

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 1 de 106

TITULO:

CURSO DE ESTADÍSTICA BÁSICA PARA RESIDENTES

ELABORADO POR:

Dr. D. FERNANDO RODRÍGUEZ CANTALEJO
 (FEA Análisis Clínicos
 S. Análisis Clínicos
 H.U. Reina Sofía de Córdoba)

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 2 de 106

AGRADECIMIENTOS:

Dr. D. RAFAEL MUÑOZ (Responsable Unidad de Garantía de la Calidad S. Análisis Clínicos H.U. Reina Sofía de Córdoba), por proponer la idea de la creación de este manual para residentes, cuya formación ha constituido una de sus preocupaciones profesionales más importantes.

Dr. D. CRISTÓBAL AGUILERA GÁMIZ (Jefe del Servicio de Análisis Clínicos del H.U. Reina Sofía de Córdoba), por transmitir la importancia del conocimiento de las herramientas matemáticas y estadísticas que permiten obtener el máximo rendimiento en la organización, desarrollo e interpretación de los datos del Laboratorio

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 3 de 106

ÍNDICE

1. CONCEPTOS BÁSICOS DE ESTADÍSTICA	2
1.1 Concepto de estadística	2
1.2 Población, muestra e individuo	2
1.3 Parámetros y estadísticos	4
1.4 Concepto de variable	5
1.5 Medios para la descripción de los datos	7
1.6 Distribución de probabilidad	9
2. VARIABLES CUALITATIVAS	16
3. VARIABLES ORDINALES	18
4. VARIABLES CUANTITATIVAS	21
5. INFERENCIA ESTADÍSTICA	31
5.1 Conceptos generales	31
5.2 Estimación	32
5.3 Test de hipótesis	33
5.4 Esquemas de test de hipótesis	34
5.4.1 Según variable clasificadora y variable a estudio	35
5.4.2 Procedimientos en estadística descriptiva e inferencial, según tipo y numero de muestras.	35
5.4.2.1 Procedimientos descriptivos y gráficos.	35
5.4.2.2 Test de comparación	35
5.4.2.3 Test de asociación	35
5.4.3 Características de los test de hipótesis más empleados	38
5.4.3.1 Prueba T de Student / F de Snedecor	
5.4.3.2 Análisis de la Varianza (ANOVA)	49
5.4.3.3 Estudios de relación entre variables	54
5.4.3.3.1 Correlación	54
5.4.3.3.2 Regresión. Tipos de regresión y ajustes	56
5.4.3.3.3 Tablas de contingencia. Método de Chi-Cuadrado. Fuentes de asociación y métodos de medida	62
6. ESTADÍSTICA EPIDEMIOLÓGICA	71
6.1 Epidemiología: aspectos generales	71
6.2 Medidas de frecuencia de enfermedad.	72

S. de Análisis Clínicos H.U. Reina Sofía	<p align="center">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 4 de 106

Prevalencia, Incidencia

6.3 Nociones sobre la calidad de los test diagnósticos empleados en epidemiología	76
6.3.1 Características de los test diagnósticos. Sensibilidad. Especificidad	78
6.3.2. Interpretación de los test diagnósticos. Valor predictivo. Teorema de Bayes. Razón de verosimilitud.	79
6.3.6 Curvas ROC	83
6.3.7 Test combinados en estudios epidemiológicos	86
6.4 Tipos de estudios epidemiológicos	87
6.4.1 Estudios transversales	88
6.4.2 Estudios de cohortes	90
6.4.3 Estudios de casos-control	94
6.4.4 Estudios experimentales	98
7. BIBLIOGRAFIA	106

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 5 de 106

1 CONCEPTOS BÁSICOS DE ESTADÍSTICA

1.1 ESTADISTICA

La estadística puede definirse como un método de razonamiento lógico matemático que estudia aquellos aspectos de la realidad en los que interviene el azar, establece los medios para obtener información y proporciona instrumentos para la toma de decisiones cuando prevalecen condiciones de incertidumbre.

Se emplea en cualquier actividad humana, algunas de las cuales han desarrollado su propia estadística (bioestadística, socioestadística, etc.)

La estadística emplea los métodos necesarios para recoger, clasificar, representar y resumir datos, así como para hacer inferencias (extraer consecuencias o estimaciones) científicas a partir de ellos.

Consta de tres partes:

- **Estadística Descriptiva**, cuyo único fin es la recogida, clasificación, representación y resumen de los datos ya existentes.
- **Cálculo de Probabilidades**, proporciona herramientas conceptuales necesarias para aplicar con éxito la inferencia estadística.
- **Estadística inferencial**, cuyo objetivo es extender las conclusiones obtenidas en una parte de la población (muestra) a toda ella (población).

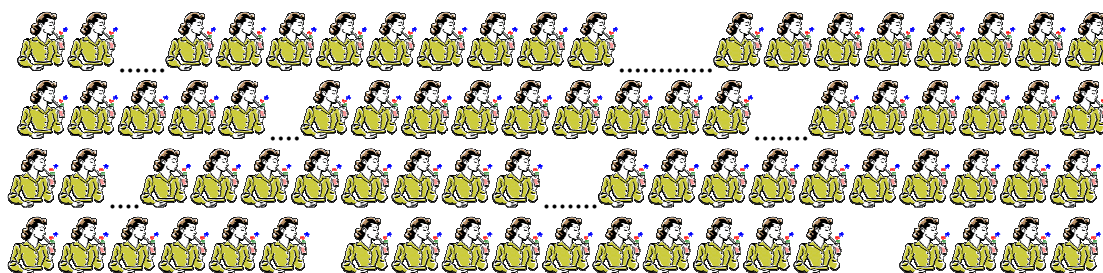
1.2 POBLACIÓN, MUESTRA E INDIVIDUO

1.2.1 POBLACIÓN O UNIVERSO

Es el conjunto de todos las personas, entidades objetos, ideas, elementos o acontecimientos que cumplen ciertas propiedades o características, entre las cuales se desea estudiar un determinado fenómeno.

Ejemplo: mujeres en edad fértil de un Distrito Sanitario, que podemos llamar "N".

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 6 de 106



En la mayoría de las ocasiones resulta imposible recoger los datos de todos y cada uno de los elementos de la población, por ello, se recurre a la selección de una muestra.

1.2.2 MARCO

Es el conjunto de información (ficheros magnéticos, directorios, listados, etc.) que posibilita la identificación sin ambigüedades de todas las unidades que componen la población, constituye la base informativa. Ej.: bases de datos, listados, etc.

1.2.3 INDIVIDUO

Se denomina individuo o unidad de investigación a cada uno de los elementos de una población. Ej.: cada mujer en edad fértil

1.2.4 MUESTRA

Es un subconjunto de la población o grupo de elementos pertenecientes a la población de estudio, escogidas al azar y tomadas como representativa de la población.

Nos puede permitir establecer una "estimación" de determinadas características de la población.

La elección al azar y el tamaño (número de elementos) se relacionan con el grado de representatividad de la muestra.

Ejemplo de muestra:

Siguiendo el ejemplo anterior obtendríamos una muestra seleccionando 30 mujeres en edad fértil de un Distrito sanitario, que podríamos llamar "n".

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 7 de 106



Existen distintas técnicas de muestreo u obtención de muestras:

- Muestreo probabilístico:
 - En el que se conoce la población objeto del estudio y la posibilidad de ser incluido en la muestra es conocida para cada individuo antes de ser o no seleccionado.
 - Muestreo aleatorio simple:
 - Garantizar que todos los elementos de la población tengan la misma probabilidad de ser seleccionados (muestra de tamaño "n", seleccionada de una población "N", de manera que cualquier posible muestra de tamaño "n" tiene la misma probabilidad de ser seleccionada).
 - Es la más empleada.
 - Muestreo estratificado:
 - Se divide la población en grupos o estratos y después se obtiene una muestra aleatoria de cada uno de ellos.
 - Muestreo sistemático:
 - Se selecciona uno de cada k individuos, siendo k la constante de muestreo (se calcula dividiendo el nº de sujetos de la población entre el tamaño calculado de la muestra)
 - Muestreo polietápico o por conglomerados:
 - Se seleccionan unidades de muestreo de la población (unidades primarias), a partir de estas se seleccionan unidades secundarias y así sucesivamente.
 - Se pueden emplear distintas técnicas de muestreo para cada etapa.
- Muestreo no probabilístico:


S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 8 de 106

- La posibilidad de ser incluido en la muestra no se conoce para cada individuo antes de ser o no seleccionado: p.e, individuos que acuden a un centro de tomas de muestras.
- No asegura la representatividad de la población.

1.2.3 INDIVIDUO

Cada uno de los elementos que componen la población es un individuo

Representa la unidad básica del muestreo.

Siguiendo con el ejemplo: cada una de las mujeres en edad fértil seleccionadas ().

1.3 PARAMETROS Y ESTADÍSTICOS.

1.3.1 PARAMETRO

Es el valor que resume determinada información o característica referente a una población

Ejemplo: la colesterolemia media de la población total de mujeres en edad fértil de un Distrito Sanitario es de 220 mg/dl. Sin embargo nosotros no hemos medido directamente este dato, sino que lo hemos estimado a partir del estudio de una muestra de 30 individuos.

1.3.2. ESTADÍSTICO

Es el valor que expresa determinada información de una muestra.

Ejemplo: la colesterolemia media que estamos estudiando es de 225 mg/dl. conociendo este dato esperamos que la media de colesterol en suero para la población total de mujeres en edad fértil sea un valor cercano a 225 mg/dl. Considerando el error que podemos tener hacemos una estimación del valor poblacional.

Podemos obtener distintos tipos de estadísticos:

- De tendencia central: moda, mediana, media

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 9 de 106

- De dispersión: desviación estándar, coeficiente de variación
- De posición de un dato: fractiles
- De forma: sesgos y curtosis

En conclusión: **el estadístico es lo que medimos en la muestra; el parámetro no lo conocemos**, lo estimamos para la población a partir del estadístico y no tiene porqué coincidir, puede variar en función del error que aceptemos y del porcentaje de probabilidad con que trabajemos.

1.3.3 MARCO

Se define como tal al soporte en el que se presenta la información sobre las características de los elementos de la muestra y que servirá para el posterior análisis estadístico (tablas, bases de datos, informes, etc.)

1.4 CONCEPTO DE VARIABLE

La variable es aquello que nos interesa medir en los individuos.

Por ejemplo, los niveles de estrógenos en la primera fase del ciclo ovárico sería una variable, los niveles de estrógenos en la segunda fase del ciclo sería otra variable, etc.

1.4.1 V. CUALITATIVA O ATRIBUTO

Representa una cualidad y, en principio, no pueden cuantificarse, ya que no toman valores numéricos.

Admiten operadores de igualdad o desigualdad ($=, \neq$)

Ejemplo: etnia de las mujeres de edad fértil de ejemplos anteriores, profesión, etc.

1.4.2 V. ORDINAL

Representa **relaciones de orden** entre los distintos valores obtenidos

Admite operadores de igualdad, desigualdad y de orden ($=, \neq, <, >, \leq, \geq$)

Ejemplo: grado de satisfacción en la atención en consulta ginecológica.

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 10 de 106

1.4.3 V. CUANTITATIVA

Toman **valores numéricos**, y se acompañan de determinadas unidades de medida.

Admite diversos operadores: $=, \neq, <, >, \leq, \geq, +, \times, :$

Ejemplo: concentración de colesterol en mg/dl

Las variables cuantitativas pueden ser **discretas** o **continuas**

1.4.3.1 V. CUANTITATIVAS DISCRETAS

Se caracterizan porque entre dos valores consecutivos no podemos encontrar ningún otro valor (corresponden a valores naturales).

Ejemplo: nº de hijos, etc. (no existen 1.2 o 3.4 hijos)

1.4.3.2. V. CUANTITATIVAS CONTINUAS

Se caracterizan porque entre dos valores continuos podemos encontrar infinitos valores (corresponden a números reales).

Ejemplo: peso (entre 60 y 61 kg: 60.1, 60.2, 60.3... kg), talla (entre 1.69 y 1.70 m: 1.691, 1.692, 1.693... m.), etc.

1.5 MEDIOS PARA LA DESCRIPCIÓN DE LOS DATOS

1.5.1 TABULACIÓN

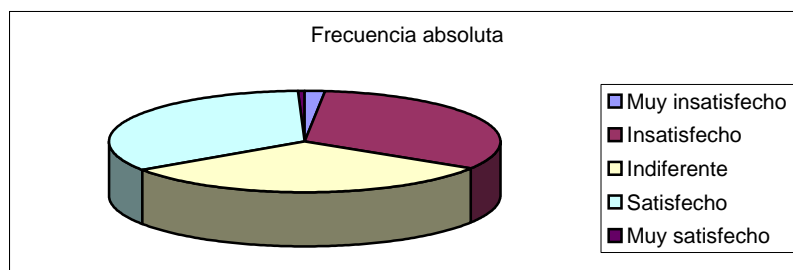
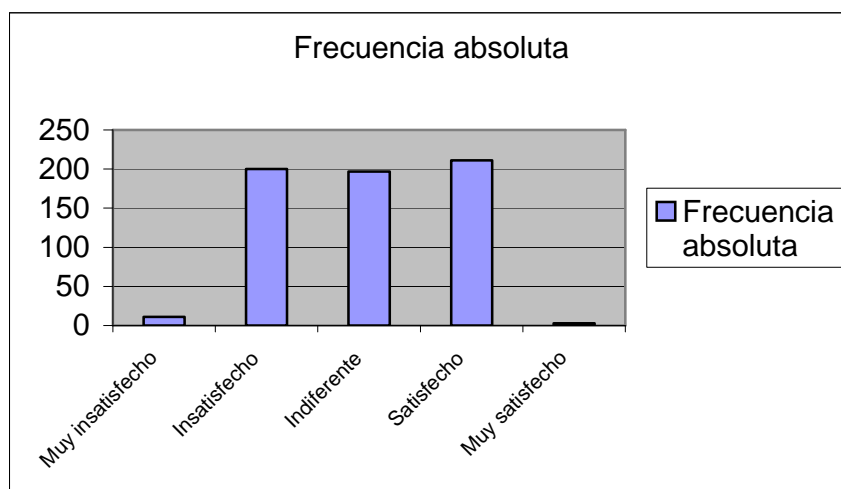
Tabular es ordenar los datos originales y presentarlos de forma que, sin perder información, sea más fácil conocer la distribución de datos.

El resultado es una tabla donde se muestran todos los valores de la variable y algunas informaciones referidas a ellas.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 11 de 106

1.5.2 REPRESENTACIÓN GRAFICA

Tras la consecución de la tabla es relativamente fácil construir una representación de los datos en un sistema de coordenadas.



La

representación gráfica de los datos refleja las características globales de la distribución.

1.5.3 OBTENCIÓN DE ESTADÍSTICOS

Los estadísticos resumen la distribución a través de medidas de:

- Tendencia central
- Variabilidad o dispersión
- Posición
- Forma

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 12 de 106

Las características de las variables definen el tipo de medida empleada.

1.5.4. ESCALAS DE MEDIDA

Los tipos de variables expuestos anteriormente presentan cuatro posibles niveles de medición. Cada uno de ellos tiene una serie de pruebas apropiadas.

1.5.4.1. ESCALA NOMINAL

Definen categorías, corresponden a las variables cualitativas.

Pueden ser expresadas por números, pero solo indican diferencias entre las entidades y su objetivo es obtener una clasificación de los mismos.

Ejemplo: 1, si es hombre; 2 si es mujer.

Permite sólo operaciones del tipo no paramétrico, como medidas de asociación. La medida de asociación más común para escalas nominales es el coeficiente de contingencia y la Q de Yule.

1.5.4.2 ESCALA ORDINAL

No sólo existen diferencias entre los distintos números, sino que pueden ser ordenados

Ejemplo: grado satisfacción de usuarios de un servicio, escala de conciencia de Glasgow, etc.

Permite operaciones del tipo no paramétrico, siendo la más empleada la mediana. Las hipótesis pueden probarse mediante estadísticos de rango, como el coeficiente de correlación de Spearman o el de Kendall).

1.5.4.3 ESCALA CUANTITATIVA

Las escalas cuantitativas se clasifican en:

- INTERVALO:

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 13 de 106

- Existe diferencia y orden entre los números, además la diferencia entre un par de entidades con números consecutivos es la misma que entre cualquier otro par.
- No existe "cero" absoluto, ni relación multiplicativa entre los números, por ejemplo: la temperatura de 0 °C no indica ausencia total de temperatura, ni 40 °C es el doble de 20 °C, puesto que si medimos la temperatura en otra escala no se mantendrá la relación 2 a 1.
- PROPORCIÓN:
 - Se caracteriza porque además de existir diferencia, orden e intervalo, el cero tiene un sentido real, por ejemplo el peso medido en kg, (onzas. Libras, etc.) si tiene un cero absoluto.
 - En ella, la proporción de un punto a otro cualquiera de la escala es independiente de la unidad de medida.
 - Al existir "cero" real puede establecerse proporción o razón entre los datos de la variable y comparar con otras categorías.

1.5.4.4 RELACIÓN ENTRE VARIABLES Y ESCALAS DE MEDIDA

La escala de medida va asociada habitualmente a una variable

Algunas variables pueden medirse en distintas escalas, por ejemplo:

La variable tensión arterial es cuantitativa, pero se puede describir usando:

- una escala de tipo nominal (hipertensos, no hipertensos),
- una escala de tipo ordinal (no hipertensos, border-line e hipertensos)
- una escala de tipo cuantitativo (mmHg)

1.6. DISTRIBUCION DE PROBABILIDAD

1.6.1 PROBABILIDAD: NOCIONES BASICAS

Podemos decir que la **probabilidad** de que ocurra un suceso corresponde a la frecuencia con que dicho suceso tendrá lugar en pruebas repetidas.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 14 de 106

También podemos calcular la probabilidad de ocurrencia de un suceso mediante el cociente de casos favorables entre los casos posibles, de manera que:

- Cuando no existen casos favorables la probabilidad es 0
- Si los casos favorables son todos los posibles la probabilidad es 1
- La suma de las probabilidades de todos los casos posibles siempre será 1
- La probabilidad de que no ocurra un suceso será el suceso complementario (1-P).

Tipo de sucesos

Sucesos elementales: son los resultados más simples de un experimento.

Suceso compuesto es el que resulta de la unión de dos o más sucesos elementales.

Suceso total es la suma de todos los sucesos elementales de un experimento (se denomina Ω).

Suceso vacío es el suceso imposible

Suceso contrario de A es A' ($A' = \Omega - A$)

Operaciones con sucesos:

Llamaremos unión a la adición de sucesos, que se puede representar con "+" o "U", e intersección a la multiplicación, representada por "x" o " \cap ":

- Probabilidad de que ocurra un suceso A o un suceso B (suma de probabilidades):
 - o A y B independientes (excluyentes): $p(A \cup B) = p(A) + p(B)$
 - o A y B dependientes (no excluyentes): $p(A \cup B) = p(A) + p(B) - p(A \cap B)$
 - es necesario descontar el suceso que se cuenta dos veces
- Probabilidad de que ocurra un suceso B dado un suceso A (probabilidad condicionada):
 - o $p(B/A) = p(A \cap B) / p(A)$
- Probabilidad de que ocurra un suceso A y un suceso B (multiplicación de sucesos):
 - o A y B dependientes: $p(A \cap B) = P(A) \times p(B/A)$
 - o A y B independientes:

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 15 de 106

- En este caso, la $P(B/A) = p(B)$, por lo que:
 - $p(A \cap B) = p(A) \times p(B)$

1.6.2 DISTRIBUCIÓN DE PROBABILIDAD

Hemos visto como a partir de la frecuencia muestral de un determinado suceso y en base a los criterios de probabilidad podemos asociar dicha frecuencia con la probabilidad de que se produzca dicho suceso.

Esta probabilidad real puede convertirse en probabilidad teórica al definir el conjunto de sucesos por sus elementos característicos: media, desviación típica y tamaño de la muestra.

La probabilidad teórica sigue un modelo definido por una función matemática de distribución $f(x)$, donde $f(x) \geq 0$ se denomina **distribución de probabilidad** (o función de densidad de probabilidad) de la variable continua aleatoria x si el área total entre la curva y el eje X es igual a 1, y si el área entre la curva, el eje X y 2 perpendiculares sobre 2 puntos a y b corresponden a la probabilidad de que x esté entre a y b .

