

TEORÍA DE LA PROBABILIDAD

Iker Caballero Bragagnini

Tabla de contenido

LOS EVENTOS Y PROBABILIDADES.....	2
LAS VARIABLES ALEATORIAS DISCRETAS	12
LAS FUNCIONES DE DISTRIBUCIÓN Y DE DENSIDAD	21
LAS DISTRIBUCIONES DE PROBABILIDAD COMUNES	34
LAS FAMILIAS DE DISTRIBUCIONES	43
LAS DISTRIBUCIONES MULTIVARIADAS DISCRETAS E INDEPENDENCIA	43
LAS DISTRIBUCIONES MULTIVARIADAS GENERALES E INDEPENDENCIA.....	49
LOS ESTADÍSTICOS DE ORDEN.....	63
LAS FUNCIONES GENERADORAS DE PROBABILIDAD	66
LOS MOMENTOS Y FUNCIONES GENERADORAS DE MOMENTOS	74
LA FUNCIÓN CARACTERÍSTICA	81
LA DISTRIBUCIÓN NORMAL MULTIVARIANTE	90
LAS DISTRIBUCIONES NORMALES MIXTAS Y DISTRIBUCIONES GENERALIZADAS HIPERBÓLICAS	102
LAS DISTRIBUCIONES ESFÉRICAS Y ELÍPTICAS	112
LA ENTROPÍA	112
LA CONVERGENCIA	112
LOS PROCESOS ESTOCÁSTICOS.....	141
LOS PROCESOS DE RAMIFICACIÓN	145
LOS CAMINOS ALEATORIOS	154
LOS PROCESOS DE POISSON.....	171
LOS PROCESOS DE RENOVACIÓN	182
LAS CADENAS DE MARKOV	182
LAS CADENAS DE MARKOV EN TIEMPO CONTINUO	220
LAS MARTINGALAS	220

Los eventos y probabilidades

- Muchas acciones tienen resultados muy impredecibles, y la teoría de la probabilidad se enfoca en esas acciones y en sus consecuencias, usando las nociones de experimentos, eventos, resultados y probabilidades para construir el espacio de probabilidad
 - La teoría matemática comienza con la idea de un experimento o intento, que se define como una acción cuya consecuencia no es predeterminada
 - Este experimento se reformula en un objeto matemático llamado espacio de probabilidad
 - Este espacio de probabilidad contiene tres cosas: el conjunto de todos los posibles resultados del experimento, una lista de todos los eventos que posiblemente pueden ocurrir como consecuencias de este experimento, y una evaluación de la posibilidad de que ocurran estos eventos
 - (i) the set $\{1, 2, 3, 4, 5, 6\}$ of possible outcomes,
 - (ii) a list of events such as
 - the result is 3,
 - the result is at least 4,
 - the result is a prime number,
 - (iii) each number 1, 2, 3, 4, 5, 6 is equally likely to be the result of the throw.
 - Dado cualquier experimento que involucre azar, hay un espacio de probabilidad correspondiente, y el estudio de estos espacios se denomina teoría de la probabilidad
 - Usando \mathcal{E} para denotar un experimento particular cuyo resultado no ha sido completamente predeterminado, el conjunto de todos los posibles resultados de \mathcal{E} se denomina espacio muestral y se denota por Ω
 - Un elemento de Ω se denomina evento elemental y se denota por ω

$$\Omega \equiv \{\omega_1, \omega_2, \omega_3, \dots\}$$

- Es posible realizar muchas preguntas sobre el resultado del experimento, las cuales se pueden reescribir en términos de subconjuntos de Ω
- La colección \mathcal{F} de subconjuntos del espacio muestral Ω se denomina espacio de eventos si $\mathcal{F} \neq \emptyset$, si $A \in \mathcal{F}$ implica que $A^c \in \mathcal{F}$ y si $A_1, A_2, \dots \in \mathcal{F}$ implica que $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

- La lista de eventos será, por tanto, una lista de subconjuntos $\mathcal{F} \equiv \{A_i: i \in I\}$ de Ω llamado espacio de eventos y que se denota por \mathcal{F} , en donde cada $A \in \mathcal{F}$ es un evento
- Un espacio de eventos \mathcal{F} tiene que contener el conjunto vacío \emptyset y el conjunto entero o espacio muestral Ω

$$A, A^c \in \mathcal{F} \Rightarrow A_1 \equiv A \text{ \& } A_i \equiv A^c \text{ for } i \geq 2 \Rightarrow \bigcup_{i=1}^{\infty} A_i = \Omega \in \mathcal{F}$$

$$\Omega \in \mathcal{F} \Rightarrow \Omega^c = \emptyset \in \mathcal{F}$$

- Un espacio de eventos está cerrado bajo la operación de uniones finitas (la unión de dos eventos en \mathcal{F} pertenece también a \mathcal{F})

$$A_1, A_2, \dots, A_m \in \mathcal{F} \text{ \& } A_i \equiv \emptyset \text{ for } i > m$$

$$\Rightarrow \bigcup_{i=1}^m A_i = \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$$

- Un espacio de eventos también está cerrado bajo la operación de intersecciones finitas o numberables (la intersección de dos eventos en \mathcal{F} pertenece también a \mathcal{F})

$$A_i \in \mathcal{F} \text{ for } i = 1, 2, \dots, m \text{ \& } A_j \equiv A_i^c \text{ for } j > m$$

$$\bigcap_{i=1}^m A_i \Rightarrow \bigcup_{j=m}^{\infty} A_j = \bigcup_{i=1}^m A_i^c \in \mathcal{F}$$

- Si Ω es un conjunto finito, entonces \mathcal{F} contiene un número par de subconjuntos de Ω

$$\Omega \equiv \{\omega_1, \omega_2, \dots, \omega_n\} \Rightarrow \{\omega_i\} \in \mathcal{F} \text{ for } i = 1, 2, \dots, n$$

$$\Rightarrow \{\omega_i\}^c \in \mathcal{F} \text{ for } i = 1, 2, \dots, n$$

$$\Rightarrow |\mathcal{F}| = 2n + 2 = 2k \text{ where } k, n \in \mathbb{N} \text{ such that } k = n + 1$$

- Un mapeado $P: \mathcal{F} \rightarrow \mathbb{R}$ se denomina medida de probabilidad en (Ω, \mathcal{F}) si $P(A) \in [0, 1]$ para $A \in \mathcal{F}$, si $P(\Omega) = 1$ y $P(\emptyset) = 0$, y si A_1, A_2, \dots , siendo eventos disjuntos ($A_i \cap A_j = \emptyset$ para toda $i \neq j$) de \mathcal{F} , implica que $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

- La medida de probabilidad P en (Ω, \mathcal{F}) solo está definida para los subconjuntos de Ω que están en \mathcal{F} , no para otros subconjuntos
- Que la probabilidad $P(\emptyset) = 0$ es consecuencia de que $P(\Omega) = 1$

$$A_1 \equiv \Omega \text{ \& } A_i \equiv \emptyset \text{ for } i \geq 2$$

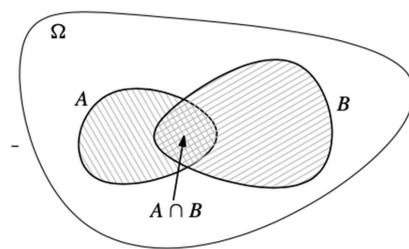
$$\Rightarrow P(\Omega) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = P(\Omega) + \sum_{i=1}^{\infty} P(\emptyset)$$

- La última condición expresa que la medida de probabilidad P es numerablemente aditiva, por lo que también es finitamente aditiva cuando A_i para $i = 1, 2, \dots$ son eventos disjuntos

$$A_1, A_2, \dots, A_m \in \mathcal{F} \text{ \& } A_j \equiv \emptyset \text{ for } j > m$$

$$\Rightarrow P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^m P(A_i) + \sum_{j=m+1}^{\infty} P(A_j) = \sum_{i=1}^{\infty} P(A_i)$$

- Siendo $\Omega = \{\omega_1, \omega_2, \dots, \omega_N\}$ un conjunto finito de exactamente N puntos y \mathcal{F} el conjunto de potencia de Ω , la función P definida por $P(A) \equiv 1/N|A|$ para $A \in \mathcal{F}$ es una medida de probabilidad de (Ω, \mathcal{F})
- Se suele utilizar un diagrama de Venn al trabajar con probabilidades con tal de hacerlas más intuitivas



- Un espacio de probabilidad es un triple (Ω, \mathcal{F}, P) de objetos tal que $\Omega \neq \emptyset$, \mathcal{F} es un espacio de eventos de subconjuntos de Ω y P es una medida de probabilidad en (Ω, \mathcal{F})
- Siendo (Ω, \mathcal{F}, P) un espacio de probabilidad, si $A, B \in \mathcal{F}$, entonces $A - B \in \mathcal{F}$

$$(A - B)^c = A^c \cup B \Rightarrow A^c \in \mathcal{F} \text{ \& } B \in \mathcal{F} \Rightarrow$$

$$\Rightarrow A^c \cup B = (A - B)^c \in \mathcal{F} \Rightarrow A - B \in \mathcal{F}$$

- Siendo (Ω, \mathcal{F}, P) un espacio de probabilidad, si $A_1, A_2, \dots \in \mathcal{F}$, entonces $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$

$$A_1, A_2, \dots \in \mathcal{F} \Rightarrow A_1^c, A_2^c, \dots \in \mathcal{F}$$

$$\Rightarrow \left(\bigcap_{i=1}^{\infty} A_i \right)^c = \bigcup_{i=1}^{\infty} A_i^c \in \mathcal{F} \Rightarrow \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$$

- Siendo (Ω, \mathcal{F}, P) un espacio de probabilidad, si $A \in \mathcal{F}$, entonces $P(A) + P(A^c) = 1$

$$A \cap A^c = \emptyset \Rightarrow (A \cap A^c)^c = A \cup A^c = \Omega$$

$$\Rightarrow P(A \cup A^c) = P(\Omega) \Rightarrow P(A) + P(A^c) = 1$$

- Siendo (Ω, \mathcal{F}, P) un espacio de probabilidad, si $A, B \in \mathcal{F}$, entonces $P(A \cup B) + P(A \cap B) = P(A) + P(B)$

$$A = (A - B) \cup (A \cap B) \text{ \& } B = (B - A) \cup (A \cap B)$$

$$\Rightarrow P(A) + P(B) = P(A - B) + P(B - A) + P(A \cap B) + P(A \cap B)$$

$$\Rightarrow P(A) + P(B) = P[(A - B) \cup (B - A) \cup (A \cap B)] + P(A \cap B)$$

$$\Rightarrow P(A) + P(B) = P(A \cup B) + P(A \cap B)$$

- De este modo, en un espacio de probabilidad (Ω, \mathcal{F}, P) , si $A, B \in \mathcal{F}$, entonces $P(A_i \cup A_j) = P(A_i) + P(A_j) - P(A_i \cap A_j)$, y esto se puede generalizar con la siguiente expresión:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \dots + (-1)^{n+1} P\left(\bigcap_{i=1}^n A_i\right)$$

- Siendo (Ω, \mathcal{F}, P) un espacio de probabilidad, si $A, B \in \mathcal{F}$ y $A \subseteq B$, entonces $P(A) \leq P(B)$

$$B = (B - A) \cup A \Rightarrow P(B) = P(B - A) + P(A) \geq P(A)$$

- Siendo \mathcal{E} un experimento con un espacio de probabilidad (Ω, \mathcal{F}, P) , la estructura de este dependerá en gran medida en si Ω es un conjunto numerable (finito o infinito numerable) o no

- De ser numerable, normalmente se toma \mathcal{F} como el conjunto de todos los subconjuntos de Ω , y, para cada $\omega \in \Omega$, es de interés

saber si ω es el resultado de \mathcal{E} , por lo que, suponiendo que $\Omega = \{\omega_1, \omega_2, \dots\}$, se necesita que cualquier $\{\omega\} \in \mathcal{F}$

- Si $A \subseteq \Omega$, entonces es numerable y A se puede expresar como la unión $\bigcup_{\omega \in A} \{\omega\} \in \mathcal{F}$ (de modo que $A = \bigcup_{\omega \in A} \{\omega\}$). Como la probabilidad del evento A se determina por las probabilidades de cada $\{\omega\}$ que pertenezca a A , y cada subconjunto es un evento disjunto, entonces se obtiene la siguiente igualdad:

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

- Si $\Omega = \{\omega_1, \omega_2, \dots, \omega_N\}$ y $P(\omega_i) = P(\omega_j)$ para toda i y j , entonces se obtienen las siguientes identidades:

$$P(\omega) = \frac{1}{N} \text{ for } \forall \omega \in \Omega$$

$$P(A) = \sum_{\omega \in A} P(\omega) = \frac{1}{N} \sum_{\omega \in A} 1 = \frac{1}{N} |A| \text{ for } A \subseteq \Omega$$

- En teoría de probabilidad es común trabajar con problemas en los que los resultados de los experimentos son equiprobables (tienen la misma probabilidad), de modo que se reducen a problemas de combinatoria
- Ahora que se ha definido la noción de espacio de probabilidad, sus componentes y lo que es un experimento, es posible introducir la noción de probabilidad condicional, independencia y continuidad de la probabilidad
 - Siendo \mathcal{E} un experimento con un espacio de probabilidad (Ω, \mathcal{F}, P) , es posible que se tenga información incompleta del resultado real de \mathcal{E} sin saber todo sobre este resultado
 - Si A y B son eventos y se dice que B ocurre, entonces, con esta información, la nueva probabilidad de que ocurra A puede no coincidir con la probabilidad $P(A)$
 - En esta nueva circunstancia, A ocurre si, y solo si, $A \cap B$ ocurre, sugiriendo que la nueva probabilidad de A puede ser proporcional a $P(A \cap B)$
 - Si $A, B \in \mathcal{F}$ y $P(B) > 0$, la probabilidad condicional de A dado B se denota por $P(A|B)$ y se define con la siguiente fórmula:

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

- La constante de proporcionalidad de la definición se ha escogido de manera que la probabilidad $P(B|B)$ satisface $P(B|B) = 1$ (lo cual es lógico)
- Si $A, B \in \mathcal{F}$ y $P(B) > 0$, entonces (Ω, \mathcal{F}, Q) es el espacio de probabilidad donde $Q: \mathcal{F} \rightarrow \mathbb{R}$ se define por $Q \equiv P(A|B)$

$$Q(A) \geq 0 \quad Q(\Omega) = P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

$A_1, A_2, \dots \in \mathcal{F}$ such that $A_i \cap A_j = \emptyset$ for $i \neq j$

$$\begin{aligned} \Rightarrow Q\left(\bigcup_i A_i\right) &= \frac{1}{P(B)} P\left(\left(\bigcup_i A_i\right) \cap B\right) = \\ &= \frac{1}{P(B)} P\left(\bigcup_i (A_i \cap B)\right) = \frac{1}{P(B)} \sum_i P(A_i \cap B) = \sum_i Q(A_i) \end{aligned}$$

- Si $A, B, C \in \mathcal{F}$ y $P(B \cap C), P(C) > 0$, entonces la probabilidad de la intersección de estos eventos es $P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$

$$\begin{aligned} P(A|B \cap C)P(B|C)P(C) &= \frac{P(A \cap B \cap C)}{P(B \cap C)} \frac{P(B \cap C)}{P(C)} P(C) = \\ &= \frac{P(A \cap B \cap C)}{P(B \cap C)} P(B \cap C) = P(A \cap B \cap C) \end{aligned}$$

- Si $A, B \in \mathcal{F}$ y $P(A), P(B) > 0$, entonces la probabilidad condicional $P(B|A) = P(A|B)P(B)/P(A)$

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

- Se llama independientes a dos eventos A y B si la ocurrencia de uno de los dos no afecta a la probabilidad (nueva) de que el otro ocurra. De manera más formal, los eventos A y B del espacio de probabilidad (Ω, \mathcal{F}, Q) son independientes si $P(A \cap B) = P(A)P(B)$, y son dependientes de otro modo
- Esto quiere decir que, si $P(A), P(B) > 0$ y A y B son eventos independientes, entonces $P(A|B) = P(A)$ y $P(B|A) = P(B)$. No obstante, en la definición se permite que $P(A) = 0$ o $P(B) = 0$

- Esta definición se puede generalizar aún más si se consideran más de dos eventos. Una familia $\mathcal{A} \equiv (A_i: i \in I)$ de eventos se llama independiente si, para todos los subconjuntos finitos J de I , se cumple la siguiente igualdad:

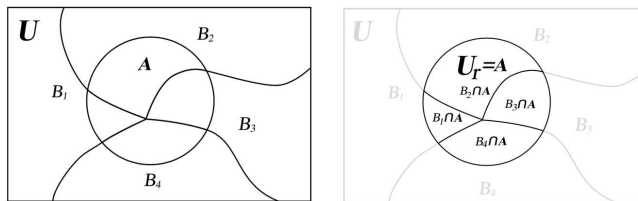
$$P\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i)$$

- La familia \mathcal{A} se denomina independiente a pares si la igualdad anterior se sostiene cuando $|J| = 2$ (para cada dos eventos)
- Tres eventos A, B y C son independientes si, y solo si, se cumplen las siguientes igualdades:

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

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

- En consecuencia, hay familias de eventos que son independientes a pares pero que no son independientes (porque hay alguna de las intersecciones que no es el producto de las probabilidades de cada evento)
- Siendo (Ω, \mathcal{F}, Q) un espacio de probabilidad, una partición de Ω es una colección $\{B_i: i \in I\}$ de eventos disjuntos (de modo que $B_i \in \mathcal{F}$ para toda i y $B_i \cap B_j = \emptyset$ si $i \neq j$) con union $\bigcup_i B_i = \Omega$. A partir de este concepto, se puede derivar el teorema de la partición (o de probabilidad total)



- Si $\{B_1, B_2, \dots\}$ es una partición de Ω (de modo que $\bigcup_i B_i = \Omega$) con $P(B_i) > 0$ para toda i , entonces se da la siguiente igualdad:

$$P(A) = \sum_i P(A|B_i)P(B_i) \quad \text{for } A \in \mathcal{F}$$

- La demostración de este teorema se basa en las expresiones equivalentes de A y la probabilidad para la unión de eventos disjuntos

$$\begin{aligned} P(A) &= P(A \cap \Omega) = P\left(A \cap \left(\bigcup_i B_i\right)\right) = P\left(\bigcup_i (A \cap B_i)\right) \\ &= \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i) \end{aligned}$$

- Hay muchas situaciones en las que se quiere deducir algo a partir de una pieza de evidencia, de modo que A sería la evidencia y B_1, B_2, \dots los posibles estados de la naturaleza. Entonces, se quiere saber $P(B_j|A)$ a partir de $P(A|B_i)$ y, para ello, se puede utilizar el teorema de Bayes

- Siendo $\{B_1, B_2, \dots\}$ una partición de Ω (de modo que $\bigcup_i B_i = \Omega$) y $P(A), P(B_i) > 0$ para toda i , entonces se da la siguiente igualdad para cualquier evento A :

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

- La demostración de este teorema se basa en el teorema de la partición anteriormente visto

$$\begin{aligned} P(B_j|A) &= \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{P(A \cap \Omega)} = \frac{P(A|B_j)P(B_j)}{P(A \cap (\bigcup_i B_i))} = \\ &= \frac{P(A|B_j)P(B_j)}{P(A \cap (\bigcup_i B_i))} = \frac{P(A|B_j)P(B_j)}{P(\bigcup_i (A \cap B_i))} = \frac{P(A|B_j)P(B_j)}{\sum_i P(A \cap B_i)} = \\ &= \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)} \end{aligned}$$

- Debido a que el numerador de $P(B_j|A)$ es la probabilidad *a priori* de B_j por la verosimilitud $P(A|B_j)$ y el denominador es la suma de productos de probabilidad *a priori* de B_i por la verosimilitud $P(A|B_i)$, si se multiplicara una constante c por cada una de las verosimilitudes, esta constante se cancelaría en la división

$$\frac{cP(A|B_j)P(B_j)}{\sum_i cP(A|B_i)P(B_i)} = \frac{cP(A|B_j)P(B_j)}{c \sum_i P(A|B_i)P(B_i)} = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

- A partir de este hecho, solo es necesario saber la verosimilitud dentro de una constante de proporcionalidad: los pesos relativos

para todas las posibilidades (la verosimilitud) es todo lo que se necesita

- De manera similar, si se multiplica cada probabilidad *a priori* por una constante, de modo que ocurriría lo mismo que antes: lo único que sería necesario son los pesos relativos que se daría a cada probabilidad *a priori*
- De este modo, se suele expresar el teorema de Bayes en su forma proporcional, dado que se multiplican todos los valores por $1/\sum_i P(A|B_i)P(B_i)$ (los pesos relativos)

$$P(B_j|A) \propto P(B_j)P(A|B_j)$$

- Se puede, por tanto, resumir el uso del teorema de Bayes para eventos con lo siguientes tres pasos:
 - El primer paso es la multiplicación de $P(B_i)P(A|B_i)$ para cada una de las B_i . Esto encuentra la probabilidad de $A \cap B_i$ por la regla de la multiplicación
 - El segundo paso consiste en sumar para $i = 1, 2, \dots, n$ todas las probabilidades $P(B_i)P(A|B_i)$. Esto encuentra la probabilidad de A a partir del teorema de probabilidad total
 - El tercer paso es la división de cada $P(B_i)P(A|B_i)$ entre la suma anterior. Esto encuentra la probabilidad condicional $P(B_i|A)$ para cada B_i posible
- Otra manera de poder lidiar con los eventos inciertos que se modelan como aleatorios es formar las *odds* de los eventos. Las *odds* para un evento C es igual a la probabilidad del evento que ocurre dividida por la probabilidad de que el evento no ocurra

$$odds(C) = \frac{P(C)}{P(C^c)} = \frac{P(C)}{1 - P(C)}$$

- Como la probabilidad de que un evento no ocurra es $P(C^c) = 1 - P(C)$, se puede ver como hay una correspondencia uno a uno entre la probabilidad $P(C)$ y las *odds*(C). Aislando para $P(C)$, se puede obtener la siguiente relación:

$$P(C) = \frac{odds(C)}{1 + odds(C)}$$

- El factor de Bayes, denotado por B , contiene la evidencia en los datos relevantes para el evento C , denotados por D , que han ocurrido. Este es el factor por el cual las *odds a priori* se han cambiado a las *odds* posteriores

$$\text{prior odds}(C) \times B = \text{posterior odds}(C)$$

$$\Rightarrow B = \frac{\text{posterior odds}(C)}{\text{prior odds}(C)}$$

- Se puede sustituir en la *ratio* de probabilidades para las *odds* posteriores y *a priori* para poder encontrar la siguiente fórmula:

$$B = \frac{P(D|C)}{P(D|C^c)}$$

- Por lo tanto, el factor de Bayes se puede interpretar como la *ratio* entre las probabilidades de obtener los datos cuando ha ocurrido un evento dado entre las probabilidades de obtener los datos cuando no ha ocurrido tal evento. Cuando este factor es mayor a uno, entonces es más probable que el evento haya ocurrido a partir de los datos, y si es menor, lo más probable es que no haya pasado
- Hay una propiedad de las medidas de probabilidad que es muy útil, la cual es la propiedad de continuidad de la medida de probabilidad
 - Una secuencia de eventos $\{A_n\}_{n=1}^{\infty}$ en un espacio de probabilidad (Ω, \mathcal{F}, P) se denomina creciente cuando $A_n \subseteq A_{n+1}$ para toda $n = 1, 2, 3, \dots$ y decreciente cuando $A_n \supseteq A_{n+1}$ para toda $n = 1, 2, 3, \dots$
 - Por lo tanto, la unión $A = \bigcup_{i=1}^{\infty} A_i$ de una secuencia creciente se denomina el límite de la secuencia y es un evento como consecuencia de los axiomas para un espacio de eventos (clausura bajo unión)
 - En el caso en que la secuencia sea decreciente, entonces la intersección $A = \bigcap_{i=1}^{\infty} A_i$ se denomina límite de la secuencia y es un evento como consecuencia de los axiomas para un espacio de eventos (clausura bajo intersección)
 - Siendo (Ω, \mathcal{F}, P) un espacio de probabilidad, si $\{A_n\}_{n=1}^{\infty}$ es una secuencia creciente de eventos en \mathcal{F} con límite A , entonces se cumple la siguiente igualdad:

$$P(A) = \lim_{n \rightarrow \infty} P(A_n) = \lim_{n \rightarrow \infty} P\left(\bigcup_{i=1}^n A_i\right)$$

- Este límite se resuelve como el límite de una secuencia, de modo que es necesaria una fórmula para la determinación de A_n

- Definiendo $B_i \equiv A_i - A_{i-1}$, se puede ver que el límite A es la unión de los eventos disjuntos $A = A_1 \cup B_2 \cup B_3 \cup \dots$, lo cual permite obtener las siguientes identidades:

$$A = A_1 \cup B_2 \cup B_3 \cup \dots \Rightarrow P(A) = P(A_1) + \sum_{i=2}^{\infty} P(B_i)$$

$$\Rightarrow P(A) = P(A_1) + \lim_{n \rightarrow \infty} \sum_{i=2}^n P(B_i)$$

$$\Rightarrow P(A) = P(A_1) + \lim_{n \rightarrow \infty} \sum_{i=2}^n [P(A_i) - P(A_{i-1})] =$$

$$= P(A_1) + \lim_{n \rightarrow \infty} P(A_n) - P(A_1) = \lim_{n \rightarrow \infty} P(A_n)$$

- De manera similar, siendo (Ω, \mathcal{F}, P) un espacio de probabilidad, si $\{B_n\}_{n=1}^{\infty}$ es una secuencia decreciente de eventos en \mathcal{F} con límite B , entonces se cumple la siguiente igualdad:

$$P(B) = \lim_{n \rightarrow \infty} P(B_n) = \lim_{n \rightarrow \infty} P\left(\bigcap_{i=1}^n B_i\right)$$

- Este límite se resuelve como el límite de una secuencia, de modo que es necesaria una fórmula para la determinación de B_n
- Definiendo $A_i \equiv \Omega - B_i$, se puede ver que el límite A es la intersección de los eventos disjuntos $A = A_1 \cap A_2 \cap A_3 \cap \dots$, lo cual permite obtener las siguientes identidades:

$$A = A_1 \cap A_2 \cap A_3 \cap \dots \Rightarrow A^c = A_1^c \cup A_2^c \cup A_3^c \cup \dots$$

$$\Rightarrow A_n^c \subseteq A_{n+1}^c \Rightarrow \emptyset \cup B_{n+1} \subseteq \emptyset \cup B_n \Rightarrow B_{n+1} \subseteq B_n$$

$$\Rightarrow P(B) = P(A^c) = \lim_{n \rightarrow \infty} \sum_{i=1}^n P(A_i^c) = \lim_{n \rightarrow \infty} \sum_{i=1}^n [P(\emptyset) + P(B_i)] =$$

$$= \lim_{n \rightarrow \infty} \sum_{i=1}^n P(B_i) = \lim_{n \rightarrow \infty} P(B_n)$$

Las variables aleatorias discretas

- Dado un espacio de probabilidad (Ω, \mathcal{F}, P) , normalmente uno se interesa en situaciones en las que una función de variable real X que actúa sobre Ω está involucrada. De este modo, se puede introducir la noción de función de masa de probabilidad y las distribuciones más comunes
 - Una variable discreta X en el espacio de probabilidad (Ω, \mathcal{F}, P) se define como el mapeado $X: \Omega \rightarrow \mathbb{R}$ tal que la imagen $X(\Omega)$ es un subconjunto numerable de \mathbb{R} y $X^{-1}(x) \equiv \{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}$ para toda $x \in \mathbb{R}$
 - Si $X: \Omega \rightarrow \mathbb{R}$ y $A \subseteq \Omega$, la imagen de A es el conjunto $X(A) \equiv \{X(\omega) : \omega \in A\}$ de valores que X toma en A
 - La palabra discreta se utiliza para referirse a la condición de que X solo puede tomar valores en \mathbb{R} numerables
 - Una variable discreta X toma valores en \mathbb{R} , pero no se puede predecir el valor actual con certidumbre dado que el experimento \mathcal{E} involucra azar. De este modo, se prefiere medir la probabilidad de que X tenga un valor x , lo cual se produce si, y solo si, ese subconjunto $X^{-1}(x)$ de Ω que se mapea en x es un evento (pertenece a \mathcal{F} y se le asignan probabilidades por P). Este es el significado de la segunda condición de la definición
 - La parte más interesante sobre la variable aleatoria discreta son los valores que puede tomar y las probabilidades asociadas con esos valores. Si X es una variable aleatoria discreta en (Ω, \mathcal{F}, P) , entonces su imagen es la imagen de Ω bajo X (que es el conjunto de valores $X(\Omega)$ que toma X)
 - La función de masa de probabilidad de una variable aleatoria discreta X es la función $p_X: \mathbb{R} \rightarrow [0,1]$ definido por $p_X(x) = P(X = x)$
 - Por lo tanto, $p_X(x)$ es la probabilidad de que el mapeado X tome un valor x
 - La imagen de X es numerable para cualquier variable aleatoria discreta X (dado que es el intervalo $[0,1]$) y se cumplen las siguientes dos condiciones:

$$p_X(x) = 0 \text{ if } x \notin \text{Im } X$$

$$\sum_{x \in \text{Im } X} p_X(x) = P\left(\bigcup_{x \in \text{Im } X} \{\omega \in \Omega : X(\omega) = x\}\right) = P(\Omega) = 1$$

- Siendo $S = \{s_i : i \in I\}$ un conjunto numerable de números reales distintos y $\{\pi_i : i \in I\}$ una colección de números reales tales que $\pi_i \geq 0$ para $i \in I$ y $\sum_{i \in I} \pi_i = 1$, existe un espacio de probabilidad (Ω, \mathcal{F}, P) y

una variable aleatoria discreta X en (Ω, \mathcal{F}, P) tal que la función de masa de probabilidad sea la siguiente:

$$\begin{cases} p_X(s_i) = \pi_i & \text{for } i \in I \\ p_X(s_i) = 0 & \text{for } s \notin S \end{cases}$$

- Este teorema se puede demostrar tomando $\Omega = S$, \mathcal{F} como el conjunto de todos los subconjuntos de Ω y $P(A) = \sum_{i: s_i \in A} \pi_i$ para $A \in \mathcal{F}$. Finalmente, se define $X: \Omega \rightarrow \mathbb{R}$ como $X(\omega) = \omega$ para $\omega \in \Omega$, y se puede ver que la función de masa de probabilidad será la susodicha
- Este teorema es muy útil debido a que permite olvidarse de los espacios muestrales, espacios de eventos y medidas de probabilidad, y solo es necesario especificar que la variable X toma el valor s_i con probabilidad π_i para $i \in I$ porque el teorema muestra que esta existe sin tener que construirla específicamente
- Ciertos tipos de variables discretas ocurren de manera frecuente, de modo que se puede crear una lista de las funciones de masa de probabilidad más frecuentes. En este caso, $n \in \mathbb{N}^+$, $p \in [0,1]$ y $q \equiv 1 - p$

- Una variable aleatoria discreta X sigue una distribución de Bernoulli con parámetro $p \in [0,1]$ si la imagen de X es $\{0,1\}$ (solo toma valores 1 o 0). La función de masa de probabilidad de esta variable:

$$\begin{cases} p_X(0) = q \\ p_X(1) = p \\ p_X(x) = 0 & \text{if } x \neq 0,1 \end{cases}$$

- Una variable aleatoria discreta X sigue una distribución binomial con parámetros n y $p \in (0,1)$ si la imagen de X es $\{0,1, \dots, n\}$. La función de masa de probabilidad de esta variable:

$$\begin{cases} p_X(k) = \binom{n}{k} p^k q^{n-k} & \text{for } k = 0,1, \dots, n \\ p_X(x) = 0 & \text{otherwise} \end{cases}$$

$$\sum_{k=0}^n \binom{n}{k} p^k q^{n-k} = (p + q)^n = 1$$

- Una variable aleatoria discreta X sigue una distribución de Poisson con parámetro $\lambda > 0$ si la imagen de X es $\{0,1,2, \dots\}$. La función de masa de probabilidad de esta variable:

$$\begin{cases} p_X(k) = \frac{1}{k!} \lambda^k e^{-\lambda} & \text{for } k = 0, 1, 2, \dots \\ p_X(x) = 0 & \text{otherwise} \end{cases}$$

$$\sum_{k=0}^{\infty} \frac{1}{k!} \lambda^k e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{1}{k!} \lambda^k = e^{-\lambda} e^{\lambda} = 1$$

- Una variable aleatoria discreta X sigue una distribución geométrica con parámetro $p \in (0,1)$ si la imagen de X es $\{0, 1, 2, \dots\}$. La función de masa de probabilidad de esta variable:

$$\begin{cases} p_X(k) = pq^{k-1} & \text{for } k = 1, 2, 3, \dots \\ p_X(x) = 0 & \text{otherwise} \end{cases}$$

$$\sum_{k=0}^{\infty} pq^{k-1} = p \sum_{k=0}^{\infty} q^{k-1} = \frac{p}{1-q} = 1$$

- Una variable aleatoria discreta X sigue una distribución binomial negativa con parámetro $p \in (0,1)$ si la imagen de X es $\{n, n+1, n+2, \dots\}$. La función de masa de probabilidad de esta variable:

$$\begin{cases} p_X(k) = \binom{k-1}{n-1} p^n q^{k-n} & \text{for } k = n, n+1, n+2, \dots \\ p_X(x) = 0 & \text{otherwise} \end{cases}$$

$$\sum_{k=n}^{\infty} \binom{k-1}{n-1} p^n q^{k-n} = p^n \sum_{k=0}^{\infty} \binom{k-1}{n-1} q^{k-n} = p^n (1-q)^{-n} = 1$$

- Es posible explicar de dónde surgen algunas de las distribuciones vistas a partir de otras

- La suma de variables discretas aleatorias que siguen una distribución de Bernoulli sigue una distribución binomial

$$S_n = X_1 + \dots + X_n \Rightarrow S_n(\omega) = X_1(\omega) + \dots + X_n(\omega)$$

$$\Rightarrow P(S_n = k) = P(\{\omega \in \Omega : X(\omega) = x\}) = \sum_{\omega: h(\omega)=k} P(\omega) =$$

$$= \binom{n}{k} p^k q^{n-k}$$

- Si en una distribución binomial n es muy grande, p es muy pequeña, pero np tiene un tamaño razonable (se asume que

$np = \lambda$), entonces esta suma de variables de Bernoulli se puede aproximar con una distribución de Poisson con parámetro λ y probabilidad $p = \frac{\lambda}{n}$

$$S_n = X_1 + \dots + X_n \Rightarrow S_n(\omega) = X_1(\omega) + \dots + X_n(\omega)$$

$$\begin{aligned} \Rightarrow P(S_n = k) &= \binom{n}{k} p^k q^{n-k} \approx \frac{n^k}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \approx \\ &\approx \frac{1}{k!} \lambda^k e^{-\lambda} \end{aligned}$$

- A partir del concepto de función de masa de probabilidad, es posible definir funciones sobre variables aleatorias discretas y la noción de valor esperado, varianza y valor esperado condicional
 - Siendo X una variable aleatoria discreta en el espacio de probabilidad (Ω, \mathcal{F}, P) y $g: \mathbb{R} \rightarrow \mathbb{R}$, es fácil ver que $Y = g(X)$ es también una variable aleatoria discreta en (Ω, \mathcal{F}, P) definida por $Y(\omega) = g(X(\omega))$ para $\omega \in \Omega$
 - Si $Y = g(X)$, entonces la función de masa de probabilidad se puede construir gracias a que x debe pertenecer al dominio de g y solo hay contribuciones diferentes de cero numerables en esta suma

$$\begin{aligned} p_Y(y) &= P(Y = y) = P(g(X) = y) = P(X \in g^{-1}(y)) = \\ &= \sum_{x \in g^{-1}(y)} P(X = x) \end{aligned}$$

- De este modo, se puede aislar X con tal de que sea una función de y y así encontrar la función de masa de probabilidad

if $Y = X^2$, then

$$\mathbb{P}(Y = y) = \begin{cases} \mathbb{P}(X = \sqrt{y}) + \mathbb{P}(X = -\sqrt{y}) & \text{if } y > 0, \\ \mathbb{P}(X = 0) & \text{if } y = 0, \\ 0 & \text{if } y < 0. \end{cases}$$

- Si X es una variable aleatoria discreta, el valor esperado o esperanza de X se denota por $E(X)$ y se define como $E(X) \equiv \sum_{x \in \text{Im } X} xP(X = x)$ cuando la suma converge absolutamente

$$E(X) \equiv \sum_{x \in \text{Im } X} xP(X = x) = \sum_{x \in \text{Im } X} xp_X(x)$$

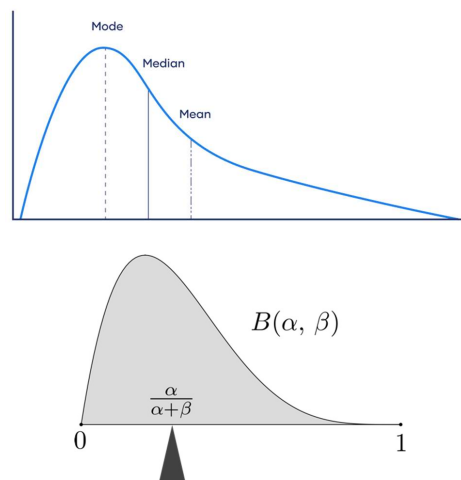
- Estas identidades se suelen escribir de manera abreviada de la siguiente manera:

$$E(X) \equiv \sum_x xP(X=x) = \sum_{x \in \text{Im } X} xp_X(x)$$

- Que una suma converja absolutamente quiere decir que la sumatoria del valor absoluto de $xP(X=x)$ es menor a infinito. Esto es necesario porque la imagen $\text{Im } X$ puede ser un conjunto infinito, y se necesita que la sumatoria tome el mismo valor sin importar el orden de los términos que se añaden

$$\sum_{x \in \text{Im } X} |xP(X=x)| = \sum_{x \in \text{Im } X} |x|P(X=x) < \infty$$

- La esperanza de una variable continua X es una indicación del centro de la distribución de X , pero no es el valor que divide la distribución en dos partes con igual probabilidad (como la mediana) o el valor más frecuente (la moda) necesariamente, sino su centro de masa (el valor en el que la masa de probabilidad agregada a la izquierda y a la derecha se equilibraría)



- Si X es una variable aleatoria discreta en un espacio de probabilidad (Ω, \mathcal{F}, P) y $g: \mathbb{R} \rightarrow \mathbb{R}$, entonces $g(X)$ es una variable aleatoria discreta y se cumple la siguiente igualdad siempre que la suma converja absolutamente (ley del estadístico subconsciente):

$$E(g(X)) = \sum_{x \in \text{Im } X} g(x)P(X = x) = \sum_{x \in \text{Im } X} g(x)p_X(x)$$

- De manera intuitiva, el resultado es claro porque $g(X)$ toma valores $g(x)$ cuando X toma un valor x , de modo que un evento con probabilidad $P(X = x)$
- Escribiendo I como la imagen de X y teniendo $Y = g(X)$ con imagen $g(I)$, se pueden derivar las siguientes equivalencias:

$$\begin{aligned} E(Y) &= E(g(X)) = \sum_{y \in g(I)} yP(Y = y) = \sum_{y \in g(I)} yP(g(X) = y) = \\ &= \sum_{y \in g(I)} yP(X \in g^{-1}(y)) = \sum_{y \in g(I)} y \sum_{x \in I: g(x)=y} P(X = x) = \\ &= \sum_{x \in I: g(x)=y} g(x) \sum_{x \in I: g(x)=y} P(X = x) = \\ &= \sum_{x \in I: g(x)=y} g(x)P(X = x) \end{aligned}$$

- Siendo X una variable aleatoria discreta, si $P(X \geq 0) = 1$ y $E(X) = 0$, entonces $P(X = 0) = 1$

$$P(X \geq 0) = 1 \Rightarrow E(X) = \sum_{x \in \text{Im } X} xP(X = x) = 0$$

$$\Rightarrow P(X = x) = 0 \text{ for } x \neq 0$$

$$\Rightarrow P(X = 0) = 1 \text{ such that } \sum_{x \in \text{Im } X} P(X = x) = 1$$

- Siendo X una variable aleatoria discreta y $a, b \in \mathbb{R}$, la esperanza cumple las propiedades lineales, por lo que $E(aX + b) = aE(X) + b$

$$\begin{aligned} E(aX + b) &= \sum_{x \in \text{Im } aX+b} (ax + b)P(aX + b = x) = \\ &= \sum_{x \in \text{Im } aX+b} \left[axP\left(X = \frac{x-b}{a}\right) + bP\left(X = \frac{x-b}{a}\right) \right] = \\ &= a \sum_{x \in \text{Im } X: ax+b= } xP(X = x) + b \sum_{x \in \text{Im } X: ax+b= } P(X = x) = \end{aligned}$$

$$= aE(X) + b$$

- La varianza $var(X)$ de una variable aleatoria discreta X es definida por $var(X) = E([X - E(X)]^2)$

$$Var(X) \equiv E([X - E(X)]^2) = \sum_{x \in Im X} (x - \mu)^2 P(X = x)$$

$$where E(X) = \mu$$

- La varianza de una variable aleatoria discreta X es una medida del grado de dispersión de X con respecto a su valor esperado $E(X)$ (el cual es un número)
- Si la dispersión de X con respecto a su valor esperado es pequeña, entonces $|X - \mu|$ tiende a ser pequeño, por lo que la varianza es más pequeña también. Si, en cambio, es grande, entonces $|X - \mu|$ es grande y la varianza también
- Expandiendo la varianza, es posible encontrar una expresión que permite calcular la varianza de manera mucho más simple

$$\begin{aligned} Var(X) &= \sum_{x \in Im X} (x - \mu)^2 P(X = x) = \\ &= \sum_{x \in Im X} (x^2 - 2x\mu + \mu^2) P(X = x) = \\ &= \sum_{x \in Im X} x^2 P(X = x) - 2\mu \sum_{x \in Im X} x P(X = x) \\ &\quad + \mu^2 \sum_{x \in Im X} P(X = x) = E(X^2) - 2[E(X)]^2 + [E(X)]^2 = \\ &= E(X^2) - [E(X)]^2 \Rightarrow var(X) = E(X^2) - [E(X)]^2 \end{aligned}$$

- Si X es una variable aleatoria discreta y $P(B) > 0$, el valor esperado o esperanza condicional de X dado B se denota por $E(X|B)$ y se define por la siguiente fórmula cuando esta suma converge absolutamente:

$$E(X|B) \equiv \sum_{x \in Im X} x P(X = x|B)$$

- Si X es una variable discreta y $\{B_i\}_{i=1}^{\infty}$ es una partición del espacio muestral tal que $P(B_i) > 0$ para cada i , entonces se

cumple la siguiente igualdad cuando la suma converge absolutamente:

$$E(X) = \sum_i E(X|B_i)P(B_i)$$

- La demostración del teorema de la partición se basa en la definición de valor esperado y de la probabilidad condicional

$$\begin{aligned} \sum_i E(X|B_i)P(B_i) &= \sum_i \sum_{x \in \text{Im } X} xP(X=x|B_i)P(B_i) = \\ &= \sum_i \sum_{x \in \text{Im } X} x \frac{P(\{X=x\} \cap B_i)}{P(B_i)} P(B_i) = \\ &= \sum_i \sum_{x \in \text{Im } X} xP(\{X=x\} \cap B_i) = \\ &= \sum_{x \in \text{Im } X} xP\left(\{X=x\} \cap \left(\bigcup_i B_i\right)\right) = \\ &= \sum_{x \in \text{Im } X} xP(\{X=x\} \cap \Omega) = \sum_{x \in \text{Im } X} xP(X=x) = E(X) \end{aligned}$$

- Siendo X una variable discreta y $g: \mathbb{R} \rightarrow \mathbb{R}$ una función, si x es un número real tal que $P(X=x) > 0$, entonces la esperanza condicional de $g(X)$ dado $X=x$ es $g(x)$

$$E(g(X)|X=x) = g(x)$$

- La demostración de este teorema se basa en el hecho de que la probabilidad de que $g(X) = g(x)$ es la misma que la de que $X=x$ y en que el único valor $g(x)$ en la imagen de $g(X)$ es, lógicamente, $g(x)$

$$\begin{aligned} E(g(X)|X=x) &= \\ &= \sum_{g(x) \in \text{Im } g(X)} g(x)P(g(X)=g(x)|X=x) = \\ &= \sum_{g(x) \in \text{Im } g(X)} g(x) \frac{P(\{g(X)=g(x)\} \cap \{X=x\})}{P(X=x)} = \end{aligned}$$

$$= \sum_{g(x) \in \text{Im } g(X)} g(x) \frac{P(X=x)}{P(X=x)} = \sum_{g(x) \in \text{Im } g(X)} g(x) = g(x)$$

Las funciones de distribución y de densidad

- Que las variables aleatorias discretas solo puedan tomar valores numerables es una condición muy restrictiva en muchas situaciones, por lo que se puede hacer una definición más extensa e introducir el concepto de funciones de distribución de probabilidad
 - Una variable aleatoria X en el espacio de probabilidad (Ω, \mathcal{F}, P) es un mapeado $X: \Omega \rightarrow \mathbb{R}$ tal que $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ para toda $x \in \mathbb{R}$
 - Si una variable aleatoria cumple estas condiciones, entonces la función X se denomina \mathcal{F} -medible
 - Se requiere que una variable cumpla estas condiciones porque a uno le interesan los valores que toma la variable aleatoria X y las probabilidades asociadas con estos valores
 - Una manera conveniente de realizar esto es fijando $x \in \mathbb{R}$ y así evaluar la probabilidad de que X tome un valor en el intervalo $(-\infty, x]$
 - Esta probabilidad solo existe si su imagen inversa $X^{-1}((-\infty, x]) = \{X \leq x\}$ está en el espacio de eventos \mathcal{F} , por lo que el postulado anterior se mantiene para toda $x \in \mathbb{R}$
 - Cualquier variable aleatoria discreta es una variable aleatoria debido a que $\{X \leq x\}$ es la unión numerable de eventos en \mathcal{F} y pertenece, por tanto, a \mathcal{F}

$$\{X \leq x\} = \bigcup_{y \in \text{Im } X: y \leq x} \{X = y\}$$

- Aunque las variables aleatorias discretas se estudiaban a través de su función de masa, las variables aleatorias se estudian a través de las funciones de distribución de probabilidad. Si X es una variable en (Ω, \mathcal{F}, P) , la función de distribución de probabilidad de X es la función $F_X: \mathbb{R} \rightarrow [0,1]$ se define por la siguiente expresión:

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

- Esta función a veces es llamada función de distribución de probabilidad acumulada de X

- La función de distribución tiene varias propiedades generales y básicas que son útiles, la primera de las cuales es que, si $x \leq y$, entonces $F_X(x) \leq F_X(y)$
 - De este modo, la función de distribución es monótonamente no decreciente
 - Esto ocurre porque $\{X \leq x\} \subseteq \{X \leq y\}$ cuando $x \leq y$, dado que si X toma un valor que no excede x , entonces tampoco excede y
- Otras propiedades elementales de $F_X(x)$ tienen relación con su comportamiento cuando x está cerca de $-\infty$ o ∞ . Los siguientes resultados son intuitivos:

$$F_X(x) \rightarrow 0 \text{ as } x \rightarrow -\infty \quad F_X(x) \rightarrow 1 \text{ as } x \rightarrow \infty$$

- En el primer caso, si $x \rightarrow -\infty$, la probabilidad del evento $\{X \leq x\}$ se hace cada vez menor, mientras que, en el segundo caso, si $x \rightarrow \infty$, entonces el evento se hace cada vez más probable
- Estas propiedades se pueden verificar a través de la continuidad de la medida de probabilidad de P
- La continuidad de P también permite demostrar que las funciones de distribuciones son continuas por la derecha

$$F_X(x) \rightarrow 0 \text{ as } \varepsilon \rightarrow 0^+$$

$$\Rightarrow P(X \leq x + \varepsilon) \rightarrow P(X \leq x) \text{ as } \varepsilon \rightarrow 0^+$$

- Estas propiedades caracterizan completamente las funciones de distribución, en el sentido que si F es una función que satisface estas condiciones, existe un espacio de probabilidad y una variable aleatoria X en este espacio tal que X tiene una función de distribución F
 - Esto hace que en muchas circunstancias se pueda omitir escribir los espacios de probabilidad y las variables aleatorias de manera explícita
- Una última propiedad útil es que, siendo $F_X = P(X \leq x)$ la probabilidad de que X tome un valor de un intervalo finito $(-\infty, x]$, para encontrar la probabilidad de que X tome un valor en intervalo $(a, b]$ en donde $a < b$ se realiza el siguiente cálculo:

$$\begin{aligned} P(a < X \leq b) &= P(\{X \leq b\} - \{X \leq a\}) = \\ &= P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a) \end{aligned}$$

- Esto ocurre dado que el evento $\{X \leq a\}$ es un subconjunto del evento $\{X \leq b\}$
- Otras propiedades útiles de las variables aleatorias y de las funciones de distribución son las siguientes:

- Para una variable aleatoria discreta X , se cumplen la siguiente identidad:

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

- Siendo m un número real, se denomina mediana de la variable X si $P(X < m) \leq 1/2 \leq P(X \leq m)$, y es posible demostrar que todas las variables aleatorias tienen al menos una mediana
- Si X es una variable aleatoria y c es un número real tal que $P(X = c) > 0$, la función de distribución $F_X(x)$ de X es discontinua en el punto $x = c$
- El número de discontinuidades que puede tomar una función de distribución de probabilidad es numerable, lo cual se puede demostrar a partir de mapear las discontinuidades y juntar la longitud total de estos intervalos

- Si F es creciente, entonces los intervalos $(F(x-), F(x+))$ para $x \in \mathbb{R}$ son mutuamente disjuntos
- Para $|x| < N$, el intervalo $(F(x-), F(x+))$ estará dentro del intervalo $(F(-N), F(N))$ porque $-N < x < N$, lo cual implica la siguiente desigualdad:

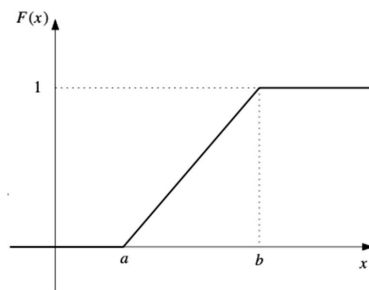
$$\sum_{|x| < N} [F(x+) - F(x-)] \leq F(-N) - F(N) < \infty$$

- Esta suma sobre un conjunto numerable es la suprema de la suma sobre todas las sumas parciales finitas. Que esta suma sea finita significa que el número de términos que no son nulos en la suma es numerable (dado que, si no, no se podría numerar la suma, no tendría sentido)
- Por lo tanto, el conjunto $D_n \equiv \{x \in (-N, N) : F(-x) \neq F(x+)\}$ es contable (el conjunto de elementos $F(x+) - F(x-)$ que no son nulos). El conjunto entero de discontinuidades es $\bigcup_N D_N$ es contable, de modo que una función de densidad $F(x)$ solo puede tomar un conjunto de discontinuidades contables

- Algunos de los ejemplos más importantes de funciones de distribución son las siguientes:

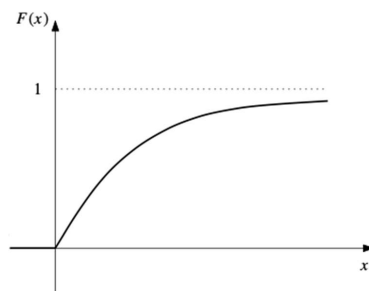
- Siendo $a, b \in \mathbb{R}$ y $a < b$, una variable aleatoria sigue una distribución uniforme en el intervalo $[a, b]$ si tiene la siguiente función de distribución:

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



- Siendo $\lambda > 0$, una variable aleatoria sigue una distribución exponencial con parámetro λ si tiene la siguiente función de distribución:

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-\lambda x} & \text{if } x > 0 \end{cases}$$



- Hay muchos tipos de variables aleatorias, pero hay dos grandes clases importantes: las variables aleatorias discretas y las variables aleatorias continuas. Las discretas toman valores numerables y su función de distribución generalmente son a trozos, mientras que las continuas son muy suaves

- Una variable aleatoria X es continua si su función de distribución F_X se puede escribir de la siguiente forma para una función f_X no negativa, la cual se denomina función de densidad de probabilidad:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(u) du \quad \text{for } x \in \mathbb{R}$$

- Siendo X una variable aleatoria continua y F_X tiene buen comportamiento, entonces la función de densidad se puede definir de la siguiente manera:

$$f_X(x) \equiv \begin{cases} F'_X(x) & \text{if this deriv. exists} \\ 0 & \text{otherwise} \end{cases}$$

- Aunque a esta definición le falta rigor matemático, esta expresión es adecuada para trabajar con funciones de densidad
- Las funciones de densidad son para las variables continuas lo que las funciones de masa de probabilidad son para las variables discretas, por lo que las propiedades generales son similares:

$$f_X(x) \geq 0 \quad \text{for } x \in \mathbb{R} \quad \int_{-\infty}^{\infty} f_X(x) dx = 1$$

- No obstante, $f_X(x)$ no es una probabilidad como en las funciones de masa de probabilidad y puede exceder 1
- En este caso, $f_X(x)$ es una medida de probabilidad en el siguiente sentido: si δx es pequeño y positivo, entonces la probabilidad de que X esté cerca de x satisface la siguiente condición:

$$\begin{aligned} P(x < X \leq x + \delta x) &= F(x + \delta x) - F(x) = \\ &= \int_x^{x+\delta x} f_X(u) du \approx f_X(x) \delta x \quad \text{for a small } \delta x \end{aligned}$$

- Si X es continua con la función de densidad f_X , entonces se cumplen las siguientes propiedades:

$$P(X = x) = 0 \quad \text{for } x \in \mathbb{R}$$

$$P(a \leq X \leq b) = \int_a^b f_X(u) du \quad \text{for } a, b \in \mathbb{R} \text{ with } a \leq b$$

- La demostración de que $P(X = x) = 0$ para toda $x \in \mathbb{R}$ es la siguiente:

$$\begin{aligned}
 P(X = x) &= \lim_{\varepsilon \rightarrow 0^+} P(x - \varepsilon < X \leq x) = \\
 &= \lim_{\varepsilon \rightarrow 0^+} [F_X(x) - F_X(x - \varepsilon)] = \lim_{\varepsilon \rightarrow 0^+} \int_{x-\varepsilon}^x f_X(u) du = 0
 \end{aligned}$$

- La demostración de que $P(X = x) = 0$ para toda $x \in \mathbb{R}$ es la siguiente:

$$\begin{aligned}
 P(a \leq X \leq b) &= P(a < X \leq b) = F_X(b) - F_X(a) = \\
 &= \int_a^b f_X(u) du
 \end{aligned}$$

- Algunas de las funciones de densidad más comunes son las siguientes:

- La distribución uniforme en el intervalo (a, b) tiene la siguiente función de densidad:

$$f_X(x) \equiv \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

- La distribución exponencial con parámetro $\lambda > 0$ tiene la siguiente función de densidad:

$$f_X(x) \equiv \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

- La distribución normal con los parámetros μ y σ^2 tiene la siguiente función de densidad:

$$f_X(x) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \text{ for } x \in \mathbb{R}$$

- La distribución de Cauchy tiene la siguiente función de densidad:

$$f_X(x) \equiv \frac{1}{\pi(1+x^2)} \text{ for } x \in \mathbb{R}$$

- La distribución gamma con los parámetros $\omega > 0$ y $\lambda > 0$ tiene la siguiente función de densidad:

$$f_X(x) \equiv \begin{cases} \frac{1}{\Gamma(\omega)} \lambda^\omega x^{\omega-1} e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

$$\text{where } \Gamma(\omega) = \int_0^{\infty} x^{\omega-1} e^{-x} dx$$

- La distribución beta con los parámetros s y $t > 0$ tiene la siguiente función de densidad:

$$f_X(x) \equiv \frac{1}{B(s,t)} x^{s-1} (1-x)^{t-1} \quad \text{for } 0 \leq x \leq 1$$

$$\text{where } B(s,t) = \int_0^1 x^{s-1} (1-x)^{t-1} dx = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}$$

- La distribución chi cuadrada con n grados de libertad (a veces se escribe χ_n^2) tiene la siguiente función de densidad:

$$f_X(x) \equiv \begin{cases} \frac{1}{2\Gamma(\frac{1}{2}n)} \left(\frac{1}{2}x\right)^{\frac{1}{2}n-1} e^{-\frac{1}{2}x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

$$\text{where } \Gamma\left(\frac{1}{2}n\right) = \int_0^{\infty} x^{\frac{1}{2}n-1} e^{-x} dx$$

- Aunque generalmente las funciones g de una variable aleatoria X no tienen por qué ser una variable aleatoria (al no satisfacer la condición anterior), pero cuando g tiene un buen comportamiento, entonces sí que lo son. Por lo tanto, es útil saber como es la distribución de $Y = g(X)$ sabiendo la distribución de la variable aleatoria X
 - Siendo X una variable aleatoria en (Ω, \mathcal{F}, P) y $g: \mathbb{R} \rightarrow \mathbb{R}$ una función con buen comportamiento, $Y = g(X)$ es un mapeado de Ω a \mathbb{R} definido por $Y(\omega) = g(X(\omega))$ para $\omega \in \Omega$ que es una variable aleatoria
 - Para que g tenga buen comportamiento, esta tiene que ser continua o monótona normalmente
 - Si X es una variable aleatoria continua con función de densidad f_X y $g: \mathbb{R} \rightarrow \mathbb{R}$ es estrictamente creciente y diferenciable, entonces $Y = g(X)$ tiene la siguiente función de densidad:

$$f_Y(y) = f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y) \quad \text{for } y \in \mathbb{R}$$

- La demostración de este teorema se basa en la función de distribución de X

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = \\ = F_X(g^{-1}(y)) \text{ since } g \text{ is incr.}$$

$$\Rightarrow f_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y)$$

- Si la función g fuera estrictamente decreciente, entonces el mismo argumento daría que $Y = g(X)$ tendría la siguiente función de densidad:

$$f_Y(y) = -f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y) \text{ for } y \in \mathbb{R}$$

- Que la igualdad se aplique de las dos maneras anteriores significa es que en verdad se coge el módulo de la derivada, por lo que queda la siguiente igualdad:

$$f_Y(y) = -f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right| \text{ for } y \in \mathbb{R}$$

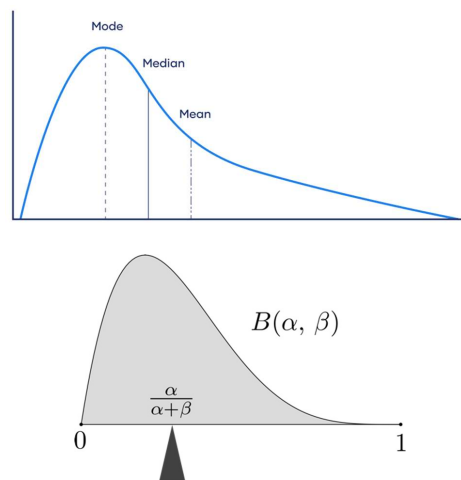
- Estos resultados se apoyan mucho en la monotonidad de g , por lo que otros casos se deben de tratar de manera concreta (si no, no aplica directamente)
- Igual que en el caso de las variables aleatorias discretas, las variables aleatorias continuas también tienen una esperanza y una varianza que se definen de manera parecida
 - Si X es una variable aleatoria continua con función de densidad f_X , el valor esperado de X se denota por $E(X)$ y se define por la siguiente expresión cuando la integral converja absolutamente:

$$E(X) \equiv \int_{-\infty}^{\infty} x f_X(x) dx \text{ where } \int_{-\infty}^{\infty} |x f_X(x)| dx < \infty$$

- Si X es una variable continua y $g: \mathbb{R} \rightarrow \mathbb{R}$ es una función con buen comportamiento, entonces $Y = g(X)$ es una variable aleatoria. Aunque no es fácil calcular el valor esperado de Y , se puede aplicar la ley del estadístico subconsciente para las variables continuas
- Si X es una variable aleatoria continua con función de densidad f_X y $g: \mathbb{R} \rightarrow \mathbb{R}$, entonces la siguiente igualdad se cumple siempre que la integral converja absolutamente:

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

- La esperanza de una variable continua X es una indicación del centro de la distribución de X , pero no es el valor que divide la distribución en dos partes con igual probabilidad (como la mediana) o el valor más frecuente (la moda) necesariamente, sino su centro de masa (el valor en el que la masa de probabilidad agregada a la izquierda y a la derecha se equilibraría)



- Igual que en el caso discreto, la varianza de X es una medida de su dispersión con respecto a su valor esperado μ , y unas medidas relacionadas a esta son la desviación estándar y la covarianza, la cual mide la relación lineal entre variables aleatorias y se define de la siguiente manera:

$$Var(X) \equiv \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx \text{ where } \mu = E(X)$$

- Cuando X toma valores alejados de μ , entonces $|X - \mu|$ suele ser grande, por lo que $E[(X - \mu)^2]$ también es grande. Del mismo modo, si se toman valores muy cercanos a μ , $|X - \mu|$ es pequeño y eso hace que $E[(X - \mu)^2]$ también lo sea
 - En el caso extremo en el que las observaciones se concentran en el valor esperado μ , entonces $Var(X) = 0$ si, y solo si, $P(X = \mu) = 1$. Esto se da porque $Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 P(X = x) \geq 0$

$$Var(X) = 0 \quad \text{iff} \quad P(X = \mu) = 1$$

- Cuando se calcula la varianza, es mucho más fácil trabajar con los momentos de X a través de la descomposición de la varianza

$$Var(X) = E(X^2) - \mu^2$$

- La varianza de una combinación lineal $aX + b$ es la varianza de X multiplicada por a^2 . Esto se puede demostrar a través de la definición de varianza

$$\begin{aligned} Var(aX + b) &= E[(aX + b - E(aX + b))^2] = \\ &= E[(aX + b - aE(X) - b)^2] = \\ &= E[(aX + b)^2 - 2(aX + b)(a\mu + b) + (a\mu + b)^2] = \\ &= E[a^2X^2 + 2abX + b^2 - 2a^2X\mu - 2a\mu b - 2abX - 2b^2 + a^2\mu^2 + 2a\mu b + b^2] = \\ &= E[a^2X^2 - 2a^2X\mu + a^2\mu^2] = E[a^2(X - \mu)^2] = a^2Var(X) \end{aligned}$$

- Como medida de dispersión, $Var(X)$ tiene una propiedad no deseable: no es lineal, en el sentido que $Var(aX) = a^2Var(X)$. Por lo tanto, se suele trabajar con la raíz $\sqrt{Var(X)}$, llamada desviación estándar de X y denotada por $sd(X)$
- A partir de las propiedades de la definición de probabilidad condicional y de las propiedades de la esperanza condicional, es posible definir el concepto de varianza condicional y descomponer la varianza en términos condicionales

- La varianza condicional de una variable aleatoria es la varianza de esta variable aleatoria condicionada a los valores que toma una o más variables aleatorias. Esta se define de la siguiente manera:

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

- La varianza se puede descomponer en dos términos condicionales debido a la definición de la esperanza condicional y a la de varianza condicional

$$Var(Y) = E_X[Var(Y|X)] + Var_X[E(Y|X)]$$

- Esta descomposición de la varianza se puede demostrar a partir de la descomposición incondicional de la varianza y de las propiedades de la esperanza

$$\begin{aligned}
\text{Var}(Y) &= E(Y^2) - E(Y)^2 = E_X[E(Y^2|X)] - E_X[E(Y|X)]^2 = \\
&= E_X[\text{Var}(Y|X) + E(Y|X)^2] - E_X[E(Y|X)]^2 = \\
&= E_X[\text{Var}(Y|X)] + E_X[E(Y|X)^2] - E_X[E(Y|X)]^2 = \\
&= E_X[\text{Var}(Y|X)] + \text{Var}_X[E(Y|X)]
\end{aligned}$$

- Es posible realizar una generalización del resultado anterior para más de una variable y para el caso de una partición del espacio muestral

- Cuando se condiciona la probabilidad a más de una variable, es posible hacer la siguiente generalización:

$$\begin{aligned}
\text{Var}(Y) &= E_{X_1, X_2}[\text{Var}(Y|X_1, X_2)] + E_{X_2}[\text{Var}[E(Y|X_1, X_2)|X_1]] + \\
&\quad + \text{Var}[E(Y|X_1, X_2)|X_1]
\end{aligned}$$

- La demostración de esta generalización proviene de las mismas propiedades y definiciones anteriormente vistas, pero extendidas a más de una variable aleatoria...

