Estadística Empresarial II

Tema 1

Variables Aleatorias

Introducción

En la asignatura anterior, se ha estudiado la **probabilidad** asociada a sucesos pertenecientes a un espacio de probabilidad o muestral, donde a partir de los experimentos se pueden obtener resultados tanto cualitativos como cuantitativos.

Ejemplo: Consideramos el experimento que consiste en lanzar una moneda y comprobar el resultado obtenido.

Espacio muestral → E = {cara,cruz}

Sucesos → *A = "salir cara", B = "salir cruz"*

$$P(A) = P(B) = 1/2$$

Sin embargo, resulta ventajoso asociar un conjunto de números reales a los resultados de un experimento aleatorio o espacio muestral, con el fin de estudiar su comportamiento aleatorio. Esto puede hacerse a través de una aplicación, definida de una manera determinada, y que se conoce por **variable aleatoria**.

Ejemplo: Para el ejemplo anterior, consideramos la aplicación: X: E → ☑ *de forma que* X("cara") = 1 y X("cruz") = 0, *pudiendo hablar de* P(X=1) y P(X=0), *respectivamente*.

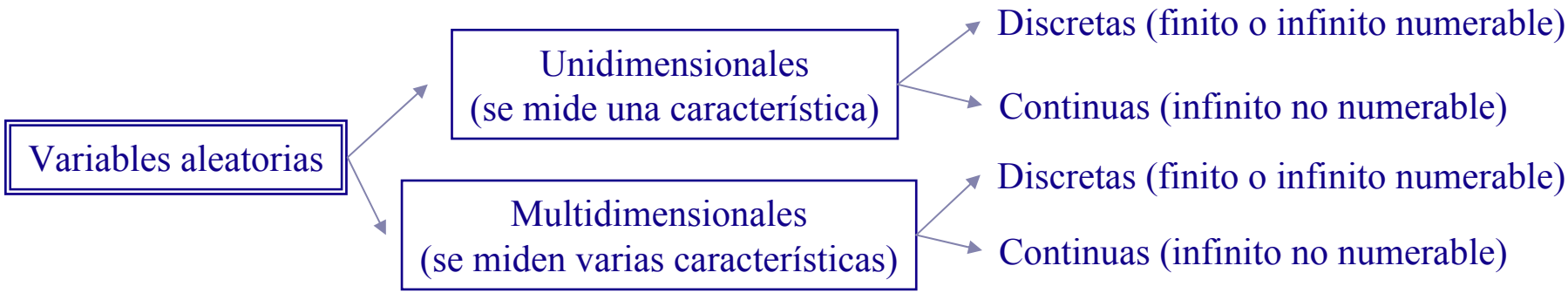
Por ello, se estudiarán las **variables aleatorias** y sus correspondientes **distribuciones de probabilidad**, que son los conceptos poblacionales que se corresponden con los de **variables estadísticas** y sus **distribuciones de frecuencias**.

Variables aleatorias unidimensionales

Se entiende por **variable aleatoria** una función real que asigna valores numéricos reales a cada uno de los sucesos elementales de un determinado experimento aleatorio.

$$X: E \rightarrow \Re$$

Las **variables aleatorias** pueden ser de carácter **finito** o **infinito**, dependiendo del número de valores que puede tomar la variable. Por otra parte, las **variables aleatorias**, al igual que las *variables estadísticas* estudiadas en Estadística Descriptiva, pueden ser:



Infinito numerable: Pueden ordenarse los valores a través de una secuencia de números naturales.

Infinito no numerable: No pueden ordenarse los valores a través de una secuencia de números naturales. Suelen ser intervalos o unión de intervalos de la recta real.

Ejercicio: Definir una variable aleatoria que sea adecuada para cada uno de los ejemplos siguientes, indicando el rango y el tipo de cada una y su distribución de probabilidad correspondiente, siempre que sea posible. Además, asignar probabilidades en función de la variable aleatoria X a los sucesos indicados.

- *Ejemplo 1: Lanzamiento de un dado.*

  Sucesos: A = "salir un 3", B = "salir un nº menor que 4", C = "salir 3 o más"

- *Ejemplo 2: Extracción de una bola de una urna que contiene 3 bolas rojas, 4 negras y 2 blancas.*

  Sucesos: A = "salir una bola roja", B = "salir una bola que no sea roja"

- *Ejemplo 3: Se observa la estatura de una serie de individuos.*

  Sucesos: A = "el individuo mide 170 cms", B = "el individuo mide más de 180 cms"

Función de distribución de probabilidad:

Dada una variable **X**, se define la **función de distribución** de **X** como:

F: $\Upsilon \rightarrow [0,1]$ $\quad F(t) = P(X \leq t)$ → Se trata de una función acumulativa que se define para variables discretas y continuas.

PROPIEDADES:

(1) Para todo **t**, $0 \leq F(t) \leq 1$.

(2) $F(+\infty) = 1$

(3) $F(-\infty) = 0$

(4) F (t) es no decreciente (si a < b entonces F (a) ≤ F (b))

$\{X \leq b\} = \{X \leq a\} \cup \{a < X \leq b\} \Rightarrow P(X \leq b) = P(X \leq a) + P(a < X \leq b) \Rightarrow$

$F(b) = F(a) + P(a < X \leq b) \geq F(a)$

(5) F (t) es continua a la derecha para todo **t**. $\lim_{t \to a^+} F(t) = F(a)$

(6) F (t) no es continua a la izquierda en todo punto **b**, de forma que P(X=b) ≠ 0. Por lo tanto, puede ocurrir que $\lim_{t \to b^-} F(t) \neq F(b)$

Ejercicio: ¿Puede ser F función de distribución?

$$F(t) = \begin{cases} 0 & \text{si } t \leq 1 \\ \frac{1}{2}(x-1) & \text{si } 1 < t \leq 2 \\ \frac{1}{2} & \text{si } 2 < t < 3 \\ 1 & \text{si } t \geq 3 \end{cases}$$

Distribuciones de probabilidad de variables aleatorias discretas:

X es **discreta** si toma un número finito o infinito numerable de valores diferentes.

Variable discreta X → **Función de probabilidad o cuantía** $P(X = x_i)$ → **Función de distribución** $F(t) = \sum_{x_i \le t} P(X = x_i)$

Condiciones: (1) $P(X = x_i) \ge 0, \forall\, x \in R_X$ (2) $\sum_{i=1}^{n} P(X = x_i) = 1$

PROP: $P(X > a) = 1 - P(X \le a) = F(a)$ $P(a < X \le b) = F(b) - F(a)$ $P(X = x_i) = F(x_i) - F(x_{i-1})$

Ejemplo: Un experimento consiste en lanzar 3 monedas y se define la variable aleatoria X = "nº de caras obtenidas". Se pide:

(a) Definir el espacio muestral y determinar el rango de X.

(b) Obtener la función de probabilidad y representarla.

(c) Obtener la función de distribución y representarla.

Distribuciones de probabilidad de variables aleatorias continuas:

Las **variables aleatorias continuas** se caracterizan porque pueden tomar cualquier valor dentro de un intervalo real de la forma (a,b), (a,+ ∞), (- ∞,b), o unión de ellos. En este caso, no tiene sentido definir la **función de probabilidad** del caso discreto, ya que P (X = x_i) = 0, ∀ x_i. Ahora, la probabilidad se concentra en intervalos, no en puntos aislados.

Variable continua X → **Función de densidad** f (x) **f: ℜ → ℜ** → **Función de distribución** $F(t) = \int_{-\infty}^{t} f(x)\,dx$

NOTA: De las propiedades que verifica F (t) resulta útil que F (∞)=1

Propiedades de la función de densidad.

(1) f (x) ≥ 0, ∀ x ∈ ℜ.

(2) $\int_{-\infty}^{+\infty} f(x)\,dx = 1$ **(3)** P (a ≤ X ≤ b) = $\int_{a}^{b} f(x)\,dx$ = F(b) – F(a) **(4)** P(X=a) = 0

(5) Si f (x) es continua en un entorno de x, se puede afirmar que $\frac{\partial F(x)}{\partial x} = f(x)$

Ejemplo: La variable aleatoria X representa el intervalo de tiempo entre dos llegadas consecutivas a una tienda, y su función de densidad viene dada por:

$$f(x) = \begin{cases} k\, e^{-x/2} & para\ \ x > 0 \\ 0 & para\ \ x \leq 0 \end{cases}$$

siendo k una constante apropiada. Determinar:

(a) El valor de k para que f (x) sea función de densidad.

(b) La función de distribución de X.

(c) El porcentaje de llegadas entre 2 y 6 minutos.

(d) El porcentaje de llegadas en como máximo 8 minutos.

Variables aleatorias bidimensionales

Una **variable aleatoria bidimensional** es una función que asigna a cada resultado de un experimento aleatorio un par de números reales, es decir:

$$(X,Y): E \rightarrow \Re^2$$

Generalizando, una **variable aleatoria n-dimensional**, será una función que asigna a cada resultado del experimento, una n-upla real.

$$(X_1, X_2, .., X_n): E \rightarrow \Re^n$$

Ejemplo 1: Realizamos el experimento de extraer dos bolas de una urna, definiendo las variables X = "asignar 1 si ambas son blancas, 2 si una es blanca y 3 si ninguna lo es" y Y = "asignar 0 si la primera es blanca y 1 si no lo es". Definir la variable aleatoria bidimensional (X,Y) y obtener su rango.

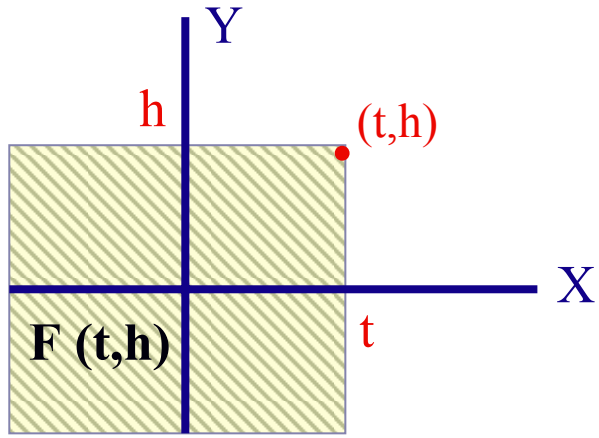
Ejemplo 2: Tenemos el experimento de lanzar tres veces un dado, de forma que X = "nº de veces que sale el 5" e Y = "nº de veces que sale el 6". Definir la variable aleatoria (X,Y) y obtener su rango.

Función de distribución:

$F: \Re^2 \rightarrow [0,1]$

Variable aleatoria bidimensional (X,Y) → **Función de distribución**
$F(t,h) = P(X \leq t, Y \leq h)$

De manera análoga al caso unidimensional, la **función de distribución** debe verificar:



(1) $0 \leq F(t,h) \leq 1$

(2) $F(+\infty,+\infty) = 1$

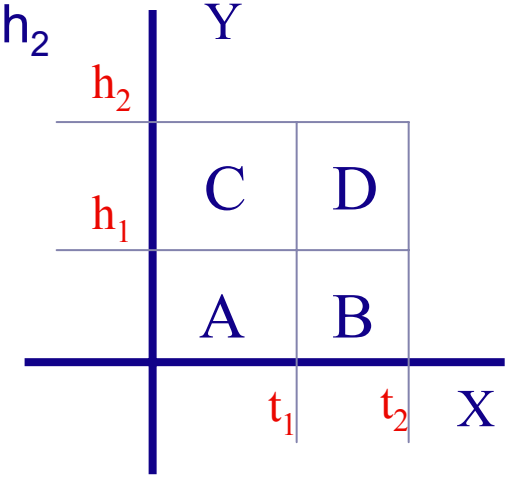
(3) $F(-\infty,h) = 0$, $F(t,-\infty) = 0$

(4) $F(t,h)$ es monótona no decreciente.

$F(t_1,h) \leq F(t_2,h)$, si $t_1 \leq t_2$; $F(t,h_1) \leq F(t,h_2)$, si $h_1 \leq h_2$

(5) $P(t_1 < X \leq t_2,\ h_1 < Y \leq h_2) = F(t_2,h_2) - F(t_2,h_1) - F(t_1,h_2) + F(t_1,h_1)$

Nota: D = (A + B + C + D) – (A + B) – (A + C) + A



(6) $F(t,h)$ es continua a la derecha para t y h.

Por el contrario, es discontinua a la izquierda para t o h en aquellos puntos en que $P(X = t, Y = h) \neq 0$.

CASO MULTIDIMENSIONAL → $F(t_1,t_2,\ldots,t_n) = P(X_1 \leq t_1, X_2 \leq t_2, \ldots, X_n \leq t_n)$ $F: \Upsilon^n \rightarrow [0,1]$

DISTRIBUCIÓN MARGINAL

Dada una **variable aleatoria bidimensional** (X,Y), se pueden obtener las **distribuciones marginales** de X e Y, prescindiendo de los valores que toma la otra variable.

Variable aleatoria bidimensional (X,Y)

$F(t,h)$

Distribución marginal de X

$F_1(t) = F(t,\infty) = P(X \leq t, Y \leq \infty)$

Distribución marginal de Y

$F_2(t) = F(\infty,h) = P(X \leq \infty, Y \leq h)$

DISTRIBUCIÓN CONDICIONADA

A partir de la **variable aleatoria bidimensional** (X,Y), podemos definir las **distribuciones condicionadas**, que nos permitirán calcular probabilidades para una de las variables marginales condicionadas a valores de la otra, como ocurre en los siguientes casos:

$$P(X = a / Y = b) = \frac{P(X = a, Y = b)}{P(Y = b)} \qquad P(Y = b / X = a) = \frac{P(X = a, Y = b)}{P(X = a)}$$

Variables bidimensionales discretas:

Una distribución de probabilidad de la variable (X,Y) es de tipo **discreto** si las dos **distribuciones marginales** lo son. En este caso, la probabilidad se va a concentrar en puntos aislados (x_i,y_j).

Variable aleatoria bidimensional discreta (X,Y)

$R_X=\{x_1,\ldots,x_n\}$, $R_Y=\{y_1,\ldots,y_m\}$

Función de distribución bidimensional o conjunta

$$F(t,h)=P(X\le t,Y\le h)=\sum_{x_i\le t}\sum_{y_j\le h}p_{ij}$$

Función de probabilidad bidimensional o conjunta

$$p_{ij}=P(X=x_i,Y=y_j)$$

Funciones de distribución marginales

$$F_1(t)=P(X\le t,Y\le\infty)=\sum_{x_i\le t}\sum_{j=1}^{m}p_{ij}$$

$$F_2(h)=P(X\le\infty,Y\le h)=\sum_{i=1}^{n}\sum_{y_j\le h}p_{ij}$$

Funciones de probabilidad marginales

$$P(X=x_i)=\sum_{j=1}^{m}p_{ij}\quad P(Y=y_j)=\sum_{i=1}^{n}p_{ij}$$

Función de distribución condicionada

$$F(t/y_j)=P(X\le t/Y=y_j)=\frac{\sum_{x_i\le t}p_{ij}}{p_{.j}}$$

$$F(h/x_i)=P(Y\le h/X=x_i)=\frac{\sum_{y_j\le h}p_{ij}}{p_{i.}}$$

Función de probabilidad condicionada

$$P(x_i/y_j)=P(X=x_i/Y=y_j)=\frac{p_{ij}}{p_{.j}}$$

$$P(y_j/x_i)=P(Y=y_j/X=x_i)=\frac{p_{ij}}{p_{i.}}$$

Las distribuciones de probabilidad bidimensionales discretas se pueden ordenar mediante una tabla de doble entrada:

X/Y	y_1	y_2	...	y_m	$p_{i.}$
x_1	p_{11}	p_{12}	...	p_{1m}	$p_{1.}$
x_2	p_{21}	p_{22}	...	p_{2m}	$p_{2.}$
...	...	...	...	...	...
x_n	p_{n1}	p_{n2}	...	p_{nm}	$p_{n.}$
$p_{.j}$	$p_{.1}$	$p_{.2}$	...	$p_{.m}$	1

<u>Ejemplo:</u> Consideremos el experimento consistente en lanzar dos monedas al aire, de forma que:

X = "Número de cruces obtenidas" Y = "Asignar un 1 si sale alguna cara y 2 si no"

$E = \{(cc), (cx), (xc), (xx)\}$ $(X,Y) : E \rightarrow \Re^2$

Obtener:

(a) Distribución de probabilidad de (X,Y)

(b) Función de distribución conjunta.

(c) Distribuciones de probabilidad y funciones de distribución marginales.

(d) Distribución de probabilidad de X condicionada a Y = 2.

(e) Distribución de probabilidad de Y condicionada a X = 1.

Variables bidimensionales continuas:

Una variable aleatoria bidimensional (X,Y) es **continua** con función de distribución continua si existe una función no negativa **f (x,y)** llamada **función de densidad**, tal que:

$$F(t,h) = \int_{-\infty}^{t}\int_{-\infty}^{h} f(x,y)\ dx\ dy$$

Además de las propiedades que verifica la función de distribución conjunta, se cumple que: $\dfrac{\partial^2 F(x,y)}{\partial x\,\partial y} = f(x,y)$

Funciones de densidad marginales

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

Funciones de distribución marginales

$$F_1(t) = F(t,\infty) = \int_{-\infty}^{t}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = \int_{-\infty}^{t} f_1(x)\,dx$$

$$F_2(h) = F(\infty,h) = \int_{-\infty}^{\infty}\int_{-\infty}^{h} f(x,y)\,dx\,dy = \int_{-\infty}^{h} f_2(y)\,dy$$

Funciones de densidad condicionadas

$$f(x/y) = \frac{f(x,y)}{f_2(y)} \qquad f(y/x) = \frac{f(x,y)}{f_1(x)}$$

Funciones de distribución condicionadas

$$F_1(t/y_0) = P(X \le t / Y = y_0) = \frac{\int_{-\infty}^{t} f(x,y_0)\,dx}{f_2(y_0)}$$

$$F_2(h/x_0) = P(Y \le h / X = x_0) = \frac{\int_{-\infty}^{h} f(x_0,y)\,dy}{f_1(x_0)}$$

Ejemplo: *Para dos coeficientes que vienen utilizándose en la división de proyectos y nuevos diseños se ha modelizado la función de densidad:*

$$f(x,y)=\begin{cases} k(6-x-y) & 2\le x\le 4,\ 0\le y\le 2 \\ 0 & resto \end{cases}$$

(a) Obtener el valor de k.

(b) Calcular la función de distribución conjunta.

(c) Obtener las funciones de densidad y distribución marginales y representarlas.

INDEPENDENCIA DE VARIABLES ALEATORIAS

Dos variables X e Y son **independientes** cuando X no influye sobre Y ni Y influye sobre X.

Si X e Y son **discretas**

$$P(X=x_i/Y=y_j)=P(X=x_i)$$
$$P(Y=y_j/X=x_i)=P(Y=y_j)$$

$$P(X=x_i,Y=y_j)=P(X=x_i).P(Y=y_j)$$

Si X e Y son **continuas**

$$f(x/y)=f_1(x)$$
$$f(y/x)=f_2(x)$$

$$f(x,y)=f_1(x).f_2(x)$$

$$F(t,h)=F_1(t).F_2(h)$$

Valor esperado o Esperanza Matemática

En el caso de las **variables estadísticas** estudiadas en la **Estadística Descriptiva** se obtenían una serie de medidas que caracterizaban y resumían la información de la distribución de frecuencias, como son las *medidas de posición* y las de *dispersión*. Algunas de estas medidas pueden obtenerse de forma teórica para las **variables aleatorias**.

Así pues, la **media** de una **variable aleatoria** puede entenderse como aquel valor al que tiende la *media aritmética* de los valores de la variable cuando el número de repeticiones del experimento es muy grande.

VALOR ESPERADO O ESPERANZA MATEMÁTICA DE h (x)

Valor medio o esperanza matemática de h (X)

$$E[h(X)] = \begin{cases} \int_{-\infty}^{\infty} h(x)\, f(x)\, dx \text{ si X es continua} \\ \sum_{i=1}^{n} h(x_i)\, P(X = x_i) \text{ si X es discreta} \end{cases}$$

Valor medio o esperanza matemática de X

$$E(X) = \mu_X = \begin{cases} \int_{-\infty}^{\infty} x\, f(x)\, dx \text{ si } X \text{ es continua} \\ \sum_{i=1}^{n} x_i\, P(X = x_i) \text{ si } X \text{ es discreta} \end{cases}$$

Ejemplo: Para el lanzamiento de 2 dados, se define una variable aleatoria **X** que toma el valor 1 si se obtienen dos resultados pares, el valor 2 si son los dos impares y 3 si se obtiene uno par y otro impar. Calcular E[X] y $E[X^2]$.

PROPIEDADES:

(1) Dada una constante **b**, se cumple que: **E [b] = b.**

(2) Dada **X** una v.a. y **b** una constante,

- (i) **E [b.X] = b.E [X]**
- (ii) **E [b+X] = b+E [X]**

(3) De la propiedad (2) se concluye que: **E [a+b.X] = a+b.E [X]**

(4) Dadas **X** e **Y** dos v.a., se verifica que: **E [X+Y] = E [X]+E [Y]**

Nota: Este resultado puede extenderse a una suma finita de v.a. Si $\mathbf{X_1}$, $\mathbf{X_2}$, ..., $\mathbf{X_n}$ son v.a., entonces:

$$\mathbf{E [X_1+X_2+...+X_n] = E [X_1]+E [X_2]+...+E [X_n]}$$

(5) Si **X** e **Y** son variables aleatorias independientes, entonces:

$$\mathbf{E [X.Y] = E [X].E [Y]}$$

Momentos

Caso unidimensional: Los **momentos** son valores que caracterizan a una distribución de frecuencias.

Momentos respecto al origen

$$\alpha_k = E[X^k] = \begin{cases} \sum_{i=1}^{n} x_i^k P(X = x_i) & si\ X\ es\ discreta \\ \int_{-\infty}^{+\infty} x^k f(x)\,dx & si\ X\ es\ continua \end{cases}$$

Casos particulares:

$$\alpha_0 = 1 \quad \alpha_1 = \mu_x$$

Momentos centrales (respecto de μ_x)

$$\mu_k = E[(X-\mu_x)^k] = \begin{cases} \sum_{i=1}^{n} (x_i - \mu_x)^k P(X = x_i) & \text{si X es discreta} \\ \int_{-\infty}^{+\infty} (x-\mu_x)^k f(x)\,dx & \text{si X es continua} \end{cases}$$

Casos particulares:

$$\mu_0 = 1 \quad \mu_1 = 0$$

RELACION

$$\mu_k = \sum_{i=0}^{k} \binom{k}{i} (-\alpha_1)^i \alpha_{k-i}$$

$$\sigma_x^2 = \mu_2 = \alpha_2 - \alpha_1^2$$

$$\mu_2 = \sigma_x^2 = E[(X-\mu_x)^2] = \begin{cases} \sum_{i=1}^{n} (x_i - \mu_x)^2 P(X = x_i) & si\ X\ es\ discreta \\ \int_{-\infty}^{+\infty} (x-\mu_x)^2 f(x)\,dx & si\ X\ es\ continua \end{cases}$$

VARIANZA

Caso bidimensional:

Momentos de orden (r,s) respecto al origen

$$\alpha_{rs} = E[X^r Y^s] = \begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{m} x_i^r y_j^s P(X = x_i, Y = y_j) & si\ (X,Y)\ es\ discreta \\ \int_{-\infty-\infty}^{+\infty+\infty} x^r y^s f(x,y)\,dx\,dy & si\ (X,Y)\ es\ continua \end{cases}$$

Casos particulares:

$\alpha_{10} = \mathrm{E}[\mathrm{X}]$ $\alpha_{01} = \mathrm{E}[\mathrm{Y}]$

$\alpha_{20} = \mathrm{E}[\mathrm{X}^2]$ $\alpha_{02} = \mathrm{E}[\mathrm{Y}^2]$

Momentos centrales de orden (r,s)

$$\mu_{rs} = E[(X-\mu_x)^r (Y-\mu_y)^s] = \begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{m} (x_i-\mu_x)^r (y_j-\mu_y)^s P(X = x_i, Y = y_j) & si\ (X,Y)\ es\ discreta \\ \int_{-\infty-\infty}^{+\infty+\infty} (x-\mu_x)^r (y-\mu_y)^s f(x,y)\,dx\,dy & si\ (X,Y)\ es\ continua \end{cases}$$

Casos particulares: $\mu_{10} = 0$ $\mu_{01} = 0$ $\mu_{20} = V[X] = \sigma_x^2$ $\mu_{02} = V[Y] = \sigma_y^2$

$$\mu_{11} = \sigma_{xy} = E[(X-\mu_x)(Y-\mu_y)] = \begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{m} (x_i-\mu_x)(y_j-\mu_y) P(X = x_i, Y = y_j) & si\ (X,Y)\ es\ discreta \\ \int_{-\infty-\infty}^{+\infty+\infty} (x-\mu_x)(y-\mu_y) f(x,y)\,dx\,dy & si\ (X,Y)\ es\ continua \end{cases}$$

Relaciones entre momentos centrales y momentos respecto al origen

$$\sigma_x^2 = \mu_{20} = \alpha_{20} - \alpha_{10}^2$$
$$\sigma_y^2 = \mu_{02} = \alpha_{02} - \alpha_{01}^2$$
$$\sigma_{xy} = \alpha_{11} - \alpha_{10}\alpha_{01}$$

PROPIEDADES:

(1) Dada **X** una v.a. y **b** una constante,

- (i) $\sigma^2_{x+b} = V(X+b) = V(X) = \sigma^2_x$
- (ii) $\sigma^2_{x.b} = V(X.b) = b^2.V(X) = b^2.\sigma^2_x$

(2) Dada **X** una v.a. y **a** y **b** constantes, $\sigma^2_{aX+b} = V(aX+b) = a^2.V(X) = a^2.\sigma^2_x$

Ejemplo: Un inversionista dispone de 100.000 dólares para una inversión de un año. Está considerando dos opciones: colocar el dinero en el mercado de valores, lo que garantiza una ganancia anual fija del 15 %, y un plan de inversión cuya ganancia anual puede considerarse como una variable aleatoria cuyos valores en tantos por uno vienen dados a continuación. Con base a la ganancia esperada, ¿cuál de los dos planes debe seleccionarse? ¿Cuál será la varianza en términos absolutos?

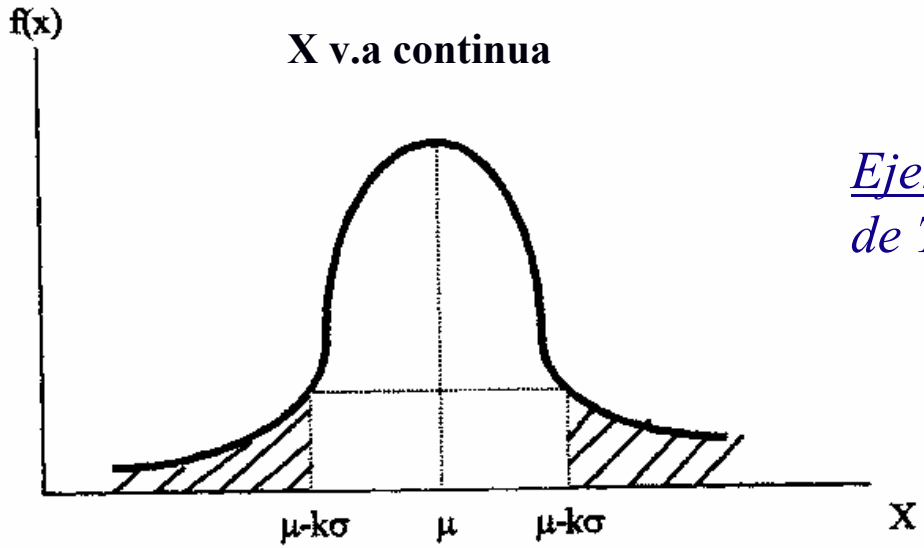
Ganancia	0.30	0.25	0.20	0.15	0.10	0.05
$P(X=x_i)$	0.20	0.20	0.30	0.15	0.10	0.05

Teorema de Tchebycheff

Sea **X** una variable aleatoria con media μ y desviación típica σ. La desigualdad de Tchebycheff nos va a poner de relieve la importancia de la desviación típica como medida de dispersión.

TEOREMA DE TCHEBYCHEFF: La probabilidad de que la variable **X** tome valores cuyas desviaciones respecto a μ sean mayores que **k** veces la desviación típica es menor o igual a $\mathbf{1/k^2}$.

$$P(|X-\mu|>k\sigma)\leq\frac{1}{k^2}\Leftrightarrow P(\mu-k\sigma\leq X\leq\mu-k\sigma)\geq 1-\frac{1}{k^2}$$



<u>Ejercicio:</u> Interpretar la desigualdad de Tchebycheff para k=2 y k=3.

Estadística Empresarial II

Tema 2

Modelos probabilísticos discretos

Introducción

En el tema anterior hemos estudiado, en términos generales, las *variables aleatorias*, tanto *discretas* como *continuas*. Ahora vamos a describir una serie de **variables aleatorias discretas** con unas distribuciones de probabilidad específicas, que ocurren con cierta frecuencia en el mundo real.

Daremos las condiciones bajo las cuales se presentan esas distribuciones específicas, y llegaremos a obtener algunas características de ellas, como la media y la varianza.

En general, una distribución de probabilidad está caracterizada por una o más cantidades que reciben el nombre de **parámetros** de la distribución, y como el *parámetro* puede tomar valores en un conjunto dado, entonces se obtendrá una familia de distribuciones de probabilidad, que tendrán la misma función de probabilidad.

Distribución de Bernouilli

Se denomina **prueba de Bernouilli** al experimento que sólo tiene dos resultados posibles: *"éxito"* o *"fracaso"*, según sea el suceso de nuestro interés.

Ejemplos: Al lanzar una moneda, salir cara o cruz; al lanzar un dado, que salga 6 o no; que un artículo esté defectuoso o no, etc.

