Modelos

- Probabilidad y variables aleatorias
  - Probabilidad y sus propiedades
  - Probabilidad condicionada
  - Variables aleatorias
- Modelos de distribución de probabilidad
  - El proceso de Bernoulli
  - El proceso de Poisson
  - Las distribuciones de duraciones de vida
  - La distribución Normal
  - La distribución Log-Normal
- Modelos Multivariantes

Modelos: Probabilidad y variables aleatorias

Probabilidad y variables aleatorias

- Probabilidad y sus propiedades
- Probabilidad condicionada
- Variables aleatorias

Probabilidad y sus propiedades: Concepto

- Cuando tenemos una muestra de una población, después de describirla, nuestro objetivo es inferir las propiedades de la población a partir de la muestra.
- El instrumento conceptual que permite la generalización es un **modelo de la población**.
- El calculo de probabilidades me permite calcular este modelo que actúa de puente entre lo observado, i.e. la muestra, y lo desconocido, i.e. la población.
- La probabilidad de una población finita homogénea de N elementos, k de los cuales tienen la característica A:
  - P(A)=k/N

Probabilidad y sus propiedades: Concepto

Probabilidad y sus propiedades: Concepto

- Hay que ser consciente que muchas veces no se puede obtener un conocimiento exacto de la probabilidad ya que:
  - Al no ser posible una experimentación definida, siempre tenemos una información limitada sobre la frecuencia relativa.
  - El sistema observado puede variar a lo largo del tiempo, y por tanto también las frecuencias relativas.
- Por tanto para poblaciones finitas la identificación de la probabilidad con la frecuencia relativa es casi inmediata, pero para poblaciones infinitas puede presentar serios problemas.
- Esto se complica mucho más para sucesos inciertos, que solamente ocurren un número de veces muy reducido.

Probabilidad y sus propiedades: Definición

- Población: conjunto de elementos homogéneos en los que se desea investigar la ocurrencia de una característica o propiedad:
  - Numero de elementos finito o infinito.
  - Debe ser posible observar sus elementos.
  - Debe ser posible saber si un elemento pertenece a ella o no.
- Sucesos elementales: es el conjunto de resultados posibles que verifican:
  - Siempre ocurre alguno de ellos.
  - Son mutuamente excluyentes.
- Sucesos compuestos: los construidos a partir de uniones de resultados elementales.
- P. ej. Tirar un dado:
  - Sucesos elementales: 1, 2, 3, 4, 5, 6
  - Sucesos compuestos: número par, número impar, menor 5, múltiplo de 2, etc.

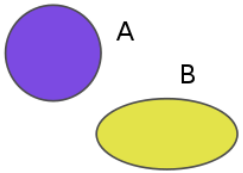
Probabilidad y sus propiedades: Definición

- Espacio muestral: conjunto de todos los resultados posibles del experimento. Así los elementos elementales y compuestos son los subconjuntos del espacio muestral.
  - Definimos el suceso seguro, E, como todo el espacio muestral: es seguro porque siempre ocurre.
  - Definimos el suceso imposible, Ø, como aquel que no ocurre nunca.

Probabilidad y sus propiedades: Definición

- Se desea asociar a cada suceso una medida de incertidumbre que llamaremos probabilidad, con las siguientes propiedades:
  1. Debido a que la $f_r(A)$ es un numero comprendido entre 0 y 1, la probabilidad: $0 \leq P(A) \leq 1$.
  2. La frecuencia del suceso seguro ocurre siempre: $P(E)=1$.
  3. Si A y B son características mutuamente excluyentes y las unimos en una nueva C=A+B. C ocurre cuando ocurre A o ocurre B. La frecuencia relativa de C es la suma de las frecuencias relativas de A y B, y por tanto la probabilidad de sucesos mutuamente excluyentes: $P(A+B)=P(A \cup B)=P(A \text{ or } B)=P(A)+P(B)$. P. ej probabilidad de sacar 1 o 2 en una tirada de dado: P(1 or 2) =1/6+1/6=1/3.

Imagen extraída de https://es.wikipedia.org/wiki/Probabilidad_condicionada



mutuamente excluyentes

Probabilidad y sus propiedades: Definición

- Se desea asociar a cada suceso una medida de incertidumbre que llamaremos probabilidad, con las siguientes propiedades:

4. Suponiendo **A y B NO son mutuamente excluyentes** definimos n_{AB}, $n_{A\overline{B}}$, $n_{\overline{A}B}$ al numero de veces que aparecen los sucesos compuestos mutuamente excluyentes (A y B), (A y no B), (no A y B), tendremos:
   - $n_A = n_{AB} + n_{A\overline{B}}$, $n_B = n_{AB} + n_{\overline{A}B}$, $n_{A+B} = n_{AB} + n_{A\overline{B}} + n_{\overline{A}B}$, de estas 3 ecuaciones se puede obtener $n_{A+B} = n_A + n_B - n_{AB}$ (sustituyendo $n_{\overline{A}B}$ y $n_{A\overline{B}}$)
   - Si dividimos por el número total de observaciones obtenemos las frecuencias relativas que pasadas a probabilidades:
     - P(A+B)=P(A)+P(B)-P(AB)
     - P(A U B)=P(A)+P(B)-P(A∩B)
     - P(A or B)=P(A)+P(B)-P(A and B)
   - **P. ej.** Probabilidad de elegir una carta al azar de una baraja sea un **corazón** o una **cara** (J, Q, K) es $13/52 + 12/52 - 3/52 = 11/26$ (de 52 13 son corazones, 12 caras , y 3 son corazones y caras).

Probabilidad y sus propiedades: Definición

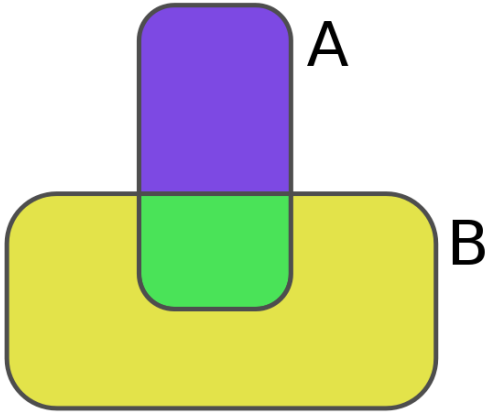
- Se desea asociar a cada suceso una medida de incertidumbre que llamaremos probabilidad, con las siguientes propiedades:

5. Suponiendo que $\overline{A}$ es el suceso complementario de A, que ocurre siempre que no ocurre A, de las propiedades anteriores se puede deducir:
   - $P(\overline{A})=1-P(A)$.

Probabilidad Condicionada

- La frecuencia relativa de A condicionada a la ocurrencia de B se define considerando los casos en los que aparece B, y viendo en cuantos de estos casos ocurre A (casos que aparecen A y B, dividido por el numero de casos que aparece B):
  - $f_r(A|B)=n_{AB}/n_B$, como $f_r(A)=n_A/n$, $f_r(B)=n_B/n$ y $f_r(AB)=n_{AB}/n$, se tiene:
  - $f_r(A|B)=f_r(AB)/f_r(B)$ o lo que es lo mismo:
  - $f_r(AB) = f_r(A|B)f_r(B)=f_r(B|A)f_r(A)$
- Así exigiremos la misma propiedad a la probabilidad y definiremos probabilidad de un suceso A condicionada a otro B como:
  - $P(A|B)=P(AB)/P(B)$, ojo es una intercesión en P(B)

Imagen extraída de https://es.wikipedia.org/wiki/Probabilidad_condicionada



Tomando los casos en los que B se cumple, la fracción en los que también se cumple A.

Probabilidad Condicionada: Independencia de sucesos

- Diremos que A y B son sucesos independientes, si el conocimiento de la ocurrencia de uno no modifica la probabilidad de aparición del otro:
  - $P(A|B)=P(A)$ o $P(B|A)=P(B)$, o lo que es lo mismo:
  - $P(AB)=P(A)\,P(B)$
- Esto se puede generalizar a cualquier numero de sucesos:
  - Diremos que los sucesos $A_1,\ldots, A_n$, son independientes si la probabilidad conjunta de cualquier subconjunto que pueda formarse con ellos es el producto de las probabilidades individuales.

Probabilidad Condicionada: Teorema de Bayes

- Consideremos un experimento que se realiza en dos etapas.
- **En la primera** los sucesos posibles, $A_1,\ldots, A_n$, son mutuamente excluyentes con probabilidades conocidas, $P(A_i)$ y con $\sum P(A_i)=1$.
- La $P(A_i)$ son las **probabilidades a priori**.
- **En la segunda** los resultados posibles, B_j, dependes de los de la primera y se conocen la probabilidades condicionadas $P(B_j|A_i)$, de obtener cada posible resultado B_j, cuando aparece en la primera etapa el suceso A_i.
- Se efectúa ahora el experimento, pero el resultado de la primera fase A_i, no se conoce, pero si el de la segunda fase que es B_j.
- Las $P(A_i|B_j)$ son las **probabilidades a posteriori**, y es lo que calcula el Teorema de Bayes.

Probabilidad Condicionada: Teorema de Bayes

- El teorema de Bayes permite calcular la probabilidades $P(A_i \mid B_i)$, de los sucesos no observados en la primera etapa, dado el resultado observado en la segunda (**probabilidades a posteriori**).
- Se calcula partiendo de la definición de probabilidad condicionada:
  - $P(A_i \mid B_i)=P(A_iB_i)/P(B_i)= P(B_i \mid A_i)P(A_i)/P(B_i)$
- Por otro lado $P(B_i)=P(B_iA_1+B_iA_2+\ldots+B_iA_n)$, ya que B_i debe ocurrir con algunos de los sucesos de A_i.
- Como los sucesos B_iA_1, B_iA_2,... son mutuamente excluyentes ya que por la definición del problema los A_i lo son, entonces:
  - $P(B_i)= \sum_i P(B_i\ A_i)= \sum_i P(B_i \mid A_i)\ P(A_i)$
- Si sustituimos arriba obtenemos el Teorema de Bayes:
  - $P(A_i \mid B_i)=P(B_i \mid A_i)P(A_i)/P(B_i)=P(B_i \mid A_i)P(A_i)/\sum_i P(B_i \mid A_i)\ P(A_i)$

Probabilidad Condicionada: Teorema de Bayes

- Dos urnas: U_1 (70% b, 30% n) y U_2 (30% b, 70% n): Sucesos A_1 y A_2
- Se selecciona una urna al azar y se saca el resultado de bolas con remplazamiento: bnbbbbnbbb. Es el suceso B,
- ¿Cuál es la probabilidad de que la muestra venga de U_1?
- Se puede considerar el experimento en dos etapas:
  - Selección de urna.
  - Extracción de la muestra de la urna seleccionada.
- Selección al azar de las urnas $P(U_1)=P(U_2)=0.5$
- Los 10 sucesos de sacar bolas son independientes (extracción con remplazamiento): $P(b|U_1)=0.7$, $P(n|U_1)=0.3$
- Se verifica $P(B|U_1)=P(bnbbbbnbbb|U_1)=P(b|U_1)\ P(n|U_1)\ P(b|U_1)\ldots P(b|U_1)=0.7^8 0.3^2$
- Se verifica $P(B|U_2)=P(bnbbbbnbbb|U_2)=P(b|U_2)\ P(n|U_2)\ P(b|U_2)\ldots P(b|U_2)=0.3^8 0.7^2$
- La probabilidad pedida es $P(U_1|B)$ que la calculamos por el Teorema de Bayes:
  - $P(U_1|B)=P(B|U_1)P(U_1)/\sum_i P(B|U_i)\ P(U_i)= P(B|U_1)P(U_1)/\ (P(B|U_1)\ P(U_1)+P(B|U_2)\ P(U_2))$
  - $P(U_1|B)=(0.7^8 0.3^2\ 0.5)/(0.7^8 0.3^2\ 0.5+\ 0.3^8 0.7^2\ 0.5)=0.994$ (para casa calcular $P(U_2|B)$)

Variables aleatorias: Discretas

- El cálculo de probabilidades utiliza variables numéricas que se denominan aleatorias, ya que sus valores se determinan al azar.
- Variables aleatorias discretas: cuando toma un numero de valores finitos o infinito numerable.
- Función de probabilidad: una variable discreta aleatoria se define por sus posibles valores discretos (su espacio muestral), junto con sus probabilidades respectivas. Así la función de probabilidad p(x) es la función que indica las probabilidades de cada posible valor:
  - $p(x_i)=P(x=x_i)$, con $\sum_i p(x_i)=1$.

