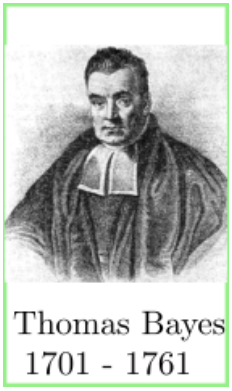
Mecánica Estadística. Clase 2. Entropía Estadística.

Prof. Juan Mauricio Matera

06/03/2025

Clase 2. Distribuciones de Probabilidad. Correlaciones. Entropía Estadística.

Probabilidad condicional e independencia



En general, interesa determinar la probabilidad de un evento A dado que se observó otro evento B. Se introduce para eso la noción de *Probabilidad condicional*

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

De la definición se sigue que

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

que se conoce como regla de Bayes. En particular, si $P(B|A) = P(B)$ decimos que B es *independiente* de A, y por lo tanto la información acerca de B no influye en la probabilidad de A y viceversa. Por ejemplo, podemos esperar el color de ojos de un individuo sea una característica independiente a la velocidad que alcanza al correr una maratón, su performance académica, o las probabilidades de obtener 7 en dos tiradas sucesivas de un par de dados:

$$P(\{\text{ojos marrones}\} \cap \{7 \text{ en los dados}\}) = P(\{\text{ojos marrones}\})P(\{7 \text{ en los dados}\})$$

El límite opuesto corresponde a variables que tienen una dependencia funcional entre sí. Por ejemplo, si sabemos que los valores de dos dados suman 7, y conocemos el valor de uno, el otro queda determinado:

$$P(\{x_1 = n_1\} \cap \{x_2 = n_2\}) = \begin{cases} P(\{x_1=n_1\}) & n_1+n_2=7 \\ 0 & n_1+n_2 \neq 7 \end{cases}$$

En este caso, como hay una relación *biyectiva*, $P(A) = P(B)$ y por lo tanto, $P(A|B) = P(B|A)$, pero $P(A|B) \neq P(A)P(B)$.

En un caso más general, podemos tener correlaciones pero no deterministas. Por ejemplo, consideremos un mazo de barajas españolas tiene 50 cartas, de las cuales 4 son ases. Si tomamos una carta y le otra cartas a otro jugador, la probabilidad de que el otro jugador tenga un as es diferente si sabemos que a nosotros nos tocó o no un as:

$$P(\{\text{tiene As}\} \cap \{\text{tengo As}\}) = \frac{6}{1225}$$

$$P(\{\text{tiene As}\}) = P(\{\text{tengo As}\}) = 4/50 = 0,08$$

$$P(\{\text{tiene As}\}|\{\text{tengo As}\}) = 3/49 = 0,0612245$$

(probar)

También podemos considerar un caso *no simétrico* $P(A|B) \neq P(B|A)$ si nos preguntamos cuál es la probabilidad de que el otro tenga un as si yo no lo tengo. La probabilidad de que ambas cosas ocurran es

$$P(\{\text{tiene As}\} \cap \{\text{no tengo As}\}) = \frac{92}{1225}$$

La probabilidad de que yo no tenga un as es de 46/50, luego

$$P(\{\text{tiene As}\}|\{\text{no tengo As}\}) = \frac{4}{49}$$

mientras que la probabilidad de que el otro tenga un as es de 4/50. Luego,

$$P(\{\text{no tengo As}\}|\{\text{tiene As}\}) = \frac{46}{49}$$

y por lo tanto $P(A|B) \neq P(B|A)$.

Nótese que la independencia es una propiedad no sólo de los conjuntos A y B, sino de la distribución de probabilidad. Sin embargo, los elementos del espacio muestral tienen que corresponder a eventos mutuamente excluyentes y por lo tanto, $P(\{\omega_i\}|\{\omega_j\}) = P(\varnothing) = 0$ si $i \neq j$.

Mezclas estadísticas

En ciertas ocasiones, podemos disponer de un modelo para las probabilidades condicionales $P(A_i|B_j)$, y la distribución de probabilidades para la condición $P(B_j)$, y nos interesa conocer la distribución $P(A_i)$. Si los eventos B_j son *mutuamente excluyentes* ($P(B_i \cap B_j) = 0$) podemos expresar $P(A_i)$ como

$$P(A_i) = \sum_j P(B_j)P(A_i|B_j)$$

Si pensamos a los $P(A_i|B_j) = P_j(A_i)$ como diferentes medidas de probabilidad, vemos que $P(A_i)$ es una *combinación lineal* de las $P_j(A_i)$ con *coeficientes no negativos* $q_j = P(B_j)$ o *combinación convexa*

$$P(A_i) = \sum_j p_j P_j(A_i)$$

Decimos entonces que $P(A_i)$ es una *mezcla estadística* de las medidas de probabilidad $P_j(A_i)$ con pesos p_j.

Si $P_j(A_j)$ es una medida de probabilidad discreta, esta queda completamente determinada por las probabilidades $p_k = P(\{e_k\})$ de los resultados individuales. Estas probabilidades se pueden expresar como las componentes de un *vector de probabilidades*. El vector de probabilidades de una mezcla estadística es entonces también una *combinación convexa* de los vectores de probabilidad de cada medida.