Al realizar la prueba, puede ocurrir el suceso **A** (*"éxito"*), con una probabilidad **p**, o $\overline{\mathbf{A}}$ (*"fracaso"*), con una probabilidad **q.**

NOTA: **q = 1 – p**, al ser **A** y $\overline{\mathbf{A}}$ dos sucesos complementarios.

X = "Número de éxitos (A) en una prueba de Bernouilli". → $X = \begin{cases} 0 & si\ ocurre\ el\ suceso\ \overline{A} \\ 1 & si\ ocurre\ el\ suceso\ A \end{cases}$

X ~ b (p) $\quad R_X = \{0,1\}$

Función de probabilidad

x_i	0	1
$P(X=x_i)$	1-p	p

$$P(X = k) = p^k \cdot (1 - p)^{1-k} \qquad k = 0, 1$$

Características:

- Sólo depende del parámetro **p**.
- **E [X] = p**
- **V (X) = p.q**

$$E(X) = 0 \,.\, P(X = 0) + 1 \,.\, P(X = 1) = 0 \,.\, q + 1 \,.\, p = p$$

$$Var(X) = E(X - p)^2 = (0 - p)^2 \,.\, P(X = 0) + (1 - p)^2 \,.\, P(X = 1) =$$
$$= p^2 \,.\, q + q^2 \,.\, p = p\, q \,.\, (p + q) = p\, q$$

- **Función generatriz de momentos:**

$$G_x(t) = \sum_{x_i = 0,1} e^{t x_i} P(X = x_i) = e^{t.0} P(X = 0) + e^{t.1} P(X = 1) = q + p.e^t$$

- **Función característica:**

$$\phi_x(t) = \sum_{x_i = 0,1} e^{it x_i} P(X = x_i) = e^{it.0} P(X = 0) + e^{it.1} P(X = 1) = q + p.e^{it}$$

Sirven para obtener momentos de distinto orden, necesarios para otro tipo de medidas.

Ejemplo: Una vendedora de cosméticos piensa que en una visita concreta a domicilio la probabilidad de conseguir una venta es 0'4. Si definimos la variable X como una variable de Bernouilli, obtener:

(a) La media y la varianza de la distribución.

(b) La probabilidad de que si hace una visita a domicilio, consiga una venta.

Distribución Binomial

Se trata de uno de los modelos probabilísticos más importantes dada su amplia aplicabilidad a muchas situaciones en las que podemos enfrentarnos en la realidad.

Partimos de la realización sucesiva de **n** pruebas independientes de Bernouilli, con parámetro **p**, y definimos la variable aleatoria siguiente:

X = "Número de éxitos (A) en n pruebas de Bernouilli".

Ejemplos: *Nº de caras o cruces obtenidas al lanzar una moneda un nº de veces, nº de artículos defectuosos de un determinado lote, etc.*

X ~ B (n, p)

$$R_X = \{0,1,2,3,...,n\}$$

Función de probabilidad

$$P(X = k) = \binom{n}{k} . p^k . (1 - p)^{n-k} \quad con \quad k = 0,1,2,3,.....n$$

La probabilidad de obtener **k** éxitos en **n** pruebas se obtiene a través del producto de $\mathbf{p^k.q^{n-k}}$ por el número de órdenes distintos que puedan establecerse, que son el número de combinaciones de **n** tomadas de **k** en **k**.

Nota: *Dadas* $X_i \sim b(p)$, *i = 1, ..., n, entonces:*

$$X_1 + X_2 + ... + X_n \sim B\ (n, p).$$

Ejercicio: Comprobar que la función anterior es de probabilidad.

Características:

- Depende de los parámetros **n** y **p**.
- **E [X] = n.p**
- **V (X) = n.p.q**

$$\mu = E(X) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} p = n * p$$

$$\begin{aligned}
&\mathrm{Var}(X) = \mathrm{E}\left[(X - np)^2\right] = \mathrm{E}\left[(X_1 + X_2 + ... + X_n - np)^2\right] = \mathrm{E}\left[[(X_1 - p) + (X_2 - p) + ... + (X_n - p)]^2\right] = \\
&= \mathrm{E}[(X_1 - p)^2 + (X_2 - p)^2 + ... + (X_n - p)^2 + \sum_{i \neq j} (X_i - p).(X_j - p)] = \\
&= \mathrm{E}[(X_1 - p)^2] + \mathrm{E}[(X_2 - p)^2] + ... + \mathrm{E}[(X_n - p)^2] + \sum_{i \neq j} \mathrm{E}\left[(X_i - p).(X_j - p)\right] = \\
&= \sum_{i=1}^{n} \mathrm{E}(X_i - p)^2 = \sum_{i=1}^{n} \mathrm{V}(X_i) = \sum_{i=1}^{n} p.q = n.p.q
\end{aligned}$$

- **Función generatriz de momentos:**

$$\begin{aligned}
G(t) &= E\left(e^{tX}\right) = E\left(e^{t.(X_1 + X_2 + ... + X_n)}\right) = E\left(e^{tX_1}.e^{tX_2}. .... .e^{tX_n}\right) = E\left(e^{tX_1}\right).E\left(e^{tX_2}\right)..... E\left(e^{tX_n}\right) = \\
&= (q + e^t p).(q + e^t p)..... = (q + e^t p)^n
\end{aligned}$$

- **Función característica:** $\phi(t) = (q + e^{it} p)^n$

Para facilitar los cálculos, se han tabulado los valores de la función de probabilidad de una **distribución binomial**, para distintos valores de **n** y **p**.

*Ejercicio: Dada una variable aleatoria $X \sim B(n,p)$, definimos la variable $Y = X / n$, que indica la "proporción de éxitos obtenidos en las **n** pruebas". Obtener la media y la variable de esta variable Y.*

Ejemplo: Algunos economistas han propuesto que haya un control de salarios y precios para combatir la inflación, pero otros consideran que esos controles no son efectivos porque tratan los efectos y no las causas de la inflación. Una reciente encuesta indica que el 40% de los españoles adultos están a favor de un control de precios y salarios. Si se seleccionan 5 adultos aleatoriamente:

(a) ¿Cuál es la probabilidad de que ninguno esté a favor de dicho control?

(b) ¿Cuál es la probabilidad de que como máximo 3 estén a favor del control de precios y salarios?

(c) Por término medio, ¿cuántos estarán a favor del control de precios y salarios?

Distribución geométrica

Partiendo de la realización de una serie de pruebas de Bernouilli independientes de parámetro **p**, definimos:

X = "Número de pruebas realizadas hasta obtener el primer éxito (A)" .

X ~ G (p) $\qquad R_X = \{1,2,3,...\}$

Función de probabilidad

$$P(X=k) = q^{k-1}.p \quad con \quad k = 1,2,3,...$$

Características:

Ejemplos: Número de lanzamientos de un dado hasta obtener un 6, número de artículos analizados hasta obtener uno defectuoso, etc.

- Depende exclusivamente del parámetro **p**.
- **Función generatriz de momentos:**

$$G(t) = E\left(e^{tX}\right) = \sum_{x=1}^{\infty} e^{tx} \,.\, p \,.\, q^{x-1} = e^t\, p + e^{2t}\, p\, q + e^{3t}\, p\, q^2 + ... = p\, e^t \,.\, \sum_{x=0}^{\infty} \left(e^t\, q\right)^x = \frac{p\, e^t}{1 - q\, e^t}$$

- **Función característica:** $\phi(t) = \dfrac{p\, e^{it}}{1 - q\, e^{it}}$

Nota: La serie $\sum_{x=0}^{\infty} \left(e^t\, q\right)^x$ *converge si la razón* $\mathbf{e^t.q < 1}$.

- **Media:**

$$G'(t) = \frac{p\,e^{t}}{\left(1 - q\,e^{t}\right)^{2}} \Rightarrow E(X) = G'(0) = \frac{1}{p}$$

- **Varianza:**

$$G''(t) = \frac{p\,e^{t}\cdot\left(1 + q\,e^{t}\right)}{\left(1 - q\,e^{t}\right)^{3}} \quad \Rightarrow \quad E(X^{2}) = G''(0) = \frac{1+q}{p^{2}}$$

$$Var(X) = E(X^{2}) - \mu^{2} = \frac{1+q}{p^{2}} - \frac{1}{p^{2}} = \frac{q}{p^{2}}$$

<u>Ejemplo:</u> En una auditoría interna sobre las facturas expedidas por una empresa, se considera que habrá que realizar una investigación más exhaustiva desde el momento en que se encuentre una factura con algún error. Por auditorías anteriores, se estima que la probabilidad de que una factura tenga algún error es del 7% ¿Cuál es la probabilidad de que se encuentre una factura con errores en la décima extracción?.

Distribución Binomial Negativa

Este modelo consiste en la generalización del modelo **geométrico** para la ocurrencia de **r** éxitos. Es decir, se van a ir realizando pruebas independientes de Bernouilli con parámetro **p** hasta que se produzcan **r** éxitos.

X = "Número de pruebas realizadas hasta obtener el r-ésimo éxito (A)" .

X ~ BN (r,p) $\quad R_X = \{r, r+1, r+2, \ldots\}$

Función de probabilidad →

$$P(X=k) = \binom{k-1}{r-1} p^r . q^{k-r} \quad con \;\; k = r, r+1, r+2, \ldots$$

Características:

- Depende de los parámetros **r** y **p**.
- **Función generatriz de momentos:**

$$G(t) = \left(\frac{p\, e^t}{1 - q\, e^t} \right)^r$$

- **Función característica:**

$$\phi(t) = \left(\frac{p\, e^{it}}{1 - q\, e^{it}} \right)^r$$

- **Media:**

$$G'(t) = r\left(\frac{p\,e^{t}}{1 - q\,e^{t}}\right)^{r-1} \cdot \frac{p\,e^{t}}{(1 - q\,e^{t})^{2}} \Rightarrow E(X) = G'(0) = r\left(\frac{p}{1-q}\right)^{r-1} \cdot \frac{p}{(1-q)^{2}} = \frac{r}{p}$$

- **Varianza:**

$$\mathrm{Var}(X) = \mathrm{E}(X^{2}) - \mu^{2} = G''(0) - \mu^{2} = \frac{r^{2} + r - rp}{p^{2}} - \frac{r^{2}}{p^{2}} = \frac{r\,q}{p^{2}}$$

<u>Ejemplo:</u> En un departamento de control de calidad de una empresa dedicada a la fabricación de teléfonos inalámbricos se inspeccionan las unidades terminadas. Se piensa que la proporción de unidades defectuosas es 0,05. ¿Cuál es la probabilidad de que la vigésima unidad inspeccionada sea la segunda que se encuentre con defectos?

Distribución de Poisson

Previamente al desarrollo de este modelo, es preciso estudiar lo que se conoce como **proceso de Poisson**. Consideremos un experimento consistente en observar sucesos discretos en un intervalo continuo, caracterizado por:

- *Estabilidad*: Genera, a largo plazo, un número medio de sucesos constante en el intervalo considerado (normalmente la unidad de tiempo, si bien puede tomar otras referencias, espaciales, regionales, etc...).
- Se puede elegir un intervalo, de amplitud **a**, lo suficientemente pequeño que verifique:
  - La probabilidad de que se de exactamente un suceso en ese intervalo de aproximadamente λ**.a**.
  - La probabilidad de que se observen más de una ocurrencia en el intervalo es aproximadamente **0**.
- Los sucesos son estadísticamente independientes, es decir, las ocurrencias en un intervalo no influyen sobre cualquier otro intervalo disjunto del anterior y de igual longitud.

Ejemplos:

- *Estudio del número de aviones que llegan a un aeropuerto de una zona turística en una unidad de tiempo (por hora, día, ...).*

  - Durante la primera semana del mes aterrizan una media de 7 aviones por hora, pudiendo fijarse un intervalo temporal (dos minutos, por ejemplo) en el que la probabilidad de que aterrice sólo un avión es 7(1/30), mientras que la probabilidad de que lo hagan 2 ó más aviones es prácticamente nula.*

  - El hecho de que entre las 10 y las 11 horas de un día cualquiera aterricen 12 aviones no afecta estadísticamente en el número de aviones que aterrizarán en cualquier otro intervalo horario.

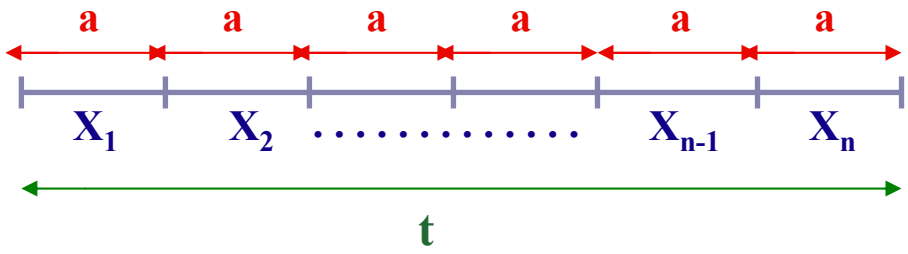
- *Número de llamadas por hora recibidas en una centralita, número de accidentados en un fin de semana, número de averías de ordenadores en una empresa por día,...*

Observando un **proceso de Poisson** en el que se da un número medio **λ.t** de ocurrencias en una unidad de tiempo **t**, definimos:

X = "Número de ocurrencias de un suceso en un periodo de tiempo t" .

X ~ P (λt)

$$R_X = \{0,1,2,...\}$$



Dividimos el intervalo de amplitud **t** en **n** subintervalos de amplitud **a=t/n**, suficientemente pequeños, como para acoplarse a las características del **proceso de Poisson.**

X_i = "Número de ocurrencias del suceso en en intervalo i-ésimo" ~ b(λ.a)

P(X_i= 1) = λ.a
P(X_i= 0) = 1 - λ.a

Si las v.a. X_i son independientes entre sí $\Rightarrow Y = \sum_{i=1}^{n} X_i \sim B(n, \lambda.a)$

$$P(Y=k) = \binom{n}{k} (\lambda a)^k . (1 - \lambda a)^{n-k} \quad , \; con \;\; k = 0, 1, 2 ....$$

$$P(Y=k) = \binom{n}{k} \left(\frac{\lambda t}{n}\right)^k . \left(1 - \frac{\lambda t}{n}\right)^{n-k} = \frac{(\lambda.t)^k}{k!} \frac{n(n-1)(n-2)....(n-k+1)}{n^k} \left(1 - \frac{\lambda t}{n}\right)^{n-k} \xrightarrow{n \to \infty} \frac{(\lambda.t)^k}{k!} e^{-\lambda.t}$$

λ.t → λ

X ~ P (λ)

$$P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad , \; con \;\; k = 0, 1, 2 ....$$

λ: nº medio de ocurrencias en un intervalo de amplitud **t**.

Características:

- Depende exclusivamente del parámetro λ.
- **Función generatriz de momentos:** $G(t) = e^{\lambda(e^t - 1)}$
- **Función característica:** $\varphi(t) = e^{\lambda (e^{it} - 1)}$
- **Media:** $G'(t) = e^{\lambda (e^t - 1)} . \lambda\, e^t \Rightarrow E(X) = G'(0) = \lambda$
- **Varianza:** $G''(t) = \lambda\, e^t . e^{\lambda (e^t - 1)} . (\lambda\, e^t + 1)$

$$Var(X) = E(X^2) - \mu^2 = G''(0) - \lambda^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda$$

- Aproximación de la distribución de Poisson a la binomial:

*Sea **X** una variable aleatoria **B(n,p)**. Si n es grande y n.p tiene un tamaño moderado, esta distribución puede aproximarse bien por una distribución **P(λ=n.p)**.*

Ejemplo: El número de clientes que llega a un banco es una variable aleatoria de Poisson. Si el número promedio es de 120 clientes por hora, ¿cuál es la probabilidad de que en un minuto lleguen por lo menos tres clientes?

Distribución Hipergeométrica

Consideremos una población finita de tamaño **N** compuesta por individuos que poseen la característica **A**, o bien la característica **A*** (**A** y **A*** son antagónicas). Sea **M** el número de individuos de esta población que tienen la característica **A**, y **M*** los que poseen la **A***.

$$POBLACION\ (N\ individuos) \rightarrow \begin{cases} M\ individuos\ de\ la\ clase\ A \\ M^*\ individuos\ de\ la\ clase\ A^* \end{cases}$$

$$p = \frac{M}{N} \quad y \quad q = \frac{M^*}{N}$$

A continuación se extrae una muestra aleatoria de **n** elementos, sin reponer los individuos ya extraídos, por lo que **p** variará de una extracción a otra (esto es la principal diferencia que existe frente a la binomial).

X = "Número de elementos de la clase A en la muestra" .

X ~ H (N,n,p)

$$R_X = \{\max\{0, n - M^*\}, \ldots, \min\{M, n\}\}$$

$$P(X = k) = \frac{\binom{M}{k} \cdot \binom{M^*}{n-k}}{\binom{N}{n}} = \frac{\binom{N.p}{k} \cdot \binom{N.q}{n-k}}{\binom{N}{n}}$$

Características:

- Depende de tres parámetros, que son **N**, **n** y **p**.
- **Media:**

$$E(X) = \sum_k k \frac{\binom{M}{k} \cdot \binom{M^*}{n-k}}{\binom{N}{n}} = \sum_k k \frac{\frac{M!}{k!\,.\,(M-k)!} \cdot \frac{M^*!}{(n-k)!\,.\,(M^*-(n-k))!}}{\frac{N!}{n!\,.\,(N-n)!}} = M \,.\, \sum_k \frac{\binom{M-1}{k-1} \cdot \binom{M^*}{n-k}}{\binom{N}{n}} =$$

$$= M \,.\, \sum_k \frac{\binom{M-1}{k-1} \cdot \binom{M^*}{n-k}}{\frac{N}{n}\binom{N-1}{n-1}} = \frac{M\,.\,n}{N} \cdot \sum_k \frac{\binom{M-1}{k-1} \cdot \binom{M^*}{n-k}}{\binom{N-1}{n-1}} = n\,.\,p$$

Suma de las probabilidades de la distribución de probabilidad de una **H(N-1,n-1,p)**, con **M-1**.

- **Varianza:**

$$\mathrm{Var}(X) = \frac{n\,.\,M\,.\,M^*}{N^2} \cdot \frac{N-n}{N-1} = n\,.\,p\,.\,q\,.\,\frac{N-n}{N-1}$$

Ejemplo: Un fabricante de automóviles compra los motores a una compañía donde se fabrican bajo estrictas especificaciones. El fabricante recibe un lote de 40 motores. Su plan para aceptar el lote consiste en seleccionar ocho de forma aleatoria y someterlos a prueba. Si encuentra que ningún motor presenta serios defectos, acepta el lote; de otra forma, lo rechaza. Si el lote contiene dos motores con serios defectos, ¿cuál es la probabilidad de que sea aceptado?

Cuadro Resumen

Distribuciones de probabilidad discretas más notables					
Variable	*Definición*	*Parámetros*	*F. de probabilidad*	*Media*	*Varianza*
Bernouilli	**Número de éxitos (A) en una prueba de Bernouilli.**	p, q=1-p	$P(X=k)=p^k.(1-p)^{1-k}$	p	q
Binomial	**Número de éxitos (A) en n pruebas de Bernouilli.**	p	$P(X=k)=\binom{n}{k}.p^k.(1-p)^{n-k}$	np	npq
Geométrica	**Número de pruebas realizadas hasta obtener el primer éxito (A).**	p	$P(X=k)=q^{k-1}.p$	$\frac{1}{p}$	$\frac{q}{p^2}$
Binomial Negativa	**Número de pruebas realizadas hasta obtener el r-ésimo éxito (A).**	r, p	$P(X=k)=\binom{k-1}{r-1}p^r.q^{k-r}$	$\frac{r}{p}$	$\frac{rq}{p^2}$
Poisson	**Número de ocurrencias de un suceso en un periodo de tiempo t.**	λ	$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$	λ	λ
Hipergeométrica	**Número de elementos de la clase A en la muestra.**	N, n, p	$P(X=k)=\frac{\binom{M}{k}.\binom{M^*}{n-k}}{\binom{N}{n}}=\frac{\binom{N.p}{k}.\binom{N.q}{n-k}}{\binom{N}{n}}$	n.p	$\text{n.p.q.}\frac{N-n}{N-1}$

Estadística Empresarial II

Tema 3

Modelos probabilísticos continuos

Introducción

Al igual que ocurre con la modelización de **variables aleatorias discretas**, existen ciertas características de las **variables aleatorias continuas** que tienden a repetirse con relativa frecuencia, por lo que se puede definir un modelo teórico que se ajuste a ellas.

Así pues, estudiaremos diversas **familias de distribuciones continuas**, que dependerán de una serie de *parámetros* en cada caso; y llegaremos a obtener algunas características de ellas, como la media y la varianza.

Algunas de las distribuciones que estudiaremos a continuación son modelos cuya importancia reside en su aplicación a diversos casos planteados por la **Inferencia Estadística**, que serán estudiados posteriormente.

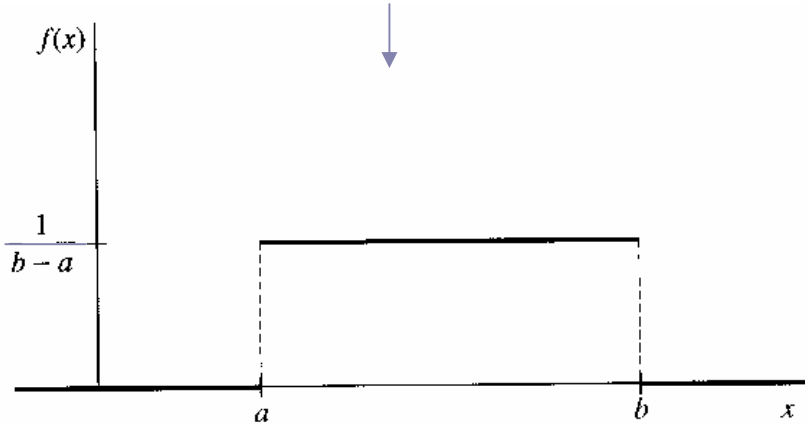
Distribución Uniforme

Esta variable es la más sencilla de las distribuciones continuas y surge al considerar una variable aleatoria que toma valores equiprobables en un intervalo finito, de manera que la probabilidad de que la variable tome un valor en cada subintervalo de la misma longitud es la misma.

X ~ U (a,b) $\qquad R_X = (a,b)$

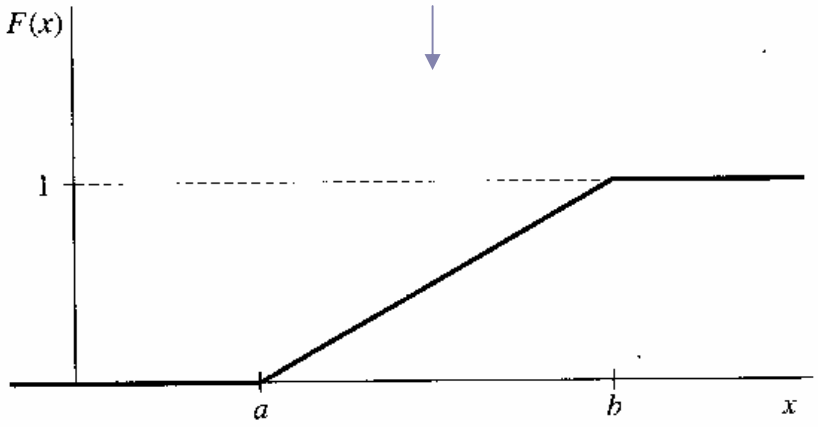
Función de densidad: Es constante a lo largo de (a,b).

$$f(x) = \begin{cases} 0 & , \ x < a \\ \dfrac{1}{b-a} & , \ a \le x \le b \\ 0 & , \ x > b \end{cases}$$



Función de distribución

$$F(x) = \begin{cases} 0 & , \ x < a \\ \dfrac{x-a}{b-a} & , \ a \le x \le b \\ 1 & , \ x > b \end{cases}$$



Características:

- Depende de los límites del rango de la variable, **a** y **b**.
- **Media:**

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx = \frac{1}{b-a}\int_{a}^{b} x\, dx = \frac{1}{b-a}\,\frac{b^2-a^2}{2} = \frac{b+a}{2}$$

- **Varianza:**

$$\mathrm{Var}(X) = \mathrm{E}(X-\mu)^2 = \mathrm{E}(X^2) - \mu^2$$

$$\mathrm{E}\left(X^2\right) = \int_{-\infty}^{\infty} x^2 f(x)dx = \frac{1}{b-a}\int_{a}^{b} x^2 dx = \frac{1}{b-a}\,\frac{b^3-a^3}{3} = \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} = \frac{b^2+ab+a^2}{3}$$

$$\mathrm{Var}(X) = \frac{b^2+ab+a^2}{3} - \frac{(b+a)^2}{4} = \frac{(b-a)^2}{12}$$

- **Función generatriz de momentos:**

$$G(t) = \mathrm{E}(e^{tx}) = \frac{e^{tb}-e^{ta}}{t\,.\,(b-a)}$$

- **Función característica:**

$$\phi(t) = \mathrm{E}(e^{itX}) = \frac{e^{itb}-e^{ita}}{i\,t\,.\,(b-a)}$$

NOTA: En el campo de la inferencia estadística, la **distribución uniforme** se utiliza para generar números aleatorios que permitirán seleccionar los individuos que formarán las **muestras aleatorias**.

<u>Ejemplo:</u> Una máquina para llenar botellas de agua llena entre 500 y 1500 botellas en una hora, de manera equiprobable para todos los posibles valores dentro del intervalo. Se pide:

(a) Obtener las funciones de densidad y de distribución del número de botellas a llenar en una hora.

(b) Calcular la media y la varianza de la distribución.

(c) ¿Cuál es la probabilidad de que se llenen entre 750 y 1000 botellas en una determinada hora?

