know that $\mathcal{B}(\mathbb{R})$ is the smallest σ -algebra containing these intervals. By definition, we can generate $\mathcal{B}(\mathbb{R})$ by using intervals $(-\infty, y]$. So, $\mathcal{B}(\mathbb{R}) \subseteq D$. Since by construction of D , $D \subseteq \mathcal{B}(\mathbb{R}) \Rightarrow \mathcal{B}(\mathbb{R}) = D$. But then, $\forall B \in \mathcal{B}(\mathbb{R})$, $B \in D$ and by construction $\mathcal{X}^{-1}(B) \in \mathcal{F} \Rightarrow$ (*By definition measurability*) \mathcal{X} is $\mathcal{F} \setminus \mathcal{B}(\mathbb{R})$ measurable.

Exercise Verify that D is a σ -algebra.

Proof To verify that D is a σ -algebra, we have to check whether it satisfies 3 properties of σ -algebra: 1.) Choose $\mathbb{R} \in \mathcal{B}(\mathbb{R})$, then $\mathcal{X}^{-1}(\mathbb{R}) = \Omega \in \mathcal{F}$, since \mathcal{F} is a σ -algebra. 2.) Pick any B , if $B \in D$, then B^C should also belong to D . Since $\mathcal{X}^{-1}(B^C) = (\mathcal{X}^{-1}(B))^C$ and \mathcal{F} is a σ -algebra, then $B^C \in D$. 3.) if $B_1, B_2, \dots \in D$, then it should hold that $\cup_{i=1}^{\infty} B_i \in D$. Since $\mathcal{X}^{-1}(\cup_{i=1}^{\infty} B_i) = \cup_{i=1}^{\infty} \mathcal{X}^{-1}(B_i) \in \mathcal{F}$, $\cup_{i=1}^{\infty} B_i \in D$. QED.

Theorem If $\mathcal{X}_1, \mathcal{X}_2, \dots$ are $\mathcal{F} \setminus \mathcal{B}(\mathbb{R})$ measurable, then

1. $\min(\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n), \max(\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n)$ are measurable.
2. $\sup_n \mathcal{X}_n, \inf_n \mathcal{X}_n$ are measurable.
3. $\liminf_n \mathcal{X}_n, \limsup_n \mathcal{X}_n$ are measurable.
4. $\lim_{n \rightarrow \infty} \mathcal{X}_n$ if it exists is measurable.

Proof (1) $\min(\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n)$ is $\mathcal{F} \setminus \mathcal{B}(\mathbb{R})$ measurable. $\forall y \in \mathbb{R}$, using the previous Theorem we know that it is enough to show that $\{\omega \in \Omega : \min(\mathcal{X}_1(\omega), \mathcal{X}_2(\omega), \dots, \mathcal{X}_n(\omega)) \leq y\} \in \mathcal{F}$. Note that, $\min(\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n) \leq y$ iff $\cup_{j=1}^n \mathcal{X}_j \leq y$. This means that $(\mathcal{X}_j \leq y)$ for at least one j . By hypothesis $(\mathcal{X}_j \leq y) \in \mathcal{F}$ for $\forall j \Rightarrow \cup_{j=1}^n (\mathcal{X}_j \leq y) \in \mathcal{F}$. The $\max(\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n)$ can be proved the same way; i.e $\{\omega \in \Omega : \max(\mathcal{X}_1(\omega), \mathcal{X}_2(\omega), \dots, \mathcal{X}_n(\omega)) \leq y\} \in \mathcal{F}$. Note that,

$\max(\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n) \leq y$ iff $\bigcap_{j=1}^n \mathcal{X}_j \leq y$. This means that $(\mathcal{X}_j \leq y)$ for all j . By hypothesis $(\mathcal{X}_j \leq y) \in \mathcal{F}$ for $\forall j \Rightarrow \bigcap_{j=1}^n (\mathcal{X}_j \leq y) \in \mathcal{F}$.

(2) since $(\sup_n \mathcal{X}_n \leq y)$ iff $\bigcap_{n=1}^\infty (\mathcal{X}_n \leq y)$. This means that $(\mathcal{X}_n \leq y)$ for all n . By hypothesis $(\mathcal{X}_n \leq y) \in \mathcal{F}$ for every $n \Rightarrow \bigcap_{n=1}^\infty (\mathcal{X}_n \leq y) \in \mathcal{F}$, similarly since $(\inf_n \mathcal{X}_n < y) = \bigcup_{n=1}^\infty (\mathcal{X}_n < y)$. This means that $(\mathcal{X}_n < y)$ for all n . By hypothesis $(\mathcal{X}_n < y) \in \mathcal{F}$ for every $n \Rightarrow \bigcup_{n=1}^\infty (\mathcal{X}_n < y) \in \mathcal{F}$

(3) $\lim_{n \rightarrow \infty} \inf_{m \geq n} \mathcal{X}_m = \sup_{n \in \mathbb{N}} (\inf_{m \geq n} \mathcal{X}_m)$ and $\limsup_n \mathcal{X}_n = \inf_{n \in \mathbb{N}} (\sup_{m \geq n} \mathcal{X}_m)$ and follows from the previous proof.

(4) $a_n = \lim_{n \rightarrow \infty} a_n$ does not always exist. It exists iff $\limsup_n a_n = \liminf_n a_n$, then follows from previous proof.

Definition A **simple random variable** is a function that takes finite number of values. (Suppose $A_i \cap A_j = \emptyset, i \neq j$)

$$\mathcal{X}(\omega) = \begin{cases} a_1 & \omega \in A_1 \\ a_2 & \omega \in A_2 \\ \cdot & \\ \cdot & \\ a_n & \omega \in A_n \end{cases}$$

Definition The **indicator function** is defined as follows: ($A \in \mathcal{F}$)

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

$$\boxed{\mathcal{X}(\omega) = \sum_{i=1}^n a_i 1_{A_i}(\omega)}$$

Example $\mathcal{X}(\omega) = \begin{cases} 1 & \omega \in A_1 \\ 2 & \omega \in A_2 \\ 3 & \omega \in A_3 \end{cases}$

$$\mathcal{X}(\omega) = 1 * 1_{A_1} + 2 * 1_{A_2} + 3 * 1_{A_3}$$

Remark A simple random variable is $\mathcal{F} \setminus \mathcal{B}(\mathbb{R})$ measurable. (Verify!)

Theorem A function $\mathcal{X}: \Omega \rightarrow \mathbb{R}$ is $\mathcal{F} \setminus \mathcal{B}(\mathbb{R})$ measurable if and only if there exists a sequence (\mathcal{X}_n) of simple random variables such that for every $\omega \in \Omega$,

$$\mathcal{X}(\omega) = \lim_{n \rightarrow \infty} \mathcal{X}_n(\omega)$$

Proof " \Leftarrow " : Let $\mathcal{X}(\omega) = \lim_{n \rightarrow \infty} \mathcal{X}_n(\omega)$ and \mathcal{X}_n be a simple random variable. Previous remark tells us \mathcal{X}_n is measurable and by point 4 of the previous theorem we know that $\lim_{n \rightarrow \infty} \mathcal{X}_n(\omega)$ is measurable, so $\mathcal{X}(\omega)$ is measurable. *QED.*

" \Rightarrow " : We want to prove that if $\mathcal{X}(\omega)$ is $\mathcal{F} \setminus \mathcal{B}(\mathbb{R})$ measurable, then $\mathcal{X}(\omega) = \lim_{n \rightarrow \infty} \mathcal{X}_n(\omega)$ is measurable, where \mathcal{X}_n is simple random variable. First we suppose $\mathcal{X} \geq 0, \forall \omega \mathcal{X}(\omega) \geq 0$. We prove the statement under this condition and then show that it also holds in the opposite case where $\mathcal{X}(\omega) \leq 0$.

$$\text{Take } \mathcal{X}_n(\omega) = \begin{cases} \frac{k}{2^n} & \text{if } \frac{k}{2^n} \leq \mathcal{X}(\omega) < \frac{k+1}{2^n} \\ 0 & \text{otherwise} \end{cases}$$

$k=0,1,2,\dots,n2^n - 1$

Fix n : \mathcal{X}_n takes values $0, \frac{1}{2^n}, \frac{2}{2^n} \dots n$

\mathcal{X}_n is simple when $n \rightarrow \infty$, $\mathcal{X}_n(\omega) \rightarrow \mathcal{X}(\omega)$. (See the figure below for $\mathcal{X}_1(\omega)$ and $\mathcal{X}_2(\omega)$) This completes the proof when $\mathcal{X}(\omega) \geq 0$.

Insert here Figure 11

To see that the result still holds in the opposite case $\mathcal{X}(\omega) \leq 0$; one should see that the function $\mathcal{X} = \mathcal{X}^+ - \mathcal{X}^-$, where $\mathcal{X}^+(\omega) = \max(\mathcal{X}, 0) \geq 0$ and $\mathcal{X}^-(\omega) = -\min(\mathcal{X}, 0) \geq 0$ (This can be best seen sketching the graph of those functions.). But, then $\mathcal{X}^+ = \lim_n \mathcal{X}_n^+$, $\mathcal{X}^- = \lim_n \mathcal{X}_n^-$ and $\mathcal{X} = \mathcal{X}^+ - \mathcal{X}^- = \lim_n (\mathcal{X}_n^+ - \mathcal{X}_n^-)$ completes the proof. *QED*.

Corollary If $\mathcal{X}_1, \mathcal{X}_2$ are $\mathcal{F} \setminus \mathcal{B}(\mathbb{R})$ measurable, then $\mathcal{X}_1 + \mathcal{X}_2, \mathcal{X}_1 - \mathcal{X}_2, \mathcal{X}_1 * \mathcal{X}_2, \mathcal{X}_1 / \mathcal{X}_2$ (Provided $\mathcal{X}_2 \neq 0$) are $\mathcal{F} \setminus \mathcal{B}(\mathbb{R})$ measurable.

Proof ($\mathcal{X}_1 + \mathcal{X}_2$): Enough to observe that $\mathcal{X}_1 = \lim_{n_1 \rightarrow \infty} \mathcal{X}_{n_1}$ and $\mathcal{X}_2 = \lim_{n_2 \rightarrow \infty} \mathcal{X}_{n_2}$, because then $\mathcal{X}_1 + \mathcal{X}_2 = \lim \mathcal{X}_{n_1} + \lim \mathcal{X}_{n_2} = \lim (\mathcal{X}_{n_1} + \mathcal{X}_{n_2})$. This follows from limit property that sum of two limits equals to the limit of the sum. Moreover, the fact that the sum of simple functions is a simple function and the above theorem guarantees that $(\mathcal{X}_1 + \mathcal{X}_2)$ is $\mathcal{F} \setminus \mathcal{B}(\mathbb{R})$ measurable.

Lecture 3 / Week 2

Random Vectors, Distributions and Integrals

Definition Suppose we have probability space (Ω, \mathcal{F}, P) : A **random vector** is a function $X : \Omega \rightarrow \mathbb{R}^K$ that is $\mathcal{F} \setminus \mathcal{B}(\mathbb{R}^K)$ measurable.

X is a K -dimensional random vector: $X(\omega) = (X_1(\omega), X_2(\omega), \dots, X_K(\omega))$, so $X = (X_1, X_2, \dots, X_K)$ is measurable. Also note that X is $\mathcal{F} \setminus \mathcal{B}(\mathbb{R}^K)$ measurable if and only if X_1, X_2, \dots, X_K are $\mathcal{F} \setminus \mathcal{B}(\mathbb{R})$ measurable. So, it is sufficient that the coordinates of the vector are measurable. This amounts to saying that $X = (X_1, X_2, \dots, X_K)$ is a random vector iff X_1, X_2, \dots, X_K are random variables.

Definition If we have a class $\mathcal{G} \subset \mathcal{F}$ (a sub- σ -algebra of \mathcal{F}) then X is a $\mathcal{G} \setminus \mathcal{B}(\mathbb{R}^K)$ **measurable** if and only if X_1, X_2, \dots, X_K are $\mathcal{G} \setminus \mathcal{B}(\mathbb{R})$ measurable. For notational simplicity we will say X is \mathcal{G} measurable if and only if X_1, X_2, \dots, X_K are \mathcal{G} measurable.

Definition The **smallest σ -algebra** w.r.t which a random vector is measurable is denoted by $\sigma(X) = \cap_{X \text{ is } \mathcal{G} \text{ measurable}} \mathcal{G}$ is a σ -algebra. By definition $X^{-1}(B) \in \mathcal{G}$. [X is \mathcal{G} measurable if $X^{-1}(B) \in \mathcal{G}$]. But then, $X^{-1}(B) \in \sigma(X)$, i.e. X is $\sigma(X)$ measurable. Note that if we take the class

of inverse images of Borel sets, this class will be contained in $\sigma(X)$, in fact it will be equal to $\sigma(X)$. Formally $\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^K)\}$. We already said that $\sigma(X) \supset \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^K)\}$, since $X^{-1}(B) \in \sigma(X)$, to show the reverse inclusion first we need to show that the class $\{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^K)\}$ is a σ -algebra. If we take a class $\mathcal{G} \subset \mathcal{F}$ such that X is \mathcal{G} measurable then $\sigma(X) \subset \mathcal{G}$. If $\sigma(X) \subset \mathcal{G} \Rightarrow X^{-1}(B) \in \mathcal{G} \Rightarrow X$ is \mathcal{G} measurable. Hence X is \mathcal{G} measurable iff $\sigma(X) \subset \mathcal{G}$.

Exercise Show that the class $\{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^K)\}$ is a σ -algebra.

Proof To show that the class $\{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^K)\}$ (Let's call it C) is a σ -algebra, we have to check the 3 properties of σ -algebra. 1.) By definition of random variable function $X^{-1}(\mathbb{R}^K) = \Omega, \mathbb{R}^K \in \mathcal{B}(\mathbb{R}^K)$, so $\Omega \in C$. 2.) If $X^{-1}(B) \in C$, then $(X^{-1}(B))^C \in C$ should hold. But $(X^{-1}(B))^C = X^{-1}(B^C)$, since $\mathcal{B}(\mathbb{R}^K)$ is a σ -algebra, $B^C \in \mathcal{B}(\mathbb{R}^K)$, $X^{-1}(B^C) \in C$. 3.) If $X^{-1}(B_1), X^{-1}(B_2), \dots \in C$, then $\cup_{i=1}^{\infty} X^{-1}(B_i) \in C$. This condition holds, since $\mathcal{B}(\mathbb{R}^K)$ is σ -algebra, then $\cup_{i=1}^{\infty} B_i \in \mathcal{B}(\mathbb{R}^K)$, but then $\cup_{i=1}^{\infty} X^{-1}(B_i) \in C$. QED.

Definition Suppose we have probability space (Ω, \mathcal{F}, P) and a random vector $X : (\Omega \rightarrow \mathbb{R}^K, \text{on } (\mathbb{R}^K, \mathcal{B}(\mathbb{R}^K)))$. Then μ_X is called the **probability distribution** of X . We take a Borel set $B \in \mathcal{B}(\mathbb{R}^K)$ and define its measure $\mu_X(B) = P(X^{-1}(B)) = P(X \in B)$. It is the measure of the probability of all ω s.t $\{\omega \in \Omega : X(\omega) \in B\}$

Insert here Figure 12

Claim μ_X is a probability measure on $(\mathbb{R}^K, \mathcal{B}(\mathbb{R}^K))$. Then we have the new probability space $(\Omega, \mathcal{F}, \mu_X)$.

Proof To prove that μ_X is a probability measure, we should check the 3 properties of probability measure. 1.) $\mu_X(B) \geq 0$, by definition and the fact that P is a probability $\mu_X(B) = P(X^{-1}(B)) \geq 0$. 2.) $\mu_X(\mathbb{R}^K) = 1$: By definition $\mu_X(\mathbb{R}^K) = P(X^{-1}(\mathbb{R}^K)) = P(\Omega) = 1$. 3.) $\mu_X(\cup_{j=1}^{\infty} B_j) = \sum_{j=1}^{\infty} \mu_K(B_j)$, $B_i \cap B_j = \emptyset$, then by definition $\mu_X(\cup_{j=1}^{\infty} B_j) = P(X^{-1}(\cup_{j=1}^{\infty} B_j)) = P(\cup_{j=1}^{\infty} X^{-1}(B_j))$, since P is a probability $P(\cup_{j=1}^{\infty} X^{-1}(B_j)) = \sum_{j=1}^{\infty} P(X^{-1}(B_j)) = \sum_{j=1}^{\infty} \mu_K(B_j)$.

Insert here Figure 13

Distribution Function

$$F_X : \mathbb{R}^K \rightarrow \mathbb{R}$$

$$F_X(x_1, x_2, \dots, x_K) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_K \leq x_K)$$

$$F_X(x_1, x_2, \dots, x_K) = \mu_X((-\infty, x_1] \times (-\infty, x_2] \times \dots \times (-\infty, x_K])$$

Example If X is a random variable, $F_X(x) = P(X \leq x) = P(X \in (-\infty, x]) = \mu_X((-\infty, x])$. If $X = (X_1, X_2)$, $F_X(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2) = P(X \in (-\infty, x_1] \times (-\infty, x_2]) = \mu_X((-\infty, x_1] \times (-\infty, x_2])$.

Insert here Figure 14

Homework Review properties of distribution functions!

Probability Distributions

Suppose we have probability space (Ω, \mathcal{F}, P) and a random vector $X = (X_1, X_2, \dots, X_K)$, then μ_X on $\mathcal{B}(\mathbb{R}^K)$ is a probability distribution of X .

Insert here Figure 15

Definition $\mu_X(B) = \sum_{s \in B} m(s)$; μ_X is **discrete** if there exists a countable (at most) set $S = \{s_1, s_2, \dots\}$ in \mathbb{R}^K such that

$$\mu_X(B) = \sum_{i=1}^{\infty} m(s_i) \mathbf{1}_B(s_i) = \sum_{i=1}^{\infty} m(s_i) \delta_{s_i}(B)$$

where $\delta_{s_i}(B) = \begin{cases} 1 & s_i \in B \\ 0 & s_i \notin B \end{cases}$ is **Dirac Measure**.

$\delta_s(A)$ is the probability measure that puts all its mass on s . Note that $\mathbf{1}_B(s_i) = \delta_{s_i}(B)$. The above equation can also be read as μ_X is a convex combination of Dirac probability measures. (convex $\Rightarrow m(s_i) \leq 1, \sum m(s_i) = 1$).

Definition Suppose we have probability space (Ω, \mathcal{F}, P) and a random vector $X = (X_1, X_2, \dots, X_K)$ with probability distribution μ_X on $\mathcal{B}(\mathbb{R}^K)$. The probability distribution μ_X is called **absolutely continuous** if there exists a nonnegative integrable function f , (**density function** of X), such that the **distribution function** F_X (equivalently μ_X) can be written as

$$\mu_X(B) = \int_B f(x) dx$$

where B is a rectangle.

$$B = (a_1, b_1] \times (a_2, b_2] \times \dots \times (a_K, b_K]$$

$$\int_B f(x) dx = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \dots \int_{a_K}^{b_K} f(x_1, \dots, x_K) dx_1 \dots dx_K$$

Riemann Integral

$I =$ rectangle, $f : \mathbb{R}^K \rightarrow \mathbb{R}$, f is continuous

Definition A collection of disjoint rectangles such that their union is I is called **partition** of I

$$\int_I f(x) dx = \sup \sum_{j=1}^n (\inf_{x \in I_j} f(x)) \cdot \lambda(I_j)$$

where I is measured with

$$\lambda(I) = \prod_{i=1}^K (b_i - a_i)$$

is called the **Riemann Integral**.

Insert here Figure 16

As we can see by the definition of Riemann Integral, that it is based on rectangles, so it is not suitable to integrate density functions of other shapes. So, we have to extend the the measure $\lambda(I)$ to Borel sets.

$$\lambda(B) = \inf \sum_{j=1}^n \lambda(I)_{j=1,2,\dots,n}$$

Note that the infimum is defined over all families of rectangles $(I)_{j=1,2,\dots,n}$ such that $B \subset \cup_{j=1}^n I_j$.