Las mezclas estadísticas permiten modelar situaciones en las que la distribución de probabilidades para los eventos dependen a su vez de otro evento, cuyo resultado es desconocido. Por ejemplo, si tenemos dos dados que lucen igual, uno de ellos perfectamente simétrico, con una distribución de probabilidades P_0 con frecuencias equiprobables, y el otro *cargado*, con una distribución de probabilidades P_c, y elegimos uno de ambos al azar, la probabilidad de obtener un cierto resultado $\omega = \{k\}$ será

$$P(\omega) = \frac{P_0(\omega) + P_c(\omega)}{2}$$

Si en lugar de tener dos dados, tuviesemos N dados, de los cuales m son *equiprobables*, y el resto tienen asociada la distribución P_c, la distribución de probabilidades para un resultado al tomar un dado de la muestra al azar sería

$$P(\omega) = \frac{mP_0(\omega) + (N-m)P_c(\omega)}{N}$$

Naturalmente, podemos considerar mezclas estadísticas más complejas, $P(\omega) = \sum p_i P_i(\omega)$ con $p_i > 0$ y $\sum_i p_i = 1$.

Por otro lado, toda distribución de probabilidad discreta p_i puede descomponerse como una *mezcla estadística* de distribuciones de probabilidad en las que un único elemento tiene probabilidad 1 (certeza):

$$P_j(\{e_i\}) = \delta_{ij},\ P(\omega) = \sum_{ij} p_i P_j(\omega)$$

Naturalmente, estas distribuciones de probabilidad no pueden *descomponerse*: decimos que son distribuciones *puras*.

Aproximación de Stirling para el factorial y la multinomial. Función Γ

Como vamos a necesitar trabajar frecuentemente con factoriales y multinomiales, conviene contar con una aproximación *analítica* para estas cantidades. Para ello, se introduce la *función* Γ

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

que, para valores enteros de x cumple

$$\Gamma(n+1) = n!$$

como puede verificarse por inducción, mediante integración por partes. La función Γ *interpola* al factorial para valores reales. Conviene ahora aproximar esta función en términos de *funciones elementales*. Para ello, vamos a estimar la integral en la *aproximación de saddle point*.

Esta aproximación consiste en expresar el integrando como el producto de una gaussiana y una serie de potencias:

$$f(t) = f(t_0) e^{-(t-t_0)^2/(2b^2)} (1 + (t-t_0)^3 \sum_{k=0}^{\infty} \frac{c_k}{(k+3)!} (t-t_0)^k)$$

El valor de t_0 se elige en el valor *máximo* de $f(t)$. En ese punto, la primera derivada se anula, la derivada segunda es negativa, y

$$f''(t_0) = -\frac{f(t_0)}{b^2}$$

El resto de los coeficientes c_k se eligen para las expansiones de Taylor en torno a t_0 coincidan. Si $f(t)$ tiene un máximo pronunciado alrededor de t_0, en la evaluación de la integral, los términos en c_k pueden ignorarse en primera aproximación. Además, si $t_0 \gg b$, podemos extender la región de integración a toda la recta real, de manera que

$$\int_0^\infty f(t) dt \approx \int_{-\infty}^{\infty} e^{-\frac{(t-t_0)^2}{2b^2}} dt = \sqrt{2\pi} b f(t_0)$$

donde se evaluó explícitamente la integral gaussiana.

Para el caso de $n!\Gamma(n+1)$, derivando el integrando $t^N e^{-t}$ encontramos que su máximo se encuentra cuando

$$(N/t - 1) t^{N-1} e^{-t} = 0 \Rightarrow t_0 = N$$

En ese punto, $f(N) = N^N e^{-N} = \left(\frac{N}{e}\right)^N$ y

$$f(t)''|_{t=N} = ((N/t-1)t^N e^{-t})'\big|_{t=N} = (-N/t^2 t^N e^{-t})\big|_{t=N} = -\frac{f(N)}{N}$$

y por lo tanto, $b = \sqrt{N}$. Finalmente,

$$N! = \Gamma(N+1) \approx \sqrt{2\pi N} \left(\frac{N}{e}\right)^N$$

que es la llamada *Aproximación de Stirling* para el factorial (y la función $\Gamma(x)$. La aproximación de Stirling es muy buena para $N \geq 9$, con un error menor al 1% del valor exacto.

N	$N!$	Stirling	Gosper
0	1	0	1.023
1	1	0.92	0.996
2	2	1.919	1.997
3	6	5.83	5.996
4	24	23.5	23.99

Una aproximación mejor consiste en modificar el argumento de la raíz, para dar cuenta del comportamiento para N pequeño. En particular, se observa que (Gosper, 1978)

$$N! = \Gamma(N+1) = \sqrt{2\pi(N+1/6)} \left(\frac{N}{e}\right)^N (1 + \mathcal{O}(1/N^2))$$

aproxima el valor correcto con un error menor al $0,5\%$ para todo N no negativo.

En todo caso, lo que se observa es que $N! \approx N^N$, por lo que crece más rápidamente que una exponencial. Por ello, conviene trabajar con su logaritmo $\ln(N!) \approx N \ln(N/e)$ que resulta en números más manejables.

