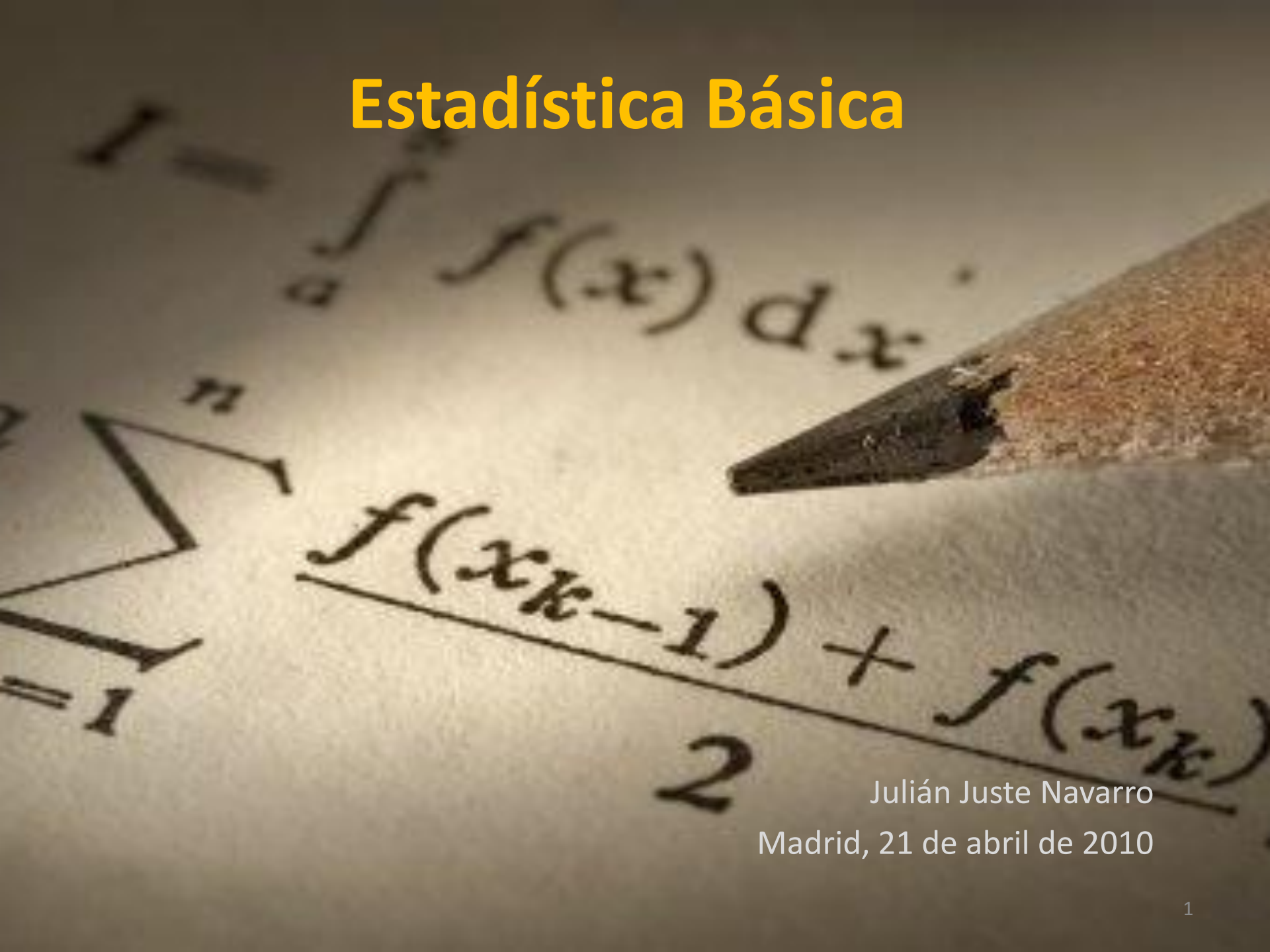


# Estadística Básica



Julián Juste Navarro

Madrid, 21 de abril de 2010

# ¿Qué es la estadística?

- \* **Ciencia con base matemática: recolección, análisis e interpretación** de datos. Busca explicar **condiciones regulares** en **fenómenos** de tipo **aleatorio**.
- \* Término que se asocia a todo proceso cuyo resultado no es previsible más que en razón de la intervención del azar.
- \* **Métodos y procedimientos:** recoger, clasificar, resumir, hallar regularidades y analizar los *datos*, siempre y cuando la *variabilidad* e *incertidumbre* sea una causa intrínseca de los mismos; así como de realizar *inferencias* a partir de ellos, con la finalidad de ayudar a la toma de *decisiones* y en su caso formular *predicciones*.

# ¿Qué es la bioestadística?

Ciencia que aplica los conocimientos de la estadística y en concreto el análisis estadístico a los problemas y objetos de estudio de la **biología**.

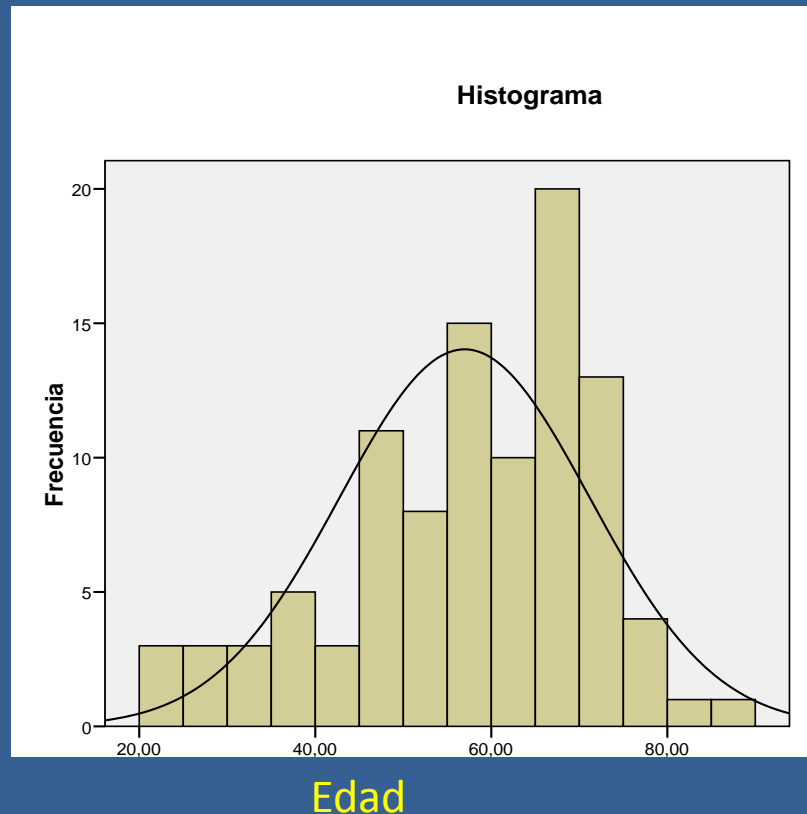
## Las dos partes de la estadística

**Estadística descriptiva:** Describe, analiza y representa un grupo de datos utilizando **métodos numéricos** y gráficos que resumen y presentan la información contenida en ellos.

**Estadística inferencial:** Se apoya en el cálculo de probabilidades y a partir de datos muestrales, efectúa **estimaciones, decisiones, predicciones** u otras **generalizaciones** sobre un conjunto mayor de datos.

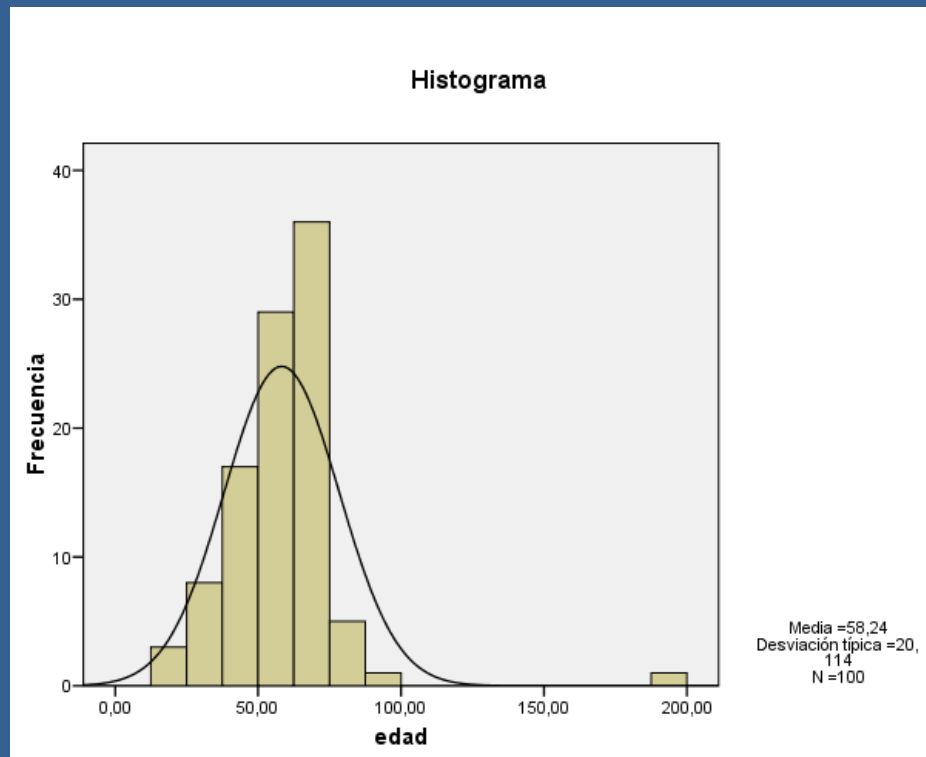
# Estadística descriptiva

Sirve para hacerse una idea de las **características** de una muestra de datos.  
Sirve para analizar cómo se **distribuyen** los datos.



# Estadística descriptiva

Sirve para identificar **valores extremos** y dentro de los mismos identificar cuáles pueden ser **erróneos**.



# Estadística descriptiva

O para analizar si existe suficiente número de datos en cada uno de los grupos de comparación, dividiendo la muestra en grupos uniformes, estratos (o clústers), con respecto a un factor determinado que queremos estudiar.

Por ejemplo, realizamos un clúster por sexos y resulta que tenemos 4 mujeres y 505 hombres. En el caso que en el diseño de la experiencia se quiera estudiar el factor sexo, nos vamos a encontrar con una muestra **descompensada**.

# Diferencia entre **población y muestra**:

- **Población:** Conjunto de todos los individuos que portan información sobre el fenómeno que se estudia.
- **Muestra: Subconjunto** que **seleccionamos** de la población.
- La **variable estadística** es una **característica** (magnitud, vector o número) que puede ser medida, adoptando diferentes valores en cada uno de los casos de un estudio.

# La variable estadística

- Según la **medición** CUANTITATIVAS Y CUALITATIVAS

## Variables cualitativas

Expresan cualidad o modalidad. Cada modalidad se denomina atributo o categoría y la medición consiste en una clasificación los mismos. Las hay ordinales y nominales. También **dicotómicas** si sólo toman dos valores posibles como *sí y no*, *hombre y mujer* o **politómicas** cuando pueden adquirir tres o más valores.

- *Variable cualitativa ordinal*: La variable puede tomar distintos valores ordenados siguiendo una escala establecida, aunque no es necesario que el intervalo entre mediciones sea uniforme, por ejemplo, *leve, moderado, grave*.
- *Variable cualitativa nominal*: En esta variable los valores no pueden ser sometidos a un criterio de orden como por ejemplo los colores o el lugar de residencia.

# La variable estadística

## Variables cuantitativas

Se expresan mediante cantidades numéricas.

- *Discretas*: Es la variable que presenta separaciones o interrupciones en la escala de valores que puede tomar. Estas separaciones o interrupciones indican la ausencia de valores entre los distintos valores específicos que la variable pueda asumir. Ejemplo: El número de hijos (1, 2, 3, 4, 5).
- *Continua*: Es la variable que puede adquirir cualquier valor dentro de un intervalo especificado de valores. Por ejemplo el peso (2,3 kg, 2,4 kg, 2,5 kg, ...) o la altura (1,64 m, 1,65 m, 1,66 m, ...), que solamente está limitado por la precisión del aparato medidor, en teoría permiten que siempre exista un valor entre dos cualesquiera.

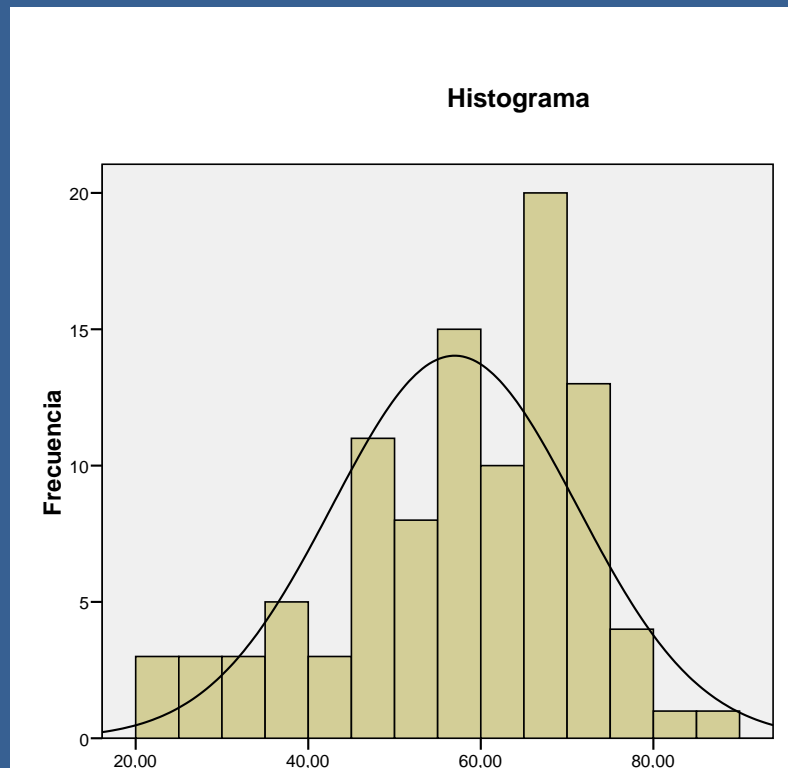
# La variable estadística

Según la influencia que asignemos a unas variables sobre otras,

- **Variables independientes:** Son las que el investigador escoge para establecer agrupaciones en el estudio, clasificando intrínsecamente a los casos del mismo. También son los factores que se deriva de una característica o propiedad del fenómeno estudiado.
- **Variables dependientes:** Son las variables de respuesta que se observan en el estudio y que podrían estar influenciadas por los valores de las variables independientes.