Definition A family of rectangles $(I)_{j=1,2,\dots,n}$ such that $B \subset \cup_{j=1}^n I_j$ is called **cover** of B .

Definition The above defined measure $\lambda(B)$ is called **Lebesgue measure**.

Insert here Figure 17

Lebesgue Integral

Now we can define the **Lebesgue Integral** on Borel sets.

$$\int_B f(x)dx = \sup \sum_{j=1}^n (\inf_{x \in B_j} f(x)) \cdot \lambda(B_j)$$

Note that for **Lebesgue Integral** we do not need a continuity assumption on density function, it is sufficient that the density function is measurable. Lebesgue Integral coincides with Riemann Integral if it is defined on rectangles; $\int_B f(x)dx = \int_{I=B} f(x)dx$

Lecture 4 / Week 3

Random Vectors and Integrals

Suppose we have the probability space (Ω, \mathcal{F}, P) : And a random vector $X : \Omega \rightarrow \mathbb{R}^K$. We defined the probability distribution of X as $\mu_X(B) = P(X^{-1}(B)) = P(X \in B)$ which can be either discrete or absolutely continuous. We also defined the the Lebesgue Integral on Borel sets $\mu_X(B) = \int_B f(x)dx$ for every Borel set B . This integral is defined for nonnegative functions, since the density function $f(x) \geq 0$. It can also be defined for negative functions in the following way:

$$\begin{aligned} f & : \mathbb{R}^K \rightarrow \mathbb{R}, \text{ measurable} \\ f^+(x) & = \max(f(x), 0) \\ f^-(x) & = -\min(f(x), 0) \\ f^+(x) & = \begin{cases} f(x) & f(x) \geq 0 \\ 0 & f(x) < 0 \end{cases} \\ f^-(x) & = \begin{cases} -f(x) & f(x) < 0 \\ 0 & f(x) \geq 0 \end{cases} \end{aligned}$$

Insert here Figure 18

Definition $\int_A f(x)dx = \int_A f^+(x) - \int_A f^-(x)$, since $f = f^+ - f^-$ (Note that we used the linearity of integral). Then we also have that

$$f^+(x) + f^-(x) = |f(x)| = \begin{cases} f^+(x) & f(x) \geq 0 \\ f^-(x) & f(x) < 0 \end{cases}$$

Exercise Show that if we integrate over B and $\int_B f(x)dx = 0$ if $\lambda(B) = 0$.

Proof Recall that $\int_B f(x)dx = \sup \sum_{j=1}^n (\inf_{x \in B_j} f(x)) \cdot \lambda(B_j) = 0$ and $\lambda(B_j) = 0 \Rightarrow \lambda(B_j) = 0 \forall j$, since $\lambda(\cup_{j=1}^n B_j) = \sum_{j=1}^n \lambda(B_j) = \lambda(B)$. Then this implies that $\sup \sum_{j=1}^n (\inf_{x \in B_j} f(x)) \cdot \lambda(B_j) = 0 = \int_B f(x)dx$.

Definition (*Almost everywhere*) The measurable functions f, g are such that $f = g$ except on a Borel set of Lebesgue measure zero ($\lambda(B) = 0$), then

$$\int_A f(x)dx = \int_A g(x)dx$$

Formally, the measurable functions f, g are **almost everywhere** equal iff there exists a Borel set $N = \{x \in \Omega : f(x) \neq g(x)\}$ with $\lambda(N) = 0$.

Proof Intuitively we can split the domain into two parts, where the two functions have equal values and where they have different values. Then we can integrate;

$$\int_{\mathbb{R}^K} (f(x) - g(x))dx = \int_{\{g=f\}} (f(x) - g(x))dx + \int_{\{g \neq f\}} (f(x) - g(x))dx$$

In the part of the domain where the two functions are the same, the integral is zero (first term in summation). Where the function values are different, then by hypothesis we have the measure 0, as we just proved that the integral is also zero (then RHS of equality 0) and hence the assertion holds, i.e. $\int_{\mathbb{R}^K} f(x)dx = \int_{\mathbb{R}^K} g(x)dx$.

Insert here Figure 19

The above proof and the figure is a crucial observation. If we change the function on countable infinite points, the integral of the two functions would not change! The picture shows the case for one point change in the domain $\int_{[0,1]} f(x)dx = \int_{[0,1]} g(x)dx$, where $\begin{cases} f(x) & x \neq \frac{1}{2} \\ 0 & x = \frac{1}{2} \end{cases}$, but then the set is $N = \{x \in \Omega : f(x) \neq g(x)\} = \frac{1}{2}$, and $\lambda(\frac{1}{2}) = 0$. Notice that it could also be the case for countable infinite points as long as they have zero Lebesgue measure. (For instance natural numbers in real line.)

Suppose we have an absolutely continuous distribution: $\mu_X(B) = \int_B f(x)dx$. An important implication what we have just seen is that the density function f is not uniquely determined, i.e. for different density functions we can still have the same probability. The following graph is such an example:

$$f(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Insert here Figure 20

Another example to see this phenomenon is the uniform distribution on $[0,1]$:

$$f(x) = \begin{cases} 1 & 1 \geq x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$f(x) = \begin{cases} 1 & 1 > x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Both are suitable density function for uniform distribution.

In general, there are many versions of the density function of an absolutely continuous probability distribution. Two versions can differ only on a set of Lebesgue measure 0. Two versions of the same probability distribution are *almost everywhere equal*.

Relation between Density and Cumulative Functions

Consider absolutely continuous probability distribution:

$$\begin{aligned}
 F(x_1, x_2, \dots, x_k) &= P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_k \leq x_k) \\
 &= P((X_1, \dots, X_k) \in (-\infty, x_1] \times (-\infty, x_2] \times \dots \times (-\infty, x_k]) \\
 &= \int_{(-\infty, x_1] \times (-\infty, x_2] \times \dots \times (-\infty, x_k]} f(s_1, s_2, \dots, s_k) ds_1 ds_2 \dots ds_k = \\
 &= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_k} f(s_1, s_2, \dots, s_k) ds_1 ds_2 \dots ds_k = \\
 \mathbf{F}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k) &= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_k} \mathbf{f}(\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_k) d\mathbf{s}_1 d\mathbf{s}_2 \dots d\mathbf{s}_k \\
 \frac{\partial F(x_1, x_2, \dots, x_k)}{\partial x_1 \partial x_2 \dots \partial x_k} &= f(x_1, x_2, \dots, x_k), \quad f \text{ is continuous in } (x_1, x_2, \dots, x_k).
 \end{aligned}$$

Independent Random Variables

Suppose we have two events A, B on a probability space (Ω, \mathcal{F}, P) . Then these two events are independent iff

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

Homework Review conditional probability and Bayes Rule.

Suppose we have a sequence of events: A_1, A_2, \dots, A_n , they are independent if for **any** choice of n and indices i_1, i_2, \dots, i_n

$$P(A_{i_1} \cap A_{i_2} \dots A_{i_n}) = P(A_{i_1}).P(A_{i_2}).\dots.P(A_{i_n})$$

Example If we have 3 events A, B, C, they are independent if $n=2$:

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

$$P(A \cap C) = P(A).P(C)$$

$$P(C \cap B) = P(C).P(B)$$

Since it holds for any n , we can also take $n=3$:

$$P(A \cap B \cap C) = P(A).P(B).P(C)$$

This will be used to define independent random variables. (discrete or abs. cont.)

Definition Suppose X_1, X_2 are random variables on (Ω, \mathcal{F}, P) are independent if for every sequence B_1, B_2, \dots of Borel sets the events $(X_1 \in B_1), (X_2 \in B_2), \dots$

are independent. Note that also the sigma-algebras generated by these random vectors $(X_1 \in B_1) \in \sigma(X_1), (X_2 \in B_2) \in \sigma(X_2), \dots$ are also independent. [recall the dice example in the book: once we roll the dice, the event of having an even number generates the sigma-algebra $F_X = \{(1, 2, 3, 4, 5, 6), \emptyset, (1, 3, 5), (2, 4, 6)\}$]

$$P(X_1 \in B_1, X_2 \in B_2, \dots, X_k \in B_k) = P(X_1 \in B_1) \times P(X_2 \in B_2) \times \dots \times P(X_k \in B_k)$$

for \forall Borel sets B_1, B_2, \dots

The σ -algebras generated by the random vectors are independent, i.e. take one event from $\sigma(X_1)$ and another event from $\sigma(X_2)$, then these events are independent. In other words the information on $X_1(\sigma(X_1))$ and the information on $X_2(\sigma(X_2))$ are independent if every event of $\sigma(X_1)$ is independent from every event in $\sigma(X_2)$.

Theorem The random variables (X_1, X_2, \dots, X_K) are independent if and only if the distribution functions of X_1, X_2, \dots, X_K

$$F_{(X_1, X_2, \dots, X_k)}(x_1, x_2, \dots, x_k) = F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdot \dots \cdot F_{X_k}(x_k)$$

for $\forall x_1, x_2, \dots, x_k$

If X_1, X_2, \dots, X_K are discrete

$$P(X_1 = x_1, X_2 = x_2, \dots, X_K = x_K) = P(X_1 = x_1) \cdot P(X_2 = x_2) \cdot \dots \cdot P(X_K = x_K)$$

If X_1, X_2, \dots, X_K have absolutely continuous probability distribution, then independence is equivalent to

$$f_{x_1, x_2, \dots, x_k} = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot \dots \cdot f_{X_k}(x_k) \text{ for } \forall x_1, x_2, \dots, x_k$$

where $f_{X_1}(x_1)$ is the density function of the random vector.

for suitable versions of densities. (\Rightarrow At least one version exists.)

Expectation of a Random Variable

The first thing to note is the expectation of the random variable does not depend whether it is discrete or absolutely continuous. We will define it for three different cases:

1. **Simple Random Variable:** Suppose we have the probability space (Ω, \mathcal{F}, P) and the simple random variable such that

$$X = \sum_{i=1}^n a_i \cdot \mathbf{1}_{A_i}$$

Then we define the expectation as

$$E(X) = \sum_{i=1}^n a_i \cdot P(A_i)$$

$$X = \left\{ \begin{array}{ll} a_1 & \omega \in A_1 \\ a_2 & \omega \in A_2 \\ \cdot & \cdot \\ a_n & \omega \in A_n \end{array} \right\}$$

Observe that $E(X)$ is the mean of the values taken by X (a_1, \dots, a_n) weighted by the probabilities of A_1, \dots, A_n .

2. **Non-negative Random Variable:** Let $X \geq 0$. Recall that we can approximate the random variable by a sequence of simple random variables. In this case it is defined as

$$E(X) = \sup E(X_*)$$

$$0 \leq X_* \leq X$$

$$X_* \text{ is simple.}$$

The idea is approximate X by a sequence of X_n of simple random variables $X_n \geq 0$, s.t $X = \lim_{n \rightarrow \infty} X_n$. (We have already proved it.)

$$E(X) = \lim_{n \rightarrow \infty} E(X_n)$$

Notice that $E(X)$ can be $+\infty$, since it is a limit, even though the terms in the limit cannot be ∞ , since they are simple function which by definition take finite values.

3. **General Random Variable:** The expectation of X can be defined as

$$E(X) = E(X^+) - E(X^-)$$

if at least one of the expectations is finite.

Convention if $E(X^+) < \infty$ and $E(X^-) = \infty \Rightarrow E(X^+) - E(X^-) = -\infty$.
if $E(X^+) = \infty$ and $E(X^-) < \infty \Rightarrow E(X^+) - E(X^-) = +\infty$.

Definition If $E(X^+)$ and $E(X^-)$ are both finite then $E(X)$ is finite and X is **integrable**.

$$E(X^+) < \infty \text{ and } E(X^-) < \infty \Rightarrow E(X) \text{ integrable}$$

$$E(X^+) < \infty \text{ and } E(X^-) = \infty \Rightarrow E(X) = -\infty$$

$$E(X^+) = \infty \text{ and } E(X^-) < \infty \Rightarrow E(X) = +\infty$$

$$E(X^+) = \infty \text{ and } E(X^-) = \infty \Rightarrow E(X) \text{ not defined.}$$

The definition of integrability is tricky, because we can integrate functions that are not integrable, but we will then obtain ∞ as the integral.

Example $\int_1^\infty \frac{1}{x} dx = \lim_{n \rightarrow \infty} \int_1^n \frac{1}{x} dx = \lim_{n \rightarrow \infty} [\log x]_1^n = \infty$. Here the function itself is not integrable by definition, but the integral is ∞ .

Insert here Figure 21

Relation between Expectation and Integrals

$$E(X) = \int_{\Omega} X dP$$

w.r.t a probability measure P .

If we have a simple function

$$\begin{aligned} \int_{\Omega} X dP &= a_1 P(A_1) + a_2 P(A_2) + \dots + a_n P(A_n) \\ &= \sum_{i=1}^n a_i \cdot P(A_i) = E(X) \end{aligned}$$

Insert here Figure 22

Definition $\int_A X dP = \int X \cdot \mathbf{1}_A dP$. This equality tells us that we integrate only over those parts of the domain where it belongs to set A and set the rest to zero. (Using the indicator function.)

Insert here Figure 23

Theorem We have the following **properties for expectation**

- (a) $E(c) = c$ (Notice that the constant function is a simple function $\Rightarrow \sum_{i=1}^n c \cdot P(A_i) = c \cdot \sum_{i=1}^n P(A_i) = c \cdot 1 = c = E(c)$)
- (b) $E(aX + bY) = aE(X) + bE(Y) \Leftrightarrow$ Expectation is a linear operator.

Proof We use the following definition of Expectation $\Rightarrow E(X) = \lim_{n \rightarrow \infty} E(X_n)$. We will prove for $X \geq 0$ and $Y \geq 0$. (General: $X = X^+ - X^-$, $Y = Y^+ - Y^-$). We take the following sequences $X_n \uparrow X$ and $Y_n \uparrow Y$. For

$a, b \geq 0$, $aX_n + bY_n$. (property of simple functions) and $\uparrow aX + bY$ (property of simple functions). Then

$$\begin{aligned}
 E(a.X + b.Y) &= \lim_{n \rightarrow \infty} E(a.X_n + b.Y_n) \stackrel{(*)}{=} \lim_{n \rightarrow \infty} (a.E(X_n) + b.E(Y_n)) \\
 &= \text{prop. lim.} \cdot a. \lim_{n \rightarrow \infty} (E(X_n)) + b. \lim_{n \rightarrow \infty} (E(Y_n)) = a.E(X) + b.E(Y). \text{ QED.} \\
 &\quad (*) \text{ since } (a.X_n + b.Y_n) \text{ simple} \\
 &= E\left(\sum_{i=1}^n a_i \cdot \mathbf{1}_{A_i} + \sum_{j=1}^m b_j \cdot \mathbf{1}_{B_j}\right) \stackrel{\text{def.}}{=} \sum_{i=1}^n a_i \cdot P(A_i) + \sum_{j=1}^m b_j \cdot P(B_j) \\
 \text{recall def. } X &= \sum_{i=1}^n a_i \cdot \mathbf{1}_{A_i}, E(X) = \sum_{i=1}^n a_i \cdot P(A_i)
 \end{aligned}$$

(c) If $X \geq 0$, then $E(X) \geq 0$

Proof Let X be non-negative simple random variable; since $X = \sum_{i=1}^n a_i \cdot \mathbf{1}_{A_i}$, $X \geq 0 \Leftrightarrow a_i \geq 0$, hence $E(X) = \sum_{i=1}^n a_i \cdot P(A_i) \geq 0$, since $P(A_i) \geq 0$ by definition. Then use the following definition of expectation: $E(X) = \sup_{0 \leq X_* \leq X} E(X_*) \geq 0$, $X_* \geq 0 \Leftrightarrow E(X_*) \geq 0 \Rightarrow \sup_{0 \leq X_* \leq X} E(X_*) \geq 0$. QED.

(d) If $X \leq Y$, then $E(X) \leq E(Y) \Leftrightarrow$ monotonicity

Proof By hypothesis $X \leq Y \Leftrightarrow Y - X \geq 0$. From previous proof $Y - X \geq 0 \Rightarrow E(Y - X) \geq 0$. Using linearity, $E(Y) - E(X) \geq 0$. QED.

(e) $|E(X)| \leq E(|X|)$

Proof We know that $x \leq |x|$. By monotonicity $E(x) \leq E(|x|)$. We also know $-x \leq |x|$. By monotonicity $E(-x) \leq E(|x|)$. By linearity, $-E(x) \leq E(|x|)$. So $E(x) \leq E(|x|)$ and $-E(x) \leq E(|x|)$ imply that $|E(x)| \leq E(|x|)$. QED.

Question If $X_{n \rightarrow \infty} \rightarrow X$, is it always true that $E(X_n) \rightarrow E(X)$?

Answer Not in general. This can be explained with the following counter example: Suppose we have $\Omega = (0, 1)$, $\mathcal{B}(0, 1)$ and $P = \lambda$

Insert here Figure 24

Then, as one can see from the above figures;

$$\begin{aligned}
 X_1(\omega) &= 1 \quad \forall \omega \in (0, 1), \\
 X_2(\omega) &\begin{cases} 2 & \omega \in (0, \frac{1}{2}) \\ 0 & \text{otherwise} \end{cases}, \\
 X_3(\omega) &\begin{cases} 4 & \omega \in (0, \frac{1}{4}) \\ 0 & \text{otherwise} \end{cases},
 \end{aligned}$$

but then $E(X_n) = 1$ for $\forall n$, $X_{n \rightarrow \infty} \rightarrow 0$ and $E(0) = 0$, therefore $E(X_n) \not\rightarrow E(0)$.

Monotone Convergence Theorem

If we have an nondecreasing sequence of non negative random variables , i.e. $X_n \geq 0$ $X_n \leq X_{n+1} \forall n$, then

$$X_{n \rightarrow \infty} \rightarrow X \quad E(X_n) \rightarrow E(X)$$

it can also diverge to ∞ .

Note that in book notation we have $E(X_n) = \lim_{n \rightarrow \infty} \int g_n(x) du(x)$, $E(X) = \int \lim_{n \rightarrow \infty} g_n(x) du(x)$. $\Rightarrow \lim_{n \rightarrow \infty} \int g_n(x) du(x) = \int \lim_{n \rightarrow \infty} g_n(x) du(x) \Leftrightarrow E(X_n) = E(X)$

Dominated Convergence Theorem

If there exists an **integrable** random variable Y such that $(X_n) \leq Y$ for every n , then

$$X_{n \rightarrow \infty} \rightarrow \textit{pointwise} X \quad E(X_n) \rightarrow E(X)$$

it converges.(finite)

Note that in book notation we have $X_{n \rightarrow \infty} \rightarrow \textit{pointwise} X \Leftrightarrow \lim_{n \rightarrow \infty} g_n(x) = g(x)$. $Y = \bar{g}(x) = \sup_{n \geq l} |g_n(x)|$ and integrable $Y \Leftrightarrow \int \bar{g}(x) du(x) < \infty$.

Lecture 5 / Week 4

OUTLINE

- 1) How to compute expectations? Book §2.3, Theorem 2.18
- 2) Inequalities. *Book* §2.6
- 3) Product of independent random variables *Book* §2.7

How to compute expectations?

Since the formula: $E(X) = \sup_{0 \leq X_* \leq X} E(X_*)$ is not very useful for computation, we need another tool to compute expectations.

Let X be a k -dimensional random vector, μ_X its probability distribution and $g : \mathbb{R}^K \rightarrow \mathbb{R}$ a measurable function ($B(\mathbb{R}^K)/B(\mathbb{R})$). Note that if g is the identity function, then $E(g(X)) = E(X)$. Notice also that the random variable $Y = g(X)$ is a function defined as $Y(\omega) = g(X(\omega))$. Since g is a measurable

function $\Rightarrow g(X)$ is measurable and $E(g(X))$ is defined for this random variable. There will be three cases for computation, namely, the general computation formula, the case where the distribution of X is discrete and the case where the distribution of X is absolutely continuous.

