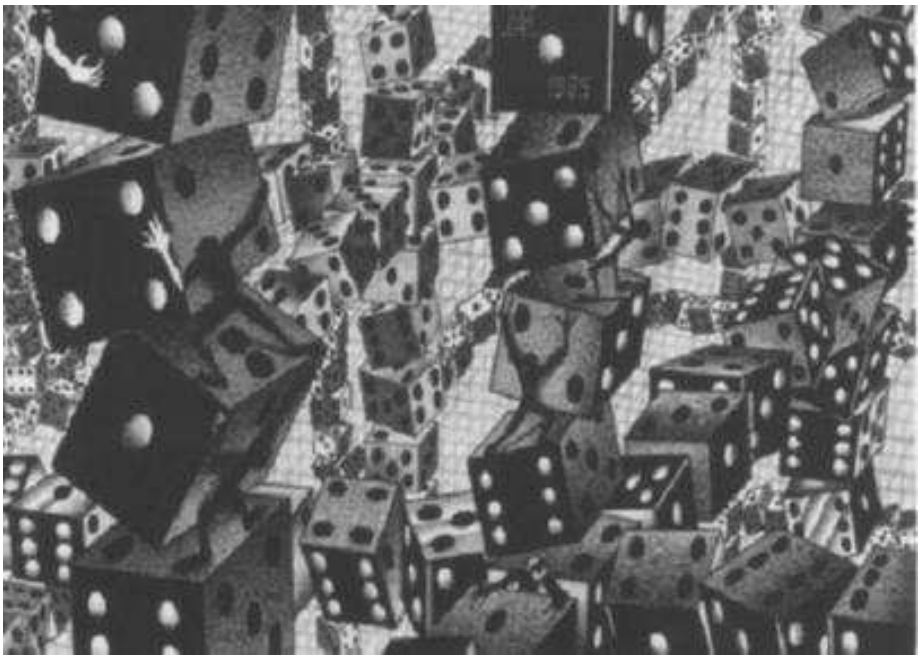
0

Introducción a los Métodos Estocásticos en Ciencias de la Computación



Versión preliminar

Centro de Investigación en Matemáticas
Guanajuato, Gto, México

J. Van Horebeek
horebeek@cimat.mx Agosto 2010

Capítulo 1

Motivación

En computación la probabilidad cumple dos fines: nos ofrece una herramienta para formalizar y manejar cierto tipo de información y al mismo tiempo nos permite salir de un marco puramente determinístico o complétamente predecible. En este capítulo describimos a través de ejemplos muy sencillos un panorama general. No pretendemos describir un marco riguroso sino dar un sabor de la gran diversidad de aplicaciones donde surge la probabilidad en computación.

1.1 Probabilidad e Información

Sin duda, la manera más común en computación para describir información y - en general - conocimiento, es a través de relaciones determinísticas usando expresiones lógicas y relacionales. Reglas específicas permiten combinar información y *deducir* nueva. Por ejemplo, si $a = b + c$ y $b < d$, entonces $a < d + c$ o si $A \Rightarrow B$ y $B \Rightarrow C$, entonces $A \Rightarrow C$.

Existen diversas situaciones importantes tales que por su propia naturaleza o por falta de un modelo analítico completo, tienen un componente de incertidumbre y no pueden ser completamente descritas de esta manera. Al mismo tiempo surge la necesidad de un *cálculo* adecuado que permite *razonar* bajo esta incertidumbre.

Ejemplo 1.1.1 **Descripción de regularidades en textos**
Tomamos el siguiente texto donde ciertas letras son ocultas:

La probab-lida- c-mple e- co-putac-ón do- fines: n-s ofr-ce un- her–mient- p-ra f-rmaliz-r y mane-ar cierto -ipo de -nf-rmació-

A pesar de ya no tener el texto completo, por la redundancia de información en un texto de lenguaje natural, cualquier persona es capaz de reconstruir la frase original. Esta redundancia se manifiesta a varios niveles: desde el nivel de carácteres hasta el semántico. Si nos fijamos al nivel más bajo, el de los carácteres, es evidente que por ejemplo para *e-*, algunas letras son más *probables* que otras de ser la letra oculta aunque

- sin leer lo demás - no podemos decir con certeza cual será. Como mostraremos en el siguiente capítulo, podemos usar el concepto de probabilidades para formalizar esta información no determinística.

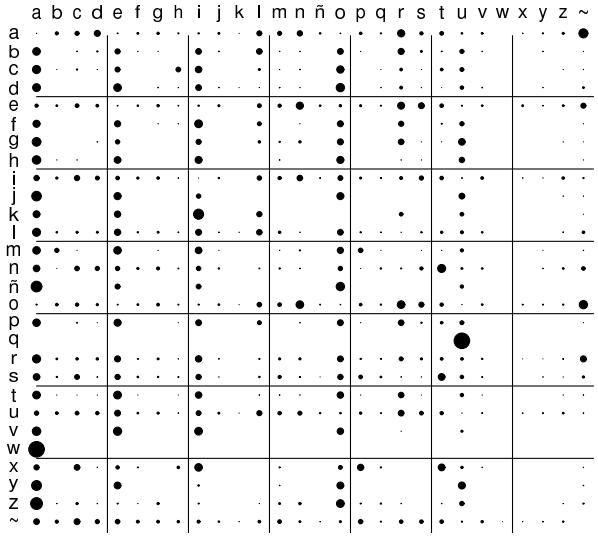


Figura 1

Como ilustración, la Figura 1 muestra gráficamente las probabilidades que una letra particular (identificada con un reglón) sea seguida por otra (identificada con una columna): la superficie del círculo es proporcional a la probabilidad de ocurrencia.

Estas regularidades forman la base de muchos sistemas de compresión de textos.

En ocasiones, no tenemos la información directamente disponible sino que la podemos accesar solamente a través de un experimento. Eso ocurre en situaciones donde la aleatoridad es un *componente intrínseco* al fénomeno de interés; ejemplos clásicos están relacionados con fénomenos físicos: pensamos en el ruido en una imagen al transmitirla sobre un canal y tenemos interés en restaurar las imágenes corruptidas por este ruido (Figura 2).

En particular en informática, la aleatoriedad tiene de vez en cuando su orígen en algún tipo de *ignorancia* o perdida de información. A continuación damos un ejemplo.

Ejemplo 1.1.2 **Métodos empíricos**
Los métodos empíricos se refieren a métodos *inductivos* de razonamiento que toman como punto de partida observaciones de algún tipo de experimento.

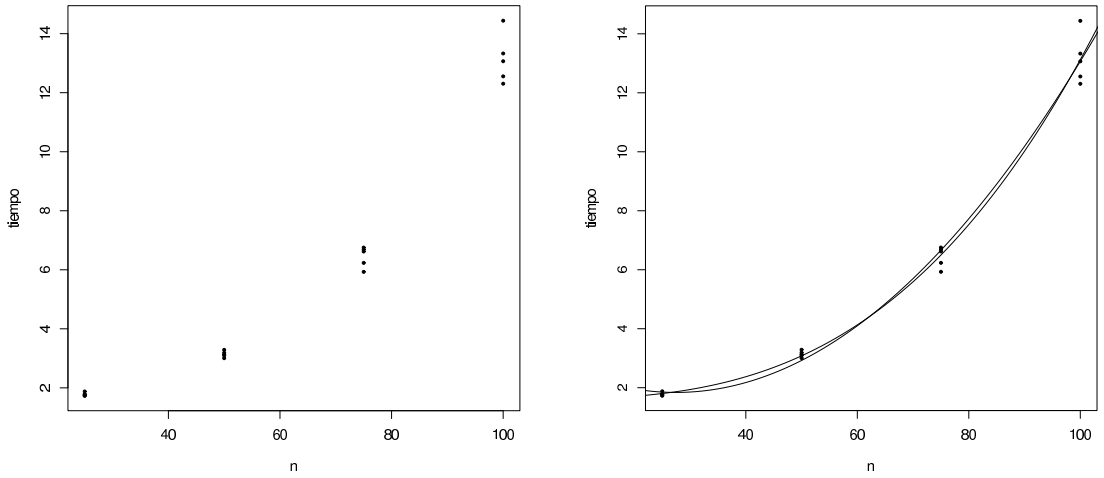
Supongamos que queremos estudiar la complejidad de un algoritmo para ordenar n números. Expresiones explícitas de la complejidad en función del tamaño son pocas veces disponibles (y además muchas veces solamente conocidas hasta un factor). Tampoco es factible correr el algoritmo para cualquier conjunto, aún si fijamos n. Por eso, elegimos *algunos* conjuntos de diferente tamaño, corremos el algoritmo y apuntamos para cada caso por ejemplo el número requerido de operaciones elementales o el tiempo en segundos. La Figura 3 (a) muestra un ejemplo de una elección particular.

Figura 2: Aleatoriedad como componente intrínsico en una fotografía antigua

A partir de estos datos, surgen varias preguntas. ¿Cómo usarlos para decir *algo* acerca del desempeño del algoritmo para un conjunto de tamaño, digamos 45, que nos llegará en un futuro, i.e., como predecir el desempeño para $n = 45$ sin saber cual conjunto de 45 elementos llegará? ¿Cómo expresar la *calidad* de esta predicción?
Veremos más adelante como usar la probabilidad tanto para la formulación de la respuesta como para el cálculo. Otra pregunta de interés es por ejemplo: ¿cómo elegir los conjuntos de prueba (y tamaño) de una manera óptima? ¿ Qué entendemos por óptimo ? etc.

Observa la diferencia en el planteamiento con métodos de interpolación usados en el área de métodos númericos. Aquí las preguntas son (1) cómo encontrar dentro de una familia de curvas, la que minimiza un cierto criterio de ajuste para las observaciones dadas (típicamente, minimizando el error cuadrático) y (2) como calcular la solución numéricamente, i.e. con precisión finita y en tiempo finito. Figura 3(b) muestra el resultado para dos funciones diferentes de interpolación.

El ejemplo anterior se extiende a muchas otras situaciones. Por ejemplo, en ingeniería de software el algoritmo puede referir ahora a una técnica particular para desarrollar un proyecto de software; n puede expresar la complejidad del proyecto y el tiempo de ejecución refiere ahora al tiempo requerido para entregar el software.



(a) *(b)*

Figura 3: (a) Tiempo de ejecución para ordenar conjuntos de tamaño n; (b) Dos funciones de interpolación para los datos: un polinomio de orden 2 y uno de orden 3.

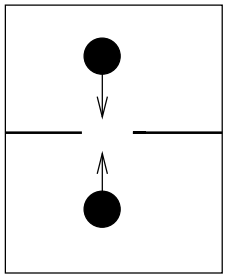
1.2 Probabilidad y Determinismo

La probabilidad nos permite que un algoritmo tome decisiones que no son predecibles de manera determinística.

Eso es de gran ayuda por ejemplo para ocultar información como se hace en criptografía, para generar datos que sirven para estudiar y simular un sistema sin necesidad de construirlo físicamente y en problemas de optimización global para poder escapar de mínimos locales.

Por supuesto un primer problema que surge aquí es como *imitar* tal comportamiento en una máquina estilo Von Neuman donde cada instrucción por definición debe ser sin ambigüedad.

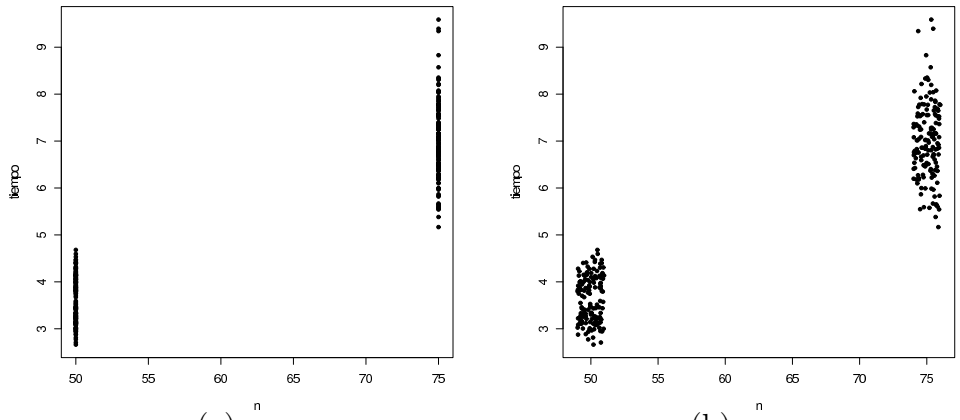
A continuación damos un ejemplo, quizás el más sencillo en su género, donde usamos la aleatoridad para que diferentes procesos se pongan de acuerdo sin recurrir a un órgano central.



Dos personas quieren pasar por una puerta angosta
Figura 4 .

Ejemplo 1.2.1 **Algoritmos aleatorizados**
El ejemplo clásico de un algoritmo aleatorizado es el problema de cómo dos personas se deben poner de acuerdo para pasar por una puerta angosta sin un órgano coordinador central (Figura 4): la puerta es tan angosta que no es posible que pasen dos personas al mismo tiempo. Si ambas personas usan el mismo protocolo determinístico, existe el peligro real de entrar en un *punto muerto (dead lock)* donde cada una se queda esperando a que la otra pase primero. Una manera fácil para evitar esto, es incluyendo un elemento aleatorio; por ejemplo, cada persona espera un tiempo aleatorio antes de hacer un intento de pasar por la puerta. Algunos protocolos de comunicaciones en redes están basado en esta idea sencilla.

Una variante de esta idea la encontramos en métodos de visualización de datos. Supongamos que en la Figura 3 (a), tenemos una gran cantidad de observaciones para un mismo valor de n. En lugar de representarlo como en la Figura 5(a), se puede añadir a cada observación en su primer coordenada un número elegido al azar entre -2 y 2. El resultado, Figura 5(b) da una mejor impresión de la distribución. En este caso, el componente aleatorio fue usado para evitar sugerir alguna estructura fantasma (artefacto); es decir una estructura determinística que no sería propia de los datos sino del algoritmo de visualización.



(a) (b)

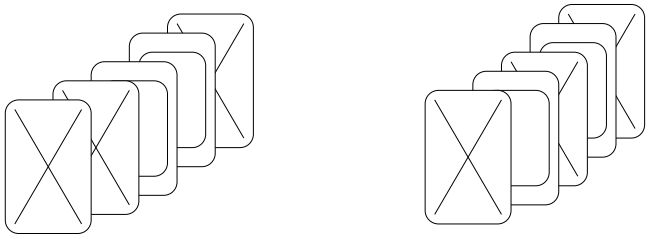
Tiempo de ejecución para ordenar conjuntos de tamaño 50 y 75
Figura 5.

Ejemplo 1.2.2 **Criptografía**
El problema central de la criptografía es como enviar mensajes entre un grupo de personas, conservando su privacidad; en primer lugar hacia personas ajenas. Algunas aplicaciones requieren que además haya privacidad dentro del grupo. Por ejemplo, al enviar una clave de acceso no solamente se debe evitar que terceros lo intercepten sino también hay que evitar que el administrador del sistema pueda abusar de su concocimiento acerca de la clave. A continuación describimos una aplicación muy simplificada con - por razones didacticas - actores humanos.

Supongamos que dos personas deben contestar una pregunta con respuestas binarias (cierto o falso), y que solamente es importante conocer si ambas contestaron *cierto*, es

decir el interés es en el resultado de la aplicación de la función *and* a sus respuestas. Por la naturaleza de la pregunta, se quiere proteger la privacidad de los contestantes de tal manera que si una persona contesta *falso*, no debería poder saber la respuesta de su oponente.

Para ese fin, usamos 5 cartas: 3 cartas tienen a un lado una cruz y 2 un sol. Se pone una carta con una cruz en la mesa y a cada uno se da una carta con una cruz y una con un sol. Si la primera persona quiere contestar *cierto*, pone primero la carta con una cruz y encima una con un sol; en caso contrario, pone primero la carta con el sol y después la carta con la cruz. Después contesta la segunda persona: si su respuesta es *cierto* pone primero la carta con el sol y encima la de la cruz, si no, en orden inverso. Se pone siempre el lado de la carta con la figura hacia abajo. El protocolo está mostrado en la Figura 6.



Configuración cuando ambas personas contestan cierto y cuando el primero contesta falso y el segundo cierto.
Figura 6.

Una tercera persona agarra las primeras n cartas donde n es un número aleatorio entre 0 y 4 y pone estas cartas abajo del mazo. Se repite eso un par de veces. Finalmente se da vuelta a las cartas: si dos soles están juntos o forman la primera y última carta, el resultado de la aplicación del operador AND a las respuestas es *cierto*; en caso que esten separados, la respuesta es *falso*. Dejamos como ejercicio verificar que efectivamente si uno contestó *falso*, no aprenderá nada de la respuesta del otro.

1.3 Probabilidad e Intuición

Probablemente no hay ninguna área en matemáticas donde abundan tantos paradojas como en la probabilidad: veremos en los siguientes capítulos varios ejemplos que mostrarán que la intuición humana no es siempre una buena guia en problemas con un componente de incertidumbre. No es diferente en computación. De ahí surgirá no solamente la necesidad de un formalismo correcto sino también de una manera de *razonar* bajo esta incertidumbre. A continuación un ejempo sencillo.

Figura 7.

Ejemplo 1.3.1 Considera los 4 dados de la Figura 7. La distribución de los números es:
dado A: $4, 4, 4, 4, 0, 0$
dado B: $3, 3, 3, 3, 3, 3$
dado C: $6, 6, 2, 2, 2, 2$
dado D: $5, 5, 5, 1, 1, 1$

Si se lanza A y B, la probabilidad de que el número que aparece en A es mayor que el de B, $P(A > B)$, es $\frac{2}{3}$; como es mayor que un medio, es más probable que $A > B$ que $B > A$ (empates no pueden ocurrir). De la misma manera, $P(B > C) = P(C > D) = \frac{2}{3}$. En otras palabras con gran probabilidad A es mayor que B, B es mayor que C, C es mayor que D. Si hubiera algo como transividad, se esperaría que sera más probable $A > D$ que $D > A$; sin embargo es facil mostrar que $P(A > D) = \frac{1}{3} < \frac{1}{2}$!

Esta falta de transitividad genera situaciones paradójicas. Por ejemplo pides a un adversario elegir un dado; tu puedes siempre elegir otro dado tal que al lanzarlos - en promedio - obtendrás un número mayor que tu adversario, o sea, independiente de la elección de tu adversario, siempre puedes ganar!

1.4 Nota histórica

En la Figura 8 vemos parte de lo que se considera como el primer libro (ensayo) sobre probabilidad bajo el nombre *Over Reeckeningh in Spelen van Geluck / De Ratiociniis in Ludo Aleae* por el holandés Christiaan Huygens publicado en el año 1657. Entre los problemas que se discuten se encuentran varios que Fermat y Pascal habian discutido en el verano de 1654.

DE
RATIOCINIIS
IN
LUDO ALEÆ.

ETſi luſionum, quas ſola ſors moderatur, incerti ſolent eſſe eventus, attamen in his, quanto quis ad vincendum quàm perdendum propior ſit, certam ſemper habet determinationem. Ut ſi quis primo jactu unâ teſſerâ ſenarium jacere contendat, incertum quidem an vincet;

ALTHOUGH in Games depending entirely upon Fortune, the Success is always uncertain; yet it may be exactly determin'd at the same time, how much more likely one is to win than lose. As, if any one shou'd lay that he wou'd throw the Number *Six* with a single Die the first throw, it is indeed uncertain whether he will win or lose; but how much more probability there is that he shou'd lose than win, is presently determin'd, and easily calculated. So likewise, if I agree with another to play the first Three Games for a certain Stake, and I have won one of my Three, it is yet uncertain which of us shall first get his third Game; but the Value of my Expectation and his likewise may be exactly discover'd; and consequently it may be determin'd, if we shou'd both agree to give over play, and leave the remaining Games unfinish'd, how much more of the Stake comes to my Share than his; [2] or, if another desired to purchase my Place and Chance, how much I might just sell it for. And from hence an infinite Number of Questions may arise between two, three, four, or more Gamesters: The satisfying of which being a thing neither vulgar nor useless, I shall here demonstrate in few Words, the Method of doing it; and then likewise explain particularly the Chances that belong more properly to Dice.

Figura 8.

Un poco de etimologia ... según el diccionario de Oxford.
Como la probabilidad es tan inherente a la vida cotidiana, no es sorprendiente que el orígen de las palabras que empleamos para denotarla, refleje diferentes percepciones a través del tiempo.

aleatorio: proveniente del latin *alea* (dado): determinado por los dados.
azar: de origen árabe; una posibilidad es que se refiere a la palabra árabe az-zahr (dado) o hace referencia a un juego que fue inventado durante la ocupación del castillo Asart o Ain Zarba en Palestina.
estocástico: la raíz es griega: apuntar hacia algo, adivinar;
chance: del latín cadere, *caer*: como caen las cosas;
probabilidad: del latín *probabilis*: lo que puede ser demostrado; el sentido antiguo de probabilidad es *la aparencia de la verdad*. Como decia el obispo Butler (1736) "Probability is the guide of life". Eso por supuesto ya lo sabíamos ....

Capítulo 2

Probabilidades

En el presente capítulo definimos el concepto de probabilidad desde diferentes puntos de vista. Derivamos las propiedades importantes y estudiamos algunos ejemplos sencillos que ilustran las aplicaciones.

2.1 El concepto de probabilidad

El punto clásico de partida para introducir el concepto de probabilidad es definiendo lo que llamamos *experimento*. Un experimento puede ser algo explícito como "lanzar un dado'' o "tomar una carta de un mazo'', o implícito, donde uno de los actores es la propia naturaleza: "medir la temperatura de mañana a mediodía en un lugar particular''. El resultado de un experimento debe ser un valor específico. Denotamos con Ω el conjunto de todos los posibles resultados del experimento que se puede obtener.

Ejemplo 2.1.1 Si el experimento es lanzar un dado, los posibles resultados son $1, 2, 3, 4$, 5 o 6. Es decir $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Un siguiente paso podría ser asignar a cada resultado del experimento un número que llamaríamos la probabilidad de este resultado. Sin embargo no es siempre suficiente (o posible). Por ejemplo, uno puede tener interés en la probabilidad de obtener un número mayor que tres; es decir en obtener un resultado que pertenece al subconjunto $\{4, 5, 6\}$.
Lo anterior motiva la definición de probabilidades sobre (ciertos) subconjuntos de valores de Ω, llamados *eventos*. De esta manera con una cierta clase $\mathcal{B}$ de subconjuntos de valores de Ω, se asocia una (función de) *probabilidad* $P(\cdot)$:

$$P : \mathcal{B} \to [0, 1] : A \to P(A), \tag{2.1}$$

donde se requiere que $P(\cdot)$ satisfaga ciertas propiedades que veremos más adelante.
En el ejemplo anterior, la probabilidad de obtener un número mayor que tres corresponde a $P(A)$ con $A = \{4, 5, 6\}$.

Ejemplo 2.1.2 Un generador de *passwords* regresa cadenas de 12 caracteres que al menos incluyen un número. Entonces Ω son todas las cadenas de tamaño 12 con al menos un número en una posición. Para calcular la probabilidad de que se genere un password que contiene la subcadena "inolvidable'', se está considerando $A = \{0inolvidable, 1inolvidable, \cdots, 9inolvidable, inolvidable0, inolvidable1, \cdots, inolvidable9\}$
La probabilidad de que se genere "123456789012'' corresponde a $P(A)$ donde $A = \{123456789012\}$.

Nótese que $P(A)$, la probabilidad de que el evento A ocurra, denota en realidad la probabilidad de obtener un elemento $\omega \in \Omega$ que pertenezca a A. Usando esta convención, tenemos por ejemplo:

1. $P(A \cap B)$ representa la probabilidad de obtener un elemento que pertenezca a ambos conjuntos, A y B,

2. $P(A \cup B)$, representa la probabilidad que el resultado pertenezca al menos a un conjunto, A o B y

3. $P(A^c)$, $= P(\Omega \setminus A) = \{\omega : \omega \in \Omega \text{ y } \omega \notin A\}$, corresponde a la probabilidad de obtener un elemento que no pertenezca al conjunto A.

A partir de (2.1) surgen dos problemas para la construcción de $P(\cdot)$. El primero corresponde a un problema técnico: dado que tenemos una función que toma como argumento un conjunto y no un elemento de un conjunto, no es trivial definir $P(\cdot)$. En lo subsecuente daremos -bastante superficialmente- diferentes métodos para construir $P(\cdot)$. Como en todo este curso, el énfasis será en el caso discreto (finito) donde Ω es contable o finito.

El segundo problema que presenta la definición presentada de probabilidad es que no es evidente cómo interpretar $P(\cdot)$. Si se puede repetir el experimento, un camino clásico es ver una probabilidad de un evento como *frecuencia de ocurrencia* en muchas repeticiones del mismo experimento. Sin embargo, en muchas aplicaciones y en particular en computación, no se puede repetir el experimento.

2.1.1 Construcción de la probabilidad cuando Ω es finito

Supongamos primero que podemos escribir Ω como el conjunto $\{\omega_1, \cdots, \omega_k\}$; entonces cada subconjunto A de Ω puede ser enumerado por sus elementos. Distinguimos los siguientes enfoques.

Por medio de la distribución de conteo o distribución uniforme

En general, cada uno tiene una noción clara de lo que significa "elegir un elemento al azar'', i.e. sin alguna preferencia. El ejemplo clásico es elegir, sin ver, un objeto de

una bolsa que contiene varios objetos más, todos indistinguibles al tacto. A partir de eso, se define para cada $\omega \in \Omega$:

$$P(\{\omega\}) = \frac{1}{\#\,\Omega}$$

y en general para cada $A \subset \Omega$:

$$P(A) = \frac{\#\,A}{\#\,\Omega}. \tag{2.2}$$

A la función $P(\cdot)$ la llamaremos *distribución de conteo* o *distribución uniforme* sobre Ω. Como definimos (2.2) para cada A, $\mathcal{B}$ es igual a $\mathcal{D}(\Omega)$, el conjunto de todos los subconjuntos de Ω. El conjunto $\mathcal{D}(\Omega)$ se conoce como el *conjunto potencia* de Ω y también suele denotarse como 2^{Ω}.

Ejemplo 2.1.3 Alguien elige al azar una letra en un teclado. De esta manera, Ω coincide con el alfabeto. Para calcular la probabilidad de que la letra elegida sea una vocal, calculamos:

$$P(\{vocales\}) = \frac{\#\,vocales}{\#\,teclas} = \frac{5}{26}.$$

Ejemplo 2.1.4 Lanzamos dos dados y queremos calcular la probabilidad de obtener el mismo valor en cada dado. Definimos $\Omega = \{(i,j), 1 \leq i,j \leq 6\}$ y A, el evento de interés, como $\{(i,i), 1 \leq i \leq 6\}$. Construimos la distribución de conteo sobre $\mathcal{B} = \mathcal{D}(\Omega)$ porque cualquier elemento de Ω es igualmente probable. Entonces, tenemos que $P(A) = \#A/\#\Omega = 6/36 = 1/6$.

Podemos extender la construcción de la distribución de conteo a una clase de $P(\cdot)$'s racionales. Si tomamos $\Omega = \{\text{hombre}, \text{mujer}\}$, se puede interpretar $P(\{\text{hombre}\}) = 0.4$ y $P(\{\text{mujer}\}) = 0.6$ como la distribución de conteo (2.2) sobre $\Omega^* = \{\text{hombre}_1, \text{hombre}_2, \text{hombre}_3, \text{hombre}_4, \text{mujer}_1, \text{mujer}_2, \text{mujer}_3, \text{mujer}_4, \text{mujer}_5, \text{mujer}_6\}$, es decir, elegimos al azar una persona de Ω^*.

El hecho de que cada conjunto A tenga un número finito de elementos, abre la posibilidad de definir la probabilidad de un conjunto como la suma de las probabilidades de los elementos que contiene.

Definición 2.1.1 *Dado $\Omega = \{\omega_1, \cdots, \omega_k\}$, si la secuencia $\{p_i\}$ satisface:*

1. $p_i \geq 0, \forall i$;
2. $\sum_{i=1}^{k} p_i = 1$;

entonces llamamos $P(\cdot)$ definido por $P(A) = \sum_{\omega_i \in A} p_i$ para cada $A \subset \Omega$ una función de probabilidad sobre Ω.

Se puede mostrar fácilmente las siguientes propiedades. Dejamos como ejercicio al lector describir con sus propias palabras el significado de cada una.

Propiedad 2.1.1 *Para cada A y B eventos,*

1. $P(A) \in [0,1]$;
2. *Si $\{A_i\}$ es una sucesión de conjuntos ajenos, i.e. $A_i \cap A_j = \emptyset$, $\forall i \neq j$, se tiene:*

$$P(\cup A_i) = \sum_i P(A_i); \tag{2.3}$$

3. $P(A) = 1 - P(A^c)$;

A partir de las propiedades anteriores se pueden derivar algunas otras. Por ejemplo, suponiendo que $A, B \in \mathcal{B}$:

$$P(A \setminus B) = P(A) - P(A \cap B). \tag{2.4}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{2.5}$$

$$\text{Si } A \subset B \text{ entonces } P(A) \leq P(B). \tag{2.6}$$

Para derivar (2.4), tomamos como punto de partida que para cualquier conjunto A y B:

$$A = (A \setminus B) \cup (A \cap B).$$

Dado que $A \setminus B$ y $A \cap B$ son ajenos, por (2.3),

$$P(A) = P(A \setminus B) + P(A \cap B),$$

o equivalente:

$$P(A \setminus B) = P(A) - P(A \cap B). \tag{2.7}$$

De una manera análoga, se derive (2.5). Dejamos la derivación de (2.6) como ejercicio.

Ejemplo 2.1.5 Se elige al azar una carta de un mazo. Queremos calcular la probabilidad de obtener un tres o un trébol. Con ese fin, definimos A como el conjunto de tres de cada palo y B como todas las cartas trébol.
Así para calcular $P(A \cup B)$, usamos (2.5):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52}.$$

Por medio de repeticiones

Se podría definir una probabilidad por medio de una repetición infinita del mismo experimento cuyos resultados no interactúan entre sí y contar cuántas veces ocurre cierto evento:

$$P(A) = \lim_{n\to\infty} \frac{\text{número de veces que ocurre } A \text{ en } n \text{ repeticiones}}{n}. \tag{2.8}$$

Es la manera conceptual más fácil pero técnicamente mas difícil. Afortunadamente como se verá más adelante, se puede mostrar que para cualquier función de probabilidad que satisface un mínimo de propiedades (axiomas), la ecuación (2.8) se cumple.

Por medio de apuestas

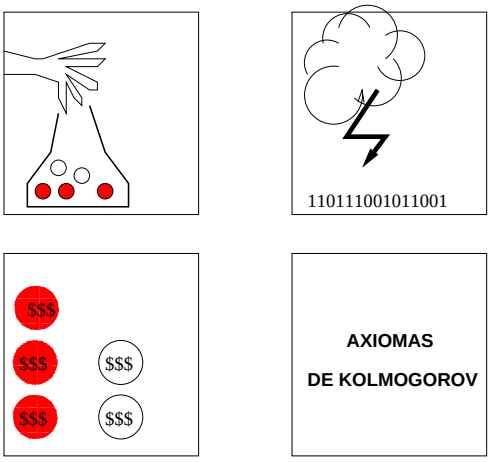
Supongamos que un asegurador necesita definir la probabilidad de que el telescopio Hubble se estrelle el siguiente año contra un meteorito. Pretender que la probabilidad de ocurrencia de este evento es igual a 0.001, puede interpretarse como que el asegurador está dispuesto a aceptar una apuesta de 1 contra 999 sobre la ocurrencia del siniestro. Es decir, el asegurador, como actor racional, pretende pagar \$999 si el telescopio es golpeado por un meteorito y recibir \$1 en caso de que esto no ocurra. Otra formulación es decir que el asegurador va a cobrar una prima de \$1 y pagar \$1000 en caso de un siniestro (suponiendo que no se usa un margen de ganancias).

En este enfoque la apuesta *a versus b*, de que un evento A ocurra, corresponde por definición a una probabilidad de ocurrencia de A igual a $a/(a+b)$. Obsérvese la diferencia con los métodos anteriores. Ahora la construcción de $P(\cdot)$ es subjetiva, dependiendo de la información que cada uno tiene disponible.

En los enfoques anteriores se trataría más bien de definir un conjunto de "elementos (o situaciones) similares'', por ejemplo contar todos los satélites que había el año pasado y calcular el porcentaje de los que se estrellaron. La dificultad es que el no tener control sobre el experimento, impide repetirlo bajo las mismas condiciones y muchas veces es difícil encontrar situaciones indistinguibles con igual probabilidad. Por ejemplo, no todos los satélites son igualmente susceptibles a accidentes o quizás el siguiente año el riesgo es de otra naturaleza al del año pasado.

Es importante entender que en este enfoque el interés es muchas veces en *razones* de probabilidades y no tanto en las probabilidades como cantidades absolutas. Como elaboraremos en el Capítulo 5, en computación se hace muchas veces aún un paso más en esta dirección, donde ni siquiera es de interés el orden relativo de las probabilidades de los eventos, sino que la localización de los eventos de máxima probabilidad.

En la Figura 1, resumimos los tres enfoques anteriores y uno que veremos más adelante.



Cuatro maneras para definir que $P(A) = 2/5$
Figura 1.

Una aplicación: probabilidades, coincidencias y superstición

En la vida cotidiana, frecuentemente nos enfrentamos a coincidencias que nos asombran porque nos parece difícil poder atribuirlas al azar. Por ejemplo al comprar un boleto de la lotería alguien se da cuenta que éste termina con los mismos últimos dígitos de su número de pasaporte, o se recibe del banco una clave para un cajero automático que contiene el día y mes de su cumpleaños, etc.