# La variable estadística

Las variables continuas, según su comportamiento, podrán estar asociadas o no a una **distribución normal**: campana de Gauss:



# LA DISTRIBUCIÓN NORMAL (Introducción)

MUY IMPORTANTE, YA QUE LA ESTADÍSTICA INFERENCIAL APLICA METODOLOGÍAS DEPENDIENDO SI TRABAJAMOS CON VARIABLES NORMALES O NO

- Para comprobar el grado de “normalidad” de una variable, existen varios test estadísticos como los de Kolmogorof-Smirnof, Shapiro-Wilk o Levene, que se aplican con cualquier paquete estadístico.
- En el caso que las distribuciones de una o varias variables se puedan asociar a una distribución normal, podremos aplicar sobre las mismas **análisis estadísticos paramétricos**.
- Una condición importante para considerar una variable como normal es que dicha variable sea continua. Esto, en la práctica (cuidados pariativos), descarta muchas variables, que analizan cuestiones psicosociales y espirituales, y que suelen establecer una gradación o puntuación, por lo que son variables cuantitativas discretas.
- En caso contrario, variables continuas que no sean normales y variables discretas, se aplicarán en general los **análisis estadísticos no paramétricos**.

# LA DISTRIBUCIÓN NORMAL (Introducción)

MUY IMPORTANTE, YA QUE LA ESTADÍSTICA INFERENCIAL APLICA METODOLOGÍAS DEPENDIENDO SI TRABAJAMOS CON VARIABLES NORMALES O NO

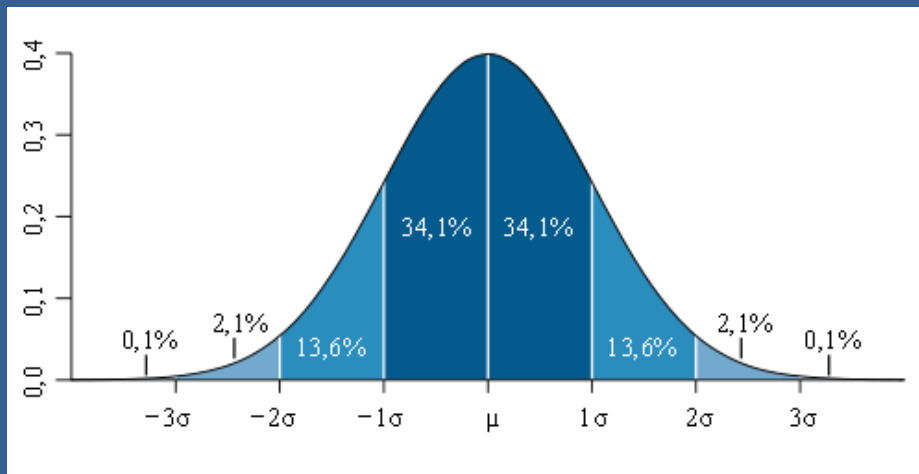
- Pero esto no es del todo cierto para las variables cuantitativas discretas, existe un teorema, llamado del límite central, que indica que, en condiciones generales, la distribución de la suma de variables aleatorias tiende a una distribución normal cuando la cantidad de variables es muy grande, es decir, si la muestra es lo suficientemente grande, se podrá aproximar la distribución de una variable a una distribución normal y, por tanto aplicar métodos paramétricos
- **¿Cuándo de grande?**, según el teorema de Moivre, una distribución binomial, propia de variables discretas, tiene una buena aproximación a una distribución normal si  $np$  y  $nq$  son  $> 15$ . Siendo  $p$  la probabilidad de acertar con la hipótesis que se busca,  $q$  la probabilidad de fallar ( $1-p$ ) y  $n$  el tamaño de la muestra, por lo que si hablamos de una probabilidad del 95%,  $q$  será el 5% (0,05) y  $n$  deberá ser **como mínimo mayor de 300**.

# Medidas del valor central de una muestra

- **Media aritmética**, que se aplica a las variables paramétricas, es una medida de tendencia central resultado de sumar todos los valores de la muestra y dividir esta suma por el número de individuos de la muestra.
- **Mediana** es el valor que se encuentra justo en el medio de todo el conjunto de datos. Es decir si ordenamos una variable de menor a mayor, la mediana dejaría el 50% de los datos a cada lado: 15, 22, 59, **84**, 105, 126.
- **Moda** es el valor que se repite más, siempre asociado a variables discretas: 1, 1, 3, 2, 1, 5, 3, 4, 7, 4, 1, 2, 1 → **1**

# Medidas de dispersión

- **Desviación estándar**, se aplica a variables paramétricas, es una medida que informa de la media de distancias que tienen los datos respecto de su media aritmética, expresada en las mismas unidades que la variable, informa de la dispersión de la media.



$$\sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

# Medidas de dispersión

**Cuartiles, se aplican a variables no paramétricas, son los valores que aparecen al dividir los datos ordenados en cuatro grupos de igual tamaño, es decir, los que dejan por delante una cuarta parte de los datos (25% de los datos), dos cuartas partes de los datos (50%) que se correspondería a la mediana, o tres cuartas partes de los datos.**

- El primer cuartil:
  - Cuando  $n$  es par:  $1 \cdot n/4$
  - Cuando  $n$  es impar:  $1(n+1)/4$
- El segundo cuartil (mediana):
  - Cuando  $n$  es par:  $2 \cdot n/4$
  - Cuando  $n$  es impar:  $2(n+1)/4$

# Medidas de dispersión

- Para el tercer cuartil
  - Cuando n es par:  $3 \cdot n / 4$
  - Cuando n es impar:  $3(n+1) / 4$

15, 18, **22**, 54, 59, **84**, 105, 126, **138**, 160, 164

**Percentiles: similar a los cuartiles pero en 100 partes:** Es cada uno de los 99 segmentos que tomamos al dividir una muestra o un conjunto de elementos ordenados por cien partes de igual frecuencia.

# Medidas de dispersión

## VALORES REPRESENTATIVOS DE LAS DIFERENTES DISTRIBUCIONES

### Paramétricas

- Media
- Desviación estándar

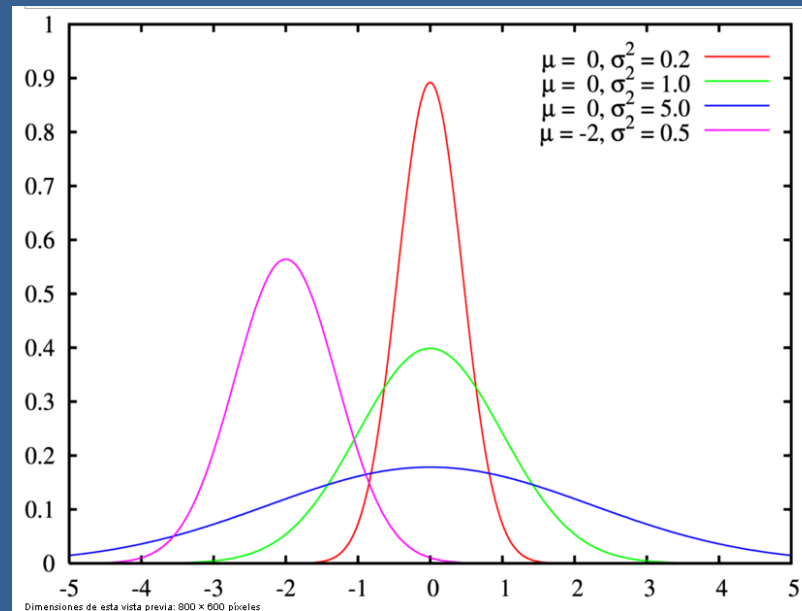
### No paramétricas

- Mediana
- Cuartiles

# DISTRIBUCIONES PARAMÉTRICAS

# Distribución normal

También llamada **campana distribución de Gauss** o **distribución gaussiana**, a una de las distribuciones de probabilidad de variable continua que con más frecuencia aparece en fenómenos reales. La gráfica de su función de densidad tiene una forma acampanada y es simétrica respecto de un determinado parámetro



# Características de la función normal

**Simetría y apuntamiento o kurtosis. Simetría respecto a la media, que la función de densidad es simétrica respecto al eje de la media. Kurtosis, grado de apuntamiento, grado en en que se agrupan los resultados alrededor de la media.**

Algunos ejemplos de variables asociadas a fenómenos naturales que siguen el modelo de la normal son:

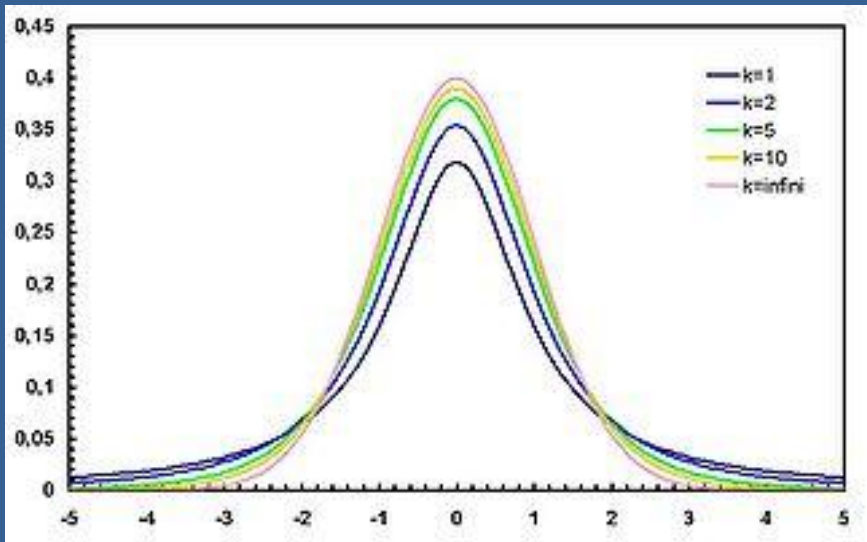
- caracteres morfológicos de individuos como la estatura;
- caracteres fisiológicos como el efecto de un fármaco;
- caracteres sociológicos como el consumo de cierto producto por un mismo grupo de individuos;
- caracteres psicológicos como el cociente intelectual;
- nivel de ruido en telecomunicaciones;
- errores cometidos al medir ciertas magnitudes;

# Distribución t de Student

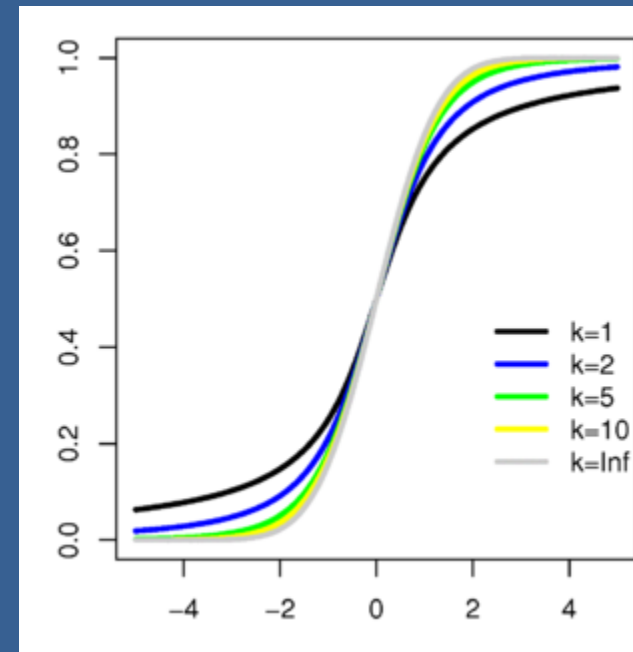
La **distribución t (de Student)**: distribución de probabilidad para estimar la media de una población normalmente distribuida cuando el tamaño de la muestra es pequeño. Aparece de manera natural al realizar la prueba t de Student para la determinación de las diferencias entre dos medias muestrales y para la construcción del intervalo de confianza para la diferencia entre las medias de dos poblaciones cuando se desconoce la desviación típica de una población y ésta debe ser estimada a partir de los datos de una muestra.

# Distribución t de Student

Función de **densidad** de probabilidad



Función de **distribución**  
de probabilidad



# Distribución t de Student

La distribución t de Student es la distribución de probabilidad del cociente:

$$\frac{Z}{\sqrt{V/\nu}}$$

Donde:

- $Z$  tiene una distribución normal de media nula y varianza 1,
- $V$  tiene una distribución chi-cuadrado con  $\nu$  grados de libertad,
- $Z$  y  $V$  son independientes.

# Distribución t de Student

- Lo importante en nuestro caso es el parámetro  $v$  representa el número de *grados de libertad*. La distribución depende únicamente de  $v$ , pero no de la media  $\mu$  o la desviación estándar  $\sigma$ .
- Para construir un intervalo de confianza hay que tener claro el concepto media, desviación estándar y grados de libertad. Los grados de libertad representan el número de posiciones que puede tomar un valor. Es decir, en el juego de las sillas, por ejemplo si tenemos una muestra de 102 sujetos y estos tienen 102 posiciones (sillas), si se levantan de su silla para moverse a cualquier otra silla, ¿a cuántas podrán ir?  $\rightarrow n - 1$  En nuestro caso a 101 sillas.

# Distribución t de Student

- Intervalo de confianza de la media

$$\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

- Siendo  $\bar{X}$  la media,  $S$  la desviación estándar,  $n$  el número de elementos de la muestra y  $\alpha$  es uno menos la probabilidad en tanto por uno.

Es decir, si buscamos un intervalo al 95% de probabilidad  $\alpha = 0,05$ . Entraríamos en la tabla para calcular  $t$  con  $\alpha / 2 = 0,025 \rightarrow p = 0,975$  y con  $n-1$  grados de libertad.

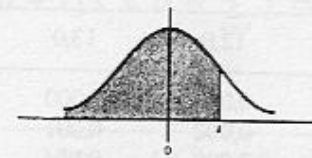
# Distribución t de Student

Si buscamos un intervalo al 95% de probabilidad  $\alpha = 0,05$ . Entraríamos en la tabla para calcular t con  $\alpha / 2 = 0,025 \rightarrow p = 0,975$  y con n-1 grados de libertad.