Theorem The *general computation formula for expectation* is the following

$$E(g(X)) = \int_{\mathbb{R}^K} g(x) \cdot \mu_X(dx)$$

Note that we have the following chain where the function g is defined $(\Omega, \mathcal{F}, P) \rightarrow^X (\mathbb{R}^K, B(\mathbb{R}^K), \mu_X) \rightarrow^g (\mathbb{R}, B(\mathbb{R}))$. Recall that

$$\begin{aligned} E(g(\omega)) &= \int_{\Omega} g(\omega) dP \rightarrow 3 \text{ cases} \\ g \text{ is simple} &\rightarrow \sum_{i=1}^n a_i \cdot P(A_i) \\ g &\geq 0 \rightarrow \sup_{0 \leq g_* \leq g} \int_{\Omega} g_*(\omega) dP \\ \text{general } g &\rightarrow \int_{\Omega} g^+(\omega) dP - \int_{\Omega} g^-(\omega) dP \end{aligned}$$

We define the integral $\int_{\Omega} g(\omega) dP$ on the probability space (Ω, \mathcal{F}, P) . Equivalently, we can define it on $(\mathbb{R}^K, B(\mathbb{R}^K), \mu_X)$ s.t we have $\int_{\mathbb{R}^K} g(x) \cdot \mu_X(dx)$. (One might also encounter the notations: $\int_{\mathbb{R}^K} g(x) \cdot \mu_X(dx) = \int_{\mathbb{R}^K} g(x) \cdot d\mu_X = \int_{\mathbb{R}^K} g(x) \cdot dF(x)$). Hence we have

$$E(g(X)) = \int_{\Omega} g(X) dP = \int_{\mathbb{R}^K} g(x) \cdot \mu_X(dx)$$

The former formula might be useful in finding the expectation of the sum of two random variables ($\int_{\Omega} (X + Y) dP$), whereas the latter one is suitable for calculations.

The expectation is well defined if and only if $\int_{\mathbb{R}^K} g(x) \cdot \mu_X(dx)$ is well defined, hence $E(g(X))$ is finite iff $\int_{\mathbb{R}^K} g(x) \cdot \mu_X(dx)$ finite and if both are finite their values are equal. Therefore we can conclude that this equality holds.

Before we proceed with the proof of the theorem, recall that when $X \geq 0$, $E(X) = \lim_{n \rightarrow \infty} E(X_n)$ is equivalent to saying $E(X) = \sup_{0 \leq X_n \leq X} E(X_n)$, where X_n is simple random variable, $0 \leq X_n \leq X$ and $X_n \uparrow X$ (limit from below).

Proof We will proof the result for 3 different cases, namely, 1.) g is simple, 2.) $g \geq 0$. 3) general case.

1.) Case 1: Suppose that g is a simple function s.t $g(x) = \sum_{i=1}^n a_i \cdot 1_{A_i}(x)$, w.l.o.g take A_i 's disjoint. On the other hand, when $g(X)$ is simple, then we have $g(X(\omega)) = \sum_{i=1}^n a_i \cdot 1_{A_i}(X(\omega)) = \sum_{i=1}^n a_i \cdot 1_{X^{-1}(A_i)}(\omega)$. But then $E(g(X)) = \sum_{i=1}^n a_i \cdot P(X^{-1}(A_i)) = \sum_{i=1}^n a_i \cdot \mu_X(A_i)$. Since $\int_{\mathbb{R}^K} g(x) \cdot d\mu_X = \sum_{i=1}^n a_i \cdot \mu_X(A_i)$ given that $g(x) = \sum_{i=1}^n a_i \cdot 1_{A_i}(x)$, we have shown that $E(g(X)) = \int_{\mathbb{R}^K} g(x) \cdot \mu_X(dx)$ if g is a simple function.

2.) Case 2: Suppose $g \geq 0$ and let g_n be simple nonnegative function s.t $g_n \uparrow g$. By the first case we know that

$$E(g_n(X)) = \int_{\mathbb{R}^K} g_n(x) \cdot \mu_X(dx)$$

By monotone convergence theorem, if take the limit, $n \rightarrow \infty$, since $g_n(X) \uparrow g(X)$, then $E(g_n(X)) \rightarrow E(g(X))$. Also by definition of integral, we know that if we take the limit, $n \rightarrow \infty$, $\int_{\mathbb{R}^K} g_n(x) \cdot \mu_X(dx) \rightarrow \int_{\mathbb{R}^K} g(x) \cdot \mu_X(dx)$. Hence, we have shown that the equality holds: $E(g(X)) = \int_{\mathbb{R}^K} g(x) \cdot \mu_X(dx)$.

3.) Case 3: The general case for g : We will use the fact that $g(x) = g^+(x) - g^-(x)$ and $g(X) = g^+(X) - g^-(X)$. Since both $g^+ \geq 0$ and $g^- \geq 0$, we can use the second case,

$$\begin{aligned} E(g(X)) &= E(g^+(X)) - E(g^-(X)) \\ \int_{\mathbb{R}^K} g(x) \cdot \mu_X(dx) &= \int_{\mathbb{R}^K} g^+(x) \cdot \mu_X(dx) - \int_{\mathbb{R}^K} g^-(x) \cdot \mu_X(dx) \end{aligned}$$

Since we have shown in case 2: $E(g^+(X)) = \int_{\mathbb{R}^K} g^+(x) \cdot \mu_X(dx)$ and $E(g^-(X)) = \int_{\mathbb{R}^K} g^-(x) \cdot \mu_X(dx)$. We have that

$$E(g(X)) = \int_{\mathbb{R}^K} g(x) \cdot \mu_X(dx)$$

Note that if both expectations are $E(g^+(X)) = \infty, E(g^-(X)) = \infty$, then $E(g(X))$ is not defined. QED.

Still we have the answer the question of how to compute the integral once we have either a discrete distribution of X and absolutely continuous distribution of X .

Case 1: μ_X is discrete distribution, i.e $\mu_X(A) = \sum_{s \in S \cap A} m(s)$, where S is at most countable and $m(s)$: the mass function. ($m(s) = p(s) = P(X = x)$, $p(s)$ =probability function). Then

$$\int_{\mathbb{R}^K} g(x) \cdot \mu_X(dx) = \sum_{s \in S} g(s) \cdot m(s) \Leftrightarrow E(X) = \sum x \cdot p(x)$$

Proof Exercise. Prove it for simple functions. (Hint use the definition of integral.)

Let $g(x) = \sum_{i=1}^n a_i \cdot 1_{A_i}(x)$, w.l.o.g take A_i 's disjoint. And let $g(s) = \sum_{i=1}^n a_i \cdot 1_{A_i}(s)$. We already know by definition of integral that $\int_{\mathbb{R}^K} g(x) \cdot d\mu_X =$

$\sum_{i=1}^n a_i \cdot \mu_X(A_i)$. Since $\mu_X(A) = \sum_{s \in S \cap A} m(s)$ and $\mu_X(A) = \sum_{i=1}^n \mu_X(A_i)$, we have $\mu_X(A_i) = \sum_{s \in S \cap A_i} m(s)$, then $\int_{\mathbb{R}^K} g(x) \cdot d\mu_X = \sum_{i=1}^n a_i \cdot \sum_{s \in S \cap A_i} m(s)$, since $\sum_{s \in S \cap A_i} m(s) = \sum_{s \in S} 1_{A_i}(s) \cdot 1_S(s) \cdot m(s)$. Then since $1_S(s) = 1 \forall s \in S$ $\int_{\mathbb{R}^K} g(x) \cdot d\mu_X = \sum_{i=1}^n a_i \cdot \sum_{s \in S} 1_{A_i}(s) \cdot m(s) = \sum_{s \in S} \sum_{i=1}^n a_i \cdot 1_{A_i}(s) \cdot m(s) = \sum_{s \in S} g(s) \cdot m(s)$. QED.

Case 2: μ_X is absolutely continuous distribution: $\mu_X(A) = \int_A f(x) dx$. (LHS: Lebesgue integral). Then

$$\int_{\mathbb{R}^K} g(x) \cdot \mu_X(dx) = \int_{\mathbb{R}^K} g(x) \cdot f(x) dx$$

Proof Exercise. Prove it for simple function g .

Let $g(x) = \sum_{i=1}^n a_i \cdot 1_{A_i}(x)$, w.l.o.g take A_i 's disjoint. We already know by definition of integral that $\int_{\mathbb{R}^K} g(x) \cdot d\mu_X = \sum_{i=1}^n a_i \cdot \mu_X(A_i)$. Since $\mu_X(A_i) = \int_{A_i} f(x) dx$, then $\int_{\mathbb{R}^K} g(x) \cdot d\mu_X = \sum_{i=1}^n a_i \cdot \int_{A_i} f(x) dx = \sum_{i=1}^n a_i \cdot \int_{\mathbb{R}^K} 1_{A_i}(x) \cdot f(x) dx = \int_{\mathbb{R}^K} \sum_{i=1}^n a_i \cdot 1_{A_i}(x) \cdot f(x) dx = \int_{\mathbb{R}^K} g(x) \cdot f(x) dx$.

Inequalities involving Mathematical Expectation

There are some inequalities involving mathematical expectation which turn out to be useful. Before proceeding to these inequalities, we mention some basic definitions:

Definition The **m'th moment** of a random variable is defined as $E(X^m)$, whereas the **m's central moment** is defined by $E(|X - \mu_X|^m)$, where $\mu_X = E(X)$. The second central moment is called **variance** of X , denoted by $\text{var}(X) = E[(X - \mu_X)^2] = E(X^2) - (E(X))^2 = \sigma_X^2$. The **covariance** of a pair of random variables (X, Y) is defined as $\text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$, where $\mu_Y = E(Y)$.

1.) Chebishev's Inequality: Let $X \geq 0$, i.e nonnegative random variable with distribution function $F(x)$ and let $\varphi(x)$ be monotonic, increasing, nonnegative measurable function. Then

$$P(X > \varepsilon) = 1 - F(\varepsilon) \leq \frac{E[\varphi(X)]}{\varphi(\varepsilon)}$$

Exercise Show that it holds for elementary case: $P(|X - \mu_X| > \varepsilon) \leq \frac{\text{var}(X)}{\varepsilon^2}$.

Proof Since $\varphi(x)$ be monotonic, increasing, nonnegative measurable function, we can let $X \rightarrow |X - \mu_X|$ and pick the following function

$$\varphi(X) = \begin{cases} X^2 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

then

$$P(|X - \mu_X| > \varepsilon) \leq \frac{E(|X - \mu_X|^2)}{\varepsilon^2} = \frac{\text{var}(X)}{\varepsilon^2}. \quad \text{QED.}$$

2.) Cauchy-Schwartz Inequality: This is a special case of Holder's inequality. It says that

$$\begin{aligned} E(|X.Y|) &\leq \sqrt{E(X^2)}.\sqrt{E(Y^2)} \\ |cov(X, Y)| &\leq \sqrt{var(X)}.\sqrt{var(Y)} \\ |E[(X - \mu_X)(Y - \mu_Y)]| &\leq E[|(X - \mu_X)(Y - \mu_Y)|] \leq \sqrt{E[(X - \mu_X)^2]}.\sqrt{E[(Y - \mu_Y)^2]} = \\ &= \sqrt{var(X)}.\sqrt{var(Y)} \end{aligned}$$

Holder's inequality is

$$\begin{aligned} E(|X.Y|) &\leq (E(|X|^p))^{\frac{1}{p}}.(E(|Y|^q))^{\frac{1}{q}} \\ \text{where } p > 1, \quad \frac{1}{p} + \frac{1}{q} &= 1 \end{aligned}$$

$|E[(X - \mu_X)(Y - \mu_Y)]| \leq E[|(X - \mu_X)(Y - \mu_Y)|]$ follows from property of expectation (Jensen's Inequality). Recall that $cov(X, Y) = E(X.Y) - E(X).E(Y)$ and $\rho(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)}.\sqrt{var(Y)}}$. From this inequality it follows that $|\rho(X, Y)| \leq 1$.

3.) Liapounov's Inequality: This also follows from Holder's inequality. For $0 < \alpha < \beta$

$$(E(|X|^\alpha))^{\frac{1}{\alpha}} \leq (E(|X|^\beta))^{\frac{1}{\beta}}$$

This inequality holds trivially if RHS is infinite and LHS is finite. Moreover, if $E(|X|^\beta)$ is finite, this implies that $E(|X|^\alpha)$ finite. Formally, if X has finite moment of order k, then X has finite moment order j $\forall j \leq k$.

Example If $E(X^2) < \infty$, then $E(X) < \infty \Rightarrow E(X)$ exists and is finite. ($\beta = 2, \alpha = 1$).

4.) Jensen's Inequality: Let $\varphi(X)$ be convex measurable function (given $E(\varphi(X))$ exists) and X simple random variable. Then

$$E(\varphi(X)) \geq \varphi(E(X))$$

Example $\varphi(X) = X^2 \Leftrightarrow E(X^2) \geq (E(X))^2$
 $\varphi(X) = |X| \Leftrightarrow E(|X|) \geq |E(X)|$

Insert here Figure 25

Expectation of Products of Independent Random Variables

Let X, Y be independent random vectors and f, g measurable functions. Then $f(X)$ and $g(Y)$ are also independent and $E(f(X).g(Y))$. Notice that

$$\begin{aligned} P(f(X) \in A, g(Y) \in B) &= P(X \in f^{-1}(A), Y \in g^{-1}(B)) = \\ &= \text{ind. } P(X \in f^{-1}(A)).P(Y \in g^{-1}(B)) = P(f(X) \in A).P(g(Y) \in B) \end{aligned}$$

Theorem If X, Y be independent random vectors, then

$$E(f(X).g(Y)) = E(f(X)).E(g(Y))$$

This theorem implies that independent random variables are uncorrelated, but the reverse is not true in general.

Proof Suppose f, g are simple functions s.t $f(X) = \sum_{i=1}^n a_i.1_{A_i}(X)$, $g(Y) = \sum_{j=1}^m b_j.1_{B_j}(Y)$. Recall that $1_{A_i}(X) = 1_{X^{-1}(A_i)}(\omega)$. Then

$$\begin{aligned} E(f(X).g(Y)) &= E\left(\sum_{i=1}^n a_i.1_{A_i}(X) \cdot \sum_{j=1}^m b_j.1_{B_j}(Y)\right) = \\ &= E\left(\sum_{i=1, j=1}^{n, m} a_i.b_j.1_{A_i}(X).1_{B_j}(Y)\right) = \\ &= E\left(\sum_{i=1, j=1}^{n, m} a_i.b_j.1_{X^{-1}(A_i) \cap Y^{-1}(B_j)}(\omega)\right) = \\ &= \sum_{i=1, j=1}^{n, m} a_i.b_j.E(1_{X^{-1}(A_i) \cap Y^{-1}(B_j)}(\omega)) = \\ &= \sum_{i=1, j=1}^{n, m} a_i.b_j.P(X^{-1}(A_i) \cap Y^{-1}(B_j)) = \\ &= \text{ind. } \sum_{i=1, j=1}^n a_i.b_j.P(X \in A_i)P(Y \in B_j) = \\ &= \sum_{i=1}^n a_i.P(X \in A_i) \cdot \sum_{j=1}^m b_j.P(Y \in B_j) = \\ &= E(f(X)).E(g(Y)). \text{ QED} \end{aligned}$$

We have shown that the equality holds for a particular case where f, g are simple functions. We have to show that it also holds for general f, g .

Exercise Show the above equality for general f, g .

Proof Let $f, g \geq 0$ and $f_n \uparrow f, g_n \uparrow g$, where f_n and g_n are simple functions, as we showed before, then $f_n(X)$ and $g_n(Y)$ are independent. The simple case we showed in the previous proof, i.e.

$$E(f_n(X).g_n(Y)) = E(f_n(X)).E(g_n(Y))$$

Then we know from the limit property $[(\lim (a_n.b_n)=\lim a_n.\lim b_n)$ that

$$\lim_{n \rightarrow \infty} [f_n(X(\omega)).g_n(Y(\omega))] = \lim_{n \rightarrow \infty} f_n(X(\omega)). \lim_{n \rightarrow \infty} g_n(Y(\omega))$$

Using the monotone convergence theorem we know

$$(LHS)_{n \rightarrow \infty} E(f(X).g(Y)) \rightarrow (RHS) E(f(X)).E(g(Y))$$

Again given the limit property, i.e. if $a_n = b_n \forall n$ and $a_n \rightarrow a$ and $b_n \rightarrow b$, then $a = b$. Hence we have

$$E(f(X).g(Y)) = E(f(X)).E(g(Y))$$

In the last part of the proof we will show the general case; exploiting the fact that

$$\begin{aligned} f(X) &= f^+(X) - f^-(X) \\ g(Y) &= g^+(Y) - g^-(Y) \end{aligned}$$

Then we can write

$$\begin{aligned} E(f(X).g(Y)) &= E[(f^+(X) - f^-(X)).(g^+(Y) - g^-(Y))] = \\ &= E[(f^+(X).(g^+(Y) - (f^+(X).g^-(Y) - f^-(X).g^+(Y) + f^-(X).g^-(Y))] \end{aligned}$$

Since the expectation operator is linear

$$= E[(f^+(X).(g^+(Y))] - E[(f^+(X).g^-(Y))] - E[f^-(X).g^+(Y)] + E[f^-(X).g^-(Y)] =$$

Since all the terms are positive valued function we can use the previous part of the proof,

$$= E(f^+(X)).E((g^+(Y)) - E(f^+(X)).E(g^-(Y)) - E(f^-(X)).E(g^+(Y)) + E(f^-(X)).E(g^-(Y)) =$$

Collecting the terms we obtain

$$E(f(X).g(Y)) = [E(f^+(X)) - E(f^-(X))].[E((g^+(Y) - E(g^-(Y))]. \quad QED.$$

Lecture 6 / Week 4

OUTLINE

- 1) Examples of Conditional Expectation.
- 2) Definition of of Conditional Expectation. *Book* 3.1
- 3) Properties of Conditional Expectation *Book* §3.2

Conditional Expectation

Example 1 Consider the following Game of Chance: A coin is tossed 10 times and the winnings of the game is denoted by $Y = 1\$$ per head. There are two cases: 1.) The player can enter the game soon. 2.) The player can enter the game after the first 6 tosses. (still receives the gain of previous tosses.)

Question: What is the fair price of the game in both cases?

Answer Case 1: The fair price of the game = $5\$ = E(Y)$: expected winnings of the game.

Case 2: Define X as the number of heads in the first 6 tosses, then the fair price would be $E(Y | X) = X + 2$: the expected winnings given the information about the outcome of the first 6 tosses, e.g. if there were 6 heads in the first 6 tosses than $E(Y | X) = 8$.

Example 2 (Information is given by a σ -algebra, not a random variable). Consider the following Game of Chance: A box contains a couple of dice: red dice and blue dice. The blue dice is numbered from 1-6 and the red dice is numbered from 7-12. We take a dice randomly and throw it. The winnings of the game is denoted by $Y =$ the score of the dice in $\$$. There are two cases: 1.) The player can enter the game at the beginning. 2.) The player can enter the game after having seen the color of the dice.

Question: What is the fair price of the game?

Answer Case 1: At the beginning of the game: $E(Y) : \frac{1+2+3+4+5+6}{12} = 6.5\$$

Case 2: After having seen the color of the dice:

$$E(Y|\mathcal{F}_0) \text{ where } \mathcal{F}_0 = \{\emptyset, \Omega, Blue, Red\}$$
$$E(Y|\mathcal{F}_0) = \left\{ \begin{array}{l} Blue: \frac{1+2+3+4+5+6}{6} = 3.5 \\ Red: \frac{7+8+9+10+11+12}{6} = 9.5 \end{array} \right\}$$

Observations

- $E(Y|\mathcal{F}_0)$ is a random variable.

- Its value is completely determined if we know which events of \mathcal{F}_0 occur, i.e. once we know the color in the above example we can calculate the conditional expectation.
- $E(Y|\mathcal{F}_0)$ has the meaning of expectation.