Ahora que disponemos de una aproximación para el factorial, podemos construir una para la multinomial[1]:

$$\left(\begin{array}{c} L \\ m_1 ... m_s \end{array} \right) \approx \exp(L\ln(L/e) - \sum_{k=1}^{s} m_k \ln(m_k/e)) \frac{\sqrt{2\pi(L+1/6)}}{\prod_{k,m_k\neq 0}\sqrt{2\pi(m_k+1/6)}}$$

Como $\sum_k m_s = L$, podemos reescribir el exponente en la exponencial como

$$\exp(L\ln(L/e) - \sum_{k=1}^{S} m_s \ln(m_s/e)) \approx e^{LS(\{m_k/L\})}$$

con

$$S(\{f_k\}) = \sum_k -f_k \ln(f_k)$$

una función de las *frecuencias relativas* $f_k = m_k/L$ conocida como *Entropía de Shannon* o *Entropía Estadística*. El factor restante puede incluirse en el exponente, y contribuye como $\mathcal{O}(\ln L)$ y en general puede despreciarse si $m_s > 0$.

Como es de esperarse, la entropía de Shannon es indiferente al orden de los f_k, se anula cuando una de las frecuencias relativas se hace = 1, y es máxima cuando todas las frecuencias son iguales:

$$S(\{1/s\}) = \sum_{k=1}^{s} -\frac{1}{s}\ln(1/s) = \ln(s)$$

Propiedades de la Entropía Estadística. Información.

La entropía estadística es una cantidad que será central en la dicusión de los temas de este curso, por lo que conviene discutir brevemente sus propiedades más importantes.

En primer lugar, es importante definirla en el límite en que alguna de las frecuencias relativas se anula. Definiendola por continuidad,

$$\lim_{s\to 0} \frac{\log(1/s)}{s} = 0$$

de manera que los términos con frecuencia 0 no contribuyen a la suma. De esta manera, para una distribución de probabilidad "pura" (determinista)

$$S(\{\delta_{ij}\}) = 0$$

que es el mínimo valor que puede tomar, ya que $1/p > 0$ y por lo tanto $S(f_i) \geq 0$.

Por otro lado, como dijimos antes, la entropía de una distribución uniforme es máxima, y toma el valor:

$$S(\{1/s\}) = \ln(s)$$

esto es, el logaritmo del número de estados accesibles.

Una forma alternativa de llegar a la expresión para la entropía consiste en definir la noción de *información* asociada a un evento. Un evento es más informativo cuando menos esperado resulte. Por ejemplo, si tenemos una bolsa de caramelos de frutilla, y alguien nos dice *saqué un caramelo, y era de frutilla* no aprendimos nada sobre el contenido de la bolsa. Por otro lado, si la bolsa tiene una mezcla de caramelos diferentes, el hecho de que el extraído se "de frutulla" será más informativo cuando mayor sea la variedad de sabores en la bolsa. Si definimos entonces que la medida de información $\mathcal{I}$ tal que

- Es una función *continua* de la probabilidad de un evento $I(A) \equiv I(P(A))$
- No negativa $I(A) \geq 0$
- Aditiva para eventos independientes $I(A \cap B) = I(A) + I(B)$ si $P(A \cap B) = P(A)P(B)$
- Se anula para eventos que se observan con certeza (I(1)=0)
- Escala: toma el valor 1 si $P(A) = P(\neg A) = 1/2$

entonces su forma funcional queda completamente determinada por

$$\mathcal{I}(A) = -\log_2(P(A))$$

Con esta definición, la entropía estadística para una distribución discreta se expresa como

$$S(P) = \langle I(\{\omega\})\rangle \times \log(2)$$

esto es, el valor medio de la información ganada al obtener un resultado[2]

[1] Los posibles factores en la productoria correspondientes a $m_k = 0$ se omiten ya que 0! = 1. En el exponente, por otro lado, como $-m\ln(m) \to 0$ para $m \to 0$ pueden mantener sin modificar la expresión.

[2] El factor log(2) se debe a la escala elegida para la medida de información: la información que ganamos al conocer la respuesta a una pregunta por sí o por no, a priori equiprobable representa un *bit* de información. Como con cualquier elección de unidades, podemos utilizar otras escalas tanto para la información como para la entropía estadística según convenga.

En general, la entropía es una cantidad *sub-aditiva*, lo que significa que si un espacio muestral es el producto de dos espacios $\Omega = \Omega_1 \times \Omega_2$, la entropía correspondiente a la distribución $p_{ij} = P(\{e_i \times e_j\})$ satisface

$$S(\{p_{ij}\}) \leq S(\{p_i\}) + S(\{p_j\})$$

con $p_i = \sum_j p_{ij}$ y $p_j = \sum_i p_{ij}$ los *marginales* de p_{ij}. La igualdad se satisface si $p_{ij} = p_i p_j$, esto es, cuando no existen *correlaciones* entre los eventos en Ω_1 y Ω_2. De forma análoga a la medida de *información*, la *entropía estadística* es la única función suave de las probabilidades que satisface esta propiedad (a menos de una constante multiplicativa).

Una consecuencia de la subaditividad es que la cantidad

$$\mathcal{I}(i;j) = S(\{p_i\}) + S(\{p_j\}) - S(\{p_{ij}\})$$

es una función no negativa que que se conoce como *Información mutua*. Esta cantidad da una medida de cuánta *información* podemos obtener de uno de los sistemas observando el otro.

