Di Tella University - 2024

Lecture Notes - PROBA

Sesto Francisco

3 de octubre de 2024

Índice

Probabilidad y conteo

Paradigmas de la probabilidad

Paradigma frecuentista

Se le **atribuye probabilidad sólo a resultados de circunstancias azarosas**. Se conceptualiza la posibilidad de repeticiones infinitas de idénticas circunstancias azarosas. La probabilidad de un resultado **se intepreta como la frecuencia relativa con la que ocurriría en las hipotéticas infinitas repeticiones**.

Si $P(evento A) \in (0,1)$ es por que el evento A puede tener distintos resultados dependientdo de circunstancias azarozas. Pero la probabilidad en un caso concreto es 0 o 1.

Paradigma bayesiano

La probabilidad **se entiende como una medida de incertidumbre acerca de la veracidad de cualquier proposición**. No requiere de la existencia de un proceso aleatorio. Hay dos grandes vertientes filosóficas:

- **Interpersonal o lógica** (JM Keynes, Rudolf Camap y H Jeffreys) considera que ante la misma evidencia, la probabilidad debe valer lo mismo para todos los individuos
- **Subjetiva** (F. Ramsey, Bruno de Finetti, LJ Savage) considera a la probabilidad como una medida de incertidumbre personal.

Definición de probabilidad según Laplace

Pierre-Simon Laplace (1749-1821) propuso la slguiente definición del concepto de probabilidad: consideremos un experimento aleatorio que tiene un número finito de resultados posibles y equiprobables

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

entonces la probabilidad de un evento $A \subset \Omega$ se define por

$$P(A) = \frac{\text{Casos favorables}}{\text{Casos posibles}} = \frac{\#(A)}{\#(\Omega)}$$

Notemos que el aparente azar en este ejemplo del dado, se debe en realidad a nuestra ignorancia. Porque la mecánica clásica (Newtoniana) nos dice que el movimiento del dado es en realidad un proceso completamente determinístico (no aleatorio). Y si conociéramos la posición y velocidad iniciales y las fuerzas que actúan, podríamos calcular (en principio) en forma exacta, cómo se va a mover el dado.

Definición de probabilidad según Kolmogorov

Un tipo diferente de azar, mucho más fundamental, aparece en la mecánica cuántica, una de las teorías fundamentales de la física moderna. Esta teoría postula que el azar es un componente esencial e irreductible de la naturaleza a nivel microscópico. Así por ejemplo, no podemos predecir con exactitud donde vamos a encontrar un electrón, sino solamente calcular la probabilidad de que el electrón esté en una cierta región del espacio.

El problema de tener espacios muestrales no necesariamente finitos hace que no se le pueda asignar probabilidad a cualquier evento que consideremos, es decir, a cualquier subconjunto de un espacio muestral n.

Modelo probabilístico de Kolmogorov

Un modelo probabilísitco se compone de

- **Espacio muestral** Ω: el conjunto de todos los resultados posibles del experimento.
- **Un evento es un subconjunto del espacio muestral**. Muchas veces, en vez de escribirlo como un conjunto lo enunciaremos con palabras.
- **Una ley de probabilidad**: asigna a un evento $A \subseteq \Omega$ de posibles resultados un número no negativo $P(A)$.
- La lista de eventos a los que se les puede asignar probabilidad se suele conocer como Σ, que es una σ-álgebra.[1]

[1] 2 Una σ-álgebra de Ω es una familia de subconjuntos de Ω que es cerrada bajo complementos, uniones e intersecciones finitas o numerables. Σ contiene a $\varnothing$ y Ω.

Axiomas de probabilidad

Definimos la ley de probabilidad axiomáticamente, es decir, a partir de supuestos que suponemos ciertos:

1. **Axioma de no negatividad:** para todo evento A se cumple que $P(A) \geq 0$.
2. **Axioma de aditividad**: Si A y B son conjuntos disjuntos, entonces la probabilidad de su unión cumple
$$P(A \uplus B) = P(A) + P(B)$$
En general, si el espacio muestral tiene numerables elementos y $A_1, A_2, ...A_n$ es una sucesión de eventos disjuntos, la probabilidad de su unión verifica
$$P(A_1 \uplus A_2 \uplus \cdots \uplus A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$
3. **Axioma de normalización**: $P(\Omega) = 1$

Propiedades de una ley de probabilidad

A partir de los axiomas podemos deducir (demostrar) varias propiedades:

1. $P(\varnothing) = 0$
2. $P(A^C) = 1 - P(A)$
3. $A \subset B \Longrightarrow P(A) \leq P(B)$ y además $P(B \setminus A) = P(B) - P(A)$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
5. $P(A \cup B) \leq P(A) + P(B)$
6. $P(A \cup B \cup C) = P(A) + P\left(A^C \cap B\right) + P\left(A^C \cap B^C \cap C\right)$

Modelos discretos

Diremos que **un modelo es discreto si su espacio muestral Ω es finito o infinito numerable**. Si el espacio muestral es discreto, entonces **la ley de probabilidad queda completamente determinada por las probabilidades de los eventos que consisten de un solo elemento**.

En particular, la probabilidad de cualquier evento $A = a_1, \ldots, a_n$ es la suma de las probabilidades de sus elementos:

$$P(\{a_1, \cdots, a_n\}) = P(\{a_1\}) + P(\{a_2\}) + \cdots + P(\{a_n\})$$

Si el espacio muestral consiste en n resultados que son equiprobables (cada uno tiene la misma probabilidad de ocurrir), entonces

$$P(A) = \frac{\text{Cantidad de elementos de A}}{n} = \frac{\#A}{\#\Omega}$$

Ejemplo 1. Consideremos el experimento de tirar una vez un dado equilibrado y observar el resultado de la cara superior.

- $\Omega = \{$⚀,⚁,⚂,⚃,⚄,⚅$\}$
- $\Sigma = \{\{$⚀$\}, \ldots, \{$⚅$\}, \{$⚀,⚁$\}, \ldots, \{$⚀,⚂,⚄$\}, \{$⚁,⚃,⚅$\}, \ldots\}$

Describimos los siugientes eventos por extensión o comprensión

- $A_1 = \{$⚀,⚂,⚄$\} = \{$resultado impar$\}$
- $A_2 = \{$⚃,⚄,⚅$\} = \{$sale más que 3$\}$

Modelos continuos

Diremos que **un modelo es continuo si el espacio muestral Ω es algún subconjunto de la recta real**.

Para modelos continuos, puede ocurrir que las probabilidades de eventos puntuales no lleguen a caracterizar las probabilidades de cualquier evento en Σ.

Ejemplo 2. Una rueda de la fortuna de longitud 2π se gira. Los valores del borde de la rueda pertenecen a $\Omega = [0, 2\pi)$. Suponiendo que la rueda es justa, es razonable suponer que todos los resultados son equiprobables. Sin embargo, ¿cuál es la probabilidad de un evento que consiste de un solo elemento? Notar que es cero esa probabilidad

Conteo

Para poder calcular probabilidades necesitamos aprender a contar cuántos elementos hay en Ω y en subconjuntos de Ω. Por lo tanto, vamos a aprender distintas formas de contar, según el caso que estemos analizando.

- **Power rule**: hay n^k formas de **elegir ordenadamente** k objetos con reposición de entre n objetos diferentes.
- **Factorial**: hay $n!$ formas de **ordenar** n objetos diferentes en fila.

- **Permutaciones**: hay $nPk = n \cdot (n-1) \cdot \cdots \cdot (n-k+1) = \frac{n!}{(n-k)!}$ formas de **extraer y ordenar** k objetos diferentes de un total de n elementos en fila.

- **Combinaciones**: hay $nCk = \binom{n}{k} = \frac{n!}{k!(n-k)!}$ formas de **extraer** k objetos diferentes de un total de n elementos. **No importa el orden.**

- **Bosones**: hay $\binom{n+k-1}{n}$ formas de **ordenar** n objetos **indistinguibles** en k grupos **distinguibles**. Notar que por propiedades de las combinaciones también se puede escribir como $\binom{n+k-1}{k-1}$.

Propiedad de las combinaciones: Para cualesquiera enteros no negativos n y k con $k \leq n$, tenemos

$$\binom{n}{k} = \binom{n}{n-k}.$$

Aplicación power rule

- **Ejemplo power rule 1**: Supongamos que quiero armar un sandwich y tengo 3 ingredientes (milanesa, tomate, queso). Tengo las siguientes posibilidades (poner ingrediente, no poner ingrediente). Entonces la cantidad de sandwiches que puedo hacer sin importar el orden de los elementos es $\#sandwich = 2 \cdot 2 \cdot 2 = 2^3$.

- **Ejemplo power rule 2**: Supongamos que Ω es el espacio muestral de todas las maneras que se pueden considerar números enteros de 7 cifras que pertenecen al conjunto $\{1, 4, 5\}$. ¿Cuántos elementos tiene Ω?. Los elementos son #números con 7 digitos $= 3^7$.

Aplicación factorial

- **Ejemplo factorial 1**: Supongamos que Ω es el espacio muestral de todas las maneras en que pueden ordenarse 10 personas en una fila. ¿Cuántos elementos tiene Ω? Son $10!$ formas de ordenar.

- **Ejemplo factorial 2**: Supongamos que Ω es el espacio muestral de todas las maneras en que pueden ordenarse las letras de la palabra "murciélago". ¿Cuántos elementos tiene Ω? Son $10!$ formas de ordenar.

Aplicación permutación y permutaciones cícliclas

- **Ejemplo permutación**: Consideramos n bolillas distintas. Supongamos que queremos elegir sucesivamente k de las n bolas sin reposición y nos importa el orden en el que se eligen, donde $k \leq n$. Sea Ω el espacio muestral de posibles resultados. ¿Cuántos posibles resultados hay en Ω? **Acá es clave que las bolillas son distinguibles y que nos importa en qué orden se sacan.** Por ejemplo, supongamos que $n = 5$ y $k = 3$, entonces Por lo tanto en este caso, la cantidad de posibles resultados en Ω es $5P3 = 5!/2! = 60$.

- **Permutaciones cíclicas**: Imaginemos que queremos ordenar n elementos diferentes, pero de manera circular (imagine una mesa redonda). En este caso lo único relevante es la ubicación relativa entre los elementos, ya no es más relevante qué elemento está primero porque están ubicados circularmente. Por ejemplo, ¿de cuántas maneras distintas pueden ubicarse los doce números -del 1 a 12- en un reloj solamente teniendo en cuenta la posición relativa de los números entre sí (es decir no nos importa qué número ocupa cada lugar sino como están relativamente ubicados)? Notemos que si estuvieran ordenados en fila habría 12! formas de ordenarlos. Ahora bien hay 12 posibles números que podrían estar primero, pero como 10 único que nos importa es el orden relativo de los números, dividimos $12!/12 = 11!$

Aplicación combinación

- **Ejemplo combinación**: Consideramos n bolillas distintas. Supongamos que queremos elegir sucesivamente k de las n bolillas sin reposición y no nos importa el orden en el que se eligen, donde $k \leq n$. Sea Ω el espacio muestral de posibles resultados. ¿Cuántos posibles resultados hay en Ω? Acá es clave que las bolillas son distinguibles y que no nos importa en qué orden se sacan. Por ejemplo, supongamos que $n = 5$ y $k = 3$, entonces $5C3 = 10 = \frac{5!}{2!3!} = \binom{5}{3}$

Coeficiente multinomial

El coeficiente multinomial nos permite conocer la cantidad de formas de ordenar n objetos, donde cada elementos pertenece a uno solo de los r grupos

Una combinación puede verse como una partición del conjunto en dos: una parte contiene k elementos y la otra contiene los $n - k$ restantes. Ahora generalizamos considerando particiones en más de dos subconjuntos.

Damos un conjunto de n elementos y unos enteros no negativos $n_1, n_2, ..., n_r$ cuya suma es igual a n. Consideramos particiones del conjunto en r subconjuntos disjuntos con el i-ésimo subconjunto conteniendo exactamente n_i elementos. Contemos de cuántas maneras se puede hacer esto.

Formamos los subconjuntos de uno en uno. Tenemos $\binom{n}{n_1}$ formas de formar el primer subconjunto. Una vez formado el primer subconjunto, nos quedamos con $n - n_1$ elementos. Tenemos que elegir entre ellos para formar el segundo subconjunto, y tenemos $\binom{n-n_1}{n_2}$ opciones, etc. Utilizando el principio de conteo para este proceso de r etapas, el número total de elecciones es

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-\cdots-n_{r-1}}{n_r},$$

lo cual es igual a

$$\frac{n!}{n_1!\,(n-n_1)!}\cdot\frac{(n-n_1)!}{n_2!\,(n-n_1-n_2)!}\cdots\frac{(n-n_1-\cdots-n_{r-1})!}{(n-n_1-\cdots-n_{r-1}-n_r)!\,n_r!}.$$

Notamos que varios terminos se cancelan por lo cual resulta en

$$\frac{n!}{n_1!\,n_2!\cdots n_r!}.$$

A esto se lo conoce como coeficiente multinomial y es denotado como

$$\binom{n}{n_1, n_2, \cdots, n_r}.$$

Teorema multinomial

Una aplicación útil del coeficiente multinomial es el teorema multinomial el cuál es una generalización del teorema del binomio; permitiendo calcular el coeficiente de polinomios elevados a un exponente natural.