Variables aleatorias: Discretas

- Función de distribución: $F(x_0)=P(x\leq x_0)$ (o Función de Distribución Acumulada)
- Si suponemos que la variable x toma los posibles valores:
  - $x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_n$
- La función distribución viene determinada por:
  - $F(x_1)=P(x\leq x_1)=p(x_1)$,
  - $F(x_2)=P(x\leq x_2)=p(x_1)+p(x_2)$,
  - $F(x_3)=P(x\leq x_3)=p(x_1)+p(x_2)+p(x_3)$,
  - ……..
  - $F(x_n)=P(x\leq x_n)= \sum_i p(x_i)=1$.

Variables aleatorias: Continuas

- Una variable aleatoria es continua, cuando puede tomar cualquier valor en un intervalo.
- Función de densidad: es una función continua que verifica las condiciones:
  - f(x)≥0.
  - $\int_{-\infty}^{+\infty} f(x)dx = 1.$
- El conocimiento de la función de densidad f(x) nos permite calcular las probabilidades a nuestra conveniencia:
  - $P(x \leq x_0) = \int_{-\infty}^{x_0} f(x)dx.$
  - $P(x_0 \leq x \leq x_1) = \int_{x_0}^{x_1} f(x)dx.$

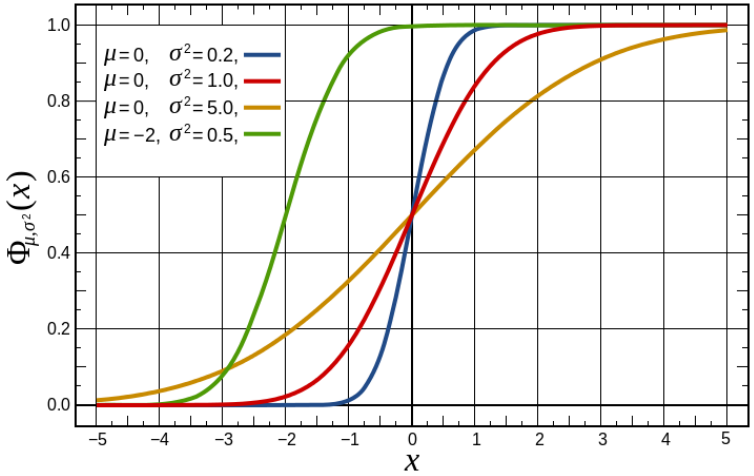
Variables aleatorias: Continuas

- La función de distribución
  - $F(x_0) = P(x \leq x_0) = \int_{-\infty}^{x_0} f(x)dx.$
  - $f(x) = \frac{dF(x)}{dx}$

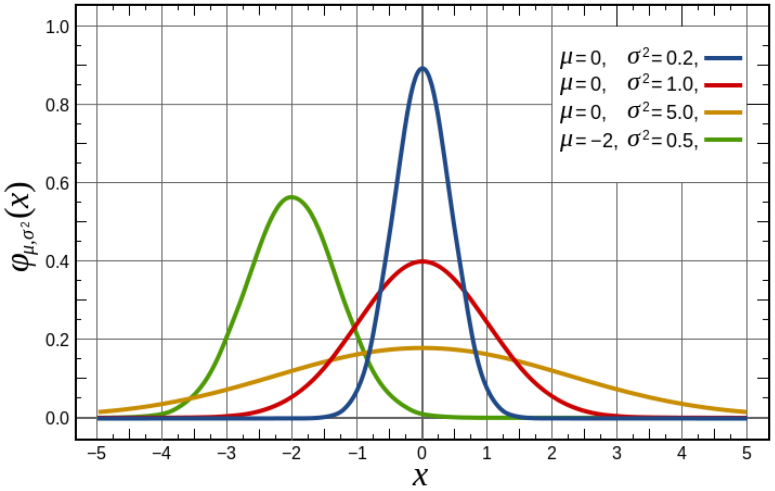
- Ejemplos para la distribución normal

Imágenes extraídas de https://es.wikipedia.org/wiki/Funci%C3%B3n_de_distribuci%C3%B3n

Función de Distribución



Función de Densidad



Variables aleatorias: Medidas de Centralización

- La mas utilizada es la media (μ) o esperanza matemática E(x):
  - $\mu=E(x)=\sum_i x_i\, p(x_i)$ (caso discreto).
  - $\mu = E(x) = \int_{-\infty}^{+\infty} xf(x)dx$ (caso continuo).
- La mediana: es aquel valor de la variable aletoria que divide la probabilidad total en dos mitades iguales:
  - Para variables discretas se define la mediana como el menor valor x_m de la variable aleatoria que satisface:
    - $F(x_m)\geq 0.5$, entonces x_m es la mediana.
  - Para variables continuas será el valor m definido por
    - $F(m)=0.5=P(x\leq m)$

Variables aleatorias: Medidas de Dispersión

- La medida de dispersión asociada a la media, es la desviación típica cuyo cuadrado es la varianza:
  - $Var(x) = \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$
- Para variables discretas las integrales se convierten en sumas y las probabilidades p(x) sustituyen a los elementos de probabilidad f(x)dx.
- El percentil p de una variable aleatoria x discreta es el valor x_p que verifica:
  - $P(x<x_p) \leq p$ y $P(x \leq x_p) \geq p$.
- Para variables continuas:
  - $F(x_p)=p$.

Variables aleatorias: Medidas de Dispersión

- Los **cuartiles** dividen la distribución en 4 partes iguales. La mediana coincide con el segundo cuartil, y con el percentil 0.5.
- La medida absoluta de dispersión más utilizada es el **rango intercuartílico**:
  - $RIC=Q_3-Q_1$, que representa la zona central donde se encuentra el 50% de la probabilidad.
  - Para distribuciones simétricas: $Q_2-Q_1=Q_3-Q_2$ y por lo tanto RIC es el doble de la distancia y los cuartiles.
- La medida de dispersión que se asocia a mediana es la Meda:
  - $Meda=Mediana(|x-Med(x)|)$

Variables aleatorias: Medidas de Dispersión

- La métrica ***Meda*** (Meda=Mediana(|x-Med(x)|)) para **distribuciones simétricas** cumple:
  - El 50% de las desviaciones son menores que Q_3-Med=Med-Q_1.
  - El otro 50% de las desviaciones son mayores que Q_3-Med=Med-Q_1.
  - Esto se entiende ya que todos los valores x que cumplen $Q_1 \leq x \leq Q_3$ tienen su meda menor que el rango: Q_3-Med=Med-Q_1. Y todos los x que están fuera del intervalo (Q1Q3), tienen su Meda mayor que el rango: Q_3-Med=Med-Q_1.
  - En consecuencia, solo para distribuciones simétricas, la Meda se puede calcular como Meda=Q_3-Med(x), i.e. la mitad de RIC.

Variables aleatorias: Otras Medidas Características

- En general definimos momento de orden k (m_k), respecto al origen de una variable continua aleatoria, x, como:
  - $m_k = \int x^k f(x) dx.$
- Si tomamos como origen la media μ:
  - $\mu_k = \int (x - \mu)^k f(x) dx.$
- El coeficiente de asimetría, CA, se define como:
  - $CA = \frac{\mu_3}{\sigma^3}$
- El apuntamiento o curtosis como:
  - $CA_p = \frac{\mu_4}{\sigma^4}$
- Y el coeficiente de variación como:
  - $CV = \frac{\sigma}{|\mu|}$

Variables aleatorias: Acotación de Tchebychev

- Si conocemos la media y desviación típica de una variable aleatoria discreta o continua dos permite calcular la proporción de la distribución que esta situada en el rango
  - $\mu \pm k\sigma$, siendo k una constante positiva.
- Al igual que hacíamos con las frecuencias relativas la acotación de Tchebychev siempre nos permite verificar que:
  - $P(\mu - k\sigma \leq x \leq \mu + k\sigma) \geq 1-(1/k^2)$, para cualquier valor de k.
  - Para cualquier variable aleatoria el intervalo $\mu \pm 3\sigma$, contiene al menos el 89% de la distribución.
  - Para cualquier variable aleatoria el intervalo $\mu \pm 4\sigma$, contiene al menos el 94% de la distribución.

Modelos: Modelos de distribución de probabilidad

Modelos de distribución de probabilidad

- El proceso de Bernoulli
- El proceso de Poisson
- Las distribuciones de duraciones de vida
- La distribución Normal
- La distribución Log-Normal

El proceso de Bernoulli

- Supongamos un experimento donde se observan elementos de una población que cumplen las siguientes características:
  - Cada observación se puede clasificar en dos posibles categorías:
    - Aceptable (A), con probabilidad de ocurrencia q=1-p.
    - Defectuoso (D), con probabilidad de ocurrencia p.
  - La proporción de elementos A y D es constante a lo largo de todo el proceso de observación.
  - Las observaciones son independientes:
    - Es decir la probabilidad de ocurrencia de los sucesos A y B no se modifican por el orden en el se observan los mismos.
- Este modelo aplica a poblaciones finitas de las que tomamos elemento al azar con reemplazamiento.
- También aplica a poblaciones conceptualmente infinitas en las cuales se generan los objetos con esas dos características con p y 1-p.

El proceso de Bernoulli: Distribución de Bernoulli

- Definimos una variable aleatoria de Bernoulli con espacio muestral {0,1}, por
  - $x = \begin{cases} 0\ si\ el\ elemento\ es\ aceptable \\ 1\ si\ el\ elemento\ es\ defectuoso \end{cases}$
- La función de probabilidades de esta variable, i.e. la probabilidad de que la variable aleatoria tome el valor 0 o 1 es:
  - $P(x)=p^x\ q^{1-x}$, $x=0,1$
- Su media es
  - $\mu=E(x)=0\ (1-p)+1\ p=p$
- La desviación típica
  - $DT(x)=\sigma=[(0-p)^2(1-p)+(1-p)^2p]^{1/2} = (pq)^{1/2}$.
- En esta distribución la media y la variabilidad depende de p, y la varianza será máxima cuando:
  - $d[p(1-p)]/dp=1-2p=0$, o lo que es lo mismo cuando $p=0.5$.

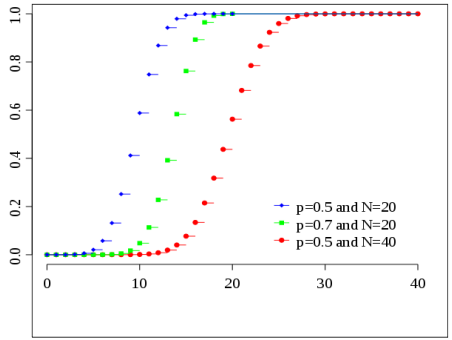
El proceso de Bernoulli: Distribución de Bernoulli

- La variable binomial en un **proceso de Bernoulli** se define como:
  - y = número de elementos defectuosos al observar n observaciones.
  - El espacio muestral de la nueva variable y es $\in(0,n)$.
- Supongamos que hemos realizado n observaciones delas cuales r son defectuosas y n-r son aceptables (da igual el orden por la hipótesis de independencia).
- La probabilidad de ocurrencia de un suceso DD…D AA…A, es:
  - pp…p (1-p)(1-p)…(1-p) $=p^r(1-p)^{n-r}$, (**conservación de proporción de 0`s t 1's**).
- Pero de cuantas formas podemos permutar las ocurrencias de cadenas de n objetos de las cuales r defectuosas y (r-n) aceptables. Son la permutaciones de n con r y (r-n) repetidos:
  - $\frac{n!}{r!(n-r)!} = \binom{n}{r}$, es decir el número combinatorio.

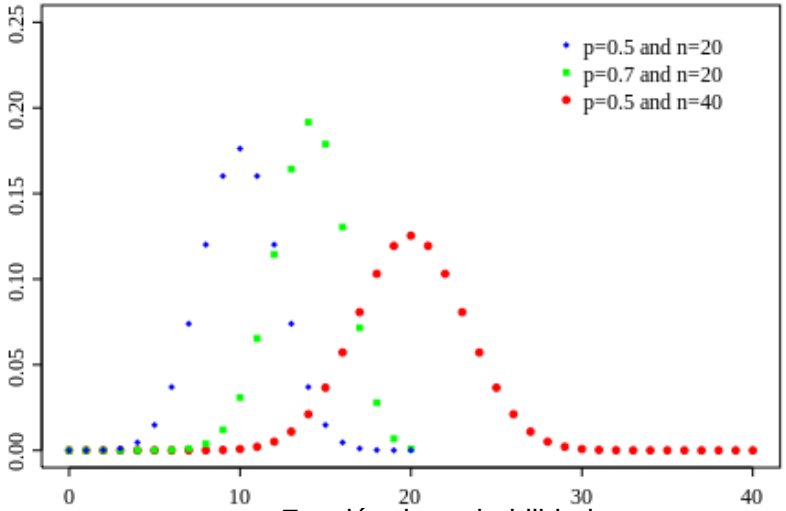
El proceso de Bernoulli: Distribución de Bernoulli

- Por tanto la probabilidad de que en una observación de n elementos r sean defectuosos viene determinado por:
  - $P(y = r) = \binom{n}{r} p^r (1-p)^{n-r}$, r=0, 1, …, n
- Es fácil comprobar que:
  - E[y]=∑rP(y=r)=np
  - DT(y)=σ=[∑(r-np)²P(y=r)]$^{1/2}$ = (npq)$^{1/2}$.