Otra propiedad muy importante de la entropía estadística es su concavidad: para dos distribuciones f_1 y f_2, su mezcla estadística tiene una entropía mayor:

$$\begin{aligned} S(\{p_1 f_{1i} + p_2 f_{2i} +\}) &= \sum_i -(p_1 f_{1i} + p_2 f_{2i}) \log((p_1 f_{1i} + p_2 f_{2i})) && (1)\\ &= \sum_i (-p_1 f_{1i} \log(p_1 f_{1i} + p_2 f_{2i}) - p_2 f_{2i} \log(p_1 f_{1i} + p_2 f_{2i})) && (2)\\ &\geq \sum_i (-p_1 f_{1i} \log(p_1 f_{1i}) - p_2 f_{2i} \log(p_2 f_{2i})) && (3)\\ &= p_1 S(\{f_{1,i}\}) + p_2 S(\{f_{2,i}\}) && (4) \end{aligned}$$

Secuencias típicas

Siempre asumiendo un espacio muestral discreto, consideremos ahora la probabilidad de obtener una cierta distribución de frecuencias relativas en los resultados de experimentos independientes, si en cada uno de ellos la probabilidad $p_k = P(\{\omega_k\})$ de observar el elemento ω_k es la misma.

Para ello, notamos que la probabilidad de obtener una secuencia de resultados particular $R = \{x_1, ..., x_L\}$ viene dada por $P(\{R\}) = \prod_j p_j$. Sin embargo, la misma distribución de frecuencias relativas se obtiene para cualquier permutación de los elementos de R. La cantidad de formas de ordenar m_1 resultados ω_1, m_2 resultados ω_2, etc en L posiciones lo construimos eligiendo m_1 posisiones posibles entre L (el binomial $\binom{N}{m_1}$) multiplicado por la formas de elegir m_2 posiciones de entre las $L - m_1$ libres, y así sucesivamente:

$$|\{R/f_L(\omega_i|R) = m_i/L\}| = \binom{N}{m_1}\binom{N-m_1}{m_2}...\binom{m_s}{m_s} = \left(\begin{array}{c} N \\ m_1 ... m_s \end{array} \right)$$

esto es, el *multinomial* de N agrupados en $m_1, m_2, ..., m_s$. Multiplicando la probabilidad de obtener un resultado particular R con la distribución dada, por el número de formas de ordenar sus elementos obtenemos la probabilidad de observar la correspondiente distribución de frecuencias relativas $f_L \equiv f_L(\{\omega_i\}) = \frac{m_i}{L}$:

$$P(f_L) = \left(\begin{array}{c} L \\ Lf_1, ..., Lf_s \end{array} \right) \prod_k p^{Lf_k}$$

o, usando la aproximación de Stirling,

$$P(f_L) \approx e^{-LS(f_L \| p)}$$

con

$$S(f\|p) = -S(\{f_k\}) - \sum_k f_k \log(p_k)$$

la *Entropía relativa* entre las distribuciones f y p. Esta cantidad se anula si $f_i = p_k$, caso en el cual $P(f_L) \approx 1$. Esto puede verse de dos maneras: por un lado, si conocemos p_k, podemos esperar que si tomamos un número suficientemente grande de muestras, las frecuencias relativas observadas se parezcan a las probabilidades correspondientes. En el sentido opuesto, si las frecuencias observadas no se parecen a la distribución de probabilidad asumida, deberíamos *sorprendernos*: por ese motivo a la entropía relativa se la conoce también como *función sorpresa*. Si la *sorpresa* es muy grande, tenemos que asumir que la probabilidad propuesta $\{p_k\}$ no es la correcta.

Variables Aleatorias

El espacio muestral Ω no es necesariamente un espacio *numérico*: puede ser por ejemplo, un conjunto de personas, y los eventos pueden identificarse con ciertas características de esas personas. Para trabajar de forma *cuantitativa* vamos a definir la noción de *Variable Aleatoria*. Una *Variable Aleatoria* es una función $X : \Omega \to \xi$ con ξ algún conjunto numérico (típicamente, $\mathbb{R}$).

En el ejemplo de las personas, la variable aleatoria puede ser su edad, su presión sanguínea, o el saldo en su cuenta bancaria.

Para una variable aleatoria real, podemos definir una Distribución de Probabilidad Acumulada $F_X : \mathbb{R} \to \mathbb{R}$ según

$$F_X(x) = P(\{\omega / X(\omega) < x, \omega \in \Omega\})$$

que es una función cresciente. Podemos representar a $F_X(x)$ en términos de una *distribución* f_X

$$F_X(x) = \int_0^x f_X(t)dt$$

En cualquier caso, $\lim_{x\to\infty} F_X(x) = 1$.

Si f_X es una función *continua* de x, decimos que X es una variable aleatoria continua, y para representarla, necesitamos de un espacio muestral continuo. Si f_X es una función *constante*, entonces decimos que X está *uniformemente distribuída*. Notemos que para variables aleatorias continuas, la *probabilidad* de un evento particular $\omega \in \Omega \Rightarrow P(\{\omega\}) = 0$, por lo que no es posible reconstruir la distribución de probabilidades a partir de la probabilidad de elementos aislados, como hicimos en el caso discreto.

A una variable aleatoria la podemos describir en términos de ciertos parámetros numéricos como

- La *Esperanza* o *valor medio* $EX = \int f_X(t)tdt$. Da una estimación del promedio de valores obtenidos al realizar varias veces el experimento.
- La *Varianza* $var(X) = EX^2 - (EX)^2$ y la *desviación estándar* $\sigma(X) = \sqrt{var(X)}$, dan una estimación de cuánto se alejan típicamente los valores obtenidos en experimentos individuales respecto a valor medio.
- Los *momentos* $M_m(X) = EX^m = \int f_X(t)t^m dt, m = 0, 1, 2, \ldots$ y la *función característica* $\phi_X(t) = Ee^{\mathbf{i}tX}$ que en ciertos casos puede reconstruirse a partir de los M_m, caracterizan completamente a la distribución original.