1.6.2.1 DISTRIBUCIÓN NORMAL

La distribución de probabilidad más conocida es la descrita por el astrónomo Karl Fiedrich Gauss (siglo XIX), denominada distribución normal, distribución de Gauss o de Laplace-Gauss, por la contribución de este otro científico.

A título informativo, expondré que la función de Gauss viene definida por la fórmula:

Función de Densidad

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ media	$\pi = 3,1415...$
σ desv. típica	$e = 2,7182...$
σ^2 varianza	x abscisa

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 16 de 106

La distribución normal se caracteriza porque los valores se distribuyen formando una campana (campana de Gauss) entorno a un valor central que coincide con el valor medio. Característicamente, **en la distribución normal coinciden la media aritmética, la mediana y la moda** en el mismo valor, que se sitúa en el centro de la curva.

Existen diversos test estadísticos para comprobar si una distribución se ajusta a la distribución teórica de Gauss: test de Kolmogorov-Smirnov, test de Shapiro-Wilks.

1.6.2.1.1. Función de densidad de probabilidad

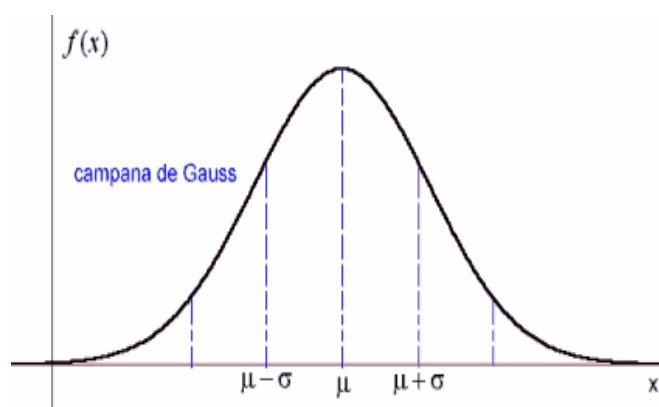
Representa la probabilidad de encontrar un valor x_i en el intervalo $0-x$ y viene determinada por el área bajo la curva de Gauss (normal) entre 0 y x .

Función de Distribución

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

$-\infty < x < +\infty$

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



1.6.2.1.2 Propiedades de la función de Gauss:

-
- Simétrica respecto a la media
- Determinada por la media y la desviación típica
- El área total bajo la curva es 1
- Si trazamos perpendiculares al eje X sobre s , $2s$ y $3s$ quedarían englobados respectivamente el 68, 95 y 99.7% del área de la curva
- Asintótica: nunca toca al eje X.
- El punto de inflexión corresponde a la desviación típica

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 17 de 106

Intervalo	Área (probabilidad)
$\mu \pm \sigma$	0.68
$\mu \pm 2 \sigma$	0.95
$\mu \pm 3 \sigma$	0.997

1.6.2.1.3. Tipificación

Si utilizáramos la fórmula de la distribución normal de Gauss necesitaríamos realizar complejos cálculos cada vez que quisiéramos calcular una probabilidad (área). Sin embargo, podemos transformar la distribución normal en una **distribución tipificada** caracterizada por:

- la media está centrada en 0
- La variable transformada z tiene una s de 1

Para transformar cualquier valor x_i de una distribución de media μ y desviación estándar σ en el correspondiente valor z de la distribución normal tipificada, se aplica la fórmula:

$$Z = \frac{X - \mu}{\sigma}$$

Mediante la tipificación de la distribución podemos estudiar el número de veces que un dato de la distribución supera la desviación típica y establecer comparaciones entre distintas distribuciones.

1.6.2.2 DISTRIBUCIÓN BINOMIAL

Las **variables de tipo cualitativo dicotómico** siguen esta distribución, como expresión de un proceso que sólo puede resultar en dos estados mutuamente excluyentes (ejemplo: ser hombre o mujer, ser fértil o estéril, etc.).

La distribución binomial se caracteriza por:

- En cada prueba del experimento sólo son posibles dos resultados: el suceso A (éxito) y su contrario A' (fracaso).

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 18 de 106

- El resultado obtenido en cada prueba es independiente de los resultados obtenidos anteriormente.
- La probabilidad del suceso A es constante (si cada vez que sacamos una bola la devolvemos a la bolsa, también llamado *reemplazamiento*). la representamos por p , y no varía de una prueba a otra. La probabilidad de \bar{A} es $1-p$ y la representamos por q .
- El experimento consta de un número n de pruebas

La distribución binomial se suele representar por $B(n,p)$ siendo n y p los parámetros de dicha distribución.

Podríamos determinar la probabilidad de éxito en n repeticiones, para ello estableceríamos un nº de combinaciones de n elementos tomados de x en x , por lo que la función binomial quedaría definida por:

Probabilidad de obtener k -éxitos

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

A la variable X , que expresa el número de éxitos obtenidos en cada prueba del experimento, la llamaremos **variable aleatoria binomial**.

1.6.2.3 DISTRIBUCIONES DISCRETAS

1.6.2.3.1 DISTRIBUCION DE POISSON

La Distribución de Poisson se llama así en honor a Simeón Dennis Poisson (1781-1840), quien observó la probabilidad de verificarse un resultado en n pruebas independientes frente a la escasa probabilidad de que el resultado se verifique en una sola prueba (**Ley de Poisson**)

La distribución de Poisson es característica de los procesos donde existe un gran nº de individuos y probabilidad pequeña de sucesos.

La distribución Poisson es ideal para predecir el número de sucesos (casos) que se producirán en un determinado período de tiempo, cuando se trata de **eventos raros** con ocurrencia aleatoria en el tiempo. La ventaja del empleo

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 19 de 106

de esta distribución se basa en que permite obtener la probabilidad de ocurrencia del evento según su comportamiento medio anterior:

$$P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

donde :

λ : es el promedio de ocurrencia del evento (media aritmética de los casos en un período determinado)

x : es el número de veces que ocurrió el suceso (número de casos).

El coeficiente de variación de variables que siguen una distribución de este tipo es alto y viene dado por la fórmula:

$$CV_{\text{Poisson}} = 1/\sqrt{N}$$

donde N es el nº de elementos de la variable (nº de eventos).

Ejemplos de distribución de Poisson:

- Nº de células por unidad de volumen de líquido orgánico (ascítico o cefalorraquídeo).
- Nº de pacientes que demandan atención por unidad de tiempo.
- Nº de coches que pasan por caseta de peaje por unidad de tiempo, etc.

1.6.2.3.2 D. HIPERGEOMETRICA

Es equivalente a la binomial sin reemplazamiento, es decir, la probabilidad cambia cada vez que se extrae un individuo.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 20 de 106

1.6.2.4. DISTRIBUCIONES CONTINUAS

- D. EXPONENCIAL
- D. UNIFORME
- DERIVADAS DE LA NORMAL
 - Chi cuadrado
 - T de Student Fisher: no conocemos la varianza poblacional pero si la muestral.
 - F de Snedecor.

2. VARIABLES CUALITATIVAS

2.1 Recordemos que las variables cualitativas reflejan **atributos o categorías** de los sujetos de la población (sexo, profesión, etc.).

2.2 Se pueden describir mediante:

2.2.1 Tabulación de datos:

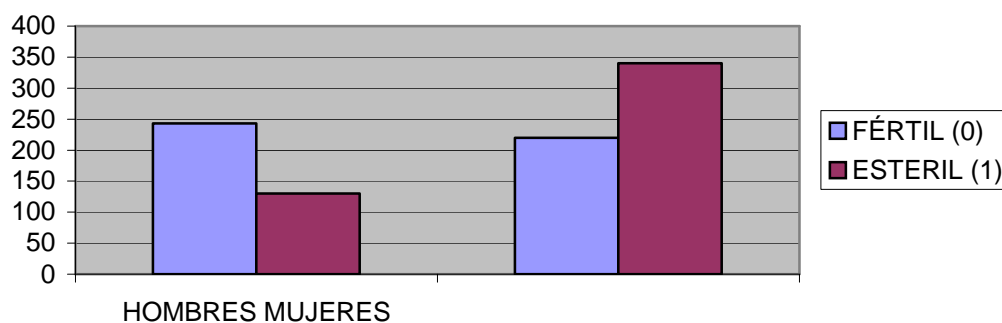
- Se utiliza una **Escala nominal**, que representa los posibles valores de la variable y, opcionalmente, su codificación, mediante:
 - **Frecuencias absolutas** n_i (número de veces que aparece en la distribución).
 - **Frecuencia relativa simple o proporción** de cada posible valor sobre el total ($f_i = n_i/N$). También se le llama probabilidad o tanto por uno.
- Tabulación de dos variables cualitativas
 - Se pueden "enfrentar" en una tabla 2 variables de tipo cualitativo en una tabla de "n" filas por "m" columnas según las categorías que presente cada variable
 - Las columnas deben representar categorías mutuamente excluyentes (hombre-mujer, estéril-fértil)
 - Corresponde a una tabla de contingencia, ejemplo:
Fertilidad por sexo

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 21 de 106

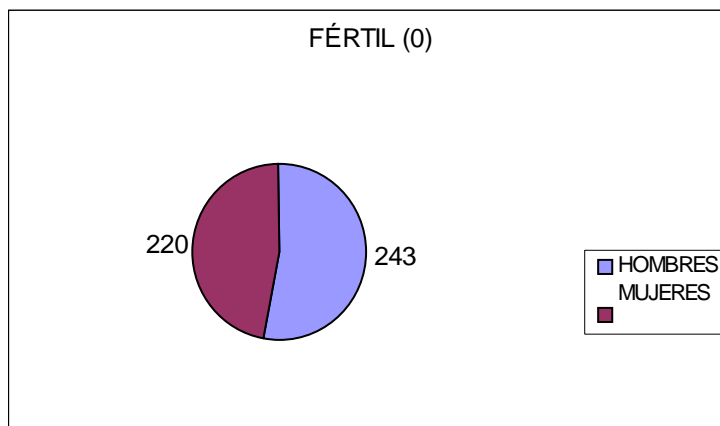
N	HOMBRE (1)	MUJER (2)	Tot fila
FÉRTIL (0)	243	220	463 (49.2)
ESTERIL (1)	130	340	470 (50.7)
Tot columna	373 (39.9)	560 (60.01)	933 (100)

2.2.2 Representación gráfica:

- **Diagrama de barras**, donde cada barra representa la frecuencia del valor de abscisas de que parte.



- **Superficies representativas**, como los **sectores circulares**, donde cada categoría de la variable ocupa un sector proporcional a la frecuencia.



S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 22 de 106

2.2.3 Estadísticos

2.2.3.1 E. De posición y tendencia central de la distribución:

- El valor más representativo es el que más se repite. La mayor frecuencia corresponde a la **moda**.
 - Si dos valores resultan ser los más repetidos la distribución es bimodal, si son más de dos es multimodal
- En variables dicotómicas, es decir, con valores mutuamente excluyentes y absolutamente exhaustivos (ser fértil o no fértil), se pueden definir las siguientes medidas de tendencia central:
 - **Proporción:** es la frecuencia relativa de un suceso, se calcula mediante el cociente del nº de sucesos (individuos) que presentan una determinada característica dividido por el total de sucesos (individuos) ($p_i = n_i / N$)
 - **Razón:** es el nº de casos que presentan una determinada característica dividido por el nº de casos que no la presentan
 - **Tasa:** es una proporción acumulada durante un periodo de tiempo determinado.

2.2.3.2 E. de dispersión:

- En realidad, los estadísticos de dispersión se aplican a variables cualitativas de tipo dicotómico, por otro lado, cualquier variable cualitativa medida en escala nominal puede ser estudiada como variable dicotómica, aunque tenga más de dos categorías
- Ejemplo: si las variables cualitativas corresponden al color del pelo: rubio, castaño, pelirrojo, moreno, etc., podemos estudiar la variable dicotómica: ser rubio/no ser rubio, etc.
- En variables dicotómicas: si np y $n(1-p)$ son mayor o igual a 5 se usan:
 - Varianza: $s^2p = \{[p(1-p)]/n\} \cdot (1-p)$
 - Desviación estándar: $DEp = sp$

2.2.3.2 E. De posición de un dato dentro de la distribución

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 23 de 106

- No tiene sentido, puesto que no tiene orden.

2.2.3.4 E. De forma: no se aplican a variables nominales

3. VARIABLES ORDINALES

Como ya comentamos, es aquella en la que los valores numéricos reflejan diferencia y relaciones de orden.

Se representan mediante escalas ordinales:

3.1 Tabulación de datos:

Podemos ordenar la frecuencias absolutas acumuladas: obtenidas mediante el sumatorio del número de veces que ocurre determinado valor de la variable y sus inferiores ordinalmente.

Podemos ordenar las frecuencias relativas acumuladas.

Ejemplo: representa la tabulación de los datos de una variable que se distribuye según una escala ordinal, corresponde al grado de "satisfacción sobre la asistencia hospitalaria" y se han establecido 5 categorías u órdenes.

Categoría	Código	Frecuencia absoluta	Frecuencia relativa	F. relativa acumulada
Muy insatisfecho	1	11	1.8	1.8
Insatisfecho	2	200	31.9	33.9
Indiferente	3	197	31.4	65.6
Satisfecho	4	211	33.7	99.5
Muy satisfecho	5	3	0.5	100
Total		627		100

La acumulación de las frecuencias relativas nos da idea de la distribución de los valores que adopta la variable. Nos introduce en el concepto del percentil.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 24 de 106

3.2 Representación gráfica de los datos:

Para representar gráficamente una escala ordinal se emplean los diagramas de barras. En este caso, es necesario mantener el *orden numérico* de los valores.

En las escalas ordinales solo sabemos si un elemento es mayor o menor pero *no sabemos cuanto mayo o menor*.

El cálculo de frecuencias acumuladas nos permite representar la curva de frecuencias acumuladas que adoptará una forma ascendente.

3.3 Estadísticos de variables ordinales :

3.3.1 E. de posición y tendencia central de la distribución

Al igual que en las escalas nominales, en las escalas ordinales también podemos usar la moda.

Interesa aprovechar la propiedad de orden, ya que si ordenamos los datos, el valor que ocupa la posición central informa de manera aceptable de la posición de la distribución. Este valor corresponde a la mediana, valor por encima y por debajo del cual se encuentran la mitad de los casos.

Cuando el número de observaciones es par, la mediana es el promedio de las dos observaciones centrales.

La mediana, por definición, no es sensible a la existencia de valores extremos

3.3.2. E. de Dispersión:

Están representados por los valores máximo y mínimo de la distribución y por la diferencia entre ellos (rango amplitud o recorrido

3.3.3. E. de posición de un dato dentro de la distribución

En base a las frecuencias acumuladas, las escalas ordinales se pueden representar por FRACTILES.

Si utilizamos tantos por ciento para definir la posición de un dato dentro de la distribución hablaríamos de percentiles.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 25 de 106

Dado un percentil K, el K% de la distribución tiene un valor igual o inferior a K y lo escribimos como $K=P50$ o bien como $C50$.

Ejemplo: un P80 de 20, indica que el 80% de la distribución tiene un valor igual o inferior a 20.

Si en lugar de utilizar tantos por ciento utilizamos tantos por diez, hablaríamos de deciles (D); si empleamos tantos por cuatro, de cuartiles (Q).

La mediana corresponde al percentil P50, al decil D5 o al cuartil Q2

3.3.4. E. De forma: resultaría conflictiva

4. VARIABLES CUANTITATIVAS

Recordemos que la variables cuantitativas son aquellas que toman **valores numéricos**, se acompañan de determinadas unidades de medida y se representan mediante escalas cuantitativas.

4.1 Tabulación de datos

Las escalas cuantitativas discretas (no existen infinitas posibilidades entre dos números) se representan de igual manera que las variables correspondientes a escalas ordinales: con las frecuencias absolutas y relativas acumuladas.

Las escalas cuantitativas continuas se representan mediante los *intervalos de clase*.

Ejemplo:

Límite inferior del intervalo	Límite superior del intervalo	Frecuencia absoluta	Frecuencia relativa	F. absoluta acumulada
25	30	11	1.7628	11
30	35	98	15.7051	109
35	40	103	16.5074	212
40	45	100	16.0256	312
45	50	99	15.8654	411

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 26 de 106

50	55	123	19.7115	534
55	60	90	14.4231	624

- *Cálculo del N° de intervalos:*
 - o N° óptimo: raíz cuadrada del n° de observaciones
 - o En la práctica, se recomienda no usar más de 10 intervalos.
- *Valores máximos y mínimo de los intervalos pueden ser:*
 - o Números reales: valores extremos que coinciden con los datos observados
 - o Límites exactos: valores imaginarios que por efecto del redondeo limitan con mayor precisión el inicio el inicio y el final de cada intervalo.
- *Amplitud de un intervalo (rango o recorrido)*
 - o Tablas con intervalos de igual amplitud
 - o Tablas con intervalos de distinta amplitud
- *Centro del intervalo:*
 - o Es la semisuma de los extremos del intervalo.

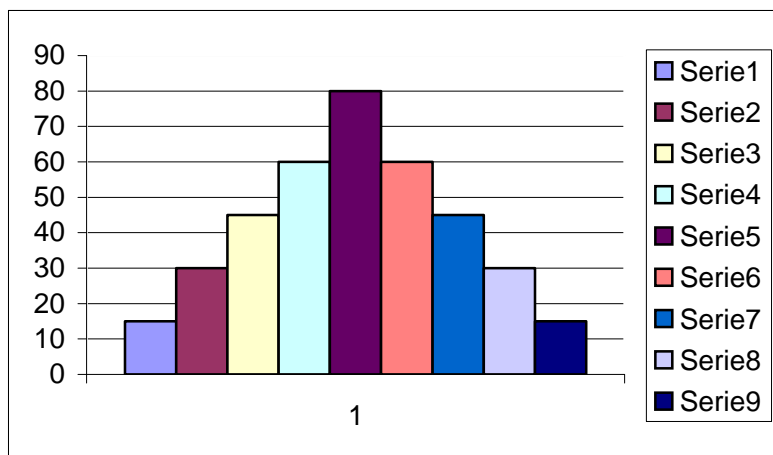
4.2 Representación gráfica de variables cuantitativas continuas

4.2.1. Histograma

Se construye con diagramas de barras a partir de la tabla de frecuencias. En él representamos los intervalos de clase en el eje horizontal (eje "x" o de abscisas) y las frecuencias correspondientes en el eje vertical (eje "y" o de ordenadas).

El área de cada barra corresponde a la frecuencia de observaciones del intervalo de clase correspondiente, aunque los intervalos de clase se muestren irregulares.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 27 de 106

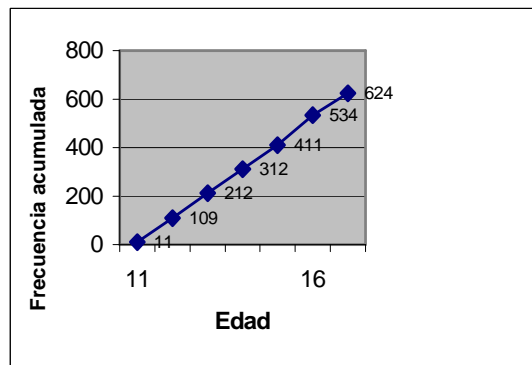
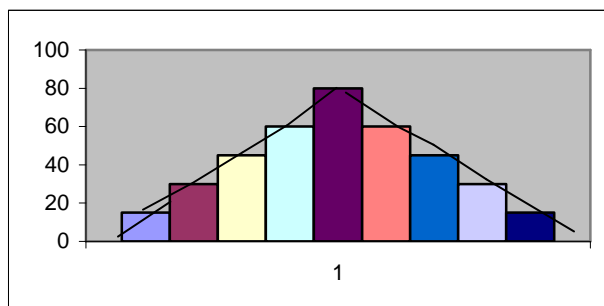


La forma del histograma es representativa de la distribución de la variable, en este caso observamos una distribución "normal" con forma de campana de Gauss.

4.2.2 Polígono de frecuencias

Uniendo con una línea el punto medio de la parte superior de cada barra del histograma obtendremos un polígono de frecuencias.

El polígono de frecuencia puede cerrarse uniendo los extremos con el eje de abscisas, correspondiente a un punto imaginario de frecuencia cero.



S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 28 de 106

4.2.3 Curva de frecuencias acumuladas

Se consigue uniendo los límites inferiores de cada intervalo de clase. Se puede obtener también con la representación lineal de los datos resultantes de la frecuencias acumuladas.