Imágenes extraídas de https://es.wikipedia.org/wiki/Distribuci%C3%B3n_binomial



Función de distribución de probabilidad



Función de probabilidad

El proceso de Bernoulli: Distribución geométrica

- Consideremos el mismo proceso de Bernoulli, pero en vez de contar número de defectos, no preguntamos por el número de elementos hasta el primer defectuoso, esto es la variable geométrica:
  - x=número de elementos hasta el primer defectuoso
- Para calcular la función de probabilidad, observemos que x tomará el valor n únicamente en el suceso:

$$\underbrace{AA\ldots A}_{n-1}\ D$$

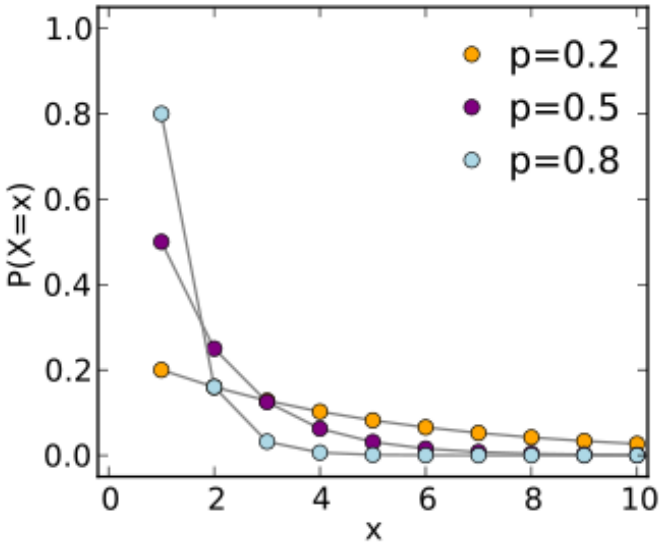
- Por tanto por la independencia:
  - $P(x=n)=p(1-p)^{n-1}$, n=1, 2, …; Con media y desviación $E[x]=1/p$ y $Var[x]=q/p^2$
- Observar que la variable geométrica tiene un conjunto ilimitado de posibles valores, aunque su probabilidad esta normalizada:
- $\sum_1^\infty P(x=n) = p\sum_{n=1}^\infty (1-p)^{n-1} = 1$ (recordar que $\sum_{n=0}^\infty ar^n = \frac{a}{1-r}$)

El proceso de Bernoulli: Distribución geométrica

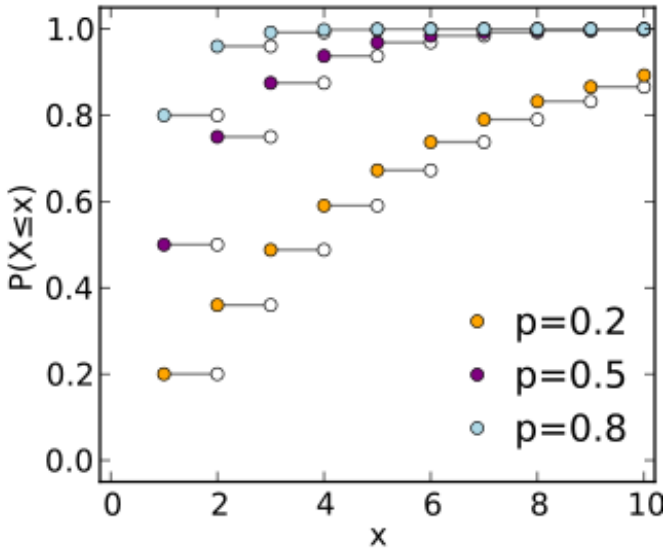
- La media y la desviación típica de la distribución geométrica se puede deducir fácilmente que es:
  - E[x]=1/p
  - Var[x]=q/p2

Imágenes extraídas de https://en.wikipedia.org/wiki/Geometric_distribution

Función de probabilidad



Función de distribución de probabilidad



El proceso de Poisson

- Supongamos un experimento en que observamos la aparición de sucesos puntuales sobre un soporte continuo.
  - p. ej. Averías de máquinas en el tiempo, llegadas de aviones a un aeropuerto, pedidos de una empresa, estrellas en el firmamento en cuadrículas del mismo tamaño, etc.
- Supondremos que el proceso que genera estos sucesos se caracteriza por:
  - Es **estable**: i.e. produce a largo plazo un número de sucesos constante por unidad de observación.
  - Los **sucesos aparecen de manera aleatoria** y de forma independiente, i.e. el proceso no tiene memoria:
    - Conocer el número de suceso en un intervalo no ayuda a predecir el número de sucesos en el siguiente.
- Este proceso es la generalización a un soporte continuo del proceso de Bernoulli.

El proceso de Poisson: La distribución de Poisson

- Dado el proceso anterior la variable aleatoria de Poisson se define como:
  - x=número de sucesos en un intervalo de longitud fija
- La distribución de Poisson aparece como límite de la distribución binomial si suponemos que el número de elementos observados es muy grande pero la probabilidad de observar la característica estudiada en cada elemento e muy pequeña:
  - Dividamos el intervalo de observación t, en n segmentos muy pequeños (así el número de segmentos es muy grande), y observemos el suceso en cada uno de ellos.
  - Si la probabilidad de este suceso, p, es muy pequeña entonces la aparición de dos o más sucesos es despreciable en el segmento.
  - Así el problema se puede plantear como la probabilidad de observar en n elementos si aparece el suceso estudiado o no, i.e. la distribución binomial.
- Es decir la distribución Poisson es un caso límite de esta distribución binomial cuando n tiende a infinito y p tienda a cero, pero de manera que el número medio de sucesos, np, permanezca constante.

El proceso de Poisson: La distribución de Poisson

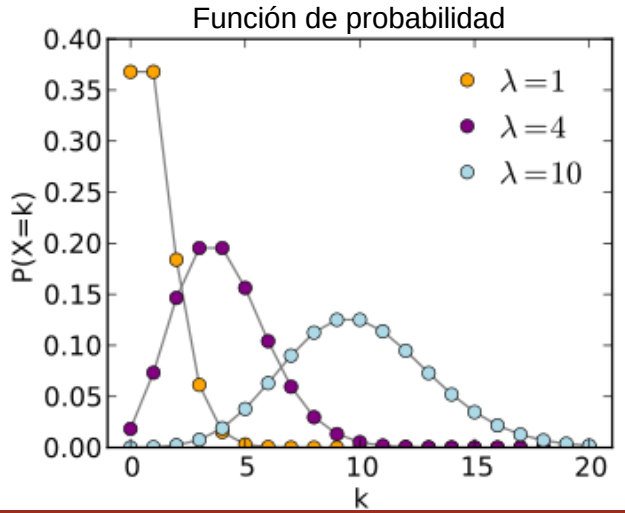
- Consideremos que el número de accidentes por 100 horas de conducción para un grupo de conductores es λ, y que los accidentes ocurren con el proceso de Poisson: aleatoria en independiente a lo largo del tiempo.
- Así la variable aleatoria x de Poisson será: # accidentes de accidentes en 100 horas de conducción.
- Como hemos dicho antes podemos convertir x en una binomial:
  - Considerando intervalos de tiempo muy pequeños (p. ej. Cada minuto), donde la probabilidad de ocurrencia de dos accidentes sea despreciable.
  - Así x puede considerarse como una variable binomial en un experimento con $n=100x60=6000$ repeticiones, de observar un accidente en un minuto.
  - La probabilidad p de accidente cumple: $E(x)= \lambda=np$, y $p= \lambda/n$.
- Así la distribución Poisson la obtendremos en el caso cuando n tiende a infinito y p tienda a cero, pero de manera que número medio de sucesos, $\lambda= np$, permanezca constante.

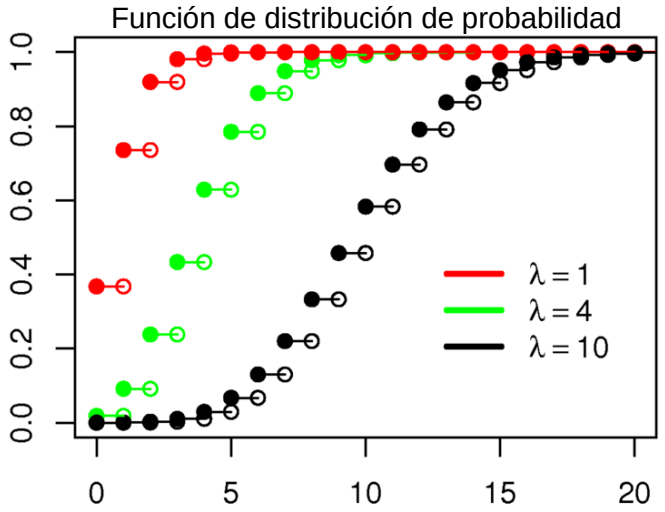
El proceso de Poisson: La distribución de Poisson

- En conclusión la distribución de Poisson puede aproximarse por:
  - $P(x=r)\binom{n}{r}\left(\frac{\lambda}{n}\right)^r\left(1-\frac{\lambda}{n}\right)^{n-r}$, así si tomamos los límites:
  - $\lim_{n\to\infty} P(x=r) = \frac{\lambda^r}{r!}\lim_{n\to\infty}\frac{n(n-1)\dots(n-r+1)}{\left(1-\frac{\lambda}{n}\right)^r n^r}\left(1-\frac{\lambda}{n}\right)^n$ $\to (n-\lambda)^r$
  - El primer termino del límite :$P(t_0<t\leq t_0+\Delta t \mid t>t_0)\frac{n}{(n-\lambda)}\frac{(n-1)}{(n-\lambda)}\dots\frac{(n-r+1)}{(n-\lambda)}=1$
  - El segundo: $\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$
  - Por tanto tenemos en el límite $n\to\infty$, con λ constante:
  - $P(x=r)=\frac{\lambda^r}{r!}e^{-\lambda}, r=0,1,2,\dots$ .Probabilidad de encontrar sucesos en una intervalo de longitud fija n que viene implícito en λ=np, donde p es la probabilidad del suceso observado.

El proceso de Poisson: La distribución de Poisson

- Las medidas características de la distribución de Poisson son:
  - $E[x]=Var[x]=\lambda$.
- Observar que los resultados son consistentes con la aproximación binomial:
  - La varianza de la binomial es npq y cuando $n\to\infty$ y $p\to 0$, pero con $np=\lambda=cte$, entonces $q\to 1$, lo que implica $npq\to np=\lambda$, que es la varianza de Poison.





Imágenes extraídas de https://es.wikipedia.org/wiki/Distribuci%C3%B3n_de_Poisson

El proceso de Poisson: La distribución de exponencial

- La variable exponencial resulta al considerar en un proceso de Poisson la variable continua:
  - t=tiempo entre la ocurrencia de dos sucesos consecutivos
- Esta variable toma valores en el intervalo $(0,\infty)$.
- Para obtener la función distribución de esta variable observemos que, la probabilidad de que en un proceso de Poisson de media λt no se produzca ningún suceso en un tiempo t es:
  - $P(t>t_0)=P[\text{cero sucesos en } (0,t_0)]=e^{-\lambda t_0}$, siendo λ la tasa media de sucesos por unidad de tiempo.
  - Así la función de distribución $F(t_0)=P(t\leq t_0)=1-e^{-\lambda t_0}$.
  - Cuya función de densidad será: $f(t)=dF(t)/dt=\lambda e^{-\lambda t}$, $\lambda>0$ y $t>0$.
  - La medidas características: $E[t]=1/\lambda=DT[t]$, $Var[t]=1/\lambda^2$

Distribuciones de duraciones de vida

- La distribución exponencial es el ejemplo más simple para distribuciones de variables aleatorios continuas que pueden tomar cualquier valor positivo no acotado.
- Se utilizan para modelar duración (vida de personas, animales o componentes físicos, duraciones de huelgas, periodos de desempleos, etc.), o el tamaño (rentas de familias, duración de discursos políticos, tamaño de yacimientos, etc.)
- Así supongamos que f(t) es la función de densidad de una variable continua positiva en $(0,\infty)$, que representa por ejemplo la duración de vida de ciertos elementos.