Commented [MOU1]: Acabar bien

- Si se considera una partición del espacio muestral, es posible utilizar la siguiente generalización:

$$\begin{aligned}
\text{Var}(Y) &= \sum_{i=1}^n \text{Var}(Y|A_i)P(A_i) + \sum_{i=1}^n E(X|A_i)^2 P(A_i)[1 - P(A_i)] - \\
&\quad - 2 \sum_{i=1}^n \sum_{j=1}^{i-1} E(X|A_i)P(A_i)E(X|A_j)P(A_j)
\end{aligned}$$

- La demostración de esta generalización proviene de las mismas propiedades y definiciones anteriormente vistas, pero extendidas a más de una variable aleatoria

Commented [MOU2]: Acabar bien

- La covarianza de dos variables aleatorias X e Y es la cantidad denotada por $\text{Cov}(X, Y)$ y definida de la siguiente manera (cuando el valor esperado existe):

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E[(X - \mu_X)(Y - \mu_Y)]$$

- La covarianza también se puede expresar equivalentemente de la siguiente manera, la cual es muy importante para operar con la esperanza en contextos en donde hay independencia de variables aleatorias:

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y)$$

- Expandiendo el producto dentro de la esperanza se puede obtener la siguiente expresión:

$$\begin{aligned} Cov(X, Y) &= E[(X - E(X))(Y - E(Y))] = \\ &= E[XY - XE(Y) - YE(X) + E(X)E(Y)] = \\ &= E(XY) - 2E(X)E(Y) + E(X)E(Y) = E(XY) - E(X)E(Y) \end{aligned}$$

- A partir de su definición, es fácil ver como la covarianza cumple con las propiedades lineales similares a las de la varianza:

$$\begin{aligned} Cov(aX + b, cY + d) &= \\ &= E[(aX + b - aE(X) - b)(cY + d - cE(Y) - d)] = \\ &= E[(aX - aE(X))(cY - cE(Y))] = \\ &= E[a(X - E(X))c(Y - E(Y))] = acCov(X, Y) \end{aligned}$$

- A partir de la covarianza, es posible obtener una expresión para la varianza de la suma de dos variables aleatorias $Var(X + Y)$

$$\begin{aligned} Var(X + Y) &= E[(X + Y - E(X + Y))^2] = \\ &= E[(X + Y - E(X) - E(Y))^2] = \\ &= E[(X - E(X) + Y - E(Y))^2] = \\ &= E[(X - E(X))^2 + 2(X - E(X))(Y - E(Y)) + (Y - E(Y))^2] = \\ &= E[(X - E(X))^2] + 2E[(X - E(X))(Y - E(Y))] + E[(Y - E(Y))^2] = \\ &= Var(X) + Var(Y) + 2Cov(X, Y) \end{aligned}$$

- Si las variables X e Y son independientes, entonces la covarianza es nula y eso hace que $Var(X \pm Y) = Var(X) + Var(Y)$

$$\begin{aligned} Cov(X, Y) &= 0 \Rightarrow \\ Var(X \pm Y) &= Var(X) + Var(Y) \pm 2Cov(X, Y) = \\ &= Var(X) + Var(Y) \end{aligned}$$

- El coeficiente de correlación de dos variables aleatorias X e Y es la cantidad $\rho(X, Y)$, definida de la siguiente manera para $\text{Var}(X)\text{Var}(Y) \neq 0$:

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

- A través de las propiedades de la covarianza y de la varianza, es posible ver como $\rho(aX + b, cY + d) = \rho(X, Y)$ tal que $ac \neq 0$

$$\begin{aligned} \rho(aX + b, cY + d) &= \frac{\text{Cov}(aX + b, cY + d)}{\sqrt{\text{Var}(aX + b)\text{Var}(cY + d)}} = \\ &= \frac{ac\text{Cov}(X, Y)}{ac\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \rho(X, Y) \end{aligned}$$

- La desigualdad de Cauchy-Schwarz para las esperanzas de variables aleatorias se utiliza para poder demostrar el rango de la correlación, y esta expresa que si U y V son dos variables aleatorias, entonces $[E(UV)]^2 \leq E(U^2)E(V^2)$
- Si X e Y son dos variables aleatorias, entonces $-1 \leq \rho(X, Y) \leq 1$ cuando esta correlación existe. Esto se puede demostrar a través de la desigualdad de Cauchy-Schwarz

$$U \equiv X - E(X) \text{ \& } V \equiv Y - E(Y)$$

$$\Rightarrow \text{Cov}(X, Y)^2 \leq \text{Var}(X)\text{Var}(Y) \Rightarrow \left| \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \right| \leq 1$$

$$\Rightarrow |\rho(X, Y)| \leq 1 \Rightarrow -1 \leq \rho(X, Y) \leq 1$$

- A través de los valores que puede tomar $\rho(X, Y)$, se puede interpretar lo siguiente:

- Si, y solo si, $\rho(X, Y) = 1$, entonces Y es una función lineal creciente de X
- Si, y solo si, $\rho(X, Y) = -1$, entonces Y es una función lineal decreciente de X
- Si X e Y son independientes, entonces $\rho(X, Y) = 0$. No obstante, lo converso no es verdad, y cuando $\rho(X, Y) = 0$ se dice que las variables no están correlacionadas

- **PROBABILIDAD GEOMÉTRICA**

Commented [MOU3]: Acabar bien

Las distribuciones de probabilidad comunes

- Las distribuciones de probabilidad se usan para modelar poblaciones, por lo que normalmente se lidia con una familia de distribuciones más que con una sola distribución
 - Esta familia está indexada por una o más parámetros, que permiten variar ciertas características de la distribución mientras se mantiene una sola forma funcional
 - Un ejemplo es la distribución normal, de modo que no se especifica precisamente la media y por tanto se lidia con una familia paramétrica de distribuciones normales con media $\mu \in (-\infty, \infty)$
 - Es posible catalogar muchas de las distribuciones de probabilidad más comunes a través de sus características y cantidades más descriptivas
 - Para cada distribución, es posible definir la esperanza y la varianza y otras medidas descriptivas que pueden mejorar el entendimiento
- Una variable aleatoria tiene una distribución discreta si el rango de esta (su espacio muestral) es numerable. En la mayoría de situaciones, una variable aleatoria discreta normalmente toma valores enteros
 - Una variable aleatoria X tiene una distribución uniforme discreta $(1, N)$ si la probabilidad se define de la siguiente manera para un entero N específico:

$$P(X = x | \theta = N) = \frac{1}{N} \quad \text{for } x = 1, 2, \dots, N$$

- Esta distribución pone la misma masa de probabilidad en todos los resultados posibles $1, 2, \dots, N$
- Para poder calcular la media y la varianza de X , se pueden utilizar dos identidades de sumas finitas:

$$\sum_{i=1}^k i = \frac{k(k+1)}{2} \quad \sum_{i=1}^k i^2 = \frac{k(k+1)(2k+1)}{6}$$

- A partir de estas, se puede obtener la esperanza y la varianza para una distribución discreta uniforme de este tipo:

$$E(X) = \sum_{x=1}^N xP(X=x|N) = \frac{1}{N} \sum_{x=1}^N x = \frac{1}{N} \frac{N(N+1)}{2} = \frac{N+1}{2}$$

$$Var(X) = \sum_{x=1}^N (x - E(X))^2 P(X=x|N) =$$

$$= \frac{1}{N} \sum_{x=1}^N x^2 - xE(X) + E(X)^2 =$$

$$= \frac{(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4} + \frac{N+1}{2} = \frac{(N+1)(N-1)}{12}$$

- Esta distribución se puede generalizar de modo que el espacio muestral es cualquier rango de enteros $N_0, N_0 + 1, \dots, N_1$ con la siguiente función de masa de probabilidad:

$$P(X=x|\theta = N_0, N_1) = \frac{1}{N_1 - N_0 + 1}$$

$$\text{for } x = N_0, N_0 + 1, \dots, N_1$$

- Una variable aleatoria X tiene una distribución hipergeométrica (N, M, K) si la probabilidad se define de la siguiente manera:

$$P(X=x|\theta = N, M, K) = \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}} \quad \text{for } x = 0, 1, \dots, K$$

- La distribución hipergeométrica tiene muchas aplicaciones en muestreo para poblaciones finitas, y se suele explicar a través del modelo urna: hay una urna con N bolas idénticas excepto porque hay M bolas rojas y $N - M$ verdes, y se quiere calcular la probabilidad de que, cogiendo K bolas al azar, se obtengan x bolas rojas
- En este caso, el número muestras de tamaño K de bolas a extraerse de las N bolas viene dado por $\binom{N}{K}$, el número total de posibilidades de coger x bolas rojas de las N bolas es $\binom{M}{x}$ y el número total de posibilidades de coger $K - x$ bolas verdes de las N bolas es $\binom{N-M}{K-x}$

- Hay una suposición subyacente sobre el espacio muestral de X , ya que los coeficientes binomiales de la forma $\binom{n}{r}$ se definen solo para $n \geq r$, por lo que el espacio muestral debe estar restringido a este par de desigualdades:

$$M \geq x \quad \text{and} \quad N - M \geq K - x \Rightarrow M - (N - K) \leq x \leq M$$

- En muchos casos, $K < M, N$, de modo que $0 \leq x \leq K$ es apropiado
- La fórmula de la distribución hipergeométrica es bastante difícil de manejar, de modo que para poder obtener la esperanza y la varianza de la distribución se utilizan las siguientes identidades para el binomio de Newton:

$$\binom{M}{x} = \frac{M!}{(M-x)!x!} = \frac{M}{x} \frac{(M-1)!}{(M-x)!(x-1)!} = \frac{M}{x} \binom{M-1}{x-1}$$

$$\binom{M-1}{x} = \frac{(M-1)!}{(M-1-x)!x!} = \frac{M-x}{M} \frac{M!}{(M-x)!x!} = \frac{M-x}{M} \binom{M}{x}$$

$$\begin{aligned} \binom{M}{x-1} &= \frac{M!}{(M-x+1)!(x-1)!} = x(M-x) \frac{M!}{(M-x)!x!} = \\ &= x(M-x) \binom{M}{x} \end{aligned}$$

- Aunque no es trivial, es posible demostrar que la suma de probabilidades es igual a la unidad:

$$\sum_{x=0}^K P(X=x|N, M, K) = \sum_{x=0}^K \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}} = 1$$

- La esperanza de la distribución hipergeométrica se puede obtener a través de las identidades anteriores

$$E(X) = \sum_{x=0}^K x \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}} = \sum_{x=0}^K \frac{M \binom{M-1}{x-1} \binom{N-M}{K-x}}{\frac{N}{K} \binom{N-1}{K-1}} =$$

$$\begin{aligned}
&= \frac{MK}{N} \sum_{x=0}^K \frac{\binom{M-1}{x-1} \binom{N-M}{K-x}}{\frac{N}{K} \binom{N-1}{K-1}} = \\
&= \frac{MK}{N} \sum_{x=0}^K \frac{\binom{M-1}{x-1} \binom{(N-1)-(M-1)}{(K-1)-(x-1)}}{\frac{N}{K} \binom{N-1}{K-1}} = \frac{MK}{N}
\end{aligned}$$

- La varianza de la distribución se puede obtener a través del resultado anterior y aplicando métodos similares a los anteriores

$$Var(X) = E(X^2) - E(X)^2 = \frac{MK}{N} \left[\frac{(N-M)(N-K)}{N(N-1)} \right]$$

- El experimento de Bernoulli es la base para muchas de las distribuciones, y este es un experimento en donde solo hay dos resultados posibles. Por lo tanto, una variable aleatoria X que sigue una distribución de Bernoulli (p) si esta se define de la siguiente manera:

$$X = \begin{cases} 1 & \text{with probab. } p \\ 0 & \text{with probab. } 1-p \end{cases} \quad \text{for } 0 \leq p \leq 1$$

- El valor de $X = 1$ normalmente se conoce como éxito y p se conoce como la probabilidad de éxito, mientras que $X = 0$ se conoce como fallo
- La media y la varianza de una variable que sigue una distribución de Bernoulli son sencillas:

$$E(X) = 1P(X=1) + 0P(X=0) = P(X=1) = p$$

$$Var(X) = (1-p)^2P(X=1) + p^2P(X=0) =$$

$$= (1-p)^2p + p^2(1-p) = p(1-p)$$

- Muchos experimentos se pueden modelar como una secuencia de experimentos de Bernoulli
 - Si n experimentos de Bernoulli idénticos e independientes se realizan, y se definen los eventos $A_i = \{X = 1 \text{ on the } i\text{th trial}\}$ para $i = 1, 2, \dots, n$, entonces es posible derivar una distribución para el número total de éxitos en n experimentos
 - Definiendo la variable Y como el número total de éxitos en n experimentos, el evento $\{Y = y\}$ solo ocurrirá si, de los eventos A_1, A_2, \dots, A_n , solo y son exitosos. A partir de esta idea y de la

independencia entre eventos, es posible derivar la probabilidad para el evento:

$$P\left[\left(\bigcap_{\{X=1\}} A_i\right) \cap \left(\bigcap_{\{X=0\}} A_i\right)\right] = P\left(\bigcap_{\{X=1\}} A_i\right) P\left(\bigcap_{\{X=0\}} A_i\right) = \\ = p p p \dots (1-p)(1-p) \dots = p^y (1-p)^{n-y}$$

- El evento $\{Y = y\}$ ocurrirá sin importar el qué eventos A_i sean exitosos y cuales no (no importa el orden) mientras que haya un número y de éxitos. Por lo tanto, hay diferentes combinaciones de eventos que producen el evento $\{Y = y\}$, y estas se pueden resumir con el binomio de Newton

$$|\{Y = y\}| = \binom{n}{y}$$

- Juntando todo esto, es posible definir una distribución para la suma total de eventos exitosos, llamada distribución binomial
- Una variable aleatoria X tiene una distribución binomial (n, p) si la probabilidad se define de la siguiente manera:

$$P(Y = y | \theta = n, p) = \binom{n}{y} p^y (1-p)^{n-y} \quad \text{for } y = 0, 1, 2, \dots, n$$

- La variable aleatoria Y se puede definir como la suma de variables independientes X_i para toda $i = 1, 2, \dots, n$ que siguen una distribución de Bernoulli idéntica

$$X_i = \begin{cases} 1 & \text{with probab. } p \\ 0 & \text{with probab. } 1-p \end{cases} \quad \text{for } 0 \leq p \leq 1$$

$$\Rightarrow Y = \sum_{i=1}^n X_i$$

- La suma de las probabilidades para todos los valores de Y es igual a uno debido al teorema binomial:

$$\sum_{i=1}^n \binom{n}{i} p^y (1-p)^{n-i} = (p + (1-p))^n \\ \Rightarrow \sum_{i=1}^n \binom{n}{i} p^y (1-p)^{n-i} = 1$$

- La media y la varianza de la distribución binomial se puede obtener a partir de la esperanza y varianza de una variable que sigue una distribución de Bernoulli

$$E(Y) = E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i) = \sum_{i=1}^n p = np$$

$$Var(Y) = Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i) = np(1-p)$$

- Calcular las probabilidades de una binomial se puede hacer rápidamente, por tanto, a partir de la siguiente relación recursiva:

$$\begin{aligned} P(Y = y) &= \binom{n}{y} p^y (1-p)^{n-y} = \\ &= \frac{n-y+1}{y} \frac{p}{1-p} \binom{n}{y-1} p^{y-1} (1-p)^{n-y+1} \\ &= \frac{n-y+1}{y} \frac{p}{1-p} P(Y = y-1) \end{aligned}$$

- La distribución de Poisson es una distribución que permite modelar varios tipos de experimentos, tales como el número de ocurrencias en un intervalo de tiempo dado. Esta distribución tiene un solo parámetro λ , el cual se denomina parámetro de intensidad, y se dice que una variable aleatoria X que tome valores no negativos sigue una distribución de Poisson (λ) si la probabilidad se define de la siguiente manera:

$$P(X = x | \theta = \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{for } x = 0, 1, 2, \dots$$

- Una de las suposiciones básicas de este tipo de distribución de probabilidad es que la probabilidad de ocurrencia es proporcional a la longitud del intervalo de tiempo siempre que estos sean intervalos pequeños
- Esta función de masa de probabilidad suma 1 para todos los valores de X debido a la definición de e^x como serie de potencias

$$\sum_{x=0}^{\infty} P(X = x | \lambda) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1$$

- La media y la varianza de X se puede calcular fácilmente utilizando unos cálculos similares:

$$E(X) = \sum_{x=0}^{\infty} xP(X=x|\lambda) = e^{-\lambda} \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} = e^{-\lambda} \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$$= e^{-\lambda} \lambda e^{\lambda} = \lambda$$

$$Var(X) = \sum_{x=0}^{\infty} (x-\lambda)^2 P(X=x|\lambda) = \lambda$$

- Calcular las probabilidades de una Poisson se puede hacer rápidamente, por tanto, a partir de la siguiente relación recursiva:

$$P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!} = \frac{\lambda e^{-\lambda} \lambda^{x-1}}{x(x-1)!} = \frac{\lambda}{x} P(X=x-1)$$

- Las relaciones recursivas para la Poisson y para la binomial se pueden relacionar a través de escoger $\lambda = np$ y considerar que p es pequeña, de modo que esto servirá para poder aproximar las probabilidades de una binomial con una Poisson y viceversa

$$\frac{n-y+1}{y} \frac{p}{1-p} = \frac{np-p(y-1)}{y-py} \approx \frac{\lambda}{y}$$

- Esto ocurre porque los términos multiplicados por p pueden ser ignorados si p es pequeña
- Para poder aproximar probabilidades solo hace falta establecer que $P(Y=0) \approx P(X=0)$, dado que todas las otras posibilidades se pueden obtener con la recursividad anterior. Esto se puede demostrar a través del límite de la probabilidad para $\{Y=0\}$

$$P(Y=0) = (1-p)^n = \left(1 - \frac{np}{n}\right)^n = \left(1 - \frac{\lambda}{n}\right)^n$$

$$\Rightarrow \lim_{n \rightarrow \infty} P(Y=0) = \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda} = P(X=0)$$

- La aproximación de probabilidades solo sirve cuando n es grande y p es pequeña, el cual es el caso más útil porque eso ahorra el cálculo de coeficientes binomiales muy grandes

- La distribución binomial cuenta el número de éxitos en un número fijo de experimentos de Bernoulli, pero para poder contar el número de intentos requeridos para obtener un número x de éxitos, se utiliza la distribución binomial negativa y se dice que una variable aleatoria X sigue una distribución binomial negativa (r, p) si la probabilidad se define de la siguiente manera:

$$P(X = x | \theta = r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \text{ for } x = r, r+1, \dots$$

- En este caso, la variable X denota el experimento en el que ha habido el éxito número r
- La derivación de esta distribución proviene de la distribución binomial. El evento $\{X = x\}$ puede ocurrir solo si hay exactamente $r - 1$ éxitos en los primeros $x - 1$ experimentos, y un éxito en el experimento x
- La probabilidad de que $r - 1$ éxitos en $x - 1$ experimentos es la probabilidad binomial $\binom{x-1}{r-1} p^{r-1} (1-p)^{x-r}$ y hay un éxito con probabilidad p en el experimento x , por lo que la distribución binomial negativa se obtiene multiplicando ambas probabilidades
- La binomial negativa a veces se define en términos de una variable aleatoria Y que cuenta el número de fallos hasta el éxito número r , esto es equivalente a $Y = X - r$, por lo que se puede utilizar la siguiente formulación:

$$P(Y = y | r, p) = \binom{r+y-1}{y} p^r (1-p)^y \text{ for } y = 0, 1, \dots$$

- Normalmente se utilizará esta formulación cuando uno se refiere a la binomial negativa. Esta recibe su nombre de la siguiente relación:

$$\binom{r+y-1}{y} = (-1)^y \binom{-r}{y}$$

$$\Rightarrow P(Y = y | r, p) = (-1)^y \binom{-r}{y} p^r (1-p)^y$$

- El hecho de que la suma de probabilidades para todos los valores de Y es uno no es trivial, pero proviene de una extensión del teorema binomial (que incluye exponentes negativos)

$$\sum_{y=0}^{\infty} P(Y=y) = \sum_{y=0}^{\infty} \binom{r+y-1}{y} p^r (1-p)^y = 1$$

- La media y la varianza de Y se pueden calcular de manera similar a las de la distribución binomial:

$$\begin{aligned} E(Y) &= \sum_{y=0}^{\infty} y \binom{r+y-1}{y} p^r (1-p)^y = \\ &= r \sum_{y=0}^{\infty} \binom{r+y-1}{y-1} p^r (1-p)^y = \\ &= r \sum_{z=1}^{\infty} \binom{r+z}{z} p^r (1-p)^{z+1} = \\ &= \frac{r(1-p)}{p} \sum_{z=1}^{\infty} \binom{(r+1)+z-1}{z} p^{r+1} (1-p)^z = \frac{r(1-p)}{p} \\ \text{Var}(Y) &= \sum_{y=0}^{\infty} (y - E(Y))^2 \binom{r+y-1}{y} p^r (1-p)^y = \frac{r(1-p)}{p^2} \end{aligned}$$

- La familia de distribuciones de binomial negativa incluye la distribución de Poisson es un caso límite. Si $r \rightarrow \infty$ y $p \rightarrow 1$ tal que $r(1-p) \rightarrow \lambda$ para $0 < \lambda < \infty$, entonces se cumplen los siguientes límites:

$$E(Y) = \frac{r(1-p)}{p} \rightarrow \lambda \text{ as } r \rightarrow \infty \text{ \& } p \rightarrow 1$$

$$\text{Var}(Y) = \frac{r(1-p)}{p^2} \rightarrow \lambda \text{ as } r \rightarrow \infty \text{ \& } p \rightarrow 1$$

- La distribución geométrica es la más simple de las distribuciones para modelar el tiempo de espera y es un caso especial de la binomial negativa en el que $r = 1$. Se dice que una variable aleatoria X sigue una distribución geométrica (p) si la probabilidad se define de la siguiente manera:

$$P(X=x|p) = p(1-p)^{x-1} \text{ for } x = 1, 2, \dots$$

- A partir de resultados de las series de potencias, es posible mostrar como la suma de probabilidades para todos los valores de X es igual a 1

$$\begin{aligned}\sum_{x=1}^{\infty} P(X=x) &= \sum_{x=1}^{\infty} p(1-p)^{x-1} = p \sum_{x=1}^{\infty} (1-p)^{x-1} = \\ &= \frac{p}{1-(1-p)} = \frac{p}{p} = 1\end{aligned}$$

- La media y la varianza de X puede ser calculada usando fórmulas de la binomial negativa y usando la equivalencia $X = Y + 1$

$$E(X) = E(Y) + 1 = \frac{1-p}{p} + 1 = \frac{1}{p}$$

$$Var(X) = Var(Y) = \frac{1-p}{p^2}$$

- La distribución geométrica cumple con la propiedad de pérdida de memoria, de modo que para números enteros $s > t$ se cumple que $P(X > s | X > t) = P(X > s - t)$ y el proceso no tiene en cuenta lo que ha ocurrido anteriormente (la probabilidad de tener $s - t$ fallos es la misma independientemente de cuantos fallos se han cometido)

$$P(X > n) = (1-p)^n$$

$$\Rightarrow P(X > s | X > t) = \frac{P(\{X > s\} \cap \{X > t\})}{P(X > t)} =$$

$$= \frac{P(X > s)}{P(X > t)} = \frac{(1-p)^s}{(1-p)^t} = (1-p)^{s-t} = P(X > s-t)$$

- Una variable aleatoria tiene una distribución continua si el rango de esta (su espacio muestral) es infinito
 - La distribución uniforme continua se define por ...

Las familias de distribuciones

Commented [MOU4]: Acabar ya

Las distribuciones multivariadas discretas e independencia

- Siendo X e Y variables aleatorias discretas en el espacio de probabilidad (Ω, \mathcal{F}, P) , hay veces que es necesario tratar X e Y como un vector aleatorio (X, Y) que toma valores en \mathbb{R}^2 . Por lo tanto, se puede introducir la noción de distribuciones bivariantes y la esperanza en el caso multivariante

- Si X e Y son variables aleatorias discretas en (Ω, \mathcal{F}, P) , la función de masa de probabilidad p_{XY} de X e Y es la función $p_{XY}: \mathbb{R}^2 \rightarrow [0,1]$ definida por la siguiente fórmula:

$$p_{XY}(x, y) = P(X = x, Y = y) = \\ = P(\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\})$$

- La imagen de X es numerable para cualquier variable aleatoria discreta X (dado que es el intervalo $[0,1]$) y se cumplen las siguientes dos condiciones:

$$p_{XY}(x, y) = 0 \text{ unless } x \in \text{Im } X \text{ and } y \in \text{Im } Y$$

$$\sum_{x \in \text{Im } X} \sum_{y \in \text{Im } Y} p_{XY}(x, y) = \\ = P\left(\bigcup_{x \in \text{Im } X} \{\omega \in \Omega : X(\omega) = x, Y(\omega) = y\}\right) = P(\Omega) = 1$$

- Las funciones de masa de probabilidad p_X y p_Y de X e Y se pueden encontrar a partir de la función de masa conjunta, las cuales se denominan funciones de masa de probabilidad marginales

$$p_X(x) = P(X = x) = \sum_{y \in \text{Im } Y} p_{XY}(x, y) = \sum_y p_{XY}(x, y)$$

$$p_Y(y) = P(Y = y) = \sum_{x \in \text{Im } X} p_{XY}(x, y) = \sum_x p_{XY}(x, y)$$

- Si se piensa en (X, Y) es un punto escogido aleatoriamente en el plano, entonces X e Y son las proyecciones de este punto en los ejes de coordenadas

	$x = 1$	$x = 2$	$x = 3$
$y = 1$	$\frac{1}{12}$	$\frac{3}{18}$	$\frac{1}{6}$
$y = 2$	$\frac{1}{18}$	0	$\frac{5}{18}$
$y = 3$	0	$\frac{3}{18}$	$\frac{1}{12}$

- Si X e Y son variables aleatorias discretas en (Ω, \mathcal{F}, P) y $g: \mathbb{R}^2 \rightarrow \mathbb{R}$ es una función, entonces $Z \equiv g(X, Y)$ es también una variable aleatoria discreta en (Ω, \mathcal{F}, P) , definida formalmente por $Z(\omega) = g(X(\omega), Y(\omega))$

para $\omega \in \Omega$. Por lo tanto, su esperanza se puede calcular directamente de la función de masa de probabilidad conjunta

- Cuando la suma converge absolutamente, la esperanza de una variable $Z \equiv g(X, Y)$ se puede calcular con la siguiente fórmula:

$$E(g(X, Y)) = \sum_{x \in \text{Im } X} \sum_{y \in \text{Im } Y} g(x, y) P(X = x, Y = y)$$

- La demostración de este teorema es análoga a la de la esperanza de $g(X)$

$$\begin{aligned} E(Z) &= E(g(X, Y)) = \sum_{z \in \text{Im } Z} z P(Z = z) = \\ &= \sum_{z \in \text{Im } Z} z \sum_{x \in \text{Im } X: g(x, y) = z} \sum_{y \in \text{Im } Y: g(x, y) = z} P(g(X, Y) = z) = \\ &= \sum_{z \in \text{Im } Z} z \sum_{x \in \text{Im } X} \sum_{y \in \text{Im } Y} P(X \in g^{-1}(x, y), Y \in g^{-1}(x, y)) = \\ &= \sum_{x \in \text{Im } X} \sum_{y \in \text{Im } Y} g(x, y) \sum_{x \in \text{Im } X} \sum_{y \in \text{Im } Y} P(X = x, Y = y) = \\ &= \sum_{x \in \text{Im } X} \sum_{y \in \text{Im } Y} g(x, y) P(X = x, Y = y) \end{aligned}$$

- Una de las propiedades más importantes de la esperanza multivariante es la linealidad de la esperanza, por lo que si X e Y son variables aleatorias discretas en (Ω, \mathcal{F}, P) y $a, b \in \mathbb{R}$, entonces se cumple la siguiente igualdad:

$$\begin{aligned} E(aX + bY) &= \sum_{x \in \text{Im } X} \sum_{y \in \text{Im } Y} (ax + by) P(X = x, Y = y) = \\ &= \sum_{x \in \text{Im } X} \sum_{y \in \text{Im } Y} ax P(X = x, Y = y) + by P(X = x, Y = y) = \\ &= a \sum_{x \in \text{Im } X} x \sum_{y \in \text{Im } Y} P(X = x, Y = y) \\ &\quad + b \sum_{y \in \text{Im } Y} y \sum_{x \in \text{Im } X} P(X = x, Y = y) = \end{aligned}$$

$$\begin{aligned}
&= a \sum_{x \in \text{Im } X} xP(X=x) + b \sum_{y \in \text{Im } Y} yP(Y=y) = \\
&= aE(X) + bE(Y) \\
&\Rightarrow E(aX + bY) = aE(X) + bE(Y)
\end{aligned}$$

- En un espacio de probabilidad (Ω, \mathcal{F}, P) , los eventos A y B se denominan independientes si $P(A \cap B) = P(A)P(B)$, pero las variables aleatorias discretas X e Y en (Ω, \mathcal{F}, P) se denominan independientes si el valor tomado por X es independiente del valor tomado por Y

- Dos variables aleatorias discretas X e Y son independientes si el par de eventos $\{X=x\}$ y $\{Y=y\}$ son independientes para toda $x, y \in \mathbb{R}$. De otro modo, las variables son dependientes

$$P(X=x, Y=y) = P(X=x)P(Y=y) \text{ for } x, y \in \mathbb{R}$$

- Esta identidad se puede expresar de manera equivalente en términos de la función de masa de probabilidad:

$$p_{XY}(x, y) = \left(\sum_{y \in \text{Im } Y} p_{XY}(x, y) \right) \left(\sum_{x \in \text{Im } X} p_{XY}(x, y) \right) \text{ for } x, y \in \mathbb{R}$$

- Las variables aleatorias discretas X e Y son independientes si, y solo si, existen funciones $f, g: \mathbb{R} \rightarrow \mathbb{R}$ tal que la función de masa de probabilidad conjunta de las variables satisface la siguiente igualdad:

$$p_{XY}(x, y) = f(x)g(y) \text{ for } x, y \in \mathbb{R}$$

- La demostración de este último teorema se basa en suponer que la proposición es real y obtener las funciones de masa de probabilidad para cada variable:

$$\begin{aligned}
p_{XY}(x, y) = f(x)g(y) &\Rightarrow \begin{cases} p_X(x) = \sum_{y \in \text{Im } Y} p_{XY}(x, y) \\ p_Y(y) = \sum_{x \in \text{Im } X} p_{XY}(x, y) \end{cases} \\
&\Rightarrow \begin{cases} p_X(x) = f(x) \sum_{y \in \text{Im } Y} g(y) \\ p_Y(y) = g(y) \sum_{x \in \text{Im } X} f(x) \end{cases}
\end{aligned}$$

$$1 = \sum_{x \in \text{Im } X} \sum_{y \in \text{Im } Y} p_{XY}(x, y) = \sum_{x \in \text{Im } X} f(x) \sum_{y \in \text{Im } Y} g(y)$$

$$\Rightarrow p_{XY}(x, y) = f(x)g(y) \sum_{x \in \text{Im } X} f(x) \sum_{y \in \text{Im } Y} g(y) = p_X(x)p_Y(y)$$

- Si dos variables aleatorias discretas X e Y son independientes y tienen esperanza $E(X)$ y $E(Y)$ entonces se cumple la siguiente igualdad:

$$E(XY) = E(X)E(Y)$$

- La demostración de este teorema se basa en la definición de independencia anterior y en la de esperanza

$$E(XY) = \sum_{x \in \text{Im } X} \sum_{y \in \text{Im } Y} xyP(X = x, Y = y) =$$

$$= \sum_{x \in \text{Im } X} \sum_{y \in \text{Im } Y} xyP(X = x)P(Y = y) =$$

$$= \sum_{x \in \text{Im } X} xP(X = x) \sum_{y \in \text{Im } Y} yP(Y = y) = E(X)E(Y)$$

- Lo converso a este teorema es falso, de modo que si $E(XY) = E(X)E(Y)$ no quiere decir que X e Y sean independientes. No obstante, el converso correcto corresponde al siguiente teorema
- Dos variables aleatorias discretas X e Y en (Ω, \mathcal{F}, P) son independientes si, y solo si, se cumple la siguiente igualdad para todas las funciones $g, h: \mathbb{R} \rightarrow \mathbb{R}$ para las que el valor esperado exista:

$$E(g(X)h(Y)) = E(g(X))E(h(Y))$$

- La demostración de este teorema se basa en la definición de independencia anterior y en la de esperanza

$$E(g(X)h(Y)) =$$

$$= \sum_{g(x) \in \text{Im } g(X)} \sum_{h(y) \in \text{Im } h(Y)} g(x)h(y)P(g(X) = g(x), h(Y) = h(y)) =$$

$$= \sum_{g(x) \in \text{Im } g(X)} g(x)P(g(X) = g(x)) \sum_{h(y) \in \text{Im } h(Y)} h(y)P(h(Y) = h(y)) =$$

$$= E(g(X))E(h(Y))$$

- Mucha de la teoría de la probabilidad se concentra en la suma de variables aleatorias, en donde se intenta averiguar la función de masa de probabilidad de una variable aleatoria discreta $Z = X + Y$. Además, también es útil estudiar las funciones indicador debido a que son muy prácticas a la hora de trabajar con problemas de probabilidad

- Si X e Y son variables aleatorias discretas independientes en (Ω, \mathcal{F}, P) , entonces $Z = X + Y$ tiene la siguiente función de masa de probabilidad:

$$P(Z = z) = \sum_{x \in \text{Im } X} P(X = x, Y = z - x) \quad \text{for } z \in \mathbb{R}$$

- Obviamente, Z debe tomar el valor z si, y solo si, $X = x$ y $Y = z - x$ para algún valor de x (debido a que $z = x + (z - x)$), de modo que se cumple la siguiente identidad:

$$\begin{aligned} P(Z = z) &= P\left(\bigcup_x (\{X = x\} \cap \{Y = z - x\})\right) = \\ &= \sum_{x \in \text{Im } X} P(X = x, Y = z - x) \quad \text{for } z \in \mathbb{R} \end{aligned}$$

- Si las variables son independientes, entonces:

$$\begin{aligned} P(Z = z) &= P\left(\bigcup_x (\{X = x\} \cap \{Y = z - x\})\right) = \\ &= \sum_{x \in \text{Im } X} P(X = x, Y = z - x) = \\ &= \sum_{x \in \text{Im } X} P(X = x)P(Y = z - x) \quad \text{for } z \in \mathbb{R} \end{aligned}$$

- En el lenguaje del análisis matemático, la función de masa de probabilidad anterior se denomina convolución de $X + Y$ (cuando se multiplican dos funciones de masa de probabilidad, en este caso preciso)
- La función indicadora indica si un evento A ocurre o no. La función indicadora de un evento A es la variable aleatoria denotada por 1_A y definida por la siguiente fórmula:

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

- La función indicadora es una variable aleatoria discreta cuyo valor esperado es el siguiente:

$$E(1_A) = P(A)$$

- La función indicadora tiene dos propiedades básicas, las cuales son las siguientes:

$$1_{A \cap B} = 1_A 1_B \quad 1_A + 1_{A^c} = P(A) + P(A^c) = 1$$

- La función indicadora proporciona una herramienta muy útil a la hora de calcular las probabilidades y valores esperados

- Por ejemplo, se puede calcular $E(A \cup B) = 1_{A \cup B}$

$$\begin{aligned} 1_{A \cup B} &= 1 - 1_{A^c \cap B^c} = 1_{A \cup B} = 1 - 1_{A^c} 1_{B^c} = \\ &= 1 - (1 - 1_A)(1 - 1_B) = 1_A + 1_B - 1_{A \cap B} \\ &\Rightarrow E(A \cup B) = E(A) + E(B) - E(A \cap B) \end{aligned}$$

- Siendo A_1, A_2, \dots, A_n eventos y 1_{A_i} una función indicador de A_i , la probabilidad de la unión de los eventos, llamada función de inclusión-exclusión, es la siguiente:

$$\begin{aligned} A \equiv A_1 \cup A_2 \cup \dots \cup A_n &\Rightarrow 1_A = 1 - \prod_{i=1}^n (1 - 1_{A_i}) \\ \Rightarrow 1_A &= \sum_i 1_{A_i} - \sum_{i < j} 1_{A_i} 1_{A_j} + \dots + (-1)^{n+1} \left(\bigcap_i A_i \right) \\ \Rightarrow E(1_A) &= \sum_i E(1_{A_i}) - \sum_{i < j} E(1_{A_i} 1_{A_j}) + \dots + (-1)^{n+1} E \left(\bigcap_i A_i \right) \\ \Rightarrow P(1_A) &= \sum_i P(1_{A_i}) - \sum_{i < j} P(1_{A_i} 1_{A_j}) + \dots + (-1)^{n+1} P \left(\bigcap_i A_i \right) \end{aligned}$$

Las distribuciones multivariadas generales e independencia

- Dadas dos variables aleatorias X e Y , actuando sobre un espacio de probabilidad (Ω, \mathcal{F}, P) , normalmente es útil pensar en ellas como un vector aleatorio (X, Y) tomando valores en \mathbb{R}^2 , de modo que se puede introducir el concepto de función de distribución y de densidad conjunta

Commented [MOU5]: Acabar bien capítulo Papoulis 7 Sequences & Intermediate Course Chapter 1 y 2 &

- Si X e Y son variables aleatorias discretas, normalmente se estudia este vector aleatorio usando la función de masa de probabilidad conjunta de X e Y , pero este método no está siempre disponible

- En el caso general de variables aleatorias arbitrarias X e Y , se estudia la función de distribución conjunta

- La función de distribución de probabilidad conjunta de un par X e Y de variables aleatorias es el mapeado $F_{XY}: \mathbb{R}^2 \rightarrow [0,1]$ definido de la siguiente manera:

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

- Las funciones de distribución conjuntas cumplen una serie de propiedades elementales análogas a las vistas anteriormente

$$\lim_{x, y \rightarrow \infty} F_{XY}(x, y) = 0 \quad \lim_{x, y \rightarrow \infty} F_{XY}(x, y) = 0$$

$$F_{XY}(x_1, y_1) \leq F_{XY}(x_2, y_2) \text{ if } x_1 \leq x_2 \text{ \& } y_1 \leq y_2$$

- La función de distribución conjunta F_{XY} contiene mucha información que las funciones de distribución F_X y F_Y , dado que permiten saber como se comportan X e Y juntas

- En particular, es posible encontrar las funciones de distribución de X e Y a partir de la función de distribución conjunta, las cuales se denominan funciones de distribución marginal de la distribución conjunta

$$F_X(x) = \lim_{y \rightarrow \infty} F_{XY}(x, y) \quad F_Y(y) = \lim_{x \rightarrow \infty} F_{XY}(x, y)$$

- La idea de independencia para variables aleatorias se puede definir de la siguiente manera: se dice que X e Y son independientes si, para toda $x, y \in \mathbb{R}$, los eventos $\{X \leq x\}$ y $\{Y \leq y\}$ son independientes. De otro modo, las variables son dependientes

- Por lo tanto, X e Y son independientes si, y solo si, se cumplen las siguientes identidades:

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \text{ for } x, y \in \mathbb{R}$$

$$F_{XY}(x, y) = F_X(x)F_Y(y) \text{ for } x, y \in \mathbb{R}$$

- Las familias de variables aleatorias también se estudian de la misma manera: si $\mathbf{X} = (X_1, X_2, \dots, X_n)$ es un vector de variables aleatorias en

(Ω, \mathcal{F}, P) , su función de distribución de probabilidad conjunta es la función $F_X: \mathbb{R}^n \rightarrow [0,1]$ definida de la siguiente manera:

$$F_X(\mathbf{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$$

$$\text{for } \mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$$

- El vector de variables $\mathbf{X} = (X_1, X_2, \dots, X_n)$ es independiente si se cumplen las siguientes identidades:

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = P(X_1 \leq x_1) \dots P(X_n \leq x_n)$$

$$\text{for } \mathbf{x} \in \mathbb{R}^n$$

$$F_X(\mathbf{x}) = F_{X_1}(x_1) \dots F_{X_n}(x_n) \quad \text{for } \mathbf{x} \in \mathbb{R}^n$$

- A partir de la función de distribución conjunta y de sus propiedades, es posible definir el concepto de función de densidad de probabilidad conjunta
 - Un par de variables aleatorias X e Y en (Ω, \mathcal{F}, P) es conjuntamente continuo si su función de distribución de probabilidad conjunta se puede expresar de la siguiente manera para $x, y \in \mathbb{R}$ y $f: \mathbb{R}^2 \rightarrow [0, \infty)$:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) = \int_{u=-\infty}^x \int_{v=-\infty}^y f(u, v) du dv$$

- Si esto se mantiene, se dice que X e Y tienen una función de densidad de probabilidad conjunta, y normalmente se denota esta función por f_{XY} . Esta se puede definir de la siguiente manera (aunque falte un poco de rigor matemático como en los casos anteriores):

$$f_{XY} = \begin{cases} \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y) & \text{if this deriv. exists at } (x, y) \\ 0 & \text{otherwise} \end{cases}$$

- Las propiedades elementales de la función conjunta densidad f_{XY} son las siguientes:

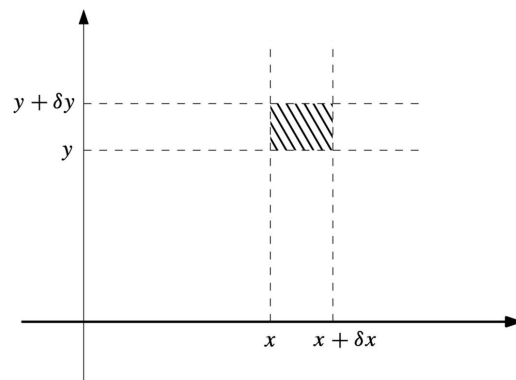
$$f_{XY}(x, y) \geq 0 \quad \text{for } x, y \in \mathbb{R}$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1$$

- Del mismo modo que antes, se puede establecer una analogía con la función de masa conjunta y la función de densidad

conjunta, de modo que para cualquier $(x, y) \in \mathbb{R}^2$ y unas δx y δy pequeñas, la probabilidad de que un vector aleatorio (X, Y) esté en un rectángulo en donde la esquina inferior izquierda sea en (x, y) y largos δx y δy es la siguiente:

$$P(x < X \leq x + \delta x, y < Y \leq y + \delta y) \approx f_{XY}(x, y) \delta x \delta y$$



- Si A es un subconjunto regular de \mathbb{R}^2 y X e Y son dos variables aleatorias continuas conjuntas con función de densidad conjunta f_{XY} , entonces se cumple la siguiente identidad:

$$P((X, Y) \in A) = \iint_{(x, y) \in A} f_{XY}(x, y) \, dx \, dy$$

- Debido a que el resultado se debe más a un resultado de cálculo integral que de probabilidad, se omite la demostración de este teorema. No obstante, la explicación intuitiva es que si se hace una partición de A , entonces $f_{XY}(x, y) \delta x \delta y$ es aproximadamente la probabilidad de que (X, Y) tome un valor en una de los conjuntos de la partición y eso hace que la suma de todas estas pequeñas probabilidades sea la probabilidad de que se tome un valor dentro de f_{XY}
- El término conjunto regular está relacionado con los conjuntos como rectángulos, discos, regiones acotadas por curvas de Jordan acotadas, y otros
- Cuando un par (X, Y) tiene una función de densidad conjunta f_{XY} , las funciones de distribución ordinarias de X e Y se puede obtener de manera inmediata (en puntos de diferenciabilidad) utilizando la primera parte del teorema fundamental del cálculo y el teorema anterior:

$$f_X(x) = \frac{d}{dx} P(X \leq x) = \frac{d}{dx} \int_{u=-\infty}^x \int_{v=-\infty}^{\infty} f_{XY}(u, v) du dv =$$

$$= \int_{u=-\infty}^x \frac{d}{dx} \int_{v=-\infty}^{\infty} f_{XY}(u, v) du dv = \int_{v=-\infty}^{\infty} f_{XY}(x, v) dv$$

$$f_Y(y) = \frac{d}{dy} P(Y \leq y) = \frac{d}{dy} \int_{u=-\infty}^{\infty} \int_{v=-\infty}^y f_{XY}(u, v) du dv =$$

$$= \int_{u=-\infty}^{\infty} \int_{v=-\infty}^y \frac{d}{dy} f_{XY}(u, v) du dv = \int_{u=-\infty}^{\infty} f_{XY}(u, y) du$$

- En este caso, el límite de integración superior cambia a ∞ para la otra variable que no es de interés (dado que la variable de interés está restringida por el valor x o y)
 - Estas funciones de densidad se llaman funciones de densidad marginales de X e Y , dado que proyectan el vector aleatorio (X, Y) en los dos ejes de coordenadas del plano
- Dos variables aleatorias continuas conjuntas X e Y son independientes si, y solo si, su función de densidad conjunta tiene la siguiente forma:

$$f_{XY}(x, y) = g(x)h(y) \text{ for } x, y \in \mathbb{R}$$

- La demostración de este teorema es muy parecida al caso en las que las variables aleatorias son discretas

$$f_{XY}(x, y) = g(x)h(y) \Rightarrow \begin{cases} f_X(x) = \int_{v=-\infty}^{\infty} f_{XY}(x, v) dv \\ f_Y(y) = \int_{u=-\infty}^{\infty} f_{XY}(u, y) dv \end{cases}$$

$$\Rightarrow \begin{cases} f_X(x) = g(x) \int_{v=-\infty}^{\infty} h(v) dv \\ f_Y(y) = h(y) \int_{u=-\infty}^{\infty} g(u) du \end{cases}$$

$$1 = \int_{u=-\infty}^{\infty} \int_{v=-\infty}^{\infty} f_{XY}(x, v) dv = \int_{u=-\infty}^{\infty} g(u) du \int_{v=-\infty}^{\infty} h(v) dv$$

$$\Rightarrow f_{XY}(x, y) = f_X(x)f_Y(y) =$$

$$= h(y) \int_{u=-\infty}^{\infty} g(u) du g(x) \int_{v=-\infty}^{\infty} h(v) dv = g(x)h(y)$$

- Además, este teorema también permite ver como este resultado solo se mantiene si, y solo si, $F_{XY}(x, y) = F_X(x)F_Y(y)$. Esto se puede demostrar a partir de la definición de función de distribución conjunta:

$$\begin{aligned} F_{XY}(x, y) &= \frac{\partial^2}{\partial x \partial y} f_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} f_X(x)f_Y(y) = \\ &= \frac{\partial}{\partial x} f_X(x) \frac{\partial}{\partial y} f_Y(y) = F_X(x)F_Y(y) \end{aligned}$$

- Todos estos resultados también aplican al caso en el que hay más de dos variables, solo haciendo falta adaptarlos a casos multivariantes (o de un vector de varias variables aleatorias)
- Teniendo nociones sobre la función de distribución y de densidad conjunta, es posible estudiar las funciones de densidad condicionales y el valor esperado de variables continuas
 - Suponiendo que X e Y son dos variables aleatorias continuas conjuntas, una situación en donde solo interesa Y y no se tuviera información sobre el valor de X , solo necesitaría la función de densidad marginal de Y
 - No obstante, cuando se tiene información completa sobre los valores que toma X , esta información tiene un efecto en la distribución, haciendo que esta sea la distribución condicional de Y dado $X = x$
 - No es posible calcular la probabilidad $P(Y \leq y | X = x)$ con la fórmula de la probabilidad condicional, dado que para una variable continua $P(X = x) = 0$
 - En vez de condicionar el evento a $X = x$, se condiciona a $x \leq X \leq x + \delta x$ y se toma el límite a la derecha $\delta x \rightarrow 0$

$$\begin{aligned} P(Y \leq y | x \leq X \leq x + \delta x) &= \frac{P(Y \leq y, x \leq X \leq x + \delta x)}{P(x \leq X \leq x + \delta x)} = \\ &= \frac{\int_x^{x+\delta x} \int_{v=-\infty}^y f_{XY}(u, v) du dv}{\int_x^{x+\delta x} f_X(u) du} = \frac{\int_x^{x+\delta x} \int_{v=-\infty}^y f_{XY}(u, v) du dv}{\int_x^{x+\delta x} f_X(u) du} \frac{\delta x}{\delta x} \end{aligned}$$

$$\begin{aligned}
&\Rightarrow \lim_{\delta x \rightarrow 0^+} P(Y \leq y | x \leq X \leq x + \delta x) = \\
&= \lim_{\delta x \rightarrow 0^+} \frac{\frac{\int_x^{x+\delta x} \int_{v=-\infty}^y f_{XY}(u, v) du dv}{\delta x}}{\frac{\int_x^{x+\delta x} f_X(u) du}{\delta x}} = \frac{\infty}{\infty} \\
&\Rightarrow \lim_{\delta x \rightarrow 0^+} \frac{\int_{v=-\infty}^y f_{XY}(x, v) dv}{f_X(x)} = \lim_{\delta x \rightarrow 0^+} \int_{v=-\infty}^y \frac{f_{XY}(x, v)}{f_X(x)} dv = G(y) \\
&\Rightarrow g(y) = \frac{\partial}{\partial y} G(y) = \frac{\partial}{\partial y} \int_{v=-\infty}^y \frac{f_{XY}(x, v)}{f_X(x)} dv = \frac{f_{XY}(x, y)}{f_X(x)} \text{ for } y \in \mathbb{R}
\end{aligned}$$

- De este modo, se puede ver que $G(y)$ es la función de distribución condicional de Y y que $g(y)$ es la función de densidad condicional de Y (siempre que $f_X(x) > 0$)
- La función de densidad condicional de Y dado $X = x$ se denota por $f_{Y|X}(\cdot | x)$ y se define de la siguiente manera para toda $f_X(x) > 0$:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} \text{ for } y \in \mathbb{R}$$

- La expresión $P(Y \leq y | X = x)$ no se puede interpretar de la manera usual (con la fórmula), sino que se interpreta como la función de distribución $G(y)$ de Y dado $X = x$
- Si X e Y son independientes, entonces $f_{XY}(x, y) = f_X(x)f_Y(y)$ y eso hace que $f_{Y|X}(y|x) = f_Y(y)$ y que los valores de X sean irrelevantes para estudiar el comportamiento de Y
- Siendo X e Y variables aleatorias continuas en (Ω, \mathcal{F}, P) y $g: \mathbb{R}^2 \rightarrow \mathbb{R}$ es una función con buen comportamiento, se puede suponer que el mapeado $Z: \Omega \rightarrow \mathbb{R}$ definido por $Z(\omega) \equiv g(X(\omega), Y(\omega))$ es una variable aleatoria y no es necesario saber la distribución explícitamente para obtener su valor esperado
- Si la integral converge absolutamente, la esperanza de $g(X, Y)$ viene dada por la siguiente expresión:

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

- Usando esta definición, se puede ver que la esperanza también actúa de manera lineal sobre el espacio de variables aleatorias continuas

$$\begin{aligned}
 E(aX + bY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (ax + by) f_{XY}(x, y) \, dx \, dy = \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ax f_{XY}(x, y) + by f_{XY}(x, y) \, dx \, dy = \\
 &= a \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{XY}(x, y) \, dx \, dy + b \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{XY}(x, y) \, dx \, dy = \\
 &= a \int_{-\infty}^{\infty} x f_X(x) \, dx + b \int_{-\infty}^{\infty} y f_Y(y) \, dy = aE(X) + bE(Y)
 \end{aligned}$$

- A partir del resultado anterior, se puede ver como no es necesario que X e Y sean independientes para que la esperanza de su suma sea la suma de sus esperanzas individuales
- Las variables aleatorias continuas conjuntas X e Y son independientes si, y solo si, se da la siguiente identidad para todas las funciones $g, h: \mathbb{R} \rightarrow \mathbb{R}$ para las cuales sus valores esperados existan:

$$E(g(X)h(Y)) = E(g(X))E(h(Y))$$

- Si X e Y son independientes, entonces $f_{XY}(x, y) = f_X(x)f_Y(y)$ y se obtiene la siguiente igualdad:

$$\begin{aligned}
 E(g(X)h(Y)) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y) f_{XY}(x, y) \, dx \, dy = \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)f_X(x)h(y)f_Y(y) \, dx \, dy = \\
 &= \int_{-\infty}^{\infty} g(x)f_X(x) \, dx \int_{-\infty}^{\infty} h(y)f_Y(y) \, dy = E(g(X))E(h(Y))
 \end{aligned}$$

- De manera converso, si la igualdad propuesta se mantiene, entonces también se mantiene para las funciones indicador para $u \leq x$ y $v \leq y$. Por lo tanto, $g(X)h(Y)$ es una variable aleatoria discreta que sigue una distribución de Bernoulli con parámetro $p_1 = P(X \leq x, Y \leq y)$ y $g(X)$ y $h(Y)$ son variables con parámetros $p_2 = P(X \leq x)$ y $p_3 = P(Y \leq y)$ respectivamente, por lo que se pueden obtener las siguientes identidades:

$$g(u) = \begin{cases} 1 & \text{if } u \leq x \\ 0 & \text{if } u < 0 \end{cases} \quad h(v) = \begin{cases} 1 & \text{if } v \leq y \\ 0 & \text{if } v < 0 \end{cases}$$

$$\Rightarrow E(g(X)h(Y)) = P(X \leq x, Y \leq y) \\ \Rightarrow E(g(X)) = P(X \leq x) \text{ \& } E(h(Y)) = P(Y \leq y)$$

$$\Rightarrow P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \text{ for } x, y \in \mathbb{R}$$

- El valor esperado condicional de Y dado $X = x$, denotado por $E(Y|X = x)$, es la media de la función de densidad condicional y se define de la siguiente manera para cualquier valor de x para el que $f_X(x) > 0$:

$$E(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy = \int_{-\infty}^{\infty} y \frac{f_{XY}(x, y)}{f_X(x)} dy$$

- La aplicación más útil de la esperanza condicional es una forma del teorema de la partición que permite calcular $E(Y)$ en situaciones donde la esperanza condicional $E(Y|X = x)$ se puede calcular fácilmente: la ley de esperanzas iteradas o de esperanza total
- Si X e Y son variables aleatorias continuas conjuntas, entonces se cumple la siguiente igualdad, en donde la integral es para todos los valores de x para los cuales $f_X(x) > 0$:

$$E(Y) = E_X[E(Y|X)] = \int_{x \in \mathbb{R}: f_X(x) > 0} E(Y|X = x) f_X(x) dx$$

- La demostración de este teorema se basa en la definición de la función de densidad marginal

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{XY}(x, y) dx dy = \\ &= \int_{x \in \mathbb{R}: f_X(x) > 0} \int_{-\infty}^{\infty} y f_{Y|X}(y|x) f_X(x) dx dy = \\ &= \int_{x \in \mathbb{R}: f_X(x) > 0} \left(\int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \right) f_X(x) dx = \\ &= \int_{x \in \mathbb{R}: f_X(x) > 0} E(Y|X = x) f_X(x) dx = E_X(E(Y|X)) \end{aligned}$$

- Una vez estudiadas las funciones de distribución y de densidad conjuntas de las variables, se pueden extender los resultados de sumas de variables continuas para el caso continuo e introducir la técnica del cambio de variables
 - Normalmente es necesario saber la función de densidad de la suma $Z \equiv X + Y$ de dos variables aleatorias continuas conjuntas, la cual será la derivada de su función de distribución

$$P(Z \leq z) = P(X + Y \leq z) = \iint_A f_{XY}(x, y) \, dx \, dy$$

$$\text{where } A \equiv \{(x, y) \in \mathbb{R}^2 : x + y \leq z\}$$

- Expresado en términos de los límites de integración, se puede obtener una expresión en términos de una doble integral y, diferenciando con respecto a z , se puede obtener la función de densidad de Z

$$P(Z \leq z) = P(X + Y \leq z) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{z-x} f_{XY}(x, y) \, dx \, dy =$$

$$= \int_{v=-\infty}^z \int_{u=-\infty}^{\infty} f_{XY}(u, v-u) \, du \, dv$$

$$\frac{d}{dz} P(Z \leq z) = \frac{d}{dz} \int_{v=-\infty}^z \int_{u=-\infty}^{\infty} f_{XY}(u, v-u) \, du \, dv =$$

$$= \int_{u=-\infty}^{\infty} f_{XY}(u, z-u) \, du \, dv$$

- Si las variables aleatorias continuas X e Y son independientes con funciones de densidad f_X y f_Y , entonces la función de densidad de $Z = X + Y$ es la siguiente:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) \, dx \quad \text{for } z \in \mathbb{R}$$

- En el lenguaje del análisis matemático, la identidad expresa que f_Z es la convolución de f_X y f_Y
- Una pregunta interesante es saber aspectos de la distribución conjunta de un par (U, V) de variables aleatorias definidas por $U = u(X, Y)$ y $V = v(x, y)$ (en donde $u, v: \mathbb{R}^2 \rightarrow \mathbb{R}$) si son X e Y son variables aleatorias

- Siendo T un mapeado de \mathbb{R}^2 a \mathbb{R}^2 definido por $T(x, y) = (u, v)$, en donde $u(x, y)$ y $v(x, y)$, y suponiendo que T es una biyección sobre algún dominio $D \subseteq \mathbb{R}^2$ y algún rango $S \subseteq \mathbb{R}^2$, entonces existe un punto $(x, y) = T^{-1}(u, v)$ en D , y se escribe $x = x(u, v)$ y $y = y(u, v)$
- El jacobiano de T^{-1} se define como el determinante de la matriz de sus derivadas (se supone que las derivadas existen y que son continuas para todos los puntos en S)

$$J(u, v) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial y}{\partial u} \frac{\partial x}{\partial v}$$

- La teoría estándar de las integrales múltiples expresa que se puede hacer un cambio de variables dentro de la integral: si $g: \mathbb{R}^2 \rightarrow \mathbb{R}$ es una función integrable, se puede cambiar de variable dentro de una integral para un subconjunto A de D lo suficientemente regular

$$\iint_A g(x, y) dx dy = \iint_{T(A)} g(x(u, v), y(u, v)) |J(u, v)| du dv$$

- Siendo X e Y variables aleatorias continuas conjuntas con función de densidad conjunta f_{XY} y $D \equiv \{(x, y): f_{XY}(x, y) > 0\}$, si el mapeado T dado por $T(x, y) = (u(x, y), v(x, y))$ es una biyección de $D \subseteq \mathbb{R}^2$ a $S \subseteq \mathbb{R}^2$, entonces el par $(U, V) = (u(X, Y), v(X, Y))$ es conjuntamente continua con la siguiente función de densidad:

$$f_{UV}(u, v) \equiv \begin{cases} f_{XY}(x(u, v), y(u, v)) |J(u, v)| & \text{if } (u, v) \in S \\ 0 & \text{otherwise} \end{cases}$$

- Suponiendo que $A \subseteq D$ y $T(A) = B$, como $T: D \rightarrow S$ es una biyección, entonces $P((U, V) \in B) = P((X, Y) \in A)$. En consecuencia, se puede encontrar las siguientes identidades que se mantienen para cualquier $B \subseteq S$

$$P((X, Y) \in A) = \iint_A f_{XY}(x, y) dx dy =$$

$$= \iint_{T(A)} f_{XY}(x(u, v), y(u, v)) |J(u, v)| du dv = P((U, V) \in B)$$

- Este teorema se puede aplicar fácilmente, pero es necesario comprobar antes que el mapeado es un biyección

- Anteriormente, se ha visto el caso de distribuciones y momentos para dos variables, pero cuando se tiene una distribución multivariante, es posible obtener expresiones análogas en términos de vectores de variables aleatorias

- La función de distribución $F_{\mathbf{x}}$ de un vector aleatorio \mathbf{x} se define de la siguiente manera:

$$F_{\mathbf{x}} \equiv \int_{-\infty}^{x_n} \int_{-\infty}^{x_{n-1}} \dots \int_{-\infty}^{x_1} f(\mathbf{x}) dx_1 \dots dx_{n-1} dx_n$$

where $\mathbf{x} = (X_1, X_2, \dots, X_n)$ is a random vector

- La función de densidad de la distribución multivariante es $f(\mathbf{x})$, y la función de distribución marginal para cada variable se puede obtener integrando sobre las otras variables (pero no sobre la variable de interés)
- El valor esperado de un vector o de una matriz es el vector o matriz de sus valores esperados, y es definido de la siguiente forma:

$$E(\mathbf{x}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \dots \\ E(X_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu} \quad \text{or}$$

$$E(\mathbf{X}) = \begin{bmatrix} E(X_{11}) & \dots & E(X_{1n}) \\ \vdots & \ddots & \vdots \\ E(X_{n1}) & \dots & E(X_{nn}) \end{bmatrix} = \begin{bmatrix} \mu_{11} & \dots & \mu_{1n} \\ \vdots & \ddots & \vdots \\ \mu_{n1} & \dots & \mu_{nn} \end{bmatrix} = \boldsymbol{\mu}$$

- En el caso en el que la esperanza estuviera condicionada a algún valor, entonces la definición es la misma pero las esperanzas están condicionadas a ese valor
- Las propiedades lineales del valor esperado también se cumplen cuando se multiplica el vector aleatorio por un vector de constantes

$$\begin{aligned} E(\mathbf{a}'\mathbf{x}) &= E(a_1X_1 + a_2X_2 + \dots + a_nX_n) = \\ &= a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n) = \mathbf{a}'\boldsymbol{\mu} \end{aligned}$$

$$E(\mathbf{A}\mathbf{X}\mathbf{B}) = \mathbf{A}E(\mathbf{X})\mathbf{B} = \mathbf{A}\boldsymbol{\mu}\mathbf{B}$$

- Tomando la esperanza de una matriz $(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'$ permite obtener la matriz de varianzas y covarianzas (respecto al par de variables en el producto)

$$(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' = \begin{bmatrix} (x_1 - \mu_1)(x_1 - \mu_1) & (x_1 - \mu_1)(x_2 - \mu_2) & \cdots & (x_1 - \mu_1)(x_n - \mu_n) \\ (x_2 - \mu_2)(x_1 - \mu_1) & (x_2 - \mu_2)(x_2 - \mu_2) & \cdots & (x_2 - \mu_2)(x_n - \mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (x_n - \mu_n)(x_1 - \mu_1) & (x_n - \mu_n)(x_2 - \mu_2) & \cdots & (x_n - \mu_n)(x_n - \mu_n) \end{bmatrix}$$

$$\Rightarrow E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'] = E(\mathbf{x}\mathbf{x}') - \boldsymbol{\mu}\boldsymbol{\mu}' = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \ddots & & \sigma_{2n} \\ \vdots & & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

- En este caso, si $i \neq j$, entonces $\sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)'] = Cov(x_i, x_j)$, y las entradas diagonales son las varianzas porque $\sigma_{ii} = E[(x_i - \mu_i)^2] = Var(x_i)$
- Esta matriz normalmente se denota como la covarianza de \mathbf{x} o \mathbf{X} , y se usa el símbolo Σ

$$Cov(\mathbf{x}) = \Sigma \quad \text{or} \quad Cov(\mathbf{X}) = \Sigma$$

- Dividendo cada entrada σ_{ij} por $\sigma_{ii}\sigma_{jj}$ se puede obtener la matriz de correlación, denotada por R

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_{ii}\sigma_{jj}} \quad \text{for } \forall i, j \Rightarrow \mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \ddots & & \rho_{2n} \\ \vdots & & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{bmatrix}$$