La **estadística** trata de estimar estos parámetros a partir de muestras finitas de resultados.

Propiedades del valor medio

El valor medio de una distribución de probabilidades satisface las siguientes propiedades matemáticas:

- $\langle cX + a\rangle = c\langle X\rangle + a$
- $X \geq 0 \Rightarrow \langle X\rangle \geq 0$ Resulta de que $\langle X\rangle = \int_0^\infty xp(x)dx > 0$
- Desigualdad de Markov: si $X, c > 0$, $P(X > c) \leq EX/c$ Basta con probarla para $c = 1$:

$$P(X \geq 1) = \int_1^\infty p(x)dx \leq \int_0^1 xp(x)dx + \int_1^\infty xp(x)dx = \langle X\rangle$$

- Desigualdad de Chebychev: $P(|X - EX| \geq c) \leq \frac{var(X)}{c^2}$. Usando la anterior sobre $(X - \langle X\rangle)^2$,

$$P(|X - \langle X\rangle| \geq c) = P((X - \langle X\rangle)^2 \geq c^2) \leq \frac{\langle (X - \langle X\rangle)^2\rangle}{c^2} = \frac{var(X)}{c^2}$$

Modelo simple: El caminante aleatorio

Antes de continuar, vamos a ver cómo usamos lo que aprendimos para construir un modelo elemental que describa el experimento numérico de la clase anterior. La idea es asumir que cada vez que una partícula alcanza un nivel en el que se encuentran los dispersores, colisiona con el más cercano. En dicha colisión, pierde toda su energía ciñetica, y sigue cayendo por acción de la gravedad, en dirección de uno de los dos dispersores más cercanos en el nivel siguiente, con probabilidades iguales:

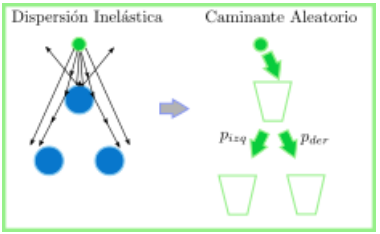


Figure 1: Modelos de dispersión inelástica y caminante al azar

Vemos que este modelo es infinitamente más simple que el original, y que introduce muchas hipótesis al menos *polémicas*: ¿por qué perdería toda la energía cinética? Si la trayectoria de la partícula al alejarse del dispersor está definida, ¿Por qué pensar que la velocidad final es al azar? ¿Por qué justo va a salir en la dirección de colisión con los dispersores

del siguiente nivel? En cualquier caso, lo que buscamos es una descripción *cualitativa* de la dinámica, y obviamente, perderemos información. Por ejemplo, cualquier correlación entre posición y velocidad se pierde en cada choque. Pero sigamos adelante.

Como resultado de los choques, el desplazamiento horizontal de la partícula respecto al eje de simetría del arreglo se expresará como $x_k = m_{der} - m_{izq} + x_0$, donde x_k es la posición horizontal al alcanzar el nivel $k-$ésimo de los dispersores, y m_{der} (m_{izq}) es la cantidad de veces que la partícula fue dispersada hacia la derecha (izquierda) en los niveles anteriores. Naturalmente, como al bajar en cada nivel sólo hay dos opciones, $m_{izq} = k - m_{der}$, de manera que $x_k = 2m_{der} - k + x_0$. Por otro lado, la distribución de probabilidades para m_{der} resulta ser una binomial equiprobable:

$$P(m_{der} = m) = \binom{k}{m} p_{der}^{m} p_{izq}^{k-m} = 2^{-k} \frac{k!}{m!(k-m)!}$$

Para $k \gg 1$, en la aproximación de Stirling obtenemos

$$P(m) \approx \frac{2^{-k}}{\sqrt{\pi k/2}} e^{-m\log(m/k) - (k-m)\log(1-m/k)}$$

Conviene ahora expandir el exponente en la llamada *aproximación de punto silla* (*saddle point*) en la que remplazamos el exponente por su expansión a segundo orden alrededor del valor máximo. Obtenemos entonces