Para cualquier entero m y cualquier entero no negativo n, se verifica que:

$$(x_1 + x_2 + \cdots + x_m)^n = \sum_{k_1+k_2+\cdots+k_m=n} \binom{n}{k_1, k_2, \ldots, k_m} \prod_{1 \le t \le m} x_t^{k_t},$$

Dicho de otro modo, la expansión de $(x_1 + x_2 + \cdots + x_m)^n$ es la suma de todos los posibles productos

$$\frac{n!}{k_1!\, k_2!\, \cdots\, k_m!}\, x_1^{k_1} x_2^{k_2} \cdots x_m^{k_m},$$

donde los exponentes $k_1, k_2, ..., k_m$ suman n

Bosones

Los bosones son **partículas que exhiben estados totalmente simétricos**. En probabilidad, llamamos bosones a objetos que no se pueden distinguir. Pensemos entonces cuántos elementos tiene Ω el espacio muestral donde se consideran todas las posibles formas de:

- considerar las combinaciones de números x_i que cumplan la ecuación $x_1 + x_2 + \cdots + x_k = n$ si $x_i \in \mathbb{N}_{\geq 0}$.
- poner n bolitas indistinguibles (iguales entre sí) en k cajitas diferentes.
- ordenar en fila n letras A y $k - 1$ letras B.

en general, si queremos contar de cuántas maneras se pueden ordenar n bolillas indistinguibles en k cajitas distinguibles ($k - 1$ separadores), hay

$$\binom{n+k-1}{k-1}$$

formas de hacerlo

Probabilidad condicional

Definición de probabilidad condicional

Definimos la probabilidad condicional del evento A dado que sabemos que ocurrió el evento B, con $P(B) > 0$ como

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

Observación 3. Si tenemos un modelo probabilístico $(\Omega, \Sigma, P(\cdot))$ y sabemos que ocurrió el evento A podemos incorporar esa información para obtener un nuevo modelo probabilístico:

$$(A, \Sigma_A, P(\cdot|A))$$

$P(\cdot|A)$ define una probabilidad sobre el nuevo espacio muestral A. Es decir, verifica los axiomas de probabilidad. O sea

- $P(A|A) = 1$
- $P(B_1|A) \geq 0$
- Para cualesquiera B_1 y B_2 eventos disjuntos que pertenencen a Σ_A vale que $P(B_1 \cup B_2|A) = P(B_1|A) + P(B_2|A)$

Definición de ley de probabilidad total (LPT)

Queremos describir $P(B)$ en como una suma de probabilidades según si ocurre A u ocurre A^c. Observemos que los eventos

$$B \cap A \quad y \quad B \cap A^c$$

son mutuamente excluyentes pues si ocurre A no puede ocurrir A^c. En este caso consideramos que Ω se puede particionar como $\Omega = A \cup A^c$. Además, podemos escribir al conjunto B como una partición de dos conjuntos disjuntos

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

Entonces,

$$P(B) = P[(B \cap A) \cup (B \cap A^c)] = P(B \cap A) + P(B \cap A^c)$$

Con la definición de prob. condicional, obtenemos un caso particular de la LPT:

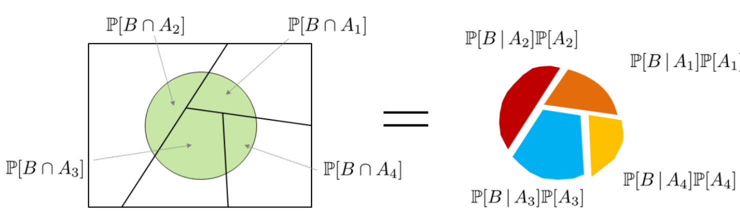
$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$$

Caso general

Más en general si $\{A_1, ..., A_n\}$ es una partición de Ω, es decir $\dot{\bigcup}_{i=1}^{n} A_i = \Omega_i$ entonces, se tiene que

$$P(B) = \overbrace{P(B|A_1)P(A_1)}^{P(A_1 \cap B)} + \cdots + \overbrace{P(B|A_n)P(A_n)}^{P(A_n \cap B)}$$

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$



Regla de multiplicación

Para el caso de dos eventos

De la definición de probabilidad condicional,

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

podemos despejar $P(A \cap B)$ obteniendo $P(A \cap B) = P(A|B)P(B)$ [2]

[2] Esta fórmula es un caso particular de la regla de multiplicación, ya que estamos considerando solamente dos eventos.

Caso general de n eventos

Dados n eventos $A_1, ..., A_n$, entonces la regla de multiplicación nos permite hallar la probabilidad de su intersección como:

$$\begin{aligned} P\left(A_1 \cap A_2 \cap \cdots \cap A_n\right) =& P\left(A_n | A_1 \cap A_2 \cap \cdots \cap A_{n-1}\right) \cdot \\ & P\left(A_{n-1} | A_1 \cap A_2 \cap \cdots \cap A_{n-2}\right) \cdots \\ & P\left(A_2 | A_1\right) \cdot P\left(A_1\right) \end{aligned}$$

La regla es útil en muchas situaciones donde el cálculo de cada probabilidad condicional es relativamente fácil e intuitivo mientras que el cálculo directo de $P\left(A_1 \cap A_2 \cap \cdots \cap A_k\right)$ puede no ser tan sencillo. **Nota**: $A_0 = \Omega$.

De manera sucinta,

$$P\left(\bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} P\left(A_k \middle| \bigcap_{j=1}^{k-1} A_j\right)$$

Regla de Bayes

Para el caso de dos eventos

La regla de Bayes relaciona las probabilidades de los eventos $P(A) \neq 0$, $P(B) \neq 0$, $P(A|B)$ y $P(B|A)$.

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

Para demostrar este resultado usamos la definición de probabilidad condicional, empezando por la siguiente igualdad:

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

Multiplicamos por $\frac{1}{P(B)}$ a ambos lados y del lado derecho multiplicamos y dividimos por $P(A)$:

$$\begin{aligned} \frac{P(A \cap B)}{P(B)} &= P(A) \cdot \frac{P(A \cap B)}{P(A)} \frac{1}{P(B)} \\ P(A|B) &= P(A) \cdot \frac{P(B|A)}{P(B)} \end{aligned}$$

Para el caso general

Sean eventos A_i de manera que $\biguplus_{i=1}^{n} A_i = \Omega$ (una partición de Ω) y B. Entonces, usando la regla de Bayes para dos eventos, sabemos que

$$P(A_i|B) = P(A_i) \cdot \frac{P(B|A_i)}{P(B)}$$

Por LPT también sabemos que

$$P(B) = \sum_{i=1}^{n} P(B|A_i) P(A_i)$$

Lo reemplazamos en la primera ecuación y queda que:

$$P(A_i|B) = P(A_i) \cdot \frac{P(B|A_i)}{\sum_{i=1}^{n} P(B|A_i) P(A_i)}$$

Probabilidad a priori y posteriori

Podemos descomponer la regla de bayes del siguiente modo

$$\underbrace{P(A|B)}_{\text{Prob a posteriori}} = \underbrace{P(A)}_{\text{Prob a priori}} \cdot \underbrace{\frac{P(B|A)}{P(B)}}_{\text{Cociente de verosimilitud}}$$

Odds a priori y posteriori

Al cociente entre la probabilidad de un evento y la probabilidad del complemento del evento, se lo denomina Odds. Por lo que los odds a priori y a posteriori resultan

$$\text{Odds a priori para el evento A} = \frac{P(A)}{P(A^c)}$$

$$\text{Odds a posteriori para el evento A dado B} = \frac{P(A|B)}{P(A^c|B)}$$

$$\underbrace{P(A|B)}_{\text{Prob a posteriori}} = \underbrace{P(A)}_{\text{Prob a priori}} \cdot \underbrace{\frac{P(B|A)}{P(B)}}_{\text{Cociente de verosimilitud}}$$

Eventos independientes

Independencia para el caso de dos eventos

Dos eventos A y B, si $0 < P(B) < 1$, son independientes cuando

$$P(A|B^c) = P(A|B)$$

Observación 4 (Afirmaciones equivalentes). Las siguientes afirmaciones son equivalentes

1. Los eventos A y B son independientes $P(A|B) = P(A|B^c)$
2. $P(A) = P(A|B)$
3. $P(A \cap B) = P(A) \cdot P(B)$

Proposición 5 (Propiedades de los eventos independientes)**.** *Valen las siguientes propiedades*

1. *Si A y B son independientes, entonces A y B^c; A^c y B^c; A^c y B son eventos independientes.*
2. *El evento $\varnothing$ es independiente de cualquier evento A.*
3. *Si $A \subseteq B$, no pueden ser independientes excepto que $P(B) = 1$.*

Independencia y exclusión mutua

Definición 6 (Eventos mutuamente excluyentes)**.** Se dice que dos eventos A y B son mutuamente excluyentes. Eso quiere decir que $A \cap B = \varnothing$ y, por lo tanto, $P(A \cap B) = 0$.

Proposición 7 (Independencia y exclusión)**.** *Sean* ***dos eventos*** *A* ***y*** *B* ***mutuamente excluyentes*** *y que cumplen que $P(A) > 0$ y $P(B) > 0$* ***no pueden ser independientes****.*

Si A y B fueran independientes tendría que valer que

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

Sin embargo, sabemos que el lado izquierdo es igual a cero, mientras que el lado derecho es positivo. Entonces, si se consideran eventos A y B con probabilidad positiva y mutuamente excluyentes, no pueden ser independientes.

Independencia para el caso de más eventos

Definición 8 (Eventos independientes)**.** Si se verifica que para cada subconjunto $S \subseteq \{1, 2, ..., n\}$ vale

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

$$P\left(A_i \cap A_j \cap \cdots\right) = P(A_i) \cdot P(A_j) \cdots$$

se dice que los eventos $A_1, ..., A_n$ son independientes (donde $i, j \in S$).

Observación 9 (Observancia práctica)*.* Es importante recordar que para demostrar que vale independencia entre los eventos $A_1, ..., A_n$ no alcanza con mostrar la independencia de a pares de eventos A_i y A_j.

Principio de inclusión-exclusión

Para el caso de 2 eventos

Si tenemos eventos A_1 y A_2 el principio de inclusión-exclusión es

$$P\left(A_1 \cup A_2\right) = P\left(A_1\right) + P\left(A_2\right) - P\left(A_1 \cap A_2\right)$$

Para el caso de 3 eventos

Si tenemos eventos A_1 , A_2 y A_3 el principio de inclusión-exclusión es

$$\begin{aligned} P\left(A_1 \cup A_2 \cup A_3\right) = P\left(A_1\right) + P\left(A_2\right) + P\left(A_3\right) - P\left(A_1 \cap A_2\right) - P\left(A_1 \cap A_3\right) \\ - P\left(A_2 \cap A_3\right) + P\left(A_1 \cap A_2 \cap A_3\right) \end{aligned}$$

Para el caso general de n eventos

Definición 10. Si tenemos eventos $A_1, ..., A_n$ el principio de inclusión-exclusión es

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P\left(A_i\right) - \sum_{i<j} P\left(A_i \cap A_j\right) + \sum_{i<j<k} P\left(A_i \cap A_j \cap A_k\right) \\ + \cdots + (-1)^{n-1} P\left(\bigcap_{i=1}^{n} A_i\right) \end{aligned}$$

que se puede reescribir como[3]

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n}\left((-1)^{k-1} \sum_{\substack{I \subseteq \{1,..,n\} \\ \#(I)=k}} P(B_I)\right)$$

[3] donde por ejemplo, si $I = \{1,2,5\}$, $B_I = A_1 \cap A_2 \cap A_5$

Observación 11 (Propiedad del principio inclusión-exclusión). Sea un elemento x que pertence a k de los n conjuntos $A_1, ..., A_n$. Del lado izquierdo del caso general $P\left(\cup_{i=1}^{n} A_i\right)$ el elemento x está contando una sola vez.

En cambio en el lado derecho es contado muchas veces:

- en el término $\sum_{i=1}^{n} P(A_i)$ es contado $+\binom{k}{1}$ veces.
- en el término $\sum_{i<j} P\left(A_i \cap A_j\right)$ es contado $-\binom{k}{2}$ veces.
- en el término $P\left(\bigcap_{i=1}^{n} A_i\right)$ es contado $(-1)^{k-1}\binom{k}{k}$ veces.

Por lo tanto vale que

$$1 = \binom{k}{1} - \binom{k}{2} + \binom{k}{3} - \cdots + (-1)^{k-1}\binom{k}{k}$$

Desarreglos

Se llama un desarreglo de un conjunto de n elementos a una permutación que no deja a ningún elemento en su posición original.

Definición 12 (Desarreglos)**.** Llamamos D_n a la cantidad de desarreglos para n elementos

$$D_n = n!\left[1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!}\right]$$

Funciones de variables aleatorias

Variables aleatorias

Una **variable aleatoria** X es una función $X : \Omega \to E$, donde:

Si E es un **conjunto discreto**, es decir, su cardinal es finito o numerable decimos que X es una **variable aleatoria discreta**.

Si E es un **conjunto continuo**, es decir, su cardinal es el cardinal de R decimos que X es una **variable aleatoria continua**.

- Ω es un espacio muestral, junto con Σ el espacio de todos los eventos y $P(\cdot)$ una probabilidad definida sobre los eventos A que están en Σ.
- E es un subconjunto de R.

Propiedades de la esperanza y varianza

Sea X una variable aleatoria con varianza finita y c un escalar. Sean $g(\cdot)$ y $h(\cdot)$ funciones. Aplicar esperanza y varianza[4] tiene las siguientes propiedades:

[4] Notar que esperanza y varianza se define en los apéndices.

Propiedades de la esperanza

Las propiedades de la esperanza son

- **Linealidad:** $E[g(X)+h(X)] = E[g(X)] + E[h(X)]$
- **Multiplicación por constantes:** $E[cX] = cE[X]$
- **Suma de constantes:** $E[X+c] = E[X] + c$

Propiedades de la varianza

Las propiedades de la varianza son

- **No negativo:** $Var(X) \geq 0$
- **Varianza de una constante:** $Var(c) = 0$
- **Multiplicación por constantes:** $Var(cX) = c^2 Var(X)$
- **Suma de constantes:** $Var(X+c) = Var(X)$

Función de una variable aleatoria

Definición de función de variable una aleatoria

Definición 13 (Función de una variable aleatoria)**.** Dada una variable aleatoria X y una función $g(\cdot)$ definimos la función de una variable aleatoria

$$\Omega \xrightarrow{X} E \xrightarrow{g} B$$

Es decir, la variable aleatoria $g(X)$ no es otra cosa que aplicarle a la variable aleatoria X la función $g(\cdot)$.