- La matriz de desviaciones estándar se denota como $\mathbf{V}^{\frac{1}{2}}$ y se define como la raíz cuadrada de la matriz diagonal \mathbf{V} que se obtiene al diagonalizar Σ

$$\mathbf{P}\Sigma\mathbf{P}' = \mathbf{V} = \begin{bmatrix} \sigma_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{nn} \end{bmatrix} \Rightarrow \mathbf{V}^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\sigma_{nn}} \end{bmatrix}$$

- A partir de la matriz de desviaciones estándar y la de correlación, es posible obtener una igualdad con Σ :

$$\mathbf{V}^{1/2}\mathbf{R}\mathbf{V}^{1/2} = \Sigma \Rightarrow \mathbf{R} = (\mathbf{V}^{1/2})^{-1}\Sigma(\mathbf{V}^{1/2})^{-1}$$

- También es posible obtener las propiedades de la esperanza, de la varianza y de la covarianza
 - Para la varianza, la propiedad de multiplicación por escalares también se mantiene, pero usando la forma cuadrática. Como Σ no es negativa, la forma cuadrática es semidefinida positiva o definida positiva (dependiendo del rango de columnas)

$$\begin{aligned} Cov(\mathbf{a}'\mathbf{x}) &= E[(\mathbf{a}'\mathbf{x} - \mathbf{a}'\boldsymbol{\mu})(\mathbf{a}'\mathbf{x} - \mathbf{a}'\boldsymbol{\mu})'] = \\ &= E[\mathbf{a}'(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\mathbf{a}] = \mathbf{a}'E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})']\mathbf{a} = \mathbf{a}'\Sigma\mathbf{a} \end{aligned}$$

$$\begin{aligned} Cov(\mathbf{A}\mathbf{X}) &= E[(\mathbf{A}'\mathbf{X} - \mathbf{A}'\boldsymbol{\mu})(\mathbf{A}'\mathbf{X} - \mathbf{A}'\boldsymbol{\mu})'] = \\ &= E[\mathbf{A}'(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'\mathbf{A}] = \mathbf{A}'E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})']\mathbf{A} = \mathbf{A}\Sigma\mathbf{A}' \end{aligned}$$

$$as \quad E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'] = \Sigma$$

- Es importante ver que, para la matriz de varianzas y covarianzas, no es lo mismo multiplicar un vector \mathbf{a} que una matriz \mathbf{A} , resultando en $\mathbf{a}'\Sigma\mathbf{a}$ para el primer caso y en $\mathbf{A}\Sigma\mathbf{A}'$ para el segundo
- Cualquier matriz de varianzas y covarianzas es simétrica y es semidefinida positiva

$$\sigma_{ij} = Cov(x_i, x_j) = Cov(x_j, x_i) = \sigma_{ji}$$

$$\begin{aligned} \mathbf{y}'\Sigma\mathbf{y} &= \mathbf{y}'E[\mathbf{y}'(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\mathbf{y}] = E[(\mathbf{x} - \boldsymbol{\mu})'\mathbf{y}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{y}] = \\ &= Var[(\mathbf{x} - \boldsymbol{\mu})'\mathbf{y}] \geq 0 \end{aligned}$$

- Las variables aleatorias centradas $(x_1 - \mu_1), (x_2 - \mu_2), \dots, (x_n - \mu_n)$ del vector aleatorio centrado $(\mathbf{x} - \boldsymbol{\mu})$ son linealmente independientes si, y solo si, $Cov(\mathbf{x})$ es positiva definida

$$\mathbf{y}'Cov(\mathbf{x})\mathbf{y} > 0$$

- Si $Z = \mathbf{y}'\mathbf{x}$, se puede demostrar el teorema anterior por contradicción:

$$\mathbf{y}'Cov(\mathbf{x})\mathbf{y} = 0 \quad for \text{ some } \mathbf{y} \neq \mathbf{0}$$

$$\Rightarrow Var(Z) = 0 \quad for \text{ some } \mathbf{y} \neq \mathbf{0}$$

$$\Rightarrow Z - m_z = 0 \quad for \text{ some } \mathbf{y} \neq \mathbf{0}$$

$$\Rightarrow \sum_{i=1}^n y_i(x_i - \mu_i) = 0 \quad for \text{ some } \mathbf{y} \neq \mathbf{0}$$

$$\Rightarrow (x_1 - \mu_1), (x_2 - \mu_2), \dots, (x_n - \mu_n) \text{ are not lin. indep.}$$

Los estadísticos de orden

- Siendo X_1, X_2, \dots, X_n una muestra aleatoria de una distribución con función de distribución F y X una variable aleatoria genérica con esta distribución, muchos objetos naturales de interés suelen ser observaciones en una posición concreta de la ordenación de una muestra
 - Para $k = 1, 2, \dots, n$ siendo $X_{(k)}$ la cantidad k -ésima más pequeña de X_1, \dots, X_n , el vector $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$ se denomina estadístico de orden, y $X_{(k)}$ es la k -ésima variable de orden

- El estadístico de orden, por lo tanto, se obtiene de la muestra original desordenada a través de permutación: las observaciones se ordenan en orden ascendente. Por lo tanto, las siguientes desigualdades se respetan:

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

- En verdad, el orden de las variables también depende de n : $X_{(k)}$ es la k -ésima observación más pequeña de las n observaciones X_1, X_2, \dots, X_n . Para ser completamente descriptivos, la notación tendría que incluir n para indicar el tamaño muestral
- Las variables de orden extremas son $X_{(1)} = \min\{X_1, X_2, \dots, X_n\}$ y $X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$ cuyas funciones de distribución se obtienen de la siguiente manera:
 - Debido a que $X_{(n)}$ es el máximo, entonces cualquier variable que no sea $X_{(n)}$ deberá respetar la desigualdad $\leq x$ (dado que son menores o iguales siempre si $X_{(n)} \leq x$). Por lo tanto, se puede expresar como la probabilidad conjunta para los valores muestrales, y como son independientes e idénticamente distribuidos, se obtiene la siguiente función de distribución:

$$\begin{aligned} F_{X_{(n)}}(x) &= P(X_1 \leq x, X_2 \leq x, \dots, X_n \leq x) = \prod_{i=1}^n P(X_i \leq x) = \\ &= [P(X_i \leq x)]^n = [F(x)]^n \end{aligned}$$

- Debido a que $X_{(1)}$ es el mínimo, entonces se puede usar la misma lógica de antes, pero usando $1 - P(X_{(1)} > x)$, dado que de esta manera se podrá usar el hecho de que, si el mínimo es mayor a x , todos los otros valores necesariamente lo son también

$$\begin{aligned}
F_{X_{(1)}}(x) &= 1 - P(X_{(1)} > x) = \\
&= 1 - P(X_1 > x, X_2 > x, \dots, X_n > x) = 1 - \prod_{i=1}^n P(X_i > x) = \\
&= 1 - [P(X_i > x)]^n = 1 - [1 - F(x)]^n
\end{aligned}$$

- En el caso continuo, además, se pueden derivar las expresiones anteriores para obtener las funciones de densidad correspondientes

$$f_{X_{(n)}}(x) = n[F(x)]^{n-1}f(x)$$

$$f_{X_{(1)}}(x) = n[1 - F(x)]^{n-1}f(x)$$

- Para $k = 1, 2, \dots, n$, la función de distribución de $X_{(k)}$ será la función de distribución de una distribución beta con parámetros k y $n + 1 - k$, de modo la función de distribución se expresa de la siguiente manera:

$$\begin{aligned}
F_{X_{(k)}}(x) &= F_{Beta(k, n+1-k)}(F(x)) \\
\Rightarrow F_{X_{(k)}}(x) &= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n+1-k)} \int_0^{F(x)} y^{k-1}(1-y)^{n-k} dy
\end{aligned}$$

- Si $X \sim U(0,1)$, entonces obviamente $X_{(k)} \sim Beta(k, n+1-k)$
- Para $i = 0, 1, \dots, n$, se define A_i como el evento de que i de las variables X_1, X_2, \dots, X_n sea menor o igual a x . Como estos eventos son disjuntos (por lo de que solo pueden haber exactamente i variables), entonces $\{A_i \leq x\} \sim Bin(n, F(x))$ y se obtiene el siguiente resultado:

$$\begin{aligned}
F_{X_{(k)}}(x) &= P(X_{(k)} \leq x) = P\left(\bigcup_{i=k}^n A_i(x)\right) = \sum_{i=k}^n P(A_i(x)) = \\
&= \sum_{i=k}^n \binom{n}{i} (F(x))^i (1-F(x))^{n-i}
\end{aligned}$$

- El resultado, por lo tanto, queda demostrado si se aplica la definición del factorial en términos de la función gamma, $\Gamma(n) = (n-1)!$, para obtener la siguiente igualdad:

$$\sum_{i=k}^n \frac{n!}{i! (n-i)!} (F(x))^i (1-F(x))^{n-i} =$$

$$= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n+1-k)} \int_0^{F(x)} y^{k-1} (1-y)^{n-k} dy$$

- En el caso continuo, la diferenciación de $F_{X_{(k)}}(x)$ permite obtener la densidad. Para $k = 1, 2, \dots, n$, la densidad de $X_{(k)}$ es la densidad de una distribución $Beta(k, n+1-k)$ para $F(x)$ multiplicada por la de la variable X

$$f_{X_{(k)}}(x) = f_{Beta(k, n+1-k)}(F(x))f(x)$$

$$\Rightarrow f_{X_{(k)}}(x) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n+1-k)} [F(x)]^{k-1} [1-F(x)]^{n-k} f(x)$$

- Esto se consigue fácilmente al integrar la expresión anterior, dado que $F_{X_{(k)}}(0) = 0$
- Para $k = 1$ y $k = n$, se redescubre, en ambos teoremas, las expresiones familiares para las funciones de distribución y de densidad del menor y el mayor valor
- Bajo la suposición de que la densidad es continua, se pueden hacer las siguientes derivaciones heurísticas: si h es "muy pequeña", entonces se pueden hacer las siguientes aproximaciones:

$$F_{X_{(k)}}(x+h) - F_{X_{(k)}}(x) = P(x < X_{(k)} \leq x+h)$$

$$\approx P(k-1 \text{ obs.} \leq x, 1 \text{ obs. in } (x, x+h], n-k \text{ obs.} > x+h)$$

- Esta aproximación se debe a que la probabilidad de al menos dos observaciones cae en el intervalo $(x, x+h]$. Esto es la probabilidad multinomial:

$$\frac{n!}{(k-1)! 1! (n-k)!} [F(x)]^{k-1} [F(x+h) - F(x)] [1 - F(x+h)]^{n-k} =$$

$$= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n+1-k)} [F(x)]^{k-1} [F(x+h) - F(x)] [1 - F(x+h)]^{n-k}$$

- Por el teorema del valor medio, $F(x+h) - F(x) = hf(\theta_{x,h})$ cuando $x \leq \theta_{x,h} \leq x+h$. Como h es pequeña y f continua, se tiene que $f(\theta_{x,h}) \approx f(x)$ y $F(x+h) \approx F(x)$ y se obtiene el siguiente resultado:

$$F_{X_{(k)}}(x+h) - F_{X_{(k)}}(x) \approx \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n+1-k)} [F(x)]^{k-1} [1-F(x)]^{n-k} f(x)$$

- Ahora que se ha visto la distribución de primer orden, se considera $X_{(1)}$ y $X_{(n)}$ conjuntamente, asumiendo que es continua durante el desarrollo

○ ...

Commented [IC6]: Acabar bien

Las funciones generadoras de probabilidad

- Igual que se puede expresar una secuencia u_0, u_1, u_2, \dots de números reales con la fórmula general para u_n , también se puede escribir la función generadora de la secuencia, definida por la siguiente suma:

$$U(s) = u_0 + u_1 s + u_2 s^2 + \dots$$

- Estas funciones generadoras son muy útiles para lidiar con secuencias de números reales porque especifican únicamente la secuencia, de modo que dada una secuencia real de números reales u_0, u_1, u_2, \dots es posible encontrar una función generadora de la forma anterior

- De manera converso, si la función generadora $U(s)$ tiene una serie de Taylor convergente definida por $U(s)$ para una s pequeña, entonces esta expansión es única y $U(s)$ solo puede generar la secuencia u_0, u_1, u_2, \dots
- Definiendo u_n de diferentes maneras, es posible poder definir otros tipos de funciones generadoras

$$u_n = \frac{1}{n!} v_n \quad U(s) = v_0 + v_1 s + \frac{1}{2!} v_2 s^2 + \frac{1}{3!} v_3 s^3 + \dots$$

- Cuando se trabaja con funciones generadoras de secuencias reales, es importante que la serie de potencias converja a una cierta $s \neq 0$
 - Mientras este sea el caso, no se tiene que especificar aquellos valores para los cuales la suma es absolutamente convergente
- Muchas variables aleatorias toman valores enteros no negativos, de modo que se puede pensar en una función de masa de probabilidad de este tipo de variable aleatoria como una secuencia de números p_0, p_1, p_2, \dots que cumple las siguientes condiciones:

$$p_k = P(X = k) \quad \text{for } k = 0, 1, 2, \dots$$

$$\text{such that } p_k \geq 0 \text{ for } \forall k \text{ \& } \sum_{k=0}^{\infty} p_k = 1$$

- La función generadora de probabilidad de X es la función $G_X(s)$ definida por $G_X(s) \equiv p_0 + p_1s + p_2s^2 + \dots$ para todos los valores de s para los que la suma converge absolutamente. En otras palabras, la función generadora de probabilidad $G_X(s)$ de X es la función generadora de la secuencia p_0, p_1, p_2, \dots

$$G_X(s) = \sum_{x \geq 0} s^x p_x = p_0 + p_1s + p_2s^2 + \dots$$

- Es fácil ver que $G_X(s)$ existe o converge para todos los valores de s tal que $|s| \leq 1$ o $s \in [-1, 1]$, dado que se cumple la siguiente desigualdad:

$$\sum_{k=0}^{\infty} |p_k s^k| = \sum_{k=0}^{\infty} p_k |s^k| = \sum_{k=0}^{\infty} p_k |s|^k \leq \sum_{k=0}^{\infty} p_k = 1$$

- De manera más general, existe un radio de convergencia R tal que $1 \leq R \leq \infty$ y que $\sum_{x \geq 0} s^x p_x$ converge absolutamente cuando $|s| < R$ y diverge cuando $|s| > R$. De este modo, la función generadora de probabilidad está bien definida para $s \in [-R, R]$
- A partir de que $\sum_{k=0}^{\infty} p_k = 1$ y la definición de $G_X(s)$, se pueden obtener las siguientes igualdades:

$$G_X(0) = \sum_{k=0}^{\infty} 0^k p_k = p_0 + \sum_{k=1}^{\infty} 0^k p_k = p_0$$

$$G_X(1) = \sum_{k=0}^{\infty} 1^k p_k = \sum_{k=0}^{\infty} p_k = 1$$

- A partir de la ley del estadístico subconsciente, se puede ver que la función $G_X(s)$ es la esperanza de s elevada a X , siempre que esta exista. Por lo tanto, la función generadora de probabilidad puede expresarse como $E(s^X)$

$$E(s^X) = \sum_{k=0}^{\infty} s^k P(X = k) = \sum_{k=0}^{\infty} s^k p_k = G_X(s)$$

$$\Rightarrow G_X(s) = E(s^X)$$

- Suponiendo que X e Y tienen la misma función generadora de probabilidad G_X y G_Y , respectivamente, entonces $G_X(s) = G_Y(s)$ para toda s si, y solo si, $P(X = k) = P(Y = k)$ para $k = 0, 1, 2, \dots$

- Debido a que $\sum_{k=0}^{\infty} |p_k s^k| \leq \sum_{k=0}^{\infty} p_k = 1$, tanto G_X como G_Y tienen una única expansión de series de potencias en el origen. Si $G_X(s) = G_Y(s)$, entonces $P(X = k) = P(Y = k)$

$$G_X(s) = \sum_{k=0}^{\infty} s^k P(X = k) \quad G_Y(s) = \sum_{k=0}^{\infty} s^k P(Y = k)$$

$$\begin{aligned} G_X(s) = G_Y(s) &\Rightarrow \sum_{k=0}^{\infty} s^k P(X = k) = \sum_{k=0}^{\infty} s^k P(Y = k) \\ &\Rightarrow P(X = k) = P(Y = k) \end{aligned}$$

- Esto demuestra, por tanto, que solo es necesario comprobar las funciones generadoras de probabilidad para saber si ambas variables siguen la misma distribución de probabilidad
- Algunos de los ejemplos más comunes de funciones generadoras de probabilidad son las siguientes:

- Si X tiene una distribución de Bernoulli con parámetro p , entonces la función generadora de probabilidad es la siguiente:

$$G_X(s) = \sum_{k=0}^1 s^k P(X = k) = q + ps$$

- Si X tiene una distribución binomial con parámetros p y n , entonces la función generadora de probabilidad es la siguiente:

$$\begin{aligned} G_X(s) &= \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} s^k = \sum_{k=0}^n \binom{n}{k} (ps)^k q^{n-k} = \\ &= (q + ps)^n \end{aligned}$$

- Si X tiene una distribución geométrica con parámetros p , entonces la función generadora de probabilidad es la siguiente:

$$\begin{aligned} G_X(s) &= \sum_{k=0}^{\infty} q^{k-1} p s^k = ps \sum_{k=1}^{\infty} (qs)^{k-1} = \frac{ps}{1 - qs} \\ G_X(s) &= \sum_{k=0}^{\infty} q^k p s^k = p \sum_{k=0}^{\infty} (qs)^k = \frac{p}{1 - qs} \\ &|s| < q^{-1} \end{aligned}$$

- Si X tiene una distribución de Poisson con parámetro λ , entonces la función generadora de probabilidad es la siguiente:

$$G_X(s) = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} s^k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda s)^k}{k!} = e^{-\lambda} e^{\lambda s} = e^{-\lambda + \lambda s} = e^{\lambda(s-1)}$$

- Si X tiene una distribución binomial negativa con parámetros p y n , entonces la función generadora de probabilidad es la siguiente:

$$\begin{aligned} G_X(s) &= \sum_{k=n}^{\infty} \binom{k-1}{n-1} p^n q^{k-n} s^k = \\ &= \sum_{y=0}^{\infty} \binom{n+y-1}{y} p^n q^y s^{n+y} = (ps)^n \sum_{y=0}^{\infty} \binom{n+y-1}{y} (qs)^y = \\ &= (ps)^n (1-qs)^{-n} = \\ &= \left(\frac{ps}{1-qs} \right)^n \quad \text{if } |s| < q^{-1} \end{aligned}$$

- Para cualquier variable aleatoria discreta X , el valor esperado $E(X)$ es una indicación central de la distribución de X , pero solo es uno de los números que contienen información de la distribución de X
 - La colección de números que contiene información sobre la distribución de X es la secuencia $E(X), E(X^2), E(X^3), \dots$ de valores esperados de las potencias de X , los cuales se llaman momentos de X
 - Siendo $k \geq 1$, el momento k de la variable aleatoria X es la cantidad $E(X^k)$
 - Uno de las cantidades más importantes que nace de los momentos de X es la varianza de X , definida de la siguiente manera:

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

- Siendo X una variable aleatoria con valores en los enteros no negativos y con función generadora de probabilidad $G_X(s)$, la derivada de r orden de $G_X(s)$ en $s = 1$ cumple la siguiente identidad:

$$G_X^{(r)}(1) = E[X(X-1) \dots (X-r+1)]$$

- Para poder ver de manera no rigurosa por qué se da esta igualdad, se puede ver como, dado que $|s| < 1$, el lema de Abel (de análisis real) aplica y se puede demostrar que la primera

derivada de $G_X(s)$ es $E(X)$ y que la lógica mostrada también aplica a una r mayor

$$\begin{aligned} G'_X(s) &= \frac{d}{ds} \sum_{k=0}^{\infty} s^k P(X=k) = \sum_{k=0}^{\infty} \frac{d}{ds} s^k P(X=k) = \\ &= \sum_{k=1}^{\infty} k s^{k-1} P(X=k) \Rightarrow G'_X(1) = \sum_{k=1}^{\infty} k P(X=k) = E(X) \end{aligned}$$

- El resultado anterior permite ver como se puede generalizar esta formulación para dar una demostración más rigurosa cuando se deriva r veces

$$\begin{aligned} G_X^{(r)}(s) &= \frac{d^r}{ds^r} \sum_{k=0}^{\infty} s^k P(X=k) = \sum_{k=0}^{\infty} \frac{d^r}{ds^r} s^k P(X=k) = \\ &= \sum_{k=r}^{\infty} k(k-1) \dots (k-r+1) s^{k-r} P(X=k) \\ \Rightarrow G_X^{(r)}(1) &= \sum_{k=r}^{\infty} k(k-1) \dots (k-r+1) P(X=k) = \\ &= E[X(X-1) \dots (X-r+1)] \end{aligned}$$

- A partir del teorema, es bastante fácil calcular los momentos de X en términos de $G_X^{(r)}(1)$

$$\begin{aligned} E(X^2) &= E[X(X-1) + X] = E[X^2 - X + X] = G_X''(1) + G_X'(1) \\ \Rightarrow \text{Var}(X) &= G_X''(1) + G_X'(1) - G_X'(1)^2 \\ &\Rightarrow \dots \end{aligned}$$

- En los casos en los que $R < 1$, se puede ver como $s \in (-1,1)$, y eso hace que no exista el límite en los puntos $s = \pm 1$. Para poder usar la función también estos casos, se puede definir la propiedad anterior de la siguiente manera:

$$G_X^{(r)}(1) = E[X(X-1) \dots (X-r+1)] = \lim_{s \rightarrow 1^-} G_X^{(r)}(s)$$

Example:

$$G_X(t) = \sum_{k=1}^{\infty} s^k P(X = k) = \sum_{k=1}^{\infty} s^k \frac{6}{\pi^2 k^2} = \frac{6}{\pi^2} \sum_{k=1}^{\infty} \frac{s^k}{k^2}$$

$$G'_X(t) = \frac{6}{\pi^2} \sum_{k=0}^{\infty} \frac{s^{k-1}}{k} = -\frac{6}{\pi^2} \frac{\ln(1-s)}{s} \rightarrow \infty \text{ as } t \rightarrow 1^-$$

- Derivando la función generadora de probabilidad $r = k$ veces, es posible ver que la función generadora de probabilidad en $s = 0$ permite obtener las probabilidades para cada k posible

$$G_X^{(k)}(s) = \sum_{k=k}^{\infty} k(k-1) \dots (k-k+1) s^{k-k} P(X = k)$$

$$\Rightarrow G_X^{(k)}(0) = \sum_{k=k}^{\infty} k(k-1) \dots (k-k+1) P(X = k) = k! P(X = k)$$

$$\Rightarrow P(X = k) = \frac{G_X^{(k)}(0)}{k!}$$

- A partir de este resultado, es posible obtener la probabilidad para cualquier valor k que tome la variable discreta
- Una forma más sencilla y equivalente de poder obtener la probabilidad para cada $P(X = k)$ es mirar cual es el coeficiente que multiplica s^k , denotado por $[s^k]G_X(s)$, ya que este es $P(X = k)$

$$G_X(s) = \sum_{k=0}^{\infty} s^k P(X = k) \Rightarrow G_X(s)[s^k] = P(X = k)$$

- Mucha de la teoría de la probabilidad se enfoca en la suma de variables independientes, por lo que se necesita una manera alternativa a la convolución para lidiar con estas sumas (dado que se necesitarían $n - 1$ para encontrar la función de masa de la suma de n variables independientes)
 - Si X e Y son variables aleatorias independientes, cada una tomando valores en el conjunto $\{0, 1, 2, \dots\}$, entonces su suma tiene la siguiente función generadora de probabilidad:

$$G_{X+Y}(s) = G_X(s)G_Y(s)$$

- La demostración de este resultado se basa en que la función generadora de probabilidad es igual a la esperanza de s^X

$$G_{X+Y}(s) = E(s^{X+Y}) = E(s^X s^Y) = E(s^X)E(s^Y) = G_X(s)G_Y(s)$$

- A partir de este teorema, la suma $S_n = X_1 + X_2 + \dots + X_n$ de n variables independientes que toman valores en $\{0,1,2, \dots\}$ tiene la siguiente función generadora de probabilidad:

$$G_{S_n}(s) = G_{X_1}(s)G_{X_2}(s) \dots G_{X_n}(s) = \prod_{j=1}^n G_{X_j}(s)$$

- La demostración de este teorema es sencilla porque sigue una lógica similar al caso de dos variables:

$$\begin{aligned} G_{S_n}(s) &= E(s^{S_n}) = E(s^{X_1+X_2+\dots+X_n}) = \\ &= E(s^{X_1} s^{X_2} \dots s^{X_n}) = E\left(\prod_{j=1}^n s^{X_j}\right) = \prod_{j=1}^n E(s^{X_j}) = \prod_{j=1}^n G_{X_j}(s) \end{aligned}$$

- Un corolario útil de este teorema de la convolución es que, si todas las variables se distribuyen igual (se pueden entender como n variables iguales X), entonces solo hace falta elevar a n la función $G_X(s)$

$$G_{S_n}(s) = E(s^{S_n}) = \prod_{j=1}^n G_{X_j}(s) = \prod_{j=1}^n G_X(s) = [G_X(s)]^n$$

- Esta fórmula permite analizar mejor la suma de variables independientes, dado que permite calcular el valor esperado de S a través del resultado anterior sobre la derivación de las funciones generadoras de probabilidad

$$G'_{S_n}(s) = n[G_X(s)]^{n-1}G'_X(s)$$

$$\Rightarrow E(S_n) = G'_{S_n}(1) = n[G_X(1)]^{n-1}G'_X(1)$$

- Siendo N y X_1, X_2, \dots, X_N variables aleatorias independientes, cada una tomando valores en $\{0,1,2, \dots\}$, si las X_i se distribuyen idénticamente con una función generadora de probabilidad G_X común, entonces la suma $S = X_1 + X_2 + \dots + X_N$ tiene la siguiente función generadora de probabilidad:

$$G_S(s) = G_N(G_X(s))$$

- Esta una extensión importante de este teorema lidia con la suma de un número aleatorio de variables aleatorias independiente, de modo que hay una variable N que determina el número de variables independientes e idénticamente distribuidas X_i
- Este teorema se puede demostrar a partir del teorema de la probabilidad total, de modo que se coge una $N = n$ y se usa como condición para la esperanza de s^S

$$\begin{aligned}
 G_S(s) &= E(s^S) = E_N[E(s^S|N = n)] = \\
 &= \sum_{n=0}^{\infty} E(s^{X_1+X_2+\dots+X_N}|N = n)P(N = n) = \\
 &= \sum_{n=0}^{\infty} E(s^{X_1+X_2+\dots+X_n})P(N = n) = \\
 &= \sum_{n=0}^{\infty} E(s^{nX})P(N = n) = \sum_{n=0}^{\infty} [E(s^X)]^n P(N = n) = G_N[G_X(s)]
 \end{aligned}$$

- Esta fórmula es muy útil para poder encontrar momentos y otros valores de interés. Debido a la regla de la cadena de la derivación, es posible obtener momentos de S a través de momentos de N y X

$$\begin{aligned}
 G'_S(s) &= G'_N(G_X(s))G'_X(s) \\
 \Rightarrow E(S) &= G'_S(1) = G'_N(G_X(1))G'_X(1) = \\
 &= G'_N(1)G'_X(1) = E(N)E(X)
 \end{aligned}$$

- Otra manera de llegar a un resultado equivalente es a través de utilizar la esperanza condicionada para los valores de N , obteniendo la siguiente identidad:

$$\begin{aligned}
 E(S) &= E_N[E(S|N = n)] = E_N[E(X_1 + X_2 + \dots + X_N|N = n)] \\
 &= E_N[E(NX|N = n)] = E_N[E(nX)] = E_N[nE(X)] = E(N)E(X)
 \end{aligned}$$

- Este teorema, sin embargo, no se mantiene si las variables no están distribuidas de manera idéntica, ya que la función generadora de probabilidad no sería la misma para cada una de las variables de la suma

- Es posible generalizar la función generadora de probabilidad para el caso multivariante
 - La función generadora de probabilidad de $\mathbf{x} = (X_1, X_2, \dots, X_n)$ es la función $G_{\mathbf{x}}(\mathbf{s})$ para todos los valores de $\mathbf{s} = (s_1, s_2, \dots, s_n)$ para los que la suma converge absolutamente. En otras palabras, la función generadora de probabilidad $G_{\mathbf{x}}(\mathbf{s})$ de \mathbf{x} es la siguiente:

$$G_{\mathbf{x}}(\mathbf{s}) = \sum_{\mathbf{x} \geq 0} \mathbf{s}^{\mathbf{x}} p_{\mathbf{x}} = \sum_{X_1 \geq 0} \sum_{X_2 \geq 0} \sum_{X_n \geq 0} s_1^{X_1} s_2^{X_2} \dots s_n^{X_n} p_{\mathbf{x}}$$

- En este caso, el vector \mathbf{s} es un vector n -dimensional de números complejos
- A partir de la ley del estadístico subconsciente, se puede ver que la función $G_{\mathbf{x}}(\mathbf{s})$ es la esperanza de \mathbf{s} elevada a \mathbf{x} , siempre que esta exista. Por lo tanto, la función generadora de probabilidad puede expresarse como $E(\mathbf{s}^{\mathbf{x}})$

$$E(\mathbf{s}^{\mathbf{x}}) = \sum_{X_1 \geq 0} \sum_{X_2 \geq 0} \sum_{X_n \geq 0} s_1^{X_1} s_2^{X_2} \dots s_n^{X_n} p_{\mathbf{x}} = G_{\mathbf{x}}(\mathbf{s})$$

$$\Rightarrow G_{\mathbf{x}}(\mathbf{s}) = E(\mathbf{s}^{\mathbf{x}})$$

- Las otras propiedades mostradas anteriormente se pueden generalizar para este caso multivariante
- Otras propiedades que cumple la función generadora de probabilidad univariante se pueden generalizar para el caso multivariante

Los momentos y funciones generadoras de momentos

- Para cualquier variable aleatoria X , el momento k de X se define para $k \in \mathbb{N}$ como el número $E(X^k)$ cuando este valor esperado existe. La secuencia $E(X), E(X^2), \dots$ permite obtener mucha información sobre X
 - Dada la función de distribución F_X de una variable aleatoria X , se pueden calcular los momentos cuando estos existen (al menos si X es continua o discreta)
 - Generalmente, no es posible reconstruir la función de distribución F_X a partir de la secuencia $E(X), E(X^2), \dots$. No obstante, es posible hacer esta reconstrucción con información extra sobre la secuencia de momentos

- Suponiendo que todos los momentos $E(X), E(X^2), \dots$ de la variable X existen, y que la serie $\sum_{k=0}^{\infty} \frac{1}{k!} t^k E(X^k)$ es absolutamente convergente para alguna $t > 0$, la secuencia de momentos determina completamente la distribución de X
 - Este resultado es un resultado del análisis matemático, y está muy relacionado al teorema de la unicidad para las funciones generadoras de probabilidad
- Anteriormente se han visto las funciones generadoras de probabilidad, las cuales eran útiles para variables discretas que toman valores enteros no negativos. Para variables aleatorias generales, normalmente se utiliza una modificación llamada función generadora de momentos
 - La función generadora de momentos de una variable aleatoria X es la función M_X definida por $M_X(t) \equiv E(e^{tX})$ para toda $t \in \mathbb{R}$ satisfaciendo $|t| < \delta$ para alguna $\delta > 0$ (para la cual el valor esperado existe y es finito)

$$M_X(t) \equiv E(e^{tX}) \text{ for } \forall t \in \mathbb{R}$$

- Esta función es una modificación de la función generadora de probabilidad, por lo que hay una conexión entre ambas funciones. Esto ocurre por la siguiente identidad:

$$M_X(t) = E(e^{tX}) = G_X(e^t)$$

- En general, la función generadora de momentos se puede expresar de la siguiente manera cuando la suma o la integral convergen absolutamente

$$M_X(t) = E(e^{tX}) = \begin{cases} \sum_x e^{tx} P(X=x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

- Como se puede ver, hay una fuerte conexión entre la transformada de Laplace de funciones de valores reales y la función generadora de momentos. Para variables aleatorias no negativas, se puede definir la transformada de Laplace como $E(e^{-t})$ y, por tanto, se puede entender $M_X(t)$ como una transformación de Laplace para ambos lados (la transformación de Laplace para la función de densidad)

$$E(e^{-t}) \text{ for } t \geq 0$$

- Debido a que $M_X(t)$ se puede definir en términos de la esperanza de e^{tX} , es posible hacer una expansión de Taylor de la función generadora de momentos a partir de la expansión de la función exponencial

$$M_X(t) \equiv E(e^{tX}) = E\left(1 + tX + \frac{1}{2!}(tX)^2 + \dots\right) =$$

$$= 1 + tE(X) + \frac{1}{2!}t^2E(X^2) + \dots + \frac{1}{k!}t^kE(X^k) + \dots = \sum_{k=0}^{\infty} \frac{t^k}{k!}E(X^k)$$

- Si la función generadora de momentos M_X satisface $M_X(t) = E(e^{tX}) < \infty$ para toda t satisfaciendo $|t| < \delta$ donde $\delta > 0$, entonces hay una única distribución con una función generadora de momentos M_X . Además, bajo esta condición, se tiene que $E(X^k) < \infty$, de modo que todos los momentos son finitos y existen
- La idea de que $E(X^k) < \infty$ se basa en que la función exponencial crece más rápido que cualquier polinomio, de modo que $|x|^r \leq e^{|tx|}$ en cuanto $|x| > x_2$ para un valor x_2 , y cuando $|x| < x_2$ entonces $|x|^r \leq Ce^{|tx|}$ para alguna C , por lo que se pueden obtener las siguientes desigualdades:

$$|x|^r \leq (C+1)e^{|tx|}$$

$$\Rightarrow E(|X|^r) \leq (C+1)E(e^{|tX|}) < \infty \text{ for } |t| < \delta$$

- Este último teorema es básicamente el teorema de la transformación inversa de Laplace, el cual expresa que si la transformada de Laplace de f_X existe de una manera adecuada, entonces f_X se puede encontrar a través de la fórmula de inversión (la serie de potencias anterior)
- Siendo X e Y variables aleatorias, si existe una $\delta > 0$ tal que $M_X(t) = M_Y(t)$ para $|t| < \delta$, entonces X e Y se distribuyen idénticamente

$$M_X(t) = 1 + tE(X) + \frac{1}{2!}t^2E(X^2) + \dots + \frac{1}{k!}t^kE(X^k) + \dots$$

$$M_Y(t) = 1 + tE(Y) + \frac{1}{2!}t^2E(Y^2) + \dots + \frac{1}{k!}t^kE(Y^k) + \dots$$

$$\Rightarrow tE(X) + \frac{1}{2!}t^2E(X^2) + \dots + \frac{1}{k!}t^kE(X^k) + \dots =$$

$$tE(Y) + \frac{1}{2!}t^2E(Y^2) + \dots + \frac{1}{k!}t^kE(Y^k) + \dots$$

$$\Rightarrow t[E(X) - E(Y)] + \frac{1}{2!}t^2[E(X^2) - E(Y^2)] + \dots = 0$$

$$\Rightarrow E(X) = E(Y), E(X^2) - E(Y^2), \dots, E(X^k) = E(Y^k), \dots$$

- Se puede demostrar que $M_X(t)$ es la función generadora de momentos exponencial de X . Si $M_X(t)$ converge en un intervalo abierto conteniendo el valor $t = 0$, entonces, para $r = 1, 2, \dots$, se cumple la siguiente igualdad:

$$E(X^k) = k! M_X^{(r)}(0) \text{ for } r = 1, 2, 3, \dots$$

$$\text{where } M_X^{(r)} = \frac{\partial^r}{\partial t^r} M_X(t)$$

- Una demostración no muy rigurosa de este último resultado es la siguiente:

$$\begin{aligned} M_X^{(r)} &= \frac{\partial^r}{\partial t^r} M_X(t) = \frac{\partial^r}{\partial t^r} E(e^{tX}) = \frac{\partial^r}{\partial t^r} \sum_{k=0}^{\infty} \left[\frac{t^k}{k!} E(X^k) \right] = \\ &= \sum_{k=0}^{\infty} \frac{\partial^r}{\partial t^r} \left[\frac{t^k}{k!} E(X^k) \right] = \sum_{k=0}^{\infty} E \left[\frac{\partial^r}{\partial t^r} \left(\frac{t^k}{k!} X^k \right) \right] = E \left(\frac{\partial^r}{\partial t^r} e^{tX} \right) = \\ &= E(X^r e^{tX}) \end{aligned}$$

$$\Rightarrow M_X^{(r)}(0) = E(X^r)$$

- Algunos de los ejemplos más comunes de funciones generadoras de momentos son las siguientes:

- Si X tiene una distribución de Bernoulli con parámetro p , entonces la función generadora de probabilidad es la siguiente:

$$M_X(t) = \sum_x e^{tx} P(X=x) = q + pe^t \text{ for } t \in \mathbb{R}$$

- Si X tiene una distribución binomial con parámetros p y n , entonces la función generadora de probabilidad es la siguiente:

$$\begin{aligned}
M_X(t) &= \sum_{k=0}^n e^{tk} \binom{n}{k} p^k q^{n-k} = \\
&= \sum_{k=0}^n (e^t)^k \binom{n}{k} p^k q^{n-k} = e^t \sum_{k=0}^n \binom{n}{k} (pe^t)^k q^{n-k} = \\
&= (q + pe^t)^n \text{ for } t \in \mathbb{R}
\end{aligned}$$

- Si X tiene una distribución de Poisson con parámetro λ , entonces la función generadora de probabilidad es la siguiente:

$$\begin{aligned}
M_X(t) &= \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} (e^t)^k \frac{\lambda^k}{k!} = \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = e^{-\lambda} e^{\lambda e^t} = e^{-\lambda + \lambda e^t} = \\
&= e^{\lambda(e^t - 1)} \text{ for } t \in \mathbb{R}
\end{aligned}$$

- Si X tiene una distribución exponencial con parámetro μ , entonces la función generadora de probabilidad es la siguiente:

$$\begin{aligned}
M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \mu e^{-\mu x} dx = \int_{-\infty}^{\infty} \mu e^{-(\mu-t)x} dx \\
&= \mu \int_{-\infty}^{\infty} e^{-(\mu-t)x} dx = \frac{\mu}{\mu-t} \int_{-\infty}^{\infty} -v e^{-v} dv = \frac{\mu}{\mu-t} \text{ for } t \in \mathbb{R}
\end{aligned}$$

- Si X tiene una distribución normal estándar $N(0,1)$, entonces la función generadora de probabilidad es la siguiente:

$$\begin{aligned}
M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2-2tx}{2}} dx = \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2-2tx+t^2}{2}} e^{\frac{1}{2}t^2} dx = e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-t)^2}{2}} dx = \\
&= e^{\frac{1}{2}t^2} \text{ for } t \in \mathbb{R}
\end{aligned}$$

- Si X tiene una distribución normal estándar $N(m, \sigma^2)$, entonces la función generadora de probabilidad es la siguiente:

$$\begin{aligned}
M_X(t) &= E(e^{tX}) = E(e^{tm + t\sigma Z}) = E(e^{tm+t\sigma Z}) = e^{tm} E(e^{t\sigma Z}) = \\
&= e^{tm} E(e^{t\sigma Z}) = e^{tm} E(e^{t\sigma Z}) = e^{tm} M_Z(t\sigma) = e^{tm} e^{\frac{1}{2}\sigma^2 t^2} =
\end{aligned}$$

$$= e^{tm + \frac{1}{2}\sigma^2 t^2} \text{ for } t \in \mathbb{R}$$

- Igual que con las funciones generadoras de probabilidad, el teorema de convolución también se puede aplicar a las funciones generadoras de momentos para sumas de variables
 - Si X e Y son variables aleatorias independientes, entonces su suma tiene la siguiente función generadora de momentos:

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

- La demostración de este resultado se basa en que la función generadora de probabilidad es igual a la esperanza de s^X y a que, si las variables son independientes, entonces $E(XY) = E(X)E(Y)$

$$\begin{aligned} M_{X+Y}(t) &= E(s^{t(X+Y)}) = E(s^{tX}s^{tY}) = E(s^{tX})E(s^{tY}) = \\ &= M_X(t)M_Y(t) \end{aligned}$$

- A partir de este teorema, la suma $S_n = X_1 + X_2 + \dots + X_n$ de n variables independientes tiene la siguiente función generadora de momentos:

$$M_{S_n}(t) = M_{X_1}(t)M_{X_2}(t) \dots M_{X_n}(t) = \prod_{j=1}^n M_{X_j}(t)$$

- La demostración de este teorema es sencilla porque sigue una lógica similar al caso de dos variables:

$$\begin{aligned} M_{S_n}(t) &= E(e^{tS_n}) = E(e^{tX_1+tX_2+\dots+tX_n}) = \\ &= E(e^{tX_1}e^{tX_2} \dots e^{tX_n}) = E\left(\prod_{j=1}^n e^{tX_j}\right) = \prod_{j=1}^n E(e^{tX_j}) = \\ &= \prod_{j=1}^n M_{X_j}(t) \end{aligned}$$

- Un corolario útil de este teorema de la convolución es que, si todas las variables se distribuyen igual (se pueden entender como n variables iguales X), entonces solo hace falta elevar a n la función $G_X(s)$

$$M_{S_n}(t) = E(e^{tS_n}) = \prod_{j=1}^n M_{X_j}(t) = \prod_{j=1}^n M_X(t) = [M_X(t)]^n$$

- Siendo N y X_1, X_2, \dots, X_N variables aleatorias independientes, si las X_i se distribuyen idénticamente con una función generadora de momentos $M_X(t)$ común, entonces la suma $S = X_1 + X_2 + \dots + X_N$ tiene la siguiente función generadora de momentos:

$$M_S(t) = G_N(M_X(t))$$

- Este teorema se puede demostrar a partir del teorema de la probabilidad total, de modo que se coge una $N = n$ y se usa como condición para la esperanza de e^{tS}

$$\begin{aligned} M_S(t) &= E(e^{tS}) = E_N[E(e^{tS}|N = n)] = \\ &= \sum_{n=0}^{\infty} E(e^{t(X_1+X_2+\dots+X_N)}|N = n)P(N = n) = \\ &= \sum_{n=0}^{\infty} E[e^{t(X_1+X_2+\dots+X_n)}]P(N = n) = \\ &= \sum_{n=0}^{\infty} E(e^{tnX})P(N = n) = \sum_{n=0}^{\infty} [E(e^{tX})]^n P(N = n) = \\ &= G_N[M_X(t)] \end{aligned}$$

- La fórmula encontrada anteriormente para la esperanza se mantiene también para la función generadora de momentos:

$$\begin{aligned} E(S) &= E_N[E(S|N = n)] = E_N[E(X_1 + X_2 + \dots + X_N|N = n)] \\ &= E_N[ENX|N = n)] = E_N[ENX] = E_N[nE(X)] = E(N)E(X) \end{aligned}$$

- Este teorema, sin embargo, no se mantiene si las variables no están distribuidas de manera idéntica, ya que la función generadora de probabilidad no sería la misma para cada una de las variables de la suma
- También se puede definir la función generadora de momentos para distribuciones multivariante, de modo que se hace más fácil operar con estas
 - Siendo $\mathbf{x} = (X_1, X_2, \dots, X_n)$ un vector aleatorio, la función generadora de momentos de \mathbf{x} se define de la siguiente manera:

$$M_{X_1, X_2, \dots, X_n}(t_1, t_2, \dots, t_n) = E[e^{t_1 X_1 + t_2 X_2 + \dots + t_n X_n}]$$

- Esto se puede escribir de manera más compacta utilizando su forma matricial

$$M_{\mathbf{x}}(\mathbf{t}) = E(e^{\mathbf{t}'\mathbf{x}})$$

- Las funciones características marginales para cada variable se pueden obtener a través de la función característica conjunta

- Dadas dos variables X e Y , la función característica marginal para cada variable viene dada por la siguiente fórmula:

$$M_X(t_1) = E(e^{t_1 X}) = E(e^{t_1 X + t_2 Y})|_{t_2=0} = M_{XY}(t_1, 0)$$

$$M_Y(t_2) = E(e^{t_2 Y}) = E(e^{t_1 X + t_2 Y})|_{t_1=0} = M_{XY}(0, t_2)$$

- Las variables aleatorias X_1, X_2, \dots, X_n son independientes si, y solo si, se cumple la siguiente identidad:

$$M_{\mathbf{x}}(\mathbf{t}) = M_{X_1}(t_1)M_{X_2}(t_2) \dots M_{X_n}(t_n) = \prod_{i=1}^n M_{X_i}(t_i)$$

- Esta identidad se puede probar fácilmente con la definición de la función característica y de la independencia

$$\begin{aligned} M_{\mathbf{x}}(\mathbf{t}) &= E[e^{t_1 X_1 + t_2 X_2 + \dots + t_n X_n}] = E(e^{t_1 X_1} e^{t_2 X_2} \dots e^{t_n X_n}) = \\ &= E(e^{t_1 X_1}) E(e^{t_2 X_2}) \dots E(e^{t_n X_n}) = M_{X_1}(t_1) M_{X_2}(t_2) \dots M_{X_n}(t_n) \end{aligned}$$

$$= \prod_{i=1}^n M_{X_i}(t_i)$$

- No obstante, es importante ver como esta expansión del producto en diferentes esperanzas no es posible si no hay independencia, dado que la covarianza no sería nula y eso haría que $E(X_i X_j) \neq E(X_i)E(X_j)$, de modo que $E(e^{t_1 X_1 + t_2 X_2}) \neq E(e^{t_1 X_1})E(e^{t_2 X_2})$

- **DOS DESIGUALDADES**

Commented [MOU7]: Acabar bien PROBABILIDAD

La función característica

- Existen distribuciones para las cuales la función generadora de momentos no existe, por lo que no es siempre posible utilizar la función generadora de momentos. No obstante, con una pequeña modificación de la definición, es posible obtener una función que existe siempre: la función característica

- La función característica ϕ_X de una variable aleatoria X se define de la siguiente manera, donde $i = \sqrt{-1}$:

$$\phi_X(t) = E(e^{itX}) \text{ for } t \in \mathbb{R}$$

- A partir de la definición de la exponencial compleja e^{itX} se puede ver como existe la esperanza para esta función

$$E(e^{itX}) = E[\cos(tX)] + iE[\sin(tX)]$$

- Esta función existe para cualquier variable aleatoria, y ϕ_X está bien definida para toda $X \in \mathbb{R}$. Además, la función $\phi_X(t)$ es uniformemente continua en \mathbb{R}
- Si la función generadora de momentos $M_X(t)$ es finita en una vecindad del origen no trivial (es un ejercicio), la función característica de X puede ser encontrada al fijar $s = it$ en la fórmula $M_X(s)$:

$$\phi_X(t) = M_X(it) \text{ for } t \in \mathbb{R}$$

- La finitud de este último es cuestionable ya que la función exponencial no está acotada, por lo que e^{itX} puede ser muy grande
- Si X es una variable aleatoria discreta, entonces la función se define de la siguiente manera:

$$\phi_X(t) = \begin{cases} \sum_x e^{itx} P(X=x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{itx} f_X(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

- Cuando X es una variable aleatoria continua, $\phi_X(t)$ es la transformada de Fourier de su densidad $f_X(x)$ (aunque la exponencial tiene el signo cambiado con respecto a la definición usual de la transformada de Fourier)
- Para variables discretas, las funciones características están relacionadas a las series de Fourier

- Esta función cumple con unas propiedades derivadas de su definición en los números complejos

- La función característica $\phi_X(t)$ yace en el círculo unitario del plano complejo, de modo que $|e^{itX}| = 1$ y $|\phi_X(t)| \leq 1$ para toda $t \in \mathbb{R}$

$$|\phi_X(t)| = |E(e^{itX})| \leq E(|e^{itX}|) = E(1) = 1$$

- La función característica evaluada en cero $\phi_X(0)$ es igual a la unidad

$$\phi_X(0) = E(e^{i0X}) = E(1) = 1$$

- El conjugado de la función $\phi_X(t)$ es $\phi_X(-t)$ debido a su definición

$$\overline{\phi_X(t)} = \overline{E(e^{itX})} = E(\overline{e^{itX}}) = E(e^{-itX}) = E(e^{i(-t)X}) = \phi_X(-t)$$

- Algunos de los ejemplos más comunes de funciones características son las siguientes:

- Si X tiene una distribución de Bernoulli con parámetro p , entonces la función generadora de probabilidad es la siguiente:

$$\phi_X(t) = \sum_x e^{itx} P(X=x) = q + pe^{it} \quad \text{for } t \in \mathbb{R}$$

- Si X tiene una distribución binomial con parámetros p y n , entonces la función generadora de probabilidad es la siguiente:

$$\begin{aligned} \phi_X(t) &= \sum_{k=0}^n e^{itk} \binom{n}{k} p^k q^{n-k} = \\ &= \sum_{k=0}^n (e^{it})^k \binom{n}{k} p^k q^{n-k} = e^{it} \sum_{k=0}^n \binom{n}{k} (pe^{it})^k q^{n-k} = \\ &= (q + pe^{it})^n \quad \text{for } t \in \mathbb{R} \end{aligned}$$

- Si X tiene una distribución de Poisson con parámetro λ , entonces la función generadora de probabilidad es la siguiente:

$$\phi_X(t) = \sum_{k=0}^{\infty} e^{itk} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} (e^{it})^k \frac{\lambda^k}{k!} =$$

$$\begin{aligned}
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^{it})^k}{k!} = e^{-\lambda} e^{\lambda e^{it}} = e^{-\lambda + \lambda e^{it}} = \\
&= e^{\lambda(e^{it}-1)} \text{ for } t \in \mathbb{R}
\end{aligned}$$

- Si X tiene una distribución exponencial con parámetro μ , entonces la función generadora de probabilidad es la siguiente:

$$\begin{aligned}
\phi_X(t) &= \int_{-\infty}^{\infty} e^{itx} \mu e^{-\mu x} dx = \int_{-\infty}^{\infty} \mu e^{-(\mu-it)x} dx \\
&= \mu \int_{-\infty}^{\infty} e^{-(\mu-it)x} dx = \frac{\mu}{\mu-it} \int_{-\infty}^{\infty} -v e^{-v} dv = \frac{\mu}{\mu-it} \\
&\text{for } t \in \mathbb{R}
\end{aligned}$$

- Si X tiene una distribución normal estándar $N(0,1)$, entonces la función generadora de probabilidad es la siguiente:

$$\begin{aligned}
\phi_X(t) &= \int_{-\infty}^{\infty} e^{itx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2+2itx}{2}} dx = \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2+2itx+t^2}{2}} e^{-\frac{1}{2}t^2} dx = \\
&= e^{-\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-it)^2}{2}} dx = e^{-\frac{1}{2}t^2} \text{ for } t \in \mathbb{R}
\end{aligned}$$

- Si X tiene una distribución normal estándar $N(m, \sigma^2)$, entonces la función generadora de probabilidad es la siguiente:

$$\begin{aligned}
\phi_X(t) &= E(e^{itX}) = E(e^{itm+it\sigma Z}) = E(e^{itm+it\sigma Z}) = \\
&= e^{itm} E(e^{it\sigma Z}) = e^{itm} E(e^{it\sigma Z}) = e^{itm} E(e^{it\sigma Z}) = \\
&= e^{itm} \phi_Z(t\sigma) = e^{itm} e^{-\frac{1}{2}\sigma^2 t^2} = e^{itm - \frac{1}{2}\sigma^2 t^2} \text{ for } t \in \mathbb{R}
\end{aligned}$$

- Es posible hacer una expansión para la función característica a través de la expansión de la exponencial compleja, siempre que todos los momentos existan y sean finitos, de modo que $E(|X^k|) < \infty$ para toda $k = 1, 2, \dots$

$$\phi_X(t) = E(e^{itX}) = E\left[1 + itX + \frac{1}{2!}(itX)^2 + \frac{1}{3!}(itX)^3 + \dots\right] =$$

$$= E \left[\sum_{k=0}^{\infty} \frac{(itX)^k}{k!} \right] = \sum_{k=0}^{\infty} \frac{E[(itX)^k]}{k!} = \sum_{k=0}^{\infty} \frac{E(X^k)}{k!} (it)^k$$

- Si $E(X^k) < \infty$ para alguna $n \geq 1$, entonces se puede expresar la función $\phi_X(t)$ en términos de sus n primeros términos y de un término de error insignificante $o(|t|^n)$

$$\phi_X(t) = \sum_{k=0}^n \frac{E(X^k)}{k!} (it)^k + o(|t|^n) \quad \text{as } t \rightarrow 0$$

- Derivando k veces y evaluando $\phi_X(t)$ en cero, se puede obtener el siguiente resultado:

$$\phi_X^{(k)}(0) = i^k E(X^k)$$

- Este resultado se demuestra derivando y utilizando la expansión de la función exponencial

$$\begin{aligned} \phi_X^{(k)}(t) &= \frac{\partial^k}{\partial t^k} \phi_X(t) = \frac{\partial^k}{\partial t^k} \sum_{k=0}^{\infty} \frac{E(X^k)}{k!} (it)^k = \\ &= \sum_{k=0}^{\infty} \frac{\partial^k}{\partial t^k} \left[\frac{E(X^k)}{k!} (it)^k \right] = E \left(\frac{\partial^k}{\partial t^k} e^{itX} \right) = E[(iX)^k e^{itX}] \end{aligned}$$

$$\Rightarrow \phi_X^{(k)}(0) = E[(iX)^k e^{i0X}] = i^k E(X^k)$$

- Esta identidad permite obtener el momento k para cualquier $k = 1, 2, \dots$ evaluando la función $\phi_X^{(k)}(t)$ en cero

$$\phi_X^{(k)}(0) = i^k E(X^k) \Rightarrow E(X^k) = \frac{\phi_X^{(k)}(0)}{i^k}$$

- Igual que con las funciones generadoras de momentos, el teorema de convolución también se puede aplicar a las funciones características para sumas de variables

- Siendo X_1, X_2, \dots, X_n variables aleatorias independientes y definiendo la suma de estas n variables como $S = X_1 + X_2 + \dots + X_n$, entonces se cumple la siguiente identidad:

$$\phi_S(t) = \phi_{X_1}(t) \phi_{X_2}(t) \dots \phi_{X_n}(t) = \prod_{i=1}^n \phi_{X_i}(t)$$

- La demostración de este teorema se basa en la definición de la función

$$\begin{aligned}\phi_S(t) &= E(e^{itS}) = E(e^{it(X_1+X_2+\dots+X_n)}) = \\ &= E(e^{itX_1} e^{itX_2} \dots e^{itX_n}) = E(e^{itX_1}) E(e^{itX_2}) \dots E(e^{itX_n}) = \\ &= \phi_{X_1}(t) \phi_{X_2}(t) \dots \phi_{X_n}(t) = \prod_{i=1}^n \phi_{X_i}(t)\end{aligned}$$

- Este es el teorema de la convolución aplicado a las transformaciones de Fourier, de modo que X e Y son continuas y variables aleatorias independientes y $Z = X + Y$, entonces $f_Z = f_X f_Y$, lo cual implica las siguientes identidades:

$$\mathcal{F}(f_Z) = \mathcal{F}(f_X) \mathcal{F}(f_Y) \Rightarrow \phi_S(t) = \phi_X(t) \phi_Y(t)$$

- Siendo X_1, X_2, \dots, X_n variables aleatorias independientes y equidistribuidas y definiendo la suma de estas n variables como $S = X_1 + X_2 + \dots + X_n$, entonces se cumple la siguiente identidad:

$$\phi_S(t) = [\phi_X(t)]^n$$

- La demostración de este teorema se basa en la definición de la función

$$\begin{aligned}\phi_S(t) &= E(e^{itS}) = E(e^{it(X_1+X_2+\dots+X_n)}) = \\ &= E(e^{itX_1} e^{itX_2} \dots e^{itX_n}) = E(e^{itX_1}) E(e^{itX_2}) \dots E(e^{itX_n}) = \\ &= \prod_{i=1}^n \phi_{X_i}(t) = \prod_{i=1}^n \phi_X(t) = [\phi_X(t)]^n\end{aligned}$$

- Siendo N y X_1, X_2, \dots, X_N variables aleatorias independientes, si las X_i se distribuyen idénticamente con una función característica $\phi_X(t)$ común, entonces la suma $S = X_1 + X_2 + \dots + X_N$ tiene la siguiente función característica:

$$\phi_S(t) = G_N(\phi_X(t))$$

- Este teorema se puede demostrar a partir del teorema de la probabilidad total, de modo que se coge una $N = n$ y se usa como condición para la esperanza de e^{itS}

$$\begin{aligned}
\phi_S(t) &= E(e^{itS}) = E_N[E(e^{itS}|N=n)] = \\
&= \sum_{n=0}^{\infty} E(e^{it(X_1+X_2+\dots+X_N)}|N=n)P(N=n) = \\
&= \sum_{n=0}^{\infty} E[e^{it(X_1+X_2+\dots+X_n)}]P(N=n) = \\
&= \sum_{n=0}^{\infty} E(e^{itnX})P(N=n) = \sum_{n=0}^{\infty} [E(e^{itX})]^n P(N=n) = \\
&= G_N[\phi_X(t)]
\end{aligned}$$

- La fórmula encontrada anteriormente para la esperanza se mantiene también para la función característica:

$$\begin{aligned}
E(S) &= E_N[E(S|N=n)] = E_N[E(X_1 + X_2 + \dots + X_N|N=n)] \\
&= E_N[E(NX|N=n)] = E_N[E(nX)] = E_N[nE(X)] = E(N)E(X)
\end{aligned}$$

- Este teorema, sin embargo, no se mantiene si las variables no están distribuidas de manera idéntica, ya que la función generadora de probabilidad no sería la misma para cada una de las variables de la suma
- Debido a la relación inherente entre la función característica y la transformada de Fourier, es posible utilizar teoremas de inversión que permitan recuperar una función de densidad o distribución a partir de $\phi_X(t)$
 - Siendo X una variable continua aleatoria con densidad $f_X(x)$ y una función característica $\phi_X(t)$, entonces se cumple la siguiente identidad para todos los puntos x en la que $f_X(x)$ es diferenciable:

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-it} \phi_X(t) dt$$

- Para obtener $f_X(x)$ a partir de $\phi_X(t)$ normalmente requiere integración de contorno en el plano complejo
- En el caso discreto, es posible utilizar las series de Fourier para poder invertir $\phi_X(t)$. Si X es una variable aleatoria que toma valores enteros, entonces se cumple la siguiente identidad:

$$P(X = x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itx} \phi_X(t) dt$$

- El resultado general relacionado a la inversión de la función característica es el siguiente: siendo X una variable con función de distribución $F_X(x)$ y función característica $\phi_X(t)$, y definiendo una función de distribución $\bar{F}_X(x) = \frac{1}{2}(F_X(x) + F_X(x^-))$, entonces se da la siguiente identidad:

$$\bar{F}_X(b) - \bar{F}_X(a) = \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \phi_X(t) dt$$

- ...
- El teorema de unicidad de la función característica es un corolario de este teorema, dado que este proporciona una fórmula para calcular explícitamente la función de distribución en términos de la función característica
- Este teorema de unicidad es el siguiente: siendo X e Y variables aleatorias con funciones características ϕ_X e ϕ_Y respectivamente, entonces X e Y tienen la misma distribución si, y solo si, $\phi_X = \phi_Y$ para toda $t \in \mathbb{R}$

Commented [MOU8]: Por qué se hace lo de la F si en el libro no sale así, y que es x-

- También se puede definir la función característica para distribuciones multivariante, de modo que se hace más fácil operar con estas
 - Siendo $\mathbf{x} = (X_1, X_2, \dots, X_n)$ un vector aleatorio, la función característica de \mathbf{x} se define de la siguiente manera:

$$\phi_{X_1, X_2, \dots, X_n}(t_1, t_2, \dots, t_n) = E[e^{i(t_1 X_1 + t_2 X_2 + \dots + t_n X_n)}]$$

- Esto se puede escribir de manera más compacta utilizando su forma matricial

$$\phi_{\mathbf{x}}(\mathbf{t}) = E(e^{i\mathbf{t}'\mathbf{x}})$$

- La función característica conjunta permite el cálculo de momentos conjuntos. Dadas dos variables X e Y , la función característica viene dada por la siguiente fórmula:

$$\phi_{XY}(t_1, t_2) = E(e^{i(t_1 X + t_2 Y)})$$

- Por lo tanto, si se evalúan las derivadas para t_1 y t_2 (la derivada cruzada) en $t_1 = t_2 = 0$, entonces se puede obtener la siguiente identidad:

$$\frac{\partial \phi_{XY}(t_1, t_2)}{\partial t_1 \partial t_2} = i^2 E(XY e^{i(t_1 X + t_2 Y)}) = -E(XY e^{i(t_1 X + t_2 Y)})$$

$$\Rightarrow \left. \frac{\partial \phi_{XY}(t_1, t_2)}{\partial t_1 \partial t_2} \right|_{t_1, t_2=0} = -E(XY)$$

- De manera más general, se puede obtener la siguiente identidad, la cual es una generalización del resultado anterior para obtener la esperanza de una sola variable:

$$E(X^k Y^r) = \frac{1}{i^{k+r}} \left. \frac{\partial^{k+r} \phi_{XY}(t_1, t_2)}{\partial t_1^k \partial t_2^r} \right|_{t_1, t_2=0}$$

- Las funciones características marginales para cada variable se pueden obtener a través de la función característica conjunta

- Dadas dos variables X e Y , la función característica marginal para cada variable viene dada por la siguiente fórmula:

$$\phi_X(t_1) = E(e^{it_1 X}) = E(e^{i(t_1 X + t_2 Y)})|_{t_2=0} = \phi_{XY}(t_1, 0)$$

$$\phi_Y(t_2) = E(e^{it_2 Y}) = E(e^{i(t_1 X + t_2 Y)})|_{t_1=0} = \phi_{XY}(0, t_2)$$

- Las variables aleatorias X_1, X_2, \dots, X_n son independientes si, y solo si, se cumple la siguiente identidad:

$$\phi_{\mathbf{x}}(\mathbf{t}) = \phi_{X_1}(t_1) \phi_{X_2}(t_2) \dots \phi_{X_n}(t_n) = \prod_{i=1}^n \phi_{X_i}(t_i)$$

- Esta identidad se puede probar fácilmente con la definición de la función característica y de la independencia

$$\begin{aligned} \phi_{\mathbf{x}}(\mathbf{t}) &= E[e^{i(t_1 X_1 + t_2 X_2 + \dots + t_n X_n)}] = E(e^{it_1 X_1} e^{it_2 X_2} \dots e^{it_n X_n}) = \\ &= E(e^{it_1 X_1}) E(e^{it_2 X_2}) \dots E(e^{it_n X_n}) = \phi_{X_1}(t_1) \phi_{X_2}(t_2) \dots \phi_{X_n}(t_n) \end{aligned}$$

$$= \prod_{i=1}^n \phi_{X_i}(t_i)$$

- No obstante, es importante ver como esta expansión del producto en diferentes esperanzas no es posible si no hay independencia, dado que la covarianza no sería nula y eso haría que $E(X_i X_j) \neq E(X_i)E(X_j)$, de modo que $E(e^{it_1 X_1 + it_2 X_2}) \neq E(e^{it_1 X_1})E(e^{it_2 X_2})$

