



Revisión

Conceptos estadísticos utilizados en el diseño e interpretación de trabajos de investigación

J.R. Martínez Alonso, I. Millán Santos, B. Ayuso Fernández

SECCIÓN DE BIOESTADÍSTICA. CLÍNICA PUERTA DE HIERRO. MADRID.

RESUMEN

El objetivo del presente trabajo es el de revisar las técnicas estadísticas generales de uso más común en el ámbito sanitario. Teniendo en cuenta que para su estudio existen en la actualidad innumerables textos de indudable calidad, se ha buscado una cierta originalidad unida a un sentido práctico, procurando presentar el trabajo con un alto nivel de comprensión para los poco iniciados. Para ello se ha intentado traducir la jerga matemática usual a un lenguaje más coloquial, sin temor a perder rigor científico en favor de una mayor facilidad de comprensión, se han omitido todo tipo de explicaciones matemáticas que puedan propiciar la confusión, se ha ilustrado el texto con sencillos ejemplos de clara interpretación y se ha prescindido totalmente del problema de cálculo de estadísticos que se resuelve en la actualidad con el uso de programas informáticos específicos. La presentación es la tradicional iniciada con un análisis de información y con el estudio estadístico descriptivo de variables. En este punto se considera fundamental definir las características de la función de distribución normal, para explicar el camino a seguir en el estudio de variables cuantitativas. La segunda parte plantea el uso de la inferencia estadística, fundamentalmente en las técnicas de estimación de parámetros y contraste de hipótesis, enunciando los problemas más usuales que son resueltos con los contrastes de significación estadística. En la tercera parte se plantea el problema de relación de variables en su doble vertiente de correlación y regresión. Por último se describe el análisis de supervivencia y la forma de resolución.

Palabras Clave: *Estadística.*

INTRODUCCIÓN

Los principales objetivos de la Estadística consisten en medir la aleatoriedad de los fenómenos, prever situaciones y conseguir una información cuantificada que permita la toma de decisiones. Dos características importantes son que cuanti-

ABSTRACT

Statistical concepts used in the design and interpretation of research works

This work aims to review the most common statistical techniques generally used in the health care environment. Considering that uncountable texts of unquestionable quality presently exist to study this, a certain originality together with a practical sense has been looked for, and an endeavor has been made to present a work which has a high level of understanding for the uninitiated individual. To do so, an attempt has been made to translate the unusual mathematical jargon into a more colloquial language without fear of losing scientific rigor in favor of a greater facility of understanding. All types of mathematical explanations that could lead to confusion have been omitted. The text has been illustrated with simple examples having clear interpretation and as the problem of calculating statistics is presently solved with the use of specific data programs, it has been totally left out. The presentation is the traditional one, beginning with an analysis of the information and with a descriptive statistical study of the variables. For this point, it is considered fundamental to define the characteristics of the normal distribution function in order to explain the method used in the study of the quantitative variables. The second part considers the used of statistical inferences, fundamentally in the techniques of estimation of parameters and contrast of hypothesis, stating the most usual problems that are solved with the contrast of statistical significance. In the third part, the problem of the relation of variables in their double slope of correlation and regression is considered. Finally, the analysis of survival and the resolution form is

Key Words: *Statistical.*

fica los resultados del azar con el uso del concepto de probabilidad y que la obtención de resultados se basa en el tratamiento de la información.

Es importante definir el concepto de población y muestra. Población es el conjunto de todos los elementos que cumplen ciertas condiciones y muestra es un subconjunto de esta po-

Correspondencia: J.R. Martínez Alonso. Sección de Bioestadística. Clínica Puerta de Hierro. San Martín de Porres 4. 28035 Madrid.

Fecha de recepción: 24-5-1999

Fecha de aceptación: 9-6-1999

blación. La información de la muestra es la que se utiliza para realizar el estudio estadístico. Las muestras suelen ser multidimensionales formando la estructura típica de base de datos.

La Inferencia Estadística utiliza la relación entre la muestra y la población para sacar conclusiones. Lógicamente es condición necesaria que la muestra utilizada sea representativa. Los valores estadísticos correspondientes a la población se denominan parámetros y los relativos a la muestra estadísticos. El valor de la diferencia entre un parámetro y el estadístico correspondiente se llama error muestral. La medida de la variabilidad de los estadísticos viene dada por el error típico.