Aunque hay ciertos factores psicológicos que juegan un papel importante en este fenómeno, éste puede explicarse en parte por las altas probabilidades de sucesos de ciertas coincidencias. Se describe a continuación un modelo genérico de coincidencia, conocido como el problema de las fechas de cumpleaños y que sirve al mismo tiempo como ejemplo de una aplicación de la distribución de conteo.

Ejemplo 2.1.6 Tenemos un grupo de n personas. ¿Cuál es la probabilidad de que al menos dos tengan la misma fecha de cumpleaños?

En este caso $\Omega = \{1, \cdots, 365\}^n$(i.e.,$\Omega$ consiste de todos los n-eadas con cada componente entre 1 y 365). Sobre este conjunto definimos la distribución de conteo. Dado que

$$P(\text{al menos dos tienen la misma fecha de cumpleaños}) =$$

$$1 - P(\text{ninguno tienen la misma fecha de cumpleaños}) = 1 - \frac{\#A}{\#\Omega},$$

donde A es el conjunto de las n-adas con entradas diferentes: $A = \{\mathbf{i} = (i_1, \cdots, i_n) : i_k \in \{1, \cdots, 365\},$ y todos diferentes$\}$.

Para cualquier $\mathbf{i}$ de A, el primer componente i_1 puede ser cualquier día del año (365 posibilidades), para i_2 quedan $365 - 1$ dias, etc., así obtenemos:

$$\#A = 365 \cdot 364 \cdot 363 \cdot \cdots \cdot (365 - n + 1).$$

Combinando todo:

$$P(\text{al menos dos tienen la misma fecha de cumpleaños}) =$$

$$1 - \frac{365 \cdot 364 \cdot 363 \cdot \cdots \cdot (365 - n + 1)}{365^n}. \tag{2.9}$$

En la siguiente tabla, se muestra la probablidad para diferentes valores de n.

tamaño del grupo	15	25	35	45	55
P(al menos dos cumpleaños coinciden)	0.25	0.56	0.8	0.94	0.98

Tabla 1.

Es fácil trasponer el ejemplo anterior a otras situaciones. Por ejemplo, para calcular la probabilidad de que en un grupo de n usuarios al menos dos elijan la misma clave o el mismo nombre de archivo en un sistema de cómputo.

Aunque la suposición de la distribución de conteo no es en general válida (ni para el ejemplo de los cumpleaños), se puede mostrar que las probabilidades de coincidencia son cotas inferiores (*worst case*) para cualquier otro tipo de distribución.

Hay que contrastar el problema anterior con el siguiente (el cual se confunde muchas veces con el problema anterior).

Ejemplo 2.1.7 Formas parte de un grupo de n personas. ¿Cuál es la probabilidad de que al menos una persona más tenga la misma fecha de cumpleaños que tú?

Una derivación similar a (2.9) conduce a:

$$P(\text{al menos un cumpleaños coincide con el tuyo}) = 1 - \frac{364^{n-1}}{365^{n-1}}.$$

Igualmente se muestra en la siguiente tabla la probabilidad para algunos valores de n.

tamaño del grupo	15	25	35	45	55
P(al menos un compleaños coincide con el tuyo)	0.03	0.06	0.08	0.11	0.13

Tabla 2.

Se observa que las probabilidades son ahora pequeñas y mucho más cercanas a lo que uno en general intuitívamente piensa. De esta manera se explica en parte la paradoja de las altas en la Tabla 1: el no suponer quiénes tienen que coincidir en las fechas de cumpleaños, aumenta sustancialmente la probabilidad de una coincidencia.

En la práctica consideramos inconscientemente varias características al mismo tiempo; por ejemplo, los últimos dígitos de la placa de su automóvil, número de casa, etc. Las probabilidades de coincidencias serán aún mayores. Por ejemplo, si consideramos

3 características, cada una con 100 posibles valores, obtenemos que en un grupo de 7 personas con probabilidad (aproximadamente) un medio, al menos dos personas tienen, al menos, dos características iguales.

Los dos ejemplos ilustran cómo calcular la probabilidad de una ocurrencia observada y decidir después si lo observado es algo excepcional (que no se puede atribuir al azar) o si es algo *común*, para lo cual no hay base de atribuirlo forzosamente a factores sobrenaturales (como por ejemplo superstición). La misma construcción encontraremos en pruebas de hipótesis que estudiaremos más adelante.

2.1.2 La probabilidad uniforme en $\mathcal{R}^n$

Supongamos que elegimos al azar un número real entre 0 y 1. Como hay un número infinito de reales entre 0 y 1, ya no podemos seguir el camino de la sección anterior para construir $P()$ que fue basado en la propiedad de la aditividad (2.3) y que permitió definir $P(\cdot)$ completamente en función de conjuntos unitarios.
En esta sección discutimos un camino alterno.

Tomamos un conjunto acotado $\Omega \subset \mathcal{R}^n$, y definimos:

$$P(A) = c \int_A 1dx, \quad \text{con } A \subset \Omega, \tag{2.10}$$

donde $c = 1/\int_\Omega 1dx$ es una constante de normalización para que $P(\Omega) = 1$. A la probabilidad expresada en (2.10) se le llama *distribución uniforme* sobre Ω y se denota como $\mathcal{U}(\Omega)$. Corresponde a la situación donde elegimos al azar un elemento del conjunto Ω.

Desde un punto de vista práctico si $n = 1$ y A representa un intervalo, $\int_A 1dx$ es la longitud de este intervalo; para el caso $n = 2$, *mutatis mutandis*, obtenemos una área y para $n = 3$ un volumen. De esta manera el cálculo de probabilidades para $n = 1, 2, 3$ se reduce a medir longitudes, áreas o volumenes de ciertas regiones en $\mathcal{R}^n$ definidas por restricciones geométricas. Por eso se habla, en este contexto, de *probabilidades geométricas*. Dejamos como ejercicio mostrar que tambien se cumplen las propiedades de la Definición 2.1.1. Observe que $P(\cdot)$ para cualquier conjunto unitario o finito será cero.

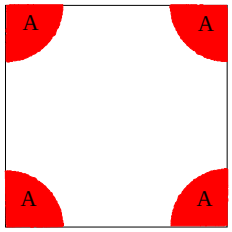


Figura 2.

Ejemplo 2.1.8 Se elige al azar un punto en el cuadrado definido en la Figura 2. Cada lado mide 4 cm. Para calcular la probabilidad de que un punto a seleccionar esté a

una distancia menor de uno cm. de alguna de las esquinas, se define la región A, como marcada en la figura, como todos los puntos a una distancia menor que un cm., de alguna de las esquinas. Usando (2.10), la probabilidad correspondiente es igual a

$$\frac{4(1^2\pi)/4}{4^2} = 0.196.$$

Ejemplo 2.1.9 Dos estudiantes quieren ir a comer juntos. Se citan entre las 7 y las 8 de la noche y están dispuestos a esperar a lo más 10 minutos. ¿Cúal es la probabilidad de que puedan ir a comer si sus horas de llegada son uniformes entre las 7 y las 8? Definimos $\Omega = [0,1]^2$, y $t = (t_1, t_2) \in \Omega$ donde t_i representa el momento de llegada para estudiante i (remapeando una hora al intervalo $[0,1]$). Definimos $P(\cdot)$ como en (2.10).

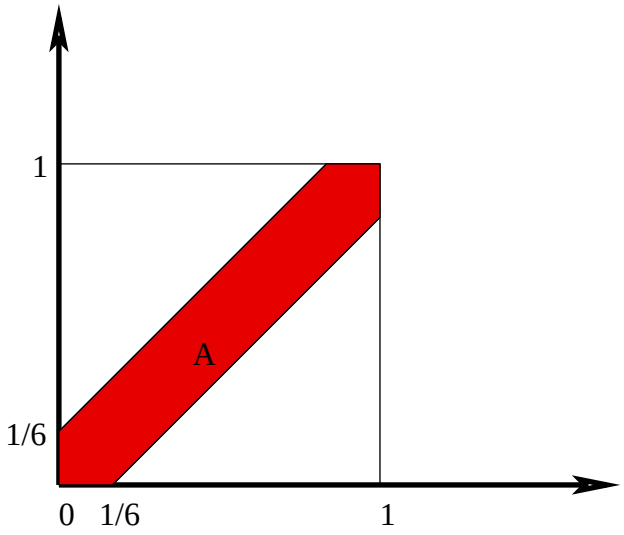


Figura 3.

Los elementos de Ω que corresponden a situaciones donde el tiempo que pasa entre sus momentos de llegada es menor que 10 minutos, definen la región, acotada por las líneas $t_1 - 1/6$, $t_1 + 1/6$ y la intersección con $[0,1]^2$ como se indica en la Figura 3. El área es 1-25/36=11/36. Así la probabilidad de que se encuentran los dos estudiantes es igual a

$$\frac{\text{área A}}{\text{área cuadro}} = \frac{11/36}{1} = \frac{11}{36}$$

Aunque la distribución anterior (2.10) corresponde bastante bien a la idea intuitiva que uno tiene de elegir un elemento al azar, la ambigüidad lingüística puede causar problemas en la formulación de los problemas; en particular cuando no se tiene una parametrización natural de Ω. Referimos a los ejercicios de este capítulo para ver algunos ejemplos.

2.1.3 El caso general: los axiomas de Kolmogorov y la teoría de la medida

Es importante subrayar que para el caso $\Omega \subset \mathcal{R}^n$ la construcción de $P(\cdot)$ como mencionada en la sección anterior, técnicamente ni es completa, ni es correcta. Por ejemplo, es facil ver que no para cualquier A, (2.10) está definida (reescríbela como $P(A) = c \int I_A(x)dx$, con $I_A(\cdot)$ la *función indicadora sobre el conjunto* A). Entonces surge la necesidad de *restringir* el conjunto de eventos pero procurando que los eventos de interés aún pertenezcan a $\mathcal{B}$.

Contrario al caso discreto donde podemos *descomponer* el problema de la definición de $P(\cdot)$ en la asignición de una probabilidad a cada conjunto unitario, en el caso continuo estamos obligados a definir una función sobre conjuntos. Eso no se puede hacer para cada conjunto *de manera independiente* porque para que la defínción de $P(\cdot)$ sea *util* (i.e. cunple con las propiedades que intuitivamente requerimos), debe satisfacer muchas restricciones (por ejemplo $P(A \cup B) = P(A) + P(B)$ en caso que la intersección de A y B es vacia). Lo anterior es una de las preguntas fundamentales de la teoría de la medida.

Desarrollada en los años 30's, una de las contribuciones de Andrey Kolmogorov es el haber deducido un sistema de axiomas muy general y que implican, bajo ciertas condiciones, la Propiedad 2.1.1 y la relación (2.8) y - igual importante - formularlos de manera matemáticamente correcto. De esta manera el estudio de la probabilidad se convirtió en el estudio de la terna $(\Omega, \mathcal{B}, P(\cdot))$.

Definición 2.1.2 *La función $P : \mathcal{B} \to \mathcal{R}$ es una probabilidad ssi*

1. $P(\cdot) \geq 0$;
2. $P(\Omega) = 1$;
3. *Si* $A_1, A_2, \cdots, \in \mathcal{B}$, *tal que* $A_i \cap A_j = \emptyset, \quad i \neq j : P(\cup_i A_i) = \sum_i P(A_i)$,

donde $\mathcal{B}$ es una colección de subconjuntos de Ω, llamado *sigma-algebra* tal que

1. $\phi \in \mathcal{B}$;
2. Si $A_1, A_2, \cdots \in \mathcal{B}$, entonces $\cup_i A_i \in \mathcal{B}$;
3. Si $A \in \mathcal{B}$, entonces $A^C \in \mathcal{B}$.

Sin duda el sigma-algebra más conocido es el *Borel sigma-algebra*: es el sigma-algebra más chiquito que contiene -entre otro- todos los intervalos abiertos y cerrados de $\mathcal{R}$, uniones contables de estos intervalos, complementos, etc. Para los ejemplos que veremos en este curso, este conjunto es suficientemente rico y además permite definir una $P(\cdot)$ que coincide con la idea intuitiva de (2.10).

2.1.4 Otros enfoques para manejar incertidumbre

Las probabilidades, como se han definido anteriormente, no forman la única manera para manejar la incertidumbre. Por ejemplo, el área de la inteligencia artificial se distingue de otras áreas por haber trabajado con varias técnicas alternas. A continuación presentamos dos de ellas.

Conjuntos difusos

La incertidumbre proveniente del concepto de probabilidad, solamente es causada por el experimento y no por la definición de los elementos de Ω. Eso es contrario a la teoría de conjuntos difusos (*fuzzy sets*), donde la incertidumbre surge por indefinición lingüística. Por ejemplo, si elegimos una persona al azar, podemos definir la probabilidad de que sea una persona alta. En probabilidad se supone que se sabe muy bien cuándo una persona es alta: la incertidumbre es causada por no saber cuál persona va a ser elegida. Al contrario, en conjuntos difusos la incertidumbre tiene su origen en la falta de un criterio excluyente que señale cuándo una persona será considerada como alta.

Por ejemplo, supongamos que clasificamos la suciedad del agua en categorías (limpia, sucia y mugrosa) basado en su transparencia (comunmente medido en NTU: Nephelometric Turbidity Units). La siguiente figura indica cómo se podría mapear la suciedad del agua a estas categorías. Como se ve, las categorías no son exclusivas. Un NTU de 8 indica que el agua es 0.1 limpia y 0.8 sucia. Con conjuntos clásicos el agua sería considerada como limpia o sucia pero no un poco de las dos categorías.

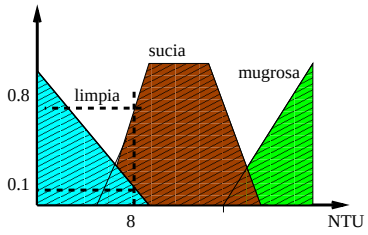


Figura 4.

Lo anterior forma la base de control borroso (fuzzy control). Por ejemplo para una máquina de lavar, algunas reglas pueden ser:

```
Si el agua está mugrosa, purifícala durante 30 minutos
Si el agua está sucia, purifícala durante 10 minutos
Si el agua está limpia, purifícala durante 2 minutos.
```

Entonces el procedimiento para tratar el agua cuando se mide 8 NTU, consistirá en purificar durante $(0.8 * 10 + 0.1 * 2)$ minutos.

Por estar muy cerca a la formulación linguistica que usamos en la vida cotidiana y por la sencillez de las especificaciones, el control borroso ha tenido gran éxito en aplicaciones industriales. Una crítica fuerte es que se usa mucha arbitrariedad en la definición de

las reglas; además no existe una manera empírica para convalidar los supuestos como se usan métodos estadísticos para convalidar modelos probabilísticos.

Dempster-Shafer

Una característica de la teoría de la probabilidad es que para cualquier evento D:

$$P(D^c) = 1 - P(D). \tag{2.11}$$

En un intento de separar los conceptos de incertidumbre e ignorancia, los seguidores de Dempster-Shafer, rechazan la relación (2.11) y a su vez introducen para cada evento D los conceptos de creencia, $be(D)$, y plausibilidad, $pl(D)$. Así ellos trabajan con el intervalo $[be(D), pl(D)]$ en lugar de un valor específico $P(D)$. Se define $pl(D) = 1 - be(D^c)$ donde en general $be(D) \neq pl(D)$.

Para ilustrar lo anterior, supongamos el conocimiento de una relación formulada a través de una regla lógica, $A \rightarrow B$. Si una persona tiene la convicción de que A ocurre con una probabilidad 0.80 (incertidumbre), se obtiene que su creencia en B, $be(B) = 0.8$. Dado que no se sabe nada sobre si $A^c \rightarrow B$ o $A^c \rightarrow B^c$ (ignorancia) , $be(B^c) = 0$, y por consecuencia $pl(B) = 1$.

Ahora ya no se exige que (2.11) sea cierta. Se paga este grado adicional de flexibilidad, con el no poder contar con una riqueza de propiedades y técnicas manejables como en el caso de la probabilidad.

2.2 Conceptos derivados

2.2.1 Probabilidad condicional

Empecemos con el siguiente ejemplo. Supongamos que queremos tomar al azar una persona de la población dibujada en la Figura 5. Usando la distribución de conteo, obtenemos que la probabilidad de obtener un hombre es $\frac{\text{número de hombres}}{\text{número de personas}} = 6/13$, es decir si tomamos con los ojos cerrados una persona, la única información que podemos decir antes de abrirlos es que la probabilidad de haber elegido un hombre es igual a 6/13.

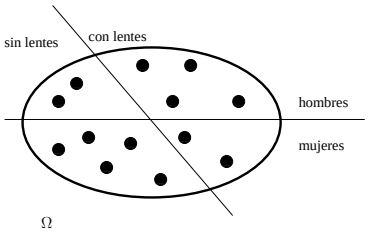


Figura 5.

Supongamos ahora que antes de abrir los ojos, alguien nos dice que la persona en frente de nosotros tiene lentes. Ahora la probabilidad de tener un hombre, condicionada al hecho de que tiene lentes es igual a:

$$\frac{\text{número de hombres con lentes}}{\text{número de personas con lentes}} = 4/6 = 2/3.$$

Esta probabilidad la denotamos como $P(A|B)$, donde A refiere al evento de elegir un hombre y B al evento de elegir una persona con lentes. Como puede observarse, la probabilidad del evento A se modifica o se actualiza al tener conocimiento de la ocurrencia de B ya que esta información restringe el espacio Ω a un nuevo espacio $\Omega^* = B$. Es fácil verificar que $P(A|B) = P(A \cap B)/P(B) = 4/13 \cdot 13/6 = 2/3$.

Lo anterior motiva la siguiente definición:

Definición 2.2.1 *Si* $P(B) > 0$,

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

Se puede considerar $P(\cdot|B)$ como una nueva función de probabilidad sobre $\Omega = B$. Por consecuencia $P(A^c|B) = 1 - P(A|B)$. Observa que no hay ninguna relación directa entre $P(A|B)$ y $P(A|B^c)$.

Ejemplo 2.2.1 Hacemos un experimento que consiste en elegir al azar dos letras consecutivas de alguna palabra. Así Ω es igual a $\{(\text{letra}_1, \text{letra}_2),\ \text{con letra}_1 \text{ y } \text{letra}_2 \in T\}$ y T denota el alfabeto.
A continuación, nos restringimos a la siguiente tabla para un lenguaje hipotético sobre $T = \{a, b, c, d, e\}$.

	a	b	c	d	e
a	0.1	0.05	0.1	0.04	0
b	0.01	0.01	0.1	0.01	0.04
c	0.02	0.05	0.05	0.1	0.01
d	0.04	0.1	0.01	0.01	0.02
e	0	0.1	0	0.01	0.02

Tabla 3.

Para saber la probabilidad de que la segunda letra seleccionada sea la "b'' dado que sabemos que la anterior fue una "e'', calculamos $P(A|B)$ donde $A = \{(\omega, b) : \omega \in T\}$ y $B = \{(e, \omega) : \omega \in T\}$:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{(e, b)\})}{\sum_{\omega \in T} P(\{(e, \omega)\})}.$$

Así

$$P(A|B) = \frac{0.1}{0.13} = 0.77.$$

Para probabilidad que la segunda letra seleccionada sea la "b'' dado que sabemos que la anterior no fue una "e'', calculamos $P(A|B^c)$:

$$P(A|B^c) = \frac{P(A \cap B^c)}{P(B^c)} = \frac{P(\{(e,b)\})}{\sum_{\omega \notin T} P(\{(e,\omega)\})}.$$

Así

$$P(A|B^c) = \frac{0.1}{1 - 0.13} = 0.11.$$

Compara lo anterior con la probabilidad que la segunda letra seleccionada no sea la "b'' dado que sabemos que la anterior fue una "e'':

$$P(A^c|B) = 1 - P(A|B) = 1 - 0.77 = 0.23.$$

A continuación damos alguna propiedades que muestran la importancia de probabilidades condicionales. Con tal fin, definiremos antes una partición del espacio.

Definición 2.2.2 *Se dice que los eventos $B_1, \ldots, B_n$ forman una partición de Ω si son ajenos y su unión es igual a Ω.*

Propiedad 2.2.1 *Dada una partición $\{B_i\}_{i=1}^n$ de Ω, tal que $P(B_i) > 0 \ \ \forall i$,*

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i). \tag{2.12}$$

En particular, si $\{C_i\}_{i=1}^m$ es una partición de D,

$$P(A|D) = \sum_{i=1}^{m} P(A|C_i)P(C_i|D). \tag{2.13}$$

Esta propiedad (llamada *Ley de la probabilidad total*) nos permite calcular $P(A)$ (y $P(A|D)$) considerando casos especiales (es decir, condicionando en conjuntos B_i o C_i).

Ejemplo 2.2.2 Retomamos el ejemplo de la Figura 5. Si llamamos M el evento de elegir una mujer (y así M^C denota un hombre) y L el evento de elegir alguien con lentes , entonces la probabilidad de elegir una persona con lentes está dada por:

$$P(L) = P(L|M)P(M) + P(L|M^C)P(M^C) = \frac{2}{7}\frac{7}{13} + \frac{4}{6}\frac{6}{13} = \frac{6}{13},$$

es decir calculamos una probabilidad relacionada con toda la población a través de probabilidades relacionadas con las subpoblaciones de hombres y mujeres.

A continuación formulamos la *regla de Bayes* que relaciona $P(A|B)$ con $P(B|A)$.

Propiedad 2.2.2 (Regla de Bayes) *Si* $P(A), P(B) > 0$,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2.14}$$

Un ejemplo de la aplicación de la regla de Bayes se da en el siguiente problema sobresimplificado del área de transmisión de información.

Ejemplo 2.2.3 Se envía un bit, 0 o 1, de un lugar a otro. El ruido sobre el canal puede cambiar el bit. La probabilidad de que un 0 se cambie a 1 o viceversa es igual a 10 %. Por experiencia se sabe que en 30 % de los casos se envía un 0, y en 70 % un 1. Supongamos que se recibe un 0, se calcula la probabilidad de que el bit original fuera 0 o 1, de la siguiente manera.

Llamamos "enviar un 0'' el evento E0 y "recibir un 0'' el evento R0. Así

$$P(R0|E0) = 0.9, \quad P(R0|E0^c) = 0.1, \quad P(E0) = 0.30.$$

La probabilidad de que el bit originalmente fuese 0 si recibimos un 0 es:

$$\begin{aligned} P(E0|R0) &= \frac{P(R0|E0)P(E0)}{P(R0)} = \frac{P(R0|E0)P(E0)}{P(R0|E0)P(E0) + P(R0|E0^c)P(E0^c)} \\ &= \frac{0.9 \cdot 0.3}{0.9 \cdot 0.3 + 0.1 \cdot 0.7} = 0.79. \end{aligned}$$

Luego, la probabilidad de que el bit original fuese 1 es igual a $P(E1|R0) = 1 - P(E0|R0) = 0.21$.

En la Figura 6, graficamos $P(E0|R0)$ en función de $x = 1 - P(R0|E0) = 1 - P(R1|E1)$ y $y = P(E0)$ para diferentes valores de x y y. A pesar de la sencillez del modelo, la relación es bastante compleja.

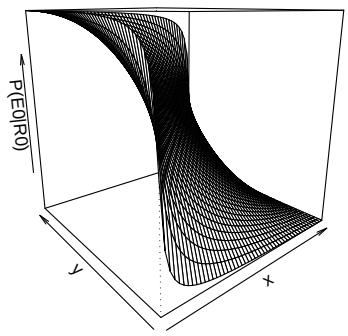


Figura 6.

Si se tiene que adivinar el valor de la señal original, una estrategia consiste en elegir el estado más probable. En este caso se tomará el valor 0. Veremos en el Capítulo 4 más ejemplos de este método de restauración.

El ejemplo anterior es representativo para un gran número de aplicaciones. A continuación damos una ilustración tomada del área de detección de enfermedades.

Ejemplo 2.2.4 Una compañía ha desarrollado una prueba para detectar la presencia de cáncer. Se pretende que $P(\text{prueba es positiva}|\text{tiene cáncer}) = 0.99$ y $P(\text{prueba es negativa}|\text{no tiene cáncer}) = 0.99$. Si el 1% de la población tiene cáncer, la probabilidad de decir que alguién tiene cáncer sin razón es:

$$P(\text{no tiene cáncer}|\text{prueba es positiva}) =$$

$$\frac{P(\text{prueba es positiva}|\text{no tiene cáncer})P(\text{no tiene cáncer})}{P(\text{prueba es positiva})}.$$

Si consideramos "tener cáncer o no'' como el equivalente a "enviar el bit 0 o 1'' y "prueba negativa o positiva'' como "recibir 0 o 1'', tenemos un problema equivalente al ejemplo anterior. Por consecuencia, usando (2.15):

$$P(\text{no tiene cancer}|\text{prueba es positivo}) = \frac{0.01 \cdot 0.99}{0.99 \cdot 0.01 + 0.01 \cdot 0.99} = \frac{1}{2}.$$

Obsérvese que aunque a primera vista la prueba parece muy confiable, el resultado obtenido muestra que es de poca utilidad práctica ya que es alta la probabilidad de un falso negativo.

2.2.2 Eventos independientes

Supongamos que lanzamos dos veces un dado. La información de que el resultado del primer tiro es 3, es irrelevante para decir algo acerca de la probabilidad de obtener un 5 en el segundo tiro. Si llamamos A y B a los eventos de que el primer tiro sea 3 y el segundo tiro sea 5 respectivamente, podemos formalizar lo anterior como:

$$P(B|A) = P(B). \tag{2.15}$$

Si la igualdad (2.15) se satisface, decimos que A y B son eventos independientes. Para incluir el caso que $P(B) = 0$ se usa la siguiente definición:

Definición 2.2.3 *Dos eventos A, B son independientes ssi*

$$P(A \cap B) = P(A)P(B). \tag{2.16}$$

Si $P(B) > 0$ la formulación (2.16) es más bien operativa en los cálculos mientras que (2.15) es más intuitiva.

Ejemplo 2.2.5 Consideremos la siguiente población:

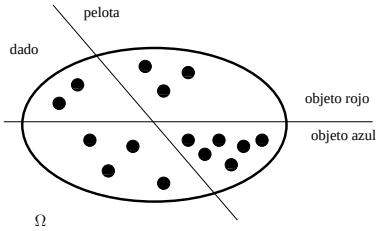


Figura 7.

La probabilidad de obtener un objeto rojo es independiente de obtener una pelota:

$$P(\text{pelota roja}) = P(\text{pelota})P(\text{objeto rojo}) = \frac{9}{15} \cdot \frac{5}{15} = \frac{1}{5}$$

o equivalente:

$$P(\text{objeto rojo}|\text{pelota}) = P(\text{objeto rojo}|\text{no es una pelota}) = P(\text{objeto rojo}).$$

Ejemplo 2.2.6 En el ejemplo 2.2.1, es claro que la aparición de la letra "c'' como primera letra y de la "a'' como segunda son eventos dependientes.

Obsérvese que la independencia es un concepto simétrico si los eventos A y B tienen probabilidad positiva. Es decir, si A y B son independientes, $P(A|B) = P(A)$ y $P(B|A) = P(B)$. Existe en la literatura algunos intentos para romper con esta simetría, y donde se prefiere hablar de (ir)relevancia de B para A como se formula en (2.15) para que $P(B|A) = P(B)$ sin que $P(A|B) = P(A)$. Por supuesto, esto implica considerar otro conjunto de axiomas para definir $P(\cdot)$ y las complicaciones de trabajar con otro sistema axiomático (al de Kolmogorov) , son en este caso, de tal naturaleza que la utilidad práctica es muy limitada.

2.3 Algunas aplicaciones más

A continuación damos dos aplicaciones sencillas de cómo usar los conceptos anteriores. El primer ejemplo ilustra el uso de la probabilidad para describir información y encontrar un estimador de una característica de interés.

En el segundo ejemplo usamos la probabilidad como sinónimo de no predictibilidad y es un ejemplo de un algoritmo aleatorio. Veremos en el Capítulo 3 otro ejemplo de esta familia en el contexto de optimización global.

2.3.1 Investigación de opinión

El siguiente ejemplo muestra cómo obtener y describir información global sobre una población sin tener que revelar la información de cada individuo usando una formu-

lación probabilística. Supongamos que una empresa quiere estimar el uso de su infraestructura de computadoras por parte de sus empleados para fines meramente personales. En este ejemplo supondremos que el interés es describir el porcentaje de empleados que abusan de la infraestructura de la empresa.

Si no se quiere investigar a todo el personal, se podría recurrir a una muestra, es decir, entrevistar a una selección arbitraria de la planta de personal. Como argumentaremos más adelante, el porcentaje de respuestas afirmativas a la pregunta de si usan la infrastructura para fines personales es, bajo algunas condiciones, un buen estimador para el porcentaje que se obtiene considerando todo el personal.

Dado que el miedo por represalias podría dificultar la implementación de una encuesta directa, una solución será el siguiente mecanismo.

```
• Cada empleado seleccionado, lanza una moneda sin mostrar el resultado
• Si es sol:
    él contesta la pregunta de si hace uso indebido de las computadoras
  Si es cruz:
    él contesta la pregunta de si tiene sangre de grupo O
```

El encuestador apunta solamente la respuesta final (sin haber sabido cuál pregunta se contestó). El porcentaje de respuestas positivas a la primera pregunta se puede recuperar de la siguiente manera.

Definimos S el evento de que se conteste "si'' y C el evento de que la moneda muestre el lado "cruz''. Usando (2.12), sabemos:

$$P(S) = P(S|C)P(C) + P(S|C^c)P(C^c).$$

Si la moneda es justa $P(C) = P(C^c) = 1/2$. Por otro lado, la probabilidad $P(S|C)$ se conoce y coincide con el porcentaje de gente con sangre de tipo O, digamos β. La cantidad que nos interesa estimar es $P(S|C^c)$, que es el porcentaje de personas que abusan de los recursos de la empresa y que llamaremos α. Sustituyendo y despejando en la relación anterior, obtenemos:

$$\alpha = 2 * P(S) - \beta.$$

Si aproximamos $P(S)$ por el porcentaje de respuestas positivas y dado que β es conocido, obtenemos un estimador para α sin haber conocido la respuesta a la pregunta original de cada persona!

2.3.2 Computación distribuida y algoritmos probabilísticos

En computación distribuida existen varios ejemplos donde agentes (procesadores) tienen que ponerse de acuerdo en la toma de una decisión, evitando problemas como que al-

guno se quede esperando la respuesta de otro, que no exista consenso en el acuerdo o que circulen diferentes versiones de lo que debería ser un mismo acuerdo.

Un ejemplo clásico es el problema de elegir un *líder* entre sí, es decir, asignar a un procesador algunos privilegios (por ejemplo para la coordinación de ciertos procesos). Supongamos que la comunicación entre n procesadores es a través de un anillo como se muestra en la Figura 8. Cada procesador sabe solamente el número de integrantes de la red. No se puede suponer que los procesadores tengan un nombre único y no existe un órgano coordinador.

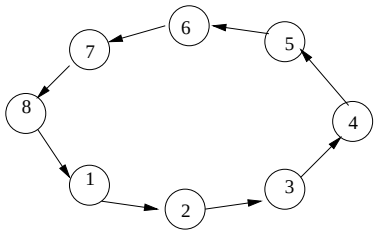


Figura 8.

Una solución consiste en que cada procesador elija al azar un número entre uno y n que funcionará como identificador y que se envia a los demás procesadores siguiendo el flujo que se presenta en la Figura 8. Cada uno guarda estos números. Si hay al menos un número único en la lista que al final todos tienen, se toma el procesador correspondiente al mayor número único como líder. En el caso contrario se repite todo de nuevo. El código de este algoritmo es el siguiente:

```
Repite
      nombre-propio =  elige un número aleatorio entre 1 y n
      lista-de-nombres =  vacio
      nombre = nombre-propio
      repite n - 1 veces:
            añade nombre a lista-de-nombres
            envia nombre al siguiente procesador
            nombre =  recibe nombre del procesador anterior
hasta que al menos un nombre en lista-de-nombres es único.
```

El líder será el que corresponda al mayor número único.

No es difícil ver que cada procesador de manera separada, puede determinar quién será el líder.

La probabilidad de que la lista generada por los procesadores no contenga algún número único, p, es menor que uno. Debido a esto y a que la ejecución del algoritmo en cada ocasión es independiente de las anteriores, se tiene que la probabilidad de que los procesadores no puedan escoger un líder en k repeticiones del algoritmo es igual a p^k y eso converge a cero conforme k crece. Este hecho garantiza que el `repite` externo en algún momento se va a terminar.

Es importante notar que en el algoritmo anterior, cada procesador ejecuta exactamente el mismo código, es decir la solución es simétrica. Se puede mostrar que no existen soluciones simétricas determinísticas.

Capítulo 3

Variables Aleatorias Discretas

En el capítulo anterior, para definir a la probabilidad, tomamos como punto de partida un experimento cuyo resultado denotamos con ω. En muchas situaciones tenemos interés solamente en algunas características particulares de ω; con este fin presentamos en este capítulo el concepto de *variable aleatoria*.

3.1 Motivación y Definición

Tomemos el siguiente ejemplo.

Ejemplo 3.1.1 Elegimos al azar una persona de un grupo. Muchas veces no tenemos tanto interés en la persona en sí, sino en su edad, altura o por ejemplo su peso. Para ese fin, construimos una función X que regresa para cada persona ω los valores que nos interesan: $X(\omega) = (X_1(\omega), \cdots, X_m(\omega))$, donde por ejemplo $X_1(\omega)$ representa su edad, $X_2(\omega)$ su altura, etc. Si el grupo de personas corresponde a una base de datos, entonces X regresa los campos de interés de cada registro. Observa que por razones que veremos más adelante, para referir a esta función no usaremos nombres como f o g sino letras en mayúsculas que se encuentran tipicamente al final del alfabeto.