La curva de frecuencias se usa también como representación de los *percentiles* cuando se usan frecuencias acumuladas relativas.

4.2.4. Representación gráfica de dos variables cuantitativas continuas.

En este caso las dos variables varían conjuntamente, ya que al variar una, también lo hace la otra.

La representación gráfica es la **nube de puntos**, que expresa las variables en los ejes y los pares de datos como puntos. Se verá más adelante cuando hablemos de correlación y regresión.

4.3 Estadísticos en representación de V. Cuantitativas.

4.3.1 E. de posición y tendencia central de la distribución para V. Cuantitativas.

- Podemos usar los estadísticos de posición y t.c. de las variables nominales, (la **moda**), de las variables ordinales (**mediana** y **moda**) y, además, la **media aritmética**.
- **Media Aritmética:**
 - o Es la suma de los valores de todas las observaciones dividida por el número de observaciones.

$$\overline{X} = \sum x_i / n$$

- o Se emplea como estadístico en distribuciones normales o paramétricas.
- o Cuando algunos valores están muy alejados del resto puede modificarse sustancialmente la media. Sin embargo la mediana no se afectaría, por esto la mediana se emplearía en variables

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 29 de 106

que no siguen una distribución normal (variables no paramétricas).

- Para datos agrupados en intervalos, la media aritmética se calcula:

$$\bar{X} = \sum x.ni / n$$

- Existen otras medias: —
 - Media geométrica: $\bar{X} = \sqrt[n]{n.x_i}$
 - Media ponderada: $\bar{X} = \sum x.p / n$
 - Media cuadrática: $\bar{X} = \sqrt{\sum x^2 / n}$
 - Media armónica: $\bar{X} = (\sum . 1/x_i)^{-1}$

4.3.2 E. de dispersión

- Los estadísticos de dispersión nos ayudan a conocer cuanto se distancian los datos entre sí.
- Para establecer los estadísticos de dispersión se necesita un **punto de referencia**. Suele emplearse una medida de tendencia central, como la media aritmética.
- Métodos para estudiar la dispersión de los datos en variables cuantitativas:
 - La media de diferencias de cada valor con la media del grupo, por sí sola, no es un buen estadístico de dispersión, ya que al obtenerse valores negativos y positivos que podrían originar una media de 0.
 - Para asegurarnos una media distinta de cero se podría emplear la media de las diferencias absolutas (eliminando directamente el signo de los valores diferenciales).
 - Otro sistema para eliminar los signos sería elevar al cuadrado las diferencias de cada valor con la media y obtendríamos una media de las diferencias cuadráticas, también llamada **varianza** (s^2), que se define como el sumatorio de las diferencias cuadráticas respecto a la media dividido por el número de datos:

$$S^2_x = \sum (x - \bar{x})^2 / n$$

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 30 de 106

Como vemos, la varianza se expresa con unidades al cuadrado, pudiendo resultar engorroso para entender la dispersión de datos, por lo que para eliminar la expresión cuadrática aplicamos la raíz cuadrada a la fórmula de la varianza y obtenemos la $\sqrt{s^2}$, que es la **desviación típica o estándar** (s o DE).

La desviación estándar es el estadístico de dispersión más usado.

- La DE de los datos de una distribución, *per se*, no es comparable con la DE de otra distribución si no conocemos las medias respectivas, por lo que, para entender el grado relativo de dispersión empleamos el **coeficiente de variación** de Pearson (CV), que se calcula dividiendo la DE por la media aritmética y multiplicando por 100:

$$CV = (DE / \bar{x}) \cdot 100$$

4.3.3 E. de posición de un dato dentro de la distribución

- La DE nos indica la dispersión de los datos en general, pero si queremos conocer si un dato se separa mucho o poco de la media podemos investigar cuantas veces ese dato supera la DE, para ello podemos calcular la diferencia entre el dato y la media y dividirla entre la DE, obteniendo la **puntuación típica o estandarizada (Z)**:

$$Z = (x_i - \mu) / s$$

4.3.4 E. de forma

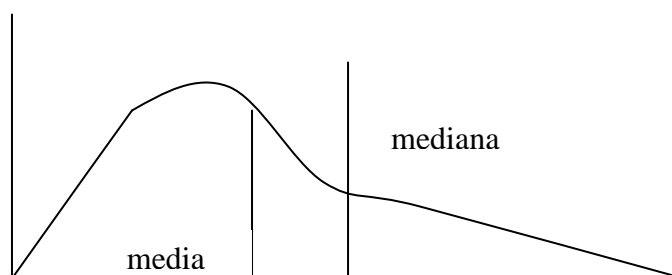
- La forma de la distribución nos orientan sobre la agrupación o situación de la mayoría de los datos.
- Asimetría o sesgo.
 - Mide la asimetría de una distribución respecto al estadístico de tendencia central (mediana).

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 31 de 106

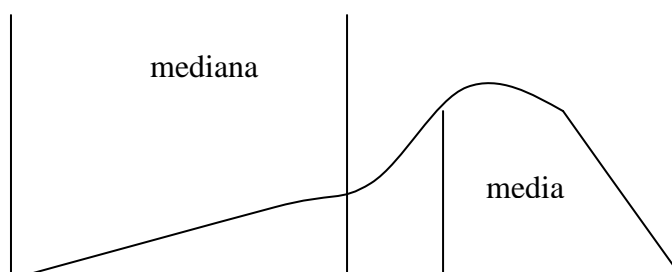
- La *asimetría positiva* indica que los valores más extremos son mayores que la media y la *asimetría negativa*, que los valores más extremos son inferiores a la media.
- El *índice de asimetría g1* viene dado por la fórmula:

$$g_1 = \{ \sum (x_i - \bar{x})^3 \} / n \cdot s^3$$

de donde, $g_1 > 0$ significa dispersión derecha: asimetría positiva



y $g_1 < 0$ corresponde a dispersión izquierda: asimetría negativa



- Apuntamiento o curtosis

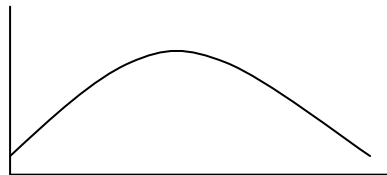
- Representa la medida en que una distribución está cargada en sus colas, comparada con una distribución normal.
- Viene determinada por la fórmula:

$$G_2 = [\{ \sum (x - \bar{x})^4 \} / n \cdot s^4] - 3$$

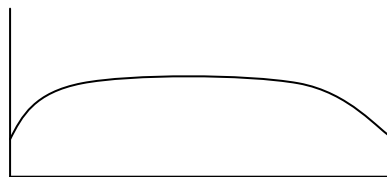
- La *curtosis positiva* indica más casos en los extremos de las colas que una distribución normal con la misma varianza.
- Se distinguen tres tipos de distribución según la curtosis:

<p>S. de Análisis Clínicos</p> <p>H.U. Reina Sofía</p>	<p>CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	<p>Código: Fecha: 01/09/2003</p>
	<p>Versión 1</p>	<p>Página 32 de 106</p>

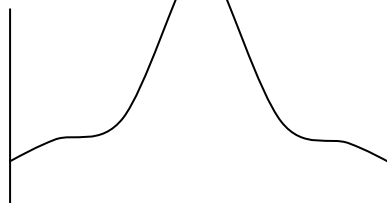
- Normal o mesocúrtica



- Aplanada o platicúrtica



- Puntiguda o leptocúrtica



S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 33 de 106

CUADRO RESUMEN REPRESENTACIÓN DE LAS VARIABLES ESTADISTICAS

	NOMINAL	ORDINAL	CUANTITATIVO
TABULACION	<u>En columnas:</u> frecuencias relativas y absolutas <u>Tablas de contingencia</u> (2 variables)	Frecuencias absolutas acumuladas (N_i) Frecuencias relativas acumuladas ($F_i = N_i/N$)	Intervalos de clase -nº óptimo: \sqrt{N} (< 10) -rango (dif. máx. y min.)
GRÁFICAS	Diagramas de barras Áreas - sectores 2 variables: agrupadas/apiladas	Diagrama de barras	Histograma Polígono de frecuencias Curva de frecuencias acumuladas Con 2 variables continuas: nube de puntos
ESTADISTICOS			
E. TENDENCIA CENTRAL	MODA (lo más frecuente) En V. Dicotómicas: -Proporción (a/N) -Razón (a/b) -Tasa: (proporción/tiempo)	MODA MEDIANA (moda ordenada) -Valor por debajo o por encima del cual está el 50% de la muestra -Ajena a extremos	MODA, MEDIANA MEDIA: <u>Aritmética</u> Geométrica Ponderada Cuadrática Armónica
E. DISPERSION	En dicotómicas con > 5 observaciones: Varianza y DE	Diferencias entre máximos y mínimos	VARIANZA (s^2) DES. ESTÁNDAR ($\sqrt{s^2}$) COEF. VARIACIÓN ($\{DE/X\}/100$)

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 34 de 106

E. POSICIÓN DATO		Frecuencias relativas acumuladas: FRACTILES PERCENTILES (% de la distribución con valores \geq que el percentil)	<u>Puntuación Z</u> (nº de veces que un dato supera la DE)
E. FORMA			ASIMETRÍA O SESGO (\longleftrightarrow) APUNTAMIENTO O CURTOSIS (\updownarrow)

4.4 DISTRIBUCIÓN DE GAUSS

- Después de revisar los conceptos estadísticos anteriores, estamos en disposición de comprender mejor el concepto y las características de la *distribución de Gauss o distribución normal*, que ya fue comentada en el punto 1.6.2.1 (páginas 9 a 11). Se menciona en este apartado para seguir con la estructura del programa de formación.
- Recordaremos las propiedades de la función de Gauss:
 - Simétrica respecto a la media.
 - Determinada por la media y la desviación típica.
 - El área total bajo la curva es 1
 - Si trazamos perpendiculares al eje X sobre s , $2s$ y $3s$ quedarían englobados respectivamente el 68, 95 y 99.7% del área de la curva.
 - Asintótica: nunca toca al eje "X" (abscisas).
 - El punto de inflexión corresponde a la desviación típica.
 -

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 35 de 106

Intervalo	Área (probabilidad)
$\mu \pm \sigma$	0.68
$\mu \pm 2 \sigma$	0.95
$\mu \pm 3 \sigma$	0.997

- Destacaremos que el área de la curva de Gauss comprendida bajo la media poblacional (μ) $\pm 2 \sigma$ contendrá el 95% de los datos, o lo que es lo mismo, cualquier dato de la población tiene un 95% de probabilidad de estar comprendido entre la media y ± 2 desviaciones estándar.

5. INFERENCIA ESTADÍSTICA

5.1 CONCEPTO

- La **estadística inferencial** es el conjunto de procedimientos por los cuales obtenemos conclusiones de tipo inductivo sobre una población en base al resultado obtenido sobre una muestra de dicha población.
- Hay dos tipos de inferencia:
 - o **Estimación**
 - Con ella solo se pretende acercarnos a la realidad de la población.
 - Podemos obtenerla calculando el rango de variación probable de los resultados obtenidos.
 - o **Test de hipótesis**
 - Se analizan estadísticos (datos muestrales) para ver si soportan o no una especulación o conjetura (hipótesis) sobre la magnitud de los parámetros (datos poblacionales).
 - Generalmente, se trata de establecer comparaciones o situaciones de igualdad o desigualdad entre variables.
- Conceptos importantes para entender la estadística inferencial:

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 36 de 106

○ Distribución muestral de medias:

- Si de una población obtenemos distintas muestras para estudiar un parámetro y de cada muestra obtenemos sus medias, podemos establecer una distribución con las medias muestrales, de manera que:
 - A mayor tamaño muestral, la variabilidad entre las medias muestrales será menor y las medias obtenidas estarán cercanas a la media poblacional.
 - Hay un gran nº de potenciales muestras aleatorias que pueden ser obtenidas de la población. Cuanto mayor sea el nº de muestras estudiadas más parecidas serán los estadísticos a los parámetros.
 - Si empleamos muestras de menos de 30 elementos la distribución de medias muestrales puede no seguir una distribución normal.
 - Las tablas de la distribución normal pueden usarse para calcular cual será la diferencia máxima entre la media muestral y la media poblacional.

- Error estándar de la media

- Cuando se realizan muestreos a partir de una población distribuida normalmente, la distribución muestral de las medias tiene varias propiedades:
 - La distribución de medias será normal.
 - La media de la distribución de medias será igual a la media de la población origen de las muestras.
 - la dispersión muestral de las medias se mide por el **error estándar** de la media, que corresponde a la desviación estándar de esta distribución.

- Teorema central del límite

- *Dada una población distribuida no normalmente de media "m" y varianza s^2 , la distribución muestral de medias de tamaño $n > 30$ seguirá una distribución normal de media m y varianza s^2/n .*

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 37 de 106

- En base a este teorema podemos realizar estimaciones válidas sobre los parámetros, aun sin conocer la distribución original de donde se obtuvo la muestra.

5.2 ESTIMACIÓN

- Tiene especial interés la ESTIMACIÓN POR INTERVALOS (intervalo de confianza o IC)
 - Definido con el rango dentro del cual se sitúa el parámetro estimado, con un determinado grado de seguridad, en el caso extremo, podríamos decir que, con un grado de seguridad del 100%, nuestro dato se sitúa entre $+\infty$ y $-\infty$.
 - Para variables cuantitativas se aplica la estimación por intervalos de confianza de medias, mientras que para variables cualitativas se emplea la estimación por intervalo de confianza de proporciones.

5.3 TEST DE HIPÓTESIS

- La hipótesis constituyen un elemento clave del método científico, mediante la cual se establece una suposición a partir de conocimientos previos para extraer determinadas conclusiones.
- En el contexto estadístico, a menudo nos interesa establecer conjeturas y comparaciones entre determinadas características de la población o entre distintas poblaciones, para ello se establecen dos tipos de hipótesis:
 - **Hipótesis nula (H_0):**
 - Supone la igualdad o que no existen diferencias entre dos variables comparadas.
 - Se formula solo para rechazarla
 - **Hipótesis alternativa (H_1):**
 - Supone la desigualdad o la existencia de diferencias.
 - Es el objetivo de la mayoría de las investigaciones y constituye una opción más arriesgada.
- Existen **test para contraste de hipótesis** que nos permiten aceptar o rechazar una determinada hipótesis en condiciones de incertidumbre.

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 38 de 106

- El Test de 2 colas o bilateral: se da si H_1 es la negación de H_0 .
- El Test de 1 colas o unilateral: se da si H_1 es una parte de la negación de H_0 .
- Además, nos informan de la probabilidad de cometer errores al tomar la decisión:
 - **Error tipo I o error alfa**: aceptar la hipótesis alternativa cuando es cierta la hipótesis nula.
 - **Error tipo II o error beta**: aceptar la hipótesis nula cuando es cierta la hipótesis alternativa.

	Ho cierta	H1 cierta
Aceptar H0	Correcto $p = 1 - \alpha$	Error tipo II $p = \beta$
Aceptar H1	Error tipo I $p = \alpha$	Correcto $P = 1 - \beta$ (potencia)

- Significado de "p"
 - El valor "p" es la probabilidad de obtener muestras aleatoriamente con una diferencia real igual o mayor a la admitida si la hipótesis nula fuera cierta.
 - Indica la probabilidad de que ocurra la H_0 y responde a la pregunta: si la H_0 fuese cierta, ¿cuál sería la probabilidad de que la diferencia observada en dos muestras seleccionadas aleatoriamente fuese igual en la población original?
 - Cuanto menor sea el valor de p es menos probable que la diferencia observada sea atribuible al azar.
 - Un valor de p es una probabilidad y, por tanto, sus valores van de 0 a 1.
 - A causa de las limitaciones en las tablas estadísticas los valores de p a menudo se dan como mayor o menor e un determinado valor umbral, arbitrariamente se ha aceptado un umbral de $p < 0.05$:
 - Si p es menor de 0.05 decimos que los resultados son **estadísticamente significativos** (no debidos al azar) y rechazamos la hipótesis nula

S. de Análisis Clínicos	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código:
H.U. Reina Sofía		Fecha: 01/09/2003
	Versión 1	Página 39 de 106

5.4 ESQUEMAS SOBRE LOS TIPOS DE TEST DE HIPÓTESIS

5.4.1 SEGÚN VARIABLE CLASIFICADORA Y VARIABLE A ESTUDIO

	COMPARACIÓN				ASOCIACIÓN
	2 CATEGORÍAS		N CATEGORÍAS		
	Independientes	Pareadas	Independientes	Pareadas	
CUANTITATIVA	t de Student	T Student pareada	ANOVA Simple/factor	ANOVA	Correlación de Pearson Regresión
ORDINAL	U de Mann Whitney	Wilcoxon	K de Kruskall Wallis	Friedman	Correlación de Spearman Passin-Bablock
CUALITATIVA	Z, Fisher, X ²	Mac Neumar	X ²	Q Cochran	

5.4.2 PROCEDIMIENTOS EN ESTADÍSTICA DESCRIPTIVA E INFERENCIAL, SEGÚN TIPO Y NÚMERO DE MUESTRAS.

5.4.2.1 PROCEDIMIENTOS DESCRIPTIVOS Y GRAFICOS

	Tipo de datos	PROCEDIMIENTOS DESCRIPTIVOS Y GRÁFICOS
1 muestra	Datos normales	Media, DE, IC, Histogramas, polígono frecuencias
	Datos no normales	Mediana, Box Plot
	Datos cualitativos	Frecuencias y barras
2 muestras relacionadas	Datos normales	Correlación Pearson, X-Y nube de puntos
	Datos no normales	Correlación Spearman, Passin-Bablock, X-Y nube de puntos.
	Datos cualitativos	Tablas de contingencias, barras

5.4.2.2 TEST DE COMPARACIÓN

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 40 de 106

		Tipos de datos	TEST DE COMPARACIÓN
2 muestras	2 muestras relacionadas	Datos normales	t de Student pareada
		Datos no normales	Friedman
		Datos dicotómicos	McNemar
	2 muestras independientes	Datos normales	t de Student independientes
		Datos no normales	U Mann Whitney
		Datos cualitativos	χ^2 Homogeneidad
> 2 muestras	Muestras relacionadas	Datos normales	ANOVA medidas repetidas
		Datos no normales	Friedman
		Datos dicotómicos	Q Cochran
	Muestras independientes	Datos normales	ANOVA grupos independientes
		Datos no normales	Kruskal-Wallis
		Datos cualitativos	χ^2

5.4.2.3 TEST DE ASOCIACIÓN

	Tipo de datos	TEST DE ASOCIACIÓN
2 muestras relacionadas	Datos normales	Correlación de Pearson Regresión lineal
	Datos no normales	Correlación de Spearman Passin-Bablock
	Datos cualitativos	χ^2 Independencia
> de 2 muestras relacinadas	Datos normales	Regresión múltiple
	Datos no normales	Correlación parcial de Kendall
	Datos cualitativos	Análisis discriminante

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 41 de 106

5.4.3. CARACTERÍSTICAS DE LOS TEST DE HIPÓTESIS MÁS EMPLEADOS

5.4.3.1. PRUEBA T DE STUDENT / F DE SNEDECOR

RECUERDO HISTÓRICO

Estas pruebas claves para los estudios de comparación de medias se desarrollaron entre los siglos IX y XX:

- La Prueba T fue desarrollada en 1899 por William Sealey Gosset (1876-1937), que publicaba sus trabajos bajo el pseudónimo de "STUDENT", para evitar la discriminación laboral de la época.
- La Prueba F, fue desarrollada por el matemático/estadístico George W. Snedecor (1882-1974)

Para su aplicación es necesario que las variables de estudio sigan una distribución normal. Otros investigadores desarrollaron pruebas para corroborar tal circunstancia, que hoy se conocen como "Pruebas de bondad de ajuste":

- La prueba más empleada es la de Kolmogorov-Smirnov, desarrollada por Nikolai Vasilyevich Smirnov (1890-1966) y Andrei Nikolaevich Kolmogorov* (1903-1987) y publicada en la obra "Teoría de la medida" (1920)
- La prueba de Shapiro-Wilks fue desarrollada por Samuel S. Shapiro y Martin .B. Wilks y dada a conocer en la obra "Biometrika", en 1965

CONCEPTO

Las pruebas T y F son aplicadas para la comparación de dos medias correspondientes a dos muestras que siguen una distribución normal, con el objeto de estudiar si éstas pertenecen a la misma población.