Distribución Normal

La **distribución normal** es indudablemente la más importante y la de mayor uso de todas las distribuciones de probabilidad continuas y, en general, de todas las usadas en Estadística. Las principales razones de su uso son:

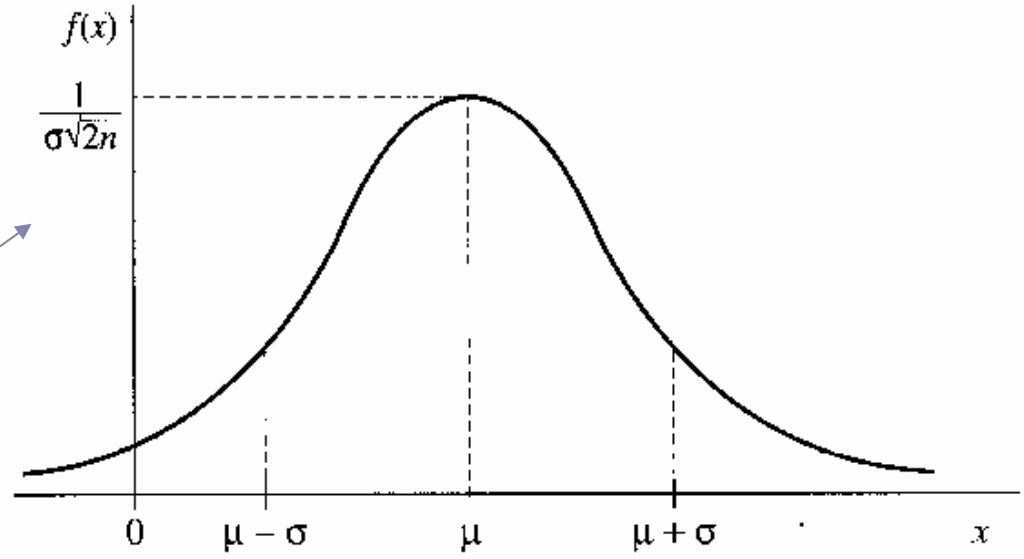
- Existen numerosas variables continuas directamente observables de la realidad que siguen una **distribución normal** o aproximadamente normal. El peso, la altura y el coeficiente de inteligencia son ejemplos de variables aleatorias aproximadamente normales.
- La **distribución normal** es una excelente aproximación de otras distribuciones, tanto discretas como continuas, hecho que queda avalado por el **Teorema Central del Límite.**
- La distribución muestral de algunos de los más importantes estadísticos muestrales, tales como la *media* o la *proporción muestral*, tienden aproximadamente a una distribución normal, si el tamaño de la muestra es lo suficientemente grande.

$$X \sim N(\mu,\sigma),\ -\infty < \mu < \infty,\ \sigma > 0.$$

$$R_X = (-\infty,+\infty)$$

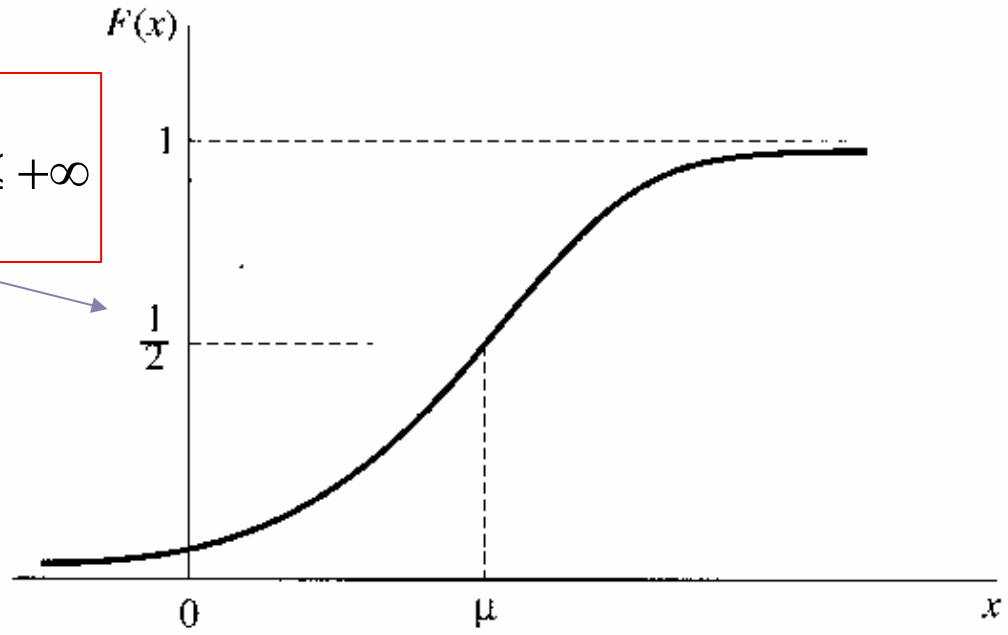
Función de densidad

$$f(x) = \frac{1}{\sigma.\sqrt{2\pi}}\, e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}},\ x \in (-\infty,+\infty)$$



Función de distribución

$$F(t) = \frac{1}{\sigma.\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\, dx\ ,\ -\infty \le t \le +\infty$$



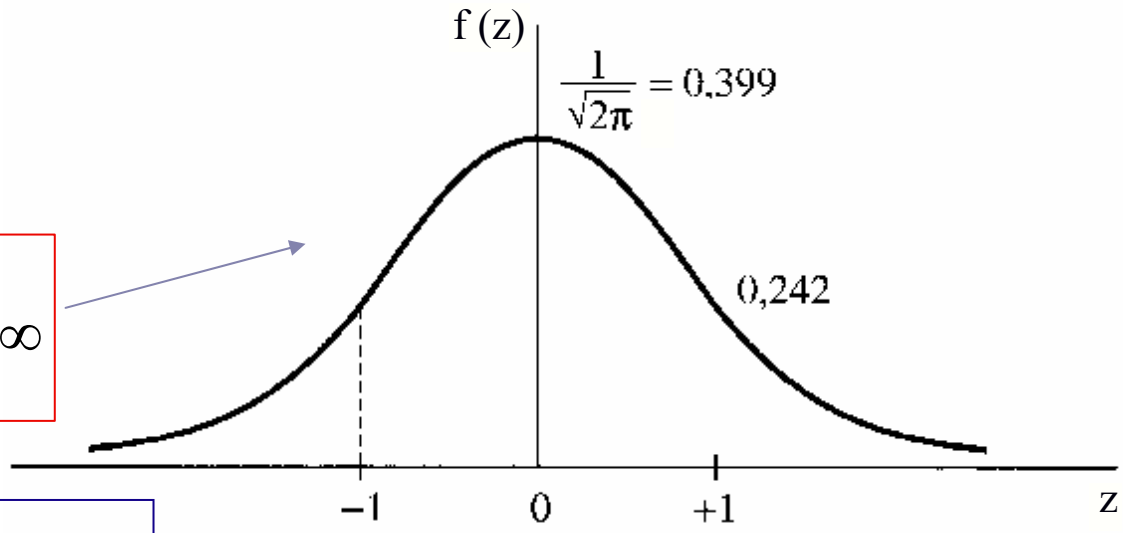
Para simplificar los cálculos probabilísticos a realizar con cualquier **N(μ,σ)**, es necesario definir la **distribución normal estandarizada Z = N(0,1).**

DISTRIBUCIÓN NORMAL ESTANDARIZADA O TIPIFICADA:

X ~ N (0,1)

Función de densidad

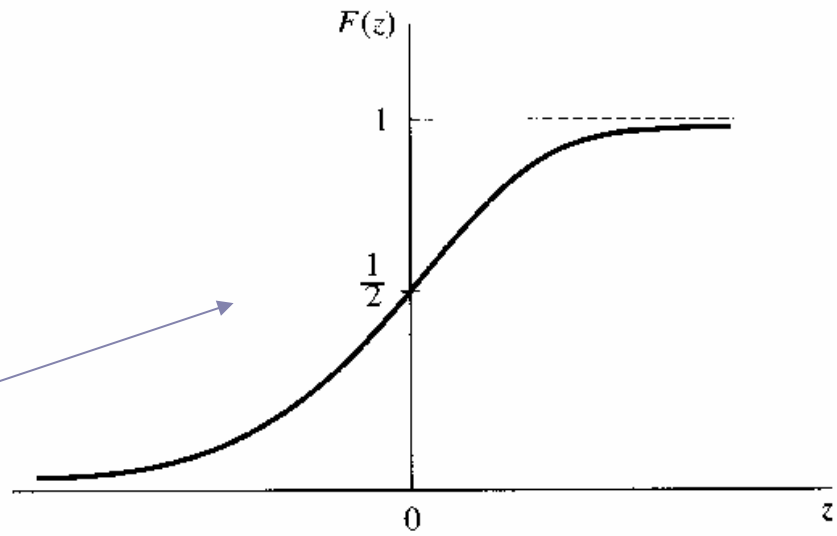
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad , -\infty \le z \le \infty$$



- Simétrica respecto al eje Y (z = 0)
- Asíntota horizontal: y = 0.
- Creciente para z < o y decreciente para z > 0.
- Máximo: (0 , $1/\sqrt{2\pi}$).
- Puntos de inflexión para z = -1 y z = 1.

Función de distribución

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{z^2}{2}} dz \quad , -\infty \le t \le +\infty$$



Características:

- No depende de ningún parámetro, al ser $\mu = 0$ y $\sigma = 1$.
- **Función generatriz de momentos:**

$$G(t) = E(e^{tZ}) = \int_{-\infty}^{+\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}+tz} dz =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(z-t)^2}{2}} e^{\frac{t^2}{2}} dz = \frac{1}{\sqrt{2\pi}} e^{\frac{t^2}{2}} \boxed{\int_{-\infty}^{+\infty} e^{-\frac{(z-t)^2}{2}} dz} = e^{\frac{t^2}{2}}$$

$\sqrt{2\pi}$ haciendo $z - t = u$

- **Media:** $G'(t) = t \,.\, e^{\frac{t^2}{2}}$

$$\mu_Z = E(Z) = G'(0) = 0$$

- **Varianza:** $G''(t) = (1 + t^2) \,.\, e^{\frac{t^2}{2}}$

$$\sigma_Z{}^2 = E[(Z - 0)^2] = E[Z^2] = G''(0) = 1$$

- La variable **Z ~ N (0,1)** se encuentra tabulada.

$$F(Z_{1-\alpha}) = P(Z \le Z_{1-\alpha}) = 1-\alpha$$

Nota: Al ser simétrica, se verifica que: $Z_{1-\alpha} = - Z_{\alpha}$

Ejemplo: Dada Z ~ N (0,1), calcular:

(a) Probabilidad de que Z sea menor que 1'23.

(b) Probabilidad de que Z sea mayor que 1'23.

(c) Probabilidad de que Z esté entre – 0'41 y 1'23.

(d) Valor de la variable Z que deja a su izquierda una probabilidad igual a 0'8212.

DISTRIBUCIÓN NORMAL GENERALIZADA:

$X \sim N(\mu,\sigma)$, $-\infty < \mu < \infty$, $\sigma > 0$.

La obtención de probabilidades para una **distribución normal generalizada** requiere una transformación de la variable para estandarizarla, lo que se conoce como **tipificación.**

$$\frac{X-\mu}{\sigma} = Z \Leftrightarrow X = \sigma . Z + \mu$$

$$(1)\ \text{Si}\ \ X \sim N(\mu,\sigma)\ \ \text{entonces}\ \ Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

$$(2)\ \text{Si}\ \ Z \sim N(0,1)\ \ \text{entonces}\ \ X = \sigma Z + \mu \sim N(\mu,\sigma)$$

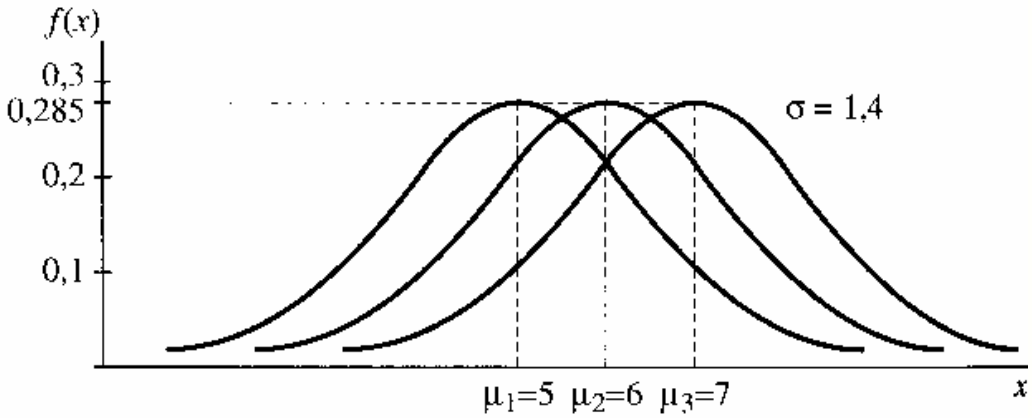
(1) Partimos de que $Z = \dfrac{X - \mu}{\sigma}$ y $z \in R$

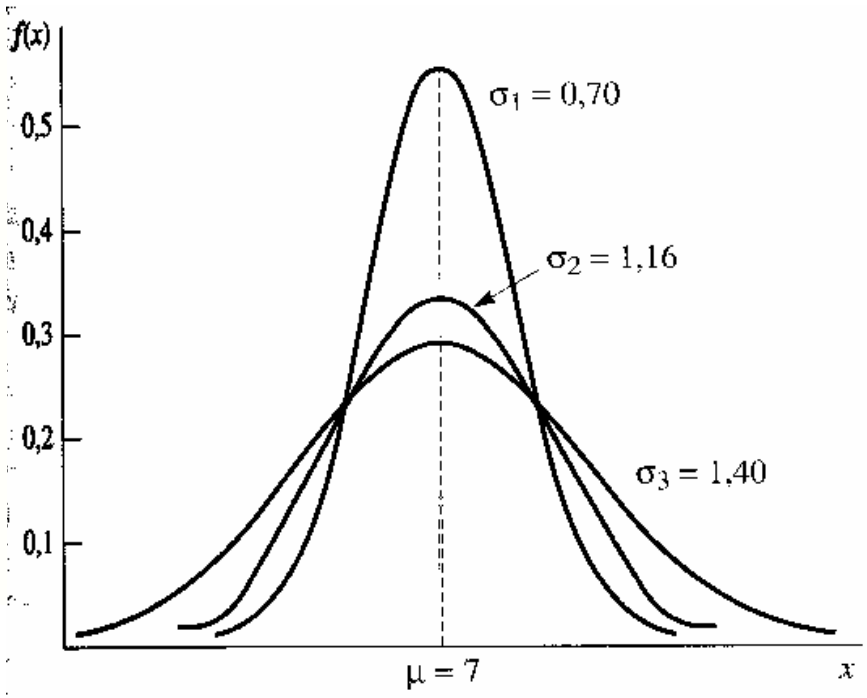
$$F_Z(z) = P(Z \le z) = P\left(\frac{X - \mu}{\sigma} \le z\right) = P(X \le \sigma z + \mu) = F_X(\sigma z + \mu)$$

Derivando :

$$f_Z(z) = \sigma f_X(\sigma z + \mu) = \sigma \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\sigma z + \mu - \mu}{\sigma}\right)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \Rightarrow Z = \frac{X - \mu}{\sigma} \to N(0,1)$$

Ejemplo: A continuación se representan distribuciones normales con diferente media e igual desviación típica, así como distribuciones normales con idéntica media y diferente desviación típica.





Características:

- Depende exclusivamente de los parámetros μ y σ.
- **Función generatriz de momentos:**

$$G_X(t) = G_{\sigma Z+\mu}(t) = e^{t\mu} \cdot G_Z(\sigma t) = e^{t\mu} \cdot e^{\frac{\sigma^2 t^2}{2}} = e^{t\mu + \frac{\sigma^2 t^2}{2}}$$

Nota: " Sea X una variable aleatoria con función generatriz $G_X(t)$ y definimos la variable $Y = a + b.X$, es demostrable que $G_Y(t) = e^{at} \cdot G_X(b.t)$ "

- **Media:** $\mu = E[X] = E[\sigma . Z + \mu] = \sigma E[Z] + \mu = \sigma . 0 + \mu = \mu$
- **Varianza:** $Var(X) = Var(\sigma Z + \mu) = \sigma^2 . Var(Z) = \sigma^2$
- El siguiente teorema permite establecer cómo se comporta una variable obtenida mediante una combinación lineal de variables normales.

Sean $\mathbf{X_i} \sim \mathbf{N(\mu_i, \sigma_i)}$, i = 1,..., n, variables aleatorias independientes y sean $\mathbf{b_i} \in \mathbb{R}-\{0\}$, i = 1,...,n y $\mathbf{a} \in \mathbb{R}$. Entonces:

$$X = \sum_{i=1}^{n} (a + b_i X_i) \approx N\left(\sum_{i=1}^{n} (a + b_i \mu_i), \sqrt{\sum_{i=1}^{n} b_i^2 \sigma_i^2} \right)$$

- *APROXIMACIÓN:* $X \sim B(n,p) \Rightarrow X \sim N(n.p, \sqrt{n.p.q})$

n.p > 4 ó **n.q > 4**

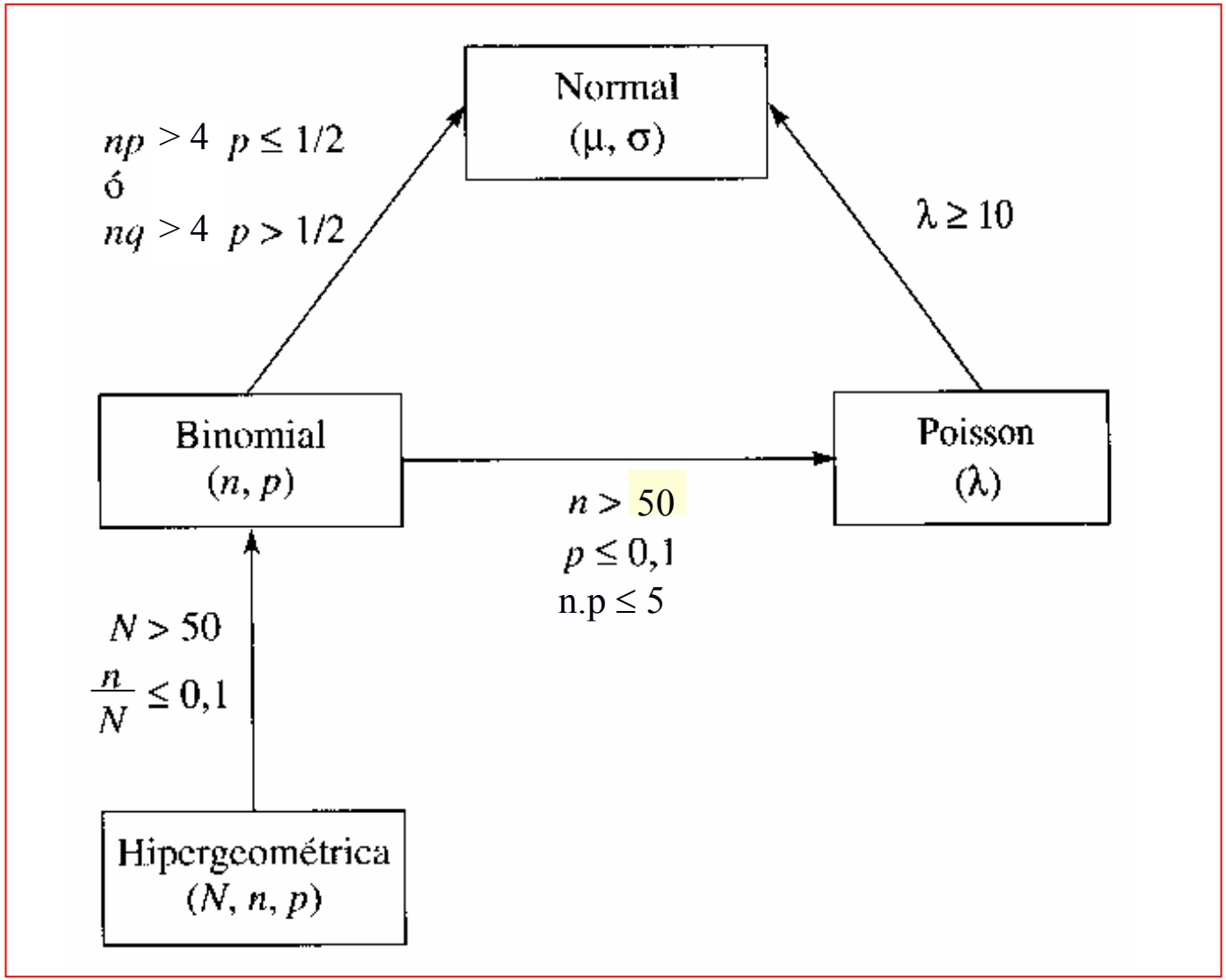
<u>Ejemplo 1:</u> El peso en kilos de las cajas de tomates preparadas en un almacén de empaquetados sigue una N (10,0'5), admitiéndose sólo las cajas con peso comprendido entre 9'5 y 11 kilos.

(a) ¿Cuál es la probabilidad de rechazar una caja?

(b) Cuánto debe pesar una caja de tomates para que el 60 % de las preparadas pesen más que ella?

<u>Ejemplo 2:</u> Se sabe que el 20 % de los pacientes que atiende un oculista no tiene miopía. Si se eligen 25 pacientes al azar, ¿cuál es la probabilidad de que como máximo 10 no tengan miopía?

Aproximaciones



Modelos derivados de la Normal

La **Inferencia Estadística** se basa en el uso de los denominados **estadísticos**, que se obtienen mediante funciones de variables aleatorias normalmente distribuidas que caracterizan a la población en estudio.

En muchas ocasiones, es posible obtener la distribución de estos **estadísticos** apoyándonos en las distribuciones de las variables en que se basan. Por esta razón, desarrollaremos a continuación algunos modelos de distribuciones para variables aleatorias continuas que son función de variables aleatorias normales.

TEOREMA CENTRAL DEL LÍMITE:

Sean $X_1, X_2, ..., X_n$ variables aleatorias independientes e idénticamente distribuidas, con media μ y desviación típica σ, entonces, para $Y = X_1 + X_2 + ... + X_n$ se verifica que:

$$\frac{Y - E[Y]}{\sqrt{V(Y)}} \sim N(0,1) \qquad \text{Por tanto,} \quad \frac{X_1 + X_2 + ... + X_n - n\mu}{\sqrt{n}\,\sigma} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Ejemplo: Supongamos que se lanza un dado 100 veces. Estimar la probabilidad de que la suma de los resultados obtenidos sea al menos de 360.

Distribución Chi-cuadrado de Pearson

Sean $\mathbf{Z_1}$, $\mathbf{Z_2}$, ..., $\mathbf{Z_n}$ variables aleatorias independientes, igualmente distribuidas según una **N(0,1)**, entonces diremos que:

$$X = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$$

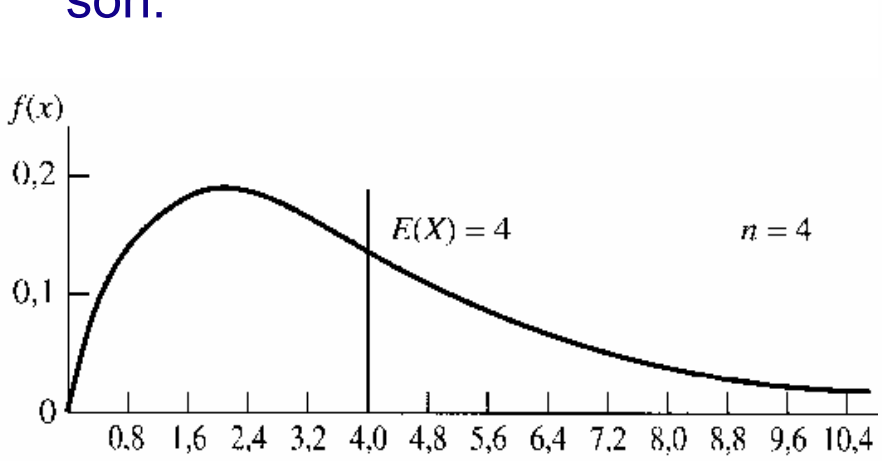
es decir, que sigue una **distribución Chi-cuadrado con n grados de libertad**.

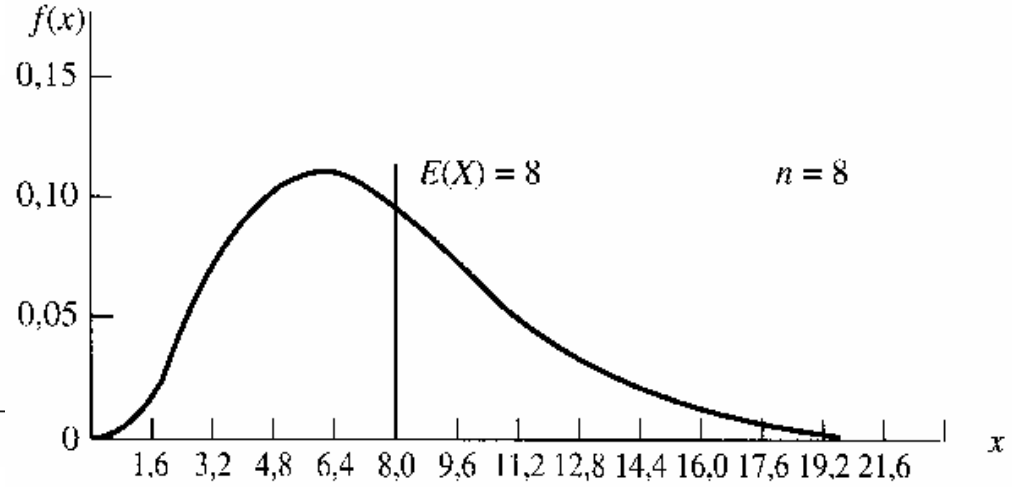
APLICACIÓN: Si $S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n}$ y $\hat{S}^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$ entonces $n\frac{S^2}{\sigma^2} = (n-1)\frac{\hat{S}^2}{\sigma^2} \sim \chi_n^2$

Características:

- Depende exclusivamente de **n**, que son los *grados de libertad*.
- $\mathbf{R_X = \Re^+}$
- Presenta una asimetría positiva (a la derecha) para valores intermedios de **n**.
- **Media**: $\mathbf{E[X] = \mu = n}$
- **Varianza: V(X) = 2n**

- La representación gráfica de la función de densidad depende del número de variables que componen la χ^2, es decir, de **n**. Algunos casos son:





- Aplicando el **Teorema Central del Límite**, cuando **n** es suficientemente grande, se puede aproximar a una **normal:**

$$X \sim \chi^2_n \xrightarrow{n\to\infty} X \sim N\left(\sqrt{2n-1}, 1\right)$$

- Se encuentra tabulada para los distintos valores de **n** ≤ **40**, de forma que:

$$P(\chi^2_n \le \chi^2_{n;1-\alpha}) = 1-\alpha$$

<u>Ejemplo:</u> Dada una variable $X \sim \chi^2_{11}$, calcular $P(\chi^2_{11} \le 4'57)$ y x tal que $P(\chi^2_{11} \le x) = 0'95$.

Distribución t de Student

Sean **Z ~ N (0,1)**, **Y ~ χ^2_n** dos variables aleatorias independientes, la variable aleatoria definida como: $X = \frac{Z}{\sqrt{Y/n}} \sim t_n$

diremos que se distribuye como una **t de Student con n grados de libertad**.

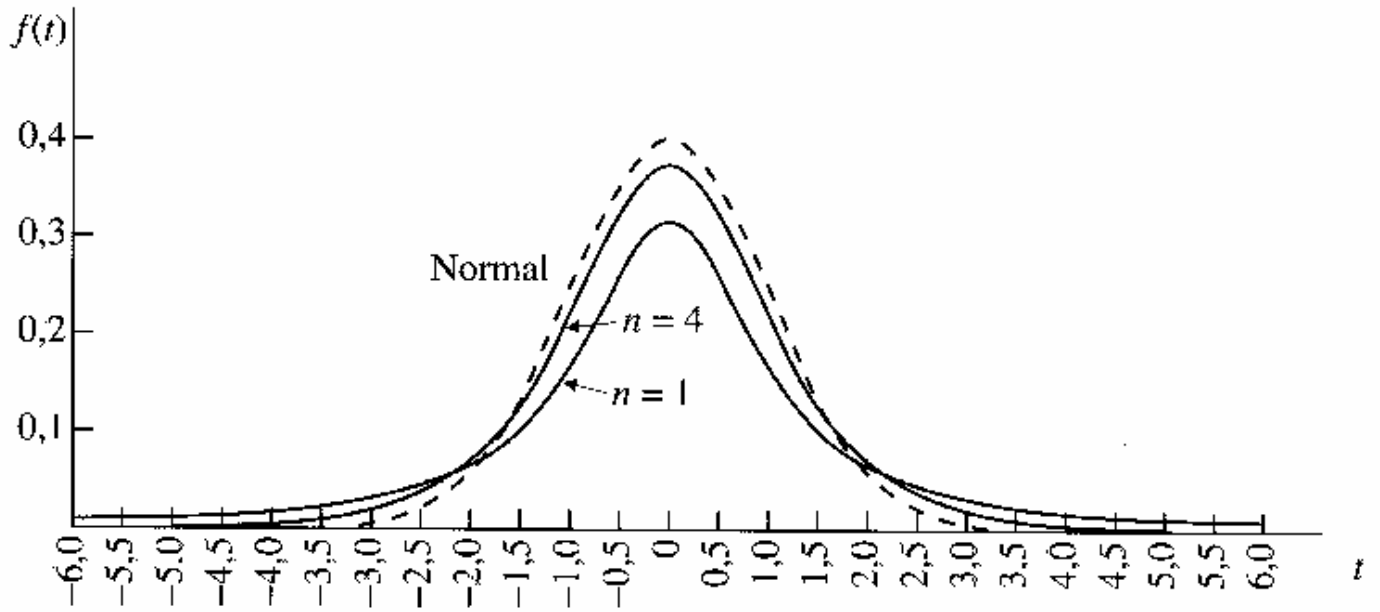
APLICACIÓN: $t_n = \frac{Z}{\sqrt{\frac{\chi^2_n}{n}}} = \sqrt{\frac{Z^2/1}{\chi^2_n/n}} = \sqrt{\frac{\frac{(\bar{X}-\mu)^2}{\sigma^2/n}}{\frac{nS^2}{\sigma^2}/n}} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ o también: $t_{n-1} = \frac{\bar{X}-\mu}{\hat{S}/\sqrt{n}}$

Características:

- Depende exclusivamente de los grados de libertad, **n**.
- **$R_X = \Re$**
- **Media: E [t_n] = 0**
- **Varianza:** Sólo existe para **n > 2**, y vale: $V(t_n) = \frac{n}{n-2}$
- Se encuentra tabulada según los distintos valores de **n**, de manera que: $P(t_n \leq t_{n;1-\alpha}) = 1 - \alpha$

Nota: Al ser simétrica, se verifica que: $t_{n;1-\alpha} = - t_{n;\alpha}$

- La representación gráfica de la función de densidad es bastante parecida a la de la **normal estándar,** aunque menos apuntada (platicúrtica).



Ejemplo: Calcular:

(a) $P(t_{10} \leq 3'169)$

(b) $t_{10,0'9}$

(c) $t_{10,0'4}$

Distribución F de Snedecor

Sean dos variables aleatorias $U \sim \chi^2_{n_1}$ y $V \sim \chi^2_{n_2}$ independientes, entonces se dice que la variable **X** definida como:

$$X = \frac{U/n_1}{V/n_2} \sim F_{n_1, n_2}$$

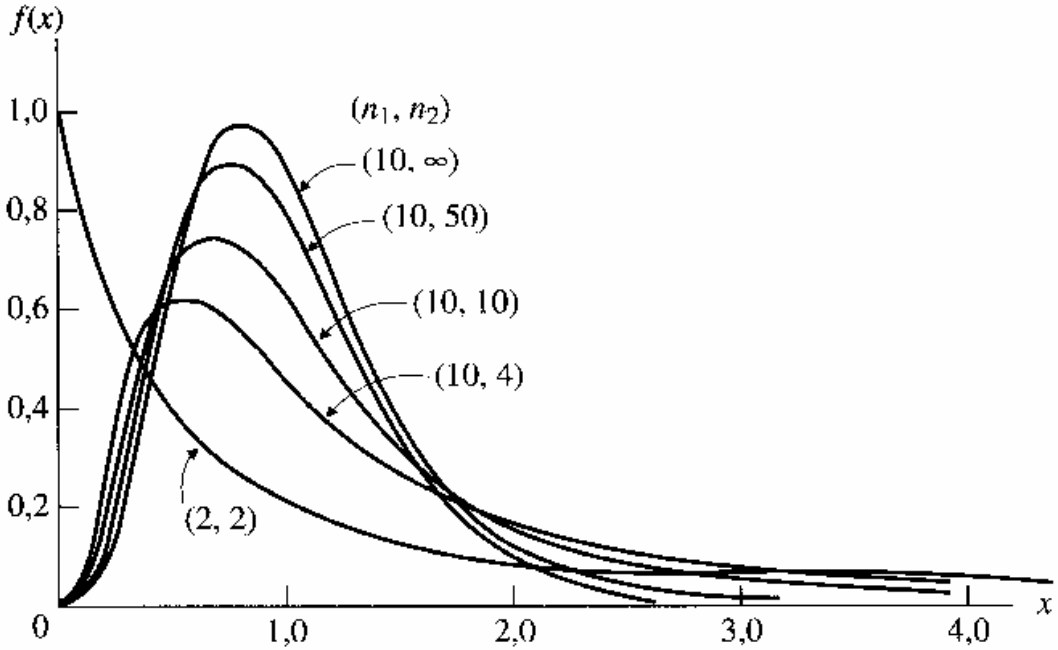
se distribuye según una **F de Snedecor con grados de libertad n_1 y n_2.**

APLICACIÓN: $F_{n_1, n_2} = \frac{\frac{\chi^2_{n_1}}{n_1}}{\frac{\chi^2_{n_2}}{n_2}} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$ o también $F_{n_1-1, n_2-1} = \frac{\frac{\hat{S}_1^2}{\sigma_1^2}}{\frac{\hat{S}_2^2}{\sigma_2^2}}$

Características:

- Depende de los parámetros n_1 y n_2, que son los grados de libertad.
- $R_X = \Re^+$.
- Para valores pequeños de n_1 y n_2 presenta una asimetría positiva.
- Si se altera el orden de n_1 y n_2, se invierte la variable.