Distribución de función de variable una aleatoria

Para poder estudiar la distribución se puede

- calcular $f_{g(X)}(u)$ la función de probabilidad puntual (si X es una variable discreta) o función de densidad (si X es una variable continua).
- usar $f_X(x)$ y pensar que $g(X)$ es una función de la variable X.

Esperanza de $Y = g(X)$

Definición 14 (Esperanza discreta)**.** La esperanza de la función $g(X)$ para X discreta es

$$E[g(X)] \underbrace{=}_{def} \sum_j y_j P\left[g(x) = y_j\right] = \sum_j y_j \sum_{i:g(x_i)=y_j} p\left(x_i\right)$$
$$= \sum_j \sum_{i:g(x_i)=y_j} g\left(x_i\right) p\left(x_i\right) \underbrace{=}_{LOTUS} \sum_{i:p(x_i)>0} g\left(x_i\right) p\left(x_i\right)$$

LOTUS significa Law of the unconscious statistician, o sea ley del estadístico inconsciente, es un teorema utilizado para calcular el valor esperado de una función $g(X)$ de una variable aleatoria X cuando se conoce la distribución de probabilidad de X pero no se conoce la distribución de $g(X)$.

Definición 15 (Esperanza continua)**.** La esperanza de la función $g(X)$ para X continua es

$$E[g(X)] \underbrace{=}_{def} \int_{-\infty}^{+\infty} x f_{g(X)}(x)\, dx \underbrace{=}_{LOTUS} \int_{-\infty}^{+\infty} g(x) f(x)\, dx$$

Proposiciones de funciones de variables continuas

Proposición 16 (Función de densidad de una función de variables continuas)**.** *Sea X una variable aleatoria continua con PDF $f(x)$ y $sop(X) \subseteq I$. Sea $Y = g(X)$ donde $g(\cdot)$ es una función derivable, biyectiva (tiene inversa g^{-1}), con $sop(Y) = g(I)$. Entonces, Y tiene función de densidad*

$$f_Y(y) = f_X\left(g^{-1}(y)\right) \left| \frac{\partial}{\partial y} g^{-1}(y) \right| \quad si\ y \in g(I)$$

Proposición 17 (Función de densidad acumulada como $g(X)$)**.** *Sea X una variable aleatoria continua. Si $V = F_X(X)$, entonces $V \sim U[0,1]$.*

$$P(V \leq v) = P(F_X(X) \leq v) = P\left(X \leq F_X^{-1}(v)\right) = F_X\left(F_X^{-1}(v)\right) = v$$

Proposición 18 (Función de densidad acumulada inversa como $g(X)$)**.** *Sea $U \sim U[0,1]$, y sea $X = F_X^{-1}(U)$. Entonces la función de distribución de X es F.* [5]

$$P(X \leq x) = P\left(F^{-1}(U) \leq x\right) = P(U \leq F(x)) = F(x)$$

[5] Este último resultado nos indica cómo generar valores aleatorios de una variable con función de distribución F_X, simplemente aplicando la función F_X^{-1} a números aleatorios generados a partir de una distribución $U[0,1]$.

Momentos de una variable aleatoria

Se pueden definir dos tipos de momentos de una variable aleatoria X que tiene $E(X) = \mu$ centrados y no centrados.

Momentos centrados

El k-ésimo momento sobre la media (o k-ésimo momento central) de una variable aleatoria X es

$$E[(X-\mu)^k]$$

donde μ es la esperanza de X.

Notar que el momento centrado

- **de orden** 1: $E(X-\mu) = 0$
- **de orden** 2: $E((X-\mu)^2) = Var(X)$
- **de orden** 3: $E((X-\mu)^3)$ mide la skewness
- **de orden** 4: $E((X-\mu)^4)$ mide la kurtosis

Momentos no centrados

El k-ésimo momento no centrado de una variable aleatoria X se puede calcular

- **usando la función de densidad** (PDF)[6]

$$E(X^k) = \int_{-\infty}^{+\infty} x^k f(x) dx$$

- **usando la función generatriz de momentos** (MGF)

$$E(X^k) = \frac{d^k}{dx^k} M_X(t)\Big|_{t=0}$$

[6] o la función de probabilidad puntual (PMF) en el caso de variables aleatorias discretas y respectivamente

$$E(X^k) = \sum_x x^k p(x)$$

Función generatriz de momentos

Definimos la función generatriz de momentos (MGF) para una variable aleatoria X como[7]

$$M_X(t) = E\left[e^{tX}\right]$$

Observación 19 (MGF para calcular momentos no centrados). Notar que $M_X(t)$ cumple las siguientes propiedades

1. $M_X(0) = 1$
2. $M_X'(0) = \frac{d}{dt}E\left[e^{tX}\right]\Big|_{t=0} = E(X)$
3. $M_X''(0) = \frac{d^2}{dt^2}E\left[e^{tX}\right]\Big|_{t=0} = E(X^2)$
4. $M_X^k(0) = \frac{d^k}{dt^k}E\left[e^{tX}\right]\Big|_{t=0} = E(X^k)$ para cualquier número $k \in \mathbb{N}$

Proposición 20 (MGF de una suma de dos variables aleatorias)**.** *Sean X e Y variables aleatorias independientes y sea $W = X + Y$ entonces*

$$M_W(t) = M_X(t)M_Y(t)$$

Proposición 21 (MGF de una suma de variables aleatorias)**.** *Sean $X_1, ..., X_n$ variables aleatorias independientes y sea $S = \sum_{i=1}^n X_i$ entonces $M_S(t)$ es*

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

Notar que si $X_i \sim_{iid}$ entonces

$$M_S(t) = (M_{X_1}(t))^n$$

Proposición 22 (MGF de transformación lineal de variable aleatoria)**.** *Si una variable aleatoria X tiene MGF $M_X(t)$, entonces la variable aleatoria $\alpha X + \beta$ tiene MGF*

$$M_{\alpha X+\beta}(t) = e^{\beta t}M_X(\alpha t)$$

Proposición 23 (MGF de transformación lineal de variables aleatorias)**.** *Si $S_n = \sum_{i=1}^n a_i X_i$, donde X_i son variables aleatorias y los a_i son constantes, entonces la MGF de S_n está dada por*

$$M_{S_n}(t) = M_{X_1}(a_1 t)\, M_{X_2}(a_2 t) \cdots M_{X_n}(a_n t)\,.$$

Distribuciones conjuntas

Vector aleatorio

Un vector aleatorio $\underline{X} = (X_1, X_2, ..., X_n)$ es una función[8]

$$\underline{X} : \Omega \to \prod_{i=1}^n E_i = E_1 \times \cdots \times E_n$$

donde, para cada $i = 1, ..., n$ vale que $X_i : \Omega \to E_i$ es una variable aleatoria.[9]

[7] Notar que para el **caso discreto** es

$$E\left[e^{tX}\right] = \sum_{x:p(x)>0} e^{tx}p(x)$$

y para el **caso continuo** es

$$E\left[e^{tX}\right] = \int_{-\infty}^{+\infty} e^{tx}f(x)dx$$

[8] Denotamos $\prod_{i=1}^n E_i = E_1 \times E_2 \times \cdots \times E_n = \{(x_1, ..., x_n) : x_i \in E_i \forall i = 1, ..., n\}$ al **producto cartesiano** de los conjunto E_i con $i = 1, ..., n$

[9] Entonces podemos decir que

- si las **variables aleatorias** X_i **son discretas** para todos los valores de $i = 1, ..., n$ el vector $\underline{X}$ es un **vector aleatorio discreto**.
- si las **variables aleatorias** X_i **son continuas** para todos los valores de $i = 1, ..., n$ el vector $\underline{X}$ es un **vector aleatorio continuo**.

Distribución de vectores aleatorios

Para dos variables aleatorias

Sea el vector aleatorio $(X, Y) : \Omega \to E_X \times E_Y$ consideramos para X e Y variables aleatorias continuas tenemos

- **PDF conjunta:**[10]

$$f_{X,Y}(x, y)$$

- **CDF conjunta:**[11]

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

- **PDF marginal de X:**[12]

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y) dy$$

- **CDF marginal de X:**[13]

$$\int_{-\infty}^{x} \int_{-\infty}^{+\infty} f_{X,Y}(u, y) \, dy \, du$$

[10] Para variables discretas es PMF conjunta

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

[11] Para variables discretas es CDF conjunta

$$F_{X,Y}(x, y) = \sum_{u<x} \sum_{v<y} p_{X,Y}(u, v)$$

[12] Para variables discretas es PMF marginal

$$p_X(x) = \sum_{y \in E_Y} p_{X,Y}(x, y)$$

[13] Para variables discretas es CDF marginal

$$F_X(x) = \sum_{u<x} \sum_{v<y} p_{X,Y}(u, y)$$

Para múltiples variables aleatorias

Dadas k variables aleatorias, definimos el vector $(X_1, X_2, ..., X_n)$ tenemos

- **PDF conjunta:**

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

- **PDF marginal de X_i:**

$$p_{X_i}(x_i^*) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} p_{X_1,\ldots X_n}(x_1, \ldots, x_{i-1}, x_i^*, x_{i+1}, \ldots, x_n)$$

Proposición 24 (Esperanza de una función de variables aleatorias). *Si $(X_1, ..., X_n)$ es un vector aleatorio y $Y = g(X_1, ..., X_n)$ entonces*

$$E(Y) = \sum_{x_1} \cdots \sum_{x_n} g(x_1, \cdots, x_n) p_{X_1,\ldots,X_n}(x_1, \cdots, x_n)$$

En particular si $(X_1, ..., X_n)$ es un vector aleatorio y $a_1, ..., a_n$ son constantes entonces

$$E(a_1 X_1 + \cdots + a_n X_n) = a_1 E(X_1) + \cdots + a_n E(X_n)$$

Propiedades de PDF y CDF

Proposición 25 (Condicion de cierre). [14] *Sean X e Y variables aleatorias continuas podemos definir la condición de cierre del siguiente modo*

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

[14] En el caso de variables discretas la condición de cierre es

$$\sum_{x \in E_X} \sum_{y \in E_Y} p_{X,Y}(x, y) = 1$$

Proposición 26 (Pasar de CDF a PDF). [15] *Sean X e Y variables aleatorias continuas podemos pasar de CDF a PDF del siguiente modo*

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial y \partial x} F_{X,Y}(x,y)$$

[15] Para el caso de variables discretas para pasar de CDF a PMF es

$$p_{X,Y}(x,y) = F_{X,Y}(x,y) - F_{X,Y}(x,\hat{y}) - F_{X,Y}(\hat{x},y) + F_{X,Y}(\hat{x},\hat{y})$$

donde $\hat{x} < x$ y además $\hat{y} < y$ son los valores inmediatamente anteriores a x e y respectivamente.

Proposición 27 (Pasar de PDF a CDF). *Sean X e Y variables aleatorias continuas podemos pasar de PDF a CDF del siguiente modo*

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}\,(u,\, v)\, dv\, du$$

y además

$$P(a \le X \le b, c \le Y \le d) = \int_a^b \int_c^d f_{X,Y}(x,y) dy dx$$

Notar que $f_{X,Y}(x,y) > 0$ si $(x,y) \in Sop(X) \times Sop(Y)$.

Para el caso de variables discretas para pasar de PMF a CDF es

$$F_{X,Y}(x,y) = \sum_{u<x} \sum_{v<y} p_{X,Y}(u,v)$$

y además

$$P(a \le X \le b, c \le Y \le d) = \\ = \sum_{a \le x \le b} \sum_{c \le v \le d} p_{X,Y}(X,y)$$

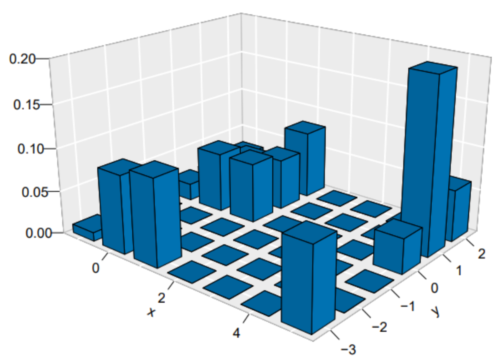
Histograma de probabilidad para la distribución conjunta

Un histograma de probabilidad para una vector aleatorio (X,Y) con componentes discretas es un gráfico de barras en tres dimensiones tal que:

- los posibles valores $(x_1,y_1), (x_2,y_2), ..., (x_n,y_n)$ de (X,Y) se ubican sobre un plano horizontal
- por cada posible valor (x_i,y_i) de (X,Y) hay un paralelepípedo centrado en (x_i,y_i), cuya base es un cuadrado de lado 1 y altura $p_{X,Y}(x_i,y_i)$

Debido a su construcción,

- El volumen del paralelepípedo sobre el valor (x_i,y_i) es igual a la probabilidad de que el vector aleatorio (X,Y) sea igual a (x_i,y_i).
- La suma de los volúmenes de todas las cajas es igual a 1.