A continuación, estudiamos el caso cuando Ω es finito o contable, suponiendo que $P(\cdot)$ está definido para cualquier subconjunto de Ω.

Definición 3.1.1 *Si Ω es finito o contable y sobre él está definido una función de probabilidad, $P(\cdot)$, para cada subconjunto de Ω, llamamos a cualquier función X de Ω a $\mathcal{R}^m$ una variable aleatoria discreta.*

A pesar de su nombre, una variable aleatoria determina una relación determinística, es decir, es una función clásica que mapea cada ω en Ω a un número real o a un vector de números reales.

Ejemplo 3.1.2 Tomemos el generador de passwords del ejemplo 2.1.2. Para cada

password ω, calculamos el número de consonantes que contiene, es decir definimos una función X como representada en la Figura 1. Entonces X es una variable aleatoria.

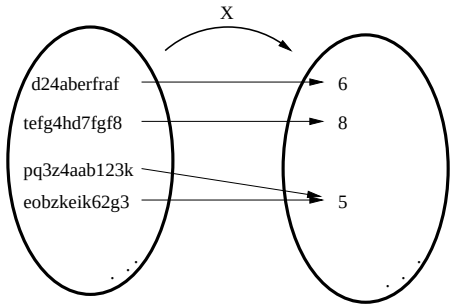


Figura 1.

Ejemplo 3.1.3 Un juego de kermesse consiste en derribar desde cierta distancia unas de las latas de la Figura 2 con una pelota. De acuerdo al diámetro de la lata derribada, se otorga un premio. Lata A corresponde a un premio de 5 pesos, lata B a 10 pesos y lata C a 50 pesos. Se excluye la posibilidad de derribar varias latas al mismo tiempo en un solo tiro.

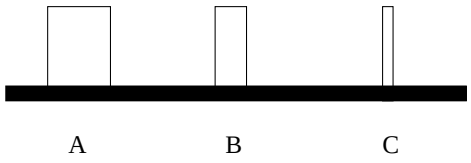


Figura 2.

En este caso, el experimento consiste en lanzar una pelota. El resultado será un elemento de $\Omega = \{A, B, C, N\}$ donde A, B, C denota la lata correspondiente y N significa que no se pegó a ninguna lata.

Definimos X como la función que mapea cada resultado con el valor del premio otorgado (Figura 3.)

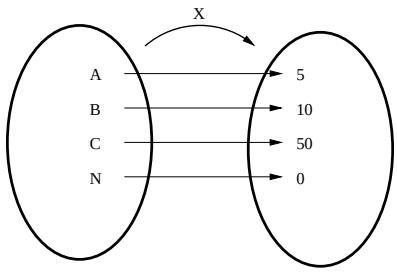


Figura 3.

Dada una distribución sobre Ω, X es una variable aleatoria.

Ejemplo 3.1.4 Una imagen discreta en escala de color de gris es un mosaico de elementos básicos (por ejemplo rectángulos), llamados pixeles, organizados sobre una

rectícula regular. Cada píxel tiene un nivel de gris particular que es codificado numéricamente, por ejemplo blanco corresponde a 0, negro a 1 y entre más oscuro, más cercano el valor a 1. Así una imagen puede ser descrita por una matriz M donde $M_{i,j}$ denota el nivel de gris de la imagen en la posición (i, j).

Recibimos una imagen binaria de tamaño $k \times l$ pixeles. Definimos a Ω como el conjunto de todas las posibles imágenes binarias de tamaño $k \times l$. Supongamos que se sabe $P(\{\omega\})$, la probabilidad de recibir la imagen particular ω.
Definimos $X(\omega) = (X_{1,1}(\omega), \cdots, X_{k,l}(\omega))$ donde $X_{i,j}(\omega) = 1$ si el pixel (i, j) en la imagen ω es negro y $X_{i,j} = 0$ en el caso que sea blanco (ver la Figura 4 para el caso $k = l = 3$). La función X es una variable aleatoria.

Obsérvese que si $Y(\omega)$ denota el número de pixeles negros en ω, podemos definirlo usando las variables aleatorias X's, considerando $Y(\omega) = \sum_{i,j} X_{i,j}(\omega)$.

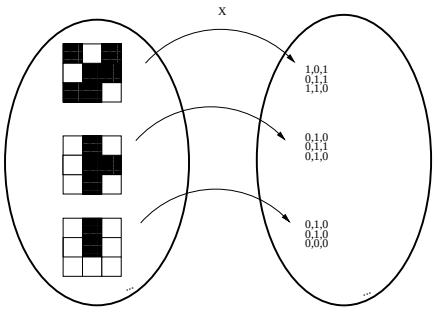


Figura 4.

Como vemos, la definición de las variables aleatorias nos permite aislar o comparar distintos aspectos de un elemento ω. Dado que su valor pertenece a $\mathcal{R}^m$, técnicamente se puede hacer cálculos con ellos pero sólo algunos pueden tener una interpretación (cfr. calcular dos veces la edad de una persona versus tomar dos veces su sexo). Las variables aleatorias denotamos con mayúsculas $X, Y, \cdots$ y los valores que toman con minúsculas $x, y, \cdots$.

3.2 La distribución de una variable aleatoria

Debido a que se tienen probabilidades $P(\cdot)$ definidas sobre Ω, se pueden asociar probabilidades con X. Definimos $P_X(\cdot)$ como

$$P_X(A) = P(\{\omega : X(\omega) \in A\}), \tag{3.1}$$

la probabilidad que se obtenga un valor que X mapea a un elemento que pertenece a A. Por las suposiciones anteriores sobre $\mathcal{B}$, esta probabilidad existe.
En la práctica se abrevia $P_X(\cdot)$ como $P(\cdot)$ y por ejemplo, se escribe $P(X = x)$ para denotar $P_X(X = x) = P(\{\omega : X(\omega) = x\})$, o $P(X < x)$ para denotar $P_X(X < x) = P(\{\omega : X(\omega) < x\})$. Es fácil mostrar que la nueva $P(\cdot)$ cumple con todas las características de una función de probabilidad (definición 2.1.2), solamente que ahora está definida sobre $\mathcal{R}^m$.

Ejemplo 3.2.1 Retomemos el ejemplo de una base de datos donde solamente tenemos interés en el número de hermanos que cada uno tiene; así $m = 1$:

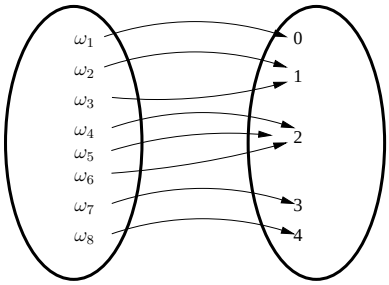


Figura 5.

A continuación calculamos $P(X = x)$ para dos conjuntos de probabilidades diferentes sobre Ω.

1. Si definimos la distribución de conteo sobre Ω, $P(X = x)$ toma los siguientes valores:

x	0	1	2	3	4
P(X=x)	1/8	1/4	3/8	1/8	1/8

Tabla 1.

2. Si definimos sobre Ω las siguientes probabilidades:

$$P(\{\omega_1\}) = 0.3, \quad P(\{\omega_2\}) = 0.1, \quad P(\{\omega_3\}) = 0.1, \quad P(\{\omega_7\}) = 0.5,$$

y para los demás elementos $P(\omega_i) = 0$, $P(X = x)$ toma ahora los siguientes valores:

x	0	1	2	3	4
P(X=x)	0.3	0.2	0	0.5	0

Tabla 2.

Para calcular $P(X \geq 3)$ (suponiendo las probabilidades de Tabla 1), podemos seguir dos caminos:

1. usando la definición de una variable aleatoria:

$$P(X \geq 3) = P(\{\omega : X(\omega) \geq 3\}) = \#\{\omega : X(\omega) \geq 3\}/\#\Omega = 1/4;$$

2. considerando $P(\cdot)$ como una probabilidad propia y usar las propiedades correspondientes:

$$P(X \geq 3) = P(X = 3 \text{ o } X = 4) = P(X = 3) + P(X = 4) = 1/8 + 1/8 = 1/4.$$

La propiedad que se utiliza en este caso es la aditividad de $P(\cdot)$ para conjuntos ajenos (en este ejemplo: nadie puede tener al mismo tiempo 3 y 4 hermanos).

En muchas situaciones trabajamos con X como si la aleatoriadad tuviera su origen ahí, olvidándonos del conjunto Ω y del experimento asociado. Lo anterior permite estudiar a las probabilidades (distribuciones) de una manera más general.

Definición 3.2.1 *Sea X una variable aleatoria discreta. Llamamos al conjunto de probabilidades $\{P(X = x)\}$ la distribución de X.*

Definición 3.2.2 *Decimos que dos variables discretas, X y Y, tienen la misma distribución ($X \sim Y$) si para cada x, $P(X = x) = P(Y = x)$.*

Si lanzamos dos veces una moneda, la distribución del resultado del primer tiro es igual a la distribución del resultado del segundo. Por supuesto, los resultados pueden ser diferentes.

Obsérvese que aunque podemos considerar diferentes variables aleatorias al mismo tiempo, el cálculo de probabilidades donde aparecen juntas, tiene solamente sentido si están definidas sobre la misma Ω. Por ejemplo para poder escribir $P(X_1 = X_2)$ necesitamos que X_1, X_2 estén definidas sobre la misma Ω porque $P(X_1 = X_2)$ se refiere a la probabilidad del conjunto $\{\omega : X_1(\omega) = X_2(\omega)\}$.

3.3 La distribución acumulativa de una variable aleatoria

En base de lo anterior se define la distribución acumulativa de X.

Definición 3.3.1 *Sea X una variable aleatoria de Ω a $\mathcal{R}^m$. Si $m = 1$, llamamos $F_X(x) = P(X \leq x)$ la distribución acumulativa de X. En general, $F_X(x_1, \cdots, x_m) = P(X_1 \leq x_1, \cdots, X_n \leq x_m)$ es la distribución acumulativa de X.*

Es fácil ver que $F_X(\cdot)$ es una función no decreciente cuyos valores se encuentran en el intervalo $[0, 1]$.

Ejemplo 3.3.1 Para las probabilidades descritas en la Tabla 1, $F_X(\cdot)$ se presenta en Tabla 3 y Figura 6 muestra la gráfica correspondiente.

x	< 0	$0 \leq x < 1$	$1 \leq x < 2$	$2 \leq x < 3$	$3 \leq x < 4$	≥ 4
$F_X(x)$	0	$\frac{1}{8}$	$\frac{1}{8} + \frac{1}{4}$	$\frac{1}{8} + \frac{1}{4} + \frac{3}{8}$	$\frac{1}{8} + \frac{1}{4} + \frac{3}{8} + \frac{1}{8}$	1

Tabla 3.

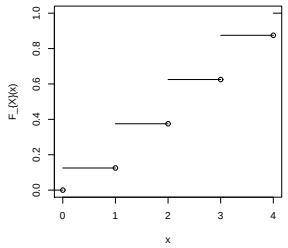


Figura 6.

Como $X \in \mathcal{N}$, entonces si $x \in \mathcal{N}$, $F_X(x) = F_X(x-1) + P(X = x)$, es decir cada uno de los brincos en la gráfica tiene como altura $P(X = x)$.

3.4 Distribuciones condicionales y variables independientes

Dado que en el fondo la distribución de una variables aleatoria está definida en términos de probabilidades sobre Ω, podemos fácilmente extender conceptos derivados como probabilidades condicionales e independencia, al caso de variables aleatorias.

Definición 3.4.1 *Para las variables aleatorias discretas X_1, X_2 y un subconjunto S, definimos la distribución condicional de X_1 dado $X_2 \in S$ como:*

$$P(X_1 = x | X_2 \in S) = \frac{P(X_1 = x, X_2 \in S)}{P(X_2 \in S)}, \tag{3.2}$$

suponiendo que $P(X_2 \in S) > 0$.

Obsérvese que lo anterior no es una definición nueva sino que se deriva directamente de la definición de probabilidad condicional que se presenta en el capítulo anterior.

$$P(X_1 = x | X_2 \in S) = P(\{\omega : X_1(\omega) = x\} | \{\omega_1 : X_2(\omega_1) \in S\}) =$$

$$\frac{P(\{\omega : X_1(\omega) = x\} \cap \{\omega_1 : X_2(\omega_1) \in S\})}{P(\{\omega_1 : X_2(\omega_1) \in S\})} = \frac{P(X_1 = x, X_2 \in S)}{P(X_2 \in S)}.$$

Ejemplo 3.4.1 Se tiene que elegir al azar un punto (O) en el triangulo representado en la Figura 7. Llamaremos X y Y a la primera y segunda coordenada del punto que se seleccionará.

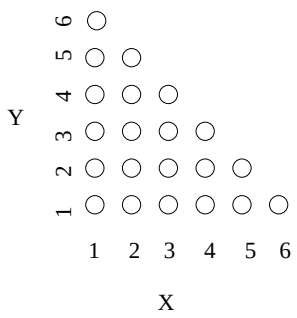


Figura 7.

Por ejemplo, se calculan las probabilidades condicionales:

$$P(Y=4|X=2)=\frac{P(X=2,Y=4)}{P(X=2)}=\frac{1/21}{5/21}=\frac{1}{5}$$

y

$$P(Y=4|X=3)=\frac{P(X=3,Y=4)}{P(X=3)}=\frac{1/21}{4/21}=\frac{1}{4}$$

Una generalización del concepto de independencia para variables aleatorias se da en la siguiente definición.

Definición 3.4.2 *Si X_1 y X_2 son variables aleatorias discretas sobre un mismo espacio de probabilidad, las dos son independientes ssi*

$$P(X_1=x,X_2=y)=P(X_1=x)P(X_2=y) \quad \textit{para cada } x,y\in\mathcal{R}^n \tag{3.3}$$

Se puede mostrar que lo anterior es equivalente a (referimos a un curso de medida para una definición rigurosa):

$$P(X_1\in M,X_2\in N)=P(X_1\in M)P(X_2\in N), \text{para (casi) cualquier conjunto } M,N \tag{3.4}$$

Por ejemplo, si X y Y son independientes,

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

donde tomamos en (3.4) $M=\{z:z<x\}$ y $N=\{z:z<y\}$.

Ejemplo 3.4.2 En el ejemplo 3.4.1, X y Y no son independientes, pero si la figura fuera un rectángulo, no es difícil verificar que se tendría que X y Y sí lo son.

Cuando se tienen dos variables aleatorias independientes X_1 y X_2 se puede recuperar u obtener la *distribución conjunta* $P(X_1=x,X_2=y)$ a partir de la multiplicación de las *distribuciones marginales*: $P(X_1=x)$ y $P(X_2=y)$.

Obsérvese que en muchas situaciones, por la propia naturaleza del experimento, podemos suponer independencia y definir la distribucion conjunta $P(X_1=x_1,\cdots,X_n=x_n)$ a través de las probabilidades marginales como en (3.3).

Ejemplo 3.4.3 Definimos X, Y como el resultado del primer y segundo tiro de un dado respectivamente. Por construcción, X y Y son independientes, así usando (3.3),

$$P(X = 2, Y = 1) = P(X = 2)P(Y = 1) = \frac{1}{6}.\frac{1}{6} \ .$$

Finalmente a continuación se extenderá el concepto de independencia a un conjunto de variables aleatorias.

Definición 3.4.3 *Decimos que las variables aleatorias* $\{X_1, \cdots, X_m\}$ *son independientes ssi*

$$P(\cap_{i\in J}\{\omega : X_i(\omega) = x_i\}) = \prod_{i\in J} P(X_i = x_i)$$

para todo conjunto de índices $J \subset \{1, \ldots, m\}$ *y cada* x_i.

3.5 Ejemplos de distribuciones discretas

A continuación daremos las distribuciones discretas más conocidas. Todas están definidas sobre los números enteros.

1. **Distribución Bernoulli**: esta distribución surge en experimentos con solamente dos resultados diferentes y que codificamos como 0 y 1. Por ejemplo el lanzamiento de una moneda, el éxito o fracaso de cierto experimento, el sexo de una persona, etc. Decimos que X tiene una distribución Bernoulli, cuando:

   $$X \sim \text{Bern}(\theta) \Leftrightarrow P(X = 1) = 1 - P(X = 0) = \theta,$$

   donde θ es un parámetro libre entre 0 y 1 que representa la probabilidad de obtener 1 como resultado.

   Si los componentes de X del ejemplo 3.1.4 son mutuamente independientes y distribuidos como Bern(θ), es decir para cada pixel lanzamos una moneda para determinar su valor (1 si es sol y 0 si es cruz), con diferentes valores de θ, $0.5, 0.3$ y 0.1 obtenemos las siguientes imagenes (en este caso $k = l = 32$).

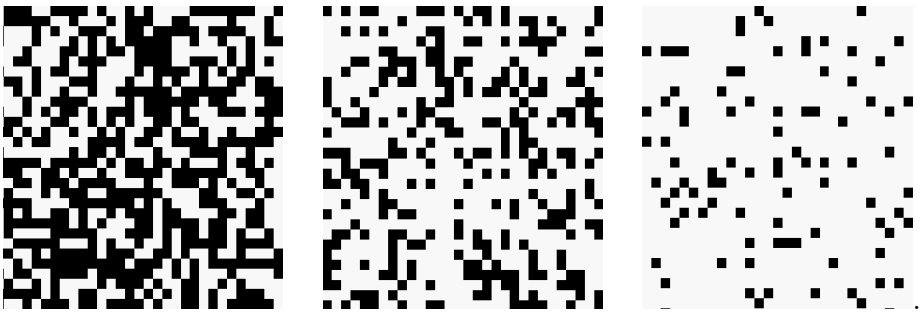


Figura 8.

La función de acumulación de X tiene la siguiente expresión:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1-\theta & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

2. **Distribución Binomial**: esta distribución surge cuando se suman n variables Bernoulli independientes con el mismo parámetro θ. Un ejemplo es el número de éxitos en n repeticiones de un experimento donde la probabilidad de éxito en cada experimento es igual a θ. Decimos que X tiene una distribución Binomial, cuando:

$$X \sim \text{Bin}(n, \theta) \Leftrightarrow P(X = x) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x}, \quad x \in \{0, \cdots, n\},$$

donde n y θ son párametros; $\theta \in [0, 1]$ denota la probabilidad de éxito y $n \in \mathcal{N}$ el número de experimentos.

Para tener una idea del efecto que realizan los parámetros sobre la forma de la distribución, en la Figura 9 se muestran las probabilidades de la distribución Binomial con parámetos (10,0.5) y (10,0.8) y en la Figura 10 se grafica la misma distribución pero ahora con parámetros (30,0.5) y (30,0.8).

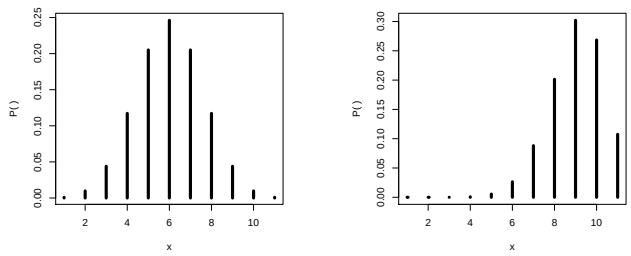


Figura 9.

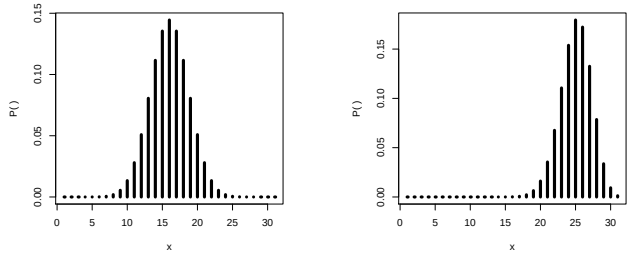


Figura 10.

La distribución acumulativa es

$$F(x) = \sum_{i=0}^{\lfloor x \rfloor} \begin{pmatrix} n \\ i \end{pmatrix} \theta^i (1-\theta)^{n-i},$$

donde $\lfloor x \rfloor = \max\{y \in \mathcal{N} : y \leq x\}$.

3. **Distribución Geométrica**: esta distribución surge cuando se determina el momento del primer éxito en una secuencia de experimentos independientes tipo Bernoulli y con el mismo parámetro θ. Decimos que X tiene una distribución Geométrica, cuando:

$$X \sim \text{Geo}(\theta) \Leftrightarrow P(X = x) = \theta(1-\theta)^{x-1}, \quad x \in \{1, 2, \ldots\},$$

donde θ es un párametro; $\theta \in [0, 1]$ y denota la probabilidad de éxito.

En la Figura 11 se muestra la distribución Geométrica con parámetros $\theta = 0.7$ y $\theta = 0.5$.

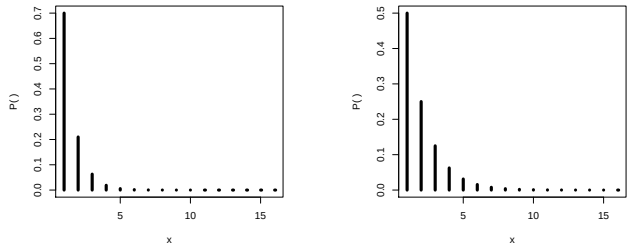


Figura 11.

Se obtiene

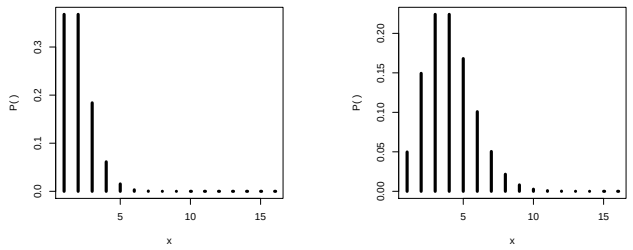
$$F(x) = 1 - (1-\theta)^{\lfloor x \rfloor}$$

4. **Distribución Uniforme**: La distribución uniforme sobre un conjunto de valores surge al elegir un elemento de este conjunto al azar, i.e. sin preferencia. Para un conjunto A fijo, escribimos:

$$X \sim \mathcal{U}(A) \Leftrightarrow P(X = x) = \frac{1}{\#A}$$

5. **Distribución Poisson**:

$$X \sim \text{Poisson}(\lambda) \Leftrightarrow P(X = x) = \frac{\exp(-\lambda)\lambda^x}{x!}, \quad x \in \{0, 1, 2, \ldots\}.$$



Poisson(1) Poisson(3)

Figura 12.

3.6 El promedio

En esta sección introducimos el concepto promedio de una distribución. Es el primero de una serie de medidas que resumen cada una algun aspecto particular de una distribución.

Nos restringimos en esta sección a variables aleatorias unidimensionales.

Promedio

Definición 3.6.1 *El promedio o esperanza de una variable discreta X, EX, se define como:*

$$EX = \sum_x xP(X = x), \tag{3.5}$$

en caso de que la suma exista.

En la vida cotidiana un promedio sobre un conjunto de numeros $\{x_i\}_1^n$ corresponde a

$$\bar{x} = \frac{\sum_i x_i}{n} \tag{3.6}$$

Se puede motivar el doble uso del nombre *promedio* de dos formas. Si X tiene la distribución de conteo sobre $\{1, \cdots, n\}$, entonces $EX = \sum_x x/n$ y así la definición 3.6.1 y el promedio aritmético coinciden. Para una distribución con probabilidades racionales, podemos extender la interpretación como explicamos en el Capítulo 2.
Por ejemplo, si $X \sim \text{Bern}(0.6)$, construimos $P(\cdot)$ como la distribución de conteo sobre $\Omega^* = \{0, 0, 1, 1, 1\}$, y X la función como en la Figura 13.

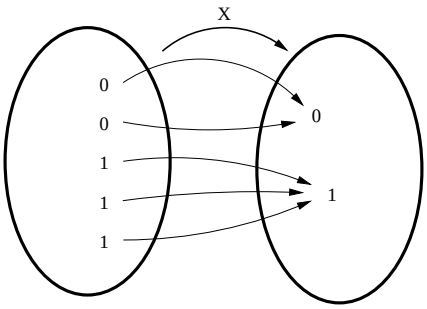


Figura 13.

Entonces $EX = 0P(X = 0) + 1P(X = 1) = 0(2/5) + 1(3/5) = (0 + 0 + 1 + 1 + 1)/5 = \sum_{x_i \in \Omega^*} x_i/n$ lo que coincide con (3.6).

Observa que

$$\sum_x (x - EX)P(X = x) = 0,$$

entonces si consideramos $P(X = x)$ como la masa de un punto en posición x, vemos que EX corresponde al centro gravitacional. Por eso el promedio es una medida de localización de la distribución.

Más adelante veremos la ley de los números grandes que ofrece otra interpretación al promedio.

Ejemplo 3.6.1 Si $X \sim \text{Bern}(\theta)$, entonces:

$$EX = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = \theta$$

Ejemplo 3.6.2 Si $X \sim \text{Geo}(\theta)$,

$$EX = \sum_{x=1}^{\infty} x\theta(1-\theta)^{x-1} = \theta \sum_{x=1}^{\infty} x(1-\theta)^{x-1}$$

Usando la relación

$$\sum_{x=1}^{\infty} xz^{x-1} = \frac{1}{(1-z)^2}, \quad |z| < 1$$

obtenemos

$$EX = \frac{\theta}{(1-(1-\theta))^2} = \frac{1}{\theta}.$$

Ejemplo 3.6.3 Supongamos que un hacker trata de adivinar un password generado como en el ejemplo 2.1.2. Escribe su propio generador aleatorio y prueba cada combinación hasta tener éxito. Si el retraso entre dos sesiones de login es de 3 segundos, se puede calcular el tiempo en promedio que va a necesitar para accesar al sistema.

Sea X_i la variable que indica el éxito o fracaso del i-ésimo intento. Sabemos que $X_i \sim \text{Bern}(\theta)$, donde $\theta = 1/\#\Omega$ y la cardinalidad de Ω es el número total de passwords con al menos un número. Esta cifra es igual a todos los posibles passwords de 12 carácteres menos los que no tienen números. Luego entonces, la cardinalidad de Ω es igual a

$$\#\Omega = (2 * 26 + 10)^{12} - (2 * 26)^{12} \sim 2^{21}$$

Si llamamos Y al número de ensayos del hacker en el que ocurre el primer éxito en la secuencia $\{X_i\}$, suponiendo independencia, obtenemos que $Y \sim Geo(\theta)$.

Así entonces el tiempo que va a tardar en promedio es 3 segundos por el promedio del número de intentos.

$$3EY = 3\frac{1}{1/2^{21}} \sim 2^{22} \text{ segundos}$$

Como comparación, se estima que el número de segundos transcuridos desde el inicio de nuestro sistema solar es del orden 2^{34}. Por supuesto, nada impide que - por casualidad - se encontre el password en pocos intentos pero la probabilidad de que eso ocurre es muy chiquita!

Definición 3.6.2 *Para una función $g(\cdot)$, se define $Eg(X)$ de una variable discreta como:*

$$Eg(X) = \sum_x g(x)P(X = x),$$

en caso de que la suma exista.

En realidad la definición anterior no es estrictamente necesaria:

$$\begin{aligned} EY &= \sum_y yP(Y = y) = \sum_y yP(g(X) = y) \sum_y y \sum_x I(g(x) = y)P(X = x) \\ &= \sum_x \left(\sum_y yI(g(x) = y) \right) P(X = x) = \sum_x g(x)P(X = x). \end{aligned}$$

Propiedad 3.6.1 *Para cada X, Y, variables aleatorias no necesariamente independientes:*

$$E(aX + bY) = aEX + bEY, \tag{3.7}$$

donde a y b son constantes.

Obsérvese que en expresiones tales como $aX + b$, son un caso particular de $aX + bY$, considerando a Y como la variable aleatoria degenerada en $y = 1$, puesto que mapea cada ω al punto y, que en este caso es igual a 1. Así, por la propiedad 3.6.1, $E(aX+b) = aEX + b$. Similarmente $E(EX) = EX$.

Ejemplo 3.6.4 Sea $X \sim \text{Bin}(n, \theta)$. Como podemos interpretar a X igual a

$$X = \sum_i Y_i,$$

con $Y_i \sim Bern(\theta)$ e independientes, usando la propiedad 3.6.1 obtenemos:

$$EX = E \sum_i Y_i = \sum_i EY_i = n\theta.$$

La contraparte de 3.6.1 para productos de variables aleatorias debe suponer independencia entre las variables.

Propiedad 3.6.2 *Si X y Y son independientes, se tiene que*

$$E(XY) = EX \cdot EY$$

3.6.1 Promedio condicional

Definición 3.6.3 *Para las variables X y Y y un evento A , se define el promedio (o la esperanza) condicional de X dado A como:*

$$E(X|A) = \sum_x xP(X = x|A),$$

y la esperanza condicional de X dado que Y es igual a un valor y, como:

$$E(X|Y=y) = \sum_x xP(X=x|Y=y),$$

cuando $P(A) > 0$ y $P(Y=y) > 0$.

En la práctica se abrevia $E(X|Y=y)$ a $E(X|Y)$. Obsérvese la diferencia entre

$$E(X+Z|Y=y) = \sum_{x,z}(x+z)P(X=x, Z=z|Y=y)$$

y

$$E(X+z|Y=y) = \sum_x (x+z)P(X=x|Y=y).$$

La regla es que se toma el promedio con respecto a las variables aleatorias cuyo valor no fue fijado. Así, por el hecho que $E(X|Y=y)$ es una función de y,

$$E(E(X|Y=y)) = \sum_y E(X|Y=y)P(Y=y).$$

A partir de la definición de la esperanza condicional, se observa que si X y Y son variables aleatorias independientes, $E(X|Y=y) = E(X)$.

Usando la Propiedad 2.2.1 se demuestran las siguientes propiedades.

Propiedad 3.6.3 *Para cualesquier X y Y se cumple que*

$$EX = E(E(X|Y=y)),$$

y para cualquier partición $\{A_i\}$,

$$EX = \sum_i E(X|A_i)P(A_i).$$

Demostración

$$\begin{aligned} E(E(X|Y=y)) &= \sum_y E(X|Y=y)P(Y=y) = \sum_y \left(\sum_x xP(X=x|Y=y)\right)P(Y=y) \\ &= \sum_x x \sum_y P(X=x|Y=y)P(Y=y). \end{aligned}$$

Por la propiedad (2.12) se tiene entonces que

$$E(E(X|Y=y)) = \sum_x xP(X=x) = EX.$$

La demostración de la segunda igualdad es análoga a la anterior.

$\diamondsuit$

Una propiedad de la esperanza simple, que se extiende a la esperanza condicional es la siguiente.

Propiedad 3.6.4 *Para cada X, Y y Z, variables aleatorias no necesariamente independientes, y a y b reales, se tiene que*

$$E(aX + bZ|Y = y) = aE(X|Y = y) + bE(Z|Y = y)$$

Ejemplo 3.6.5 Las propiedades anteriores juegan un papel importante en el cálculo de la complejidad de un algoritmo. Consideramos el siguiente código:

```
{ --A --
  if (p)
    --B--
  else
    --C--
 }
```

donde p es `true` con probabilidad 0.3 y A, B y C son grupos de comandos.
Supongamos que el tiempo de ejecución de A,B y C son variables aleatorias con promedio igual a 8, 10 y 12 segundos, respectivamente.

Definimos X como el tiempo total de ejecución del programa. Usando la propiedad 3.6.1 se obtiene que el promedio es igual a:

$$\begin{aligned} EX &= EA + E(B|p = true)P(p = true) + E(B|p = false)P(p = false) \\ &= 8 + 10 \cdot 0.3 + 12 \cdot 0.7 = 19.4. \end{aligned}$$

Una extensión es el siguiente programa.

```
read (N)
For i=1 to N do
  {
    --A --
    if (p)
      --B--
    else
      --C--
  }
```

Considera N como una variable aleatoria con promedio 5 e independiente de los tiempos de ejecución de A, B y C.

Definimos X_i como el tiempo para ejecutar los comandos dentro del `For` la i-ésima vez. Supongamos que todas son independientes entre sí y con la misma distribución que X del ejemplo anterior.

Así el tiempo promedio de ejecución es $E(\sum_{i=1}^{N} X_i)$.

Usando propiedad 3.6.3, y las suposiciones de independencia, obtenemos:

$$\begin{aligned} E(\sum_{i=1}^{N} X_i) &= E(E(\sum_{i=1}^{N} X_i | N = n)) \\ &= E(E(\sum_{i=1}^{n} X_i | N = n)) \\ &= E(E(nX | N = n)) \\ &= E(E(NX | N = n)) \\ &= E(NE(X)) \\ &= EN \cdot EX = 5 \cdot 19.4. \end{aligned}$$

3.7 La varianza y la entropía

El promedio es una característica importante de una distribucíon pero no refleja nada sobre la variabilidad en los valores que se obtienen. A continuación daremos las medidas más usadas. En general, entre más variabilidad una variable tiene, más dificil es predicir su valor.

3.7.1 La varianza

La varianza es una medida de la dispersión de la distribución sobre los valores. A continuación se da su definición.

Definición 3.7.1 *Para una variable aleatoria X, con $E(X^2) < \infty$, se define la varianza, Var(X) como:*

$$Var(X) = E(X - EX)^2. \tag{3.8}$$

Como se puede interpretar $(x-a)^2$ como (el cuadrado de) la distancia euclidiana entre x y a, vemos que la varianza es el error en promedio si aproximamos X por su promedio EX. Si $\text{Var}(X) = 0$, tenemos que X es la función constante que mapea todo al valor EX con probabilidad 1.