- La representación gráfica de la función de densidad es bastante similar a la de la **Chi-cuadrado**, tal y como se muestra a continuación:



- **Media:** Existe sólo si $n_2 > 2$, y su valor es: $E(X) = \frac{n_2}{n_2 - 2}$
- **Varianza:** Existe sólo si $n_2 > 4$, y su valor es: $V(X) = \frac{2\,n_2^2 \cdot (n_1 + n_2 - 2)}{n_1 \cdot (n_2 - 2)^2 \cdot (n_2 - 4)}$
- Se encuentra relacionada con la **t de Student**: $t_n^2 = \frac{Z^2}{\frac{\chi_n^2}{n}} = \frac{\frac{\chi_1^2}{1}}{\frac{\chi_n^2}{n}} = F_{1,n}$

- Se encuentra tabulada para los distintos valores de n_1 y n_2, de manera que: $P\left(F_{n_1,n_1} \leq F_{n_1,n_2;1-\alpha}\right) = 1-\alpha$

Nota: Para otros valores, usar: $F_{n_1,n_2,1-\alpha} = \dfrac{1}{F_{n_2,n_1,\alpha}}$

Ejemplo: Obtener: (a) $P(F_{12,10} \leq 2{,}91)$ (b) $F_{7,9;0,9}$ (c) $F_{5,8;0,01}$

Distribuciones de probabilidad continuas más notables			
Variable	*Parámetros*	*Media*	*Varianza*
Uniforme	a, b	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Normal	μ, σ	μ	σ^2
Chi-cuadrado	n	n	2n
t de Student	n	0	$\dfrac{n}{n-2}$
F de Snedecor	n_1, n_2	$\dfrac{n_2}{n_2-2}$	$\dfrac{2\,n_2^2.(n_1+n_2-2)}{n_1.(n_2-2)^2.(n_2-4)}$

Estadística Empresarial II

Tema 4

Modelos probabilísticos multivariantes

Introducción

En gran cantidad de fenómenos de la vida real hay que analizar **varios caracteres a la vez**, ya sean discretos o continuos, por lo que es necesario asociarle una **variable aleatoria multidimensional**.

Por ello, vamos a estudiar algunos ejemplos de familias de distribuciones de probabilidad multivariantes, concretamente, la **distribución multinomial** y la **distribución normal multivariante**.

Distribución multinomial

Es la generalización de la **distribución binomial** al caso **n-dimensional**, es decir, cuando en cada prueba de Bernouilli se consideran **k** sucesos excluyentes A_1, A_2, ..., A_k, asociados a un experimento, con probabilidades p_1, p_2, ..., p_k, siendo $\sum_{i=1}^{k} p_i = 1$

Consideramos **k** variables aleatorias X_1, X_2, ..., X_k, de manera que:

X_i: "Número de veces que se presenta el suceso A_i al realizar n pruebas independientes"

X: "Número de veces que se presentan los sucesos A_1, A_2, ..., A_k al realizar n pruebas independientes"

$X \sim M(n, p_1, ..., p_k)$

$$R_X = \{(x_1, x_2, ..., x_k) \in \{0,1,...,n\}^k / x_1 + x_2 + ... + x_k = n\}$$

Función de probabilidad

$$P(X_1 = x_1, ..., X_k = x_k) = \frac{n!}{x_1! ..... x_k!} p_1^{x_1} p_2^{x_2} .... p_k^{x_k}$$

Características:

- Depende del número de pruebas independientes **n** y de $\mathbf{p_1, ..., p_k}$.
- **Función generatriz de momentos:**

$$G(t_1\, t_2 \,..\, t_k) = (p_1 e^{t_1} + p_2 e^{t_2} + ... + p_k e^{t_k})^n$$

- **Función de probabilidad de las distribuciones marginales:**

$$P(X_i = x_i) = \binom{n}{x_i} p_i^{x_i} (p_1 + p_2 + ... + p_{i-1} + p_{i+1} + ... + p_k)^{n-x_i} = \binom{n}{x_i} p_i^{x_i} (1 - p_i)^{n-x_i}$$

- **Medias y varianzas de las distribuciones marginales:**

$$\mu_{X_1} = n\, p_1 \quad \mu_{X_2} = n\, p_2 \quad \mu_{X_h} = n\, p_k$$

$$\sigma^2_{X_1} = n\, p_1\, q_1 \quad \sigma^2_{X_2} = n\, p_2\, q_2 \quad \sigma^2_{X_k} = n\, p_k\, q_k$$

- **Covarianzas:** $\mathrm{Cov}(X_i, X_j) = -n\, p_i p_j$

Ejemplo: Tres empresas controlan la totalidad de las ventas de paneles solares en el mercado de Tenerife, de forma que la empresa A controla el 60 %, B el 30 % y C el 10 %. Si elegimos 4 compradores, ¿cuál será la distribución de probabilidad de las posibles compras?

Distribución multinormal

La **distribución multinormal** es una generalización al caso multidimensional del modelo normal estudiado en los modelos probabilísticos continuos unidimensionales. Estudiaremos el caso bidimensional y el n-dimensional, obteniendo las distribuciones **normal bivariante** y **normal multivariante** (o **n-variante**).

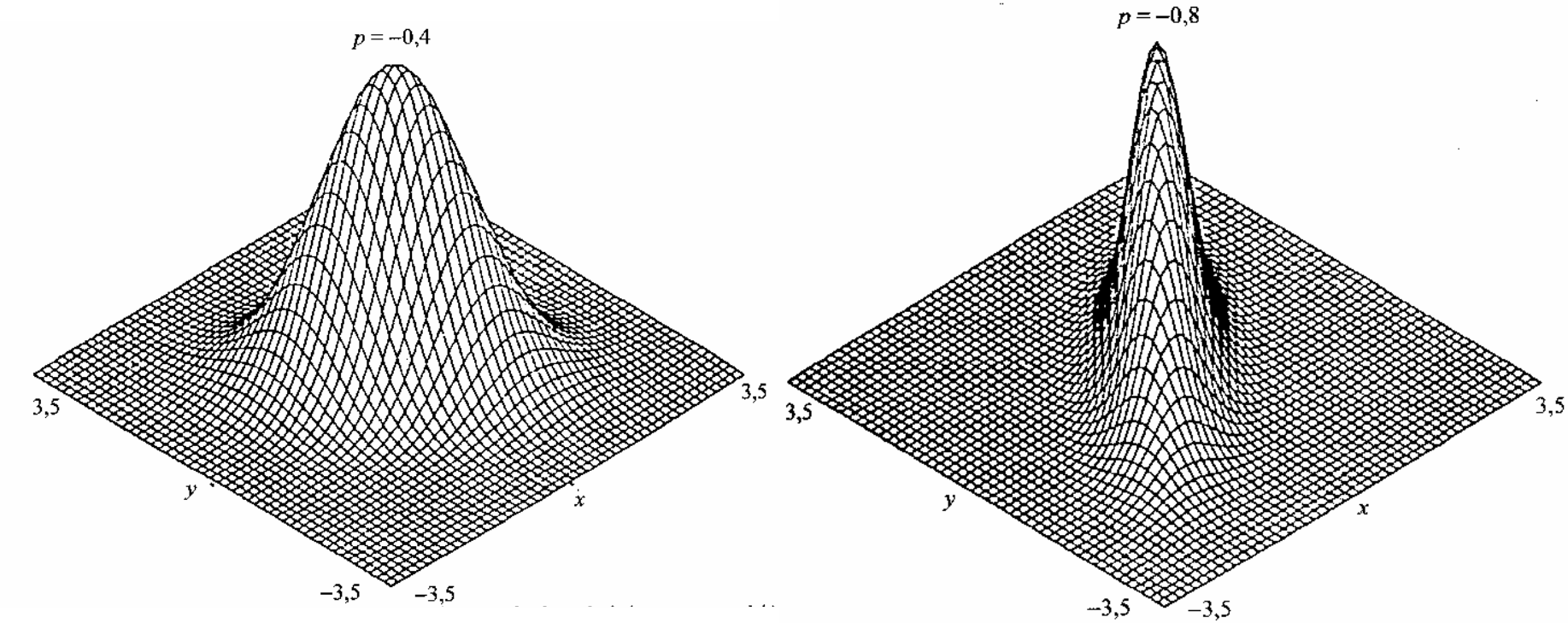
DISTRIBUCIÓN NORMAL BIVARIANTE

$$\mathbf{X} = (X_1, X_2) \sim N_2(\mu, \Sigma)$$

Una variable aleatoria bidimensional continua **X**, sigue una **distribución normal bivariante**, si su función de densidad viene dada por:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}\frac{1}{(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]} \quad \text{siendo} \quad \rho = \frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}$$

Para algunos valores del coeficiente de correlación lineal ρ se obtiene la siguiente representación gráfica de la **función de densidad**:



Características:

- **Media:** $\mu_X = E[X] = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$
- **Matriz de varianzas-covarianzas:** $\Sigma_X = V(X) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$

- Si la variables $\mathbf{X_1}$ y $\mathbf{X_2}$ son incorreladas, entonces son independientes.
- Las **distribuciones marginales** de una **distribución normal bivariante** son **distribuciones normales univariantes**.

$$\mathbf{X_1 \sim N(\mu_1, \sigma_1)} \text{ y } \mathbf{X_2 \sim N(\mu_2, \sigma_2)}$$

<u>Ejemplo:</u> Sean X e Y las desviaciones horizontal y vertical (sobre un plano), respectivamente, de la estación espacial MIR respecto al punto de aterrizaje de éste en el océano Pacífico. Si X e Y son dos variables aleatorias independientes cada una, con distribución normal bivariante y medias $\mu_x = \mu_y = 0$ y varianzas iguales, ¿cuál es la máxima desviación típica permisible de X e Y, que cumpla con el requisito de la Agencia Espacial Rusa de tener una probabilidad de 0,99 de que el vehículo aterrice a no más de 500 millas del punto elegido, tanto en dirección vertical como horizontal?

DISTRIBUCIÓN NORMAL n-VARIANTE:

$$\mathbf{X = (X_1, X_2, ..., X_n) \sim N_n(\mu, \Sigma)}$$

$$f_{\bar{x}}(x) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{n/2}} e^{-\frac{1}{2}\left((x-\mu)'(A^{-1})'\left(A^{-1}(x-\mu)\right)\right)} = \frac{1}{|\Sigma|^{1/2}(2\pi)^{n/2}} e^{-\frac{1}{2}\left[(x-\mu)'\Sigma^{-1}(x-\mu)\right]}$$

siendo $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, y A una matriz regular de orden n tal que $A.A' = \Sigma$

Características:

- **Media:** $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$
- **Matriz de varianzas-covarianzas:** $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}$

- Las **distribuciones marginales** de una **distribución normal multivariante** son **distribuciones normales univariantes**.

- Dada $\mathbf{X \sim N_n(\mu, \Sigma)}$, y $\mathbf{B_{mxn}}$ una matriz cualquiera, entonces:

$$\mathbf{B.X \sim N_m(B.\mu, B.\Sigma.B')}$$

<u>Ejemplo:</u> Para una selección de personal, se hace un examen a los candidatos que consta de tres partes que se califican por separado, de forma que las puntuaciones de cada una siguen aproximadamente una distribución normal multivariante con vector de medias y matriz de covarianzas siguientes:

$$\mu = \begin{pmatrix} 60 \\ 65 \\ 40 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 81 & 18 & 3 \\ 18 & 64 & 3 \\ 3 & 3 & 81 \end{pmatrix}$$

(a) Si se exige para aprobar cada parte por lo menos 50 puntos, ¿Cuál es la probabilidad de aprobar cada parte?

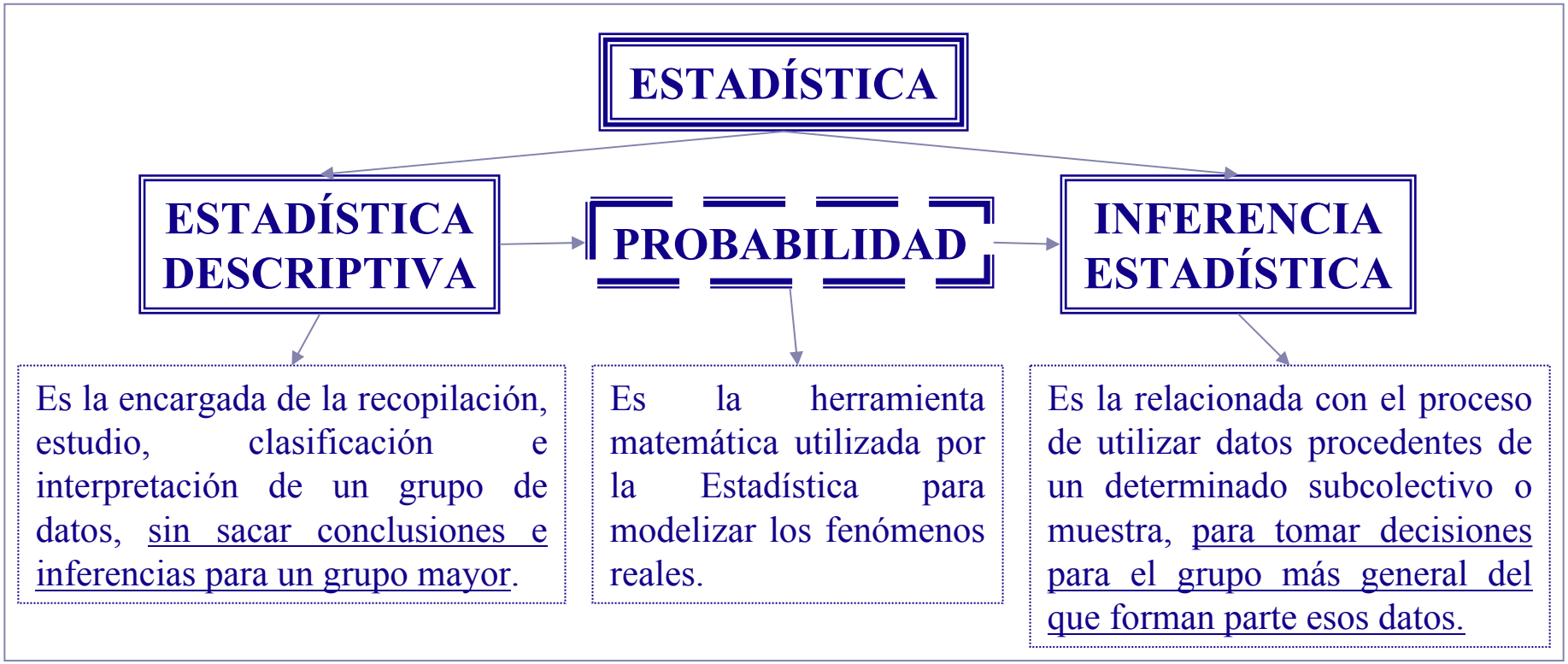
(b) Si para aprobar el examen es suficiente obtener 50 puntos de media entre las tres partes, ¿qué probabilidad hay de aprobar?.

Estadística Empresarial II

Tema 5

Muestreo Aleatorio

Inferencia Estadística



La **Inferencia Estadística** se encargará de inferir o inducir propiedades desconocidas de una **población** (parámetros o tipo de distribución), a partir de la información proporcionada por una **muestra**.

Para medir el grado de certeza de las conclusiones a las que se llegue, se necesitará utilizar algunos conocimientos sobre los diferentes modelos probabilísticos ya estudiados.

La **Inferencia Estadística** puede clasificarse en función de su objetivo y del tipo de información a utilizar:

CLASIFICACIÓN (según su objetivo)

Métodos paramétricos: Son aquellos en los que se supone que los datos provienen de una distribución conocida, centrándose las inferencias en sus parámetros.

Métodos no paramétricos: Son aquellos en los que no se supone conocida la distribución poblacional, introduciéndose hipótesis muy generales respecto a ellas (continuidad, simetría, ...).

CLASIFICACIÓN (según el tipo de información)

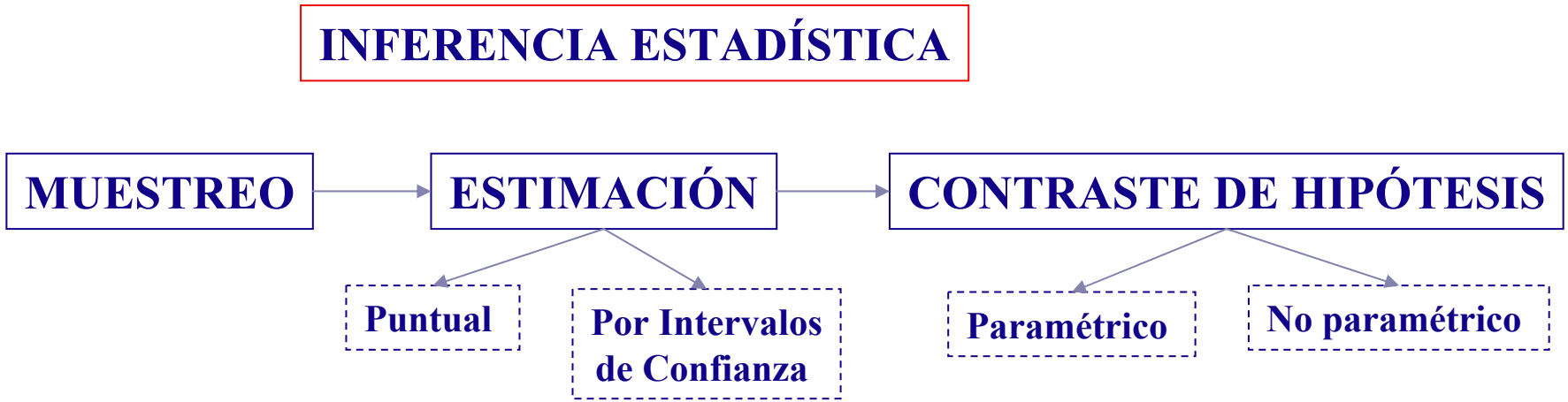
Inferencia Clásica: Se caracteriza porque los parámetros son considerados como valores fijos desconocidos, siendo la única información existente sobre ellos la contenida en la muestra.

Inferencia Bayesiana: En ella, los parámetros son considerados como variables aleatorias, permitiéndose introducir información "a priori" sobre los mismos, además de la obtenida a partir de la muestra.

Muchas veces interesa estimar alguna característica, contrastar alguna hipótesis o tomar alguna decisión respecto a una **población** con un determinado modelo probabilístico; para ello, se procede utilizando la información obtenida a partir de la **muestra**.

Este planteamiento es de gran importancia en el **mundo empresarial y económico**, campos en los que la toma de decisiones lleva asociado un coste o beneficio determinado, y no se puede contar con toda la información existente debido a problemas de tiempo, monetarios, etc.

El esquema a seguir a partir de ahora es el siguiente:



Muestreo

CONCEPTOS:

Los *conceptos fundamentales* que debemos plantear al introducirnos en la Inferencia Estadística son:

Población: Es cualquier colección finita o infinita de individuos o elementos, que no tienen que ser necesariamente seres vivos.

Muestra: Es un subconjunto de la **población**, elegido de forma representativa.

A partir de ellos podemos definir:

Muestreo: Es el procedimiento mediante el cual se obtienen las **muestras** a partir de la **población**.

Tamaño muestral: Se trata del número de individuos que forman la **muestra**, denotándose mediante **n**.

CONVENIENCIA Y LIMITACIONES DEL MUESTREO:

Un **censo** completo de los elementos de una **población** sólo será necesario en algunos casos concretos, ya que, en general, una buena **muestra** puede suministrar información poblacional más precisa y a un coste muy inferior al del **censo**.

¿Cuándo conviene realizar un muestreo?

- Cuando la **población** es demasiado grande.
- Cuando la **población** es suficientemente homogénea desde cierto punto de vista, careciendo de sentido examinar toda la **población**.

¿Cuáles son las ventajas del muestreo?

- *Economía:* El coste de en el **muestreo** es inferior al de un **censo**.
- *Calidad:* Al considerar una **muestra**, se cuida más la precisión de cada observación asociada a cada elemento.

¿Cuándo no se aconseja realizar un muestreo?

- Cuando se necesite información sobre todos los elementos de la **población**.
- Cuando la información deba extenderse a grupos o áreas muy pequeñas de la **población**.

Muestra aleatoria

Partiremos de una **población** de tamaño **N** de la que nos interesa inferir alguna de sus características o tomar alguna decisión sobre la misma. Para ello, recabaremos información sobre ella a través de una **muestra** de tamaño **n**.

$(x_1, x_2, ..., x_n) \rightarrow$ **n** observaciones sucesivas e independientes de una variable **X**.

Muestra genérica: Si consideramos todos los posibles valores de la **muestra**, sin particularizar ninguno, entonces $(x_1, x_2, ..., x_n)$ será una variable aleatoria n-dimensional, denominada **muestra genérica**.

Muestra específica: Es la obtenida asignándole unos valores particulares a la **muestra genérica**.

Muestra aleatoria: Se denominará así cuando la forma de seleccionarla permite conocer la distribución de probabilidad de la **muestra genérica**.

Una **muestra aleatoria** puede ser tomada con reposición (cada elemento analizado se devuelve a la **población**, luego cada extracción será independiente de las anteriores) o sin reposición (los elementos seleccionados no son devueltos, luego cada extracción dependerá de los elementos seleccionados en las anteriores).

Diseño del muestreo

Los **diseños muestrales** se plantean en función de las características de la **población** en estudio y a las pretensiones del mismo. Para evaluar un **diseño muestral** nos basaremos en los siguientes criterios:

- Fiabilidad: El **error del muestreo** es la diferencia entre el valor del **estadístico** obtenido mediante una **muestra aleatoria** y el valor del parámetro poblacional correspondiente. Se mide mediante la **fiabilidad o precisión** del muestreo, que está relacionada con la varianza del estadístico (a mayor varianza, menor fiabilidad).
- Efectividad: Viene asociada al costo del muestreo. Se considera que un diseño es **efectivo** si permite obtener el mayor grado de fiabilidad con el menor costo posible.

Se considera como **estadístico** a cualquier función obtenida a partir de los datos de la **muestra**. Puede definirse como la variable aleatoria unidimensional que es función de la **muestra genérica** $(x_1, x_2, ..., x_n)$.

Estadístico → $g(x_1, x_2, ..., x_n)$

Los principales **diseños muestrales** son:

1.- MUESTREO ALEATORIO SIMPLE: Es aquél en el que todas las muestras de **n** elementos tienen la misma probabilidad de ser escogidas. Por tanto, los elementos de la **población** tendrán la misma probabilidad de ser seleccionados para formar parte de la **muestra**.

2.- MUESTREO ALEATORIO ESTRATIFICADO: Se divide la **población** en **k** subpoblaciones o **estratos**, obtenidos incluyendo en cada uno de ellos elementos parecidos entre sí. De esta manera se obtendrán estratos homogéneos internamente pero con una gran heterogeneidad entre ellos.

Dentro de cada estrato se obtiene una **muestra aleatoria simple**, de manera que, uniendo todas estas submuestras se obtiene la **muestra**.

$$m^* = m_1 \cup m_2 \cup ... \cup m_k$$

3.- MUESTREO POR CONGLOMERADOS: Se va a dividir la **población** en **M** conglomerados (que sean heterogéneos internamente), obteniendo una muestra de **m** de ellos. Se analizan todos los elementos que los componen.

$$m^* = C_1 \cup C_2 \cup ... \cup C_m$$

4.- MUESTREO BIETÁPICO: Es una modificación del anterior, en la que, una vez seleccionados los **conglomerados**, se realiza un **muestreo aleatorio simple** en cada uno de ellos.

$$m^* = m_1 \cup m_2 \cup ... \cup m_m \quad \text{con} \quad m_i \subset C_i, \forall i = 1,...,m$$

5.- MUESTREO POLIETÁPICO: Se trata de una generalización del anterior que se realiza en **k** etapas, combinando **muestreos por conglomerados** con **muestreos aleatorios simples**.

6.- MUESTREO BIFÁSICO: En este tipo de muestreo se toma una **muestra**, de forma rápida, sencilla y poco costosa, a fin de que su información sirva de base para la selección de otra más pequeña, relativa a la característica de estudio.

7.- MUESTREO POLIFÁSICO: Es una generalización del caso anterior que se lleva a cabo en más de dos fases.

8.- MUESTREO SISTEMÁTICO: Esta forma de muestreo se puede emplear cuando los miembros de la **población** están ordenados. Consiste en seleccionar la **muestra** tomando valores cada **k** elementos.

9.- MUESTREO DIRIGIDO: Suele ser de gran utilidad si el investigador está bien familiarizado con la **población** y puede elegir de forma coherente elementos representativos para integrarlos en la **muestra**.

En la práctica, es frecuente el empleo de **métodos mixtos** y **diseños complejos**, obtenidos como combinación de los propuestos anteriormente.

Ejemplos:

1.- Se quiere hacer un estudio sobre la población formada por los estudiantes de Empresariales, obteniendo información sobre:

(a) Opinión acerca de la LOU.

(b) Limpieza de los baños del centro.

Indicar algunos tipos de muestreo que podrían llevarse a cabo.

2.- Se pretende realizar una investigación sobre la opinión de los habitantes de Tenerife sobre el conflicto bélico en Afganistán. Indicar algún tipo de muestreo que podría realizarse.

Distribuciones asociadas al muestreo

En la **Inferencia Estadística** intervienen distribuciones diferenciadas que precisan ser explicadas:

- **Distribución poblacional:** Se trata de la distribución de probabilidad que presenta la variable estudiada **X** para los individuos de la **población**.

$$F(x) = P(X \leq x)$$

- **Distribución de la muestra genérica:** Es la distribución de probabilidad de la variable aleatoria n-dimensional **$(x_1, x_2, ..., x_n)$** (muestra genérica):

En el caso de una muestra aleatoria simple, las variables X_i son i.i.d. que la X.	$F(x_1, x_2, ..., x_n) = F(x_1).F(x_2)...F(x_n)$

- **Distribución de la muestra específica:** Al tomar la **muestra genérica** unos valores concretos, se obtiene la **muestra específica**, que posee una distribución de frecuencias asociada (y no de probabilidad).

- **Distribución del estadístico:** Es la distribución de probabilidad de la variable aleatoria unidimensional dada por el **estadístico** $g(x_1, x_2, ..., x_n)$.

Estadísticos: características y distribuciones

Supongamos que se ha extraído una **muestra** de **n** observaciones de una **población** con media μ y varianza σ^2 para la variable **X.** Representaremos los elementos de la muestra por $(X_1, X_2, ..., X_n)$, donde cada uno de los miembros X_i tendrá media μ y varianza σ^2.

MEDIA MUESTRAL: $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ → **Se trata de una v.a. unidimensional, al ser función de las variables aleatorias X_i.**

$$E[\overline{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\, n\, \mu = \mu$$

X_i y X_j independientes

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

$Cov(X_i, X_j) = 0$

$$= \frac{1}{n^2} E\left(\sum_{i=1}^{n}(X_i - \mu)^2 + \sum_{i\neq j}(X_i - \mu).(X_j - \mu)\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n} E(X_i - \mu)^2 + \sum_{i\neq j} E[(X_i - \mu).(X_j - \mu)]\right) =$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} E(X_i - \mu)^2 = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}\, n\, \sigma^2 = \frac{\sigma^2}{n}$$

En caso de **poblaciones finitas** de tamaño **N**, la varianza de la **media muestral** se verá afectada por un *factor de corrección*, obteniendo:

$$V(\overline{X}) = \frac{\sigma^2}{n} \cdot \frac{N - n}{N - 1}$$

Si el tamaño muestral **n** no es una fracción muy pequeña del tamaño poblacional **N**, no podremos asumir la independencia de los $\mathbf{X_i}$.

Ejemplo: Sean 4 lotes A, B, C y D de 20 latas de atún, cada uno de los cuales tiene 1, 3, 4 y 5 unidades defectuosas, respectivamente. Se eligen al azar 2 de los 4 lotes. Calcular el número medio muestral de unidades defectuosas, obteniendo su media y su varianza.