Distribuciones de funciones de variables aleatorias

Distribución de $g(X,Y) = X + Y$ (Convolución)

Sean X e Y variables aleatorias continuas para calcular $p_S(s)$ o $F_S(s)$, donde $S = X + Y$, se tiene que utiliza la convolución entre las PDF marginales de X y de

Y.[16]

$$F_S(s) = \int_{-\infty}^{+\infty} \int_{-\infty}^{s-y} f_{X,Y}(x,y)dxdy = \int_{-\infty}^{+\infty} \int_{-\infty}^{s} f_{X,Y}(x, v-x)dvdx$$

[16] Para variables discretas notar que es

$$p_S(s) = \sum_{Sop(X)} p_{X,Y}(x, s-x)$$

Proposición 28 (Convolución). *La convolución entre dos función f y g es la función $f * g$ que cumple que*

$$f * g(v) = \int_{-\infty}^{+\infty} f(x)g(v-x)dx$$

Sean X e Y variables aleatorias independientes. Si queremos calcular la función de distribución acumulada de $X+Y$ donde f y g son las funciones de densidad respectivamente entonces

$$F_{X+Y}(s) = \int_{-\infty}^{+\infty} \int_{-\infty}^{s} f(x)g(v-x)dvdx = \int_{-\infty}^{s} f * g(v)dv$$

Vale que $F * g = g * f$ y que $f * (g * h) = (f * g) * h$,

Distribución de $g(X,Y) = Y/X$

Sean X e Y variables aleatorias continuas, para calcular $p_C(c)$ o $F_C(c)$, donde $C = Y/X$ tenemos [17]

$$F_C(c) = \int_{-\infty}^{0} \int_{cx}^{+\infty} f_{X,Y}(x,y)dydx + \int_{0}^{+\infty} \int_{-\infty}^{cx} f_{X,Y}(x,y)dydx$$

[17] Para variables discretas notar que

$$p_C(c) = \sum_{Sop(X)} p_{X,Y}(x, cx)$$

Distribución de $g_1(X,Y)$ y $g_2(X,Y)$

Sean X e Y variables aleatorias continuas y sean las variables aleatorias U y V que cumplen que

$$U = g_1(X,Y) \text{ } y \text{ } V = g_2(X,Y)$$

donde se pueden obtener funciones h_1 y h_2 que cumplen que

$$X = h_1(X,Y) \text{ } y \text{ } Y = h_2(X,Y)$$

Entonces vale que

$$f_{U,V}(u,v) = f_{X,Y}(h_1(u,v), h_2(u,v)) \cdot |J(h_1(u,v); h_2(u,v))|$$

donde

$$|J(h_1(u,v); h_2(u,v))| = \left| det \begin{pmatrix} \frac{\partial}{\partial u}h_1 & \frac{\partial}{\partial v}h_1 \\ \frac{\partial}{\partial u}h_2 & \frac{\partial}{\partial v}h_2 \end{pmatrix} \right|$$

Construcción de distribuciones conjuntas a partir de marginales

Familia de distribuciones bivariadas (Farlie-Gumbel-Morgenstern)

Se puede demostrar que si $F_X(x)$ y $F_Y(y)$ son dos CDF de variables aleatorias X e Y respectivamente y $|\alpha| < 1$, entonces las funciones

$$H(x,y) = F_X(x)F_Y(y) + \alpha F_X(x)F_Y(y)[1 - F_X(x)F_Y(y)]$$

son posibles funciones de distribución conjunta para las variables X e Y.

Esperanza de una función de dos variables aleatorias

Esperanza para un $g(X,Y)$

Si (X,Y) es un vector aleatorio discreto[18] con función de masa de probabilidad conjunta $p_{X,Y}(x,y)$ y $W = g(X,Y)$ entonces

$$E(W) = \sum_{(x,y)} g(x,y) \cdot p_{X,Y}(x,y)$$

En particular, si $g(X,Y) = aX + bY$ se tiene que

$$E(aX + bY) = aE(X) + bE(Y) \quad a, b \in \mathbb{R}$$

[18] Note que para el caso de un vector aleatorio continuo tiene que considerar $f_{X,Y}(x,y)$ e integrales en vez de sumas.

Esperanza para $g(X,Y) = XY$

Sean X e Y dos variables aleatorias discretas[19]. Entonces la esperanza del producto $XY = g(X,Y)$ es

$$E(XY) = \sum_{y \in \mathrm{Sop}(Y)} \sum_{x \in \mathrm{Sop}(X)} xy \cdot p_{X,Y}(x,y)$$

[19] Si X e Y son variables aleatorias continuas.

$$E(XY) = \int_{y \in \mathrm{Sop}(Y)} \int_{x \in \mathrm{Sop}(X)} xy \cdot f_{X,Y}(x,y)\, dx\, dy$$

Independencia de variables aleatorias

Decimos que dos variables aleatorias X e Y son independientes si y sólo si[20]

$$f_{X.Y}(x,y) = f_X(x) f_Y(y),$$

[20] para el caso discreto

$$p_{X.Y}(x,y) = p_X(x) p_Y(y)$$

Proposición 29 (Propiedad para la esperanza)**.** *Cuando X e Y son independientes, calcular $E(g(X,Y))$ es más sencillo si $g(X,Y)$ se puede escribir como $g(X,Y) = h(X) \cdot q(Y)$. En ese caso vale que*[21]

$$E(Z) = E[h(X)q(Y)] = E[h(X)] \cdot E[q(Y)]$$

[21] Note que para el caso de un vector aleatorio continuo tiene que considerar $f_{X,Y}(x,y)$ e integrales en vez de sumas.

Corolario 30. *Notar que si X e Y son independientes entonces*

- ***Esperanza de la multiplicación de variables aleatorias:*** *Tomando $g(X) = X$ y $q(Y) = Y$ notamos que si X e Y son independientes tiene que valer que*

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

- ***Varianza de la suma de variables aleatorias:*** *Si X e Y son variables aleatorias discretas independientes, entonces*

$$Var(X+Y) = Var(X) + Var(Y)$$

Covarianza y correlación

Dadas dos variables aleatorias X e Y, con sendas esperanzas μ_X y μ_Y, definimos

$$\mathrm{Cov}(X,Y) = E\left((X - \mu_X)(Y - \mu_Y)\right)$$

$$\mathrm{Corr}(X,Y) = \frac{E\left((X-\mu_X)(Y-\mu_Y)\right)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

La covarianza[22] y la correlación[23] miden la ***tendencia a una relación lineal*** entre las variables aleatorias X e Y.

Proposición 31 (Identidades derivadas). *Notar las siguientes identidades que se derivan*

- ***Covarianza:*** *Sean X e Y variables aleatorias entonces*

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

- ***Varianza:*** *Sean X e Y variables aleatorias entonces*

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

- ***Esperanza de v.a. independientes:*** *Sean X e Y variables aleatorias independientes entonces*

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

  y en particular $Cov(X,Y) = 0$. En general, ***no es cierto el recíproco*** *que si $Cov(X,Y) = 0$ entonces X e Y sean independientes.*

[22] La covarianza depende de la escala de medición (por ejemplo, euros, kilos, etc). Además notar que Notar que

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X)$$

.

[23] La correlación es una medida que no depende de la escala de medición y satisface $|\mathrm{Corr}(X,Y)| \leq 1$.

- Si $\mathrm{Corr}(X,Y) = 1$, existen $a, b \in \mathbb{R}$ tales que $Y = aX + b$.
- Si $\mathrm{Corr}(X,Y) > 0$, la tendencia es a una asociación lineal creciente entre X e Y.
- Si $\mathrm{Corr}(X,Y) < 0$, la tendencia es a una asociación lineal decreciente entre X e Y.

La covarianza es una función lineal,

$$\mathrm{Cov}(X+bW,Y) = \mathrm{Cov}(X,Y) + b\mathrm{Cov}(W,Y)$$

.

Muestra aleatoria y estadísticos de orden

Variables aleatorias independientes e identicamente distribuidas

Decimos que una colección de v.a. $X_1, \ldots, X_n$ es independiente e idénticamente distribuida (i.i.d.) si

- $X_1, \ldots, X_n$ son independientes y
- $X_1, \ldots, X_n$ tienen la misma distribución i.e., $f_{X_1}(x) = \cdots = f_{X_n}(x)$.

Distribución de variables aleatorias iid

Si $X_1, \ldots, X_n$ son i.i.d., se tiene que

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) = [f_{X_1}(x_1)]^n$$

Estadísticos de orden

Dadas v.a.i.i.d. $x_1, \ldots, x_n$, sea la **muestra aleatoria ordenada** $X_{(1)}, \ldots, X_{(n)}$. Definimos los **estadísticos de orden** desde la muestra aleatoria ordenada.

$$X_{(1)} = \min\{X_1, \ldots, X_n\}$$

$$\mathrm{mediana}(X) = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{si } n \text{ es impar} \\ \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} & \text{si } n \text{ es par} \end{cases}$$

$$X_{(n)} = \max\{X_1, \ldots, X_n\}$$

Distribución de los estadísticos de orden

La **distribución del máximo** es

$$F_{\text{máx}\{X_i\}} = F_{X_{(n)}}(x) = (F_{X_1}(x))^n$$

La **distribución del mínimo** es

$$F_{\text{mín}\{X_i\}} = F_{X_{(1)}}(x) = 1 - (1 - F_{X_1}(x))^n$$

Distribución del estadístico k

Note que la densidad de $X_{(k)}$, si $1 < k < n$, es

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

- Sin recurrir a una demostración, consideremos con qué probabilidad ocurre el evento $A = \{\omega \in \Omega : x \leq X_{(k)}(\omega) \leq x + dx\}$.[24] La probabilidad de que ocurra para una configuración específica es $f(x)[F(x)]^{k-1}[1 - F(x)]^{n-k}$.
- Ahora, hay $\frac{n!}{(k-1)!(n-k)!}$ formas de ordenar los n valores aleatorios de las v.a.i.i.d., dividiendo por las permutaciones de las $(k-1)$ v.a. que tomen valores menores que x y dividiendo por las permutaciones de las $(n-k)$ v.a. que tomen valores mayores a $x + dx$.

[24] Este evento ocurre si $k-1$ observaciones son menores que x, una observación cae en el intervalo $[x, x + dx]$ y las restantes observaciones son mayores a $x + dx$.

Distribución normal bivariada

Decimos que $f_{X,Y}(x, y)$ es la densidad de una *distribución normal bivariada* cuando

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho\frac{(x-\mu_X)}{\sigma_X}\frac{(y-\mu_Y)}{\sigma_Y}\right)\right]}$$

para $\mu_X \in \mathbb{R}, \mu_Y \in \mathbb{R}, \sigma_X > 0, \sigma_Y > 0$ y ρ tal que $|\rho| \leq 1$.

Escribimos

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$$

Notar que si consideramos $v^T = (a, b)$, la variable aleatoria

$$v^T \begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(v^T\mu, v^T\Sigma v\right)$$

Distribuciones condicionales y cadenas de Markov

Distribución condicional

Función de densidad condicional

Sean X e Y dos v.a. discretas[25] con soporte $Sop(X)$ e $Sop(Y)$ respectivamente. La PMF de X dado $Y = y$ es

$$p_{X|Y=y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} = P(X = x \mid Y = y).$$

[25] Para variables continuas la PDF de X dado $Y = y$ es

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

Función de probabilidad acumulada condicional

La CDF de X condicional a $Y = y$ es

$$F_{X|Y=y}(x \mid y) = P(X \leq x \mid Y = y) = \sum_{u \leq x} p_{X|Y=y}(u \mid y).$$

Independencia según probabilidad condicional

Notemos que con esta definición podemos decir que dos variables aleatorias son independientes si vale una de las siguientes afirmaciones $\forall x \in Sop(X), \forall y \in Sop(Y)$

$$p_{X|Y=y}(x \mid y) = p_X(x) \quad \text{ó} \quad p_{Y|X=x}(y \mid x) = p_Y(y) \quad \text{para v.a. discretas}$$

$$f_{X|Y=y}(x \mid y) = f_X(x) \quad \text{ó} \quad f_{Y|X=x}(y \mid x) = f_Y(y) \quad \text{para v.a. continuas.}$$

No negatividad y condición de cierre

Si $X|Y = y$ es una variable discreta, para cada valor de y, $p_{X|Y=y}(x, y)$ es una función de masa de probabilidad. Por lo tanto, si $Sop(X)$ es el soporte de X se verifica que entonces se verifica que fijado un valor de y cualquiera

$$p_{X|Y=y}(x \mid y) \geq 0 \quad \forall x \in Sop(X)$$

y por la condición de cierre para variables discretas[26]

$$\sum_{x \in Sop(X)} p_{X|Y=y}(x|y) = 1$$

[26] Notar que para variables continuas

$$f_{X|Y=y}(x \mid y) \geq 0 \quad \forall x \in Sop(X)$$

y

$$\int_{x \in Sop(X)} p_{X|Y=y}(x|y) = 1$$

Regla de multiplicación

Para dos variables

Para dos eventos vimos $P(A \cap B) = P(A)P(B|A)$. Si consideramos los eventos $A = \{X = x\}$ y $B = \{Y = y\}$ obtenemos

$$P(X = x, Y = y) = P(X = x)P(Y = y|X = x)$$

es decir que

$$p_{X,Y}(x, y) = p_X(x)p_{Y|X=x}(y \mid x)$$

Para k variables

Si tenemos k variables aleatorias $X_1, ..., X_k$, se verifica que

$$p_{X_1,...,X_k}(x_1, ..., x_k) = p_{X_1}(x_1) \cdot p_{X_2|X_1=x_1}(x_2|x_1) \cdots p_{X_k|X_1=x_1,...,X_{k-1}=x_{k-1}}(x_k|x_1, ..., x_{k-1})$$

Esperanza condicional

Definición de esperanza condicional

Dadas dos variables aleatorias X e Y, con Y una variable discreta, la esperanza condicional de Y dado $X = x$

$$E(Y \mid X = x)$$

se define como la esperanza de la distribución $P_{Y|X=x}(y \mid x)$, [27] es decir

$$E(Y \mid X = x) = \sum_{y \in \mathrm{Sop}(Y)} y \cdot P_{Y|X=x}(y \mid x)$$

donde $\mathrm{Sop}(Y)$ es el soporte de Y.

[27] Si Y fuera continua tendríamos que $E(Y \mid X = x) = \int_{\mathrm{Sop}(Y)} y \cdot f_{Y|X=x}(y \mid x)\, dy$

Observación 32. Notar que $E(Y \mid X = x)$ depende de x. Por lo tanto, $E(Y \mid X)$ es una variable aleatoria que depende de X. Podemos escribir $g(X) = E(Y \mid X)$.