La distribución normal multivariante

- La distribución normal multivariante es la distribución multivariante más importante debida a sus aplicaciones, propiedades y métodos. Hay varias definiciones posibles para esta distribución, por lo que se presenta la primera, que hace referencia a la normalidad de la combinación lineal de los componentes
 - El vector aleatorio n -dimensional \mathbf{x} es normal si, y solo si, para todo vector n -dimensional \mathbf{a} , la variable aleatoria unidimensional $\mathbf{a}'\mathbf{x}$ es normal
 - La notación $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \Sigma)$ se usa para denotar que \mathbf{x} tiene una distribución normal multivariante con media $\boldsymbol{\mu}$ y matriz de varianza y covarianza Σ
 - La distribución $\mathbf{a}'\mathbf{x}$ depende, por supuesto, de \mathbf{a} . La distribución normal degenerada (con varianza nula) también se incluye en una distribución posible de $\mathbf{a}'\mathbf{x}$
 - No se hace ninguna suposición sobre la independencia de los componentes de \mathbf{x}
 - Esta definición es muy aplicable, y se pueden derivar unas propiedades que son consecuencias inmediatas de la definición. Si el vector aleatorio es normal, entonces se cumplen las siguientes propiedades:
 - (a) *Every component of \mathbf{x} is normal*
 - (b) $x_1 + x_2 + \dots + x_n$ is normal
 - (c) *Every marginal distribution is normal*
 - Para poder demostrar que x_k es normal para $k = 1, 2, \dots, n$ se escoge \mathbf{a} tal que $a_k = 1$ y $a_j = 0$ de otro modo. Además, se puede ver como la suma de los componentes es normal escogiendo \mathbf{a} tal que $a_k = 1$ para $k = 1, 2, \dots, n$
 - También se sabe que una combinación lineal de variables aleatorias normales independientes es normal (a través del teorema de la

convolución y la función característica), por lo que, si el vector aleatorio es normal, se cumple la siguiente propiedad:

(d) *x is normal if x has independent normal components*

- Por lo tanto, si x tiene componentes normales independientes entonces x es normal. Esta propiedad, permite ver, entonces, demostrar que un vector aleatorio es normal si sus componentes son independientes
- Como se verá más adelante, la independencia entre componentes se puede comprobar a través de la correlación, dado que esta tiene que ser nula para que los componentes sean independientes
- Para que todas las distribuciones marginales de un vector aleatorio X normal sean normales, la suposición de normalidad conjunta de todos los componentes de X es vital
 - Dos o más variables pueden ser marginalmente normales, pero no conjuntamente, de modo que el converso general del lema no se mantiene y existen variables normales que no son conjuntamente normales
 - Si las propiedades (a) – (c) se cumplen, no quiere decir que el vector aleatorio sea normal o que las variables sean conjuntamente normales, por lo que no se puede demostrar la normalidad de un vector con estas
 - Un contraejemplo famoso se presenta definiendo $X \sim N(0,1)$, Z como una variable binaria tal que $P(Z = 1) = P(Z = -1) = 1/2$ y $Y = ZX$. La función de distribución de Y se puede obtener de la siguiente manera:

$$\begin{aligned}
 P(Y \leq x) &= \frac{1}{2}P(X \leq x) + \frac{1}{2}P(-X \leq x) = \\
 &= \frac{1}{2}P(X \leq x) + \frac{1}{2}P(X \geq -x) = \frac{1}{2}\Phi(x) + \frac{1}{2}[1 - \Phi(-x)] = \\
 &= \frac{1}{2}\Phi(x) + \frac{1}{2}[1 - \Phi(-x)] = \Phi(x)
 \end{aligned}$$

- No obstante, aunque $Y \sim N(0,1)$, $X + Y$ no es normal, dado que se puede obtener la probabilidad positiva para el valor concreto 0 (cuando tendría que ser nula por definición). De este modo, (X, Y) no es un vector aleatorio normal porque la suma de sus

componentes no es normal (en consecuencia, X e Y no son conjuntamente normales)

$$P(X + Y = 0) = P(Z = -1) = \frac{1}{2}$$

- Suponiendo que $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$ y se escoge $\mathbf{y} = \mathbf{B}\mathbf{x} + \mathbf{c}$, en donde \mathbf{B} es una matriz $n \times n$ y \mathbf{d} es un vector de constantes n -dimensional, entonces $\mathbf{y} \sim N(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\Sigma\mathbf{B}')$

- La combinación lineal de los componentes de \mathbf{y} es otra combinación lineal de los componentes de \mathbf{x}

$$\mathbf{a}'\mathbf{y} = \mathbf{a}'\mathbf{B}\mathbf{x} + \mathbf{a}'\mathbf{b}$$

- Fijando $\mathbf{c} = \mathbf{a}'\mathbf{B}$ y $\mathbf{d} = \mathbf{a}'\mathbf{d}$, como $\mathbf{c}'\mathbf{x}$ es normal acorde a la definición (y \mathbf{d} es una constante), entonces $\mathbf{a}'\mathbf{y}$ es normal (es una combinación lineal de un vector normal)
- Una consecuencia obvia del teorema anterior es que si $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$ y las matrices \mathbf{A} y \mathbf{B} son tales que $\mathbf{A} = \mathbf{B}$, entonces $\mathbf{A}\mathbf{x} = \mathbf{B}\mathbf{x}$ (tienen la misma distribución). No obstante, el converso de este no se mantiene
 - Esto ocurre porque se pueden encontrar dos matrices diferentes \mathbf{A} y \mathbf{B} tal que $\mathbf{A}\boldsymbol{\mu} = \mathbf{B}\boldsymbol{\mu}$ y que $\mathbf{A}\Sigma\mathbf{A}' = \mathbf{B}\Sigma\mathbf{B}'$ (hay diferentes combinaciones lineales, desde el punto de vista de las filas, que pueden llevar al mismo resultado)
 - Un ejemplo claro es utilizando variables normales estándar $N(0,1)$, ya que de esta manera la esperanza siempre será igual pero solo es necesario que $\mathbf{A}'\mathbf{A} = \mathbf{B}'\mathbf{B}$ (porque $\Sigma = \mathbf{I}_{n \times n}$)

- Otra manera de poder definir la distribución normal multivariante es a través de su función característica

- Tal y como se ha visto anteriormente, la función característica de un vector aleatorio es $\phi_{\mathbf{x}}(\mathbf{t}) = E(e^{i\mathbf{t}'\mathbf{x}})$. Suponiendo que $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$, entonces se puede ver como $\mathbf{t}'\mathbf{x}$ sigue una distribución normal univariante por la definición anterior
 - Los parámetros de esta distribución normal serán $m = E(\mathbf{t}'\mathbf{x}) = \mathbf{t}'\boldsymbol{\mu}$ y $\sigma^2 = \text{Cov}(\mathbf{t}'\mathbf{x}) = \mathbf{t}'\Sigma\mathbf{t}$, de modo que se obtiene la siguiente función:

$$\phi_{\mathbf{x}}(\mathbf{t}) = \phi_{\mathbf{t}'\mathbf{x}}(1) = E(e^{it'x}) = e^{im} \frac{1}{\sqrt{2\sigma^2}}$$

- A partir de eso, se puede establecer que, para $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$, se obtiene el siguiente resultado:

$$\phi_{\mathbf{x}}(\mathbf{t}) = E(e^{it'\mathbf{x}}) = e^{it'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}}$$

- También se puede establecer un resultado converso a este último y así obtener una definición equivalente de la distribución normal multivariante

- Para cualquier matriz semidefinida positiva Σ , la función característica de un vector aleatorio \mathbf{x} con $E(\mathbf{x}) = \boldsymbol{\mu}$ y $Cov(\mathbf{x}) = \Sigma$ es la siguiente:

$$\phi_{\mathbf{x}}(\mathbf{t}) = e^{it'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}}$$

- Siendo \mathbf{y} un vector aleatorio cuyos componentes y_1, y_2, \dots, y_n son independientes y distribuidos siguiendo una $N(0,1)$ y fijando $\mathbf{x} = \Sigma^{1/2}\mathbf{y} + \boldsymbol{\mu}$ se pueden obtener las siguientes igualdades:

$$E(\mathbf{x}) = \Sigma^{1/2}E(\mathbf{y}) + \boldsymbol{\mu} = \boldsymbol{\mu}$$

$$Cov(\mathbf{x}) = \Sigma^{1/2}Cov(\mathbf{y})\Sigma^{1/2} = \Sigma^{1/2}\Sigma^{1/2} = \Sigma$$

- Por lo tanto, es posible obtener las siguientes expresiones para $\phi_{\mathbf{y}}(\mathbf{t})$ y $\phi_{\mathbf{x}}(\mathbf{t})$ sin asumir en ningún momento normalidad para \mathbf{x} ni que los componentes de \mathbf{x} son independientes:

$$y_1, \dots, y_n \text{ iid } N(0,1) \Rightarrow \phi_{\mathbf{y}}(\mathbf{t}) = e^{-\frac{1}{2}\mathbf{t}'\mathbf{t}}$$

$$\Rightarrow \phi_{\mathbf{x}}(\mathbf{t}) = E(e^{it'\mathbf{x}}) = E(e^{it'(\Sigma^{1/2}\mathbf{y} + \boldsymbol{\mu})}) = E(e^{it'\Sigma^{1/2}\mathbf{y} + it'\boldsymbol{\mu}}) =$$

$$= e^{it'\boldsymbol{\mu}} E(e^{i(\Sigma^{1/2}\mathbf{t})'\mathbf{y}}) = e^{it'\boldsymbol{\mu}} \phi_{\mathbf{y}}(\Sigma^{1/2}\mathbf{t}) = e^{it'\boldsymbol{\mu}} e^{-\frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}} =$$

$$= e^{it'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}}$$

- Si ahora se considera una combinación lineal $\mathbf{a}'\mathbf{x}$ para un vector \mathbf{a} n -dimensional, se puede ver como $\mathbf{a}'\mathbf{x} \sim N(m, \sigma^2)$ porque es una combinación lineal que satisface (a) – (d), lo cual demuestra a su vez que \mathbf{x} se distribuye como una normal multivariante

$$\phi_{\mathbf{a}'\mathbf{x}}(u) = E(e^{iua'\mathbf{x}}) = e^{iu(\mathbf{a}'\boldsymbol{\mu}) - \frac{1}{2}(u\mathbf{a})'\Sigma(u\mathbf{a})} = e^{ium - \frac{1}{2}u^2\sigma^2}$$

where $m = \mathbf{a}'\boldsymbol{\mu}$ & $\sigma^2 = \mathbf{a}'\Sigma\mathbf{a} \geq 0$

- Alternativamente, se puede justificar esto viendo que $\mathbf{a}'\mathbf{x}$ es una combinación lineal que resulta en una combinación lineal de los componentes de \mathbf{y} (que es un vector aleatorio normal porque tiene componentes independientes como en la condición (d))

$$\mathbf{a}'\mathbf{x} = \mathbf{a}'\Sigma^{1/2}\mathbf{y} + \mathbf{a}'\boldsymbol{\mu} = (\Sigma^{1/2}\mathbf{a})'\mathbf{y} + \mathbf{a}'\boldsymbol{\mu}$$

- Por lo tanto, un vector aleatorio \mathbf{x} es normal si, y solo si, su función característica tiene la siguiente forma para un vector $\boldsymbol{\mu}$ y matriz semidefinida positiva Σ :

$$\phi_{\mathbf{x}}(\mathbf{t}) = e^{it'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}}$$

- La primera definición de normalidad multivariante y esta segunda definición son equivalentes
- La definición y la expresión para la función generadora de momentos son las obvias

$$M_{\mathbf{x}}(\mathbf{t}) = E(e^{\mathbf{t}'\mathbf{x}}) = e^{\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}}$$

- El procedimiento descrito anteriormente para demostrar el teorema, permite demostrar que un vector aleatorio es normal si sigue se pueden crear los componentes del vector \mathbf{x} a partir del vector de componentes normales estándar independientes \mathbf{y}
- Siendo $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$, si $\det(\Sigma) = 0$, entonces la distribución es singular y no existe una distribución de densidad, pero si $\det(\Sigma) > 0$, entonces existe una función de densidad que está únicamente determinada por $\boldsymbol{\mu}$ y Σ . Esto motiva una tercera definición alternativa

- Con tal de determinar la densidad, es suficiente encontrarla para una distribución normal construida de manera conveniente

- Siendo \mathbf{y} un vector aleatorio n -dimensional con componentes normales estándar e independientes y $\mathbf{x} = \Sigma^{1/2}\mathbf{y} + \boldsymbol{\mu}$ un vector aleatorio normal $N_n(\boldsymbol{\mu}, \Sigma)$, es posible encontrar la función de densidad para \mathbf{y} :

$$f_{\mathbf{y}}(\mathbf{y}) = \prod_{k=1}^n f_{y_k}(y_k) = \prod_{k=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y_k^2}{2}} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{\sum_{k=1}^n y_k^2}{2}} =$$

$$= \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\mathbf{y}'\mathbf{y}} \text{ for } \mathbf{y} \in \mathbb{R}^n$$

- Debido a que $\det(\Sigma) > 0$, se sabe que Σ es una matriz invertible y se pueden considerar las siguientes igualdades:

$$\mathbf{x} = \Sigma^{1/2}\mathbf{y} + \boldsymbol{\mu} \Rightarrow \mathbf{y} = \Sigma^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$$

$$\Rightarrow \det(\Sigma^{1/2}) = [\det(\Sigma)]^{1/2}$$

- Por lo tanto, para $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \Sigma)$ con $\det(\Sigma) > 0$, se puede ver que la expresión de la función de densidad de \mathbf{x} es la siguiente:

$$f_{\mathbf{y}}(\mathbf{y}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2}\mathbf{y}'\mathbf{y}}$$

$$\Rightarrow f_{\mathbf{x}}(\mathbf{x}) = \left(\frac{1}{2\pi\det(\Sigma)}\right)^{\frac{n}{2}} e^{-\frac{1}{2}[\Sigma^{-1/2}(\mathbf{x}-\boldsymbol{\mu})]'[\Sigma^{-1/2}(\mathbf{x}-\boldsymbol{\mu})]}$$

$$\Rightarrow f_{\mathbf{x}}(\mathbf{x}) = \left(\frac{1}{2\pi\det(\Sigma)}\right)^{\frac{n}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

- Para poder calcular la función de densidad marginal, se puede integrar sobre los otros componentes

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \left(\frac{1}{2\pi\det(\Sigma)}\right)^{\frac{n}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})} d\mathbf{x}_{-i}$$

- Un vector aleatorio n -dimensional \mathbf{x} con $E(\mathbf{x}) = \boldsymbol{\mu}$ y $Cov(\mathbf{x}) = \Sigma$ y con $\det(\Sigma) > 0$ se distribuye como una normal multivariante $N_n(\boldsymbol{\mu}, \Sigma)$ si, y solo si, la densidad es la siguiente función:

$$f_{\mathbf{x}}(\mathbf{x}) = \left(\frac{1}{2\pi\det(\Sigma)}\right)^{\frac{n}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})} \text{ for } \mathbf{x} \in \mathbb{R}^n$$

- Se tiene que denotar el hecho de que se eleva el término constante a $n/2$ debido a que el vector \mathbf{x} tiene n componentes, por lo que la definición de la función de densidad variará dependiendo de los componentes aleatorios
- Todas las definiciones dadas, incluyendo esta, son equivalentes, dado que esta definición es consecuencia del teorema de la unicidad de las funciones características en el caso en que $\det(\Sigma) > 0$ (solo hay una única función característica para cada

distribución, de modo que solo puede haber una única función de densidad)

- El procedimiento descrito anteriormente para demostrar el teorema, permite demostrar que un vector aleatorio es normal si se pueden crear los componentes del vector \mathbf{x} a partir del vector de componentes normales estándar independientes \mathbf{y}
- Para calcular explícitamente la función de densidad...
- Siendo $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \Sigma)$ y suponiendo que $\det(\Sigma) > 0$, entonces la densidad de probabilidad existe y se puede definir el concepto de función de densidad condicional para la distribución normal multivariante
 - Se considera un vector columna n -dimensional $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ que sigue una distribución normal multivariante, y en donde los subvectores \mathbf{x}_1 y \mathbf{x}_2 tienen dimensiones r y $n - r$ respectivamente
 - También se considera el vector de medias n -dimensional \mathbf{m} , en donde $\mathbf{m}_1 = E(\mathbf{x}_1)$ y $\mathbf{m}_2 = E(\mathbf{x}_2)$, y \mathbf{K} la matriz de varianzas y covarianzas con submatrices de varianzas y covarianzas \mathbf{K}_{11} y \mathbf{K}_{22} para \mathbf{x}_1 y \mathbf{x}_2 respectivamente (no consideran componentes del otro vector), y también con una submatriz \mathbf{K}_{12} de tamaño $r \times (n - r)$ cuyos elementos son las covarianzas a pares entre los componentes de \mathbf{x}_1 y \mathbf{x}_2

$$\mathbf{m} = \begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{pmatrix} \quad \mathbf{K} = \begin{pmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{12}' & \mathbf{K}_{22} \end{pmatrix}$$

- Asumiendo que $\det(\mathbf{K}_{22}) > 0$, entonces la distribución del vector aleatorio \mathbf{x}_1 condicionada a $\mathbf{x}_2 = \mathbf{z}_2$ es una distribución normal multivariante con un vector de medias y matriz de varianzas y covarianzas dadas por las siguientes expresiones:

$$\mathbf{m}^* = \mathbf{m}_1 + \mathbf{K}_{12}\mathbf{K}_{22}^{-1}(\mathbf{z}_2 - \mathbf{m}_2)$$

$$\mathbf{K}^* = \mathbf{K}_{11} - \mathbf{K}_{12}\mathbf{K}_{22}^{-1}\mathbf{K}_{12}'$$

- Siendo $\mathbf{A} = \begin{pmatrix} \mathbf{I}_{r \times r} & -\mathbf{K}_{12}\mathbf{K}_{22}^{-1} \\ \mathbf{0} & \mathbf{I}_{(n-r) \times (n-r)} \end{pmatrix}$, es posible encontrar el vector de medias y la matriz de varianzas y covarianzas para $\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})$

$$\begin{aligned} \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) &= \begin{pmatrix} \mathbf{I}_{r \times r} & -\mathbf{K}_{12}\mathbf{K}_{22}^{-1} \\ \mathbf{0} & \mathbf{I}_{(n-r) \times (n-r)} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 - \mathbf{m}_1 \\ \mathbf{x}_2 - \mathbf{m}_2 \end{pmatrix} = \\ &= \begin{pmatrix} \mathbf{x}_1 - \mathbf{m}_1 - \mathbf{K}_{12}\mathbf{K}_{22}^{-1}(\mathbf{x}_2 - \mathbf{m}_2) \\ \mathbf{x}_2 - \mathbf{m}_2 \end{pmatrix} \end{aligned}$$

Commented [MOU9]: Acabar bien Intermediate Probability

$$\Rightarrow E[A(x - \mu)] = \mathbf{0}$$

$$\Rightarrow \text{Cov}[A(x - \mu)] = \text{Cov}(Ax) = AK A' =$$

$$\begin{aligned} &= \begin{pmatrix} I_{r \times r} & -K_{12}K_{22}^{-1} \\ \mathbf{0} & I_{(n-r) \times (n-r)} \end{pmatrix} \begin{pmatrix} K_{11} & K_{12} \\ K_{12}' & K_{22} \end{pmatrix} \begin{pmatrix} I_{r \times r} & \mathbf{0} \\ -K_{22}^{-1}K_{12}' & I_{(n-r) \times (n-r)} \end{pmatrix} \\ &= \begin{pmatrix} K_{11} - K_{12}K_{22}^{-1}K_{12}' & \mathbf{0} \\ K_{12}' & K_{22} \end{pmatrix} \begin{pmatrix} I_{r \times r} & \mathbf{0} \\ -K_{22}^{-1}K_{12}' & I_{(n-r) \times (n-r)} \end{pmatrix} = \\ &= \begin{pmatrix} K_{11} - K_{12}K_{22}^{-1}K_{12}' & \mathbf{0} \\ \mathbf{0} & K_{22} \end{pmatrix} \end{aligned}$$

- Debido a que la matriz de varianzas y covarianzas de la parte superior del vector gaussiano $A(x - \mu)$, y este vector sigue una distribución normal multivariante (al ser combinación lineal de un vector gaussiano), se puede ver que esta parte superior sigue una distribución normal multivariante de la siguiente forma:

$$x_1 - m_1 - K_{12}K_{22}^{-1}(x_2 - m_2) \sim N_r(\mathbf{0}, K_{11} - K_{12}K_{22}^{-1}K_{12}')$$

- Además, se puede ver como la matriz de varianzas y covarianzas tiene $\mathbf{0}$ en las covarianzas, de modo que los vectores aleatorios $x_1 - m_1 - K_{12}K_{22}^{-1}(x_2 - m_2)$ y $x_2 - m_2$ son vectores aleatorios normales independientes
- Por lo tanto, si $x_2 = z_2$, entonces el vector aleatorio $x_1 - m_1 - K_{12}K_{22}^{-1}(z_2 - m_2)$ también sigue una distribución normal. Entonces, condicional a $x_2 = z_2$, el vector aleatorio x_1 sigue la siguiente distribución:

$$x_1 \sim N_r(m_1 + K_{12}K_{22}^{-1}(z_2 - m_2), K_{11} - K_{12}K_{22}^{-1}K_{12}')$$

- La forma funcional de la densidad condicional, en este caso, sería la siguiente:

$$f_{x_1|x_2=z_2}(z_1) = \frac{f_{x_1,x_2}(z_1, z_2)}{f_{x_2}(z_2)} = \frac{\left(\frac{1}{2\pi \det(K)}\right)^{\frac{n}{2}} e^{-\frac{1}{2}(x-m)'K^{-1}(x-m)}}{\left(\frac{1}{2\pi \det(K_{22})}\right)^{\frac{n}{2}} e^{-\frac{1}{2}(z_2-m_2)'K_{22}^{-1}(z_2-m_2)}}$$

- En esta fórmula, z_2 sería un vector de constantes, de modo que la función de densidad condicional debe de ser una función de densidad normal multivariante de algún tipo, por lo que las

distribuciones condicionales de vectores aleatorios normales son normales

- Esta fórmula es la manera más sencilla de poder obtener la función de densidad condicional, de modo que solo hace falta sustituir el valor de x_2 o los valores a los que se condiciona y dividir
- En el caso bivalente, se considera un vector aleatorio $(Y, X)'$ normal, se aplica el resultado de arriba para poder obtener una interpretación más sencilla de los términos del resultado anterior

- La distribución de Y condicionada a $X = x$ es una distribución normal univariante con parámetros los siguientes parámetros:

$$m' = E(Y|X = x) = m_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - m_X)$$

$$\sigma^2 = Var(Y|X = x) = (1 - \rho^2)\sigma_Y^2$$

- Estableciendo similitudes, se puede ver como $\rho\sigma_Y$ sería K_{12} en el caso bivalente, mientras que σ_X sería K_{22}^{-1} . Además, se puede ver como $K_{12}K_{22}^{-1}K'_{12}$ sería $\rho^2\sigma_Y^2$ (debido a que $K_{12}K_{22}^{-1}K'_{12}$ es la forma cuadrática de $K_{12}K_{22}^{-1}$) y, por tanto, K_{11} sería σ_Y
- La propiedad más especial de la distribución normal multivariante es que si los componentes no están correlacionados, entonces los componentes son independientes entre ellos
 - Siendo x un vector aleatorio normal, los componentes de x son independientes si, y solo si, no están correlacionados
 - Como el converso siempre es cierto (es una condición suficiente), entonces suponiendo que $Cov(X_i, X_j) = 0$ para $i \neq j$, esto implica que la matriz de varianzas y covarianzas es una matriz diagonal con elementos $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$
 - Si alguna $\sigma_k^2 = 0$ para $k = 1, 2, \dots, n$, entonces el componente es degenerado y entonces es independiente de los otros componentes, por lo que se asume que la varianza es positiva para todos los componentes
 - En este caso, el inverso de la matriz de varianzas y covarianzas Σ^{-1} existe y es una matriz diagonal con elementos en la diagonal $1/\sigma_1^2, 1/\sigma_2^2, \dots, 1/\sigma_n^2$. Por lo tanto, la función de densidad correspondiente muestra como hay independencia entre

componentes (porque es el producto de las funciones de densidad individuales)

$$f_{\mathbf{x}}(\mathbf{z}) = \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\prod_{k=1}^n \sigma_k} e^{\sum_{k=1}^n \frac{(x_k - \mu_k)^2}{2\sigma_k^2}} =$$

$$= \prod_{k=1}^n \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{(x_k - \mu_k)^2 / 2\sigma_k^2}$$

- Suponiendo que $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \Sigma)$ donde Σ se puede partir en matrices $\Sigma_1, \Sigma_2, \dots, \Sigma_n$, que están a lo largo de la diagonal Σ , entonces \mathbf{x} puede partirse en vectores $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}$ con $Cov(\mathbf{x}^{(i)}) = \Sigma_i$ para $i = 1, 2, \dots, n$ de tal manera que los vectores aleatorios son independientes

$$\Sigma = \begin{pmatrix} \Sigma_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \Sigma_n \end{pmatrix}$$

- Una consecuencia importante del teorema anterior sobre la independencia es que es posible llevar a cabo transformaciones lineales de vectores normales de modo que el vector resultante tenga componentes independientes
 - Siendo $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \Sigma)$ y fijando $\mathbf{y} = \mathbf{C}'\mathbf{x}$, en donde la matriz ortogonal \mathbf{C} es tal que $\mathbf{C}'\Sigma\mathbf{C} = \mathbf{D}$, entonces $\mathbf{y} \sim N_n(\mathbf{C}'\boldsymbol{\mu}, \mathbf{D})$ y los componentes de \mathbf{y} son independientes y $Var(y_k) = \lambda_k$ para $k = 1, 2, \dots, n$, en donde los valores propios $\lambda_1, \lambda_2, \dots, \lambda_n$ de Σ
 - Debido a que la transformación lineal de un vector normal es un vector normal, si se considera $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \Sigma)$, como Σ es una matriz semidefinida positiva, existe una matriz ortogonal \mathbf{C} tal que $\mathbf{C}'\Sigma\mathbf{C} = \mathbf{D}$, en donde \mathbf{D} tiene los valores propios de Σ como elementos en la diagonal
 - Si se fija $\mathbf{y} = \mathbf{C}'\mathbf{x}$, entonces se puede ver a través del teorema sobre la combinación lineal que se obtendría una distribución $N_n(\mathbf{C}'\boldsymbol{\mu}, \mathbf{D})$
 - En particular, puede ser que $\lambda_k = 0$ para alguna k , de modo que el componente correspondiente es una variable normal degenerada
 - Un corolario especial de este teorema es que los componentes del vector aleatorio son independientes y siguen una distribución normal estándar si, y solo si, $\mathbf{x} \sim N_n(\mathbf{0}, \mathbf{I})$

- Otra situación de considerable importancia en estadística son las transformaciones ortogonales de variables normales independientes con la misma, de modo que las variable transformadas también son independientes

- Este es el caso debido al teorema anterior sobre una combinación lineal ortogonal, por lo que se puede postular otro teorema relacionado

- Siendo $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ con $\sigma^2 > 0$, \mathbf{C} una matriz ortogonal arbitraria e $\mathbf{y} = \mathbf{C}\mathbf{x}$, entonces $\mathbf{y} \sim N_n(\mathbf{C}\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ y, en particular, Y_1, Y_2, \dots, Y_n son variables normales independiente con la misma varianza σ^2

- La matriz de varianza y covarianzas es la misma que para las variables normales en el vector \mathbf{x} debido a que se cumple la siguiente identidad:

$$\text{Cov}(\mathbf{y}) = \mathbf{C}\sigma^2\mathbf{I}\mathbf{C}' = \sigma^2\mathbf{I}$$

- Siendo $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ y $\bar{X}_n = (1/n)\sum_{i=1}^n x_k$, si la función $g(\mathbf{x})$ es invariante a las translaciones (de modo que $g(\mathbf{x} + a\mathbf{1}) = g(\mathbf{x})$ para toda $\mathbf{x} \in \mathbb{R}^n$ y a), entonces $g(\mathbf{x})$ y \bar{X}_n son independientes

- Este teorema se conoce como el teorema de Daly, y este se puede demostrar de manera más sencilla si se asume que $\boldsymbol{\mu} = \mathbf{0}$ y que $\sigma^2 = 1$ (sin pérdida de generalidad)

- Si $g(\mathbf{x})$ es invariante a las translaciones, entonces está en el hiperplano $(n-1)$ -dimensional $x_1 + x_2 + \dots + x_n = c$, donde c y \bar{X}_n son constantes (porque $\bar{X}_n = c/n$). Por lo tanto, se puede hacer un cambio de variable y definir una matriz ortogonal \mathbf{C} tal que los elementos de la primera fila sean iguales a $1/n$ y fijar $\mathbf{y} = \mathbf{C}\mathbf{x}$

- Esto hace que, por construcción $Y_1 = \bar{X}_n\sqrt{n}$ y que, por tanto, $\mathbf{y} \sim N_n(\mathbf{0}, \mathbf{I})$. La invarianza a la translación implica necesariamente que g solo dependa de Y_2, Y_3, \dots, Y_n y que sea independiente de Y_1 (debido a que la matriz \mathbf{I} se puede partir en una matriz con \mathbf{I}_1 y $\mathbf{I}_{2,n}$ en la diagonal, y por tanto los dos tipos de componentes son independientes)

- Las formas cuadráticas de los vectores normales son de gran importancia en estadística, tal y como los métodos de mínimos cuadrados, el análisis de la varianza, el análisis de regresión o el diseño experimental

- La idea general es dividir la suma de cuadrados de las observaciones en un número de formas cuadráticas, cada una correspondiente a una causa de la variación
 - La división de la suma de cuadrados separa las causas de la variabilidad de tal manera que cada forma cuadrática corresponde a una causa, con una forma residual que mide los errores aleatorios en el experimento
 - La conclusión del teorema de Cochran es que, bajo la suposición de normalidad, las diferentes formas cuadráticas son independientes y distribuidas siguiendo una χ^2 (excepto por el factor constante). Esto se puede usar para contrastar hipótesis sobre la influencia de diferentes tratamientos
- Suponiendo que $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \Sigma)$ con $\det(\Sigma) > 0$, entonces la siguiente forma cuadrática se distribuye como una chi cuadrada con n grados de libertad (donde n es la dimensión del vector \mathbf{x}):

$$(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi_n^2$$

- Fijando $\mathbf{y} = \Sigma^{-1/2} (\mathbf{x} - \boldsymbol{\mu})$, entonces tanto la esperanza y la varianza permiten ver que $\mathbf{y} \sim N_n(\mathbf{0}, \mathbf{I})$

$$E(\mathbf{y}) = \mathbf{0} \quad \text{Cov}(\mathbf{y}) = \text{Cov}(\Sigma^{-1/2} \mathbf{x}) = \mathbf{I}_{n \times n}$$

- Por lo tanto, se puede ver que la distribución de $\mathbf{y}' \mathbf{y}$ es una chi cuadrada con n grados de libertad

$$\begin{aligned} (\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) &= \left(\Sigma^{-1/2} (\mathbf{x} - \boldsymbol{\mu}) \right)' \left(\Sigma^{-1/2} (\mathbf{x} - \boldsymbol{\mu}) \right) = \\ &= \mathbf{y}' \mathbf{y} \sim \chi_n^2 \end{aligned}$$

- Suponiendo que X_1, X_2, \dots, X_n es una muestra con $X \sim N(0, \sigma^2)$ y fijando $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$, se puede considerar la siguiente identidad:

$$\sum_{i=1}^n X_i^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 + n \bar{X}_n^2$$

- El primer término en el lado de derecha es igual a $(n-1)s_n^2$, donde s_n^2 es la varianza muestral, que está distribuida como una $\sigma^2 \chi^2(n-1)$. Además, el segundo término está distribuido como una $\sigma^2 \chi^2(1)$ y los términos son independientes

- En el lado izquierdo, el término se distribuye como una $\sigma^2 \chi^2(n)$, por lo que se ha dividido la suma de cuadrados de las observaciones en la suma de dos formas cuadráticas diferentes que siguen una distribución chi cuadrada
 - El significado estadístico de esta división de la suma de cuadrados $\sum_{i=1}^n X_k^2$ es que el primer término es grande si la muestra está muy dispersa, y el segundo término es grande si la media no está cerca de cero. Por lo tanto, si la suma de cuadrados es grande, se puede encontrar la causa a través de la descomposición
 - Como los dos términos son independientes, esto lleva al contraste t , usado para saber si la media es nula. Más generalmente, las representaciones de la suma de cuadrados como una suma de formas cuadráticas semidefinidas positivas juegan un papel crucial en estadística
- Siendo x_1, x_2, \dots, x_n son ...

Commented [MOU10]: Acabar ya

Las distribuciones normales mixtas y distribuciones generalizadas hiperbólicas

- Es posible generalizar esta distribución normal multivariante para obtener distribuciones mixtas de normal multivariante. La idea crucial es la introducción de aleatoriedad en la primera matriz de varianzas y covarianzas y del vector de medias a través de una mezcla con una variable positiva W
- El vector aleatorio \mathbf{X} tiene una distribución mixta de normal multivariante si este se define de la siguiente manera:

$$\mathbf{X} \sim \boldsymbol{\mu} + \sqrt{W} \mathbf{A} \mathbf{Z}$$

- En este caso, el vector \mathbf{Z} sigue una distribución normal multivariante estándar $N_k(\mathbf{0}, \mathbf{I}_k)$, $W \geq 0$ es una variable no negativa con valores escalares que es independiente del vector \mathbf{Z} , y \mathbf{A} es una matriz $d \times k$ con valores reales y $\boldsymbol{\mu}$ es el vector de medias constantes
- Estas distribuciones se conocen como mezclas de varianzas porque si se condiciona el vector aleatorio a W , se puede comprobar que $w \equiv \mathbf{X} | W$ sigue una distribución normal multivariante $N_d(\boldsymbol{\mu}, w\Sigma)$, donde $\Sigma = \mathbf{A}\mathbf{A}'$ y donde la varianza queda condicionada a w
- La distribución de \mathbf{X} se puede interpretar como una distribución compuesta construida tomando un conjunto de distribuciones normales

multivariantes con el mismo vector de medias y con la misma matriz de varianzas y covarianzas multiplicada por la constante w

- El caso interesante es donde la matriz A tiene rango $\text{rank}(A) = d \leq k$ (rango completo) y Σ tiene rango completo y es positiva definida. Esto permite obtener una mezcla de varianza normal no singular
 - La distribución mixta se construye obteniendo muestras aleatorias de este conjunto de distribuciones normales multivariantes acorde a un conjunto de ponderaciones determinadas por la distribución de W , por lo que la distribución resultante no es una distribución normal multivariante (si no se condiciona a un valor)
 - En el contexto de modelaje de rendimientos de factores de riesgo, la distribución de mezcla W se puede interpretar como el *shock* que surge de nueva información y de impactos de volatilidad en todos los factores de riesgo
- Suponiendo que W tiene una expectación finita, se pueden obtener la esperanza y la matriz de varianzas y covarianzas para \mathbf{X}

$$E(\mathbf{X}) = E(\boldsymbol{\mu} + \sqrt{W}\mathbf{AZ}) = \boldsymbol{\mu} + E(\sqrt{W})A E(\mathbf{Z}) = \boldsymbol{\mu}$$

$$\begin{aligned} \text{Cov}(\mathbf{X}) &= E[(\sqrt{W}\mathbf{AZ})(\sqrt{W}\mathbf{AZ})'] = E[\sqrt{W}\mathbf{AZZ}'A'\sqrt{W}] = \\ &= E[\sqrt{W}\sqrt{W}\mathbf{AZZ}'A'] = E(W)A E(\mathbf{ZZ}')A' = E(W)AA' = E(W)\Sigma \end{aligned}$$

- Normalmente $\boldsymbol{\mu}$ se denomina vector de localización y Σ se denomina matriz de dispersión de la distribución
 - La matriz Σ es la matriz de varianzas y covarianzas de \mathbf{AZ} es la matriz de varianzas y covarianzas de \mathbf{X} si $E(W) = 1$, y $\boldsymbol{\mu}$ es solo el vector de medias cuando $E(W)$ está definido, lo cual requiere que $E(\sqrt{W}) < \infty$
- Estas distribuciones proporcionan un buen ejemplo de que la falta de correlación no necesariamente implica independencia de los componentes de \mathbf{X}
- Suponiendo que (X_1, X_2) tienen una distribución mixta de normal multivariante con $A = I_{2 \times 2}$ y $E(W) < \infty$ tal que $\text{Cov}(X_1, X_2) = 0$, entonces X_1 y X_2 son independientes si, y solo

si, W es casi seguro constante (las variables se distribuyen normalmente, por ejemplo)

- Si W es casi seguro una constante, entonces X_1 y X_2 siguen una distribución bivalente normal y son independientes. Conversamente, si X_1 y X_2 son independientes, se tiene que $E(|X_1||X_2|) = E(|X_1|)E(|X_2|)$

$$\begin{aligned} E(|X_1||X_2|) &= E(W|Z_1||Z_2|) = E(W)E(|X_1|)E(|X_2|) \geq \\ &\geq E(\sqrt{W})E(|Z_1|)E(\sqrt{W})E(|Z_2|) = E(|X_1|)E(|X_2|) \end{aligned}$$

- Solo es posible obtener una igualdad en el medio si W es una constante
- Usando la ecuación característica vista para las normales multivariantes, se puede calcular la función característica de una mezcla de varianza normal:

$$\phi(\mathbf{X}) = E[E(e^{it'\mathbf{X}}|W)] = E\left(e^{it'\boldsymbol{\mu} - \frac{1}{2}Wt'\boldsymbol{\Sigma}t}\right) = e^{it'\boldsymbol{\mu}} - \hat{H}\left(\frac{1}{2}t'\boldsymbol{\Sigma}t\right)$$

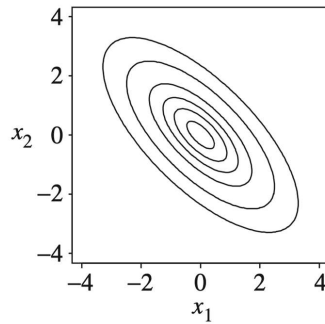
$$\text{where } \hat{H}(x) = \int_0^\infty e^{-xv} dH(v)$$

- A partir de este resultado, la notación para denotar a una distribución mixta de normal multivariante es $\mathbf{X} \sim M_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \hat{H})$
- Asumiendo que $\boldsymbol{\Sigma}$ es definida positiva y que la distribución de W no tiene un punto de masa en cero, se puede derivar la función de densidad conjunta de esta distribución

$$\begin{aligned} f(\mathbf{x}) &= \int f_{\mathbf{X}|W}(\mathbf{x}|w) dH(w) = \\ &= \int \frac{w^{-\frac{d}{2}}}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})}{2w}\right] dH(w) \end{aligned}$$

- En este caos, $f_{\mathbf{X}|W}(\mathbf{x}|w)$ es la función de densidad conjunta gaussiana condicional a W y la función de densidad conjunta de \mathbf{x} se expresa en términos de la integral de Lebesgue-Stieltjes
- Cuando H tiene densidad h , esta función de densidad se convierte en la integral de Riemann de $\int_0^\infty f_{\mathbf{X}|W}(\mathbf{x}|w)h(w) dw$. Todas estas densidades dependen de \mathbf{x} solo a través del término

cuadrático $(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$, de modo que son las densidades de las distribuciones elípticas



- Si $\mathbf{X} \sim M_d(\boldsymbol{\mu}, \Sigma, \hat{H})$ y $\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b}$, donde $\mathbf{B} \in \mathbb{R}^{k \times d}$ y $\mathbf{b} \in \mathbb{R}^k$, entonces $\mathbf{Y} \sim M_k(\mathbf{B}\boldsymbol{\mu} + \mathbf{b}, \mathbf{B}\Sigma\mathbf{B}', \hat{H})$

- La función característica mostrada anteriormente se puede utilizar para demostrar el siguiente resultado:

$$\phi(\mathbf{X}) = E(e^{it'(\mathbf{B}\mathbf{X} + \mathbf{b})}) = e^{it'\mathbf{b}} \phi(\mathbf{B}\mathbf{X}) = e^{it'(\mathbf{B}\boldsymbol{\mu} + \mathbf{b})} \hat{H}\left(\frac{1}{2} \mathbf{t}' \mathbf{B} \Sigma \mathbf{B}' \mathbf{t}\right)$$

- Como se puede ver, la subclase de distribuciones de mezcla especificada por \hat{H} es cerrado bajo transformaciones lineales (no cambia)

- Estas distribuciones se pueden simular

Commented [MOU11]: Acabar bien

- A partir de las distribuciones mixtas de normal multivariante, se pueden obtener diferentes distribuciones importantes al escoger una forma para la variable W

- Si se toma W como una variable con una distribución gamma inversa $W \sim Ig\left(\frac{1}{2}\nu, \frac{1}{2}\nu\right)$ (que es equivalente a $\nu/W \sim \chi_\nu^2$), entonces \mathbf{X} tiene una distribución t multivariante con ν grados de libertad

- La función de densidad para la variable W de mezcla es la siguiente:

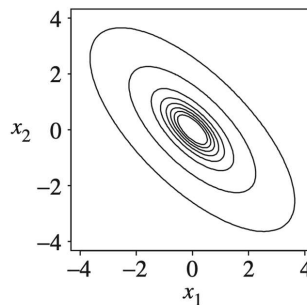
$$f(w) = \frac{\beta^\alpha}{\Gamma(\alpha)} w^{-(\alpha+1)} e^{-\frac{\beta}{w}} = \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} w^{-\left(\frac{\nu}{2}+1\right)} e^{-\frac{\nu}{2w}}$$

where $\alpha = \frac{\nu}{2}$ & $\beta = \frac{\nu}{2}$

- La notación para esta distribución con ν grados de libertad es $\mathbf{X} \sim t_d(\nu, \boldsymbol{\mu}, \Sigma)$ (sin que Σ sea la matriz de varianzas y covarianzas en esta definición). Debido a que $E(W) = \nu/(\nu - 2)$ y a que $Cov(W) = [\nu/(\nu - 2)]\Sigma$, la matriz de varianzas y covarianzas de esta distribución solo está definida para $\nu > 2$
- La densidad de esta distribución se puede calcular a través de la función de densidad derivada anteriormente para la mixta de normal multivariante

$$f(\mathbf{x}) = \frac{\Gamma\left(\frac{1}{2}(\nu + d)\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\left[1 + \frac{(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})}{\nu}\right]^{-\frac{(\nu+d)}{2}}$$

- El locus de puntos con igual densidad es un elipsoide con la ecuación $(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c$ para alguna $c > 0$. Como se puede ver, a comparación de la normal multivariante, los niveles de densidad incrementan más rápido en el centro de la distribución y decaen más lentamente cuanto más se alejan del centro



- A comparación de la normal multivariante, esta distribución tiene las colas más anchas y, por tanto, una tendencia más pronunciada a generar valores extremos
 - Cuando W es una variable discreta que asume los valores positivos k_1 y k_2
 - ...
- Todas las distribuciones anteriormente vistas tienen simetría elíptica, lo cual puede simplificar mucho el modelo para los datos de los rendimientos de los

Commented [MOU12]: Acabar bien

factores de riesgo reales, por lo que se pueden ver otros modelos para introducir simetría

- La simetría elíptica implica que las distribuciones marginales unidimensionales son rígidamente simétricas, lo cual contradice la observación frecuente de los rendimientos que los rendimientos negativos tienen colas más anchas que los positivos
 - Los modelos que introducen asimetría a la clase de mezcla de normales lo hacen a través de mezclar las distribuciones normales con diferentes medias y diferentes varianzas
- El vector aleatorio \mathbf{X} tiene una distribución normal media-varianza multivariante si se define de la siguiente manera:

$$\mathbf{X} \sim \mathbf{m}(W) + \sqrt{W}\mathbf{A}\mathbf{Z}$$

- En este caso, el vector \mathbf{Z} sigue una distribución normal multivariante estándar $N_k(\mathbf{0}, \mathbf{I}_k)$, $W \geq 0$ es una variable no negativa con valores escalares que es independiente del vector \mathbf{Z} , \mathbf{A} es una matriz $d \times k$ con valores reales y $\mathbf{m}: [0, \infty) \rightarrow \mathbb{R}^d$ es una función medible
- En este caso, si se condiciona el vector aleatorio a W , se puede comprobar que $w \equiv \mathbf{X} | W$ sigue una distribución normal multivariante $N_d(\mathbf{m}(w), w\Sigma)$, donde $\Sigma = \mathbf{A}\mathbf{A}'$ y, debido a que ahora hay mezcla tanto en la media como en la varianza, estas se denominan distribuciones de mezcla de media-varianza de normales
- Una posible especificación concreta para la función $\mathbf{m}(W)$ suele ser $\mathbf{m}(W) = \boldsymbol{\mu} + W\boldsymbol{\gamma}$ cuando $\boldsymbol{\mu}$ y $\boldsymbol{\gamma}$ son vectores de parámetros en \mathbb{R}^d
 - El vector $\boldsymbol{\gamma}$ se puede interpretar como el parámetro de asimetría, de modo que si $\boldsymbol{\gamma} = \mathbf{0}$, la distribución es simétrica alrededor de $\boldsymbol{\mu}$, y si $\boldsymbol{\gamma} \neq \mathbf{0}$, entonces la distribución es asimétrica alrededor de $\boldsymbol{\mu}$
 - Debido a que $E(\mathbf{X}|W) = \boldsymbol{\mu} + W\boldsymbol{\gamma}$ y $Cov(\mathbf{X}|W) = W\Sigma$, se puede ver que la esperanza y la matriz de varianzas y covarianzas de \mathbf{X} vienen dadas por las siguientes expresiones cuando la variable de mezcla W tiene varianza finita:

$$E(\mathbf{X}) = E[E(\mathbf{X}|W)] = E(\boldsymbol{\mu} + W\boldsymbol{\gamma}) = \boldsymbol{\mu} + E(W)\boldsymbol{\gamma}$$

$$Cov(\mathbf{X}) = E[Cov(\mathbf{X}|W)] + Cov[E(\mathbf{X}|W)] =$$

$$\begin{aligned}
&= E(W)\Sigma + Cov[(\mu + W\gamma)'(\mu + W\gamma)] = \\
&= E(W)\Sigma + Cov[\gamma'W'W\gamma] = E(W)\Sigma + Var(W)\gamma\gamma'
\end{aligned}$$

- Como se puede ver, los parámetros μ y Σ no son, en general, el vector de medias y la matriz de varianzas y covarianzas, respectivamente, de X (o un múltiplo). Esto solo es el caso si $\gamma = \mathbf{0}$, de modo que la distribución es una normal de mezcla de varianzas y las fórmulas anteriormente vistas aplican
- Una clase especial de distribuciones hiperbólicas generalizadas (GH) es la familia de mezclas de normal multivariante conocidas como las distribuciones hiperbólicas generalizadas simétricas

- Esta familia de distribuciones se obtiene tomando W como una variable aleatoria que sigue una distribución gaussiana inversa generalizada (GIG) $W \sim N^-(\lambda, \chi, \psi)$

- Los parámetros λ , χ y ψ determinan la forma de la distribución o el peso de las colas y del centro de la distribución. En general, mientras más grandes sean estos parámetros, más cerca estarán de la normal
- La función de densidad para la variable W de mezcla es la siguiente:

$$f(w) = \frac{\chi^{-\lambda} (\sqrt{\chi\psi})^\lambda}{2K_\lambda(\sqrt{\chi\psi})} w^{-\lambda-1} e^{-\frac{1}{2}(\chi w^{-1} + \psi)} \quad \text{when } w > 0$$

- Los parámetros satisfacen las siguientes condiciones para los posibles valores de λ :

$$\lambda < 0 \Rightarrow \chi > 0 \text{ \& } \psi \geq 0$$

$$\lambda = 0 \Rightarrow \chi > 0 \text{ \& } \psi > 0$$

$$\lambda > 0 \Rightarrow \chi \geq 0 \text{ \& } \psi > 0$$

- Usando la densidad de la mezcla de normal multivariante, se puede obtener la función de densidad conjunta para esta familia de distribuciones:

$$f(\mathbf{x}) = \frac{(\sqrt{\chi\psi})^{-\lambda} \psi^{\frac{d}{2}}}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}} K_{\lambda}(\sqrt{\chi\psi})} \frac{K_{\lambda - (\frac{d}{2})} \sqrt{[\chi + (\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})] \psi}}{[\sqrt{[\chi + (\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})] \psi}]^{\frac{d}{2} - \lambda}}$$

K_{λ} = Bessel function of third kind

- Esta distribución es un caso especial ($\boldsymbol{\gamma} = \mathbf{0}$) de una familia más general de distribuciones hiperbólicas generalizadas multivariante, la cual puede ser obtenida como una mezcla de media-varianza de normales, las cuales no son necesariamente distribuciones elípticas

$$f(\mathbf{x}) = c \frac{K_{\lambda - (d/2)} \sqrt{[\chi + (\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})] (\psi + \boldsymbol{\gamma}' \Sigma^{-1} \boldsymbol{\gamma})}}{[\sqrt{[\chi + (\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})] (\psi + \boldsymbol{\gamma}' \Sigma^{-1} \boldsymbol{\gamma})}]^{\frac{d}{2} - \lambda}}$$

$$\text{where } c \equiv \frac{(\sqrt{\chi\psi})^{-\lambda} \psi^{\frac{d}{2}} (\psi + \boldsymbol{\gamma}' \Sigma^{-1} \boldsymbol{\gamma})^{\frac{d}{2} - \lambda}}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}} K_{\lambda}(\sqrt{\chi\psi})}$$

- En este caso general, la distribución no es elíptica y tiene márgenes (distribuciones marginales) asimétricos
- Se suele adoptar la notación $\mathbf{X} \sim GH_d(\lambda, \chi, \psi, \boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$ para la GH
- El vector de medias y la matriz de varianzas y covarianzas se puede calcular fácilmente a través de las fórmulas anteriormente vistas para mezclas de media-varianza de normales utilizando la información sobre la distribución GIG y sus momentos
- Además, la función de la distribución GH se puede calcular a través de la función característica anteriormente vista

$$\phi(\mathbf{X}) = E(e^{it'\mathbf{X}}) = e^{it'\mathbf{b}} \phi(B\mathbf{X}) = e^{it'\boldsymbol{\mu}} \hat{H}\left(\frac{1}{2} \mathbf{t}' \Sigma \mathbf{t} - it' \boldsymbol{\gamma}\right)$$

- Si $\mathbf{X} \sim GH_d(\lambda, \chi, \psi, \boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$ y $\mathbf{Y} = B\mathbf{X} + \mathbf{b}$, donde $B \in \mathbb{R}^{k \times d}$ y $\mathbf{b} \in \mathbb{R}^k$, entonces $\mathbf{Y} \sim GH_d(\lambda, \chi, \psi, B\boldsymbol{\mu} + \mathbf{b}, B\Sigma B', B\boldsymbol{\gamma})$
- Usando la función característica de esta distribución, se puede obtener el siguiente resultado:

$$\phi(\mathbf{Y}) = E(e^{it'(B\mathbf{X} + \mathbf{b})}) = e^{it'(B\boldsymbol{\mu} + \mathbf{b})} \hat{H}\left(\frac{1}{2} \mathbf{t}' B\Sigma B' \mathbf{t} - it' B\boldsymbol{\gamma}\right)$$

- Por lo tanto, la clase de distribuciones GH son cerrados bajo operaciones lineales y los parámetros heredados de la distribución de mezcla GIG no cambian bajo operaciones lineales
- Esto causa que los márgenes de \mathbf{X} sean fáciles de calcular, ya que estos se distribuirán como $\mathbf{X}_i \sim GH_1(\lambda, \chi, \psi, \boldsymbol{\mu}_i, \Sigma_{ii}, \boldsymbol{\gamma}_i)$
- La distribución de mezcla GIG es muy flexible y contiene varios tipos de distribuciones concretas, las cuales son casos especiales de la GH
 - Si $\lambda > 0$ y $\chi = 0$, se obtiene una distribución gamma, y si $\lambda < 0$ y $\psi = 0$, se obtiene una distribución gamma inversa. En estos casos la función de densidad anterior se puede interpretar como el límite cuando $\chi \rightarrow 0$ y $\psi \rightarrow 0$ (respectivamente), de modo que son casos límites especiales
 - La distribución de mezcla gamma resulta en distribuciones de Laplace o los llamados modelos de varianza-gamma simétrica (VG), mientras que la mezcla de gamma inversa resulta en una distribución t multivariante con $\lambda = -\nu/2$ y $\chi = \nu$
 - Si $\lambda = (d + 1)/2$, la distribución es una distribución hiperbólica multivariante (pero no generalizada) cuyas marginales univariantes son distribuciones hiperbólicas
 - Si $\lambda = -1/2$, la distribución que se obtiene es la inversa gaussiana normal simétrica (NIG), mientras que si $\lambda = 1$ la distribución es una GH multivariante cuyas marginales univariantes son distribuciones hiperbólicas
- Hay varias parametrizaciones alternativas para la distribución GH en la literatura, y es muy común encontrar esta distribución en su forma reparametrizada. Las parametrizaciones más comunes son la parametrización $(\lambda, \chi, \psi, \boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$, la $(\lambda, \bar{\alpha}, \boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$ y la $(\lambda, \alpha, \boldsymbol{\mu}, \Delta, \delta, \boldsymbol{\beta})$

	$(\lambda, \chi, \psi, \mu, \Sigma, \gamma)$ -Parametrization					
	λ	χ	ψ	μ	Σ	γ
ghyp	$\lambda \in \mathbb{R}$	$\chi > 0$	$\psi > 0$	$\mu \in \mathbb{R}^d$	$\Sigma \in \mathbb{R}^\Sigma$	$\gamma \in \mathbb{R}^d$
hyp	$\lambda = \frac{d+1}{2}$	$\chi > 0$	$\psi > 0$	$\mu \in \mathbb{R}^d$	$\Sigma \in \mathbb{R}^\Sigma$	$\gamma \in \mathbb{R}^d$
NIG	$\lambda = -\frac{1}{2}$	$\chi > 0$	$\psi > 0$	$\mu \in \mathbb{R}^d$	$\Sigma \in \mathbb{R}^\Sigma$	$\gamma \in \mathbb{R}^d$
t	$\lambda < 0$	$\chi > 0$	$\psi = 0$	$\mu \in \mathbb{R}^d$	$\Sigma \in \mathbb{R}^\Sigma$	$\gamma \in \mathbb{R}^d$
VG	$\lambda > 0$	$\chi = 0$	$\psi > 0$	$\mu \in \mathbb{R}^d$	$\Sigma \in \mathbb{R}^\Sigma$	$\gamma \in \mathbb{R}^d$

	$(\lambda, \bar{\alpha}, \mu, \Sigma, \gamma)$ -Parametrization				
	λ	$\bar{\alpha}$	μ	Σ	γ
ghyp	$\lambda \in \mathbb{R}$	$\bar{\alpha} > 0$	$\mu \in \mathbb{R}^d$	$\Sigma \in \mathbb{R}^\Sigma$	$\gamma \in \mathbb{R}^d$
hyp	$\lambda = \frac{d+1}{2}$	$\bar{\alpha} > 0$	$\mu \in \mathbb{R}^d$	$\Sigma \in \mathbb{R}^\Sigma$	$\gamma \in \mathbb{R}^d$
NIG	$\lambda = \frac{1}{2}$	$\bar{\alpha} > 0$	$\mu \in \mathbb{R}^d$	$\Sigma \in \mathbb{R}^\Sigma$	$\gamma \in \mathbb{R}^d$
t	$\lambda = -\frac{\nu}{2} < -1$	$\bar{\alpha} = 0$	$\mu \in \mathbb{R}^d$	$\Sigma \in \mathbb{R}^\Sigma$	$\gamma \in \mathbb{R}^d$
VG	$\lambda > 0$	$\bar{\alpha} = 0$	$\mu \in \mathbb{R}^d$	$\Sigma \in \mathbb{R}^\Sigma$	$\gamma \in \mathbb{R}^d$

	$(\lambda, \alpha, \mu, \Delta, \delta, \beta)$ -Parametrization					
	λ	α	δ	μ	Δ	β
ghyp	$\lambda \in \mathbb{R}$	$\alpha > 0$	$\delta > 0$	$\mu \in \mathbb{R}^d$	$\Delta \in \mathbb{R}^\Delta$	$\beta \in \{x \in \mathbb{R}^d : \alpha^2 - x' \Delta x > 0\}$
hyp	$\lambda = \frac{d+1}{2}$	$\alpha > 0$	$\delta > 0$	$\mu \in \mathbb{R}^d$	$\Delta \in \mathbb{R}^\Delta$	$\beta \in \{x \in \mathbb{R}^d : \alpha^2 - x' \Delta x > 0\}$
NIG	$\lambda = -\frac{1}{2}$	$\alpha > 0$	$\delta > 0$	$\mu \in \mathbb{R}^d$	$\Delta \in \mathbb{R}^\Delta$	$\beta \in \{x \in \mathbb{R}^d : \alpha^2 - x' \Delta x > 0\}$
t	$\lambda < 0$	$\alpha = \sqrt{\beta' \Delta \beta}$	$\delta > 0$	$\mu \in \mathbb{R}^d$	$\Delta \in \mathbb{R}^\Delta$	$\beta \in \mathbb{R}^d$
VG	$\lambda > 0$	$\alpha > 0$	$\delta = 0$	$\mu \in \mathbb{R}^d$	$\Delta \in \mathbb{R}^\Delta$	$\beta \in \{x \in \mathbb{R}^d : \alpha^2 - x' \Delta x > 0\}$

- La parametrización $(\lambda, \chi, \psi, \mu, \Sigma, \gamma)$ se obtiene de la distribución de mezcla de media-varianza de normal cuando $W \sim N^-(\lambda, \chi, \psi)$, la cual presenta problemas de identificación
 - Debido a que los vectores $X \sim GH_d(\lambda, \chi/k, k\psi, \mu, k\Sigma, k\gamma)$ y $X \sim GH_d(\lambda, \chi, \psi, \mu, \Sigma, \gamma)$ son idénticas para cualquier $k > 0$, ocurre un problema de identificación cuando se intentan estimar los parámetros en la práctica
 - Esto se puede resolver restringiendo el determinante de Σ a un valor particular cuando se ajusta el modelo (normalmente a 1). Esta restricción tiene un efecto en los valores de χ y ψ que se estiman, pero no tienen un efecto en el valor de su producto $\chi\psi$, el cual es un buen parámetro resumen de la distribución
- Para eliminar el grado de libertad de manera más elegante que fijando el determinante de Σ , se puede requerir que el valor esperado de la variable W que sigue una GIG sea 1, derivando a la parametrización $(\lambda, \bar{\alpha}, \mu, \Sigma, \gamma)$
 - Esto hace que la interpretación de γ sea más fácil y que, además, el proceso de ajuste sea más rápido
 - Como el valor esperado de W existe mientras $\psi > 0$, si $\psi \rightarrow 0$, la distribución GIG se aproxima a la distribución de gamma inversa, para el cual la esperanza solo existe si $\gamma < -1$. De este modo, si $E(W) = 1$, se puede obtener ψ y χ a partir de $\bar{\alpha}$ y K_λ

$$E(W) = \sqrt{\frac{\chi K_{\lambda+1}(\sqrt{\chi\psi})}{\psi K_{\lambda}(\sqrt{\chi\psi})}} = 1$$

$$\text{If } \bar{\alpha} \equiv \sqrt{\chi\psi} \Rightarrow \psi = \bar{\alpha} \frac{K_{\lambda+1}(\bar{\alpha})}{K_{\lambda}(\bar{\alpha})} \text{ \& } \chi = \frac{\bar{\alpha}^2}{\psi} = \bar{\alpha} \frac{K_{\lambda}(\bar{\alpha})}{K_{\lambda+1}(\bar{\alpha})}$$

- La desventaja de esta parametrización es que no existe en el caso en que $\bar{\alpha} = 0$ y $\lambda \in [-1, 0]$, lo que corresponde a la distribución t -Student con varianza inexistente
- En la introducción de la distribución GH se utilizó la parametrización $(\lambda, \alpha, \boldsymbol{\mu}, \Delta, \delta, \boldsymbol{\beta})$ para el caso multivariante
 - Igual que en la parametrización $(\lambda, \chi, \psi, \boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$, hay un problema de identificación que se puede resolver restringiendo el determinante de Δ a 1
 - La representación de la mezcla perteneciente a esta parametrización es la siguiente:

$$X|W = w \sim N_d(\boldsymbol{\mu} + w\boldsymbol{\beta}\Delta, w\Delta)$$

$$\text{where } W \sim N^-(\lambda, \delta^2, \alpha^2 - \boldsymbol{\beta}'\Delta\boldsymbol{\beta})$$

Las distribuciones esféricas y elípticas

Commented [MOU13]: Acabar bien Quant Risk Management

La entropía

Commented [MOU14]: Acabar bien Papoulis

La convergencia

Commented [IC15]: Estaría interesante extender esto a multivariante (donde lo de convergencia Op está en el libro de Time Series)

- Existen varios conceptos de convergencia en la teoría de la probabilidad, de modo que se presentan estos conceptos a partir de asumir una secuencia X_1, X_2, X_3, \dots de variables aleatorias
 - Se dice que X_n converge de manera casi segura o *almost surely* (denotado como a.s.) a la variable aleatoria X cuando $n \rightarrow \infty$ si, y solo si, se cumple la siguiente condición:

$$P(\{\omega \in \Omega: X_n(\omega) \rightarrow X(\omega) \text{ as } n \rightarrow \infty\}) = 1$$

- Cuando se lidia con este tipo de convergencia, se consideran todos los eventos $\omega \in \Omega$ en el espacio muestral de un experimento aleatorio y se evalúa si los números reales $X_n(\omega)$ convergen al número real $X(\omega)$ cuando $n \rightarrow \infty$ o no

- Se tiene convergencia casi segura si el conjunto ω para el cual hay convergencia tiene probabilidad unitaria o, equivalentemente, si el conjunto ω para el cual no se tiene convergencia tiene probabilidad nula. Por lo tanto, se tiene una definición alternativa con la cual es más sencillo operar:

$$P(\bar{R}) = 0 \quad \text{where } \bar{R} = \{\omega \in \Omega: X_n(\omega) \nrightarrow X(\omega) \text{ as } n \rightarrow \infty\}$$

- La notación que se utiliza para denotar este tipo de convergencia es $X_n \xrightarrow{a.s.} X$ cuando $n \rightarrow \infty$
- Se dice que X_n converge en probabilidad a la variable aleatoria X cuando $n \rightarrow \infty$ si, y solo si, se cumple la siguiente condición:

$$P(|X_n - X| \geq \varepsilon) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

- Este concepto de tendencia nace de considerar el valor X al que tiende una secuencia X_n cuando $n \rightarrow \infty$ al mirar la probabilidad de los posibles valores (se tiende en probabilidad a aquellos valores más probables, intuitivamente)
- Esta condición se puede expresar equivalentemente considerando una distancia menor a ε

$$P(|X_n - X| < \varepsilon) \rightarrow 1 \quad \text{as } n \rightarrow \infty$$

- La notación que se utiliza para denotar este tipo de convergencia es $X_n \xrightarrow{p} X$ cuando $n \rightarrow \infty$
- Dos conceptos derivados de la convergencia en probabilidad es el orden en probabilidad y la acotación en probabilidad
- El orden en probabilidad, denotado como $X_n = o_p(a_n)$, se da si, y solo si, se cumple la siguiente convergencia en probabilidad:

$$\frac{X_n}{a_n} \xrightarrow{p} 0$$

- El término acotación en probabilidad, denotado como $x_n = O_p(a_n)$, significa que para toda $\varepsilon > 0$, existe una $\delta(\varepsilon) > 0$ tal que se cumpla la siguiente condición para toda n :

$$P\left(\left|\frac{X_n}{a_n}\right| > \delta(\varepsilon)\right) \leq \varepsilon \quad \text{for } \forall n$$