1.- Distribución de la media muestral de una población normal.

$$\mathbf{X \sim N(\mu, \sigma)} \longrightarrow \mathbf{X_i \sim N(\mu, \sigma),\ i = 1, 2, ..., n}$$

(A) Con varianza poblacional conocida:

$$\text{Si } X_i \sim N(\mu, \sigma), \text{ entonces } \frac{X_i}{n} \sim N\left(\frac{\mu}{n}, \frac{\sigma}{n}\right) \quad \text{Luego, } \overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} \sim N\left(n \cdot \frac{\mu}{n}, \sqrt{\sum_{i=1}^{n} \left(\frac{\sigma}{n}\right)^2}\right) = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Por lo tanto, se obtiene que:

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = Z \sim N(0,1)$$

Ejemplo: *Se tiene una máquina de llenado para vaciar 500 gramos de gofio en una bolsa. Supóngase que la cantidad de gofio que se coloca en cada bolsa es una variable aleatoria normalmente distribuida con media de 500 gramos y desviación típica igual a 20 gramos. Para verificar que el peso promedio de cada bolsa se mantiene en 500 gramos se toma una muestra aleatoria de 25 de éstas en forma periódica y se pesa el contenido de cada bolsa. El gerente de la planta ha decidido detener el proceso y encontrar el fallo cada vez que el valor promedio de la muestra sea mayor de 510 gramos o menor de 490 gramos. Obtener la probabilidad de detener el proceso.*

(B) Con varianza poblacional desconocida:

σ^2 desconocida ⇒ Usaremos como aproximación la **cuasivarianza:** $\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

$$\text{Si } \bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow \frac{(\bar{X}-\mu)}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Habrá que determinar la distribución de:

$$\frac{\bar{X}-\mu}{\frac{\hat{s}}{\sqrt{n}}} = \frac{\bar{X}-\mu}{\frac{1}{\sqrt{n}}\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}} = \frac{\sqrt{n}\,.(\bar{X}-\mu)}{\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}}$$

$$\sum_{i=1}^{n}(X_i-\bar{X})^2 = \sum_{i=1}^{n}(X_i-\mu+\mu-\bar{X})^2 = \sum_{i=1}^{n}\left[(X_i-\mu)-(\bar{X}-\mu)\right]^2 = \sum_{i=1}^{n}\left[(X_i-\mu)^2+(\bar{X}-\mu)^2-2(X_i-\mu)(\bar{X}-\mu)\right] =$$

$$= \sum_{i=1}^{n}(X_i-\mu)^2 + n(\bar{X}-\mu)^2 - 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) = \sum_{i=1}^{n}(X_i-\mu)^2 + n(\bar{X}-\mu)^2 - 2(\bar{X}-\mu)(n\bar{X}-n\mu) =$$

$$= \sum_{i=1}^{n}(X_i-\mu)^2 + n(\bar{X}-\mu)^2 - 2n(\bar{X}-\mu)^2 = \sum_{i=1}^{n}(X_i-\mu)^2 - n(\bar{X}-\mu)^2$$

Dividiendo ambos miembros por la varianza σ^2, se obtiene que:

$$\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}-\frac{n(\bar{X}-\mu)^2}{\sigma^2}=\sum_{i=1}^{n-1}Z_i^2\sim\chi^2_{n-1}$$

Entonces:

$$\frac{\sqrt{n}\,(\bar{X}-\mu)}{\sqrt{\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n-1}}}=\sqrt{\frac{n\,(\bar{X}-\mu)^2}{\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n-1}}}=\sqrt{\frac{n\,\frac{(\bar{X}-\mu)^2}{\sigma^2}}{\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{(n-1)\,\sigma^2}}}=\sqrt{\frac{\frac{(\bar{X}-\mu)^2}{\sigma^2/n}}{\frac{1}{n-1}\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma^2}}}=\sqrt{\frac{Z^2}{\frac{\chi^2_{n-1}}{n-1}}}=\sqrt{\frac{X_1^2}{\frac{\chi^2_{n-1}}{n-1}}}=t_{n-1}$$

Por tanto:

$$\frac{\bar{X}-\mu}{\frac{\hat{s}}{\sqrt{n}}}=t_{n-1}$$

En el caso de que el tamaño muestral **n** fuera lo suficientemente grande, podría aplicarse el Teorema Centra de Límite y plantear que:

$$\frac{\bar{X}-\mu}{\frac{\hat{s}}{\sqrt{n}}}=Z\sim N(0,1)$$

Ejemplo: *En un estudio publicado en el diario "El País", se asegura que, para un coche compacto particular, el consumo de gasolina en carretera es de 1 litro cada 15 kilómetros. Una organización independiente de consumidores adquiere uno de estos coches y lo somete a prueba con el propósito de verificar la cifra indicada por el diario "El País". El coche recorrió una distancia de 100 kilómetros en 25 ocasiones. En cada recorrido se anotó el número de litros necesarios para realizar el viaje. En los 25 ensayos, la cuasidesviación típica tomó un valor de 1,5 kilómetros por litro respectivamente. Si se supone que el número de kilómetros que se recorren por litro es una variable aleatoria distribuida normalmente, con base en esta prueba, ¿cuál es la probabilidad de que la media muestral sea inferior a la poblacional?*

2.- Diferencia de las medias muestrales de dos poblaciones normales independientes.

$X_1 \sim N(\mu_1, \sigma_1)$ $\quad$ $X_2 \sim N(\mu_2, \sigma_2)$

Son dos variables aleatorias definidas para dos poblaciones independientes.

(A) Con varianzas poblacionales conocidas:

Sabemos que:

$$\left.\begin{array}{l} \overline{X}_1 \sim N\left(\mu_1, \dfrac{\sigma_1}{\sqrt{n_1}}\right) \\ \overline{X}_2 \sim N\left(\mu_2, \dfrac{\sigma_2}{\sqrt{n_2}}\right) \end{array}\right\} \Rightarrow \overline{X}_1 - \overline{X}_2 \approx N\left((\mu_1 - \mu_2), \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Por tanto: $$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = Z \sim N(0,1)$$

Ejemplo: *El Servicio Canario de Salud está realizando un estudio sobre el consumo de cigarrillos en la Provincia de Santa Cruz de Tenerife y en la Provincia de Las Palmas de Gran Canaria. Por trabajos realizados anteriormente, conoce que las varianzas poblacionales son, respectivamente, 100 y 64. En cuanto a los consumos medios en ambas provincias, no los conoce por lo que toma dos muestras de tamaño 49 y 36. Dicha entidad piensa que el consumo medio de cigarrillos en la Provincia de Santa Cruz de Tenerife es igual al de la Provincia de Las Palmas de Gran Canaria ¿Cuál es la probabilidad de que el consumo medio muestral en Santa Cruz de Tenerife supere al de Las Palmas de Gran Canaria?*

(B) Con varianzas poblacionales desconocidas:

- Si las **muestras** son **grandes** (**n** > 30):

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} = Z$$

- Si las **muestras** son **pequeñas**:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} = t_{n_1+n_2-2}$$

NOTA: En este segundo caso, las varianzas son estimadas con poca información, lo que les confiere una relativamente baja fiabilidad. En el caso de que las varianzas poblacionales fueran supuestamente iguales, el riesgo asumido no sería excesivo, ya que ambas poblaciones están dispersas de manera semejante. Pero si fueran distintas, los riesgos de las inferencias que se pretenden realizar con este estadístico serían, cuando menos, preocupantes.

- Varianzas supuestamente iguales: Dado que las varianzas poblacionales se suponen iguales, se considera que sus estimadores (**cuasivarianzas**) también lo deberán ser, aunque, dada las aleatoriedad de las muestras, no es normal que ocurra, por lo que sustituiremos las **cuasivarianzas muestrales** por una media ponderada de las mismas.

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\hat{S}_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = t_{n_1 + n_2 - 2}$$

siendo $\hat{S}_p^2 = \frac{(n_1 - 1) \cdot \hat{S}_1^2 + (n_2 - 1) \cdot \hat{S}_2^2}{n_1 + n_2 - 2}$

- Varianzas supuestamente distintas: En este caso, se intenta compensar el riesgo asumido modificando la distribución del estadístico, reduciendo los grados de libertad del mismo mediante la **Aproximación de Welch**.

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} = t_f$$

siendo la Aproximación de WELCH: $f = \frac{\left(\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{S}_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{\hat{S}_2^2}{n_2}\right)^2}{n_2 + 1}} - 2$

Ejemplo: El gerente de la refinería de CEPSA en Tenerife piensa modificar el proceso para producir gasolina a partir de petróleo crudo. El gerente hará la modificación sólo si la gasolina promedio que se obtiene por este nuevo proceso (expresada como porcentaje del crudo) aumenta su valor con respecto al proceso en uso. Con base en un experimento de laboratorio y mediante el empleo de dos muestras aleatorias de tamaño 12, una para cada proceso, presentando la cantidad de gasolina del proceso en uso una cuasidesviación típica de 2,3, y para el proceso propuesto, de 2,7. El gerente piensa que los resultados proporcionados por los dos procesos son variables aleatorias independientes normalmente distribuidas con varianzas iguales. ¿Cuál es la probabilidad de que la cantidad de gasolina media muestral sea mayor para el nuevo proceso?

Ejemplo: La empresa Agroman desea estimar el número total de horas/hombre perdidas debido a accidentes de sus obreros y técnicos en un mes determinado. Por experiencia, sabe que la dispersión es distinta para esos dos tipos de trabajadores, aunque no conoce las varianzas poblacionales, y considera que las medias poblacionales coinciden. Para ello se tomó una muestra de 18 obreros y 10 técnicos.

Obreros	8	0	6	7	9	18	24	16	0	4	5	2	0	32	16	4	8	0

Técnicos	4	0	8	3	1	5	4	7	2	8

¿Cuál es la probabilidad de que el número medio muestral de horas/hombre perdidas sea mayor en el caso de los obreros que para los técnicos?

VARIANZA MUESTRAL: $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

$$E(S^2) = E\left(\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n}\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}(X_i-\mu-\bar{X}+\mu)^2\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}[(X_i-\mu)-(\bar{X}-\mu)]^2\right) =$$

$$= \frac{1}{n}E\left[\sum_{i=1}^{n}(X_i-\mu)^2 + \sum_{i=1}^{n}(\bar{X}-\mu)^2 - 2\sum_{i=1}^{n}(X_i-\mu)(\bar{X}-\mu)\right] = \frac{1}{n}\left[\sum_{i=1}^{n}E(X_i-\mu)^2 + nE(\bar{X}-\mu)^2 - 2E(\bar{X}-\mu)\left(n\sum_{i=1}^{n}\frac{X_i}{n} - n\mu\right)\right] =$$

$$= \frac{1}{n}\left[n\sigma^2 + n\frac{\sigma^2}{n} - 2E(X-\mu)n(\bar{X}-\mu)\right] = \frac{1}{n}\left[n\sigma^2 + n\frac{\sigma^2}{n} - 2nE(X-\mu)^2\right] = \frac{1}{n}\left[n\sigma^2 + n\frac{\sigma^2}{n} - 2n\frac{\sigma^2}{n}\right] = \frac{1}{n}\left[n\sigma^2 - n\frac{\sigma^2}{n}\right] =$$

$$= \sigma^2 - \frac{\sigma^2}{n}$$

$$V(S^2) = \frac{2(n-1)\sigma^4}{n}$$

Distribución de la varianza muestral:

$X \sim N(\mu, \sigma) \longrightarrow X_i \sim N(\mu, \sigma)$, $i = 1, 2, ..., n$

$$\frac{n}{\sigma^2}S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} - \frac{n(\bar{X}-\mu)^2}{\sigma^2} = \sum_{i=1}^{n-1}Z_i^2 \sim \chi^2_{n-1}$$

Ejemplo: Un proceso produce lotes de un producto químico cuyos niveles de concentración de impurezas siguen una distribución normal con varianza 1'75. Se extrae una muestra aleatoria de 20 lotes. Hallar la probabilidad de que la varianza muestral sea mayor que 3'10.

CUASIVARIANZA MUESTRAL:

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$E(\hat{s}^2) = E\left(\frac{n}{n-1}\cdot s^2\right) = \frac{n}{n-1}\cdot E(s^2) = \frac{n}{n-1}\cdot\frac{n-1}{n}\cdot\sigma^2 = \sigma^2$$

$$Var(\hat{s}^2) = \frac{2\,\sigma^4}{n-1}$$

Distribución de la cuasivarianza muestral:

$$X \sim N(\mu, \sigma) \longrightarrow X_i \sim N(\mu, \sigma)\ ,\ i = 1, 2, ..., n$$

$$\frac{(n-1)\cdot\hat{S}^2}{\sigma^2} = \frac{(n-1)\cdot\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{(n-1)}}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} = \frac{n}{\sigma^2}S^2 \sim \chi^2_{n-1}$$

Ejemplo: En un proceso de llenado de bolsas de café, se quiere estudiar la variabilidad en la medición de su peso. Con base en la experiencia, se sabe que dicha medición es una variable aleatoria normalmente distribuida con media 1000 y cuasidesviación típica igual a 10 gramos. Si se toma una muestra aleatoria procedente del proceso de manufactura de los instrumentos de tamaño 25, ¿ cuál es la probabilidad de que el valor de la cuasivarianza muestral sea mayor de 14 unidades cuadradas?.

Distribución del cociente de cuasivarianzas muestrales:

$X_1 \sim N(\mu_1, \sigma_1)$ $X_2 \sim N(\mu_2, \sigma_2)$

Son dos variables aleatorias definidas para dos poblaciones independientes.

Sabemos que:

$$\frac{(n_1 - 1)\hat{S}_1^2}{\sigma_1^2} \sim \chi^2_{n_1 - 1}$$

$$\frac{(n_2 - 1)\hat{S}_2^2}{\sigma_2^2} \sim \chi^2_{n_2 - 1}$$

$$\frac{\dfrac{(n_1 - 1)\hat{S}_1^2}{\sigma_1^2} \Big/ n_1 - 1}{\dfrac{(n_2 - 1)\hat{S}_2^2}{\sigma_2^2} \Big/ n_2 - 1} \sim F_{(n_1 - 1),(n_2 - 1)}$$

Ejemplo: Se quiere estudiar la variabilidad del número de turistas que se alojan en dos zonas distintas de la isla de Tenerife. Con base en la experiencia, se sabe que el número de turistas es una variable aleatoria normalmente distribuida, por lo que se toman dos muestras aleatorias de tamaños 8 y 7 respectivamente en las zonas A y B. Sabiendo que para ambas zonas la varianza poblacional es la misma, ¿cuál es la probabilidad de que la cuasivarianza muestral del número de turistas alojados en la zona A sea menor que la de la zona B?.

PROPORCIÓN MUESTRAL:

$$\hat{p} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Indica la **proporción de éxitos** que tienen lugar en la muestra de tamaño **n**.

$X_i \sim b(p)$, i = 1, ..., n, independientes. → **$X \sim B(n, p)$**

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = E\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\, n\, p = p$$

$$Var(\hat{p}) = E(\hat{p} - p)^2 = E\left(\sum_{i=1}^{n} \frac{X_i}{n} - p\right)^2 = E\left(\frac{\sum_{i=1}^{n} X_i - n\,p}{n}\right)^2 = \frac{1}{n^2} E\left(\sum_{i=1}^{n} (X_i - p)\right)^2 = \frac{1}{n^2} E\left(\sum_{i=1}^{n} (X_i - p)^2 + \sum_{i \neq j} (X_i - p).(X_j - p)\right) =$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n} E(X_i - p)^2 + \sum_{i \neq j} \underbrace{E(X_i - p).(X_j - p)}_{\| \; 0 \text{ (al ser los } X_i \text{ independientes)}}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\, n\, p\, q = \frac{p\, q}{n}$$

1.- Distribución de la proporción muestral cuando el tamaño muestral es grande:

Aplicando el **Teorema Central del Límite**:

$$\hat{p} = \sum_{i=1}^{n} \frac{X_i}{n} \sim N\left(p, \frac{p\,q}{n}\right)$$

Ejemplo: Tenemos una muestra de 250 hoteles de la población de hoteles de cuatro estrellas en Canarias para estimar la proporción de hoteles de este tipo que tengan instalados paneles solares para agua caliente.. Supongamos que, de hecho, el 30% de todos los hoteles de esta población tiene instalados paneles solares para agua caliente. Hallar la probabilidad de que la proporción de hoteles de la muestra con instalación de paneles solares para agua caliente esté entre 0,25 y 0,35.

2.- Distribución de la diferencia de proporciones muestrales cuando el tamaño de las muestras es grande:

$$\hat{p}_1 \sim N\left(p_1, \sqrt{\frac{p_1 \cdot q_1}{n_1}}\right) \quad y \quad \hat{p}_2 \sim N\left(p_2, \sqrt{\frac{p_2 \cdot q_2}{n_2}}\right) \Rightarrow \hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1\,q_1}{n_1} + \frac{p_2\,q_2}{n_2}}\right)$$

Ejemplo: Tenemos una muestra de 250 consumidores del área metropolitana de Santa Cruz de Tenerife y 150 consumidores del sur de la isla. Estudios anteriores indican que el porcentaje de consumidores del área metropolitana de Santa Cruz que compra en grandes superficies es del 80%, mientras que en el sur es del 60%. Hallar la probabilidad de que la diferencia de proporciones de las muestras de consumidores oscile entre 0,2 y 0,25.

Estadística Empresarial II

Tema 6

Estimación puntual y por intervalos de confianza

Introducción

El objetivo principal de la **Inferencia Estadística** es inferir o inducir propiedades desconocidas de una **población** (parámetros o tipo de distribución), a partir de la información proporcionada por una **muestra**.

Generalmente, nos encontramos ante características o **parámetros** de la población que son desconocidos, y que habrá que **estimarlos** usando la información suministrada por una **muestra** representativa de la **población**.

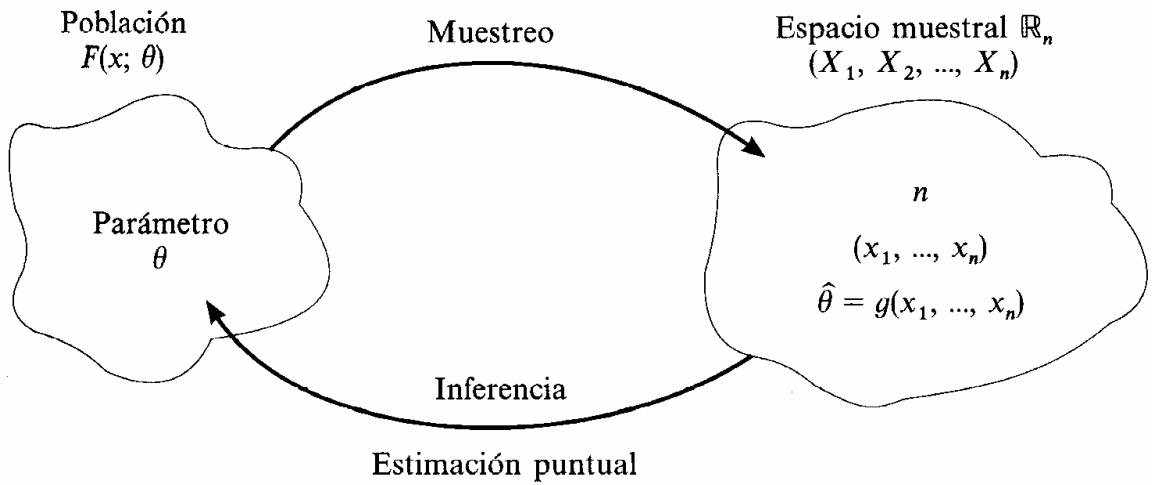
De esta manera, habrá que obtener **estimadores**, que serán estadísticos aplicados a la determinación de los parámetros poblacionales que se pretenden inferir. A las realizaciones específicas de los **estimadores** (que, al ser estadísticos, serán v.a. que dependen de los valores de la muestra) se les denomina **estimaciones**.

Se puede dividir la **estimación** en base a su precisión, dando lugar a la **estimación puntual** y a la **estimación por intervalos de confianza**.

Estimación Puntual

Supongamos que disponemos de una **población** para la que una variable aleatoria **X** se distribuye según una determinada función de distribución **$F(x, \theta_1, \theta_2, ..., \theta_k)$**, que depende de una serie de **parámetros**. Si estos parámetros son *desconocidos*, habrá que estimarlos, usando para ello una muestra aleatoria.

La **Estimación Puntual** consiste en determinar unos valores numéricos que hagan el papel de los valores de **$(\theta_1, \theta_2, ..., \theta_k)$**, para lo cual utilizaremos funciones de los valores de la muestra **$g(X_1, X_2, ..., X_n)$** (estadísticos) que denominaremos **estimadores** de los **parámetros**.



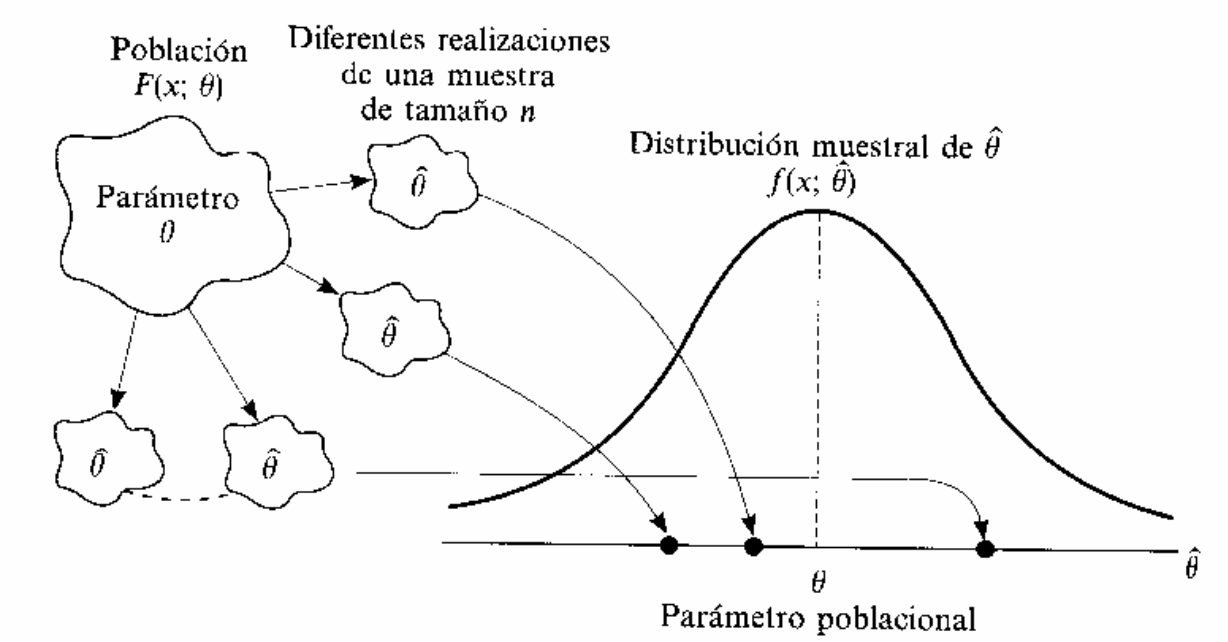
Estimador puntual

$$\hat{\theta} = g(X_1, X_2, ..., X_n)$$

Estimación puntual

$$\hat{\theta} = g(x_1, x_2, ..., x_n)$$

La **estimación puntual** proporciona el valor numérico del **parámetro** a estimar, pero éste dependerá de la **muestra aleatoria** que haya sido seleccionada. Así pues, considerando distintas muestras, se podría producir diferencias en los valores estimados.



Al ser la extracción de una muestra un fenómeno aleatorio, su utilización para estimar una característica poblacional puede conducir a resultados erróneos. Por tanto, sería conveniente utilizar un procedimiento que permita conocer el riesgo de cometer errores en cada caso, como puede ser la **Estimación por Intervalos**.

Propiedades de los Estimadores

Las propiedades que debe reunir un buen **estimador** son:

- **SUFICIENCIA:** Un estimador $\hat{\theta}$ será **suficiente** para un parámetro θ cuando utiliza toda la información relevante contenida en la muestra aleatoria con respecto a θ.

Ejemplo: Para **X ~ N (μ,σ)**, $\hat{\mu} = \overline{X}$ *es un estimador suficiente para* μ.

- **CONSISTENCIA:** Un estimador $\hat{\theta}$ de un parámetro θ es **consistente** si converge en probabilidad al valor de θ cuando el tamaño muestral **n** crece. Para un número pequeño ε **> 0:**

$$\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| < \varepsilon\right) = 1 \text{ o bien } \lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) = 0$$

Teorema:

Si $\lim_{n\to\infty} E[\hat{\theta}] = \theta$ y $\lim_{n\to\infty} V[\hat{\theta}] = 0$ ⟹ entonces $\hat{\theta}$ es un estimador **consistente** de θ.

Ejemplo: *Dada* **X** una v.a. de media poblacional μ y varianza poblacional σ^2, entonces $\overline{X}$ es un estimador consistente de μ. ($E(\overline{X}) = \mu$ y $V(\overline{X}) = \sigma^2/n$)

- **INSESGADEZ:** Un estimador $\hat{\theta}$ de θ es **insesgado** o **centrado** cuando su valor esperado con el parámetro a estimar θ. Es decir:

$$E(\hat{\theta}) = \theta$$

En caso contrario, el estimador será **sesgado** y a la desviación entre su valor esperado y el parámetro θ se le denomina **sesgo**.

$$E(\hat{\theta}) \neq \theta \Rightarrow E(\hat{\theta}) = \theta + \delta(\theta) \;/\; \delta(\theta) = \text{Sesgo}$$

Ejemplos:

$$\begin{cases} E(\overline{X}) = \mu \Rightarrow \overline{X} \text{ es insesgado} \\ E(S^2) = \dfrac{n-1}{n}\sigma^2 = \sigma^2 - \dfrac{\sigma^2}{n} \Rightarrow S^2 \text{ es sesgado, con sesgo igual a } -\dfrac{\sigma^2}{n} \end{cases}$$

- **EFICIENCIA:**

Si un estimador $\hat{\theta}$ es **insesgado**, el valor promedio que toma es el parámetro a estimar θ, sin embargo, esto no implica que el estimador tenga una alta probabilidad de estar cerca de θ para una muestra dada.

Dados $\hat{\theta}_1$ y $\hat{\theta}_2$ dos estimadores insesgados de θ; el estimador que tenga menor varianza tendrá mayor probabilidad de estar cerca del parámetro desconocido θ. Así:

$$\hat{\theta}_1 \text{ es más eficiente que } \hat{\theta}_2 \Leftrightarrow V(\hat{\theta}_1) < V(\hat{\theta}_2)$$

Al estimador insesgado con menor varianza de todos, $\hat{\theta}$, se le llama **estimador más eficiente** o **estimador insesgado de mínima varianza**. Este estimador posee una varianza igual a:

$$V(\hat{\theta}) = \frac{1}{E\left(\frac{\partial \log L}{\partial \theta}\right)^2} = \frac{1}{n \cdot E\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^2}$$

→ **COTA DE CRAMER-RAO**

siendo **L** la **función de verosimilitud**, que para una muestra aleatoria simple **(X_1, X_2, ..., X_n)** de una población para la que **X** tiene función de densidad **f(x,θ)**, tiene la forma:

$$L(X_1, X_2, \ldots, X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

Ejercicio: Comprobar que dada una población para la que **X** *es* **N(μ, σ)**, *se verifica que la media muestral es el estimador* ***más eficiente*** *de* **μ**.

Obtención de estimadores

A continuación se desarrollarán los principales **métodos de obtención de estimadores** para los parámetros desconocidos de una población, a partir de la información suministrada por una muestra.

1.- MÉTODO DE LOS MOMENTOS: Fue propuesto por K. Pearson en 1894. Se trata de un método muy sencillo que genera estimadores razonables.

El **método de los momentos** consiste en igualar los **k** primeros momentos respecto al origen de la variable **X** representativa de la población (dependientes de **k** parámetros desconocidos θ_j , **j = 1, ..., k**) a los **k** primeros momentos respecto al origen en la muestra. Los **estimadores** buscados se obtendrán resolviendo el sistema resultante.

Sea **X** con función de densidad **f (X, θ_1, θ_2, ..., θ_n)** y **(X_1, X_2, ...,X_n)** una muestra aleatoria de tamaño **n:**

$$\alpha_1 = \int_{-\infty}^{\infty} X \cdot f(X;\theta_1,....,\theta_k) \quad \alpha_2 = \int_{-\infty}^{\infty} X^2 \cdot f(X;\theta_1,....,\theta_k) \quad ..... \quad \alpha_k = \int_{-\infty}^{\infty} X^k \cdot f(X;\theta_1,....,\theta_k)$$

$$a_1 = \frac{\sum_{i=1}^{n} X_i}{n} \qquad a_2 = \frac{\sum_{i=1}^{n} X_i^2}{n} \qquad \ldots\ldots\ldots \qquad a_k = \frac{\sum_{i=1}^{n} X_i^k}{n}$$

Igualando los α_i a los $\mathbf{a_i}$ se obtendrá un sistema de ecuaciones, de cuya resolución saldrán los estimadores $\hat{\theta}_j$:

$$\begin{aligned}
&\hat{\theta}_1 = G_1(a_1, a_2, \ldots\ldots a_k) \\
&\hat{\theta}_2 = G_2(a_1, a_2, \ldots\ldots a_k) \\
&\ldots\ldots\ldots\ldots \\
&\hat{\theta}_k = G_k(a_1, a_2, \ldots\ldots a_k)
\end{aligned}$$

Los **estimadores** obtenidos por este método verifican las siguientes propiedades:

- Son **consistentes**.
- No son, en general, ni **insesgados** ni de **mínima varianza**.
- No tienen en cuenta la distribución de la la variable en la población.
- Su **eficiencia** no da resultados satisfactorios, por lo que se recurre al **método de la máxima verosimilitud**.
- Sus distribuciones son **asintóticamente normales**.

2.- MÉTODO DE LA MÁXIMA VEROSIMILITUD: Desarrollado por R. Fisher se trata de una técnica muy poderosa y de amplio uso.

Sea **X** una v.a. con función de probabilidad o de densidad dependiente de los parámetros $\theta = (\theta_1, \theta_2, ..., \theta_k)$, y consideraremos una m.a.s. de tamaño **n**, $(X_1, X_2, ..., X_n)$.

Función de Verosimilitud

$(X_1, X_2, ..., X_n)$ discreta

$$L(X_1, X_2, ..., X_n; \theta) = \prod_{i=1}^{n} P(X_i; \theta)$$

$(X_1, X_2, ..., X_n)$ continua

$$L(X_1, X_2, ..., X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

Los **estimadores de máxima verosimilitud** de θ serán aquellos valores que maximizan la **función de verosimilitud**. Muchas veces resulta más sencillo obtener los valores que maximizan $\Phi(\theta) = \log L(X;\theta)$.

$$\frac{\partial \Phi(X;\theta)}{\partial \theta_1} = 0 \quad \frac{\partial \Phi(X;\theta)}{\partial \theta_2} = 0 \quad ..... \quad \frac{\partial \Phi(X;\theta)}{\partial \theta_k} = 0$$

Se resuelve el sistema y se obtienen los **estimadores:**

$$\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n$$

Las propiedades que verifican los **estimadores de máxima verosimilitud** son:

- Si existe un estimador **suficiente** para el parámetro, cualquier solución del sistema es función del mismo.
- Si existe un estimador **eficiente** que haga accesible la cota de Cramer-Rao, se obtendrá por este método como solución única.
- Son **asintóticamente insesgados**.
- Se distribuyen **asintóticamente normales**.
- En condiciones muy generales, son **consistentes**.