$$P(m) \approx \frac{1}{\sqrt{\pi k/2}} e^{-2k(m/k - 1/2)^2}$$

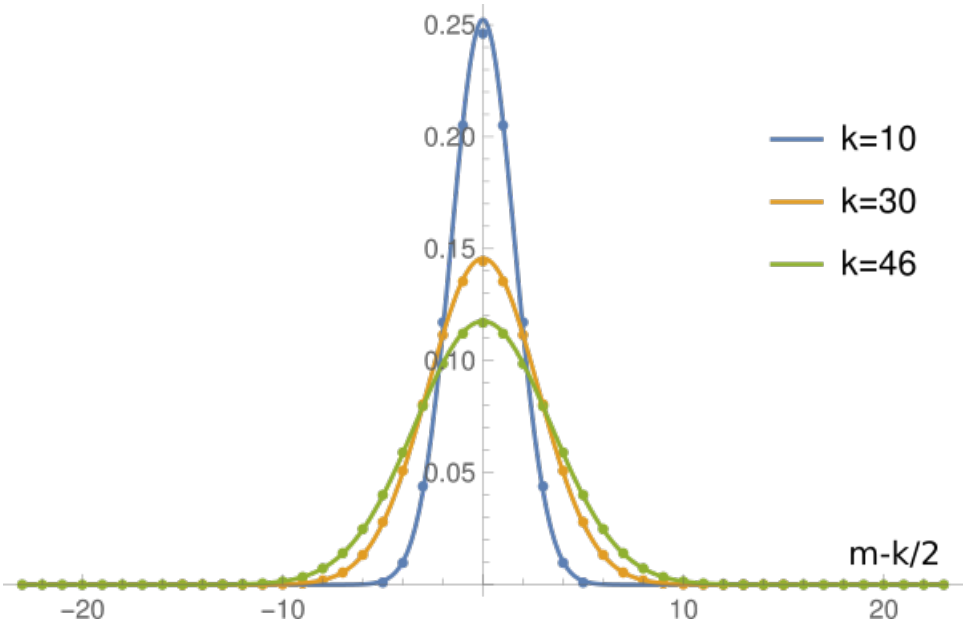


Figure 2: Distribución binomial y aproximación gaussiana

que luce como una gaussiana con un ancho $1/\sqrt{4k}$. Usando la relación lineal entre m_{der} y x_k, obtenemos finalmente

$$P(x) = 2^{-k} e^{-m\log(m/k) - (k-m)\log(1-m/k)} \approx \frac{1}{\sqrt{\pi k/2}} e^{-(x-x_0)^2/(2k))}$$

de lo que resulta que luego de suficientes niveles, la distribución de las posiciones es aproximadamente Gaussiana, con media $\langle x \rangle \approx x_0$ y $\sigma_x \approx \sqrt{k}$. A medida que bajamos niveles, el ancho de la distribución crece con $\sqrt{k}$, pero siempre centrada en x_0. Observamos que si entendemos a k como una cantidad proporcional al tiempo, la dinámica de la distribución de probabilidad se parece a la de un proceso difusivo: una perturbación localizada se *desparrama* de acuerdo a la ecuación

$$\frac{\partial}{\partial t}\psi(x,t) = \frac{1}{2}\nabla^2 \psi(x,t)$$

lo que resulta en una distribución *gaussiana*, centrada en la posición inicial, con un ancho que crece con $\sqrt{t}$. Esto es interesante ya que el modelo original de choques inelásticos da origen a la misma dinámica difusiva (Ley de Fick, modelo de Drude). Volveremos a esto más adelante.

Funciones de variables aleatorias y cambio de variables

Volviendo al modelo original de las disperisiones inelásticas, debido a las leyes de la mecánica, dada la posición y velocidad inicial *exacta* de una partícula, podríamos reconstruir su trayectoria completa en forma *determinista*. Sin embargo, como su posición y velocidades iniciales no se pueden conocer con precisión infinita, conviene modelar estas cantidades como variables aleatorias. Como resultado, la posición a un tiempo t será una función de la posición inicial que es una variable aleatoria, y por lo tanto, también será una variable aleatoria. Veamos cómo se relacionan.

Toda función real y aplicada a una variable aleatoria X da una nueva variable aleatoria $Y = y(X)$ con una distribución de probabilidad acumulada

$$F_Y(z) = P(\{\omega / y(X(\omega)) < z, \omega \in \Omega\})$$

Si $y(x)$ es una función *invertible*, se verifica que

$$f_X(x) = f_Y(y(x)) \left| \frac{dy}{dx} \right|$$

Una consecuencia interesante de esta relación es que si un espacio muestral está definido en términos de una variable aleatoria continua, siempre es posible encontrar una variable aleatoria tal que su distribución es *uniforme*. Basta para ello *definir* la nueva variable como $Y = y(X)$ con

$$y(x) = \int_{-\infty}^{x} f_X(x')dx'$$

Por construcción, $y : \mathbb{R} \Rightarrow (0, 1)$. Si además $f_X(x) > 0$, la función es invertible (si no, sería invertible a trozos). Esto nos permite "generar" variables aleatorias con distribuciones arbitrarias a partir de una variable aleatoria uniformemente distribuída en el intervalo $(0, 1)$.

Volviendo a nuestro modelo continio *detallado*, resulta que dependiendo de cómo es la dinámica *exacta*, una distribución inicial muy bien definida puede evolucionar a otra distribución muy diferente, en la que algunos de los grados de libertad pasan a *desparramarse* mientras otros toman valores cada vez mejor definidos. En particular, en un sistema mecánico, el *teorema de Liouville* establece que el *volumen* en el espacio de fases para una evolución Hamiltoniana se *conserva*. Esto significa que si a lo largo de la evolución ciertos grados de libertad tienden a desparramarse, otros tenderán a concentrarse para mantener el volumen de la distribución. Para poder discutir en más detalle este fenómeno, necesitamos herramientas para describir distribuciones de probabilidad para más de una variable.