- Bajo estas convenciones, $X_n \xrightarrow{p} X$ se convierte en $X_n - X = o_p(1)$. Además, estas definiciones se pueden comparar con sus contrapartes no aleatorias (para una secuencia fija $x_n = o(1)$ si $x_n \rightarrow 0$ y $x_n = O(1)$ para $n = 1, 2, \dots$ si esta está acotada)
- Algunas propiedades útiles, parecidas a las de sus contrapartes no aleatorias, son las siguientes:
 - (a) If $x_n = o_p(a_n)$ & $y_n = o_p(b_n)$ then $x_n y_n = o_p(a_n b_n)$
and $x_n + y_n = o_p(\max(a_n, b_n))$
 - (b) If $x_n = o_p(a_n)$ & $y_n = O_p(b_n)$ then $x_n y_n = o_p(a_n b_n)$
and $x_n + y_n = o_p(\max(a_n, b_n))$
 - (c) If $x_n = O_p(a_n)$ & $y_n = O_p(b_n)$ then $x_n y_n = O_p(a_n b_n)$
and $x_n + y_n = O_p(\max(a_n, b_n))$
- Se dice que X_n converge en media- r a la variable aleatoria X cuando $n \rightarrow \infty$ si, y solo si, se cumple la siguiente condición:

$$E(|X_n - X|^r) \rightarrow 0 \text{ as } n \rightarrow \infty$$

- Para entender mejor este concepto, se tiene que considerar un espacio de probabilidad (Ω, \mathcal{F}, P) y un espacio vectorial \mathcal{H} r -dimensional, cuyos vectores sean variables aleatorias X tales que $E(X^r) < \infty$. Definiendo el producto interior en \mathcal{H} como $\langle X_1, X_2, \dots, X_n \rangle = E(X_1 X_2 \dots X_n)$, la norma inducida es $\|X\| = \sqrt[r]{E(|X|^r)}$ y la distancia entre dos puntos será $\|X_i - X_j\| = \sqrt[r]{E(|X_i - X_j|^r)}$, por lo que la convergencia en media r es una convergencia de la distancia entre dos variables en un espacio vectorial (de vectores aleatorios) como \mathcal{H}
- La convergencia en $r = 2$ normalmente se denomina convergencia en media cuadrática, y $|X_n - X|^2 = (X_n - X)^2$. En verdad, para cualquier valor $r = 2k$ para $k \in \mathbb{Z}$, se cumple que $|X_n - X|^r = (X_n - X)^r$
- El caso más sencillo para operar con esta definición de convergencia es cuando la variable es discreta, de modo que se puede descomponer la esperanza en una suma de términos multiplicados por las probabilidades

$$E(|X_n - X|^r) = \sum_k |k - X|^r P(X_n = k) \text{ for } n = 1, 2, \dots$$

- La notación que se utiliza para denotar este tipo de convergencia es $X_n \xrightarrow{r} X$ cuando $n \rightarrow \infty$. A este criterio también se conoce como convergencia en L^r , ya que si la secuencia de variables aleatorias X_1, X_2, \dots converge, esta pertenece al espacio L^r (resultado de la completitud de L^r más adelante)
- Se dice que X_n converge en distribución a la variable aleatoria X cuando $n \rightarrow \infty$ si, y solo si, se cumple la siguiente condición:

$$F_{X_n}(x) \rightarrow F_X(x) \text{ as } n \rightarrow \infty \text{ for } \forall x \in C(F_X)$$

- En este caso, $C(F_X)$ es el conjunto de continuidad de la función de distribución de X (el conjunto de puntos en donde la función es continua)

$$C(F_X) = \{x: F_X(x) \text{ is continuous at } x\}$$

- Uno puede demostrar que una función de distribución tiene como mucho un número numerable de discontinuidades. En consecuencia, $C(F_X)$ es igual a \mathbb{R} , excepto, posiblemente, por un número numerable de puntos
- En esta definición, las variables independientes solo se presentan en términos de sus funciones de distribución. De este modo, no necesitan estar definidas en el mismo espacio de probabilidad (Ω, \mathcal{F}, P) (solo importa la forma funcional)
- La notación que se utiliza para denotar este tipo de convergencia es $X_n \xrightarrow{d} X$ cuando $n \rightarrow \infty$. No obstante, se suele hacer un abuso conveniente de la notación y se suele denotar con la distribución misma de X , tal como $X_n \xrightarrow{d} F_X$ cuando $n \rightarrow \infty$
- La desigualdad de Chebyshev generalizada permite estudiar la convergencia en probabilidad de manera más sencilla y, además, permite derivar desigualdades complementarias, tales como la desigualdad de Chebyshev y la de Markov
- La desigualdad de Chebyshev generalizada expresa que, siendo $u: \mathbb{R} \rightarrow [0, \infty)$ una función no negativa, si $E[u(Y)]$ existe, entonces para cualquier $\varepsilon > 0$ se cumple la siguiente desigualdad

$$P(u(Y) \geq \varepsilon) \leq \frac{E[u(Y)]}{\varepsilon}$$

Commented [MOU16]: Preguntar

- Si $\mu = E(Y)$ existe, tomando $u(y) \equiv |y - \mu|^r$ y $\tilde{c} \equiv c^r$ se obtiene la desigualdad de Markov. Esta expresa que, para cualquier $r > 0$ y $\varepsilon > 0$, se obtiene la siguiente desigualdad:

$$P(|Y - \mu|^r \geq \varepsilon^r) \leq \frac{E[|Y - \mu|^r]}{\varepsilon^r}$$

- Si $r = 2$ y $Var(Y)$ es constante y existe, entonces se puede obtener la desigualdad de Chebyshev. Esta expresa que, para $r = 2$, se obtiene la siguiente desigualdad:

$$P((Y - \mu)^2 \geq \varepsilon^2) = P(|Y - \mu| \geq \varepsilon) \leq \frac{E[(Y - \mu)^2]}{\varepsilon^2} = \frac{Var(Y)}{\varepsilon^2}$$

- Como estas son consecuencia de la desigualdad de Chebyshev generalizada, solo hace falta demostrar esta para demostrar las otras dos. La demostración se basa en la definición de esperanza y de función de distribución

$$\begin{aligned} E[u(Y)] &= \int_{\mathbb{R}} u(y)f(y) dy = \\ &= \int_{\{y: u(y) \geq c\}} u(y)f(y) dy + \int_{\{y: u(y) < c\}} u(y)f(y) dy \geq \\ &\geq \int_{\{y: u(y) \geq c\}} u(y)f(y) dy \end{aligned}$$

$$u(y)f(y) \geq 0 \Rightarrow E[u(Y)] \geq c \int_{\{y: u(y) \geq c\}} f(y) dy = cP(u(Y) \geq c)$$

- Para la convergencia en probabilidad, la convergencia en media r y la convergencia casi segura, uno puede demostrar que la convergencia de Cauchy es equivalente
 - Para la convergencia en probabilidad y la convergencia en media r , uno puede demostrar este resultado a través del resultado de la completitud de los espacios L^r y a través de la implicación de la convergencia de probabilidad si se converge en media r
 - La convergencia casi segura se deriva del resultado correspondiente para números reales
- La convergencia de variables aleatorias es única, de modo que la variable aleatoria en el límite se define únicamente

- Si $X_n \rightarrow X$ y $X_n \rightarrow Y$ de manera casi segura, en probabilidad o en media r , entonces $X = Y$ de manera casi segura, por lo que se cumple la siguiente condición:

$$P(\{\omega \in \Omega: X(\omega) = Y(\omega)\}) = 1$$

- En el caso de convergencia en distribución, la unicidad significa que $F_X(x) = F_Y(x)$ para toda x

$$F_X(x) \xrightarrow{d} F_Y(x)$$

- En análisis matemático se dice que, para una secuencia de números a_1, a_2, \dots un límite es único cuando existen dos números reales a y b tales que se cumpla la siguiente condición:

$$a_n \rightarrow a \quad \& \quad a_n \rightarrow b \quad \text{as } n \rightarrow \infty$$

- En este caso, necesariamente ocurre que $a = b$, dado que, a través de la desigualdad triangular, se pueden obtener los siguientes resultados:

$$|a - b| \leq |a - a_n| + |b - a_n|$$

$$\Rightarrow |a - a_n| + |b - a_n| \rightarrow 0 + 0 \quad \text{as } n \rightarrow \infty$$

$$\Rightarrow |a - b| \leq 0 \Rightarrow |a - b| = 0 \Rightarrow a = b$$

- Siendo X_1, X_2, \dots una secuencia de variables aleatorias, si X_n converge de manera casi segura, en probabilidad, en media r o en distribución cuando $n \rightarrow \infty$, entonces la variable aleatoria en el límite es única

- Suponiendo que $X_n \xrightarrow{a.s.} X$ y que $X_n \xrightarrow{a.s.} Y$ cuando $n \rightarrow \infty$ y definiendo los conjuntos $\bar{R}_X = \{\omega: X_n(\omega) \nrightarrow X(\omega) \text{ as } n \rightarrow \infty\}$ y $\bar{R}_Y = \{\omega: X_n(\omega) \nrightarrow Y(\omega) \text{ as } n \rightarrow \infty\}$, entonces se puede ver como $P(\bar{R}_X) = P(\bar{R}_Y) = 0$ y, estableciendo $\omega \notin (\bar{R}_X \cup \bar{R}_Y)$ se obtienen los siguientes resultados:

$$|X(\omega) - Y(\omega)| \leq |X(\omega) - X_n(\omega)| + |X_n(\omega) - Y(\omega)|$$

$$\Rightarrow |X(\omega) - X_n(\omega)| + |X_n(\omega) - Y(\omega)| \rightarrow 0 + 0 \quad \text{as } n \rightarrow \infty$$

$$\Rightarrow |X(\omega) - Y(\omega)| \leq 0 \quad \text{as } n \rightarrow \infty$$

$$\Rightarrow X(\omega) = Y(\omega) \quad \text{for } \forall \omega \notin (\bar{R}_X \cup \bar{R}_Y) \quad \text{as } n \rightarrow \infty$$

$$\Rightarrow P(X \neq Y) \leq P(\bar{R}_X \cup \bar{R}_Y) \leq P(\bar{R}_X) + P(\bar{R}_Y) = 0$$

- Suponiendo que $X_n \xrightarrow{p} X$ y que $X_n \xrightarrow{p} Y$ cuando $n \rightarrow \infty$ y definiendo $\varepsilon > 0$, entonces, si $|X - Y| > \varepsilon$, no puede ser que $|X - X_n| > \varepsilon/2$ ni que $|Y - X_n| > \varepsilon/2$ (si no, $|X - Y| \leq \varepsilon$), de modo que se obtienen las siguientes implicaciones:

$$|X - Y| \leq |X - X_n| + |X_n - Y|$$

$$\Rightarrow \{\omega: |X - Y| > \varepsilon\} \subseteq \left\{\omega: |X - X_n| > \frac{\varepsilon}{2}\right\} + \left\{\omega: |X_n - Y| > \frac{\varepsilon}{2}\right\}$$

$$\Rightarrow P(|X - Y| > \varepsilon) \leq P\left(|X - X_n| > \frac{\varepsilon}{2}\right) + P\left(|X_n - Y| > \frac{\varepsilon}{2}\right)$$

$$\Rightarrow \begin{cases} P\left(|X - X_n| > \frac{\varepsilon}{2}\right) = 0 & \text{as } n \rightarrow \infty \\ P\left(|X_n - Y| > \frac{\varepsilon}{2}\right) = 0 & \text{as } n \rightarrow \infty \end{cases}$$

$$\Rightarrow P(|X - Y| > \varepsilon) = 0 \text{ as } n \rightarrow \infty \text{ for } \forall \varepsilon > 0$$

$$\Rightarrow P(|X - Y| > 0) = 0 \text{ as } n \rightarrow \infty \text{ for } \forall \varepsilon > 0$$

$$\Rightarrow P(X = Y) = 1 \text{ as } n \rightarrow \infty \text{ for } \forall \varepsilon > 0$$

- Suponiendo que $X_n \xrightarrow{d} X$ y que $X_n \xrightarrow{d} Y$ cuando $n \rightarrow \infty$ y suponiendo que $x \in C(F_X) \cap C(F_Y)$, entonces se obtienen las siguientes implicaciones:

$$|F_X(x) - F_Y(x)| \leq |F_X(x) - F_{X_n}(x)| + |F_{X_n}(x) - F_Y(x)|$$

$$\Rightarrow |F_X(x) - F_{X_n}(x)| + |F_{X_n}(x) - F_Y(x)| \rightarrow 0 + 0 \text{ as } n \rightarrow \infty$$

$$\Rightarrow |F_X(x) - F_Y(x)| = 0 \text{ as } n \rightarrow \infty$$

$$\Rightarrow F_X(x) = F_Y(x) \text{ for } \forall x \in C(F_X) \cap C(F_Y)$$

- Debido a que $\{C(F_X) \cap C(F_Y)\}^c$ tiene como máximo un número numerable de puntos (ya que solo puede haber una cantidad numerable de puntos de discontinuidad), entonces se puede demostrar que $F_X(x) = F_Y(x)$ para toda x (los puntos de discontinuidad tienen asignada una probabilidad nula)
- Para poder llevar a cabo la demostración del teorema anterior para la media r , es necesario utilizar una desigualdad concreta. Siendo $r > 0$, y suponiendo que U y V son variables aleatorias con media finita, entonces se cumple la siguiente desigualdad:

$$E(|U + V|^r) \leq 2^r [E(|U|^r) + E(|V|^r)]$$

- Siendo a y b dos números reales, entonces se puede obtener la siguiente desigualdad:

$$\begin{aligned} |a + b|^r &\leq (|a| + |b|)^r \leq (2 \max\{|a|, |b|\})^r = \\ &= 2^r \max\{|a|^r, |b|^r\} \leq 2^r (|a|^r + |b|^r) \end{aligned}$$

- Por lo tanto, considerando $\omega \in \Omega$, se puede obtener la desigualdad deseada:

$$|U(\omega) + V(\omega)|^r \leq 2^r (|U(\omega)|^r + |V(\omega)|^r)$$

- Suponiendo que $X_n \xrightarrow{r} X$ y que $X_n \xrightarrow{r} Y$ cuando $n \rightarrow \infty$ y suponiendo que $r, c > 0$, entonces se puede obtener el siguiente resultado que demuestra la unicidad en media r :

$$\begin{aligned} E(|Y - X|^r) &\leq 2^r [E(|X - X_n|^r) + E(|X_n - Y|^r)] \\ &\Rightarrow \begin{cases} E(|X - X_n|^r) \rightarrow 0 & \text{as } n \rightarrow \infty \\ E(|X_n - Y|^r) \rightarrow 0 & \text{as } n \rightarrow \infty \end{cases} \\ &\Rightarrow (|X - X_n|^r) + E(|X_n - Y|^r) \rightarrow 0 + 0 \quad \text{as } n \rightarrow \infty \\ &\Rightarrow E(|Y - X|^r) \rightarrow 0 \quad \text{as } n \rightarrow \infty \\ &\Rightarrow P(|X - Y|^r \geq \varepsilon^r) \leq \frac{E[|X - Y|^r]}{\varepsilon^r} \rightarrow 0 \quad \text{as } n \rightarrow \infty \\ &\text{as } \varepsilon, r > 0 \Rightarrow P(|X - Y|^r \geq 0) \rightarrow 0 \quad \text{as } n \rightarrow \infty \\ &\Rightarrow P(X = Y) = 1 \quad \text{as } n \rightarrow \infty \end{aligned}$$

- Los teoremas de unicidad para la convergencia casi segura y la convergencia en media r tienen como corolario el teorema de unicidad para la convergencia en probabilidad

- Suponiendo explícitamente que $X_n \xrightarrow{a.s.} X$ y que $X_n \xrightarrow{a.s.} Y$ cuando $n \rightarrow \infty$, entonces también se tiene que $X_n \xrightarrow{P} X$ y que $X_n \xrightarrow{P} Y$ cuando $n \rightarrow \infty$ (debido a la relación entre los conceptos de convergencia, que se presenta en la siguiente sección) y, por el teorema de unicidad, que $P(X = Y) = 1$
- Lo mismo ocurre con la convergencia en media r , de modo que si se supone que $X_n \xrightarrow{r} X$ y que $X_n \xrightarrow{r} Y$, entonces también se tiene que $X_n \xrightarrow{P} X$ y que $X_n \xrightarrow{P} Y$ cuando $n \rightarrow \infty$ (debido a la

relación entre los conceptos de convergencia, que se presenta en la siguiente sección)

- Los diferentes conceptos de convergencia presentados tienen relación entre ellos y es posible establecer un orden entre estos conceptos

- La convergencia casi segura $X_n \xrightarrow{a.s.} X$ cuando $n \rightarrow \infty$ implica la convergencia en probabilidad $X_n \xrightarrow{P} X$ cuando $n \rightarrow \infty$

- Si $X_n \xrightarrow{a.s.} X$ cuando $n \rightarrow \infty$, significa que el conjunto de puntos $\bar{R} = \{\omega : X_n(\omega) \not\rightarrow X(\omega) \text{ as } n \rightarrow \infty\}$ tiene medida nula. Fijando un número $\varepsilon > 0$ y considerando una secuencia de conjuntos como $A_n = \bigcup_{m \geq n} \{|X_m - X| \geq \varepsilon\}$, se puede ver como la secuencia de conjuntos es decreciente $A_n \supseteq A_{n+1} \supseteq \dots$ cuanto más aumenta n y tiende al conjunto $A^* = \bigcap_{n \geq 1} A_n = \bigcap_{n \geq 1} \bigcup_{m \geq n} \{|X_m - X| \geq \varepsilon\}$
- Para esta secuencia decreciente de eventos, sus probabilidades también son una secuencia decreciente, y decrece hacia $P(A^*)$. Cualquier punto $\omega \in R$ es tal que $X_n(\omega) \not\rightarrow X(\omega)$ cuando $n \rightarrow \infty$, lo que implica que $|X_n - X| < \varepsilon$ para toda $n < N$, en donde N es un número entero

$$\{|X_n - X| \geq \varepsilon\} \subseteq A_n \Rightarrow P(|X_n - X| \geq \varepsilon) \leq P(A_n)$$

- Por lo tanto, para toda $n \geq N$, el punto ω no pertenecerá al conjunto A_n y, consecuentemente, no pertenecerá a A^* . Eso significa que $A^* \cap R = \emptyset$ o, equivalentemente, $A^* \subset \bar{R}$ y eso implica que $P(A^*) < P(\bar{R}) = 0$ y que $P(A^*) = 0$, por lo tanto, se obtiene el siguiente resultado:

$$\begin{aligned} \lim_{n \rightarrow \infty} P(|X_n - X| \geq \varepsilon) &\leq \lim_{n \rightarrow \infty} P(A_n) = \sum_{n=m}^{\infty} P(A_m) \\ &\Rightarrow \lim_{n \rightarrow \infty} P(A_n) = \sum_{n=m}^{\infty} P(A_m) \leq \sum_{n=m}^{\infty} P(\bar{R}) = 0 \\ &\Rightarrow \lim_{n \rightarrow \infty} P(|X_n - X| \geq \varepsilon) \leq \lim_{n \rightarrow \infty} P(A_n) \leq \lim_{n \rightarrow \infty} P(\bar{R}) = 0 \\ &\Rightarrow \lim_{n \rightarrow \infty} P(|X_n - X| \geq \varepsilon) = 0 \end{aligned}$$

- Lo converso no se da debido a que, aunque $\lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) = 0$, la desigualdad $\lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) \leq \lim_{n \rightarrow \infty} P(A_n)$ se sigue cumpliendo sin necesidad de que $\lim_{n \rightarrow \infty} P(A_n) = 0$

- La convergencia de media r $X_n \xrightarrow{r} X$ para una $r \geq 1$ cualquiera cuando $n \rightarrow \infty$ implica la convergencia en probabilidad $X_n \xrightarrow{P} X$ cuando $n \rightarrow \infty$

- Esta implicación se puede demostrar a través de la desigualdad generalizada de Markov

$$E[|X_n - X|^r] \rightarrow 0 \text{ as } n \rightarrow \infty$$

$$\Rightarrow \lim_{n \rightarrow \infty} P(|X_n - X| \geq \varepsilon) \leq \lim_{n \rightarrow \infty} \frac{E[|X_n - X|^r]}{\varepsilon^r} = 0$$

$$\Rightarrow \lim_{n \rightarrow \infty} P(|X_n - X| \geq \varepsilon) = 0$$

- Lo converso no se da debido a que $E[|X_n - X|^r]$ puede no existir. No obstante, puede ser que haya casos en donde $X_n \xrightarrow{P} X$ cuando $n \rightarrow \infty$, mientras que $E[|X_n - X|^r] \nrightarrow 0$ cuando $n \rightarrow \infty$
- Esta relación permite demostrar la convergencia en probabilidad de manera sencilla, dado que se puede utilizar la desigualdad de Chebyshev o de Markov junto a $E[|X_n - X|^r]$ para una $r \geq 1$ cualquiera (normalmente se escoge $r = 2$ por simplicidad), pero primero se tiene que demostrar $E[|X_n - X|^r] \rightarrow 0$
- Los conceptos de convergencia casi segura y convergencia en media r no se puede ordenar, dado que ninguno implica el otro (no hay relación entre ambas)
 - Hay veces en donde se da uno y no el otro, veces en donde se dan los dos y veces en las que no se da ninguna
- La convergencia de media r $X_n \xrightarrow{r} X$ cuando $n \rightarrow \infty$ implica la convergencia en media $X_n \xrightarrow{s} X$ para $s \in [1, r)$ cuando $n \rightarrow \infty$
 - Este teorema se puede demostrar a través de la desigualdad de Hölder:

$$E(|AB|) \leq E(|A|^p)^{\frac{1}{p}} E(|B|^q)^{\frac{1}{q}} \text{ for } 1 < p, q < \infty \text{ \& } \frac{1}{p} + \frac{1}{q} = 1$$

$$\text{Choosing } A = |X_n - X|^r, Y = 1 \text{ \& } p = \frac{s}{r} > 1:$$

$$\Rightarrow E(|X_n - X|^r) \leq E(|X_n - X|^s)^{\frac{1}{p}}$$

$$E(|X_n - X|^r) \rightarrow 0 \text{ as } n \rightarrow \infty \Rightarrow E(|X_n - X|^s)^{\frac{1}{p}} \rightarrow 0 \text{ as } n \rightarrow \infty$$

$$\Rightarrow X_n \xrightarrow{s} X$$

- Por lo tanto, solo hace falta demostrar que se converge en media r para demostrar que se converge en media $s \in [1, r)$, lo cual simplifica algunos cálculos y procedimientos

- La convergencia en probabilidad $X_n \xrightarrow{P} X$ cuando $n \rightarrow \infty$ implica la convergencia en distribución $X_n \xrightarrow{D} X$ cuando $n \rightarrow \infty$

- Siendo $\varepsilon > 0$, entonces se puede obtener el siguiente resultado:

$$F_{X_n}(x) = P(X_n \leq x) = P(\{X_n \leq x + \varepsilon\} \cap \{|X_n - X| < \varepsilon\}) +$$

$$P(\{X_n \leq x\} \cap \{|X_n - X| \geq \varepsilon\}) \leq P(\{X \leq x + \varepsilon\} \cap \{|X_n - X| < \varepsilon\}) +$$

$$P(\{|X_n - X| \geq \varepsilon\}) \leq P(X \leq x + \varepsilon) + P(|X_n - X| \geq \varepsilon)$$

$$\Rightarrow F_{X_n}(x) \leq F_X(x + \varepsilon) + P(|X_n - X| \geq \varepsilon)$$

By switching X_n for X and x to $x - \varepsilon$, analogously:

$$F_X(x - \varepsilon) = P(X \leq x - \varepsilon) = P(\{X \leq x - \varepsilon\} \cap \{|X - X_n| < \varepsilon\}) +$$

$$P(\{X \leq x - \varepsilon\} \cap \{|X - X_n| \geq \varepsilon\}) \leq P(\{X_n \leq x\} \cap \{|X - X_n| < \varepsilon\}) +$$

$$P(\{|X - X_n| \geq \varepsilon\}) \leq P(X_n \leq x) + P(|X - X_n| \geq \varepsilon)$$

$$\Rightarrow F_X(x - \varepsilon) \leq F_{X_n}(x) + P(|X_n - X| \geq \varepsilon)$$

- Como $X_n \xrightarrow{P} X$ cuando $n \rightarrow \infty$, se obtiene el siguiente resultado cuando se toma el límite en las dos desigualdades derivadas anteriormente:

$$F_X(x) \leq F_X(x + \varepsilon) \text{ as } n \rightarrow \infty$$

$$F_X(x - \varepsilon) \leq F_X(x) \text{ as } n \rightarrow \infty$$

$$\Rightarrow F_X(x - \varepsilon) \leq \liminf_{n \rightarrow \infty} F_{X_n}(x) \leq \limsup_{n \rightarrow \infty} F_{X_n}(x) \leq F_X(x + \varepsilon)$$

- Este resultado se mantiene para toda x y $\varepsilon > 0$, por lo que, para demostrar la convergencia en distribución, solo es necesario suponer que $x \in C(F_X)$ (un conjunto de continuidad arbitrario) y que $\varepsilon \rightarrow 0$ cuando $n \rightarrow \infty$

$$F_X(x) \leq \liminf_{n \rightarrow \infty} F_{X_n}(x) \leq \limsup_{n \rightarrow \infty} F_{X_n}(x) \leq F_X(x)$$

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x)$$

- Si F_X tiene un salto en x , entonces solo se puede concluir la siguiente desigualdad, la cual explica por qué solo los puntos de continuidad se incluyen en la definición de distribución de convergencia (en este caso, $F_X(x) - F_X(x-)$ es el tamaño del salto):

$$F_X(x-) \leq \liminf_{n \rightarrow \infty} F_{X_n}(x) \leq \limsup_{n \rightarrow \infty} F_{X_n}(x) \leq F_X(x)$$

- Siendo X y X_1, X_2, \dots variables aleatorias, las siguientes implicaciones se mantienen cuando $n \rightarrow \infty$, siendo todas las implicaciones estrictas:

$$\begin{array}{c} X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{D} X \\ \uparrow \\ X_n \xrightarrow{r} X \Rightarrow X_n \xrightarrow{s} X \text{ for } s \in [1, r) \end{array}$$

- Tal y como se ha mostrado antes, los conversos de estas implicaciones no se mantienen
- Debido a que $X_n \xrightarrow{a.s.} X$ y $X_n \xrightarrow{r} X$ implican $X_n \xrightarrow{P} X$ cuando $n \rightarrow \infty$, es lógico ver que se cumplen las siguientes implicaciones:

$$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{D} X$$

$$X_n \xrightarrow{r} X \Rightarrow X_n \xrightarrow{D} X$$

- Debido a que la contraposición es equivalente a la implicación, entonces se puede demostrar como se cumplen las siguientes implicaciones, las cuales permiten demostrar la no convergencia y sus relaciones:

$$\left[X_n \xrightarrow{P} X \right]^c \Rightarrow \left[X_n \xrightarrow{a.s.} X \right]^c \wedge \left[X_n \xrightarrow{r} X \right]^c$$

$$\left[X_n \xrightarrow{D} X \right]^c \Rightarrow \left[X_n \xrightarrow{P} X \right]^c$$

$$\left[X_n \xrightarrow{D} X \right]^c \Rightarrow \left[X_n \xrightarrow{a.s.} X \right]^c \wedge \left[X_n \xrightarrow{r} X \right]^c$$

- Además de estos resultados sobre las relaciones entre conceptos de convergencia, se pueden considerar algunos resultados adicionales a través de suposiciones adicionales
 - Siendo X_1, X_2, \dots variables aleatorias y c una constante, entonces se da la siguiente implicación:

$$X_n \xrightarrow{P} c \text{ as } n \rightarrow \infty \Leftrightarrow X_n \xrightarrow{D} \delta(x - c) \text{ as } n \rightarrow \infty$$

$$\text{where } \delta(x - c) = \begin{cases} \infty & \text{if } x = c \\ 0 & \text{if } x \neq c \end{cases}$$

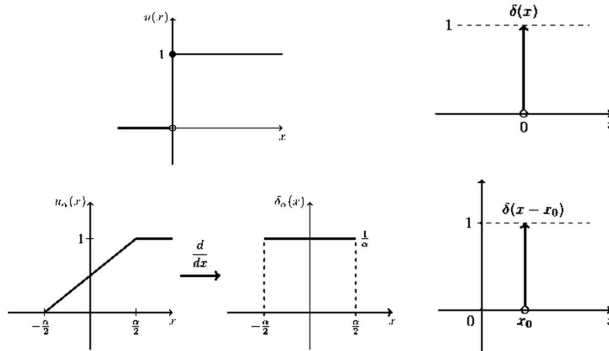
- La función $\delta(x - c)$ es la función delta de Dirac, la cual se define a partir del límite cuando $\alpha \rightarrow 0$ de la derivada de una función de paso unitario o *unitary step function*

$$u_\alpha(x) = \begin{cases} 1 & \text{for } x > \frac{\alpha}{2} \\ \frac{1}{\alpha} \left(x + \frac{\alpha}{2} \right) & \text{for } -\frac{\alpha}{2} \leq x \leq \frac{\alpha}{2} \\ 0 & \text{for } x < -\frac{\alpha}{2} \end{cases}$$

$$\Rightarrow \delta_\alpha(x) = \frac{\partial u_\alpha(x)}{\partial \alpha} = \begin{cases} \frac{1}{\alpha} & \text{for } |x| < \frac{\alpha}{2} \\ 0 & \text{for } |x| > \frac{\alpha}{2} \end{cases}$$

$$\Rightarrow \delta(x) = \lim_{\alpha \rightarrow 0} \frac{\partial u_\alpha(x)}{\partial \alpha} = \begin{cases} \infty & \text{if } x = 0 \\ 0 & \text{if } x \neq 0 \end{cases}$$

$$\Rightarrow \delta(x - x_0) = \begin{cases} \infty & \text{if } x = x_0 \\ 0 & \text{if } x \neq x_0 \end{cases}$$



- Debido a que la implicación hacia la derecha siempre se mantiene, solo hace falta demostrar el converso. Asumiendo que $X_n \xrightarrow{D} \delta(x - c)$ cuando $n \rightarrow \infty$ y que $\varepsilon > 0$, entonces se obtienen los siguientes resultados:

$$P(|X_n - c| > \varepsilon) = 1 - P(c - \varepsilon \leq X_n \leq c + \varepsilon) =$$

$$= 1 - F_{X_n}(c + \varepsilon) + F_{X_n}(c - \varepsilon) \leq 1 - F_{X_n}(c + \varepsilon) + F_{X_n}(c - \varepsilon)$$

$$\Rightarrow \lim_{n \rightarrow \infty} 1 - F_{X_n}(c + \varepsilon) + F_{X_n}(c - \varepsilon) = 1 - 1 + 0 = 0$$

- Esto ocurre debido a que $X_n \xrightarrow{D} \delta(x - c)$ cuando $n \rightarrow \infty$, entonces $F_{X_n}(c + \varepsilon) \rightarrow F_X(c + \varepsilon) = 1$ y $F_{X_n}(c - \varepsilon) \rightarrow F_X(c - \varepsilon) = 0$, y $c + \varepsilon$ y $c - \varepsilon \in \mathcal{C}(F_X) = \{x: x \neq c\}$
- Si $X_n \xrightarrow{P} X$ y existe una constante C tal que $P(|X_n| \leq C) = 1$ para toda n , entonces $X_n \xrightarrow{r} X$ para toda $r \geq 1$

- Debido a que $X_n \xrightarrow{P} X$, se ve como $P(|X_n - X| \leq \varepsilon) = 1 \dots$

Commented [MOU17]: Proof

- Si $P_n(\varepsilon) = P(|X_n - X| > \varepsilon)$ satisface que $\sum_n P_n(\varepsilon) < \infty$ para toda $\varepsilon > 0$, entonces $X_n \xrightarrow{a.s.} X$

Commented [MOU18]: No seria >=?

- ...

Commented [MOU19]: Proof

- Este teorema permite demostrar de manera simple la convergencia casi segura de una variable si se tiene una expresión explícita para las probabilidades $P_n(\varepsilon)$. Solo será necesario evaluar si $\sum_n P_n(\varepsilon) < \infty$ para esta expresión
- Siendo X_1, X_2, \dots una secuencia de variables aleatorias en L^r (con $r \geq 1$), entonces existe una X en L^r tal que $X_n \xrightarrow{r} X$ si, y solo si, se cumple la siguiente igualdad:

$$\lim_{m \rightarrow \infty} \sup_{n \geq m} E(|X_n - X_m|^r) = 0$$

- Aquellas secuencias que satisfacen esta condición se conocen como secuencias de Cauchy en L^r , y este criterio se conoce como el criterio de Cauchy para L^2 . Como son secuencias de Cauchy que cumplen esta condición, entonces deben pertenecer al espacio L^r
- La proposición es un resultado que se conoce como la completitud de L^r , dado que aquellos espacios métricos en donde todas las secuencias de Cauchy convergen se denomina completo
- Este resultado es muy útil porque permite ver que, si es posible demostrar que X_1, X_2, \dots es una secuencia de Cauchy en L^r (cuya

condición para serlo es la condición de arriba), entonces es posible demostrar que la secuencia converge a alguna X (aunque esta sea desconocida)

- Igual que las transformaciones de variables aleatorias eran útiles para determinar la distribución de nuevas variables aleatorias, estas también se pueden utilizar para demostrar la convergencia en distribución

- Siendo X, X_1, X_2, \dots variables aleatorias enteras no negativas y suponiendo que $G_{X_n}(s) \rightarrow G_X(s)$ cuando $n \rightarrow \infty$, entonces X_n converge en distribución a X cuando $n \rightarrow \infty$

$$G_{X_n}(s) \rightarrow G_X(s) \text{ as } n \rightarrow \infty \Rightarrow X_n \xrightarrow{D} X \text{ as } n \rightarrow \infty$$

▪ ...

Commented [MOU20]: Proof

- Siendo X_1, X_2, \dots variables aleatorias tal que $M_{X_n}(t)$ exista para $|t| < h$ para alguna $h > 0$ y para toda n y suponiendo que X es una variable aleatoria cuya función generadora de momentos $M_X(t)$ existe para $|t| < h_1 < h$ para $h_1 > 0$ y que $M_{X_n}(t) \rightarrow M_X(t)$ cuando $n \rightarrow \infty$ para $|t| \leq h_1$, entonces X_n converge en distribución a X cuando $n \rightarrow \infty$

$$M_{X_n}(t) \rightarrow M_X(t) \text{ as } n \rightarrow \infty \text{ for } |t| \leq h_1 \Rightarrow X_n \xrightarrow{D} X \text{ as } n \rightarrow \infty$$

▪ ...

Commented [MOU21]: Proof

- Siendo X, X_1, X_2, \dots variables aleatorias y suponiendo que $\phi_{X_n}(t) \rightarrow \phi_X(t)$ cuando $n \rightarrow \infty$, entonces X_n converge en distribución a X cuando $n \rightarrow \infty$

$$\phi_{X_n}(t) \rightarrow \phi_X(t) \text{ as } n \rightarrow \infty \text{ for } t \in (-\infty, \infty)$$

$$\Rightarrow X_n \xrightarrow{D} X \text{ as } n \rightarrow \infty$$

- El converso de esta implicación también es cierto, dado que uno puede demostrar que si X_1, X_2, \dots es una secuencia de variables aleatorias tal que $X_n \xrightarrow{D} X$ cuando $n \rightarrow \infty$ para una variable aleatoria X , entonces se cumple la siguiente implicación:

$$X_n \xrightarrow{D} X \text{ as } n \rightarrow \infty$$

$$\Rightarrow \phi_{X_n}(t) \rightarrow \phi_X(t) \text{ as } n \rightarrow \infty \text{ for } t \in (-\infty, \infty)$$

- Por lo tanto, X_n converge en distribución a X cuando $n \rightarrow \infty$ si, y solo si, $\phi_{X_n}(t) \rightarrow \phi_X(t)$ cuando $n \rightarrow \infty$ para $t \in (-\infty, \infty)$

$$X_n \xrightarrow{D} X \text{ as } n \rightarrow \infty$$

$$\Leftrightarrow \phi_{X_n}(t) \rightarrow \phi_X(t) \text{ as } n \rightarrow \infty \text{ for } t \in (-\infty, \infty)$$

- De este modo, este teorema se puede utilizar para poder demostrar si $X_n \xrightarrow{D} X$ a través de la convergencia de las funciones características
- El teorema para las funciones características se puede refinar a través de añadir una suposición sobre la continuidad de $\phi_X(t)$ en $t = 0$, obteniendo así el teorema de continuidad de Lévy
 - Si $\phi_X(t) = \lim_{n \rightarrow \infty} \phi_{X_n}(t)$ existe y es continua en $t = 0$, entonces $\phi_X(t)$ es la función característica de una función de distribución $F_X(x)$ y $F_{X_n}(x) \rightarrow F_X(x)$ en cada punto x donde la función límite $F_X(x)$ es continua
 - Si para alguna función de distribución $F_X(x)$ con una función característica $\phi_X(t)$ se tiene que $F_{X_n}(x) \rightarrow F_X(x)$ en cada punto x donde $F_X(x)$ es continua, entonces $\phi_{X_n}(t) \rightarrow \phi_X(t)$ para toda $t \in \mathbb{R}$
 - La formulación de este teorema supone que el límite de $\phi_{X_n}(t)$ es una función característica y que esta es una función característica de una variable X conocida
 - Este teorema es de gran utilidad porque permite demostrar la convergencia en distribución a través de los límites comunes (el límite para cada punto x o t), y se puede usar para demostrar la ley de grandes números y el teorema del límite central
- Siendo X_1, X_2, \dots variables aleatorias y suponiendo que, para un número real c , $\phi_{X_n}(t) \rightarrow e^{itc}$ cuando $n \rightarrow \infty$ para $-\infty < t < \infty$, entonces X_n tiende en probabilidad a c cuando $n \rightarrow \infty$

$$\phi_{X_n}(t) \rightarrow e^{itc} \text{ as } n \rightarrow \infty \text{ for } -\infty < t < \infty$$

$$\Rightarrow X_n \xrightarrow{P} c \text{ as } n \rightarrow \infty$$

- Para demostrar y operar con este teorema, es necesario saber cuál es la función de distribución y de densidad de probabilidad de la función delta de Dirac y su función característica

$$f_X(x) = \delta(x - x_0) \text{ for } -\infty < x < \infty \text{ \& } -\infty < x_0 < \infty$$

$$\Rightarrow F_X(x) = \begin{cases} 1 & \text{for } x \geq x_0 \\ 0 & \text{for } x < x_0 \end{cases}$$

$$\Rightarrow \phi_X = E(e^{itX}) = \sum_{x=-\infty}^{\infty} e^{itx} P(X=x) = e^{it \circ} P(X=x_0) = e^{it \circ}$$

- La demostración de este teorema se basa en el hecho de que la convergencia en distribución es equivalente a la convergencia en probabilidad en el caso de una constante. Debido a que $\phi_{X_n}(t)$ tiende a la función característica de X (que es e^{itc}), eso quiere decir que $X_n \xrightarrow{D} \delta(x-c)$, lo cual equivale, por el teorema anterior a que $X_n \xrightarrow{P} c$
- A partir de los teoremas de convergencia para transformaciones, se puede demostrar que la distribución binomial se puede aproximar a través de la distribución de Poisson cuando $n \rightarrow \infty$
- Dada una secuencia de variables $X_n \sim \text{Bin}(n, \lambda/n)$ para $n = 1, 2, \dots$ se puede ver como la función generadora de probabilidad converge a la función generadora de probabilidad de la Poisson cuando $n \rightarrow \infty$

$$G_{X_n}(s) = (q + ps)^n = \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n}s\right)^n = \left(1 + \frac{\lambda(s-1)}{n}\right)^n$$

$$\Rightarrow \lim_{n \rightarrow \infty} G_{X_n}(s) = \lim_{n \rightarrow \infty} \left(1 + \frac{\lambda(s-1)}{n}\right)^n = e^{\lambda(s-1)} = G_{Po(\lambda)}(s)$$

$$\Rightarrow G_{X_n}(s) \rightarrow G_{Po(\lambda)}(s) \text{ as } n \rightarrow \infty$$

- Al demostrar la unicidad para la convergencia en distribución había problemas con los puntos de continuidad, pero cuando se utiliza el teorema anterior sobre la convergencia de funciones características, estos problemas no se tienen en cuenta
- La razón es que los teoremas de unicidad y los de convergencia para las transformaciones implican la unicidad en distribución y la convergencia en distribución de las variables aleatorias, y los eventos en conjuntos con probabilidad nula no importan en estos últimos teoremas
- Asumiendo que X_1, X_2, \dots son variables aleatorias tal que $X_n \xrightarrow{D} X$ y que $X_n \xrightarrow{D} Y$ cuando $n \rightarrow \infty$, entonces se puede obtener el siguiente resultado:

$$\begin{aligned}
& \phi_{X_n}(t) \rightarrow \phi_X(t) \text{ \& } \phi_{X_n}(t) \rightarrow \phi_Y(t) \text{ as } n \rightarrow \infty \\
& \Rightarrow |\phi_X(t) - \phi_Y(t)| \leq |\phi_X(t) - \phi_{X_n}(t)| + |\phi_{X_n}(t) - \phi_Y(t)| \\
& \Rightarrow |\phi_X(t) - \phi_{X_n}(t)| + |\phi_{X_n}(t) - \phi_Y(t)| \rightarrow 0 \text{ as } n \rightarrow \infty \\
& \Rightarrow \phi_X(t) = \phi_Y(t) \Rightarrow X \sim Y
\end{aligned}$$

- Los dos resultados más fundamentales en la teoría de la probabilidad son la ley de grandes números y el teorema del límite central, de modo que estos se tienen que estudiar con más detalle

- Siendo X_1, X_2, \dots variables aleatorias i.i.d con esperanza finita μ y fijando $S_n = X_1 + X_2 + \dots + X_n$ para $n \geq 1$, entonces se cumple la siguiente convergencia en probabilidad:

$$\bar{X}_n = \frac{S_n}{n} \xrightarrow{P} \mu \text{ as } n \rightarrow \infty$$

- A partir del corolario sobre la convergencia en probabilidad para una constante a través de la función característica, se puede obtener el siguiente resultado:

$$\begin{aligned}
\phi_{\bar{X}_n}(t) &= E(e^{it\bar{X}_n}) = E\left(e^{i\left(\frac{t}{n}\right)S_n}\right) = \phi_{S_n}\left(\frac{t}{n}\right) = \\
&= E\left(e^{i\left(\frac{t}{n}\right)(X_1 + \dots + X_n)}\right) = E\left(e^{i\left(\frac{t}{n}\right)X_1}\right) \dots E\left(e^{i\left(\frac{t}{n}\right)X_1}\right) = \left[\phi_{X_1}\left(\frac{t}{n}\right)\right]^n
\end{aligned}$$

$$\Rightarrow \phi_{\bar{X}_n}(t) = \left[\phi_{X_1}\left(\frac{t}{n}\right)\right]^n = \left(1 + i\frac{t}{n}\mu + O\left(\frac{t}{n}\right)\right)^n$$

$$\Rightarrow \lim_{n \rightarrow \infty} \left(1 + \frac{it\mu}{n} + O\left(\frac{t}{n}\right)\right)^n = e^{it\mu} \text{ as } O\left(\frac{t}{n}\right) \rightarrow 0$$

$$\Rightarrow \phi_{\bar{X}_n}(t) \rightarrow e^{it\mu} \text{ as } n \rightarrow \infty \Rightarrow \bar{X}_n \xrightarrow{P} \mu \text{ as } n \rightarrow \infty$$

- También se puede demostrar el teorema a partir de la desigualdad de Chebyshev y de la varianza de \bar{X}_n si se asume adicionalmente que esta es finita

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\text{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\text{Var}(X_1)}{n\varepsilon^2}$$

$$\Rightarrow \lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \lim_{n \rightarrow \infty} \frac{\text{Var}(X_1)}{n\varepsilon} = 0$$

$$\Rightarrow \bar{X}_n \xrightarrow{P} \mu \text{ as } n \rightarrow \infty$$

- Siendo X_1, X_2, \dots variables aleatorias i.i.d y fijando $S_n = X_1 + X_2 + \dots + X_n$ para $n \geq 1$, entonces se cumple la siguiente convergencia casi segura si, y solo si, μ es finita:

$$\bar{X}_n = \frac{S_n}{n} \xrightarrow{a.s.} \mu \text{ as } n \rightarrow \infty$$

- Este resultado se puede demostrar con métodos diferentes y necesita la suposición de media finita, y da lugar a la llamada ley de grandes números fuerte, mientras que la anterior se denomina ley de grandes números débil
- La ley de grandes números expresa que $\bar{X}_n - \mu$ es pequeña (con una gran probabilidad) cuando n es grande, lo cual se puede interpretar como una proposición cualitativa
- Siendo X_1, X_2, \dots variables aleatorias i.i.d con esperanza μ y varianza σ^2 finita, y fijando $S_n = X_1 + X_2 + \dots + X_n$ para $n \geq 1$, entonces se cumple la siguiente convergencia en distribución:

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{D} N(0,1)$$

- Este teorema se conoce como el teorema del límite central o *central limit theorem* (CLT), y también se puede expresar de manera equivalente a través de \bar{X}_n y de la función de densidad normal:

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$$

$$\lim_{n \rightarrow \infty} F_{Z_n}(x) = F_N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$$

- Con tal de demostrar este teorema, se usan las funciones características. A partir de una expansión de Taylor para la función característica de X_1 , se puede obtener el siguiente resultado:

$$\phi_{Z_n}(t) = e^{-\left(\frac{i\mu\sqrt{n}}{\sigma}\right)t} E\left(e^{S_n\left(\frac{t}{\sigma\sqrt{n}}\right)}\right) = e^{-\left(\frac{i\mu\sqrt{n}}{\sigma}\right)t} \phi_{S_n}\left(\frac{t}{\sigma\sqrt{n}}\right) =$$

$$\begin{aligned}
&= e^{-\left(\frac{i\mu\sqrt{n}}{\sigma}\right)t} \left[\phi_{X_1} \left(\frac{t}{\sigma\sqrt{n}} \right) \right]^n = \\
&= e^{-\left(\frac{i\mu\sqrt{n}}{\sigma}\right)t} \left[1 + \frac{i\mu}{\sigma\sqrt{n}}t - \frac{t}{2\sigma^2n} E(X_1^2) + O\left(\frac{t^2}{n}\right) \right]^n = \\
&= e^{-\left(\frac{i\mu\sqrt{n}}{\sigma}\right)t} \left[1 + \frac{i\mu}{\sigma\sqrt{n}}t - \frac{t(\sigma^2 + \mu)}{2\sigma^2n} + O\left(\frac{t^2}{n}\right) \right]^n
\end{aligned}$$

- Explotando la relación $x = e^{\ln(x)}$ y la expansión de Taylor de la función $\ln(1+x)$ (válida para todos los números complejos z tales que $|z| < 1$, la última expresión se convierte en la siguiente:

$$\begin{aligned}
&e^{-\left(\frac{i\mu\sqrt{n}}{\sigma}\right)t} e^{n \ln \left[1 + \frac{i\mu}{\sigma\sqrt{n}}t - \frac{t(\sigma^2 + \mu)}{2\sigma^2n} + O\left(\frac{t^2}{n}\right) \right]} = \\
&= e^{-\left(\frac{i\mu\sqrt{n}}{\sigma}\right)t + n \left[\frac{i\mu}{\sigma\sqrt{n}}t - \frac{(\sigma^2 + \mu)}{2\sigma^2n} + O\left(\frac{t^2}{n}\right) - \frac{1}{2} \left(\frac{i\mu}{\sigma\sqrt{n}}t - \frac{(\sigma^2 + \mu)}{2\sigma^2n} + O\left(\frac{t^2}{n}\right) \right)^2 + O\left(\frac{t^2}{n}\right) \right]} \\
&= e^{-\left(\frac{i\mu\sqrt{n}}{\sigma}\right)t + n \left[\frac{i\mu}{\sigma\sqrt{n}}t - \frac{(\sigma^2 + \mu)}{2\sigma^2n} + \frac{1t^2\mu^2}{2\sigma^2n} + O\left(\frac{t^2}{n}\right) \right]} = \\
&= e^{-\left(\frac{i\mu\sqrt{n}}{\sigma}\right)t + n \left[\frac{i\mu}{\sigma\sqrt{n}}t - \frac{(\sigma^2 + \mu)}{2\sigma^2n} + \frac{1t^2\mu^2}{2\sigma^2n} + O\left(\frac{t^2}{n}\right) \right]} = e^{-\frac{t^2}{2} + nO\left(\frac{t^2}{n}\right)}
\end{aligned}$$

- Por lo tanto, tomando el límite cuando $n \rightarrow \infty$, se puede ver como la función característica de Z_n tiende a la de una variable aleatoria normal estándar para toda $t \in \mathbb{R}$

$$\begin{aligned}
\phi_{Z_n}(t) &= e^{-\frac{t^2}{2} + nO\left(\frac{t^2}{n}\right)} \rightarrow e^{-\frac{t^2}{2}} \quad \text{as } n \rightarrow \infty \\
&\Rightarrow Z_n \xrightarrow{D} N(0,1)
\end{aligned}$$

- Escalando la variable Z_n por σ , entonces se obtiene una forma equivalente que resulta útil para diferentes aplicaciones y operaciones

$$\sigma Z_n = \frac{1}{\sqrt{n}}(S_n - n\mu) = \sqrt{n}(\bar{X}_n - \mu) \Rightarrow \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$$

- En todo caso, siempre se tiene que tener en cuenta que μ y σ^2 pertenecen a la distribución de X_1 y no a la de \bar{X}_n

- Con tal de demostrar este teorema, se usan las funciones características. A partir de una expansión de Taylor para la función característica de X_1 , se puede obtener el siguiente resultado:

$$\begin{aligned}
 \phi_{\sigma Z_n}(t) &= e^{-it\mu\sqrt{n}} E\left(e^{S_n\left(\frac{t}{\sqrt{n}}\right)}\right) = e^{-i\mu\sqrt{n}t} \phi_{S_n}\left(\frac{t}{\sqrt{n}}\right) = \\
 &= e^{-it\mu\sqrt{n}} \left[\phi_{X_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n = \\
 &= e^{-it\mu\sqrt{n}} \left[1 + \frac{i\mu}{\sqrt{n}}t - \frac{t}{2n}E(X_1^2) + O\left(\frac{t^2}{n}\right)\right]^n = \\
 &= e^{-it\mu\sqrt{n}} \left[1 + \frac{i\mu}{\sqrt{n}}t - \frac{t(\sigma^2 + \mu)}{2n} + O\left(\frac{t^2}{n}\right)\right]^n
 \end{aligned}$$

- Explotando la relación $x = e^{\ln(x)}$ y la expansión de Taylor de la función $\ln(1+x)$ (válida para todos los números complejos z tales que $|z| < 1$, la última expresión se convierte en la siguiente:

$$\begin{aligned}
 &e^{-it\mu\sqrt{n}} e^{n \ln\left[1 + \frac{i\mu}{\sqrt{n}}t - \frac{t(\sigma^2 + \mu)}{2n} + O\left(\frac{t^2}{n}\right)\right]} = \\
 &= e^{-it\mu\sqrt{n} + n\left[\frac{i\mu}{\sqrt{n}}t - \frac{t(\sigma^2 + \mu)}{2n} + O\left(\frac{t^2}{n}\right) - \frac{1}{2}\left(\frac{i\mu}{\sqrt{n}}t - \frac{t(\sigma^2 + \mu)}{2n} + O\left(\frac{t^2}{n}\right)\right)^2 + O\left(\frac{t^2}{n}\right)\right]} \\
 &= e^{-it\mu\sqrt{n} + n\left[\frac{i\mu}{\sqrt{n}}t - \frac{t(\sigma^2 + \mu)}{2n} - \frac{1}{2}\frac{t^2\sigma^2 - 2t^2\sigma\mu + t^2\mu^2}{n} + O\left(\frac{t^2}{n}\right)\right]} \\
 &= e^{-\frac{t^2}{2}\sigma^2 + nO\left(\frac{t^2}{n}\right)}
 \end{aligned}$$

- Por lo tanto, tomando el límite cuando $n \rightarrow \infty$, se puede ver como la función característica de Z_n tiende a la de una variable aleatoria normal estándar

$$\begin{aligned}
 \phi_{\sigma Z_n}(t) &= e^{-\frac{t^2}{2}\sigma^2 + nO\left(\frac{t^2}{n}\right)} \rightarrow e^{-\frac{t^2}{2}\sigma^2} \text{ as } n \rightarrow \infty \\
 &\Rightarrow \sigma Z_n \xrightarrow{D} N(0, \sigma^2)
 \end{aligned}$$

- Siendo X_1, X_2, \dots variables aleatorias i.i.d con media nula, varianza unitaria y función característica común $\phi_X(t)$, si $\int_{-\infty}^{\infty} |\phi_X(t)|^r dt < \infty$ para algún entero $r \geq 1$, entonces la función de densidad $f_n(u)$ de

$U_n = S_n/\sqrt{n}$ existe para toda $n \geq r$ y esta tiende a una función de densidad normal, en donde la convergencia es uniforme en \mathbb{R}

$$U_n = \frac{S_n}{\sqrt{n}} = \frac{X_1 + X_2 + \dots}{\sqrt{n}} \quad \& \quad \exists r \geq 1, \int_{-\infty}^{\infty} |\phi_X(t)|^r dt < \infty$$

$$\Rightarrow \begin{cases} \exists f_n(u) & \text{for } \forall n \geq r \\ f_n(u) \rightarrow \frac{1}{\sqrt{2\pi}} e^{-u^2/2} & \text{as } n \rightarrow \infty \end{cases}$$

- Esta es una versión local del teorema del límite central, la cual sirve para...

Commented [MOU22]: Para qué sirve la versión local

- El teorema del límite central expresa que $Z_n \xrightarrow{D} N(0,1)$, lo cual significa que $Z_n \sim N(0,1)$ y que $\sigma Z_n \sim N(0, \sigma^2)$ cuando n es grande, lo cual se puede interpretar como una proposición cuantitativa

- De este modo, el teorema del límite central proporciona información sobre la tasa de convergencia en la ley de grandes números, ya que proporciona un resultado numérico, a través de la distribución normal, sobre $P(|\bar{X} - \mu| > \varepsilon)$
- Si se supone que $E(|X_1|^3) < \infty$, entonces se puede calcular una estimación de la constante C , que permite analizar que tan rápido ocurre esta convergencia en distribución

$$\sup_x |F_{Z_n}(x) - \Phi(x)| \leq C \frac{E(|X_1|^3)}{\sigma^3 \sqrt{n}}$$

- Una aplicación importante del teorema del límite central es el teorema de Moivre-Laplace, el cual muestra que si $S_n \sim \text{Bin}(n, p)$, entonces se cumple la siguiente convergencia:

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{D} N(0,1) \quad \text{as } n \rightarrow \infty$$

- Para que $S_n \sim \text{Bin}(n, p)$ las X_1, X_2, \dots que conforman la suma tienen que seguir una distribución de Bernoulli y ser i.i.d
- Otra aplicación importante del CLT es sobre la convergencia de la función de distribución empírica. Siendo X_1, X_2, \dots, X_n es una muestra de la variable aleatoria X , la función de distribución de X es F y F_n es la función de distribución empírica de la muestra, definida de la siguiente manera:

$$F_n(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n} \quad \text{where } I(X \leq x) = \begin{cases} 1 & \text{with } p = F(x) \\ 0 & \text{with } q = 1 - F(x) \end{cases}$$

- Debido a que $\sum_{i=1}^n I(X_i \leq x) \sim \text{Bin}(n, F(x))$, se puede obtener la siguiente convergencia en probabilidad:

$$P(|F_n(x) - F(x)| \leq \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2} = \frac{F(x)(1 - F(x))}{n\varepsilon^2}$$

$$\Rightarrow P(|F_n(x) - F(x)| \leq \varepsilon) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

$$\Rightarrow F_n(x) \xrightarrow{P} F(x) \quad \text{as } n \rightarrow \infty$$

- A partir del teorema del límite central, es posible mostrar como $\sqrt{n}(F_n(x) - F(x)) \xrightarrow{D} N(0, \sigma^2(x))$ ya que $E(F_n(x)) = F(x) < \infty$ y $\text{Var}(F_n(x)) = F(x)(1 - F(x)) < \infty$:

$$Z_n = \frac{F_n(x) - F(x)}{\sqrt{\frac{F(x)(1 - F(x))}{n}}} = \frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1 - F(x))}}$$

$$\Rightarrow \sigma Z_n = \sqrt{n}(F_n(x) - F(x))$$

$$\Rightarrow \sqrt{n}(F_n(x) - F(x)) \xrightarrow{D} N(0, \sigma^2(x))$$

- Utilizando la ley de grandes números fuerte, uno puede demostrar que $F_n(x) \xrightarrow{a.s.} F(x)$ cuando $n \rightarrow \infty$. Otro teorema más fuerte es el teorema de Glivenko-Cantelli, que muestra el siguiente resultado:

$$\sup_x |F_n(x) - F(x)| \xrightarrow{a.s.} 0$$

- La función de distribución empírica es una herramienta útil para estimar la verdadera distribución desconocida

- De manera más precisa, la convergencia en probabilidad mostrada indica que la distribución empírica en un punto x está cerca del valor verdadero de $F(x)$ para muestras grandes, y la segunda parte da una estimación de la desviación del valor verdadero (la varianza de la distribución normal)
- Otro uso de esta distribución es para contrastar las hipótesis de que la muestra o la serie de observaciones se ha obtenido de

una distribución especificada (el contraste de Kolmogorov usa esta idea)

- Chap 7: non iid case CLT y LLN

Commented [MOU23]: Acabar bien

- LA GCLT (Lévy-Mandelbrot)

Commented [IC24]: Acabar bien con apuntes y cosas del TFM

- Siendo X_1, X_2, \dots y Y_1, Y_2, \dots secuencias de variables aleatorias, se puede determinar el alcance en el que se puede concluir que $X_n + Y_n \rightarrow X + Y$ cuando $n \rightarrow \infty$

- Siendo X_1, X_2, \dots y Y_1, Y_2, \dots secuencias de variables aleatorias tales que $X_n \xrightarrow{a.s.} X$ y $Y_n \xrightarrow{a.s.} Y$ cuando $n \rightarrow \infty$, entonces $X_n + Y_n \xrightarrow{a.s.} X + Y$

$$X_n \xrightarrow{a.s.} X \text{ \& } Y_n \xrightarrow{a.s.} Y \text{ as } n \rightarrow \infty \Rightarrow X_n + Y_n \xrightarrow{a.s.} X + Y$$

- Introduciendo los conjuntos \bar{R}_X y \bar{R}_Y anteriormente vistos y escogiendo $\omega \notin (\bar{R}_X \cup \bar{R}_Y)$, se puede obtener el siguiente resultado:

$$\begin{aligned} & |X_n(\omega) - X(\omega) + Y_n(\omega) - Y(\omega)| \\ & \leq |X_n(\omega) - X(\omega)| + |Y_n(\omega) - Y(\omega)| \end{aligned}$$

$$\Rightarrow |X_n(\omega) - X(\omega)| + |Y_n(\omega) - Y(\omega)| \rightarrow 0 + 0 \text{ as } n \rightarrow \infty$$

$$\Rightarrow |X_n(\omega) - X(\omega) + Y_n(\omega) - Y(\omega)| \leq 0 \text{ as } n \rightarrow \infty$$

$$\Rightarrow |X_n(\omega) + Y_n(\omega) - (X(\omega) + Y(\omega))| \leq 0 \text{ as } n \rightarrow \infty$$

$$\Rightarrow X_n + Y_n \xrightarrow{a.s.} X + Y$$

- Siendo X_1, X_2, \dots y Y_1, Y_2, \dots secuencias de variables aleatorias tales que $X_n \xrightarrow{P} X$ y $Y_n \xrightarrow{P} Y$ cuando $n \rightarrow \infty$, entonces $X_n + Y_n \xrightarrow{P} X + Y$

$$X_n \xrightarrow{P} X \text{ \& } Y_n \xrightarrow{P} Y \text{ as } n \rightarrow \infty \Rightarrow X_n + Y_n \xrightarrow{P} X + Y$$

- La demostración de este teorema se puede hacer directamente con la del teorema anterior, dado que la convergencia casi segura implica la convergencia en probabilidad

- Siendo X_1, X_2, \dots y Y_1, Y_2, \dots secuencias de variables aleatorias tales que $X_n \xrightarrow{r} X$ y $Y_n \xrightarrow{r} Y$ cuando $n \rightarrow \infty$, entonces $X_n + Y_n \xrightarrow{r} X + Y$

$$X_n \xrightarrow{r} X \text{ \& } Y_n \xrightarrow{r} Y \text{ as } n \rightarrow \infty \Rightarrow X_n + Y_n \xrightarrow{r} X + Y$$

- La demostración de este teorema es parecida al caso para la convergencia casi segura, ya que, utilizando la desigualdad vista anteriormente, es posible obtener el siguiente resultado:

$$E(|X_n - X + Y_n - Y|^r) \leq 2^r [E(|X_n - X|^r) + E(|Y_n - Y|^r)]$$

$$\Rightarrow E(|X_n - X|^r) + E(|Y_n - Y|^r) \rightarrow 0 + 0 \text{ as } n \rightarrow \infty$$

$$\Rightarrow E(|X_n - X + Y_n - Y|^r) \leq 0 \text{ as } n \rightarrow \infty$$

$$\Rightarrow E(|X_n - X + Y_n - Y|^r) \leq 0 \text{ as } n \rightarrow \infty$$

$$\Rightarrow X_n + Y_n \xrightarrow{r} X + Y$$

- Siendo X_1, X_2, \dots y Y_1, Y_2, \dots secuencias de variables aleatorias tales que $X_n \xrightarrow{D} X$ y $Y_n \xrightarrow{D} Y$ cuando $n \rightarrow \infty$, y suponiendo que X_n e Y_n y X e Y son independientes, entonces $X_n + Y_n \xrightarrow{D} X + Y$

$$X_n \xrightarrow{D} X \text{ \& } Y_n \xrightarrow{D} Y \text{ as } n \rightarrow \infty \Rightarrow X_n + Y_n \xrightarrow{D} X + Y$$

- Para este resultado es necesario asumir la independencia, dado que, de otro modo, esta convergencia no ocurriría (se pueden encontrar varios contraejemplos)
- La demostración de este teorema se basa en la convergencia de la función característica:

$$\phi_{X_n+Y_n}(t) = \phi_{X_n}(t)\phi_{Y_n}(t) \text{ as } n \rightarrow \infty \text{ for } -\infty < t < \infty$$

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) \text{ as } n \rightarrow \infty \text{ for } -\infty < t < \infty$$

$$\text{given independence \& } X_n \xrightarrow{D} X \text{ and } Y_n \xrightarrow{D} Y:$$

$$\Rightarrow \phi_{X_n}(t)\phi_{Y_n}(t) \rightarrow \phi_X(t)\phi_Y(t) \text{ as } n \rightarrow \infty \text{ for } -\infty < t < \infty$$

$$\Rightarrow \phi_{X_n+Y_n}(t) \rightarrow \phi_{X+Y}(t) \text{ as } n \rightarrow \infty \text{ for } -\infty < t < \infty$$

- Chap 7: Sums of dependent random variables

- Distribuciones estables

Commented [MOU25]: Acabar bien

- Para poder operar con los diferentes conceptos de convergencia cuando se opera con variables aleatorias, existen varios teoremas y resultados que permiten simplificar los cálculos:

- Siendo X_1, X_2, \dots y Y_1, Y_2, \dots secuencias de variables aleatorias tales que $X_n \xrightarrow{D} X$ y $Y_n \xrightarrow{P} a$ cuando $n \rightarrow \infty$, entonces se dan las siguientes convergencias en distribución cuando $n \rightarrow \infty$:

$$X_n \pm Y_n \xrightarrow{D} X \pm a$$

$$X_n Y_n \xrightarrow{D} Xa \quad \frac{X_n}{Y_n} \xrightarrow{D} \frac{X}{a} \text{ for } a \neq 0$$

- Este teorema se conoce como el teorema de Crámer o el teorema de Slutsky, y permite operar de manera sencilla con la convergencia en distribución
- Siendo X_1, X_2, \dots una secuencia de variables aleatorias tal que $X_n \xrightarrow{P} a$ cuando $n \rightarrow \infty$ y suponiendo que g es una función continua en a , entonces se da la siguiente convergencia en probabilidad:

$$g(X_n) \xrightarrow{P} g(a) \text{ as } n \rightarrow \infty$$

- La demostración de este teorema se basa en la suposición de que $X_n \xrightarrow{P} a$, de modo que $P(|X_n - a| > \delta) \rightarrow 0$ cuando $n \rightarrow \infty$ para cualquier $\delta > 0$. La continuidad de g implica lo siguiente:

$$\forall \varepsilon > 0, \exists \delta > 0 \text{ such that } |x - a| < \delta \Rightarrow |g(x) - g(a)| < \varepsilon$$

or

$$\forall \varepsilon > 0, \exists \delta > 0 \text{ such that } |g(x) - g(a)| > \varepsilon \Rightarrow |x - a| > \delta$$