En la tabla la probabilidad p es F y los grados de libertad n se denominan  $\gamma$  (gamma)

Fuente:

*Estadística para biología y ciencias de la salud.*  
McGraw-Hill.



$$F(t) = P[T \leq t]$$

$\gamma \backslash F$	0,60	0,75	0,90	0,95	0,975	0,99	0,995	0,9995
1	0,325	1,000	3,078	6,314	12,706	31,821	63,657	636,619
2	0,289	0,816	1,886	2,920	4,303	6,965	9,925	31,598
3	0,277	0,765	1,638	2,353	3,182	4,541	5,841	12,924
4	0,271	0,741	1,533	2,132	2,776	3,747	4,604	8,610
5	0,267	0,727	1,476	2,015	2,571	3,365	4,032	6,869
6	0,265	0,718	1,440	1,943	2,447	3,143	3,707	5,959
7	0,263	0,711	1,415	1,895	2,365	2,998	3,499	5,408
8	0,262	0,706	1,397	1,860	2,306	2,896	3,355	5,041
9	0,261	0,703	1,383	1,833	2,262	2,821	3,250	4,781
10	0,260	0,700	1,372	1,812	2,228	2,764	3,169	4,587
11	0,260	0,697	1,363	1,796	2,201	2,718	3,106	4,437
12	0,259	0,695	1,356	1,782	2,179	2,681	3,055	4,318
13	0,259	0,694	1,350	1,771	2,160	2,650	3,012	4,221
14	0,258	0,692	1,345	1,761	2,145	2,624	2,977	4,140
15	0,258	0,691	1,341	1,753	2,131	2,602	2,947	4,073
16	0,258	0,690	1,337	1,746	2,120	2,583	2,921	4,015
17	0,257	0,689	1,333	1,740	2,110	2,567	2,898	3,965
18	0,257	0,688	1,330	1,734	2,101	2,552	2,878	3,922
19	0,257	0,688	1,328	1,729	2,093	2,539	2,861	3,883
20	0,257	0,687	1,325	1,725	2,086	2,528	2,845	3,850
21	0,257	0,686	1,323	1,721	2,080	2,518	2,831	3,819
22	0,256	0,686	1,321	1,717	2,074	2,508	2,819	3,792
23	0,256	0,685	1,319	1,714	2,069	2,500	2,807	3,767
24	0,256	0,685	1,318	1,711	2,064	2,492	2,797	3,745
25	0,256	0,684	1,316	1,708	2,060	2,485	2,787	3,725
26	0,256	0,684	1,315	1,706	2,056	2,479	2,779	3,707
27	0,256	0,684	1,314	1,703	2,052	2,473	2,771	3,690
28	0,256	0,683	1,313	1,701	2,048	2,467	2,763	3,674
29	0,256	0,683	1,311	1,699	2,045	2,462	2,756	3,659
30	0,256	0,683	1,310	1,697	2,042	2,457	2,750	3,646
40	0,255	0,681	1,303	1,684	2,021	2,423	2,704	3,551
60	0,254	0,679	1,296	1,671	2,000	2,390	2,660	3,460
120	0,254	0,677	1,289	1,658	1,980	2,358	2,617	3,373
$\infty$	0,253	0,674	1,282	1,645	1,960	2,326	2,576	3,291

# Distribución F

La **distribución F** es una distribución de probabilidad continua. También se la conoce como **distribución F de Snedecor** o como **distribución F de Fisher-Snedecor**.

Una variable aleatoria de distribución  $F$  se construye como el siguiente cociente:

$$F = \frac{U_1/d_1}{U_2/d_2}$$

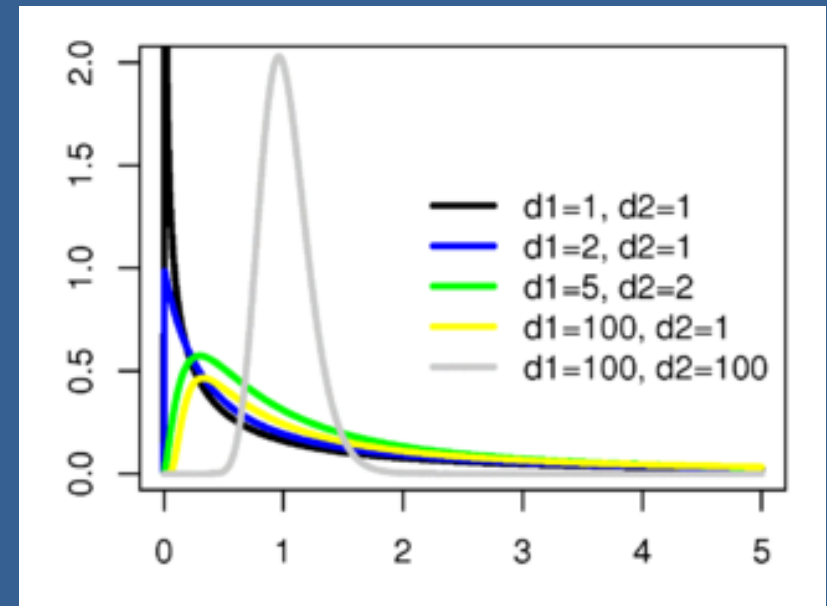
Donde:

- $U_1$  y  $U_2$  siguen una distribución chi-cuadrada con  $d_1$  y  $d_2$  grados de libertad respectivamente, y
- $U_1$  y  $U_2$  son estadísticamente independientes.
- La distribución  $F$  aparece frecuentemente como la *distribución nula* de una prueba estadística, especialmente en el análisis de varianza.

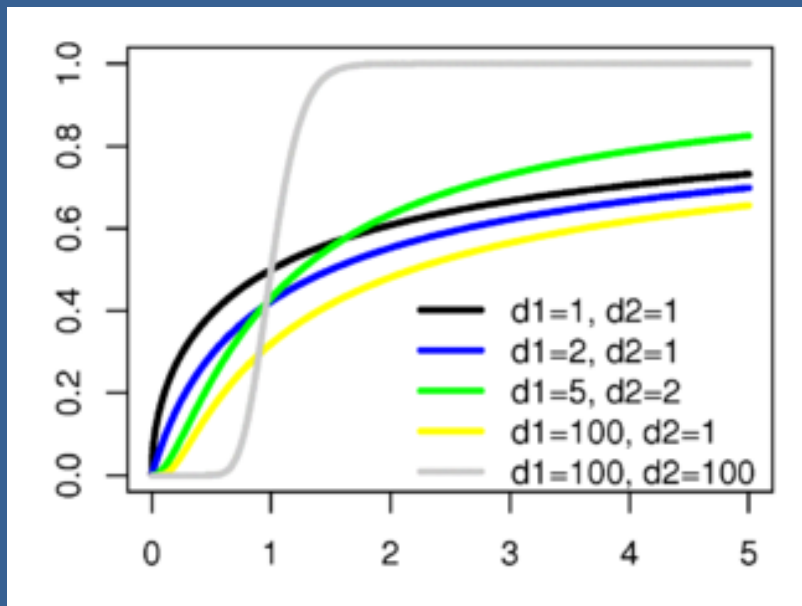
# Distribución F

La distribución  $F$  aparece frecuentemente como la *distribución nula* de una prueba estadística, especialmente en el análisis de varianza.

Función de densidad de probabilidad

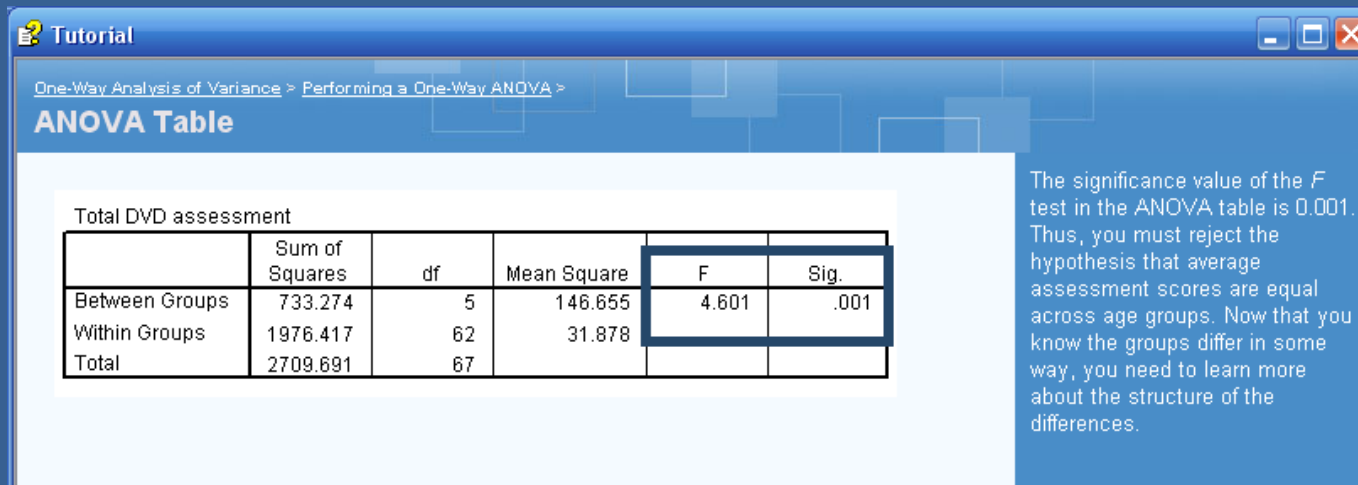


Función de distribución de probabilidad



# Prueba F de Fisher

- **Hipótesis nula en el test F:** las medias de múltiples poblaciones normalmente distribuidas y con la misma desviación estándar son iguales. Esta es, la más conocida de las hipótesis verificada mediante el test F y el problema más simple del análisis de varianza.
- Si el test F sale significativo, quiere decir que la hipótesis nula se refuta, por lo que las medias de las poblaciones son diferentes, por tanto existen diferencias significativas.



The screenshot shows a SPSS window titled 'Tutorial' with a breadcrumb trail: 'One-Way Analysis of Variance > Performing a One-Way ANOVA > ANOVA Table'. The main content is an ANOVA table for 'Total DVD assessment'. The table has columns for Sum of Squares, df, Mean Square, F, and Sig. The 'Between Groups' row shows a significant result with F = 4.601 and Sig. = .001. To the right of the table, a text box explains that the significance value of the F test is 0.001, leading to the rejection of the null hypothesis that average assessment scores are equal across age groups.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	733.274	5	146.655	4.601	.001
Within Groups	1976.417	62	31.878		
Total	2709.691	67			

The significance value of the  $F$  test in the ANOVA table is 0.001. Thus, you must reject the hypothesis that average assessment scores are equal across age groups. Now that you know the groups differ in some way, you need to learn more about the structure of the differences.

# Distribución chi-cuadrado

- Distribución de probabilidad continua (pero aplicable tanto a variables cuantitativas continuas como a variables cuantitativas discretas y tanto a variables con distribución paramétrica como a no paramétricas).
- Posee un parámetro  $k$  que representa los grados de libertad de la variable aleatoria:

$$X = Z_1^2 + \cdots + Z_k^2$$

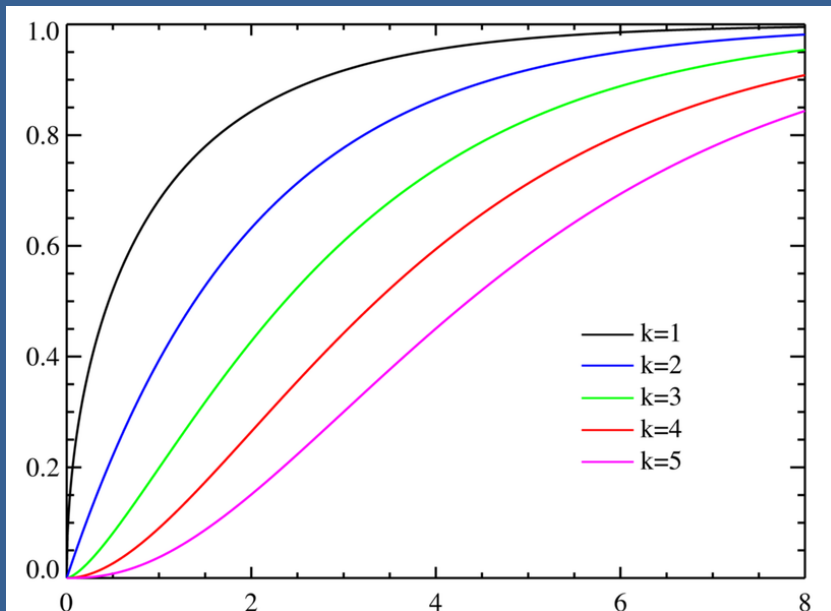
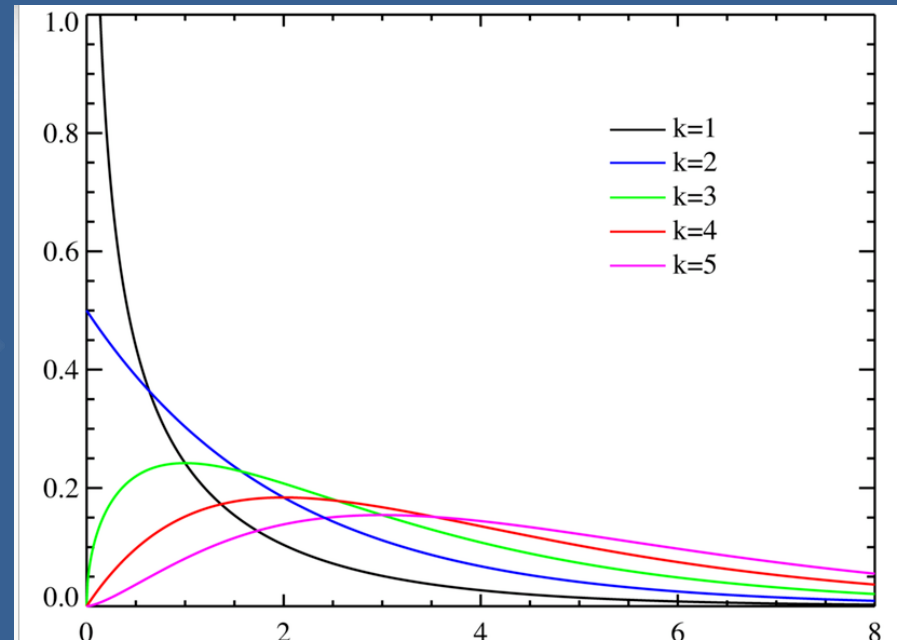
- Donde  $Z_i$  son variables de distribución normal, de media cero y varianza uno.

# Distribución chi-cuadrado

Función de probabilidad



Función de distribución de probabilidad



**Parámetros:**

**K > 0** grados de libertad

**Función de densidad:**

$$f(x; k) = \begin{cases} \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} & \text{para } x \geq 0, \\ 0 & \text{para } x < 0 \end{cases}$$

donde  $\Gamma$  es la función gamma, para  $n = k/2$ :

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

# Distribución chi-cuadrado

- La distribución  $\chi^2$  tiene muchas aplicaciones en inferencia estadística, por ejemplo en la denominada prueba  $\chi^2$  utilizada como prueba de independencia y como prueba de bondad de ajuste y en la estimación de varianzas.
- Cuando  $k$  es suficientemente grande, como consecuencia del teorema central del límite, puede aproximarse por una distribución normal.