Las muestras pueden ser *independientes*, no están relacionadas (por ejemplo, cuando comparamos los resultados de colesterol de una muestra de hombres y de una muestra de mujeres), o *dependientes o pareadas*, están relacionadas (por ejemplo, cuando queremos saber el efecto de un fármaco

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 42 de 106

hipolipemiente sobre la concentración de colesterol en sangre de un grupo de pacientes, las muestras son pareada por que se realizan sobre los mismos sujetos, antes y después de la toma del fármaco)

La diferencia que se plantea se establece en términos estadísticos (diferencia significativa estadísticamente o no) y debe interpretarse en el contexto de la probabilidad. La prueba T da un valor de p y un intervalo de confianza, de forma que cuando la probabilidad de ocurrencia de una diferencia como la obtenida sea inferior al 5% se dice que el resultado es estadísticamente significativo y que existe diferencia entre las medias.

Para el cálculo de la prueba T es necesario establecer una variable como independiente (variable explicativa, de naturaleza cualitativa dicotómica) y la otra como variable dependiente (variable resultado, de naturaleza cuantitativa normal)

Ejemplo: comparación de dos variables (dos tratamientos, dos métodos analíticos, etc.).

Características de la prueba T de Student

- Es un modelo en el que una *variable explicativa* (var. independiente) dicotómica intenta explicar una *variable respuesta* (var. dependiente) cuantitativa
- En cada grupo *la variable estudiada debe seguir una distribución normal*
- *La dispersión en ambos grupos debe ser homogénea* (hipótesis de homocedasticidad = igualdad de varianzas).
- Debemos realizar el cálculo de *estadísticos descriptivos previos* (e.d.p.): el número de observaciones, la media y la desviación típica en cada grupo
- A partir de e.d.p. se calcula el *estadístico de contraste experimental* (e.c.e.).
- A partir de e.c.e se busca en las tablas el *p-valor asociado*
 - si el valor p asociado es < de 0,05 se considera la existencia de diferencia significativa

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 43 de 106

Para contrastar la normalidad de los datos se pueden aplicar varias pruebas estadísticas, como la de Kolmogorov-Smirnov (la más empleada) o la de Shapiro-Wilks

-Si se comprobara que las muestras no siguen una distribución normal podríamos hacer dos cosas:

- Normalización de las muestras, mediante
 - Transformación logarítmica (neperiana)
 - Elevación al cuadrado de los datos
 - Extracción de raíz cúbica de los datos
- Utilizar pruebas estadísticas no paramétricas
 - U de Mann Whitney (muestras independientes)
 - Wilcoxon (muestras dependientes)

Para comprobar si la dispersión de ambos grupos es homogénea se recurre a los *test de homogeneidad de las varianzas*, como el test de la F se Snedecor, test de Levene, etc.

-Si se concluye que las varianzas son homogéneas se podrá aplicar la prueba T.

-Si se concluye que las varianzas no son homogéneas no se debe aplicar el T, ya que la diferente dispersión entre grupos no los hace comparables, en este caso habría que aplicar una corrección aplicando la prueba t de Student debida a Satterthwaite.

La prueba t de Student es muy utilizada en la práctica, sin embargo a menudo su aplicación se hace sin excesivo cuidado, no comprobando las asunciones que requiere.

-Ejemplo:

Se supone que se quiere comparar dos tratamientos con relación a una variable cuantitativa. Los datos experimentales son:

Trat A: 25, 24, 25, 26

Trat B: 23, 18, 22, 28, 17, 25, 19, 16

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 44 de 106

Si se aplica la t de Student directamente se obtiene una $p=0,096 > 0,05$ con lo que se concluye que no se puede demostrar diferencias entre los dos tratamientos.

Por otro lado, la prueba de Levene pone de manifiesto que $p=0,014 < 0,05$ y se concluye que en estos datos no se verifica la igualdad de varianzas (heterocedasticidad), con lo que la conclusión anterior queda en suspenso.

Tras aplicar Satterthwaite, que es válido en este caso de heterocedasticidad, se obtiene que $p=0,032 < 0,05$ con lo que la conclusión correcta es que sí hay diferencia entre los dos tratamientos.

Intervalo de confianza para la diferencia de medias y prueba t-student para dos medias.

Cálculo de los estadísticos descriptivos básicos

Si se llama n_1 y n_2 a los tamaños muestrales del primer y del segundo grupos, las medias y las desviaciones típicas para los dos grupos son:

$$\bar{x}_1 = \frac{\sum x_{1i}}{n_1}$$

$$\bar{x}_2 = \frac{\sum x_{2i}}{n_2}$$

$$s_1 = \sqrt{\frac{1}{n_1 - 1} \sum (x_{1i} - \bar{x}_1)^2}$$

$$s_2 = \sqrt{\frac{1}{n_2 - 1} \sum (x_{2i} - \bar{x}_2)^2}$$

donde x_{1i} indica los valores de la variable respuesta para el grupo 1 y x_{2i} indica los valores de la variable respuesta para el grupo 2.

Cálculo del Intervalo de Confianza (IC) $(1 - \alpha)\%$ para la diferencia de medias suponiendo igualdad de varianzas

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 45 de 106

Para calcular el IC(1 - α)% para la diferencia de medias se necesita calcular el error estándar de la diferencia de medias que, en el supuesto de igualdad de varianzas, tiene la expresión:

$$EE(\bar{X}_1 - \bar{X}_2) = \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

donde s^2 recibe el nombre de varianza conjunta ("pooled variance")

En segundo lugar para calcular el IC deseado se necesita el valor de la T-Student $T_{1-\alpha/2;gl}$ con grados de libertad $gl = (n_1 - 1) + (n_2 - 1) = (n_1 + n_2 - 2)$, con lo que:

$$IC(1 - \alpha)\%(\bar{X}_1 - \bar{X}_2) = \left[(\bar{X}_1 - \bar{X}_2) \pm t_{1-\alpha/2,gl} EE(\bar{X}_1 - \bar{X}_2) \right]$$

Cálculo del IC(1 - α)% para la diferencia de medias suponiendo no igualdad de varianzas

El error estándar de la diferencia medias en el caso de no igualdad de varianzas:

$$EE(\bar{X}_1 - \bar{X}_2) = \sqrt{EE(\bar{X}_1)^2 + EE(\bar{X}_2)^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

con lo que IC sería:

$$IC(1 - \alpha)\%(\bar{X}_1 - \bar{X}_2) = \left[(\bar{X}_1 - \bar{X}_2) \pm t_{1-\alpha/2,gl} EE(\bar{X}_1 - \bar{X}_2) \right]$$

Cálculo de la prueba T-Student para la diferencia de medias suponiendo igualdad de varianzas

Para llevar a cabo el contraste:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

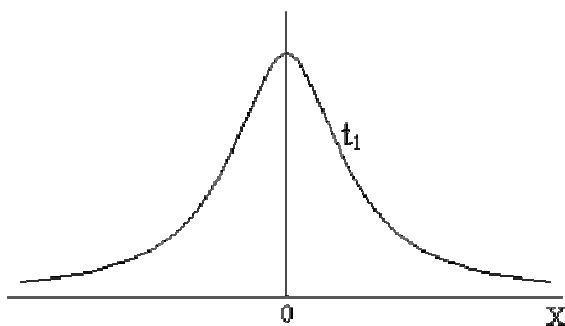
suponiendo igualdad de varianzas poblacionales, se construye el estadístico de contraste experimental T dado por:

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 46 de 106

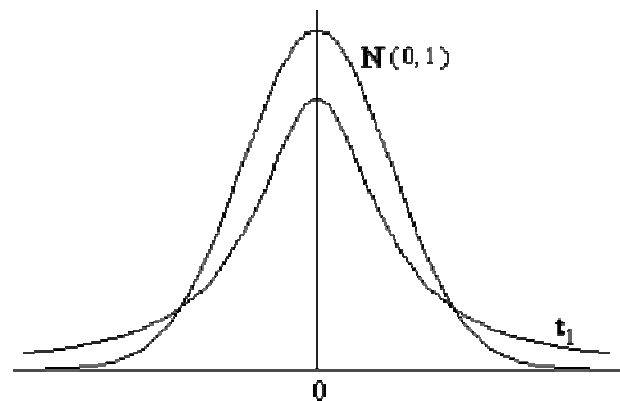
$$t = \frac{\bar{X}_1 - \bar{X}_2}{EE(\bar{X}_1 - \bar{X}_2)} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

que bajo la hipótesis nula sigue una distribución T-Student con grados de libertad $gl = (n_1 - 1) + (n_2 - 1) = (n_1 + n_2 - 2)$.

Comparación entre las funciones de densidad de t_1 y $N(0,1)$.



Función de densidad de una de Student



- Es de media cero, y simétrica con respecto a la misma;
- Es algo más dispersa que la normal, pero la varianza decrece hasta 1 cuando el número de grados de libertad aumenta;

Cálculo de la prueba T-Student para la diferencia de medias suponiendo no igualdad de varianzas

Para llevar a cabo el contraste:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

suponiendo no igualdad de varianzas poblacionales, se construye el estadístico de contraste experimental t dado por:

<p>S. de Análisis Clínicos</p> <p>H.U. Reina Sofía</p>	<p>CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	<p>Código: Fecha: 01/09/2003</p>
	<p>Versión 1</p>	<p>Página 47 de 106</p>

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

que bajo la hipótesis nula sigue una distribución T-Student con grados de libertad "gl" dados por:

$$gl = \frac{[EE(\bar{X}_1 - \bar{X}_2)]^4}{\frac{1}{n_1 - 1} [EE(\bar{X}_1)]^4 + \frac{1}{n_2 - 1} [EE(\bar{X}_2)]^4}$$

que recibe el nombre de **grados de libertad de Satterthwaite**.

Intervalo de confianza para el cociente de varianzas y prueba f-Snedecor para dos varianzas

Cálculo del IC(1 - a)% para el cociente de varianzas

La expresión para calcular el IC(1 - a)% para el cociente de varianzas es:

$$IC95\% \left(\frac{\sigma_1^2}{\sigma_2^2} \right) = \left(\frac{S_1^2 / S_2^2}{F_{1-\alpha/2; gl_1; gl_2}}, \frac{S_1^2}{S_2^2} F_{1-\alpha/2; gl_1; gl_2} \right)$$

Cálculo de la prueba F-Snedecor para la igualdad de varianzas

Para llevar a cabo el contraste: $H_0: \sigma_1 - \sigma_2 = 0$

$H_1: \sigma_1 - \sigma_2 \neq 0$

mediante la prueba F-Snedecor de comparación de varianzas se construye el estadístico de contraste experimental F dado por:

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 48 de 106

$$F = \frac{\max \{s_1^2; s_2^2\}}{\min \{s_1^2; s_2^2\}}$$

que bajo la hipótesis nula sigue una distribución F-Snedecor. En el caso de no poder rechazar la hipótesis nula (p-valor > 0.05) se considera que las dos varianzas son iguales (homogéneas).

Ejemplos:

En un ensayo clínico para evaluar un hipotensor se compara un grupo placebo con el grupo tratado. La variable medida es la disminución de la presión sistólica y se obtiene: grupo placebo n = 35; $\bar{x} = 3,7$ mm de Hg. y $s^2 = 33,9$; grupo tratado n = 40; $\bar{x} = 15,1$ mm de Hg. y $s^2 = 12,8$. ¿Es eficaz el tratamiento?

Se trata de un contraste sobre diferencias de medias :

$$H_0: m_T - m_P = 0$$

$$H_1: m_T - m_P \neq 0$$

Como no conocemos las varianzas, para realizarlo debemos decidir si son iguales o distintas, para ello se plantea el contraste

$$H_0: \sigma_T^2 = \sigma_P^2$$

$$H_1: \sigma_T^2 \neq \sigma_P^2$$

El estadístico es $F = s_P^2 / s_T^2 = 33,9 / 12,8 = 2,65$

para el que $p < 0,05$, en consecuencia rechazamos la H_0 y concluimos que las varianzas son distintas. Por lo tanto usaríamos la t para varianzas distintas. Haciendo los cálculos $t = -10,2$ $p < 0,05$ rechazamos la H_0 y concluimos que las medias son distintas

POR SUERTE, LOS MODERNOS PROGRAMAS ESTADÍSTICOS REALIZAN TODOS LOS CÁLCULOS REPRESENTADOS POR LA FÓRMULAS ANTERIORES

LO IMPORTANTE ES CONOCER LOS CONCEPTOS NECESARIOS PARA APLICAR UNOS ESTADÍSTICOS U OTROS EN CADA CASO

<p>S. de Análisis Clínicos</p> <p>H.U. Reina Sofía</p>	<p>CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p> <p>Versión 1</p>	<p>Código: Fecha: 01/09/2003</p> <p>Página 49 de 106</p>
--	--	--

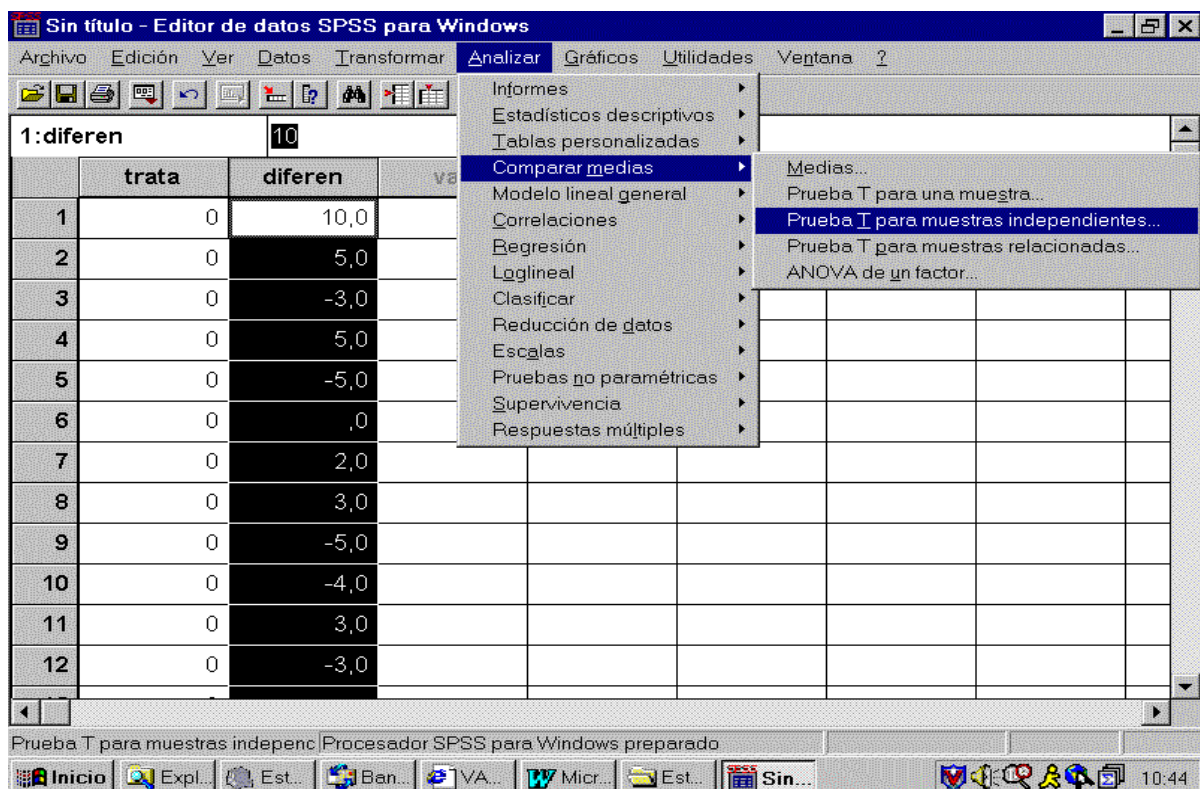
Este problema se podría resolver empleando un paquete estadístico del tipo del SPSS:

Resolución del problema anterior con paquete estadístico SPSS

1) Deberíamos crear un archivo con 2 variables:

- "Tratamiento", con un código distinto para cada grupo, p.e. 0 para placebo y 1 para tratado
- "Diferencia", con la diferencia de presión arterial para cada individuo al empezar el estudio y al acabar.

2) Para calcular la t desplegamos los menús que se ven en la gráfica:



Y el programa calcula la t para varianzas iguales y distintas y realiza el contraste para las varianzas.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES		Código: Fecha: 01/09/2003
	Versión 1		Página 50 de 106

Para el contraste sobre las varianzas el SPSS no usa la prueba descrita más arriba, sino la **de Levene** que no asume normalidad y se puede usar para comparar varias varianzas

Estadísticos del grupo

	TRATAMIENTO	N	Media	Desviación típ.	Error típ. de la media
DIFERENCIA	0	35	3,729	5,666	,958
	1	40	15,075	3,576	,565

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Intervalo de confianza para la diferencia	
									Inferior	Superior
DIFERENCIA	Se han asumido varianzas iguales	10,431	,002	-10,503	73	,000	-11,346	1,080	-13,500	-9,193
	No se han asumido varianzas iguales			-10,201	55,909	,000	-11,346	1,112	-13,575	-9,118

Aplicación de prueba T de hoja de cálculo EXCEL

Suponga que en los pasados seis meses algunos de sus empleados nuevos hayan asistido a un seminario de entrenamiento en Boston y otros en New

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 51 de 106

York. Al final del seminario, todos los empleados tomaron un examen para obtener el certificado. El seminario en Boston es más caro, pero en general usted piensa que el entrenamiento que se ofrece en Boston es mejor que el que se ofrece en New York. Usted ha recibido los resultados de las calificaciones de 15 empleados que estudiaron en Boston y de 15 empleados que estudiaron en New York. Basadas en es estas calificaciones, ¿se puede comprobar que el programa de Boston es mejor que el programa de New York?

Persona	Boston	New York
1	99	98
2	99	96
3	98	96
4	97	95
5	90	85
6	85	80
7	84	79
8	82	78
9	81	75
10	79	73
11	79	72
12	68	69
13	61	67
14	60	62
15	56	60
Promedio	81.2	79
Desv. Est.	14.4973	12.6152

La función de TTEST calcula la probabilidad asociada con la prueba t de Student para determinar la probabilidad de que dos muestras procedan de dos poblaciones subyacentes. La función pide lo siguiente: **TTEST(Array1, Array 2, tails, type) [PRUEBA.T(matriz1, matriz2, colas, tipo)]**:

Array 1 es el primer conjunto de datos, el cual en este ejemplo son las calificaciones de Boston.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 52 de 106

Array 2 es el segundo conjunto de datos, el cual en este ejemplo son las calificaciones de New York.

Tails especifica el número de colas de distribución. Si el argumento colas = 1, TTEST utiliza la distribución de una cola. Si colas = 2, TTEST utiliza la distribución de dos colas. En este ejemplo se supone 2 colas ya que la diferencia puede ser positiva o negativa.

Type es el tipo de prueba t que se realiza: 1 = Observaciones por pares; 2 = Observaciones de dos muestras con varianzas iguales; y 3 = Observaciones de dos muestras con varianzas diferentes. En este ejemplo se supone dos muestras con varianzas iguales.

Como resultado, la función de este ejemplo es la siguiente: TTEST(B2:B16, C2:C16, 2, 2). La probabilidad asociada con el valor t es de 0.6609. Ya que el valor no es menor de 0.05, no podemos decir que el entrenamiento en Boston es significativamente mejor que el entrenamiento de New York. Además, basada en esta información sería difícil justificar el entrenamiento más caro de Boston

5.4.3.2. ANALISIS DE VARIANZA (ANOVA)

El análisis de la varianza (**analysis of variance o ANOVA**) es una técnica estadística de contraste de hipótesis y puede ser vista como una generalización del test de Student

ANOVA y Regresión Lineal Múltiple (RLM) marcan el comienzo de técnicas multivariantes.

¿Por qué ANOVA y no repeticiones de pruebas T dos a dos?

- Evita incrementar del riesgo de dar un resultado falso positivo
- Es difícil interpretar la verdadera influencia de la variable que actúa como factor de clasificación

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 53 de 106

- Se analiza globalmente la influencia de cada variable independiente (único nivel de significación)

¿Qué estudia ANOVA?