- Debido a que $\{\omega: |g(X_n(\omega)) - g(a)| > \varepsilon\}$ es un subconjunto de $\{\omega: |X_n(\omega) - a| > \delta\}$, lo cual implica lo siguiente:

$$\{\omega: |g(X_n(\omega)) - g(a)| > \varepsilon\} \subseteq \{\omega: |X_n(\omega) - a| > \delta\}$$

$$\Rightarrow P(|g(X_n) - g(a)| > \varepsilon) \leq P(|X_n - a| > \delta)$$

$$X_n \xrightarrow{P} a \Rightarrow P(|X_n - a| > \delta) \rightarrow 0 \text{ as } n \rightarrow \infty$$

$$\Rightarrow g(X_n) \xrightarrow{P} g(a) \text{ as } n \rightarrow \infty$$

- El método delta es un teorema que permite demostrar la convergencia siguiente convergencia si se cumplen las siguientes condiciones:

$$a_n(g(X_n) - g(\theta)) \xrightarrow{D} N(0, |g'(\theta)|\sigma)$$

- La secuencia $\{a_n: n \geq 1\}$ es una secuencia de números reales tales que $a_n \rightarrow \infty$ cuando $n \rightarrow \infty$ y $a_n \neq 0, \forall n$
- La secuencia $\{X_n: n \geq 1\}$ es una secuencia de variables aleatorias y θ es un número real tal que $a_n(X_n - \theta) \xrightarrow{D} N(0, \sigma)$
- La función real g tiene derivada continua en un intervalo que contiene θ y tal que $g'(\theta) \neq 0$
- Siendo X_1, X_2, \dots y X variables aleatorias tomando valores enteros no negativos, una condición necesario y suficiente para que $X_n \xrightarrow{D} X$ es la siguiente:

$$\lim_{n \rightarrow \infty} P(X_n = k) = P(X = k) \text{ for } \forall k \geq 0$$

- Este teorema permite demostrar la convergencia en distribución para variables discretas que tomen valores enteros no negativos a través de la expresión para la probabilidad
- Siendo X_1, X_2, \dots y X variables aleatorias tales que $X_n \xrightarrow{D} X$ cuando $n \rightarrow \infty$, entonces existe un espacio de probabilidad $(\Omega', \mathcal{F}', P')$ y unas variables aleatorias Y_1, Y_2, \dots y Y que mapean Ω' en \mathbb{R} tal que Y_1, Y_2, \dots y Y tienen las mismas funciones de distribución que X_1, X_2, \dots y X y se cumple que $Y_n \xrightarrow{a.s.} Y$
- Este teorema se denomina teorema de representación de Shorokhod, y sirve para poder demostrar la existencia de un conjunto de variables Y_1, Y_2, \dots y Y las cuales se distribuyen igual que X_1, X_2, \dots y X y convergen de manera casi segura $Y_n \xrightarrow{a.s.} Y$ en espacio de probabilidad alternativo
- Uno de los resultados más importantes para la convergencia casi segura son los lemas de Borel-Cantelli
 - Siendo $\{A_n: n \geq 1\}$ es una secuencia de eventos (subconjuntos de Ω), se definen los siguientes eventos:

$$A_* = \liminf_{n \rightarrow \infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m$$

$$A^* = \limsup_{n \rightarrow \infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m$$

- Es importante ver que $\bigcup_{m=n}^{\infty} A_m$ es una secuencia decreciente en n , dado que para una mayor n , se toma la unión de menos eventos A_m . Además, $\bigcap_{m=n}^{\infty} A_m$ es una secuencia creciente, debido a que los elementos en la intersección irán aumentando cada vez que se interseccione menos eventos (cuando n aumenta)
- Siendo $\{A_n: n \geq 1\}$ eventos arbitrarios, entonces se cumple la siguiente implicación:

$$\sum_{n=1}^{\infty} P(A_n) < \infty \Rightarrow P(A^*) = 0$$

- La demostración del teorema es bastante sencilla aplicando la siguiente lógica:

$$P(A^*) = P\left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m\right) \leq \lim_{n \rightarrow \infty} P\left(\bigcup_{m=n}^{\infty} A_m\right) \leq \lim_{n \rightarrow \infty} \sum_{m=n}^{\infty} P(A_m)$$

$$\sum_{n=1}^{\infty} P(A_n) < \infty \Rightarrow \lim_{n \rightarrow \infty} \sum_{m=n}^{\infty} P(A_m) = 0 \text{ (if finite)}$$

$$\Rightarrow P(A^*) \leq \lim_{n \rightarrow \infty} \sum_{m=n}^{\infty} P(A_m) = 0 \Rightarrow P(A^*) = 0$$

- El converso de este teorema no se mantiene en general, habiendo varios contraejemplos
- Entonces, si $\omega \in \Omega$ pertenece al conjunto $A_* = \liminf_{n \rightarrow \infty} A_n$, entonces ω pertenece a $\bigcap_{m=n}^{\infty} A_m$ para alguna n , lo que quiere decir que existe una n tal que $\omega \in A_m$ para toda $m \geq n$. En particular si A_n es el evento de que ocurra algo “especial” en el momento n , entonces $\liminf_{n \rightarrow \infty} A_n^c$ significa que a partir de una n hacia adelante, esta propiedad ya no ocurre (solo ocurre un número finito de veces)
- El primer lema de Borel-Cantelli es bastante útil para poder demostrar convergencia casi segura, y también se puede utilizar si se utilizan desigualdades como la de Chebyshev o la de Markov (en donde se fija que un evento $\{|X_n - X| \geq \varepsilon\}$ para

$n \rightarrow \infty$ no ocurrirá o tiene una probabilidad nula, de modo que $\sum_{n=1}^{\infty} E(|X_n - X|^r)/\varepsilon^r \rightarrow 0$ y $\sum_{n=1}^{\infty} P(\{|X_n - X| \geq \varepsilon\}) \rightarrow 0$

- Siendo $\{A_n: n \geq 1\}$ eventos independientes, entonces se cumple la siguiente convergencia:

$$\sum_{n=1}^{\infty} P(A_n) = \infty \Rightarrow P(A^*) = 1$$

- La demostración del teorema se basa en aplicar la fórmula de De Morgan y la independencia de eventos. Además, se usa la suposición de que $\sum_{n=1}^{\infty} P(A_n) = \infty$ y que para $0 < x < 1$ se tiene que $e^{-x} \geq 1 - x$

$$\begin{aligned} P(A^*) &= P\left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m\right) = 1 - P\left(\bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m^c\right) = \\ &= 1 - \lim_{n \rightarrow \infty} P\left(\bigcap_{m=n}^{\infty} A_m^c\right) = 1 - \lim_{n \rightarrow \infty} \lim_{N \rightarrow \infty} P\left(\bigcap_{m=n}^N A_m^c\right) \end{aligned}$$

by independence:

$$\begin{aligned} \lim_{n \rightarrow \infty} \lim_{N \rightarrow \infty} P\left(\bigcap_{m=n}^N A_m^c\right) &= \lim_{n \rightarrow \infty} \lim_{N \rightarrow \infty} \prod_{m=n}^N [1 - P(A_m)] \\ &\Rightarrow \lim_{N \rightarrow \infty} \prod_{m=n}^N [1 - P(A_m)] \leq \lim_{N \rightarrow \infty} e^{-\sum_{m=n}^N P(A_m)} = 0 \\ &\Rightarrow P(A^*) = 1 - \lim_{n \rightarrow \infty} \lim_{N \rightarrow \infty} \prod_{m=n}^N [1 - P(A_m)] = 1 - 0 = 1 \end{aligned}$$

- Similarmente, si $\omega \in \Omega$ pertenece a $A^* = \limsup_{n \rightarrow \infty} A_n$, entonces ω pertenece a $\bigcup_{m=n}^{\infty} A_m$ para toda n , lo que quiere decir que no importa lo grande que se escoja m , siempre habrá una $n \geq m$ tal que $\omega \in A_n$ o, equivalentemente, que $\omega \in A_n$ para valores arbitrariamente grandes de n (el evento puede ocurrir infinitas veces). Una manera conveniente de expresar esto último es a través de las siguientes implicaciones:

$$\omega \in \{A_n \text{ infinitely often (i.o.)}\} \Leftrightarrow \omega \in A^*$$

- Existen versiones más generales de estos resultados, los cuales permiten una cierta dependencia entre eventos (la independencia no es necesaria para que el converso se mantenga)
- Un corolario de este teorema es la ley 0-1 de Kolmogorov, que expresa que, para eventos independientes $\{A_n: n \geq 1\}$ y siendo A^* el límite superior de esta secuencia, entonces $P(A^*) = 0$ o $P(A^*) = 1$ dependiendo de si $\sum_{m=n}^{\infty} P(A_m)$ converge o no, respectivamente

$$\sum_{m=n}^{\infty} P(A_m) < \infty \Rightarrow P(A^*) = 0$$

$$\sum_{m=n}^{\infty} P(A_m) = \infty \Rightarrow P(A^*) = 1$$

- Por lo tanto, este corolario es muy útil para poder determinar la probabilidad de A^* para eventos A_n independientes, solo necesitando evaluar la serie $\sum_{m=n}^{\infty} P(A_m)$
- Este corolario se ha usado anteriormente para demostrar que la convergencia casi segura implica la convergencia en probabilidad
- Siendo X_1, X_2, \dots
- Chap 7: Dominios de atracción

Commented [MOU26]: Acabar bien

Commented [MOU27]: Acabar bien

Los procesos estocásticos

- Hasta ahora, se ha desarrollado términos y resultados básicos de teoría de probabilidad, pero se pueden considerar aplicaciones. El paso del tiempo juega un papel esencial, por lo que las aplicaciones de probabilidad involucran cantidades que se desarrollan aleatoriamente mientras pasa el tiempo
- Estos procesos evolutivos aleatorios se denominan procesos aleatorios o procesos estocásticos, y hay muchos tipos diferentes
 - Como la mayoría de procesos naturales son muy complejos para describirse exactamente y los modelos probabilísticos también los son
 - De este modo, se presentan diversos modelos simples para poder trabajar mejor con estos

- Los procesos principales que se consideran son los procesos de ramificación, los caminos aleatorios y los procesos de Poisson y relacionados
 - Los procesos de ramificación modelan el crecimiento de una población que se reproduce por ella misma
 - Los caminos aleatorios modelan el movimiento de un individuo que se mueve erráticamente dentro de un medio
 - Los procesos de Poisson y relacionados modelan cosas como la emisión de partículas radioactivas de una fuente con una pequeña caída o *decay* o la longitud de una cola
- Un proceso estocástico es un modelo matemático de una magnitud que evoluciona mientras el tiempo pasa, de modo que uno tiene una variable aleatoria para cualquier instante fijo en el tiempo
 - Siendo (Ω, \mathcal{F}, P) un espacio de probabilidad y siendo $R_t \subseteq \mathbb{R}$, un proceso estocástico X es un mapeado $X: R_t \times \Omega \rightarrow \mathbb{R}$ (lo cual implica que $(t, \omega) \mapsto X(t, \omega)$) tal que $X(t, \omega)$ es una variable aleatoria para toda $t \in R_t$
 - Siendo $t \in R_t$ y, como $X(t, \omega)$ es una variable aleatoria, el subconjunto de resultados $\{\omega \in \Omega: X(t, \omega) \leq x\}$ es un evento (un elemento de \mathcal{F}) para toda $x \in \mathbb{R}$
 - En muchas aplicaciones, t significa tiempo, y la dependencia sobre ω normalmente se omite, escribiendo X , $X(t)$ o X_t para denotar el proceso estocástico
 - Por lo tanto, un proceso estocástico X es una colección de variables aleatorias indexadas por t , aunque también es una colección de funciones de tiempo, en donde $X_\omega(t)$ es la realización del proceso asociado al resultado $\omega \in \Omega$

$$\{X(t, \omega) \equiv X_t(\omega), t \in R_t\} \quad \text{or} \quad \{X(t, \omega) \equiv X_\omega(t), \omega \in \Omega\}$$
 - Si R_t es un conjunto discreto (finito o infinitamente numerable), se dice que X es un proceso de tiempo discreto, mientras que, si el conjunto no es numerable, se dice que X es un proceso de tiempo continuo
 - Si $R_t = \{t_1, t_2, \dots, t_n\}$ es finito, entonces X , definido de la siguiente manera, es un vector aleatorio:

$$X = [X_{t_1}, X_{t_2}, \dots, X_{t_n}]$$

- Si $R_t = \{t_1, t_2, \dots, t_n, \dots\}$ es infinitamente numerable, entonces X , definido de la siguiente manera, es una secuencia de variables aleatorias:

$$\{X_t\}_{t=1}^{\infty} = \{X_{t_1}, X_{t_2}, \dots, X_{t_n}, \dots\}$$

- Los casos más usuales de procesos de tiempo continuo son los siguientes:

$$X_t, t \in [a, b]$$

$$X_t, t \in (0, \infty)$$

$$X_t, t \in \mathbb{R}$$

- Acorde a los valores que toma el proceso, se puede dividir el proceso en proceso con espacio de estado discreto o con espacio de estado continuo
 - El proceso X_t es un proceso con espacio de estado discreto si el conjunto de imagen del mapeado X es finito o infinitamente numerable. Por lo tanto, para cada $t \in R_t$, uno tiene una variables aleatoria discreta
 - El proceso X_t es un proceso con espacio de estado continuo si el conjunto de imagen del mapeado X es no numerable. Por lo tanto, para cada $t \in R_t$, uno tiene una variables aleatoria continua
- También se pueden definir los conceptos de función de distribución, de masa de probabilidad, de densidad, de media y de autocovarianza para procesos estocásticos
 - Uno puede considerar, para cada $t \in R_t$, la función de distribución de probabilidad de la variable aleatoria X_t es la siguiente:

$$F_{X_t}(x) \equiv F_X(x; t) = P(X_t \leq x)$$

- En este caso, $F_X(x; t)$ es la función de distribución de primer orden del proceso, y describe la evolución temporal de la ley de probabilidad de la colección de variables aleatorias $\{X_t, t \in R_t\}$
- Para un proceso estocástico de estado continuo, X_t es una variable aleatoria continua para cualquier $t \in R_t$, de modo que se puede considerar la densidad de primer orden:

$$f(x; t)\Delta x \approx P(x \leq X_t \leq x + \Delta x)$$

$$\Rightarrow f_{X_t}(x) \equiv f(x; t) = \lim_{\Delta x \rightarrow 0} \frac{P(x \leq X_t \leq x + \Delta x)}{\Delta x}$$

- La función de densidad de primer orden $f(x; t)$ es, para cada $t \in R_t$, la densidad de probabilidad $f_{X_t}(x)$ de la variable continua X_t
- La relación entre la función de densidad de primer orden y la función de distribución es la misma que la considerada anteriormente:

$$f_{X_t}(x) = \frac{\partial}{\partial x} F_{X_t}(x) \quad F_{X_t}(x) = \int_{-\infty}^x f_{X_t}(u) du$$

- De manera similar, si X_t es un proceso de espacio de estado discreto tomando valores x_1, \dots, x_k, \dots , se puede definir la función de masa de probabilidad de primer orden de la siguiente manera:

$$p_X(x_k; t) \equiv p_{X_t}(x_k) = P(X_t = x_k) \quad \text{for } k \geq 1$$

- Por lo tanto, se puede obtener la siguiente relación con la función de distribución:

$$F_X(x; t) = \sum_{x_k \leq x} p_X(x_k; t)$$

- La función de distribución de orden n $F_X(x_1, \dots, x_n; t_1, \dots, t_n)$ del proceso estocástico X_t da la función de distribución conjunta de las n variables aleatorias $X_{t_1}, X_{t_2}, \dots, X_{t_n}$ que nacen de X_t en los momentos $t_1, t_2, \dots, t_n \in R_t$

$$\begin{aligned} F_X(x_1, \dots, x_n; t_1, \dots, t_n) &\equiv F_{X_{t_1}, X_{t_2}, \dots, X_{t_n}}(x_1, \dots, x_n) = \\ &= P(X_{t_1} \leq x_1, X_{t_2} \leq x_2, \dots, X_{t_n} \leq x_n) \end{aligned}$$

- Para un proceso de espacio de estado continuo, la función de densidad de orden n es la densidad conjunta de las variables aleatorias $X_{t_1}, X_{t_2}, \dots, X_{t_n}$, definida de la siguiente manera:

$$\begin{aligned} f_X(x_1, \dots, x_n; t_1, \dots, t_n) &\equiv f_{X_{t_1}, X_{t_2}, \dots, X_{t_n}}(x_1, \dots, x_n) = \\ &= \frac{\partial F_X(x_1, \dots, x_n; t_1, \dots, t_n)}{\partial x_1 \dots \partial x_n} \end{aligned}$$

- Análogamente, si X_t es un proceso de espacio de estado discreto, la función de masa de probabilidad será la siguiente:

$$p_X(x_1, \dots, x_n; t_1, \dots, t_n) = P(X_{t_1} = x_1, X_{t_2} = x_2, \dots, X_{t_n} = x_n)$$

- Las relaciones usuales y las propiedades de estas funciones se siguen manteniendo:

$$\int_{-\infty}^{\infty} f_X(x; t) dx = 1$$

$$\int_{-\infty}^{\infty} f_X(x_1, x_2; t_1, t_2) dx_1 = f_X(x_2; t_2)$$

$$\lim_{x_2 \rightarrow \infty} F_X(x_1, x_2; t_1, t_2) = F_X(x_1; t_1)$$

$$p_X(x_{k_1}; t_1) = \sum_{k_2} p_X(x_{k_1}, x_{k_2}; t_1, t_2)$$

- La función de media de un proceso...
- La autocorrelación...
- La autocovarianza...

- Un proceso estocástico es estrictamente estacionario si...

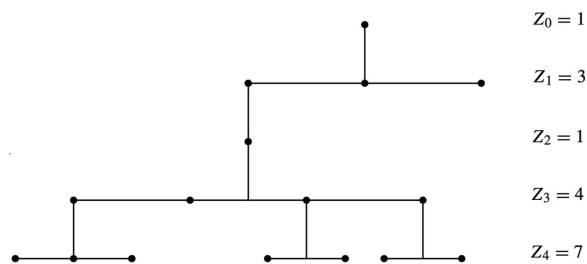
Commented [MOU28]: Acabar bien

Los procesos de ramificación

- Un modelo de ramificación simple que permite modelar el crecimiento poblacional es el proceso de Galton-Watson, el cuál es un proceso de ramificación general que permitirá desarrollar la teoría para procesos de ramificación
 - Se define el término nómada como un tipo de objeto hipotético que es capaz de reproducirse acorde a las siguientes normas:
 - En el momento $t = 0$ existe algún nómada, el cual vive para una unidad de tiempo y entonces, en el momento $n = 1$, este muere en el momento de parto y se reemplaza por la familia nómada descendiente
 - Estos nómadas tienen una biografía similar, en donde todos sobreviven hasta el momento $n = 2$ y entonces se mueren y son reemplazados por sus descendientes. Esto continúa siendo el caso para todos los descendientes hasta $n = 3, 4, \dots$
 - Si se sabe el tamaño de todas las familias de los nómadas individuales, se sabe todo sobre el desarrollo del nómada, pero el problema es que diferentes nómadas pueden tener un

diferente número de descendientes, y este número no se sabe antes de su realización

- Por lo tanto, se tienen que asumir que el tamaño de las familias son variables aleatorias que satisfacen las siguientes condiciones:
 - Los tamaños de las familias son variables aleatorias independientes que toman valores discretos $\{0, 1, 2, \dots\}$
 - Los tamaños de la familias son variables aleatorias idénticamente distribuidas con función de masa de probabilidad conocida p , de modo que el número de niños C de un nómada genérico tiene función de masa $P(C = k) = p_k$ para $k = 0, 1, 2, \dots$
- Una vez establecido el modelo, el problema básico es poder decir algo sobre como el desarrollo del proceso depende de la función de masa p del tamaño familiar



- Para poder evitar trivialidades, se supone que $p_k \neq 1$ para toda $k = 1, 2, 3, \dots$
- El conjunto de nómadas nacidos en el momento n se denomina n -ésima generación del proceso de ramificación, y la variable Z_n es el número de nómadas en esa generación
- La evolución del proceso se describe por la secuencia Z_0, Z_1, \dots de variables aleatorias, y uno trabaja en esta secuencia
- El primer paso para estudiar el proceso de ramificación es explicar como encontrar distribuciones de Z_i en términos de la función de masa del tamaño de la familia p
 - La función de masa de Z_2 es difícil de derivar directamente, dado que Z_2 es la suma de un número Z_1 aleatorio de tamaños familiares

- Para generaciones anteriores, $Z_0 = 1$ y $P(Z_1 = k) = p_k$ para $k = 0, 1, 2, \dots$, ya que Z_1 es el número de descendientes del nómada inicial ($Z_1 = C_1$)
- Esto se puede traducir como que, escribiendo C_i para el número de descendientes del nómada i en la primera generación, Z_2 es la suma de tamaños familiares de los Z_1 nómadas

$$Z_2 = C_1 + C_2 + \dots + C_{Z_1}$$

- De manera general, para $n = 1, 2, 3, \dots$ se cumple lo mismo, en donde C'_1, C'_2, \dots son los números de descendientes de los nómadas en la generación $n - 1$ (que se distribuyen igual que C_1, C_2, \dots pero no son las mismas variables al no pertenecer a los mismos individuos)

$$Z_n = C'_1 + C'_2 + \dots + C'_{Z_{n-1}}$$

- La suma de un número aleatorio de variables aleatorias se trata de manera más sencilla usando las funciones generadoras de probabilidad, por lo que se quiere expresar $G_n(s)$ en términos de $G(s)$ (la función para toda C_i , ya que están distribuidas de manera idéntica)
- En general, si $r \geq 1$, se puede expresar la probabilidad condicional de que Z_{n+1} tome un valor dado un valor Z_n de la siguiente manera, en donde las C'_i son variables independientes e idénticamente distribuidas:

$$\begin{aligned} P(Z_{n+1} = z | Z_n = r) &= P(C'_1 + C'_2 + \dots + C'_r = z | Z_n = r) = \\ &= P(C'_1 + C'_2 + \dots + C'_r = z) \end{aligned}$$

- Esta eliminación de la condición se debe a que, aunque la realización de Z_n determine cuantas variables C' pueda haber en Z_{n+1} , la realización de Z_{n+1} no depende de Z_n porque las variables C y C' son i.i.d
- Si $r = 0$, entonces se puede comprobar que se cumple $P(Z_{n+1} = 0 | Z_n = 0) = 1$
- Por lo tanto, a través del teorema de probabilidad total se puede demostrar que se cumple la siguiente identidad, ya que el evento de que la suma de variables C dé z y que $Z_n = r$ son independientes (se multiplican las probabilidades):

$$P(Z_{n+1} = z) = \sum_{r \geq 0} P(Z_{n+1} = z | Z_n = r) P(Z_n = r)$$

$$= \sum_{r \geq 0} P(C'_1 + C'_2 + \dots + C'_r = z) P(Z_n = r)$$

- Las funciones generadoras de probabilidades G_C, G_0, G_1, \dots satisfacen que $G_0(s) = s$ y $G_n = G_{n-1}(G(s))$ para $n = 1, 2, \dots$ y, entonces, G es la iteración número n de G

$$G_n = G(G(\dots G(s) \dots)) \quad \text{for } n = 0, 1, 2, \dots$$

- Teniendo $Z_0 = 1$, se puede ver como $G_0(s) = \sum_{k=1}^{\infty} s^k P(Z_0 = k) = s$, y debido a que la suma Z_n es de Z_{n-1} variables aleatorias, se puede usar el teorema anteriormente visto para convoluciones y la independencia entre variables C_i para obtener que $G_n(s) = G_{n-1}(G(s))$

$$\begin{aligned} G_n(s) &= E(s^{Z_n}) = E_{Z_n}[E(s^{Z_n} | Z_{n-1} = k)] = \\ &= E_{Z_{n-1}}[E(s^{C_1 + \dots + C_k} | Z_{n-1} = k)] = E_{Z_{n-1}}[E(s^{C_1} \dots s^{C_k} | Z_{n-1} = k)] = \\ &= E_{Z_{n-1}}[E(s^{C_1} | Z_{n-1} = k) E(s^{C_2} | Z_{n-1} = k) \dots E(s^{C_k} | Z_{n-1} = k)] = \\ &= E_{Z_{n-1}}[E(s^{C_1}) \dots E(s^{C_k})] = E_{Z_{n-1}}([E(s^C)]^k) = E_{Z_{n-1}}([G(s)]^k) \\ &= E([G(s)]^{Z_{n-1}}) = G_{n-1}[G(s)] \end{aligned}$$

- Por lo tanto, por iteración se obtiene la siguiente igualdad:

$$G_n(s) = G_{n-1}(G(s)) = G_{n-2}(G(G(s))) = \dots =$$

$$= G_1(G(G(\dots G(s) \dots))) = G(G(G(\dots G(s) \dots)))$$

$$\text{as } G_1 = G \text{ because } P(Z_1 = k) = p_k \text{ for } k = 0, 1, 2, \dots$$

- Este teorema contiene la información necesaria para poder estudiar el desarrollo del proceso, de modo que se puede encontrar la función generadora por composición encontrando solo $G(s)$ de una C genérica

$$G(s) = \sum_{k=0}^{\infty} s^k P(C = k) = P(C = 0) + sP(C = 1) + \dots + s^j P(C = j) + \dots$$

- La media de los valores Z_n es $E(Z_n) = \mu^n$ donde $E(X) = \sum_k k p_k = \mu$ es la media de la distribución del tamaño de la familia

$$E(Z_n) = \mu^n$$

- La demostración de este teorema se basa en que $G'_X(1) = E(X)$ y en la recursividad anterior:

$$\begin{aligned} E(Z_n) &= G'_n(1) = G'_{n-1}(G(1))G'(1) = G'_{n-1}(1)G'(1) = \\ &= E(Z_{n-1})\mu \end{aligned}$$

$$\Rightarrow E(Z_n) = E(Z_{n-1})\mu = E(Z_{n-2})\mu^2 = \dots = E(Z_0)\mu^n = \mu^n$$

- La varianza de Z_n puede derivarse de modo similar en términos de μ y de la varianza σ^2 de la distribución del tamaño de familia
- A partir de este corolario, se puede ver como la esperanza tiende a tres valores distintos dependiendo sustancialmente de μ

$$E(Z_n) \rightarrow \begin{cases} 0 & \text{if } \mu < 1 \\ 1 & \text{if } \mu = 1 \\ \infty & \text{if } \mu > 1 \end{cases}$$

- En las siguientes secciones se demostrará como, si $\mu \leq 1$, la población nómada se extinguirá, mientras que si $\mu > 1$, hay una probabilidad estrictamente positiva que la línea de descenso de los nómadas continúe para siempre
- Esta dependencia en la media del tamaño familiar μ es natural porque $\mu < 1$ significa que cada nómada tiene (de media) estrictamente menos descendientes que los necesarios para llenar el vacío por su muerte, mientras que si $\mu > 1$ significaría que cada muerte produciría más nómadas de media
- El caso en que $\mu = 1$ se denomina crítico porque la media del tamaño poblacional es 1 todo el tiempo. En este caso, las fluctuaciones aleatorias aseguran que el tamaño de la población tomará el valor 0 en algún momento, y por tanto los nómadas se extinguirán
- Tal y como se ha mencionado, la última extinción del proceso de ramificación es cierta (con probabilidad 1) si, y solo si, la media del tamaño de las familias es μ satisface $\mu \leq 1$, la cual es una propiedad cierta para cualquier proceso de ramificación
 - La probabilidad de extinción e se define como la probabilidad que el tamaño de la familia de nómadas en la generación $n \geq 0$ sea igual a cero:

Commented [MOU29]: Seguir poneindo todo bien por aqui

$$e = P(Z_n = 0 \text{ for } \forall n \geq 0)$$

- Siendo $E_n = \{Z_n = 0\}$ el evento de que el proceso de ramificación se extingue en la generación n se puede deducir lo siguiente: si $Z_n = 0$, entonces necesariamente $Z_{n+1} = 0$ (pero no es una condición suficiente), de modo que $E_n \subseteq E_{n+1}$ y en particular $e_n \leq e_{n+1}$

$$\emptyset = E_0 \subseteq E_1 \subseteq \dots \subseteq E_{n-1} \subseteq E_n$$

$$0 = e_0 \leq e_1 \leq \dots \leq e_{n-1} \leq e_n$$

- En el caso en el que $n = 0$, entonces $Z_0 = 1$ y $E_0 = \emptyset$ porque no se puede extinguir (debido al supuesto), de modo que $e_0 = 0$
- Tiene sentido, por tanto, considerar el evento límite que corresponde a la última extinción de la población, definida de la siguiente manera:

$$\{Z_n = 0 \text{ for } \forall n \geq 0\} \equiv \lim_{n \rightarrow \infty} E_n \equiv \bigcup_{n=0}^{\infty} E_n$$

- Siendo $e_n = P(E_n)$, se puede ver que la probabilidad de última extinción e se puede expresar de la siguiente manera:

$$e = P\left(\lim_{n \rightarrow \infty} E_n\right) = \lim_{n \rightarrow \infty} e_n$$

- Este teorema se puede demostrar de la siguiente manera, en donde se usa la resta entre eventos (una resta de conjuntos) y el hecho de que $\{E_n - E_{n-1}\} \cap \{E_{n+j} - E_{n+j-1}\} = \emptyset$ por la independencia de las variables C_i (todos los elementos comunes que puedan ocurrir se eliminan en esta resta de conjuntos):

$$e = P\left(\lim_{n \rightarrow \infty} E_n\right) = P\left(\bigcup_{n=0}^{\infty} E_n\right) = P\left[\bigcup_{n=1}^{\infty} \{E_n - E_{n-1}\}\right]$$

$$\Rightarrow e = P\left(\lim_{n \rightarrow \infty} E_n\right) = \sum_{n=1}^{\infty} [P(E_n) - P(E_{n-1})] =$$

$$= \lim_{n \rightarrow \infty} \sum_{i=1}^n [P(E_i) - P(E_{i-1})] = \lim_{n \rightarrow \infty} \sum_{i=1}^n [P(E_i) - P(E_{i-1})] =$$

$$= \lim_{n \rightarrow \infty} [P(E_n) - P(E_0)] = \lim_{n \rightarrow \infty} P(E_n) = \lim_{n \rightarrow \infty} e_n$$

- Cuando $p_0 = 0$, un nómada siempre tendrá descendientes, de modo que $e = 0$, y esta lógica se puede generalizar para poder calcular la probabilidad de e con el siguiente teorema
- La probabilidad e de última extinción es la raíz no negativa más pequeña de la ecuación $s = G(s)$ (siendo la e la solución de la ecuación), donde G es la función generadora de probabilidad de C (la variable genérica del número de descendientes)

$$s = G(s) \Rightarrow s = P(C = 0) + sP(C = 1) + \dots + s^j P(C = j) + \dots$$

\Rightarrow *smallest root is e*

- Debido a que $e_n = P(Z_n = 0)$, se tiene que $e_n = G_n(0)$ y, de resultados anteriores, se puede demostrar que se da la siguiente igualdad:

$$G_n(s) = G_{n-1}(G(s)) = G(G(\dots G(s) \dots)) = G(G_{n-1}(s))$$

- Fijando $s = 0$, se puede ver que $e_n = G_n(0)$ satisface la recursividad $e_n = G_n(e_{n-1})$ para $n = 1, 2, \dots$ con la condición inicial de que $e_0 = 0$. Esta condición permite obtener cual es la probabilidad de extinción en una generación n a través de comenzar con $e_0 = 0$ y seguir hasta e_{n-1}

$$\begin{aligned} e_n = P(E_n) &= \sum_{k \geq 0} P(E_n | Z_1 = k) P(Z_1 = k) = \\ &= \sum_{k \geq 0} (e_{n-1})^k P(C = k) = G(e_{n-1}) \end{aligned}$$

- Tomando el límite cuando $n \rightarrow \infty$, se puede ver como $e_n \rightarrow e$ y, como G es una serie de potencias con radio de convergencia $|s| \leq 1$, es continua en $[0, 1]$ y e sería raíz de la ecuación $e = G(e)$ en este límite

$$e_n = G(e_{n-1}) \Rightarrow s = G(s) \Rightarrow e = G(e)$$

- Suponiendo que η es cualquier raíz no negativa de $s = G(s)$ y viendo que G es una secuencia no decreciente de $[0, 1]$ (al no tener coeficientes negativos), se puede comprobar que, por argumentos de inducción, e es la raíz más pequeña de la ecuación anterior:

$$e_1 = G(e_0) = G(0) \leq G(\eta) = \eta$$

$$e_2 = G(e_1) \leq G(\eta) = \eta$$

...

$$e_n = G(e_{n-1}) \leq \eta \text{ for } n = 1, 2, \dots$$

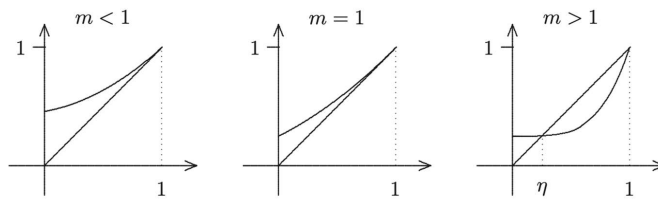
$$\Rightarrow e = \lim_{n \rightarrow \infty} e_n = \lim_{n \rightarrow \infty} G(e_{n-1}) \leq \eta$$

- Se puede ver como, cuando $s = 1$, la raíz de $s = G(s)$ será $1 = G(1)$

$$s = 1 \Rightarrow G(1) = \sum_{k \geq 0} 1^k P(C = k) = \sum_{k \geq 0} P(C = k) = 1$$

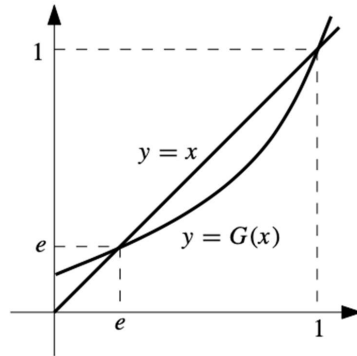
- Asumiendo que $p_1 \neq 1$, la probabilidad e de última extinción satisface $e = 1$ si, y solo si, la media μ del tamaño de la familia Z satisface $\mu \leq 1$

$$\begin{cases} e = 1 & \text{if } \mu < 1 \\ e = 1 & \text{if } \mu = 1 \\ e < 1 & \text{if } \mu > 1 \end{cases}$$

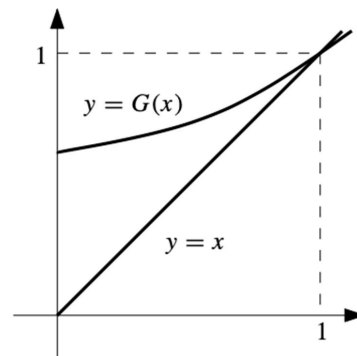


- Se elimina el caso especial de $p_1 = 1$ porque en este caso $Z_n = 1$ para toda n y eso hace que $\mu = 1$ y que $e = 0$. No obstante, se supone que $p_0 > 0$ porque si no $e = 0$ y $\mu > 1$
- En el intervalo $[0, 1]$, G es continua, no decreciente y convexa (la primera y la segunda derivada son no negativas), por lo que se puede ver generalmente hay al menos una o dos intersecciones entre las curvas $y = G(s)$ y la línea $y = s$ en este intervalo
- Si la derivada $G'(1)$ satisface que $G'(1) > 1$, se puede ver que habría dos intersecciones distintas (de las cuales una está en $s = 1$) y que $e < 1$, ya que e es la raíz más pequeña de la ecuación $s = G(s)$. Debido a que $G'(1) = E(C) = \mu$, entonces la condición es equivalente a que $\mu > 1$

Commented [MOU30]: Seguir poniend puntos apuntes en la misma distancia



- En el caso contrario, si $G'(1) \leq 1$, se puede ver que habría una única intersección en el intervalo $[0,1]$, la cual es $s = 1$ es la única (dado que la otra raíz cuando $G'(1) < 1$ está fuera del intervalo), por lo que, por definición de e como raíz más pequeña, hace que $e = 1$. Debido a que $G'(1) = E(C) = \mu$, entonces la condición es equivalente a que $\mu \leq 1$



- En consecuencia, las iteraciones $e_n = G(e_{n-1})$ convergerán a la menor de las dos raíces (si se comienza desde la condición inicial $G(0) = p_0$). Además, $G'(1) = \partial G(s)/\partial s = \mu$ es la pendiente de la línea tangente de $G(s)$ en el punto $s = 1$
- Si se considera T_n como el número total de nómaditas hasta la generación n , entonces se pueden obtener resultados útiles condicionando para $Z_1 = r$ (los individuos en la primera generación)

$$T_n = Z_0 + Z_1 + \dots + Z_n = 1 + Z_1 + \dots + Z_n$$

- En este caso, T_n se puede expresar en términos de T_1, T_2, \dots, T_{n-1} debido a que, para el número descendientes en $n = 1$, se puede coger

cada uno como el nómada inicial y, entonces, el número total de nómadas se contará para cada descendiente hasta la generación $n - 1$

$$T_n = 1 + \sum_{k=1}^{Z_1} T_{n-1}(k)$$

- En este caso $T_{n-1}(k)$ son variables i.i.d debido a que los descendientes de cada individuo en $n \geq 1$ (y los siguientes) no dependen de los de la otra ramificación
- Se puede encontrar la función generadora de probabilidad para T_n a través de esta relación:

$$G_{T_n}(s) = sG_X(G_{T_{n-1}}(s))$$

- Suponiendo que $\mu \leq 1$, entonces $e = 1$ y existe una variable aleatoria T que describe el total de la población, en donde T se define de la siguiente manera:

$$T = \lim_{n \rightarrow \infty} T_n$$

- De manera más precisa, la familia de variables aleatorias T_n converge a T cuando $n \rightarrow \infty$, que, en particular, implica que las funciones generadoras convergen. Haciendo que $n \rightarrow \infty$, se obtiene la siguiente igualdad:

$$G_T(s) = sG_X(G_T(s))$$

- Los momentos para T se pueden obtener a partir de las primeras derivadas de la función generadora de probabilidad para T

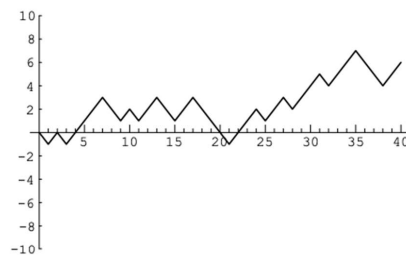
Los caminos aleatorios

- Un camino aleatorio es una formalización matemática de la trayectoria que resulta de hacer sucesivos pasos aleatorios, lo cual se puede encontrar en muchos sistemas complejos y otros contextos
 - Siendo $\{X_k\}_{k=1}^{\infty}$ una secuencia de variables aleatorias i.i.d, se define S_n como la suma $X_1 + X_2 + \dots + X_n$ para cada número entero positivo n , de modo que la secuencia $\{S_n\}_{n=1}^{\infty}$ se conoce como camino aleatorio

$$S_n = \begin{cases} 0 & \text{if } n = 0 \\ X_1 + X_2 + \dots + X_n & \text{if } n \geq 1 \end{cases}$$

Commented [IC31]: Acabar con libro Introduction to ... RW (es más formal que esto y se explica mucho mejor)

- Por conveniencia, se asume que S_n comienza siempre en el estado nulo 0. Como se asume esto, no hace falta condicionar la probabilidad todo el rato a $X_0 = 0$ (además, se verá que, al cumplirse un proceso de Markov, esto no siempre es necesario)
 - Si el rango de las X_k es \mathbb{R}^d , entonces se dice que $\{S_n\}_{n=1}^{\infty}$ es un camino aleatorio en \mathbb{R}^d
 - De este modo, el camino aleatorio definido es un proceso estocástico de tiempo discreto. El espacio de estado de este proceso definido es \mathbb{Z}
- Una realización de este proceso se puede interpretar como el movimiento de una partícula que se mueve aleatoriamente en una dimensión, cuando comienza desde el origen



- En este caso, el eje vertical representa la línea de los enteros \mathbb{Z} y el eje horizontal denota el tiempo n
 - Siendo $m > 0$ y x enteros, un camino (S_1, S_2, \dots, S_m) desde el punto de origen $(0,0)$ al punto (m, x) es una línea poligonal cuyos vértices tienen abscisas $0, 1, \dots, m$ y ordenadas S_0, S_1, \dots, S_m satisfaciendo las siguientes identidades con $s_m = x$
- $$S_0 = 0 \quad S_k - S_{k-1} = \pm 1 \quad \text{for } k = 1, 2, \dots, n$$
- Si en un momento n la partícula está en la coordenada k , entonces en el momento $n + 1$ estará en la coordenada $k + 1$ con probabilidad p o en la coordenada $k - 1$ con probabilidad q

$$P(S_{n+1} = k + 1 | S_n = k) = P(X_1 = 1 | X_0 = 0) = p$$

$$P(S_{n+1} = k - 1 | S_n = k) = P(X_1 = -1 | X_0 = 0) = q$$

- Si $p = q = 1/2$, se dice que el proceso es un camino aleatorio simétrico, mientras que si $p \neq q$, se dice que el proceso es un camino aleatorio asimétrico

- Como se asume que S_n comienza siempre en el estado nulo 0, si $m > n$, entonces se cumple la siguiente propiedad:

$$P(S_m = b | S_n = a) = P(S_{m-n} = b - a)$$

- La demostración de este teorema se basa en que cada término es igual a $P(\sum_{i=1}^{m-n} X_i = b - a)$, dado que S_{m-n} considera las mismas $X_n + X_{n+1} + \dots + X_m$ que consideraría un proceso que comienza en n con un valor a

$$\begin{aligned} P(S_m - S_n = s) &= P\left(\sum_{i=1}^m X_i - \sum_{i=1}^n X_i = s\right) = P\left(\sum_{i=1}^{m-n} X_i = s\right) \\ &= P(S_{m-n} = s) \end{aligned}$$

- Por lo tanto, para toda $r > 0$, se da la propiedad de homogeneidad espacial (no importa el punto inicial y el final mientras aumenten o disminuyan en la misma cantidad)

$$\begin{aligned} P(S_m = b + r | S_n = a + r) &= P(S_{m-n} = (b - r) - (a - r)) = \\ &= P(S_{m-n} = b - a) = P(S_m = b | S_n = a) \end{aligned}$$

- Además, para toda $p > 0$, se da la propiedad de homogeneidad temporal (no importa el momento inicial y el final mientras aumenten o disminuyan en la misma cantidad)

$$\begin{aligned} P(S_{m+p} = b | S_{n+p} = a) &= P(S_{m+p-(n+p)} = b - a) = \\ &= P(S_{m-n} = b - a) = P(S_m = b | S_n = a) \end{aligned}$$

- En este proceso se cumple la propiedad de Markov: condicionando en el presente, el futuro es independiente del pasado. Para $n < m < r$, se cumple la siguiente identidad:

$$P(S_r = c | S_m = b, S_n = a) = P(S_r = c | S_m = b)$$

- Como el futuro es independiente del pasado cuando se condiciona en el presente, la probabilidad condicional no dependerá de valores pasados de X , solo en el valor presente y en el futuro, lo cual permite eliminar otros valores en el condicionamiento
- Por lo tanto, para $n < m < r$, se puede obtener el siguiente resultado:

$$P(S_n = a, S_m = b, S_r = c) =$$

$$\begin{aligned}
&= P(S_r = c | S_m = b, S_n = a) P(S_m = b | S_n = a) P(S_n = a) = \\
&= P(S_r = c | S_m = b) P(S_m = b | S_n = a) P(S_n = a) = \\
&= P(S_{r-m} = c - b) P(S_{m-n} = b - a) P(S_n = a)
\end{aligned}$$

- La suma S_n representa la posición de la partícula en n , por lo que en el intervalo de tiempo $[n-1, n]$, la partícula se mueve (o salta) de la posición S_{n-1} a S_n . No obstante, se puede demostrar que estos saltos serán independientes

- En este caso, la resta $S_n - S_{n-1}$ equivale a X_n , de modo que, en un camino aleatorio, los saltos o incrementos son independientes e idénticamente distribuidos
- La propiedad de Markov y la expresión de los saltos se pueden usar para mostrar como los incrementos son independientes, considerando unos momentos $0 < n_1 < \dots < n_l$:

$$\begin{aligned}
&P(S_{n_1} = a_1, S_{n_2} = a_2, \dots, S_{n_l} = a_l) = \\
&= P(S_{n_l} = a_l | S_{n_{l-1}} = a_{l-1}, \dots) P(S_{n_{l-1}} = a_{l-1} | S_{n_{l-2}} = a_{l-2}, \dots) \dots P(S_{n_1} = a_1) \\
&= P(S_{n_l} = a_l | S_{n_{l-1}} = a_{l-1}) P(S_{n_{l-1}} = a_{l-1} | S_{n_l} = a_{l-1}) \dots P(S_{n_1} = a_1) \\
&= P(S_{n_1} = a_1) P(S_{n_2 - n_1} = a_2 - a_1) \dots P(S_{n_l - n_{l-1}} = a_l - a_{l-1})
\end{aligned}$$

$$\begin{aligned}
&\text{because, trivially } P(S_{n_1} = a_1, S_{n_2} = a_2, \dots, S_{n_l} = a_l) = \\
&= P(S_{n_1} - S_0 = a_1, S_{n_2} - S_{n_1} = a_2 - a_1, \dots, S_{n_l} - S_{n_{l-1}} = a_l - a_{l-1}) \\
&\Rightarrow P(S_{n_1} = a_1, \dots, S_{n_l} - S_{n_{l-1}} = a_l - a_{l-1}) = \\
&= P(S_{n_1} = a_1) \dots P(S_{n_l} - S_{n_{l-1}} = a_l - a_{l-1})
\end{aligned}$$

- La independencia de los saltos, no obstante, no se cumple si se consideran unos intervalos temporales superpuestos (en donde un intervalo temporal es un subconjunto de otro), dado que se está en el mismo proceso y las realizaciones no serán independientes (una depende de otra para llegar a ese punto)
- Otra interpretación que se puede dar para este proceso es el de un juego entre dos jugadores que repetidamente lanzan una moneda en la que sale cara con probabilidad p

- Siendo H_n y T_n el número de caras y cruces en las primeras n tiradas, respectivamente, se considera un proceso de camino aleatorio definido de la siguiente manera:

$$S_n = H_n - T_n$$

- Si en el momento n $S_n > 0$, entonces el primer jugador lleva ventaja, mientras que si $S_n < 0$, el segundo jugador la lleva. Si S_n fuera 0, entonces el juego está en empate
- A partir de esta interpretación, un truco sencillo para pasar de un camino aleatorio bidimensional a uno unidimensional es restar los componentes, dado que se formará un camino aleatorio unidimensional
- Esta interpretación es muy útil porque permite obtener una expresión para la probabilidad de estar en una coordenada k en el momento n
 - Uno se refiere a n como la longitud del camino, en donde hay 2^n caminos posibles de longitud n . Si r tiradas son cara y s son cruces, entonces se cumplen las siguientes igualdades:

$$n = r + s \quad k = r - s$$

- Un camino del origen a un punto arbitrario (n, k) existe solo si n y k tienen la forma anterior. En este caso, los r lugares para la tirada positiva pueden escogerse de los $n = p + q$ lugares posibles de $N_{n,k}$ maneras diferentes

$$N_{n,k} = \binom{n}{r} = \binom{r+s}{r} = \binom{r+s}{s} = \binom{n}{s}$$

- Por conveniencia se define $N_{n,k} = 0$ cuando n y k no son de la forma anterior. Con esta convención, solo existen $N_{n,k}$ caminos diferentes desde el punto de origen a (n, k)
- Se puede ver como $\{S_n = k\}$ si, y solo si, en las primeras n tiradas, el número r de caras y el número s de cruces satisface la siguiente relación trivial:

$$r - s = r - (n - r) = 2r - n = k$$

$$\Rightarrow r = \frac{n+k}{2} \text{ for } r = 0, 1, \dots, n$$

- Por lo tanto, como este proceso sigue una distribución binomial en donde el número de éxitos (el número de caras) es r , se puede obtener una expresión para la probabilidad y relacionarla

con la interpretación de la partícula. En el momento n , la partícula estará en una coordenada k con la siguiente probabilidad:

$$P(S_n = k) = \binom{n}{\frac{n+k}{2}} p^{\frac{n+k}{2}} q^{\frac{n-k}{2}}$$

where $k = -n, -n+2, \dots, n-2, n$ so that $r = 0, 1, \dots, n$

- Además, como se sabe que la distribución que sigue H_n y T_n es binomial, también es posible obtener la esperanza y la varianza para S_n y resultados en el límite

- A partir de la distribución de H_n y T_n , se pueden obtener los siguientes resultados:

$$E(S_n) = E(H_n) - E(T_n) = np - nq = n(p - q)$$

$$Var(S_n) = Var(2H_n - n) = 4Var(H_n) = 4npq$$

- Debido a que la esperanza es finita y la varianza también, el teorema del límite central implica que el siguiente estadístico tiende en distribución:

$$Z_n = \frac{S_n - n(p - q)}{\sqrt{4npq}} \xrightarrow{D} N(0,1)$$

$$\Rightarrow Z_n \sqrt{4npq} = \sqrt{n}[S_n - n(p - q)] \xrightarrow{D} N(0,1)$$

- Un tipo de eventos de interés en relación a los caminos aleatorios son los de tomar una coordenada negativa en un momento determinado, de modo que se analizan estos en profundidad
 - Siendo T un momento aleatorio en el que la partícula visita una coordenada negativa por primera vez, este momento se define de la siguiente manera:

$$T = \min\{n \geq 1: S_n < 0\}$$

- Alternativamente, T es el primer momento en el que el segundo jugador toma la ventaja
- Calculando probabilidades para todos los posibles valores de T , se puede obtener el patrón de que una T par tiene probabilidad nula

$$P(T = 1) = q$$

$$P(T = 2) = 0$$

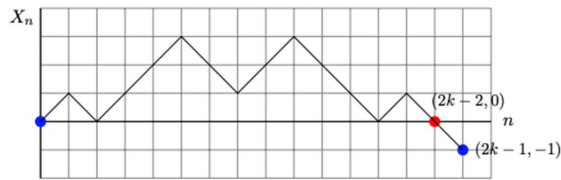
$$P(T = 3) = pq^2$$

$$P(T = 4) = 0$$

$$P(T = 5) = 2p^2q^3$$

...

- Como se puede ver, si $T = 2k$ para toda $k \geq 1$, entonces $P(T = 2k) = 0$. Esto se puede demostrar de manera gráfica, dado que como solo se puede tomar un valor $+1$ y -1 , entonces solo puede tomar un valor negativo un número impar de veces, porque en los momentos pares serían $S_n \geq 0$



- Puede ser que la partícula nunca visite una coordenada negativa, de modo que $P(T = \infty) > 0$ y, en tal caso, se dice que la variable T es degenerada
- Para números impares $n \geq 3$ se puede obtener la siguiente relación, en donde $r_n = P(T = n)$:

$$P(T = n | X_1 = 1) = \frac{P(T = n, X_1 = 1)}{P(X_1 = 1)} = \frac{P(T = n, X_1 = 1)}{p}$$

$$\Rightarrow r_n = P(T = n, X_1 = 1) = pP(T = n | X_1 = 1)$$

for $n = 2k - 1$ when $k \geq 1$

- Considerando el primer retorno al origen $Z = \min\{m \geq 2 : X_m = 0\}$, se puede desarrollar esta igualdad para obtener la siguiente:

$$P(T = n | X_1 = 1) = \sum_{m=2}^{n-1} P(T = n | Z = m, X_1 = 1) P(Z = m | X_1 = 1)$$

- A partir de la propiedad de Markov y de resultados anteriores, se puede obtener una relación recursiva para $r_n = P(T = n)$ de la siguiente manera:

$$r_n = p \sum_{m=2}^{n-1} r_{n-m} r_{m-1} \quad \text{for } n = 2k - 1 \text{ when } k \geq 1$$

- Siendo $G_T(s) = \sum_{n \geq 0} r_n s^n$ la función generadora de probabilidad de la secuencia $\{r_n\}_{n=1}^{\infty}$ (asumiendo que $r_0 = 0$), se puede observar que $G_T(s)$ tiene la siguiente forma:

$$G_T(s) = r_1 s + r_3 s^3 + r_5 s^5 + \dots$$

- Debido a que $r_1 s = qs$, se puede comprobar que $ps(G_T(s))^2$ es igual a $G_T(s) - qs$, de modo que se puede obtener una expresión para $G_T(s)$ resolviendo la siguiente ecuación:

$$ps(G_T(s))^2 - G_T(s) + qs = 0$$

$$\Rightarrow G_T(s) = \frac{1 - \sqrt{1 - 4pqs^2}}{2ps} = \frac{2qs}{1 + \sqrt{1 - 4pqs^2}}$$

- El radio de convergencia de $G_T(s)$ es $R = 1/2\sqrt{pq}$, de modo que si $p = q = 1/2$, entonces $R = 1$, pero si $p \neq q$, entonces $R > 1$
- En general, se puede hacer una expansión para poder obtener la probabilidad de cada n impar (los coeficientes de las s)

$$G_T(s) = \frac{2qs}{1 + \sqrt{1 - 4pqs^2}} = qs + pq^2 s^3 + 2p^2 q^3 s^5 + \dots + 14p^4 q^5 s^9$$

- La probabilidad de una visita eventual a una coordenada negativa (sin especificar un momento concreto) se puede encontrar a través de la función generadora de probabilidad de T

$$\sum_{n=1}^{\infty} P(T = n) = \sum_{n=1}^{\infty} r_n = G_T(1) = \frac{1 - \sqrt{1 - 4pq}}{2p} =$$

$$\frac{1 - |q - p|}{2p} = \begin{cases} 1 & \text{if } p \leq q \\ q/p & \text{if } p > q \end{cases}$$

- Si $p \leq q$, entonces la partícula visitará una coordenada negativa con probabilidad 1. Debido a que la suma de las probabilidades es unitaria, T es una variable aleatoria con función de masa de probabilidad $r_n = P(T = n)$

- Si $p > q$, la variable T es degenerada, ya que hay una probabilidad positiva de que la partícula nunca visite una coordenada negativa:

$$P(T = \infty) = 1 - \sum_{n=1}^{\infty} P(T = n) = 1 - \frac{q}{p}$$

- Si T no es una variable aleatoria degenerada, entonces se puede obtener una expresión de la esperanza a través de $G'_T(1)$:

$$E(T) = G'_T(1) = \begin{cases} 1/(q-p) & \text{if } p < q \\ \infty & \text{if } p = q \end{cases}$$

- Si se hace una expansión de una parte de la función generadora de probabilidad $G_T(s)$, es posible encontrar una expresión explícita para r_n
- La probabilidad r_{2k-1} se puede encontrar (para un número impar $2k-1$) a través de expandir $(1-4pqs^2)^{1/2}$

$$(1+x)^\alpha = \sum_{m=0}^{\infty} \binom{\alpha}{m} x^m$$

$$\Rightarrow (1-4pqs^2)^{1/2} = \sum_{k=0}^{\infty} \binom{1/2}{k} (-4pqs^2)^k$$

- De este modo, se puede obtener la siguiente expresión, expandiendo el resultado y comprobando que se cumple la igualdad:

$$r_{2k-1} = -\frac{1}{2p} \binom{1/2}{k} (-4pqs^2)^k = \frac{1}{2k-1} \binom{2k-1}{k} p^{k-1} q^k$$

- Otro tipo de eventos importantes en un camino aleatorio son los relacionados con el retorno al origen, destacando la primera vez que ocurre, la no ocurrencia del evento y la última vez que ocurre
- La probabilidad de que la partícula visite el origen en un momento n es la siguiente:

$$P(S_n = 0) = \begin{cases} \binom{2k}{k} p^k q^k & \text{if } n = 2k \\ 0 & \text{if } n = 2k+1 \end{cases} \quad \text{for } k \geq 0$$

- Si $S_n = S_0 = 0$, el número de pasos a la derecha es el mismo que el de la izquierda en los primeros n pasos, lo cual es

imposible si n es impar. Por lo tanto, la probabilidad solo puede ser positiva para n impares

- Suponiendo que $n = 2k$, entonces $S_{2k} = X_1 + \dots + X_{2k}$ y $X_{2m} = 0$ si, y solo si, k de las X_i son iguales a $+1$ y k son iguales a -1 . Como hay $\binom{2k}{k}$ maneras de dividir las X_i en dos conjuntos y la probabilidad (dada la independencia) de caer en uno o en otro será p^k y q^k , respectivamente, se obtiene la expresión anterior
- La aproximación de Stirling para $m! \approx \sqrt{2\pi m}(m/e)^m$ permite obtener una expresión aproximada para la probabilidad:

$$P(S_{2k} = 0) = \frac{\sqrt{2\pi 2k}(2k/e)^{2k}}{2\pi k(k/e)^{2k}} p^k q^k = \frac{(4pq)^k}{\sqrt{\pi k}}$$

- Siendo A^* el límite superior de la secuencia de eventos $A_k = \{S_{2k} = 0\}$ para $k \geq 0$ (el evento de que el retorno al origen ocurra infinitamente seguido), se puede demostrar que la probabilidad de este evento es nula
- A través de la aproximación anterior se puede demostrar que la serie $\sum_{k=1}^{\infty} (4pq)^k / \sqrt{\pi k} < \infty$, y que, por tanto, $\sum_{k=1}^{\infty} P(A_k) < \infty$. A partir del primer lema de Borel-Cantelli, se puede ver que esto implica que $P(A^*) = 0$

$$P(\{S_{2k} = 0\} \text{ i. o.}) = 0$$

- Por lo tanto, si el camino aleatorio no es simétrico, $\{S_{2k} = 0\}$ solo ocurrirá un número finito de veces (con probabilidad 1)
- Siendo $f_n = P(Z = n)$, en donde $Z = \min\{n \geq 1: S_n = 0\}$ es el primer retorno al origen, se puede ver como f_n se puede definir de la siguiente manera:

$$f_0 = f_1 = 0$$

$$f_n = P(S_1 \neq 0, S_2 \neq 0, \dots, S_{n-1} \neq 0, S_n = 0) \text{ for } n \geq 2$$

- Para $n \geq 2$, se puede establecer la siguiente relación recursiva con las probabilidades de tomar una coordenada negativa y las de una positiva en el momento n (denotada por r'_n , y cuya expresión es la misma que para una coordenada negativa, pero intercambiando p por q):

$$f_n = P(Z = n) = pP(Z = n|X_1 = 1) + qP(Z = n|X_1 = -1) =$$

- $$\begin{aligned} G_Z(s) &= \sum_{n=0}^{\infty} s^n f_n = ps \sum_{n=2}^{\infty} s^{n-1} r_{n-1} + qs \sum_{n=2}^{\infty} s^{n-1} r'_{n-1} \\ &= ps \sum_{n=0}^{\infty} s^n r_n + qs \sum_{n=0}^{\infty} s^n r'_n = ps G_T(s) + qs G_T(s) \end{aligned}$$

$$\begin{aligned}\Rightarrow G_Z(s) &= psG_T(s) + qsG_T(s) = 1 - \sqrt{1 - 4pq s^2} = \\ &= 2pq s^2 + 2p^2 q^2 s^4 + 4p^3 q^3 s^6 + \dots\end{aligned}$$

- $$\begin{aligned} P(Z = 2k) &= f_{2k} = pr_{n-1} + qr'_{n-1} = \\ &= p \frac{1}{2k-1} \binom{2k-1}{k} p^{k-1} q^k + q \frac{1}{2k-1} \binom{2k-1}{k} p^k q^{k-1} = \\ &= \frac{2}{2k-1} \binom{2k-1}{k} p^k q^k = \frac{1}{2k-1} \binom{2k}{k} p^k q^k = \\ &= \frac{1}{2k-1} P(S_{2k} = 0) \text{ for } k \geq 1 \end{aligned}$$

- 164

$$\sum_{n=0}^{\infty} P(Z = n) = \sum_{n=0}^{\infty} f_n = G_Z(1) = 1 - \sqrt{1 - 4pq} = 1 - |q - p|$$

- Si $p \neq q$, entonces hay una probabilidad $|q - p|$ de que la partícula nunca vuelva al origen (ya que $\sum_{n=0}^{\infty} f_n \neq 1$ y eso quiere decir que la probabilidad de que no vuelva es positiva)
- Si $p = q$ (el camino aleatorio es simétrico), entonces Z no es una variable degenerada y el retorno al origen se produce con probabilidad 1, haciendo que Z sea una variable aleatoria con función de masa de probabilidad $f_n = P(Z = n)$
- No obstante, en el caso simétrico se puede calcular la esperanza, pero esta no es finita

$$E(Z) = G'_Z(1) = \frac{d}{ds} \left(1 - \sqrt{1 - s^2} \right) \Big|_{s=1} = \frac{s}{\sqrt{1 - s^2}} \Big|_{s=1} = \infty$$

- Para un camino aleatorio simétrico, se puede demostrar que la probabilidad de volver al origen en un momento n es la misma que la probabilidad de no volver al origen hasta el momento n

$$P(S_{2k} = 0) = P(S_1 \neq 0, S_2 \neq 0, \dots, S_{2k} \neq 0)$$

- Siendo $f_{2k} = P(Z = 2k)$ y $u_{2k} = P(S_{2k} = 0)$, si el camino aleatorio es simétrico, entonces se cumplen las siguientes identidades para las probabilidades:

$$f_{2k} = \frac{1}{2k-1} \binom{2k}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^k = \frac{1}{2k-1} \binom{2k}{k} \frac{1}{2^{2k}}$$

$$u_{2k} = \binom{2k}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^k = \binom{2k}{k} \frac{1}{2^{2k}}$$

$$\Rightarrow u_{2k-2} - u_{2k} = f_{2k} \text{ for } k \geq 1$$

- Por lo tanto, $P(S_1 \neq 0, S_2 \neq 0, \dots, S_{2k} \neq 0) = P(Z > 2k)$ y se puede obtener la siguiente igualdad:

$$P(S_1 \neq 0, S_2 \neq 0, \dots, S_{2k} \neq 0) = P(Z > 2k) = 1 - P(Z \leq 2k)$$

$$= 1 - \sum_{m=1}^k f_{2m} = 1 - \sum_{m=1}^k (u_{2m-2} - u_{2m}) = 1 - u_0 + u_{2k} =$$

$$= 1 - 1 + u_{2k} = u_{2k} = P(S_{2k} = 0)$$

- Este resultado también se puede demostrar de manera alternativa utilizando funciones generadoras de probabilidad

- Siendo $\alpha_0 = 1$ y $\alpha_k = P(S_1 \neq 0, S_2 \neq 0, \dots, S_{2k} \neq 0)$ para $k = 2m$ para $m \geq 1$ y $\alpha_k = 0$ para $k = 2m + 1$, se puede obtener la siguiente función generadora de probabilidad:

$$\begin{aligned}\alpha_{2k} &= 1 - \sum_{m=1}^k f_{2k} \text{ for } k \geq 0 \\ \Rightarrow U(s) &= \sum_{k=0}^{\infty} \alpha_{2k} s^{2k} = \sum_{k=0}^{\infty} \left(1 - \sum_{m=1}^k f_{2k} \right) s^{2k} = \\ &= \sum_{k=0}^{\infty} (s^{2k}) - \sum_{k=0}^{\infty} \sum_{m=1}^k f_{2k} s^{2k} = \frac{1}{1-s^2} - \sum_{k=0}^{\infty} \left(f_{2k} \sum_{m=k}^{\infty} s^{2k} \right) = \\ &= \frac{1}{1-s^2} - \frac{1}{1-s^2} \sum_{k=0}^{\infty} f_{2k} s^{2k} = \frac{1}{1-s^2} (1 - G_Z(s)) = \\ &= \frac{1}{1-s^2} \sqrt{1-s^2} = \frac{1}{\sqrt{1-s^2}}\end{aligned}$$