Definition of Conditional Expectation

Definition Suppose that Y is an *integrable* random variable defined on the probability space (Ω, \mathcal{F}, P) . Let \mathcal{F}_0 be a sub σ -algebra of \mathcal{F} . Then, the **conditional expectation** of Y given \mathcal{F}_0 is a random variable Z on same probability space (Ω, \mathcal{F}, P) such that

1. Z is \mathcal{F}_0 -measurable.
2. For every event $A \in \mathcal{F}_0 \rightarrow \int_A Z dP = \int_A Y dP$

The first property can be interpreted using the previous example, once i know the color (an event in \mathcal{F}_0), i can find which value Z takes, i.e. $Z = z$. The second property just says that the mean value of Y and Z is the same.

Definition If we know completely about the experiment, (Blue dice and 5), then we have complete information, formally

$$\mathbf{Full\ information} := E(Y|\mathcal{F})$$

while we have partial information if we only know the color of the dice, but not the number itself, formally

$$\mathbf{Partial\ Information} := E(Y|\mathcal{F}_0)$$

Example We have the following sample space $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. The numbers 1-6 are in the blue dice, 7-12 in the red dice. We can define two different random variables.

$$\mathit{Random\ Variable1} \quad : \quad Y = 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 10 \ 11 \ 12$$

$$\mathit{Random\ Variable2} \quad : \quad E(Y|\mathcal{F}_0) = 3, 5 \ 3, 5 \ 3, 5 \ 3, 5 \ 3, 5 \ 3, 5 \ 3, 5 \ 3, 5 \ 3, 5 \ 3, 5 \ 3, 5 \ 3, 5 \ 3, 5$$

The first random variable can takes values from 1 to 12, whereas the second one (which is the conditional expectation, but a random variable itself, let's say

Z) takes either the value 3.5 or 9.5. Then the mean values depending on the event (blue or red)

| | |
|------------------------|----------------------------|
| On Blue: mean value of | $E(Y \mathcal{F}_0) = 3.5$ |
| On Blue: mean value of | $Y = 3.5$ |
| On Red: mean value of | $E(Y \mathcal{F}_0) = 9.5$ |
| On Red: mean value of | $Y = 9.5$ |

This explains the second property of the definition that the mean values are the same w.r.t the same event in the σ -algebra.

Notation Almost surely (**a.s.**) = with probability 1

Examples

$$\begin{aligned} Z_1 &= Z_2 \text{ a.s.} \Leftrightarrow P\{\omega : Z_1(\omega) = Z_2(\omega)\} = 1 \\ Z_1 &\leq Z_2 \text{ a.s.} \Leftrightarrow P\{\omega : Z_1(\omega) \leq Z_2(\omega)\} = 1 \\ Z_{n \rightarrow \infty} &\rightarrow Z \text{ a.s.} \Leftrightarrow P\{\omega : Z_{n \rightarrow \infty}(\omega) \rightarrow Z(\omega)\} = 1 \end{aligned}$$

Proposition $E(Y|\mathcal{F}_0)$ exists whenever Y is integrable.

Proof Hint: Take a random variable Y and a σ -algebra \mathcal{F}_0 it is possible to find a Z that satisfies the properties of the definition of conditional expectation, i.e Z is \mathcal{F}_0 -measurable and every event $A \in \mathcal{F}_0 \rightarrow \int_A Z dP = \int_A Y dP$.

The question is, is Z uniquely determined?

The answer is almost, we can find Z_1 and Z_2 that satisfy the properties of the definition of conditional expectation and it can be proved that $Z_1 = Z_2$ **a.s.**

Properties of Conditional Expectation

Theorem The following properties hold:

- 1.) $E(c|\mathcal{F}_0) = c$ **a.s.**
- 2.) $E(a_1.Y_1 + a_2.Y_2|\mathcal{F}_0) = E(a_1.Y_1|\mathcal{F}_0) + E(a_2.Y_2|\mathcal{F}_0)$ **a.s.** (i.e. it is *linear.*)

Example First property. Take $E(c|\mathcal{F}_0)$ and let $Z_1 = c$, then Z_1 satisfies the two properties of the definition of conditional expectation. Let $Z_2 = c + \mathbf{1}_A$ $A \in \mathcal{F}_0$ $P(A) = 0$. Z_2 also satisfies the two properties of the definition of conditional expectation.

Insert here Figure 26

Exercise Show that if $Z_1 = c$, then Z_1 satisfies the two properties of the definition of conditional expectation.

Proof 1.) c is \mathcal{F}_0 -measurable. 2.) For every event $A \in \mathcal{F}_0 \rightarrow \int_A c dP = c = \int_A Y dP$

3.) Conditional expectation is *monotone*:

If $Y_1 \leq Y_2$ **a.s** then $E(Y_1|\mathcal{F}_0) \leq E(Y_2|\mathcal{F}_0)$ **a.s**

4.) $|E(Y|\mathcal{F}_0)| \leq E(|Y| |\mathcal{F}_0)$ **a.s.**

5.) *Dominated Convergence Theorem*:

If $Y_{n \rightarrow \infty} \rightarrow Y$ **a.s** and there exists an integrable random variable Z s.t $|Y_n| \leq Z$, for every n , then

$$E(Y_n|\mathcal{F}_0)_{n \rightarrow \infty} \rightarrow E(Y|\mathcal{F}_0) \quad \mathbf{a.s}$$

and $E(Y|\mathcal{F})$ must be integrable. If we take Y_n integrable random variables and $Y_n \uparrow Y$ and Y is integrable by definition, we can use the dominated convergence theorem instead of monotone convergence theorem. It becomes redundant since in case of conditional expectation Y must be integrable.

Theorem $E(E(Y|\mathcal{F}_0)) = E(Y)$.

Proof We use the second property of conditional expectation

$$\begin{aligned} \int_A E(Y|\mathcal{F}_0) dP &= \int_A Y dP \quad \forall A \in \mathcal{F}_0 \\ \text{since } \Omega &\in \mathcal{F}_0 \text{ pick } A = \Omega \\ \int_{\Omega} E(Y|\mathcal{F}_0) dP &= \int_{\Omega} Y dP \quad \forall A \\ E(E(Y|\mathcal{F}_0)) &= E(Y). \text{ QED.} \end{aligned}$$

Theorem If Y is \mathcal{F}_0 -measurable, then $E(Y|\mathcal{F}_0) = Y$ **a.s.**

Proof We have to prove that Y satisfies the definition of the conditional expectation.

1. Y is \mathcal{F}_0 -measurable. (Y as Z in the definition).
2. $\int_A Y dP = \int_A Y dP$. (Y in LHS refers to RHS in the theorem, viceversa)

Proposition If Y is \mathcal{F}_0 -measurable, then Y behaves like a constant. $E(Y|\mathcal{F}_0) = Y \Leftrightarrow E(c) = c$.

Theorem If Y is \mathcal{F}_0 -measurable, then

$$E(Y.Z|\mathcal{F}_0) = Y.E(Z|\mathcal{F}_0) \quad \mathbf{a.s}$$

Proof Observe that Y behaves like a constant $\Rightarrow E(c.X) = c.E(X)$

Lecture 7 / Week 5

OUTLINE

- 1) Further Properties of Conditional Expectation *Book* §3.2
- 2) Conditional Probability and its distribution. *Book* 3.1, 3.3

Conditional Expectation

Recall the definition, i.e suppose that Y is an *integrable* random variable defined on the probability space (Ω, \mathcal{F}, P) . Let \mathcal{F}_0 be a sub σ -algebra of \mathcal{F} . Then, the *conditional expectation* of Y given \mathcal{F}_0 is a random variable Z on same probability space (Ω, \mathcal{F}, P) such that

1. Z is \mathcal{F}_0 -measurable.
2. For every event $A \in \mathcal{F}_0 \rightarrow \int_A Z dP = \int_A Y dP$

We have also seen the following properties if Y is \mathcal{F}_0 -measurable, then;

$$\begin{aligned} E(Y|\mathcal{F}_0) &= Y \text{ a.s} \\ E(Y.Z|\mathcal{F}_0) &= Y.E(Z|\mathcal{F}_0) \text{ a.s} \end{aligned}$$

Now suppose we have $\mathcal{F}_0, \mathcal{F}_1$ and the random variable Y . The question is whether the following equality holds in general;

$$E(Y|\mathcal{F}_0) \text{ is r.v} \rightarrow E(E(Y|\mathcal{F}_0)|\mathcal{F}_1) \stackrel{?}{=} E(E(Y|\mathcal{F}_1)|\mathcal{F}_0)$$

So in other words, does the order matter? The answer is in general yes, the order matters, i.e they are not equal.

Theorem Let $\mathcal{F}_0 \subseteq \mathcal{F}_1$, then

$$E(E(Y|\mathcal{F}_0)|\mathcal{F}_1) = E(E(Y|\mathcal{F}_1)|\mathcal{F}_0) = E(Y|\mathcal{F}_0)$$

Proof First we will proof the equality between first and third term; Given that $E(Y|\mathcal{F}_0)$ is \mathcal{F}_0 -measurable

$$E(Y|\mathcal{F}_0)^{-1}(B) \in \mathcal{F}_0 \subseteq \mathcal{F}_1 \Rightarrow E(Y|\mathcal{F}_0) \text{ is also } \mathcal{F}_1\text{-measurable.}$$

But then

$$E(E(Y|\mathcal{F}_0)|\mathcal{F}_1) = E(Y|\mathcal{F}_0) \text{ a.s}$$

It is left to prove that

$$E(E(Y|\mathcal{F}_1)|\mathcal{F}_0) = E(Y|\mathcal{F}_0).$$

Call $Z = E(Y|\mathcal{F}_0)$ and we'll check whether it satisfies the definition of the conditional expectation of $E(Y|\mathcal{F}_1)$ given \mathcal{F}_0 .

Z is \mathcal{F}_0 -measurable by definition.

Take an event $A \in \mathcal{F}_0$, so it is also true that $A \in \mathcal{F}_0 \subseteq \mathcal{F}_1$, but we want to show that the following holds

$$\int_A Z dP = \int_A E(Y|\mathcal{F}_1) dP$$

since the event belongs to both σ -algebras and following the definition of conditional expectation, the RHS above equals

$$\int_A E(Y|\mathcal{F}_0) dP = \int_A Y dP$$

On the other hand from the definition of conditional expectation Z , we know that it equals to the LHS below. Hence it follows that

$$\int_A Y dP = \int_A Y dP. \quad \text{QED.}$$

Theorem Suppose we have $E(Y|\mathcal{F}_0)$, Y and \mathcal{F}_0 are independent. This is same as saying $\sigma(Y)$ and \mathcal{F}_0 are independent. Then

$$E(Y|\mathcal{F}_0) = E(Y) \quad \text{a.s}$$

Proof $E(Y)$ is \mathcal{F}_0 -measurable. (Because it behaves like a constant \rightarrow constant random variable)

$$c^{-1}(B) = \{\omega \in \Omega : c \in B\} = \begin{cases} \emptyset & c \notin B \\ \Omega & c \in B \end{cases}$$

since every σ -algebra contains (Ω, \emptyset) . So, in general every constant variable is measurable w.r.t. every σ -algebra.

Definition $\sigma(Y)$ and \mathcal{F}_0 are independent. (or Y and \mathcal{F}_0 are independent.). For every $A \in \sigma(Y)$ and $B \in \mathcal{F}_0 \Rightarrow A, B$ are independent

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

So knowing that Y and \mathcal{F}_0 are independent,

$$\begin{aligned} A &\in \mathcal{F}_0 \\ \int_A E(Y)dP &= \int_A YdP = E(Y).P(A) \\ \int_A YdP &= \int \mathbf{1}_A.YdP = E(\mathbf{1}_A.Y) = E(\mathbf{1}_A).E(Y) = E(Y).P(A) \quad (\text{RHS}) \end{aligned}$$

Exercise We exploited the fact that $\mathbf{1}_A$ and Y are independent. Show that it's true.

Proof We know that $\sigma(Y)$ and \mathcal{F}_0 are independent. Take an event $A \in \mathcal{F}_0$. Notice that

$$\{\omega \in \Omega : \mathbf{1}_A(\omega) \in B_1\} = \left\{ \begin{array}{lll} A & 1 \in B_1 & 0 \notin B_1 \\ A^c & 1 \notin B_1 & 0 \in B_1 \\ \emptyset & 1 \notin B_1 & 0 \notin B_1 \\ \Omega & 1 \in B_1 & 0 \in B_1 \end{array} \right\}$$

Insert here Figure 27

But then

$$B_1, B_2 \quad P(\mathbf{1}_A(\omega) \in B_1, Y \in B_2)$$

since the event $(\mathbf{1}_A(\omega) \in B_1)$ belongs \mathcal{F}_0 , which follows from the fact that $A \in \mathcal{F}_0$ and \mathcal{F}_0 is a σ -algebra, then it also follows

$$P(\mathbf{1}_A(\omega) \in B_1, Y \in B_2) = P(\mathbf{1}_A(\omega) \in B_1).P(Y \in B_2)$$

since $\sigma(Y)$ and \mathcal{F}_0 are independent and this completed the proof that $\mathbf{1}_A$ and Y are independent.

Example Toss a coin 10 times. Let Y = the number heads and X = the number heads after 8 trials. We need to formalize $E(Y|\sigma(X))$. Call

$$\begin{aligned} Z &= Y - X := \text{the number heads in the last 2 trials} \\ Y &= X + Z \end{aligned}$$

Note that X and Z are independent, while X and Y are not. Then

$$\begin{aligned} E(Y|\sigma(X)) &= E(X + Z|\sigma(X)) \stackrel{\text{lin of E.}}{=} E(X|\sigma(X)) + E(Z|\sigma(X)) = \\ &= \stackrel{\text{prev.theorem}}{=} X + E(Z) = X + 1 \end{aligned}$$

So in fact the conditional expectation $E(Y|X)$ is a function in X

$$\begin{aligned} E(Y|X) &= X + 1 = g(X) \\ \text{where } g(x) &= x + 1 \end{aligned}$$

Note also that

$$E(Y|\sigma(X)) = E(Y|X)$$

Theorem Let Y be a random variable and X be a random vector. If Y is measurable w.r.t $\sigma(X)$, then there exists a function g s.t

$$Y = g(X)$$

Proof Before proving the theorem, we will show that the converse of the theorem holds, i.e.

$$Y = g(X) \Rightarrow Y \text{ is measurable w.r.t } \sigma(X)$$

We take the inverse of the r.v

$$\begin{aligned} Y^{-1}(B) &= \{\omega \in \Omega : Y(\omega) \in B\} = \{\omega \in \Omega : g(X(\omega)) \in B\} = \\ &= \{\omega \in \Omega : X(\omega) \in g^{-1}(B)\} \in \sigma(X) = \\ &= \{\omega \in \Omega : X^{-1}(g^{-1}(B))\} \in \sigma(X) \\ &\text{i.e } Y \text{ is measurable w.r.t. } \sigma(X). \end{aligned}$$

On the other hand, the theorem says that

$$E(Y|\sigma(X)) = E(Y|X) \text{ is a function of } X.$$

Suppose \mathcal{F}_0 is σ -algebra. We have the triple (Ω, \mathcal{F}, P) and $\mathcal{F}_0 \subseteq \mathcal{F}$. Also suppose that the event $B \in \mathcal{F}$. We use the following definition

Definition

$$\begin{aligned} P(B|\mathcal{F}_0) &= E(\mathbf{1}_B|\mathcal{F}_0) \\ E(\mathbf{1}_B) &= P(B) \end{aligned}$$

We observe that

1. $P(B|\mathcal{F}_0)$ is \mathcal{F}_0 -measurable random variable. (by definition of conditional expectation.)
2. $A \in \mathcal{F}_0$, $\int_A P(B|\mathcal{F}_0)dP \stackrel{def. E(\mathbf{1}_B|\mathcal{F}_0)}{=} \int_A \mathbf{1}_B dP = \int \mathbf{1}_A \mathbf{1}_B dP = \int \mathbf{1}_{A \cap B} dP = P(A \cap B)$. Hence

$$\begin{aligned} \forall A \in \mathcal{F}_0, \\ \int_A P(B|\mathcal{F}_0)dP &= P(A \cap B) \end{aligned}$$

Since the conditional expectation is monotone we have

$$\begin{aligned} 0 &\leq \mathbf{1}_B \leq 1 \\ E(0|\mathcal{F}_0) &\leq E(\mathbf{1}_B|\mathcal{F}_0) \leq E(1|\mathcal{F}_0) \quad \text{a.s} \\ 0 &\leq P(B|\mathcal{F}_0) \leq 1 \\ \text{since } P(B|\mathcal{F}_0) &= E(\mathbf{1}_B|\mathcal{F}_0) \text{ and } E(0|\mathcal{F}_0) = 0, E(1|\mathcal{F}_0) = 1 \end{aligned}$$

Definition We can generalize the previous result. Let both X and Y be random vectors. Then

$$\begin{aligned} P(Y \in B|X) &= g(X) \quad B \in \mathcal{B}(\mathbb{R}^K) \\ P(Y \in B|X) &= P(\{\omega \in \Omega : Y(\omega) \in B\}|X) \\ g(x) &= P(Y(\omega) \in B|X = x) \end{aligned}$$

this function is called **the conditional probability distribution** of Y given $X = x$. In fact, this is more general formulation of the wellknown conditional probability

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

which requires that $P(B) > 0$, whereas the new definition well defined over the whole domain.

We will introduce the conditional probability distribution function in two different cases, namely; discrete and absolutely continuous cases

1. (X, Y) *discrete random vector*

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

Then we can define the function as follows

$$p_{Y|X}(y|x) = \left\{ \begin{array}{ll} \frac{p_{X,Y}(x,y)}{p_X(x)} & p_X(x) \neq 0 \\ 0 & p_X(x) = 0 \end{array} \right\}$$

Then

$$P(Y \in B|X = x) = \sum_{y \in B} p_{Y|X}(y|x)$$

Then the **conditinal expectation** will be calculated by

$$E(g(Y)|X = x) = \sum_y g(y)p_{Y|X}(y|x)$$

2. (X, Y) *absolutely continuous random vector*

In this case we will have the following density function

$$f_{Y|X}(y|x) = \left\{ \begin{array}{ll} \frac{f_{X,Y}(x,y)}{f_X(x)} & f_X(x) \neq 0 \\ 0 & f_X(x) = 0 \end{array} \right\}$$

And consequently

$$P(Y \in B|X = x) = \int_B f_{Y|X}(y|x)dy$$

Then the **conditinal expectation** will be calculated by

$$E(g(Y)|X = x) = \int g(y)f_{Y|X}(y|x)dy$$

Definition Let A_1, A_2, \dots be events and let \mathcal{F}_0 be σ -algebra. A_1, A_2, \dots are said to be **conditionally independent** given \mathcal{F}_0 if for every n and i_1, i_2, \dots, i_n

$$P(\cap_{j=1}^n A_{i_j} | \mathcal{F}_0) = \prod_{j=1}^n P(A_{i_j} | \mathcal{F}_0) \quad \text{a.s}$$

Proposition Y_1, Y_2, \dots are conditionally independent given \mathcal{F}_0 if for every B_1, B_2, \dots (*Borel Sets*) the events $(Y_1 \in B_1), (Y_2 \in B_2)$ are conditionally independent given \mathcal{F}_0 .

Example Toss a coin 100 times. Let

- X_1 = the number of heads after first 10 tosses
- X_2 = the number of heads after first 50 tosses
- X_3 = the number of heads after first 70 tosses

Note that these three variables are not independent, but X_1 and X_3 are conditionally independent given X_2 . In other words, knowing X_2 (the number of heads after first 50 tosses), X_1 (the number of heads after first 10 tosses) does not give any information on X_3 . (the number of heads after first 70 tosses.)

