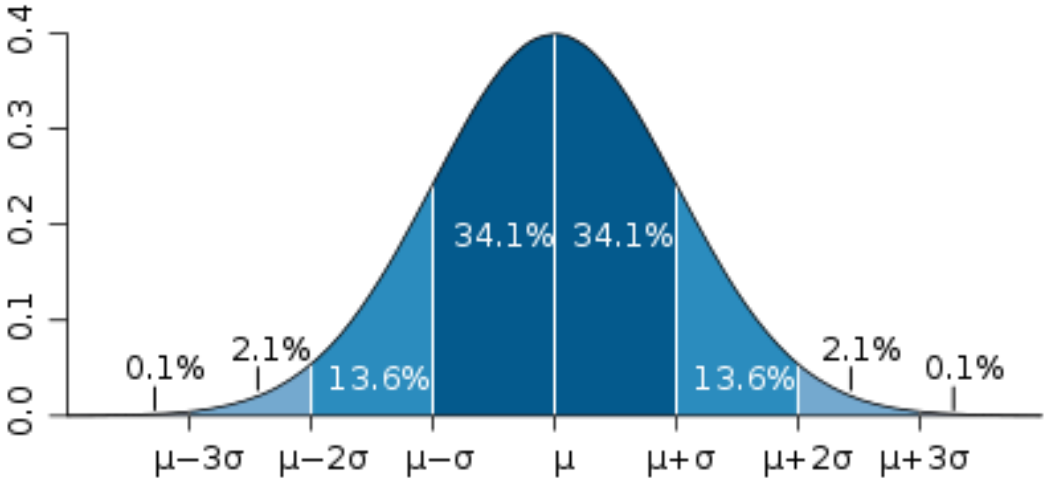
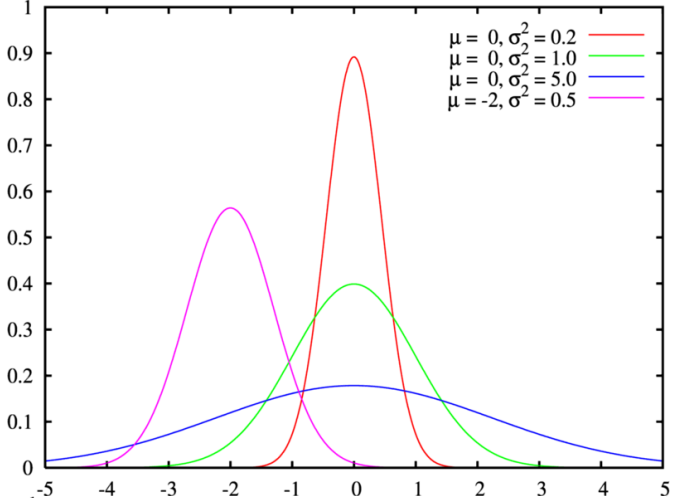
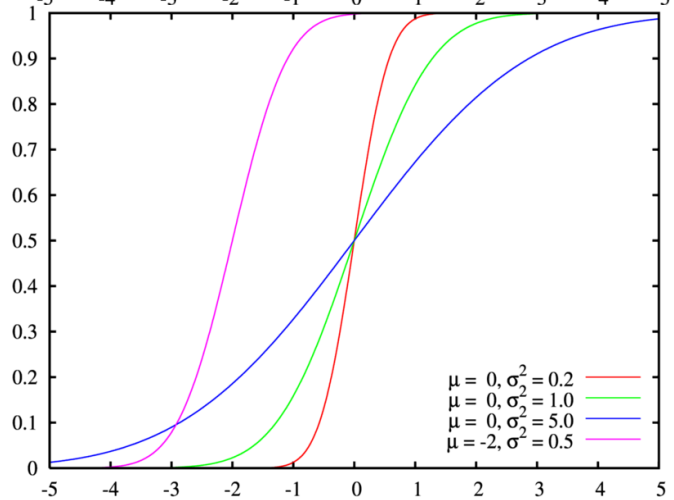
La distribución normal