- Por lo tanto, se puede utilizar la expansión binomial para poder obtener la siguiente igualdad:

$$\begin{aligned}U(s) &= (1-s^2)^{-\frac{1}{2}} = \sum_{k=0}^{\infty} (-1)^k \binom{-1/2}{k} s^{2k} = \\ &= \sum_{k=0}^{\infty} \binom{2k}{k} \frac{1}{2^{2k}} s^{2k} \Rightarrow \alpha_{2k} = \binom{2k}{k} \frac{1}{2^{2k}} = P(S_{2k} = 0) \\ &\text{as } (-1)^k \binom{-1/2}{k} = \binom{2k}{k} \frac{1}{2^{2k}}\end{aligned}$$

- Este lema se puede reescribir de una forma más conveniente para dar un signo a cada una de las sumas parciales

$$\begin{cases} P(S_1 > 0, S_2 > 0, \dots, S_{2k} > 0) = \frac{1}{2} P(S_{2k} = 0) \\ P(S_1 < 0, S_2 < 0, \dots, S_{2k} < 0) = \frac{1}{2} P(S_{2k} = 0) \end{cases}$$

- Considerando todos los posibles valores de S_{2k} , se puede obtener la siguiente igualdad, en donde los términos con $r > 2k$ desaparecen:

$$P(S_1 < 0, S_2 < 0, \dots, S_{2k} < 0) = P(S_1 > 0, S_2 > 0, \dots, S_{2k} > 0)$$

$$= \sum_{r=1}^{\infty} P(S_1 > 0, S_2 > 0, \dots, S_{2k-1} > 0, S_{2k} = 2r)$$

- Por el teorema de la urna, el número de caminos que satisface la condición indicada en la derecha de la ecuación será $N_{2k-1, 2r-1} - N_{2k-1, 2r+1}$, y el término r de la suma dará el siguiente resultado:

$$\frac{1}{2}(P(S_{2k-1} = 2r - 1) + P(S_{2k-1} = 2r + 1))$$

- La parte negativa del término r se cancela con la parte positiva del término $r + 1$, con el resultado de que la suma se reduce a $(1/2)P(S_{2k-1} = 1)$. Se puede verificar como, por lo tanto, $P(S_{2k-1} = 1) = P(S_1 \neq 0, S_2 \neq 0, \dots, S_{2k} \neq 0) = P(S_{2k} = 0)$
- Otra manera de reescribir el lema es considerando que las sumas parciales son iguales o mayores o menores que cero

$$P(S_1 \geq 0, \dots, S_{2k} \geq 0) = P(S_{2k} = 0)$$

- Un camino de longitud $2k$ con todos los vértices estrictamente sobre el eje x pasa a través del punto $(1,1)$. Tomando el punto como nuevo origen se obtiene un camino de longitud $2k - 1$ con los vértices sobre o cruzando el eje x
- Por lo tanto, se obtiene, pero el número S_{2k-1} es un número impar, lo cual implica que $S_{2k} \geq 0$. La probabilidad de la derecha, por lo tanto, es igual a $P(S_1 \geq 0, \dots, S_{2k} \geq 0)$

$$P(S_1 > 0, \dots, S_{2k} > 0) = \frac{1}{2}P(S_1 \geq 0, \dots, S_{2k-1} \geq 0)$$

- Finalmente, un análisis útil adicional que se puede hacer es el de último retorno al origen
 - La probabilidad $\alpha_{2k, 2m}$ de que un camino aleatorio simétrico de longitud $2m$ tenga un último retorno al origen en $2k$ es la siguiente:

$$\alpha_{2k, 2m} = \binom{2k}{k} \binom{2m-2k}{m-k} \frac{1}{2^{2m}} = P(S_{2k} = 0)P(S_{2m-2k} = 0)$$

for $0 \leq k \leq m$

- El caso $k = 0$ corresponde a $\{S_i \neq 0: 1 \leq i \leq 2m\}$, de modo que $\alpha_{0,2m} = P(S_1 \neq 0, S_2 \neq 0, \dots, S_{2m} \neq 0)$
- Siendo $A_{2k,2m}$ denota el evento de que el camino aleatorio simétrico de longitud $2m$ tiene un último retorno al origen en el momento $2k$ para $0 \leq k \leq m$. Entonces, se puede obtener la siguiente probabilidad:

$$\begin{aligned}\alpha_{2k,2m} &= P(A_{2k,2m}) = \\ &= P(\{S_{2k} = 0\} \cap \{S_{2k+1} \neq 0, S_{2k+2} \neq 0, \dots, S_{2m} \neq 0\}) = \\ &= P(S_{2k} = 0)P(S_{2k+1} \neq 0, S_{2k+2} \neq 0, \dots, S_{2m} \neq 0 | S_{2k} = 0)\end{aligned}$$

- Debido a que un camino aleatorio es temporalmente homogéneo, se puede obtener la siguiente identidad:

$$\begin{aligned}P(S_{2k+1} \neq 0, S_{2k+2} \neq 0, \dots, S_{2m} \neq 0 | S_{2k} = 0) &= \\ &= P(S_1 \neq 0, S_2 \neq 0, \dots, S_{2m-2k} \neq 0 | S_0 = 0) = \\ &= P(S_1 \neq 0, S_2 \neq 0, \dots, S_{2m-2k} \neq 0) = P(S_{2k} = 0) = \\ &= P(S_{2m-2k} = 0)\end{aligned}$$

- Por lo tanto, se pueden obtener los siguientes resultados, a partir del comportamiento asintótico:

$$\alpha_{2k,2m} = P(A_{2k,2m}) = P(S_{2k} = 0)P(S_{2m-2k} = 0)$$

$$\lim_{k \rightarrow \infty} P(S_{2k} = 0) = \frac{1}{\sqrt{\pi k}}$$

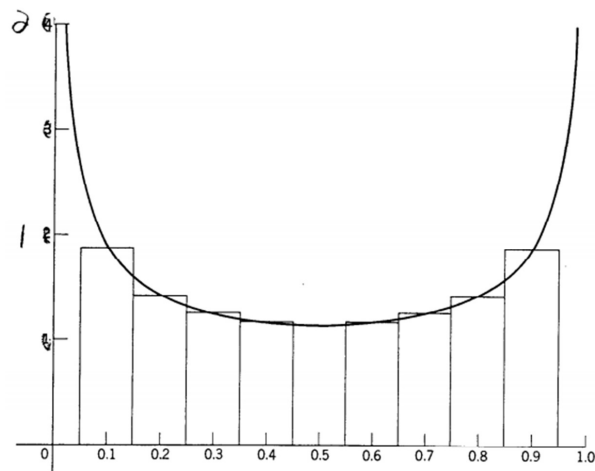
$$\lim_{k \rightarrow \infty} P(S_{2(m-k)} = 0) = \frac{1}{\sqrt{\pi(m-k)}}$$

$$\Rightarrow \alpha_{2k,2m} \rightarrow \frac{1}{\sqrt{\pi k(m-k)}} \text{ as } n \rightarrow \infty$$

- Debido a que $P(S_{2k} = 0) = P(S_{2m-2k} = 0)$, se puede ver como se cumple la siguiente identidad:

$$\alpha_{2k,2m} = \alpha_{2m-2k,2m}$$

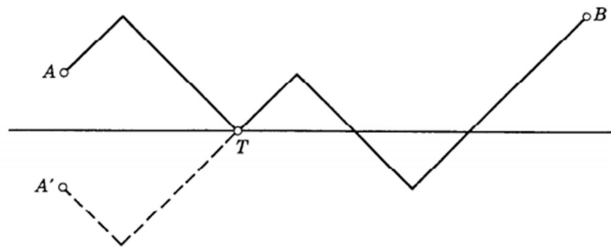
- Como cada trayectoria posible de longitud $2m$ ocurre con probabilidad $1/2^{2m}$, la simetría quiere decir que el número de caminos que tienen un retorno al origen en el momento $2k$ es el mismo que el número de caminos que tienen un retorno al origen en el momento $2m - 2k$
- Entonces, la probabilidad de que un camino aleatorio simétrico de longitud $2m$ no tenga un retorno al origen durante los últimos m pasos es $1/2$ (porque en cada paso hay una probabilidad $1/2$ de divergir hacia una dirección diferente a la que esté 0)
- Volviendo a la interpretación del camino aleatorio como un juego de lanzar la moneda entre dos jugadores, si los jugadores A y B juegan a un juego de longitud $2m$, la probabilidad de que A tome la ventaja en el momento $2k$ es $\alpha_{2k,2m}$



	$k = 0$ $k = 10$	$k = 1$ $k = 9$	$k = 2$ $k = 8$	$k = 3$ $k = 7$	$k = 4$ $k = 6$	$k = 5$
$\alpha_{2k,20}$	0.1762	0.0927	0.0736	0.0655	0.0617	0.0606

- Aunque lo intuitivo sería decir que el número de veces que es más probable que A tome ventaja sería m , este no es el caso. El valor menos probable de momentos es $2k = m$, mientras que los números más probables en los que A tome ventaja son $2k = 0$ y $2k = 2m$

- Dos de los resultados más importantes para poder operar con este marco teórico son el principio del reflejo y el teorema de la urna
 - Siendo $A = (a, \alpha)$ y $B = (b, \beta)$ puntos enteros en el cuadrante positivo (de modo que $b > a \geq 0$, $\alpha > 0$ y $\beta > 0$), la reflexión de A en el eje t es el punto $A' = (a, -\alpha)$. Por lo tanto, el número de caminos de A a B que tocan o cruzan el eje x iguala el número de caminos de A' a B



- A este lema se le conoce como el principio del reflejo
- Considerando un camino $(S_a = \alpha, S_{a+1}, \dots, S_b = \beta)$ de A a B teniendo uno o más vértices en el eje t o cruzándolo. Siendo t la abscisa del primero de esos vértices (se escoge t tal que $S_a > 0, \dots, S_{t-1} > 0, S_t = 0$)
- Entonces $(-S_a, -S_{a+1}, \dots, -S_{t-1}, S_t = 0, S_{t+1}, \dots, S_n)$ es un camino de A' a B y teniendo $T = (t, 0)$ como primer vértice que cruza el eje t . Las secciones AT y $A'T$ son reflexiones la una de la otra, y existe una correspondencia entre los caminos de A' a B y los caminos de A a B que tienen un vértice en el eje x
- A partir del resultado anterior, se puede considerar el teorema de la urna. Siendo m y k enteros positivos, hay exactamente $(k/m)N_{n,k}$ caminos $(S_1, \dots, S_m = k)$ desde el punto de origen al punto (m, k) tal que $S_1 > 0, S_2 > 0, \dots, S_m > 0$
 - En este caso, se define $n = r + s$ y $k = r - s$ (como antes)
 - Claramente existen tantos caminos admisibles como caminos desde $(1, 1)$ a (n, k) (porque $S_1 > 0$) que no cruzan o tocan el eje t . Por el lema anterior, el número de caminos posibles es igual a la siguiente cantidad:

$$N_{n-1, k-1} - N_{n-1, k+1} = \binom{r+s-1}{r-1} - \binom{r+s-1}{r}$$

$$N_{n-1, k-1} = n^{\circ} \text{ paths from } (1, -1) \text{ to } (n-1, x-1)$$

$$N_{n-1,k+1} = n^{\circ} \text{ paths from } (1, -1) \text{ to } (n-1, x+1)$$

- Calculando esta cantidad, se puede obtener que el lado derecho es igual a $N_{n,x}(r-s)/(r+s)$

$$\begin{aligned} \binom{r+s-1}{r-1} - \binom{r+s-1}{r} &= \frac{(r+s-1)!}{s!(r-1)!} - \frac{(r+s-1)!}{r!(s-1)!} = \\ &= \frac{(r+s-1)!r!(s-1)!}{s!(r-1)!r!(s-1)!} - \frac{(r+s-1)!s!(r-1)!}{s!(r-1)!r!(s-1)!} = \\ &= \binom{r+s}{r-s} \frac{k}{n} = N_{n,k} \frac{r-s}{r+s} \end{aligned}$$

- Debido a que el número total de caminos diferentes posibles es $N_{n,k}$, a partir del teorema también es posible obtener la probabilidad de se produzca un camino con ordenadas positiva únicamente

$$P(S_1 > 0, S_2 > 0, \dots, S_n > 0) = \frac{N_{n,k} \frac{k}{n}}{N_{n,k}} = \frac{\binom{r+s}{r-s} \frac{r-s}{r+s}}{\binom{r+s}{r-s}} = \frac{k}{n}$$

Los procesos de Poisson

Commented [IC32]: Acabar bien Intermediate Probability

- Los procesos de Poisson son procesos estocásticos de tiempo continuo y espacio discreto que se pueden definir de diversas maneras, por las que se estudian y se analizan a partir de sus propiedades
 - Se supone que un evento E puede ocurrir en cualquier punto en el tiempo y que el número de ocurrencias de E durante intervalos de tiempo disjuntos son independientes
 - Algunos ejemplos de este tipo de experimentos son las llegadas de clientes a una tienda (E es la llegada del cliente), llamadas a un teléfono, accidentes en una calle, u otros
 - La característica común en todos estos ejemplos es que se realizan muchas repeticiones de experimentos de Bernoulli independientes y la probabilidad de éxito de cada experimento es muy pequeña
 - Imaginando un intervalo de tiempo $(0, t]$ dividido en n partes $(0, t/n], (t/n, 2t/n], \dots, ((n-1)t/n, t]$, donde n es "muy grande". Por lo tanto, la probabilidad de ocurrencia del evento es muy pequeña en cada intervalo de tiempo pequeño, los

eventos en intervalos de tiempo disjuntos son independientes, y el número de intervalos de tiempo es grande

- La aproximación de Poisson de la distribución binomial permite ver que el número total de ocurrencias en $(0, t]$ se distribuye aproximadamente como una Poisson. En todo caso, se descarta la probabilidad de más de una ocurrencia en un intervalo pequeño de tiempo
- Denotando X_t como el número de ocurrencias en $(0, t]$, un proceso de Poisson es un proceso estocástico $\{X_t, t \geq 0\}$ con incrementos independientes y estacionarios distribuidos como una Poisson. Además, $X_0 = 0$
 - Los incrementos $\{X_{t_k} - X_{t_{k-1}}, 1 \leq k \leq n\}$ son variables aleatorias independientes para toda $0 \leq t_0 \leq t_1 \leq \dots \leq t_{n-1} \leq t_n$ y para toda n
 - Se fija que $X_0 = 0$ y existe una $\lambda > 0$ tal que $X_t - X_s \sim \text{Pois}[\lambda(t-s)]$ para $0 \leq s < t$, en donde λ se denomina la intensidad del proceso
 - Por la ley de grandes números (esencialmente) o por la desigualdad de Chebyshev, se pueden obtener fácilmente de la definición que $X_t/t \xrightarrow{P} \lambda$ cuando $t \rightarrow \infty$. Por lo tanto, la intensidad mide la frecuencia media o densidad de las ocurrencias para un intervalo de tiempo $(0, t]$

$$\bar{X}_n = \frac{S_n}{n} \xrightarrow{a.s.} \mu \text{ as } n \rightarrow \infty$$

$$\Rightarrow E(X_t - X_0) = E(X_t) = \lambda \Rightarrow \frac{X_t}{t} \xrightarrow{a.s.} \lambda \text{ as } t \rightarrow \infty$$

$$\Rightarrow \frac{X_t}{t} \xrightarrow{P} \lambda \text{ as } t \rightarrow \infty$$

- Además de la independencia entre diferentes intervalos de tiempo disjuntos, se ha remarcado que es “casi imposible” que dos o más ocurrencias se den un intervalo de tiempo pequeño. Para un intervalo de tiempo arbitrario $((i-1)t/n, it/n]$ para $i = 1, 2, \dots, n$, por lo que es razonablemente probable que E ocurra una vez y es esencialmente imposible que E ocurra más de una vez
 - Esto se puede demostrar matemáticamente para el proceso de Poisson, de modo que estas propiedades (junto a la de

independencia de los incrementos) caracterizan el proceso de Poisson

- Primero se puede observar que $0 < 1 - e^{-x} < x$ para $x > 0$, por lo que se puede obtener la siguiente igualdad:

$$\begin{aligned} P(E \text{ occurs once during } (t, t+h]) &= e^{-\lambda h} \frac{(\lambda h)^1}{1!} = e^{-\lambda h} \lambda h = \\ &= \lambda h - \lambda h(1 - e^{-\lambda h}) = \lambda h + O(h) \text{ as } h \rightarrow 0 \end{aligned}$$

- Además, la desigualdad $\sum_{k=2}^{\infty} \frac{x^k}{k!} \leq \frac{1}{2} \sum_{k=2}^{\infty} x^k \leq \frac{1}{2} \frac{x^2}{1-x} \leq x^2$ para toda $0 < x < 1/2$ implica lo siguiente:

$$\begin{aligned} P(\text{at least 2 occurrences of } E \text{ during } (t, t+h]) &= \\ &= \sum_{k=2}^{\infty} e^{-\lambda h} \frac{(\lambda h)^k}{k!} \leq (\lambda h)^2 = O(h) \text{ as } h \rightarrow 0 \end{aligned}$$

- Un proceso de Poisson $\{X_t, t \geq 0\}$ es un proceso estocástico entero no negativo tal que $X_0 = 0$ y que se cumplen las siguientes proposiciones:

- Los incrementos del proceso son independientes (como visto anteriormente)
- Se cumplen las siguientes igualdades para las probabilidades de ocurrencia:

$$P(1 \text{ occurrence during } (t, t+h]) = \lambda h + O(h) \text{ as } h \rightarrow 0$$

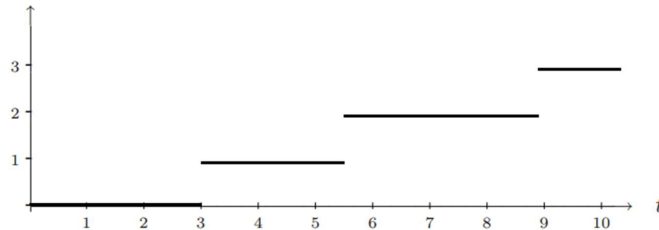
$$P(\geq 2 \text{ occurrences during } (t, t+h]) = O(h) \text{ as } h \rightarrow 0$$

- A partir de esta definición $\{X_t, t \geq 0\}$ no es decreciente, y esto también se puede inferir del hecho de que el proceso cuenta el número de ocurrencias. No obstante, hay veces que el proceso se define en términos de saltos en vez de ocurrencias, y entonces la suposición de que el proceso no es decreciente se tiene que incluir como suposición
- Es posible demostrar como la primera definición para los incrementos y esta segunda definición son equivalentes

- Como se ha visto en la derivación de las probabilidades, se ha demostrado como la primera definición implica la segunda. Por lo tanto, para demostrar lo converso, se quiere demostrar que...

Commented [IC33]: Acabar bien

- Una realización típica de un proceso de Poisson es, por tanto, una función a trozos que comienza en cero, en donde se mantiene constante por un periodo de tiempo aleatorio, y después salta a uno y se queda constante por un tiempo, y así



- En este caso, se considera que T_1, T_2, \dots son puntos sucesivos en el tiempo de las ocurrencias de un evento E (cuándo ocurre). Fijando $\tau_1 = T_1$ y $\tau_k = T_k - T_{k-1}$ para $k \geq 2$. En este caso, las τ son las duraciones (la longitud del intervalo para un número de ocurrencias hasta que se da otra)
- Siendo $t > 0$, como $\{T_1 > t\} = \{X_t = 0\}$, entonces se puede obtener que T_1 y τ_1 se distribuyen como una $Exp(1/\lambda)$

$$F_{\tau_1}(t) = F_{T_1}(t) = P(T_1 \leq t) = 1 - e^{-\lambda t}$$

$$\Rightarrow 1 - F_{\tau_1}(t) = 1 - F_{T_1}(t) = P(T_1 > t) = P(X_t = 0) = e^{-\lambda t}$$

- La distribución exponencial es famosa por su propiedad de falta o pérdida de memoria. Si en un momento s se sabe que no ha habido ninguna ocurrencia, entonces el tiempo de espera residual hasta que ocurra se distribuye igual (se tiene la misma distribución que para el tiempo de espera inicial)

$$P(T_1 > t + s | T_1 > s) = e^{-\lambda t} = P(T_1 > t)$$

$$P(T_1 > t + s | T_1 > s) = \frac{P(T_1 > t + s)}{P(T_1 > s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t}$$

- Como la distribución exponencial, y, en consecuencia, la distribución de Poisson (por su forma exponencial) no tienen memoria (comienzan de nuevo) en cualquier punto temporal fijo observado, uno puede pensar que también ocurre lo mismo para momentos aleatorios
- Si esto fuera verdad, entonces el tiempo hasta la primera ocurrencia debería distribuirse $Exp(1/\lambda)$ y $\tau_2 \sim Exp(1/\lambda)$ y, además, τ_1 y τ_2 deben ser independientes, por lo que $T_2 \sim \Gamma(2, 1/\lambda)$ y estos argumentos se pueden repetir

- Para $k \geq 1$, siendo T_k el momento de la ocurrencia k en un proceso de Poisson, y fijando $\tau_1 = T_1$ y $\tau_k = T_k - T_{k-1}$ para $k \geq 2$, entonces τ_k para $k \geq 1$ son variables aleatorias independientes distribuidas $Exp(1/\lambda)$ y $T_k \sim \Gamma(k, 1/\lambda)$

▪ La demostración...

Commented [IC34]: Acabar bien

- Siendo $\{X_t, t \geq 0\}$ un proceso estocástico con $X_0 = 0$, y considerando que τ_1 es el momento de primera ocurrencia y que τ_k es el tiempo entre la ocurrencia $k-1$ y la k (para $k \geq 2$), entonces si $\{\tau_k, k \geq 1\}$ son variables aleatorias independientes distribuidas como $Exp(\theta)$ para alguna $\theta > 0$ y X_t es el número de ocurrencias en $(0, t]$, entonces $\{X_t, t \geq 0\}$ es un proceso de Poisson con intensidad $\lambda = \theta^{-1}$

- Esta tercera definición es equivalente a la primera y a la segunda definición, y esto se puede demostrar a través de demostrar que un proceso estocástico como el definido tiene incrementos independientes, estacionarios y distribuidos como una Poisson

$$P(X_t = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!} \text{ for } k = 0, 1, 2, \dots \text{ where } \lambda = \theta^{-1}$$

- El primer paso consiste en obtener las siguientes probabilidades a través de la distribución de $\{\tau_k, k \geq 1\}$

$$P(X_t = 0) = P(\tau_1 > t) = \int_t^\infty \lambda e^{-\lambda x} dx = e^{-\lambda t}$$

$$\text{as } T_k = \tau_1 + \tau_2 + \dots + \tau_k \text{ \& } T_k \sim \Gamma(k, 1/\lambda)$$

$$\begin{aligned} P(X_t = k) &= P(X_t \geq k) - P(X_t \geq k+1) = \\ &= P(T_k \leq t) - P(T_{k+1} \leq t) = P(T_{k+1} > t) - P(T_k > t) = \\ &= \int_t^\infty \frac{1}{\Gamma(k+1)} \lambda^{k+1} x^k e^{-\lambda x} dx - \int_t^\infty \frac{1}{\Gamma(k)} \lambda^k x^{k-1} e^{-\lambda x} dx = \\ &= \sum_{j=0}^k e^{-\lambda t} \frac{(\lambda t)^j}{j!} - \sum_{j=0}^{k-1} e^{-\lambda t} \frac{(\lambda t)^j}{j!} = e^{-\lambda t} \frac{(\lambda t)^k}{k!} \end{aligned}$$

- Considerando dos intervalos de tiempo $(0, s]$ y $(s, s+t]$ conjuntamente, si $i \geq 0$ y $j \geq 2$ son enteros no negativos, entonces...

Commented [IC35]: Acabar bien

- Ahora que se tienen tres definiciones equivalentes para el proceso de Poisson, se pueden utilizar estas a conveniencia para poder establecer varias propiedades del proceso
 - Usando la propiedad de pérdida de memoria se puede asegurar que un proceso de Poisson que comience en un punto (posterior) fijo en el tiempo es un proceso de Poisson
 - Como el proceso de Poisson siempre comienza en cero, se tiene que sustraer el valor del nuevo punto inicial
 - Si $\{X_{t+s} - X_s, t \geq 0\}$ es un proceso de Poisson, entonces también se cumple que $\{X_{t+s} - X_s, t \geq 0\}$ para toda $s > 0$ fija
 - Si $Y_t = X_{t+s} - X_s$ para $t \geq 0$, el proceso Y_t tiene incrementos independientes y una ocurrencia en el proceso Y durante $(t, t + h]$ corresponde a una ocurrencia en el proceso X durante $(t + s, t + s + h]$. Por lo tanto, las propiedades de la segunda definición se satisfacen y se demuestra el resultado
 - Si $\{X_{t+s} - X_s, t \geq 0\}$ es un proceso de Poisson, entonces también se cumple que $\{X_{t+s} - X_s, t \geq 0\}$ para toda $s > 0$ fija
 -
 - Random times
 - Some further topics...
- También es posible investigar como un número dado de ocurrencias de un proceso de Poisson durante un intervalo de tiempo fijo se distribuyen dentro de ese intervalo de tiempo. Por simplicidad, se asume que el intervalo de tiempo es $(0,1]$
 - Todos los resultados son independientes de la intensidad del proceso de Poisson, dado que la intensidad solo actúa como un factor de escala, y el condicionamiento hace que estos efectos de escala desaparezcan
 - Además, si $Y \sim \text{Exp}(\theta)$, para $\theta > 0$, entonces $aY \sim \text{Exp}(a\theta)$ para cualquier $a > 0$. Explotando estos hechos y la propiedad de falta de memoria, es fácil formular y demostrar los resultados correspondientes para intervalos generales
 - El problema más simple es determinar la distribución de T_1 dado que $X_1 = 1$: debido a la propiedad de pérdida o falta de memoria, el proceso no debería ser capaz de recordar cuando

Commented [IC36]: Acabar bien

hubo una ocurrencia en el intervalo de tiempo $(0,1]$. Todos los puntos deberían, en algún sentido, ser igualmente probables

- La distribución condicional de T_1 dado que $X_1 = 1$ es la distribución uniforme $U(0,1)$

$$F_{T_1|X_1=1}(t) = P(T_1 \leq t | X_1 = 1) = \begin{cases} 0 & \text{for } t < 0 \\ t & \text{for } 0 \leq t \leq 1 \\ 1 & \text{for } t > 1 \end{cases}$$

$$F_{T_1|X_1=1}(t) = \begin{cases} 1 & \text{for } 0 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- La demostración de este resultado se puede llevar a cabo de la siguiente manera:

$$\begin{aligned} P(T_1 \leq t | X_1 = 1) &= \frac{P(T_1 \leq t, X_1 = 1)}{P(X_1 = 1)} = \frac{P(X_t = 1, X_1 = 1)}{P(X_1 = 1)} = \\ &= \frac{P(X_t = 1, X_1 - X_0 = 1)}{P(X_1 = 1)} = \frac{P(X_t = 1)P(X_1 - X_0 = 1)}{P(X_1 = 1)} = \\ &= \frac{\lambda t e^{-\lambda} e^{-\lambda(1-t)}}{\lambda e^{-\lambda}} = t \end{aligned}$$

- Los casos $t < 0$ y $t > 1$ son triviales, de modo que se demuestran los resultados
- Si $X_1 = n$, intuitivamente, se tienen n puntos, los cuales se comportan acorde al teorema anterior, y debido a la propiedad de pérdida de memoria, es razonable creer que se comportan independientemente unos de otros
- Para $k = 1, 2, \dots, n$ se cumple que $T_k | X_1 = n \sim \text{Beta}(k, n + 1 - k)$, lo cual es equivalente a la siguiente distribución:

$$f_{T_k | X_1=n}(t) = \begin{cases} \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n+1-k)} t^{k-1} (1-t)^{n-k} & \text{for } 0 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- Para $k = n = 1$, el teorema anterior sobre la distribución uniforme $U(0,1)$

$$P(T_k \leq t | X_1 = n) = \frac{P(T_k \leq t, X_1 = n)}{P(X_1 = n)} =$$

$$\begin{aligned}
&= \frac{\int_0^t P(X_1 = n | T_k = s) f_{T_k}(s) ds}{P(X_1 = n)} = \\
&= \frac{\int_0^t P(X_1 - X_s = n - k) f_{T_k}(s) ds}{P(X_1 = n)} = \\
&= \frac{\int_0^t e^{-\lambda(1-s)} \frac{(\lambda(1-s))^{n-k}}{(n-k)!} \frac{1}{\Gamma(k)} \lambda^k s^{k-1} e^{-\lambda s} ds}{e^{-\lambda} \frac{\lambda^n}{n!}} = \\
&= \frac{n!}{\Gamma(k)(n-k)!} \int_0^t s^{k-1} (1-s)^{n-k} ds = \\
&= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n+1-k)} \int_0^t s^{k-1} (1-s)^{n+1-k-1} ds
\end{aligned}$$

- La densidad conjunta condicional de T_1, T_2, \dots, T_n dado que $X_1 = n$ es la siguiente:

$$f_{T_1, T_2, \dots, T_n | X_1 = n}(t_1, t_2, \dots, t_n) = \begin{cases} n! & \text{for } 0 < t_1 < t_2 < \dots < t_n < 1 \\ 0 & \text{otherwise} \end{cases}$$

- Primero se determina la distribución de (T_1, T_2, \dots, T_n) : para τ_k con $1 \leq k \leq n$, se obtienen los siguientes resultados del teorema anterior (para $u_k > 0$):

$$f_{\tau_1, \tau_2, \dots, \tau_n}(u_1, u_2, \dots, u_n) = \prod_{k=1}^n \lambda e^{-\lambda u_k} = \lambda^n \exp\left(-\lambda \sum_{k=1}^n u_k\right)$$

$$\Rightarrow f_{T_1, T_2, \dots, T_n}(t_1, t_2, \dots, t_n) = \lambda^n e^{-\lambda t_n} \text{ for } 0 < t_1 < \dots < t_n$$

$$\Rightarrow f_T(\mathbf{t}) = \lambda^n e^{-\lambda t_n} \text{ for } 0 < t_1 < \dots < t_n$$

- Procediendo como en demostraciones anteriores, se puede obtener el siguiente resultado:

$$\begin{aligned}
&P(T_1 \leq t_1, \dots, T_n \leq t_n | X_1 = n) = \\
&= \frac{P(T_1 \leq t_1, \dots, T_n \leq t_n, X_1 = n)}{P(X_1 = n)} = \\
&= \frac{\int \int \dots \int P(X_1 - X_{s_n} = 0) f_T(\mathbf{s}) ds_1 ds_2 \dots ds_n}{P(X_1 = n)} =
\end{aligned}$$

$$= \frac{\int_0^{t_1} \int_{s_1}^{t_2} \dots \int_{s_{n-1}}^{t_n} e^{-\lambda(1-s_n)} \lambda^n e^{-\lambda s_n} ds_n ds_{n-1} \dots ds_1}{e^{-\lambda} \frac{\lambda^n}{n!}} =$$

$$= n! \int_0^{t_1} \int_{s_1}^{t_2} \dots \int_{s_{n-1}}^{t_n} ds_n ds_{n-1} \dots ds_1$$

- Esto establece que la distribución condicional conjunta de los tiempos de ocurrencia sea la misma que la de los estadísticos de orden de una muestra $U(0,1)$
- Siendo U_1, U_2, \dots, U_n variables aleatorias independientes distribuidas como una $U(0,1)$, y siendo $U_{(1)} \leq U_{(2)} \leq \dots \leq U_{(n)}$ el orden de las variables, el siguiente vector se distribuye de la siguiente manera:

$$((T_1, T_2, \dots, T_n) | X_1 = n) \sim (U_{(1)}, U_{(2)}, \dots, U_{(n)})$$

- Un problema relacionado con estos resultados es el cálculo de la probabilidad condicional $P(X_s = k | X_t = n)$ para $k = 0, 1, \dots, n$ y $0 \leq s \leq t$. Una solución se puede encontrar usando el teorema anterior
- Como las ocurrencias se distribuyen uniformemente en $(0, t]$, la probabilidad de que una ocurrencia preceda s (que haya una ocurrencia antes o en s) es s/t para $0 \leq s \leq t$. Debido a la independencia, se concluye que para $0 \leq s \leq t$ se cumple lo siguiente:

$$\{n^o \text{ occurrences } (0, s] | X_t = n\} \sim \text{Bin}(n, s/t)$$

$$\Rightarrow P(X_s = k | X_t = n) = \binom{n}{k} \left(\frac{s}{t}\right)^k \left(1 - \frac{s}{t}\right)^{n-k} \text{ for } k = 0, 1, \dots, n$$

- Se puede demostrar la siguiente igualdad utilizando la distribución del número de ocurrencias en $(0, s]$ condicionado a $X_t = n$:

$$P(T_k \leq t | X_1 = n) = P(X_t \geq k | X_1 = n)$$

$$\text{for } k = 1, \dots, n \text{ \& } 0 \leq t \leq 1$$

$$\Rightarrow \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n+1-k)} \int_0^t x^{k-1} (1-x)^{n-k} dx =$$

$$= \sum_{j=k}^n \binom{n}{j} t^j (1-t)^{n-j}$$

- Este resultado se puede generalizar a varios subintervalos. Explícitamente, por argumentos similares, uno puede demostrar que la distribución condicional conjunta para el siguiente vector es una multinomial de parámetros $(n; p_1, \dots, p_k)$ cuando $p_j = t_j - s_j$ para $j = 1, 2, \dots, k$:

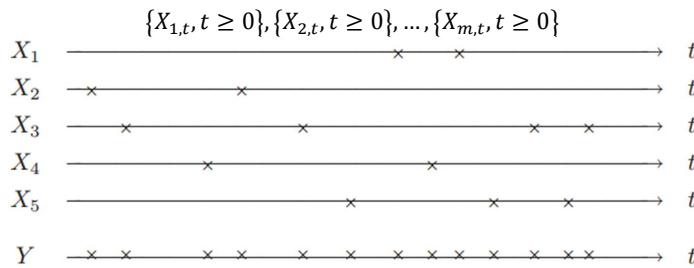
$$(X_{t_1} - X_{s_1}, X_{t_2} - X_{s_2}, \dots, X_{t_k} - X_{s_k}) \mid X_1 = n$$

$$\text{for } 0 \leq s_1 < t_1 \leq s_2 < t_2 \leq \dots \leq s_k < t_k \leq 1$$

• Condicionando al tiempo

Commented [IC37]: Acabar bien

- Suponiendo que se tienen m procesos de Poisson independientes con intensidades $\lambda_1, \lambda_2, \dots, \lambda_m$ respectivamente, se considera un nuevo proceso $\{Y_t, t \geq 0\}$ en donde las ocurrencias de Y se definen como la unión de todas las ocurrencias de X_k para $k = 1, 2, \dots, m$ (toda ocurrencia de Y corresponde con una ocurrencia X_k para alguna k y viceversa)



- El proceso Y se denomina proceso de Poisson superposicionado, y la inclusión del proceso de Poisson en el nombre se motiva por el siguiente resultado:

- El proceso $\{Y_t, t \geq 0\}$ es un proceso de Poisson con intensidad $\lambda = \lambda_1 + \lambda_2 + \dots + \lambda_m$
- El proceso Y tiene incrementos independientes porque todos los procesos X los tienen y porque los procesos son independientes entre si. Además, como la suma de variables independientes distribuidas Poisson es una variable Poisson con parámetro de intensidad igual a la suma de las individuales, se obtiene la siguiente igualdad:

$$Y_{t+s} - Y_s = \sum_{k=1}^m (X_{k,t+s} - X_{k,s}) \sim \text{Pois} \left(\sum_{k=1}^m \lambda_k t \right) \quad \text{for } \forall s, t \geq 0$$

- Las duraciones en el proceso Y se distribuyen como una $Exp((\sum_k^m \lambda_k)^{-1})$

- Siendo $T^{(k)}$ el tiempo hasta la primera ocurrencia para el proceso X_k para $k = 1, 2, \dots, m$, entonces $T^{(1)}, T^{(2)}, \dots, T^{(m)}$ son independientes, $T^{(k)} \sim Exp(1/\lambda_k)$ para $k = 1, 2, \dots, m$ y T_y se define de la siguiente manera:

$$T_y = \min_{1 \leq k \leq m} T^{(k)}$$

- Por lo tanto, se puede obtener el siguiente resultado:

$$\begin{aligned} P(T_y > t) &= P\left(\bigcap_{k=1}^m \{T^{(k)} > t\}\right) = \prod_{k=1}^m P(T^{(k)} > t) = \\ &= \prod_{k=1}^m e^{-\lambda_k t} = \exp\left[-\left(\sum_{k=1}^m \lambda_k\right)t\right] \text{ for } t \geq 0 \end{aligned}$$

$$T_y \sim Exp\left(\left(\sum_k^m \lambda_k\right)^{-1}\right)$$

- Considerando una j fija, y fijando $\tilde{T}^{(j)} = \min\{T^{(i)}: i \neq j\}$, entonces $T_y = \min\{T^{(j)}, \tilde{T}^{(j)}\}$ y $\tilde{T}^{(j)}$ son independientes del proceso X_j , y eso hace que $\{X_{j, T_y+t} - X_{j, T_y}, t \geq 0\}$ es un proceso de Poisson con intensidad λ_j
- Como j arbitraria, la misma conclusión se mantiene para j , lo que implica que el tiempo entre la primera y la segunda ocurrencia de Y es la misma porque el tiempo hasta que la primera ocurrencia de este proceso generado por los procesos X se reinicien en T_y
- ...

- CUANDO OCURRE EL PRIMER EVENTO

- Extra

- Thinning (marked process) (ENTERO)
- Proceso compuesto de Poisson
- Extensiones

Commented [IC38]: Acabar bien

Los procesos de renovación

Las cadenas de Markov

- Las cadenas de Markov son un tipo de procesos estocásticos muy importantes debido a sus amplias aplicaciones y los diferentes tipos de procesos que cumplen con la propiedad de Markov
 - Un proceso estocástico tiene la propiedad de Markov si, condicional a su valor presente, su futuro es independiente de su pasado. Esta es una suposición fuerte, pero que tiene dos beneficios
 - El primer beneficio es que se pueden modelar muchos procesos naturales a través de un proceso que cumpla esta propiedad, y el segundo es que la teoría matemática de este tipo de procesos es muy completa
 - Siendo S un conjunto numerable llamado espacio de estados, y siendo $\mathbf{X} = (X_n: n \geq 0)$ una secuencia de variables aleatorias tomando valores en S , las funciones X_n son funciones sobre un espacio de probabilidad común y se puede definir una cadena de Markov
 - La secuencia \mathbf{X} se denomina cadena de Markov si satisface la propiedad de Markov para toda $n \geq 0$ y toda $i_0, i_1, \dots, i_{n+1} \in S$
$$P(X_{n+1} = i_{n+1} | X_0 = i_0, \dots, X_n = i_n) = P(X_{n+1} = i_{n+1} | X_n = i_n)$$
 - La cadena de Markov se denomina homogénea si, para toda $i, j \in S$, la probabilidad condicional $P(X_{n+1} = j | X_n = i)$ no depende del valor de n . Por simplicidad, se asume que todas las cadenas de Markov que se verán a lo largo del desarrollo teórico son homogéneas
$$P(X_{n+1} = j | X_n = i) = P(X_1 = j | X_0 = i)$$
 - Una manera informal de definir una cadena de Markov con esta notación es que una secuencia \mathbf{X} es una cadena de Markov si, condicional al valor presente de X_n , el futuro $(X_r: r > n)$ es independiente del pasado $(X_m: m < n)$
 - Algunos ejemplos de cadenas de Markov son las siguientes, las cuales tienen una teoría coherente apoyándose en la suposición de independencia de la propiedad de Markov:
 - ...

Commented [IC39]: Acabar bien

- Para poder calcular las probabilidades asociadas con la cadena, es necesario saber dos cantidades: la matriz de transición y la distribución inicial

- La matriz de transición $\mathbf{P} = [p_{ij}]_{i,j \in S}$ es una matriz que viene dada por las probabilidades $p_{ij} = P(X_1 = j | X_0 = i)$, mientras que el vector fila de la distribución inicial $\boldsymbol{\lambda} = (\lambda_i : i \in S)$ dada por $\lambda_i = P(X_0 = i)$
- Esta última puede ser un vector con solo un uno en donde comience la cadena (solo cuando se sabe ciertamente donde empieza la cadena). No obstante, suele ser un vector de probabilidades de tamaño $1 \times |S|$ siempre que S sea finito (asigna probabilidad a los diferentes estados) que indica la posibilidad de que se comience en un estado i
- La matriz \mathbf{P} , tendrá un tamaño $|S| \times |S|$ siempre que $|S| \neq \infty$, donde las filas i denotan el estado de inicio y las columnas j el estado final

- Debido a la suposición de homogeneidad, $P(X_1 = j | X_0 = i) = P(X_{n+1} = j | X_n = i)$ para toda $n \geq 0$, de modo que se puede caracterizar el par $(\mathbf{P}, \boldsymbol{\lambda})$ de la siguiente manera:

- El vector $\boldsymbol{\lambda}$ es una distribución en el sentido de que $\lambda_i \geq 0$ para $i \in S$ y $\sum_{i \in S} \lambda_i = 1$

$$\sum_{i \in S} \lambda_i = \sum_{i \in S} P(X_0 = i) = P(X_0 \in S) = 1$$

- La matriz \mathbf{P} es una matriz estocástica en el sentido de que $p_{ij} \geq 0$ para $i, j \in S$ y $\sum_{j \in S} p_{ij} = 1$ para toda $i \in S$, de modo que las filas de \mathbf{P} tienen que sumar 1

$$\sum_{j \in S} p_{ij} = \sum_{j \in S} P(X_{n+1} = j | X_n = i) = \sum_{j \in S} P(X_{n+1} \in S | X_n = i) = 1$$

- La matriz \mathbf{P} es una matriz doblemente estocástica en el sentido de que $p_{ij} \geq 0$ para $i, j \in S$ y $\sum_{j \in S} p_{ij} = 1$ para toda $i \in S$ y $\sum_{i \in S} p_{ij} = 1$ para toda $j \in S$, de modo que las filas y las columnas de \mathbf{P} tienen que sumar 1

$$\sum_{j \in S} p_{ij} = \sum_{j \in S} P(X_{n+1} = j | X_n = i) = \sum_{j \in S} P(X_{n+1} \in S | X_n = i) = 1$$

$$\sum_{i \in S} p_{ij} = \sum_{i \in S} P(X_{n+1} = j | X_n = i) = \sum_{i \in S} P(X_{n+1} = j | X_n \in S) = 1$$

- Siendo λ una distribución y P una matriz estocástica, la secuencia aleatoria $X = (X_n: n \geq 0)$ es una cadena de Markov con distribución inicial λ y matriz de transición P si, y solo si, se cumple la siguiente igualdad para toda $n \geq 0$ y $i_0, i_1, \dots, i_n \in S$:

$$P(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = \lambda_{i_0} p_{i_0, i_1} p_{i_1, i_2} \dots p_{i_{n-1}, i_n}$$

▪ La demostración

Commented [IC40]: Acabar bien

- La propiedad de Markov proclama que el pasado solo afecta al futuro a través del presente, lo cual se puede formalizar a través de la propiedad de Markov extendida, en donde X_n es el valor presente, F un evento futuro y H un evento histórico o pasado

- Siendo X una cadena de Markov, para $n \geq 0$, para cualquier evento H dado en términos de su historia pasada, X_0, X_1, \dots, X_{n-1} y cualquier evento F dados en términos del futuro X_{n+1}, X_{n+2}, \dots , se cumple la siguiente igualdad:

$$P(F | X_n = i, H) = P(F | X_n = i) \quad \text{for } i \in S$$

- Una pequeña complicación nace del hecho de que F pueda depender del futuro infinito. Hay un argumento general en teoría de la probabilidad que permite restringirse al caso en que F depende de los valores del proceso un número finito de veces, el cual no se explica aquí
- Por definición de la probabilidad condicional y por el teorema anterior, se puede obtener la siguiente igualdad:

$$\begin{aligned} P(F | X_n = i, H) &= \frac{P(F, X_n = i, F)}{P(H, X_n = i)} = \\ &= \frac{\sum_{<n} \sum_{>n} \lambda_{i_0} p_{i_0, i_1} p_{i_1, i_2} \dots p_{i_{n-1}, i} p_{i, i_{n+1}} \dots}{\sum_{>n} \lambda_{i_0} p_{i_0, i_1} p_{i_1, i_2} \dots p_{i_{n-1}, i}} = \\ &= \sum_{<n} p_{i, i_{n+1}} p_{i_{n+1}, i_{n+2}} \dots = P(F | X_n = i) \end{aligned}$$

- Ahora que se ha introducido teoría básica sobre las cadenas de Markov, es posible hablar de las probabilidades de transición, dado que son los elementos más importantes

- Siendo \mathbf{X} una cadena de Markov con matriz de transición $\mathbf{P} = [p_{ij}]_{i,j \in S'}$, los elementos p_{ij} se denominan probabilidades de transición de un paso
 - De manera más general, las probabilidades de transición de n pasos se dan por $p_{ij}(n) = P(X_n = j | X_0 = i)$, y forman una matriz llamada matriz de transición de n pasos, definida como $\mathbf{P}(n) = [p_{ij}(n) : i, j \in S]$
 - Las matrices $\mathbf{P}(n)$ satisfacen una colección de ecuaciones llamadas las ecuaciones de Chapman-Kolmogorov
 - Se asume, por convención de que $p_{ij}(0) = 1$ si $i = j$, pero que $p_{ij}(0) = 0$ si $i \neq j$
 - A partir de esta matriz $\mathbf{P}(n)$, es posible saber la media $E(X_n = j | X_0 = i)$ y el estado más probable después de n pasos
- Para toda $i, j \in S$ y $m, n \geq 0$, se cumplen las siguientes igualdades, llamadas ecuaciones de Chapman-Kolmogorov:

$$p_{ij}(m+n) = \sum_{k \in S} p_{ik}(m)p_{kj}(n) \Rightarrow \mathbf{P}(m+n) = \mathbf{P}(m)\mathbf{P}(n)$$

- Por la definición de probabilidad condicional, es posible obtener las siguientes igualdades:

$$\begin{aligned} p_{ij}(m+n) &= P(X_{m+n} = j | X_0 = i) = \\ &= \sum_{k \in S} P(X_{m+n} = j | X_m = k, X_0 = i) P(X_m = k | X_0 = i) = \\ &= \sum_{k \in S} P(X_{m+n} = j | X_m = k) P(X_m = k | X_0 = i) = \\ &= \sum_{k \in S} (X_m = k | X_0 = i) P(X_n = j | X_0 = k) P = \sum_{k \in S} p_{ik}(m)p_{kj}(n) \end{aligned}$$

- Debido a este teorema, las probabilidades de transición de n pasos forman una matriz $\mathbf{P}(n) = [p_{ij}(n)]_{i,j \in S}$ que satisface $\mathbf{P}(n) = \mathbf{P}(1)^n = \mathbf{P}^n$

$$\mathbf{P}(m+n) = \mathbf{P}(m)\mathbf{P}(n)$$

$$\Rightarrow \mathbf{P}(n) = \mathbf{P}(n-1)\mathbf{P}(1) = \mathbf{P}(n-2)\mathbf{P}(1)\mathbf{P}(1) =$$

$$= \mathbf{P}(1) \dots \mathbf{P}(1) = \mathbf{P}(1)^n = \mathbf{P}^n$$

- Por lo tanto, una manera de poder calcular las probabilidades de transición es encontrar la potencia n de la matriz \mathbf{P} . Cuando el espacio de estado es finito, \mathbf{P} también lo será, y el cálculo de \mathbf{P}^n se realiza más fácilmente diagonalizando

- Para diagonalizar \mathbf{P} , primero es necesario calcular los valores propios de la matriz, por lo que habrá que obtener las raíces de la ecuación $\det(\mathbf{P} - \kappa \mathbf{I}) = 0$. Si n es un entero pequeño, multiplicando la matriz \mathbf{P} directamente se puede obtener el resultado deseado sin complicaciones
- Después, se descompone la matriz \mathbf{P} en la multiplicación de tres matrices, en donde la del medio es una matriz diagonal con los valores propios de \mathbf{P} de tamaño $|S| \times |S|$ y \mathbf{U} es una matriz invertible de vectores propios (en sus columnas)

$$\mathbf{P} = \mathbf{U}^{-1} \begin{bmatrix} \kappa_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \kappa_{|S|} \end{bmatrix} \mathbf{U}$$

- Al elevar a la n , se puede obtener una ecuación para la descomposición de \mathbf{P}^n , lo cual permitirá obtener una expresión para cada $p_{ij}(n)$ para recuperar la matriz \mathbf{P}^n

$$\mathbf{P}^n = \mathbf{U}^{-1} \begin{bmatrix} \kappa_1^n & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \kappa_{|S|}^n \end{bmatrix} \mathbf{U}$$

- Lo único que hace falta es encontrar una matriz invertible \mathbf{U} a través de encontrar los vectores propios, y así recuperar \mathbf{P}^n . Estos se pueden encontrar a través de la siguiente ecuación:

$$(\mathbf{P} - \kappa_i \mathbf{I}) \mathbf{u}_i = \mathbf{0} \text{ for } i = 1, 2, \dots, |S|$$

- La mayoría de la teoría de las cadenas de Markov requiere la manipulación de vectores y matrices. Las ecuaciones normalmente son lineales, y por tanto mucha parte del tema requerirá de álgebra lineal

- Por ejemplo, si X_0 tiene distribución inicial λ (un vector fila), entonces se cumple la siguiente igualdad, de modo que la distribución de X_1 es el vector fila $\lambda \mathbf{P}$:

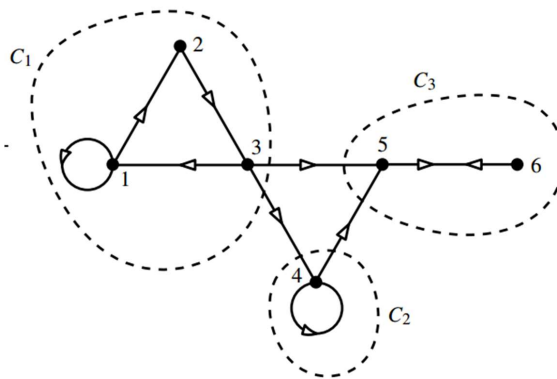
$$P(X_1 = j) = \sum_{i \in S} \lambda_i p_{ij} \text{ for } j \in S$$

- Por iteración, X_2 tiene distribución λP^2 y así. Por lo tanto, se adopta la convención de que las distribuciones de probabilidad son, por defecto, vectores fila, ya que actúan en el costado izquierdo de las matrices
 - En consecuencia, λ' denota la transpuesta del vector fila, siendo este un vector columna
- Un elemento importante de la teoría de las cadenas de Markov es la interacción entre el estado de espacio S y el mecanismo de transición P
 - Siendo X una cadena de Markov homogénea de espacio de estado S y matriz de transición P , para $i, j \in S$, se dice que i lleva a j (denotado por $i \rightarrow j$) si $p_{ij}(n) > 0$ para alguna $n \geq 0$
 - Es importante denotar que se considera $p_{ij}(n)$, de modo que no tiene por qué a ver una conexión directa entre estados de S , solo que se pueda conectar después de n pasos
 - Al fijar $n = 0$, se tiene que $i \rightarrow i$ para toda $i \in S$
 - Se escribe $i \leftrightarrow j$ si $i \rightarrow j$ y $j \rightarrow i$, y en este caso se dice que i y j se comunican (ya sea directamente o a través de otros elementos de S)
 - Es posible demostrar que la relación \leftrightarrow entre elementos del espacio de estado es una relación de equivalencia
 - Si la relación es de equivalencia, tiene que ser reflexiva, simétrica y transitiva. Debido a que $i \rightarrow i$, se tiene que $i \leftrightarrow i$, y, además, la relación es simétrica debido a que $i \leftrightarrow j$ cuando $j \leftrightarrow i$
 - Suponiendo que $i, j, k \in S$ satisfacen $i \leftrightarrow j$ y $j \leftrightarrow k$, debido a que $i \rightarrow j$ y $j \rightarrow k$, existe unas $m, n \geq 0$ tal que $p_{ij}(m) > 0$ y $p_{jk}(n) > 0$. Debido a las ecuaciones de Chapman-Kolmogorov, se puede ver que $i \rightarrow k$ y, de manera análoga, $k \rightarrow i$ y por tanto $i \leftrightarrow k$, lo cual quiere decir que \leftrightarrow es transitiva

$$p_{ik}(m+n) = \sum_{l \in S} p_{il}(m)p_{lk}(n) \geq p_{ij}(m)p_{jk}(n) > 0$$

- Debido a que \leftrightarrow es una relación de equivalencia, tiene clases de equivalencia, que son subconjuntos de S de la forma $C_i = \{j \in S : i \leftrightarrow j\}$ los cuales se denominan clases comunicadoras

- Una cadena X o el espacio de estado S es irreducible si solo hay una clase comunicadora, lo cual quiere decir que $i \leftrightarrow j$ para toda $i, j \in S$
- Un subconjunto $C \subseteq S$ es cerrado si $i \in C$ y $i \rightarrow j$ implica que $j \in C$. Si la cadena da en algún momento a un elemento de un conjunto cerrado C , entonces se queda en C para siempre



- Si un conjunto de un solo elemento $\{i\}$ es cerrado, entonces i se denomina estado absorbente o *absorbing*
- Un subconjunto C de estados es cerrado si, y solo si, $p_{ij} = 0$ para $i \in C$ y $j \notin C$
 - Siendo $C \subseteq S$, si $p_{ij} \neq 0$, entonces $i \in C$ e $i \rightarrow j$ no implica que $j \in C$ porque $j \notin C$, y eso haría que C no fuera cerrado
 - Se supone que $p_{ij} = 0$ para $i \in C$ y $j \notin C$ se mantiene y se considera $k \in C$ y $l \in S$ tal que $k \rightarrow l$. Debido a que $k \rightarrow l$, existe una $m \geq 0$ tal que $P(X_m = l | X_0 = k) > 0$, por lo que existe una secuencia $k_0 (= k), k_1, k_2, \dots, k_m (= l)$ con $p_{k_r, k_{r+1}} > 0$ para $r = 0, 1, \dots, m-1$ y eso hace que $k_r \in C$ para toda r , de modo que $l \in C$ y eso hace que $k \in C$ e $k \rightarrow l$ no implica que $l \in C$
- Uno de los conceptos más importantes es el de transitoriedad y recurrencia en una cadena de Markov, de modo que se tienen que analizar y estudiar los aspectos más importantes de estos
 - Se asume que X es una cadena de Markov homogénea con estado de espacio S y matriz de transición P

- El tiempo de primer paso al estado j se define como $T_j = \min\{n \geq 1 : X_n = j\}$, mientras que las probabilidades de primer paso se dan por $f_{ij}(n) = P(T_j = n | X_0 = i)$
- Un estado i es recurrente si $P(T_i < \infty) = 1$, mientras que un estado i se denomina transitorio si no es recurrente
- La propiedad de recurrencia es una propiedad de clase, en el sentido en que cualquier par de estados comunicados son ambos o recurrentes o transitorios
- Un criterio para la recurrencia en términos de la matriz de transición \mathbf{P} y sus potencias es que el estado i es recurrente si, y solo si, $\sum_{n=0}^{\infty} p_{ii}(n) = \infty$
 - La demostración de este teorema requiere de un teorema adicional y de la definición de unas funciones generadoras de probabilidad concretas
 - Las funciones generadoras de probabilidad necesarias se definen de la siguiente manera para $i, j \in S$

$$P_{ij}(s) = \sum_{n=0}^{\infty} p_{ij}(n)s^n \quad F_{ij}(s) = \sum_{n=0}^{\infty} f_{ij}(n)s^n$$

$$\text{where } f_{ij}(0) = 0 \quad \& \quad p_{ij}(0) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

- Se define $f_{ij} = F_{i,j}(1) = P(T_j < \infty | X_0 = i) = \sum_{n=0}^{\infty} f_{ij}(n)$ y se indica que i es recurrente si, y solo si, $f_{ij} = 1$
- Para $i, j \in S$, se tiene que $P_{ij}(s) = \delta_{ij} + F_{ij}(s)P_{jj}(s)$ para $s \in (-1, 1]$
 - Condicionando en el valor de T_j se obtiene la siguiente igualdad para $n \geq 1$:

$$p_{ij}(n) = \sum_{m=1}^{\infty} P(X_n = j | X_0 = i, T_j = m) P(T_j = m | X_0 = i)$$

- Este sumando es cero para $m > n$, ya que el momento mínimo en el que se pasa a j aún no se ha dado (no se puede dar si se va a dar en m). Para $m \leq n$, en cambio, se obtiene la siguiente igualdad:

$$P(X_n = j | X_0 = i, T_j = m) = P(X_n = j | X_m = j, H)$$

where $H = \{X_r \neq j \text{ for } 1 \leq r \leq m\}$

- Por la homogeneidad de las cadenas y la propiedad de Markov extendida, se puede obtener la siguiente igualdad:

$$\begin{aligned} P(X_n = j | X_0 = i, T_j = m) &= P(X_n = j | X_0 = i, X_m = j) = \\ &= P(X_{n-m} = j | X_0 = j) \end{aligned}$$

- Sustituyendo el resultado anterior, se puede obtener la siguiente ecuación:

$$\begin{aligned} p_{ij}(n) &= \sum_{m=1}^{\infty} P(X_{n-m} = j | X_0 = j) P(T_j = m | X_0 = i) = \\ &= \sum_{m=1}^{\infty} p_{jj}(n-m) f_{ij}(m) \text{ for } n \geq 1 \end{aligned}$$

- Multiplicando por s^n y sumando sobre $n \geq 1$, se obtiene la ecuación a demostrar:

$$\begin{aligned} P_{ij}(s) - p_{ij}(0) &= P_{jj}(s) F_{ij}(s) \\ \Rightarrow P_{ij}(s) &= p_{ij}(0) + P_{jj}(s) F_{ij}(s) \end{aligned}$$

- A partir de esto, se puede demostrar el teorema anterior sobre el criterio de transitoriedad

- Cuando $i = j$, por el teorema anterior, $P_{ii}(s) - 1 = P_{ii}(s) F_{ii}(s)$, y esto permite obtener la siguiente identidad:

$$P_{ii}(s) = \frac{1}{1 - F_{ii}(s)} \text{ for } |s| < 1$$

- Cuando $s \rightarrow 1^-$, se tiene que, por el lema de Abel, las siguientes proposiciones:

$$\lim_{s \rightarrow 1^-} F_{ii}(s) = F_{ii}(1) = f_{ii} \quad \lim_{s \rightarrow 1^-} P_{ii}(s) = \sum_{n=0}^{\infty} p_{ii}(n)$$

- Debido a la ecuación para $P_{ii}(s)$, se tiene que $\sum_{n=0}^{\infty} p_{ii}(n) = \infty$ si, y solo si, $f_{ii} = F_{ii}(1) = 1$, por lo que se demuestra el teorema

- Siendo C una clase comunicadora, cualquier estado en C es recurrente o transitorio, y, si C contiene un estado recurrente, entonces C es cerrado

- Si $i \leftrightarrow j$ y $i \neq j$, por el teorema anteriormente demostrado, solo hace falta demostrar que $\sum_{n=0}^{\infty} p_{ii}(n) = \infty$ si, y solo si, $\sum_{n=0}^{\infty} p_{jj}(n) = \infty$. Debido a que $i \leftrightarrow j$, entonces existen unas $m, n \geq 1$ tal que $\alpha = p_{ij}(m)p_{ji}(n) > 0$
- Por las ecuaciones de Chapman-Kolmogorov, se pueden obtener las siguientes desigualdades:

$$p_{ii}(m+r+n) \geq p_{ij}(m)p_{jj}(r)p_{ji}(n) = \alpha p_{jj}(r) \text{ for } r \geq 0$$

$$\Rightarrow \sum_{r=0}^{\infty} p_{ii}(m+r+n) \geq \alpha \sum_{r=0}^{\infty} p_{jj}(r)$$

- Por lo tanto, $\sum_{r=0}^{\infty} p_{ii}(r) = \infty$ cuando $\sum_{r=0}^{\infty} p_{jj}(r) = \infty$. Como el converso se puede obtener de manera similar, se demuestra la primera proposición del teorema
- Asumiendo que $i \in C$ es recurrente y C no es cerrado, entonces, por lo segundo, existe una $j \in C$ y una $k \notin C$ tal que $p_{jk} > 0$. Dado que C es una clase comunicadora (ya que i es recurrente) y $k \notin C$, entonces $k \nrightarrow j$
- Esto hace que j sea recurrente, pero se obtiene una contradicción evaluando la probabilidad de que no se vuelva al estado j , que será positiva. Consecuentemente, C es una clase cerrada

$$P(X_n \neq j \text{ for } \forall n \geq 1 | X_0 = j) \geq P(X_1 = k | X_0 = j) = p_{jk} > 0$$

- Suponiendo que el espacio de estados S es finito, entonces existe al menos un estado recurrente y, si la cadena es irreducible, entonces todos los estados son recurrentes

- Un resultado preliminar que se usará para la demostración de este teorema es el siguiente: siendo $i, j \in S$, si j es transitorio, entonces $p_{ij}(n) \rightarrow 0$ cuando $n \rightarrow \infty$
- Debido al teorema de Abel, $P_{jj}(1) = \sum_{n=0}^{\infty} p_{jj}(n) < \infty$, por lo que, por el teorema sobre la recurrencia, $P_{ij}(1) < \infty$ y, por tanto, el n -ésimo término de la suma infinita, $p_{ij}(n)$, tiende a cero cuando $n \rightarrow \infty$

- Suponiendo que $|S| < \infty$, entonces $\sum_{j \in S} P(X_n = j | X_0 = i) = \sum_{j \in S} p_{ij}(n) = 1$ (trivialmente). Si se asume que todos los estados son transitorios, entonces para toda $j \in S$ se da que $p_{ij}(n) \rightarrow 0$ cuando $n \rightarrow \infty$, lo cual sería una contradicción y eso quiere decir que al menos debe de haber un estado recurrente
 - Si la cadena es irreducible, entonces todos los estados son recurrentes o transitorios, y como debe de haber al menos un estado recurrente, eso quiere decir que todos los estados deben de ser recurrentes
- Para una cadena de Markov con espacio de estados S finito, se cumplen los siguientes hechos, los cuales se verán más claros con el subsiguiente desarrollo:
 - Una cadena de Markov se puede descomponer en una o más clases recurrentes, y también es posible descomponer teniendo en cuenta algunos estados transitorios (si hay)
 - Un estado recurrente es accesible por todos los estados de su misma clase, pero no lo es para estados recurrentes de otras clases. Un estado transitorio, en cambio, no es accesible por ningún estado recurrente
 - Al menos un estado recurrente es accesible a partir de un estado transitorio
 - Una vez una cadena entra o comienza en una clase recurrente, se queda en esa clase, dado que todos los estados serán accesibles del uno al otro y se visitarán infinitas veces
 - Si el estado inicial es transitorio, entonces la trayectoria del estado contiene una proporción inicial consistiendo de estados transitorios y una porción final consistiendo de estados recurrentes de la misma clase (entre los estados recurrentes solo)
- Un tipo de cadenas muy importantes son las cadenas absorbentes, y el análisis de este tipo de cadenas se puede hacer de manera matricial
 - Una cadena de Markov con espacio de estados S finito es absorbente si existe al menos un estado i tal que $p_{ii} = 1$ (si hay al menos un estado absorbente) y si, desde cada estado, es posible llegar al estado absorbente (no necesariamente en un paso)

- Los estados absorbentes, definidos anteriormente, son estados recurrentes en los que, una vez la cadena llega a ellos (por lo que $X_n = j$), su futura evolución está totalmente determinada

$$X_{n+1} = j, X_{n+2} = j, X_{n+3} = j, \dots$$

- Estos estados se pueden reconocer fácilmente en una matriz, dado que serán los únicos para los cuales su fila i será todo ceros menos una entrada, que será $p_{ii} = 1$

$$P = \begin{pmatrix} 1/4 & 1/2 & 1/4 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- Si la cadena absorbente tiene r estados absorbentes y t estados transitorios, se puede arreglar la matriz P de forma canónica en diferentes matrices de la siguiente manera:

$$P = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix}$$

- En este caso, Q es una matriz $t \times t$ que contiene las probabilidades de transición entre los estados transitorios, R es una matriz $t \times r$ conteniendo las probabilidades de transición de un estado transitorio a un estado absorbente, I es la matriz identidad de orden r y 0 es la matriz nula $r \times t$
- Aunque las matrices de transición no se den descompuestas de esta manera, se pueden arreglar para ponerlas en esta forma:

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \left(\begin{array}{ccc|cc} 0 & 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array} \right)$$

$$Q = \begin{pmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \quad R = \begin{pmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 1/2 \end{pmatrix}$$

- El límite de Q^n cuando $n \rightarrow \infty$ es 0 , y la cadena se absorberá eventualmente con probabilidad 1 (se visitará eventualmente un estado absorbente)

$$\lim_{n \rightarrow \infty} Q^n = 0 \Leftrightarrow \lim_{n \rightarrow \infty} p_{ij}(n) = 0 \text{ for } \forall i, j \in Q$$

- Debido a que, como los estados transitorios cumplen $p_{ij}(n) \neq 1$, entonces $p_{ij}(n) \rightarrow 0$ cuando $n \rightarrow \infty$
- Para una cadena de Markov absorbente, la matriz $I - Q$ tiene una inversa $N = (I - Q)^{-1}$, llamada matriz fundamental, la cual se define de la siguiente manera:

$$N = \sum_{n=0}^{\infty} Q^n$$

- La convergencia de la serie para N es por elementos, de modo que se puede obtener la siguiente equivalencia:

$$n_{ij} = \sum_{n=0}^{\infty} p_{ij}(n) \quad \text{for } i \in Q \text{ \& } j \in A$$

- Esta matriz es fundamentalmente diferente a la vista anteriormente para los pasos antes de llegar a un estado j si se inicia en i (la matriz para m_{ij} que se verá posteriormente), dado que en este caso $j \in A$
- Se puede demostrar que $I - Q$ tiene una inversa si se demuestra que el sistema homogéneo $(I - Q)x = 0$ tiene una solución única $x = 0$. Como $(I - Q)x = 0$ es equivalente a $x = Qx$, debido a la matriz identidad, se puede ver que $x = Qx = Q^2x = \dots = Q^n x = \dots$, pero como $Q^n x \rightarrow 0$, entonces $x = 0$ (por las igualdades)
- Si se considera $I - Q^{n+1} = (I - Q)(I + Q + Q^2 + \dots + Q^n)$, se pueden multiplicar ambos lados por $N = (I - Q)^{-1}$ para obtener $N - NQ^{n+1} = I + Q + Q^2 + \dots + Q^n$ y, usando el límite cuando $n \rightarrow \infty$, se puede obtener la siguiente serie:

$$\lim_{n \rightarrow \infty} N - NQ^{n+1} = N \Rightarrow N = I + Q + Q^2 + \dots = \sum_{n=0}^{\infty} Q^n$$

- Siendo \mathcal{T} el conjunto de estados transitorios e $i \in \mathcal{T}$, la entrada n_{ij} de la matriz N es el número esperado de pasos antes de que un estado i inicial sea absorbido por un en un estado $j \in A$
- Por lo tanto, τ_i , el número esperado de pasos hasta que la cadena sea absorbida, será la suma de las entradas n_{ij} para toda $j \in \mathcal{T}$

$$\tau_i = \sum_{j \in \mathcal{T}} n_{ij}$$

- Si $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots, \tau_t)^T$ y $\mathbf{1} = (1, 1, \dots, 1)^T$, entonces $\boldsymbol{\tau} = \mathbf{N}\mathbf{1}$
- Si $I(X_n = j)$ es la variable aleatoria indicadora de $\{X_n = j\}$, entonces $T_j = \sum_{n=0}^{\infty} I(X_n = j)$ es el número de visitas a j . Por lo tanto, usando la esperanza condicionada a $X_0 = i$, se puede obtener el siguiente resultado:

$$\begin{aligned} E(T_j | X_0 = i) &= \sum_{n=0}^{\infty} E[I(X_n = j) | X_0 = i] = \\ &= \sum_{n=0}^{\infty} P(X_n = j | X_0 = i) = \sum_{n=0}^{\infty} p_{ij}(n) = n_{ij} \end{aligned}$$

- Una vez se obtiene esta igualdad, entonces se puede utilizar una variable T tal que $T = \min\{n \geq 1: X_n \in \mathcal{A}\}$ y, por tanto, que $\tau_i = E(T | X_0 = i)$. En este caso, $\min\{n \geq 1: X_n \in \mathcal{A}\}$ es igual a la suma de T_j para $j \in \mathcal{T}$ porque el paso n en el que se llega al estado absorbente por primera vez es equivalente a la cantidad de pasos en los que se estaba en un estado transitorio (pasos hasta la absorción)

$$T = \sum_{j \in \mathcal{T}} T_j$$

$$\Rightarrow \tau_i = E(T | X_0 = i) = E\left(\sum_{j \in \mathcal{T}} T_j | X_0 = i\right) = \sum_{j \in \mathcal{T}} E(T_j | X_0 = i)$$

$$\Rightarrow \tau_i = \sum_{j \in \mathcal{T}} n_{ij}$$

- Siendo b_{ij} la probabilidad de que una cadena absorbente sea absorbida en el estado absorbente j si comienza en el estado transitorio i , si \mathbf{B} es una matriz de tamaño $t \times r$ con entradas b_{ij} , se cumple la siguiente igualdad:

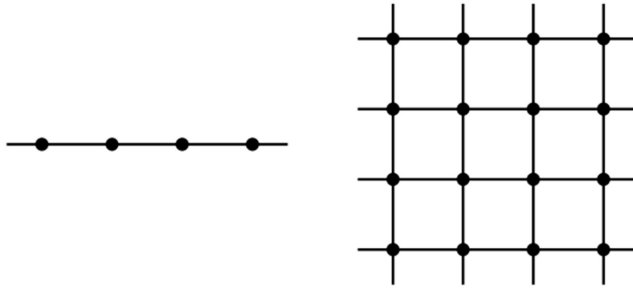
$$\mathbf{B} = \mathbf{NR}$$

- Condicionando en el primer paso, se tiene que, para cualquier $i \in \mathcal{T}$ y $j \in \mathcal{A}$, se pueden obtener los siguientes resultados:

$$b_{ij} = p_{ij} + \sum_{k \in J} p_{ik} b_{kj} \Rightarrow B = R + QB$$

$$\Rightarrow (I - Q)B = R \Rightarrow B = (I - Q)^{-1}R = NR$$

- Aunque los caminos aleatorios unidimensionales se han explicado de manera más detallada anteriormente, con la teoría de cadenas de Markov, se puede extender la teoría a mayores dimensiones dentro del contexto de las cadenas de Markov
 - Los gráficos que se considerarán serán rejillas o *lattices* d -dimensionales. Siendo \mathbb{Z}^d el conjunto de todos los vectores d -dimensionales de enteros, denotados como $x = (x_1, x_2, \dots, x_d)$ con cada $x_i \in \mathbb{Z}$



- El conjunto \mathbb{Z}^d se puede interpretar como un gráfico con un conjunto de vértices \mathbb{Z}^d y con arcos juntando cualquier par de vectores x e y , separados por una distancia euclidiana de 1. Estos vértices se declaran adyacentes y se dice que son vecinos
- El gráfico resultante también se denota por \mathbb{Z}^d , y se acentúa que cada vértice tiene exactamente $2d$ vecinos
- Siendo $d \geq 1$, un camino aleatorio en \mathbb{Z}^d es una cadena de Markov con espacio de estados \mathbb{Z}^d que, en cada paso, salta uniformemente a un vecino concreto
 - La matriz de transición se determina por las probabilidades p_{xy} , las cuales se definen de la siguiente manera:

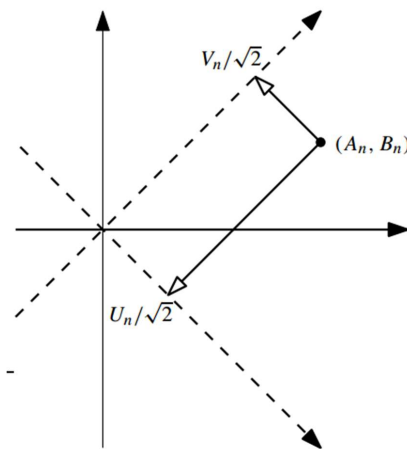
$$p_{xy} = \begin{cases} 1/2d & \text{if } y \text{ is a neighbour of } x \\ 0 & \text{otherwise} \end{cases}$$
 - Como el grafo está conectado, la cadena es irreducible. Por lo tanto, o cada estado es recurrente o cada estado es transitorio

- El camino aleatorio simétrico en \mathbb{Z}^d es recurrente si $d = 1$ o $d = 2$, pero es transitorio si $d \geq 3$

- Siendo $d = 1$ y $X_0 = 0$, el camino solo puede volver a cero...