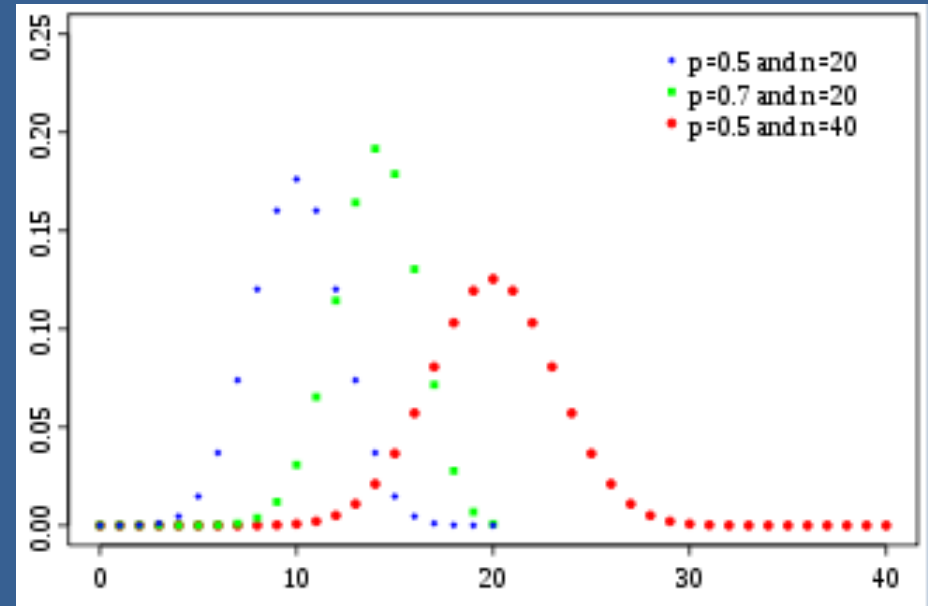
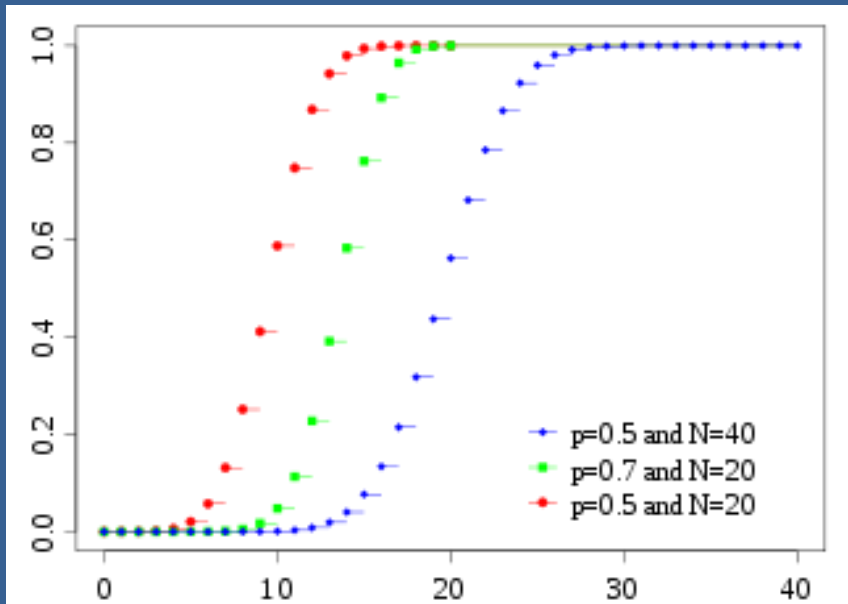
# DISTRIBUCIONES NO PARAMÉTRICAS

# Distribución binomial

Función de probabilidad



Función de distribución de probabilidad



**Parámetros:**

$n \geq 0$  número de ensayos (entero)

$0 \leq p \leq 1$  probabilidad de éxito (real)

# Distribución binomial

- Distribución de probabilidad discreta (aplicable a variables discretas) que mide el número de éxitos en una secuencia de  $n$  ensayos independientes de Bernoulli con una probabilidad fija  $p$  de ocurrencia del éxito entre los ensayos.
- Un **experimento de Bernoulli** se caracteriza por ser dicotómico, esto es, sólo son posibles dos resultados. A uno de estos se denomina éxito y tiene una probabilidad de ocurrencia  $p$  y al otro, fracaso, con una probabilidad  $q = 1 - p$ . En la distribución binomial el anterior experimento se repite  $n$  veces, de forma independiente, y se trata de calcular la probabilidad de un determinado número de éxitos.

# Distribución binomial

$$p(X = k) = \binom{n}{k} p^k \cdot q^{n-k}$$

**n** es el número de pruebas.

**k** es el número de éxitos.

**p** es la probabilidad de éxito.

**q** es la probabilidad de fracaso.

El número combinatorio  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

## Ejemplos:

- Se lanza un dado diez veces y se cuenta el número de treses obtenidos:  $X \sim B(10, 1/6)$
- Se lanza una moneda dos veces y se cuenta el número de caras obtenidas.

# Distribución binomial

## Ejemplos:

- Se lanza un dado diez veces y se cuenta el número de treses obtenidos:  $X \sim B(10, 1/6)$ . ¿Cuál es la probabilidad de obtener 2 treses?  $N=10$   $p=1/6$   $q=5/6$   $k=2$
- Se lanza una moneda 10 veces y se cuenta el número de caras obtenidas. ¿Cuál es la probabilidad de obtener 4 caras?  $N=10$   $p=1/2$   $q=1/2$   $k=4$
- Un examen consta de 10 preguntas a las que hay que contestar SI o NO. Suponiendo que a las personas a las que se le aplica no saben contestar a ninguna de las preguntas y, en consecuencia contestan al azar hallar la probabilidad de obtener cinco aciertos.  $N=10$   $p=1/2$   $q=1/2$   $k=5$

Es una distribución binomial, la persona sólo puede acertar o fallar la pregunta.

Suceso A (éxito) = acertar la pregunta  $\Rightarrow p = p(A) = 0,5$

Suceso  $\bar{A}$  = no acertar la pregunta  $\Rightarrow q = p(\bar{A}) = 0,5$

Distribución binomial de parámetros  $n = 10, p = 0,5 \Rightarrow \mathbf{B(10; 0,5)}$

### a) Probabilidad de obtener cinco aciertos

Obtener exactamente cinco aciertos  $k = 5$ , aplicamos la fórmula:

$$P(x=k) = \binom{n}{k} \cdot p^k \cdot q^{n-k} \Rightarrow \begin{matrix} k=5 \\ n=10 \\ p=0,5 \\ q=0,5 \end{matrix} \Rightarrow P(x=5) = \binom{10}{5} \cdot (0,5)^5 \cdot (0,5)^{10-5}$$

$$\binom{n}{k} = \frac{n!}{k! (n-k)!} \quad \text{Números combinatorios} \Rightarrow \binom{10}{5} = \frac{10!}{5!(10-5)!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot \cancel{5!}}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot \cancel{5!}} = 252$$

$$P(x=5) = \binom{10}{5} \cdot (0,5)^5 \cdot (0,5)^{10-5} \Rightarrow P(x=5) = 252 \cdot (0,5)^5 \cdot (0,5)^5 = 0,2461$$

# Distribución de Poisson

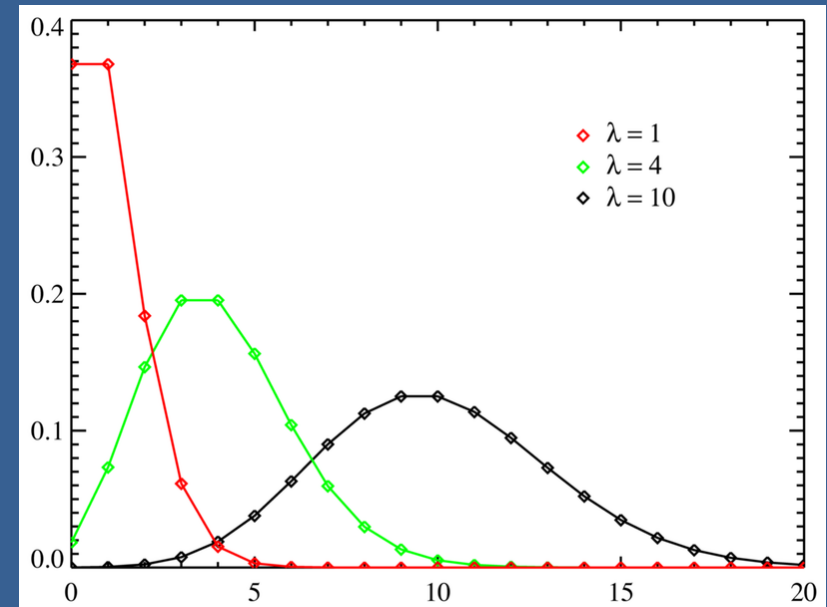
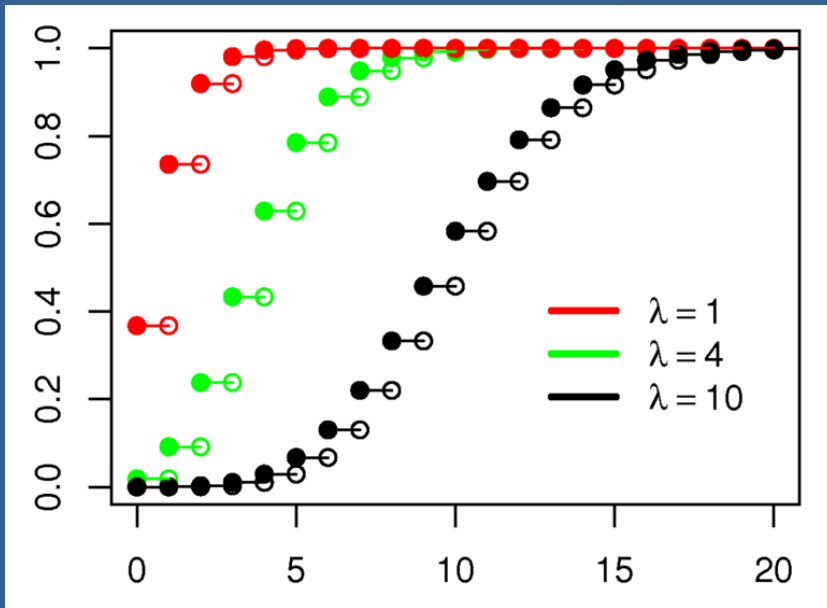
- La **distribución de Poisson** es una distribución de probabilidad discreta. Expresa la probabilidad de un número  $k$  de eventos ocurriendo en un tiempo fijo si estos eventos ocurren con una frecuencia media conocida y son independientes del tiempo discurrido desde el último evento.
- Fue descubierta por Siméon-Denis Poisson, que la dio a conocer en 1838 en su trabajo *Recherches sur la probabilité des jugements en matières criminelles et matière civile* (Investigación sobre la probabilidad de los juicios en materias criminales y civiles).

# Distribución de Poisson

Función de probabilidad



Función de distribución de probabilidad



Parámetros:

$K \geq 0$  número de eventos (entero)

$0 \leq p \leq 1$  probabilidad de éxito (real)

El eje horizontal es el índice  $k$ .  
La función solamente está definida en valores enteros de  $k$ .

# Distribución de Poisson

Función de densidad

$$f(k; \lambda) = \frac{e^{-\lambda} \lambda^k}{k!}$$

Ejemplos:

- Si el 2% de los libros encuadernados en cierto taller tiene encuadernación defectuosa, obtener la probabilidad de que 5 de 400 libros encuadernados en este taller tengan encuadernaciones defectuosas puede calcularse usando la distribución de Poisson. En este caso concreto,  $k$  es 5 y  $\lambda$  (lambda), el valor esperado de libros defectuosos es el 2% de 400, es decir, 8. Por lo tanto, la probabilidad deseada es

$$P(5; 8) = \frac{8^5 e^{-8}}{5!} = 0,092.$$

- Este problema también podría resolverse recurriendo a una distribución binomial de parámetros  $k = 5$ ,  $n = 400$ ,  $p = 0,02$ ,  $q = 0,98$ .

# Distribución de Poisson

La distribución de Poisson se aplica a varios fenómenos discretos de la naturaleza (esto es, aquellos fenómenos que ocurren 0, 1, 2, 3, ... veces durante un periodo definido de tiempo o en un área determinada) cuando la probabilidad de ocurrencia del fenómeno es constante en el tiempo o el espacio. Ejemplos de estos eventos que pueden ser modelados por la distribución de Poisson incluyen:

- El número de vehículos que pasan a través de un cierto punto en una ruta (suficientemente distantes de los semáforos) durante un periodo definido de tiempo.
- El número de errores de ortografía que uno comete al escribir una única página.
- El número de llamadas telefónicas en una central telefónica por minuto.
- El número de animales muertos encontrados por unidad de longitud de ruta.
- El número de mutaciones de determinada cadena de ADN después de cierta cantidad de radiación.
- El número de núcleos atómicos inestables que decayeron en un determinado período en una porción de sustancia radiactiva. La radiactividad de la sustancia se debilitará con el tiempo, por lo tanto el tiempo total del intervalo usado en el modelo debe ser significativamente menor que la vida media de la sustancia.
- La distribución de receptores visuales en la retina del ojo humano.

# REPRESENTACIONES GRÁFICAS DE UNA MUESTRA

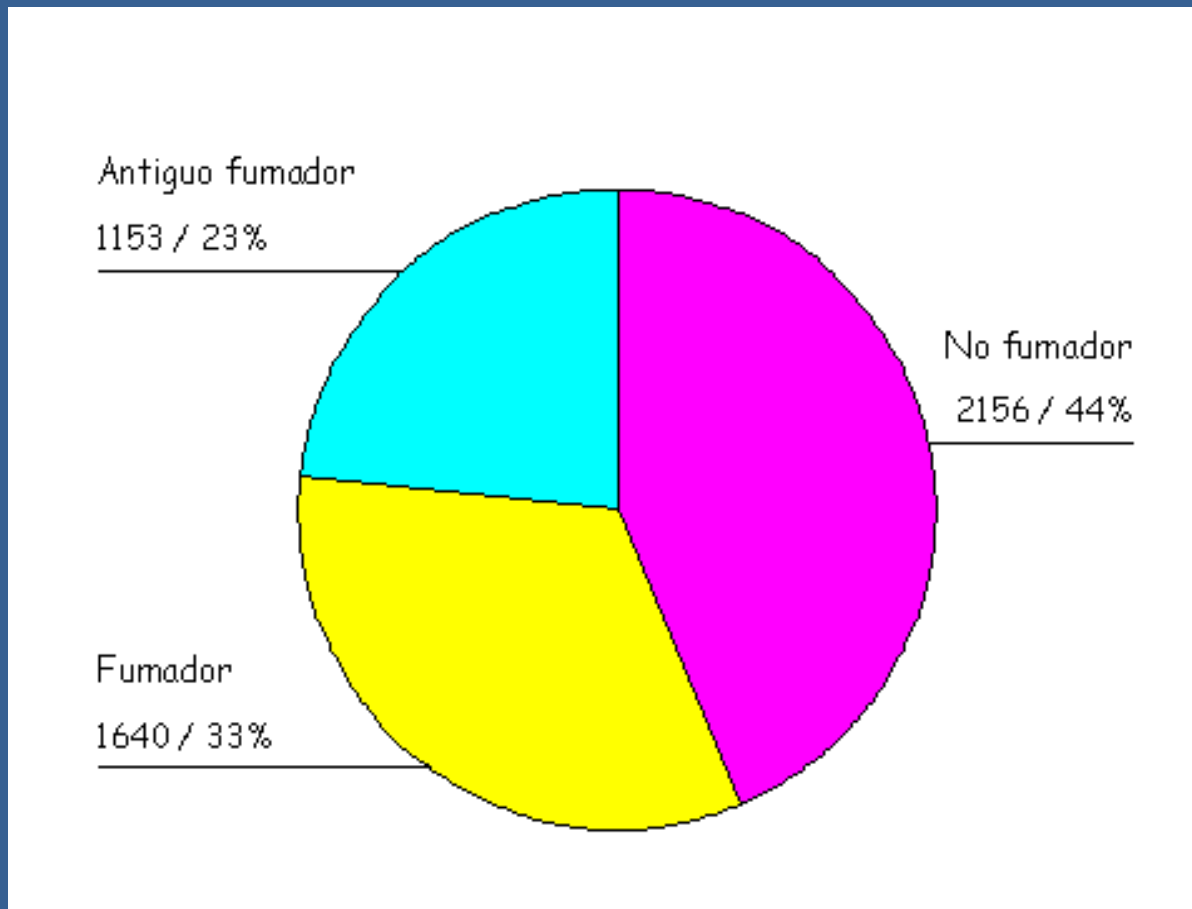
# Gráficos de sectores

También conocidos como diagramas de "tartas", se divide un círculo en tantas porciones como clases tenga la variable, de modo que a cada clase le corresponde un arco de círculo proporcional a su frecuencia absoluta o relativa. Este gráfico es interesante para variables cualitativas discretas.