- El modelo más importante para variables continuas, es la distribución normal:
  - $f(x)=1/\sigma(2\pi)^{0.5}\ \exp\{-(1/2\sigma^2)(x-\mu)^2\}$







Imágenes extraídas de
https://es.wikipedia.org/wiki/Distribuci%C3%B3n_normal

La distribución normal

- Depende de dos parámetros:
  - μ que es al mismo tiempo la media, mediana y moda.
  - σ que es la desviación típica.
- **Ver ejemplo 5.6**
- La distribución normal aproxima lo observado en muchos procesos de medición sin errores sistemáticos. P. ej. Las medidas físicas de cuerpo humano en una población, las características psíquicas medidas por el test de inteligencia o personalidad, las medidas de calidad de muchos procesos industriales, o los errores de las observaciones astronómicas, etc.
- Una justificación de la frecuente aparición de la distribución normal es el teorema central del límite: cuando los resultados de un experimento son debidos a un conjunto grande de causas independientes, que actúan sumando sus efectos (siendo cada efecto individual de poca importancia respecto al conjunto), es esperable que los resultados sigan una distribución normal.

La distribución normal

- N(0,1) es una normal estándar de $\mu=0$ y $\sigma=1$.
- Para convertir la variable x en la normal estándar z:
  - $z=(x-\mu)/\sigma$, que sustituyendo en la normal:
  - $f(z)=1/\sigma(2\pi)^{0.5}\ \exp\{-(z)^2/2\}$.
- El cálculo de probabilidades de x se realiza utilizando la expresión:
  - $F(x_0)=P(x\leq x_0)=P(\mu+\sigma z\leq x_0)=P(z\leq(x_0-\mu)/\sigma)=\Phi((x_0-\mu)/\sigma)$, donde $\Phi(.)$ representa la función distribución de la normal estándar.
- Esto quiere decir que podemos calcular el valor de la distribución de cualquier variable normal en cualquier punto si conocemos la función distribución de la normal estándar, f(z) (ver tabla 4 del apéndice de tablas).
- Solo tenemos que convertir el punto x_0 en un punto de la normal estándar: restándole la media y dividiendo por la desviación típica.

La distribución normal

- Se comprueba que, en toda distribución normal:
  - En el intervalo $\mu \pm 2\sigma$ se encuentra el 95,5% de la distribución.
  - En el intervalo $\mu \pm 3\sigma$ se encuentra el 99,7% de la distribución.
- Es decir conocer que ciertos datos siguen una distribución normal nos permite dar intervalos más precisos que los de la acotación de Tchebycheff:
  - $P(\mu - k\sigma \leq x \leq \mu + k\sigma) \geq 1-(1/k^2)$, para cualquier valor de k.
  - Para cualquier variable aleatoria el intervalo $\mu \pm 2\sigma$, contiene al menos el 75% de la distribución.
  - Para cualquier variable aleatoria el intervalo $\mu \pm 3\sigma$, contiene al menos el 89% de la distribución.
- La distribución normal se toma como referencia para juzgar muchas otras distribuciones, por ejemplo como el coeficiente de apuntamiento de la normal es 3, se define como
  - $CA_p = \frac{\mu_4}{\sigma^4} - 3.$

La normal como aproximación de otras distribuciones: El teorema Central del Límite

- El teorema establece que si $x_1, \ldots, x_n$ son variables aleatorias independientes con media μ_i, varianza σ_i^2 y distribución cualquiera (y no necesariamente al misma) y formamos la variable aleatoria suma Y:
  - $Y = x_1 + \ldots + x_n$,entonces si n crece $\sigma_i^2 / \Sigma\sigma_i^2 \rightarrow 0$, que implica que el efecto de una variable es pequeño respecto al efecto total, entonces la variable Z cumple:
    - $Z = (Y - \Sigma\mu_i)/(\Sigma\sigma_i^2)^{0.5}$ tiende a una distribución N(0,1).
- Darse cuenta que la variable Z está en la forma de normal estándar.
- El resultado anterior implica que si n es grande, podemos aproximar las probabilidades de la variable aleatoria Y utilizando que:
  - $Y \sim N(\Sigma\mu_i, (\Sigma\sigma_i^2)^{0.5})$