En el ANOVA se comparan medias, pero se analizan varianzas (Entre subgrupos generados por factores de clasificación)

De forma parecida a lo que ocurre en la prueba T, en el ANOVA se establecen varios tipos de variables:

- Variable(s) independiente(s) o factor: variables categóricas que agrupan datos.
- Variable dependiente: variable aleatoria cuantitativa continua (a examen)

La hipótesis alternativa es múltiple, ya que la pregunta que se plantea es si existen diferencias entre alguno de los subgrupos

- Se trata de estudiar si ¿La variabilidad observada en los datos se debe al azar o es debida al efecto del factor?
 - o En cada grupo existe una *varianza intraclass o varianza residual*
 - o Entre cada grupo existe una *varianza interclase o varianza explicativa*
 - o Se compararan las varianzas empíricas de cada muestra con la varianza de la muestra global
 - Se puede cuantificar la ley de las diferentes componentes de la varianza, empleando el hecho de que la suma de dos variables independientes que siguen dos leyes de chi-cuadrado sigue también una ley de chi-cuadrado, y que su cociente ponderado sigue una ley de Fisher.
 - o Si los tratamientos tienen efectivamente un efecto, se espera que la varianza explicada sea grande en comparación con la varianza residual.

Si se encuentran diferencias entre subgrupos ($p < 0,05$), se deben aplicar test "a posteriori", para estudiar entre qué grupos se establecen diferencias.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 54 de 106

- Bonferroni (Método de ajuste para contrastes múltiples más utilizado rutinariamente)
 - o Ajuste de Bonferroni: Técnica estadística que ajusta el nivel de significación en relación al número de pruebas estadísticas realizadas simultáneamente sobre un conjunto de datos. El nivel de significación para cada prueba se calcula dividiendo el error global de tipo I entre el número de pruebas a realizar.
- Tukey (Método de Tukey o Método de la Diferencia Significativa Honesta de Tukey (DSH))
 - o En esta prueba se utiliza un sólo valor con el cual se comparan todos los posibles pares de medias. El método de comparación de Tukey fue reformado por Kramer (1956) para casos en el que el número de réplicas no es igual (Tukey-Kramer)
- Scheffé (Tamaño de muestras desiguales)
- Dunnett

Tipos de ANOVA

- Análisis de la varianza (con factor)
 - o Un factor : ANOVA simple (de una vía)
 - o Varios factores : ANOVA factorial (de dos vías)
 - Estudio de "efecto aditivo"
 - Estudio de "efecto multiplicativo"
 - Estudio de "efecto interacción"
 - Análisis de COVARIANZA (corrección o ajuste)
- Análisis de medidas repetidas
 - o Se emplea para estudiar los factores "intra sujetos"
 - Todos los niveles del factor se aplican a los mismos sujetos.
 - Pueden tener más de dos medidas y más de un factor

Condiciones para aplicar ANOVA

- Las observaciones proceden de poblaciones normales
- Las muestras son aleatorias e independientes. Además, dentro de cada nivel las observaciones son independientes entre sí.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 55 de 106

- Hipótesis de homocedasticidad

Ejemplo: análisis de varianza, un factor (i)

Se quiere evaluar la eficacia de distintas dosis de un fármaco contra la hipertensión arterial, comparándola con la de una dieta sin sal. Para ello se seleccionan al azar 25 hipertensos y se distribuyen aleatoriamente en 5 grupos. Al primero de ellos no se le suministra ningún tratamiento, al segundo una dieta con un contenido pobre en sal, al tercero una dieta sin sal, al cuarto el fármaco a una dosis determinada y al quinto el mismo fármaco a otra dosis. Las presiones arteriales sistólicas de los 25 sujetos al finalizar los tratamientos son:

Grupos

1	2	3	4	5
180	172	163	158	147
173	158	170	146	152
175	167	158	160	143
182	160	162	171	155
181	175	170	155	160

La tabla de ANOVA es:

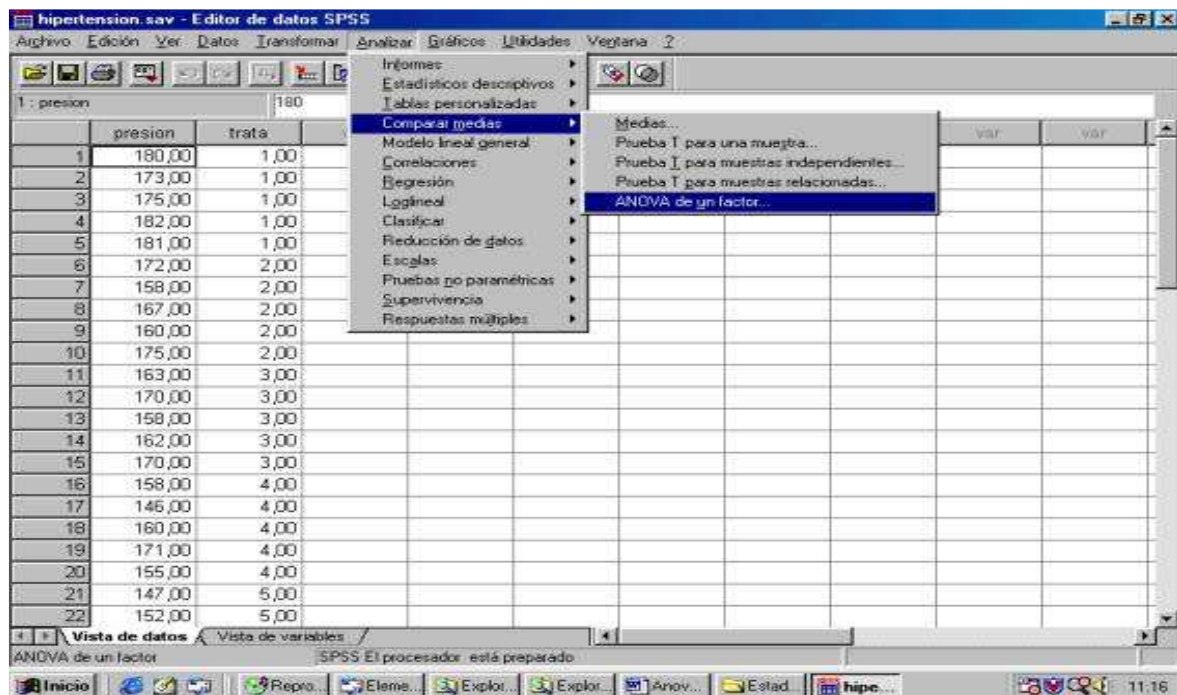
Fuente de variación	GL	SS	MS	F
Tratamiento	4	2010,64	502,66	11,24
Error	20	894,4	44,72	
Total	24	2905,04		

Como $F_{0,05}(4,20) = 2,87$ y $11,24 > 2,87$ rechazamos la hipótesis nula y concluimos que los resultados de los tratamientos son diferentes.

Ejemplo con el programa SPSS: análisis de varianza, un factor

Para hacerlo con un paquete estadístico, p.e. el **SPSS**, deberíamos crear un archivo con 2 variables: **Tratamiento** (con un código distinto para cada grupo, p.e. de 1 a 5) y **Presión** con la presión arterial de cada individuo al acabar el estudio. Para calcular el ANOVA desplegamos los menús que se ven en la imagen:

<p>S. de Análisis Clínicos</p> <p>H.U. Reina Sofía</p>	<p>CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	<p>Código: Fecha: 01/09/2003</p>
	<p>Versión 1</p>	<p>Página 56 de 106</p>



La tabla de *ANOVA* que presenta el programa es:

PRESION

	Suma de cuadrados	gl	Media cuadrática	F	Sig
Inter-grupos	2010,640	4	502,660	11,240	,000
Intra-grupos	894,400	20	44,720		
Total	2905,040	24			

que incluye también el "valor p" asociado al contraste.

Por otro lado, podemos observar que el cociente entre las medias uadráticas "intergrupos" e "intragrupos" ($502,660 / 44,720$) da el valor F (11,24) que lleva asociado un valor p (0,000).

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 57 de 106

5.4.3.3. ESTUDIO DE RELACIÓN ENTRE VARIABLES

Muy a menudo, en la vida real, se encuentra en la práctica que existe una relación entre dos (o más) variables y se necesita una función o fórmula que represente y cuantifique esa relación.

Ejemplos: $L = 2\pi r$
 $E = m_0 c^2$

La estadística puede dar respuesta a la necesidad de conocer relación en condiciones de incertidumbre, aplicando diversas "Medidas de relación":

- CORRELACIÓN: para encontrar la medida de su relación
- REGRESIÓN: para encontrar una "fórmula" que explique en qué medida varía uno en función de las variaciones de la otra
 - o Pueden ser:
 - Lineal / no lineal
 - Simple / múltiple

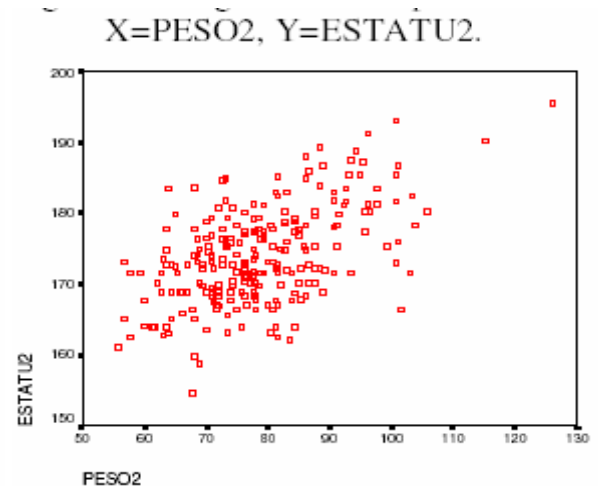
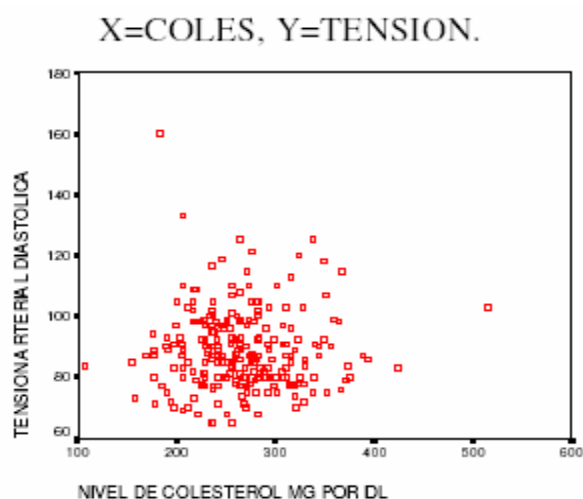
5.4.3.3.1. CORRELACIÓN

Su objetivo es estudiar la relación existente entre dos **variables de tipo cuantitativo**.

Métodos de expresión y cuantificación

- Diagrama de dispersión o nube de puntos (Representación gráfica)
 - o El diagrama de dispersión o nube de puntos es la representación gráfica de los pares de valores $X_i Y_j$.
 - o Permite observar la dirección de la posible relación (relación positiva o negativa)
 - o También podemos intuir la forma de la relación.
 - La forma más sencilla es la lineal
 - Las demás formas serán no lineales, parábolas, curvas potenciales, exponenciales

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 58 de 106



- Coefficiente de correlación

- Este coeficiente *es una medida* (un número) que expresa la relación lineal entre dos variables cuantitativas X e Y y se representa por la letra " r ":

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

$Cov(X,Y)$ es la covarianza entre X e Y
 S_X es la desviación típica de X
 S_Y es la desviación típica de Y

- El valor de " r " puede tomar cualquier valor entre -1 y 1
 - Si r toma valores cercanos a 0, entonces no existe relación lineal entre las dos variables. Hay que pensar que aunque sea nulo X e Y pueden mantener otro tipo de relación no lineal.
 - Si r toma valores cercanos a 1, entonces existe una relación lineal directa o positiva entre X e Y
 - Si r toma valores cercanos a "-1", entonces existe una relación lineal inversa o negativa entre X e Y .
- Existen distintos métodos en función de la distribución de las variables:
 - El C. de correlación empleado en condiciones paramétricas es el de PEARSON
 - El C. de correlación empleado en condiciones no paramétricas es el de SPEARMAN

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 59 de 106

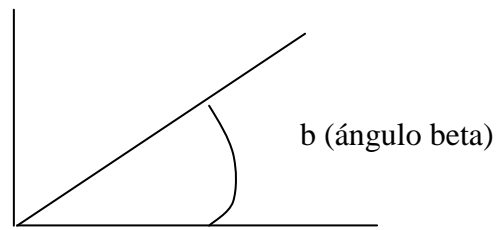
5.4.3.3.2. REGRESIÓN

La regresión es una técnica estadística que permite el estudio de la relación entre dos variables cuantitativas X e Y .

La relación entre dos variables puede estar representada por una función matemática: $Y = f(X)$, que puede ser exponencial, lineal, etc.

- La función matemática más frecuentemente representativa de la relación entre dos variables cuantitativas en el entorno biológico es la *ecuación de la línea recta*:

$$y = bx + a$$



a ($= 0$ en este caso, pero puede adoptar cualquier valor)

donde " b " representa la pendiente de la recta y " a " el punto en el origen

- En el modelo de regresión que utiliza la ecuación de la línea recta se emplea el criterio de los "*mínimos cuadrados*" para construir la recta que mejor prediga los valores de Y en función de los de X , es decir, el criterio de los mínimos cuadrados consiste en construir una línea recta (determinar a y b) de tal forma que la suma de los errores (diferencias) al cuadrado sea mínima.
 - Se llamaría "*residuo*" *ei* del individuo i a la diferencia entre el valor de Y_i que realmente toma y el que predice la recta que va a tomar, $i \hat{Y}$. El residuo puede ser negativo, positivo o nulo.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 60 de 106

REGRESIÓN LINEAL

- Consiste en obtener una función lineal en la que se relaciona una variable que se quiere explicar (v. dependiente o explicada) con otra/s ya conocidas (v. independiente o explicativa).
- El nº de variables explicativas define el modelo de REGRESIÓN LINEAL
 - o Una sola variable explicativa se aplica un modelo de R.L. SIMPLE
 - o Varias variables explicativas se aplica un modelo de RL MÚLTIPLE
- Todas las variables deben ser cuantitativas.
- Antes de aplicar una regresión lineal es conveniente comprobar la correlación existente entre las variables

Ejemplo de regresión lineal:

Se desea saber la cantidad de ingresos hospitalarios que tendrá un hospital en un mes (variable explicada) en función de la población que existe, edad media de la misma, etc. (variable explicativas)

REGRESIÓN LINEAL SIMPLE (RLS)

- Modelo de regresión con una sola variable independiente que explica a la v. dependiente.
- Se desea encontrar la función lineal que mejor prediga los valores de la variable Y en función de los valores de la variable X
 - o Para ello la diferencia entre los valores de Y calculados (\hat{Y}) en función de X deben ser lo más parecidos a los valores observados de Y:

$$\hat{y} = a + bx \text{ (ecuación de la línea recta)}$$

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 61 de 106

El objetivo es encontrar la mínima distancia $d = y - \hat{y}$, para conseguirlo se aplica el criterio de los mínimos cuadrados anteriormente expuesto

$$d = (y - \hat{y})^2 = (y - (a + bx))^2 \text{ sea un mínimo}$$

Ejemplo de regresión lineal simple:

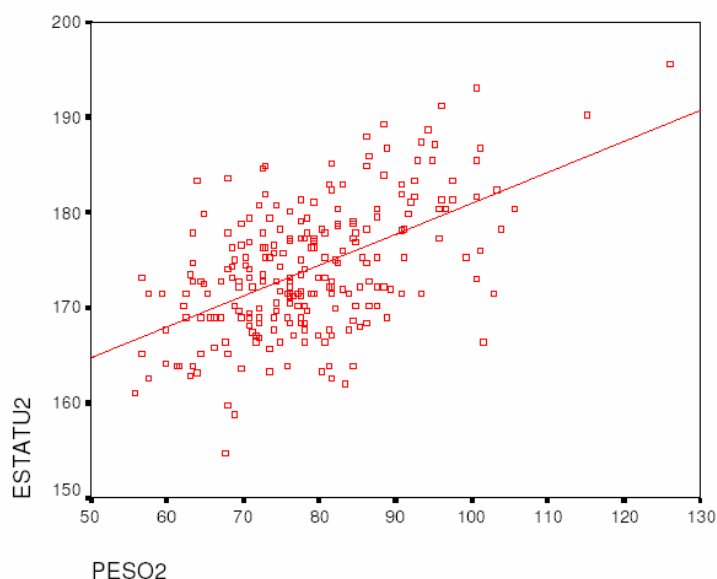
Recta de regresión mínimo cuadrática entre PESO2 (variable independiente o explicativa) y ESTATU2 (variable dependiente o explicada), se quiere estudiar la diferencia de estatura en función del peso, con SPSS y el diagrama de dispersión con la recta superpuesta

MODEL: MOD_1.

Independent: PESO2

Dependent	Mth	Rsq	d.f.	F	Sigf	b0	b1
ESTATU2	LIN	.287	238	95.7	0.000	148.564	.3237

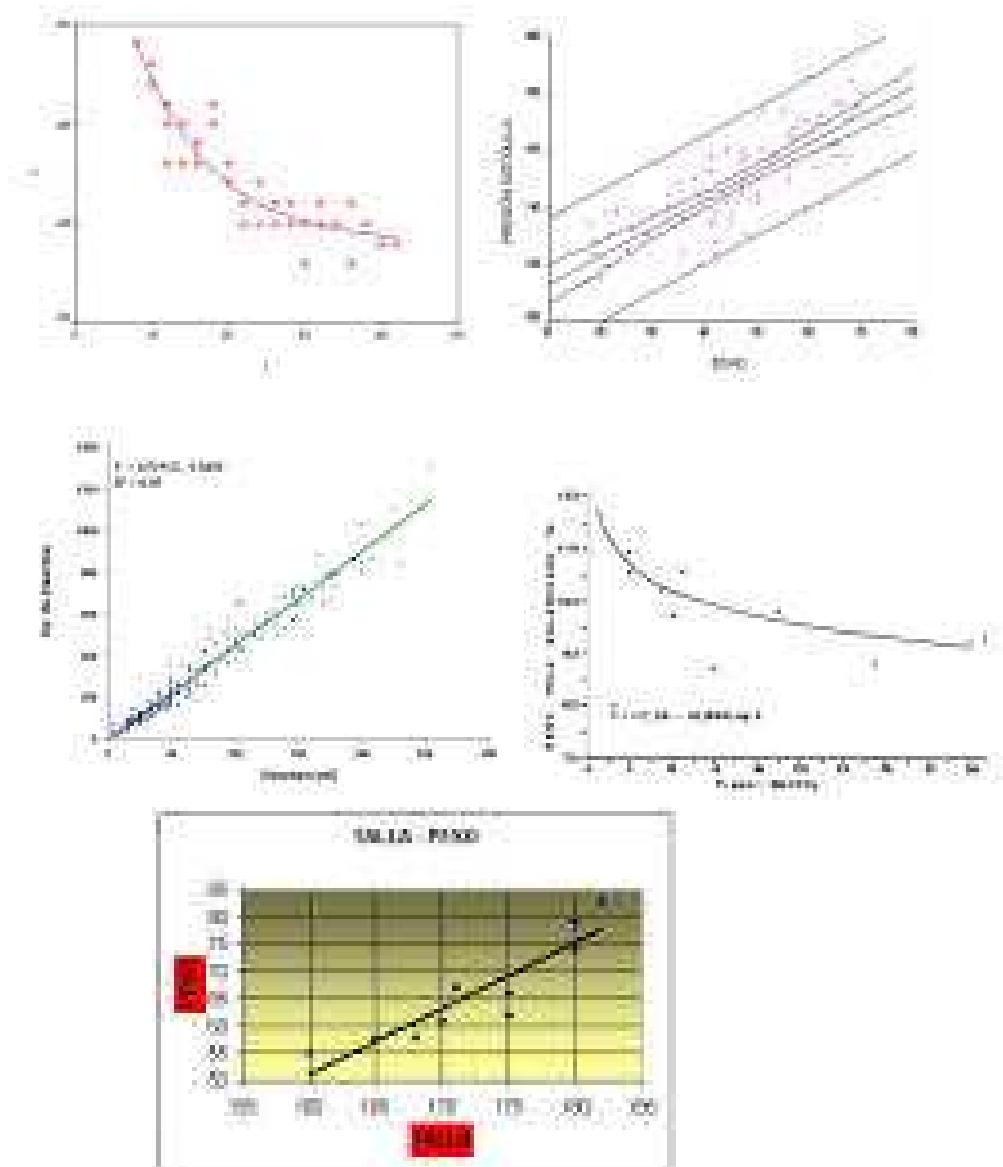
$$Y = 148.564 + (0.3237)x$$



<p>S. de Análisis Clínicos</p> <p>H.U. Reina Sofía</p>	<p>CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	<p>Código: Fecha: 01/09/2003</p>
	<p>Versión 1</p>	<p>Página 62 de 106</p>

Los distintos modelos de regresión lineal no siempre siguen la función de la línea recta.