- Para ilustrar gran parte de las técnicas estadísticas del texto se va a utilizar un trabajo de J.M. Arribas y cols. (F.I.S. 92/0400) de título "Estudio del valor predictivo de la hipercolesterolemia en la detección de hipotiroidismo en individuos mayores de 60 años". La población está constituida por individuos de edad avanzada y la muestra fue tomada mediante muestreo aleatorio simple seleccionando individuos mayores de 60 años utilizando el censo de un pueblo del norte de Madrid. El tamaño de la muestra es 964 (306 hombres y 658 mujeres).

ANÁLISIS DE LA INFORMACIÓN

La información para un tratamiento estadístico se estructura en forma de una base de datos que es una colección de información organizada en columnas y filas. Cada fila es una entrada o registro en la base de datos. Cada columna contiene los campos o variables que componen cada registro. Un dato aislado es cualquier hecho, objeto, fenómeno u observación que es la respuesta para un registro concreto de una variable planteada.

La valoración de cualquier dato puede hacerse a través de la definición de una variable que en una primera aproximación debe ser cualitativa o cuantitativa. Las variables cualitativas o categóricas hacen referencia a atributos, su valor es generalmente alfabético y pueden clasificarse como puras, ordenadas y procedentes de numéricas. Las variables cuantitativas se caracterizan por tener un valor numérico y pueden ser discretas (números aislados) o continuas (permite todos los valores de un intervalo).

- En el estudio de referencia, se elaboró un cuestionario que incluía preguntas sobre antecedentes y fármacos consumidos y el registro de ciertos datos clínicos y analíticos. Se presenta a continuación la descripción de algunas de las variables utilizadas en el estudio.

Variables cuantitativas: Edad (años), Peso (Kgs), Talla (cm), Índice de Masa Corporal = $\text{Peso}/(\text{Talla})^2$, Colesterol (mg/dl), TSH (Medida de TSH), T4 (Medida de T4).

Variables cualitativas: Sexo (M y F), Body Mas Index (Normal, Sobrepeso y Obeso), Diabetes (SI y NO), Hipertensión Arterial (SI y NO), Hipercolesterolemia (SI >240, NO ≤ 240), Hipotiroidismo (SI y NO).

DISTRIBUCIÓN DE FRECUENCIAS

La Distribución de Frecuencias presenta en forma resumida la información procedente de una o varias variables mediante tablas o gráficos. Se basa en la estructuración de la información en categorías o clases de frecuencias que unas veces vienen definidas de forma natural (atributos) y otras se definen especificando intervalos de valores. Requisito fundamental de las clases de frecuencias es que cualquier valor posible de una variable debe pertenecer a una de las clases previstas y solamente a una.

Los resultados obtenidos tras la operación de "contar" se ofrecen en forma de frecuencia absoluta de una clase, que indica el número de elementos de la muestra que tienen como valor el definido por dicha clase, o en forma de porcentaje, que estandariza el valor de la frecuencia sobre un total de 100. A veces resulta interesante presentar este tipo de información con el uso de frecuencias acumuladas, que indica el número de elementos menores o iguales al representante de la clase.

Las representaciones gráficas más usuales se basan en establecer una correspondencia biunívoca entre la frecuencia representada y el área de la figura geométrica elegida. El gráfico de sectores o ciclograma es una representación muy adecuada para escalas nominativas de variables cualitativas puras. El histograma o diagrama de rectángulos representa las frecuencias por las áreas de rectángulos contruidos con una base constante y alturas según los valores de las frecuencias. El gráfico de barras utiliza la misma idea que el histograma pero reduce el rectángulo correspondiente a una barra central. El polígono de frecuencias se construye uniendo los extremos del diagrama de barras.

El estudio de frecuencias para muestras multidimensionales se resuelve con tablas cruzadas de las frecuencias de presentación de las variables y tiene una representación clara en dos dimensiones (dos variables) en una tabla de frecuencia de doble entrada. El estudio con más variables puede realizarse en pasos sucesivos, fijando valores en las variables restantes. Los gráficos para representación de este tipo de información son una adaptación de los modelos para una variable (histogramas desglosados).