Proof We will show that

$$P(X_1 = x_1, X_3 = x_3 | X_2 = x_2) = P(X_1 = x_1 | X_2 = x_2) \cdot P(X_3 = x_3 | X_2 = x_2)$$

First note that the random variables $X_1, X_2 - X_1$, and $X_3 - X_2$ are independent. Then

$$\frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3)}{P(X_2 = x_2)} \stackrel{\text{same event}}{=} \frac{P(X_1 = x_1, X_2 - X_1 = x_2 - x_1, X_3 - X_2 = x_3 - x_2)}{P(X_2 = x_2)}$$

We multiply and divide by $P(X_2 = x_2)$ and using independence

$$\begin{aligned}
 &= \frac{P(X_1 = x_1, X_2 - X_1 = x_2 - x_1)}{P(X_2 = x_2)} \cdot P(X_3 - X_2 = x_3 - x_2) = \\
 &= \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)} \cdot \frac{P(X_2 = x_2, X_3 - X_2 = x_3 - x_2)}{P(X_2 = x_2)}
 \end{aligned}$$

Also noting that $P(X_1 = x_1, X_2 - X_1 = x_2 - x_1)$ and $P(X_1 = x_1, X_2 = x_2)$ are the same events (strictly speaking probability of the same event), we can conclude that

$$P(X_1 = x_1 | X_2 = x_2) \cdot P(X_3 = x_3 | X_2 = x_2)$$

Lecture 8 / Week 5

OUTLINE

- 1) An example
- 2) Best Forecast Scheme
- 3) Conditioning on Increasing σ -Algebras
- 4) Distributions of Transformations of Random Vectors
- 5) Noteworthy Distributions

Example Suppose we have two random variables X, Y . We know that X has the exponential (λ) distribution with the density function (absolutely continuous)

$$f_X(x) = \lambda \cdot e^{-\lambda \cdot x} \mathbf{1}_{[0, \infty)}(x)$$

We also know that the conditional probability distribution is discrete: $(Y | X = x) \sim \text{Poisson}(x)$ Note $\lambda = x$.

$$P(Y = y | X = x) = \frac{e^{-x} \cdot x^y}{y!} \quad y = 1, 2, \dots$$

Suppose we want to compute $E(X \cdot Y) = E(g(X, Y))$. The problem is the function g has one discrete and one abs. continuous component, and we do not know how to compute the expectation in such a case. BUT, we can exploit the properties of the conditional expectation, i.e

$$E(X \cdot Y) = E(E(X \cdot Y | X)) = E(X \cdot E(Y | X))$$

But we know that the expected value of a variable with poisson distribution is $E(X) = \lambda$. Consequently, $E(Y|X = x) = x$, or $E(Y|X) = X$, but then

$$E(X.E(Y|X)) = E(X.X) = E(X^2)$$

Then using the formula

$$var(X) = E(X^2) - [E(X)]^2$$

and recalling that the expected value and the variance of a variable with exponential distribution; $E(X) = \frac{1}{\lambda}$, $var(X) = \frac{1}{\lambda^2}$, respectively, we have

$$E(X^2) = var(X) + [E(X)]^2 = \frac{1}{\lambda^2} + \frac{1}{\lambda^2} = \frac{2}{\lambda^2}. \quad \text{QED.}$$

This example is a good illustration where we have to exploit the properties of conditional expectation.

Best Forecast Scheme=Conditional Expectation

The conditional expectation $E(Y|X)$ has a special interpretation once the second moment of the variable Y has finite moment, formally $E(Y^2) < \infty$. Namely, it is the best forecast of Y based on X ; i.e suppose we have a random variable Y and a random vector X and we want to predict Y as a function of X . Then we claim that, which will prove in a second, $g(X)$ is the best predictor of Y , in the sense that it minimizes the prediction error. We need the following definition to clarify the problem

Definition We define the **mean square error of prediction (MSEP)** as follows $E[(Y - g(X))^2]$. In words, this is the expected error we are trying to minimize, naturally it is the expectation of the square of difference between the actual random value(Y) and our prediction($g(X)$). Note that it is squared so that positive and negative errors cannot cancel out.

Theorem The function g that minimizes MSEP is $g(X) = E(Y|X)$. Formally, for every function $g(X)$

$$E((Y - g(X))^2) \geq E((Y - E(Y|X))^2)$$

Proof We take $E((Y - g(X))^2)$ and do the add/subtract trick, i.e

$$E((Y - E(Y|X) + E(Y|X) - g(X))^2)$$

we group the terms and take the square

$$E((Y - g(X))^2) = E[((Y - E(Y|X)) + (E(Y|X) - g(X)))^2] =$$

$$= E[(Y - E(Y|X))^2] + E[(E(Y|X) - g(X))^2] + 2E[(Y - E(Y|X)) \cdot (E(Y|X) - g(X))] \quad (*)$$

Then we show that

$$\begin{aligned} E[(Y - E(Y|X)) \cdot (E(Y|X) - g(X))] &= 0 \\ E[E((Y - E(Y|X)) \cdot (E(Y|X) - g(X)) | X)] &= 0 \end{aligned}$$

In the second line we used the conditional expectation property, i.e. the expectation of the conditional expectation is the expectation itself, since the second term of the multiplication is fixed at X , behaves like a constant, we can take it out,

$$E[(E(Y|X) - g(X)) \cdot E(Y - E(Y|X) | X)]$$

but then

$$E(Y - E(Y|X) | X) = E(Y|X) - E(Y|X) = 0$$

Since in (*), the only term left that is dependent on $g(X)$ is $E[(E(Y|X) - g(X))^2]$, since it is a square, we can minimize the error by setting

$$E(Y|X) = g(X). \quad QED.$$

The previous result has important implication in econometrics and particularly it is fundamental in regression analysis.

Conditioning on Increasing σ -Algebras

We might want to make predictions in more general settings. Assume that we have the following sequence of random variables

$$\{Y_t\}_{t=-\infty}^{\infty}$$

Then we might want to know what the best prediction Y_t is based on the Y 's at times $t-1, t-2, \dots, t-m$.

$$E(Y_t | Y_{t-1}, Y_{t-2}, \dots, Y_{t-m})$$

as m goes larger ($m \rightarrow \infty$) the σ -algebra (the information set) becomes larger as well. We interested in answering questions such as; what happens to the conditional expectation as the sequence of σ -algebra's goes larger, i.e

$$\text{What is } \lim_{m \rightarrow \infty} E(Y_t | Y_{t-1}, Y_{t-2}, \dots, Y_{t-m})?$$

The answer lies in the following theorem.

Theorem Let \mathcal{F}_n be an increasing sequence of σ -algebras. (i.e. $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$). Let $\mathcal{F}_\infty = \bigvee_{n=1}^{\infty} \mathcal{F}_n$, which is the σ -algebra that contains the union of σ -algebras (since the union itself is not necessarily a σ -algebra.) Then

$$E(Y | \mathcal{F}_n)_{n \rightarrow \infty} \rightarrow E(Y | \mathcal{F}_\infty) \quad \mathbf{a.s}$$

Proof It is complicated and thus omitted.

As mentioned above we are interested in

$$E(Y_t|Y_{t-1}, Y_{t-2}\dots)$$

but now following the theorem we know that

$$E(Y_t|Y_{t-1}, Y_{t-2}\dots Y_{t-m}) \approx E(Y_t|Y_{t-1}, Y_{t-2}\dots)$$

if we have m large enough, in other words if we can collect as much past data as possible (m large depends on the data and application, e.g whether data is quarterly, weekly, or so) than we can be almost sure that we have all the past information about the random variable and the expectation conditioned (based on this information set) on this past data is going to be the best prediction of Y_t .

Distributions of Transformations of Random Vectors

We have the random vector X and the function g s.t $Y = g(X)$. We want to know how to compute the distribution of Y once we know the distribution of X . We will again analyse two different cases; discrete and absolutely continuous.

Case 1 We have a **discrete random vector** X :

$$p_X(x) = P(X = x) \quad S = \{x_1, x_2, \dots\} \text{ at most countable}$$

Then we will also have a discrete Y s.t $Y = g(X)$ and $S_y = \{y_1, y_2, \dots\}$ at most countable. Note that in fact we have a vector X so the correct notation should be

$$p_{X_1, \dots, X_k}$$

but we slightly abuse the notation. Then we will also have $g(x_i); \{g(x_1), g(x_2), \dots, g(x_k)\}$, but note that this set can contain less elements than k distinct functions, if some functions have the same value, in fact it only includes functions g with distinct values. Then

$$p_Y(y_i) = P(Y = y_i) = P(X \in g^{-1}(y_i)) = \sum_{g(x_j)=y_i} p_X(x_j)$$

We can better see this with an example:

Example We have the random vector $X = (X_1, X_2)$ where X_1, X_2 have independent Poisson(λ) distributions. We also have

$$\begin{aligned} g(x_1, x_2) &= x_1 + x_2 \\ Y &= X_1 + X_2 \\ S_x &= \{(0, 0), (0, 1), (1, 0), \dots\} \\ s_1 &= 0, 1, 2, \dots \\ s_2 &= 0, 1, 2, \dots \\ S_y &= \{0, 1, 2, \dots\} \end{aligned}$$

We are trying to calculate

$$P(Y = y) = P(X_1 + X_2 = y) \quad y = 0, 1, 2, \dots$$

Using the previous formula and the independence of two components

$$\sum_{x_1+x_2=y} P(X_1 = x_1, X_2 = x_2) \stackrel{ind.}{=} \sum_{x_1+x_2=y} P(X_1 = x_1).P(X_2 = x_2)$$

we use the poisson distribution

$$\begin{aligned} \sum_{x_1+x_2=y} P(X_1 = x_1).P(X_2 = x_2) &= \sum_{x_1+x_2=y} \frac{e^{-\lambda} \cdot \lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda} \cdot \lambda^{x_2}}{x_2!} = \\ P(Y = y) &= \frac{e^{-2\lambda} \cdot (2\lambda)^y}{y!} \quad y = 0, 1, 2, \dots \end{aligned}$$

Exercise Make the above calculation

$$\sum_{x_1+x_2=y} \frac{e^{-\lambda} \cdot \lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda} \cdot \lambda^{x_2}}{x_2!} = e^{-2\lambda} \cdot \lambda^y \cdot \sum_{x_1+x_2=y} \frac{1}{x_1! \cdot x_2!}$$

Hint multiply divide by $y!$ and use $(1+1)^y = \sum \binom{y}{x} 1^x \cdot 1^{y-x}$.

An important point to note from this example is that the sum of two independent poisson variables ended up to have a poisson distribution with the parameter $(2 \cdot \lambda)$.

Case 2 X has **absolutely continuous distribution**: We know that then we have

$$P(X \in B) = \int_B f_X(x) dx$$

Suppose we have $Y = g(X)$. Does the absolute continuity of X imply the same for Y ? The answer is no in general. Here we have a counterexample;

Example $X \sim N(0, 1)$, so it has standard normal distribution and Y is the following function of X :

$$\begin{aligned} Y &= \left\{ \begin{array}{ll} 0 & X \geq 0 \\ 1 & X < 0 \end{array} \right\} \\ Y &= \mathbf{1}_{(-\infty, 0)}(X) \end{aligned}$$

but Y is a discrete random variable. In fact Y has Bernoulli($\frac{1}{2}$) distribution.

$$\begin{aligned} P(Y = 0) &= P(X > 0) = \frac{1}{2} \\ P(Y = 1) &= P(X \leq 0) = \frac{1}{2} \end{aligned}$$

In general, once Y has a absolutely continuous distribution we have

$$\begin{aligned} P(Y \in B) &= P(g(X) \in B) = P(X \in g^{-1}(B)) = \\ &= \int_{g^{-1}(B)} f_X(x) dx \\ P(Y \leq y) &= \int_{g^{-1}(-\infty, y]} f_X(x) dx \end{aligned}$$

Still we have to impose some conditions to be able to find the density of Y directly from the density of X .

Theorem Let X be a k -dimensional random vector with absolutely continuous distribution and the density function $f_X(x)$. Let $g : \mathbb{R}^k \rightarrow \mathbb{R}^k$ be a

one to one function such that its **inverse** function g^{-1} is **differentiable**. Let $Y = g(X)$. Then Y has absolutely continuous distribution with density function

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \cdot |\det J(y)| \text{ on the set} \\ \{y &: y = g(x) \text{ with } f_X(x) > 0\} \\ \text{where } J_{ij}(y) &= \frac{\partial g_i^{-1}}{\partial y_j} \text{ (Jacobian)} \end{aligned}$$

The following example illustrates such a case

Example Suppose X_1, X_2 independent $N(0, 1)$, i.e.

$$f_{X_{1,2}}(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$$

Also suppose that we have (Y_1, Y_2) , so 1-1 requirement satisfied and

$$\begin{aligned} Y_1 &= X_1 + X_2 \Rightarrow X_1 = \frac{Y_1 + Y_2}{2} \\ Y_2 &= X_1 - X_2 \Rightarrow X_2 = \frac{Y_1 - Y_2}{2} \end{aligned}$$

so the functions $g_{1,2}$ are invertible (1-1)

$$\begin{aligned} X_1 &= g_1^{-1}(Y_1, Y_2) = \frac{Y_1 + Y_2}{2} \\ X_2 &= g_2^{-1}(Y_1, Y_2) = \frac{Y_1 - Y_2}{2} \end{aligned}$$

so they are also differentiable (condition). Then we can compute the Jacobian

$$\begin{aligned} J &= \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \\ |\det(J)| &= \frac{1}{2} \end{aligned}$$

Applying the formula and exploiting the independence we have

$$\begin{aligned} f_{Y_1, Y_2}(y_1, y_2) &= f_{X_1}\left(\frac{Y_1 + Y_2}{2}\right) \cdot f_{X_2}\left(\frac{Y_1 - Y_2}{2}\right) \cdot \frac{1}{2} \\ &= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{y_1 + y_2}{2}\right)^2} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{y_1 - y_2}{2}\right)^2} \cdot \frac{1}{2} = \\ &= \frac{1}{\sqrt{2\pi} \cdot \sqrt{2}} \cdot e^{-\frac{1}{2}\frac{y_1^2}{2}} \cdot \frac{1}{\sqrt{2\pi} \cdot \sqrt{2}} \cdot e^{-\frac{1}{2}\frac{y_2^2}{2}} \end{aligned}$$

So we have found that Y_1 and Y_2 are independent and have the normal distribution $N(0, 2)$.

Noteworthy Distributions

1. **Normal Distribution:** It is denoted by $N(\mu, \sigma^2)$

$$\begin{aligned} E(X) &= \mu \\ \text{var}(X) &= \sigma^2 \\ f_X(x) &= \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \end{aligned}$$

2. **Standard Normal Distribution:** It is denoted by $N(0, 1)$. It is a special case of normal distribution

$$\begin{aligned} E(X) &= 0 \\ \text{var}(X) &= 1 \\ E(X^4) &= 3 \\ f_X(x) &= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \end{aligned}$$

Exercise If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, then $Y \sim N(a\mu + b, a^2 \cdot \sigma^2)$

Proof It directly follows from the properties of expectation and variance.

$$\begin{aligned} E(X) &= \mu \\ E(Y) &= a \cdot E(X) + E(b) = a\mu + b \\ \text{var}(X) &= \sigma^2 \\ \text{var}(Y) &= a^2 \text{var}(X) = a^2 \cdot \sigma^2 \quad \text{var}(b)=0 \end{aligned}$$

3. **Chi-Square Distribution:** It is denoted by \mathcal{X}_n^2 . Let X_1, X_2, \dots, X_n be independent identically distributed random variables with standard normal distributions and set $Y = X_1^2 + X_2^2 + \dots + X_n^2$. This new random variable Y has chi-square distribution with n degrees of freedom.

$$\begin{aligned} f_Y(y) &= \frac{1}{\Gamma(\frac{n}{2}) \cdot 2^{\frac{n}{2}}} \cdot y^{\frac{n}{2}-1} \cdot e^{-\frac{y}{2}} \cdot \mathbf{1}_{(0, \infty)}(y) \\ E(Y) &= n \cdot E(X_1^2) = n \\ \text{var}(Y) &= n \cdot \text{var}(X_1^2) = n \cdot [E(X_1^4) - (E(X_1^2))^2] = n(3 - 1) = 2n \end{aligned}$$

where $\Gamma(\alpha)$ is called gamma function.

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} \cdot e^{-x} dx \quad \text{for } \alpha > 0$$

Exercise Show that $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.

Proof Since $\Gamma(\alpha - 1) = \int_0^\infty x^{\alpha-2} \cdot e^{-x} dx$ for $\alpha > 0$, we integrate by parts $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \cdot e^{-x} dx = [-e^{-x} \cdot x^{\alpha-1}]_0^\infty + \int_0^\infty (\alpha - 1)x^{\alpha-2} \cdot e^{-x} dx = (\alpha - 1)\Gamma(\alpha - 1)$.

Exercise $\Gamma(1) = 1$.

Proof Since $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \cdot e^{-x} dx$ for $\alpha > 0$, then $\Gamma(1) = \int_0^\infty x^0 \cdot e^{-x} dx = 1$

Exercise $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

Note that $\Gamma(n) = (n - 1)!$, this function is a generalization of the factorials where n can be real numbers not just integers.

4. **t Distribution** X, Y are independent $X \sim N(0, 1)$ and $Y \sim \chi_n^2$.

$$T = \frac{X}{\sqrt{\frac{Y}{n}}} \quad t - \text{distribution}$$

5. **F(Fisher) Distribution** $X \sim \chi_m^2, Y \sim \chi_n^2$

$$F = \frac{\frac{X}{m}}{\frac{Y}{n}} \quad F_{m,n}$$

Lecture 9 / Week 6

Multivariate Normal Distribution

OUTLINE

- 1) Expectation and Variance of Random Vectors
- 2) Definition of MND
- 3) Properties of MND

Expectation and Variance of Random Vectors

In this section we will introduce vector and matrix notation. From now on, a vector is always defined as a **column** vector. The superscript T, will be used to denote the transpose of the vector which will be row vector. Hence,

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ v_m \end{bmatrix}_{(m \times 1)} \quad \mathbf{v}^T = (v_1 \ v_2 \ \cdot \ \cdot \ \cdot \ v_m)_{(1 \times m)}$$

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}_{(nx1)} \quad \mathbf{E}(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{bmatrix}_{(nx1)}$$

where \mathbf{X} is a random vector and $\mathbf{E}(\mathbf{X})$ is its expected value. (again a vector)
Then we can also define the variance which will turn out to be a matrix.

$$\mathbf{var}(\mathbf{X})_{(n \times n)} = \mathbf{E}((\mathbf{X} - \mathbf{E}(\mathbf{X}))_{(nx1)} \cdot (\mathbf{X} - \mathbf{E}(\mathbf{X}))^T_{(1 \times n)})$$

To see how this matrix looks like we will analyse a simple case with $n=2$;

Example Let \mathbf{X} be $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$. Then the matrix will be

$$\begin{aligned} & \mathbf{E} \left(\begin{pmatrix} X_1 - E(X_1) \\ X_2 - E(X_2) \end{pmatrix} \cdot \begin{pmatrix} (X_1 - E(X_1)) & (X_2 - E(X_2)) \end{pmatrix} \right) = \\ \mathbf{E} \left(\begin{pmatrix} (X_1 - E(X_1))^2 & ((X_1 - E(X_1)) & (X_2 - E(X_2))) \\ ((X_1 - E(X_1)) & (X_2 - E(X_2))) & (X_2 - E(X_2))^2 \end{pmatrix} \right) = \\ & \begin{pmatrix} \mathit{var}(X_1) & \mathit{cov}(X_1, X_2) \\ \mathit{cov}(X_1, X_2) & \mathit{var}(X_2) \end{pmatrix} = \Sigma \end{aligned}$$

where $\sum_{ii} = \mathit{var}(X_i)$ and $\sum_{ij} = \mathit{cov}(X_i, X_j)$. This Σ which quite often used in econometrics is called the Σ = **variance covariance matrix**.

The following is a numerical example:

Example Let \mathbf{X} be $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$. Let $\mathit{var}(X_1) = 2$, $\mathit{var}(X_2) = 1$, $\mathit{cov}(X_1, X_2) = 1$.

Then $\mathbf{var}(\mathbf{X}) = \Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$.