La normal como aproximación de otras distribuciones: Relación entre binomial, Poisson y normal

- Imaginemos que la variable $Y= x_1+\ldots+x_n$ es la suma de n variables de Bernoulli, x_i, que toman el valor 1 cuando el elemento es defectuoso y 0 en caso contrario entonces por el TCL:
  - Como $E[x_i]=p$ y $Var[x_i]=pq$, la variable aleatoria Y tenderá hacia la normal con parámetros $\Sigma\mu_i = np$ y $(\Sigma\sigma_i^2)^{0.5}=(npq)^{0.5}$, i.e. $N(np, (npq)^{0.5})$.
- En general la aproximación por una normal es buena para $npq>5$.

La normal como aproximación de otras distribuciones: Relación entre binomial, Poisson y normal

- Esta misma situación pasa con variables de Poisson, supongamos que Y(0,T) es una variable de Poisson que cuenta el número de sucesos entre 0 y T, con media del proceso en ese intervalo de λ. Si dividimos el intervalo en n partes iguales:
  - $Y(0,T)=x_1(0,t_1)+x_2(t_1,t_2)+\ldots+x_n(t_{n-1},T)$, donde $x_i(t_{i-1},t_i)$ cuenta el número de sucesos en el intervalo (t_{i-1},t_i).
  - Así por tanto Y es una suma de variables aleatorias $x_i(t_{i-1},t_i)$ distribuidas según Poisson independientes con media $\mu_i=\lambda/n$ varianza $\sigma_i^2=\lambda/n$. Por tanto podemos aplicar el TCL si n crece. Aquí suponemos que la media del número de sucesos λ de Poisson se reparte aproximadamente igual por todos los intervalos $x_i(t_{i-1},t_i)$ y que en cada uno de ellos seguirá habiendo un proceso de Poisson, por lo tanto la media en cada intervalo es igual a su varianza (esta suposición se puede hacer cuando λ es grande según crece n).
- Por tanto, se verifican las condiciones del TCL cuando n aumenta (i.e. λ grande), y la distribución de Poisson se puede aproximar por una distribución normal de parámetros $\Sigma\mu_i=\lambda$ y $(\Sigma\sigma_i^2)^{0.5}=(\lambda)^{0.5}$ (recordar que la suma es sobre n).
- La aproximación se puede demostrar que es buena cuando $\lambda>5$.

La distribución lognormal

- Una consecuencia del TCL, es que si un determinado efecto es el producto de muchas causas, cada una de poca importancia en relación a las demás y además las causas son independientes entre ellas de tal forma que $y=x_1x_2\ldots x_n$
- Entonces el **logaritmo de y** seguirá una distribución normal por el TCL.
- Así se llama la distribución lognormal a la distribución de la variable $x=\log y$, que se puede deducir que sigue la siguiente distribución: $g(y)=1/\sigma(2\pi)^{0.5}\ \exp\{-(1/2\sigma^2)(\log y-\mu)^2\}(1/y)$, para $y>0$.

Distribuciones deducidas de la normal: distribución χ^2 de Pearson

- Es una de la herramientas de análisis más utilizadas en ciencia actual.
- Supongamos que generamos n variables aleatorias distribuidas según una normal, con media cero y varianza la unidad, y definimos la siguiente operación:
  - $\chi_n^2 = z_1^2 + \cdots + z_n^2$
- Si aplicamos este procedimiento múltiples veces (elevamos los *n* valores generados al cuadrado y los sumamos), al final obtenemos la distribución de una variable que solo depende del número de sumandos (n grados de libertad).
- Esta distribución se denomina χ^2 con n grados de libertad.