- Los primeros resultados estadísticos obtenidos en nuestro estudio de referencia fueron los siguientes:

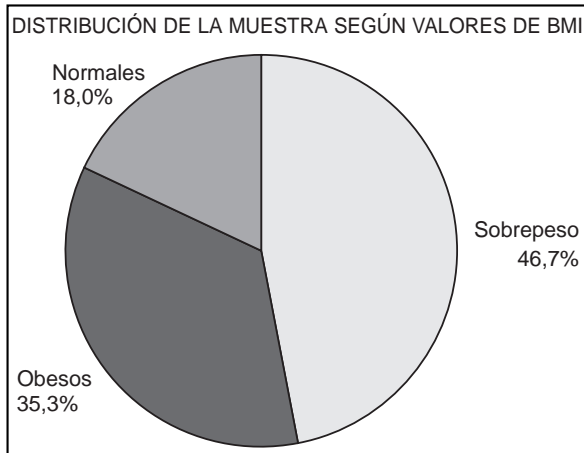


Figura 1. Representación de Frecuencias. Gráfico de Sectores.

La muestra estaba formada por 658 mujeres y 306 hombres. Las frecuencias absolutas son 658 para la clase Mujeres y 306 para Hombres. Los porcentajes son 68.3% y 31.7%. Esta información ha sido obtenida analizando la variable SEXO.

El análisis de la variable BMI (Body max index) indica unos valores de frecuencias de 174 (Normales), 450 (Sobrepeso) y 340 (Obesos), que suponen unos valores en porcentajes de 18%, 46.7% y 35.3%. Su representación en diagrama de sectores puede observarse en la figura 1.

El estudio de la variable HTA (Hipertensión arterial) indica SI para 371 personas y NO para 593. De forma similar se observa que hay 65 individuos diagnosticados de hipotiroidismo frente a los 899 restantes.

En la Tabla 1 puede observarse el formato típico de Tabla de Frecuencias ofrecido por un programa informático y cuyo significado es fácilmente comprensible. En este caso presenta la distribución de valores de COLESTEROL cuya representación gráfica se corresponde con la figura 2.

El estudio de frecuencias para algunas muestras bidimensionales se resuelve con una tabla de doble entrada y un gráfico tal como puede observarse en las figuras 3 y 4. Su interpretación inmediata (el hipotiroidismo es más frecuente en

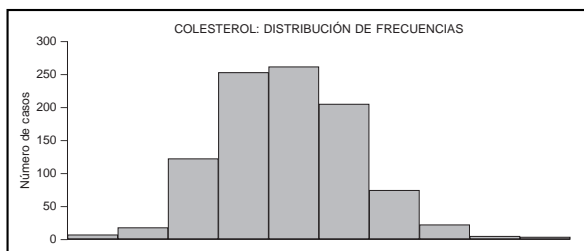


Figura 2. Representación de Frecuencias. Histograma.

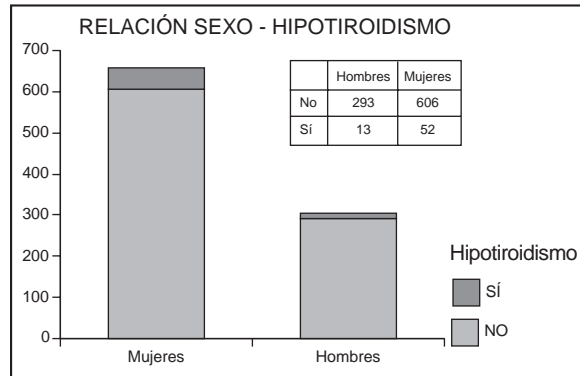


Figura 3. Frecuencias cruzadas. Barras apiladas.

mujeres y la hipertensión arterial está asociada con sobrepeso), se verá reforzada con el uso posterior de contrastes de significación estadística.

ESTUDIOS DESCRIPTIVOS DE VARIABLES CUANTITATIVAS

Los programas informáticos de Estadística ofrecen una colección de estadísticos que pueden asociarse a una variable cuantitativa. En la práctica la definición de una variable con la serie de estadísticos citados no tiene sentido y debe razonarse la explicación de su uso. Generalmente el resumen de la información se realiza a través de una medida de centralización y la medida de dispersión asociada. En este apartado se explicará brevemente el significado de los estadísticos más comunes asociados a una variable cuantitativa.