Note also that if X_1 and X_2 are independent then $\mathit{cov}(X_1, X_2) = 0$ which amounts to saying that $\mathbf{var}(\mathbf{X})$ is a **diagonal matrix**; i.e.

$$\Sigma = \begin{pmatrix} \mathit{var}(X_1) & 0 \\ 0 & \mathit{var}(X_2) \end{pmatrix}$$

In general if X_1, X_2, \dots, X_n are independent then

$$\Sigma = \begin{pmatrix} \mathit{var}(X_1) & 0 & 0 & 0 \\ 0 & \mathit{var}(X_2) & 0 & 0 \\ 0 & 0 & \mathit{var}(X_3) & 0 \\ 0 & 0 & 0 & \mathit{var}(X_4) \end{pmatrix}$$

One should be cautious, because even though independence implies no covariance (no correlation) between two random variables, the reverse does not hold, i.e. no covariance does not imply independence. In other words, independence is a

stronger condition than covariance. This idea can be better understood given the following example;

Example Consider the following table $X \setminus Y$

| $X \setminus Y$ | -1 | 0 | 1 | Mar. Prob. |
|-----------------|---------------|---------------|---------------|---------------|
| -1 | 0 | $\frac{1}{4}$ | 0 | $\frac{1}{4}$ |
| 0 | $\frac{1}{4}$ | 0 | $\frac{1}{4}$ | $\frac{1}{2}$ |
| 1 | 0 | $\frac{1}{4}$ | 0 | $\frac{1}{4}$ |
| Mar. Prob. | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ | |

This table should be read as follows;

$$P(X = 0, Y = -1) = P(X = 0, Y = 1) = P(X = -1, Y = 0) = P(X = 1, Y = 0) = \frac{1}{4}.$$

$$P(X = -1, Y = -1) = P(X = 0, Y = 0) = P(X = 1, Y = -1) = P(X = 1, Y = 1) = 0.$$

$$P(X = 0) = \frac{1}{2}, P(X = -1) = P(X = 1) = \frac{1}{4}.$$

$$P(Y = 0) = \frac{1}{2}, P(Y = -1) = P(Y = 1) = \frac{1}{4}.$$

Given this information we can see that X and Y are not independent since for instance

$$P(X = -1, Y = -1) = 0$$

$$P(X = -1).P(Y = -1) = \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16}$$

On the other hand we can also calculate the expectation using the information given in the table:

$$E(X) = 0 \cdot \frac{1}{2} + (-1) \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = 0$$

$$E(Y) = 0 \cdot \frac{1}{2} + (-1) \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = 0$$

$$E(X.Y) = E(g(X, Y))$$

$$g(X, Y) = X.Y$$

$$E(X.Y) = \sum_{x,y} x.y.p_{X,Y}(x, y) = 0.$$

Proposition Given the random vector $X_{(n \times 1)}$ and the vector $a_{(n \times 1)}$ and the scalar b .

$$E(a^T . X + b) = a.E(X) + b$$

Proof Given the linearity of the expectation we have

$$E(a_1.X_1 + a_2.X_2 + \dots + a_n.X_n + b) = a_1.E(X_1) + a_2.E(X_2) + \dots + a_n.E(X_n) + b = a^T . E(X) + b$$

In general we have given the random vector $X_{(n \times 1)}$, the matrix $A_{(m \times n)}$ and the vector $b_{(m \times 1)}$, we have

$$E(A.X + b) = A.E(X) + b$$

Proof Exercise.

Now we will see the counterpart of the previous result for $\text{var}(Y) = \sum$, where $Y = a^T \cdot X$. Then we have

$$\begin{aligned} \text{var}(Y) &= E((a^T \cdot X - E(a^T \cdot X)) \cdot (a^T \cdot X - E(a^T \cdot X))^T) = \\ &= E(a^T \cdot (X - E(X)) \cdot (X - E(X))^T \cdot a) = \\ &= a^T \cdot E((X - E(X)) \cdot (X - E(X))^T) \cdot a = \\ &= a_{(1 \times n)}^T \cdot \text{var}(X)_{(n \times n)} \cdot a_{(n \times 1)}. \quad (\text{Note it's a positive scalar.}) \end{aligned}$$

We have the assumption that $E(Y) = E(a^T \cdot X) = 0$, then the above result simplifies to

$$\text{var}(Y) = \text{var}(a^T \cdot X) = E((a^T \cdot X) \cdot (a^T \cdot X)^T) = E(a^T \cdot X \cdot X^T \cdot a) = a^T \cdot E(X \cdot X^T) \cdot a$$

Proposition Let $X_{(n \times 1)}$ a random vector, $A_{(m \times n)}$ a matrix and $b_{(m \times 1)}$ another vector. Then

$$\text{var}(Y)_{(m \times m)} = A_{(m \times n)} \cdot \text{var}(X)_{(n \times n)} \cdot A_{(n \times m)}^T$$

Proof Exercise.

Definition If \sum is a variance covariance matrix, then for every vector a

$$a^T \cdot \sum \cdot a \geq 0 \Rightarrow \sum \text{ is nonnegative definite.}$$

$$a^T \cdot \sum \cdot a > 0 \Rightarrow \sum \text{ is positive definite.}$$

$$\sum \text{ is positive definite.} \Rightarrow \text{The inverse of } \sum \text{ exists.}$$

$$\text{The inverse of } \sum \text{ exists.} \Leftrightarrow \det \sum \neq 0.$$

recall that $\sum^{-1} \sum = \mathbf{I}$ (identity matrix.) \sum – the variance covariance matrix is non-negative definite and symmetric matrix. (i.e. $\sum^T = \sum$).

Definition Let X and Y be random vectors. Then

$$\text{cov}(X, Y) = E((X - E(X)) \cdot (Y - E(Y))^T)$$

Example Assume we have $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ and $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$. Then

$$\begin{aligned} \text{cov}(X, Y) &= \mathbf{E} \left(\begin{pmatrix} X_1 - E(X_1) \\ X_2 - E(X_2) \end{pmatrix}_{(2 \times 1)} \cdot ((Y_1 - E(Y_1)) \quad (Y_2 - E(Y_2)) \quad (Y_3 - E(Y_3)))_{(1 \times 3)} \right) = \\ &= \begin{pmatrix} \text{cov}(X_1, Y_1) & \text{cov}(X_1, Y_2) & \text{cov}(X_1, Y_3) \\ \text{cov}(X_2, Y_1) & \text{cov}(X_2, Y_2) & \text{cov}(X_2, Y_3) \end{pmatrix}_{(2 \times 3)} \end{aligned}$$

Question What is the relationship between $\text{var}(X_1), \text{var}(X_2)$ and $\text{cov}(X_1, X_2)$ if we have two random vectors such as $X_1_{(k \times 1)}$ and $X_2_{((n-k) \times 1)}$ such that we have $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}_{(n \times 1)}$.

Proposition In the above case we will have

$$\text{var} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}_{(n \times n)} = \begin{pmatrix} \text{var}(\mathbf{X}_1)_{(k \times k)} & \text{cov}(X_1, X_2)_{(k \times (n-k))} \\ \text{cov}(X_1, X_2)^T_{((n-k) \times k)} & \text{var}(\mathbf{X}_2)_{((n-k) \times (n-k))} \end{pmatrix}$$

Proof Exercise.

Example Assume that we have $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ and $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$, then $\begin{pmatrix} X \\ Y \end{pmatrix}$

will be $\begin{pmatrix} X_1 \\ X_2 \\ Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$. The following term will be its variance

$$\text{var} \begin{pmatrix} X_1 \\ X_2 \\ Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} \text{var}(\mathbf{X}_1) & \text{cov}(\mathbf{X}_1, \mathbf{X}_2) & \text{cov}(X_1, Y_1) & \text{cov}(X_1, Y_2) & \text{cov}(X_1, Y_3) \\ \text{cov}(\mathbf{X}_1, \mathbf{X}_2) & \text{var}(\mathbf{X}_2) & \text{cov}(X_2, Y_1) & \text{cov}(X_2, Y_2) & \text{cov}(X_2, Y_3) \\ \text{cov}(X_1, Y_1) & \text{cov}(X_2, Y_1) & \text{var}(\mathbf{Y}_1) & \text{cov}(\mathbf{Y}_1, \mathbf{Y}_2) & \text{cov}(\mathbf{Y}_1, \mathbf{Y}_3) \\ \text{cov}(X_1, Y_2) & \text{cov}(X_2, Y_2) & \text{cov}(\mathbf{Y}_1, \mathbf{Y}_2) & \text{var}(\mathbf{Y}_2) & \text{cov}(\mathbf{Y}_2, \mathbf{Y}_3) \\ \text{cov}(X_1, Y_3) & \text{cov}(X_2, Y_3) & \text{cov}(\mathbf{Y}_1, \mathbf{Y}_3) & \text{cov}(\mathbf{Y}_2, \mathbf{Y}_3) & \text{var}(\mathbf{Y}_3) \end{pmatrix}$$

to simplify the notation generally $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{21} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}$ is used.

Multivariate Normal Distribution

Recall that in the univariate case we have the normal distribution $N(\mu, \sigma^2)$ of a random variable X that has the following properties

$$\begin{aligned} E(X) &= \mu \\ \text{var}(X) &= \sigma^2 \\ f_X(x) &= \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}} \end{aligned}$$

Now we will analyze the analogous case for the random vector $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$.

Recall that we have an absolutely continuous distribution. Since the density function of a random vector is the joint density of its components X_1, X_2, \dots, X_n . Then we will have

$$\begin{aligned} E(X) &= \mu_{(n \times 1)} \\ \text{var}(X) &= \sum_{(n \times n)} \\ f_{X_{(n \times 1)}}(\mathbf{x}_{(n \times 1)}) &= (2\pi)^{-\frac{n}{2}} \cdot \det \Sigma^{-\frac{1}{2}} e^{-\frac{1}{2}\{(\mathbf{x}-\boldsymbol{\mu})^T \cdot \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\}} \end{aligned}$$

where Σ is not singular $\Rightarrow \det \Sigma \neq 0 \Rightarrow \exists \Sigma^{-1}$.

Example Suppose we have $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $E(X) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \mu$, $\text{var}(X) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} = \Sigma$. Hence a random vector with normal distribution $N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right)$. Then we can find $\det \Sigma$ and Σ^{-1}

$$\begin{aligned} \det \Sigma &= 1 \\ \Sigma^{-1} &= \frac{1}{1} \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \end{aligned}$$

we can always check $\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} = \mathbf{I}$. Then plugging in these results into our formula

$$\begin{aligned} f(x_1, x_2) &= (2\pi)^{-\frac{2}{2}} \cdot 1^{-\frac{1}{2}} \cdot \exp^{-\frac{1}{2}\{x_1 \quad x_2 \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\}} \\ f(x_1, x_2) &= \frac{1}{2\pi} \cdot \exp^{-\frac{1}{2}\{x_1^2 - 2x_1x_2 + 2x_2^2\}} \end{aligned}$$

Since we have found the density function then we can calculate for instance the following probability by integrating;

$$P(0 < X_1 < 1, X_2 > 2) = \int_0^1 \left(\int_2^\infty \frac{1}{2\pi} \cdot \exp^{-\frac{1}{2}\{x_1^2 - 2x_1x_2 + 2x_2^2\}} dx_2 \right) dx_1$$

$N_n(\mu, \Sigma)$ is the usual notation for multivariate normal distributions.

In the following part, we will analyze the analogous case for standard normal distribution:

Recall that in the univariate case we have $N(0, 1)$.

$$\begin{aligned} E(X) &= 0 \\ \text{var}(X) &= 1 \\ f_X(x) &= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \end{aligned}$$

The multivariate standard normal distribution will be denoted by $N_n(0, \mathbf{I}_n)$, where

$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}_{(n \times n)}$$

note that $\det \mathbf{I}_n = 1$ and $\mathbf{I}_n^{-1} = \mathbf{I}_n$. Keeping this in mind we can write the density

$$\begin{aligned} f_X(\mathbf{x}) &= (2\pi)^{-\frac{n}{2}} \exp^{-\frac{1}{2}\{\mathbf{x}^T \mathbf{I}_n \mathbf{x}\}} = (2\pi)^{-\frac{n}{2}} \cdot \exp^{-\frac{1}{2}\{\mathbf{x}^T \mathbf{x}\}} = \\ &= (2\pi)^{-\frac{n}{2}} \cdot \exp^{-\frac{1}{2}\sum_{j=1}^n x_j^2} = \\ \text{since } \mathbf{x}^T \mathbf{x} &= \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1^2 + x_2^2 + \dots + x_n^2 \\ &= (2\pi)^{-\frac{n}{2}} \cdot \prod_{j=1}^n \exp^{-\frac{1}{2}x_j^2} = \\ &= \prod_{j=1}^n \left(\frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}x_j^2} \right) \end{aligned}$$

notice that each term of the product is the density of the univariate standard normal distribution, hence we have X_1, X_2, \dots, X_n are i.i.d (independently and identically distributed) with $N(0, 1)$.

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{j=1}^n \varphi(x_j)$$

where φ is the density of the univariate standard normal distribution. So we have actually proved the following proposition:

Proposition The random vector $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$ has $N_n(0, \mathbf{I}_n)$ distribution

if and only if X_1, X_2, \dots, X_n are i.i.d.

Lecture 10 / Week 6

Properties of Multivariate Normal Distribution

Example Assume we have the random vector $X \sim N_n(\mu, \Sigma)$ and $Y_{(n \times 1)} = AX + b$, where A is a $(n \times n)$ nonsingular matrix and b is a $(n \times 1)$ vector. Find the probability distribution of Y .

We know that X has multivariate normal distribution, hence its density function is the following

$$f_X(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \cdot \det \Sigma^{-\frac{1}{2}} e^{-\frac{1}{2}\{(\mathbf{x}-\boldsymbol{\mu})^T \cdot \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\}}$$

To find the density function of Y , we have to do the following transformation

$$\begin{aligned} Y &= g(X) \Leftrightarrow X = g^{-1}(Y) \\ X_{(n \times 1)} &= A_{(n \times n)}^{-1} \cdot (Y - b)_{(n \times 1)} = g^{-1}(Y) \end{aligned}$$

Then we can apply the transformation formula

$$f_Y(y) = f_X(g^{-1}(y)) \cdot |\det J(y)|$$

First of all, notice that $J(y) = A^{-1}$. We plug it into the formula

$$\begin{aligned} f_Y(\mathbf{y}) &= (2\pi)^{-\frac{n}{2}} \cdot \det \Sigma^{-\frac{1}{2}} e^{-\frac{1}{2}\{(A^{-1} \cdot (\mathbf{y}-\mathbf{b})-\boldsymbol{\mu})^T \cdot \Sigma^{-1}(A^{-1} \cdot (\mathbf{y}-\mathbf{b})-\boldsymbol{\mu})\}} \cdot |\det A^{-1}| = \\ &= (2\pi)^{-\frac{n}{2}} \cdot |A|^{-1} \cdot |\Sigma|^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}\{(\mathbf{A}^{-1}\mathbf{y}-\mathbf{A}^{-1}\mathbf{b}-\boldsymbol{\mu})^T \cdot \Sigma^{-1}(\mathbf{A}^{-1}\mathbf{y}-\mathbf{A}^{-1}\mathbf{b}-\boldsymbol{\mu})\}} \end{aligned}$$

notice that $|\det A^{-1}| = |A|^{-1}$, $\det \Sigma^{-\frac{1}{2}} = |\Sigma|^{-\frac{1}{2}}$

$$|A|^{-1} = |A^2|^{-\frac{1}{2}} = |A \cdot A^T|^{-\frac{1}{2}} \Rightarrow |A \cdot A^T|^{-\frac{1}{2}} \cdot |\Sigma|^{-\frac{1}{2}} = |A \cdot \Sigma \cdot A^T|^{-\frac{1}{2}}$$

$$= (2\pi)^{-\frac{n}{2}} \cdot |A \cdot \Sigma \cdot A^T|^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}\{(\mathbf{A}^{-1}\mathbf{y}-\mathbf{A}^{-1}\mathbf{b}-\mathbf{A}^{-1} \cdot \mathbf{A} \cdot \boldsymbol{\mu})^T \cdot \Sigma^{-1}(\mathbf{A}^{-1}\mathbf{y}-\mathbf{A}^{-1}\mathbf{b}-\mathbf{A}^{-1} \cdot \mathbf{A} \cdot \boldsymbol{\mu})\}}$$

notice that we multiplied $\boldsymbol{\mu}$ with $A^{-1} \cdot A = \mathbf{I}_{(n \times n)}$

$$= (2\pi)^{-\frac{n}{2}} \cdot |A \cdot \Sigma \cdot A^T|^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}\{(\mathbf{y}-\mathbf{b}-\mathbf{A} \cdot \boldsymbol{\mu})^T (\mathbf{A}^{-1})^T \cdot \Sigma^{-1} \mathbf{A}^{-1} \cdot (\mathbf{y}-\mathbf{b}-\mathbf{A} \cdot \boldsymbol{\mu})\}}$$

we take out \mathbf{A}^{-1} . (Note transpose gets out as transpose)

$$= (2\pi)^{-\frac{n}{2}} \cdot |A \cdot \Sigma \cdot A^T|^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}\{(\mathbf{y}-\mathbf{b}-\mathbf{A} \cdot \boldsymbol{\mu})^T (A \cdot \Sigma \cdot A^T)^{-1} \cdot (\mathbf{y}-\mathbf{b}-\mathbf{A} \cdot \boldsymbol{\mu})\}}$$

$$= (2\pi)^{-\frac{n}{2}} \cdot |A \cdot \Sigma \cdot A^T|^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}\{(\mathbf{y}-(\mathbf{b}+\mathbf{A} \cdot \boldsymbol{\mu}))^T (A \cdot \Sigma \cdot A^T)^{-1} \cdot (\mathbf{y}-(\mathbf{b}+\mathbf{A} \cdot \boldsymbol{\mu}))\}}$$

Thus we have shown that $Y \sim N_n(\mathbf{A} \cdot \boldsymbol{\mu} + \mathbf{b}, A \cdot \Sigma \cdot A^T)$.

Exercise Show that given $g^{-1}(Y) = X = A^{-1} \cdot (Y - b)$, then $J(y) = A^{-1}$.

Proof Recall that $J_{ij}(y) = \frac{\partial g_i^{-1}}{\partial y_j}$. Since $X_{(n \times 1)}$ and $Y_{(n \times 1)}$, then $i=n, j=n$, so we will have an $n \times n$ matrix. (A^{-1} being a $n \times n$ matrix satisfies this.) Let's see

with $n=2$ case. Then we will have $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \cdot \begin{pmatrix} Y_1 - b_1 \\ Y_2 - b_2 \end{pmatrix} = \begin{pmatrix} a_{11}(Y_1 - b_1) + a_{12}(Y_2 - b_2) \\ a_{21}(Y_1 - b_1) + a_{22}(Y_2 - b_2) \end{pmatrix}$. Then

$$\frac{\partial g_i^{-1}}{\partial y_j} = \frac{\partial X_i}{\partial Y_j} = \begin{pmatrix} \frac{\partial X_1}{\partial Y_1} & \frac{\partial X_2}{\partial Y_1} \\ \frac{\partial X_1}{\partial Y_2} & \frac{\partial X_2}{\partial Y_2} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix} = A^{-1}.$$

Properties

Theorem Let $X \sim N_n(\mu, \Sigma)$ and $Y = AX + b$ with $A_{(n \times n)}$ non-singular matrix. Then $Y \sim N_n(A\mu + \mathbf{b}, A\Sigma A^T)$.

Proof We have just proved in the above example.

Assume we have $Y \sim N_n(\mu, \Sigma)$ and also $\text{rank}(\Sigma) = K$. We wonder if we can define a normal distribution once we have a singular $\Sigma (\equiv n > K)$. Recall that

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{21} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}$$

We relax the assumption that the whole Σ is nonsingular, but we will only assume that Σ_{11} is a $(k \times k)$ nonsingular matrix, i.e. Σ itself can be singular or not. Then we claim that the vector $Y = \begin{pmatrix} Y_{1(n \times 1)} \\ Y_{2(n-k \times 1)} \end{pmatrix}_{(n \times 1)} \sim N_n \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{21} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix} \right)$ if $Y_1 \sim N_k((\mu_1), (\Sigma_{11}))$ where Σ_{11} is nonsingular and $Y_2 = AY_1 + b$ for some A and b .