Si se expande la expresión $(X-EX)^2$ y se usa la Propiedad 3.6.1, se obtiene la siguiente definición equivalente de la varianza:

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

Ejemplo 3.7.1 Sea $X \sim \text{Bern}(\theta)$, entonces $\text{Var}(\theta) = (1-\theta) \cdot (0-\theta)^2 + \theta \cdot (1-\theta)^2 = \theta(1-\theta)$. En este caso se observa que la varianza se maximiza para $\theta = 0.5$.

Ejemplo 3.7.2 En la Figura 15, se muestran dos imágenes donde los valores de los pixeles se distribuyen independientemente entre sí y según dos distribuciones d1 y d2 con promedios iguales pero varianzas distintas

La contraparte de la propiedad 3.6.1, ahora para la varianza, se presenta en seguida.

Propiedad 3.7.1 *Si X, Y son independientes:*

$$Var(aX + bY) = a^2\, Var(X) + b^2\, Var(Y)$$

Ejemplo 3.7.3 Si X y Y son variables aleatorias independientes con la misma varianza, entonces

$$\mathrm{Var}(\frac{X+Y}{2}) = \frac{\mathrm{Var}(X) + \mathrm{Var}(X)}{4} = \frac{\mathrm{Var}(X)}{2} < \mathrm{Var}(X).$$

En consequencia, promediar siempre disminuye la varianza.

Ejemplo 3.7.4 Si $X \sim \mathrm{Bin}(n, \theta)$, usando la propiedad 3.7.1, se obtiene que $\mathrm{Var}(X) = n(1-\theta)\theta$.

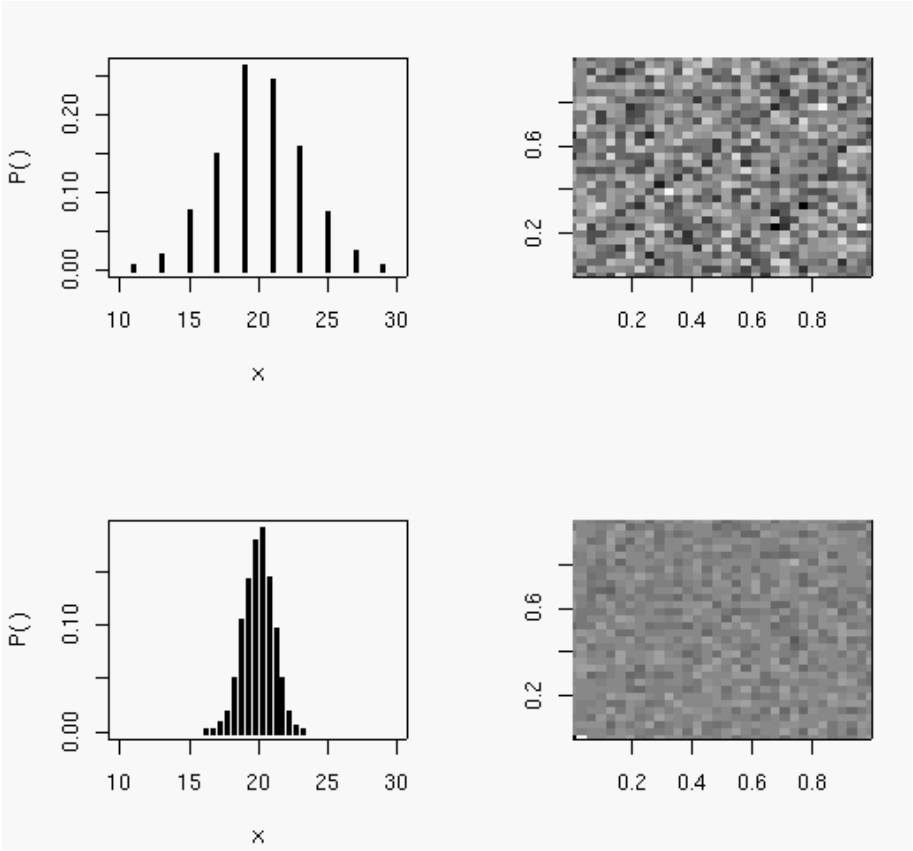


Figura 15.

3.7.2 La entropía

El punto de partida en la derivación de la varianza fue una medida de distancia entre una observación y el promedio. Depende de la variable si esta distancia tiene una

interpretación en términos de las variables: muy comúnmente los valores son solamente una codificación sin significado adicional.

A diferencia de la varianza, la entropía toma como punto de partida el concepto de información y no depende de la codificación.

Supongamos que X representa al número de errores tipográficos en un texto. ¿Cuánta información recibimos si alguién nos proporciona el valor x para un texto particular? Sin duda, la información obtenida dependerá de $P(X = x)$. Si x es un valor que ocurre con gran frecuencia (digamos un caso "normal''), nos sorprenderá menos que recibir un valor con poca probabilidad, en ese caso eso significa recibir un valor muy chico o muy muy grande.

Así cualquier medida de información, $I(x)$, debe satisfacer a:

$$I(x) > I(y) \quad \text{si} \quad P(X = x) < P(X = y). \tag{3.9}$$

Una elección muy popular es

$$I(x) = -\log(P(X = x)). \tag{3.10}$$

A continuación se define la entropía de una distribución como la información promedio obtenida.

Definición 3.7.2 *Dada la variable aleatoria discreta X, la entropía de X, $H(X)$ es*

$$H(X) = EI(X) = -\sum_x \log(P(X = x))P(X = x).$$

donde $(\log 0)0$ *se define igual a* 0.

Obsérvese que la entropía se define para variables de cualquier dimensión.

Ejemplo 3.7.5 Para una distribución degenerada en un punto x la entropía es igual a 0.

Ejemplo 3.7.6 Para una distribución de conteo sobre Ω con $\#\Omega = n$:

$$H(X) = -\sum_x \frac{1}{n} \log \frac{1}{n} = \log n. \tag{3.11}$$

Se puede mostrar que para cualquier otra distribución sobre Ω con $\#\Omega = n$, la entropía es acotada por (3.11).

Dos propiedades importantes son

$$H(X) \geq 0 \tag{3.12}$$

y si X y Y son independientes,

$$H(X, Y) = H(X) + H(Y) \tag{3.13}$$

De hecho se puede puede mostrar que la única familia de funciones que satisfacen las propiedades anteriores, son de la forma $cH(X)$ con $c > 0$.

Otra elección popular para $I(x)$ es

$$I_2(x) = 1 - P(X = x). \tag{3.14}$$

la que conduce a la siguiente medida de variabilidad:

$$EI_2(X) = \sum_x P(X = x)(1 - P(X = x)). \tag{3.15}$$

La cantidad $I_2(X)$ tiene la siguiente interpretación. Supongamos que tenemos que adivinar el resultado x. Para eso generamos un valor de la distribucíon X. La probabilidad que nos equivoquemos es igual a

$$P(\text{nos equivocamos}) = \sum_x p(X = x)P(\text{nos equivocamos}|X = x) =$$

$$\sum_x P(X = x)P(\text{no generar x}) = \sum_x P(X = x)(1 - P(X = x)) = EI_2$$

Obsérvese que para el caso de una distribución Bernoulli, esta medida coincide con la varianza.

3.8 Transformaciones

A partir de una variables aleatoria X podemos construir otras. Por ejemplo, si X representa la longitud del lado de un cuadrado, podemos tener interés en su área, i.e., $Y = X^2$.
En el caso particular de variables discretas, es relativamente sencillo determinar la distribución de Y a partir de la distribución de X por la aditividad de las probabilidades.
Considera como ejemplo la Figura 16.

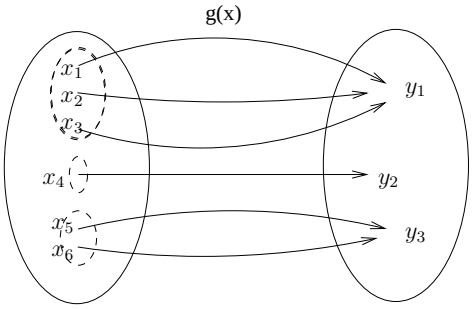


Figura 16.

Si $Y = g(X)$, no es dificil ver que

$$P(Y = y_1) = P(X = x_1) + P(X = x_2) + P(X = x_3).$$

En general se obtiene:

$$P(Y = y) = P(g(X) = y) = \sum_x P(X = x) I(g(x) = y), \tag{3.16}$$

donde $I()$ es la función indicador.

Ejemplo 3.8.1 Para la distribución de Tabla 1 del capítulo anterior, si definimos $Y = |X - 2|$, obtenemos:

$$P(Y = 0) = P(X = 2) = 3/8; P(Y = 1) = P(X = 1) + P(X = 3) = 3/8;$$

$$P(Y = 2) = P(X = 0) + P(X = 4) = 2/8.$$

Probablemente el ejemplo más conocido y usado es la transformación de dos variables independientes X y Y en su suma $Z = X + Y$.
Usando la ley de la probabilidad total obtenemos que

$$P(Z = z) = \sum_x P(Z = z|X = x) P(X = x). \tag{3.17}$$

Dado que

$$P(Z = z|X = x) = P(X + Y = z|X = x) = P(Y = z - x|X = x) \stackrel{*}{=} P(Y = z - x),$$

donde usamos en (*) la independencia entre X y Y, (3.17) se convierte en

$$P(Z = z) = \sum_x P(Y = z - x) P(X = x),$$

es decir, la distribución de la suma es la convolución de las distribuciones subyacentes.

3.9 Aplicaciones

3.9.1 Codificación de señales

Supongamos que queremos enviar de forma digital un mensaje entre dos lugares. El mensaje es una secuencia formada por cinco diferentes símbolos. Denotamos el conjunto de posibles símbolos, *el alfabeto*, por $\{\omega_1, \cdots, \omega_5\}$.

La tabla 4 muestra una codificación directa, asociando con cada ω una palabra de código formada por una cadena de bits para la cual se necesita tres bits para símbolo y convirtiendo el mensaje en una cadena larga de bits.

ω_1	000
ω_2	001
ω_3	010
ω_4	011
ω_5	100

Tabla 4.

Supongamos que se sabe por experiencia las probabilidades de ocurrencia (O) de cada ω_i y estas corresponden a los valores dados en la tabla 5.

ω_1	0.25
ω_2	0.25
ω_3	0.2
ω_4	0.15
ω_5	0.15

Tabla 5.

Recurriendo a una codificación de longitud variable donde se asignan no a todos los ω's el mismo número de bits, podemos aprovechar este conocimiento apriori.

Usando codificación de longitud variable, hay que tener cuidado para poder saber en cuales lugares de la cadena de bits empieza un nuevo símbolo. Para tal fin, se puede usar un *código prefix* que en general usa menos bits que el método directo poniendo un símbolo de separación entre las diferentes partes.

Un código prefix se caracteriza por el hecho que ninguna palabra del código coincide con los primeros tantos carácteres de otra, es decir no es un prefix. Todos los códigos de longitud fija son de tipo prefix. Aparte, si codificamos ω_5 por 1 en la Tabla 4, se conduce de nuevo a un código prefix. Un contraejemplo es codificar ω_1 por 0; recibiendo un 0 no se puede decidir si la intención fue transmitir ω_1 o si es el primer bit de alguna otra ω.

Considera el código prefix de la Tabla 6.

ω_1	00
ω_2	10
ω_3	11
ω_4	010
ω_5	011

Tabla 6.

Para este código, la longitud promedio, EL, es

$$EL = 2 \cdot P(L = 2) + 3 \cdot P(L = 3) = 2 \cdot (0.7) + 3 \cdot (0.3) = 2.3 < 3.$$

La codificación de la Tabla 6 se obtuvo usando *el método de Huffman*: consiste en construir el código para cada ω_i de atrás hacia adelante:

define cada ω_i como un estado;
repite hasta que haya solamente un estado:
• busca dos estados con menor probabilidad;
• asígnales respectivamente 0 y 1;
• agrupa los dos estados obtenidos en uno solo con probabilidad la suma de los dos estados;

Si llamamos $L(C)$ la longitud esperado de código C, y $H(O)$ la entropía de las probabilidades de ocurrencia de los elementos del alfabeto dado, Shannon mostró que para cualquier códificación invertible:

$$H(O) \leq L(C).$$

Se puede mostrar que para el código de Huffman se tiene que

$$H(O) \leq L(C) \leq H(O) + 1.$$

Así si la entropía no es cercana a 1, el código de Huffman es bastante eficiente (la entropía de las probabilidades de la Tabla 5 es 2.2855).

3.9.2 Quicksort

Un algoritmo muy usado para ordenar un conjunto de n números es *Quicksort*. Por razones de simplificación, a continuación supongamos que todos los números son diferentes.

Quicksort es un método recursivo con la siguiente estructura:

1. Si $n = 1$: regresa el número
2. Si $n > 1$: elige un número s del arreglo al azar (puede ser el primero). Divide el arreglo en dos subconjuntos: $A = \{x : x < s\}$ y $B = \{x : x > s\}$. Aplica el algoritmo para ordenar A y B. Llamamos los resultados ord(A) y ord(B). Finalmente regresa $(\text{ord}(A), s, \text{ord}(B))$.

Supongamos que estamos interesados en la complejidad promedio de este algoritmo (en función de n), es decir en EY_n con Y_n el número de operaciones simples. Condicionando en el rango del elemento que elegimos al azar y usando la propiedad 3.6.3, obtenemos:

$$EY_n = \sum_{i=1}^{n} E(Y_n|\text{s tiene rango i})P(\text{s tiene rango i})$$

Por la construcción,

$$E(Y_n|\text{s tiene rango i}) = (n-1) + EY_{i-1} + EY_{n-i}$$

y por el hecho que elegimos s al azar,

$$P(\text{s tiene rango i}) = \frac{1}{n}$$

Si definimos $a_n = EY_n$ obtenemos la siguiente recursión que se puede resolver:

$$a_n = \sum_i \frac{(n-1) + a_{i-1} + a_{n-i}}{n}.$$

Un poco de historia ...

Primera ocurrencia en la literatura de algunos términos que vimos en este capítulo (según H.A. David, Am. Stat., 1995, 49, 2).

Distribution function: von Mises (1919)
Cumulative distribution: Wilks (1943)
Geometric distribution: Feller (1950)
Random Variable, Cramer (1937)
Variance: Fisher (1918)
Moment: Pearson (1893)

Capítulo 4

Variables Aleatorias Continuas

Igual al caso discreto, una variable aleatoria continua es en primer lugar una función que mapea el resultado de un experimento a una característica de interés.

Además de eso, queremos *asociar probabilidades de ocurencia* a esta característica. Como suponemos en este cápitulo que la característica toma valores en una escala continua (típicamente un número real), ya no podemos simplemente enumerar todos los posibles valores que la característica tome y asociar con cada uno una probabilidad.

4.1 Definición

Discutimos en la sección 2.1.2 el experimento de elegir un punto al azar en el intervalo $[a, b]$. Para definir probabilidades sobre el resultado de este experimento, tuvimos que recurrir a un integral; si llamamos X el punto elegido al azar en el intervalo $[a, b]$, obtuvimos:

$$P(X \in A) = \int_A \frac{1}{b-a} dx. \tag{4.1}$$

Para el caso particular cuando A es de la forma $[a, x]$, lo anterior se traduce en:

$$P(X \leq x) = \int_a^x \frac{1}{b-a} dz = \frac{x-a}{b-a} \quad x \in [a, b]. \tag{4.2}$$

Figura 1 ilustra (4.1) gráficamente: la probabilidad de obtener un valor que pertenezca a A es igual al área de este intervalo bajo la curva $f_X(x) = 1/(b-a)$.

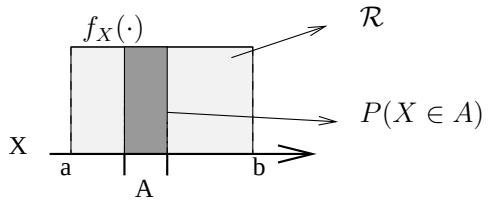


Figura 1.

Lo anterior sirve como punto de partida para el caso general donde queremos poder expresar que valores en algunas partes del intervalo son más probables que en otras. Eso significa que $f_X(\cdot)$ y $P(X \leq x)$ puedan ser funciones más general.

Definición 4.1.1 *Sea Ω un espacio sobre la cual está definido una función de probabilidad, la función X de Ω a $\mathcal{R}^n$ es una variable aleatoria continua si existe una función $f_X(\cdot)$ tal que:*

$$P(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(z)dz. \tag{4.3}$$

Llamamos $f_X(\cdot)$ la función de densidad de X y $F_X(\cdot)$ su función acumulativa de distribución.

Se puede mostrar que para una clase de conjuntos A suficientemente amplia, ecuación (4.3) es equivalente a

$$P(X \in A) = \int_A f_X(x)dx, \tag{4.4}$$

es decir, definir la distribución acumulativa es suficiente para determinar $P(\cdot)$.

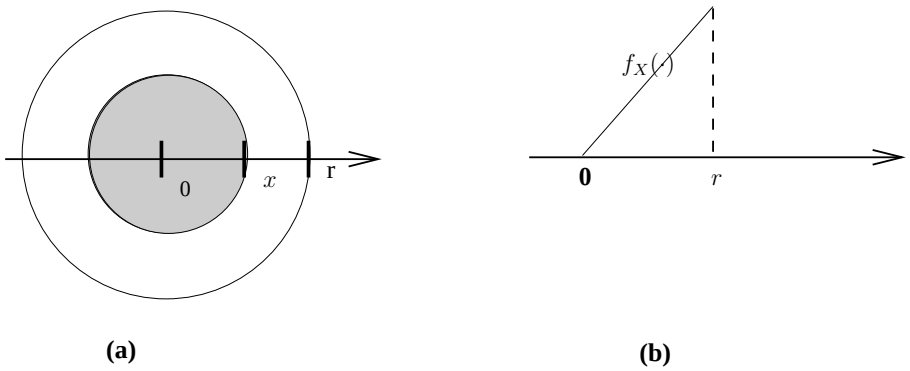
Tomando como $A = [x, x + \epsilon]$ un intervalo muy pequeño, obtenemos que $P(X \in [x, x+\epsilon])$ es aproximandamente igual a $\epsilon * f_X(x)$. Usando las propiedades de un integral, se verifica que $P(\cdot)$ satisfaga las propiedades de cualquier función de probabilidad.

Ejemplo 4.1.1 Probablemente el ejemplo más conocido para ilustrar el concepto de densidad es cuando uno lanza al azar una flecha a un disco con un radio r y el castigo que uno recibe es igual a la distancia entre la flecha y el centro del disco. Llama X el castigo: es una función que mapea una posición en el disco al monto del castigo.

De esta manera, $P(X \leq x)$ corresponde a obtener un resultado dentro del círculo con radio x como mostrado en la figura 2 (a). Como supongamos que se lanza al azar,

$$P(X \leq x) = \frac{\text{superficie del disco con radio x}}{\text{superficie del disco con radio r}} = \frac{\pi * x^2}{\pi * r^2} = x^2/r^2 = \int_0^x 2y/r^2 dy.$$

Así la densidad de X es $f_X(x) = 2x/r^2$ como graficada en la figura 2 (b).



(a) **(b)**

Figura 2.

No es dificil ver que como $P(X \in A)$ es siempre positivo y $P(X \in \mathcal{R}^n) = 1$, la densidad debe cumplir con:

1. $f_X(x) \geq 0$
2. $\int f_X(x)dx = 1$

Se observa que, contrario al caso discreto, no es necesario que $f_X(x) \leq 1$.

Una variante del ejemplo anterior es tomar como objecto la región $\mathcal{R}$ acotada por la curva de una función y el eje horizontal con área total a 1. Elegimos al azar un punto en $\mathcal{R}$ y llamamos X su primera coordenada. Usando (4.4), no es dificil ver que la densidad de X debe ser la curva.

Lo anterior nos ofrece un mecanismo para interpretar $f_X(\cdot)$ a través de un experimento explícito.

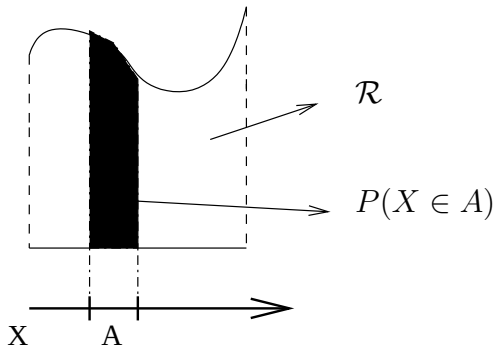


Figura 3.

Igual al caso discreto, se puede definir probabilidades condicionales como muestra el siguiente ejemplo.

Ejemplo 4.1.2 Llama X el tiempo de vida de un aparato. De esta manera, $F_X(x) = P(X \leq x)$ representa la probabilidad que el aparato funcionará menos de x tiempo y $P(X > x)$ representa la probabilidad que el aparato funcionará más de x tiempo.

Podemos calcular eventos condicionales: por ejemplo para $x > x_0$,

$$P(X \leq x | X > x_0) = \frac{P(x_0 < X \leq x)}{P(X > x_0)} = \frac{\int_{x_o}^{x} f_X(y)dy}{P(X > x_0)}$$

es la probabilidad que el componente trabaja menos de x tiempo dado que no falló hasta momento x_0.

Podemos considerar $P(X \leq x | X > x_0)$ como una nueva función acumulativa, y tomando la derivada con respeto a x, obtenemos su densidad:

$$f_{X|X>x_0}(x) = \frac{f_X(x)}{P(X > x_0)}, \quad x > x_0. \tag{4.5}$$

En confiabilidad una característica muy importante es el factor de riesgo $R(x)$:

$$R(x) = f_{X|X>x}(x) = \frac{f_X(x)}{P(X > x)}.$$

Para una ϵ chiquita, $R(x)\epsilon$ es la probabilidad de que el componente falla entre x y $x+\epsilon$ dado que en el momento x aún estaba funcionando.

Como se muestra en la Figura 4, típicamente se pueden distinguir tres fases en el comportamiento de $R(x)$. En los primeros momentos de vida de un aparato predominan fallas por errores en la producción: conforme el tiempo transcurre, la probabilidad de su occurencia baja. En la última parte de la vida útil, predominan errores por desgaste: conforme el tiempo transcurre, la probabilidad de su occurencia sube. Entre estos dos extremos $R(x)$ es relativamente constante: la probabilidad de fallar no dependerá mucho del tiempo de vida que tiene.

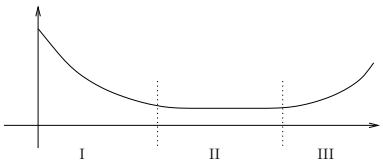


Figura 4.

Ejemplo 4.1.3 Supongamos que X tiene densidad $f_X(x) = \lambda \exp(-\lambda x), x \geq 0$ (lo que se conoce como la distribución exponential). Se obtiene que:

$$F_X(x) = \int_0^x \lambda \exp(-\lambda y) dy = 1 - \exp(-\lambda x)$$

y

$$R(x) = \frac{\lambda \exp(-\lambda x)}{\exp(-\lambda x)} = \lambda.$$

Lo anterior significa que si un componente tiene un tiempo de vida que sigue una distribución exponential, el tiempo que ya transcurrió funcionando bien, no cambia la probabilidad de fallar en los siguientes momentos.

En lo anterior condicionamos en un evento con probabilidad positiva. Para variables continuas X_1, X_2, se puede extender lo anterior condicionando en eventos de la forma $X_2 = x_2$. Tomando en

$$P(X_1 \in A | X_2 \in [x, x+\epsilon]) = \frac{\int_A \int_x^{x+\epsilon} f_{X1,X2}(x_1, x_2) dx_2 dx_1}{P(X_2 \in [x, x+\epsilon])}$$

el límite a 0, se llega a la siguiente definición.

Definición 4.1.2 *Para las variables aleatorias X_1, X_2, definimos la distribución condicional de X_1 dado $X_2 = y$ como:*

$$f_{X_1|X_2=x_2}(x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)}, \tag{4.6}$$

suponiendo que $f_{X_2}(x_2) > 0$.

Finalemente, formulamos la contraparte de la *Ley de la probabilidad total* para variables continuas.

Si X_1, X_2 son variables aleatorias continuas, entonces:

$$P(X_1 \in A) = \int_A \int_{x_2} f_{X_1|X_2=x_2}(x_1) f_{X_2}(x_2) dx_2 dx_1.$$

Definición 4.1.3 *Si X_1 y X_2 son variables aleatorias continuas sobre un mismo espacio de probabilidad, las dos son independientes ssi*

$$f_{X_1,X_2}(x,y) = f_{X_1}(x) f_{X_2}(y) \quad \textit{para cada } x, y \in \mathcal{R}^n \tag{4.7}$$

En analogía con el caso de variables discretas, se definen el promedio y la varianza.

Definición 4.1.4 *Sea X una variable aleatoria continua con densidad $f_X(x)$, el promedio de X, EX es*

$$EX = \int x f_X(x) dx.$$

La varianza $Var(X)$ es:

$$Var(X) = \int (x - EX)^2 f_X(x) dx.$$

suponiendo que los integrales existan.

Como antes, el promedio es un operador lineal, y si las variables son independientes, la varianza es aditiva.

Ejemplo 4.1.4 Sea X un punto elegido al azar de $[a, b]$,

$$EX = \int_a^b x f_X(x)\, dx \int_a^b \frac{x}{b-a}\, dx = \frac{a+b}{2}.$$

Ejemplo 4.1.5 Sea X una variable aleatoria con densidad $f_X(x) = \lambda \exp(-\lambda x), x \geq 0$. Entonces,

$$EX = \int_0^\infty x\lambda \exp(-\lambda x)\, dx = -\int_0^\infty x\, d\exp(-\lambda x) = -x\exp(-\lambda x)|_0^\infty + \int_0^\infty \exp(-\lambda x)\, dx = \frac{1}{\lambda}.$$

4.2 Transformaciones

4.2.1 Transformaciones monótomas

Contrario al caso discreto, en el caso de variables continuas ya no se puede usar la aditividad de la función de probabilidad para encontrar la distribución de una transformación de una variable aleatoria, porque la función de densidad no es propiamente una probabilidad. Dado que la función acumulativa sí es una probabilidad, una manera para resolver el problema, es buscar la función acumulativa de la variable transformada.

Por ejemplo, si X es la altura de una persona en centímetros, y si definimos $Y = X/100$ la altura en metros, entonces

$$F_Y(y) = P(Y \leq y) = P(X/100 \leq y) = P(X \leq (100y)) = F_X(100y).$$

Aplicando la regla de la cadena para la derivada, obtenemos:

$$f_Y(y) = \frac{dF_Y(y)}{y} = \frac{dF_X(100y)}{y} = f_X(100y).100$$

En general si $Y = g(X)$, no es dificil mostrar que para cualquier función $g(\cdot)$, invertible, creciente y derivable:

$$f_{g(X)}(y) = f_X(g^{-1}(y)).\frac{d(g^{-1}(y))}{dy}. \tag{4.8}$$

En el caso que $g(\cdot)$ es decreciente, invertible y derivable, se obtiene:

$$f_{g(X)}(y) = -f_X(g^{-1}(y)).\frac{d(g^{-1}(y))}{dy}.$$

En general si $g(\cdot)$ es invertible y derivable, se obtiene:

$$f_{g(X)}(y) = f_X(g^{-1}(y))|\frac{d(g^{-1}(y))}{dy}|.$$

4.2.2 Mínima y Máxima

Empezamos con un ejemplo.

Ejemplo 4.2.1 Considera un sistema (de producción) compuesto de dos componentes que están en paralelo (Figura 5 (a)) o en serie (Figura 5 (b)). Supongamos que cada componente i tiene un tiempo de vida X_i, independientes y de la misma distribución.

Para el caso del sistema paralelo, supongamos que el sistema funciona correctamente mientras hay al menos un componente trabajando bien. Si no consideramos reparaciones, el tiempo de vida del sistema está dada por

$$Y = \max(X_1, X_2).$$

Para el caso del sistema serial, supongamos que el sistema funciona correctamente si ambos componentes trabajan bien. De nuevo, si no consideramos reparaciones, el tiempo de vida del sistema está dada por

$$Y = \min(X_1, X_2).$$

El ejemplo anterior muestra la utilidad de poder calcular la distribuciones del máximo o mínimo de variables aleatorias independientes.

Dado que

$$\min(x_1, \cdots, x_n) > x \text{ ssi cada } x_i > x$$

obtenemos que

$$P(\min(X_1, \cdots, X_n) \leq x) = 1 - P(\min(X_1, \cdots, X_n) > x) = 1 - P(X_1 > x, \cdots, X_n > x).$$

Usando la independencia, obtenemos:

$$P(\min(X_1, \cdots, X_n) \leq x) = 1 - (1 - F_X(x))^n.$$

De esta manera la densidad de $min(X_1, \cdots, X_n)$ es igual a

$$n(1 - F_X(x))^{n-1} f_X(x).$$

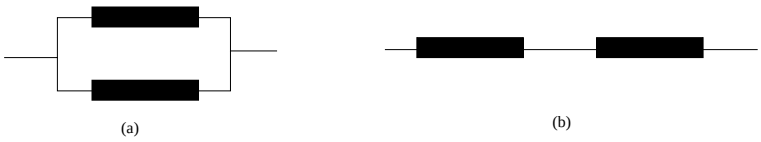


Figura 5.

Dejamos la derivación de la distribución de $\max(X_1, \cdots, X_n)$ como ejercicio.

4.2.3 Sumas de variables independientes

Como veremos más adelante, sumar variables aleatorias independientes es una transformación muy importante. A continuación determinamos la distribución de $Y = X + Z$ donde X y Z son variables independientes. Para calcular la densidad de Y, calculamos primero la función acumulativa.

$$P(Y \leq y) = P(X + Z \leq y).$$

Usando la ley de la probabilidad total, obtenemos:

$$P(X + Z \leq y) = \int P(X + Z \leq y | Z = z) f_Z(z) dz = \int P(X \leq y - z | Z = z) f_Z(z) dz.$$

Usando la independencia entre X y Z:

$$P(X+Z \leq y) = \int P(X \leq y-z) f_Z(z) dz = \int F_X(y-z) f_Z(z) dz.$$

Tomando la derivada con respecto a y, se obtiene:

$$f_{X+Z}(y) = \int f_X(y-z) f_Z(z) dz. \tag{4.9}$$

La integral se conoce como la *convolución* de $f_X(\cdot)$ con $f_Z(\cdot)$.

Ejemplo 4.2.2 Sean X y Z variables aleatorias independientes con distribución $\exp(\lambda)$, entonces:

$$f_{X+Z}(y) = \int_0^y \lambda \exp(-\lambda(y-z)) \lambda \exp(-\lambda z) dz = \int_0^y \lambda^2 \exp(-\lambda y) \lambda dz = \lambda^2 y \exp(-\lambda y).$$

4.3 La distribución normal

En esta sección definimos una familia de distribuciones que jugará un papel importante en el estudio de la distribución de sumas de variables aleatorias.

Definición 4.3.1 *La variable aleatoria X tiene una distribución normal estandar, $\mathcal{N}(0,1)$ si la densidad tiene la siguiente forma:*

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}), x \in \mathcal{R}. \tag{4.10}$$

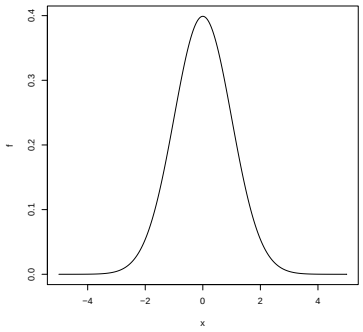


Figura 6.

Se observa que la densidad es simétrica alrededor de 0, así,

$$EX = \int_{-\infty}^{\infty} x f_X(x) dx = \int_{-\infty}^{0} x f_X(x) dx + \int_0^{\infty} x f_X(x) dx = \int_{-\infty}^{0} x f_X(x) dx + \int_{-\infty}^{0} (-x) f_X(-x) dx =$$

$$\int_{-\infty}^{0} x f_X(x)dx - \int_{-\infty}^{0} x f_X(x)dx = 0.$$

De la misma manera, se puede mostrar que $EX^k = 0$ si k es impar. Por otro lado, $Var(X) = 1$.

A partir de la distribución normal estandar, se define una familia de distribuciones, a través de la transformación

$$Y = \mu + \sigma X, X \sim \mathcal{N}(0, 1),$$

donde $\sigma > 0$ y $\mu \in \mathcal{R}$. Se escribe $Y \sim \mathcal{N}(\mu, \sigma^2)$.

Usando (4.8), dado que $\sigma > 0$, se obtiene:

$$f_Y(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{-(y-\mu)^2}{2\sigma^2}), \quad \mu \in \mathcal{R}, \sigma > 0, x \in \mathcal{R}.$$

Algunas de las propiedades más importantes de esta distribución, son:

1. Si $X \sim \mathcal{N}(\mu, \sigma^2)$, se tiene que $(X - \mu)/\sigma \sim \mathcal{N}(0, 1)$
2. Si $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ y son independientes, entonces
$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$
3. Si $X \sim \mathcal{N}(0, \sigma^2)$, entonces $-X \sim \mathcal{N}(0, \sigma^2)$

4.3.1 El Teorema del Límite Central

Para la variable X con una densidad graficada en la Figura 7 (a), se muestra en figura 19 (b) la densidad de la suma de dos variables independientes y de la misma distribución que X y, de la misma manera, en la Figura 7 (c) la densidad de la suma de tres variables. Se aplicó el mismo procedimiento para una distribución discreta en la Figura 8.