Las medidas de tendencia central tratan de resumir una variable cuantitativa por su "valor más representativo". Los diferentes criterios darán origen a las diferentes medidas de cen-

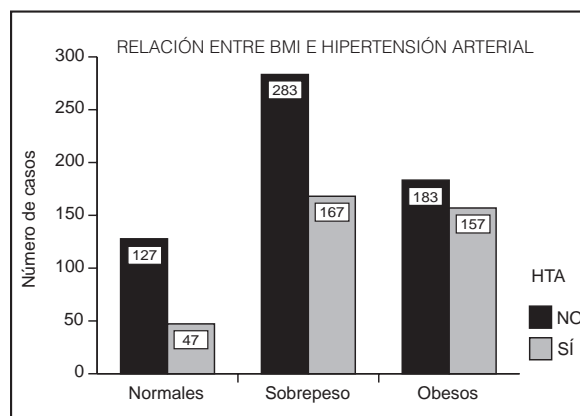


Figura 4. Relación de dos variables cualitativas.

tralización. La mediana es el valor que en la serie ordenada de los datos ocupa la posición central, de tal forma que la mitad de los datos son menores que la mediana y la otra mitad son mayores; es considerada el mejor estadígrafo de tendencia central en distribuciones asimétricas. La media es aquel valor tal que la suma de las desviaciones de todos los datos es cero; es la medida de tendencia central más utilizada y la más completa en distribuciones simétricas. Es importante destacar el riesgo de uso incorrecto de la media (distribuciones asimétricas) que puede dar lugar a interpretaciones falsas y erróneas.

Las medidas de dispersión complementan el uso de una medida de tendencia central con una medida de la desviación general de los datos respecto a la medida de centralización. El recorrido intercuartílico se asocia con la mediana y viene dado por la mitad del intervalo entre los dos cuartiles (valores que tienen por debajo el 25 y 75 % de casos). La desviación típica es el estadígrafo más completo que complementa a la media en distribuciones simétricas. Es importante señalar en este punto que la varianza es el cuadrado de la desviación típica, por lo que su significado se corresponde exactamente con el de ésta.

El objetivo de las medidas de asimetría consiste en determinar la simetría de los datos respecto a una distribución normal. El uso del coeficiente de asimetría indica que cuanto mayor sea su valor mayor es la asimetría. Valores positivos o negativos indican asimetría positiva o negativa. Valor cero es el valor ideal para indicar simetría.

Como medida de curtosis se usa el coeficiente de curtosis, que indica la forma de la distribución de los datos con respecto a una distribución normal. Un valor cero indica que la curva es mesocúrtica (curva normal), si es positivo indica que la

curva es leptocúrtica (apuntada) y si es negativo platocúrtica (achatada).

• El tratamiento informático de datos sobre la información recogida en nuestro estudio de referencia ofrece los valores de los estadísticos descritos de las variables cuantitativas de una forma similar a la presentada en la Tabla 2. Debe hacerse constar que estos resultados son específicos de la muestra obtenida. Así, por ejemplo, el valor medio de colesterol para este grupo es exactamente 231.3, y de momento no realizamos ninguna extrapolación de la información aplicando los resultados obtenidos sobre la población en general.

Se observa también (Tabla 3) que el estudio global de cada variable puede presentar una distribución que enmascara distribuciones diferentes, tal y como puede observarse en las variables Peso por Sexo. El estudio estadístico de valores de tensión arterial diastólica en las diferentes clases de BMI puede permitir sacar este mismo tipo de conclusiones.

FUNCIÓN DE DISTRIBUCIÓN NORMAL

Es una función de distribución estadística muy importante por ser considerada como una distribución de referencia para variables cuantitativas. Es una distribución simétrica en la que los estadísticos media y mediana coinciden y en la que las probabilidades de los intervalos van decreciendo a medida que se alejan del intervalo central. Su representación gráfica es la llamada curva normal o campana de Gauss. La media define la posición de la curva y coincide con el eje de simetría y el máximo de la curva. La desviación típica mide la distancia sobre el eje de abscisas entre el valor máximo de la curva y su punto de inflexión.