Ejemplo: Sea ***X ~ N*** (μ,σ)*, con media y varianza desconocidas, obtener los estimadores de máxima verosimilitud de ambos parámetros.*

3.- MÉTODO DE LOS MÍNIMOS CUADRADOS: Dado un modelo matemático referente a una población en estudio: $Y^* = \Psi(X;\theta_1,\theta_2,....,\theta_k)$

Se pretenden estimar los valores de los θ_j , **j = 1, ..., k**, utilizando una m.a.s de **n** pares (X_i,Y_i), minimizando la expresión:

$$D = \sum_{i=1}^{n}\left[Y_i - \Psi(X_i;\theta_1,\theta_2,....,\theta_k)\right]^2$$

Derivando respecto de cada parámetro e igualando a 0, resulta un sistema, del que se obtienen los **estimadores por mínimos cuadrados**

En el caso de que las variables aleatorias **(Y_1, Y_2, ..., Y_n)** sean normales, este método es un caso particular del **método de la máxima verosimilitud**, por lo que las propiedades de los **estimadores mínimo cuadráticos** coinciden con las obtenidas para el otro método.

ESTIMADORES PUNTUALES MÁS USADOS

(1) Para la **media poblacional μ**: $\hat{\mu} = \overline{X}$

$$E(\overline{X}) = \mu \quad Var(\overline{X}) = \frac{\sigma^2}{n}$$

(2) Para la **varianza poblacional σ^2:** $\hat{\sigma}^2 = \hat{S}^2$

$$E(\hat{S}^2) = \sigma^2 \quad Var(\hat{S}^2) = \frac{2\,\sigma^4}{n-1}$$

(3) Para la **proporción o probabilidad de éxito:** $\hat{p} = \sum_{i=1}^{n} \frac{X_i}{n}$

$$E(\hat{p}) = p \quad Var(\hat{p}) = \frac{p\,(1-p)}{n}$$

(4) Parámetro λ de una distribución de Poisson: $\hat{\lambda} = \sum_{i=1}^{n} \frac{X_i}{n}$

$$E(\hat{\lambda}) = \lambda \quad Var(\hat{\lambda}) = \frac{\lambda}{n}$$

Propiedades:

- Insesgado
- Eficiente

Estimación por Intervalos de Confianza

La **Estimación Puntual** tiene la ventaja de que permite obtener una estimación unívoca del parámetro, pero tiene el gran inconveniente de que dependerá de la muestra considerada, por lo que la fiabilidad de la misma será desconocida.

Por ello, es mejor realizar la **estimación** a través de **intervalos de confianza**, ya que permite dar una idea acerca del valor real del parámetro y, además, establecer el grado de fiabilidad o **nivel de confianza** que, en base a la muestra empleada, se puede tener de que el verdadero valor del mismo se encuentre dentro del intervalo considerado.

Por tanto, la **Estimación por Intervalos** tiene por objetivo estimar el parámetro a través de un intervalo de la recta real para el cual exista una alta y conocida probabilidad de que el valor del mismo se encuentre entre sus límites. Se considera que cualquier valor del intervalo puede tomarse como **estimación** del parámetro.

Sea **X** una variable aleatoria cuya distribución depende de un parámetro desconocido θ. Para estimarlo, se selecciona una muestra aleatoria de tamaño **n**.

Diremos que **$[U_1,U_2]$** constituye un **intervalo de confianza** para el parámetro θ, con un **nivel de confianza** del **(1- α) %**, si existe una probabilidad **1-** α de que el verdadero valor de θ se encuentre entre **U_1** y **U_2**. Es decir:

$$P(U_1 \leq \theta \leq U_2) = 1 - \alpha$$

Intervalo de confianza **$[U_1,U_2]$**

- **Amplitud del intervalo:** Es la diferencia entre los límites superior e inferior.
- **Error de estimación:** Es la diferencia entre el estimador muestral y el parámetro poblacional.

Cuanto menor sea el **error de estimación** mayor será la precisión obtenida en la **estimación por intervalos de confianza**.

A partir de los datos muestrales que se poseen y para un **nivel de confianza** establecido, se puede obtener el **tamaño muestral** necesario para conseguir un máximo **error de estimación**.

Ejemplo 1: *Sea* **X ~ N (μ,σ)**, *con media desconocida y varianza conocida. Para estimar* μ *se toma una muestra aleatoria simple de tamaño* **n**.

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \text{Estimador de máxima verosimilitud de } \mu$$

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = Z \sim N(0,1)$$

$$P\left(-Z_{1-\alpha/2} \le \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le Z_{1-\alpha/2}\right) = P\left(\frac{|\overline{X}-\mu|}{\frac{\sigma}{\sqrt{n}}} \le Z_{1-\alpha/2}\right) = 1-\alpha \Rightarrow P\left(|\overline{X}-\mu| \le Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(-Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \overline{X}-\mu \le +Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \Rightarrow P\left(\overline{X}-Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}+Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Intervalo de confianza para μ con un **nivel de confianza** del **(1- α) %.**

Amplitud: $A = 2 * Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$

Error de estimación: $E = \pm Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow n = \sigma^2\left(\frac{Z_{1-\alpha/2}}{E}\right)^2$

Con esta expresión se puede calcular el tamaño muestral **n** necesario para obtener un error **E** y para un nivel de confianza del **(1- α) %.**

De la expresión del ***error de estimación*** *se deduce que al aumentar el tamaño muestral* ***n****, disminuye el* ***error****, y por tanto, aumenta la precisión del intervalo de confianza.*

Ejemplo 2: Sea **X ~ N (μ,σ)**, *con varianza desconocida. Obtener un intervalo de confianza para dicho parámetro.*

$$\hat{\sigma}^2 = \hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad \frac{(n-1)\cdot \hat{S}^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$P\left(\chi^2_{n-1;\frac{\alpha}{2}} \le \frac{(n-1)\hat{S}^2}{\sigma^2} \le \chi^2_{n-1;1-\frac{\alpha}{2}}\right) = 1-\alpha \rightarrow P\left(\frac{\chi^2_{n-1;\frac{\alpha}{2}}}{(n-1)\hat{S}^2} \le \frac{1}{\sigma^2} \le \frac{\chi^2_{n-1;1-\frac{\alpha}{2}}}{(n-1)\hat{S}^2}\right) = 1-\alpha \rightarrow$$

$$\rightarrow P\left(\frac{(n-1)\hat{S}^2}{\chi^2_{n-1;1-\frac{\alpha}{2}}} \le \sigma^2 \le \frac{(n-1)\hat{S}^2}{\chi^2_{n-1;\frac{\alpha}{2}}}\right) = 1-\alpha$$

Ejemplo 3: *La duración de las bombillas de bajo consumo producidas por una fabrica sigue una distribución N(μ, σ) de desviación típica igual a 180 horas. El departamento de producción desea estimar la duración media poblacional de las bombillas fabricadas, para lo que examina 100 unidades que se mantienen encendidas hasta que se funden, resultando una duración media muestral de 1600 horas.*

(a) *Construir un intervalo de confianza del 95 % para μ.*

(b) *Obtener el tamaño muestral necesario para un error máximo de 25 horas.*

Estadística Empresarial II

Tema 7

Contrastes de hipótesis

Introducción

Cuando se extrae una muestra aleatoria de una población, la información obtenida de ésta, puede usarse para realizar inferencias sobre las características de una determinada población.

- Una posibilidad consiste en estimar los parámetros desconocidos de la población mediante el cómputo de **estimadores puntuales** o **intervalos de confianza**.

- Alternativamente, la información muestral puede emplearse para verificar la validez de una conjetura o **hipótesis**, que el investigador realiza sobre alguna característica desconocida de la población.

Así pues, los **contrastes de hipótesis** permitirán confirmar la veracidad o falsedad de cualquier afirmación realizada sobre alguna característica desconocida de la población, en base a la información obtenida a partir de la muestra aleatoria.

Conceptos Básicos

Una **hipótesis estadística** es cualquier afirmación, verdadera o falsa, sobre alguna característica desconocida de una población. Si la hipótesis se refiere al valor de un parámetro desconocido θ de la población (cuya distribución se supone conocida), diremos que se trata de un **contraste paramétrico**. Sin embargo, si la hipótesis se refiere a la forma de la distribución poblacional, se hablará de **contraste no paramétrico**.

En este capítulo se estudiarán los contrastes paramétricos, dejando los no paramétricos para otro posterior.

Partamos de una población caracterizada por cierta variable **X**, cuya distribución es conocida y dada por $\mathbf{F(x;\theta)}$, pero dependiente de un parámetro desconocido θ.

Espacio paramétrico: Se trata del conjunto de valores que puede tomar el parámetro θ y se suele denotar por Ω.

Las **hipótesis** formuladas en los **contrastes paramétricos** pueden ser de dos tipos:

Simples: Si la hipótesis se refiere a un único valor del parámetro θ. En este caso, quedaría totalmente especificada la distribución poblacional **F(x;θ)**.

Compuestas: Si la hipótesis se refiere a un rango de valores que constituyen una región del espacio paramétrico Ω.

Ejemplo: Dada una población N(μ,σ) con μ desconocida, indicar si las siguientes hipótesis planteadas son simples o compuestas:

(a) $\mu = 5$ (b) $\mu > 5$ (c) $\mu \leq 5$ (d) $\mu \neq 5$

Cuando se realiza un **contraste de hipótesis**, se enfrentan dos hipótesis:

Hipótesis nula H_0: Es la hipótesis (simple o compuesta) que se pretende contrastar, luego será la que se rechace o acepte a la conclusión del test.

***Hipótesis alternativa H_1 (ó H_a)*:** Es la hipótesis (simple o compuesta) que se contrapone a la **hipótesis nula** y frente a la que se compara.

Ejemplos: *La especificación apropiada de las hipótesis nula y alternativa depende de la naturaleza del problema en cuestión. Algunas formas básicas son:*

$H_0: \theta = \theta_0$
$H_1: \theta = \theta_1$

$H_0: \theta = \theta_0$
$H_1: \theta \neq \theta_0$

$H_0: \theta \leq \theta_0$
$H_1: \theta > \theta_0$

$H_0: \theta \geq \theta_0$
$H_1: \theta < \theta_0$

Nota: *Si* H_1 *plantea que* θ *pertenece a un intervalo, el contraste será* ***unilateral o de una cola****, mientras que, si plantea que pertenece a dos intervalos, se denominará* ***bilateral o de dos colas.***

Después de especificar las **hipótesis nula** y **alternativa**, y de recoger información muestral, debe tomarse una decisión sobre H_0 (rechazarla o no). Para ello, se determina un estimador $\hat{\theta}$ del parámetro θ (sobre el que se establece H_0) cuya distribución sea conocida, y a partir del valor que tome la estimación para la muestra considerada, se decidirá si rechazar o no la **hipótesis nula**. Así pues, el **espacio muestral** (conjunto de todas las muestras posibles) quedará dividido en dos partes complementarias:

Región crítica o de rechazo R: Es la región del espacio muestral en la que se rechaza la **hipótesis nula** H_0. Puede estar formada

Región de aceptación R*: Es la región del espacio muestral en la que no se rechaza (se acepta) H_0.

Si denotamos a la muestra específica por **M(X) = (x_1, x_2, ..., x_n)**, entonces podemos concluir que:

Si $M(X) \in R$ ⇒ Se rechaza H_0
Si $M(X) \in R^*$ ⇒ Se acepta H_0

En el **contraste de hipótesis** tenemos que decidir si aceptamos la **hipótesis nula** o si la rechazamos, por lo que siempre correremos el riesgo de equivocarnos. La aparición de los errores tiene un carácter aleatorio, lo que exigirá su medición en términos probabilísticos.

Tipos de errores en un contraste de hipótesis		**H_0**	
		H_0 cierta	**H_0 falsa**
DECISIÓN	**Rechazar H_0**	**Error de tipo I**	Decisión correcta
	Aceptar H_0	Decisión correcta	**Error de tipo II**

La probabilidad de cometer **error de tipo I** se denomina **nivel de significación** y se denota por α, mientras que la probabilidad de cometer el **error de tipo II** se denota por β.

$$\alpha = P(E\ I) = P(\text{Rechazar } H_0 / H_0 \text{ cierta}) = P(M(X) \in R / H_0 \text{ cierta})$$
$$\beta = P(E\ II) = P(\text{Aceptar } H_0 / H_0 \text{ falsa}) = P(M(X) \in R^* / H_0 \text{ falsa})$$

Las repercusiones de estos errores pueden ser de diferente índole. Se suele admitir la mayor gravedad del **error de tipo I**, ya que rechazar una hipótesis cierta reviste, al menos en principio, mayor trascendencia que aceptar una falsa.

Así pues, el **contraste de hipótesis** que se elegirá será aquel que presente un **nivel de significación** α aceptable y que sea obtenido minimizando β, o lo que es lo mismo, maximizando **1-**β, denominada **potencia del contraste** (poder que tiene el contraste para reconocer que H_0 es falsa y debe ser, por tanto, rechazada).

$$\text{Potencia} = 1-\beta = P\,(\text{Rechazar } H_0 \,/\, H_0 \text{ falsa}) = P\,(M(X) \in R \,/\, H_0 \text{ falsa})$$

A la **región crítica** del contraste en el que para un α dado se maximiza la potencia **1-**β, se le denomina **región crítica prepotente**.

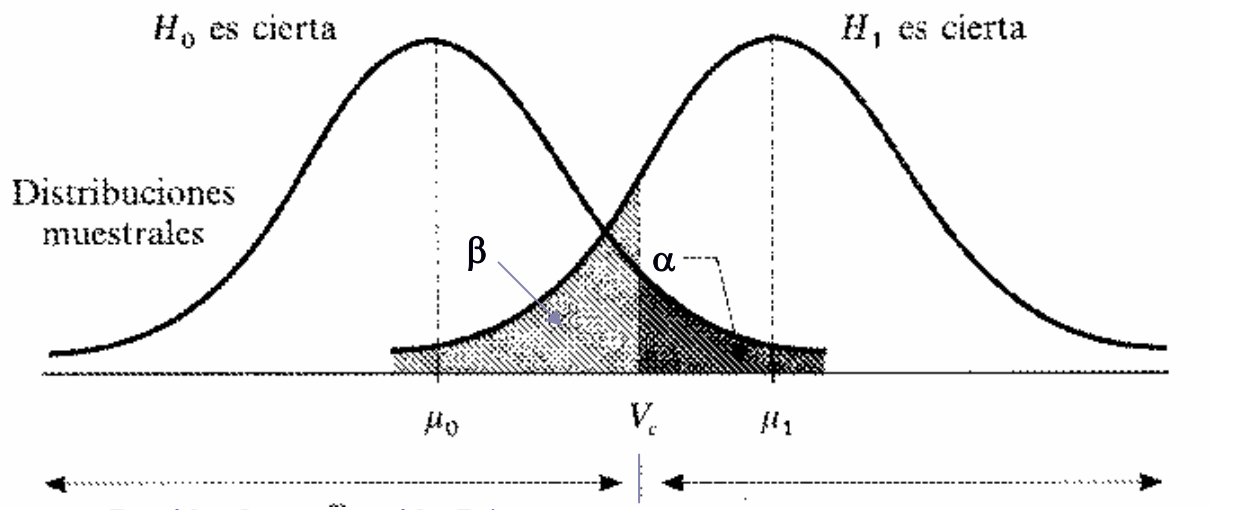
Nota 1: Cuando la hipótesis alternativa es compuesta, β*, y por tanto, la potencia* **1-**β *dependerá del valor específico del parámetro* θ*. En este caso, se define la* ***función de potencia*** $\mathbf{Q(\theta) = 1 - b(\theta)}$.

Nota 2: En lugar de decir que "se acepta H_0*" sería más correcto formular que "no existe evidencia suficiente para rechazar* H_0 *con los datos disponibles".*

Ejemplo: *Para una determinada población, se desea estudiar una característica X ~ N(μ,σ), con varianza conocida.Se pretenden contrastar dos posibles valores de μ.*

$H_0 : \mu = \mu_0$
$H_1 : \mu = \mu_1$

A partir de una muestra aleatoria (X_1, X_2, ..., X_n) de la población obtendríamos el estadístico de contraste, que en este caso será la media muestral $\overline{X}$. Suponiendo que la región crítica viene dada por todos los puntos situados a la derecha del **punto crítico V_c**, a continuación se muestran las áreas representativas de α y β.

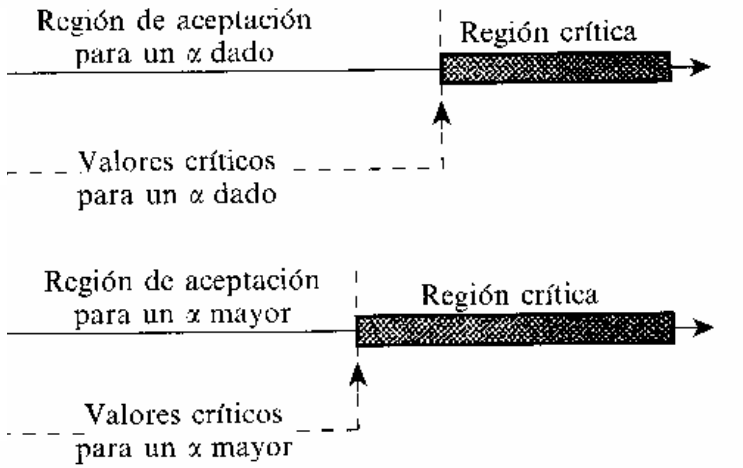


Desplazando el valor crítico V_c a la derecha, aumenta β y disminuye α; y desplazándolo a la izquierda, disminuye β y aumenta α, pero estas variaciones no se producen en la misma cuantía.

Del gráfico anterior se deduce que la **potencia** depende de varios factores:

- Del nivel de significación α: Al aumentar α, disminuye β, por lo que aumenta la potencia **1-** β, siempre que mantengamos inalterados los demás factores. Inversamente, si α disminuye, β aumenta.
- Del tamaño muestral **n**: Al aumentar **n**, α y β tienden a disminuir, ya que se reduce el error del muestreo. En este caso, aumentará la potencia.
- De la dispersión poblacional: Una menor desviación típica permite obtener un menor valor de β, incrementándose así la potencia.

El aumento o la disminución de α tiene una incidencia directa sobre la región crítica, ya que un mayor valor de α lleva consigo una mayor región crítica.



Contraste de hipótesis simples: Teorema de Neyman-Pearson

En un **contraste de hipótesis simples**, la distribución de la variable queda totalmente especificada, pudiéndose obtener la **potencia** del contraste de manera exacta. Para este tipo de situaciones existe un teorema que permite encontrar la región crítica de máxima potencia de entre todos los contrastes con un nivel de significación α dado.

TEOREMA DE NEYMAN-PEARSON:

Sea $(X_1, X_2, ..., X_n)$ una m.a.s. obtenida de una población con función de densidad $f(x,\theta)$, $\theta \in \Omega = \{\theta_0, \theta_1\}$, siendo la función de verosimilitud de la muestra: $L(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$

y si queremos contrastar las hipótesis simples: $\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta = \theta_1 \end{cases}$

siendo **k** un número positivo prefijado y $R \subset \Re^n$ tal que:

$$(1)\ \frac{L(x;\theta_0)}{L(x;\theta_1)} \le k,\ si\ (x_1,...,x_n) \in R \quad (2)\ \frac{L(x;\theta_0)}{L(x;\theta_1)} > k,\ si\ (x_1,...,x_n) \notin R \quad (3)\ P((X_1,...,X_n) \in R / \theta = \theta_0) = \alpha$$

Entonces, podemos asegurar que **R** es la **región crítica prepotente** (es decir, para un nivel de significación α dado, la de máxima potencia) para el contraste de hipótesis simples planteado.

<u>Ejemplo:</u> Sea (X_1, X_2, ..., X_n) una m.a.s. de tamaño n procedente de una población N(μ,σ), con μ desconocida y σ conocida. Obtener la región crítica prepotente con un nivel de significación α, para el contraste:

$$\begin{cases} \mathbf{H_0: \mu = \mu_0} \\ \mathbf{H_1: \mu = \mu_1} \end{cases}$$

Según el Teorema de Neyman-Pearson, la región crítica prepotente **R** quedará definida en base al cumplimiento de que el cociente de funciones de verosimilitud sea igual o inferior a una constante, cuando la muestra específica $(x_1, ..., x_n) \in \mathbf{R}$

$$\frac{L_0}{L_1} = \frac{L(x;\mu_0)}{L(x;\mu_1)} = \frac{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2}}{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_1)^2}} = \frac{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2}}{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_1)^2}} = e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\mu_0)^2-\sum_{i=1}^{n}(x_i-\mu_1)^2\right]} \le k$$

Tomando logaritmos neperianos:

$$\ln\frac{L_0}{L_1} = \ln e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\mu_0)^2-\sum_{i=1}^{n}(x_i-\mu_1)^2\right]} = -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\mu_0)^2-\sum_{i=1}^{n}(x_i-\mu_1)^2\right] \le \ln k = k_1$$

Operando:

$$\left[\sum_{i=1}^{n}(x_i-\mu_0)^2-\sum_{i=1}^{n}(x_i-\mu_1)^2\right]=\left(\sum_{i=1}^{n}x_i^2-2\mu_0\sum_{i=1}^{n}x_i+n\mu_0^2\right)-\left(\sum_{i=1}^{n}x_i^2-2\mu_1\sum_{i=1}^{n}x_i+n\mu_1^2\right)=$$

$$=n\left(\mu_0^2-\mu_1^2\right)-2\left(\mu_0-\mu_1\right)n\overline{X}$$

Así: $\ln\frac{L_0}{L_1}=-\frac{1}{2\sigma^2}\left[n\left(\mu_0^2-\mu_1^2\right)-2\left(\mu_0-\mu_1\right)n\overline{X}\right]\le k_1\Rightarrow-\frac{n}{2\sigma^2}\left(\mu_0^2-\mu_1^2\right)+\frac{\left(\mu_0-\mu_1\right)}{\sigma^2}n\overline{X}\le k_1\Rightarrow$

$$\Rightarrow\frac{\left(\mu_0-\mu_1\right)}{\sigma^2}n\overline{X}\le k_1+\frac{n}{2\sigma^2}\left(\mu_0^2-\mu_1^2\right)\Rightarrow\left(\mu_0-\mu_1\right)\overline{X}\le\frac{\sigma^2}{n}k_1+\frac{1}{2}\left(\mu_0^2-\mu_1^2\right)=k_2\Rightarrow$$

$$\Rightarrow\begin{cases}\overline{X}\le\dfrac{k_2}{\mu_0-\mu_1}=k' & \text{si }\mu_0>\mu_1\Rightarrow R=\left\{(x_1,x_2,\ldots,x_n)\in\Re^n/\ \overline{X}\le k'\right\}\\ \overline{X}\ge\dfrac{k_2}{\mu_0-\mu_1}=k' & \text{si }\mu_0<\mu_1\Rightarrow R=\left\{(x_1,x_2,\ldots,x_n)\in\Re^n/\ \overline{X}\ge k'\right\}\end{cases}$$

Para calcular el valor de **k'** habrá que imponer que el nivel de significación es **α**, luego, suponiendo que $\mu_0<\mu_1$:

$$\alpha=P\left(\overline{X}\ge k'/\mu=\mu_0\right)=P\left(Z\ge\frac{k'-\mu_0}{\sigma/\sqrt{n}}\right)\Rightarrow P\left(Z\le\frac{k'-\mu_0}{\sigma/\sqrt{n}}\right)=1-\alpha\Rightarrow\frac{k'-\mu_0}{\sigma/\sqrt{n}}=Z_{1-\alpha}\Rightarrow$$

$$\Rightarrow k'=\mu_0+Z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

Ahora, calcularemos la potencia del contraste:

$$\beta = P\left(\overline{X} \le k'/\mu = \mu_1\right) = P\left(Z \le \frac{k'-\mu_1}{\sigma/\sqrt{n}}\right) \Rightarrow 1-\beta = 1-P\left(Z \le \frac{k'-\mu_1}{\sigma/\sqrt{n}}\right)$$

<u>Ejercicio:</u> Dada una característica X de una población que sigue una distribución normal de desviación típica σ = 40, se extrae una muestra aleatoria simple de tamaño n = 100.

(a) Para un nivel de significación α = 0'01 obtener la región crítica asociada al contraste de máxima potencia.

$$\begin{cases} H_0 : \mu = 60 \\ H_1 : \mu = 80 \end{cases}$$

(b) Calcular la potencia del contraste.

(c) Determinar el tamaño de la muestra necesario para determinar la región crítica de manera que α sea 0'05 y la potencia sea 0'95.

Extensión del Teorema de Neyman-Pearson para hipótesis compuestas

Vamos a considerar que contrastamos una hipótesis nula simple frente a una hipótesis alternativa compuesta. El Teorema de Neyman-Pearson podrá ser aplicado a cada valor del parámetro θ incluido en H_1, obteniéndose así la la mejor región crítica para cada uno de ellos.

Puesto que el número de valores que toma θ (en H_1) suele ser infinito, será preciso encontrar un método que permita determinar si existe una única región crítica para los diferentes valores, que pasaremos a denominar **región crítica uniformemente más potente.** Así, un contraste será **uniformemente más potente** (UMP) si la región crítica es independiente del valor de θ.

El método de contrastación que se propondrá puede ser empleado para obtener la región crítica para el contraste de H_0 frente a H_1, incluso en el caso en que ambas hipótesis sean compuestas. El método utilizado se conoce como el **contraste de razón de verosimilitud**.

Partiremos de una muestra aleatoria $(X_1, X_2, ..., X_n)$ de una población caracterizada por la variable **X**, cuya distribución depende de **k** parámetros desconocidos $\theta = (\theta_1, \theta_2, ..., \theta_k)$, $\theta \in \Omega \subseteq \Re^k$.

Se pretende contrastar: $\begin{cases} H_0: \theta \in \Omega_0 \subset \Omega \\ H_1: \theta \in \Omega_1 = \Omega - \Omega_0 \end{cases}$ siendo $\begin{array}{l} \Omega_0 \cup \Omega_1 = \Omega \\ \Omega_0 \cap \Omega_1 = \varnothing \end{array}$

y pudiendo ser ambas hipótesis tanto simples como compuestas.

A partir de la muestra aleatoria extraída, se construye la función de verosimilitud:

$$L(x;\theta) = L(x_1, x_2, ..., x_n;\theta) = \begin{cases} \prod_{i=1}^{n} P(x_i;\theta) & \text{si X es discreta} \\ \prod_{i=1}^{n} f(x_i;\theta) & \text{si X es continua} \end{cases}$$

Puesto que las hipótesis pueden ser compuestas, se calculan los siguientes máximos de la función de verosimilitud:

$$L(x;\theta_{\Omega_0}) = \max_{\theta \in \Omega_0} L(x;\theta) \qquad L(x;\theta_{\Omega}) = \max_{\theta \in \Omega} L(x;\theta)$$

El estadístico a usar será:

$$\lambda(x) = \frac{L(x;\theta_{\Omega_0})}{L(x;\theta_{\Omega})}$$

La **región crítica** vendrá dada por: $R = \{x = (x_1, x_2, ..., x_n) / \lambda(x) \leq k\}$ y el valor de **k** se obtendrá en base a que:

$$\alpha = P(M(X) = (x_1, ..., x_n) \in R / H_0 \text{ cierta}) = P(\lambda(x) \leq k / \theta \in \Omega_0)$$

El valor de $\lambda(\mathbf{x})$ depende de la muestra seleccionada y verifica las siguientes propiedades:

- $\lambda(\mathbf{x}) \geq \mathbf{0}$, ya que $\mathbf{L}(\mathbf{x};\theta) \geq \mathbf{0}$ al ser producto de funciones de densidad o de probabilidad, que toman siembre valores no negativos.
- $\lambda(\mathbf{x}) \leq \mathbf{1}$, puesto que $\mathbf{L}(\mathbf{x};\theta_{\Omega o}) \leq \mathbf{L}(\mathbf{x};\theta_{\Omega})$ al estar $\Omega_0 \subset \Omega$.

Así, si $\mathbf{H_0}$ es cierta, $\theta_{\Omega o}$ tenderá a estar próximo a θ_{Ω}, por lo que $\lambda(\mathbf{x})$ se aproximará a 1. Sin embargo, si $\mathbf{H_0}$ es falsa, $\theta_{\Omega o}$ tenderá a estar muy alejado de θ_{Ω}, por lo que $\lambda(\mathbf{x})$ se aproximará a 0.

El problema fundamental se centrará en obtener la distribución de $\lambda(\mathbf{x})$ cuando $\mathbf{H_0}$ es cierta, lo que permitirá fijar la **región crítica** del contraste.