A pesar de empezar con dos distribuciones totalmente diferente, se observa que en ambos casos la distribución de 3 variables independientes tiende a la forma de una campana. El *Teorema Central del Límite* da el soporte teórico para este fenómeno.

Propiedad 4.3.1 *Supongamos que $X_1, X_2, \cdots$ son variables independientes y de la misma distribución con un promedio μ y varianza σ^2, ambos finitos, entonces*

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq y\right) = \int_{\infty}^{y} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) dx, \tag{4.11}$$

donde $S_n = X_1 + \cdots + X_n$.

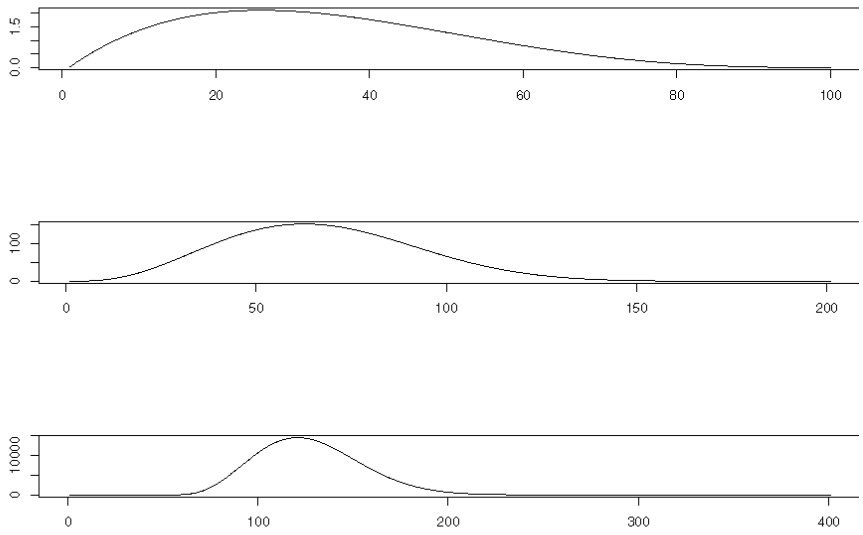


Figura 7.

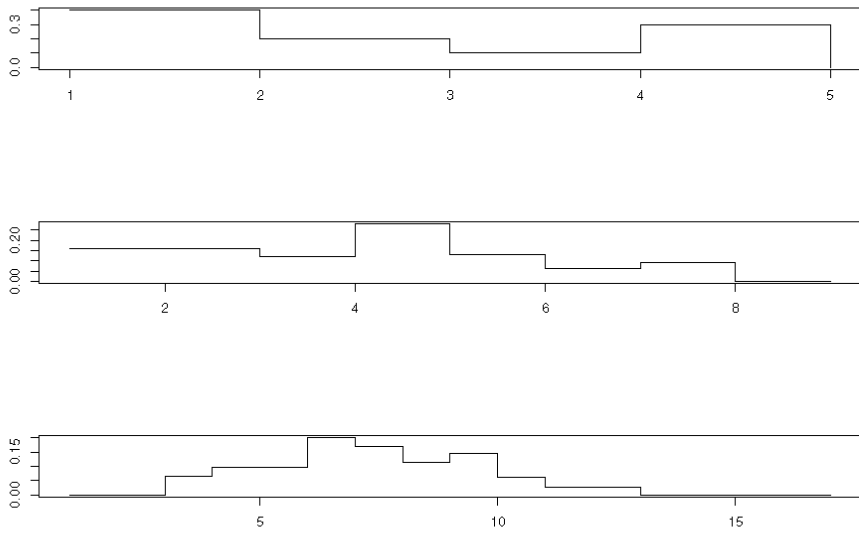


Figura 8.

Una consecuencia de (4.11) es que para n suficientemente grande,

$$S_n/n \sim \mathcal{N}(\mu, \sigma^2/n), \tag{4.12}$$

o equivalente

$$S_n \sim \mathcal{N}(n\mu, n\sigma^2).$$

La relación (4.12) implica que si $n \to \infty$, la varianza tiende a 0. Una distribución con varianza cero significa que toma siempre el mismo valor (que a su vez es también el promedio). Así se obtiene que S_n/n converge a $\mu = EX$. Lo anterior suponía que $var(X) < \infty$. A continuación damos la ley (debil) de los números grandes que no impone una restricción sobre $var(X)$.

Propiedad 4.3.2 *Sea X una variable aleatoria tal que $EX = \mu < \infty$. Si las variables $X_1, X_2, \cdots$ son independientes y de la misma distribución que X entonces para cada $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|\frac{X_1 + \cdots + X_n}{n} - EX| < \epsilon) = 1$$

En otras palabras, esta propiedad nos dice que para n valores $\{x_1, \cdots, x_n\}$ obtenidos como resultados de un experimento con la misma distribución que X, su promedio aritmético converge a EX si n crece a infinito.

En caso que $X \sim Bern(\theta)$, $EX = P(X = 1)$, entonces se ve que la probabilidad de un evento surge como el limite de la frecuencia relativa de ocurrencia de este evento.

4.3.2 Ejemplos de distribuciones continuas

1. Distribución Normal:

$$X \sim \mathcal{N}(\mu, \sigma^2) \Leftrightarrow f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{-(x-\mu)^2}{2\sigma^2}), \quad \mu \in \mathcal{R}, \sigma > 0, x \in \mathcal{R}.$$

   Por su gran importancia, a la distribución que aparece en (4.10) se le llama normal estándar y corresponde a la distribución normal con parámetros $\mu = 0$ y $\sigma^2 = 1$.

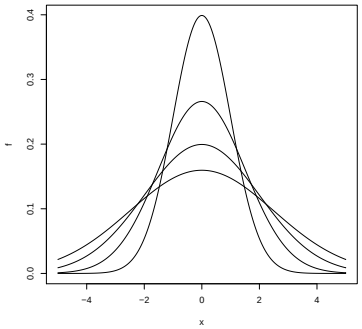


Figura 9.
La densidad de una distribución normal para diferentes valores de σ.

Algunas de las propiedades más importantes de esta distibución, son:

(a) Si $X \sim \mathcal{N}(\mu, \sigma^2)$, se tiene que $(X - \mu)/\sigma \sim \mathcal{N}(0, 1)$

(b) Si $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ y son independientes, entonces

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

(c) Si $X \sim \mathcal{N}(0, \sigma^2)$, entonces $-X \sim \mathcal{N}(0, \sigma^2)$

2. Distribución Exponencial:

$$X \sim \text{Exp}(\lambda) \Leftrightarrow f_X(x) = f_X(x) = \lambda \exp(-\lambda x), \quad \lambda > 0, x > 0.$$

Esta distribución forma la contraparte continua de la distribución geométrica. Su promedio es $\frac{1}{\lambda}$ y varianza $\frac{1}{\lambda^2}$.

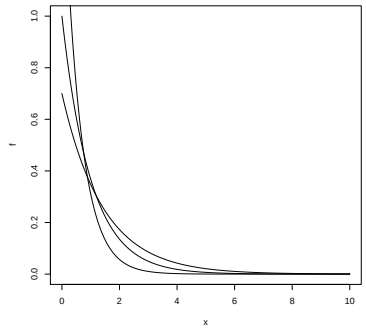


Figura 10.
La densidad de una distribución exponential para diferentes valores de λ.

Es la única distribución continua que tiene la propiedad de *p*érdida de memoria, que consiste en $P(X > a + b|X > a) = P(X > b)$.

3. Distribución Cauchy:

$$X \sim \text{Cauchy}(\alpha, \beta) \Leftrightarrow f_X(x) = \frac{1}{\pi(\beta^2 + (x - \alpha)^2)}, \quad \alpha \in \mathcal{R}, \beta > 0, x \in \mathcal{R}.$$

Al igual que la distribución Normal, la Cauchy es simétrica, pero ésta no tiene promedio ni varianza finita.

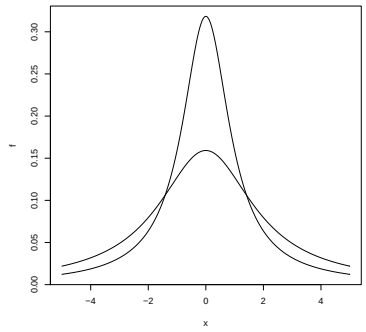


Figura 10.
La densidad de una distribución Cauchy para diferentes valores de β.

4. Distribución χ-cuadrada:

$$X \sim \chi^2_k \Leftrightarrow f_X(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} \exp(-x/2), \quad k \in \{1, 2, \ldots\}, x > 0.$$

Con $\Gamma(\cdot)$ se refiere a la función Gamma. En especial, al parámetro de la distribución χ^2_n, n, se le conoce como grados de libertad.

A partir de la suma de n variables estandar normal, independientes y elevadas al cuadrado, se obtiene la distribución χ-cuadrada con n grados de libertad.

El promedio de una distribución χ^2_n es n y la varianza es $2n$.

5. Distribución Gamma:

$$X \sim \text{Gamma}(\lambda, \alpha) \Leftrightarrow f_X(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(n)}, \quad \lambda, \alpha, x > 0.$$

La distribución Gamma generaliza la distribucion exponencial y la χ-cuadrada en el sentido que una Gamma con parámetros $(1, \lambda)$ es una Exponencial con parámetro λ y una Gamma con parámetros $(1/2, n/2)$ es una χ^2 con n grados de libertad, cuando n es un entero.

El promedio de una distribución Gamma(λ, α) es $\frac{\alpha}{\lambda}$ y la varianza es $\frac{\alpha}{\lambda^2}$.

6. Distribución t de Student:

$$X \sim t_k \Leftrightarrow f_X(x) = \frac{\Gamma[(k+1)/2]}{\Gamma(k/2)} \frac{1}{\sqrt{k\pi}(1 + x^2/k)^{(k+1)/2}}, \quad k \in \{1, 2, \ldots\}, x > 0.$$

La importancia de esta distribución es práctica y corresponde a una variable aleatoria originada por la operación de otras dos: la normal estándar y la χ cuadrada. Si $Z \sim \mathcal{N}(0, 1)$ y $U \sim \chi^2_k$ son independientes, entonces

$$X = \frac{Z}{\sqrt{U/k}} \quad \text{tiene distribución } t \text{ de Student con } k \text{ grados de libertad.}$$

El promedio de una distribución t_k es 0 y la varianza es $\frac{n}{n-2}$ si $n > 2$.

7. Distribución F:

$$X \sim \text{F}(a, b) \Leftrightarrow f_X(x) = \frac{\Gamma[(a+b)/2]}{\Gamma(a/2)\Gamma(b/2)} \left(\frac{a}{b}\right)^{a/2} \frac{x^{(a-2)/2}}{[1 + (a/b)x]^{(a+b)/2}}, \quad a, b \in \{1, 2, \ldots\}, x > 0.$$

La importancia de la distribución F, al igual que la t, radica en su utilidad práctica y también surge como la densidad de la operacion de dos variables aleatorias con

distribución ya conocida. Si U y V son variables aleatorias independientes y con distribución χ cuadrada, con a y b grados de libertad, respectivamente, entonces

$$X = \frac{U/a}{V/b} \quad \text{tiene distribución } \mathrm{F}(a, b).$$

El promedio de una distribución $\mathrm{F}(a, b)$ es $\frac{b}{b-2}$, si $b > 2$.

Capítulo 5

Dependencia, Independencia y Distribuciones Multivariadas

En la práctica es indispensable poder estudiar diferentes características o variables aleatorias al mismo tiempo y no solamente una a la vez como hemos hecho principalmente hasta ahora. En este capítulo nos enfocamos a diferentes métodos para construir distribuciones multivariadas y a los nuevos problemas que eso implica. Como introdución, empezamos con el caso bidimensional y hacemos la distincción entre dependencia y causalidad.

5.1 Introducción

5.1.1 Interacción para el caso de dos variables binarias

Supongamos que X, Y sean dos variables aleatorias binarias. Denotamos con p_{ij} la probabilidad que $X = i$ y $Y = j$ y suponemos que no haya combinaciones con probabilidad 0.

En caso que X, Y son independientes sabemos que

$$P(X = 1|Y = 1) = P(X = 1|Y = 0) \quad \text{y} \quad P(X = 0|Y = 1) = P(X = 0|Y = 0)$$

y entonces

$$\frac{P(X = 1|Y = 1)}{P(X = 0|Y = 1)} = \frac{P(X = 1|Y = 0)}{P(X = 0|Y = 0)}. \tag{5.1}$$

El *oddsratio* expresa la interacción a través de la desviación de la igualdad (5.1):

$$R = \frac{\frac{P(X=1|Y=1)}{P(X=0|Y=1)}}{\frac{P(X=1|Y=0)}{P(X=0|Y=0)}}. \tag{5.2}$$

No es dificil ver que

$$R = \frac{p_{00}p_{11}}{p_{01}p_{10}} \tag{5.3}$$

Si las variables son independientes, $R = 1$. En general $R \in (0, \infty)$. Mientras más diferente a 1, más dependencia hay entre X y Y.

Dos propiedades son:

1. R es invariante al intercambiar X y Y;
2. Si cambiamos la codificación de una variable (1 en lugar de 0 y viceversa), el nuevo oddsratio=$1/R$ con R el oddsratio de la codificación original.

Muchas veces, se prefiere trabajar con $\theta = \log R$, el log-oddsratio. Independencia implica ahora que $\theta = 0$. Asi, el cambiar la codificación de una variable, implica solamente un cambio de signo.

Un siguiente paso consiste en normalizar el (log-)oddsratio para que tome valores en un intervalo acotado. Dos ejemplos son

$$Q = \frac{R-1}{R+1} \quad Y = \frac{\sqrt{R}-1}{\sqrt{R}+1}.$$

Lo interesante y **excepcional** del caso de 2 variables binarias, es que se puede resumir la distribucion conjunta por medio de 3 parámetros:

$$P(X = 1), P(Y = 1) \text{ y } R \text{ (o alguna transformacion de } R),$$

es decir, se puede separar la interacción entre las dos variables de las características marginales.

Covarianza

Como vimos, el oddsratio está basado en tomar la fracción de dos números para expresar hasta que punto estos difieren entre sí. Otra manera es tomar su diferencia. Es decir convertir (5.3) en

$$p_{11}p_{00} - p_{10}p_{01}. \tag{5.4}$$

Dejamos como ejercicio mostrar que lo anterior es igual a

$$E((X - EX)(Y - EY)) = E(XY) - EX \cdot EY,$$

lo que definimos como la *covarianza* entre X y Y

Para no depender de la escala de X y Y, se introduce la *correlación*:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \tag{5.5}$$

Esta estandarización asegura que la correlación tendrá sus valores dentro del intervalo $[-1, 1]$.

Contrario al caso del oddsratio, esta definición permite trabajar con variables que toman más de dos valores. Por otro lado, la covarianza/correlación ya no es capaz de resumir en un solo número toda la interacción entre dos variables. En Figura 1 vemos una muestra de dos distribuciones diferentes para (X, Y). La correlación es la misma en ambas distribuciones pero la estructura de interacción es claramente diferente.

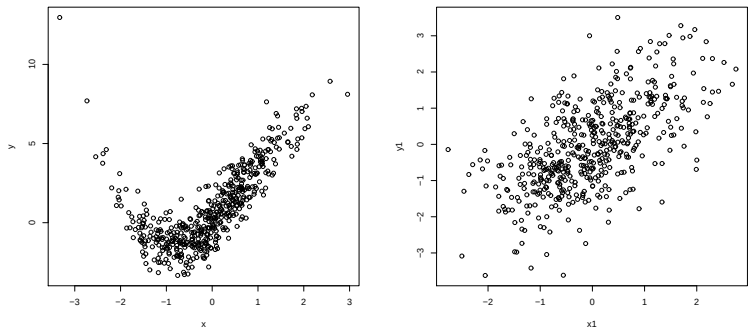


Figura 1.

Como ilustración de este concepto consideremos dos variables independientes X y Z, ambas con promedio 0 y varianza 1. Definimos la variable Y como

$$Y = aX + Z.$$

La covarianza entre X y Y es el coeficiente a. Más adelante veremos que en general la covarianza determina el coeficiente de regresión de una variable contra la otra. Obsérvese que la covarianza es cero si las dos variables son independientes y que la covarianza de una variable con su misma es la varianza.

Información mutua

Definamos primero la entropía condicional entre dos variables

Definición 5.1.1 *Sean X, Y dos variables aleatorias, la entropía condicional de Y dado X es igual a:*

$$H_X(Y) = EH(Y|X = x) = -\sum_x (\sum_y P(Y = y|X = x) \log P(Y = y|X = x)) P(x = x) \tag{5.6}$$

En base de lo anterior definimos la *información mutua*.

Definición 5.1.2 *Sean X, Y dos variables aleatorias, la información mutua está dada por:*

$$I(X, Y) = H(Y) - H_X(Y).$$

Dejamos como ejercicio demostrar que $I(X, Y) = I(Y, X)$. Podemos interpretar $I(X, Y)$ como la disminuición de incertidumbre en X por conocer Y y vice versa.

5.1.2 El caso multivariado

Aunque técnicamente podemos extender la covarianza a tres variables a través de

$$E(X - EX)(Y - EY)(Z - EZ),$$

ya no se tiene una interpretación tan clara. Además se pierden muchas propiedades. Por ejemplo la covarianza es invariante a los promedios de las variables; pero $E(X - EX)(Y - EY)(Z - EZ)$ ya no es invariante a la covarianza.

Otro enfoque consiste en reducir el caso n dimensional a muchos casos bidimensional. Si denotamos con $X_i \perp X_k$ independencia entre X_i y X_k, podemos distinguir dos caminos resumidos en la siguiente tabla.

abstracción	versus	*especificación*
descripción marginal	versus	*descripción conditional*
$X_i \perp X_k$		$X_i \perp X_k \mid X_{-i,-k}$
$P(X_i)$		$P(X_i \mid X_{-i})$
$Cov(X_i, X_j)$		$Cov(X_i, X_j \mid X_{-i,-j})$

Desafortunadamente no es evidente como reconstruir através de características marginales o condicionales la distribución/interacción conjunta. Más adelante veremos como usar grafos para ese fin.

5.1.3 Dependencia y causalidad

En particular en situaciones con un componente de incertidumbre, es importante distinguir los conceptos de dependencia y causalidad. Si X_1 denota si una persona elegida al azar fuma o no y X_2 denota si tiene (o ha tenido) cancer, es muy probable que habrá una fuerte dependencia entre ambas pero no es - en sí - un argumento suficiente para deducir que fumar genera cancer. Por ejemplo, puede ser que una característica X_3 no observada causa que con alta probabilidad uno fuma y que desarrolla cancer. Lo anterior es una consequencia de que un modelo basado en $P(X_1, X_2)$ no excluye la existencia de otras caracteristicas X_i o - más formalmente dicho - las probabilidades formulan regularidades de una manera abstracta sobre una familia de situaciones *similares* y donde con *similar* referimos al hecho que definiendo $P(X_1, X_2)$ no distinguimos explicitamente entre $P(X_1, X_2, X_3 = 0)$ y $P(X_1, X_2, X_3 = 1)$. A continuación damos otro ejemplo de este fenómeno.

Tomamos como ejemplo la población de la Figura 2. Una version "blanco-negro'' está representada en la Figura 3.

La población Ω de la Figura 2 consiste de objetos de diferentes formas (círculos o rectángulos), cada uno con o sin una marca (un '1' o no) y dos posibles colores (verde o rojo).

Supongamos que tenemos interés en saber si la presencia o ausencia de un círculo (C), implica cierta información acerca de la presencia de una marca (M), es decir la relación $C \to M$ y (no C) $\to M$. Se pueden transponer muchas situaciones de la vida cotidiana a esta esquema; por ejemplo los objetos son personas, el color indica el sexo, la forma corresponde a tomar o no una medicina y la presencia o ausencia de una marca representa curarse de una enfermedad o no.

Construimos la distribución de conteo sobre Ω. En la tabla 1 (a) y (b) las probabilidades condicionales correspondientes están resumidas restringiendo Ω a los objetos rojos y verdes, respectivamente.

Para ambas poblaciones

$$P(M|C) > P(M|C^c). \tag{5.7}$$

Ahora supongamos que recibimos ambas poblaciones juntas y además en una imagen de blanco negro, lo que impide distinguir las dos subpoblaciones (ver figura 3).

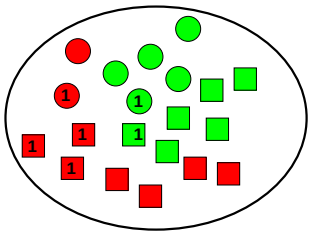


Figura 2.

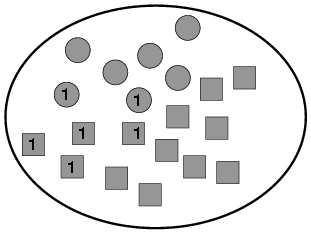


Figura 3.

Para esta situación, las probabilidades condicionales están resumidas en la última parte de la Tabla 1. Vemos que ahora

$$P(M|C) < P(M|C^c), \tag{5.8}$$

lo que es el contrario a la relación (5.7).

(a) Subpoblación obj. rojos:	
$P(M^c\|C) = 1/2 = 0.50$	$P(N\|C) = 1/2 = 0.50$
$P(M^c\|C^c) = 4/7 = 0.57$	$P(M\|C^c) = 3/7 = 0.43$
(b) Subpoblación obj. verdes:	
$P(M^c\|C) = 4/5 = 0.80$	$P(N\|C) = 1/5 = 0.20$
$P(M^c\|C^c) = 5/6 = 0.83$	$P(M^c\|C^c) = 1/6 = 0.17$
(c) Población completa:	
$P(M^c\|C) = 5/7 = 0.71$	$P(M\|C) = 2/7 = 0.29$
$P(M^c\|C^c) = 9/13 = 0.69$	$P(M\|C^c) = 4/13 = 0.31$

Tabla 1.

Las ecuaciones (5.7) y (5.8) parecen, a primera vista, una paradoja. Sin embargo como se definen sobre diferentes poblaciones (aunque relacionadas a los mismos objetos físicos) y como son relaciones no determinísticas, no se contradicen. De esta manera, vemos que por las diferencias en lo que se consideran como "situaciones similares'', en lo anterior con y sin tomar en cuenta el color, la relación entre las probabilidades condicionales puedan cambiar y en consecuencia, si las consideríamos como causalidades, las causalidades subyacentes también cambian.

Este fenómeno no aparece con reglas de lógica porque ahí son relaciones funcionales a nivel de individuos sin alguna abstracción como en el caso probabilístico; Es imposible que al mismo tiempo se tiene:

$$C \to M^C \quad , \quad (C \ \& \ \text{rojo}\ ) \to M \quad , \quad (C \ \& \ \text{verde}\ ) \to M$$

Lo anterior es conocido en la literatura como *la paradoja de Simpson*.

5.2 Distribuciones multivariadas a través de grafos: redes probabilísticas

5.2.1 Independencias

El punto de partida en la construcción de grafos para definir distribuciones multivariadas, es simplificar la estructura de interacción entre los componentes de X por la suposición de (ciertas) independencias.

Dada una distribución (positiva) $P(X_1, \cdots, X_n)$ y denotando con $X_A, A \subset \{1, \cdots, n\}$, el conjunto $\{X_i, i \in A\}$, decimos que

$$X_A \perp X_B | X_C \text{ ssi } X_A \text{ es independiente de } X_B \text{ dada } X_C$$

No es dificil mostrar las siguientes equivalencias (por simplificación, usamos la formulación para el caso de variables discretas).

$$X_A \perp X_B | X_C, \quad \text{ssi} \tag{5.9}$$

$$P(X_A, X_B|X_C) = P(X_A|X_C)P(X_B|X_C), \quad \text{ssi} \tag{5.10}$$

$$\exists f_1, f_2 : \quad P(X_A, X_B|X_C) = f_1(X_A, X_C)f_2(X_B, X_C), \quad \text{ssi} \tag{5.11}$$

$$\exists g_1, g_2 : \quad P(X_A, X_B, X_C) = g_1(X_A, X_C)g_2(X_B, X_C), \quad \text{ssi} \tag{5.12}$$

$$P(X_A|X_B, X_C) = P(X_A|X_C), \quad \text{ssi} \tag{5.13}$$

$$\exists h : \quad P(X_A|X_B, X_C) = h(X_A, X_C). \tag{5.14}$$

A continuación usaremos seguido la relación (usando la formulación para el caso discreto):

$$P(X_1, \cdots, X_n) = \prod_i P(X_i|X_1, \cdots, X_{i-1}). \tag{5.15}$$

5.2.2 Grafos en forma de cadenas

Empezamos con un ejemplo.

Ejemplo 5.2.1 Tenemos n personas. Persona 1 dice un mensaje a persona 2; a su vez persona 2 pasa un mensaje a 3 etc., por un malentendido y/o malas intenciones, el mensaje que alguién pasa puede ser diferente al que recibió. Supongamos que al pasar un mensaje solamente las dos personas involucradas lo pueden escuchar y consideramos que no hay acuerdos entre las personas.

Si llamamos X_i el mensaje que persona i recibe, la definición de $P(X_1, \cdots, X_n)$ directamente no es evidente. Sin embargo, por las suposiciones, se puede aceptar que:

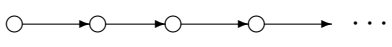
$$P(X_i|\{X_j, j < i\}) = P(X_i|X_{i-1}). \tag{5.16}$$

Es fácil de mostrar que usando (5.15), (5.16) implica que:

$$P(X_1, \cdots, X_n) = \prod_i P(X_i|X_{i-1}). \tag{5.17}$$

En consecuencia, lo que hay que determinar es $P(X_i|X_{i-1})$, es decir, convertimos un problema n-dimensional en $2 * n$ problemas unidimensionales (suponiendo que los X_i son binarios).

Se puede representar lo anterior a través de la siguiente gráfica



Si llamamos una variable X_j un *padre* de X_i cuando hay una flecha de X_j a X_i y denotamos con $pa(i)$ el (en general: los) padre(s) de nodo i, (5.16) es equivalente a:

$$P(X_i|\{X_j, j < i\}) = P(X_i|X_{pa(i)}). \tag{5.18}$$

Ciertas independencias implican otras independencias. Por ejemplo, dejamos como ejercicio mostrar que el modelo anterior implica:

$$X_i \perp X_{i-3}|X_{i-2}.$$

Una de las ventajas del uso de grafos es la facilidad de detectar todas las independencias implicadas.

Con ese fin definimos que dos conjuntos A, B de nodos son *separados* en una gráfica no dirigida, por un tercer conjunto C ssi no se puede caminar de un nodo de A a otro de B (y vice versa) sin pasar por un nodo de C.

Propiedad 5.2.1 *Dada una cadena que representa las independencias de una familia de distribuciones positivas a través de (5.18). Para conjuntos A, B, C de nodos se tienen:*

$$X_A \perp X_B|X_C \leftrightarrow C \text{ separa } A \text{ de } B \text{ en el grafo no dirigido correspondiente.}$$

Observa que si $P(X_i|X_{i-1})$ no depende de i, se llama $\{X_i\}$ una cadena de Markov. En la última sección discutimos este caso brevemente.

En lo siguiente, damos una generalización a gráficas arbitrarias (empezando con árboles).

5.2.3 Arboles de Markov

Retomamos primero ejemplo 6.2.1.

Ejemplo 5.2.2 (continuación) Supongamos que cada persona i también transmite de manera independiente su mensaje a un coordenador. Por errores en la transmisión lo que se recibe puede ser distincto a lo que se envió. Si Y_i denota lo que se recibe de mensaje X_i, lo anterior se traduce en:

$$P(Y_i|X_j, Y_k, j \leq i, k < i) = P(Y_i|X_i),$$

$$P(X_i|X_j, Y_k, j < i, k < i) = P(X_i|X_{i-1}).$$

Como veremos más adelante eso se traduce en un grafo de la forma (suponiendo que hay 4 personas):

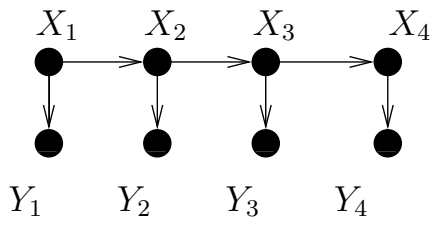


Figura 4..

Empecemos con una gráfica dirigida que forma un árbol (i.e. en cada nodo diferente de la raiz llega una sola flecha). Numeremos los nodos empezando con el raíz y bajando nivel por nivel hacia las hojas y llamamos los *precedentes* de un nodo i todos los nodos $j < i$.

Definición 5.2.1 *Un árbol de Markov es un árbol dirigido que define una familia de distribuciones sobre variables $\{X_i\}$ asociadas con los nodos, las cuales satisfacen:*

$$P(X_i|\{X_j, j \text{ precedentes de } i\}) = P(X_i|X_{pa(i)}). \tag{5.19}$$

El ejemplo anterior es claramente un caso especial de un arbol de Markov; de hecho es uno de lo más usados que se llama un model de cadena oculta.

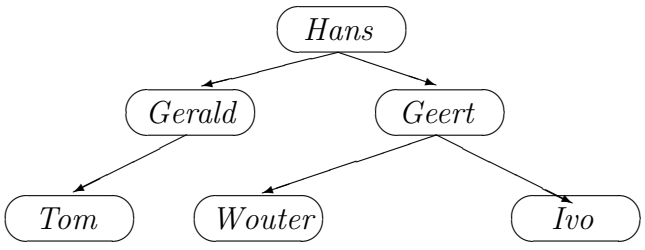
Una propiedad base es la que para un árbol dirigido dado, cualquier conjunto de distribuciones positivas $P(X_i|X_{pa(i)})$, define de una manera única la distribución conjunta:

$$P(X) = \prod_i P(X_i|X_{pa(i)}). \tag{5.20}$$

Lo anterior es una consecuencia de (5.15).

Ejemplo: *Tenemos una enfermedad E que es transmitible solamente por la línea masculina. La probabilidad de que se transmite de padre a hijo es* 0.05*. La probabilidad de desarrollar la enfermedad por otra fuente es* 0.001*; se supone que lo último ocurre independiente para cada individuo.*

Llamamos $X_i = 0$ resp. 1 *si la persona tiene resp. no tiene la enfermedad. Para una familia con el siguiente árbol geneológico, tenemos que $\{X_i\}$ forma un árbol de Markov. Dejamos el cálculo de las probabilidades condicionales de un hijo dado su padre, como ejercicio.*



No es sorprendente que el árbol va a reflejar mucho más independencias que las de forma (5.20).

Así, usando (5.20) se puede demostrar:

Propiedad 5.2.2 *Dado un árbol de Markov de distribuciones positivas. Para conjuntos A, B, C de nodos se tienen:*

$$X_A \perp X_B | X_C \leftrightarrow C \text{ separa } A \text{ de } B \text{ en el árbol no dirigido correspondiente.}$$

Hay que tener cuidado con la interpretación de $\leftrightarrow$. Si los conjuntos son separados, tenemos automáticamente independencia. Si no, existe al menos un miembro en la familia de distribuciones que son representados por el árbol, para lo cual X_A y X_B no son condicionalmente independiente.

Lo anterior permite definir un árbol de Markov de otra manera:

Propiedad 5.2.3 *Una distribución positiva es representada por un árbol de Markov si:*

$$X_A \perp X_B | X_C \leftrightarrow C \text{ separa } A \text{ de } B \text{ en el árbol no dirigido correspondiente.}$$

En la práctica, muchas veces se requiere calcular las probabilidades marginales dada cierta información. Por ejemplo, dado que Geert no tiene E y Tom si, calcula la probabilidad que Hans la tiene, $P(X_1 = 1 | X_3 = 0, X_4 = 1)$.

En teoría se pueden calcular estas probabilidades a partir de la distribución conjunta:

$$\frac{\sum_{x_2,x_5,x_6} P(X_1 = 1, X_2 = x_2, X_3 = 0, X_4 = 1, X_5 = x_5, X_6 = x_6)}{\sum_{x_1,x_2,x_5,x_6} P(X_1 = x_1, X_2 = x_2, X_3 = 0, X_4 = 1, X_5 = x_5, X_6 = x_6)}. \tag{5.21}$$

Sin embargo el número de elementos en las sumas ya no va a permitir un cálculo rápido (o factible). A continuación mostramos como evitar los cálculos de (5.21), usando las independencias de la gráfica.

Propagación de probabilidades

Supongamos que tenemos la información I que para cada v, $X_v \in S_v$ (S_v puede ser de un singletón hasta el universum para esta variable). Denotamos con I_v^- resp. I_v^+ la información de I relacionada con nodos (variables) k y sus descendientes resp. no descendientes de v.

Usando la regla de Bayes y Propiedad 5.2.2, sabemos que para cualquier nodo v

$$P(X_v = x_v | I) \sim P(I_v^- | X_v = x_v) P(X_v = x_v | I_v^+). \tag{5.22}$$

Demostramos primero como calcular $P(I_v^- | X_v)$ por medio de una sola corrida a través del árbol desde las hojas hacia la raíz. Por supuesto, si v es una hoja $P(I_v^- | X_v = x_v) = I(x_v \in S_v)$. En general:

$$P(I_v^- | X_v = x_v) = I(x_v \in S_v). \prod_{\text{hijos } w \text{ de } v} P(I_w^- | X_v = x_v),$$

y

$$P(I_w^-|X_v = x_v) = \sum_{x_w} P(I_w^-|X_w = x_w).P(X_w = x_w|X_v = x_v). \tag{5.23}$$

En consequencia, podemos expresar $P(I_v^-|X_v = x_v)$ en términos de $P(I_w^-|X_w = x_w)$ asociados con sus hijos.