Distribuciones de duraciones de vida

- Así podemos calcular la **probabilidad de muerte o fallo** en el intervalo $(t_0, t_0+\Delta t)$ para los elementos que ya han vivido hasta t_0, mediante la probabilidad condicionada:
  - $P(t_0<t\leq t_0+\Delta t \mid t > t_0) = P(t_0<t\leq t_0+\Delta t, t>t_0)/P(t>t_0) = P(t_0<t\leq t_0+\Delta t)/P(t>t_0)$. A la función $P(t > t_0)$ se le suele llamar función de **fiabilidad o supervivencia** y se define como $P(t > t_0) = \int_{t_0}^{\infty} f(t)\, dt = 1 - F(t_0)$.
  - Arriba tomamos $P(t_0<t\leq t_0+\Delta t, t>t_0)=P(t_0<t\leq t_0+\Delta t)$ ya que la probabilidad conjunta de los dos sucesos
    - $t > t_0$ y $t_0< t \leq t_0+\Delta t$
  - Coinciden en la probabilidad del segundo, i.e. los intervalo solapan.
- Si definimos $F(t_0)$ como la función distribución de la variable t_0, recordemos que la función de distribución es $F(t_0) = P(t \leq t_0)$, así podemos aproximar:
  - $P(t_0< t \leq t_0+\Delta t \mid t > t_0) \approx f(t_0)\Delta t/1\text{-}F(t_0)$, ya que la probabilidad en un intervalo viene dado por el área que encierra la densidad de probabilidad, f(x). i.e. un rectángulo de base Δt y de altura $f(t_0)$.

Distribuciones de duraciones de vida

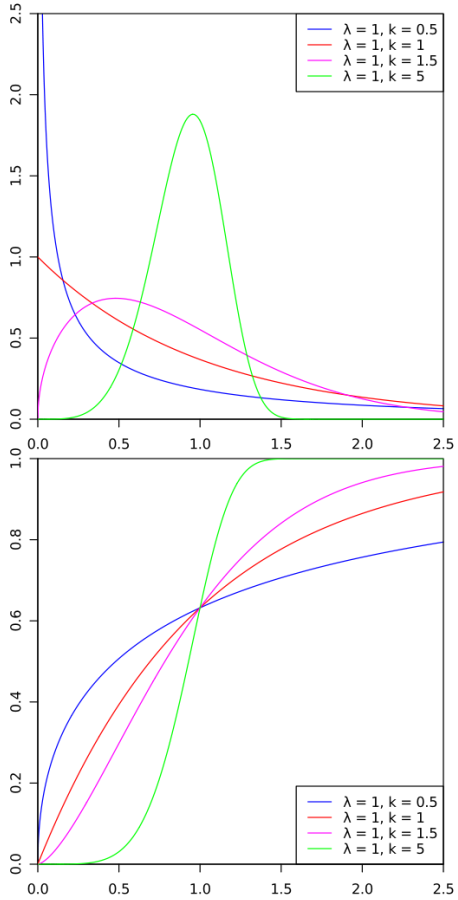
- Así podemos definir $\lambda(t)$ como la tasa de fallo en el limite $\Delta t \to 0$:
  - $\lambda(t)=f(t)/1-F(t)$, la tasa de fallos se define como $\lim_{\Delta t \to 0} P(t_0<t\leq t_0+\Delta t \mid t>t_0)/\Delta t$.
- Esta esta cantidad representa la **probabilidad de muerte en cada instante para los elementos que han sobrevivido hasta dicho instante**.
- Para obtener la función densidad f(t) en función de la tasa de fallo, $\lambda(t)$, sabiendo que F(0)=0, integramos:
  - $\Lambda(t) = \int_0^t \lambda(x)dx = \int_0^t \frac{f(x)}{1-F(x)}dx = -\ln[1-F(x)]_0^t = -\ln[1-F(t)]$.
- Así $1 - \mathrm{F(t)} = \exp\{-\Lambda(t)\}$ y $\mathrm{F(t)} = 1 - \exp\{\Lambda(t)\}$ y derivando
- $f(t) = \lambda(t)\exp\{-\Lambda(t)\}$, que es la forma habitual de las distribuciones continuas para variables positivas.
- Recordemos que f(t) era la función de densidad de una variable continua positiva en $(0,\infty)$, que representaba la duración de vida de ciertos elementos. Aquí la hemos calculado de la tasa de fallos y su función distribución.

Distribuciones de duraciones de vida

- La distribución exponencial se caracteriza por una tasa de fallo constante: la probabilidad de morir en cualquier intervalo no depende de la vida anterior, i.e. $f(t)= \lambda e^{-\lambda t}$, $\lambda>0$ y $t>0$, λ es constante.
- Darse cuenta que para la distribución exponencial $\Lambda(t) = \int_0^t \lambda \, dt = \lambda t + c$ y tomando arbitrariamente c=0, tenemos que $\Lambda(t) = \lambda t$, para particular de la distribución exponencial.
- Así la distribución exponencial se puede así como una distribución de duración de vida para una tasa de fallo λ constante: $f(t)=\lambda \exp\{-\Lambda(t)\}= \lambda e^{-\lambda t}$
- Por tanto la distribución exponencial es adecuada para describir aparición de muertes al azar, no debidas a desgaste o deterioro.
- No obstante podemos suponer una tasa de fallo no constante del tipo: $\lambda(t)=ht^{c-1}$
  - Tenemos una tasa de fallo que aumentará con el tiempo si $c>1$.
  - Será constante (distribución exponencial) si $c=1$ (distribución exponencial).
  - Y disminuirá si $c<1$.
  - Siendo la función densidad dada por: $f(t)= ht^{c-1} \exp\{(-h/c)t^c\}$.
  - Esta es **distribución de Weibull** (ver ejemplos en el libro).

Distribuciones de duraciones de vida: distribución de Weibull

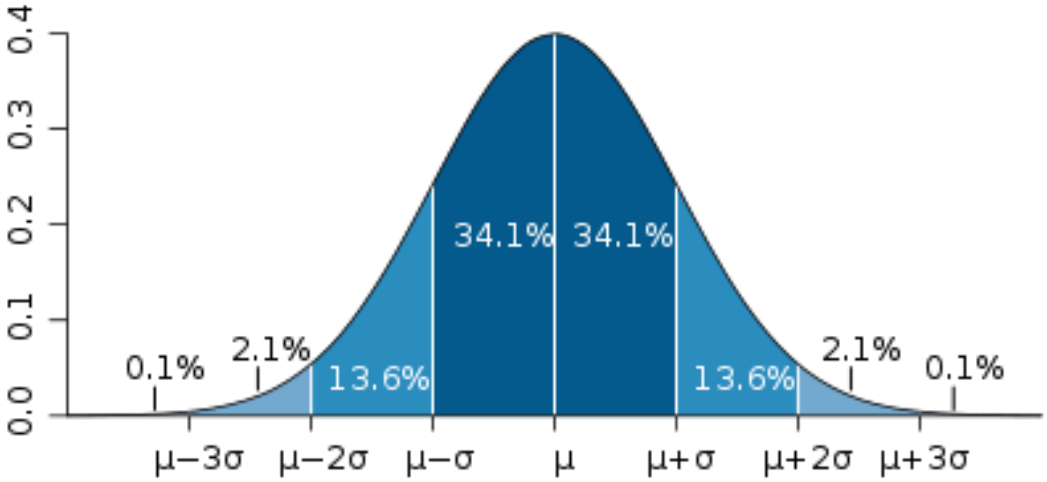
- Normalmente se pueden definir unos parámetros de forma y escala para la distribución de Weibull en la tasa acumulada, siendo el parámetro de **forma** k y parámetro de **escala** λ . Así se puede definir la función de tasas acumulada $\Lambda(x) = \left(\frac{x}{\lambda}\right)^{\mathrm{k}}$ y por tanto la función de distribución queda para x≥0: $f(x; \lambda, k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{\mathrm{k}-1} \exp\left\{-\left(\frac{x}{\lambda}\right)^{\mathrm{k}}\right\}$, recordar que la tasa de fallos es la derivad de la tasa de fallos acumulada, i.e. $\mathrm{d}\Lambda(x)/dx$.



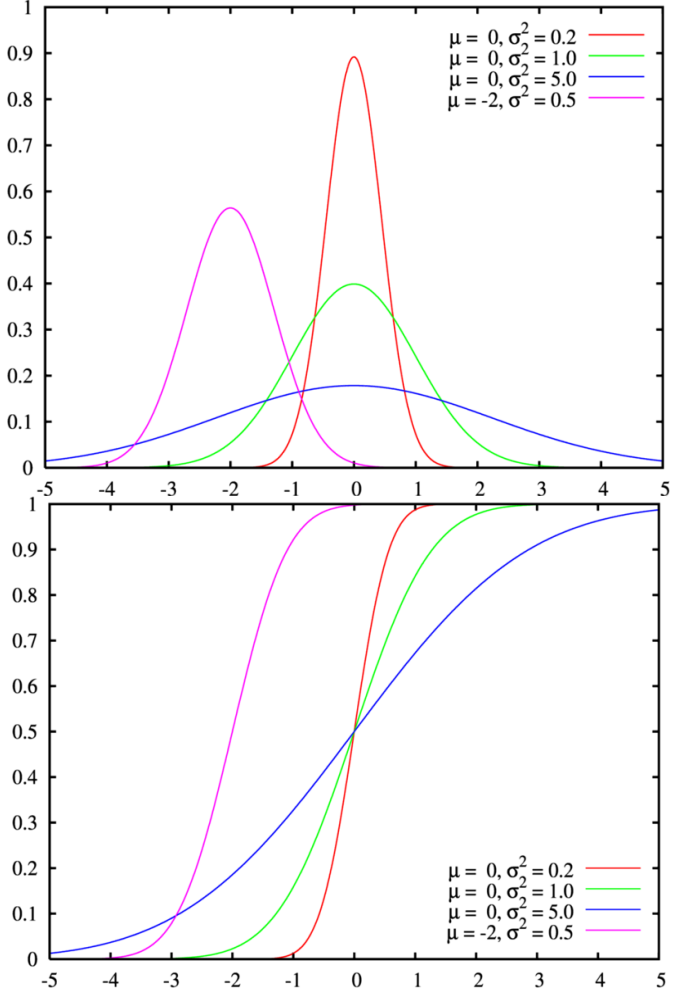
Imágenes extraídas de https://es.wikipedia.org/wiki/Distribuci%C3%B3n_de_Weibull

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.
- 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:
  - $F(x_0)=P(x\leq x_0)=P(\mu+\sigma z\leq x_0)=P(z\leq(x_0-\mu)/\sigma)=\Phi((x_0-\mu)/\sigma)$, 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.
- 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: por desigualdad de Tchebycheff se convierte en:
  - 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.
- 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 suma:
  - $Y = x_1 + \ldots + x_n$,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, y además la variable:
    - Y la variable $(Y - \Sigma\mu_i)/(\Sigma\sigma_i^2)^{0.5}$ tiende a una distribución N(0,1).
- El resultado anterior implica que si n es grande, podemos aproximar las probabilidades de 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

- Si la variable $Y = x_1 + \ldots + x_n$ es la suma de n variables de Bernoulli, x_i, que toman el valor1cuando 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 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, sea Y(0,T) la variable de Poisson que cuenta el número de sucesos entre 0 y T. 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 $x_i(t_{i-1},t_i)$ y se 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 crece.
- 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 $\lambda>5$.