Ejemplo: Sea X una v.a. cuya distribución se a una N(μ,σ) desconocidos. Se toma una muestra aleatoria de tamaño n, a partir de la cual se pretende contrastar, con un nivel de significación α:

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases} \qquad \Omega = \{ (\mu,\sigma) \in \Re^2 / \sigma > 0 \}$$

$$\Omega_0 = \{ (\mu,\sigma) \in \Re^2 / \mu = \mu_0 , \sigma > 0 \}$$

La función de verosimilitud viene dada en función de $\theta = (\mu,\sigma)$

$$L(x_1,...,x_n;\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

y habrá que determinar: $L(x;\theta_{\Omega_0}) = \underset{\theta \in \Omega_0}{máx}\, L(x;\theta)$ $\qquad L(x;\theta_{\Omega}) = \underset{\theta \in \Omega}{máx}\, L(x;\theta)$

Los estimadores de máxima verosimilitud obtenidos son:

$$\mu_{\Omega_0} = \mu_0 \qquad \sigma^2_{\Omega_0} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2$$

$$\mu_{\Omega} = \overline{X} \qquad \sigma^2_{\Omega} = S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

El estadístico razón de verosimilitudes será:

$$\lambda(x) = \frac{L(x;\theta_{\Omega_0})}{L(x;\theta_{\Omega})} = \frac{L(x;\mu_{\Omega_0},\sigma_{\Omega_0})}{L(x;\mu_{\Omega},\sigma_{\Omega})}$$

$$\lambda(x)=\frac{L(x;\mu_{\Omega_0},\sigma_{\Omega_0})}{L(x;\mu_{\Omega},\sigma_{\Omega})}=\frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n\left(\frac{1}{n}\sum_{i=1}^n(x_i-\mu_0)^2\right)^{-\frac{n}{2}}e^{-\frac{2}{n}\frac{\sum_{i=1}^n(x_i-\mu_0)^2}{\sum_{i=1}^n(x_i-\mu_0)^2}}}{\left(\frac{1}{\sqrt{2\pi}}\right)^n\left(\frac{1}{n}\sum_{i=1}^n(x_i-\bar{X})^2\right)^{-\frac{n}{2}}e^{-\frac{2}{n}\frac{\sum_{i=1}^n(x_i-\bar{X})^2}{\sum_{i=1}^n(x_i-\bar{X})^2}}}=\frac{\left(\frac{1}{n}\sum_{i=1}^n(x_i-\mu_0)^2\right)^{-\frac{n}{2}}}{\left(\frac{1}{n}\sum_{i=1}^n(x_i-\bar{X})^2\right)^{-\frac{n}{2}}}=\left(\frac{\sum_{i=1}^n(x_i-\bar{X})^2}{\sum_{i=1}^n(x_i-\mu_0)^2}\right)^{\frac{n}{2}}$$

Operando con el denominador:

$$\sum_{i=1}^n(x_i-\mu_0)^2=\sum_{i=1}^n\left[(x_i-\bar{X})+(\bar{X}-\mu_0)\right]^2=\sum_{i=1}^n\left[(x_i-\bar{X})^2+(\bar{X}-\mu_0)^2+2(x_i-\bar{X})(\bar{X}-\mu_0)\right]=$$

$$=\sum_{i=1}^n(x_i-\bar{X})^2+\sum_{i=1}^n(\bar{X}-\mu_0)^2+2(\bar{X}-\mu_0)\sum_{i=1}^n(x_i-\bar{X})=\sum_{i=1}^n(x_i-\mu_0)^2+n(\bar{X}-\mu_0)^2+2(\bar{X}-\mu_0)(n\bar{X}-n\bar{X})=$$

$$=\sum_{i=1}^n(x_i-\bar{X})^2+n(\bar{X}-\mu_0)^2$$

Así:

$$\lambda(x)=\left(\frac{\sum_{i=1}^n(x_i-\bar{X})^2}{\sum_{i=1}^n(x_i-\bar{X})^2+n(\bar{X}-\mu_0)^2}\right)^{\frac{n}{2}}=\left(\frac{1}{1+\frac{n(\bar{X}-\mu_0)^2}{\sum_{i=1}^n(x_i-\bar{X})^2}}\right)^{\frac{n}{2}}=\left(\frac{1}{1+\frac{n(\bar{X}-\mu_0)^2}{\frac{n-1}{n-1}\sum_{i=1}^n(x_i-\bar{X})^2}}\right)^{\frac{n}{2}}=\left(\frac{1}{1+\frac{n(\bar{X}-\mu_0)^2}{(n-1)\hat{S}^2}}\right)^{\frac{n}{2}}=\left(\frac{1}{1+\frac{1}{n-1}\frac{(\bar{X}-\mu_0)^2}{\hat{S}^2/n}}\right)^{\frac{n}{2}}\leq k$$

$$\left(\frac{1}{1+\frac{1}{n-1}\frac{(\bar{X}-\mu_0)^2}{\hat{S}^2/n}}\right)^{\frac{n}{2}} \le k \Rightarrow \frac{1}{1+\frac{1}{n-1}\frac{(\bar{X}-\mu_0)^2}{\hat{S}^2/n}} \le k^{\frac{2}{n}} \Rightarrow 1+\frac{1}{n-1}\frac{(\bar{X}-\mu_0)^2}{\hat{S}^2/n} \ge k^{-\frac{2}{n}} \Rightarrow$$

$$\Rightarrow \frac{(\bar{X}-\mu_0)^2}{\hat{S}^2/n} \ge (n-1)\left(k^{-\frac{2}{n}}-1\right) = k' \Rightarrow R = \left\{x=(x_1,x_2,\ldots,x_n) / \frac{(\bar{X}-\mu_0)^2}{\hat{S}^2/n} \ge k'\right\}$$

*Ahora, hay que determinar **k'**:*

$$\alpha = P(M(X)=(x_1,\ldots,x_n)\in R / H_0 \text{ cierta}) = P\left(\frac{(\bar{X}-\mu_0)^2}{\hat{S}^2/n} \ge k' / \mu=\mu_0\right)$$

Sabemos que: $\dfrac{(\bar{X}-\mu_0)}{\frac{\hat{S}}{\sqrt{n}}} \sim Z$, si $n>30$ $\qquad$ $\dfrac{(\bar{X}-\mu_0)}{\frac{\hat{S}}{\sqrt{n}}} \sim t_{n-1}$, si $n\le 30$

Luego, por ejemplo, si n > 30: $\alpha = P\left(\frac{(\bar{X}-\mu_0)^2}{\hat{S}^2/n} \ge k'\right) = P(Z^2 \ge k') =$

$$= P\left(|Z| \ge \sqrt{k'}\right) \Rightarrow 1-\alpha = P\left(|Z| \le \sqrt{k'}\right) = P\left(-\sqrt{k'} \le Z \ge \sqrt{k'}\right) \Rightarrow \sqrt{k'} = Z_{1-\frac{\alpha}{2}}$$

<u>Ejemplo1:</u> De una muestra de 200 emprendedores cuya empresa fracasó, 100 no hicieron plan de viabilidad al iniciar su actividad. A un nuvel de significación del 1%, contrastar la hipótesis de una asociación de empresarios que dice que como máximo el 35% de las empresas que fracasan no hacen plan de viabilidad.

<u>Ejemplo 2:</u> Se pidió a una muestra de 16 consumidores de un barrio de Santa Cruz de Tenerife que valorasen en una escala de 1 (completamente en desacuerdo) a 5 (completamente de acuerdo), el interés de poner un supermercado de segunda generación en el barrio. La empresa que piensa ponerlo lo hará siempre que la puntuación sea como mínimo 3,75. Los resultados muestrales indicaban que la media muestral es de 3,68 y la cuasidesviación típica muestral 1,21. A un nivel de significación del 5%, ¿La empresa se decidirá a abrir el superpercado en el barrio?.

Medición de la potencia de un contraste de hipótesis compuestas

Cuando la **hipótesis alternativa** del contraste planteado es **compuesta**, el error β y la **potencia 1-β** dependerán del parámetro desconocido θ.

En este caso, para obtener la **potencia** del contraste, se siguen los siguientes pasos:

(1) Determinar el rango de valores del estimador $\hat{\theta}$ de θ que conducen a la aceptación o al rechazo de la hipótesis nula.

(2) Para un valor de interés θ_1 de θ, calcular la probabilidad de que $\hat{\theta}$ pertenezca al intervalo de aceptación determinado en (1) para muestras de n observaciones de una población con valor θ_1 para el parámetro.

(3) Como estamos ante hipótesis compuestas, obtendremos varios valores de β mediante los que determinaremos los correspondientes valores de la potencia del contraste **1-** β, que, en conjunto, nos darán la función de potencia **$Q(\theta) = 1 - \beta(\theta)$**.

Ejemplo: En un proceso de selección de tomates para exportación de un almacén de empaquetado, se sabe que el peso de los tomates sigue una distribución normal de media 50 gramos y desviación típica 10 gramos. Se ha cambiado de fincas suministradoras y el encargado del almacén sospecha que se ha incrementado el peso medio de los tomates sin modificarse la desviación típica. Para probar esta sospecha, se toma una muestra de 100 tomates, comprobándose que su peso medio es de 50,38 gramos. Realizar un contraste de hipótesis a un nivel de significación del 5% y determinar la potencia del mismo.

El contraste planteado será:

$$\begin{cases} H_0 : \mu \le 50 \\ H_1 : \mu > 50 \end{cases}$$

$$\text{Estadístico}: \quad Z = \frac{(\overline{X} - 50)}{\frac{10}{\sqrt{100}}}$$

$$\text{Punto crítico}: \quad Z_{1-\alpha} = Z_{1-0,05} = Z_{0,95} = 1,645$$

(1) *Determinar el rango de valores del estimador* $\hat{\mu} = \overline{X}$ *de* μ *que conducen a la aceptación o al rechazo de* H_0.

Se rechaza H_0 si $Z > Z_{1-\alpha} \Leftrightarrow \frac{(\overline{X} - 50)}{\frac{10}{\sqrt{100}}} \succ 1,645 \Leftrightarrow \overline{X} > 51'645$

Se acepta H_0 si $Z \le Z_{1-\alpha} \Leftrightarrow \frac{(\overline{X} - 50)}{\frac{10}{\sqrt{100}}} \le 1,645 \Leftrightarrow \overline{X} \le 51'645$

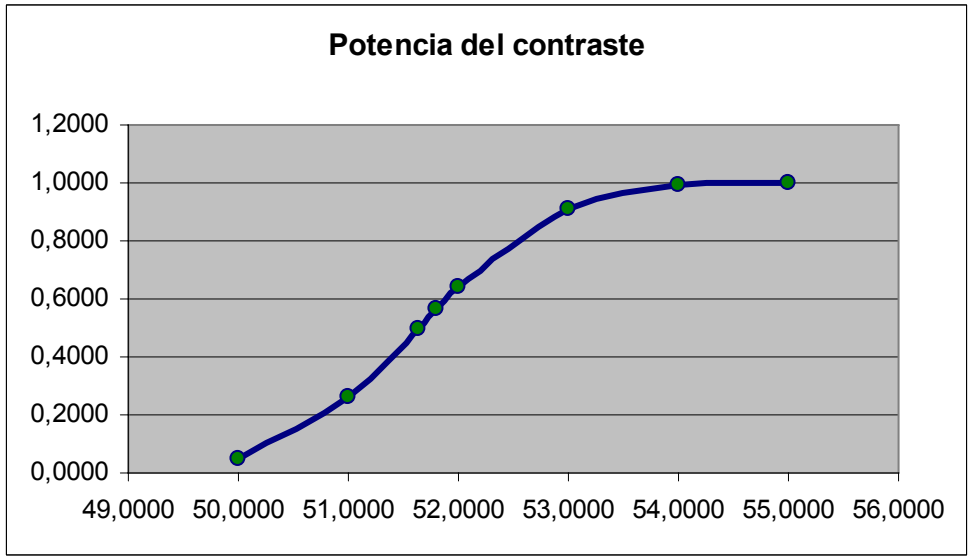
*(2) Para un valor de interés μ_1 = **51'8** calculamos la probabilidad de que pertenezca al intervalo de aceptación determinado en (2) para muestras de n observaciones de una población con media poblacional μ_1.*

$$\beta(\mu_1) = P(\text{aceptar } H_0 / \mu = \mu_1) = P(\overline{X} \le 51'645 / \mu_1 = 51'8) = P\left(Z \le \frac{51,645 - 51,8}{\frac{10}{\sqrt{100}}}\right) =$$

$$= P(Z \le -0,155) = 0,437 \Rightarrow 1 - \beta(51'8) = 1 - 0,4372 = 0,5628$$

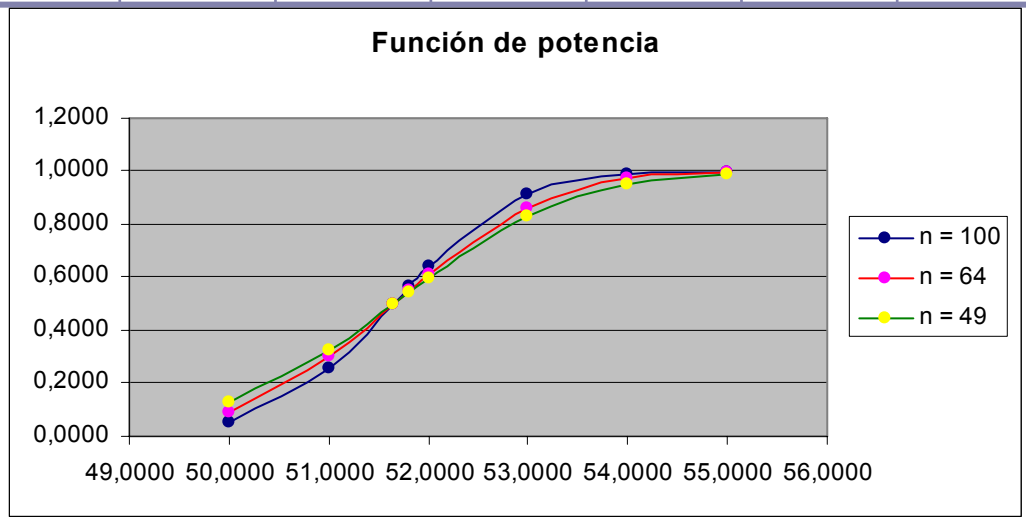
*(3) Dándole diversos valores a μ_1, obtendremos varios valores de β mediante los que determinaremos los correspondientes valores de la potencia del contraste **1- β**.*

μ_1	β	1- β
50	0'95	0'05
51	0'7406	0'2595
51'645	0'5	0'5
51'8	0'4372	0'5628
52	0'3613	0'6387
53	0'0877	0'9123
54	0'0091	0'9909
55	0	1



Nota: *Cuanto mayor sea el tamaño de la muestra, **n**, mayor será la potencia del contraste, **1-β**, ya que cuanta más información se tenga de la población, hay más posibilidad de detectar cualquier desviación de la hipótesis nula, como podemos observar en la siguiente tabla:*

n = 100			n = 64		n = 49	
μ_1	β	1- β	β	1- β	β	1- β
50	0’95	0’05	0’9057	0’0943	0’8749	0’1251
51	0’7406	0’2595	0’6985	0’3015	0’6736	0’3264
51’645	0’5	0’5	0’5	0’5	0’5	0’5
51’8	0’4372	0’5628	0’4502	0’5498	0’4562	0’5438
52	0’3613	0’6387	0’3878	0’6122	0’4013	0’5987
53	0’0877	0’9123	0’139	0’861	0’1711	0’8289
54	0’0091	0’9909	0’0298	0’9702	0’0495	0’9505
55	0	1	0’0036	0’9964	0’00904	0’99096



Estadística Empresarial II

Tema 8

Aplicaciones de los contrastes de hipótesis

Introducción

En el capítulo anterior se estudiaron los **contrastes paramétricos**, en los que se formulaban hipótesis relativas a los valores de algunos parámetros desconocidos de una distribución poblacional conocida. Sin embargo, en muchos casos ocurre que la distribución poblacional se desconoce, por los que las hipótesis se centrarán en el tipo de distribución de probabilidad que presenta.

Estos contrastes mencionados anteriormente, son conocidos como **pruebas de bondad de ajuste**, que, al ser la distribución poblacional desconocida, formarán parte de los **contrastes no paramétricos**. En los test de bondad de ajuste no es apropiado el uso del test de razón de verosimilitudes del caso paramétrico, al no quedar especificada la función de verosimilitud cuando la hipótesis alternativa sea cierta.

Además, se estudiará otro tipo de **contrastes no paramétricos**, como la **prueba de independencia**, en la que se pretenderá verificar si existe dependencia entre dos características de una determinada población.

Prueba Chi-cuadrado de Pearson de Bondad de Ajuste

Esta prueba se emplea para decidir si una determinada distribución de probabilidad se ajusta a un conjunto de datos. Para ello, se comparan los valores observados de una muestra aleatoria con los que se espera obtener si la hipótesis nula fuese correcta (valores esperados), en cada una de las categorías en que son clasificados los individuos.

TEOREMA: Sea $(X_1, X_2, ..., X_k)$ una v.a. multinomial de parámetros n, p_1, p_2, ..., p_k, entonces el estadístico siguiente:

$$U = \sum_{i=1}^{k} \frac{(X_i - n.p_i)^2}{n.p_i} = \sum_{i=1}^{k} \frac{X_i^2}{n.p_i} - n \underset{n\to\infty}{\sim} \chi^2_{k-1}$$

Frecuentemente, el estadístico **U** aparece expresado en función de los valores observados $\mathbf{O_i}$ y los valores esperados $\mathbf{E_i}$, obteniéndose:

$$U = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n$$

En esta **prueba de bondad de ajuste**, se va a contrastar la hipótesis nula de que un grupo de datos de una muestra procede de una v.a. con una distribución de probabilidad especificada.

Nota: Para cada una de las categorías, se debe verificar que $\mathbf{E_i > 5}$, *por lo que habría que reagrupar en caso de no cumplirse para alguna.*

CASO 1: ***Distribuciones de probabilidad con parámetros conocidos o especificados.***

Se quiere resolver el siguiente contraste: $\begin{cases} H_0 : F(x) = F_0(x) \\ H_1 : F(x) \neq F_0(x) \end{cases}$

Para ello, seleccionamos una muestra aleatoria de tamaño **n**, cuyas observaciones serán clasificadas en **k** categorías (o intervalos), denotando por $\mathbf{O_i}$ a la frecuencia absoluta observada de cada categoría. Sabiendo que $\mathbf{p_i}$ es la probabilidad de que la variable **X** (bajo H_0) tome valores en la categoría o intervalo i-ésimo, las frecuencias absolutas esperadas serán: $E_i = n \cdot p_i \ , i = 1, ..., k$

La región crítica vendrá dada por: $R = \left\{ (x_1, x_2, ..., x_n) \in E \ / \ U > \chi^2_{k-1,1-\alpha} \right\}$

Ejemplo: *El número de erratas por página de un libro suele considerase como una variable de Poisson. Se contaron errores en 100 páginas de una novela, obteniéndose:*

Número de erratas	Número de páginas
0	65
1	25
2	8
≥3	2

Contrastar si el número de erratas se distribuye según una Poisson P(0'4), para un nivel de significación del 10%.

Solución: *El contraste planteado vendrá dado por:* $\begin{cases} H_0 : F(x) = F_0(x) \\ H_1 : F(x) \neq F_0(x) \end{cases}$ *, siendo* **$F_0(x)$** *la función de distribución de* **X ~ P(0'4)**.

Al ser ***X*** *discreta, las categorías o intervalos se consideran centrados en los valores de la variable. A continuación, se muestran en la siguiente tabla:*

Nº de erratas	Categorías	Nº de páginas: O_i	p_i	$E_i = n \cdot p_i$
0	X < 0'5	65	0'6703	67'03
1	0'5 ≤ X <1'5	25	0'2681	26'81
2	1'5 ≤ X <2'5	8	0'0536	5'36
≥3	X ≥ 2'5	2	0'0080	0'80
Total		**100**	**1**	**100**

Al ser ***E_4 = 0'80 < 5****, habrá que reagrupar, uniendo las categorías 3 y 4 en una sola.*

$p_1 = P(X < 0'5) = P(X = 0)$ $p_2 = P(0'5 \leq X < 1'5) = P(X = 1)$ $p_3 = P(1'5 \leq X < 2'5) = P(X = 2)$

$p_4 = P(X \geq 2'5) = P(X = 3) + P(X = 4) + P(X = 5)$

Categorías	Nº de páginas: O_i	p_i	$E_i = n \cdot p_i$	$(O_i - E_i)^2 / E_i$
X < 0'5	65	0'6703	67'03	0'061
0'5 ≤ X <1'5	25	0'2681	26'81	0'122
X ≥ 1'5	10	0'0616	6'16	2'394
Total	**100**	**1**	**100**	**U = 2'577**

Estadístico de contraste:

$$U = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 2'577$$

Punto crítico:

$$\chi^2_{k-1,1-\alpha} = \chi^2_{2,0'9} = 4'60$$

Región crítica: $R = \left\{(x_1, x_2, ..., x_n) \in E \ / \ U > \chi^2_{2,0'9} = 4'60\right\}$

Por tanto, a la vista de los datos, se concluye que no tenemos evidencia suficiente para rechazar la hipótesis nula de que X ~ P(0'4).

CASO 2: ***Distribuciones de probabilidad con parámetros desconocidos o no especificados.***

Muchas veces, nos puede interesar probar si una variable se distribuye de una forma determinada, sin especificar los valores de los parámetros de los que depende.

En este caso, las frecuencias esperadas $\mathbf{E_i}$ no podrán ser determinadas al desconocer los parámetros de la distribución, por lo que habrá que estimarlos a partir de la información de la muestra aleatoria.

Así pues, los valores esperados vendrán dados por:

$E_i = n \cdot p_i(\hat{\theta})$, siendo $\hat{\theta}$ el estimador de máxima verosimilitud de θ.

La región crítica vendrá dada por: $R = \left\{(x_1, x_2, ..., x_n) \in E \ / \ U > \chi^2_{k-r-1,1-\alpha}\right\}$ siendo **r** el número de parámetros que han de ser estimados.

*Nota: En este caso, se están utilizando los valores observados **O_i** para estimar los parámetros a partir de los que se calcularán posteriormente los **E_i**. Esta manera de actuar distorsiona las diferencias reales que pueden existir entre las frecuencias, por lo que se reajusta la región crítica.*

Ejemplo: Con el fin de estudiar la distribución que siguen las estaturas de los alumnos de la Escuela de Empresariales de la ULL, se selecciona una muestra aleatoria de 100 alumnos, cuyas estaturas podemos agrupar en la siguiente tabla:

Estatura en metros	Número de alumnos
[1'50 , 1'60)	6
[1'60 – 1'70)	28
[1'70 –1'80)	40
[1'80 – 1'90)	22
[1'90 – 2'00)	4

Contrastar la hipótesis de que la muestra procede de una población normal, estudiando la bondad de ajuste para un nivel de significación del 5 %.

Solución: *Consideramos la variable* **X = "estatura de los alumnos de Empresariales de la ULL"**

El contraste planteado vendrá dado por $\begin{cases} H_0 : F(x) = F_0(x) \\ H_1 : F(x) \neq F_0(x) \end{cases}$

siendo $\mathbf{F_0(x)}$ *la función de distribución de la variable* $\mathbf{X \sim N(\mu,\sigma)}$. Al ser μ y σ^2 *desconocidos, habrá que estimarlos.*

$L_i - L_{i+1}$	n_i	X_i	$n_i \cdot X_i$	$n_i \cdot X_i^2$
1'50 - 1'60	6	1'55	9'3	14'415
1'60 – 1'70	28	1'65	46'2	76'23
1'70 –1'80	40	1'75	70	122'5
1'80 – 1'90	22	1'85	40'7	75'295
1'90 – 2'00	4	1'95	7'8	15'21
Total	**100**		**174**	**303'65**

Estimadores de máxima verosimilitud:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} X_i\, n_i = 1'74$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{k} (X_i - \overline{X})^2\, n_i = \frac{1}{n}\sum_{i=1}^{k} X_i^2\, n_i - \overline{X}^2 = 0'0089$$

$$s = \sqrt{0'0089} = 0'094$$

Categorías	O_i	p_i	$E_i = n \cdot p_i$
1'50 - 1'60	6	0'0627	6'27
1'60 – 1'70	28	0'2655	26'55
1'70 –1'80	40	0'4053	40'53
1'80 – 1'90	22	0'2165	21'65
1'90 – 2'00	4	0'0418	4'18
Total	**100**		

Por tanto, el contraste podría expresarse como:

$$\begin{cases} H_0 : X \sim N(1'74, 0'094) \\ H_1 : X \not\sim N(1'74, 0'094) \end{cases}$$

Al ser $\mathbf{E_5 = 4'18 < 5}$*, habrá que reagrupar, uniendo las categorías 4 y 5 en una sola.*

Las probabilidades p_i se obtienen bajo H_0 como sigue:

$$p_1 = P(1'50 \le X < 1'60) = P(-2'55 \le Z < -1'49) = 0'0627$$
$$p_2 = P(1'60 \le X < 1'70) = P(-1'49 \le Z < -0'43) = 0'2655$$
$$p_3 = P(1'70 \le X < 1'80) = P(-0'43 \le Z < 0'64) = 0'4053$$
$$p_4 = P(1'80 \le X < 1'90) = P(0'64 \le Z < 1'70) = 0'2165$$
$$p_5 = P(1'90 \le X < 2'00) = P(1'70 \le Z < 2'77) = 0'0418$$

Una vez agrupadas las categorías, se obtienen los siguientes resultados:

Categorías	O_i	p_i	$E_i = n \cdot p_i$	$(O_i - E_i)^2 / E_i$
1'50 - 1'60	6	0'0627	6'27	0'0116
1'60 – 1'70	28	0'2655	26'55	0'0792
1'70 –1'80	40	0'4053	40'53	0'0069
1'80 – 2'00	26	0'2583	25'83	0'0011
Total	**100**			**0'0988**

Estadístico de contraste:

$$U = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 0'0988$$

Punto crítico:

$$\chi^2_{k-r-1,1-\alpha} = \chi^2_{4-2-1,0'95} = \chi^2_{1,0'95} = 3'84$$

La región crítica será: $R = \{(x_1, x_2, \ldots, x_n) \in E \ / \ U > \chi^2_{1,0'95} = 3'84\}$

Por tanto, a la vista de los datos, se concluye que no tenemos evidencia suficiente para rechazar la hipótesis nula de que X ~ N (1'74, 0'094).

Prueba de Kolmogorov - Smirnov de bondad de ajuste

En la **prueba Chi-cuadrado** anterior, cuando la distribución propuesta es continua, hay que aproximar $F_0(x)$ con el agrupamiento de los datos en un número finito de categorías o intervalos, lo que precisa muestras relativamente grandes. Una prueba más adecuada cuando $F_0(x)$ es continua es la **prueba de Kolmogorov-Smirnov**, ya que no necesita que los datos estén agrupados y es aplicable a muestras pequeñas.

Esta prueba se basa en la comparación de la función de distribución (acumulativa) observada de los datos muestrales ordenados y de la función de distribución propuesta en H_0.

El contraste planteado será: $\begin{cases} H_0 : F(x) = F_0(x) \\ H_1 : F(x) \neq F_0(x) \end{cases}$

siendo $\mathbf{F_0(x)}$ una función de distribución con todos sus parámetros conocidos o especificados (sólo en esta situación es aplicable esta prueba).

Dada una muestra aleatoria $(X_1, X_2, ..., X_n)$, se ordenan los valores de menor a mayor, obteniendo $(X_{(1)}, X_{(2)}, ..., X_{(n)})$. A continuación, se construye la función de distribución para el caso de equiprobabilidad, que viene dada por:

$$S_n(X) = \begin{cases} 0 & \text{si } X < X_{(1)} \\ \dfrac{k}{n} & \text{si } X_{(k)} \le X < X_{(k+1)} \\ 1 & \text{si } X \ge X_{(n)} \end{cases}$$

Para comparar $F_0(x)$ y $S_n(x)$, se define el **estadístico de Kolmogorov-Smirnov**, que se define como:

$$D_n = \max_{-\infty < x < \infty} |S_n(x) - F_0(x)|$$

La distribución de D_n se evalúa sólo en función del tamaño muestral **n**, pudiendo usarse para cualquier distribución $F_0(x)$ especificada en H_0.

La región crítica vendrá dada por: $R = \{(x_1, x_2, ..., x_n) \in E \ / \ D_n > D_{n,1-\alpha}\}$

Ejemplo: Se ha elegido una muestra de 16 ejecutivos de empresas de Tenerife, a los que se les ha preguntado por su renta familiar mensual en miles de pesetas constantes del año 2000. Los resultados obtenidos, una vez ordenados, fueron: 852, 875, 910, 933, 957, 963, 981, 998, 1007, 1010, 1015, 1018, 1023, 1035, 1048 y 1063. En años anteriores, la renta familiar de los ejecutivos de Tenerife seguía una distribución N (985, 50). ¿Para un nivel de significación del 5 %, podemos decir que ha cambiado el modelo?

Solución: Sea la variable ***X = "Renta familiar mensual de los ejecutivos de Tenerife en miles de ptas"***, el contraste que habrá que resolver será:

$$\begin{cases} H_0 : F(x) = F_0(x) \\ H_1 : F(x) \neq F_0(x) \end{cases}$$

siendo $\mathbf{F_0(x)}$ la función de distribución de una N (985, 50).

$X_{(k)}$	n_k	$S_n(x) = k/16$	$F_0(x) = P(X \leq x)$	$\lvert S_n(x) - F_0(x) \rvert$
852	1	0'0625	0'0039	0'0586
875	1	0'125	0'0139	0'1111
910	1	0'1875	0'0668	0'1207
933	1	0'25	0'1492	0'1008
957	1	0'3125	0'2877	0'0248
963	1	0'375	0'33	0'045
981	1	0'4375	0'4681	0'0306
998	1	0'5	0'6026	0'1026
1007	1	0'5625	0'67	0'1075
1010	1	0'625	0'6915	0'0665
1015	1	0'6875	0'7257	0'0382
1018	1	0'75	0'7454	0'0046
1023	1	0'8125	0'7764	0'0361
1035	1	0'875	0'8413	0'0337
1048	1	0'9375	0'8962	0'0413
1063	1	1	0'9406	0'0594

$$F_0(x) = P(X \leq x) = P\left(Z \leq \frac{x - 985}{50}\right)$$

Región crítica:

$$R = \{(x_1, x_2, ..., x_n) \in E \ / \ D_n > D_{n,1-\alpha}\}$$

Estadístico de contraste:

$$D_n = \max |S_n(x) - F_0(x)| = 0'1207$$

Punto crítico:

$$D_{n,1-\alpha} = D_{16,0'95} = 0'328$$

Por tanto, se concluye que no tenemos evidencia suficiente para rechazar H_0. Luego, no ha cambiado el modelo.