Se puede derivar algo similar para el segundo término $P(X_v = x_v|I_v^+)$ en (5.22) (pero ahorita desde la raíz hacia las hojas):

$$P(X_v = x_v|I_v^+) = \sum_{\text{padres } p \text{ de } v} P(X_v = x_v|X_p = x_p)P(X_p = x_p|I_v^+),$$

y

$$P(X_p = x_p|I_v^+) \sim I(x_p \in S_p).P(X_p|I_p^+). \prod_{\text{hijos } b \neq v \text{ de } p} P(I_b^-|X_p = x_p).$$

Lo último se puede calcular por medio de (5.23).

5.2.4 Redes Bayesianas

El siguiente paso es definir un campo sobre cualquier gráfica dirigida no cíclica.

Definición 5.2.2 *Una red bayesiana es una gráfica dirigida y no cíclica, que define una familia de distribuciones sobre variables $\{X_i\}$ asociadas con los nodos, las cuales satisfacen:*

$$P(X_i|\{X_j, j \text{ precedentes de } i\}) = P(X_i|X_{pa(i)}). \tag{5.24}$$

Otra vez, se tiene que:

$$P(X) = \prod_i P(X_i|X_{pa(i)}). \tag{5.25}$$

Las independencias implicadas son ahora más complicadas. Con ese fin se introduce la *gráfica moral*.

Definición 5.2.3 *Dada una gráfica dirigida $\mathcal{G}$, la gráfica moral $\mathcal{G}^m$, es la gráfica nodirigida donde, si es necesario, se añade una conexión entre los padres de cada nodo.*

Si se define los *antepasados* de un nodo i como todo los nodos j desde donde se puede llegar a i, se obtiene:

Propiedad 5.2.4 *Dado $A, B, C \subset N$:*
B separa A de C en la gráfica moral asociada y restringida a los antepasados de $A \cup B \cup C$ ssi

$$X_A \perp X_C|X_B.$$

5.3 La Distribución Multivariada Gaussiana

Entre todas las distribuciones multivariadas, la multivariada gaussiana tiene un lugar especial por las propiedades que se describen a continuación.

Una manera para introducirla es através de transformaciones de variables gaussianas unidimensionales.
Supongamos que $Y = (Y_1, \cdots, Y_d)$ es un vector con n componentes independientes y $Y_i \sim \mathcal{N}(0,1)$. Para una matriz A de rango n y μ un vector de dimensión n, definimos:

$$X = A.Y^T + \mu.$$

Se puede mostrar que la densidad es de la forma:

$$f_{X_1,\cdots,X_d}(x_1,\cdots,x_d) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} \exp \frac{-(x-\mu)\Sigma^{-1}(x-\mu)^T}{2},$$

donde $\mu_i = EX_i$ y $\Sigma_{i,j} = \mathrm{Cov}(X_i, X_j) = A.A^T$.

Gráficamente la densidad tiene la forma de una *campana* y las curvas de nivel son elipsoides concentricos.

Decimos que X tiene una distribución multivariada con promedio μ y covarianza Σ, $X \sim \mathcal{N}(\mu, \Sigma)$.

En lo que sigue, particionamos el vector X en $(X(1), X(2))$ con $X(1)$, $X(2)$ de dimensión k resp. $d-k$, $\mu = (\mu_1, \mu_2)$ y $\Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}$. Obtenemos:

1. $X(1) \sim \mathcal{N}(\mu_1, \Sigma_{1,1})$
2. $X(2)|X(1) = x(1) \sim \mathcal{N}(\mu_{2|1}, \Sigma_{2|1})$ donde
   $\mu_{2|1} = \mu_2 + \Sigma_{2,1}\Sigma_{1,1}^{-1}(x_1 - \mu_1)$ y $\Sigma_{2|1} = \Sigma_{2,2} - \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}$.

Desde punto geométrico, lo anterior significa que tanto en la intersección de la densidad con un plano como en la proyección en un plano, siempre se obtiene de nuevo una distribución gaussiana.

Una particularidad de la multivariada gaussiana es que

$$\Sigma_{i,j}^{-1} = \mathrm{cov}(X_i, X_j | X_k, k \neq i, j),$$

Es decir, obtenemos una relación sencilla entre las características condicionales y marginales que junto con el promedio determinan toda la distribución.

Capítulo 6

Simulación

El problema central de este capítulo es como generar valores de una distribución usando una computadora, es decir usando un algoritmo deterministico con precision finita. Aparte discutimos algunas estrategias para evaluar la calidad de los diferentes métodos.

6.1 Generar muestras de la distribución $\mathcal{U}(0,1)$

6.1.1 El generador lineal congruencial

Aparte de algunos modestos experimentos basados en la arbitrariedad de ciertos procesos electrónicos (físicos), se usan en la vida cotidiana algoritmos determinísticos para generar valores de variables aleatorias. Como veremos más adelante, para la mayoría de los casos, eso es suficiente.

El algoritmo más popular para generar datos uniformes en el intervalo (0,1) es el *generador lineal congruencial*. Este se basa en la operación *módulo* que sirve para mapear números naturales de una manera no evidente a un cierto intervalo, a través de la generación de una secuencia:

$$x_{i+1} = (ax_i + c) \bmod b. \tag{6.1}$$

En pseudo código su formulación es:

$x = x_0$
`Repite`
  calcula $x = (ax + c) \bmod b$
regresa $u = x/b$

Obsérvese que si b es de la forma 2^k, usando una representación binaria de x, se puede implementar el operador *mod* tomando los últimos k bits de x.

A continuación daremos diversos criterios de calidad que sirven como guía en la determinación de los valores a, c y b.

6.1.2 Determinar la calidad de un generador

La probabilidad permite formular rigurosamente el concepto de una distribución uniforme a nivel de una variable aleatoria. Sin embargo, no nos ofrece una metodología natural o contundente para verificar si un conjunto de números fueron elegidos al azar o para cuantificar la *aleatoriedad* de un conjunto.

Consideramos el siguiente ejemplo, aparentemente más sencillo, donde uno quiere determinar si una secuencia de ceros y unos proviene de una distribución Bern(0.5). Las secuencias

$$01010110100100111010 \quad 10011100111001110011 \quad 10101010101010101010$$

tienen la misma probabilidad de ocurrencia, sin embargo pocas personas estarán dispuestas a aceptar la segunda y tercera como resultado del azar por mostrar *patrones* aparentes.

Como la definición de un patrón no es única, es mejor hablar de aleatoridad *con respecto a un cierto conjunto de patrones*, es decir, como una característica relativa.

Dado un modelo genérico de computación, Kolmogorov considera una secuencia (string) de longitud n, aleatorio si no se puede describir (generar) por un algoritmo con menos que n bits. En este sentido un string s_1 es más aleatorio que s_2 si el mínimo algoritmo necesario para generar s_1 es más largo que el mínimo algoritmo necesario para s_2. Por ejemplo, un algoritmo para generar la segunda secuencia es `repite 4 veces '10011'` lo que es más largo que el algoritmo de la tercera secuencia `repite 10 veces '10'`.

Desafortunadamente el enfoque de Kolmogorov no nos ofrece un mecanismo constructivo para determinar la calidad de un generador, aún si nos restringieramos a la clase de algoritmos, con un limite de tiempo de ejecución. Sin embargo podemos considerar el grado de aleatoridad como una medida de la complejidad, lo que resulta un concepto muy útil en el área de aprendizaje (y que forma la base de criterios como *mimimum description length*).

Otro enfoque para evaluar la calidad de un generador consiste en pasar los datos por *detectores de patrones* ("pruebas''). Las pruebas más sencillas consisten en gráficar (visualizar) ciertas características, como más adelante veremos. Con este mismo objetivo, en la teoría estadística existen un sinnúmero de pruebas llamadas pruebas de hipótesis. Aunque éstas son mucho más operacionales, este enfoque tampoco dará una respuesta contundente. Para el caso del generador congruencial se puede fácilmente mostrar que siempre existirá una prueba que rechazará el generador.

A continuación nos restringiremos a pruebas gráficas y las dividiremos en dos clases: las que evaluan la independencia entre los valores generados y las que se concentran en la distribución propia. Aunque ambas estan intrínsicamente relacionadas, las trataremos de manera separada.

La calidad obtenida de los generadores que se "aprueben'' a través de este método, será suficiente para las aplicaciones que veremos más adelante. Para aplicaciones de muy alta seguridad en criptografía, es posible que se deba recurrir a otros métodos.

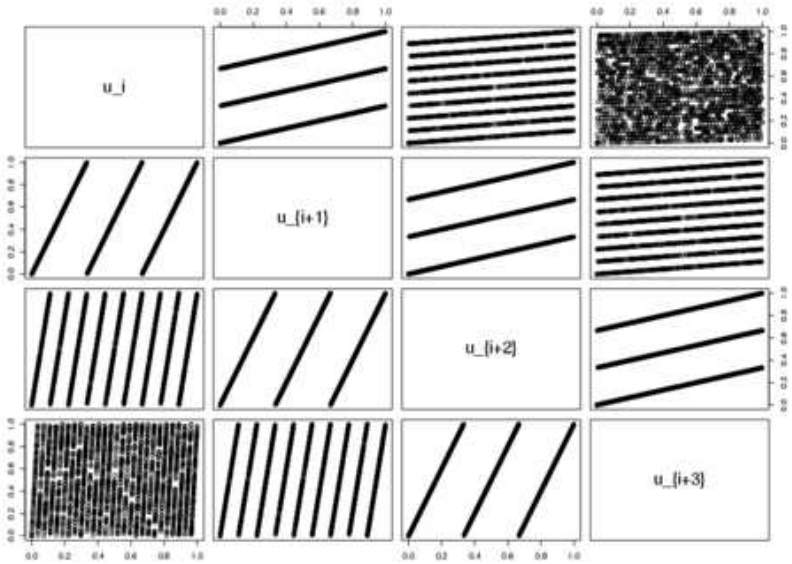


Figura 1: Prueba visual de independencia para la muestra generada con (6.2).

Verificación de la independencia

Para ese fin, es común recurrir a representaciones gráficas usando por ejemplo un *scatter-plot*, donde se grafican pares de observaciones separadas por no más que una distancia dada. La aparición de ciertas estructuras es un indicador en contra de la independencia.

Ejemplo 6.1.1 Consideramos los siguientes dos generadores:

$$x_i = (3x_{i-1} + 2) \bmod (2^{31} - 1); \tag{6.2}$$

$$x_i = (13^{13}x_{i-1} + 1) \bmod (2^{31} - 1). \tag{6.3}$$

que originan las muestras $u_i = x_i/(2^{31} - 1)$.

Como se observa en la Figura 1, el scatter-plot de (6.2) muestra un alto grado de regularidad en la gráfica de u_i versus u_{i+1} y u_i versus u_{i+2}, lo que no ocurre en figura 2, que correspone al generador (6.3).

Verificación de la distribución

Por la propia naturaleza de una computadora digital, los números generados serán siempre elementos de un conjunto finito y no de un intervalo continuo [0,1], como debe ser en teoría. Debido a esto para el caso del generador congruencial, una corrida típica es por ejemplo la que se presenta en la Figura 3. Por tanto entre menor sea la periodicidad del generador, mayor será *la granularidad* de los valores obtenidos.

La teoría de números nos ofrece varias propiedades que nos pueden guiar en la selección de los parámetros para que el periodo seá máximo; sin embargo, no existe ninguna propiedad necesaria y suficiente. Un ejemplo representativo es el que a continuación se expone.

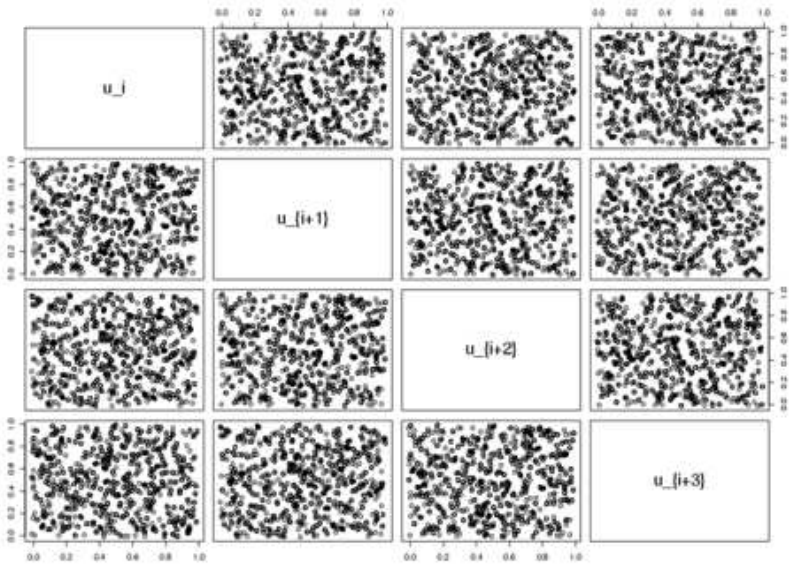


Figura 2: Prueba visual de independencia para la muestra generada con (6.3).

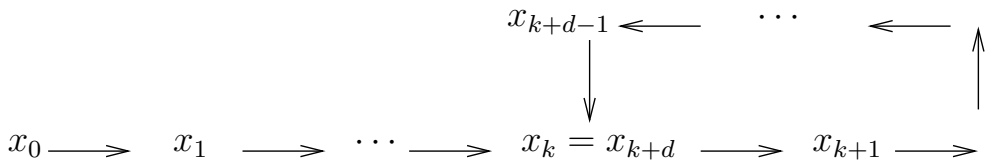


Figura 3.

Propiedad 6.1.1 *Para la secuencia*

$$x_{i+1} = (ax_i + c) \ mod \ b$$

si

1. *el único divisor común entre* b *y* c *es* 1*;*
2. $a - 1$ *es un múltiplo de cada factor primo de* b*;*
3. $a - 1$ *es un múltiplo de* 4 *si* b *lo es;*

entonces el périodo de $\{x_i\}$ *será* b*.*

Por supuesto un periodo máximo (acotado por la precisión de la máquina), no determina por sí solo la calidad de la muestra generada. Uno de los métodos clásicos para

investigar si la distribución de los datos es parecida a la distribución deseada, es comparar el histograma de los datos generados con la distribución teórica. Otra alternativa que es cada vez más popular, es dibujar los cuantiles de la distribución empíricos asociada con los puntos generados, contra los cuantiles de la distribución, usando un Q-Q plot.

Para una distribución con distribución acumulativa continua $F_X()$, define el $\alpha-$ cuantil como aquel valor x_α tal que $F_X(x_\alpha) = \alpha$. En caso de la distribución $\mathcal{U}(0,1)$, el $\alpha-$ quantil es simplemente α.
Por otro lado para un conjunto de observaciones $\{x_1, \cdots, x_n\}$, definimos el $i/n-$cuantil empírico como la i-ésima observación ordenada de chica a grande, lo que denotamos como $x_{(i)}$.

La gráfica de cuantil-cuantil o Q-Q plot consiste en graficar los cuantiles empíricos contra los teóricos y en el caso de la distribución uniforme coincide en graficar $(x_{(i)}, i/n)$. Si la muestra es suficientemente grande, estos puntos deben estar cerca de la primera bisectrica (suponiendo que los datos provienen de $\mathcal{U}(0,1)$).

En la figura 4 se muestra, para 100 números generados, el Q-Q plot junto con 3 histogramas con diferentes números de intervalos o categorías. Como se observa, el número de intervalos afecta la percepción de la forma del histograma, dificultando una comparación con la distribución teórica, que en este caso es la uniforme. En cambio esta subjetividad no se presenta en la construcción de un Q-Q plot.

Ejemplo de valores de parámetros

Algunas elecciones de parámetros en la literatra para el generador lineal congruencial son:

a	c	b
742938285	0	2^{31} -1
950706376	0	2^{31} -1
68909602460261	0	2^{48}
33952834046453	0	2^{48}

Tabla 1

Extensiones

A partir de los generadores anteriores se pueden construir otros con un comportamiento más complejo. Un ejemplo es el uso de una mezcla de generadores congruenciales.

Si $\{X_i^1\}, \{X_i^2\}$ son números generados de una manera independiente y uniforme en $\{0, , \cdots, b-1\}$, se calcula:

$$w = (x_i^1 + x_i^2) \bmod b.$$

Otra extensión consiste en considerar un grado de recursión más alta:

$$x_{i+1} = (a_1 x_i + \cdots + a_{k+1} x_{i-k} + c) \bmod b. \tag{6.4}$$

En muchas situaciones, las técnicas anteriores permiten reducir el valor de b sin una pérdida de calidad notable. Por otro lado, no es por tener más parámetros en el generador que la calidad automáticamente va a ser mejor que un generador lineal congruencial clásico.

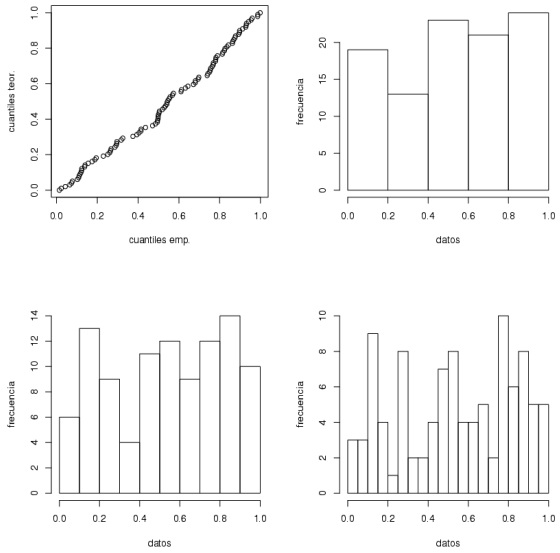


Figura 4.

Es importante subrayar que todos los generadores lineales congruenciales, generan valores en hiperplanos del espacio, lo cual significa que existe una relación lineal entre un valor y el conjunto de valores anteriores. Entre los remedios que existen en la literatura, se encuentra el generador inverso congruencial:

$$x_{i+1} = (a\bar{x}_i + c) \bmod b$$

donde $\bar{x}_{i-1}$ es $1/x_{i-1}$ para un x_{i-1} positivo y cero en el otro caso. Se puede mostrar que para b primo, $x_{i-1}^{-1} = x_{i-1}^{b-2}$.

Ejemplo de valores para a, c y b son respectivamente $1288490188, 1$ y 2147444483647.

6.1.3 Generar muestras arbitrarias

A partir de la generación de muestras de una distribución uniforme, se pueden obtener muestras de cualquier otra distribución a través de la función cumulativa inversa, como se formula en la siguiente propiedad.

Propiedad 6.1.2 *Sea $U \sim \mathcal{U}(0,1)$ y $F(\cdot)$ una distribución dada, entonces $F^{-1}(U)$, tiene distribución $F(\cdot)$.*

Es fácil ver que para el caso discreto, la propiedad 6.1.2 es equivalente a la siguiente.

Propiedad 6.1.3 *Dada una distribución discreta* $\{p_k\}_{k>0}$*, si* U *es una variable con una distribución uniforme sobre* $(0,1)$*, entonces,*

$$Y = \min\{i : U \leq \sum_{k=1}^{i} p_k\},$$

tiene la distribución $\{p_k\}_{k>0}$.

Demostración Demostramos esta última propiedad. Dado que Y es igual a i si y solo si

$$\sum_{k=1}^{i-1} p_k \leq U \leq \sum_{k=1}^{i} p_k,$$

entonces

$$P(Y = i) = P(\sum_{k=1}^{i-1} p_k < U \leq \sum_{k=1}^{i} p_k) = P(U \leq \sum_{k=1}^{i} p_k) - P(U \leq \sum_{k=1}^{i-1} p_k) = p_i.$$

♢

Consideramos los siguientes ejemplos.

Ejemplo 6.1.2 Para generar valores de una distribución Exp(λ), calculamos primero la función de cuantiles. Dado que la función de acumulación tiene la expresión:

$$F_{\exp}(x) = 1 - \exp(-\lambda x)$$

la función inversa es igual a

$$F^{-1}_{\exp}(y) = -\frac{\ln(1-y)}{\lambda}.$$

Como $1 - U$ y U tienen la misma distribución, la propiedad 6.1.2 se reduce a:

`genera` $U \sim \mathcal{U}(0,1)$
`regresa` $-\frac{\ln(U)}{\lambda}$

Ejemplo 6.1.3 Para el caso de una distribución Bern(θ), se obtiene:

`genera` $U \sim \mathcal{U}(0,1)$
`si` $U < 1 - \theta$ `regresa` 0
`si no, regresa` 1

La disposición de una expresión explícita para los cuantiles es más bien una excepción que la regla. En particular en el caso discreto, con distribución dada por $\{p_k\}_1^n$siempre se va a tener que recurrir a una enumeración del tipo:

genera $U \sim \mathcal{U}(0,1)$
cumul<-0; i<-0;
repite
i<-i+1;
cumul <- cumul + p_i;
hasta cumul < U
regresa i;

Una gran clase de distribuciones se definen por medio de una transformación explícita la cual nos permite generar directamente una muestra. En el caso discreto, vimos en el capítulo anterior varios ejemplos de distribuciones derivada de la distribución Bernoulli.

Por ejemplo, dado que la distribución $\text{Bin}(n,p)$ es la suma de bernoullis independientes, se obtiene un valor de $\text{Bin}(n,p)$, sumando n realizaciones de un $\text{Bern}(p)$. Para una distribución geométrica, generamos realizaciones de una Bernoulli hasta encontrar el primer éxito y regresamos el momento de esta ocurencia, etc.

En el caso continuo, la distribución normal juega un papel similar. Desafortunadamente, no existe una expresión explícita para los cuantiles normales. La transformación de *Box-Muller* define una manera para generar realizaciones gaussianas a partir de una distribución uniforme:

Propiedad 6.1.4
Si $U_1, U_2 \sim \mathcal{U}(0,1)$ mutuamente independientes. Entonces

$$X = \sqrt{-2\ln(U_1)}\cos(2\pi U_2), \quad Y = \sqrt{-2\ln(U_1)}\sin(2\pi U_2)$$

son independientes y tienen distribución $\mathcal{N}(0,1)$.
Usando las propiedadades de la distribución normal, se obtiene valores para cualquier distribución normal.

Si el tiempo de cálculo es muy importante y la precisión no tiene que ser alta, un alternativa es generar n números uniformes y promediarlos ya que el Teorema Central de Límite nos garantiza la normalidad si n es suficientemente grande.

A partir de lo anterior, podemos por ejemplo generar valores de una distribución χ^2_n, sumando los cuadrados de los valores de una muestra de tamaño n proveniente de una distribución normal estándar.

6.2 Aplicaciones

6.2.1 Cálculo de integrales

El cálculo de integrales definidas es una aplicación típica de como usar una simulación. Consideramos la integral:

$$I = \int_a^b g(x)dx.$$

Tomando una densidad positiva $f_X(\cdot)$ sobre $[a, b]$, podemos reescribir I como

$$I = \int_a^b \frac{g(x)}{f_X(x)} f_X(x)dx = E(g(X)/f_X(X)),$$

con $X \sim f_X(\cdot)$. Así el problema se convierte en calcular el promedio de $g(X)/f_X(X)$. A partir de una muestra $\{X_i\}$ de X, construimos el estimador

$$\hat{I_n} = \frac{1}{n} \sum_i \frac{g(X_i)}{f(X_i)}.$$

La ley de los números grandes nos garantiza que $\hat{I_n}$ converge a $I = E(g(X)/f_X(X))$. Dado que $\hat{I_n}$ es una función de una muestra, es a su vez una variable aleatoria. Su promedio es :

$$E\hat{I_n} = E(\frac{1}{n} \sum_i \frac{g(X_i)}{f(X_i)}) = \frac{1}{n} nE(\frac{g(X_i)}{f(X_i)}) = I.$$

Ejemplo 6.2.1 Consideremos la integral

$$I = \int_0^1 x^2 dx.$$

Elegimos como $f_X(\cdot)$ la densidad uniforme sobre $(0, 1)$, $\mathcal{U}(0, 1)$. Así para una muestra $\{X_i\}$ de $\mathcal{U}(0, 1)$, el estimador es:

$$\hat{I_n} = \frac{1}{n} \sum_i (X_i)^2$$

En la figura 5 se grafica $\hat{I_n}$ para diferentes muestras como función de n. Se puede observar la convergencia al verdadero valor $I = 1/3$.

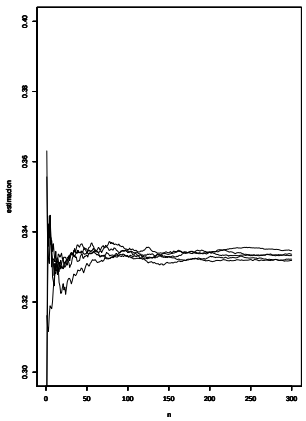


Figura 5.

El hecho que $E\hat{I}_n = I$, justifica el uso de la varianza para cuantificar la variabilidad (incertidumbre) del estimador. Esta es igual a

$$\mathrm{Var}(\hat{I}_n) = \frac{\mathrm{Var}(\frac{g(X_i)}{f(X_i)})}{n} = \frac{E(\frac{g(X_i)^2}{f(X_i)^2}) - I^2}{n}.$$

Aunque no podemos obtener una expresión explícita (por desconocer a I), se observa que entre más similar sea $f_X(\cdot)$ a $g(\cdot)$, menor será la varianza.

Existen diversos métodos para adaptar un estimador para que su varianza sea menor. Si $\hat{I}^1_{n_1}$ y $\hat{I}^2_{n_2}$ son dos estimadores insesgados basados en las muestras $\{X_i\}_1^{n_1}$ y $\{Y_i\}_1^{n_2}$ respectivamente (las muestras que pueden ser o no compartidas), entonces

$$\hat{I}^3_m = \frac{\hat{I}^1_{n_1} + \hat{I}^2_{n_2}}{2},$$

donde m es la cardinalidad del conjunto $\{\hat{I}^1_{n_1} \cup \hat{I}^2_{n_2}\}$, será insesgado y con varianza

$$\mathrm{Var}(\hat{I}^3_m) = \frac{\mathrm{Var}(\hat{I}^1_{n_1})}{4} + \frac{\mathrm{Var}(\hat{I}^2_{n_2})}{4} + \frac{\mathrm{Cov}(\hat{I}^1_{n_1}, \hat{I}^2_{n_2})}{2}.$$

Si logramos que $\hat{I}^1_{n_1}$ y $\hat{I}^2_{n_2}$ tengan correlación negativa, es posible disminuir la varianza, obteniendo un mejor estimador que cada uno separado (suponiendo que las varianza son iguales) ya que la varianza del nuevo estimador será menor que la mínima entre los dos estimadores, utilizando toda la información de la muestra. Es decir,

$$\mathrm{Var}(\hat{I}^3_m) \leq \min\left\{\mathrm{Var}(\hat{I}^1_m), \mathrm{Var}(\hat{I}^2_m)\right\}.$$

Ejemplo 6.2.2 (continuación) Si $X \sim \mathcal{U}(0,1)$, también $(1-X) \sim \mathcal{U}(0,1)$. Claramente estas dos variables aletorias tienen correlación negativa. Usando esta propiedad, para una muestra $\{X_i\}_1^n$ de $\mathcal{U}(0,1)$, construimos los estimadores $\hat{I}^1_n = \sum_i (X_i)^2/n$ y $\hat{I}^2_n = \sum_i (1-X_i)^2/n$. La varianza del promedio de los dos es:

$$\mathrm{Var}(\hat{I}^3_n) = 2\frac{\mathrm{Var}(\hat{I}^1_n)}{4} - \epsilon < \mathrm{Var}(\hat{I}^1_n).$$

En la figura 6 se muestra $\hat{I}^3_n$ para diferentes muestras, con el objetivo de compararse con la figura 5.

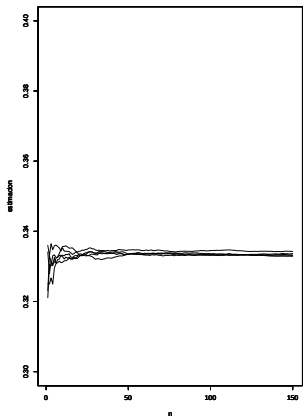


Figura 6.

6.2.2 Optimización global y probabilidad

El problema genérico que consideramos en esta sección es buscar en un espacio Ω un óptimo (máximo o mínimo) de una función $f(\cdot)$ definada sobre Ω.

Situación clásica

El caso más fácil y mejor estudiado es sin duda cuando Ω es parte de $\mathcal{R}^m$ y $f(\cdot)$ es suficiente suave (por lo menos derivable) por la existencia de una caracterización necesaria de los óptimos

$$x \text{ es óptimo, entonces } \frac{df(x)}{dx} = 0.$$

En caso de no tener una solución explícita para x se puede recurrir a métodos iterativos de búsqueda local de los cuales el *método del gradiente* es el más conocido. En este consecutívamente se avanza en el espacio Ω con pasos pequeños en la dirección del gradiente (en caso de buscar un máximo) o en la dirección opuesta al gradiente (en caso de buscar un mínimo) hasta quedar atrapado en un valor que se toma como (aproximación) a un óptimo.

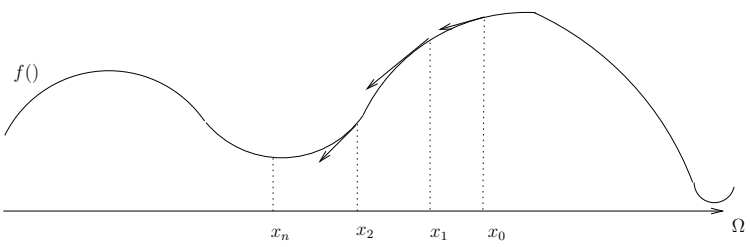


Figura 7.

Un método iterativo de busqueda local, genera una partición natural del espacio, agrupando en la misma clase los puntos que se usan como punto de arranque y a partir de los cuales se converge al mismo óptimo local. Ver Figura 8.

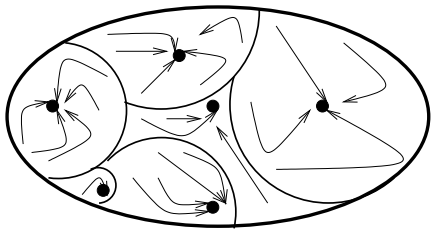


Figura 8.

De esta manera, se observa que si el interés es en óptimos globales y si la función no posee propiedades muy particulares (como convexidad), el problema es encontrar puntos de arranque interesantes.

Para evitar quedar atrapado siempre en los mismos óptimos, un remedio es incluir un elemento aleatorio en el algoritmo. Los ejemplos más sencillos son los algoritmos *multi-start* que eligen una cierta distribución sobre los puntos de Ω para escoge unos y correr el algoritmo usandolos como puntos de arranque. El mejor óptimo local encontrado dentro del conjunto obtenido se elige como aproximación a un óptimo global.

La elección más sencilla para la distribución de sorteo de los puntos es la uniforme pero esquemas avanzados para la elección toman en cuenta la ubicación de los puntos entre sí (es de esperar que puntos muy cercanos van a conducir al mismo óptimo) y la calidad de los óptimos locales alcanzados.

Cuando el espacio es finito

Si Ω es finito, el panorama cambia completamente. En primer lugar porque ya no existe el concepto de derivada que resume de una manera compacta (analítica) el comportamiento de una función derivable en una vecindad alrededor de un punto. Además ya no existe el concepto *vecindad natural*; así métodos basados en diferencias finitas implican una decisión (subjetiva) que influye en la calidad de la solución obtenida.

Tomemos como ejemplo el problema del agente viajero, que es representativo de una gran familia de problemas de optimización combinatoria que en complejidad son equivalentes entre sí. Por ejemplo, ciertos problemas de ruteo, de "scheduling", ordenamientos, etc.

Definición 6.2.1 *El problema del agente viajero consiste en buscar un orden para visitar un conjunto dado de ciudades de las cuales se tienen las distancias entre sí. Se quiere que la distancia recorrida sea mínima y que cada ciudad sea visitada exactamente una vez.*

Para este ejemplo, cada elemento de Ω corresponde a un recorrido o itinerario específico y $f(\cdot)$ su distancia total.

Podemos decir que dos recorridos son vecinos si difieren solamente en el orden de visitar dos pares de ciudades consecutivas como se ilustra en la Figura 9.

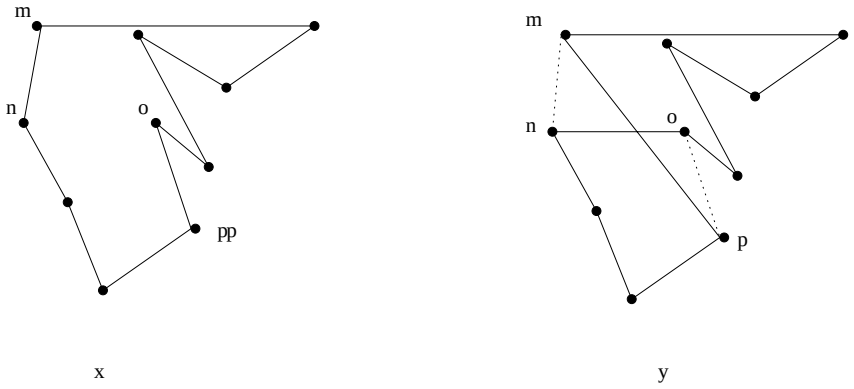


Figura 9.

Obsérvese que

$$f(y) - f(x) = d(m,p) + d(n,o) - d(m,n) - d(o,p),$$

donde $d(a,b)$ representa la distancia entre a y b.