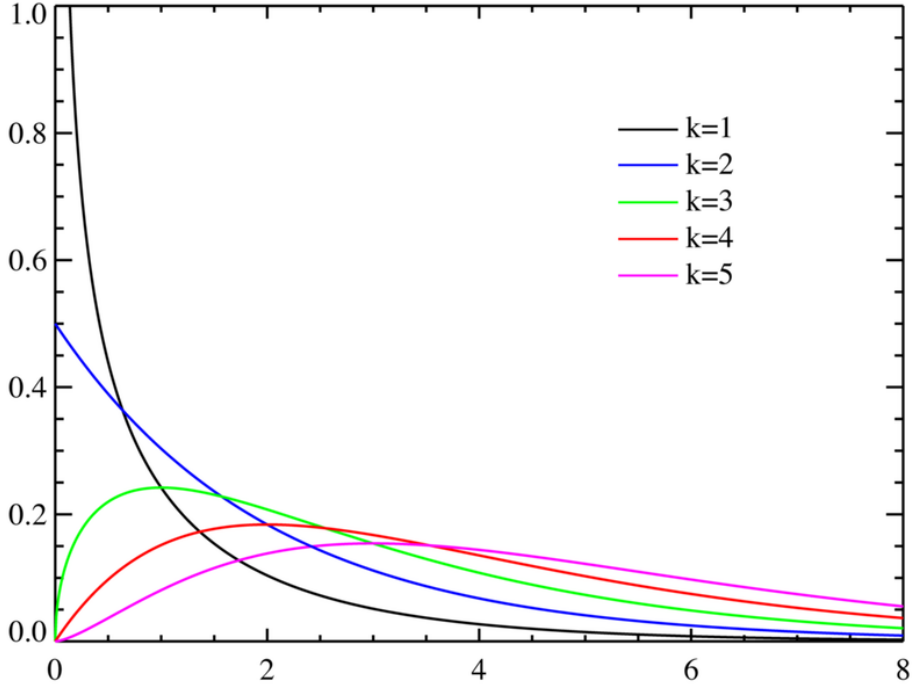
La distribución lognormal

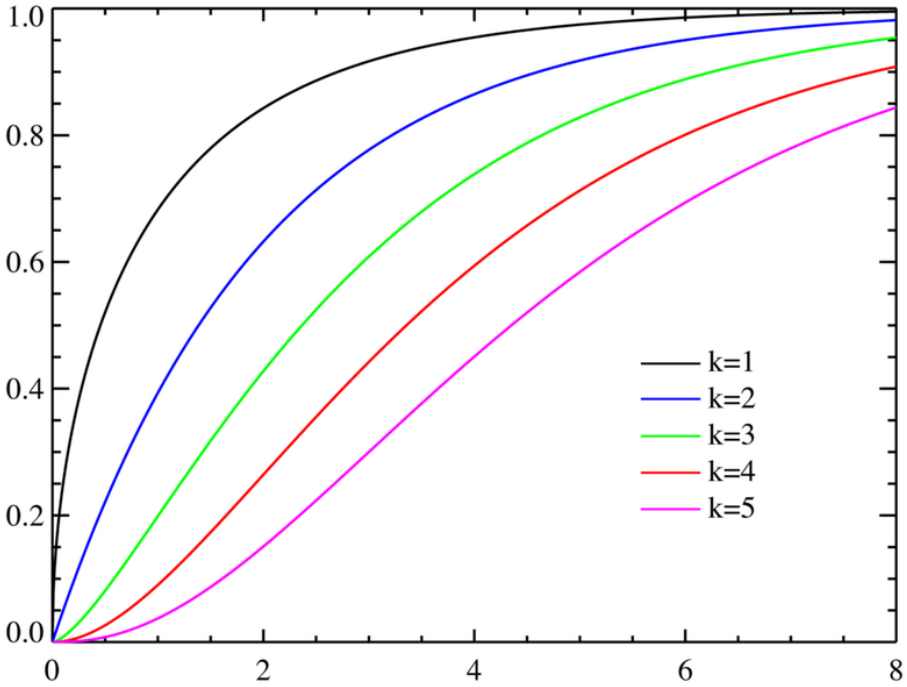
- Una consecuencia del TCL, es que si un determinado efecto es el producto de muchas causas, cada una de poca importancia en relación a las demás y además las causas son independientes entre ellas de tal forma que $y=x_1x_2\ldots x_n$
- Entonces el logaritmo de y seguirá una distribución normal por el 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 (grados de libertad).
- Esta distribución se denomina χ^2 con n grados de libertad.
- Los parámetros de la distribución se sacan fácilmente haciendo uso de la independencia de variables:
  - $E[z_1^2] = 1$, ya que $\sigma^2 = 1$ (recordemos que $\mathrm{E}[\mathrm{x}^k] = m_k = \int x^k f(x)dx$ y $\mathrm{E}[(x-\mu)^k] = \mu_k = \int (x-\mu)^k f(x)dx$).
  - $E[z_1^4] = 3$ (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





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

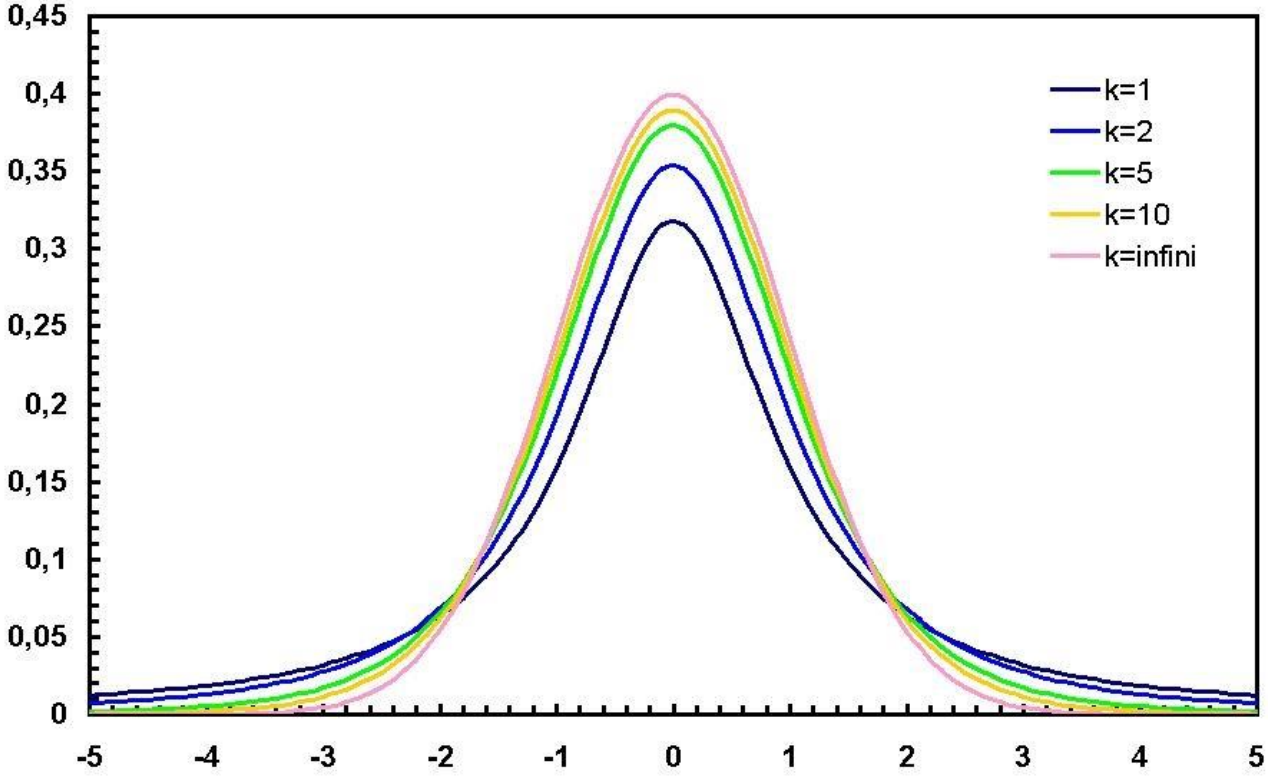
Distribuciones deducidas de la normal: distribución t de Student

- 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 $t_n = \frac{z}{\left(\frac{\chi_n^2}{n}\right)^{1/2}}$, siendo z una variable aleatoria normal estándar independiente del denominador que es la raíz cuadrada de Chi-cuadrado dividido por el número de sus grados de libertad.
- Así la distribución t de Student se genera 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 las variables x, 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.

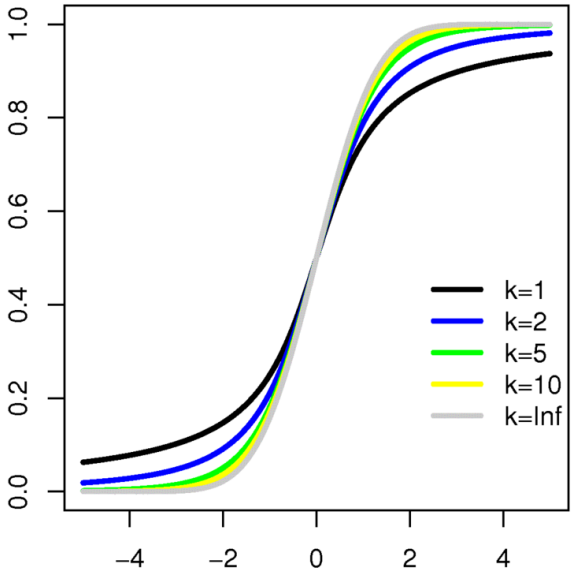
Distribuciones deducidas de la normal: distribución t de Student

- 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



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.
- Hemos definido 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).

Inferencia Estadística

- Inferencia Estadística
  - Estimación puntual
  - Estimación por intervalos
  - Estimación bayesiana
  - Contraste de hipótesis

Inferencia Estadística: Estimación puntual

Estimación puntual

- Introducción a la inferencia estadística
- Métodos de muestreo
- La estimación puntual
- Distribución de un estimador en el muestreo
- Propiedades de los estimadores
- Estimadores de máxima verosimilitud

Introducción a la inferencia estadística

- La creación de modelos probabilísticos es un caso típico de razonamiento deductivo donde se generan las hipótesis generales sobre el mecanismo que origina los datos, generando así las distribuciones de probabilidades que originan los datos:
  - Por ejemplo definimos e un proceso de Bernoulli la variable binomial como: y = número de elementos defectuosos al observar n observaciones. Ahora suponemos que hemos realizado n observaciones de las cuales r son defectuosas y n-r son aceptables (da igual el orden por la hipótesis de independencia). Con estas hipótesis dedujimos en capítulos anteriores la distribución de probabilidad binomial (razonamiento deductivo), y así con resto de distribuciones estudiadas anteriormente.
- El procedimiento inverso se realiza mediante la inferencia estadística, i.e. mediante las frecuencias observadas de una variable, extraer o inferir el modelo probabilístico que han generado los datos mostrando esas frecuencias (razonamiento inductivo).

Introducción a la inferencia estadística

- Existen muchos tipos de inferencia estadística:
  - Según el objetivo del estudio: muestreo frente a diseño.
    - **Describir** variables y sus relaciones entonces se utilizan técnicas de muestreo.
    - **Contrastar** relaciones entre variables y **predecir** valores futuros se utilizan técnicas de diseño experimental (se fijan valores de cierta variables y se miden la respuesta que inducen otras).

Introducción a la inferencia estadística

- Existen muchos tipos de inferencia estadística:
  - Por el método utilizado: métodos paramétricos v.s. no paramétricos.
    - **Paramétrico**: se supone que los datos provienen de una cierta distribución y se tienen muestras para estimar los parámetros de la misma.
    - **No paramétrico**: supones aspectos generales de la distribución (continua simétrica, etc.) y tratan de estimar o contrastar su estructura. Generalmente se estiman su forma mediante el suavizado los histogramas de los datos muestrales.

Introducción a la inferencia estadística

- Existen muchos tipos de inferencia estadística:
  - Por la información considerada: enfoque clásico v.s. bayesiano.
    - **Clásico**: los parámetros son cantidades fijas desconocidas (sin información sobre ellos), y la inferencia utiliza solo la información de los datos muéstrales.
    - **Bayesiano**: considera los parámetros como variables aleatorias y permite introducir información adicional sobre los mismos a través de una probabilidad a priori.

Métodos de muestreo: muestra y población

- **Población**: conjunto homogéneo de elementos en los que se estudia una característica dada. Normalmente no es posible estudiar toda la población:
  - Destrucción de los elementos: ej. estudiar la tensión de rotura de cables.
  - Lo elementos pueden existir conceptualmente, pero no en la realidad: ej. Población de piezas defectuosas que producirá una máquina.
  - Inviable económicamente estudiar toda la población.
  - El estudio llevaría tanto tiempo que sería impracticable.
- Se suele elegir un conjunto representativo que es la **muestra,** y si esta se selecciona bien podemos obtener una información similar de la población.
- La clave es seccionar la muestra **representativa** de la población.