Prueba Chi-cuadrado de independencia

Muchas veces, surge la necesidad de analizar si existe relación entre dos características (con un cierto nº de categorías o modalidades) mediante las que una población ha sido clasificada. Cuando una muestra aleatoria se obtiene de una población que se clasifica de esta manera, el resultado es una **tabla de contingencia**.

Partimos de una tabla de contingencia con dos características **A** y **B**, subdivididas en **h** y **k** categorías, respectivamente.

		B				**Marginal**
		B_1	**B_2**	**...**	**B_k**	
A	**A_1**	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
	A_2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
	:	:	:	...	:	:
	A_h	n_{h1}	n_{h2}	...	n_{hk}	$n_{h.}$
Marginal		$n_{.1}$	$n_{.2}$	...	$n_{.k}$	**n**

Sea $\mathbf{p_{ij}}$ la probabilidad de que una observación elegida al azar corresponda a la categoría **(i,j)**, y sean $\mathbf{p_{i.}}$ y $\mathbf{p_{.j}}$ las probabilidades marginales.

El contraste de hipótesis será: $\begin{cases} H_0 : p_{ij} = p_{i.}\, p_{.j} \quad i = 1,...,h,\ j = 1,...,k \\ H_1 : \exists (i,j) / p_{ij} \neq p_{i.}\, p_{.j} \end{cases}$

CASO 1: Las probabilidades marginales están especificadas.

El estadístico a considerar será:

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{\left(n_{ij} - np_{i.}p_{.j}\right)^2}{np_{i.}p_{.j}} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{O_{ij}^2}{E_{ij}} - n \sim \chi^2_{hk-1}$$

y la región crítica vendrá dada por: $R = \left\{(x_1, x_2, ..., x_n) \in E \ / \ \chi^2 > \chi^2_{hk-1,\,1-\alpha}\right\}$

CASO 2: Las probabilidades marginales no están especificadas.

En este caso, habrá que estimar previamente las probabilidades marginales, mediante: $\hat{p}_{i.} = \frac{n_{i.}}{n} \quad \hat{p}_{.j} = \frac{n_{.j}}{n}$. En este caso, los valores esperados serán $E_{ij} = n\,\hat{p}_{i.}\,\hat{p}_{.j}$

Región crítica: $R = \left\{(x_1, x_2, ..., x_n) \in E \ / \ \chi^2 > \chi^2_{(h-1)(k-1),\,1-\alpha}\right\}$

Ejemplo: *Una compañía evalúa una propuesta para fusionarse con una corporación. El consejo de administración desea conocer si la opinión de los accionistas al respecto es independiente de la participación que posean en el capital de la compañía. Para ello, se toma una muestra aleatoria de 250 accionistas. En base a esta información, ¿existe alguna razón para dudar de que la opinión con respecto a la propuesta es independiente de la participación que posea la compañía? Usar un nivel de significación del 10 %.*

		Opinión sobre la fusión			Totales	$\hat{p}_{i.}$
	O_{ij}	A favor	En contra	Indecisos		
Número de acciones	Menos de 200	38	29	9	**76**	**0'304**
	200 – 1000	30	42	7	**79**	**0'316**
	Más de 1000	32	59	4	**95**	**0'38**
Totales		**100**	**130**	**20**	**250**	
$\hat{p}_{.j}$		**0'4**	**0'52**	**0'08**		

Solución: *El contraste planteado será:*

$$\begin{cases} H_0 : p_{ij} = p_{i.}\, p_{.j} \quad i = 1,\ldots,h,\ j = 1,\ldots,k \\ H_1 : \exists (i,j) / p_{ij} \neq p_{i.}\, p_{.j} \end{cases}$$

Al ser las probabilidades marginales desconocidas, habrá que estimarlas. Los valores obtenidos se encuentran en la tabla anterior.

Los valores esperados E_{ij} *se muestran en la siguiente tabla:*

		Opinión sobre la fusión			Totales
	E_{ij}	A favor	En contra	Indecisos	
Número de acciones	Menos de 200	30'4	39'52	6'08	**76**
	200 – 1000	31'6	41'08	6'32	**79**
	Más de 1000	38	49'40	7'60	**95**
Totales		**100**	**130**	**20**	**250**

		Opinión sobre la fusión			Totales
	$(O_{ij} - E_{ij})^2 / E_{ij}$	A favor	En contra	Indecisos	
Número de acciones	Menos de 200	1'90	2'80	1'40	**6'10**
	200 – 1000	0'08	0'02	0'07	**0'17**
	Más de 1000	0'95	1'87	1'71	**4'52**
Totales		**2'93**	**4'69**	**3'18**	**10'80**

Estadístico de contraste:

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 10'80$$

Punto crítico: $\chi^2_{(3-1)(3-1),\,1-0'1} = \chi^2_{4,\,0'9} = 13'28$

Por tanto, no tenemos evidencia suficiente para rechazar H_0, luego se acepta la independencia.

Estadística Empresarial II

Tema 9

Contrastes de hipótesis no paramétricos

Introducción (I)

En el tema anterior se han estudiado las **pruebas de bondad de ajuste**, que son contrastes no paramétricos cuyo objetivo es comprobar si un conjunto de datos que componen una muestra proceden de una determinada distribución de probabilidad. Sin embargo, hay situaciones en las que no se puede presuponer la distribución de probabilidad de la variable de la que procede la muestra, por lo que tendremos que emplear técnicas no paramétricas que permitan realizar contraste de hipótesis acerca de algunas características poblacionales.

Aunque deben cumplir algunos supuestos, como la independencia de las observaciones muestrales, los **contrastes no paramétricos de características poblacionales** son generalmente válidos cualquiera que sea la distribución de la población de la que se ha obtenido la muestra, por lo que son de gran utilidad en campos de la economía y la empresa, en los que es difícil que se cumpla la hipótesis de normalidad, exigida en la mayoría de los contrastes paramétricos estudiados.

Introducción (II)

En este tema, se desarrollarán algunas de las técnicas más utilizadas en el campo no paramétrico para probar hipótesis sobre características promedio de dos o más poblaciones independientes o relacionadas.

TEST NO PARAMÉTRICOS		
2 MUESTRAS	**Independientes**	U de Mann-Whitney
	Relacionadas	Wilcoxon
K MUESTRAS	**Independientes**	Kruskall-Wallis
	Relacionadas	Friedman

Test de Mann-Whitney

Se trata de una prueba de homogeneidad cuyo objetivo es contrastar si los datos muestrales proceden de dos poblaciones independientes con el mismo promedio.

Para este caso, existen los contrastes paramétricos de igualdad de medias de dos poblaciones independientes, basados en estadísticos con distribución normal Z o t de Student. Sin embargo, estas pruebas paramétricas requieren una serie de supuestos, como son la normalidad de las distribuciones poblacionales o la necesidad de datos cuantitativos.

Cuando los supuestos de las pruebas paramétricas no se verifican, existe **la prueba de Mann-Whitney**, que unicamente requiere dos condiciones:

- Observaciones extraídas de dos muestras aleatorias independientes.
- Valores que se puedan ordenar (ya sean cuantitativos o cualitativos en escala ordinal).

Como medida promedio o de tendencia central para comparar en ambas poblaciones se selecciona la **mediana**, ya que permite la aplicación al caso de que los datos sean ordinales. Así pues, el contraste a plantear será:

$$H_0: Me_1 = Me_2$$
$$H_1: Me_1 \neq Me_2$$

La **prueba de Mann-Whitney** se basa en la combinación de las **n** y **m** observaciones procedentes de las dos poblaciones ordenadas en orden creciente de magnitud. A cada una de ellas se le asigna un rango, de manera que a la observación más pequeña le corresponde el 1 y a la mayor, **n+m**. En el caso de que coincidan algunas observaciones, se asignará a cada una el valor medio de los rangos que les corresponden.

Si las muestras provienen de poblaciones con igual tendencia central se espera que los rangos se encuentren suficientemente dispersos, por lo que la suma de los rangos de ambas muestras no debería diferir demasiado.

El **estadístico U de Mann-Whitney** tiene la siguiente expresión:

$$U = n.m + \frac{n.(n+1)}{2} - T$$

siendo **T** la suma de rangos de la muestra de tamaño **n** (la más pequeña).

Las características de la distribución de probabilidad de **U** son:

$$E[U] = \frac{n.m}{2} \quad y \quad V(U) = \frac{n.m.(n+m+1)}{12}$$

Cuando los tamaños muestrales **n** y **m** son mayores que 10, la distribución de **U** se aproxima a una **normal** de media y varianza las indicadas anteriormente. Así pues, para el contraste bilateral propuesto inicialmente, la región crítica vendría dada por:

$$R = \left\{ (x_{11}, x_{12}, ..., x_{1n}), (x_{21}, x_{22}, ..., x_{2m}) \in E \,/\, |Z| > Z_{1-\alpha/2} \right\}$$

y el estadístico de contraste será: $$Z = \frac{U - E[U]}{\sqrt{V(U)}}$$

<u>Ejemplo:</u> Se quieren comparar dos dietas distintas para engorde de cerdos. Para ello, se seleccionan 11 cerdos de 6 meses de edad de la granja A que los alimenta con la primera dieta, y a 12 cerdos de la misma edad de la granja B que usa la segunda dieta, obteniéndose el incremento de peso en el último mes. Los resultados se reflejan en la tabla adjunta. Comprobar, con 5 % de significación, que existen diferencias significativas en el promedio de incremento de peso de los cerdos en el último mes entre las dos dietas. (Se ha comprobado previamente la no normalidad de los incrementos de peso de ambas granjas).

Granja A	22’3	18	14’7	19’1	21’7	22’9	21’5	19’3	17’2	23’1	16’5	
Granja B	15’2	18’1	14’7	15’4	17’5	15’6	24’8	12	20’9	13’6	14’6	13’3

Test de Wilcoxon

Se trata de una prueba no paramétrica que tiene como objetivo contrastar si existen diferencias significativas entre las medidas promedio de dos distribuciones poblaciones relacionadas. Para ello, se consideran dos muestras relacionadas, obtenidas exponiendo a un mismo grupo de individuos a dos situaciones diferentes.

Para este caso también existen pruebas paramétricas con el mismo objetivo, pero que requiere ciertos supuestos, que muchas veces no se satisfacen, de ahí que resulte conveniente la aplicación de una prueba no paramétrica como la **prueba de rangos y signos de Wilcoxon**. Esta prueba exige que los datos vengan dados en escala cuantitativa (no ordinal).

El contraste considerado será:

$$H_0: Me_1 = Me_2$$
$$H_1: Me_1 \neq Me_2$$

La **prueba de Wilcoxon** se basa en analizar tanto el signo como la magnitud de las diferencias entre cada par de observaciones de la muestra, procediendo de la siguiente manera:

- Se obtienen las diferencias $\mathbf{d_i}$ para los **n** pares de observaciones, que se ordenan sin importar el signo. Si alguna diferencia es nula, se prescinde de la misma.
- Se asigna un rango $\mathbf{r_i}$ a cada diferencia en base a ese orden, desde 1 hasta n (considerando el rango promedio en caso de igualdad en las diferencias), añadiéndole a cada rango el correspondiente signo de la diferencia.

El estadístico **W de Wilcoxon** se obtiene mediante la suma de todos los rangos positivos.

$$W = \sum_{r_i > 0} r_i$$

Las características de la distribución del estadístico **W** son:

$$E[W] = \frac{n.(n+1)}{4} \quad y \quad V(W) = \frac{n.(n+1).(2n+1)}{24}$$

Si la hipótesis nula es cierta, se espera que **W** tenga aproximadamente igual valor que la suma de los rangos negativos.

Cuando el tamaño muestral **n** es mayor que 10, la distribución de **W** se aproxima a una **normal** de media y varianza las indicadas anteriormente. Así pues, para el contraste bilateral, la región crítica vendría dada por:

$$R = \left\{ (x_1, x_2, \ldots, x_n) \in E \,/\, |Z| > Z_{1-\alpha/2} \right\}$$

y el estadístico de contraste será: $Z = \dfrac{W - E[W]}{\sqrt{V(W)}}$

<u>Ejemplo:</u> En un experimento para comparar dos materiales distintos, A y B, que se deben utilizar para fabricar tacones de zapatos de caballero, se seleccionó a 15 hombres y se les proporcionó un par de zapatos nuevos de los cuales un tacón estaba hecho con el material A y el otro con el material B. Al principio del experimento, cada tacón tenía un grosor de 10 mm. Después de usar los zapatos durante un mes, se midió el grosor restante, resultando:

PAR	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Material A	6’6	7’0	8’3	8’2	5’2	9’3	7’9	8’5	7’8	7’5	6’1	8’9	6’1	9’4	9’1
Material B	7’4	5’4	8’8	8’0	6’8	9’1	6’3	7’5	7’0	6’6	4’4	7’7	4’2	9’4	9’1

Verificar que no existen diferencias significativas en el grosor resultante de los tacones entre ambos materiales, con una significación del 5 %.

Test de Kruskall-Wallis

Se trata de una prueba no paramétrica que permite comprobar si varias muestras independientes (más de 2) proceden de poblaciones con distribución de igual medida promedio.

En este caso, la técnica paramétrica correspondiente es el **análisis de la varianza**, que requiere el cumplimiento de algunos supuestos. Cuando estos supuestos no se verifican, se puede optar por una prueba no paramétrica, como el **test de Kruskall-Wallis**.

La prueba de **Kruskall-Wallis** se considera como una extensión de la de Mann-Whitney cuando se trata de verificar la homogeneidad de promedios de **k** (mayor que 2) muestras aleatorias independientes, siendo requisito, al menos, que los datos sean ordinales.

El contraste planteado sería:

$$H_0: Me_1 = Me_2 = \ldots = Me_k$$
$$H_1: \exists\ (i,j) \,/\, Me_i \neq Me_j$$

Se procede como sigue:

- Se parte de un conjunto de **n** observaciones (**n** = $n_1 + n_2 + \ldots + n_k$) que se ordenan de forma creciente.
- Se asignan rangos, de manera que el valor 1 le corresponde a la observación menor y el n a la mayor. Los empates se resuelven de análoga forma a los test anteriores.
- Se suman los rangos de cada muestra, denotando por $\mathbf{R_j}$ a la suma de los rangos de la muestra j-ésima.
- Se eleva al cuadrado cada $\mathbf{R_j}$ y se divide entre el tamaño muestral correspondiente a dicha muestra, $\mathbf{n_j}$. A continuación, se procede a la suma de todos los $\mathbf{R_j^2}$.

El estadístico de **Kruskall-Wallis** presenta la siguiente expresión:

$$KW = \frac{12}{n.(n+1)}\left(\sum_{j=1}^{k}\frac{R_j^2}{n_j}\right) - 3.(n+1)$$

y sigue una distribución χ^2_{k-1} cuando los tamaños muestrales $\mathbf{n_j}$ son todos mayores que 5. Por tanto, los valores de **KW** serán siempre positivos.

La región crítica será:

$$R = \left\{(x_{11,}\ldots,x_{1n_1}),(x_{21,}\ldots,x_{2n_2}),\ldots,(x_{k1,}\ldots,x_{kn_k}) \in E \,/\, KW > \chi^2_{k-1,1-\alpha}\right\}$$

Si se rechaza la hipótesis nula, la conclusión será que existen diferencias entre los **k** grupos, por lo que al menos uno de ellos presentará diferencias con respecto a los demás. Este test no informa dónde se encuentran las diferencias ni cuántos grupos son diferentes entre sí.

Ejemplo: Se desea saber si la renta familiar influye en el grado de cultura de los hijos. Para verificarlo se toman cuatro niveles de renta y en cada uno un cierto número de familias con niños comprendidos entre ciertas edades. Se somete a los niños a una serie de tests cuyos resultados, expresados en la tabla adjunta, reflejan el grado de cultura. Usar un 5 % de significación.

Nivel renta 1	171	146	117	191	164	137	126	182	155	121		
Nivel renta 2	121	144	164	196	125	155	137	191				
Nivel renta 3	108	108	108	178	149	117	119	89	155	129	98	98
Nivel renta 4	121	108	96	72	121	96	72					

Test de Friedman

Se trata de una prueba no paramétrica que permite comprobar si **k** muestras dependientes o relacionadas (más de 2) proceden de poblaciones dependientes con igual medida promedio (mediana), siempre que la escala de medida sea, al menos, ordinal. Las muestras relacionadas se obtienen estudiando el comportamiento de un mismo grupo de individuos bajo **k** situaciones o tratamientos diferentes.

La contrapartida paramétrica para esta prueba corresponde al **análisis de la varianza** (concretamente, el caso de bloques aleatorizados o medidas repetidas).

El contraste no paramétrico planteado sería:

H_0: $Me_1 = Me_2 = \ldots = Me_k$
H_1: $\exists\ (i,j) / Me_i \neq Me_j$

Los datos, en este caso, se presentan en una tabla de doble entrada con **n** filas (individuos) y **k** columnas (tratamientos o situaciones). A partir de la misma, el **test de Friedman** procede según los siguientes pasos:

- De forma independiente, se asignan rangos a cada una de las observaciones de cada fila, correspondiendo el valor 1 a la observación menor y **k** a la mayor. En caso de empate, se procede como en los anteriores casos.
- Se calculan las sumas de los rangos de cada columna o muestra, que denotaremos mediante $\mathbf{R_j}$.
- Se eleva al cuadrado cada $\mathbf{R_j}$ y se determina la suma de todos ellos.

El estadístico de **Friedman** se obtiene como se indica a continuación:

$$F_R = \frac{12}{n.k.(k+1)}\left(\sum_{j=1}^{k} R_j^2\right) - 3.n.(k+1)$$

Si **k ≥ 5** o **n ≥ 10**, la distribución de $\mathbf{K_R}$ se aproxima a una χ^2_{k-1}. Por tanto, los valores de $\mathbf{F_R}$ serán siempre positivos.

La región crítica asociada al contraste considerado será:

$$R = \left\{(x_{11,}...,x_{1n}),(x_{21,}...,x_{2n}),...,(x_{k1,}...,x_{kn}) \in E \,/\, F_R > \chi^2_{k-1,1-\alpha}\right\}$$

<u>Ejemplo:</u> Cuatro jueces se encargan de calificar en una competición de salto que incluye a 10 finalistas. Los datos que figuran en la tabla siguiente son calificaciones, dónde un 10 indica un salto perfecto. Para una significación del 1 %, determinar si existe diferencia significativa en las calificaciones que otorgan cada uno de los cuatro jueces.

	Juez			
Competidor	1	2	3	4
1	8’5	8’6	8’2	8’4
2	9’8	9’7	9’4	9’6
3	7’9	8’1	7’5	8’2
4	9’7	9’8	9’6	9’6
5	6’2	6’8	6’9	6’5
6	8’9	9’2	8’1	8’7
7	9’2	9’2	8’7	8’9
8	8’4	8’5	8’4	8’6
9	9’2	9’6	8’9	9’5
10	8’8	9’2	8’6	9’3

Estadística Empresarial II

Tema 10

Análisis de la varianza

Introducción

El **Analisis de la Varianza** es una prueba estadística de homogeneidad de los comportamientos medios de una determinada característica o **variable respuesta**, para **k** poblaciones independientes, correspondientes a **k** condiciones distintas de un determinado **factor**.

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_1: \exists\ (i,j) \,/\, \mu_i \neq \mu_j$$

Esta prueba paramétrica puede considerarse como una extensión del contraste paramétrico de igualdad de medias para dos poblaciones independientes, ya estudiado anteriormente. Al igual que éste, el **Análisis de la Varianza** requiere la verificación de una serie de supuestos, como puede ser la normalidad, homocedasticidad, etc.

En caso de que no se cumpla algunos de estos supuestos, hemos estudiado algunos procedimientos alternativos, encuadrados dentro de los contrastes no paramétricos.

Conceptos

A continuación se van a definir los principales términos que intervienen en el **análisis de la varianza**:

● **Variable respuesta:** Es la variable dependiente o característica objeto de nuestro estudio y que cuantifica el efecto de una serie de condiciones que influyen sobre ella. Por tanto, se necesita que dicha variable pueda medirse en escala cuantitativa.

● **Factor:** Es cada una de las variables independientes o explicativas que influyen en la característica de estudio o **variable respuesta**. Cada factor debe incluir las diferentes condiciones a las que se somete a los individuos para analizar el efecto diferencial de las mismas.

A las distintas modalidades que presenta un **factor** se les denomina **niveles**. Estos suelen diferenciarse en **tratamientos** (cuando se pueden manipular las condiciones del factor) o **modos de clasificación** (cuando las condiciones del factor no son susceptibles de manipulación).

● **Supuestos básicos:** Para poder emplear el **análisis de la varianza** se necesita el cumplimiento de los siguientes supuestos:

(1) Las muestras han de ser extraídas de forma aleatoria.

(2) Las puntuaciones u observaciones han de ser independientes entre sí.

(3) Las observaciones del j-ésimo grupo ($\mathbf{X_{ij}}$, $i=1,\ldots,n_j$) deben tener distribución Normal de media μ_j.

(4) Todos los grupos deben tener la misma varianza poblacional σ^2, lo que se conoce como **homocedasticidad**.

(5) La variable respuesta debe ser cuantitativa, mientras que la variable independiente o factor se establece a modo de categorías, pudiendo ser cuantitativa o cuantitativa.

● **Diseños según el tipo de factores:** Los factores que influyen en una determinada variable respuesta pueden ser **fijos** o **aleatorios**. Se dice que un factor es **fijo** cuando los niveles observados del mismo incluyen todos los posibles, o bien, todos los que interesan. Sin embargo, se dirá que es **aleatorio** cuando el número de posibles niveles del factor es elevado y se seleccionan aleatoriamente algunos para realizar el estudio.

Estas consideraciones dan lugar a distintos modelos:

(1) Modelo de efectos fijos: Intervienen únicamente factores fijos.

(2) Modelo de efectos aleatorios: Intervienen únicamente factores aleatorios.

(3) Modelo de efectos mixtos: Intervienen factores tanto fijos como aleatorios.

El **Análisis de la Varianza** permite separar el efecto que sobre la variable respuesta ejerce uno o varios factores controlados del de otros no controlados, contrastando la influencia de los factores controlados sobre los resultados.

La variabilidad total de la variable respuesta se puede dividir en dos partes. Por un lado, estaría la causada por el factor controlable y sus niveles; por otro, la originada por el resto de factores, conocidos o no, que influyen sobre ella, llamada variabilidad debida al error experimental. Esta división daría lugar a dos tipos de varianzas:

(1) **Varianza dentro de los grupos:** Representa la variabilidad debida al error experimental, causante de las posibles diferencias existentes entre los elementos de cada grupo.

(2) **Varianza entre grupos:** Representa la variabilidad existente entre los grupos debida al efecto de los diferentes niveles del factor.

Para decidir si existen diferencias entre o no como consecuencia de los diferentes niveles del factor, esta técnica se basará en la comparación de los estimadores de las dos varianzas definidas.

Modelo factorial simple o ANOVA I

Este modelo se caracteriza porque la variable respuesta considerada depende de un único factor con **k** niveles, quedando el resto de las causas de variación englobadas en el error experimental.

El objetivo del mismo será contrastar la homogeneidad de promedios de la variable respuesta para **k** poblaciones independientes, pudiendo expresarse de la siguiente manera:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_1: \exists\ (i,j) \,/\, \mu_i \neq \mu_j$$

Si rechazamos la hipótesis nula, concluiremos que existen diferencias significativas entre los comportamientos promedio, ya que, al menos uno de ellos es diferente a los demás.

El modelo considerado en el ANOVA I es el siguiente: $\mathbf{X_{ij} = \mu + A_j + \varepsilon_{ij}}$

- $\mathbf{X_{ij}}$ → Valor de la variable respuesta para el i-ésimo individuo del j-ésimo grupo.
- μ → Constante común para todas las observaciones que representa a la media poblacional.
- $\mathbf{A_j}$ → Es la aportación cuantitativa del j-ésimo nivel del factor a la puntuación total, que refleja la diferencia entre la puntuación esperada del j-ésimo grupo μ_j y la puntuación esperada para toda la población, μ.
- ε_{ij} → Error experimental correspondiente a cada puntuación, que indica la parte de $\mathbf{X_{ij}}$ no explicada por las otras dos componentes. Se verifica que $\varepsilon_{ij} \sim \mathbf{N}(0, \sigma^2)$.

Al ser los valores de μ, A_j y ε_{ij} desconocidos, habrá que estimarlos, por ejemplo, utilizando el método de los mínimos cuadrados, dando lugar a:

$$\hat{\mu} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij} \qquad \hat{A}_j = \bar{X}_j - \bar{X} \qquad \hat{\varepsilon}_{ij} = X_{ij} - \bar{X}_j$$

Por tanto, el modelo quedaría: $X_{ij} = \bar{X} + (\bar{X}_j - \bar{X}) + (X_{ij} - \bar{X}_j) \Rightarrow X_{ij} - \bar{X} = (\bar{X}_j - \bar{X}) + (X_{ij} - \bar{X}_j)$

Así pues:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X})^2=\sum_{j=1}^{k}n_j(\bar{X}_j-\bar{X})^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)^2 \;\; ya\ que \;\; \sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)=0$$

En términos de las sumas de cuadrados se tiene que: **SCT = SCF + SCE**

A partir de estas sumas de cuadrados se obtienen los estimadores de las varianzas:

(1) Cuasivarianza total: $\hat{S}_T^2=\dfrac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X})^2}{n-1}=\dfrac{SCT}{n-1}$

(2) Cuasivarianza debida al factor (entre grupos): $\hat{S}_F^2=\dfrac{\sum_{j=1}^{k}(\bar{X}_j-\bar{X})^2}{k-1}=\dfrac{SCF}{k-1}$

(3) Cuasivarianza debida al error (dentro de los grupos):

$$\hat{S}_E^2=\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)^2}{n-k}=\frac{SCE}{n-k}$$

Se puede demostrar que $\mathbf{S^2_T}$, $\mathbf{S^2_F}$ y $\mathbf{S^2_E}$ son estimadores insesgados de σ^2.

Además:

$$\frac{(k-1)\hat{S}_F^2}{\sigma^2} \square \chi^2_{k-1}$$

$$\frac{(n-k)\hat{S}_E^2}{\sigma^2} \square \chi^2_{n-k}$$

$$\Rightarrow \quad \frac{\hat{S}_F^2}{\hat{S}_E^2} \square F_{k-1,n-k}$$

La región crítica asociada al contraste será:

$$R = \left\{ (x_{11},\ldots,x_{n_1 1}),(x_{12},\ldots,x_{n_2 2}),\ldots,(x_{1k},\ldots,x_{n_k k}) \in E \,/\, F > F_{k-1,\,n-k,\,1-\alpha} \right\}$$

siendo el estadístico de contraste: $F = \dfrac{\hat{S}_F^2}{\hat{S}_E^2}$

Para facilitar los cálculos a la hora de determinar las sumas de cuadrados, se expresarán las expresiones en función de las sumas de las puntuaciones, ya sea la total **T** o la de cada grupo, $\mathbf{T_j}$.

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij} = \frac{T}{n} \qquad \bar{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij} = \frac{T_j}{n_j} \qquad n = \sum_{j=1}^{k} n_j$$

Por tanto, las sumas de cuadrados se puede expresar como sigue:

$$SCT = \sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij}^2 - \frac{T^2}{n} \qquad SCF = \sum_{j=1}^{k}\frac{T_j^2}{n_j} - \frac{T^2}{n} \qquad SCE = \sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij}^2 - \sum_{j=1}^{k}\frac{T_j^2}{n_j}$$

La tabla de los cálculos a realizar será la siguiente:

Niveles	**1**	**2**	**…**	**k**	
	X_{11}	X_{12}	:	X_{1k}	
	X_{21}	X_{22}	:	X_{2k}	
	:	:	:	:	
	X_{n11}	X_{n22}	…	X_{nkk}	
$\mathbf{T_j}$	T_1	T_2	…	T_k	**T**
$\mathbf{T_j^2}$	T_1^2	T_2^2	…	T_k^2	$\mathbf{\Sigma_j T_j^2}$
$\mathbf{T_j^2 / n_j}$	T_1^2 / n_1	T_2^2 / n_2	…	T_k^2 / n_k	$\mathbf{\Sigma_j T_j^2/n_j}$
$\mathbf{\Sigma_i X_{ij}^2}$	$\Sigma_i X_{i1}^2$	$\Sigma_i X_{i2}^2$	…	$\Sigma_i X_{ik}^2$	$\mathbf{\Sigma_j\Sigma_i X_{ij}^2}$

Y el cuadro resumen de todo el proceso vendrá dado por:

Variabilidad	Suma de cuadrados	g. l.	Estimador	Estadístico	Punto crítico
FACTOR	SCF	k-1	$S_F^2 = SCF/(k-1)$	$F = S_F^2 / S_E^2$	$F_{k-1,n-k,1-\alpha}$
ERROR	SCE	n-k	$S_E^2 = SCF/(n-k)$		
TOTAL	SCT	n-1	$S_T^2 = SCT/(n-1)$		

<u>Ejemplo:</u> Se ha realizado un experimento con el fin de comparar los precios de la barra de pan de molde en cuatro ciudades diferentes. La muestra está formada por ocho almacenes seleccionados aleatoriamente para las tres primeras ciudades, mientras que la cuarta está formada por siete almacenes.

1	2	3	4
139	138	134	149
143	141	139	150
145	144	135	148
141	143	138	150
144	137	139	146
138	140	136	151
140	143	140	149
141	140	135	

Partiendo de la hipótesis de normalidad para los precios de la barra de pan de molde en cada una de las ciudades, ¿los datos nos proporcionan suficiente evidencia para indicar que hay diferencias significativas en el precio medio en las cuatro ciudades, con un nivel de significación del 5%?