Example $\Sigma = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix}$, $\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$. We can see that Σ is singular, since the third row is a linear combination of the first two rows. ($\det \Sigma = 0$). In fact $\text{rank}(\Sigma) = 2$. Then we can split the Y vector into $Y = \begin{pmatrix} Y_{1(2 \times 1)} \\ Y_{2(1 \times 1)} \end{pmatrix}_{(3 \times 1)}$ and define $Y_1 \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)$ and $Y_2 = A_{(1 \times 2)} \cdot Y_{1(2 \times 1)} + b$. Notice that $\Sigma_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ is nonsingular. ($n=K=2$), where $E(Y_2) = 0$, $\text{var}(Y_2) = 2$.

Theorem Let $X \sim N_n(\mu, \Sigma)$ and $Y = AX + b$ and $A_{(m \times n)}$. (notice not a square matrix.) and b be vector of size m , s.t

$$Y_{(m \times 1)} = A_{(m \times n)} \cdot X_{(n \times 1)} + b_{(m \times 1)}$$

Then $Y \sim N_m(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$.

The above result holds whatever dimension and whatever rank of A .

Example $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $Y_{(3 \times 1)} = A_{(3 \times 2)}X_{(2 \times 1)} + b_{(3 \times 1)}$. Then $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$

has a normal distribution. Note that variance covariance matrix of Y is singular. ($\text{rank}(A_{(3 \times 2)} \cdot \Sigma_{(2 \times 2)} \cdot A_{(2 \times 3)}^T)_{(3 \times 3)} = 2$, because $\text{rank}(\Sigma_{(2 \times 2)}) = 2$).

Or another example would be $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$, $Y = A_{(2 \times 3)}X_{(3 \times 1)} + b_{(2 \times 1)}$,

then $Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2(A\boldsymbol{\mu} + b, (A\Sigma A^T)_{(2 \times 2)})$.

Another interesting case is when we have $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ X_n \end{bmatrix}_{(n \times 1)}$ and

$$A = (I_k | 0) = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \end{pmatrix}_{(k \times n)}.$$

$$\text{Then } Y = A \cdot X = (I_k | 0) \cdot \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \cdot \\ \mathbf{X}_k \\ X_{k+1} \\ \cdot \\ X_n \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ X_k \end{bmatrix} \sim N_k \left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \mu_k \end{bmatrix}, \text{var} \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ X_k \end{bmatrix} \right).$$

Theorem Let X_1 and X_2 be random vectors of dimensions k and $(n - k)$, respectively. If $\begin{pmatrix} \mathbf{X}_1_{(n \times 1)} \\ \mathbf{X}_2_{(n-k \times 1)} \end{pmatrix}_{(n \times 1)} \sim N_n \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sum_{11} & \sum_{21} \\ \sum_{12} & \sum_{22} \end{pmatrix} \right)$. Then $X_1 \sim N_k((\mu_1), (\sum_{11}))$ and $X_2 \sim N_{n-k}((\mu_2), (\sum_{22}))$.

Example $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N_3$ and then $X_1 \sim N$, $X_2 \sim N$, $X_3 \sim N$. $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2$, $\begin{pmatrix} X_1 \\ X_3 \end{pmatrix} \sim N_2$, $\begin{pmatrix} X_2 \\ X_3 \end{pmatrix} \sim N_2$.

This theorem asserts that every subvector of a random vector with a normal distribution has a normal distribution.

Now we will state two important properties of multivariate normal distribution:

1. As we have shown in the first example, normality is preserved by linear transformation: $Y = AX + b$.
2. Even though in general no covariance does not apply independence (recall yesterday's example), in normal distribution case $cov(X_1, X_2) = 0 \Rightarrow$ independence of X_1 and X_2 (i.e when $\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim N_n$). This property will be formalized and proved in the following theorem.

Theorem Let X_1 and X_2 be random vectors such that $\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim N_n$ (i.e it has multivariate normal distribution.). Then X_1 and X_2 are independent if and only if $cov(X_1, X_2) = 0$.

Proof The theorem says $cov(\mathbf{X}_1, \mathbf{X}_2) = 0 \Leftrightarrow$ independence of X_1 and X_2 under normal distribution, so we have to prove both directions, but " \Leftarrow " is already shown yesterday and it holds in general regardless of the underlying distribution. So it left to prove that under multivariate normal distribution " \Rightarrow " holds.

" \Rightarrow " Suppose $cov(X_1, X_2) = 0$, under normality assumption we have

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}_{(n \times 1)} \sim N_n \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \right)$$

Recall the properties of diagonal matrices:

$$\left| \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \right| = \left| \Sigma_{11} \right| \cdot \left| \Sigma_{22} \right|$$

$$\begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix}$$

Be cautious because the above properties do not hold in general, only if we have diagonal matrix (i.e off-diagonal elements are 0's.)

Then we can write the joint density of X_1 and X_2 in the following way,

$$f_{X_1, X_2}(\mathbf{x}_1, \mathbf{x}_2) = (2\pi)^{-\frac{n}{2}} \cdot |\Sigma_{11}|^{-\frac{1}{2}} \cdot |\Sigma_{22}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \begin{pmatrix} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\ (\mathbf{x}_2 - \boldsymbol{\mu}_2) \end{pmatrix}^T \cdot \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \begin{pmatrix} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\ (\mathbf{x}_2 - \boldsymbol{\mu}_2) \end{pmatrix} \right\}$$

Note that we use the previous result that X_1 and X_2 are normally distributed. Then multiplying out the term in the power of the exponential we have

$$f_{X_1, X_2}(\mathbf{x}_1, \mathbf{x}_2) = (2\pi)^{-\frac{n}{2}} \cdot |\Sigma_{11}|^{-\frac{1}{2}} \cdot |\Sigma_{22}|^{-\frac{1}{2}} \exp^{-\frac{1}{2} \{ (\mathbf{x}_1 - \boldsymbol{\mu}_1)^T \cdot \Sigma_{11}^{-1} \cdot (\mathbf{x}_1 - \boldsymbol{\mu}_1) + (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \cdot \Sigma_{22}^{-1} \cdot (\mathbf{x}_2 - \boldsymbol{\mu}_2) \}}$$

Also notice that $((\mathbf{x}_1 - \boldsymbol{\mu}_1))^T \cdot \Sigma_{11}^{-1} \cdot (\mathbf{x}_1 - \boldsymbol{\mu}_1)$ and $(\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \cdot \Sigma_{22}^{-1} \cdot (\mathbf{x}_2 - \boldsymbol{\mu}_2)$ are scalars. ($A \cdot \Sigma \cdot A^T = \text{matrix}$, $A^T \cdot \Sigma \cdot A = \text{scalar}$.) Now we can split it into a product of two terms which will turn out to be the densities of both vectors X_1 and X_2 , which completes the proof, so it follows

$$= (2\pi)^{-\frac{k}{2}} \cdot |\Sigma_{11}|^{-\frac{1}{2}} \cdot \exp^{-\frac{1}{2}} \{((\mathbf{x}_1 - \boldsymbol{\mu}_1))^T \cdot \Sigma_{11}^{-1} \cdot (\mathbf{x}_1 - \boldsymbol{\mu}_1)\} \cdot (2\pi)^{-\frac{n-k}{2}} \cdot |\Sigma_{22}|^{-\frac{1}{2}} \exp\{(\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \cdot \Sigma_{22}^{-1} \cdot (\mathbf{x}_2 - \boldsymbol{\mu}_2)\}$$

$$f_{X_1, X_2}(\mathbf{x}_1, \mathbf{x}_2) = f_{X_1}(\mathbf{x}_1) \cdot f_{X_2}(\mathbf{x}_2) \Rightarrow X_1 \text{ and } X_2 \text{ are independent. QED.}$$

Keep in mind this result because it has nice consequences that will be useful in most of the applications.

Exercise Show that $\begin{pmatrix} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\ (\mathbf{x}_2 - \boldsymbol{\mu}_2) \end{pmatrix}^T \cdot \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \cdot \begin{pmatrix} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\ (\mathbf{x}_2 - \boldsymbol{\mu}_2) \end{pmatrix} = \left\{ ((\mathbf{x}_1 - \boldsymbol{\mu}_1))^T \cdot \Sigma_{11}^{-1} \cdot (\mathbf{x}_1 - \boldsymbol{\mu}_1) + (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \cdot \Sigma_{22}^{-1} \cdot (\mathbf{x}_2 - \boldsymbol{\mu}_2) \right\}$.

Theorem Let $\mathbf{X} \sim N_n(0, \mathbf{I}_n)$ and $\mathbf{Y} = B_{(k \times n)} X_{(n \times 1)}$, $\mathbf{Z} = C_{((l \times n)} X_{(n \times 1)}$. Then Y and Z are independent if and only if $BC^T = 0$.

Proof We know that $\begin{pmatrix} \mathbf{Y} \\ \mathbf{Z} \end{pmatrix}_{((k+l) \times 1)} = \begin{pmatrix} \mathbf{B} \\ \mathbf{C} \end{pmatrix}_{((k+l) \times n)} \cdot \mathbf{X}$. Then $\begin{pmatrix} \mathbf{Y} \\ \mathbf{Z} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{B}\boldsymbol{\mu} \\ \mathbf{C}\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \mathbf{B} \\ \mathbf{C} \end{pmatrix} \cdot \mathbf{I}_n \cdot \begin{pmatrix} \mathbf{B}^T & \mathbf{C}^T \end{pmatrix}\right)$.

By multiplying the variance term we find

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{Z} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{B}\boldsymbol{\mu} \\ \mathbf{C}\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \mathbf{B} \cdot \mathbf{B}^T & \mathbf{B} \cdot \mathbf{C}^T \\ \mathbf{C} \cdot \mathbf{B}^T & \mathbf{C} \cdot \mathbf{C}^T \end{pmatrix}\right)$$

Notice that this theorem uses both properties of the multivariate normal distribution. $\begin{pmatrix} \mathbf{Y} \\ \mathbf{Z} \end{pmatrix}$ is normally distributed because of the first property and the second property tells us that Y and Z are independent if and only if $\text{cov}(\mathbf{Y}, \mathbf{Z}) = 0 \Leftrightarrow \mathbf{B} \cdot \mathbf{C}^T = 0 = \mathbf{C} \cdot \mathbf{B}^T$.

We might also be interested if the above theorem holds in case of quadratic transformation.

Let $\mathbf{X} \sim N_n(0, \mathbf{I}_n)$ and let C be a nonsingular, symmetric matrix such that $\mathbf{Y} = B_{(k \times n)} X_{(n \times 1)}$ and $\mathbf{Z} = X_{(1 \times n)}^T \cdot C_{(n \times n)} X_{(n \times 1)}$. Note that the first one is a linear and the second one is a quadratic transformation. (Intuitively, we can think in terms of scalars $y=b \cdot x$ vs. $z=c \cdot x^2$). Then we have

$$\mathbf{Z} = X^T \cdot C \cdot X = X^T \cdot C \cdot C^{-1} \cdot C \cdot X = (C \cdot X)^T \cdot C^{-1} \cdot C \cdot X = g(C \cdot X)$$

the last equality tells us that the quadratic transformation is a linear function. Then if $B^T \cdot C = B \cdot C = 0 \Rightarrow Y$ and $C \cdot X$ are independent $\Rightarrow Y$ and

$g(C.X)$ are independent $\Rightarrow Y$ and Z are independent. The following theorem formalizes this:

Theorem Let $\mathbf{X} \sim N_n(0, \mathbf{I}_n)$ and let C be symmetric matrix such that $\mathbf{Y} = B_{(k \times n)} \mathbf{X}_{(n \times 1)}$ and $\mathbf{Z} = X_{(1 \times n)}^T \cdot C_{(n \times n)} \mathbf{X}_{(n \times 1)}$. Then if $B.C = 0$, then Y and Z are independent.

Corollary If (X_1, X_2, \dots, X_n) is a random sample from $N(\mu, \sigma^2)$, then the sample mean \bar{X} and sample variation S^2 are independent. Note that $\bar{X} = B.X$ and $S^2 = X^T.C.X$.

Theorem Let $\mathbf{X} \sim N_n(0, \mathbf{I}_n)$ and let B, C be symmetric matrices such that $Y = \mathbf{X}_{(1 \times n)}^T \cdot \mathbf{B}_{(n \times n)} \mathbf{X}_{(n \times 1)}$ and $Z = \mathbf{X}_{(1 \times n)}^T \cdot \mathbf{C}_{(n \times n)} \mathbf{X}_{(n \times 1)}$. If $B.C = 0$, then Y and Z are independent.

Example Let $(X_1, X_2, \dots, X_n) \sim^{i.i.d} N(0, 1)$, $\mathbf{B} = \begin{pmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, $\mathbf{C} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-k} \end{pmatrix}$. Then $\mathbf{B.C} = \mathbf{0}_{(n \times n)}$.

$$Y = \mathbf{X}_{(1 \times n)}^T \cdot \mathbf{B}_{(n \times n)} \mathbf{X}_{(n \times 1)} = (X_1 \quad \dots \quad X_n) \begin{pmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \sum_{i=1}^k X_i^2$$

$$Z = \mathbf{X}_{(1 \times n)}^T \cdot \mathbf{C}_{(n \times n)} \mathbf{X}_{(n \times 1)} = (X_1 \quad \dots \quad X_n) \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-k} \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \sum_{i=k+1}^n X_i^2$$

Notice from the summation indices that Y and Z are independent. Furthermore, in this special case where we had $(X_1, X_2, \dots, X_n) \sim^{i.i.d} N(0, 1)$ assumption, we have that

$$Y = \mathbf{X}^T \cdot \mathbf{B} \cdot \mathbf{X} = \mathbf{X}^T \cdot \begin{pmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \cdot \mathbf{X}$$

$$Y = \mathbf{X}^T \cdot \mathbf{M} \cdot \mathbf{X} \sim \mathcal{X}_k^2$$

We also want to know under which conditions of \mathbf{M} , the random variable Y has a \mathcal{X}^2 -distribution. The following theorem answers this question.

Theorem Let $\mathbf{X} \sim N_n(0, \mathbf{I}_n)$ and let \mathbf{M} be a symmetric *idempotent* matrix. (i.e. $\mathbf{M}^2 = \mathbf{M}$, also called *projection*, e.g identity matrix: $\mathbf{I}_n \cdot \mathbf{I}_n = \mathbf{I}_n$). Let K be the rank of \mathbf{M} . Then $\mathbf{X}^T \cdot \mathbf{M} \cdot \mathbf{X} \sim \mathcal{X}_K^2$.

Proof Omitted.

Assume we have $\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}_{(n \times 1)} \sim N_n \left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sum_{XX} & \sum_{XY} \\ \sum_{YX} & \sum_{YY} \end{pmatrix} \right)$. What is the conditional distribution of Y given $X = x$?

The following theorem provides an answer to this question.

Theorem $Y | X = x \sim N\left(\mu_Y + \sum_{YX} \cdot \sum_{XX}^{-1}(x - \mu_X), \sum_{YY} - \sum_{YX} \sum_{XX}^{-1} \sum_{XY}\right)$

Proof Only an outline of the proof will be sketched. Let $U = Y - \mu_Y - \sum_{YX} \sum_{XX}^{-1}(X - \mu_X)$. Then we have the following observations

$$\begin{aligned} E(U) &= 0 \\ cov(U, X) &= 0 \Rightarrow U \text{ and } X \text{ are independent} \\ E(U | X) &= E(U) \\ var(U | X) &= var(U) \\ \begin{pmatrix} U \\ X \end{pmatrix} &\sim N \\ U | X &\sim N \rightarrow Y | X \sim N \end{aligned}$$

Lecture 11 / Week 7

Convergence

OUTLINE

- 1.) In Probability
- 2.) Almost Surely

Convergence in Probability

Definition Suppose we have a sequence $(X_n)_{n=1}^{\infty}$ of random variables. Let X be a random variable. We say that \mathbf{X}_n **converges to X in probability** if

$$\lim_{n \rightarrow \infty} P(|X_n - X| < \varepsilon) = 1, \forall \varepsilon > 0.$$

For instance take an event $\{\omega : |X_n - X| < \varepsilon\}$. This event says that X_n is near X . Here the distance is expressed as absolute value. Then for n large enough

$$P(|X_n - X| < \varepsilon) \approx 1$$

Example Toss a coin infinitely many times. Call $F_n :=$ the frequency of heads in the first n trials. It can be proved that F_n converges in probability to $\frac{1}{2}$. (It can also be shown with a computer simulation.) Fix an $\varepsilon > 0$. (it can be as small as we want) and then take the event

$$\{|F_n - \frac{1}{2}| < \varepsilon\}$$

what the above definition says that

$$P\{\frac{1}{2} - \varepsilon \leq F_n \leq \frac{1}{2} + \varepsilon\} \geq 0.9999\dots$$

if n is large enough.

The above definition can be expressed equivalently in terms of complement event, i.e.

$$\begin{aligned} \lim_{n \rightarrow \infty} P(|X_n - X| < \varepsilon) &= 1, \forall \varepsilon > 0 \\ (|X_n - X| < \varepsilon)^c &= (|X_n - X| \geq \varepsilon) \end{aligned}$$

Then the equivalent definition will be

$$\begin{aligned} \lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) &= 0, \forall \varepsilon > 0 \\ &or \\ \lim_{n \rightarrow \infty} P(|X_n - X| \geq \varepsilon) &= 0, \forall \varepsilon > 0 \end{aligned}$$

There are different notation for expressing this concept such as

$$\begin{aligned} X_n &\xrightarrow{P} X \\ \text{plim}_{n \rightarrow \infty} X_n &= X \end{aligned}$$

Weak Laws of Large Numbers

Suppose we have X_1, X_2, \dots . Then the sample mean $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ follows

$$\bar{X}_n \xrightarrow{P} E(X)$$

under certain assumptions. The following theorem tells us under which assumptions it holds.

Theorem (WLLN for uncorrelated r.v) Let X_1, X_2, \dots be a sequence of random variables such that $E(X^2) < \infty$ (i.e. second moment is *finite*) for every n , then $E(X_n) = \mu, \text{var}(X_n) = \sigma^2$, where μ and σ^2 do not depend on n . Moreover let $\text{cov}(X_n, X_m) = 0 \quad \forall n, m$. Then

$$\bar{X}_n \xrightarrow{P} \mu \quad n \rightarrow \infty$$

Note that this theorem makes the following assumptions: X_n are uncorrelated and $E(X_n), \text{var}(X_n)$ are constant w.r.t n . (independent of n)

Proof The theorem says

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0, \forall \varepsilon > 0$$

$$E(\bar{X}_n) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} n\mu = \mu$$

We will use Chebishev Inequality to prove the result, i.e.

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - E(\bar{X}_n)| > \varepsilon) \leq^{Cheb.} \frac{var(\bar{X}_n)}{\varepsilon^2} = \frac{\frac{\sigma^2}{n}}{\varepsilon^2} \rightarrow 0, \forall \varepsilon > 0$$

$$var(\bar{X}_n) = var\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \stackrel{uncorr.}{=} \frac{1}{n^2} \sum_{i=1}^n var(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}$$

Example Let $F_n :=$ the frequency of heads, $F_n := \frac{\text{the number of heads in the first } n \text{ trials}}{n}$,
 $F_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \bar{X}_n$.

$$X_i \left\{ \begin{array}{l} 1 \text{ head at toss } i \\ 0 \text{ tail at toss } i \end{array} \right\}$$

Note that X_i are independent and independence implies no correlation.

$$P(X_i = 1) = \frac{1}{2}, P(X_i = 0) = \frac{1}{2}$$

$$E(X_i) = 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{2}$$

$$var(X_i) = \frac{1}{2} \left(1 - \frac{1}{2}\right)^2 + \frac{1}{2} \left(0 - \frac{1}{2}\right)^2 = \frac{1}{4}$$

$$E(X_n) = \mu = \frac{1}{2}$$

$$var(X_n) = \sigma^2 = \frac{1}{4}$$

we can apply the theorem, hence

$$\bar{X}_n \xrightarrow{P} E(X) = \frac{1}{2}$$

Suppose we have X_1, X_2, \dots i.i.d random variables. $E(X_n^2) < \infty$. Then we can apply the previous theorem $\Rightarrow \bar{X}_n \xrightarrow{P} \mu$. ($n \rightarrow \infty$), since X_n satisfy the hypothesis of the theorem. The independence implies no correlation and since X_n are identically distributed $\Rightarrow E(X_n), var(X_n)$ do not depend on n . In the following theorem we will strive for a stronger result, in the sense that we will not require a finite second moment as an assumption.