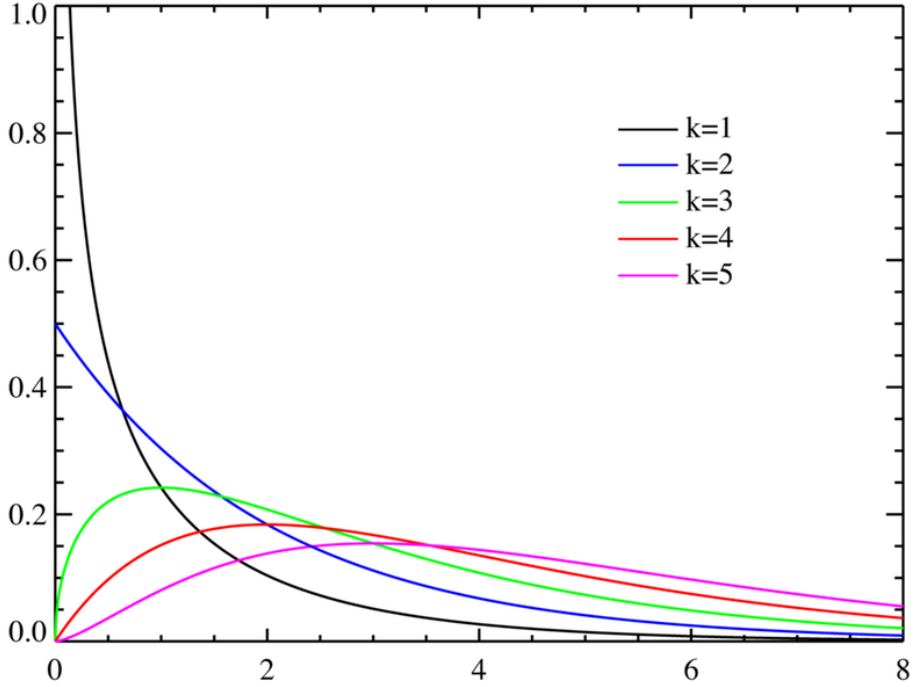
Distribuciones deducidas de la normal: distribución χ^2 de Pearson

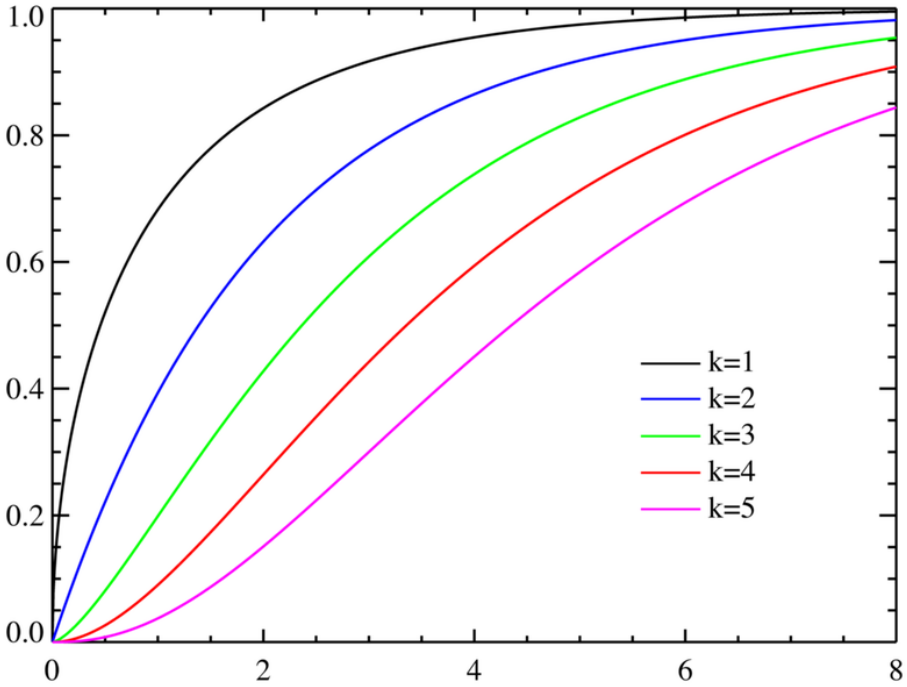
- Los parámetros de la distribución se sacan fácilmente haciendo uso de la independencia de variables:
  - $E[z_i^2] = 1$, ya que $\sigma^2 = 1$ (recordemos que momento de orden k respecto al origen es $\mathrm{E}[\mathrm{x}^k] = m_k = \int x^k f(x)dx$ y el momento de orden k respecto la media $\mathrm{E}[(x-\mu)^k] = \mu_k = \int (x-\mu)^k f(x)dx$, estos momentos permiten describir otros aspectos relevantes de una distribución de probabilidad).
  - $E[z_i^4] = 3$ (coeficiente de apuntamiento o curtosis de una normal).
- Así se puede comprobar que $E[\chi_n^2] = n$ y $Var[\chi_n^2] = 2n$.