Podemos encontrar distintas representaciones gráficas de modelos de regresión lineal:



S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 63 de 106

BONDAD DEL AJUSTE. COEFICIENTE DE DETERMINACIÓN (R^2)

La recta mínimo cuadrática no siempre da buenos resultados, sobre todo, si las variables no mantienen relación lineal.

El C. DE DETERMINACIÓN (R^2) nos indica si el ajuste (la obtención de la recta) es bueno:

- Si R^2 es próximo a 0 el ajuste es malo.
- Si R^2 es próximo a 1 el ajuste es bueno

R^2 se calcula mediante el cociente entre la varianza de las \hat{y} calculadas y las varianzas de las y observada ($s^2_{\hat{y}} / s^2_y$), de manera que cuanto mayor sea este cociente mejor será la regresión.

En el ejemplo anterior, observamos que R^2 toma el valor 0.287 que es muy bajo por tanto el ajuste es muy malo. No nos sirve de mucho.

PREDICCIÓN

Una de las mayores utilidades de hacer regresión es predecir que va a ocurrir con la variable Y si conocemos el comportamiento de la variable X .

REGRESIÓN LINEAL MULTIPLE

Es un modelo de regresión lineal en el que existen más de dos variables explicativas

En el caso de que el coeficiente de determinación R^2 salga bajo (digamos menor de un 30%), considerando además que su valor no se ha visto afectado por datos anormales, deduciremos que el modelo es pobre y para mejorarlo hay tres alternativas que frecuentemente se usan:

- *Transformar* la variable predictora, o la variable de respuesta Y , o ambas y usar luego un modelo lineal.
- Usar *regresión polinómica* con una variable predictora.

<p>S. de Análisis Clínicos</p> <p>H.U. Reina Sofía</p>	<p>CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	<p>Código: Fecha: 01/09/2003</p>
	<p>Versión 1</p>	<p>Página 64 de 106</p>

- Conseguir *más variables predictoras* y usar una regresión lineal múltiple.

Método:

- Se pretende encontrar una función que haga mínimas las diferencias entre las "y" observadas y las "y" calculadas.
- El modelo sería: $Y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n$
- Se debe conseguir que la diferencia $d = (y - \hat{y})^2 = (y - (a + b_1x_1 + b_2x_2 + \dots + b_nx_n))^2$ sea mínima
- El modelo de regresión lineal múltiple con p variables predictoras y basado en n observaciones podría representarse por:

$$y_i = \beta_o + \beta_1x_{1i} + \beta_2x_{2i} + \dots + \beta_px_{pi} + e_i$$

$$y_1 = \beta_o + \beta_1x_{11} + \beta_2x_{21} + \dots + \beta_px_{p1} + e_1$$

$$y_2 = \beta_o + \beta_1x_{12} + \beta_2x_{22} + \dots + \beta_px_{p2} + e_2$$

.....

$$y_n = \beta_o + \beta_1x_{1n} + \beta_2x_{2n} + \dots + \beta_px_{pn} + e_n$$

- que puede ser escrita en forma matricial como:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 65 de 106

INCONVENIENTES DE LA REGRESIÓN LINEAL

Autocorrelación

- Existencia de relación entre los residuos
- Se detecta mediante el test de Durbin-Watson:
 - o Un resultado entre 1.75 y 2 descarta la autocorrelación

Multicolinealidad

- Cuando dos o más variables están muy relacionadas entre sí. Produce sesgos.
 - o Ej: emplear como variables explicativas la edad y años de vida laboral.

Nº de datos en la regresión

- Los autores más estrictos recomienda un mínimo de 100 datos por variable
- En general: se considera aceptable a partir de 25

Causalidad y correlación

- Se deben de utilizar variables entre las que se prevea algún tipo de relación y no que ésta se produzca a través de otra variable principal.

5.4.3.3.3. TABLAS DE CONTINGENCIA

Las tablas de contingencia se utilizan para conocer la existencia de relación entre **variables de tipo cualitativo** (nominal u ordinal)

Permiten responder a las siguientes preguntas:

- ¿Existe relación entre dos variables de tipo cualitativo?
 - o Se emplea el método de χ^2
- ¿Qué grado de relación existe entre las variables?
 - o Se emplean los métodos de coeficientes de asociación: contingencia, Q de Yule, La Gamma, Tau-b y c, de Kendall y D de Sommers.
- ¿En qué sentido se produce la relación?
 - o Se puede responder con la Técnica de residuos estandarizados

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 66 de 106

En una tabla de contingencia existe independencia entre los dos factores cuando las frecuencias observadas (O_{ij}) son iguales a las estimaciones de las frecuencias esperadas (E_{ij})

Método:

- Si dos atributos o factores (Nivel 1, Nivel 2) se estudian sobre una misma población y se miden sus frecuencias absolutas se obtendrán dos series representativas de ambos, que puede representarse en un tabla 2x2 o de doble entrada:

	Nivel 2.1	Nivel 2.1	Total
Nivel 1.1	A	B	A+B
Nivel 1.2	C	D	C+D
Total	A+C	B+D	N

Estas tablas se usan para mostrar la dependencia o independencia entre dos factores (para el caso de muestras independientes)

Las hipótesis que se plantean son:

- H_0 : las variables son independientes
- H_1 : las variables no son independientes

Ejemplo:

Supongamos estudiar la relación entre la existencia o no de una patología y el sexo

	Hombre	Mujer	Total
Patología	30	20	50
No patología	10	30	40
Total	40	50	90

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 67 de 106

MÉTODO DE CHI-CUADRADO

Partimos de la base de que dos muestras independientes no guardan relación, lo cual constituye la hipótesis nula, que se demuestra por la fórmula:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde,

- "r" es el nº de filas y "k" el nº de columnas, con grados de libertad = (r-1)(k-1).
- O_{ij} : es la frecuencia observada o nº de casos observados clasificados en la fila i de la columna j
- E_{ij} : es la frecuencia esperada (teórica) o nº de casos esperados de esa fila en esa columna. Se definiría como aquella frecuencia que se daría si los sucesos fueran independientes

Partiendo de este planteamiento, la elevación de la diferencia al cuadrado convierte en positivo cualquier tipo de diferencia

Este es un test no dirigido: sólo indica diferencia.

¿Qué medimos con el estadístico χ^2 ?

Se mide la diferencia entre el valor que debiera resultar si los dos factores fueran totalmente independientes (*frecuencia esperada o E_{ij}*) y el que se ha observado en realidad (*frecuencia observada O_{ij}*).

Para calcular E_{ij} se emplea la fórmula:

$$E_{ij} = \frac{r_i * k_i}{N}$$

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 68 de 106

Donde " r_i " indica el total de individuos de la fila i , " k_j " el total de individuos de la columna " j " y " N " el total de individuos (basado en la probabilidad de sucesos independientes).

La representación gráfica del cálculo de E_{ij} sería:

	Nivel 1	Nivel 2	Total
Nivel 1	A	B	A+B
Nivel 2	C	D	C+D
Total	A+C	B+D	N

El resultado del estadístico Chi-cuadrado se compara con el resultado teórico proporcionado por las tablas para una distribución χ^2 (tabla 2x2), para un nivel de significación determinado ($\alpha = 0.05$) y 1 grados de libertad:

- Si el valor calculado por χ^2 es mayor que el valor teórico de las tablas: RECHAZAMOS H_0 .
- Ante un nivel de significación $< 0,05 \rightarrow$ RECHAZAMOS H_0

Requisitos para aplicar χ^2 (recomendaciones de Cochran)

- Para hallar correctamente el valor de χ^2 , la tabla de 2x2 debe estar integrada por valores de una muestra aleatoria, con distribución multinomial y los valores esperados no deben ser < 5 .
- Si el N casos es pequeño (< 5), se debe utilizar la prueba exacta de Fisher para obtener el valor de chi cuadrado (χ^2).
- Si N está entre 20 y 40, se puede usar χ^2 si todos los valores son > 5
- Si el N > 40 casos se puede utilizar la corrección de continuidad de Yates para obtener el estadístico χ^2 .

$$\chi^2 = \frac{N[|AD-BC| - (N/2)]^2}{(A+B)(C+D)(A+C)(B+D)} \quad \text{Con qd de 1}$$

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 69 de 106

- Los métodos estadísticos más usados para hallar el valor del X^2 son el método de Pearson y el de razón de verosimilitud, funcionan muy bien para muestras grandes.

Ejemplo:

¿es el coma al ingreso un factor de riesgo para la mortalidad?

Emplearemos la prueba de *Chi Cuadrado* de Independencia.

Esta prueba contrasta la hipótesis: ¿las categorías de las dos variables son independientes entre sí o no?.

El análisis del "*Chi cuadrado*" arroja un valor de p determinado, que si es inferior a 0.05, indica que existe una relación entre las categorías estudiadas, o sea que las variables no son independientes entre sí.

Pruebas de chi-cuadrado

	Valor	gl	Valor p
Chi-cuadrado de Pearson	133,353	1	,000
Corrección de continuidad	130,857	1	
Razón de verosimilitud	120,913	1	
Estadístico exacto de Fisher			
Asociación lineal por lineal	133,170	1	
N de casos válidos	728		

Tabla de contingencia EVOL * COMA

Recuento

		COMA		Total
		NO	SI	
EVOL	SV ^a	484	37	521
	NS ^b	118	89	207
Total		602	126	728

a. SV = Sobreviviente

b. NS = No Sobreviviente

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 70 de 106

Según el análisis el valor del estadístico "*Chi cuadrado*" es de 133, 353 correspondiendo a un valor de $p = 0.000$ (según el test de Fisher), es decir una $p < 0.001$, sumamente significativa, lo cual indica que existe una relación entre coma al ingreso y sobrevida en pacientes críticos.

TABLAS DE CONTINGENCIA CON g.l. MAYOR DE 1

Cuando el K (nº de niveles) es mayor de 2 (y g.l. > 1) puede usarse la prueba χ^2 si :

- menos del 20% de las celdillas tienen una frecuencia esperada menor de 5.
- si no hay ninguna celdilla con una frecuencia esperada menor de uno.

Test de Independencia en r (filas) x k (columnas)

Las tablas de contingencia con g.l. > 1 nos proporcionan menor información que en 2x2., entre las causa destacan:

- Las frecuencias esperadas son demasiado pequeñas
- La estructura de la tabla no permite identificar las fuentes de asociación, es decir, no permite conocer las partes de la tabla causantes de dicha asociación

En muchas ocasiones se necesita recurrir a análisis más precisos

Ejemplo:

Relación "nivel económico" y "opinión sobre SNS"

Nivel de renta	Bueno	Malo	Regular	Total
Bajo	75 (51)	40 (51)	35(48)	150
Medio	60(61.2)	50(61.2)	70(57.6)	180
Alto	20(30.6)	40(30.6)	30(28.8)	90
Muy alto	15(27.2)	40(27.2)	25(25.6)	80
Total	170	170	160	500

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 71 de 106

La 1ª frecuencia corresponde a O y la cifra entre paréntesis corresponde a la frecuencia E

$$E_{11} = 150 * 170 / 500 = 51$$

$$E_{12} = 150 * 170 / 500 = 51$$

$$E_{13} = 160 * 150 / 500 = 48$$

$$\chi^2 = \frac{(75-51)^2}{51} + \frac{(40-51)^2}{51} + \frac{(35-48)^2}{48} + \dots + \frac{(25-25.6)^2}{25.6} = 40.05$$

El valor de χ^2 es igual a 40.05, mientras que una χ^2 con 6 grados de libertad ((4-1)(3-1)) y un nivel de significación del 55 da un valor tabulado de 12.59, resulta altamente significativo, lo cual no lleva a rechazar la hipótesis nula (independencia) y concluiríamos la dependencia de las variables.

DETERMINACIÓN DE LAS FUENTES DE ASOCIACIÓN

El test usual estudiado en el análisis de las tablas 2 x 2 solo suministran información sobre la existencia o no de asociación entre las variables cualitativas, pero no informan sobre el sentido de la misma

Medios para detectar sentido de asociación

- Métodos directos: ANÁLISIS DE RESIDUOS (Haberman)
 - o Se observa el patrón de los residuos o el de los residuos ajustados a una distribución teórica conocida

$$E_{ij} = \frac{N_{ii} * E_{ii}}{\sqrt{E_{ii}}}$$

o

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 72 de 106

- Métodos indirectos: PARTICIÓN DE LA TABLA ORIGINAL EN TABLAS 2 x 2.
 - o Se considera un nivel como permanente y se divide la tabla original en subtablas dependientes
 - o Se analizan cada una de las tablas 2 x2 originadas obteniendo el X2 asociado a cada una
 - o Se halla el valor que corresponde en la tabla X2 para un nivel de significación igual a:
 - o $\beta = \alpha / 2(c-1)$

MEDIDAS DE ASOCIACIÓN EN TABLAS DE CONTINGENCIA

Estudian: ¿Cuál es la intensidad de la asociación?

Para Tablas 2 x2:

- **Q de Yule**
 - o Parte de que las diferencias entre las frecuencias observadas y las frecuencias esperadas pueden darnos una primera medida del grado de asociación existente entre los factores y del sentido de la misma.
 - o El estadístico Q de Yule elimina el inconveniente derivado de los cambios en los valores de N y las frecuencias absolutas:
 - Q = 0, independencia
 - Q > 0, asociación positiva
 - Q < 0, asociación negativa
- Características:
 - Alcanza sus valores extremos (+1 ó -1) bajo condiciones de asociación perfecta.
 - Es invariante ante cambios de escala en filas y columnas.
 - Si se intercambian filas por columnas se mantiene la magnitud de la asociación pero cambia la dirección de la misma.

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 73 de 106

Para tablas cualesquiera $R \times C$

- **Coefficiente de contingencia**
 - Medida del grado de asociación entre dos conjuntos de atributos.
 - Especialmente útil en variables ordinales.
 - Presentará el mismo valor al margen del orden de categorías en las filas y las columnas.
 - Presenta un problema:
 - Cuando haya carencia de asociación el C. de contingencia es nulo pero cuando hay completa dependencia no llega a 1

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 74 de 106

6. ESTADISTICA EPIDEMIOLOGICA

6.1 EPIDEMIOLOGÍA: ASPECTOS GENERALES

8.1.1 CONCEPTO

La epidemiología se encarga del estudio de la frecuencia y distribución de las enfermedades en las poblaciones humanas, incluyendo las relaciones entre la frecuencia de enfermedad y diferentes factores relacionados con la mismas.

Los objetivos generales de la epidemiología son conseguir:

- Describir el estado de salud de la población, organizando datos del tipo de las frecuencias relativas y las tendencias.
- Explicar la etiología de las enfermedades a través de la observación y determinación de factores causales .
- Predecir la aparición de enfermedad y su distribución en la población.

6.1.2 FINES TEÓRICOS DE LA EPIDEMIOLOGIA

- Exactitud: reduciendo los errores aleatorios y sistemáticos
- Clasificación: proporcionar la información básica necesaria para desarrollar buenos sistemas de taxonomía.
- Razonamiento: reforzar los criterios de juicio en clínica y salud comunitaria
- Normalidad: mejorar los conceptos y técnicas de normalidad de la salud para lograr una estandarización de métodos que permita el contraste de resultados.
- Representatividad: establecer la representatividad de las observaciones respecto a la población de referencia.

6.1.3. FINES PRÁCTICOS DE LA EPIDEMIOLOGÍA

- Contribuir a la elección de los mejores métodos diagnósticos, para poder definir mejor la enfermedades y su clasificación .
- Definir y cuantificar la morbilidad de la población.
 - o Identificar los grupos de riesgo.
 - o Definir los programas de salud a desarrollar
 - Tratamientos
 - Intervenciones comunitarias.

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 75 de 106

- Estudiar la causa de aparición y persistencia de un enfermedad en una población: fundamento de cualquier medida preventiva.
- Evaluar la eficacia de los programas de salud.
 - o Prevención, tratamiento, atención, cambio de conducta, rehabilitación.
- Estudiar la evolución temporal de los fenómenos de salud sometidos a "condiciones inestables" (enfermedades, características biológicas) y la realizar una evaluación dinámica dela enfermedad (estudiar su variabilidad).

6.2 MEDIDAS DE FRECUENCIA DE ENFERMEDAD

6.2.1 FRECUENCIAS ABSOLUTAS, FRECUENCIAS RELATIVAS y FRECUENCIAS COMPARADAS.

Las **frecuencias absolutas** (p.e.: nº de casos de enfermedad en un año) no nos aportan mucha información si no sabemos a que población se refiere, por ello, las frecuencias de casos se deben evaluar siempre con respecto a la población a la que se refiere (p.e.: 125 casos de enfermedad de una población de 1000 habitantes), en forma de **frecuencia relativa** de enfermedad, que se obtiene con el cociente de la frecuencia absoluta (numerador) entre el conjunto total de observaciones (denominador), de la que puede o no formar parte (proporción o razón).

Con la frecuencia relativa podemos expresar la prevalencia y es sinónimo de la probabilidad de encontrar enfermos en una muestra aleatoria de población de referencia.

En epidemiología es frecuente comparar entre sí a grupos de sujetos, para ello pueden emplearse ciertos estadísticos de posición y tendencia central de escalas nominales ya vistos anteriormente:

- Razón: cociente de dos términos excluyentes entre sí ($R = a/b$)
- Proporción: cociente cuyo numerador forma parte del denominador ($P = a/a+b$).
- Tasa: es una proporción acumulada en un intervalo de tiempo $T = \{\text{suceso/población}/\text{tiempo}\}$.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 76 de 106

6.2.2 MEDIDAS DE FRECUENCIA RELATIVA EN UN GRUPO.

6.2.2.1 CONCEPTOS PREVIOS

- **Prevalencia:** casos existentes
- **Incidencia:** aparición de casos nuevos
- **Medidas de prevalencia:** proporción de la población con la enfermedad en un momento determinado
- **Medidas de incidencia:** Números de casos nuevos que aparecen en un periodo de tiempo:
 - o Incidencia acumulada
 - o Tasa de incidencia.

6.2.2.2 PREVALENCIA

La prevalencia de una enfermedad es la proporción de individuos con la enfermedad en un momento dado:

$$p = \text{nº enfermos en un momento dado} / \text{nº población en ese momento}$$

6.2.2.3 INCIDENCIA ACUMULADA

Vendría definida por **I_a** :

$$\frac{\text{nº individuos en que aparece la enfermedad en un periodo de tiempo determinado}}{\text{nº de individuos de la población libres de enfermedad al comienzo de ese periodo y en riesgo de contraerla}}$$

$$\frac{\text{nº de individuos de la población libres de enfermedad al comienzo de ese periodo y en riesgo de contraerla}}{\text{nº de individuos de la población libres de enfermedad al comienzo de ese periodo y en riesgo de contraerla}}$$

Equivaldría a la proporción de individuos que durante el periodo de estudio pasan de estar sanos a estar enfermos.

Requiere que la población sea cerrada o fija (cohorte) y el periodo igual para todos.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 77 de 106

Un ejemplo de población cerrada o cohorte serían los pasajeros de un avión, en el que todos permanecen durante todo el viaje y en las mismas condiciones.

8.2.2.4 TASA DE INCIDENCIA O DENSIDAD DE INCIDENCIA

En muchos casos resulta muy difícil mantener una cohorte o población cerrada en las mismas condiciones, por lo que, para salvar este problema, podríamos estudiar un modelo de población abierta en el que se calculan tasas de incidencia (una cohorte abierta podría representarse por los pasajeros de un autobús, que pueden variar a lo largo del trayecto, sin saber cuantos quedan, cuantos se bajan, ni por qué).

La **densidad de incidencia** es sinónimo de **tasa de incidencia** y de velocidad de incidencia, se calcula mediante la fórmula de **I**:

$I = \frac{\text{Nº de casos de la enfermedad que aparecen en la población durante un periodo de tiempo}}{\text{personas} \times \text{tiempo}}$
--

Donde el Tiempo en riesgo, sería aquel en el que se está libre de enfermedad pero en riesgo de contraerla

Difiere respecto a la fórmula de la Incidencia acumulada en el denominador, ya que, en la densidad de incidencia, vendría dado por la suma de los periodos de tiempo libre de enfermedad pero en riesgo de contraerla correspondientes a cada individuo de la población.