Theorem (WLLN for i.i.d r.v) Let X_n be a sequence of *independent and identically distributed* random variables such that $E(|X_n|) < \infty$. (*integrable*). This theorem guarantees convergence even though the random variable has infinite variance, i.e.

$$\text{Let } \mu = E(X_n)$$

$$\text{Then } \bar{X}_n \xrightarrow{P} \mu \quad (n \rightarrow \infty)$$

Suppose $\bar{X}_n \rightarrow E(X_1)$. $E(X)$ is the limit of the mean of X_n where X_n is a sequence of i.i.d random variables distributed as X . We have shown that the above result holds for arithmetic mean.

Example Dice score $E(X)=3.5$

We wonder whether we can make any conclusion about the convergence of geometric mean. This is an important consideration since in economics and finance one often needs to use geometric averages. (interest rates, compounded returns in the long term.) Suppose we have a sequence of nonnegative random variables $X_1, X_2, \dots, X_n \geq 0$ and we consider the geometric mean

$$\begin{aligned} \text{Geometric Mean} & : (X_1 \cdot X_2 \cdot \dots \cdot X_n)^{\frac{1}{n}} \rightarrow ? \\ (X_1 \cdot X_2 \cdot \dots \cdot X_n)^{\frac{1}{n}} & = (e^{\log X_1} \cdot e^{\log X_2} \cdot \dots \cdot e^{\log X_n})^{\frac{1}{n}} \end{aligned}$$

Note that we can use this transformation since we have nonnegative random variables.

$$(e^{\log X_1} \cdot e^{\log X_2} \cdot \dots \cdot e^{\log X_n})^{\frac{1}{n}} = \left(e^{\sum_{i=1}^n \log X_i} \right)^{\frac{1}{n}} = e^{\frac{1}{n} \sum_{i=1}^n \log X_i}$$

Suppose X_i i.i.d $\Leftrightarrow \log X_i$ i.i.d

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \log X_i & \rightarrow^P E(\log X_1). \\ e^{\frac{1}{n} \sum_{i=1}^n \log X_i} & \rightarrow^P e^{E(\log X_1)}. \end{aligned}$$

So we have shown that the result of the theorem also holds in case of geometric mean.

Theorem (Slutsky) Let X_n be a sequence of random variables converging in probability to a nonrandom constant c . Let ψ be a *continuous* function, then

$$\begin{aligned} X_n & \rightarrow^P c \\ \psi(X_n) & \rightarrow^P \psi(c) \quad n \rightarrow \infty \end{aligned}$$

Example $\psi(x) = e^x$. Since the *exponential function* is continuous, then

$$\psi\left(\frac{1}{n} \sum \log X_i\right) \rightarrow^P \psi(E(\log X_1))$$

Example $\psi(x) = \sqrt{x}$, $x > 0$, since square root is continuous, then

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n X_i^2 & \rightarrow^P E(X_1^2) = \mu \\ \sqrt{\frac{1}{n} (X_1^2 + X_1^2 + \dots + X_n^2)} & \rightarrow^P \sqrt{\mu} = \sqrt{E(X_1^2)} \end{aligned}$$

Almost Sure Convergence

Definition Suppose we have a sequence $(X_n)_{n=1}^{\infty}$ of random variables. Let X be a random variable. We say that \mathbf{X}_n **converges almost surely to X** if

$$P(\lim_{n \rightarrow \infty} X_n = X) = 1$$

Since

$$P(\lim_{n \rightarrow \infty} \sup_{m \geq n} |X_m - X| = 0) = 1 \Leftrightarrow P(\lim_{n \rightarrow \infty} \sup_{m \geq n} |X_m - X| < \varepsilon) = 1, \forall \varepsilon > 0.$$

In the following we will show the relationship between the convergence in probability and a.s convergence. The following holds

$$\begin{aligned} \text{a.s convergence} &\Rightarrow \text{p.convergence} \\ P(\lim_{n \rightarrow \infty} \sup_{m \geq n} |X_m - X| < \varepsilon) = 1, \forall \varepsilon > 0. &\Rightarrow \lim_{n \rightarrow \infty} P(|X_n - X| < \varepsilon) = 1, \forall \varepsilon > 0. \end{aligned}$$

To see this, consider

$$\sup_{m \geq n} |X_m - X| \geq |X_n - X|$$

then the event

$$\begin{aligned} \sup_{m \geq n} |X_m - X| < \varepsilon &\subset |X_n - X| < \varepsilon \\ P(|X_n - X| < \varepsilon) &\geq P(\sup_{m \geq n} |X_m - X| < \varepsilon) \end{aligned}$$

then we can see that the fact that the supremum is smaller than ε implies that all the distances are smaller than ε , therefore a.s convergence \Rightarrow p.convergence, but the reverse is not true. Note that the strong law of large numbers (**SLLN**) is based on a.s convergence, whereas weak law of large (**WLLN**) numbers is based on convergence in probability.

Theorem (Kolmogorov SLLN) Let $\{X_n\}$ be a sequence of independent and identically distributed random variables with $E(|X_n|) < \infty$. Let $E(X_1) = \mu$ and $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Then

$$\bar{X}_n \rightarrow \mu \quad \mathbf{a.s}$$

Proof Complicated, hence omitted.

Example Let $F_n :=$ the frequency of heads, then

$$\begin{aligned} F_n &\rightarrow \frac{1}{2} \quad \mathbf{a.s} \\ F_n &= \frac{1}{n} \sum_{i=1}^n X_i, \quad X_i \text{ i.i.d} \end{aligned}$$

In the following theorem we will see that Slutsky theorem carries over to a.s convergence, it is even stronger since $X_n \rightarrow c$ is not required.

Theorem (Slutsky for a.s convergence) Let $X_n \rightarrow X$ a.s and let ψ be a continuous function. Then

$$\psi(X_n) \rightarrow \psi(X) \quad a.s$$

Proof Take a set $M = \{\omega : X_n(\omega) \rightarrow X(\omega)\}$. By a.s. hypothesis $P(M) = 1$.

$$\begin{aligned} \omega &\in M : X_n(\omega) \rightarrow X(\omega) \quad n \rightarrow \infty \\ \psi - cont. & : \psi(X_n(\omega)) \rightarrow \psi(X(\omega)) \quad n \rightarrow \infty \\ \omega &\in M \text{ then } \psi(X_n(\omega)) \rightarrow \psi(X(\omega)) \quad n \rightarrow \infty \\ \psi(X_n(\omega)) &\rightarrow \psi(X(\omega)) \quad (n \rightarrow \infty) \text{ a.s} \end{aligned}$$

Generalizations

Within the WLLN we saw two cases:

1. **X'_n s are uncorrelated:** This case can be generalized to *weakly correlated* X'_n s. \Rightarrow **weakly stationary processes**
2. **X'_n s are i.i.d:** This case can also be generalized to *weakly dependent* X'_n s. \Rightarrow **strictly stationary processes.**

Definition A **time series process** is a sequence of random variables $X_t (t = 1, 2, 3, \dots)$ where the random variables are indexed by time.

Definition A sequence of random variables X_t is called **strictly stationary** if for every m , the probability distribution of $(X_t, X_{t+1}, \dots, X_{t+m})$ does not depend on t .

$$\begin{aligned} m = 1 & : \text{prob. dist } X_1 = p.d \ X_2 = p.d \ X_3 \\ m = 2 & : \text{prob. dist } (X_1, X_2) = p.d \text{ of } (X_2, X_3) = p.d \text{ of } (X_3, X_4) \end{aligned}$$

Example X_n are i.i.d. $\Rightarrow X_n$ is strictly stationary

$$f_{X_t, X_{t+1}, \dots, X_{t+m}}(x_0, \dots, x_m) = \prod_{i=0}^m f(x_i)$$

so the the distribution does not depend on t . What matters is the time lag. (m)

Example Suppose Z_n are i.i.d and $X_n = Z_n + Z_{n+1}$

X_n is strictly stationary $(X_t, X_{t+1}, \dots, X_{t+m}) = (Z_t + Z_{t+1}, Z_{t+1} + Z_{t+2}, \dots, Z_{t+n} + Z_{t+n+1})$

If $\{X_n\}$ is strictly stationary, then X_1, X_2, \dots are identically distributed. The probability distribution of X_t does not depend on t .

$$\begin{aligned} \text{i.i.d} &\Rightarrow \text{strict stationarity} \\ \text{strict stationarity} &\Rightarrow \text{identical distribution} \end{aligned}$$

So note that, strict stationarity does not imply independence.

Definition A sequence X_t of random variables is **weakly stationary** if $E(X_t^2) < \infty \forall t$ and $E(X_t) = \mu$, $cov(X_t, X_{t+m}) = \gamma(m)$ ($m = 0, 1, \dots$). (Note that $m=0 \Rightarrow \text{variance}$)

when μ and $\gamma(m)$ do not depend on t .

Example Suppose we have weakly stationary process, then

$$\begin{aligned} m = 1 &: E(X_1) = E(X_2) = E(X_3) \dots \\ m = 1 &: var(X_1) = Var(X_2) = Var(X_3) \dots \\ m = 1 &: cov(X_1, X_2) = cov(X_2, X_3) \dots = cov(X_{17}, X_{18}) \\ m = 0 &: cov(X_t, X_t) = var(X_t) \\ m = 2 &: cov(X_1, X_3) = cov(X_{14}, X_{16}) \end{aligned}$$

Lecture 12 / Week 7

Convergence in Probability

The first part of the lecture on the "Extension of the WLLN to Stationary Time Series" can be found in the handout given by Prof. Fortini.

Sofar we defined the theorems on convergence in probability for random variables, but these theorems can be extended to random vectors. Then

X, X_n : random vectors

convergence in probability can be expressed now

$$P(\|X_n - X\| > \varepsilon) \rightarrow 0 \quad n \rightarrow \infty$$

where we used Euclidean norm ($\|x\| = \sqrt{x^T \cdot x} = \sqrt{\sum x_i^2}$) instead of the absolute value. Slutsky Theorem and Law of Large Numbers carry over to random vectors.

Convergence in Distribution

We will use a completely different notation than we used before in case of convergence in probability or a.s.

Suppose X_n are random vectors and X is a random vector. We define

$$\begin{aligned} F_n & : = \text{the distribution function of } X_n \\ F & : = \text{the distribution function of } X \\ F(x) & = P(X \leq x) \end{aligned}$$

Definition X_n **converges in distribution** to X if $F_n(x) \rightarrow F(x)$ $n \rightarrow \infty$ for every x where F is continuous. Note that the distribution of X_n converges to the distribution of X , not X_n to X . in other words, suppose X_n, X are random variables then

$$P(a < X_n \leq b) = F_n(b) - F_n(a) \rightarrow F(b) - F(a) = P(a < X \leq b)$$

if F is continuous in a and b .

$$\begin{aligned} P(a < X_n \leq b) & \rightarrow P(a < X \leq b) \\ P(X_n \in A) & \rightarrow P(X \in A) \end{aligned}$$

for most of the Borel sets A . (not every, because we have the condition that b is a continuity point, recall that the distribution function is right-continuous.)

If n is large enough

$$P(X_n \in A) \approx P(X \in A)$$

for most of the Borel sets.

Example Assume that X_n have exponential distribution(n).

$$\begin{aligned} f_n(x) & = n \cdot e^{-nx} \cdot 1_{[0, \infty)}(x) \\ F_n(x) & = \int_{-\infty}^x f_n(s) ds = (1 - e^{-nx}) \cdot 1_{[0, \infty)}(x) \\ X_n & \rightarrow^d 0, \quad n \rightarrow \infty \\ n & \rightarrow \infty \quad \lim_{n \rightarrow \infty} (1 - e^{-nx}) \cdot 1_{[0, \infty)}(x) = 1_{(0, \infty)}(x) \end{aligned}$$

Insert here Figure 28

Note that the limit distribution is a constant function ($1_{(0, \infty)}(x)$) which is not continuous at $x=0$, but we are not concerned.

Insert here Figure 29

$$F_n(x)_{n \rightarrow \infty} \rightarrow 1_{[0, \infty)}(x) \text{ where } F \text{ is continuous}$$

We will use the following notation

$$\begin{aligned} X_n &\rightarrow {}^d X \\ X_n &\rightarrow {}^d X \sim N(0, 1) \\ &\text{or directly} \\ X_n &\rightarrow {}^d N(0, 1) \end{aligned}$$

Note that

$$n \rightarrow \infty \quad \lim_{n \rightarrow \infty} (1 - e^{-nx}) \cdot 1_{[0, \infty)}(x) = 1_{(0, \infty)}(x)$$

is not a distribution function since it is left-continuous but not right-continuous, i.e. $F(0^-) = F(0)$, that's why we define as

$$F_n(x)_{n \rightarrow \infty} \rightarrow 1_{[0, \infty)}(x) \text{ where } F \text{ is continuous}$$

Insert here Figure 30

The Relation between Convergence in Distribution and Probability

Convergence in Probability \Rightarrow Convergence in Distribution

$$X_n \rightarrow^P X \Rightarrow X_n \rightarrow^d X$$

The converse is not true in general. It is true if X is *constant*. Therefore

$$X_n \rightarrow^d c \Rightarrow X_n \rightarrow^P c$$

In other words when the limit is constant

$$\rightarrow^P = \rightarrow^d$$

Continuous Mapping Theorem If $X_n \rightarrow^d X$ and ϕ is a *continuous* function, then

$$\phi(X_n) \rightarrow^d \phi(X)$$

Example $X_n \rightarrow^d N(0, 1)$

$$X_n^2 = \phi(X_n) \rightarrow^d \phi(X) \text{ with } X \sim N(0, 1)$$

since $\phi(x) = x^2$ is a continuous function, the assumption of CMT holds.

$$X_n^2 \rightarrow X_1^2$$

Example Let $X_n \rightarrow^d N_k(0, \mathbf{I}_k)$ (i.e. $X \sim (0, \mathbf{I}_k)$) and M an idempotent, symmetric matrix. Consider the following transformation.

$$X_n^T . M . X_n = \phi(X_n)$$

note that ϕ is a continuous function.

$$X_n^T . M . X_n = \phi(X_n) \rightarrow^d \phi(X) = X^T . M . X \sim \mathcal{X}_r^2$$

where $r = \text{rank}(M)$, this is an example of an asymptotic result often used in statistical inference.

Suppose we have

$$\begin{aligned} X_n &\rightarrow^d X \\ Y_n &\rightarrow^d Y \end{aligned}$$

then we have a continuous function $\phi(x, y)$. Is it true that

$$\phi(X_n, Y_n) \rightarrow^d \phi(X, Y)$$

the answer is no in general. But, the following theorem tells us under which condition it holds.

Theorem Let

$$\begin{aligned} X_n &\rightarrow^d X \\ Y_n &\rightarrow^d c \end{aligned}$$

and let $\phi(x, y)$ be a continuous function. Then $\phi(X_n, Y_n) \rightarrow^d \phi(X, c)$.

Example Let $T_n \sim \text{Student} - t(n)$. Then

$$T_n \rightarrow^d N(0, 1)$$

this result is a consequence of the previous theorem. (One can check that looking at the tables at the end of any statistics book, for high n the distributions are very similar.) Consider

$$\begin{aligned} T_n &= \frac{X_n}{\sqrt{\frac{Y_n}{n}}} \\ X_n &\sim N(0, 1) \\ Y_n &\sim \mathcal{X}^2(n) \end{aligned}$$

Exercise Let $X_n \rightarrow^d N(0, 1)$ and show that $\frac{Y_n}{n} \rightarrow^d 1$. Hint: use the fact that $Y_n = Z_1^2 + Z_2^2 + \dots$, together with the law of large numbers.

So, we can express it i.t.o the previous theorem, i.e

$$\phi(x, y) = \frac{x}{\sqrt{y}}$$

note that ϕ is a continuous function. Then the theorem says

$$T_n = \phi\left(X_n, \frac{Y_n}{n}\right) \rightarrow^d \phi(X, 1) = \frac{X}{\sqrt{1}} = X \sim N(0, 1)$$

note that we used the result in the previous exercise.

Central Limit Theorem

Theorem (Lévy) Let X_n be a sequence of independent and identically distributed random variables with $E(X_n) = \mu$ and $\text{var}(X_n) = \sigma^2 < \infty$. Then

$$\sqrt{n}(\bar{X}_n - \mu) \rightarrow^d N(0, \sigma^2)$$

This is one of the most important, hence mostly applied theorems in statistics (See **CLT simulation**.)

We would like to generalize this result in case of an continuous function ϕ . Does $\sqrt{n}(\phi(\bar{X}_n) - \phi(\mu))$ converge in distribution?

Suppose ϕ is *continuously differentiable* function, s.t. $\phi : \mathbb{R} \rightarrow \mathbb{R}$. By properties of derivative

$$\sqrt{n}(\phi(\bar{X}_n) - \phi(\mu)) = \sqrt{n}(\phi'(\mu + \lambda(\bar{X}_n - \mu))(\bar{X}_n - \mu))$$

where we used *Taylor expansion*

$$\phi(x) = \phi(x_0) + \phi'(x_0 + \lambda(x - x_0))(x - x_0)$$

Then

$$\begin{aligned} \sqrt{n}(\phi(\bar{X}_n) - \phi(\mu)) &= (\phi'(\mu + \lambda(\bar{X}_n - \mu)) \cdot \sqrt{n}(\bar{X}_n - \mu)) \\ n &\rightarrow \infty \quad \mathbf{CLT}: \sqrt{n}(\bar{X}_n - \mu) \rightarrow^d N(0, \sigma^2) \\ n &\rightarrow \infty \quad \mathbf{WLLN}: \bar{X}_n \xrightarrow{P} \mu \\ \sqrt{n}(\phi(\bar{X}_n) - \phi(\mu)) &= (\phi'(\mu + \lambda(\bar{X}_n - \mu)) \cdot \sqrt{n}(\bar{X}_n - \mu)) \rightarrow^d \phi'(\mu) \cdot Z \sim N(0, \sigma^2 \cdot \phi'(\mu)^2) \end{aligned}$$

where $Z \sim N(0, \sigma^2)$.

This result on mean can be generalized also for other moments, using **CLT+Delta Method** as long as $\phi(\cdot)$ is a continuous function.

Theorem Let X_n be a sequence of i.i.d random variables of size k with $E(X_n) = \mu$, $var(X_n) = \Sigma$. (i.e. the second moment is finite.) Then

$$\sqrt{n}(\bar{X}_n - \mu) \rightarrow^d N_k(0, \Sigma)$$

Suppose we have $\phi : \mathbb{R}^k \rightarrow \mathbb{R}^m$ continuously differentiable. Then

$$\sqrt{n}(\phi(\bar{X}_n) - \phi(\mu)) \rightarrow^d N_m(0, \Delta\phi(\mu) \Sigma \Delta\phi(\mu)^T)$$

where $\Delta\phi(\mu)$ is computed as follows

$$\Delta\phi(x) = \begin{bmatrix} \frac{\partial\phi_1}{\partial x_1} & \cdot & \cdot & \frac{\partial\phi_1}{\partial x_k} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \frac{\partial\phi_m}{\partial x_1} & \cdot & \cdot & \frac{\partial\phi_m}{\partial x_k} \end{bmatrix}$$