Métodos de muestreo: muestreo aleatorio simple

- Una muestra es aleatoria simple si (m.a.s.):
  - Cada elemento de la población tiene la misma probabilidad de ser elegido.
  - Las observaciones se realizan con reemplazamiento (población idéntica en todas las extracciones).
- La primera condición asegura representatividad de la muestra (si A esta en el 20% y todos los elementos tienen idéntica probabilidad de ser seleccionados, la muestra tendrá un 20% también).
- La segunda se impone por simplicidad.

Métodos de muestreo: muestreo aleatorio simple

- En una muestra aleatoria cada observación tiene la distribución de probabilidad de la población.
- Sea la muestra observada $X'=(x_1, \ldots, x_n)$, donde x_i representa el valor de x en el elemento i-ésimo.
- Llamamos $f_1,\ldots,f_n$ a las funciones de densidad de esas variables que verifican en el muestreo aleatorio simple que $f_1=\ldots=f_n=f$.
- Como las observaciones son independientes en una muestra aleatoria simple, entonces la distribución conjunta de la muestra se puede poner $f_c(x_1, \ldots, x_n)= f_1(x_1)\, f_2(x_2)\ldots\, f_n(x_n) = f(x_1)\, f(x_2)\ldots\, f(x_n)$.

Métodos de muestreo: otros tipos de muestreo

- Muestreo **estratificado**: El método anterior se utiliza cuando la población es homogénea. Cuando se tiene información heterogénea de la población hay que dividir la población en estratos o clases, realizando un muestreo aleatorio simple dentro de cada estrato (ej. Encuestas de opinión, que se divide por sexo, edad, profesión, etc.).
- Supongamos k estratos de tamaños $N_1,\ldots N_k$, con $N=N_1+\ldots N_k$.
- La muestra que tomemos debe garantizar la presencia adecuada de cada estrato.
- Existen criterios básicos para dividir el tamaño total de muestra (n) entre los estratos (n_i):
  - Proporcional: $n_i=n(N_i/N)$.
  - Proporcional a la variabilidad del estrato: los estratos variables están más representados. Si σ_i es la variabilidad dele estrato i, entonces:

$$n_i=n\ (\sigma_i N_i)/(\sum^k_{i=1}\ \sigma_i N_i)$$

Métodos de muestreo: otros tipos de muestreo

- Muestreo por **conglomerados**: Hay situaciones en las cuales donde ni el muestreo aleatorio simple ni el estratificado pueden darse. En estos casos la población se encuentra agrupada en conglomerados, cuyo número se conoce.
- Ej. La población se distribuye en provincias, los habitantes de provincias en ciudades, etc.
- Si se supones los conglomerados independientes se pueden analizar con la metodología anterior.

Métodos de muestreo: otros tipos de muestreo

- Muestreo **sistemático**: Cuando los elementos están ordenados en listas. Supongamos que queremos una muestra n de una población N. Calculamos k=N/n. Se coge un elemento entre los primeros k, supongamos que el orden del elegido es n_1, tomamos a continuación los elementos n_1+k, n_1+2k, etc., hasta completar la muestra, es decir n veces (siendo k el número de grupos de tamaño n en la población).
  - Si el orden en la lista es al azar este procedimiento es equivalente al muestreo aleatorio simple.
  - Si el orden en la lista es de la forma que elementos cercanos son más similares que los más alejados, entonces se puede demostrar que este procedimiento cubre más homogéneamente toda la población, siendo más preciso que el muestreo aleatorio simple.

La estimación puntual: fundamentos

- **Supongamos que se observa una muestra aleatoria simple de una variable aleatoria x siguiendo una distribución conocida** como las que hemos estudiado: distribución normal, Poisson, etc.
- Lo que **no conocemos son los parámetros** de esas distribuciones conocidas.
- A las cantidades que estiman los parámetros de la distribución de la población a través de datos muestrales se le llaman **estimadores** estadísticos.
- ¿Cómo estimamos esos parámetros de los datos muestrales recogidos?

La estimación puntual: fundamentos

- En primera aproximación supondremos que **no tenemos ningún tipo de información del parámetro a ajustar**, ϑ, de la distribución supuesta.
- **Si hubiese algún tipo de evidencia sobre el parámetro a estimar se utiliza el enfoque bayesiano** (más adelante).
- Así el enfoque que vamos a ver ahora es el paramétrico, que dependiendo del tipo de variable a estudiar supondrá un modelo u otro a ajustar sus parámetros.

La estimación puntual: la identificación del modelo

- La primera operación a realizar con la muestra en un análisis descriptivo, para así ver si el modelo que consideramos es consistente con la muestra.
- Si tenemos muestra pequeñas (menos que 30), es más complicado y se suelen hacer ciertos tipos de gráficos que nos sacan de dudas.
- Para chequear visualmente si una muestra pequeña la podemos asociar a una distribución de Poisson:
  - Si **siguen una distribución de Poisson** entonces $E[f_{ob}(x)]=nP(x)=n\frac{\lambda^x}{x!}e^{-\lambda}$, siendo n el tamaño de la muestra.
  - Si sacamos logaritmos neperianos: $\ln E[f_{ob}(x)]=\ln n-\lambda+x\ln\lambda-\ln x!$, por lo tanto $\ln E[f_{ob}(x)]+\ln x!=\ln n-\lambda+x\ln\lambda=A+xB$.
  - Por tanto si dibujamos $\ln f_{ob}(x)+\ln x!$ respecto de x tiene que salir casi una recta si los valores esperados se distribuyen según una distribución de Poisson (ver ejemplo numérico en el libro).
  - La recta debería tener una pendiente $\ln\lambda$ y ordenada en el origen $\ln n-\lambda$.

La estimación puntual: la identificación del modelo

- Existen otros métodos para comprobar otras distribuciones, como por ejemplo para la distribución normal:
  - Que se puede utilizar un papel probabilístico normal (ver ejemplo en el libro) para dibujar los datos. Sino se ajustan a una recta, los datos no se distribuyen según una normal.
  - También se pueden usar los gráficos Q-Q plots con la misma idea anterior, que ya hemos visto anteriormente.
- En general los gráficos Q-Q plots se pueden utilizar con cualquier distribución de probabilidad, como ya comentamos anteriormente.

La estimación puntual: el método de los momentos

- Es el primer método que se utilizó para obtener el estimador de un parámetro de una distribución dada (formalizado por K. Pearson).
- Se toma como estimador de la varianza de la población la varianza de la muestra, de la media de la población la media muestral, y así sucesivamente con todos lo momentos que se quieran incluir en la estimación paramétrica.
- Se trata de estimar un vector de parámetros $\underline{\vartheta} = (\vartheta_1, \dots, \vartheta_k)$, cuyos componentes se pueden expresar en función de los k momentos de la población, m_k, siendo $\vartheta_1 = g_1(m_1,\dots,m_k)$, …, $\vartheta_k = g_k(m_1, \dots, m_k)$.

La estimación puntual: el método de los momentos

- Así estimamos los correspondientes momentos muestrales, $\widehat{\mathrm{m}}_1, \ldots, \widehat{\mathrm{m}}_k$, sustituyéndolos en el sistema de ecuaciones anteriores, obteniendo los parámetros estimados en la población: $\hat{\vartheta}_1, \ldots, \hat{\vartheta}_k$.
- Es muy importante estimar cual es la bondad de ajuste de estos estimadores de los parámetros y sus propiedades deseables. Esto es lo que vamos a ver a continuación.

Distribución de un estimador en el muestreo: concepto

- **Podemos ver el estimador como una variable aleatoria**, cuyo valor cambia de muestra en muestra.
- Por ejemplo supongamos la muestra (2, 4, 9, 1) de una distribución uniforme en el intervalo (0, b), para estimar b de la muestra (el valor esperado de una distribución uniforme en a,b, es a+b/2):
  - E[x]=(0+b)/2, por tanto $\hat{b} = 2\bar{x}$=2(2+4+9+1/4)=8, este estimador no es el más preciso, ya que pudiéramos elegir el máximo, 9. Además si tenemos otra muestra cambia.
- Consideremos una población de la que se toman muestras con remplazamiento de tamaño n y calculamos en cada muestra la media $\bar{x}$.

Distribución de un estimador en el muestreo: concepto

- Así si tomamos k muestras obtendremos en general k valores de las medias muestrales: $\bar{x}_1, \ldots, \bar{x}_k$. Si k es muy grande tendiendo a infinito los valores los valores $\bar{x}_i$ tendrán una distribución que llamaremos **distribución muestral de la media** en el muestreo.
- Esta distribución en el muestreo de un estadístico depende de:
  - La población base.
  - El tamaño de la muestra n.

Distribución de un estimador en el muestreo: concepto

- Recordemos que el estudio de determinadas características de una población se efectúa a través de diversas muestras que pueden extraerse de ella (no olvidar).
- En general el muestreo se puede realizar con o sin reposición.
- La población de partida puede ser infinita o finita.
- Una población finita en la que se efectúa muestreo con reposición podría considerarse infinita, aunque también una población muy grande puede considerarse como infinita.
- Aquí vamos a limitarnos a una población de partida infinita o a muestreo con reposición.

Distribución de un estimador en el muestreo: concepto

- Si consideramos todas las posibles muestras de tamaño n en una población, para cada muestra se puede calcular estadísticos como la media, desviación típica, proporción, etc., que variarán de una muestra a otra.
- Así obtenemos una distribución del estadístico que consideremos que se es lo que se llama **distribución muestral**, en general.
- Un **estadístico** (refiriéndose a datos muestrales, i.e. estadístico muestral) es una medida cuantitativa, derivada de un conjunto de datos de una muestra, con el objetivo de estimar o inferir características de una población o modelo estadístico.

Distribución de un estimador en el muestreo: concepto

- En general podemos definir un **Estadístico** como una función de los valores de la muestra. Es una variable aleatoria, cuyos valores dependen de la muestra seleccionada.
- Su distribución de probabilidad, se conoce como **Distribución muestral del estadístico**.
- Por ejemplo para el estadístico media de las diferentes muestras de una población se obtiene la distribución muestral de la media, para la varianza la distribución muestral de la varianza.

Distribución de un estimador en el muestreo: distribución en el muestreo de una proporción

- Supongamos una población donde observamos la presencia o no de un atributo. Y sea p la proporción desconocida de elementos con dicho atributo en la población.
- La distribución del muestreo del estimador de $\hat{p}$ en la muestra, viene determinada por la distribución binomial:
  - $P\left(\hat{p} = \frac{r}{n}\right) = PB(r) = \binom{n}{r} p^r (1-p)^{n-r}$, r=0, 1, …, n
- Por lo tanto la **probabilidad de que la proporción** en la muestra sea r/n es igual a la probabilidad de obtener r elementos con una característica determinada en una muestra de tamaño n, que es directamente la **distribución binomial**.

Distribución de un estimador en el muestreo: distribución en el muestreo de una proporción

- Así las propiedades de la distribución en el muestreo del estimador $\hat{p}$ vendrán dadas por la propiedades del valor esperado y la varianza del estimador de la proporción:
  - **E[$\hat{p}$]**=E[r/n]=(1/n)E[r]=np/n=**p,**
    - (recordar que en una binomial E[r]=np).
  - **Var[$\hat{p}$]**=E[r/n] =$(1/n)^2$ Var[r]= **pq/n,**
    - (recordar que en una binomial Var[r]=npq).

Distribución de un estimador en el muestreo: distribución en el muestreo de una proporción

- Cuando k es grande, la distribución de muestreo de $\hat{p}$ será aproximadamente normal con la media y varianza de las dos expresiones anteriores, ya que es un caso particular de la distribución muestral de una media, ya que $\hat{p}$ se calcula por: $\hat{p}=(x_1+\ldots+x_n)/n$ y entonces se puede aplicar aplica el TCL.
- Cada x_i toma el valor 1 si el elemento tiene el atributo estudiado, y 0 en cualquier otro caso.
- **Así $\hat{p}$ es la media muestral de las variables de Bernoulli, x_i.**

Distribución de un estimador en el muestreo: distribución muestral de la media

- Para calcular la distribución muestral de la media tenemos que tener en cuenta que cada muestra de tamaño n que podemos extraer de una población proporciona una media.
  - Podemos considerar cada una de estas medias como valores de una variable aleatoria y podemos estudiar su distribución que llamaremos **distribución muestral de las medias**.
- Vamos a calcular la media y varianza de la distribución muestral de la media, en el caso general en el que la variable aleatoria x tiene media μ y varianza σ^2.