Ley de expectativas iteradas (Adam's Law)

Sea (X, Y) un vector aleatorio de v.a. discretas. LEI también vale para v.a. continuas.

$$E(Y) = \sum_{x \in Sop(X)} E(Y|X = x) p_X(x) = E[E(Y|X)]$$

Varianza condicional

Definición de varianza condicional

Dadas dos variables aleatorias X e Y, con Y una variable discreta[28], la varinza condicional de Y dado $X = x$

$$Var(Y \mid X = x)$$

se define como la varianza de la distribución $P_{Y|X=x}(y \mid x)$, es decir

$$Var(Y \mid X = x) = \sum_{y \in Sop(Y)} (y - E(Y|X = x))^2\, p_{Y|X=x}(y|x)$$

donde $\mathrm{Sop}(Y)$ es el soporte de Y.

[28] Notar que el mismo resultado vale para variables continuas

Observación 33. Recordando la fórmula alternativa para la varianza obtenemos

$$Var(Y \mid X = x) = E(Y^2 \mid X = x) - [E(Y \mid X = x)]^2$$

Descomposición de la varianza (Eve's Law)

Como $Var(Y|X = x)$ depende de x, $Var(Y|X)$ es una variable aleatoria $g(X) = Var(Y|X)$ entonces

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$$

Cadenas de Markov discretas

Definimos un proceso estocástico discreto como una familia de variables aleatorias X_t, con $t \in N$ que toma valores en un espacio de estados $\Omega = \{s_1, ...s_n\}$. Supongamos que $\#(\Omega) = n$. Un proceso estocástico tiene la propiedad de Markov si para cualquier t, X_{t+1} depende de X_t pero no de $X_{t\text{-}1}, X_{t\text{-}2}, ..., X_1, X_0$. Es decir,

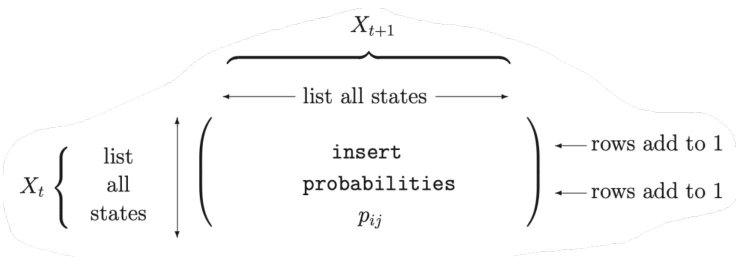
$$P(X_{t+1} = s | X_t = s_{(t)}, \cdots, X_1 = s_{(1)}, X_0 = s_{(0)}) = P(X_{t+1} = s | X_t = s_{(t)})$$

Definimos una cadena de Markov como un proceso estocástico que cumple la propiedad de Markov para todo $t = 0, 1, 2, ...$y todos posibles valores $s, s_{(t)}, ..., s_{(1)}, s_{(0)}$.

Matriz de transición

Definimos la matriz de transición P como la matriz donde, en el lugar ij, se escriben las probabilidades de transición de pasar en el período t del estado s_i a pasar en el período t + 1 al estado s_j y se denotan

$$p_{ij} = Prob(X_{t+1} = s_j | X_t = s_i).$$



Notemos que

- La suma de los elementos de cada una de las filas de P suman 1 porque $\sum_{j=1}^{n} p_{ij} = \sum_{j=1}^{n} P(X_{t+1} = s_j | X_t = s_i) = 1$.
- Las suma de los elementos de cada una de las columnas de P no tiene por qué sumar 1.

Probabilidad de transición entre k períodos

¿Cuál es la probabilidad de hacer una transición del estado s_i al s_j a lo largo de dos períodos?

$$P(X_2 = s_j | X_0 = s_i) = \sum_{k=1}^{n} P(X_2 = s_j | X_1 = s_k, X_0 = s_i) P(X_1 = s_k | S_0 = s_i)$$

$$= \sum_{k=1}^{n} P(X_2 = s_j | X_1 = s_k) P(X_1 = s_k | X_0 = s_j)$$

$$= \sum_{k=1}^{n} p_{kj} \cdot p_{ik} = \sum_{k=1}^{n} p_{ik} \cdot p_{kj}$$

$$= (P^2)_{ij}$$

De manera inductiva se puede probar que las probabilidades de transición de pasar del estado s_i al s_j a lo largo de k períodos es:

$$P(X_3 = s_j | X_0 = s_j) = (P^3)_{ij}$$

$$P(X_t = s_j | X_0 = s_i) = P(X_{k+t} = j | X_t = i) = (P^k)_{ij}$$

Probabilidades marginales

Consideremos por otro lado las distribuciones no condicionadas de las variables $X_0, X_1, ..., X_n$. Llamamos a la **distribución de probabilidades iniciales**

$$\pi^T = (\pi_1, \pi_2, \cdots, \pi_n) = (P(X_0 = s_1), P(X_0 = s_2), \cdots, P(X_0 = s_n))$$

Por lo tanto, para calcular

$$P(X_1 = s_j) = \sum_{k=1}^{n} P(X_1 = s_j | X_0 = s_k) P(X_0 = s_k)$$

$$= \sum_{k=1}^{n} p_{kj} \cdot \pi_k = \sum_{k=1}^{n} \pi_k \cdot p_{kj} = (\pi^T P)_j$$

De manera inductiva se puede probar que:

$$P(X_k = s_j) = (\pi^{\mathsf{T}} P^k)_j$$

Probabilidad conjunta (historia parcial)

Si queremos calcular la probabilidad conjunta de una trayectoria (también llamada *historia parcial*) $S_{(0)}, S_{(1)}, S_{(2)}, \ldots, S_{(t)}$ con $X_0 \sim \pi^T$ se tiene que

$$P\left(X_0 = s_{(0)}, X_1 = s_{(1)}, \ldots, X_t = s_{(t)}\right)$$

$$= P\left(X_t = s_{(t)} \mid X_{t-1} = s_{(t-1)}, \ldots, X_0 = s_{(0)}\right) \cdot P\left(X_{t-1} = s_{(t-1)}, \ldots, X_0 = s_{(0)}\right)$$

$$= P\left(X_t = s_{(t)} \mid X_{t-1} = s_{(t-1)}\right) \cdot P\left(X_{t-1} = s_{(t-1)}, \ldots, X_0 = s_{(0)}\right) \quad \text{(Markov Property)}$$

$$= p_{s_{(t-1)}, s_{(t)}} P\left(X_{t-1} = s_{(t-1)} \mid X_{t-2} = s_{(t-2)}, \ldots, X_0 = s_{(0)}\right) \cdot P\left(X_{t-2} = s_{(t-2)}, \ldots, X_0 = s_{(0)}\right)$$

$$= p_{s_{(t-1)}, s_{(t)}} \cdot p_{s_{(t-2)}, s_{(t-1)}} \cdot \ldots \cdot p_{s_{(0)}, s_{(1)}} \cdot P\left(X_0 = s_{(0)}\right)$$

$$= p_{s_{(t-1)}, s_{(t)}} \cdot p_{s_{(t-2)}, s_{(t-1)}} \cdot \ldots \cdot p_{s_{(0)}, s_{(1)}} \cdot \pi_{s_{(0)}}.$$

Clasificación de los estados

Definición 34 (Estado recurrente)**.** Un estado s_i es recurrente si, empezando desde el estado s_i, la probabilidad de que la cadena de Markov vuelva eventualmente al estado *si* es igual a 1. Es decir,

$$P(X_t = s_i \text{ para infinitos } t | X_0 = s_i) = P(\exists t \in \mathbb{N} : X_t = s_i | X_0 = s_i) = 1$$

Definición 35 (Estado transitorio)**.** Un estado s_i es transitorio si

$$P(\exists t \in \mathbb{N} : X_t = s_i | X_0 = s_i) < 1$$

Observación 1 En un estado s_i transitorio llamemos p a la probabilidad de que, empezando desde el estado s_i la cadena nunca vuelva al estado s_i. Por lo tanto, partiendo desde el estado s_i la cantidad de veces que la cadena vuelve al estado s_i antes de no volver nunca más tiene distribución Geométrica con parámetro p.

Observación 2 Un caso particular en el que sabremos que todos los estados serán recurrentes es cuando la cadena de Markov sea irreducible.

Definición 36 (Cadenas irreducibles y reducibles)**.** Una cadena de Markov con matriz de transición P **es irreducible si para dos estados cualesquiera** s_i y s_j **es posible ir del estado** i **al estado** j **en una cantidad finita de períodos con probabilidad positiva.** Es decir, si $\forall s_i$ y s_j existe un número $n \in \mathbb{N}$ de manera que la matriz $(P^n)_{ij} > 0$. **Una cadena** de Markov **que no es irreducible se dice reducible**.

Proposición 37. *Si la cadena de Markov es irreducible sucede que todos sus estados son transitorios (si tiene infinitos estados) o todos son recurrentes (si son finitos)*

Observación 38 (Consideración práctica)*.* Si no se aclaran las probabilidades de transición entonces son equiprobables (contar de cada estado cuantas flechitas salen)

Período de una cadena de Markov

Definición 39 (Período de un estado)**.** Sea $(X_t)_{t \geq 0}$ una cadena de Markov con espacio de estados S. Sea $s_i \in S$ un estado con la probabilidad de que existe $n \geq 1$ de manera que $(P^n)_{ii} > 0$. Definimos

$$c(i) = mcd\{n \geq 1 : (P^n)_{ii} > 0\}$$

donde *mcd* calcula el máximo común divisor del conjunto $n \in \mathbb{N}$ de las probabilidades $(P^n)_{ii} > 0$

Definición 40 (Periodicidad de estado)**.** Dado un estado $s_i \in S$ entonces

- si $c(i) \geq 2$ entonces s_i se llama estado periódico con período $c(i)$.
- si $c(i) = 1$ entonces s_i se llama estado aperiódico.

Proposición 41 (Periodicidad de una cadena de Markov)**.** *Dada una cadena de Markov, llamamos*

- ***Cadena de Markov periódica*** *con período c si todos los estados $s_i \in S$ tienen el mismo período c.*
- ***Cadena de Markov aperiódica*** *si todos los estados $s_i \in S$ son aperiódicos..*

Distribución estacionaria

Definimos una distribución estacionaria como r donde

$$r^T P = r^T$$

Queremos comprender el comportamiento a largo plazo de una cadena de Markov. A largo plazo, sabemos que la cadena de Markov solamente estará en estados que sean recurrentes pero **¿qué "fracción del tiempo pasará en cada uno de ellos"? Esto lo determina la distribución estacionaria.**

Proposición 42 (Existencia del estado estacionario)**.** *Para cualquier cadena de Markov existe al menos una distribución de estado estacionario. En general esa(s) distribucion(es) de estado estacionario podrían potencialmente tener probabilidades iguales a 0 y podrían no ser únicas.*

Demostración. Buscamos números λ de manera que

$$det(P - \lambda I) = 0 = det((P - \lambda I)^T) = det(P^T - \lambda I)$$

etnonces P y P^T tienen los mismos autovalores.

Considerando que r^T es la distribución estacionaria,

$$si \ \ \exists r^T/1 \ r^T = r^T P$$

(o sea r^T es autovector de P de autovalor $\lambda = 1$). Esto sucede

$$\iff \exists v/1 \ v^T = v^T P^T$$

(o sea v^T es autovector de P^T de autovalor $\lambda = 1$) y esto sucede

$$\iff (v^T)^T = (v^T P^T)^T \iff v = (v^T)^T (P^T)^T$$

$$\iff v = vP$$

en resumen $\exists r^T$ distribución estacionaria de $P \iff \exists v \in \mathbb{R}^{n \times 1}/v = Pv$. Podemos probar que este v existe pues hay infinitos v que cumplen esto por ende existe al menos un estado estacionario.

Por ejemplo consideremos $v = \begin{pmatrix} 1/n \\ \vdots \\ 1/n \end{pmatrix}$ como P tiene filas que suman 1

$$\implies P \begin{pmatrix} 1/n \\ \vdots \\ 1/n \end{pmatrix} = \begin{pmatrix} 1/n \\ \vdots \\ 1/n \end{pmatrix}$$

□

Proposición 43 (Existencia y unicidad de distribución estacionaria)**.** *Para una cadena de Markov irreducible existe una única distribución estacionaria. En esta distribución todos los estados tienen probabilidad positiva por el teorema de Perron-Frobenius.*

Proposición 44 (Convergencia a la distribución estacionaria)**.** *Si la cadena de Markov es **irreducible y aperiódica**, la distribución estacionaria es única y la distribución de largo* *plazo converge a la distribución estacionaria cuando $t \to \infty$. O sea*

$$\lim_{t\to\infty} P^t = \begin{pmatrix} \dots r^T \dots \\ \vdots \\ \dots r^T \dots \end{pmatrix}$$

Además, la distribución estacionaria es el único autovector de autovalor 1.

Hitting probabilities y expected hitting time

Definición 45 (Estado absorbente)**.** Sea $S = \{s_1, ..., s_n\}$ el conjunto de posibles estados en una cadena de Markov, llamamos $A \subseteq S$ de estados absorbentes si $\exists t_0 \in \mathbb{N} / P(s_t \in A|s_0) = 1 \forall t \geq t_0$.

Definición 46 (Hitting probabilities)**.** Consideremos una familia $A \subseteq S$ de nodos, definimos la **probabilidad de pegarle** a la familia A empezando desde el estado s_i como

$$h_{iA} = P(X_t \in A|X_0 = s_i)$$

y se puede interpretar como

- la probabilidad de quedarnos en estados "s" absorbentes a partir de s_i.
- la probabilidad de "pegarle"[29] a un conjunto A a partir de s_i.

[29] notar que pegarle es pegar e irse no que se queda.