Commented [IC41]: Acabar bien

- Escribiendo $X_n = (A_n, B_n)$ para la posición del camino en el momento n , y siendo $Y_n = (U_n, V_n)$, donde $U_n = A_n - B_n$ y $V_n = A_n + B_n$, Y_n se deriva de X_n como si fuera un sistema de coordenadas rotado y reescalado



- El hecho clave es que $U = (U_n)$ y $V = (V_n)$ son caminos aleatorios independientes y simétricos en la línea \mathbb{Z} . Realizando cálculos como el siguiente para $Y_{n+1} - Y_n$ igual a $(1,1)$, $(1,0)$, $(0,1)$ y $(0,0)$, se puede ver que U y V son caminos aleatorios simétricos

$$\begin{aligned}
 P(Y_{n+1} - Y_n = (1,1)) &= P(U_{n+1} - U_n = 1)P(V_{n+1} - V_n = 1) \\
 &= P(A_{n+1} - A_n + B_n - B_{n-1} = 1)P(A_{n+1} - A_n + B_{n+1} - B_n = 1) \\
 &= P(A_{n+1} - A_n = 1) = P(X_{n+1} - X_n = (1,0)) = 1/4
 \end{aligned}$$

- Además, hay independencia, debido a que se cumple la siguiente igualdad:

$$\begin{aligned}
 P(Y_{n+1} - Y_n = (1,1)) &= P(U_{n+1} - U_n = u)P(V_{n+1} - V_n = v) \\
 &\text{for } u, v = \pm 1
 \end{aligned}$$

- Finalmente, $X_n = 0$ si, y solo si, $Y_n = 0$, y esto ocurre si, y solo si, $U_n = 0$ y $V_n = 0$, por lo que se puede obtener la siguiente fórmula:

$$p_{00}(2n) = P(U_n = 0 | U_0 = 0)P(V_n = 0 | V_0 = 0) =$$

$$= \left[\left(\frac{1}{2} \right)^{2n} \binom{2n}{n} \right]^2$$

- Todo este argumento es inválido cuando se consideran tres o más dimensiones

- Hitting times

Commented [IC42]: Acabar bien

- ...
- La propiedad de Markov requiere que la evolución futura de la cadena sea independiente de su pasado, condicional al valor de la cadena en el presente n . Normalmente, se requiere una extensión de esta propiedad para un momento aleatorio en el tiempo n , pero esta no siempre puede ser verdad para todos los momentos, por lo que se utiliza el concepto de *stopping times*
 - Una variable aleatoria $T : \Omega \rightarrow \{0, 1, 2, \dots\} \cup \{\infty\}$ se denomina *stopping time* para una cadena X si, para toda $n \geq 0$, el evento $\{T = n\}$ se da en términos solamente de X_0, X_1, \dots, X_n
 - Esto quiere decir que un momento aleatorio T es un *stopping time* si uno puede decir si es igual a un tiempo concreto solo examinando el pasado y el presente de la cadena. Los tiempos aleatorios que "miran hacia el futuro" no son *stopping times*
 - Por ejemplo, el momento $T_i = \min\{n \geq 1 : X_n = i\}$ de retorno al estado i es un *stopping time* debido a que se puede expresar en términos de $X_1 \neq i, X_2 \neq i, \dots, X_n = i$
 - Siendo X una cadena de Markov con matriz de transición P , y siendo T un *stopping time*, dado que $T < \infty$ y que $X_T = i$, la secuencia $Y = (Y_k : k \geq 0)$ dada por $Y_k = X_{T+k}$ es una cadena de Markov con matriz de transición P y estado inicial $Y_0 = i$. Además, dado que $T < \infty$ y $X_T = i$, Y es independiente de X_0, X_1, \dots, X_{T-1}
 - A este teorema se le conoce como la propiedad fuerte de Markov
 - Siendo H un evento dado en términos de X_0, X_1, \dots, X_{T-1} , se requiere demostrar la siguiente identidad, que proviene de la independencia entre Y y H :

$$P(X_{T+1} = i_1, X_{T+2} = i_2, \dots, X_{T+n} = i_n, H | T < \infty, X_T = i) =$$

$$P(H | T < \infty, X_T = i) P(X_{T+1} = i_1, X_{T+2} = i_2, \dots, X_{T+n} = i_n)$$

- El evento $H \cap \{T = m\}$ se da en términos solo de X_0, X_1, \dots, X_m . Además, como $X_T = X_m$ cuando $T = m$, se condiciona al evento $H \cap \{T = m\} \cap \{X_m = i\}$ y se usa la propiedad de Markov en el momento m para deducir la siguiente igualdad:

$$P(X_{T+1} = i_1, X_{T+2} = i_2, \dots, X_{T+n} = i_n, H, T = m, X_T = i) =$$

$$= P(H, T = m, X_T = i) P(X_{T+1} = i_1, X_{T+2} = i_2, \dots, X_{T+n} = i_n)$$

- Sumando sobre $m = 0, 1, 2, \dots$ y dividir por $P(T < \infty, X_T = i)$, se puede obtener la identidad a demostrar
- Anteriormente se ha visto que un estado i es recurrente si, empezando desde i , la cadena vuelve a i con probabilidad 1, y es transitorio si no es recurrente. Si el estado inicial i es recurrente, la cadena siempre volverá a él y volverá infinitas veces o *infinitely often*
 - Suponiendo que $X_0 = i$ y siendo $V_i = |\{n \geq 1: X_n = i\}|$ el número de visitas subsiguientes por la cadena de Markov a i , V_i tiene una distribución geométrica de la siguiente forma, donde f_{ii} es la probabilidad de retorno $f_{ii} = P(X_n = i \text{ for some } n \geq 1 | X_0 = i)$

$$P(V_i = r | X_0 = i) = (1 - f_{ii}) f_{ii}^r \quad \text{for } r = 0, 1, 2, \dots$$

- En particular, $P(V_i = \infty | X_0 = i) = 1$ si i es recurrente (se visita un número infinito de veces), y $P(V_i < \infty | X_0 = i) = 1$ si i es transitorio (se visita solo un número finito de veces). Además, se resalta que $E(V_i | X_0 = i)$ es un número finito siempre que $f_{ii} \neq 1$

$$E(V_i | X_0 = i) = \sum_{r=0}^{\infty} r(1 - f_{ii}) f_{ii}^r = (1 - f_{ii}) \sum_{r=0}^{\infty} r f_{ii}^r = \frac{f_{ii}}{1 - f_{ii}} < \infty$$

- La probabilidad de un retorno eventual a i es $f_{ii} = P(T_i < \infty | X_0 = i)$, de modo que i es recurrente si $f_{ii} = 1$ y transitorio si $f_{ii} < 1$
- El momento de la visita número r a i es T_i^r (elevado a r), con $T_i^r = \infty$ si $V_i < r$. Debido a que los T_i^r son crecientes en r (cuanto más aumenta r , mayor es el tiempo necesario para esa visita), entonces se puede obtener la siguiente igualdad por la propiedad fuerte de Markov:

$$\begin{aligned}
P(V_i \geq r | X_0 = i) &= P(T_i^r < \infty | X_0 = i) = \\
&= P(T_i^r < \infty | T_i^{r-1} < \infty, X_0 = i) P(T_i^{r-1} < \infty | X_0 = i) = \\
&= f_{ii} P(T_i^{r-1} < \infty | X_0 = i) \text{ for } r \geq 1
\end{aligned}$$

- Por iteración al sustituir $P(T_i^{r-1} < \infty | X_0 = i)$ de la misma forma, se obtiene que $P(V_i \geq r | X_0 = i) = f_{ii}^r$. Si $r \rightarrow \infty$, entonces se puede ver que se dan los siguientes casos:

$$P(V_i \geq \infty | X_0 = i) = P(V_i = \infty | X_0 = i) = \begin{cases} 1 & \text{if } f_{ii} = 1 \\ 0 & \text{if } f_{ii} < 1 \end{cases}$$

- Para poder clasificar los estados, se definen cuatro conceptos importantes: el tiempo de recurrencia medio, la nulidad, el periodo y la ergodicidad

- El tiempo de recurrencia medio μ_i del estado i se define por μ_i se define de la siguiente manera:

$$\mu_i = E(T_i | X_0 = i) = \begin{cases} \sum_{n=1}^{\infty} n f_{ii}(n) & \text{if } i \text{ is recurrent} \\ \infty & \text{if } i \text{ is transient} \end{cases}$$

- Si i es recurrente, se clasifica como nulo si $\mu_i = \infty$ y positivo o no nulo si $\mu_i < \infty$ (la serie no tiene por qué converger siempre)
- El periodo d_i del estado i se da por $d_i = \gcd\{n: p_{ii}(n) > 0\}$ (el mayor divisor común del conjunto), y se dice que el estado i es periódico si $d_i > 1$ y aperiódico si $d_i = 1$
- Esto último ocurre siempre si $p_{ii} > 0$, y la manera más sencilla de saber la periodicidad es mirando directamente al diagrama de la cadena y buscar bucles (aperiodicidad) o usando P^n y viendo un patrón para las $p_{ii}(n) > 0$
- El estado i se denomina ergódico si es aperiódico y recurrente no nulo o positivo
- Se demostró anteriormente que la recurrencia es una propiedad de clase, y esta conclusión se puede extender. Si $i \leftrightarrow j$, entonces se cumplen las siguientes proposiciones:
 - Los estados i y j tienen el mismo periodo (todos los estados son periódicos o aperiódicos)

- El estado i es recurrente si, y solo si, j es recurrente también (todos los estados son recurrentes o transitorios)
- El estado i es recurrente positivo o no nulo si, y solo si, j es recurrente positivo o no nulo también (todos los estados son recurrentes positivos o recurrentes nulos)
- El estado i es ergódico si, y solo si, j es ergódico también (todos los estados son ergódicos o no son ergódicos)
- Por lo tanto, es posible hablar de una clase comunicadora C que puede ser recurrente, transitoria, ergódica y así (las diferentes clasificaciones que tiene)
 - Una cadena de Markov irreducible solo tiene una clase comunicadora, y por lo tanto se le pueden atribuir estos objetivos (si son apropiados) a la cadena misma (más que a la clase)
- La demostración del teorema anterior se puede realizar para cada proposición
 - Debido a que $i \leftrightarrow j$, deben existir unas $m, n \geq 1$ tal que $\alpha = p_{ij}(m)p_{ji}(n) > 0$. Por las ecuaciones de Chapman-Kolmogorov, se puede obtener la siguiente desigualdad:

$$p_{ii}(m+r+n) \geq p_{ij}(m)p_{jj}(r)p_{ji}(n) = \alpha p_{jj}(r) \text{ for } r \geq 0$$
 - En particular, $p_{ii}(m+n) \geq p_{ij}(m)p_{ji}(n) = \alpha > 0$, de modo que d_i divide $m+n$ (porque $p_{ii}(m+n) > 0$). Por lo tanto, si d_i no divide r , entonces tampoco divide $m+r+n$, implicando que $p_{ii}(m+n+r) = 0$
 - Como $p_{ii}(m+n+r) = 0$, $p_{jj}(r) = 0$ obligatoriamente, y entonces d_i no divide r . Esto quiere decir que $d_i | d_j$ y, usando el mismo argumento de manera inversa, se obtendría que $d_j | d_i$, implicando que $d_i = d_j$ (demostrando la primera proposición)
 - Suponiendo que i es recurrente no nula y C es una clase comunicadora de estados conteniendo i , C será una clase cerrada. Si $X_0 \in C$, entonces $X_n \in C$ para toda n (al ser cerrada), y la cadena es irreducible en el espacio de estados $S = C$
 - Por el teorema de existencia de las distribuciones estacionarias o invariantes, esta cadena tiene una distribución estacionaria y todo estado de C es recurrente no nulo. Si $i \leftrightarrow j$, entonces $j \in C$, por lo que j es recurrente no nulo, y por un argumento inverso,

si $j \in C$ fuera recurrente no nulo, entonces $i \in C$ también lo es (demostrando así la tercera proposición)

- La última proposición se demuestra a través de las tres anteriores, dado que si un estado $i \in C$ es aperiódico y recurrente no nulo, entonces $j \in C$ debe tener el mismo periodo y ser también recurrente no nulo, haciendo que ambos i y j sean ergódicos
- A partir de las propiedades vistas, es posible derivar un teorema de equivalencia para cadenas aperiódicas. Para una cadena de Markov finita e irreducible, las siguientes proposiciones son equivalentes:

- La cadena es aperiódica, por lo que $d_i = 1$. Esto es condición suficiente para decir que existe al menos una $p_{ii} > 0$ (un *loop*)
- La cadena no se puede descomponer en una estructura periódica, de modo que no se puede hacer una partición mutuamente excluyente tal que $p_{kt} > 0$ para $k \in S_j$ y $t \in S_{j+1}$ y d es el mayor entero por el cual una descomposición de S es posible

$$S = \bigcup_{j=0}^{d-1} S_j \text{ such that } S_i \cap S_j = \emptyset \text{ for } i \neq j$$

$$p_{kt} > 0 \text{ for } k \in S_j \text{ \& } t \in S_{j+1}$$

- Existe un número entero $l \geq 1$ y un estado j tal que $p_{ij}(l) > 0$ para toda $i \in S$. Esto quiere decir que todas las entradas de la columna j de \mathbf{P}^l son mayores a cero (el estado j se puede alcanzar a través de cualquier $i \in S$ en l pasos)
- Existe un número entero $m \geq 1$ tal que $p_{ij}(m) > 0$ para toda $i, j \in S$. Esto quiere decir que todas las entradas de \mathbf{P}^m son mayores a cero (cualquier estado j se puede alcanzar desde cualquier estado i en m pasos como mínimo)
- Si la cadena es irreducible y $j \in C$ es recurrente, entonces la probabilidad de que se llegue a un estado j para alguna $n \geq 1$ es 1, sin importar la distribución de X_0

$$P(X_n = j \text{ for some } n \geq 1) = 1$$

- Siendo i y j estados distintos y recordando que la probabilidad de paso es $f_{ij} = P(X_n = j \text{ for some } n \geq 1 | X_0 = i)$, como la cadena es irreducible, existe al menos un número entero menor

$m \geq 1$ tal que $p_{ji}(m) > 0$. Debido a que este número m es el menor, entonces la probabilidad se puede expresar de la siguiente manera:

$$p_{ji}(m) = P(X_m = i, X_r \neq j \text{ for } 1 \leq r < m | X_0 = j)$$

- Suponiendo que $X_0 = j$, $X_m = i$ y ningún retorno a j ocurre después del momento m , debido a $p_{ji}(m)$, la probabilidad condicional de que no ocurra j nunca será 1. A partir de la propiedad de Markov, por tanto, en el momento m se puede obtener la siguiente desigualdad:

$$p_{ji}(m)(1 - f_{jj}) \leq 1 - f_{jj}$$

- Si f_{jj} es recurrente, entonces $f_{jj} = 1$, de modo que $f_{ij} = 1$ para toda $i \in S$. Siendo $\lambda_i = P(X_0 = i)$ para $i \in S$ y definiendo $T_j = \inf\{n \geq 1 : X_n = j\}$ de la manera usual, se puede obtener la siguiente probabilidad condicionando en X_0 , que demuestra el teorema:

$$\begin{aligned} P(T_j < \infty) &= \sum_{i \in S} P(X_n = j \text{ for some } n \geq 1 | X_0 = i) P(X_0 = i) \\ &= \sum_{i \in S} \lambda_i f_{ij} = 1 \end{aligned}$$

- Ahora se estudia el comportamiento a largo plazo de las cadenas de Markov cuando $n \rightarrow \infty$ a través de las distribuciones estacionarias o invariantes
 - Debido a que la secuencia tiene $(X_n : n \geq 0)$ está sujeta a fluctuaciones aleatorias, normalmente no converge a ningún estado en particular
 - No obstante, la distribución de la secuencia sí converge a un equilibrio. Antes de poder ver el teorema, se necesita explorar los límites posibles
 - Cualquier límite distribucional es necesariamente invariante o estacionario bajo la evolución de la cadena
 - Siendo X una cadena de Markov con una matriz de transición P , el vector fila $\pi = (\pi_i : i \in S)$ se denomina distribución estacionario o invariante de la cadena si $\pi_i \geq 0$ para toda $i \in S$ y $\sum_{i \in S} \pi_i = 1$ y $\pi = \pi P$
 - Una distribución estacionaria o invariante es invariante al paso del tiempo: si X_0 tiene una distribución π , entonces X_n tendrá

una distribución πP^n , y $\pi P^n = \pi P P^{n-1} = \pi P^{n-1} = \dots = \pi$, de modo que toda X_n tiene una distribución π

- Debido a que $\pi = \pi P$, esto permite hacer un sistema de ecuaciones para poder obtener el vector π a partir de P . Al resolverlo, no obstante, siempre se tiene que tener en cuenta la condición de normalización $\sum_{i \in S} \pi_i = 1$ al final del sistema (es decir, una vez encontradas las igualdades o relaciones simplificadas en el sistema, se utiliza esta condición, pero no antes)

$$\begin{array}{rcl} \pi_1 & = & 0.2 \pi_1 + 0.1 \pi_2 + 0.4 \pi_3 \\ \pi_2 & = & 0.7 \pi_1 + 0.8 \pi_2 + 0.4 \pi_3 \\ \pi_3 & = & 0.1 \pi_1 + 0.1 \pi_2 + 0.2 \pi_3 \end{array} \quad \begin{array}{rcl} 0.8 \pi_1 & = & 0.1 \pi_2 + 0.4 \pi_3 \\ 0.2 \pi_2 & = & 0.7 \pi_1 + 0.4 \pi_3 \\ 0.8 \pi_3 & = & 0.1 \pi_1 + 0.1 \pi_2 \end{array}$$

$$\pi_1 + \pi_2 + \pi_3 = 1$$

$$\pi = \left(\frac{4}{27}, \frac{20}{27}, \frac{3}{27} \right) \approx (0.148, 0.741, 0.111)$$

- Si $\pi_i = 0$ para toda $i \in S$, entonces $\pi = \mathbf{0}$ y no existe una distribución estacionaria
- Considerando una cadena de Markov irreducible, existe una distribución invariante π si, y solo si, algún estado es recurrente no nulo, y, si existe una distribución invariante π , entonces todos los estados son recurrentes no nulos y $\pi_i = 1/\mu_i$ para $i \in S$. En particular, π es la única distribución invariante
 - Por lo tanto $\pi_i = 1/\mu_i$ es la fracción de tiempo esperada en el largo plazo de que el estado sea i
 - Debido a que la cadena es irreducible, entonces todos los estados serán o recurrentes o transitorios. Por lo tanto, si los estados son transitorios o recurrentes nulos, entonces $\mu_i = \infty$ y eso hace que $\pi_i = 0$
- Este teorema se demuestra exhibiendo una solución explícita a la ecuación vectorial $\rho = \rho P$. Al considerar una solución, es normal considerar que el vector ρ indica la proporción de tiempo gastado en cada uno de los estados, por lo que se fija una $k \in S$ y se comienza la cadena desde este estado
 - Siendo W_i el número de visitas subsiguientes al estado i antes del primer retorno al estado inicial k , se puede expresar de la siguiente manera, donde $T_k = \inf\{n \geq 1: X_n = k\}$

$$W_i = \sum_{m=1}^{\infty} 1(X_m = i, T_k \geq m) = \sum_{m=1}^{T_k} 1(X_m = i) \text{ for } i \in S$$

- Si $T_k < \infty$ (en algún momento se vuelve a k), $W_k = 1$, dado que solo habrá un momento en el cual $X_m = k$ cuando $T_k \geq m$
- Los elementos del vector candidato ρ se dan por $\rho_i = E(W_i | X_0 = k)$
- Para una cadena irreducible, recurrente y dado cualquier $k \in S$, el vector $\rho = (\rho_i : i \in S)$ satisface las siguientes propiedades:
 - El elemento de ρ en el estado k es $\rho_k = 1$
 - La sumatoria de los elementos de ρ es $\sum_{i \in S} \rho_i = \mu_i$, independientemente de que $\mu_i < \infty$ o no
 - El vector ρ cumple la equivalencia $\rho = \rho P$
 - El valor de cualquier elemento de ρ está en el intervalo $0 < \rho_i < \infty$ para $i \in S$
- La demostración de esta última proposición funciona de la siguiente manera:
 - ...
- Un lema necesario para la demostración del teorema es el siguiente: siendo $\lambda = (\lambda_i : i \in S)$ una distribución en el conjunto numerable S , haciendo que $\alpha_i(n)$ cumpla $|\alpha_i(n)| \leq M < \infty$ para toda $i \in S$ y toda n , y asumiendo que $\alpha_i(n) \rightarrow 0$ cuando $n \rightarrow \infty$, se cumple la siguiente convergencia:

$$\sum_{i \in S} \lambda_i \alpha_i(n) \rightarrow 0 \text{ as } n \rightarrow \infty$$

- Siendo F un subconjunto finito de S , se pueden obtener los siguientes resultados:

$$\sum_{i \in S} |\lambda_i \alpha_i(n)| = \sum_{i \in S} \lambda_i |\alpha_i(n)| \leq \sum_{i \in F} \lambda_i |\alpha_i(n)| + M \sum_{i \notin F} \lambda_i$$

$$\Rightarrow \sum_{i \in F} \lambda_i |\alpha_i(n)| + M \sum_{i \notin F} \lambda_i \rightarrow M \sum_{i \notin F} \lambda_i \text{ as } n \rightarrow \infty$$

Commented [IC43]: Por qué

Commented [IC44]: Acabar bien

as F is finite

$$\Rightarrow \lim_{F \rightarrow S^-} M \sum_{i \notin F} \lambda_i = 0 \text{ as } \sum_{i \in S} \lambda_i < \infty$$

- A partir de esta proposición y del lema, es posible obtener la demostración del teorema sobre las probabilidades de la distribución invariante π

▪ ...

Commented [IC45]: Acabar bien

- Un corolario del teorema es el siguiente: siendo $i, k \in S$ estados distintos de una cadena de Markov irreducible y recurrente no nula con una distribución invariante π , el número medio de visitas al estado i entre dos visitas consecutivas a k (entre una visita a k y la siguiente más próxima) es equivalente a π_i / π_k
 - Como la cadena es no nula, se tiene que $\mu_k < \infty$, y por el teorema sobre distribuciones invariantes, se sabe que existe una única distribución invariante π , y que $\pi_i = 1/\mu_i$
 - Utilizando la proposición anterior, definiendo $v = \pi_k \rho$, se satisface $v = vP$, dado que $\rho = \rho P$ implica $v = \pi_k \rho = \pi_k \rho P = vP$, y además $\sum_{i \in S} v_i = 1$
 - Debido a la unicidad de la distribución invariante, es necesario que $\pi = v$, y, por lo tanto, $\rho_i = \frac{v_i}{\pi_k} = \frac{\pi_i}{\pi_k}$
- Siendo $T_i = \min\{n \geq 1 : X_n = i\}$ el primer tiempo de paso a i , si $X_0 = i$, entonces T_i es el tiempo de recurrencia de i , con media $\mu_i = E(T_i | X_0 = i)$. A partir de esto, se puede plantear un sistema de ecuaciones que permitirá obtener μ_i , lo cual se conoce como *first step analysis*
 - Siendo m_{ij} el número esperado de pasos hasta que la cadena visita un estado j por primera vez dado $X_0 = i$, entonces $m_{ii} = 0$ y para $i \neq j$ se tiene $m_{ij} = E(T_j | X_0 = i)$. Si se pasa de i a j , entonces $T_j = 1$, pero si se pasa a un estado intermedio k , entonces el valor de T_{ij} esperado sería $\sum_{k \neq i} p_{ik} m_{ki}$ (se multiplica la probabilidad por el número esperado de pasos en ese estado intermedio k)

$$m_{ij} = \sum_{k \in S} P(T_j | X_0 = i, X_1 = k) P(X_1 = k | X_0 = i)$$

$$\Rightarrow m_{ij} = 1 + \sum_{k \neq j} p_{ik} m_{kj} \text{ for } i, j \in S \text{ and } i \neq j$$

$$\Rightarrow m_{ii} = \mu_i = 1 + \sum_{k \neq i} p_{ik} m_{ki} \text{ for } i \in S$$

- Una vez hecho esto, se puede plantear un sistema de ecuaciones para m_{ij} (donde $i \neq j$) para una j concreta, obteniendo una solución única de $|S|$ ecuaciones e incógnitas, teniendo así una solución única

if $|S| = 3$ & $j = 0$:

$$\begin{cases} m_{00} = 1 + \sum_{k \neq 0} p_{0k} m_{k0} \\ m_{10} = 1 + \sum_{k \neq 1} p_{1k} m_{k0} \\ m_{20} = 1 + \sum_{k \neq 2} p_{2k} m_{k0} \end{cases} \Rightarrow \begin{cases} m_{00} = 1 + p_{01} m_{10} + p_{02} m_{20} \\ m_{10} = 1 + p_{10} m_{00} + p_{12} m_{20} \\ m_{20} = 1 + p_{20} m_{00} + p_{21} m_{10} \end{cases}$$

- Este sistema de ecuaciones permitirá evaluar la recurrencia nula o no nula de una matriz, dado que, para un estado recurrente, se comparará $m_{ii} = \mu_i$ con ∞
- Una propiedad muy útil para simplificar el sistema es la de simetría: si en π hay algunos elementos que son equivalentes (por ejemplo $\pi_i = \pi_j$ para algún par de $i \neq j$), entonces se pueden eliminar todas menos una de las ecuaciones de estos elementos para m_{ij} (dado que se puede obtener igualmente para los otros elementos por equivalencia)

$$\pi_1 = \pi_2 = \pi_3 \text{ & } \pi_4$$

$$\Rightarrow m_{41} = m_{42} = m_{43} ; m_{14} = m_{24} = m_{34}$$

$$\Rightarrow m_{12} = m_{23} = m_{31} ; m_{21} = m_{32} = m_{13}$$

- Considerando una cadena de Markov finita con $|S|$ estados, se define \mathbf{M} como una matriz cuadrada $|S| \times |S|$ tal que $\mathbf{M} = (m_{ij})_{i,j \in S}$ con $m_{ii} = 0$, $\mathbf{D} = (\mu_1, \mu_2, \dots, \mu_{|S|})$ como una matriz diagonal $|S| \times |S|$ y \mathbf{C} como una matriz cuadrada $|S| \times |S|$ con todas sus entradas iguales a 1

- Las ecuaciones anteriores relacionando el primer paso medio a un estado y los tiempos de recurrencia medios se pueden escribir como $(\mathbf{I} - \mathbf{P})\mathbf{M} = \mathbf{C} - \mathbf{D}$

- Si la cadena es irreducible y tiene una distribución estacionaria (se introducirá en las siguientes secciones) π , como $\pi = \pi P$, entonces $\pi(I - P) = 0$. Por lo tanto, se obtienen los siguientes resultados:

$$0 = \pi(I - P)M = \pi(C - D) = \pi C - \pi D$$

$$\Rightarrow \pi C = (1, 1, \dots, 1) \text{ and } \pi D = (\pi_1 \mu_1, \pi_2 \mu_2, \dots, \pi_{|S|} \mu_{|S|})$$

$$\Rightarrow \pi_i = \frac{1}{\mu_i} \text{ for all } \mu_i > 0$$

- Definiendo Π como una matriz $|S| \times |S|$ cuyas filas son los vectores π , entonces se obtienen los siguientes resultados:
 - Para una cadena de Markov finita e irreducible, la matriz $I - P + \Pi$ tiene una matriz inversa llamada matriz fundamental de la cadena Z
 - Los primeros pasos medios m_{ij} para una cadena finita e irreducible se determinan por la matriz fundamental Z y la distribución estacionaria π de la siguiente manera:

$$m_{ij} = \frac{z_{jj} - z_{ij}}{\pi_j}$$

- El principal resultado de las cadenas de Markov es que, bajo condiciones razonables, su distribución converge a una única distribución invariante
 - Considerando una cadena de Markov que es aperiódica, irreducible y recurrente positiva, para $i, j \in S$, se obtiene la siguiente convergencia, en donde π es la única distribución invariante de la cadena

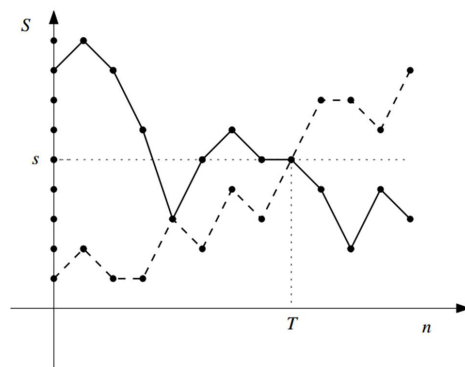
$$p_{ij}(n) \rightarrow \pi_j \text{ as } n \rightarrow \infty$$

- Este teorema se denomina teorema de convergencia de las cadenas de Markov, y se puede apreciar como el límite es independiente del inicio $i \in S$
- La demostración de este teorema utiliza una técnica importante llamada *coupling* o acoplamiento, que se basa en construir un par ordenado $Z = (X, Y)$ de cadenas de Markov independientes $X = (X_n: n \geq 0)$ e $Y = (Y_n: n \geq 0)$ las cuales tienen el mismo espacio de estado S y la misma matriz de transición P . Por lo tanto, $Z = (Z_n: n \geq 0)$ se da por $Z_n = (X_n, Y_n)$ y es fácil

comprobar que Z es una cadena de Markov con espacio de estado $S \times S$ y probabilidades de transición $p_{ik}p_{jl}$

$$\begin{aligned} p_{ij,kl} &= P(Z_{n+1} = (k,l) | Z_n = (i,j)) = \\ &= P(X_{n+1} = k | X_n = i) P(Y_{n+1} = l | Y_n = j) = p_{ik}p_{jl} \end{aligned}$$

- Debido a que X es irreducible y aperiódica, para $i, j, k, l \in S$, existe una $N = N(i, j, k, l)$ tal que $p_{ik}(n)p_{jl}(n) > 0$ para $n \geq N$ (como $i \leftrightarrow k$ y $j \leftrightarrow l$, entonces i, j, k, l deben ser aperiódicos porque las cadenas son irreducibles, haciendo que los estados sean aperiódicos en cada espacio, y además la n es la misma, por la que debe existir una N común para la cual se cumple la condición). Esto hace que Z sea irreducible, demostrado solo por aperiodicidad de las cadenas acopladas
- Suponiendo que X es recurrente, X tendrá una única distribución π ...



Commented [IC46]: Acabar bien

- Un criterio para la transitoriedad o la recurrencia se presentó anteriormente, pero también se puede presentar una para la recurrencia nula
 - Siendo X una cadena de Markov irreducible y recurrente, la existencia de un estado i tal que $p_{ii}(n) \rightarrow 0$ cuando $n \rightarrow \infty$ y que todos los estados de la cadena sean recurrentes nulos son proposiciones equivalentes
 - Como una aplicación, se puede considerar un camino aleatorio simétrico en los gráficos \mathbb{Z} y \mathbb{Z}^2 . En este caso, $p_{0,0}(n) \rightarrow 0$ cuando $n \rightarrow \infty$, por lo que se deduce que los caminos aleatorios unidimensionales y bidimensionales son recurrentes nulos

- Que la primera proposición implique la segunda se puede demostrar considerando una cadena de Markov X recurrente no nula y aperiódica, de modo que $p_{ii}(n) \rightarrow 1/\mu_i > 0$ cuando $n \rightarrow \infty$ (debido a que $p_{ii}(n) \rightarrow \pi_i$ y $\pi_i \rightarrow 1/\mu_i > 0$), haciendo que no se cumpla la primera proposición
- El mismo argumento se puede usar en el caso periódico si se considera una cadena $Y_n = X_{nd}$, donde d es la periodicidad de la cadena

$$p_{ii}(nd) = P(Y_n = i | Y_0 = i) \rightarrow \frac{d}{\mu_i} \text{ as } n \rightarrow \infty$$

- Como la primera proposición no se cumple en ninguno de los casos si $p_{ii}(n) \rightarrow 0$ cuando $n \rightarrow \infty$, por lo que se demuestra que la primera proposición implica la segunda (tiene que ser que la cadena sea recurrente nula si se da la primera proposición)
- Se puede discutir cual es la proporción de veces en que la cadena de Markov está en un estado concreto en el largo plazo. Siendo $i \in S$ y definiendo $V_i(n) = \sum_{k=1}^n 1(X_k = i)$ como el número de visitar a i hasta el momento n , se da el siguiente teorema:

- Siendo $i \in S$, si la cadena es irreducible y recurrente no nula, entonces se produce la siguiente convergencia en distribución, independientemente de la distribución inicial de la cadena:

$$\frac{1}{n} V_i(n) \xrightarrow{D} \frac{1}{\mu_i}$$

- Hay varios modos de convergencia de variables aleatorias, pero se ha escogido la convergencia en distribución por simplicidad (aunque es equivalente a la convergencia en probabilidad en este caso). Resultados más fuertes son válidos, y se apoyan en la ley fuerte de grandes números
- La demostración de este teorema se puede realizar mediante funciones características
- Una observación importante en física es que muchas ecuaciones son válidas sin tener en cuenta si el flujo del tiempo es hacia adelante o hacia atrás. La invarianza bajo una reversión del tiempo es una propiedad importante de algunas cadenas de Markov
 - Se supone una cadena de Markov $X = (X_n : 0 \leq n \leq N)$ irreducible y recurrente no nula con matriz de transición P y distribución estacionaria π

- Además, se supone que X_0 tiene una distribución π , de modo que X_n tiene una distribución π para toda n
- La cadena revertida $Y = (Y_n : 0 \leq n \leq N)$ viene dada por el tiempo de reversión $Y_n = X_{N-n}$ para $0 \leq n \leq N$
- Debido al teorema visto anteriormente, $\pi_i > 0$ para $i \in S$
- La secuencia Y es una cadena de Markov irreducible con matriz de transición $\hat{P} = (\hat{p}_{ij} : i, j \in S)$ dado por $\hat{p}_{ij} = (\pi_j / \pi_i) p_{ji}$ para $i, j \in S$ y con la distribución invariante π

- Primero es necesario comprobar que \hat{P} es una matriz estocástica. Sus entradas no son negativas, y, además, se puede ver que suman 1 por fila, dado que $\pi = \pi P$

$$\sum_{j \in S} \hat{p}_{ij} = \frac{1}{\pi_i} \sum_{j \in S} \pi_j p_{ji} = \frac{1}{\pi_i} \pi_i = 1$$

- Además, π es estacionaria para \hat{P} a través de la ecuación del teorema, dado que las filas de P suman 1. Estas ecuaciones se denominan ecuaciones de balance

$$\sum_{i \in S} \pi_i \hat{p}_{ij} = \sum_{j \in S} \pi_j p_{ji} = \pi_j$$

- Ahora, para demostrar que la cadena de Markov tiene distribución inicial π y matriz de transición \hat{P} , se obtiene el siguiente resultado:

$$\begin{aligned} P(Y_0 = i_0, Y_1 = i_1, \dots, Y_n = i_n) &= \\ = P(X_{N-n} = i_n, X_{N-n+1} = i_{n-1}, \dots, X_N = i_0) &= \\ = \pi_{i_n} p_{i_n, i_{n-1}} \dots p_{i_1, i_0} = \pi_{i_0} \hat{p}_{i_0, i_1} \dots \hat{p}_{i_{n-1}, i_n} \end{aligned}$$

- Por el mismo teorema referido, Y tiene matriz de transición \hat{P} y distribución inicial π
- Se dice que la cadena Y es una reversión temporal de la cadena X , y se dice que X es reversible si su reversión temporal tiene las mismas probabilidades de transición
- Siendo $X = (X_n : 0 \leq n \leq N)$ una cadena de Markov irreducible tal que X_0 tiene una distribución invariante π , a

cadena es reversible si X y su reversión temporal Y tienen las mismas matrices de transición, lo cual se expresa en la siguiente identidad:

$$P = \hat{P} \Rightarrow p_{ij} = \frac{\pi_j}{\pi_i} p_{ji} \text{ for } i, j \in S$$

$$\Rightarrow \pi_i p_{ij} = \pi_j p_{ji} \text{ for } i, j \in S$$

- Estas ecuaciones se denominan las ecuaciones de balance detallado, y son pivotaes para el estudio de las reversiones temporales. Más generalmente, se dice que la matriz de transición P y la distribución λ están en el balance detallado si se cumple la siguiente condición:

$$\lambda_i p_{ij} = \lambda_j p_{ji} \text{ for } i, j \in S$$

- Una cadena irreducible X con distribución invariante π se llama reversible en equilibrio si la matriz de transición P está en balance detallado con π
- Ocurre que, para una cadena irreducible, P está en balance detallado con una distribución λ si, y solo si, λ es la única distribución invariante. Esto proporciona una buena manera de encontrar una distribución invariante para la cadena reversible
- Siendo P la matriz de transición de una cadena irreducible X y supón que π es una distribución satisfaciendo $\pi_i p_{ij} = \pi_j p_{ji}$ para $i, j \in S$, entonces π es la única distribución invariante de la cadena. Además, X es reversible en equilibrio
- Suponiendo que π es una distribución que cumpla la identidad anterior, se puede obtener la siguiente igualdad debido a que P tiene sumas de filas 1:

$$\sum_{i \in S} \pi_i p_{ij} = \sum_{i \in S} \pi_j p_{ji} = \pi_j \sum_{i \in S} p_{ji} = \pi_j$$

- Entonces, $\pi = \pi P$ cuando π es invariante. La reversibilidad en equilibrio de X se deriva de la definición anterior
- Siendo P una matriz doblemente estocástica de una cadena de Markov irreducible y con espacio de estados S finito con $|S|$ estados, entonces la distribución de probabilidad invariante π viene dada por las siguientes probabilidades:

$$\pi_j = \frac{1}{|S|} \text{ for } j = 1, 2, \dots, |S|$$

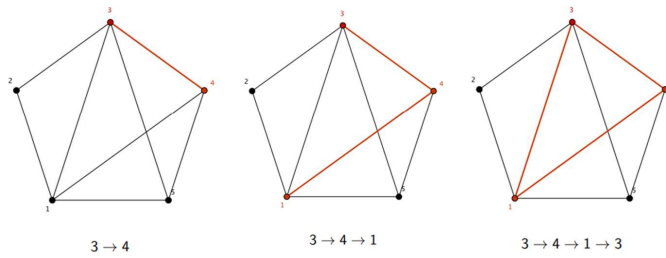
- Se tiene que $\sum_{j=1}^{|S|} \pi_j = \sum_{j=1}^{|S|} 1/|S| = 1$. Además, las ecuaciones de balance se satisfacen, por lo que se demuestra el teorema

$$\sum_{i=1}^{|S|} \pi_i p_{ij} = \sum_{i=1}^{|S|} \frac{1}{|S|} p_{ij} = \frac{1}{|S|} \sum_{i=1}^{|S|} p_{ij} = \frac{1}{|S|} = \pi_j$$

- La discusión anterior sobre la reversibilidad se restringe únicamente a cadenas de Markov irreducibles y recurrentes no nulas con puntos de tiempo numerables y finitos $(0, 1, 2, \dots, N)$
 - Una cadena de este tipo en el conjunto de tiempo singular infinito $0, 1, 2, \dots$ se llama reversible si las subsecuencias finitas X_0, X_1, \dots, X_N son reversibles para toda $N \geq 0$. La discusión también se puede extender al conjunto de tiempo doblemente infinito $\dots, -1, 0, 1, \dots$ a la suposición de que X_n tiene una distribución invariante π para n
- La reversibilidad temporal es un concepto útil en la teoría de redes aleatorias, y hay una analogía útil usando el lenguaje de los flujos. Siendo X una cadena de Markov con espacio de estado S y distribución invariante π , a esta cadena le corresponde la siguiente red dirigida:
 - Los vértices de la red son los estados de la cadena, y una flecha se coloca del vértice i al j si $p_{ij} > 0$
 - Una unidad de material nocional (en este caso, de probabilidad) se distribuye alrededor de los vértices y se permite que fluya a través de las flechas
 - Una proporción π_i del material se coloca inicialmente en el vértice i . En cada momento del tiempo y para cada vértice i , una proporción p_{ij} del material en i se transporta a cada vértice j
- Es inmediato que la cantidad de material en el vértice i después de un momento es $\sum_j \pi_j p_{ji} = \pi_i$, dado que $\pi = \pi P$. Por lo tanto, el flujo determinístico de probabilidad está en equilibrio: hay un “balance global” en el sentido de que la cantidad total dejando cada vértice se balancea por una cantidad igual al llegar ahí
 - Puede que no haya un “balance local” en el sentido de que, para toda $i, j \in S$, la cantidad fluyendo de i a j iguala a la cantidad fluyendo de j a i . El balance local ocurre si, y solo si, $\pi_i p_{ij} =$

$\pi_j p_{ji}$ para $i, j \in S$, lo cual quiere decir que \mathbf{P} y $\boldsymbol{\pi}$ están en balance detallado

- Los gráficos de rejillas \mathbb{Z}^d vistos para los caminos aleatorios se pueden entender como grafos infinitos, de modo que se puede usar la teoría de grafos para estudiarlas



- Un grafo $G = (V, E)$ es un conjunto de V vértices, pares los cuales están juntos por ejes o arcos. Por lo tanto, el conjunto de arcos o ejes E es un conjunto de pares desordenados distintos $\langle u, v \rangle$ de elementos distintos de V
 - Un grafo es conexo o está conectado si, para un par distinto de $u, v \in V$, existe un camino de arcos de u a v . Se escribe $u \sim v$ si $\langle u, v \rangle \in E$, y en este caso se dice que u y v son vecinos
 - El grado $d(v)$ del vértice v es el número de ejes o arcos conteniendo v , por lo que $d(v) = |\{u \in V : v \sim u\}|$
- Siendo $G = (V, E)$ un grafo conectado con $d(v) < \infty$ para toda $v \in V$, una partícula se mueve por los vértices de G , tomando pasos a lo largo de los arcos
 - Siendo X_n la posición de la partícula en el momento n , en el momento $n + 1$, la partícula se mueve a un vecino aleatorio uniforme de X_n (cada vecino tiene la misma probabilidad)
 - De manera más precisa, un camino aleatorio es la cadena de Markov $\mathbf{X} = (X_n : n \geq 0)$ con espacio de estado V y matriz de transición determinada por las siguientes probabilidades:

$$p_{uv} = \begin{cases} \frac{1}{d(u)} & \text{if } v \sim u \\ 0 & \text{otherwise} \end{cases}$$

- Cuando G es infinito, la pregunta está en entender el comportamiento a largo plazo del camino, refiriéndose a ver si es recurrente o transitorio. En este apartado, se considera un grafo finito conexo G

- Es útil notar que $\sum_{v \in V} d(v) = 2|E|$, dado que cada eje o arco contribuye 2 a la suma
- El camino aleatorio en el grado conexo finito $G = (V, E)$ es una cadena de Markov irreducible con una distribución única invariante con probabilidades π_v , y la cadena es reversible en equilibrio

$$\pi_v = \frac{d(v)}{\sum_{v \in V} d(v)} = \frac{d(v)}{2|E|} \quad \text{for } v \in V$$

- Dado que G es conectado, la cadena es irreducible, y el vector π es una distribución ya que $\pi_v \geq 0$ para $v \in V$, y $\sum_{v \in V} \pi_v = 1$ por la fórmula para π_v mostrada
- Por las ecuaciones de balance detallado, se puede obtener la siguiente identidad, la cual se mantiene por la definición de las probabilidades de transición para el camino aleatorio

$$\frac{d(u)}{2|E|} p_{uv} = \frac{d(v)}{2|E|} p_{vu} \quad \text{for } u, v \in V$$

- A partir del grafo, es posible obtener una matriz de transición \mathbf{P} , la cual permite poder llevar a cabo un análisis del camino aleatorio como una cadena de Markov

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \quad \mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}$$

- Este análisis se vuelve complicado dependiendo de la cardinalidad del espacio de estados S y de los arcos por nodo (que puede variar de estado a estado, por ejemplo)
- Sin embargo, para análisis de caminos aleatorios en n pasos, cuando n es un número pequeño, y para caminos aleatorios con evolución idéntica (con el mismo número de arcos por nodo), el análisis se vuelve sencillo
- Ahora que la teoría sobre las cadenas de Markov se ha desarrollado de manera detallada, se pueden ver algunos ejemplos interesantes de analizar utilizando los diferentes resultados

- El problema de la ruina del apostador o *Gambler's Ruin* se basa en considerar una cadena de Markov con espacio de estados $S = \{0, 1, \dots, m\}$ con las siguientes probabilidades de transición:

$$p_{ij} = \begin{cases} 1 & \text{for } i = j = 0 \\ p & \text{for } j = i + 1, 1 \leq i \leq m - 1 \\ q & \text{for } j = i - 1, 1 \leq i \leq m - 1 \\ 1 & \text{for } i = j = m \end{cases}$$

where $p \in (0, 1)$ & $q = 1 - p$

- En este caso, si se está en el estado 0 o m , la cadena se absorbe en estos. Para $1 \leq i \leq m - 1$, la probabilidad de ir al siguiente estado $j + 1$ es p , mientras que la de ir hacia atrás al estado $j - 1$ es q
- En este problema, lo interesante es calcular la probabilidad de ser absorbido en 0 y en m y calcular el tiempo esperado hasta la absorción (usando el caso simétrico $p = q$)
- Denotando la probabilidad de ser absorbido en m como $v(j) = P(X_n = m | X_0 = j)$, se pueden plantear las siguientes igualdades:

$$v(0) = 0 \quad \& \quad v(m) = 1$$

$$v(j) = pv(j + 1) + qv(j - 1) \quad \text{for } 1 \leq j \leq m - 1$$

- La ecuación anterior se puede expresar de manera equivalente para poder obtener los siguientes resultados:

$$v(j) = pv(j + 1) + qv(j - 1)$$

$$\Rightarrow pv(j) + qv(j) = pv(j + 1) + qv(j - 1)$$

$$\Rightarrow p[v(j + 1) - v(j)] = q[v(j) - v(j - 1)]$$

$$\Rightarrow v(j + 1) - v(j) = \left(\frac{q}{p}\right) [v(j) - v(j - 1)] =$$

$$= \left(\frac{q}{p}\right)^2 [v(j - 1) - v(j - 2)] = \dots = \left(\frac{q}{p}\right)^j [v(1) - v(0)] =$$

$$= \left(\frac{q}{p}\right)^j v(1)$$

- Debido a que $v(m) - v(0) = 1$, se puede utilizar la siguiente igualdad para obtener la probabilidad $v(1)$ y $v(j)$

$$1 = v(m) - v(0) = \sum_{j=0}^{m-1} [v(j+1) - v(j)] = v(1) \sum_{j=0}^{m-1} \left(\frac{q}{p}\right)^j$$

$$\Rightarrow 1 = \begin{cases} v(1)m & \text{for } p = q \\ v(1) \frac{1 - (q/p)^m}{1 - q/p} & \text{for } p \neq q \end{cases}$$

$$\Rightarrow v(1) = \begin{cases} 1/m & \text{for } p = q \\ \frac{1 - q/p}{1 - (q/p)^m} & \text{for } p \neq q \end{cases}$$

$$v(m) - v(0) = \sum_{j=0}^{m-1} [v(j+1) - v(j)]$$

$$\Rightarrow v(j) = \sum_{k=0}^{j-1} [v(k+1) - v(k)] = v(1) \sum_{k=0}^{j-1} \left(\frac{q}{p}\right)^k$$

$$\Rightarrow v(j) = \begin{cases} j/m & \text{for } p = q \\ \frac{1 - (q/p)^j}{1 - (q/p)^m} & \text{for } p \neq q \end{cases}$$

- Debido a que $v(m) - v(0) = 1$, entonces $1 - v(j)$ es la probabilidad de ser absorbido en el estado 0, y se puede obtener simplemente a partir de $v(j)$

$$1 - v(j) = \begin{cases} 1 - j/m & \text{for } p = q \\ \frac{(q/p)^m - (q/p)^j}{1 - (q/p)^m} & \text{for } p \neq q \end{cases}$$

- Siendo $T = \min\{n \geq 0: X_n = 0 \text{ or } X_n = m\}$ y definiendo $\tau_i = E(T|X_0 = i)$, se puede realizar un análisis del primer paso para obtener las siguientes igualdades:

$$\tau(1) = p\tau(m-1) + q$$

$$\Rightarrow \tau(2) = \frac{1}{2}[\tau(m-2) + 1] + \frac{1}{2}[\tau(1) + 1]$$

$$\Rightarrow \dots$$

$$\Rightarrow \tau(i) = \frac{1}{2}[\tau(i+1) + 1] + \frac{1}{2}[\tau(i-1) + 1]$$

$$\text{for } 1 \leq i \leq m-1$$

$$\tau(0) = \tau(m) = 0$$

- Definiendo $d(i) = \tau(i) - \tau(i-1)$, entonces se pueden obtener los siguientes resultados:

$$d(i+1) = d(i) - 2 \quad \text{for } i = 1, 2, \dots, m-1$$

$$\Rightarrow d(2) = d(1) - 2$$

$$\Rightarrow \dots$$

$$\Rightarrow d(m) = d(m-1) - 2(m-1)$$

$$0 = \tau(m) - \tau(0) = d(m) + \dots + d(1)$$

$$= md(1) - 2 \sum_{j=1}^m (j-1) = md(1) - 2 \frac{m(m-1)}{2} =$$

$$= m[d(1) - (m-1)] \Rightarrow d(1) = m-1$$

$$\Rightarrow \tau(i) = \sum_{n=1}^{i-1} d(n) = \frac{[(m-1)(m-2i+1)](i-1+1)}{2} =$$

$$= i(m-i)$$

- El problema del camino aleatorio con barreras considera una cadena de Markov con espacio de estados $S = \{0, 1, \dots, m\}$ y con probabilidades de transición definidas de la siguiente manera:

$$p_{ij} = \begin{cases} 1-d & \text{for } i=j=0 \\ b & \text{for } j=i+1, 1 \leq i \leq m-1 \\ d & \text{for } j=i-1, 1 \leq i \leq m-1 \\ 1-b-d & \text{for } 1 \leq j=i \leq m-1 \\ 1-b & \text{for } i=j=m \end{cases}$$

$$\text{where } b, d > 0 \text{ \& } b+d \leq 1$$

- En este caso, la cadena es finita e irreducible, dado que $i \leftrightarrow j$ para $i, j \in S$ (todos los estados están conectados y no hay estados absorbentes), por lo que es recurrente no nula
- En este problema, lo interesante es obtener las probabilidades de la distribución estacionaria π
 - En esta ecuación, las ecuaciones de balance detallado se cumplen debido a que la cadena es irreducible y es recurrente no nula, de modo que la existencia de π está garantizada. Por lo tanto, se pueden obtener los siguientes resultados:

$$\pi_i b = \pi_{i+1} d \quad \text{for } 0 \leq i \leq m-1$$

$$\rho = \frac{b}{d} \Rightarrow \pi_i \rho = \pi_{i+1} \quad \text{for } 0 \leq i \leq m-1$$

$$\Rightarrow \pi_i = \pi_0 \rho^i \quad \text{for } 0 \leq i \leq m$$

$$\sum_{i=0}^m \pi_i = 1 \Rightarrow 1 = \pi_0 \sum_{i=0}^m \rho^i = \begin{cases} \pi_0(m+1) & \text{for } \rho = 1 \\ \pi_0 \frac{1-\rho^{m+1}}{1-\rho} & \text{for } \rho \neq 1 \end{cases}$$

$$\Rightarrow \pi_i = \begin{cases} \frac{1}{m+1} & \text{for } \rho = 1 \\ \frac{\rho^i(1-\rho)}{1-\rho^{m+1}} & \text{for } \rho \neq 1 \end{cases} \quad \text{for } 0 \leq i \leq m$$

- Por lo tanto, si $\rho = 1$, entonces la matriz de probabilidades de transición es doblemente estocástica y se obtiene una distribución uniforme. Si $\rho \neq 1$, entonces se pueden estudiar los casos en los que $m \rightarrow \infty$

$$\rho < 1 \Rightarrow \lim_{m \rightarrow \infty} \pi_i = \lim_{m \rightarrow \infty} \frac{\rho^i(1-\rho)}{1-\rho^{m+1}} = \rho^i(1-\rho)$$

$$\rho > 1 \Rightarrow \lim_{m \rightarrow \infty} \pi_i = \lim_{m \rightarrow \infty} \frac{\rho^i(1-\rho)}{1-\rho^{m+1}} = 0$$

- El problema del proceso de nacimiento y muerte o *birth and death process* es similar al del camino aleatorio con barreras, y este considera una cadena de Markov con espacio de estados $S = \mathbb{N}$ y con probabilidades de transición definidas de la siguiente manera:

$$p_{ij} = \begin{cases} 1-b & \text{for } j=i=0 \\ b & \text{for } j=i+1, i \geq 1 \\ d & \text{for } j=i-1, i \geq 1 \\ 1-b-d & \text{for } j=i, i \geq 1 \end{cases}$$

where $b, d > 0$ & $b + d \leq 1$

- En este caso, la cadena no es finita, pero si es aperiódica e irreducible, porque $p_{00} > 1$
- En este problema, lo interesante es obtener las probabilidades de la distribución estacionaria π
- En esta ecuación, las ecuaciones de balance detallado se cumplen debido a que la cadena es irreducible y es recurrente no nula, de modo que la existencia de π está garantizada. Por lo tanto, se pueden obtener los siguientes resultados:

$$\pi_i b = \pi_{i+1} d \quad \text{for } 0 \leq i \leq m-1$$

$$\rho = \frac{b}{d} \Rightarrow \pi_i \rho = \pi_{i+1} \quad \text{for } 0 \leq i \leq m-1$$

$$\Rightarrow \pi_i = \pi_0 \rho^i \quad \text{for } 0 \leq i \leq m$$

$$\sum_{i=0}^{\infty} \pi_i = 1 \Rightarrow 1 = \pi_0 \sum_{i=0}^{\infty} \rho^i = \begin{cases} \infty & \text{for } \rho \geq 1 \\ \frac{\pi_0}{1-\rho} & \text{for } \rho < 1 \end{cases}$$

$$\Rightarrow \pi_i = \begin{cases} 0 \text{ or } \infty & \text{for } \rho \geq 1 \\ (1-\rho)\rho^i & \text{for } \rho < 1 \end{cases} \quad \text{for } 0 \leq i \leq m$$

$$(\rho \geq 1) \text{ if } \pi_0 = 0 \Rightarrow \pi_i = 0 \quad \text{for } i \geq 0$$

$$(\rho \geq 1) \text{ if } \pi_1 > 0 \Rightarrow \pi_i = \infty \quad \text{for } i \geq 0$$

Las cadenas de Markov en tiempo continuo

Las martingalas

- Chap 7: Martingales