Distribuciones deducidas de la normal: distribución χ^2 de Pearson

- Más propiedades importantes:
- Si tenemos la suma de dos variables Chi-cuadrado independientes $\chi^2_{n_1}$ y $\chi^2_{n_2}$, el resultado es otra variable Chi-cuadrado que tiene de grados de libertad las suma de los grados de libertad: $\chi^2_{n_1+n_2}$.
- La distribución $\frac{\chi^2_n}{n}$ representa **la distribución de la varianza de n variables normales independientes:**
  - $\frac{\chi^2_n}{n}=\frac{z_1^2+\cdots+z_n^2}{n}$, y como z_i es una variable normal tiene media 0, entonces este cociente es una varianza de n muestras normales independientes (recordar que la varianza es $\sigma^2=\sum(z_i-\mu)^2/n=\sum z_i^2/n=\frac{\chi^2_n}{n}$), ya aquí para cada z_i su media es $\mu=0$).
- La distribución $\frac{\chi^2_n}{n}$ tiene media 1 y varianza $2/n$, $(\mathrm{Var}[\frac{\chi^2_n}{n}]=\frac{1}{n^2}\mathrm{Var}[\chi^2_n]=2/n)$.
  - Recordar las propiedades de la esperanza y varianza: $E[aX]=aE[X]$ y $Var[aX]=a^2Var[X]$.

Distribuciones deducidas de la normal: distribución χ^2 de Pearson





Imágenes extraídas de https://es.wikipedia.org/wiki/Distribuci%C3%B3n_%CF%87%C2%B2

Distribuciones deducidas de la normal: distribución t de Student

- Distribución utilizada por el químico Gosset (1908) para ver como diferentes tratamientos a la cervecería Guinnes de Dublin podían afectar a la calidad de la misma. Se publico bajo el seudónimo de Student (Guinnes no dejaba divulgar resultados a sus empleados).
- La expresión es $t_n = \frac{z}{\left(\frac{\chi^2_n}{n}\right)^{1/2}}$, siendo z una variable aleatoria normal estándar, independiente del denominador, siendo el denominador la raíz cuadrada de Chi-cuadrado dividido por el número de sus grados de libertad.