Definición 47 (Expected hitting time)**.** Definimos el tiempo esperado de parada al conjunto A empezando desde el estado inicial s_i como $m_{iA} = E(\tau_A|X_0 = s_i)$ donde $\tau_A = \inf\{t \geq 0 : s_{(t)} \in A\}$ Además notar que vale que

$$m_{iA} = \begin{cases} 0 & \text{si} \quad s_i \in A \\ 1 + \sum_{j\notin A} p_{ij} m_{jA} & \text{si} \quad s_i \notin A \end{cases}$$

- el "tiempo de parada" condicional a s_0 el cual es $m_0 A = E(\tau|s_0)$, donde $\tau = \inf_t\{s_t \in A\}$
- el "hitting time" condicional a s_0 el cual es $m_0 A = E(\tau|s_0)$, donde $\tau = \inf_t\{s_t \in B\}$ donde $B \subseteq S$

Desigualdades

Desigualdad de Cauchy-Schwarz

Sean X e Y dos variables aleatorias. Entonces:[30]

$$E[XY]^2 \leq E[X^2]E[Y^2]$$

[30] Este resultado sirve para demostrar que el coeficiente de correlación $\rho \in [-1, 1]$. Esto se prueba tomando $X = \widetilde{X} - E(\widetilde{X})$ y $Y = \widetilde{Y} - E(\widetilde{Y})$ y aplicando la desigualdad.

Funciones cóncavas y convexas

Una función $g(\cdot)$ es **convexa** si para cualquiera x e y, $\alpha \in [0,1]$ vale que

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y)$$

Si $g(x)$ es una función derivable dos veces, es convexa si $g''(x) > 0$.

Una función $g(x)$ es **cóncava** si $-g(x)$ es convexa.

Ejemplos de funciones cóncavas o convexas:

- $g(x) = \ln(x)$ es cóncava pues $g''(x) = -\frac{1}{x^2} < 0$
- $g(x) = x^2$ es convexa pues $g''(x) = 2 > 0$
- $g(x) = e^{-x}$ es convexa pues $g''(x) = e^{-x} > 0$

Desigualdad de Jensen

Sea X una variable aleatoria y sea $g : \mathbb{R} \to \mathbb{R}$ una función convexa. Entonces:

$$E(g(X)) \geq g(E(X))$$

Sea X una variable aleatoria y sea $g : \mathbb{R} \to \mathbb{R}$ una función cóncava. Entonces:

$$E(g(X)) \leq g(E(X))$$

Entonces, usando Jensen, podemos afirmar que:

$$E(\ln(X)) \leq \ln(E(X)), \text{ tomando } g(x) = \ln(x)$$
$$E(X^2) \geq (E(X))^2, \text{ tomando } g(x) = x^2$$
$$E(e^{-x}) \geq e^{-E(X)}, \text{ tomando } g(x) = e^{-x}$$
$$E\left(\frac{1}{X}\right) \geq \frac{1}{E(X)}, \text{ tomando } g(x) = \frac{1}{x}$$

Desigualdad de Markov

Sea X una variable aleatoria, si $E(X) < \infty$, para todo $\varepsilon > 0$:

$$P(|X| \geq \varepsilon) \leq \frac{E(|X|)}{\varepsilon}$$

Entonces la Markov afirma que la probabilidad de que X asuma valores muy por encima de $E(X)$ es relativamente pequeña.

Las desigualdades de Markov, Cheby, Mill, Chernoff, Hoeffding sirven para acotar la probabilidad de que una variable aleatoria (bajo los supuestos de cada desigualdad) tome valores extremos en su distribución.

Ejemplo 48. Por ejemplo, si X es el ingreso mensual de un individuo elegido al azar de la población argentina y tomamos $\varepsilon = 2E(X)$, la desigualdad de Markov nos dice que:

$$P(X \geq 2E(X)) \leq \frac{1}{2}$$

Es decir, no es posible que más de la mitad de la población tenga un ingreso de al menos el doble del ingreso promedio.

Observación 49 (Resultado general). Si g es una función no negativa, $P(g(X) \leq 0) = 1$ donde $E(g(X)) < \infty$, luego

$$P(g(X) \geq \varepsilon) \leq \frac{E(g(X))}{\varepsilon}$$

Desigualdad de Chebyshev

Sea X una variable aleatoria, si $E(X) < \infty$ y $\text{Var}(X) < \infty$, para todo $\varepsilon > 0$:

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}$$

Equivalentemente, para todo $\varepsilon > 0$:

$$P\left(|X - E(X)| \geq \varepsilon\sqrt{\text{Var}(X)}\right) \leq \frac{1}{\varepsilon^2}$$

Notar que Markov vale para cualquier variable que cumpla E(X) sea finita y Chebyshev vale para cualquier vairable que cumpla Var(X) finita. Markov es un caso más general que Chebyshev.

Intuitivamente, la desigualdad de Chebyshev dice que $Var(X)$ nos da una cota de la probabilidad de que X tome valores alejados de su esperanza. Si $Var(X)$ es pequeña, entonces es poco probable que X tome valores alejados de $E(X)$.

Observación 50 (Para la distribución de medias muestrales). Para la variable aleatoria $\overline{X}_n$, que recordemos verifica $E(\overline{X}_n) = E(X)$ y $\text{Var}(\overline{X}_n) = \frac{\text{Var}(X)}{n}$, por Chebyshev se cumple que:

$$P\left(|\overline{X}_n - E(X)| \geq \epsilon\right) \leq \frac{\text{Var}(X)}{n} \cdot \frac{1}{\epsilon^2}$$

Interpretación: En otras palabras, cuando n es grande, la distribución de probabilidad de la variable aleatoria $\overline{X}_n$ está "concentrada" en torno a la constante $\mu = E(X)$. **¡Notar que este resultado no dice nada sobre ninguna realización de la media muestral en particular!**

Desigualdad de Hoeffding

Si $X_1, \ldots, X_n$ son independientes, $E(X_i) = 0$ para cualquier $i = 1, \ldots, n$, además $a_i \leq X_i \leq b_i$ para cualquier $i = 1, \ldots, n$, entonces, dado un valor de $\epsilon > 0$, vale que:

$$P\left(\sum_{i=1}^{n} X_i \geq \varepsilon\right) \leq e^{-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

Observación 51 (Caso particular de Bernoulli). En particular, si $X_i \sim_{iid} Be(p)$, usando la desigualdad de Hoeffding para las variables $Y_i = X_i - p$, vale que

$$P\left(|\overline{X}_n - p|\right) \leq 2e^{-2n\varepsilon^2}$$

donde $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

Desigualdad de Mill

Si $Z \sim N(0, 1)$ entonces

$$P(|Z| \geq \varepsilon) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{\varepsilon^2}{2}}}{\varepsilon}.$$

Observación 52 (Para la distribución de medias muestrales). Notar que también se puede decir que si $X_i \sim_{iid} N(0, 1)$, entonces $\overline{X}_n\sqrt{n} \sim N(0, 1)$. y por lo tanto,

$$P(|\overline{X}_n| \geq \varepsilon) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{n\varepsilon^2}{2}}}{\sqrt{n}\varepsilon}.$$

Desigualdad de Chernoff

Sea X una variable aleatoria y constantes $a > 0$ y $t > 0$

Notar que esta también es un caso particular de Markov

$$P(X \geq a) \leq \frac{E\left(e^{tX}\right)}{e^{ta}}$$

Comparación de cotas

Comparemos las tres cotas para la probabilidad $P(|Z| > 1{,}96)$ si $Z \sim N(0,1)$ usando las desigualdades de Markov, Chebyshev, Mill y Chernoff y calculando la probabilidad.

- **Markov:** $P(|Z| > 1{,}96) \leq 0{,}4$
- **Chebyshev:** $P(|Z| > 1{,}96) \leq \frac{1}{1{,}96^2} \approx 0{,}26$
- **Mill:** $P(|Z| > 1{,}96) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-1{,}96^2/2}}{1{,}96} \approx 0{,}0596$
- **Chernoff:** $P(|Z| > 1{,}96) \leq 2 \cdot e^{\frac{(t^*)^2}{2} - 1{,}96 t^*} \approx 0{,}293$
- **Cálculo exacto:** $P(|Z| > 1{,}96) = 0{,}05$

Donde en la desigualdad de Chernoff usamos que la normal es una distribución simétrica respecto del valor $z = 0$ y que el valor t^* que minimiza esa cota es $t^* = 1{,}96$.

Sucesiones y convergencias

Sucesiones de números

Una sucesión de números $(a_n)_{n \in \mathbb{N}}$ es una lista de números indexados por $n \in \mathbb{N}$.

Decimos que $(a_n)_{n \in \mathbb{N}}$ es **convergente** si existe $\lim_{n \to \infty} a_n = a$. Es decir, existe un número a tal que para cualquier $\varepsilon > 0$ que se considere, existe $n(\varepsilon)$ de manera que si $n > n(\varepsilon)$ entonces

$$|a_n - a| < \varepsilon$$

Sucesiones estocásticas

Una sucesión estocástica o sucesión de variables aleatorias $(Y_n)_{n \in \mathbb{N}}$ es una lista de variables aleatorias indexadas por $n \in \mathbb{N}$.

Observación 53 (Para la media muestral). Por ejemplo, si $Y_n = \overline{X}_n$

$$(\overline{X}_n)_{n \in \mathrm{N}} = \left(X_1, \frac{X_1 + X_2}{2}, \frac{X_1 + X_2 + X_3}{3}, \dots \right)$$

Observación 54 (Para una medida en general). Más en general, $Y_n = \widehat{\theta}_n$

$$(\widehat{\theta}_n)_{n \in \mathrm{N}} = \left(\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3, ... \right)$$

Muestra aleatoria y media muestral

Llamamos una ***muestra aleatoria*** de v.a. $X_1, \dots, X_n$ si las variables $\{X_i\}_{i=1}^n$ son variables aleatorias independientes e idénticamente distribuidas.

Definición 55 (Media muestral)**.** Definimos a la variable aleatoria $\overline{X}_n$ la ***media muestral*** definida por

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\left(X_1 + \cdots + X_n\right)$$

Proposición 56 (Media muestral y media poblacional)**.** *La esperanza de las medias muestrales es igual a la media poblacional.*[31]

$$E(\overline{X}_n) = E(X)$$

Definición 57 (Varianza muestral)**.** Definimos a la variable aleatoria S_n^2 la ***varianza muestral*** definida por

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\left(\overline{X}_n\right)^2$$

Proposición 58 (Varianza de la media muestral y varianza poblacional)**.** *La varianza de las medias muestrales es igual a la varianza poblacional.*[32]

$$Var(\overline{X}_n) = \frac{Var(X)}{n}$$

Proposición 59 (Esperanza de la varianza muestral y varianza poblacional)**.** *La esperanza de la varianza muestral es igual a la varianza poblacional.*

$$E(S_n^2) = Var(X)$$

[31] Para esta prueba notar lo siguiente

$$E(\overline{X}_n) = \frac{1}{n}E(X_1 + \cdots + X_n) =$$

$$= \frac{1}{n}\left(E(X_1) + \dots E(X_n)\right) \underbrace{=}_{X_i id} E(X)$$

[32] Para esta prueba notar lo siguiente

$$Var(\overline{X}_n) = \frac{1}{n^2}Var(X_1 + \cdots + X_n)$$

$$\underbrace{=}_{X_i indep} \frac{1}{n^2}\left(Var(X_1) + \dots Var(X_n)\right) \underbrace{=}_{X_i id} \frac{Var(X)}{n}$$

Ley de los grandes números

Ley de los grandes números débil

Sea $X_1, \dots, X_n$ una sucesión de variables aleatorias (que provienen de una muestra) independientes e idénticamente distribuidas (i.i.d.) con tal que $E(X) = \mu$. Entonces para todo $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| \geq \varepsilon\right) = 0$$

Notacionalmente escribimos $\overline{X}_n \xrightarrow{p} \mu$ o plim $\overline{X}_n = \mu$.

Notar que la ley de los grande números débil utiliza la convergencia en probabilidad

Notemos que la conclusión es más general y débil que la que obtuvimos antes usando Chebyshev, pero también que las hipótesis son más débiles ya que para que valga la LGN no requerimos que exista Var(X). Con Cheby al pedir que exista Var(X) sabemos que la probabilidad se achica con velocidad $\frac{1}{n}$.

Observación 60 (Interpretación frecuentista)*.* Si n es grande con probabilidad alta el estimador $\overline{X}_n$ tomará valores "cerca" de $E(X)$. Notemos que es una propiedad deseable de $\overline{X}_n$ como estimador de μ.

Ley de los grandes números fuerte

Sea $\{X_n\}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas (i.i.d.) con $E[X_i] = \mu$. Entonces si $E(X) < \infty$ para cada elemento ω se intenta calcular $\lim_{n\to\infty} \overline{X}_n(\omega)$ y se define la variable aleatoria $\lim_{n\to\infty} \overline{X}_n$. Entonces vale que

$$P\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1$$

Notar que la ley de los grande números fuerte utiliza la convergencia casi seguro

Comparación

Notemos las diferencias entre LGN débil y LGN fuerte:

- En la ley de grandes números fuerte se afirma que existe la v.a. $\lim_{n\to\infty} \overline{X}_n$ y que ésta tomando el valor μ con probabilidad 1.
- Se intercambia el orden en el que se consideran las operaciones tomar probabilidad y límite.
- Por eso, es que la LGN débil se refiere una convergencia de probabilidades (números) mientras que la LGN fuerte se refiere a una convergencia de variables aleatorias.

Convergencia de variables aleatorias

Convergencia en Distribución

Sea $V_1, \ldots, V_n, \ldots$ una sucesión de variables aleatorias y sea V una variable aleatoria. **Decimos que V_n converge en distribución a** V si, para todo v tal que la función de distribución acumulada de V, F_V es continua en v, vale que

$$F_{V_n}(v) = P(V_n \leq v) \xrightarrow{n\to\infty} P(V \leq v) = F_V(v).$$

La convergencia en distribución, más que una convergencia de variables aleatorias, es ***una convergencia de funciones de distribución***.[33]

Observación 61. **Que converga en distribución no quiere decir que V_n esté cerca de V!** De hecho, $V_1, \ldots, V_n$ y V pueden estar definidas todas sobre espacios muestrales distintos!