Las características de la tasa de incidencia:

- Mide la fuerza instantánea de ocurrencia de enfermedad
- No requiere poblaciones cerradas
- No es una proporción
- Cada individuo aporta al denominador su propio tiempo en riesgo.

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 78 de 106

No siempre es posible conocer el tiempo en riesgo para cada individuo, en este caso, el denominador representa la media de personas al principio y al final del estudio por el tiempo medio de seguimiento.

6.2.2.4 TASA DE ATAQUE

Es la proporción de afectados en un periodo de riesgo definido y limitado (conocido).

La limitación puede derivarse de que los factores de riesgo actúan durante un periodo de tiempo corto o que el riesgo está restringido a ciertos grupos de edad.

6.2.2.5 RELACIÓN PREVALENCIA - INCIDENCIA

La prevalencia se nutre de la incidencia y depende de la duración de la enfermedad:

$P = I \times D$, donde D es la duración de la enfermedad

$I_a = 1 - e^{(-I \times t)}$, donde t es la duración del periodo.

En enfermedades de larga duración, la existencia de una incidencia nula con una prevalencia elevada indica el fin de la propagación.

6.2.3 ESTANDARIZACIÓN DE TASAS

En ocasiones existen factores diferenciales que dificultan la aplicación de métodos de comparación, por ejemplo, la edad. Para salvar este problema se recurre a los métodos de estandarización de tasas, que pueden ser directos o indirectos.

- Estandarización directa
 - Se conocen las tasas específicas de los subgrupos y el nº total de sujetos.
 - Población de referencia: suma de las poblaciones
 - Resultado: razón estandarizada de tasa de prevalencias
- Estandarización indirecta:
 - Se conocen las tasas específicas de la población de referencia pero no se conoce las tasas específicas de los subgrupos.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 79 de 106

- Tasa de referencia: la de la otra población
- Resultado: razón de estandarización de tasas, de prevalencias, etc.

Tras la estandarización de tasas, por un método u otro es posible valorar estadísticamente la diferencia, para ello calculamos primero el error estándar de la diferencia: **Sdif**

Después dividimos la diferencia real por dicho error estándar. Si el resultado es > 1.96 o < -1.96 la diferencia será significativa con $p < 0.05$.

8.3 NOCIONES SOBRE LA CALIDAD DE LOS TEST DIAGNÓSTICOS EMPLEADOS EN EPIDEMIOLOGÍA

8.3.1 CONCEPTOS PREVIOS

El punto de partida del desarrollo de los conceptos que veremos más adelante es la **variable aleatoria**:

- Se entendemos por variable aleatoria el objeto matemático que representa cierta propiedad del mundo real que, a priori, no se conoce con certeza.
- Hemos de considerar que los valores de la variable han de ser **exclusivos** (mutuamente excluyentes, es decir, dos de ellos no pueden existir simultáneamente) y **exhaustivos** (han de cubrir todos los casos posibles).
- En función de la exclusividad de los valores de la variable, los resultados de los test diagnósticos desarrollados son **dicotómicos**, en el sentido de: positivo/negativo, enfermo/sano, etc.
Las variables aleatorias con las que trabajamos son **discretas** (toman un número finito de valores)

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 80 de 106

En los test diagnósticos que ofrecen resultados dicotómicos se pueden presentar cuatro situaciones:

Verdaderos positivos

- Nº de pacientes que tienen la enfermedad y que a la vez tienen una prueba positiva

Verdaderos negativos:

- Nº de sujetos que no tienen la enfermedad y tienen una prueba negativa

Falsos Positivos:

- Nº de sujetos que no tienen la enfermedad, pero tienen una prueba positiva

Falsos negativos:

- Nº de pacientes que tienen la enfermedad, pero tienen una prueba negativa

Si contrastamos los resultados de nuestro test con los resultados de un test de referencia (supuestamente perfecto) podremos obtener los falsos positivos y negativos.

Para expresar los resultados de forma más clara, habitualmente recurrimos a la expresión de los problemas mediante tablas de este tipo:

	ENFERMOS	SANOS
TEST +	Verdadero positivo VP	Falso positivo FP
TEST -	Falso negativo FN	Verdadero negativo VN

Esta tabla puede interpretarse en función de la probabilidad de verificar un test de hipótesis y de cometer errores de tipo α o β :

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 81 de 106

	ENFERMOS	SANOS
TEST +	Sensibilidad VP Potencia ($1-\beta$)	Proporción de falsos positivos: Error I (α)
TEST -	Proporción de falsos negativos: Error II (β)	Especificidad VN Seguridad ($1-\alpha$)

8.3.2. CARACTERÍSTICAS DE LOS TEST DIAGNÓSTICOS

8.3.2.1 SENSIBILIDAD.

Representa la **probabilidad de verdadero positivo** o probabilidad de identificar a los enfermos. En la tabla anterior, la probabilidad vendría dada por: $VP / (VP + FN)$

Aplicada a test diagnósticos, la sensibilidad representa la probabilidad condicionada de que estando enfermo el test sea positivo: $p(T+ / E+)$:

8.3.2.2 ESPECIFICIDAD

Representa la **probabilidad de verdadero negativo**: probabilidad de identificar a los sanos. En la tabla anterior la probabilidad vendría dada por: $VN / (VN+FP)$.

Representa una probabilidad condicionada, de que no estando enfermo el test se a negativo $p(T- / E-)$.

8.3.2.3 PROPORCIÓN DE FALSOS POSITIVOS: PFP

Según la tabla anterior, vendría dada por la fórmula:

$$FP / (FP+VN) = 1-\text{especificidad}$$

$$1 - p(T-/E_-) = p(T/E-)$$

8.3.2.4 PROPORCIÓN DE FALSOS NEGATIVOS: PFN

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 82 de 106

$$pFN = FN / (VP+FN) = 1-\text{sensibilidad}$$

$$1-p(T+/E+) = p(T-/E)$$

8.3.2.5 PREVALENCIA (P)

Frecuencia con que la enfermedad se da en el grupo de estudio.

Representa la probabilidad de la enfermedad previa a la prueba diagnóstica (probabilidad preprueba)

$$P = (VP+FN) / N, \text{ donde } N = VP+VN+FP+FN \text{ (es decir, toda la población)}$$

8.3.2.6 EFICIENCIA DIAGNÓSTICA

Es la probabilidad de obtener un test positivo o negativo dependiendo de que el individuo sea enfermo o no. Indica la proporción de individuos correctamente clasificados:

$$VP+VN / (VP+FP+VN+FN)$$

8.3.4 INTERPRETACIÓN DE LOS TEST DIAGNÓSTICOS

8.3.4.1 VALOR PREDICTIVO POSITIVO

Representa la **probabilidad de que un test positivo sea un enfermo**

Viene dado por la fórmula: VPP: $VP / (VP+FP)$

Es una probabilidad condicionada: siendo el test positivo esté enfermo ($p(E+/T+)$).

8.3.4.2 VALOR PREDICTIVO NEGATIVO

Representa la **probabilidad de que un test negativo sea un sano**

Viene dado por la fórmula: VPN: $VN / (VN+FN)$.

Es una probabilidad condicionada: siendo el test negativo esté sano $\{p(E-/T-)\}$

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 83 de 106

8.3.5. TEOREMA DE BAYES.

El teorema de Bayes refleja la relación existente entre sensibilidad, especificidad y prevalencia.

- La **prevalencia** de un fenómeno es equivalente a la probabilidad previa, determinada, a su vez, por la frecuencia del fenómeno en la población de estudio.
- la prevalencia del fenómeno estudiado por el test y las características operativas del mismo condicionan la **eficacia de un test diagnóstico**:
 - Si la prevalencia es baja un resultado positivo tiene más probabilidades de ser falso
 - Si la prevalencia es alta, un resultado negativo tiene más probabilidades de ser falso.

El **Teorema de Bayes** introduce el concepto de **Valor Predictivo** (proporción de verdaderos positivos de entre los considerados positivos por el test diagnóstico) en función de una prevalencia dada, entendiendo ésta como probabilidad pre-test, a partir de la cual se puede calcular la probabilidad post-test o valor predictivo positivo.

En teoría, si consideramos la prevalencia de una enfermedad solo por el resultados de los test diagnósticos estaríamos cometiendo un error al considerar como enfermos a los falsos positivos y como no enfermos a los falsos negativos.

Conociendo las características operacionales de los test diagnósticos sabremos cual sería el nivel de error que se cometería al clasificar a los individuos. Teniendo en cuenta este error se puede estimar la verdadera prevalencia del fenómeno.

Si llamamos P' a la prevalencia del fenómeno según la prueba diagnóstica, ésta tendría dos componentes:

- Verdaderos positivos (VP)
- Falsos positivos (FP)

Así mismo, la verdadera prevalencia (P) también tendría dos componentes:

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 84 de 106

- Verdaderos positivos (VP)
- Falsos negativos (FN)

Con estos datos podemos construir la siguiente tabla

	ENFERMOS	SANOS
TEST +	VP Sensibilidad	FP (1 - especificidad)
TEST -	FN (1 - sensibilidad)	VN Especificidad
TOTAL	Prevalencia	1 - prevalencia

$$VP = P \times \text{sensibilidad}$$

$$FP = (1-P) \times (1-\text{especificidad})$$

$$VN = (1-P) \times \text{especificidad}$$

$$FN = P \times (1-\text{sensibilidad})$$

La proporción de enfermos clasificados como enfermos es $P' = VP+FP$, es decir, $= (P \times \text{sensibilidad}) + \{(1-P) \times (1-\text{especificidad})\}$, si despejamos P:

$$P = P' + \text{especificidad} - 1 / \text{sensibilidad} + \text{especificidad} - 1.$$

De este grupo, la fracción de verdaderos positivos es: $P \times \text{sensibilidad}$, por ello, el **Valor predictivo** o *proporción de enfermos de entre los que presentan un test positivo* sería $VP / VP+FP$, o lo que es lo mismo:

$$\text{Valor Predictivo} = \frac{P \times \text{sensibilidad}}{P \times \text{sensibilidad} + (1-P) \times (1 - \text{especificidad})}$$

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 85 de 106

El **Valor Predictivo Positivo (VPP)** es la probabilidad de que una persona con test positivo tenga la enfermedad y es igual a la *probabilidad de presentar un resultado positivo verdadero dividido por la probabilidad de presentar un resultado positivo verdadero o un falso positivo*:

$$\text{VPP} = P(\text{VP}) / p(\text{VP}) + p(\text{FP})$$

En base a estos planteamientos:

El Teorema de Bayes nos dice que la probabilidad a posteriori es proporcional a la probabilidad a priori (prevalencia) y la verosimilitud (concordancia con los hallazgos)

En definitiva, cualquier médico, al diagnosticar una enfermedad se basa en el teorema de Bayes, ya que valora la concordancia de hallazgos (síntomas y signos) con el modelo patológico más afín y más frecuente o prevalente.

8.3.5.2 RAZON DE VEROSIMILITUD

Nos da idea del grado de concordancia de hallazgos y nos orienta sobre la probabilidad de ocurrencia de los mismos en relación a determinadas situaciones (sano, enfermo, etc)

También se llama *razón o cociente de probabilidades (CP)* o "*likelihood ratio*".

Encontramos tres modalidades:

8.3.5.2.1 Razón de probabilidades positiva

Determina cuántas veces es más probable hallar un resultado positivo en un individuo enfermo que en otro sano.

Viene determinado por el cociente pVP / pFP , por lo que

$CP+ = \text{sensibilidad} / (1 - \text{especificidad})$

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 86 de 106

8.3.5.2.2 Razón de probabilidades negativa

Determina cuántas veces es más probable hallar un resultado negativo en un individuo enfermo que en otro sano

Viene dada por el cociente p_{FN} / p_{VN} , por lo que

$CP^- = (1 - \text{sensibilidad}) / \text{especificidad}$

Para algunos autores (Sackett et al., 1997), los conceptos de sensibilidad y especificidad son anticuados y es mejor trabajar con razones de verosimilitud, sin embargo, para Díez FJ (2003) esto sería cierto en una situación en la que una enfermedad X, que puede darse o no en una población (variable binaria), presenta un hallazgo Y, que puede darse o no. En cambio cuando X puede tomar múltiples valores, el método de las razones de verosimilitud es inaplicable y no queda más remedio que aplicar el teorema de Bayes.

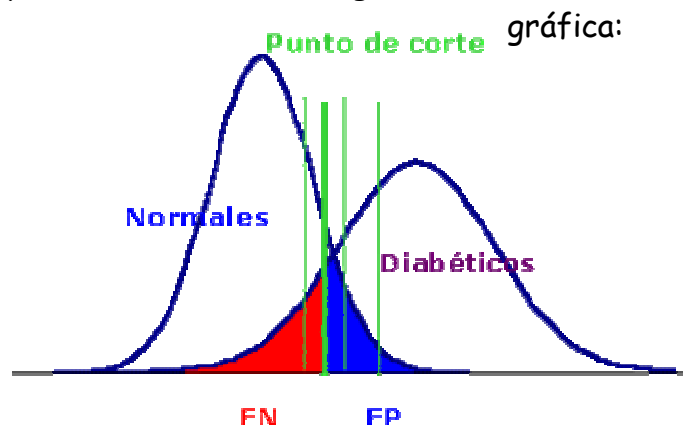
8.3.6 CURVAS DE RENDIMIENTO DIAGNÓSTICO

8.3.6.1 Concepto

Habitualmente existe un cierto grado de solapamiento en la función de probabilidad de la variable en los grupos de enfermos y no enfermos.

Cuando se utilizan variables continuas en nuestro método diagnóstico, las características operacionales de una prueba cambian según donde se sitúe el punto de corte. Además, la sensibilidad y la especificidad se mueven en sentidos opuestos.

Por ejemplo, en el caso de la glucosa la situación se esquematiza en la



S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 87 de 106

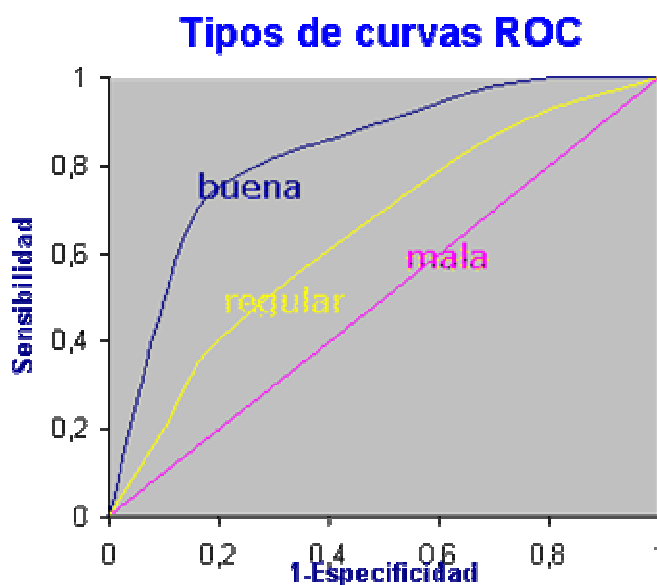
Si se desplaza el punto de corte a la derecha (valores mayores de glucosa) disminuyen los falsos positivos (región azul) pero aumentan los falsos negativos (región roja) o, en otros términos, disminuye la sensibilidad y aumenta la especificidad e inversamente si se desplaza a la izquierda, de modo que un problema en estas pruebas es la selección del punto de corte óptimo.

Para caracterizar el comportamiento (características operacionales) de los test diagnósticos se usan las llamadas curvas ROC (*Receiver Operating Characteristic*) o **curvas de rendimiento diagnóstico**, desarrolladas por los operadores de radar e introducidas en la investigación clínica por los radiólogos (Hanley y McNeil): como curvas que presenta la sensibilidad en función de los falsos positivos (complementario de la especificidad: 1-especificidad) para distintos puntos de corte.

8.3.6.2 Utilidad de las curvas ROC

La mayor utilidad de las curvas ROC radica en la **comparación de distintos métodos diagnósticos para seleccionar el más eficaz** (es más eficaz el que presenta mayor área bajo la curva).

- Las curvas ROC son un instrumento muy útil para **seleccionar el punto de corte óptimo para un determinado test diagnóstico**.
 - o El punto más cercano al ángulo superior izquierdo es el que corresponde al mejor par sensibilidad-especificidad.



S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 88 de 106

Un parámetro para evaluar la bondad de la prueba es el área bajo la curva que tomará valores entre 1 (prueba perfecta) y 0,5 (prueba inútil). Puede demostrarse que éste área puede interpretarse como la probabilidad de que ante un par de individuos, uno enfermo y el otro sano, la prueba los clasifique correctamente (Hanley y McNeil).

Ventajas de la figura de la curva ROC

- Es una representación simple, y fácilmente comprensible, de la precisión de una prueba, o sea, de su habilidad de discriminar a través de todo el rango de valores.
- No requiere seleccionar un umbral de decisión particular porque es incluido todo el rango de posible umbrales.
- Es independiente de la prevalencia, no necesita obtener muestras con prevalencia representativa, de hecho usualmente es preferible tener igual número de sujetos con ambas condiciones. Sin embargo, se ha planteado que estudios en los que se recluta a los pacientes con la enfermedad y sin ella, por separado, sobrestiman la precisión, en relación con aquellos en que los sujetos son obtenidos como una muestra representativa de la población en la cual el proceder diagnóstico fue realizado, sin selección previa, según el estado de la enfermedad.
- Proporciona una comparación visual directa entre pruebas sobre una escala común.
- Puede ser aplicado para pruebas diagnósticas cuyos resultados son medidos en escala tanto ordinal, como por intervalo o continua.

Desventajas de la figura de la curva ROC:

- No se muestra los umbrales de decisión reales.
- No se muestra el número de sujetos, y a medida que el tamaño de la muestra decrece, la representación gráfica tiende a volverse progresivamente mellada y desigual.
- La generación de la figura y el cálculo de los parámetros es difícilmente manejable sin programas de computación, los que no están ampliamente disponibles.
- No tiene aplicación cuando los resultados de la prueba no son medidos en una escala dicotómica.

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 89 de 106

- Limitaciones de su uso:
 - Sólo contemplan dos estados clínicos posibles (sano, enfermo) y no sirven para situaciones en que se trata de discernir entre más de dos enfermedades.

8.3.7 TEST COMBINADOS EN ESTUDIOS EPIDEMIOLÓGICOS

8.3.7.1 Test secuenciales (en serie)

8.3.7.1.1 Test de Screening

Los test de screening se realizan en las primeras fases de los estudios, ya que presentan mayor sensibilidad, pocos falsos negativos y no importa que haya más falsos positivos, ya que en una segunda fase se confirmarán estos positivos con test más específicos.

8.3.7.1.2 Test de confirmación en positivos

- Los test de confirmación se emplean tras los test de screening.
- Son meneos sensibles, pero más específicos
- La sensibilidad final disminuye cuando la sensibilidad de cualquier test individual es menor de 100% y la especificidad final aumenta.

8.3.7.2 Test simultáneos (en paralelo)

- Se aplican varios test a la vez
- Si alguno resulta positivo, el paciente se considera como enfermo
- Aumenta la sensibilidad y disminuye la especificidad
- Pueden resultar adecuado como inicio de una serie de test
- Pueden realizarse múltiples determinaciones con el mismo test
 - La sensibilidad y especificidad no varían
 - Los test son totalmente diferentes entre sí
 - Sensibilidad combinada = aprox. Suma de las sensibilidades
 - Especificidad combinada = aprox. Producto de especificidades

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 90 de 106

- La sensibilidad y especificidad combinada debe ser calculada de forma empírica.

CUADRO RESUMEN DE TEST COMBINADOS

	SECUENCIALES	SIMULTÁNEOS
SENSIBILIDAD	Baja	Alta
ESPECIFICIDAD	Alta	Baja

Estudios observacionales.

Se caracterizan porque no interviene el investigador: las personas presentan exposición o no a diferentes factores.

E.O. Descriptivos

- Transversales (son los más característicos)
- Casos y series de casos (sin valor epidemiológico, parten de muestras de conveniencia)
- Ecológicos (se incluyen en los "e.o. de datos secundarios", en ellos existe información de grupos, no de personas, que pueden compararse) Son válidos en la medida en que se realizan muy rápidamente, prácticamente sin coste y con información que suele estar disponible. Ejemplo: correlacionar la mortalidad por enfermedad coronaria con el consumo per cápita de cigarrillos.