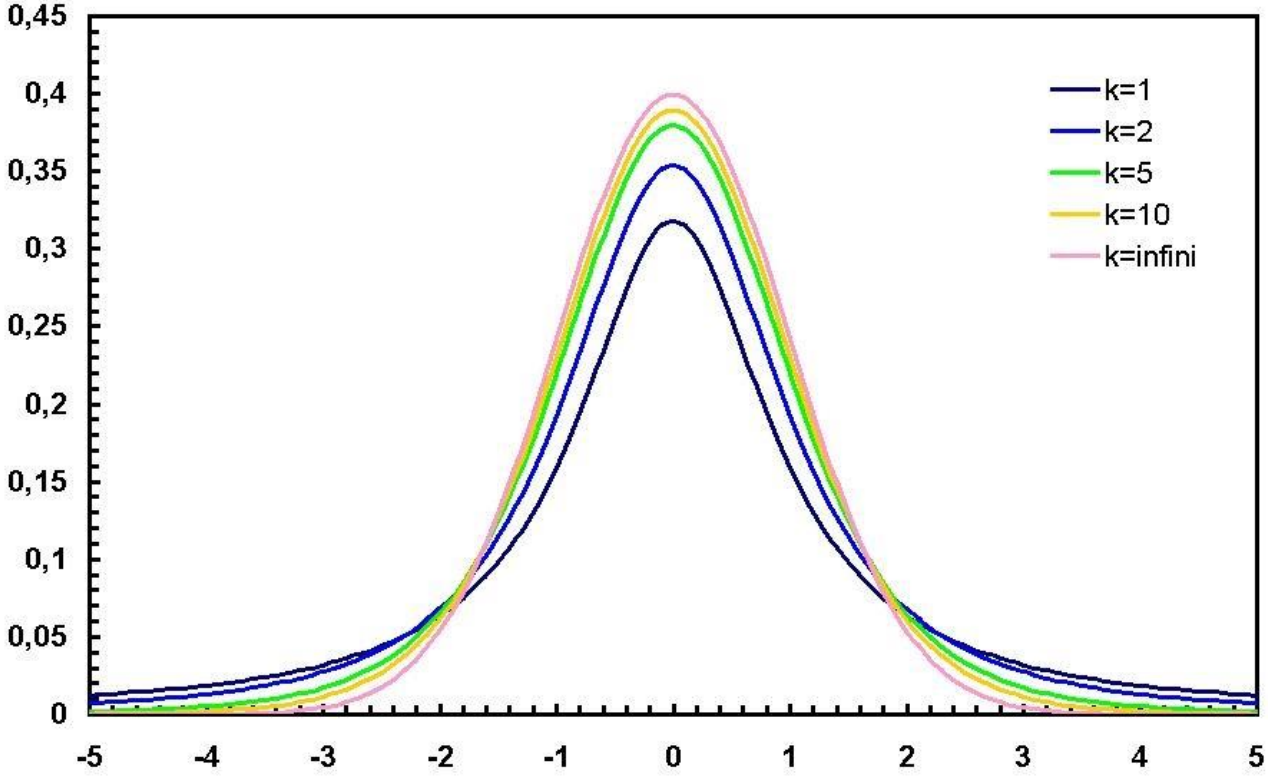
Distribuciones deducidas de la normal: distribución t de Student

- Así la distribución t de Student se forma mediante la generación de una variable normal estándar, y de manera independiente generamos n variables aleatorias distribuidas según una normal, con media cero y varianza la unidad para generar Chi-cuadrado. Al final dividimos la primera generación por la segunda de Chi-cuadrado.
- El denominador $\left(\frac{\chi_n^2}{n}\right)^{1/2} = \left(\left(\frac{1}{n}\right)(x_1^2 + \cdots + x_n^2)\right)^{1/2}$ representa la desviación típica muestral de n variables normales x independientes, ya que estas tienen media cero (recordar Chi-cuadrado).
- Así la distribución t de Student es el resultado de comparar una variable de media cero con una estimación de su desviación típica construida con n datos independientes (observación clave para intervalos de confianza y contraste de hipótesis).