# Gráficos de sectores

Ejemplo de gráfico de sectores:

Distribución de una muestra de pacientes según el hábito de fumar.

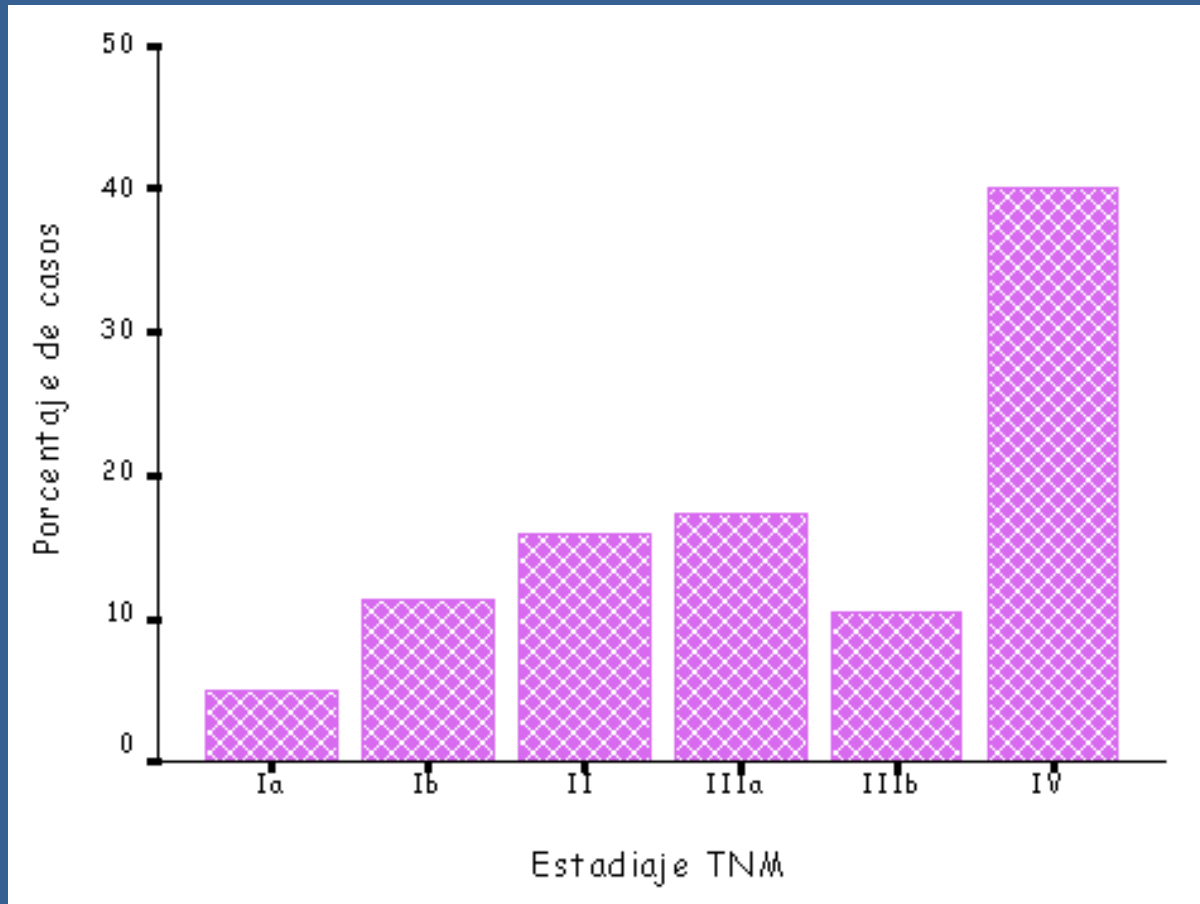


# Diagramas de barras

Son similares a los gráficos de sectores. Se representan tantas barras como categorías tiene la variable, de modo que la altura de cada una de ellas sea proporcional a la frecuencia o porcentaje de casos en cada clase. Suele utilizarse para variables discretas.

# Diagramas de barras

Ejemplo de diagrama de barras:  
Estadio TNM en el cáncer gástrico.



# Histogramas

Es una representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados.

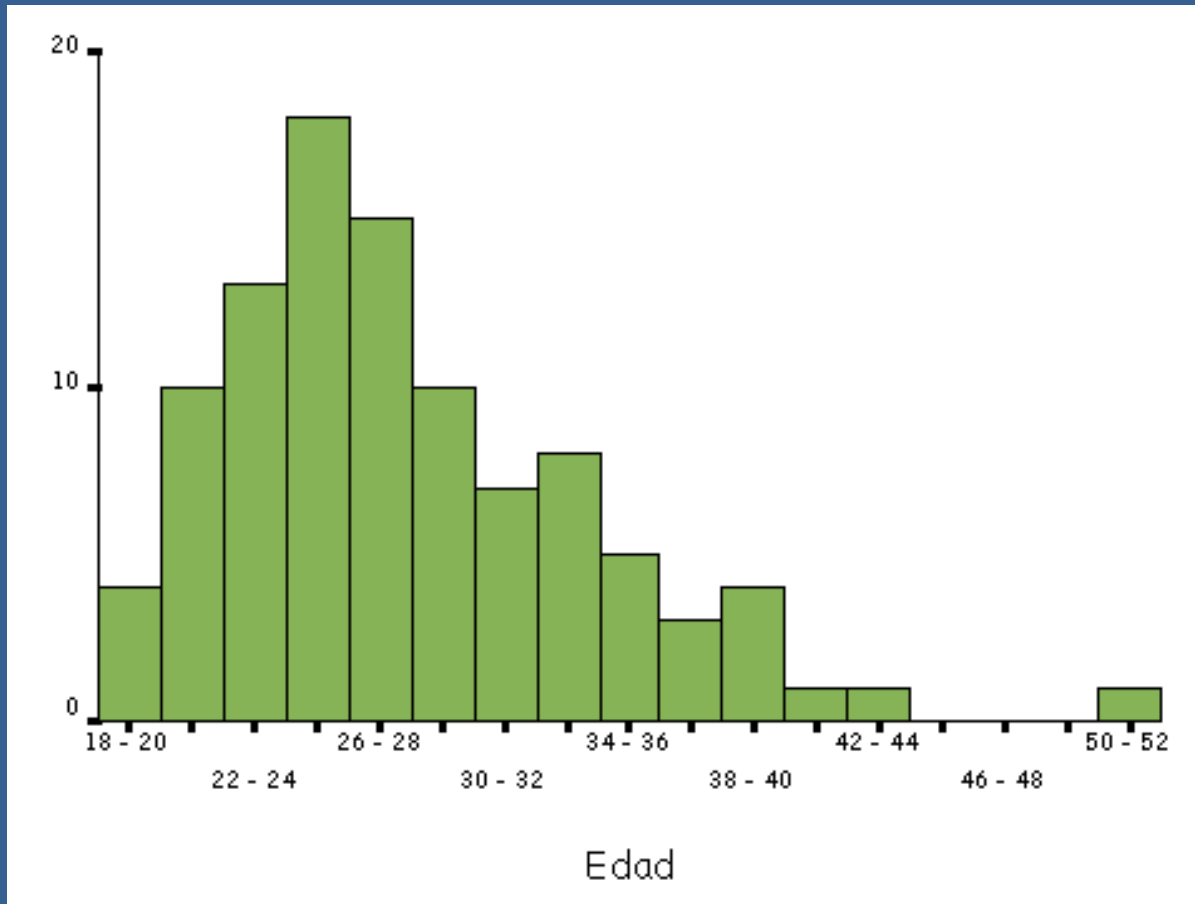
Eje vertical: frecuencias.

Eje horizontal: valores de las variables, normalmente señalando las marcas de clase, es decir, la mitad del intervalo en el que están agrupados los datos.

Se utiliza cuando se estudian ***variables numéricas continuas***, tales como la edad, la tensión arterial o el índice de masa corporal, peso o altura, y, por comodidad, sus valores se agrupan en clases.

# Histogramas

Distribución de frecuencias de la edad en 100 pacientes.



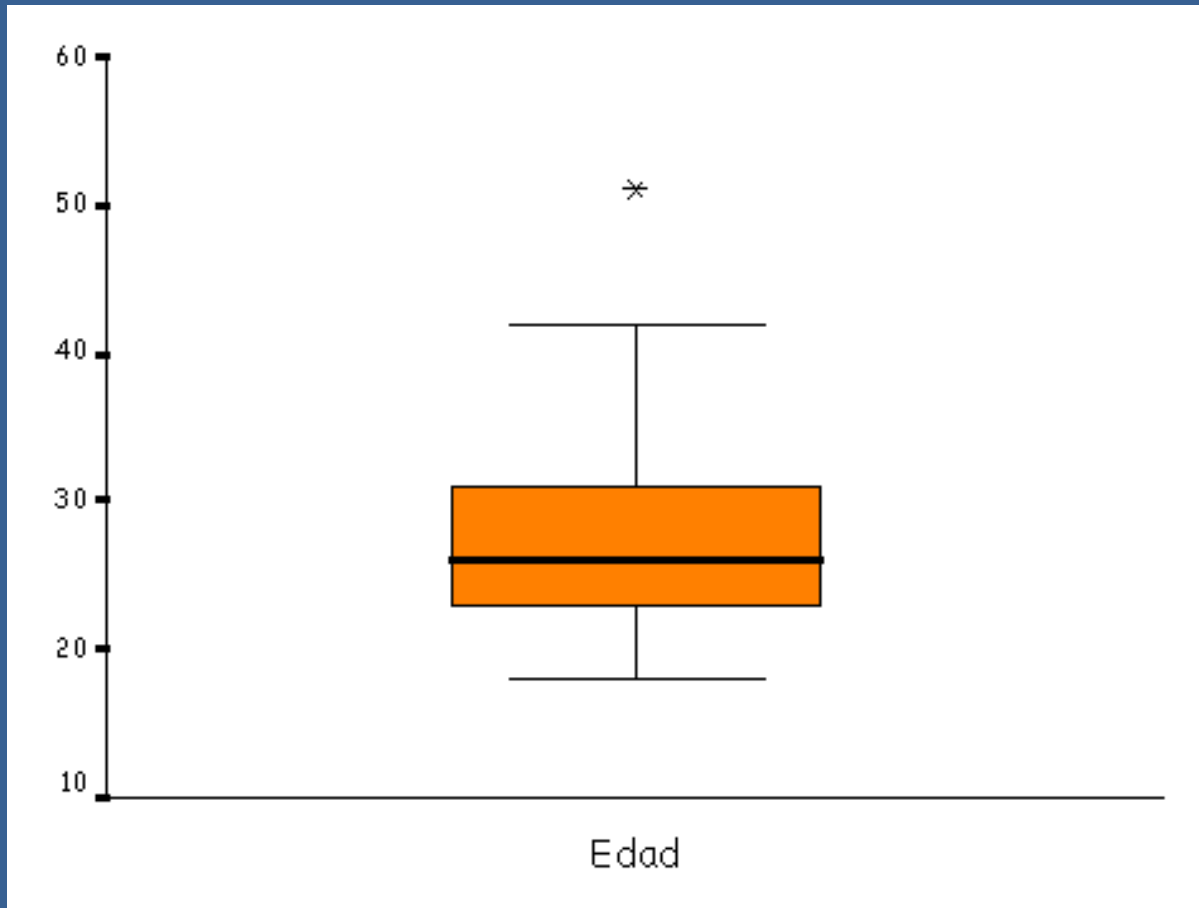
Edad	Nº de pacientes
18	1
19	3
20	4
21	7
22	5
23	8
24	10
25	8
26	9
27	6
28	6
29	4
30	3
31	4
32	5
33	3
34	2
35	3
36	1
37	2
38	3
39	1
41	1
42	1

# Diagramas de cajas

Otro modo habitual, y muy útil, de resumir una variable de tipo numérico es utilizando el concepto de percentiles, mediante el **diagramas de cajas**. La figura **con la caja naranja** muestra un gráfico de cajas correspondiente a los datos de la tabla siguiente. La caja central indica el rango en el que se concentra el 50% central de los datos. Sus extremos son, por lo tanto, el 1<sup>er</sup> y 3<sup>er</sup> cuartil de la distribución. La línea central en la caja es la mediana. De este modo, si la variable es simétrica, dicha línea se encontrará en el centro de la caja. Los extremos de los "bigotes" que salen de la caja son los valores que delimitan el 95% central de los datos, aunque en ocasiones coinciden con los valores extremos de la distribución. Se suelen también representar aquellas observaciones que caen fuera de este rango (outliers o valores extremos). Esto resulta especialmente útil para comprobar, gráficamente, posibles errores en nuestros datos.

# Diagramas de cajas

Distribución de frecuencias de la edad en 100 pacientes.

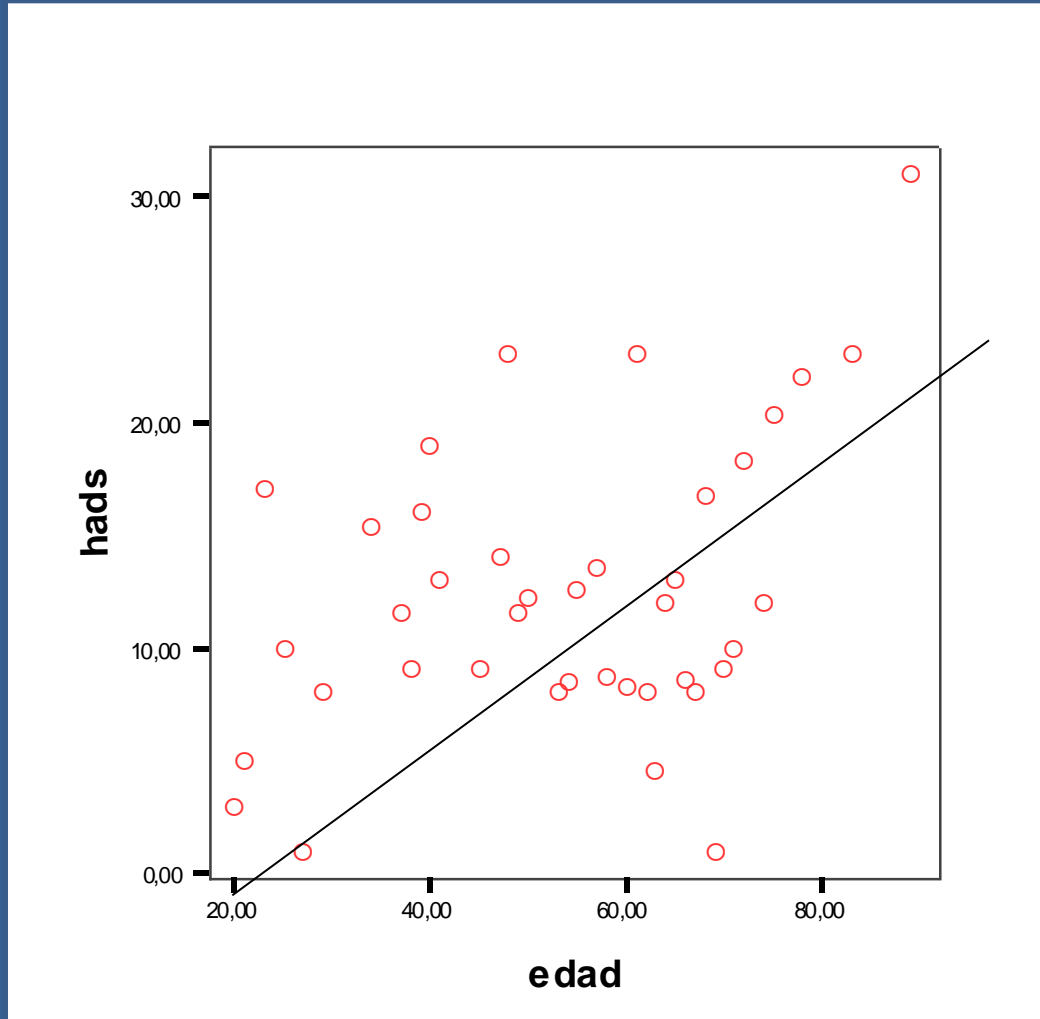


Edad	Nº de pacientes
18	1
19	3
20	4
21	7
22	5
23	8
24	10
25	8
26	9
27	6
28	6
29	4
30	3
31	4
32	5
33	3
34	2
35	3
36	1
37	2
38	3
39	1
41	1
42	1

# Gráficos de dispersión

son el resultado de estudiar la relación entre dos variables. Se sitúa una de ellas en el eje X y la otra en el eje Y, situando un punto en cada ocurrencia. De esta manera se puede observar gráficamente si existe algún tipo de relación entre las variables, en el caso que los puntos se agrupen de una forma determinada; o descartar cualquier relación entre las dos variables en el caso que la nube de puntos tienda a distribuirse de manera uniforme por todo el gráfico.