Observación 62 (Notación). Notamos $V_n \xrightarrow{d} V$. También haremos el siguiente abuso de notación. Por ejemplo, si $V \sim N(0,1)$, escribiremos $V_n \xrightarrow{d} N(0,1)$.

[33] Si $V_n \xrightarrow{d} V$, la función de distribución acumulada de V_n está "cerca" de la distribución de V para valores de n grandes.

Convergencia en Probabilidad

Convergencia a un número

Sea $V_1, ..., V_n$ una sucesión de variables aleatorias. Decimos que la sucesión $(V_n)_{n\in\mathbb{N}}$ converge en probabilidad a $a \in \mathbb{R}$ si

$$\lim_{n\to\infty} P\left(\{\omega \in \Omega : |V_n(\omega) - a| > \varepsilon\}\right) = 0$$

Escribimos $V_n \xrightarrow{p} a$ para denotar convergencia en probabilidad.[34]

[34] En la LGN débil $V_n = \overline{X}_n$ y $a = E(X)$

Convergencia a una variable aleatoria

Decimos que la sucesión $(V_n)_{n\in\mathbb{N}}$ converge en probabilidad a V (una variable aleatoria) si

$$\lim_{n\to\infty} P\left(\{\omega \in \Omega : |V_n(\omega) - V(\omega)| > \varepsilon\}\right) = 0$$

Escribimos $V_n \xrightarrow{p} V$. Notar que en cierta bibliográfia se escribe $plimV_n = V$.[35]

[35] Notemos que si $V_n \xrightarrow{p} V$ y $V_n \xrightarrow{p} W$ entonces $P(V = W) = 1$

Relación con convergencia en distribución

Si se cumple convergencia en probabilidad en variables aleatorias entonces también se cumple en distribución. Si $V_n \xrightarrow{p} V$ entonces $V_n \xrightarrow{d} V$. Pero el recíproco no es necesariamente verdadera.

Propiedades

Sea a_n es una sucesión de números reales tal que $a_n \xrightarrow{n\to\infty} a$. Sean V_n y W_n dos variables aleatorias tales que $V_n \xrightarrow{P} V$ y $W_n \xrightarrow{P} W$. Entonces:

1. $a_n + V_n \xrightarrow{P} a + V$.
2. $a_n V_n \xrightarrow{P} aV$.
3. $V_n + W_n \xrightarrow{P} V + W$.
4. $V_n W_n \xrightarrow{P} VW$.
5. Si $P(W = 0) = 0$, $V_n/W_n \xrightarrow{P} V/W$.
6. Si $f : \mathbb{R} \to \mathbb{R}$ es continua luego $f(V_n) \xrightarrow{P} f(V)$.
7. Si $f : \mathbb{R}^2 \to \mathbb{R}$ es continua luego $g(V_n, W_n) \xrightarrow{P} g(V, W)$.

Convergencia en Media Cuadrática

Convergencia a un número

Decimos que las variables V_n convergen en media cuadrática al número a, si: [36]

$$\lim_{n\to\infty} E\left[(V_n - a)^2\right] = 0$$

Escribimos $V_n \xrightarrow{m.c.} a$ para denotar convergencia en media cuadrática.

[36] Es equivalente a que cumpla que

1. $E(V_n) \xrightarrow{n\to\infty} a$
2. $\mathrm{Var}(V_n) \xrightarrow{n\to\infty} 0$

Convergencia a una variable aleatoria

Decimos que las variables V_n convergen en media cuadrática a V (una variable aleatoria), si:

$$\lim_{n\to\infty} E\left[(V_n - V)^2\right] = 0$$

Escribimos $V_n \xrightarrow{m.c.} V$ para denotar convergencia en media cuadrática.

Relación con convergencia en probabilidad

Si se cumple convergencia en media cuadrática entonces también se cumple en probabilidad. Si $V_n \xrightarrow{m.c.} a$ entonces vale $V_n \xrightarrow{p} a$. Del mismo modo sucede con la convergencia a una variable aleatoria si $V_n \xrightarrow{m.c.} V$ entonces vale $V_n \xrightarrow{p} V$.

Intuición: Si el centro del histograma (densidad) de V_n se está acercando a a y la variabilidad de V_n se está acercando a cero cuando n crece, entonces toda la distribución de probabilidad de V_n se está concentrando sobre a cuando n crece (por tanto lo tanto, $V_n \xrightarrow{P} a$).

Convergencia Casi Segura

Convergencia a un número

Sea $V_1, \ldots, V_n$ una sucesión de variables aleatorias. Decimos que la sucesión $(V_n)_{n\in\mathbb{N}}$ **converge casi seguro** a $a \in \mathbb{R}$ si

$$P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} V_n(\omega) = a\right\}\right) = 1$$

Análogamente podríamos escribir,

$$P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} V_n(\omega) \neq a\right\}\right) = 0$$

Escribimos $V_n \xrightarrow{\text{c.s.}} a$ o $V_n \xrightarrow{\text{a.s.}} a$ para denotar convergencia casi seguro.

Convergencia a una variable aleatoria

Similarmente $V_n \xrightarrow{\text{a.s.}} V$ si $V_n - V \xrightarrow{\text{a.s.}} 0$, es decir,

$$P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} V_n(\omega) - V(\omega) = 0\right\}\right) = 1$$

Relación con convergencia en probabilidad

Si se cumple convergencia casi seguro entonces también se cumple en probabilidad. Vale que si $V_n \xrightarrow{\text{c.s.}} a$ entonces $V_n \xrightarrow{p} a$. No vale la implicación recíproca. Es decir, si $V_n \xrightarrow{p} a$ podría llegar a suceder que $V_n \not\xrightarrow{\text{c.s.}} a$.

Lema de Borel-Cantelli

Para demostrar que variables aleatorias V_n convergen casi seguro a un número a, se puede utilizar el lema de **Borel-Cantelli** que dice:

Lema 63 (Borel-Cantelli)**.** *Para cada $\varepsilon > 0$ definimos A_n eventos en el espacio de probabilidad (Ω, Σ, P) como*

$$A_n = \{\omega \in \Omega : |V_n(\omega) - a| > \epsilon\}$$

Si $\sum_{n=1}^{\infty} P(A_n) < \infty$ para cualquier $\varepsilon > 0$, entonces

$$P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} V_n(\omega) \neq a\right\}\right) = 0 \quad (V_n \xrightarrow{a.s.} a).$$

Hay otro resultado, conocido como el **recíproco de Borel-Cantelli**[37] que dice:

Lema 64 (recíproco de Borel-Cantelli)**.** *Si $\sum_{n=1}^{\infty} P(A_n) = \infty$ y los A_n son independientes, entonces*

$$P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} V_n(\omega) \neq a\right\}\right) = 1.$$

[37] Esto quiere decir que si la suma de las probabilidades de los eventos diverge entonces no converge en ningún lugar

Teorema Central del Límite

Proposición del TCL

Sea $X_1, \ldots, X_n \overset{\text{iid}}{\sim} X$ con $E(X)$ y $\text{Var}(X)$, y sea

$$W_n = \frac{\overline{X}_n - E(X)}{\sqrt{\frac{\text{Var}(X)}{n}}} = \frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{\text{Var}(\overline{X}_n)}}$$

Notar que en este caso $E(\overline{X}_n)$ es la esperanza de la distribución de las medias muéstrales y $\sqrt{\text{Var}(\overline{X}_n)}$ es el desvío estándar de la distribución de las medias muéstrales o dicho de otro modo el error estándar

para todo $z \in \mathbb{R}$ vale que

$$\lim_{n\to\infty} P\left(W_n \leq z\right) = P(Z \leq z)$$

donde $P(Z \leq z)$ con $Z \sim N(0,1)$. A veces se escribe $\Phi(z) = P(Z \leq z)$. . Entonces escribimos

$$W_n = \frac{\left(\overline{X}_n - E(X)\right)}{\sqrt{\frac{\text{Var}(X)}{n}}} \xrightarrow{d} N(0,1)$$

Esto quiere decir que la distribución de W_n converge en distribución a $N(0,1)$

Observaciones sobre el TCL

Podemos notar lo siguiente sobre el TCL

- El TCL implica que si $X_i \sim_{iid}$ entonces para cualesquiera $z_1, z_2 \in \mathbb{R}$ que cumplen $z_1 \leq z_2$ vale

$$P\left(z_1 \leq \frac{\left(\overline{X}_n - E(X)\right)}{\sqrt{\frac{\text{Var}(X)}{n}}} \leq z_2\right) \underset{n\to\infty}{\longrightarrow} P(Z \leq z_2) - P(Z \leq z_1)$$

- El TCL permite cuantificar la incertidumbre (dar un margen de error que tenemos sobre nuestra estimación de $E(X)$ usando la media muestral $\overline{X}_n$.
- El teorema se refiere a la distribución de $\overline{X}_n$, no a la de $X_1, ..., X_n$.
- Si $X_i \sim_{iid} N(\mu, \sigma^2)$, no es necesario usar TCL porque vale que, para cualquier n que: $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

Notación correcta sobre el TCL

Podemos distinguir la notación **para la distribución asintótica** (versión estandarizada)

$$\frac{\left(\overline{X}_n - E(X)\right)}{\sqrt{\frac{\text{Var}(X)}{n}}} \xrightarrow{d} N(0,1)$$

También podemos distinguir la notación **para la distribución aproximada** para tamaños de muestra grande ($n \gg 0$), la distribución de $\overline{X}_n$ es aproximadamente normal. Primero para su versión estandarizada

$$\frac{\left(\overline{X}_n - E(X)\right)}{\sqrt{\frac{\text{Var}(X)}{n}}} \sim_a N(0,1)\; si\; n \gg 0$$

Luego para su versión no estandarizada

$$\overline{X}_n \sim_a N\left(E(X), \frac{\text{Var}(X)}{n}\right)\; si\; n \gg 0$$

Propiedades de la convergencia en distribución

Lema de Slutsky

Sea $c \in \mathbb{R}$ una constante. Si $W_n \xrightarrow{p} c$ y $V_n \xrightarrow{d} V$ entonces

$$W_n V_n \xrightarrow{d} cV \quad \text{y} \quad V_n + W_n \xrightarrow{d} V + c$$

Teorema de la función continua

Las funciones continuas conservan los límites aún cuando sus argumentos son (sucesiones de) variables aleatorias. En otras palabras, si $g : \mathbb{R} \to \mathbb{R}$ es una función continua y:

$$V_n \xrightarrow{p} c \text{ entonces } g(V_n) \xrightarrow{p} g(c).$$

$$V_n \xrightarrow{d} V \text{ entonces } g(V_n) \xrightarrow{d} g(V).$$

Ejemplo 65. Sea $W_n = (\sqrt{S^2})^{-1} \xrightarrow{p} (\sqrt{Var(x)})^{-1}$ y sea $V_n = \sqrt{n}(\overline{X} - E(x)) \xrightarrow{d} N(0, Var(x))$

Ojo! No es cierto que si $V_n \xrightarrow{d} V$ y $W_n \xrightarrow{d} W$ entonces $V_n \cdot W_n \xrightarrow{d} V \cdot W$ o $V_n + W_n \xrightarrow{d} V + W$. Por ejemplo, si $V_n = W_n \sim N(0,1)$ y $V = W \sim N(0,1)$ se tiene que $V_n, W_n \xrightarrow{d} V, W$, pero $V_n + W_n \sim N(0,4)$ mientras que $V + W = 0$.

Implicaciones de las convergencias

Podemos resumir las implicaciones de las convergencias del siguiente modo

$$\begin{array}{ccc}
V_n \xrightarrow{c.s.} V & & \\
& \underbrace{\Searrow}_{(2)} & \\
\underbrace{\not\Nwarrow}_{(4)}, \underbrace{\not\Swarrow}_{(5)} & & V_n \xrightarrow{P} V \underbrace{\Longrightarrow}_{(3)} V_n \xrightarrow{d} V \\
& \underbrace{\Nearrow}_{(1)} & \\
V_n \xrightarrow{m.c.} V & &
\end{array}$$

Método delta univariado

Supongamos que V_n cumple que

$$\sqrt{n}(V_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

donde μ y σ^2 son constantes. Supongamos que $g(x)$ es una función con derivada continua y que $g'(\mu) \neq 0$. Entonces

$$\sqrt{n}(g(V_n) - g(\mu)) \xrightarrow{d} N(0, \sigma^2[g'(\mu)]^2)$$

.

Apéndice I: Distribución de variables aleatorias discretas

Distribución Bernoulli

Una variable aleatoria con distribución Bernoulli es una variable aleatoria binaria que toma dos valores posibles, 0 y 1, con probabilidades $1 - p$ y p respectivamente. Su función de probabilidad puntual es:

$$p_X(1) = p, \quad p_X(0) = 1 - p, \quad pX(x) = 0\,(si\, x \neq 0\, y\, x \neq 1)$$

La función de probabilidad puntual se puede reescribir como:

$$P(X=x)=p_X(x)=\begin{cases} p^x(1-p)^{1-x} & \text{si } x=1 \quad \text{ó} \quad x=0 \\ 0 & \text{en otro caso} \end{cases}$$

Se denota a su distribución como $Be(p)$.

Las variables de Bernoulli se pueden escribir como variables indicadoras: Si A es un evento, la variable indicadora I_A toma el valor 1 si A ocurre y 0 si A no ocurre. De este modo, $I_A(\omega)=1$ si $\omega \in A$ y $I_A(\omega)=0$ si $\omega \notin A$. Es decir que I_A es una variable Bernoulli con parámetro $p=p(A)$.

Ejemplo: De un envío grande de mercadería, elegimos un lote al azar, y nos interesa averiguar si en el lote hay alguna mercadería fallada. Entonces $X=1$ si el lote tiene una mercadería fallada y $X=0$ en caso contrario.