Independencia y correlaciones entre variables aleatorias. Distribuciones conjuntas

En un experimento, suele interesar definir diferentes variables aleatorias que dan cuenta de diferentes aspectos de los elementos del espacio muestral. En el caso de una persona, puede interesar saber su altura, su peso corporal, o la concentración de una cierta substancia en sangre, o en el caso de una partícula, el valor de sus coordenadas, su masa, etc. Más aún, lo que suele importar es si las variables son o no *independientes* entre sí o si están *correlacionadas*. Para esto se introduce la noción de *distribución conjunta*. Consideremos primero el caso de dos variables X e Y. Su distribución conjunta acumulada se define como

$$F_{X,Y}(x, y) = P(\{X < x\} \cap \{Y < y\})$$

donde $X < x$ es una forma abreviada de expresar $\{\omega | X(\omega) < x, \omega \in \Omega\}$. Esto es, la probabilidad de que (X, Y) sea un punto del "rectángulo" $(-\infty, x) \times (-\infty, y)$ En general,

$$F_{X_1,...,X_k}(x_1, ..., x_k) = P(\bigcap_k \{X_k < x_k\})$$

Definida esta función, no es difícil expresar la probabilidad de que $X = \{X_1, ..., X_k\}$ se encuentre en una región arbitraria $\mathcal{A}$, que podemos expresar como la integral

$$P(X \in \mathcal{A}) = \int_{\mathcal{A}} f_X(x_1, ..., x_k)d^d x$$

con $f_X(x_1, ..., x_k)$ la *densidad de probabilidad conjunta* asociada a las variables $X_1, ..., X_k$.

En lo que sigue, consideremos el caso de dos variables, X, Y, que pueden ser "compuestas" de varias variables reales.

Distribuciones marginales y distribución condicional

A partir de la *densidad de probabilidad conjunta* $f_{XY}(x, y)$ es posible obtener las *densidad de probabilidad marginal* asociada a una de las variables integrando sore la otra:

$$f_X(x) = \int f_{XY}(x, y)dy$$

De manera análoga,

$$F_X(x) = \lim_{y\to\infty} F_{XY}(x,y)$$

Por otro lado, fijando el valor de una de las variables,

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

obtenemos la *densidad de probabilidad condicional* de X dado Y (Notar la semejanza con la regla de Bayes). En particular, si $f_{XY}(x,y) = f_X(x)f_Y(y)$, $f_{X|Y=y}(x) = f_X(x)$ para cualquier y. Decimos entonces que X e Y son *variables independientes*.

Observamos que la distribución *marginal* puede expresarse entonces como una *mezcla estadística* de las distribuciones condicionales:

$$f_X(x) = \int f_{X|Y=y}(x) f_Y(y) dy$$

Una medida de cuán dependientes son dos variables aleatorias es la *covarianza*:

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

que se anula si las variables son independientes. Por otro lado, se verifica que

$$cov(X,Y)^2 \leq var(X)var(Y)$$

se define entonces el *coeficiente de correlación lineal*

$$r(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$$

Estimadores estadísticos

Ahora que contamos con las nociones básicas para describir experimentos probabilísticos, podemos introducir la noción de *estimadores estadísticos*, que son las herramientas necesarias para estimar las propiedades de las variables aleatorias de interés con el resultado de experimentos.

Un *estimador estadístico* es una cantidad que construímos a partir de valores obtenidos de la observación de experimentos repetidos. Por ejemplo, el *promedio estadístico* o *media muestral* $\overline{X} = \sum_{i=1}^{L} \frac{X_i}{L}$ se construye a partir de los valores indivituales obtenidos al medir una cierta variable aleatoria L veces. De esta manera, un estimador estadístico es a su vez una variable aleatoria. Si el promedio se realiza sobre realizaciones independientes del experimento, en las que las variables X_i tienen la misma media $\mu = \langle X_i \rangle$ y la misma varianza $\sigma^2 = \langle X_i^2 - \mu^2 \rangle$, el promedio estadístico cumple

$$\langle \overline{X} \rangle = \mu$$

$$\langle \overline{X}^2 - \mu^2 \rangle = \sigma^2/L$$

Decimos entonces que a medida que el tamaño L de la muestra se vuelve más grande, el promedio *concentra* su valor en torno a la media asociada a la variable observada. Del teorema de Tchebychev se sigue entonces que la probabilidad de que el promedio se aleje del valor medio cae con el tamaño de la muestra como $L^{-1/2}$.

El Teorema del límite central da una descripción más precisa sobre cómo es la distribución del promedio como variable estadística, estableciendo que independientemente de la forma de la forma de la distribución de probabilidad de las variables individuales, para L suficientemente grande, el promedio tiende a una distribución normal. Notemos que este resultado vale independientemente de si la distribución original era una variable discreta o continua: en el límite de $L \to \infty$, su distribución tiende a la gaussiana, que es una distribución continua.

Observamos además que el promedio empírico es una función de las frecuencias relativas:

$$\langle X \rangle = \sum_i f_i X(\omega_i)$$

por lo que para un número suficientemente grande de *observaciones independientes*, converge *en probabilidad* a la esperanza.

Otros estimadores estadísticos. Momentos

Así como el *promedio* es un estimador estadístico, podemos considerar otros estimadores. Por ejemplo, la *mediana* es un estimador estadístico que se obtiene al ordenar la muestra de menor a mayor, y tomar el valor que queda en el centro, o la *moda*, que se define como el valor con mayor frecuencia. Una familia importante de estimadores son los *momentos de orden m*, que se definen como el *promedio* de la m-ésima potencia de la variable aleatoria. Por ejemplo, $\mu_3 X = \frac{1}{N}\sum_i X_i$ es el *tercer momento central* de la variable X. Naturalmente, el momento central de orden "0" es siempre $\mu_0 = 1$, el de órden 1 es el promedio $\mu_1 = \overline{X}$, etc. Una característica interesante de los momentos es que la distribución de frecuencias en una muestra de n elementos queda completamente caracterizada por sus momentos hasta orden n.