Distribución de un estimador en el muestreo: distribución muestral de la media

- Para el cálculo de distribución muestral de la media suponemos que cada muestra es de tamaño n, y suponemos que todas las variables x_i de una muestra aleatoria simple tiene la misma distribución de la población.
- Así el valor esperado la distribución muestral de las medias y su varianza vienen determinados por las expresiones:
  - $E[\bar{x}]=E[(1/n)\sum x_i]=1/n \sum E[x_i]= 1/n \sum \mu= \mu.$
  - $Var[\bar{x}]=(1/n)^2 \sum Var[x_i]= n\sigma^2 /n^2= \sigma^2 /n.$
- Hemos aplicado el hecho de que para la variable aleatoria x_i se cumple que $E[x_i]=\mu$ y $Var[x_i]= \sigma^2$, como hemos dicho antes.

Distribución de un estimador en el muestreo: distribución muestral de la media

- Resumiendo, al tomar una muestra de tamaño n de una variable con media μ y varianza σ^2 y distribución cualquiera, la distribución muestral de la media verifica que $E[\bar{x}]=\mu$ y $Var[\bar{x}]=\sigma^2/n$.
- En el caso de la distribución de muestreo de la proporción $E[\hat{p}]=p$ y $Var[\hat{p}]=pq/n$, es un caso especial de este que acabamos de ver con media p y varianza pq.
- Si tenemos una población normal $N(\mu, \sigma)$ y extraemos de ella muestras de tamaño n, la distribución muestral de medias sigue también una distribución normal $N(\mu, \sigma/(n)^{0.5})$.
- Si la población no sigue una distribución normal pero n>30, aplicando el llamado el TCL la distribución muestral de medias se aproxima también a la normal anterior.

Distribución de un estimador en el muestreo: distribución muestral de la varianza

- Para la **distribución muestral de varianzas,** podemos seguir los mismos razonamientos anteriores y suponemos de nuevo una variable aleatoria x que se observa en la muestra que tiene media μ y varianza σ^2.
- Se puede calcular que esperanza de la distribución de varianza de la muestra es $E[s^2]=\sigma^2(n-1)/n$. En consecuencia el valor medio de s^2 es menor que σ^2, aunque la diferencia tiende a cero a aumentar el tamaño de la muestra n.
- Se puede definir la varianza muestral corregida como $\hat{s}^2=(n/n-1)\ s^2$, de tal forma que $E[\hat{s}^2]= \sigma^2$.
- Estas propiedades se verifican siempre, cualquiera que sea la distribución de la variable x.
- Así se pueden calcular la distribución de cualquier estimador en el muestreo, en libro vienen más ejemplos (capítulo 7).

Propiedades de los estimadores: centrado o insesgado

- Diremos que un **estimador** $\hat{\vartheta}$ es **centrado** o insesgado para ϑ si para cualquier tamaño muestral tenemos que $E[\hat{\vartheta}] = \vartheta$.
- Cuando no es centrado se define el **sesgo** del estimador como $\text{sesgo}\left(\hat{\vartheta}\right) = E\left[\hat{\vartheta}\right] - \vartheta$.
- **Pueden existir muchos estimadores centrados para un parámetro:**
  - Para estimar μ en una distribución cualquiera todos los estimadores del tipo siguiente son centrados: $\hat{\mu} = a_1 x_1 + \cdots + anxn$ con $\sum a_i = 1$.
- Anteriormente hemos comprobado que $\bar{x}$ (y como caso particular el estimador de la proporción $\hat{p}$) es siempre centrado para estimar μ.

Propiedades de los estimadores: centrado o insesgado

- También hemos visto que s^2 es no es centrado para estimar σ^2 (recordar que lo corregimos).
- Una ventaja fundamental de los **estimadores centrados** es que **los podemos combinar para obtener nuevos estimadores centrados:**
  - Si tenemos dos muestras independientes y calculamos en cada una de ellas un estimador centrado $\hat{\vartheta}_i$ para un para un parámetro determinado, cualquier estimador del tipo $\hat{\vartheta}_t = a_1\hat{\vartheta}_1 + a_2\hat{\vartheta}_2$, con $a_1 + a_2 = 1$, es un estimador centrado.
- Los **estimadores centrados no tienen porque ser los mejores**, alguna veces es preferible tener uno sesgado con poca varianza, que uno no sesgado pero con mucha varianza (**eficiencia de estimadores**).

Propiedades de los estimadores: eficiencia o precisión

- La **eficiencia** o precisión se define en función del **inverso de la varianza**: $\text{precisión}\left(\hat{\vartheta}\right) = 1/Var\left[\hat{\vartheta}\right]$.
- Diremos que un estimador $\hat{\vartheta}_2$ es más eficiente que un estimador $\hat{\vartheta}_1$ si para cualquier tamaño muestral se cumple que $\text{Var}\left(\hat{\vartheta}_2\right) \leq \text{Var}\left(\hat{\vartheta}_1\right) \Leftrightarrow \text{Efic}\left(\hat{\vartheta}_2\right) \geq \text{Efic}\left(\hat{\vartheta}_1\right)$.
- Se define la eficiencia relativa de $\hat{\vartheta}_2$ respecto a $\hat{\vartheta}_1$ al cociente entre sus eficiencias: $\text{ER}\left(\hat{\vartheta}_2/\hat{\vartheta}_1\right) = \text{Efic}\left(\hat{\vartheta}_2\right)/\text{Efic}\left(\hat{\vartheta}_1\right) = \text{Var}\left(\hat{\vartheta}_1\right)/\text{Var}\left(\hat{\vartheta}_2\right)$.
- La eficiencia de estimadores esta completamente ligada a la varianza de los mismos.
- Mirar en el libro como combinar linealmente estimadores centrados para minimizar la varianza y el ejemplo 7.4.

Propiedades de los estimadores: error cuadrático medio

- Muchas veces se nos presenta el problema de elegir entre dos estimadores uno centrado y con varianza no muy grande y otro sesgado y con varianza un poco más pequeña. En estos casos se elige aquel que tiene el menor error cuadrático medio con el parámetro que esta intentando estimar.
- Así tenemos $\mathrm{ECM}(\vartheta) = E\left[\left(\hat{\vartheta} - \vartheta\right)^2\right]$, tomando el promedio respecto a la distribución en el muestreo del estimador.
- Se puede demostrar que $\mathrm{ECM}(\vartheta) = \left[sesgo\left(\hat{\vartheta}\right)\right]^2 + Var\left(\hat{\vartheta}\right)$.

Propiedades de los estimadores: consistencia y robustez

- Cuando lo único que podemos tener son estimadores segados y no con mucha eficiencia, se le pide al estimador que sea **consistente**.
- Un estimador se dice que es consistente si cuando crece el tamaño de la muestra el estimador tiende al parámetro que se está estimando: $\lim_{n\to\infty} E[\hat{\vartheta}_n] \to \vartheta$.
- Es decir la esperanza del estimador es asintóticamente el valor del parámetro.

Propiedades de los estimadores: consistencia y robustez

- En este caso la varianza del estimador va a cero también con el tamaño de la muestra: $\lim_{n\to\infty} Var[\hat{\vartheta}_n] \to 0$.
- Un buen estimador es **robusto** para un parámetro ϑ en el modelo f(x), si variando débilmente el modelo este estimador experimenta una pequeña modificación (ver ejemplo en el libro, capítulo 7 de contaminación de un modelo normal como decrece la eficiencia).

Estimadores de máxima verosimilitud: distribución conjunta de la muestra

- Los conceptos de funciones de verosimilitud se deben a Fisher, y es fundamental en inferencia estadística.
- Este concepto se define a partir de la distribución conjunta de la muestra.
- Supongamos una variable discreta x con distribución $P(x; \vartheta)$ que es conocida.
- Supongamos que tomamos muestras independientes de tamaño n, representando está por el vector **X**.
- Así podemos definir la distribución conjunta de la muestra en función de esta variable, y para el caso de una muestra aleatoria simple tenemos:
  - $P(\mathbf{X} = \mathbf{X}_0) = P(x_1 = x_{10}, x_2 = x_{20}, \ldots, x_n = x_{n0}) = P(x_{10}) \ldots P(x_{n0})$
- Así conociendo la distribución $P(x; \vartheta)$ podemos calcular fácilmente la probabilidad de cualquier muestra.

Estimadores de máxima verosimilitud: distribución conjunta de la muestra

- En el **caso continuo** (función de densidad $f(x; \vartheta)$), la probabilidad del intervalo $x_1 - 1/2$, $x_1 + 1/2$, la podemos aproximar por el rectángulo de altura $f(x_i)$ y base unidad:
  - $\mathbf{P(x_i) = f(x_i) \cdot 1}$
- Por tanto la probabilidad de la muestra aleatoria simple:
  - $P(x_1, ..., x_n) = \prod f(x_i)$
- Así la función de densidad conjunta de la muestra $f(x_1, ..., x_n)$ se interpreta como la probabilidad de obtener los valores muestrales $x_1 \pm 0{,}5$, ..., $x_n \pm 0{,}5$.

Estimadores de máxima verosimilitud: la función de verosimilitud

- Sea una variable aleatoria continua x con función de densidad que $f(x \mid \vartheta)$ para indicar que depende de un vector de parámetros ϑ. Es decir dado que conozco ϑ, representa cual es la función de densidad de la variable aleatoria x.
- Si tenemos una muestra aleatoria simple $\mathbf{X} = (x_1, \ldots, x_n)$, entonces la función de densidad conjunta de la muestra es:
  - $f(\mathbf{X} \mid \vartheta) = \prod f(x_i \mid \vartheta)$
- Es decir cuando conozco ϑ la expresión anterior determina la probabilidad de aparición de cada muestra.

Estimadores de máxima verosimilitud: la función de verosimilitud

- En inferencia para un problema de estimación se conoce un valor particular de una muestra **X**, siendo desconocido el parámetro ϑ.
- Así si sustituimos **X** por el valor observado de una muestra, $\mathbf{X}_0 = (x_{10}, \ldots, x_{n0})$, entonces la función $f(\mathbf{X}_0 | \vartheta)$ puede ser vista como una función del parámetro.
- Es decir $f(\mathbf{X}_0 | \vartheta)$ puede ser visto y proporciona, para cada valor de ϑ, la probabilidad de obtener el valor muestral $\mathbf{X}_0$ para ese ϑ.

Estimadores de máxima verosimilitud: la función de verosimilitud

- Así esta nueva función que obtenemos cuando variamos ϑ, mientras mantenemos $\mathbf{X}_0$ fijo (no variamos la muestra), define la **función de verosimilitud**, $\ell(\vartheta \mid \boldsymbol{X})$ (dado que conozco la muestra como varia la probabilidad en función del parámetro ϑ)
  - $\ell(\vartheta \mid \boldsymbol{X})$, o $\ell(\vartheta)$: Es decir $\ell(\vartheta \mid \boldsymbol{X})=\ell(\vartheta)=f(\mathbf{X}_0 \mid \vartheta)$, con $\mathbf{X}_0$ fijo y ϑ variable.
- La óptica cambia, en vez de tener un parámetro fijo ϑ y calcular para ese parámetro la probabilidad de obtener distintas muestras **X**, lo que fijamos es una determinada muestra $\mathbf{X}_0$ y estimamos que valor del parámetro ϑ hace más verosímil la muestra que se observa $\mathbf{X}_0$.

Estimadores de máxima verosimilitud: la función de verosimilitud

- Este enfoque cambia completamente la forma de la función.
- Si tenemos una variable x que distribuye según una Poisson:
  - $P(x = r) = \frac{\lambda^r}{r!} e^{-\lambda}, r = 0, 1, 2, ...$, y observamos el valor de la muestra x=5, entonces $\ell(\lambda) = \frac{\lambda^5}{5!} e^{-\lambda}$, es la función de verosimilitud para una muestra de un solo valor de x=5.

Estimadores de máxima verosimilitud: la función de verosimilitud

- Esta función $\ell(\lambda)$ es continua en λ y proporcional a la probabilidad de observar x=5 para cada valor posible de λ.
- El valor de la verosimilitud no es único: $\ell(\vartheta_1) = f(\boldsymbol{X}_0|\vartheta_1) > f(\boldsymbol{X}_0|\vartheta_2) = \ell(\vartheta_2)$.
- Esto quiere decir que a la vista de los datos muestrales el valor del parámetro ϑ_1 es más verosímil que el valor del parámetro ϑ_2 ya que la probabilidad de obtener la muestra observada $\mathbf{X}_0$ es mayor con ϑ_1 que con ϑ_2.