Notemos que la varianza de una variable $X \sim Be(p)$ depende de p. Notemos que $\mathrm{Var}(X)=p(1-p)\leq 0{,}25$ para cualquier $p\in[0,1]$.

Distribución Binomial

Una variable aleatoria se llama binomial de parámetros n y p si es igual al número total de éxitos en n ensayos de Bernoulli independientes. Si X denota una variable aleatoria con distribución $Bin(n,p)$, su función de probabilidad de masa es:

$$p_X(x)=\begin{cases} \binom{n}{x}p^x(1-p)^{n-x} & \text{si } x=0,1,2,...,n. \\ 0 & \text{en otro caso} \end{cases}$$

Donde $\binom{n}{x}$ es el coeficiente binomial, que se define como:

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

Esta fórmula representa la probabilidad de obtener x éxitos en n ensayos, donde cada ensayo tiene probabilidad de éxito p y probabilidad de fracaso $1-p$.

La distribución binomial se denota como $Bin(n,p)$.

Distribución Geométrica

Una variable aleatoria geométrica $Ge(p)$ cuenta la cantidad total de ensayos de Bernoulli, con probabilidad de éxito p, que se realizan para obtener el primer éxito. Su función de probabilidad de masa es:

$$P(X=x)=p_X(x)=\begin{cases} p(1-p)^{x-1} & \text{si } x\in\{1,2,3,\dots\} \\ 0 & \text{en otro caso} \end{cases}$$

Esta función representa la probabilidad de que se necesiten x ensayos para obtener el primer éxito en una secuencia de ensayos independientes, donde cada ensayo tiene probabilidad de éxito p y probabilidad de fracaso $1-p$.

La distribución geométrica se denota como $Ge(p)$.

Distribución Binomial Negativa

Una variable aleatoria binomial negativa $BN(p,r)$ cuenta la cantidad total de ensayos de Bernoulli, con probabilidad de éxito p, necesarios para obtener r éxitos. Su función de probabilidad de masa es:

$$p_X(x) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r}, & \text{si } x \in \{r, r+1, r+2, \ldots\} \\ 0, & \text{en otro caso} \end{cases}$$

Esta función representa la probabilidad de que se necesiten x ensayos para obtener r éxitos en una secuencia de ensayos independientes, donde cada ensayo tiene probabilidad de éxito p y probabilidad de fracaso $1-p$.

Es importante notar cómo se define una variable con distribución negativa ya que difiere según la bibliografía que se consulte. Es una generalización de la variable geométrica, ya que $BN(p,1) = Ge(p)$.

Notemos que podríamos escribir que si $X \sim BN(p,r)$ y $W_i \sim Ge(p)$ independientes, con $i = 1, 2, \ldots, r$, entonces:

$$X = W_1 + W_2 + \ldots + W_r$$

Distribución Hipergeométrica

Una variable aleatoria hipergeométrica $H(n,r,m)$ cuenta la cantidad de bolillas azules x que se obtienen de una urna cuando se extraen m bolillas al azar sin reemplazo. En la urna hay un total de n bolillas, r de las cuales son de color azul y $n-r$ son de color rojo. Su función de probabilidad de masa es:

Donde $\binom{a}{b}$ representa el coeficiente binomial, que se define como:

$$\binom{a}{b} = \frac{a!}{b!(a-b)!}$$

$$P_X(x) = \begin{cases} \frac{\binom{r}{x}\binom{n-r}{m-x}}{\binom{n}{m}} & \text{si } x \in \{0, 1, 2, \ldots, r\} \\ 0 & \text{en otro caso} \end{cases}$$

Esta función representa la probabilidad de obtener exactamente x bolillas azules en una muestra de m bolillas extraídas de la urna sin reemplazo.

Aproximación de la distribución binomial a la distribución hipergeométrica: Si la cantidad total de bolillas n cumple que $r \gg m$, se puede aproximar la probabilidad de que $X \sim H(n,r,m)$ tome el valor x como:

$$p_X(x) \approx \binom{m}{x} \left(\frac{r}{n}\right)^x \left(1 - \frac{r}{n}\right)^{m-x}$$

Esta aproximación se basa en la idea de que la probabilidad de extraer una bolilla azul es constante en cada ensayo y que las extracciones son independientes.

Distribución de Poisson

Una variable aleatoria se llama Poisson de parámetro λ, $Pois(\lambda)$, si puede tomar cualquier valor en el conjunto $\{0, 1, 2, 3, \ldots\}$ y su función de masa de probabilidad es:

$$p_X(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{si } x \in \{0, 1, 2, 3, \ldots\} \\ 0 & \text{en otro caso} \end{cases}$$

Donde λ es un número real positivo.

Las variables de Poisson modelan el número de veces que ocurre algún evento en una franja de tiempo (o espacio) bajo algunos supuestos. Por ejemplo, puede representar el número de errores de tipeo en una página, el número de llamados que llegan a un call-center, o el número de autos que pasan por una calle en un minuto.

Aproximación de Poisson por una binomial: Si $n \gg 0$ y $p := \frac{\lambda}{n}$ se verifica que

$$p_X(x) = e^{-\lambda}\frac{\lambda^x}{x!} \approx \binom{n}{x} p^x (1-p)^{n-x}$$

O sea, si $n \geq 100$ y $p \leq \frac{1}{100}$, $\lambda = np$ vale que $\text{Pois}(\lambda) \approx \text{Bin}(n, \lambda/n)$

Resumen de distribuciones discretas

Resumen de momentos de variables aleatorias

Distribución	Esperanza	Varianza
Be(p)	$E(X) = p$	$\text{Var}(X) = p(1-p)$
Bin(n, p)	$E(X) = np$	$\text{Var}(X) = np(1-p)$
Ge(p)	$E(X) = \frac{1}{p}$	$\text{Var}(X) = \frac{1-p}{p^2}$
BN(n, r)	$E(X) = \frac{r}{p}$	$\text{Var}(X) = \frac{r(1-p)}{p^2}$
H(n, r, m)	$E(X) = \frac{mr}{n}$	$\text{Var}(X) = \frac{(m-1)(r-1)nr+1-mn}{n^2(n-1)}$
Poi(λ)	$E(X) = \lambda$	$\text{Var}(X) = \lambda$

Resumen de funciones de probabilidad puntual (PMF) y FGM

Distribución	F. de masa de prob.	F. generatriz de momentos
Be(p)	$P(x) = p(1-p)^{1-x}$	$M_X(t) = pe^t + 1 - p$
Bin(n, p)	$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$	$M_X(t) = (pe^t + 1 - p)^n$
Ge(p)	$P(x) = p(1-p)^{x-1}$	$M_X(t) = \frac{pe^t}{1-(1-p)e^t}$
BN(p, r)	$P(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$	$M_X(t) = (\frac{pe^t}{1-(1-p)e^t})^r$
H(n, r, m)	$P(x) = \frac{\binom{r}{x}\binom{n-r}{m-x}}{\binom{n}{m}}$	$M_X(t) = \sum_{x=0}^{r} \frac{\binom{r}{x}\binom{n-r}{m-x}}{\binom{n}{m}} e^{tx}$
Poi(λ)	$P(x) = e^{-\lambda} \frac{\lambda^x}{x!}$	$M_X(t) = e^{\lambda(e^t-1)}$

Apéndice II: Distribución de variables aleatorias continuas

Distribución Uniforme

Una variable aleatoria continua X tiene distribución uniforme sobre un intervalo $E = [a, b]$, $X \sim \text{Unif}[a, b]$, cuando su función de densidad de probabilidad es:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{de otro modo} \end{cases}$$

La función de distribución acumulada es:

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

Si $a = 0$ y $b = 1$, $X \sim U[0, 1]$, es un caso particular de la distribución Beta cuando $\alpha = 1$ y $\beta = 1$. Es decir $U[0, 1] \equiv Beta(1, 1)$.

La distribución uniforme está caracterizada por dos parámetros a y b. Notemos que se puede escribir la PDF como $f(x) = \frac{1}{b-a} I_{[a,b]}(x)$

Es importante destacar que si $X \sim U(a, b)$ o $X \sim U(a, b]$ o $X \sim U[a, b)$, solo afecta al soporte de X, pero no a la expresión de la función de densidad.

Distribución exponencial

Una variable aleatoria continua X tiene distribución exponencial en el intervalo $(0, \infty)$, $X \sim \text{Exp}(\lambda)$, cuando su función de densidad de probabilidad es:

Donde λ es el parámetro de escala de la distribución, que representa la tasa promedio de ocurrencia de eventos. A medida que λ es mayor, la función de densidad cae más abruptamente.

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

La función de distribución acumulada es:

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

Una característica particular de la distribución exponencial es que representa procesos que no tienen memoria. Eso quiere decir que:

$$P(T \geq t + s | T \geq s) = P(T \geq t) = e^{-\lambda t}$$

En la práctica, esta distribución se utiliza en modelos relacionados a tiempos de espera o vida útil bajo ciertos supuestos. La aplicación clásica de esta distribución es la de modelar el tiempo de vida de algún componente electrónico.

En economía, una aplicación podemos, bajo ciertos supuestos, modelizar el tiempo de duración del desempleo como una variable exponencial.

Relación entre exponencial y poisson

La relación entre la variable aleatoria Poisson y la exponencial se puede ver en el siguiente ejemplo: imagina que N mide la cantidad de personas que llegan (de manera independiente) a una parada de colectivo a una tasa de λ personas por hora. La cantidad promedio de personas que llegan en t horas es $\lambda \cdot t$. La variable aleatoria T que mide el tiempo entre la llegada de dos personas consecutivas es una variable aleatoria que depende de $N \sim Poi(\lambda)$ la variable aleatoria Poisson. Entonces $T \sim Exp(\lambda)$ La función de densidad de probabilidad de T es $\lambda e^{-\lambda t}$.

Distribución Normal

Una variable aleatoria continua X tiene distribución normal o Gaussiana en el intervalo $(-\infty, +\infty)$, $X \sim N(\mu, \sigma^2)$, cuando su función de densidad de probabilidad es:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

La función de distribución acumulada es:

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(u-\mu)^2} du$$

Las características de $f(x)$ son:

- Toma valores positivos en toda la recta real (pero asigna valores muy pequeños a valores de x que difieren de μ en más de $\pm 3\sigma$).
- Tiene un solo máximo global en $x = \mu$ y es simétrica alrededor de μ.
- El valor de σ controla la dispersión de la curva alrededor de $x = \mu$. Tiene puntos de inflexión (cambio de concavidad) en $x = \mu - \sigma$ y $x = \mu + \sigma$.
- Cuando $\mu = 0$ y $\sigma = 1$, X tiene una **distribución normal estándar** $N(0, 1)$.

Para una variable aleatoria con distribución $N(\mu, \sigma^2)$, la integral

$$\int_a^b f_X(x)dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

no se puede calcular "a mano". Se calcula numéricamente.

Distribución Gamma

Una variable aleatoria continua X tiene distribución Gamma en el intervalo $(0, +\infty)$, $X \sim \Gamma(\alpha, \lambda)$, cuando su función de densidad de probabilidad es:

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

donde $\Gamma(\alpha)$ es la función Gamma que está definida como:

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

En particular, $\Gamma(n) = (n-1)!$ si n es un número entero positivo.

Una propiedad importante de la función Gamma es que:

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$

La distribución Gamma generaliza a la distribución exponencial, con $Exp(\lambda) \equiv \Gamma(1, \lambda)$. También, la distribución chi-cuadrado con $\chi^2(\nu)$ es un caso particular con $\Gamma(\nu/2, 1/2)$.

Distribución Beta

Una variable aleatoria continua X tiene distribución Beta en el intervalo $[0, 1]$, $X \sim \text{Beta}(\alpha, \beta)$, con parámetros $\alpha > 0$ y $\beta > 0$, cuando su función de densidad de probabilidad es:

$$f_X(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} & \text{si } 0 \leq x \leq 1 \\ 0 & \text{de otro modo} \end{cases}$$

donde $B(\alpha, \beta)$ es la función Beta, definida como:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

y $\Gamma(\cdot)$ es la función Gamma.

La distribución Beta es flexible para modelar variables aleatorias en el intervalo [0,1] y tiene varias aplicaciones, como en pruebas Bayesianas o modelado de proporciones.

Resumen de distribuciones continuas

Resumen de momentos de variables aleatorias continuas

Distribución	Esperanza	Varianza
Unif(a, b)	$E(X) = \frac{a+b}{2}$	$\text{Var}(X) = \frac{(b-a)^2}{12}$
Exp(λ)	$E(X) = \frac{1}{\lambda}$	$\text{Var}(X) = \frac{1}{\lambda^2}$
N(μ, σ^2)	$E(X) = \mu$	$\text{Var}(X) = \sigma^2$
Gamma(α, λ)	$E(X) = \frac{\alpha}{\lambda}$	$\text{Var}(X) = \frac{\alpha}{\lambda^2}$
Beta(α, β)	$E(X) = \frac{\alpha}{\alpha+\beta}$	$\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Resumen de funciones de densidad de probabilidad (PDF) y FGM

Distribución	F. de densidad de prob. (PDF)	F. generatriz de momentos (MGF)
Unif(a, b)	$f(x) = \frac{1}{b-a}$	$M_X(t) = \frac{e^{tb}-e^{ta}}{t(b-a)}$
Exp(λ)	$f(x) = \lambda e^{-\lambda x}$	$M_X(t) = \frac{\lambda}{\lambda-t}, \quad t < \lambda$
N(μ, σ^2)	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$M_X(t) = e^{\mu t+\frac{\sigma^2 t^2}{2}}$
Gamma(α, λ)	$f(x) = \frac{\lambda^\alpha x^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)}$	$M_X(t) = \left(\frac{\lambda}{\lambda-t}\right)^\alpha, \quad t < \lambda$
Beta(α, β)	$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	No tiene FGM cerrada