Dada una distribución de probabilidad, los valores esperados de los momentos se expresan como $E[\mu_n X] = \int f(x)x^n dx$. En particular, si la función $E[e^{tX}]$ existe y es *suave* en torno a $t = 0$, $E[\mu_n X] = \frac{\partial^n}{\partial t^n} E[e^{tX}]\big|_{t=0}$ por lo que $E[e^{tX}]$ es la llamada *función generadora de momentos* para la distribución.

$$\begin{aligned} E[e^{tX}] &= \int f(x)e^{tx}dx \\ &= \sum_m \int f(x)\frac{t^m x^m}{m!}dx \\ &= \sum_m \frac{t^m}{m!}E[X^m] \\ \frac{d^n}{dt^n}E[e^{tX}] &= \frac{d^{n-1}}{dt^{n-1}}\sum_{m=1}\frac{mt^{m-1}}{m!}E[X^m] = ... \\ &= \sum_{m=n}\frac{n!t^{m-n}}{m!}E[X^m] \end{aligned}$$

Como la exponencial es una función analítica, podemos *rotarla* en el plano complejo $tx \to \mathbf{i}tx$, con lo que

$$E[e^{itX}] = \int f(x)e^{-\mathbf{i}tx}dx = \mathcal{F}[f](t)$$

es la *Transformada de Fourier* de la densidad de probabilidad f. De esta manera, si se conocen *todos* los momentos de la distribución de probabilidades, podemos reconstruir la distribución de probabilidades, y por lo tanto, los momentos determinan completamente a la distribución.

Promedios temporales e hipótesis ergódica

Supongamos ahora que la distribución de probabilidades para una variable aleatoria X depende del tiempo. A los efectos de calcular probabilidades, el tiempo es otra variable aleatoria, por lo que podemos incluir al tiempo en el espacio muestral. La distribución de probabilidad de X un *marginal* de la distribución $f_{X,T}(\xi, t)$.

Si la evolución es *determinista*, podemos expresar la distribución de probabilidades a un dado tiempo t en términos de la distribución a un tiempo anterior:

$$f(\xi(\xi_0, t), t) = f(\xi_0, t_0)$$

donde $\xi(\xi_0, t)$ representa el valor de ξ al tiempo t si partió de ξ_0 a t_0, esto es, $\xi(\xi_0, t)$ es la trayectoria del sistema si a t_0 se encontraba en ξ_0. Derivando esta relación respecto de t obtenemos

$$\frac{d}{dt}f(\xi(\xi_0, t), t) = \frac{\partial f}{\partial t} + \frac{d\xi}{dt}\cdot\nabla_\xi f = 0$$

donde el segundo miembro se anula ya que no depende de t. En mecánica clásica, la evolución de un sistema mecánico no depende sólo de sus coordenadas en su espacio de configuraciones, sino también de sus velocidades. En la ecuación anterior debemos entonces interpretar por x las coordenadas en el espacio de fases $\xi \equiv (x, p)$. Luego, el teorema de Liouville establece que para t fijo, $f(\xi, t)$ satisface una *ecuación de continuidad local*

$$\frac{\partial}{\partial t}f(\xi, t) + \nabla\cdot\vec{\mathcal{J}}$$

con $\vec{\mathcal{J}} = \xi f(\xi, t)$ una *corriente conservada* en el espacio de fases.

Prueba: remplazando la definición de $\vec{\mathcal{J}}$ en la ecuación de continuidad,)

$$\frac{\partial f}{\partial t} + \nabla\cdot\dot{\xi}f = (\frac{\partial f}{\partial t} + \dot{\xi}\cdot\nabla f) + f\nabla\cdot\dot{\xi} = \frac{df}{dt} + f\nabla\cdot\dot{\xi}$$

Como la evolución es determinista, la derivada total se anula. Por otro lado, el término restante se expresa como

$$f\nabla \cdot \dot{\xi} = f(\nabla_x \dot{x} + \nabla_p \dot{p})$$

Pero por las ecuaciones de Hamilton,

$$\nabla_x \dot{x} + \nabla_p \dot{p} = (\nabla_x \nabla_p H - \nabla_p \nabla_x H) = 0$$

De este teorema se sigue que el *volumen* en el espacio de fases que ocupa una distribución de probabilidad se mantiene *constante* como función del tiempo. Sin embargo, si observamos el sistema en un instante de tiempo al azar dentro de un intervalo *suficientemente largo*, ese volumen se desparramará a lo largo de la trayectoria del sistema. Por otro lado, en un sistema cerrado, estas trayectorias se encuentran restringidas por las leyes de conservación de la energía, la cantidad de movimiento y el momento angular. Si además las trayectorias están confinadas en el espacio de coordenadas, el *volumen accesible* por dichas trayectorias es finito. Luego, o bien la trayectoria es *periódica*, o bién estarán confinadas a alguna región del *volumen accesible*, de manera que si se espera un tiempo suficientemente largo, el sistema pasará arbitrariamente cerca de cualquier punto de esa región.

En todo caso, la densidad de probabilidad (marginal) para encontrar al sistema en una dada región del espacio de fases resultará proporcional a la cantidad de tiempo que el sistema transite en el entorno de esa región. Si la región es todo el volumen accesible, y la densidad de probabilidad resulta *uniforme*, decimos que el sistema satisface la *hipótesis ergódica*.