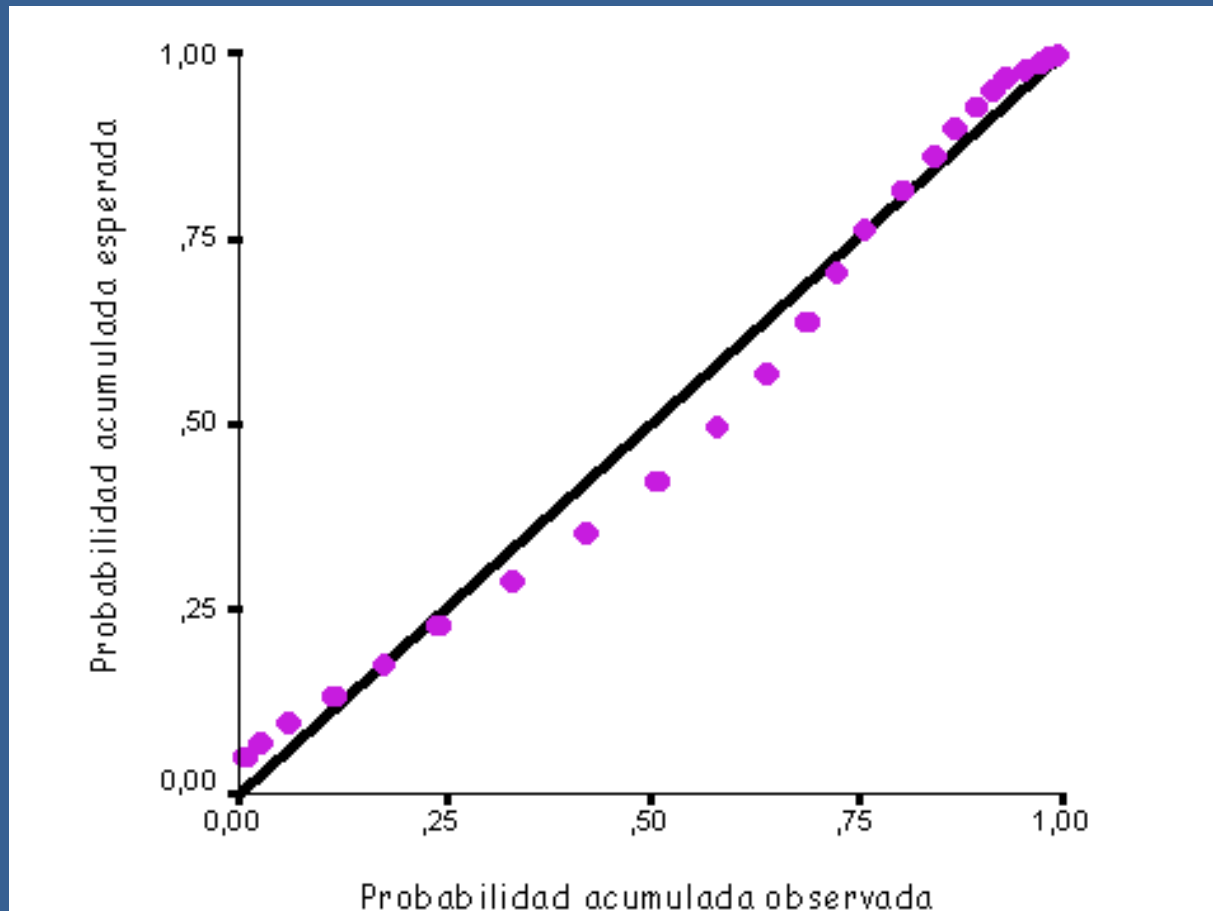
# Gráficos de dispersión



# Gráficos P-P o Q-Q

Por último, y en lo que respecta a la descripción de los datos, suele ser necesario, para posteriores análisis, comprobar la normalidad de alguna de las variables numéricas de las que se dispone. Un diagrama de cajas o un histograma son gráficos sencillos que permiten comprobar, de un modo puramente visual, la **simetría** y el **apuntamiento** de la distribución de una variable y, por lo tanto, valorar su desviación de la normalidad. Existen otros métodos gráficos específicos para este propósito, como son los **gráficos P-P o Q-Q**. En los primeros, se confrontan las proporciones acumuladas de una variable con las de una distribución normal. Si la variable seleccionada coincide con la distribución de prueba, los puntos se concentran en torno a una línea recta.

# Gráficos P-P o Q-Q



# DISEÑO DE EXPERIMENTOS

# EXPERIMENTO

Creación y preparación de lotes de prueba (unidades de muestreo) que verifiquen la validez de las hipótesis establecidas sobre las causas de un determinado problema o defecto, objeto de estudio.

# DISEÑO DE EXPERIMENTOS

Metodología estadística destinada a la planificación y análisis del experimento.

# SUJETO O UNIDAD EXPERIMENTAL

El sujeto o unidad experimental es la unidad básica sobre la que se efectúa el proceso de medida.

Ejemplo: El contenido de azúcar en el zumo de naranja producido se medirá recogiendo cada hora una unidad experimental de 1 litro de zumo.

# OBSERVACIÓN

Una observación es una toma de medida de una variable y consta entonces de un valor de la misma. Dependiendo del tipo de Diseño, las observaciones pueden tomarse a diferentes sujetos o al mismo sujeto de manera secuencial.

# REPETICIÓN

Reiteración de una observación o medida al mismo nivel de tratamiento.

Proporciona una oportunidad para que los efectos de las variables extrañas, incontroladas se compensen y permite, además, medir el error experimental.

# ALEATORIZACIÓN

Técnica utilizada para reducir la influencia no predeterminable de variables extrañas sobre los resultados del Experimento

La aleatorización consiste en asignar los sujetos a los distintos niveles de tratamiento al azar, con la esperanza de que los efectos extraños se contrarresten entre los distintos sujetos y observaciones que componen cada nivel de tratamiento (condición experimental). Que la toma de muestras sea al azar.

# ALEATORIZACIÓN

Fundamental en el Diseño de Experimentos

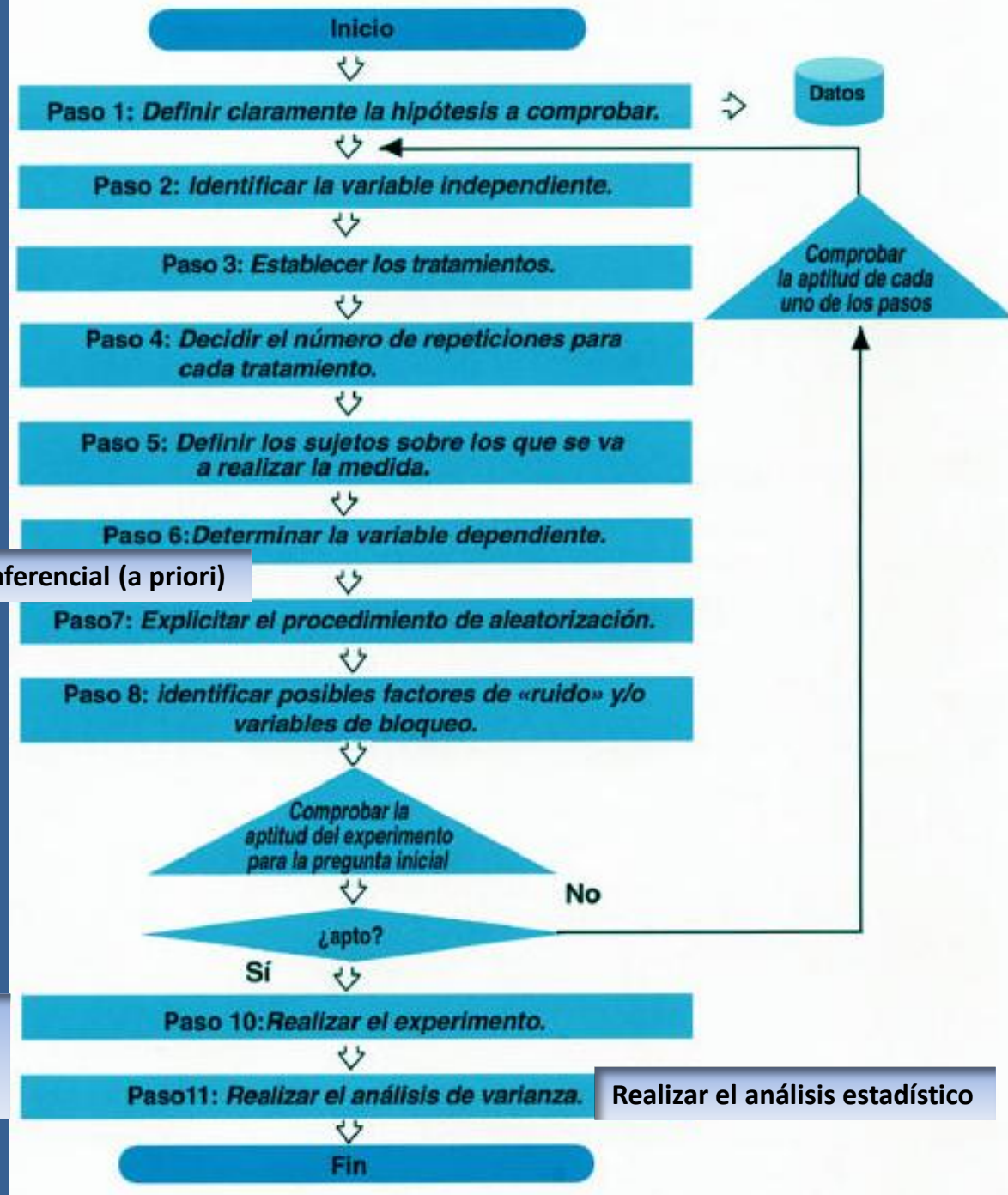
- a) Previene la existencia de sesgo.
- b) Evita la dependencia entre observaciones.
- c) Confirma la adecuación de los procedimientos estadísticos para el análisis de los resultados del Experimento.

# PROCESO

Fuente:  
Fundación Iberoamericana  
para la Gestión de la Calidad.

Entre 6 y 7. Nuevo paso: Elección del método inferencial (a priori)

Verificar la aptitud de los datos para el análisis predefinido y en su caso replanificar el método inferencial (análisis estadístico)



Realizar el análisis estadístico

# Proceso

## *Paso 1: Definir claramente la hipótesis a comprobar*

Es de importancia fundamental identificar de forma muy específica el objetivo del Experimento, es decir, la pregunta exacta que se quiere contestar o la hipótesis que se necesita contrastar.

Esta pregunta básica se formulará por escrito.

# Proceso

## *Paso 2: Identificar la variable independiente*

La variable independiente representa la característica que, suponemos, influye sobre los valores de la variable dependiente.

Puesto que, para la realización del Experimento, se le asignarán diferentes valores, hay que asegurarse que esté en nuestro poder manipularla.

# Proceso

## *Paso 3: Establecer los tratamientos*

En base a la naturaleza de la variable, las condiciones reales del proceso o situación y la pregunta específica que se quiere contestar, se identificarán los valores o el recorrido de valores de la variable independiente, relevantes para el Experimento y se establecerán los tratamientos a efectuar.

# Proceso

## *Paso 4: Decidir el número de repeticiones para cada tratamiento o el tamaño muestral*

Es absolutamente aconsejable realizar varias observaciones para cada nivel de tratamiento (condición experimental), en caso de experimentos de repeticiones, o bien tener un número de muestras suficiente para los requisitos del análisis estadístico que queramos realizar, para que los errores de medida e influencias no controladas de variables extrañas puedan contrarrestarse entre sí.

# Proceso

## *Paso 4: Decidir el número de repeticiones para cada tratamiento o el tamaño muestral*

### EJEMPLO DESDE INICIO

Deseamos realizar un estudio de la ansiedad en los pacientes de cáncer de colon. La idea que tenemos en la cabeza es que una determinada terapia psicológica disminuye el grado de ansiedad del paciente. Queremos testar si es realmente efectiva para este fin.

Hipótesis: Hay diferencias en el grado de ansiedad entre los pacientes tratados y los que no lo han sido.

Nuestra variable independiente es la ocurrencia o no de esta asistencia psicológica al paciente, hablamos de una variable de Bernuilli (binomial).

Para lo cual vamos a distinguir dos tipos de pacientes, los que han utilizado esta terapia y los que no. Tenemos 2 “tratamientos”.

Para que la muestra sea suficientemente representativa, hemos consultado la bibliografía y otros estudios similares y hemos concluido un número de 50 pacientes para cada “tratamiento”.

# Proceso

***Paso 5: Definir los sujetos sobre los que se va a realizar la medida***

En el ejemplo, los pacientes de los hospitales de la Comunidad de Madrid

# Proceso

## *Paso 6: Determinar la variable dependiente*

- Sólo puede existir una única variable dependiente. Esta deberá tener, necesariamente, un nivel de medida continuo, o lo más próximo a ese extremo que sea posible. Cuantas más posibilidades de apreciar diferencias entre distintas observaciones ofrezca la variable dependiente, más se favorecerá la sensibilidad de la misma a los distintos tratamientos.
- Aconsejable que esta variable sea de carácter continua y de distribución paramétrica.
- Métodos estadísticos paramétricos suelen tener una mayor consistencia.
- Siempre existe la posibilidad de aplicar métodos no paramétricos.