E.O. Analíticos

- Cohortes
- Casos-control

Estudios experimentales

Se caracterizan por la **intervención del investigador**: éste decide a que grupo de comparación será asignado cada individuo participante.

Ensayo clínico aleatorio

Estudios pseudo-aleatorios

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 91 de 106

8.4.1 ESTUDIOS TRANSVERSALES

Concepto

Son estudios observacionales descriptivos

Estudian **prevalencias**, por eso también se llaman estudios de prevalencia

- Pueden estudiar la prevalencia global en la población y se asemejan a los estudios descriptivos analíticos si se comparan la prevalencia en sujetos expuestos y en no expuestos a determinadas variables supuestamente predoctoras: *razón de prevalencias*, sin embargo se diferencian de los estudios analíticos en que no pueden verificar la hipótesis de causalidad, puesto que no hay seguimiento temporal (no son direccionales).

La población a estudio debe partir de muestreo aleatorio simple de la población general, la muestra es representativa de la población y por lo tanto sus resultados.

Eficaces para estudiar enfermedades de largo periodo de inducción

Métodos estadísticos empleados en estudios transversales.

Cuando la razón de prevalencias es superior a 1, emplearemos el método de χ^2 para estudiar la probabilidad de cometer error aleatorio mediante el valor de p:

- si encontramos diferencias significativas ($p < 0.05$) podemos verificar que existe asociación estadística entre la exposición al riesgo y la presentación de la enfermedad.

Es importante fijar un nivel de seguridad para el cálculo del tamaño muestral (habitualmente del 95 o del 99%) y un intervalo de confianza para el que admitimos dicha seguridad, siendo p_0 la prevalencia preestimada y $q_0 = 1 - p_0$

- Cuando se realiza el muestreo sobre una población de tamaño conocido, es necesario hacer una corrección sobre el cálculo del tamaño muestral.

Secuencia de trabajo

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 92 de 106

1. Seleccionar una muestra de la población
2. Medir las variables:
 - 2.1 Resultado (enfermedad)
 - 2.2 Predictora (factor de riesgo) si vamos a calcular razón de prevalencias

Ventajas de los estudios transversales

- Estudia varias variables de resultado
- Poco tiempo de ejecución
- Buen paso inicial para estudios de cohortes
- Proporcionan estimadores de prevalencia

Inconvenientes de los estudios transversales

- No establece la secuencia de eventos
- No es útil para estudios de eventos raros
- No estima incidencia ni riesgo relativo.

Series de estudios transversales

- Consiste en la realización de series de estudios transversales sobre una misma población a lo largo del tiempo.
- Cada estudio se realiza sobre diferentes muestras
- Permite evaluar la evolución de la prevalencia y /o los factores de riesgo sin estar afectados por sesgos de información sobre las muestras (a diferencia de los de cohortes)
- Sirven para valorar el impacto de los programas de intervención comunitaria

8.4.2 ESTUDIOS DE COHORTES

Conceptos

Cohorte:

- es un grupo de individuos que se siguen conjuntamente en el tiempo y son **seleccionados antes de que presenten el efecto** (enfermedad).
- Posteriormente se hace un seguimiento en el tiempo para detectar el efecto:

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 93 de 106

- Durante el tiempo de seguimiento, los efectos detectados se cuantifican por la incidencia de los mismos.
- En poblaciones cerradas (fijas) se emplean las *incidencias acumuladas*
- En poblaciones abiertas (individuos pueden entrar o salir del estudio en cualquier momento) se emplean la *tasa de incidencias*
- El objetivo del estudio es la comparación entre grupos con diferentes niveles del factor predictor, aunque puede limitarse a describir la incidencia en un grupo.
- Los estudios de cohortes son de tipo **longitudinal** (seguimiento), **prospectivos** (la enfermedad aun no se ha producido en el momento de la asignación). Son útiles en la **comprobación de hipótesis de causalidad** formuladas previamente a partir de estudios transversales.

Métodos estadísticos: parámetros y cálculo

	Enfermos	No enfermos	
Expuestos	a	B	N1
No expuestos	c	D	N0

Riesgo Relativo (RR)

Valora la **fuerza de asociación** establecida mediante la **razón de incidencias acumuladas**:

- Incidencia en individuos expuestos / incidencia en individuos no expuestos.
- $RR = IC_e / IC_0 \{ (a/N1) / (c / N2) \}$

Los valores del RR pueden determinar si el factor de exposición es un factor de riesgo o un factor protector:

RR	INTERPRETACIÓN
<0.85	Factor protector
0.85 - 1.2	Sin asociación
> 1.2	Factor de riesgo

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 94 de 106

Riesgo Atribuible (RA)

Valora la **cantidad de casos que pueden atribuirse al factor de riesgo** mediante la **diferencia entre los riesgos absolutos** de enfermedad en expuestos y no expuestos:

- $RA = IC_e - IC_0$
- Considerando que $IC_e = RR \times IC_0$, tenemos que $RA = IC_0 (RR - 1)$

Fracción etiológica (FE)

Valora el **porcentaje de riesgo de enfermedad que se eliminaría del grupo de expuestos si se eliminase el FR (factor de riesgo)**, es decir, el % de causalidad atribuible al FR en la producción de la enfermedad en los expuestos.

Se calcula como la razón de incidencia acumulada de enfermedad debida al FR respecto al total de IC_e ($FE_e = (IC_e - IC_0) / IC_e$)

Tasa de incidencia (TI)

Valora el **nº de casos nuevos que se producen en en relación con el nº de personas por unidad de tiempo y grado de exposición**.

Ventajas del estudio de cohortes

- Permite el estudio de incidencias, riesgo relativo y riesgo atribuible
- Es una estrategia eficaz para descubrir relaciones causales, por cuanto permite la observación en el tiempo de la aparición de eventos
- Permite medir variables de exposición y controlar su influencia a lo largo del tiempo.

Inconvenientes de los estudios de cohortes.

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 95 de 106

- Baja eficiencia en el estudio de enfermedades infrecuentes (requeriría grandes cohortes)
- Generalmente resultan caros.

Sesgos más frecuentes en el estudio de cohortes

Sesgo de información

- Cuando el seguimiento es diferente para el grupo de expuestos y no expuestos.
- Para evitarlo se recurre a técnicas de enmascaramiento (en las que el investigador desconoce cual es cada grupo)

Sesgo de confusión

- Se puede dar en base a que puede haber diferencias de distribución de enfermedad en función de factores contundentes, como puede ser la edad y el sexo.
- Se evitan recurriendo a técnicas de análisis estratificado y análisis multivariante.

Medidas de asociación en los estudios de cohortes

Incidencia acumulada

Dada la tabla de contingencia

	Enfermedad	No enfermedad	Total
Expuesto	a	b	a+b
No expuesto	c	d	C+d
Total	a+c	b+d	N

- *El riesgo en expuestos viene dado: $Re = a/(a+b)$*
- *El riesgo en no expuestos viene dado: $Rne = c/(c+d)$*
- *El riesgo relativo es la razón de riesgos entre los expuestos y no expuestos:*
 - o $RR = [a/(a+b)] / [c/ (c+d)]$
- *La fracción atribuible (AF_e), representa la proporción de*

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 96 de 106

enfermedad en los expuestos que es atribuible a la exposición.

- $AF_e = \{(Re - R_{ne}) / Re\} * 100$
- La fracción atribuible no es significativa en ausencia de asociación causal.

Densidad de incidencia

	Enfermedad		
Expuesto	a	-	PTa
No expuesto	b	-	PTb

- *Densidad de incidencia en expuestos:* $Re = a/PTa$
- *Densidad de incidencia en no expuestos:* $R_{ne} = b/PTb$
- *El Riesgo Relativo (RR)* viene dado por la razón entre la densidad de incidencia en expuestos entre la densidad de incidencia en no expuestos: $RR = (a/PTa) / (b/PTb)$
- *Diferencia de riesgos (RD)* es la diferencia entre la densidad de incidencia en expuestos menos la densidad de incidencia en no expuestos
- *Fracción atribuible en expuestos (AF_e)* es la razón de la diferencia de riesgos sobre la densidad de incidencia en el grupo expuesto, expresado como porcentaje:
 - $AF_e = (Re - R_{ne}/Re)*100 = (1 - 1/RR)*100$

8.4.3 ESTUDIOS DE CASO - CONTROL

Son **estudios observacionales** en los que se agrupan a los individuos según presenten o no la enfermedad.

- Son analíticos, longitudinales y retrospectivos
- Se parte de dos grupos de sujetos, de los cuales en uno tienen la enfermedad (casos) y en otro no la tienen (controles)
- Posteriormente se analiza de forma retrospectiva cada una de los grupos buscando antecedentes de exposición al FR.

S. de Análisis Clínicos H.U. Reina Sofía	<p style="text-align: center;">CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES</p>	Código: Fecha: 01/09/2003
	Versión 1	Página 97 de 106

Se emplean para ensayar hipótesis de prevención y explorar las características de los casos y de los controles de interés para aclarar la etiología de la enfermedad.

Selección de los controles

- Los controles deben tener las mismas características que los casos, salvo en la enfermedad.
- Se deben obtener las mismas informaciones y de la misma forma que en los casos
- Es necesario eliminar el efecto de posibles factores contundentes y diferentes al estudio en ambos grupos
- Para evitar sesgos debería desconocerse la identidad de casos y controles

Secuencia de operación:

- Seleccionar una muestra de la población con la enfermedad
- Seleccionar una muestra de la población sin la enfermedad pero en riesgo de enfermar
- Medir variables predictoras (factor de riesgo, cálculo del riesgo, estimación de la asociación).

Ventajas del estudio de casos y controles:

- Son eficientes en el estudio de enfermedades o condiciones raras o con largos periodos de latencia
- Permite evaluar el efecto de múltiples exposiciones sobre una enfermedad
- Resultan relativamente rápidos y baratos: poco tiempo de ejecución y tamaños muestrales relativamente pequeños.

Inconvenientes de estudio de casos y controles

- Mayor susceptibilidad de sufrir sesgos (de selección y de información)
- No estiman incidencias ni prevalencias

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 98 de 106

- Los *casos* no son una muestra representativa de todos los enfermos, quedando fuera, generalmente, los casos fatales y los no diagnosticados.
- Es difícil establecer un buen grupo de control.
- No establecen claramente la secuencia de eventos de interés
- Sólo permiten el estudio de una enfermedad

Sesgos en los estudios de casos y control

Sesgos de selección.

Se da cuando existe una desigualdad en la selección de casos y controles (excepto en la presentación de enfermedad), en función de su clasificación como expuestos o no expuestos o en nivel de exposición.

Casos más frecuentes de sesgo de selección en casos y controles:

Paradoja de Berkson: se produce cuando se escogen controles hospitalarios que pueden diferir significativamente de la población general. Existe sesgo si la probabilidad de hospitalización es diferente para casos que para controles, o para los expuestos y los no expuestos.

Falacia de Neyman: se produce ante la utilización de casos prevalentes cuando la exposición es un factor pronóstico.

Sesgos de información

Se dan cuando la información aportada en la anamnesis no es veraz o completa.

Puede existir:

- sesgo de memoria (los casos, que sufren la enfermedad, recuerdan mejor los antecedentes).
- Sesgos por el entrevistador (diferencia en las encuestas de casos y de controles)

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 99 de 106

- Sesgos por aplicación de diferente pauta exploratoria según grupo de pertenencia.

Cálculo del riesgo en estudio de casos y controles

Los estudios de casos y controles no establecen la prevalencia o la incidencia de la enfermedad, por lo tanto no sirven para calcular la fuerza de asociación (RR), sin embargo si permiten una estimación indirecta próxima al valor del riesgo relativo cuando la prevalencia de la enfermedad es baja, mediante el cálculo de la **Odds Ratio (OR)**:

El OR se usa como **estimador del riesgo relativo** y tiene una relación matemática con las probabilidades, de forma que:

$$\text{Odds} = \text{probabilidad} / (1 - \text{probabilidad})$$

Así, para calcular la odds ratio, se divide la odds de los casos entre la odds de los controles $\{(a/c) / (b/d)\}$

Dada la siguiente tabla, que relaciona casos y controles con exposición a FR:

	Casos	Controles
FR	a	b
No FR	c	d

El cálculo de la OR sería: $OR = a \times d / b \times c$

La interpretación de los valores de la OR es igual que la de los valores del RR.

Medidas de asociación en estudios caso-control

En la mayoría de los estudios de casos-control la diferencia de riesgos no puede ser calculada, ya que las tasas de incidencia en los grupos de expuestos y no expuestos no se conocen. En

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 100 de 106

cualquier caso la fracción atribuible en los expuestos puede ser estimada usando el odds ratio como medida aproximada de riesgo relativo:

$$AF: \{(OR-1) / OR \times 100$$

Las medidas de asociación empleadas variarán en base al tipo de estudio:

No pareados

	Casos	Controles
Exposición +	a	B
Exposición -	c	D
	A+c	B+d

Proporción de exposición en los casos: $a/a+c$

Proporción de exposición en controles: $b/a+d$

Odds de exposición para los casos: $\{a/(a+c)\} / \{c/(a+c)\} = a/c$

Odds de exposición para controles: $\{b/(b+d)\} / \{d/(b+d)\} = b/d$

La OR en exposición es: $a*d / b*c$

Pareados (1:1)

	Controles expuestos	Controles no expuestos	
Casos expuestos	a	b	A+b
Casos no expuestos	c	d	C+d
	A+c	B+d	N

Para el análisis de asociación en estudios de casos-control pareados 1:1 se emplea el test de **McNemar para datos apareados**, mediante "chi cuadrado" (test de comparación entre 2 variables cualitativas pareadas):

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 101 de 106

$$X^2 = (b-c)^2 / (b+c)$$

$$X^2 \text{ corregida} = (b-c - 1)^2 / (b+c)$$

La estimación del odds ratio:

$$OR = b / c$$

$$100\% * (1-\alpha) \text{ IC: } Pl / 1-Pl \quad OR: Pu / 1-Pu, \text{ donde}$$

Pl : límite inferior del IC de la proporción b/c

Pu: Límite superior del IC de la proporción b/c
(método de Fleiss)

8.4.4 ESTUDIOS EXPERIMENTALES

Los estudios experimentales en ciencias de la salud tienen unas características especiales:

- El investigador no puede controlar completamente todas las variables.
- Por motivos éticos, no se puede investigar en humanos

Por estos motivos, es necesario plantear los estudios en otros términos:

Ensayo clínico aleatorio (ECA)

Es el tipo de estudio experimental más importante

Tiene como objetivo la evaluación de la eficacia o efectividad de las intervenciones preventivas, curativas o rehabilitadoras.

Se define como:

Toda evaluación experimental de una sustancia o medicamento, a través de su aplicación a seres humanos, orientada a alguno de los siguientes fines:

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 102 de 106

- Revelar sus efectos u obtener datos referentes a sus características farmacocinéticas.
- Establecer su eficacia para una indicación terapéutica, profiláctica o diagnóstica
- Estudiar sus reacciones adversas y establecer su seguridad

Pasos a seguir en ECA:

1º Definir los objetivos concretos del estudio

2º Elección de la muestra

- La población diana o de referencia es aquella de la que salen los individuos en los que se va a realizar la intervención
- La población de estudio es la que cumple los criterios de inclusión marcados a la población diana.
 - Generalmente los criterios de selección se introducen para dar una mayor homogeneidad la muestra.
 - Se deben evitar los sesgos de información
 - La evaluación de las intervenciones debe realizarse siempre en poblaciones similares a las que se va aplicar la intervención
- El tamaño muestral se determina en función del efecto esperado y del error de la estimación que consideremos admisible.

3º Asignación aleatoria al grupo de tratamiento o de control

- En este tipo de estudios el investigador, además de observar los fenómenos, también **manipula las condiciones naturales de la investigación** y asigna, de forma aleatoria, unos sujetos al grupo de experimentación y otros al grupo de control.
- En este punto puede cometerse un sesgo de selección, para evitarlo se aplica un proceso de "**aleatorización**", con ello se consigue que, en muestras suficientemente grandes, las características de ambos grupos tiendan a ser iguales y que las variables que pudieran influir se distribuyan de forma equilibrada, facilitándose la posibilidad de realizar inferencias causales.

4º Observación

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 103 de 106

- Evitar el **sesgo de observación**, mediante métodos para la ocultación o enmascaramiento de resultados, tanto a sujetos de estudio como a investigadores
 - o Simple ciego: una de las partes (paciente) ignora el tratamiento asignado
 - o Doble ciego: ambas partes (pacientes e investigadores) ignoran el tratamiento asignado
 - o Evaluación ciega por terceros: útil en estudios con tratamientos difícil de enmascarar o cuando el enmascaramiento no puede mantenerse por razones éticas.
 - El enmascaramiento debe afectar a todas las fases del estudio, incluyendo la fase de análisis final de los resultados.

5º Análisis e interpretación de los resultados:

- El análisis supone la comparación de los resultados obtenidos en los grupos
- El tipo de análisis dependerá de los parámetros estudiados (media, proporciones, etc)
- Debe valorarse siempre si el protocolo ha sido fielmente seguido o si se ha modificado de manera que pueda invalidar los resultados
- Siempre se deben comparar los resultados con los de otros estudios y, en caso de discrepancias, buscar posibles explicaciones.
- Se evaluará la importancia clínica de la tecnología evaluada, considerando la posibilidad de aplicación a poblaciones distintas de la del estudio.

6º Intervención

- Intervenciones preventivas:
 - o Pretenden disminuir o eliminar el riesgo de enfermar de una población.
- Intervenciones diagnósticas:
 - o Evalúan la utilidad de nuevos métodos para el diagnóstico
- Intervenciones curativas:

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 104 de 106

- Evalúan la aplicación terapéutica más efectiva y segura en una enfermedad.

7º Seguimiento

- La pauta de exploración a que son sometidos los pacientes deben ser independientes del grupo a que pertenezcan.
- El tiempo de seguimiento deberá ser lo suficientemente largo como para detectar los resultados que se quieran estudiar.
- En todo momento, las exploraciones y pruebas que se realicen serán las más indicadas para medir los parámetros que se quieran estudiar.

Estructuración de las fases de los ensayos clínicos:

- Fase I: estudio de la farmacología clínica y toxicidad
- Fase II: investigación clínica inicial del efecto terapéutico. Valoración inicial de la eficacia y seguridad.
- Fase III: Evaluación clínica completa. Ensayo propiamente dicho
- Fase IV: Farmacovigilancia tras comercializar el producto

Ventajas de los ECA

- Permiten un mayor control de los factores de estudio.
- La asignación aleatoria tiende a controlar los factores de confusión conocidos y desconocidos, aislando el efecto de la intervención.
- Permiten una mayor evidencia de la relación causa/efecto, lo que los convierte en el diseño que proporciona resultados más sólidos.

Inconvenientes de los ECA

- Las razones éticas condicionan algunos aspectos de su desarrollo y de la obtención de resultados
- Se suelen realizar sobre muestras muy seleccionadas, lo que impide la generalización de resultados.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 105 de 106

- Suelen tener elevado coste; según duración, nº de sujetos involucrados, modalidad de estudio.
- Las intervenciones suelen ser estandarizadas y, en ocasiones, alejados de la práctica clínica. Normalmente estudian una sola intervención y un solo resultado.

S. de Análisis Clínicos H.U. Reina Sofía	CONCEPTOS BASICOS DE ESTADISTICA PARA RESIDENTES	Código: Fecha: 01/09/2003
	Versión 1	Página 106 de 106

BIBLIOGRAFIA

- Martín A y Luna JD. Resúmenes de Bioestadística. Ed Norma, SA. 1997. Pág: 3-39.
- Juez P: Herramientas estadísticas para la investigación en Medicina y economía de la salud. Ed. Centro de estudios Ramón Areces, S.A. 2001. Pág: 1-69.
- Fernandez-Carreira J. Metodología Estadística. Epidemiología aplicada, en: Manual intensivo para oposiciones de médicos de Atención Primaria, coordinador Villacampa T. Editorial MAD SL, Vol. 1: 21-93. 1996
- Díez FJ. Probabilidad y la teoría de la decisión médica. UNED. Edición 2003.
- Hanley J.A., McNeil B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. **143**: 29-36. 1982