A partir de lo anterior se construye el siguiente algoritmo:

$x \leftarrow x_0$;
`mientras no-suficientemente-bien(`x`)` $\wedge$ `no atrapado()`
    $y \leftarrow$ `genera-vecino-de(`x`);`
    `IF` $f(y) < f(x)$
        $x \leftarrow y$;
`regresa` x;

El algoritmo va a quedarse atrapado en un mínimo local donde un punto x es "local'' si la función `genera-vecino-de(`x`)` no puede generar una solución mejor que x.

En general se tiene que buscar un compromiso entre el número de óptimos locales y la facilidad de cómputo para verificar si $f(y) < f(x)$. Un extremo corresponde a permitir tranformar un recorrido a cualquier otro con la consecuencia de ya no poder usar el valor de $f(x)$ en el cálculo de $f(y)$. Este algoritmo se conoce con el nombre de *búsqueda aleatoria* (random search).

A continuación describimos dos soluciones que hacen uso de un componente aleatorio que permite alcanzar óptimos globales sin tener que caer en el extremo (ineficiente) de búsqueda aleatoria.

En ambos casos, el componente probabilístico va a servir para *explorar* el espacio de búsqueda de una manera arbitraria, contrario al componente determinístico que se basa en, por ejemplo, información analítica (el componente explotación).

Simulated Annealing

La característica principal de este algoritmo iterativo es que al buscar una nueva aproximación y dado x, se acepta de vez en cuando una solución peor a x por medio de un mecanismo probabilístico.

El esquema base es:

$x \leftarrow x_0$
$T \leftarrow T_0$
`mientras no-suficientemente-bien(`x`)`
    $y \leftarrow$ `genera-vecino-de(`x`)`
    $\Delta f \leftarrow f(y) - f(x)$
    `If` $\Delta f < 0$
        $x \leftarrow y$
    `else`
        $x \leftarrow y$ `con probabilidad` $e^{(-\Delta f)/T}$
    $T \leftarrow$ `baja-temperatura(`T, x`)`

donde:

- `genera-vecino-de():` genera un vecino y de x tal que se pueda calcular fácilmente $f(y) - f(x)$; esta función determina el número de óptimos locales. Las restricciones teóricas son que se debería poder generar cualquier estado a partir de cualquier otro estado, usando eventualmente un número de estados intermediarios y que la probabilidad de generar x a partir de y es igual a la probabilidad de generar y a partir de x.

- `baja-temperatura():` Típicamente la nueva temperatura es α veces la anterior, con α entre 0.9 y 0.99; un valor demasiado grande implica una pérdida de cálculos porque se acepta casi cualquier otro valor. Por otro lado un valor demasiado pequeño impide escapar de óptimos locales.

- T_0**:** Es tal que acepta, en las primeras iteraciones del algoritmo, la mayoría de los cambios (en general, se determina el valor de una manera empírica corriendo el algoritmo a cierta tempertura y dependiendo del número de aceptaciones se incrementa o disminuye el valor de T).

En general se observa que fijando Δf, entre más baja sea la temperatura, menor será la probabilidad de aceptar una solución peor; es decir al inicio el énfasis está en la exploración del espacio y al final a la explotación.

Por otro lado, fijando T, se ve que entre peor sea y con respecto a x, menor será la probabilidad de aceptarla.

Unas de las grandes ventajas del algoritmo es el reducido número de parámetros. Para la descripción teórica de este algoritmo (usando cadenas de Markov), nos referimos a un curso más especializado.

Figura 10.

Algoritmos Genéticos

Los algoritmos genéticos extienden lo anterior trabajando en cada paso t con un conjunto (*población*) de aproximaciones y permitiendo más libertad en la transición de una población a la siguiente. El trabajar con una población implica que hay un paralelismo implícito en estos algoritmos.

La transición de una población de aproximaciones a la siguiente se lleva a cabo en diferentes etapas (inspirada en la evolución biológica donde se considera la función $f(\cdot)$ como la adaptación al medio (*Fitness*) de cada individuo).

1. **Selección**: Se selecciona una nueva población, por medio de la elección de elementos de la población actual tal que para un elemento x, entre mejor sea $f(x)$, más grande será la probabilidad de ser seleccionado.
   Muchas veces se hace una muestra independiente dónde la probabilidad de seleccionar x es $f(x)/\sum_y f(y)$.

2. **Cruzamiento**: A partir de la población del paso anterior, se elige por azar parejas x, y y se forma *hijos* combinando partes de la solución que representan x y y. Muchas veces x es de la forma $(x_1, \cdots, x_n)$ con cada componente binario (es decir un bit string). Así, una implementación eficiente es tomar los primeros k bits de x (resp. y) y pegarlos con los últimos $l-k$ bits de y (resp. x) donde k es elegido arbitrariamente, siempre y cuando los nuevos elementos pertenezcan a Ω.

3. **Mutación**: Con cierta probabilidad se hace pequeñas perturbaciones en x. Por ejemplo si x se codifica como un bitstring, se cambia cada bit con una probabilidad que, en general, es muy pequeña.

Lo anterior se presta a muchas variantes. Igual a métodos de simulación, los algoritmos genéticos son - contrario a métodos analíticos de optimización - muy flexibles y generales, muchas veces a cambio de un tiempo de cálculo mayor.

Capítulo 7

Aplicaciones

7.1 Búsquedas en Strings

Ante el crecimiento exponencial de la información guardada de forma digital (desde bases de datos, archivos de textos, hasta páginas WWW), existe una gran necesidad de disponer de algoritmos eficientes para encontrar una o más ocurrencias de una palabra clave en un texto.

En lo que sigue, P representará la palabra o cadena que deseamos buscar en el texto T. En general, T es una cadena muy grande. Su longitud denotaremos por m, puede ser del orden de 10000 - 100000000 caracteres. En contraste P es relativamente chico y su longitud n, es en general del orden de 5-100 caracteres.

7.1.1 Soluciones determinísticas

El método más fácil pero sumamente ineficiente, consiste en alinear P con cada caracter de T y verificar si los caracteres de P coinciden con los caracteres correspondientes de T como se ilustra en la figura 1. La complejidad de este método es de orden $n \times m$.

```
D: amaaimagdaa...
T: maaide
    maaide
     maaide
      maaide
       maaide
        maaide
         maaide
        ---------
        12345678
```

Figura 1.

A partir de la idea anterior surgen dos grupos de algoritmos: unos donde se trata de detectar casos en los cuales se puede avanzar la cadena P con más de una posición, usando cierta información obtenida por un preprocesamiento de T, y otros donde se busca sobre una estructura de datos más adecuada que un string.

En el ejemplo representado en la figura 1 se empieza a comparar P y T caracter por caracter. Supongamos que se verifica si P alinea con T a partir del segundo caracter de T; no es difícil ver que al detectar una diferencia en la quinta posición de P, eso implica que la próxima comparación puede empezar a partir de la posición 6 en T. Métodos derivados de ese algoritmo son el de Knuth-Morris-Pratt y el de Boyer-Moore.

La figura 2 representa un ejemplo de un sufix-árbol: un sufix-árbol para T tiene m hojas y su principal característica es que para cada hoja i, la concatenación de las etiquetas asociadas con las aristas que forman el camino de la raíz hasta la hoja i, forman la subcadena de T que empieza en la posición i.

La búsqueda de algún texto consiste en comparar los carácteres de P con las etiquetas que se encuentran en un camino desde la raíz, hasta agotar los caracteres de P o hasta ya no encontrar una arista por donde se pueda continuar. En el último caso, eso indica que P no aparece en el texto T.

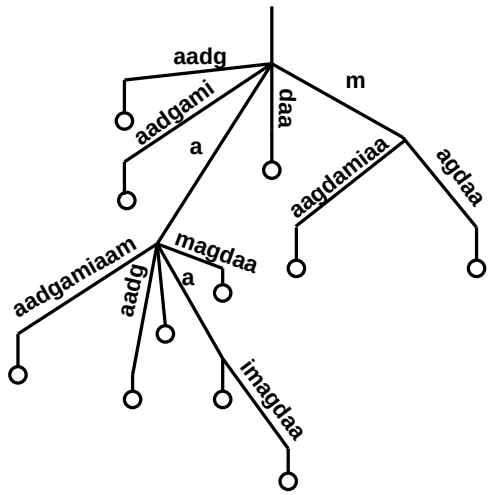


Figura 2.

7.1.2 Finger Printing: una solución probabilística

Verificación de una igualdad

Consideremos primero un problema más sencillo: el verificar si P y T son iguales. Con ese fin, mapeamos de una manera única cada string a un número entero.
Si definimos P_i como el i-ésimo bit en la expansión binaria de P, definimos $H(P)$ como:

$$H(P) = \sum_{i=1}^{n} P_i 2^i,$$

y de una manera análoga definimos $H(T)$. De esta forma $H(P)$ permite relacionar un único número entero a cada string.

Definimos "x mod p" como el residuo de la división de x entre p. Por ejemplo: $10 \bmod 3 = 1, 7 \bmod 3 = 1, 6 \bmod 7 = 6$.

El método de búsqueda del string, basado en la definición de $H(\cdot)$, se describe en el siguiente esquema:

Variables por determinar:
k y N, números enteros.

Algoritmo:
- `contador=0; igualdad=true;`
- `Mientras contador es menor que` k `e igualdad es true:`
  `se elige un número primo` p `menor que` N `al azar;`
  `se calcula` $y_1 = H(T)$ `mod` p `y` $y_2 = H(P)$ `mod` p`;`
  `Si` $y_1 \neq y_2$`, igualdad=false;`
  `en otro caso, contador=contador+1;`
- `Si igualdad es false, concluye que son diferentes,`
  `en el otro caso, concluye que son iguales.`

Consideramos primero una sola ejecución de la repetición ($k = 1$). Si se detecta que $y_1 \neq y_2 \bmod p$, se decide correctamente que T y P son diferentes. Sin embargo la igualdad resultante del algoritmo no permite deducir que T y P son idénticos.

Si el primo p fuera fijo, cometeríamos errores sistemáticos. El hecho de elegir p al azar, nos permite calcular la probabilidad de equivocarnos sobre la igualdad de P y T, y esta probabilidad será en general pequeña. Para calcularla, llamamos P a la probabilidad de tener que $y_1 \bmod p$ es igual a $y_2 \bmod p$. Usando el hecho que $y_1 \bmod p = y_2 \bmod p$ si y solo si p divide $|y_1 - y_2|$, entonces:

$$P = \frac{\#\{p : p < N, p \text{ primo y divide } |y_1 - y_2|, \text{ con } y_1 \neq y_2\}}{\Pi(N)},$$

donde $\Pi(N)$ es el número de primos menor que N. Por otro lado se tiene la propiedad que si $0 < a < m$, el número de primos que dividen a es acotado por $\Pi(\log m)$. Entonces,

$$P \leq \frac{\Pi(\log m)}{\Pi(N)}$$

Dado que $\Pi(\cdot)$ es una función creciente, si $N > \log m$, lo anterior es mucho menor que 1.

Considerando varias iteraciones, dado que cada una es independiente de las anteriores, se obtiene que la probabilidad de sacar una conclusión equivocada es igual a $P^k < 1$. Para cualquier ϵ podemos encontrar un k tal que $P^k < \epsilon$.
Regresamos ahora al problema original.

Búsqueda en un string

Definamos $T_{l,n}$ como el substring de T que empieza en la posición l, y que tiene longitud n.

Aplicando el algoritmo anterior para cada $T_{l,n}$, con $1 \leq l \leq m-n$, dado que

$$H(T_{l,n}) = (H(T_{l-1,n}) - T_{l-1})/2 + 2^n T_{l+n-1},$$

se pueden reducir sustancialmente los cálculos.

El algoritmo anterior es otro ejemplo de un *algoritmo aleatorio* (randomized algorithm). Comparando con métodos determinísticos, se observa que :

1. Se trata de un método en general mucho más sencillo (aunque no siempre más eficiente);

2. Incluye un elemento iterativo;

3. Permite un análisis de complejidad;

4. Son más generales (en el ejemplo anterior se podría trabajar con patrones en 2 o más dimensiones sin ninguna adaptación substancial).

7.2 Identificación de un usuario

La manera clásica para restringir el acceso a un recurso o servicio (como a un archivo, a una cuenta que se accesa a través de un cajero automático o a una computadora), es el introducir de un nombre de usuario y una clave secreta correspondiente que el usuario U debe enviar a un verificador V. Así surgen, entre otros, tres aspectos a cuidar:

1. Evitar que la clave sea previsible por un extraño.

2. Evitar que al interceptar la comunicación entre U y V, algún otro usuario pueda presentarse como U.

3. En caso de requerir el respaldo de una clave, evitar algun abuso de este respaldo en el lugar donde está V.

Para que no sea fácil predecir la clave, lo mejor es elegirla de una manera aleatoria (al menos en teoría, ya que algunas personas suelen escribir sus claves difíciles en un papelito que luego se pierde ...). Es decir elegir la clave C de la distribución $\mathcal{U}(\Omega)$.

Con el objetivo de impedir la intercepción y el reuso del valor de C, se puede trabajar con una función Hash $h(.,.)$ cuya característica principal es que no es invertible y tal que cambios pequeños en los valores de los parámetros impliquen cambios grandes de valor para $h(.,.)$.

Distribución de la información:
U y V conocen una función hash $h(.,.)$ y V tiene c, la clave de U.
U pretende que su llave es x.
Algoritmo:
- `U envia a V un mensaje que quiere entrar al sistema`
- `V envia un mensaje` m `que depende del momento y un número aleatorio`
- `U envia` $y = h(x, m)$
- `V verifica si` $y = h(c, m)$.
  `Si no son iguales, V no reconoce la persona como U`
  `en el otro caso, V reconoce la persona como U.`

Public Key Systems

En el siguiente método para identificar a un usuario, se recurre a ciertas propiedades de la teoría de números para construir una función que, sin un conocimiento adicional, no es fácil invertir.

El método que a continuación se expondrá está basado en la siguiente popiedad.

Propiedad 7.2.1 *Si p y q son dos números primos mayores que 2, y a y b son dos números enteros que satisfacen*

$$ab = 1 \ mod \ (p-1)(q-1),$$

entonces si definimos $n = pq$, para cualquier $x \in \mathcal{Z}_n = \{0, \cdots, n-1\}$

$$(x^b)^a = x \ mod \ n. \tag{7.1}$$

La transmisión, en este caso, sigue la secuencia:

Distribución de la información:
el valor de n y b son públicos
V conoce a y c, la clave de U
U pretende que su llave es x.
Algoritmo:
- `U envia a V` $y = x^b$ `mod` n
- `V verifica si` $z = y^a$ `mod` n

  `Si no son iguales, V no reconoce la persona como U`
  `en el otro caso, V reconoce la persona como U.`

La propiedad 7.2.1 garantiza que V puede recuperar el mensaje original de U. En teoría un espía podría resolver $y = x^b$ mod n para y, b y n dadas o tratar de factorizar n para calcular a. Sin embargo, si p y q son suficientemente grandes, esto no es factible computacionalmente.

Métodos de Conocimiento Cero

Todos los métodos anteriores resolvieron hasta cierto punto el problema de la inseguridad durante la transmisión. Una desventaja común es que el receptor debe conocer (guardar) la llave secreta o alguna transformación de ella. Para evitar eso, los métodos de conocimiento cero (*zero knowledge proofs*) fueron desarrollados .

La idea básica es poder convencer a alguién de conocer cierta información sin revelar nada de ella. Se observará que es aplicable en un contexto mucho más general.

Como introducción considera la cueva de la figura 3. La puerta se abre con una clave secreta. Supongamos que U pretende conocer la clave y que V lo quiere verificar. Consideramos el siguiente método:

Repite k veces:
• U entra a la cueva y camina por una de los dos pasillos mientras V se queda afuera;
• V entra a la cueva y elige también al azar un de los dos pasillos;
• Dependiendo de la elección de V, U usa su clave para abrir la puerta y salir al otro pasillo para que V no le encuentre;
• Si V encuentra a U, decide que U no conoce la clave y se termina todo.

Si V no encontró ninguna vez a U, concluye que U conoce la clave.

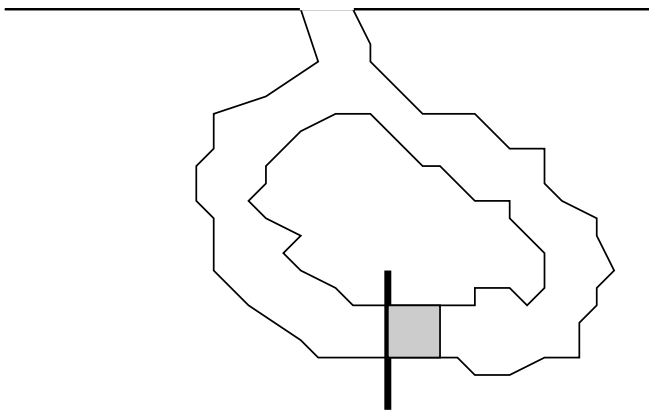


Figura 3.

Dejamos como ejercicio verificar que (1) V no aprenda nada de la clave, (2) como V elige su camino siempre al azar, para k suficientemente grande, alguién que trata de imitar a U, con gran probabilidad, no siempre se va poder esconder de V y (3) V puede asegurarse con gran probabilidad de si U conoce la clave o no.

Consideramos el siguiente algorítmo como ejemplo genérico de una implementación de lo anterior:

Distribución de la información:
$v = c^2 \bmod n$ es público, junto con n.
V elige k.
U pretende que su llave es x.

Algoritmo:
- `igual=true;`
- `contador=0;`
- `Mientras contador` $< k$ `e igual:`
  `U elige un número aleatorio` r `de` $\{1, \cdots, n-1\}$ `y envia` $y_1 = r^2$ `mod` n
  `V elige al azar un bit, b (0 o 1), y envia el resultado a U`
  `U envia` $y_2 = r$, `si` $b = 0$, `y` $y_2 = rx$ `mod` n, `si` $b = 1$.
  `V verifica si` $y_2^2 = y_1.v^b$, `en caso negativo, igual=false`
  `en el otro caso, contador=contador+1.`
- `Si no son iguales, V no reconoce la persona como U`
  `en el otro caso, V reconoce la persona como U.`

Es fácil entender que si U conoce la verdadera c, entonces V va a sacar la conclusión correcta. Pero ¿qué dificil es engañar a V?.

Consideramos los 2 escenarios, b igual a 0 y b igual a 1:

$b = 0$**:** U puede elegir y_2 al azar de $\{1, \cdots, n-1\}$, envia primero $y_1 = y_2^2$ y después y_2. Como $y_2^2 = y_1.v^b$, `igual` va a ser true.

$b = 1$**:** U puede elegir y_2 al azar de $\{1, \cdots, n-1\}$, envia primero $y_1 = y_2^2/v \bmod n$ (hay algorítmos eficientes para calcularlo) y después y_2. Como de nuevo $y_2^2 = y_1.v^b$, `igual` va a ser true.

El problema consiste en que U no conoce a priori el valor de b. Si piensa que V va a elegir $b = 1$ y entonces opta erroneamente por el segundo escenario, implica que debe calcular la raiz de $y_2^2/v \bmod n$ para poder contestar bien. Hasta el momento no existe un algoritmo rápido (si n es grande de la forma $n = pq$ con p y q números primos grandes). En el otro caso, pensar que $b = 0$ y V elige $b = 1$, implica que conozca z o equivalente, la raiz de v que es igual de dificil. Así, si U elige con igual probabilidad 0 y 1, la probabilidad que igual =true si no se conoce z es igual a $1/2$ (acceptando que no existe un algoritmo eficiente para factorizar o calcular raices sobre $\mathcal{Z}_p$).

Dado que se itera lo anterior k veces, la probabilidad de sacar una conclusión equivocada es igual a $1/2^k$ lo que a su vez para k suficiente grande es tan chico como uno quiere.

Observa que si la secuencia fuera conocida de antemano, alguién podría grabar la comunicación entre U y R y luego presentarse a U como V usando las respuestas que U había dado anteriormente. Dado que b es aleatorio, la probabilidad que la secuencia de b's va a coincider con la grabada (y que implique que no se va detectar la trampa) será $1/2^k$.

Lo más característico del algoritmo anterior es que V no aprende nada acerca de z, al menos si uno está dispuesto de acceptar que no existe un algoritmo que en un tiempo razonable poder factorizar. Para ese fin considera el siguiente escenario: V genera arbitrariamente b y dependiendo del valor genera (y_1, y_2) como indicado en los dos escenarios anteriores.

No es difícil ver que la distribución correspondiente de (y_1, y_2) es igual a si fuera el resultado con un dialogo con U (conociendo z): si $b = 0$, en ambos casos tenemos algo de la forma (R^2, R) donde R tiene una distribución uniforme. Si $b = 1$, hay que comparar la distribución de (R^2, Rz) versus $(R^2/v, R)$. Dado que U no conoce z la distribución de Rz es igual a R (ambos de una distribución de conteo). De esta manera es suficiente verificar que la distribución de la primera coordenada dada la segunda es igual en ambos casos: si $y = rz$ automáticamente $r^2 = y^2/v$, entonces las primeras coordenadas coincidirán.

7.3 Cadenas de Markov

7.3.1 Definición

Considera el siguiente grafo (dirigido).

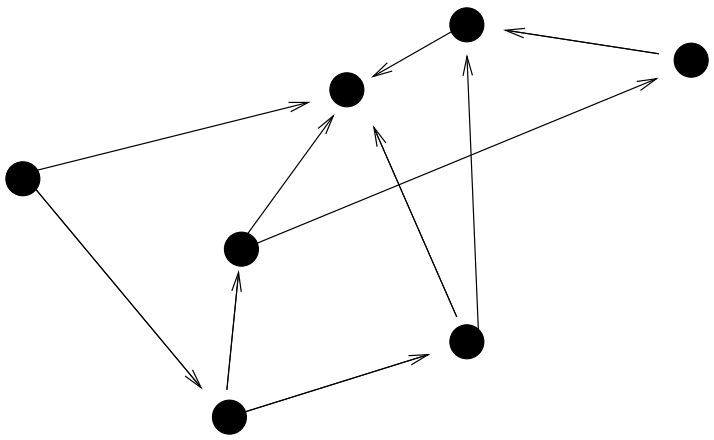


Figura 4

En cada momento, n, uno está en uno de los nodos; denotamos con X_n el lugar específico. Para determinar a donde ir en el siguiente momento $n+1$, se genera un valor de una distribución que solamente depende de donde uno está en ese momento; por ejemplo se elige al azar uno de los nodos vecinos. Se repite lo anterior sucesivamente y de esta manera se obtiene una secuencia $\{X_n\}$.

Por supuesto en general $\{X_n\}$ no forman variables aleatorias independientes, sin embargo no es dificil ver que por construcción:

$$P(X_n|X_i = x_i, i < n) = P(X_n|X_{n-1} = x_{n-1}).$$

Eso forma la base de lo que se llama una cadena de Markov.

Definición 7.3.1 *Séan $\{X_n\}_{n\geq 0}$ variables aleatorias sobre un conjunto finito Ω, $\{X_n\}$ forma una cadena de Markov ssi:*

$$P(X_n|X_i = x_i, i < n) = P(X_n|X_{n-1} = x_{n-1}).$$

Llamamos a la cadena homogenea si estas probabilidades no dependen de n y en este caso definimos una matriz M por medio de $M_{x,y} = P(X_n = y|X_{n-1} = x)$. Se llama M la matriz de transición.

Una consecuencia de esta definición es que, si definimos B como un evento dependiente de $X_{n-2}, \cdots, X_0$, entonces:

$$P(X_n|X_{n-1}, B) = P(X_n|X_{n-1}). \tag{7.2}$$

La matriz de transición define las transiciones entre estados en un solo paso. Calculamos las probablidades de transiciones entre estados en dos pasos usando (2.13):

$$P(X_{n+2} = x_{n+2}|X_n = x_n) = \sum_{x_{n+1}} P(X_{n+2} = x_{n+2}|X_{n+1} = x_{n+1}, X_n = x_n)P(X_{n+1} = x_{n+1}|X_n = x_n)$$

Por (7.2), lo anterior es igual a

$$\sum_{x_{n+1}} P(X_{n+2} = x_{n+2}|X_{n+1} = x_{n+1},)P(X_{n+1} = x_{n+1}|X_n = x_n) = (M.M)_{x_n,x_{n+2}}.$$

En general se tiene:

$$P(X_n = y|X_0 = x) = (M^n)_{x,y}.$$

Para una clase muy amplia de cadenas de Markov, M^n va a converger. Eso no implica que la cadena va a converger a un valor, más bien que la distribución correspondiente converge!

Definición 7.3.2 *Dada una cadena, llamaremos a dos estados x, y conectados, si existen m, n tal que $M^m_{x,y}, M^n_{y,x} > 0$. Si todos los estados son conectados, entonces llamaremos a la cadena irreducible.*

Definición 7.3.3 *El periodo de un estado está definido como el máximo común divisor de todos los naturales n para los cuales $M^n_{x,x} > 0$. Cuando el período es 1, se llama el estado no periódico. Si todos los estados lo son, se llama a la cadena no periódica.*

Podemos definir una relación de equivalencia por medio del concepto de *conectividad*. Se puede mostrar que el período de un estado son propiedades de clase, es decir todos los elementos de una clase tienen el mismo período.

Propiedad 7.3.1 *Séa $\{X_n\}$ una cadena de Markov sobre un espacio finito. Si la cadena es no periódica e irreducible, entonces $P(X_n = \cdot | X_0 = x_0)$ converge a una distribución π que no depende de x_0. La distribución π es la solución única del sistema de ecuaciones:*

$$\pi_x = \sum_y \pi_y M_{y,x} \tag{7.3}$$

con

$$\sum_x \pi_x = 1.$$

Para una cadena irreducible (no necesariamente aperiodico), si definimos $N^n_{x,y}$ como el número de visitas a y en n movimientos saliendo de x, entonces:

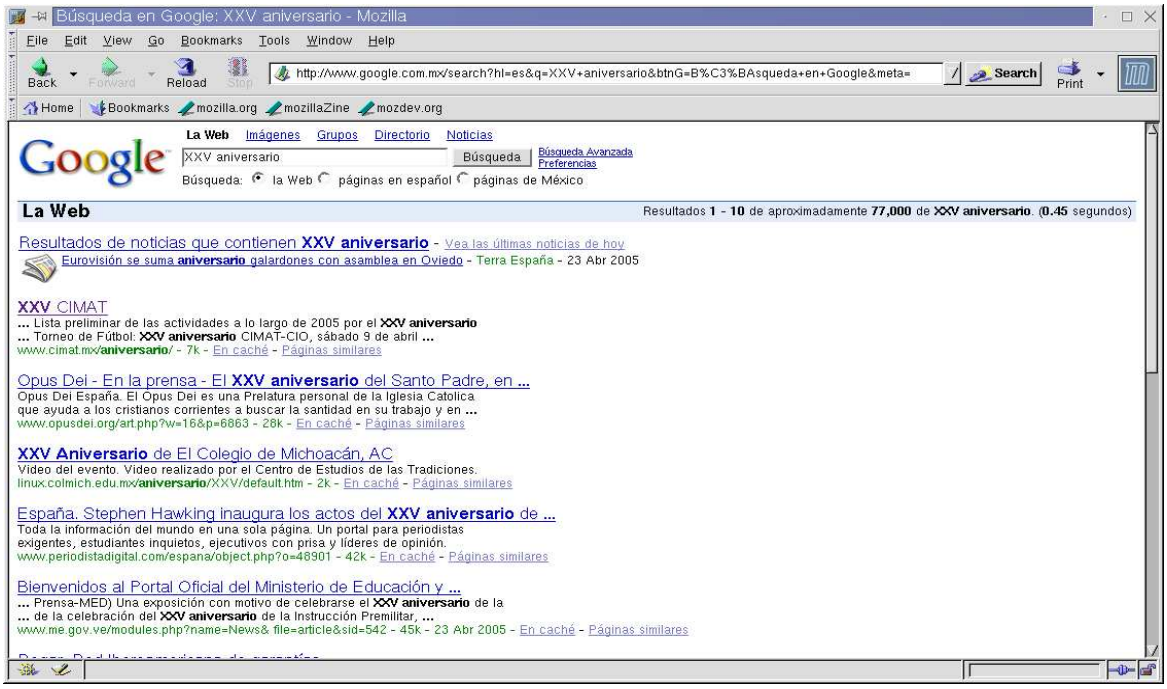
$$\lim_{n\to\infty} E\frac{N^n_{x,y}}{n} = \pi_y,$$

entonces podemos interpretar π_y como el promedio del número de visitas a y.

Observa que (7.3) implica que π es un vector propio de M con valor propio igual a 1.

7.3.2 Google meets Markov

La figura 4 podemos considerar como un (mini)sitio WWW: los nodos representan las páginas WWW y las conexiones reflejan las ligas. Construye una caminata sobre las páginas WWW de la siguiente manera:
supongamos que en momento n uno está viendo la página x; con probabilidad α la siguiente página que se visitará será elegida al azar entre **todas** las páginas; con probabilidad $1-\alpha$ se elige al azar una página y al cual hay en la página x una liga.



El algoritmo Pagerank calcula la distribución de equilibro de esta cadena de Markov para la red WWW; asigna el valor de π_x como peso (rank) a cada página x. Google usa este algoritmo para ordenar las páginas WWW que cumplen con una cierta búsqueda (querry).

Capítulo 8

Inferencia Estadística

8.1 Modelación, Inferencia y Predicción

Datos y computación son estrechamente ligados entre sí. La problemática de respaldo, transmisión y manipulación de datos de manera eficiente, forma una piedra angular del área de Ciencias de la Computación.
En muchas aplicaciones el interés no es tanto en los datos mismos sino en ciertas regularidades subyacentes que pueden ser de utilidad por ejemplo en la toma de decisiones a futuro. En particular en el área de aprendizaje máquina los ejemplos abundan.

Piensa por ejemplo en una base de digitos escritos a mano. Un problema de gran interés es como usar la información en estos datos para construir un algoritmo que clasifica de manera automática dígitos nuevos. Para que los datos actuales tendrán relevancia para datos que se presentarán en el futuro, es muy natural suponer que la distribución de ambos es igual. Es decir, existe una mecanismo subyacente que llamaremos *un modelo* que generó los datos actuales y generará los futuros. En este modelo hay un componente intrínseco de incertidumbre: por ejemplo ningún digito escrito a mano es igual como vemos en la Figura 1. Por eso vamos a suponer que el modelo es probabilístico. Eso es representado en la Figura 2 (a).

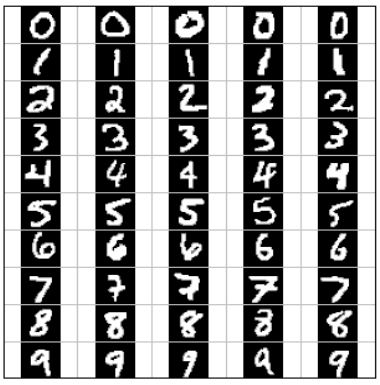


Figura 1

Compara lo anterior con la siguiente situación: un software toma un cierto tiempo para efectuar una cierta tarea (por ejemplo ordenar un conjunto de 100 números). Por la naturaleza de la tarea y del algoritmo, supongamos que el tiempo requerido es una variabla aleatoria con una cierta distribución. Por razones de control de calidad, usando datos de varias corridas del software, queremos conocer un aspecto específico del modelo: por ejemplo el tiempo promedio o verificar si hay mucha evidencia para creer que el tiempo promedio no sobrepasa un cierto límite. Eso es representado en la Figura 2(b).

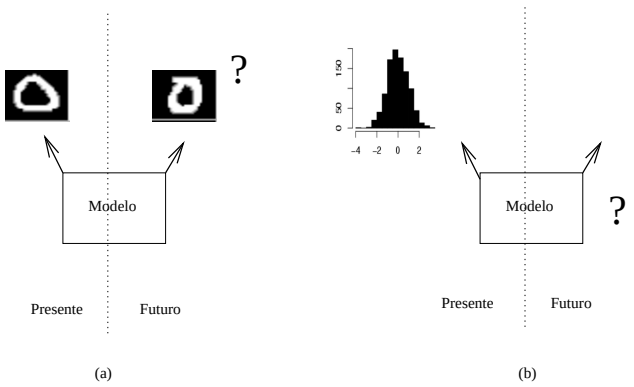
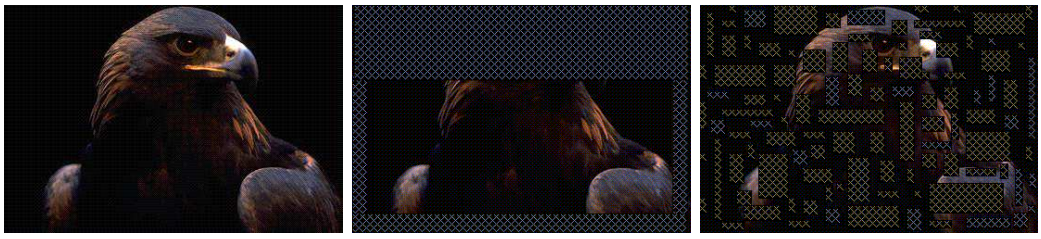


Figura 2

Aunque los dos tipos de problemas no son totalmente desconectados (ambos suponen un modelo subyacente) en este capítulo nos enfocamos al segundo y que es conocido como *inferencia estadística*.

Otra manera para describir la idea principal de la inferencia está resumido en las tres imágenes de la Figura 3. La población de la cual tenemos solamente una parte observada es representada por la imagen (a). Un conjunto de observaciones corresponde a una parte (ventana) de esta imagen (ver (b) y (c)). El problema es estimar (a) a través de esta parte. Es claro que observando solamente una parte, nunca podemos estar seguros de lo que la imagen representará. La probabilidad nos permitirá formular y cuantificar de manera correcta este incertidumbre.