# Proceso

## ***Paso 6: Determinar la variable dependiente***

En nuestro ejemplo, ya que no existe ningún parámetro continuo medible para la ansiedad, hemos consultado la bibliografía y hemos encontrado una metodología de entrevistas a pacientes que da como resultado una valoración de 1 a 100 del nivel de estrés del paciente.

# Proceso

## *Paso 7: Elección del método inferencial*

- Elegir cual es el método estadístico que mejor se adapta a las condiciones de nuestro estudio, **razonando convenientemente los motivos de esa elección.**
- **Los principales factores para esta elección son:**
  - Las características de las variables (**dependiente e independiente**): **Si son** continuas o discretas, **si son** paramétricas o no.
  - El objetivo del estudio, **de la hipótesis de partida, en relación a si lo que buscamos son** diferencias entre grupos o relaciones entre variables.

**Más adelante veremos los distintos tratamientos estadísticos que se pueden aplicar.**

# Proceso

## *Paso 7: Elección del método inferencial*

En nuestro ejemplo, vamos a estudiar las diferencias entre dos poblaciones. La variable independiente la utilizamos como una “categoría” para dividir las dos poblaciones. La variable dependiente (valoración de 0 a 100 del estrés) se puede asumir como variable cuantitativa continua y a priori pensamos que pueda comportarse como una variable normal. Para asegurarnos de ello, cuando tengamos los datos a posteriori, antes de dividirlos por estratos, utilizaremos la estadística descriptiva para analizarlos y les someteremos a test de normalidad con nuestro programa de análisis estadístico. En caso que nuestra muestra supere estos test, vamos a utilizar el método del Análisis de Varianza (ANOVA) para estudiar las diferencias entre las poblaciones. En caso contrario utilizaremos un test de Kruskal-Wallis.

# Proceso

## *Paso 8: Explicitar el procedimiento de aleatorización*

Esta es una parte muy importante del Diseño, ya que asegurará que las diferencias que se encuentren entre los tratamientos son debidas a ellos mismos y no a efectos laterales no deseados. Para ello sobre la población objeto de estudio, es necesario que los grupos que constituyan la muestra hayan sido elegidos al azar.

# Proceso

## *Paso 8: Explicitar el procedimiento de aleatorización*

En nuestro ejemplo, hemos consultado en cada uno de los hospitales de la Comunidad de Madrid, cuales de ellos utilizan la técnica que queremos estudiar. Esto nos ha dado dos estratos: En el primer estrato están todos los hospitales que no utilizan la técnica, que son 15; en el segundo estrato aparecen los hospitales que utilizan la técnica, que son únicamente 2.

Bien en cada uno de los dos estratos queremos “sacar” una muestra de 50 pacientes. La primera pregunta, que suele ser bastante frecuente en muchos tratamientos “novedosos” es ¿Hay más de 50 pacientes en los 2 hospitales del segundo estrato? Si no es así habrá que reequilibrar las muestras. Supongamos que tenemos más de 50.

# Proceso

## *Paso 8: Explicitar el procedimiento de aleatorización*

Ahora hay que utilizar algún procedimiento para aleatorizar la elección de los pacientes de cada estrato. Una técnica interesante puede ser realizar papeletas o meter registros en una tabla con los posibles sujetos. Por ejemplo, como vamos a trabajar con pacientes hospitalizados, vamos a utilizar un código con el nombre del hospital y el número de habitación. Realizaremos dos tablas, una por estrato (hemos elegido un muestreo estratificado). En la primera incluiremos todos los registros de sujetos que no se han sometido al tratamiento. En la segunda todos los que si se han sometido al tratamiento. Ya tenemos las dos poblaciones.

Ahora se trata de escoger de cada una 50 sujetos de forma aleatoria. Esto lo podemos hacer con muchos programas estadísticos a partir de cada tabla, generando números aleatorios que seleccionarán individuos. También se puede hacer a la antigua usanza (de forma democrática), realizando papeletas, metiéndolas en una saca y cogiéndolas al azar.

# Proceso

## *Paso 9: Identificar posibles factores de "ruido" y/o variables de bloqueo*

Analizar la futura situación experimental e identificar los factores que puedan, además de la variable independiente, influir sobre los valores de la variable dependiente.

# Proceso

## *Paso 9: Identificar posibles factores de "ruido" y/o variables de bloqueo*

Con estos factores hay varias estrategias posibles:

- Tenerlos bajo control (constantes), a lo largo de todas las observaciones.
- Integrarlos en el Diseño, como variable de bloqueo.
- Transformarlos en una variable independiente. Esto será necesario, aunque se complique notablemente el Diseño de Experimento desde el punto de vista estadístico, cuando su influencia sobre la variable dependiente resulte ser relevante.
- O bien esperar que la aleatorización sea suficiente para que sus efectos se contrarresten en las repeticiones de cada tratamiento. Esta posibilidad será aceptable sólo si la variable en cuestión está fuera de nuestro control y se considera que su influencia es bastante limitada. Su efecto se englobará dentro del "error experimental", o "ruido".

# Proceso

## ***Paso 9: Identificar posibles factores de "ruido" y/o variables de bloqueo***

En nuestro ejemplo, un posible factor de bloqueo puede ser el grado de avance del cáncer. Parece razonable pensar que una persona con un estadio muy avanzado, tenga más ansiedad que otro.

Nosotros consideramos que esta variable está fuera de nuestro control y confiamos que al elegir un tamaño muestral grande y un procedimiento de aleatorización alto, el "ruido" se diluya, es decir, sea homogéneo en los distintos estratos.

Otra alternativa sería incorporar esta variable dependiente a nuestro estudio. Si optáramos por esta alternativa, podríamos, por ejemplo trabajar con cuatro estratos en lugar de dos: sujetos sometidos al tratamiento y con cáncer avanzado, sujetos sometidos al tratamiento y con cáncer incipiente; sujetos no sometidos al tratamiento y con cáncer avanzado y, por último, sujetos no sometidos al tratamiento y con cáncer incipiente. En este caso, tendríamos que volver a revisar la planificación desde el principio, por ejemplo, aumentando el tamaño muestral total para que a cada estrato le correspondan 50 sujetos.

# Proceso

## *Paso 10: Asegurarse de la aptitud del Diseño del Experimento para contestar la pregunta inicial*

Comprobar que el tipo de resultados que obtendremos del Experimento tal y como lo hemos planificado, nos proporcionará efectivamente la información que necesitamos.

En nuestro ejemplo, la planificación realizada hasta el momento sí nos permite contestar la hipótesis de partida: Hay diferencias en el grado de ansiedad entre los pacientes que han sido tratados por la terapia psicológica y los que no lo han sido.

# Proceso

## *Paso 11: Realización del Experimento*

Se crearán las condiciones experimentales (tratamientos) y se efectuarán las observaciones según el plan establecido, teniendo un cuidado particular en evitar posibles influencias extrañas sobre los valores de la variable dependiente.

# Proceso

## *Paso 11: Realización del Experimento*

En nuestro ejemplo, escogemos a un equipo de diez técnicos de muestreo que se desplacen a los distintos hospitales y, con el consentimiento de los pacientes realicen la encuesta planificada. Es importante, para que no haya influencias extrañas sobre la variable dependiente (grado de estrés), que el estado en el que el paciente realice la entrevista sea el normal, no debe estar sobreescitado, ni sobredeprimido.

# Proceso

- Con la realización iterativa y secuencial de estos pasos nos aseguraremos de que estamos aplicando correctamente el método.
- En el caso de que haya cambios en algún paso, será necesario volver a revisar secuencialmente todos los pasos anteriores.

# ESTADÍSTICA INFERENCIAL

# Estadística inferencial

- Efectúa estimaciones, decisiones, predicciones u otras generalizaciones sobre un conjunto mayor de datos.
- Sobre el pilar construido en base a un buen diseño de experimentos, de cuya disciplina hablaremos adelante, de una buena muestra y de la estadística descriptiva, podremos aplicar los test estadísticos que nos permitan “inferir” conclusiones.
- Queremos conseguir conclusiones sobre nuestras muestras que puedan ser extrapolables a toda la población.

# Estadística inferencial

Los siete factores para la elección (BIBLIA DE LA ESTADISTICA INFERENCIAL):

- 1) Si estamos trabajando con **variables continuas o discretas**.
- 2) Si estamos trabajando con **variables cuantitativas o cualitativas (categóricas)**.
- 3) Si estamos ante variables con distribución **paramétrica** (continua cuantitativa, que además tiene una distribución normal) **o no paramétrica**.
- 4) Si buscamos **diferencias entre muestras** o estratos o por el contrario estamos buscando **relaciones entre variables**.

# Estadística inferencial

- 5) Si tenemos una sola muestra de sujetos, que queremos comparar con un valor teórico o con el valor poblacional. O si tenemos dos muestras de sujetos a comparar entre sí. O por último, si tenemos más de dos muestras de sujetos a comparar entre sí.
- 6) Si las **muestras**, o estratos son **independientes** o están **relacionadas** entre sí. Si dos o más muestras están formadas, al menos parcialmente por los mismos sujetos hablaremos de muestras relacionadas.
- 7) Las pruebas no paramétricas son menos exigentes y por lo tanto, también se pueden aplicar en caso de necesidad a las variables paramétricas. Pero hay que tener en cuenta que las pruebas no paramétricas tienen menos potencia relativa (suelen acertar algo menos). Esto quiere decir que si quieres la misma potencia que una prueba paramétrica, tendrás que aumentar el tamaño muestral.

Pasemos a estudiar las principales pruebas estadísticas que se emplean en Bioestadística:

# PRUEBAS ESTADÍSTICAS

# Comparaciones entre muestras

## Tipos de test estadísticos para realizar inferencias sobre comparaciones entre muestras

Distribución	Número de muestras *	Variable dependiente	Relación entre las muestras	Prueba Estadística
Paramétrica	Es una sola muestra (Se compara contra un valor teórico)	Cuantitativa	No relacionadas	t-student para una muestra
	Dicotómica	Categórica	No relacionadas	No existe, pero se puede usar Chi-cuadrado de Pearson.
		Cuantitativa	Relacionadas	No existe. Hay que usar no paramétricas.
			No relacionada	t-student para muestras independientes. ANOVA
	Policotómica (cualitativa no dicotómica)	Categórica	Relacionadas	t-student para muestras relacionadas.
			No relacionadas	No existe, pero se puede usar Chi-cuadrado.
		Cuantitativa	Relacionadas	ANOVA de una vía.
			No relacionadas	ANOVA de medidas repetidas.

\* En el caso de dos o más muestras, esta columna se podría llamar también la de la variable dependiente, que en el caso de la comparación de muestras se identifica con el factor de “clusterización”, que divide las muestras. Suele ser una variable discreta y cualitativa.

## Tipos de test estadísticos para realizar inferencias sobre comparaciones entre muestras

Distribución	Número de muestras *	Variable dependiente	Relación entre las muestras	Prueba Estadística
No paramétrica	Es una sola muestra (Se compara contra un valor teórico)	Categórica	Relacionadas	Binomial
			No relacionadas	Chi-cuadrado de Pearson
		Cuantitativa	Relacionadas	Chi-cuadrado de Mantel-Haenzel
			No relacionadas	Kolmogorow-Smirnov
	Dicotómica	Categórica	Relacionadas	Test exacto de McNemar
			No relacionadas	Test Exacto de Fisher Chi-cuadrado de Pearson
		Cuantitativa	Relacionadas	Test de Wilcoxon
			No relacionadas	Mann-Whitney Kolmogorow-Smirnov Valores extremos de Moses
	Policotómica	Categórica	No relacionadas	Prueba Q de Cochran
		Cuantitativa	Relacionadas	Friedman W de Kendall (concordancia)
			No relacionadas	Kruskal-Wallis Mediana de K variables.

\* En el caso de dos o más muestras, esta columna se podría llamar también la de la variable dependiente, que en el caso de la comparación de muestras se identifica con el factor de “clusterización”, que divide las muestras. Suele ser una variable discreta y cualitativa.

# Comparaciones entre variables (correlaciones)

# Correlaciones

En el caso que interese comparar muestras en las cuales la variable independiente sea cuantitativa, siempre será más interesante estudiar la relación entre estas dos variables, ya que, si resultan estar relacionadas entre sí querrá decir también que existen diferencias entre muestras extraídas de “clusterizar” o agrupar cada muestra en un rango de valores de la variable independiente. Ej. Diferencias con respecto al tiempo en realizar una maratón entre individuos altos y bajos.

## Tipos de test estadísticos para realizar inferencias sobre comparaciones entre variables (correlaciones)

Distribución	Variable independiente	Variable dependiente	Prueba Estadística
Paramétrica	Cuantitativa	Cuantitativa	Correlación de Pearson. Con ANOVA.
No paramétrica	Cuantitativa	Cuantitativa	Correlación de Spearman

Cada una de las pruebas estadísticas está profusamente explicada en Internet (recomiendo la Wikipedia), además de en la ayuda de cada uno de los paquetes estadísticos que posteriormente recomendaremos. La aplicación de cada prueba es sistemática. Juzgo más importante para este curso que aprendáis a elegir cada test en función de las circunstancias y objetivos concretos.

# OTROS TEST ESTADISTICOS

Existen otros tratamientos estadísticos más sofisticados, que normalmente no se utilizan en cuidados paliativos, pero que es interesante al menos citarlos:

# Análisis multivariante

- Es un método estadístico utilizado para determinar la contribución de varios factores en un simple evento o resultado.
- Sirve para separar el trigo de la paja.

Por ejemplo: En la variabilidad existente (la varianza total existente, en términos de calidad ) en el fruto de una determinada variedad de castaño, intervienen varias variables, el peso, la longitud, el grosor y la anchura.

Queremos saber cuales de estas variables son imprescindibles para poder distinguir entre castañas de primera calidad y castañas de segunda calidad. Realizo un análisis multivariante (que se podría considerar una metodología compuesta, pues es el resultado de realizar de forma ordenada y por fases una serie de análisis estadísticos individuales) y el resultado es que con solo el grosor se explica el 80% de la variabilidad de la calidad del fruto.