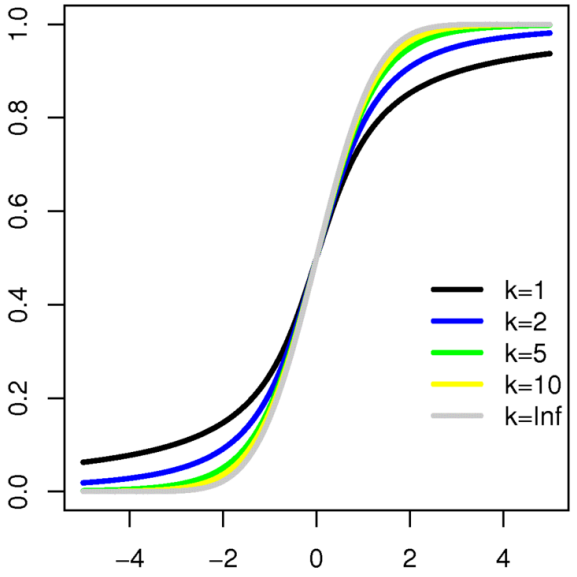
Distribuciones deducidas de la normal: distribución t de Student

- La variable t de Student es simétrica con mayor dispersión que la distribución normal.
- La variable t de Student tiende rápidamente a una normal estándar a medida que aumentamos n (para n mayor que 100 la aproximación a la normal es buena).
- La variable t de Student tiene una media 0 y varianza (para n>2):
  - $Var[t] = \frac{n}{n-2}$

Distribuciones deducidas de la normal: distribución t de Student



Imágenes extraídas de
https://es.wikipedia.org/wiki/Distribuci%C3%B3n_t_de_Student



Ejemplos para realizar en casa

- Realizar para casa los ejemplos del libro:
  - 5-1, 5-2, 5-3, 5-4, 5-5, 5-6, 5-7 y 5-8.

Modelos multivariantes

Modelos multivariantes

- Variables aleatorias vectoriales
- Distribución conjunta
- Distribuciones marginales
- …

Modelos multivariantes: variables aleatorias vectoriales

- Cuando en lugar de observar una característica de una población se observan n características a la de vez de la población en cada elemento de la misma, entonces diremos que tenemos acceso a una **variable aleatoria vectorial o multidimensional**.
- Cada valor de la variable aleatoria esta compuesta de n valores numéricos.
- Tenemos definida una distribución conjunta de una variable aleatoria multidimensional si se especifica:
  - El espacio muestral, siendo cada punto del mismo un vector n-dimensional.
  - Las probabilidades de cada posible resultado de esos vectores n-dimensionales.

Modelos multivariantes: distribución conjunta

- Dada una variable aleatoria vectorial discreta (supongamos que es bidimensional para simplificar), la función de probabilidad conjunta $p(X)=p(x_1,x_2)$ proporciona las probabilidades de cada posible valor de la pareja en este caso.
- Debe verificar (al igual que el caso unidimensional):
  - $p(X_i)=p(x_{1i},x_{2i})\geq 0 \ \forall i$
  - $\sum_i p(X_i)= \sum_i p(x_{1i},x_{2i})=1$
- Cuando las variables son continuas las probabilidades vienen dadas por función densidad y los sumatorios por integrales:
  - $f(X)=f(x_1,x_2)\geq 0$
  - $\int\int f(x_1,x_2)\, dx_1 dx_2=1$
- Las probabilidades se calculan por integración en los intervalos correspondientes:
  - $P(a<x_1\leq b, c<x_2\leq d) = \int_a^b \int_c^d f(x_1,x_2)\, dx_1 dx_2$

Modelos multivariantes: distribuciones marginales

- Variables discretas:
  - $p(x_1)= \sum_{x_2} p(x_1,x_2)$
  - $p(x_2)= \sum_{x_1} p(x_1,x_2)$
- Variables continuas:
  - $f(x_1)= \int f(x_1,x_2)\, dx_2$
  - $f(x_2)= \int f(x_1,x_2)\, dx_1$
- Probabilidad de pertenecer a un intervalo (a,b):
  - $P(a<x_1\leq b) = P(a<x_1\leq b, -\infty<x_2\leq\infty) = \int_a^b dx_1 \int_{-\infty}^{\infty} f(x_1,x_2)dx_2 = \int_a^b f(x_1)dx_1$
- Así se puede seguir con probabilidad condicional, teorema de Bayes, esperanzas, correlaciones, distribuciones varias variables, etc. (mirar el capítulo 6 del libro).