Por otro lado, comparando (b) y (c), es evidente que el elegir las observaciones al azar (lo llamaremos más adelante tomar una muestra), ayuda mucho mas en formarnos una idea de lo que está presentado que hacerlo de manera sistemática.



(a) (b) (c)

Figura 3.

8.2 Estimación estadística paramétrica

En estimación paramétrica, supongamos que los datos provienen de una distribución $P_\theta(\cdot)$, donde θ son parámetros desconocidos y supongamos que las preguntas de interés son traducibles en terminos de estos parámetros.

Un primer aspecto importante es el hecho que tenemos sólamente una muestra a nuestra disposición, implica que a través de esta muestra particular, nunca será posible deducir con toda la certeza el verdadero valor de θ, ni será posible obtener una cota superior de la diferencia entre el verdadero valor y el estimado. Para ilustrar lo anterior, tomamos dos muestras de la misma distribución normal con varianza 1 y promedio desconocido. En la Figura 7, las dos muestras están marcadas con $\circ$ y $\bullet$. Por el azar, las dos muestras mostradas son de una forma muy diferente. En consecuencia, también las dos estimadores correspondientes serán probablemente muy diferente. El usuario nunca puede distinguir de antemano en cual caso esté y así no se puede cuantificar la calidad de un estimación en particular.

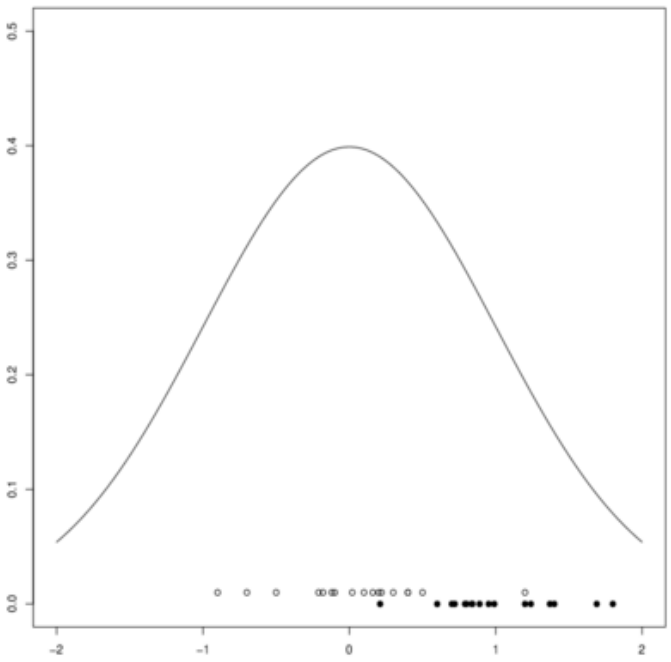


Figura 4.

Una salida para este problema es no estudiar las propiedades de una estimación particular, sino, usar el hecho que un estimación de θ será siempre alguna función de las observaciones. Como las observaciones son variables aleatorias con cierta distribución, a su vez podemos considerar una estimación como una variable aleatoria con una propia distribución. Para ese fin, useremos la notación $\hat{\Theta}(\mathbf{X})$ para denotar la función que mapea una muestra particular a una estimación para θ.

De esta manera, podemos formular algunas características deseables de $\hat{\Theta}(\mathbf{X})$:

1. *ser insesgado*: llamamos $\hat{\Theta}(\mathbf{X})$ insesgado si

$$E\hat{\Theta}(\mathbf{X}) = \theta.$$

2. *tener varianza mínima*: en caso de ser insesgada, $var(\hat{\Theta}(\mathbf{X})) = E(\hat{\Theta}(\mathbf{X}) - \theta)^2$ y mide la variabilidad del estimador con respecto al verdadero valor. Entre menor la variabilidad, mejor.

3. *ser consistente*: llamamos $\hat{\Theta}(\mathbf{X})$ consistente si el estimador converge al verdadero valor θ cuando el tamaño de la muestra crece a infinito.

Otro aspecto es que no es siempre evidente encontrar la familia de distribuciones $P_\theta(\cdot)$. Equivocarse puede ser otro factor de error en la estimación.

Es instructivo resumir estos dos aspectos en el problema de la estimación en la siguiente figura:

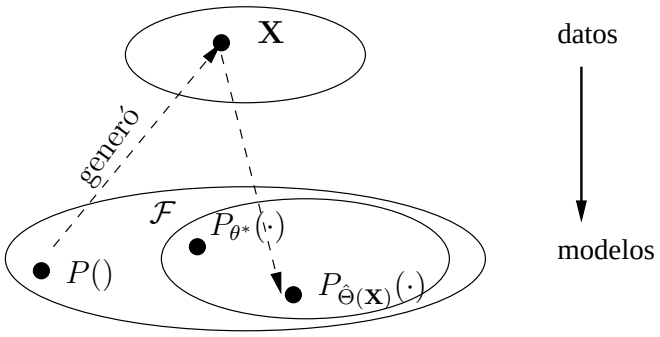


Figura 5.

Al nivel superior, se encuentran los datos observados donde se puede buscar algún patrón. Abajo se encuentra el espacio donde vive la distribución, $P()$, que generó estos datos y que puede pero no debe pertenecer a una cierta clase $\mathcal{F}$.

Basado en nuestras datos $\mathbf{X}$, buscamos $P_{\hat{\Theta}(\mathbf{X})}() \in \mathcal{F}$. La distancia o mejor dicho el error entre $P()$ y $P_{\hat{\Theta}(\mathbf{X})}()$ se divide en dos partes: lo que se puede atribuir al hecho que conocemos solamente un conjunto finito de datos y lo que se puede atribuir a la clase $\mathcal{F}$ que en general contiene una familia restringida de distribucione, que denotamos en la figura con la distancia entre $P_\theta()$ y $P_{\hat{\Theta}(\mathbf{X})}()$ por un lado y $P()$ y $P_\theta()$ por otro lado.

A continuación, derivamos un método particular para construir $P_{\hat{\Theta}(\mathbf{X})}()$.

8.2.1 Estimación de máxima verosimilitud para el caso paramétrico

Para una distribución y una observación, llamamos la verosimilitud la probabilidad de observar esta observación. Si estamos dispuesto de aceptar que:

1. la verosimilitud resume toda la información relevante en $\mathbf{X}$ sobre el modelo;
2. lo que observamos es más probable algo representativo (ocurre con alta frecuencia) que una excepción;

es razonable tomar un valor de θ como estimador para θ que maximiza la verosimilitud, es decir un valor que da mayor probabilidad a observar los datos recibidos.

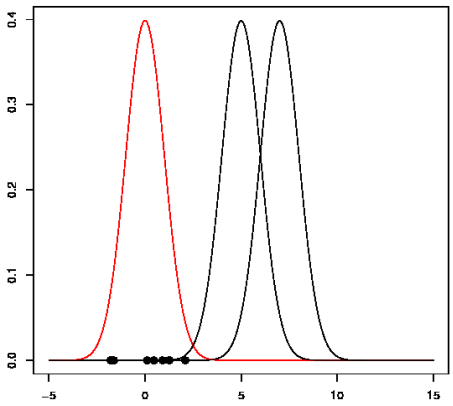
Hablamos de la verosimilitud en lugar de la probabilidad porque fijamos ahora la observación y consideramos diferentes distribuciones (definidas por diferentes valores de θ); lo anterior es la explicación de la frase: probabilidades suman uno; verosimilitudes no.

Definición 8.2.1 *El estimador de máxima verosimilitud, $\hat{\Theta}(X)$, es un valor que para X dada, maximiza la verosimilitud $P_\theta(\mathbf{X})$. En el caso de observaciones independientes, lo anterior es equivalente a maximizar con respecto a θ:*

$$\prod_i P_\theta(X_i = x_i), \tag{8.1}$$

en caso de que el máximo exista.

Usando el principio de maximizar la verosimilitud, para los siguientes datos, entre las tres densidades dibujadas, la primera es la de mayor verosimilitud.



Tomamos el siguiente ejemplo numérico.

Ejemplo 8.2.1 Supongamos que tenemos las siguientes observaciones independientes de una distribución normal con varianza 1: La función de verosimilitud es:

$$P(X_1 = x_1, \cdots, X_n = x_n) = \prod_i P(X_i = x_i) = \prod_i \frac{1}{\sqrt{\pi}} \exp(-\frac{(x_i - \theta)^2}{2}) =$$

$$\frac{1}{(\sqrt{\pi}}^n \exp(-\frac{\sum_i (x_i - \theta)^2}{2}) \tag{8.2}$$

Dado que el logaritmo es una función que conserva el orden (i.e. si $x < y$, entonces $log(x) < log(y)$), podemos tomar en (8.2) el logaritmo, y calcular θ que maximice:

$$-n \log \sqrt{\pi} - \frac{\sum_i (x_i - \theta)^2}{2}.$$

Derivar con respecto a θ e igualando a 0, resulta en:

$$\hat{\theta} = \frac{\sum x_i}{n},$$

y en general:

$$\hat{\Theta}(X) = \frac{\sum X_i}{n}.$$

Bajo algunas condiciones de regularización se puede mostrar que los estimadores de máxima verosimilitud son consistentes, asinptóticamente insesgados y tienen mínima varianza.

Para el ejemplo anterior obtenemos las siguientes propiedades:

Ejemplo 8.2.2 *(continuación)*: Usando la indepencia de los X_i y el hecho que $EX_i = \theta$, obtenemos que el estimador es insesgado porque

$$E\hat{\Theta}(X) = E\frac{\sum X_i}{n} = \frac{n\theta}{n} = \theta.$$

La varianza es

$$Var(\hat{\Theta}(X)) = Var(\frac{\sum X_i}{n}) = \frac{\sum Var(X_i)}{n^2} = \frac{1}{n}.$$

Así, por ejemplo basarnos en solamente $\tilde{n}$ de las n observaciones ($\tilde{n} < n$), resulta en un estimador (insesgado) con mayor varianza.

La consistencia es una consecuencia de la ley debil de los números grandes.

El estimador de máxima verosimilitud nısiempre existe, y si existe, no siempre es único. A continuación damos un ejemplo:

Ejemplo 8.2.3 Supongamos que $X \sim \mathcal{U}(\theta, \theta+1)$. La verosimilitud es:

$$P_\theta(\mathbf{X}) = \begin{cases} (\frac{1}{(\theta+1)-\theta})^n = \frac{1}{1} & \text{si } \theta < \min\{X_i\} \text{ y } \max\{X_i\} < \theta+1 \\ 0 & \text{otro caso} \end{cases}$$

Todos las θ's que satisfagan $\theta < \min\{X_i\}$ y $\max\{X_i\} < \theta+1$, conducen al mismo valor de la verosimilitud y pueden ser tomado como estimador de máxima verosimilitud.

Un último ejemplo es la estimación de la varianza de una distribucón normal donde el estimador será sesgado (para una muestra finita).

Ejemplo 8.2.4 Supongamos que los datos $\mathbf{X}$ son independientes de una distribución normal $\mathcal{N}(\theta_1, \theta_2)$. Para calcular los estimadores de máxima verosimilitud para (θ_1, θ_2), maximizamos:

$$P(X_1 = x_1, \cdots, X_n = x_n) = \prod_i P(X_i = x_i) = \prod_i \frac{1}{\sqrt{\pi\theta_2}} \exp(-\frac{(x_i-\theta_1)^2}{2\theta_2}) \tag{8.3}$$

$$= (\frac{1}{\sqrt{\pi\theta_2}})^n \exp(-\frac{\sum_i (x_i-\theta)^2}{2\theta_2})$$

Tomando el logaritmo, se obtiene:

$$-\frac{n}{2}\log(\pi\theta_2) - \frac{\sum_i (x_i-\theta_1)^2}{2\theta_2}.$$

Derivando con respecto a θ_1 y θ_2 e igualando a 0, obtenemos respecitivamente:

$$2\frac{\sum_i (x_i-\theta_1)^2}{2\theta_2} = 0$$

y

$$-\frac{n}{2\theta_2} + \frac{\sum_i (x_i-\theta_1)^2}{2\theta_2^2} = 0.$$

Resolviendo este sistema obtenemos:

$$\hat{\theta}_1 = \frac{\sum x_i}{n} \text{ y } \hat{\theta}_2 = \frac{\sum (x_i-\hat{\theta}_1)^2}{n}$$

Un poco cálculo, conduce a

$$E(\hat{\Theta}_2(\mathbf{X})) = \frac{n-1}{n}\theta_2,$$

es decir el estimador para la varianza es sesgado. Por lo anterior se prefiere más bien trabajar con el estimador

$$\hat{\Theta}_2(\mathbf{X})) = \frac{\sum (x_i-\theta_1)^2}{n-1}.$$

Aplicación 1: rotor

Hasta la segunda guerra mundial, el uso de una máquina rotor fue un método muy popular para codificar mensaje de una manera (electro-)mecánica. Podemos sistematizar el funcionamiento de un rotor como en figura 6. A lado izquierda se presenten las letras consecutivas del mensaje secreto por codificar y donde a lado izquierda se genera como salida el mensaje codificado. Cada letra (caracter), l es mapeado a través de la función $\theta()$ a una letra, $\theta(l)$. En la práctica el mapeo es relizado por cilindros rotantes. Simplificamos el problema suponiendo que la función $\theta()$ es fija.

Ilustramos la decodificación del rotor, i.e. estimar la función $\theta()$, usando una formulación de máxima verosimilitud. Supongamos que sabemos las estadísticas de ocurrencia de cada letra en el idioma donde fue escrito el mensaje original. Para simplificar la notación supongamos en lo que sigue, que cada letra está codificada con su número de secuencia del alfabeto (1 representa a "a'', 2 a "b'', etc.). Dada que la función $\theta()$ es complétamente descrita por los parámetros $\theta_1, \cdots, \theta_{25}$ donde $\theta_i = k$ denota que $\theta(i) = k$, el problema de la decodificación se convierte en un problema de la estimación de los parámetros $\{\theta_i\}$.

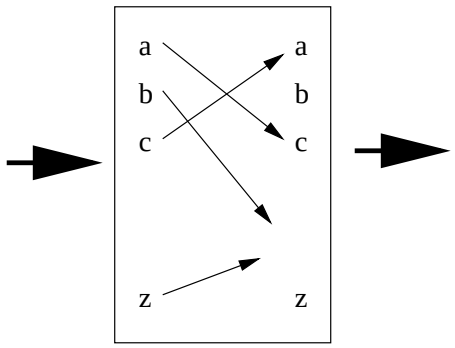


Figura 6.

Denotamos con p_i la probabilidad que una letra elegida al azar es i, y con X el output del rotor para un input elegida al azar según las probabilidades $\{p_i\}$. La probabilidad de observar la letra k es:

$$P(X = k) = p_l \text{ donde } \theta_l = k.$$

Baja la suposición (poca realista) que las letras (y en consecuencia el output correspondiente) son independientes entre sí,

$$P(X_1 = x_1, \cdots, X_n = x_n) = \prod_k p_k^{n(\theta(k))}, \tag{8.4}$$

donde $n(k)$, denota el número de letras tipo k encontradas en el output $\{x_1, \cdots, x_n\}$.

Siguiendo el principio de máxima verosimilitud, buscamos estos valores para $\{\theta_l\}$ que maximicen (8.4), bajo la restricción que definan una permutación.

8.3 Pruebas de Hipótesis

A continuación denotamos con H_0 y H_1 dos suposiciones disjuntas acerca de $P()$ y creemos que uno de las dos debe ser cierta (un caso extremo es cuando H_1 es la negación de H_0).

Tomamos el siguiente ejemplo.

Ejemplo 8.3.1 Un proceso de fabricación de chips, genera en 20 por ciento de los casos un chip defectuoso. Se pretende que implementando una modificación se pueda obtener una mejora donde solamente 10 por ciento salgan defectuosos. Para una muestra de 50 chips con el nuevo procedimiento, obtuvimos los siguientes datos (D denota que el chip fue defectuoso y C indica la ausencia de un defecto):

C,D,C,C,C,C,C,C,C,C,D,C,C,C,C,C,C,C,C,C,C,C,D,C,C,C,C,C,C,C,C,C,D,C,
C,C,C,C,C,C,C,C,C,D,D,C,C,C,C,C,C,C

Así, si X denota la presencia (1) o ausencia (0) de un defecto para un chip elegido al azar, $X \sim Bern(\theta)$ y

$$H_0 : \theta = 0.2 \text{ vs } H_1 : \theta = 0.1. \tag{8.5}$$

En general, las hipótesis están denotadas como:

$$H_0 : \theta \in S_0 \text{ vs } H_1 : \theta \in S_1.$$

con caso especial:

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta = \theta_1. \tag{8.6}$$

Igual al caso de estimación, nunca vamos a poder determinar si $P()$ es elemento de un cierto conjunto o no. Más bien, usamos de nuevo la probabilidad para formular una respuesta - aunque con incertidumbre - que es matemáticamente correcta.

Así al tomar una decisón cual de las dos hipótesis es cierta, se pueden asociar dos errores:

1. type I: rechazar H_0 si H_0 es cierto
2. type II: no rechazar H_0 si H_0 no es cierto

8.3.1 Enfoque basado en verosimilitud

Supongamos primero que la hipótesis es de la forma (8.6). Por consecuencia, tanto bajo H_0 como H_1, la verosimilitud es completamente descrita.

Este primer enfoque se basa exclusivamente en la razón de la verosimilitud:

$$R(\theta_0; \theta_1; \mathbf{X}) = \frac{P_{\theta_0}(\mathbf{X})}{P_{\theta_1}(\mathbf{X})}.$$

Para los datos anteriores, obtenemos:

$$R(0.2; 0.1; \mathbf{X}) = \frac{(1-0.2) * 0.2 * (1-0.2) * (1-0.2) * \cdots}{(1-0.1) * 0.1 * (1-0.1) * (1-0.1) * \cdots} = \frac{(1-0.2)^{44} 0.2^6}{(1-0.1)^{44} 0.1^6} = 0.36$$

La variable $R(\theta_0; \theta_1; \mathbf{X})$ expresa cuantas veces H_0 es más probable que H_1 (por supuesto, puede ser menor que 1!). Dado que el promedio de una distribucíon geométrica con parámetro θ es igual a $1/\theta$, podemos al mismo tiempo interpretarla como cuantas veces más en promedio tenemos que esperar hasta observar $\mathbf{X}$ bajo H_1 comparado con bajo H_0.

En la práctica, el usuario determina un número natural k. Si $R(\theta_0; \theta_1; \mathbf{X}) > k$, se concluye que hay más *evidencia* en favor de H_0 que en favor de H_1, mientras $R(\theta_0; \theta_1 \mathbf{X}) < \frac{1}{k}$ o equivalente $R(\theta_1; \theta_0; \mathbf{X}) > k$ da más *evidencia* en favor de H_1. No se toma ninguna decisión si $\frac{1}{k} < R(\theta_0; \theta_1; \mathbf{X}) < k$.

Se muestra en apendice que bajo H_1, la probabilidad que R *por casualidad* es mayor que k, es acotado por $\frac{1}{k}$ i.e.

$$P_{\theta_1}(R(\theta_0; \theta_1; \mathbf{X}) > k) \leq \frac{1}{k}.$$

Así, la probabilidad de un error de tipo I o II es acotado por $\frac{1}{k}$.

Ejemplo 8.3.2 El determinar el autor de un texto en baso de características comparativas de estilo, siempre ha sido una aplicación llamativa de la estadística. Un ejemplo histórico son los *Federal Papers*. Sin embargo, no es un enfoque sin problemas: es crucial y muy dificil el determinar de las características comparativas por usar.

Tomamos el siguiente ejemplo didáctico. Se quiere determinar cual de dos programadores escribió un cierto programa. De cada uno se tiene muchos lineas de código de mano conocida. Con ese fin, se define la variable L, la longitud de los nombres de las variables definidas en los programas.

Supongamos que para el primer programador, la distribución de L, $P_{\theta_0}()$ es:

	1	2	3	4	5	6	7	> 7
$P_{\theta_0}(L =)$	0.2	0.15	0.15	0.15	0.1	0.05	0.1	0.1

y para el segundo programador, $P_{\theta_1}()$:

	1	2	3	4	5	6	7	> 7
$P_{\theta_1}(L =)$	0.05	0.05	0.1	0.1	0.05	0.05	0.2	0.4

Tomamos ahora un programa cuyo autor es desconocido y observamos las siguientes frecuencias (total es 1000):

	1	2	3	4	5	6	7	> 7
frecuencia	150	150	100	100	150	100	100	150

Así,

$$R(\theta_0, \theta_1, \mathbf{X}) = \frac{0.2^{150} * 0.15^{150} * 0.15^{100} \cdots}{0.05^{150} * 0.05^{150} * 0.1^{100} \cdots} = \exp(2406 - 2126).$$

En baso de lo cual podemos concluir que hay más evidencia que el programa fue escrito por el primer programador.

La gran ventaja de este método es que no se usa algo más que los datos. Uno de las desventajas es resumido con el siguiente ejemplo: tomamos al azar una carta de juego. Si tomamos como H_0 que todas las cartas son iguales al as de corazones, y H_1 la hipótesis que tenemos un juego normal, es fácil ver que si la carta elegida es un as de corazones, $R = 52$. Sin embargo, pocas estarán dispuesto de acceptar H_0.

8.3.2 Enfoque basado en repitibilidad

Retomamos ejemplo 8.3.1. Se puede atribuir el hecho que el porcentaje de elementos defectuosas, $\frac{\sum_i x_i}{n} = \bar{x}$ es diferente de 0.2, a dos factores:

1. un elemento sistemático ($\theta \neq 0.2$)
2. un elemento aleatorio en la muestra por el hecho que n es pequeña (finita).

La idea fundamental en este enfoque es verificar si el elemento aleatorio del muestreo puede explicar esta diferencia bajo el modelo H_0.

Tomamos primero el caso:

$$H_0 : \theta = 0.2 \qquad H_1 : \theta \neq 0.2 \tag{8.7}$$

Intuitívamente, se quiere no rechazar H_0 si

$$|\bar{x} - 0.2|$$

no es demasiado grande.

Determinar que es grande o chico no es una *sine cura*: interviene el tamaño de la muestra y la distribución de X. Aquí se lo resuelve expresando la distancia en la escala de probabilidad (acotado por uno y cero), es decir se calcula bajo H_0,

$$p = P(|\bar{X} - 0.2| \geq |\bar{x} - 0.2|), \tag{8.8}$$

la probabilidad de obtener lo que se observó o algo peor. Se llama *p el valor de p.*

Bajo H_0 sabemos que $\bar{X} \sim Bin(0.2, n)$. En general, si n es suficientemente grande, se aproxima la distribución binomial por una normal. Bajo H_0, y n suficientemente grande: $\bar{X} \sim \mathcal{N}(0.2, \sigma^2/n)$, con $\sigma^2 = 0.2 * (1 - 0.2)$.

Por consecuencia, si definimos $Z = \frac{\bar{X}-0.2}{\sigma/\sqrt{n}}$ sabemos que bajo H_0,$Z \sim \mathcal{N}(0,1)$, tal que (8.8) es igual a :

$$P(|Z| > \frac{|\bar{x}-0.2|}{\sigma/\sqrt{n}}) = P(Z < -\frac{|\bar{x}-0.2|}{\sigma/\sqrt{n}}) + P(Z > \frac{|\bar{x}-0.2|}{\sigma/\sqrt{n}}). \tag{8.9}$$

En general se llama Z *la estadística de prueba* y debe tener una distribucíon totalmente conocida bajo H_0 para que se pueda calcular (8.9). Por otro lado debe poder capturar de manera eficiente las diferencias entre H_0 y H_1.

Para el caso:

$$H_0 : \theta = 0.2 \qquad H_1 : \theta = 0.1 \tag{8.10}$$

usamos la misma estadística de prueba pero en lugar de (8.8), tomamos:

$$p = P(\bar{X} - 0.2 < \bar{x} - 0.2); \tag{8.11}$$

si esta probabilidad es demasiado chica, aporta soporte en favor de H_1.

Formulación de la conclusión

Dependiendo del enfoque que uno prefiere, hay dos maneras para formular una conclusión.

1. *Como una decisíon*

   Se pide antes de hacer el análisis al usuario el nivel de significancia α, i.e. la probabilidad de un error de tipo I que el está dispuesto de acceptar.
   Si el *valor de p* es menor que α, se rechaza H_0, en el otro caso, se accepta H_0, o formulado de una manera completa, " a nivel de significancia α se accepta (o rechaza) H_0''.

   Lo anterior es el enfoque clásico de pruebas de hipótesis, basada en repitibilidad, donde uno se forza a tomar una decision (H_0 ó H_1) midiendo la calidad de la decisón con una visión de plazo largo (long run): α representa el número de veces de rechazar incorrectamente H_0 bajo aplicación repetida del procedimiento.

   Para el caso anterior se puede calcular explícitamente la probabilidad de un error de tipo II. Dado que

$$Z = \frac{\bar{X}-0.2}{\sqrt{0.2*(1-0.2)}/\sqrt{n}} = \frac{\sqrt{0.1*(1-0.1)}}{\sqrt{0.2*(1-0.2)}}(\frac{\bar{X}-0.1}{\sqrt{0.1*(1-0.1)}/\sqrt{n}} + \frac{0.1-0.2}{\sqrt{0.1*(1-0.1)}/\sqrt{n}}),$$

   vemos que bajo H_1, Z tiene una distribución normal y se puede calcular la probabilidad que Z toma un valor en el area de aceptación de H_0.

   En general, cuando H_1 no especifica completamente la distribución (por ejemplo: $H_1 : \theta < 0.2$), no se puede calcular explícitamente la probabilidad de un error de

tipo II. Se elige la estadística de prueba y el área de rechazo tal que el error sea mínima para un nivel de significancia dada pero sin tener control sobre el valor mismo. La derivación es bastante teórica. En Tabla I y II resumimos algunos casos.

Dado que solamente se tiene el control explícito del error de tipo I, la decisión cual será H_0 y H_1 debe ser tal que el error más grave para el usuario coincida con el error de tipo I.

2. *como medida de evidencia*

   Se comunica al usuario *el valor de p* como medida de la evidencia que hay en favor de H_0.

H_0	H_1	est. de prueba	area de acept.	distr. bajo H_0
$p = p_0$	$p \neq p_0$	$\frac{\bar{X}-p_0}{\sqrt{p_0(1-p_0)/n}}$	$[-z_{\alpha/2}, z_{\alpha/2}]$	$\mathcal{N}(0,1)$ (para n grande)
$p = p_0$	$p > p_0$	$\frac{\bar{X}-p_0}{\sqrt{p_0(1-p_0)/n}}$	$[-\infty, z_\alpha]$	$\mathcal{N}(0,1)$ (para n grande)
$\mu = \mu_0$, σ^2 conocida	$\mu \neq \mu_0$	$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$	$[-z_{\alpha/2}, z_{\alpha/2}]$	$\mathcal{N}(0,1)$
$\mu = \mu_0$, σ^2 conocida	$\mu > \mu_0$	$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$	$[-\infty, z_\alpha]$	$\mathcal{N}(0,1)$
$\mu = \mu_0$, σ^2 desconocida:	$\mu \neq \mu_0$	$\frac{\bar{X}-\mu_0}{S/\sqrt{n}}$	$[-t_{\alpha/2}, t_{\alpha/2}]$	t_{n-1}
$\mu = \mu_0$, σ^2 desconocida:	$\mu > \mu_0$	$\frac{\bar{X}-\mu_0}{S/\sqrt{n}}$	$[-\infty, t_\alpha]$	t_{n-1}
$\sigma = \sigma_0$	$\sigma \neq \sigma_0$	$\frac{(n-1)S^2}{\sigma}$	$[\chi_{1-\alpha/2}, \chi_{\alpha/2}]$	χ_{n-1}
$\sigma = \sigma_0$	$\sigma > \sigma_0$	$\frac{(n-1)S^2}{\sigma}$	$[0, \chi_\alpha]$	χ_{n-1}
$X \overset{\mathcal{D}}{=} X_0$, con $X \in A$, $\#A < \infty$	$X \overset{\mathcal{D}}{\neq} X_0$	$\sum_{i\in A} \frac{(0_i - nP(X_0=i))^2}{nP(X_0=i)}$	$[0, \chi_\alpha]$	$\chi_{\#A-1-p}$

donde $S^2 = \frac{\sum_i (x_i-\bar{x})^2}{n-1}$, O_i es el número de observaciones igual a i y p el número de parámetros por estimar bajo H_0.

Tabla 1

H_0	H_1	est. de prueba	area de acept.	distr. bajo H_0
$\mu_1 = \mu_2$, var. σ_1^2, σ_1^2 conocidas	$\mu \neq \mu_0$	$\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$[-z_{\alpha/2}, z_{\alpha/2}]$	$\mathcal{N}(0,1)$
$\mu_1 = \mu_2$, var. σ_1^2, σ_1^2 conocidas	$\mu > \mu_0$	$\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$[-\infty, z_\alpha]$	$\mathcal{N}(0,1)$
$\mu_1 = \mu_2$, var. $\sigma_1^2 = \sigma_1^2$ desconocidas	$\mu \neq \mu_0$	$\frac{\bar{X}_1-\bar{X}_2}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$[-t_{\alpha/2}, t_{\alpha/2}]$	$t_{n_1+n_2-2}$
$\mu_1 = \mu_2$, var. $\sigma_1^2 = \sigma_1^2$ desconocidas	$\mu > \mu_0$	$\frac{\bar{X}_1-\bar{X}_2}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$[-\infty, t_\alpha]$	$t_{n_1+n_2-2}$
$\sigma_1^2 = \sigma_1^2$	$\sigma_1^2 \neq \sigma_1^2$	$\frac{S_1-S_2}{S_p\sqrt{\frac{1}{2n_1}+\frac{1}{2n_2}}}$	$[-z_{\alpha/2}, z_{\alpha/2}]$	$\mathcal{N}(0,1)$

donde n_1, n_2 son los tamaños de las dos poblaciones (independientes); S_1, S_2 y S_p son estimadores de la desviación estandar en la primera población, la segunda población y en ambas respectivamente.

Tabla 2

8.3.3 Pruebas de hipótesis noparamétricas

Todas las pruebas anteriores supusieron normalidad de los datos (o al menos aproximadamente). A continuación damos - con fines ilustrativos - un ejemplo de una prueba

de hipótesis que no supone una forma paramétrica particular de la distribución de la muestra (por eso la llaman pruebas noparamétricas).

Supongamos que tenemos dos muestras independientes de variables continuas X y Y de tamaño n_1 resp. n_2; queremos verificar si vienen de la misma población versus la hipotesis que la observaciones de X suelen ser mayores (resp. menores, mayores o menores) que las de Y.

Con ese fin, definimos:

$$U_X = \#\text{ de parejas } (X_i, Y_j) : X_i < Y_j$$
$$U_Y = \#\text{ de parejas } (X_i, Y_j) : X_i > Y_j$$
$$U = \min(U_X, U_Y)$$

Bajo H_0, $P(X_i > Y_j) = 1/2$, entonces el promedio de U_X y U_Y es igual a $\frac{n_1 n_2}{2}$. Una desviación demasiado grande de este promedio puede aportar evidencia para rechazar H_0.

Se puede mostrar que si n es suficientemente grande, la distribución de la estadística de prueba es aproximadamente normal con $\mu = \frac{n_1 n_2}{2}$ y $\sigma = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$.

Para n pequeño se puede recurrir a pruebas exactas. Supongamos $n_1 = 2$, $n_2 = 3$; las posibles combinaciones son:

orden	U_X	U_Y
yyyxx	6	0
yyxyx	5	1
yyxxy	4	2
yxyyx	4	2
yxyxy	3	3
yxxyy	2	4
xyyyx	3	3
xyyxy	2	4
xyxyy	1	5
xxyyy	0	6

Bajo H_0 todas las realizaciones tienen igual probabilidad 1/10, tal que por ejemplo $P(U_X > 4) = 1/5$.

La prueba anterior es conocida como la estadística de *Mann-Whitney*. Es intrínsicamente relacionada con la estadística de *Wilcoxon* que a continuación describimos.

Definimos para cada observación de las dos poblaciones su rango con respecto a la unión de las dos poblaciones. Sea R_X resp. R_Y la suma de los rangos de los elementos de X resp. Y. Se puede mostrar que:

$$U_Y = n_1 * n_2 + \frac{n_1(n_1+1)}{2} - R_X.$$

Extensiones existen que incluyen distribuciones discretas etc.

8.3.4 Pruebas de randomización

Pruebas de randomización forman parte de una familia cada vez más popular de métodos de computación intensiva. Tomamos una vez más un ejemplo como punto de partida.

Supongamos que los tiempos de ejecución de programa I son:

$$4.1, 3.9, 3.5, 5.4, 3.7, 2.5, 2.9,$$

los de un programa alternativa:

$$3.9, 3.8, 3.5, 4.7, 7.2, 6.2, 4.0, 4.1$$

Se quiere probar si los promedios son iguales o no. Se puede calcular $z = |\bar{x}_1 - \bar{x}_2|$. Se toman al azar reordenamientos, r_k, de estos 15 datos juntos; para cada reordenamiento se consideran los primeros 7 como del grupo I, los demás como del grupo II y se calcula $z^k = |\bar{x}_1^k - \bar{x}_2^k|$.
Se estima el valor de p como el porcentaje de valores z^k mayor que z.

Sin duda la ventaja de este método es su generalidad (se puede considerar datos nominales mixtos con métricos, etc). Sin embargo por falta de mucho soporte teórico para poder controlar el error de tipo II, se debe considerar más bien como un último refugio cuando métodos clásicos fallan.