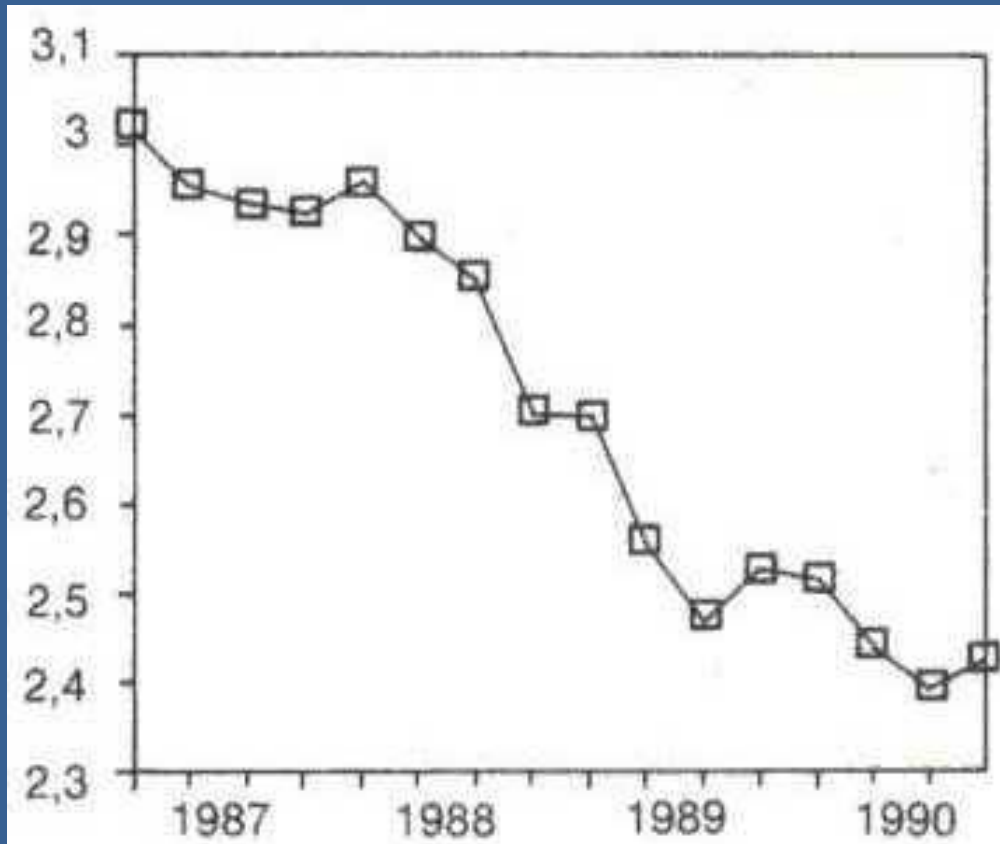
- Los factores de estudio son los llamados *factores de riesgo* (bioestadística), *variables independientes* o *variables explicativas*.
- El resultado estudiado es el *evento*, la *variable dependiente* o la *variable respuesta*.

# Análisis temporal

La serie temporal es una secuencia de datos, observaciones o valores, medidos en determinados momentos del tiempo, ordenados cronológicamente y, normalmente, espaciados entre sí de manera uniforme. El **análisis de series temporales** comprende métodos que ayudan a interpretar este tipo de datos, extrayendo información representativa, tanto referente a los orígenes o relaciones subyacentes como a la posibilidad de extrapolar y predecir su comportamiento futuro.

# Análisis temporal

Para representar de una serie temporal se debe realizar una gráfica de dispersión x-y:



# Análisis temporal

El análisis temporal se basa en el análisis de tendencias, las correlaciones entre un año y el año anterior, entre el año y dos años antes, y así sucesivamente.

Estas correlaciones se llaman autocorrelaciones parciales. Sobre las mismas se aplica una metodología compuesta de análisis que puede dar lugar a predicciones a pocos años vista, con una determinada probabilidad de éxito.

Se usa bastante en la bolsa, pero también por ejemplo para predecir el cambio climático (con respecto a un parámetro medido de manera continuada en el tiempo).

# EJERCICIOS

# Ejercicios

- Queremos estudiar las diferencias en la longitud del fémur entre individuos de la etnia Hutu, en Burundi y la tribu Mahorí de Nueva Zelanda.  
(ANOVA)
- Queremos saber si existen diferencias entre pacientes de distintos sexos con respecto a un test de medición del dolor cuyo rango es discreto de 1 a 10.  
(Kolmogorow-Smirnov)
- Queremos saber si existen diferencias entre pacientes con cáncer de pulmón de distintos sexos con respecto a su grado de depresión, que se mide con una simple pregunta, al médico, ¿Considera usted que su paciente posee síntomas de depresión?  
(CHI-Cuadrado)

# LA SIGNIFICACIÓN ESTADÍSTICA

# ¿Qué es la significación estadística?

- La significación estadística mide el grado en que la hipótesis que quieres comprobar, además de cumplirse en la muestra, puede ser extrapolable al total de la población.
- Los test de significación contrastan la posibilidad de esta extrapolación. Estos test utilizan un nivel de significación, que suele ser el 95% de probabilidad o el 99% de probabilidad. Es decir, la hipótesis se cumple para la totalidad de la población al 95% de probabilidad, si recogieras 100 muestras de población y validaras para cada una de ellas el cumplimiento de la hipótesis, no más de 5 muestras fallarían.

# ¿Qué es la significación estadística?

- Los test estadísticos trabajan con el parámetro “p”, que es la probabilidad en tanto por 1, y es  $1-q$ , siendo  $q$  la probabilidad real en tanto por uno. Es decir  $p=0,05$  equivale a una probabilidad del 95% y  $p=0,01$  equivale a una probabilidad del 99%.
- En los test estadísticos también aparecen los parámetros alfa  $\alpha$  y beta  $\beta$ . Ambos se miden en tanto por uno y se miden de forma análoga a  $p$ , es decir  $\alpha = 0,05$  se corresponde al 95% de probabilidad.
- Alfa es la probabilidad de tipo I, es decir, la probabilidad de admitir como cierto, a partir de unos datos limitados, un resultado que es falso en el conjunto de la población.
- Beta, también llamada probabilidad de tipo II, es la probabilidad de admitir como falso un fenómeno que sí sucede en el conjunto de la población.

# ¿Qué es la significación estadística?

- En bioestadística siempre se suele buscar que la probabilidad o riesgo alfa, sea lo más baja posible, ya que normalmente buscas contrastar que tu hipótesis se cumple en la vida real.
- Un ejemplo de búsqueda de probabilidad de tipo beta muy bajo, sería la legislación penal, en la que pesa más que un inocente no vaya a la cárcel antes que un culpable sea declarado inocente.

# PAQUETES ESTADÍSTICOS (SOFTWARE)

# Paquetes estadísticos

- **SPSS**
- **MICROSOFT EXCEL**
- **KNIME**
- **EPIDAT**

# Paquetes estadísticos

## SPSS

Descargar de:

[www.spss.com/statistics](http://www.spss.com/statistics)

Existe la posibilidad de un **Trial** de 22 días. Interesante para aprender a utilizarlo.

Ayuda:

[www.spsfree.com/indice.html](http://www.spsfree.com/indice.html)

# EJERCICIO SPSS (SOFTWARE)

# Paquetes estadísticos: SPSS

- Ejemplo. Diferencias en ritmo deposicional entre pacientes tratados para lactulosa y pacientes no tratados.
  - En primer lugar accedemos a los datos. Estos también se pueden importar de un Excel, un fichero de texto o una base de datos.
  - Vamos a ver dos opciones muy interesantes: Transformar Recodificar y Datos Segmentar (para crear clusters distintos, para incorporar una nueva variable independiente, por ejemplo) .
  - Vemos que el ritmo deposicional puede variar entre 0 y 28. Aunque se trata de una variable discreta, vamos a intentar aplicar test paramétricos, que como hemos indicado son mas potentes. Por el teorema del límite, sabemos que una variable discreta de un número elevado de rangos, tiende a distribuirse como una variable normal si además la muestra es elevada.
  - Vamos a realizar un test de normalidad. Para ello vamos a usar las opciones explorar y frecuencias.
  - Tenemos un “resut coach” para interpretar los resultados.
  - Al no cumplir la prueba, hemos de descartar los test paramétricos.
  - Vamos al cuadro a ver que elegimos.
  - Preguntamos en la ayuda por la prueba seleccionada.
  - Utilizamos pruebas no paramétricas, dos muestras independientes.
  - Utilizamos Mann-Witney

#### 454 □ METODOS DE DISTRIBUCION LIBRE

Volvamos ahora a la Tabla X del Apéndice B entrando con  $m = 12$  y  $n = m + 3 = 15$ . El punto crítico para un contraste con cola a la derecha, y  $\alpha = 0,05$ , es 202. Como  $212 > 202$ , rechazamos  $H_0$  y concluimos que los fumadores tienden a tardar más tiempo en caer dormidos que los no fumadores.

---

Si ambas poblaciones  $X$  e  $Y$  se suponen normales, entonces el test de los rangos con signo de Wilcoxon contrasta la misma hipótesis que el test  $T$  conjunto en la teoría normal.

Muchos otros contrastes para distribuciones libres son equivalentes al test de los rangos con signo de Wilcoxon. La alternativa más conocida es el test de Mann-Withney. Depende también de las observaciones  $X$  e  $Y$  linealmente ordenadas. El estadístico en este caso es  $U$ , número de veces que un valor  $X$  precede a un valor  $Y$ . Si la población  $X$  está situada por encima de la población  $Y$ , entonces  $U$  será grande; si es cierto lo contrario,  $U$  será pequeño. También se ha tabulado la distribución de probabilidad de  $U$  para tamaños muestrales seleccionados. Puesto que el test es equivalente al de Wilcoxon, no es necesario insistir en su manejo.

al nivel  $\alpha = 0,05$  para cada grupo.

### CONTRASTES DE POSICIÓN: DATOS NO APAREADOS

**13.3**  En esta sección analizamos un test de distribución libre que puede utilizarse para comparar la posición de dos poblaciones continuas, basado en muestras independientes de tamaños  $m$  y  $n$  extraídas de aquellas poblaciones. Se llama *test de la suma de los rangos de Wilcoxon*.

#### Test de la suma de los rangos de Wilcoxon

Sean  $X$  e  $Y$  variables aleatorias continuas. Sean  $X_1, X_2, \dots, X_m$  e  $Y_1, Y_2, \dots, Y_n$  muestras aleatorias independientes, de tamaños  $m$  y  $n$ , de las distribuciones de  $X$  e  $Y$ , respectivamente. Supongamos que  $m < n$ . Esto es, supongamos que las  $X$  representan la muestra más pequeña. La hipótesis nula es que las poblaciones  $X$  e  $Y$  son idénticas. Queremos contrastar estas hipótesis con un test que es especialmente idóneo para rechazar  $H_0$  si las poblaciones difieren en posición. Las  $m + n$  observaciones se funden para formar una única muestra. Las observaciones se ordenan linealmente y se les atribuye un rango de 1 a  $m + n$  conservando su identidad de grupo. Se asigna a las coincidencias que aparezcan la media de los rangos que les corresponderían, como en los test de Wilcoxon.

El estadístico es  $W_m$  la suma de los rangos asociados con las observaciones que originalmente constituyeron la muestra menor (valores  $X$ ). La lógica que está detrás de esta elección del estadístico es la siguiente: si la población  $X$  está situada por debajo de la población  $Y$  entonces los rangos menores tenderán a asociarse con los valores  $X$ . Ello producirá un valor pequeño para  $W_m$ . Si es cierto lo contrario (la población  $X$  está situada por encima de la población  $Y$ ), entonces los rangos mayores se encontrarán entre las  $X$ , dando lugar a un gran valor de  $W_m$ . De este modo rechazaremos  $H_0$  si el valor observado de  $W_m$  fuera demasiado pequeño o demasiado grande para que se debiera al azar. La Tabla X del Apéndice B da las probabilidades para valores seleccionados de  $m$  y  $n$ . Indicamos cómo utilizar esta tabla en el Ejemplo 13.3.1.

**EJEMPLO 13.3.1**  En un estudio sobre el hábito de fumar y sus efectos sobre las pautas del sueño, una de las variables importantes es el tiempo que se tarda en caer dormido. Se extrae una muestra de tamaño 12 de la población

	FUMADORES (S)		NO FUMADORES (N)	
	69,3	52,7	28,6	30,6
	56,0	34,4	25,1	31,8
	22,1	60,2	26,4	41,6
	47,6	43,8	34,9	21,1
	53,2		29,8	36,0
	48,1		28,4	37,9
	23,2		38,5	13,9
	13,8		30,2	

¿Indican estos datos que los fumadores tienden a tardar más tiempo en caer dormidos que los no fumadores?

Para responder a esta pregunta fundimos las dos muestras, ordenamos las observaciones de menor a mayor conservando su identidad de grupo y les atribuimos un rango de 1 a 27:

OBSERVACION	13,8	13,9	21,1	22,1	23,2	25,1	26,4	28,4	28,6
Grupo	S	N	N	S	S	N	N	N	N
Rango	1	2	3	4	5	6	7	8	9

OBSERVACION	29,8	30,2	30,6	31,8	34,4	34,9	36,0	37,9	38,5
Grupo	S	N	N	S	S	N	N	N	N
Rango	10	11	12	13	14	15	16	17	18

OBSERVACION	41,6	43,8	47,6	48,1	52,7	53,2	56,0	60,2	69,3
Grupo	S	N	N	S	S	N	N	N	N
Rango	19	20	21	22	23	24	25	26	27

Puesto que la muestra de la población de fumadores ( $m = 12$ ) es menor que la de no fumadores ( $n = 15$ ) el estadístico  $W_m$  es la suma de los rangos asociados a los fumadores. Como sospechamos que los fumadores tardan más tiempo en caer dormidos que los no fumadores rechazamos la hipótesis nula de que no existe diferencia entre los dos grupos si el valor observado de  $W_m$  es demasiado grande para que se deba al azar. Para estos datos

$$W_m = 1 + 4 + 5 + 14 + 20 + 21 + 22 + 23 + 24 + 25 + 26 + 27 = 212$$

# Paquetes estadísticos: Knime

## Knime

Ayuda:

<http://laurel.datsi.fi.upm.es/media/docencia/cursos/inapejemplodm.pdf>

[www.knime.org/documentation/getting\\_started](http://www.knime.org/documentation/getting_started)

Descargar de:

[www.knime.org](http://www.knime.org)

# Paquetes estadísticos: Knime

Para conectarse a Excel → Nodo Database Reader

```
jdbc:odbc:Driver={Microsoft Excel Driver (*.xls)};DBQ =  
C:/Documents and Settings/HP_Administrator/My  
Documents/Julian/excel.xls
```

- Es la ruta del archivo

Para conectarse a Access -> Nodo Database Reader

```
jdbc:odbc:Driver={Microsoft Access Driver  
(* .mdb)};DBQ=C:/Documents and  
Settings/HP_Administrator/My  
Documents/bd2.mdb;DriverID=22;READONLY=false}
```

- Ver el ejemplo [ejemplodm.pdf](#)

# Bibliografía

- Estadística para Biología y Ciencias de la Salud. McGraw-Hill
- Estadística modelos y Métodos. Alianza Universal Textos.
- Curso de Postgrado en Estadística Aplicada. Departamento de Matemáticas da Universidade da Coruña. Juan Manuel Vilar Fernández. Enero 2001.
- Bioestadística para Clínicos. Mas allá de “ $p < 0.05$ ”. A.S. Rubiales, M.L. del Valle. A.Vecino, L.A. Flores. Hospital Clínico Universitario “Centro de Salud Medina del Campo”. Valladolid.
- Estadística para Clínicos. Enrique de Ramón Garrido, Oscar Fernández Fernández, Gloria Luque Fernández, Unidad de Investigación Clínico-Experimental. Servicio de Neurología. Hospital Regional de SAS “Carlos Haya” de Málaga.

# Bibliografía

- Wikipedia <http://es.wikipedia.org>
- Fundación Iberoamericana para la Gestión de la Calidad [www.fundibeq.org](http://www.fundibeq.org)

# ¡GRACIAS!

$$\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$