Estimadores de máxima verosimilitud: la función de verosimilitud

- Observar que la verosimilitud tiene unidades, las de la variable x, entonces la diferencia de verosimilitudes no tiene sentido ya que varia arbitrariamente en función de las unidades de la variable x.
- Por lo tanto para comparar verosimilitudes lo mejor es el cociente de las mismas ya que este es invariante frente a las diferentes unidades de la variable:
  - $\ell(\vartheta_1 \mid \boldsymbol{X})/\ell(\vartheta_2 \mid \boldsymbol{X})$, este cociente es invariante hacia cambios de escalas en la variable x que se esta observando.
  - El cociente $\ell(\vartheta_1)/\ell(\vartheta_2)$ se pude sustituir por la diferencia de logaritmos: $\ln\ell(\vartheta_1) - ln\ell(\vartheta_2)$
- Así se puede definir la **función soporte** por el logaritmo de la verosimilitud: $\mathrm{L}(\vartheta) = ln\ell(\vartheta)$.

Estimadores de máxima verosimilitud: la función de verosimilitud

- Se define como la discriminación contenida en la muestra **X** entre ϑ_1 y ϑ_2 a la siguiente expresión (diferencia de soporte de ambos valores): $\mathrm{L}(\vartheta_2) - \mathrm{L}(\vartheta_1) = ln\ell(\vartheta_2) - ln\ell(\vartheta_1)$.
- Si ϑ es un parámetro cuyos valores posibles pertenecen a un intervalo ϑ_1 y ϑ_2, llamaremos **discriminación relativa** entre ϑ_2 y ϑ_1 a:
  - $\mathrm{L}(\vartheta_2) - \mathrm{L}(\vartheta_1)/\vartheta_2 - \vartheta_1 = ln\ell(\vartheta_2) - ln\ell(\vartheta_1)/\ \vartheta_2 - \vartheta_1$.
- En el limite cuando $\vartheta_2 \to \vartheta_1$ obtenemos la tasa de discriminación para la muestra **X** respecto el parámetro ϑ valorada en el punto ϑ_1
- $\mathrm{d}(\vartheta_1) = \lim_{\vartheta_2 \to \vartheta_1} \frac{\mathrm{L}(\vartheta_2) - L(\vartheta_1)}{\vartheta_2 - \vartheta_1} = \left.\frac{\mathrm{dL}(\vartheta)}{\mathrm{d}\vartheta}\right|_{\vartheta\ =\ \vartheta_1}$, fue introducida por Fisher, que la denominó "Score".

Estimadores de máxima verosimilitud: la función de verosimilitud

- Si este "score" cumple que $d(\vartheta_1)>0$, la verosimilitud aumenta para valores superiores a ϑ_1,
  - es decir, la muestra tiene mayor probabilidad de ocurrir con valores mayores que ϑ_1,
- Mientras que si $d(\vartheta_1)< 0$ el razonamiento es el contrario, la verosimilitud aumenta para valores inferiores de ϑ_1,
  - es decir, la muestra tiene mayor probabilidad de ocurrir con valores menores que ϑ_1.

Estimadores de máxima verosimilitud: la función de verosimilitud

- Resumiendo:

1. La función de verosimilitud es la herramienta básica que nos permite juzgar la compatibilidad entre los valores muestrales observados y los posibles valores del parámetro de la distribución de probabilidad.
2. Si queremos comparar dos posibles valores del parámetro, ϑ, se debe utilizar el cociente de sus verosimilitudes, y no su diferencia, ya que la diferencia depende de la escala de medida de las variables, como hemos dicho antes.

Estimadores de máxima verosimilitud: la función de verosimilitud

- Por ejemplo para estimar el parámetro, λ, de Poisson de la muestra observada $x_1,\ldots, x_n$ hacemos:
  - $P(x) = \frac{\lambda^x}{x!} e^{-\lambda}$, distribución de Poisson.
  - $\ell(\lambda) = \Pi\ P(xi|\lambda) = P(x_1|\lambda)P(x_2|\lambda) \ldots P(xn|\lambda) = \frac{\lambda^{\Sigma x_i}}{\Pi x_i!} e^{-n\lambda}$
- Como el término $1/(\Pi x_i!)$ es una constante la podemos eliminar y escribir la función de verosimilitud para la distribución de Poisson como:
  - $\ell(\lambda) = e^{-n\lambda}\ \lambda^{\Sigma x_i} = e^{-n\lambda}\ \lambda^{n\bar{x}}$
  - Así la función soporte será para la distribución de Poisson vendrá dada por: $L(\lambda) = -n\lambda + n\bar{x}\ ln\lambda$ (mirar los ejercicios para el resto de distribuciones).

Estimadores de máxima verosimilitud: el método de máxima verosimilitud

- Una vez que tenemos calculada un función de verosimilitud para un vector de parámetros $\boldsymbol{\vartheta}$, $\ell(\boldsymbol{\vartheta})$, un procedimiento intuitivo para estimar los parámetros de la distribución a partir de los valores observados muestralmente es maximizar el valor del parámetro que sea más verosímil, es decir el que maximice la verosimilitud.
- Así podemos resolver el sistema de ecuaciones:
  - $\partial\ell(\boldsymbol{\vartheta})/\partial\vartheta_1 = 0, \ldots, \partial\ell(\boldsymbol{\vartheta})/\partial\vartheta_p = 0$
- El valor que resuelve el sistema de ecuaciones, $\widehat{\boldsymbol{\vartheta}}$, corresponderá a un máximo si el valor de la matriz hessiana de segundas derivadas en ese punto es definida negativa: $\mathrm{H}\left(\widehat{\boldsymbol{\vartheta}}\right) = \left(\partial^2\ell(\boldsymbol{\vartheta})/\partial\vartheta_i\partial\vartheta_j\right)_{\boldsymbol{\vartheta}\,=\,\hat{\boldsymbol{\vartheta}}}$.

Estimadores de máxima verosimilitud: el método de máxima verosimilitud

- A la hora de hacer los cálculos de estimadores máximo-verosímiles (MV) se obtienen derivando la función soporte: $\mathrm{L}(\vartheta) = ln\ell(\vartheta)$, ya que la transformación logarítmica es monótona y por tanto tiene el mismo máximo.
- Recordemos que la derivada de la función soporte la habíamos definido como la tasa de discriminación: $\mathrm{d}(\vartheta_1) = \lim_{\vartheta_2 \to \vartheta_1} \frac{\mathrm{L}(\vartheta_2) - L(\vartheta_1)}{\vartheta_2 - \vartheta_1} = \left.\frac{\mathrm{dL}(\vartheta)}{\mathrm{d}\vartheta}\right|_{\vartheta = \vartheta_1}$, así podemos definir el estimador máximo-verosímil como aquel valor de los parámetros para los que se anulan la tasa de discriminación de la muestra.

Estimadores de máxima verosimilitud: Ejemplo

- Para estimar los parámetros de una normal de la muestra observada $x_1,\ldots, x_n$:
  - $f(x)=1/\sigma(2\pi)^{0.5}\ \exp\{-(1/2\sigma^2)(x-\mu)^2\}$, densidad de probabilidad normal.
  - $\ell(\mu,\sigma^2) = \Pi\ f(xi|\mu,\sigma^2) = f(x_1|\mu,\sigma^2)f(x_2|\mu,\sigma^2)\ldots f(xn|\mu,\sigma^2) =$
    $\Pi_i(1/\sigma(2\pi)^{0.5})\ \exp\{-(1/2\sigma2)(x_i-\mu)^2\} =$
    $(1/(\sigma(2\pi)^{0.5})^n)\ \exp\{-(1/2\sigma^2)\Sigma(xi-\mu)2\}.$
- Así la función soporte será para la distribución normal vendrá dada por: $L(\mu,\sigma^2) = -n\ln(\sqrt{2\pi}\,\sigma)-(1/2\sigma^2)\Sigma(xi-\mu)2 =$
  $-(n/2)\ln(2\pi\sigma^2)-(1/2\sigma^2)\Sigma(xi-\mu)2.$
- Así ahora derivamos la función soporte de la muestra e igualamos a cero para sacar cuales son los estimadores de máxima verosimilitud para μ,σ^2 de la población.

Estimadores de máxima verosimilitud: Ejemplo

- $\partial L(\mu,\sigma^2)/\partial\mu = 0 = (2/2\sigma^2)\Sigma(xi-\mu)=\Sigma(xi-\mu)/\sigma^2=(n\bar{x}-n\mu)/\sigma^2$, así $\hat{\mu} = \bar{x}$, por lo tanto parece lógico que el estimador máxima verosimilitud para la media de la normal es la media aritmética de la muestra observada.
- $\partial L(\mu,\sigma^2)/\partial\sigma^2 = 0 = -(n/2)\,(2\pi/2\pi\sigma^2)+(2\,\Sigma(xi-\mu)2)/(2\sigma^2)^2= -(n/2)\,(1/\sigma^2)+\Sigma(xi-\mu)2/\sigma^4$, así despejando $\hat{\sigma}^2 = \Sigma(xi-\mu)2/\sigma^2=s^2$, por lo tanto parece lógico que el estimador máxima verosimilitud para la varianza de la normal es la desviación típica de la muestra observada.

Estimadores de máxima verosimilitud: propiedades de los estimadores máximo-verosímiles

- Si $\widehat{\boldsymbol{\vartheta}}_{MV}$ son los estimadores de máxima verosimilitud de un modelo de una población a través de sus muestras estos cumplen las siguientes propiedades:
  - Asintóticamente centrados.
  - Asintóticamente normales.
  - Asintóticamente eficientes.
  - Suficiencia.
  - Invariancia.
  - Robustez.

Estimadores de máxima verosimilitud: propiedades de los estimadores máximo-verosímiles - invariancia

- Invariancia:
  - Si $\widehat{\boldsymbol{\vartheta}}_{MV}$ es el estimador máximo verosímil de $\boldsymbol{\vartheta}$, entonces h($\widehat{\boldsymbol{\vartheta}}_{MV}$) es el estimador máximo verosímil de h($\boldsymbol{\vartheta}$).
  - Ejemplo:
    - Sea $x_1, \ldots, x_n$ una muestra aleatoria simple de $x \sim N(\mu, \sigma)$.
    - Sabemos que $\hat{\mu}_{MV} = \bar{x}$, lo hemos demostrado anteriormente.
    - ¿Quiénes serán los estimadores de máxima verosimilitud para 3μ, μ^2 y $1/\mu^2$?
    - Por el principio de invariancia tenemos que:
    - $\widehat{3\mu}_{MV} = 3\bar{x}$
    - $\widehat{\mu^2}_{MV} = \bar{x}^2$
    - $\widehat{1/\mu}_{MV} = 1/\bar{x}$

Estimadores de máxima verosimilitud: propiedades de los estimadores máximo-verosímiles – consistencia y centrado

- Consistencia:
  - Bajo ciertas condiciones generales, $\widehat{\boldsymbol{\vartheta}}_{MV}$ es un estimador consistente de $\boldsymbol{\vartheta}$.

- Asintóticamente centrado:
  - Se verifica que el $\lim_{n\to\infty} E\left[\widehat{\boldsymbol{\vartheta}}_{MV}\right] = \boldsymbol{\vartheta}$.

Estimadores de máxima verosimilitud: propiedades de los estimadores máximo-verosímiles – normalidad asintótica

- Normalidad asintótica :
  - Para tamaños muestrales grandes, desarrollando en serie la función soporte en un entorno del estimador $\widehat{\boldsymbol{\vartheta}}_{MV}$:
  - $L(\boldsymbol{\vartheta}) \cong L\left(\widehat{\boldsymbol{\vartheta}}_{MV}\right) + (1/2)\left(\frac{d^2L\left[\widehat{\boldsymbol{\vartheta}}_{MV}\right]}{d\vartheta^2}\right)(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}_{MV})^2.$
  - Si llamamos $\widehat{\boldsymbol{\sigma}}^2_{MV} = \left(\frac{d^2L\left[\widehat{\boldsymbol{\vartheta}}_{MV}\right]}{d\vartheta^2}\right)^{-1}$, entonces la verosimilitud se puede escribir: $\ell(\boldsymbol{\vartheta}) = \ell(\boldsymbol{\vartheta}|\boldsymbol{X}) = \mathrm{k}\exp\left(\frac{1}{2\widehat{\boldsymbol{\sigma}}^2_{MV}}(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}_{MV})^2\right)$, como el soporte es el log de la verosimilitud, la verosimilitud es la exp, la constante k es $\exp(L\left(\widehat{\boldsymbol{\vartheta}}_{MV}\right))$.
  - Así la verosimilitud tiene la forma de una normal, con media $\widehat{\boldsymbol{\vartheta}}_{MV}$, y varianza $\widehat{\boldsymbol{\sigma}}^2_{MV}$.