Al ser considerada una distribución de referencia conviene

TABLA 1. Forma de presentación de una tabla de frecuencias

	Frecuencia	Porcentaje	Porcentaje acumulado
Válidos 100-130	2	.2	.2
130-160	20	2,1	2,3
160-190	121	12,6	14,8
190-220	253	26,2	41,1
220-250	262	27,2	68,3
250-280	207	21,5	89,7
280-310	74	7,7	97,4
310-340	22	2,3	99,7
340-370	2	.2	99,9
370-400	1	.1	100,0
Total	964	100,0	

TABLA 2. Formato de presentación de estadísticos

	Peso	Talla	Edad	Colesterol
N	964	964	964	964
Media	68.080	1.5621	70.76	231.31
Desv. Típ.	12.641	9.430E-02	7.66	38.92
Error típ. de la media	.407	3.038E-03	.25	1.25
Varianza	159.794	8.898E-03	58.628	1514.693
Mediana	67.000	1.5500	69.00	230.00
Mínimo	33.0	1.31	60	117
Máximo	115.0	1.85	97	386
Curtosis	.450	-.285	-.220	-.060
Error típ. de la curtosis	.157	.157	.157	.157
Asimetría	.473	.241	.756	.207
Error típ. de la asimetría	.079	.079	.079	.079

**TABLA 3. Formatos de estadísticos de una variable desglosados en subseries de datos**

			Descriptivos					
			N	Media	Desviación típica	Error típico	Mínimo	Máximo
Peso	Sexo	Mujeres	658	64.752	11.581	.451	33.0	114.0
		Hombres	306	75.236	11.848	.677	43.0	115.0
		Total	964	68.080	12.641	.407	33.0	115.0
			Descriptivos					
			N	Media	Desviación típica	Error típico	Mínimo	Máximo
TAD	BMI	Normales	174	76.25	10.94	.83	50	110
		Sobrepeso	450	81.17	10.62	.50	50	110
		Obesos	339	84.42	10.13	.55	55	110
		Total	963	81.43	10.87	.35	50	110

señalar los medios para confirmarla. Estos se concretan en:

- Las propias características del experimento a estudiar.
- La forma de la gráfica de su distribución de frecuencias.
- La interpretación de los valores estandarizados de asimetría y curtosis.
- La realización de un contraste estadístico de normalidad.

La confirmación de la distribución normal y la determinación de sus dos parámetros (media y desviación típica) permite controlar la información de la variable correspondiente a niveles de probabilidad.

• En nuestro estudio de referencia la variable colesterol (Tabla 1 y Figura 2) parece cumplir los requisitos de distribución normal. En cuanto a la variable edad (Figura 5) ya sabíamos por lógica que no seguía una distribución normal.

ESTUDIO DE UNA VARIABLE CUANTITATIVA

El análisis de una variable cuantitativa debe iniciarse comprobando si la distribución de la variable a estudiar sigue la distribución normal. Su confirmación llevará hacia un camino que se inicia con la elección de la media y desviación típica para definir la variable. El rechazo de la hipótesis de normalidad vetará el uso de la media, cuyo sentido estará distorsionado y su uso incorrecto, aconsejando el uso de otros parámetros tales como la mediana y los centiles.

Para el estudio de una variable que sigue la distribución

normal, el significado de la desviación típica permite conocer la variabilidad de los datos sabiendo que el 95% de los casos están comprendidos entre la media más/menos 2 desviaciones típicas, información que puede ser ampliada en ambos sentidos (amplitud del intervalo y probabilidad) tanto como se desee.

Una representación gráfica de una variable cuantitativa es la denominada de "caja" o más común box-plot. Consiste en un rectángulo de anchura cualquiera y altura igual al recorrido intercuartílico, dentro del cual se traza un segmento en el punto correspondiente a la mediana; por último los segmentos que van desde las caras laterales del rectángulo hasta los valores adyacentes (valores mínimo y máximo de la distribución definidos según ciertos criterios) completan la caja. Según el criterio de Tuckey, los valores de la distribución que estén fuera del rango comprendido entre los valores adyacentes son observaciones extremas o outliers.

• En nuestro estudio de referencia utilizamos una representación box-plot para observar de forma gráfica (Figura 6) la variable talla en hombres y mujeres.

INFERENCIA ESTADÍSTICA

Constituye un procedimiento inductivo que va de lo particular a lo general y que permite obtener conclusiones de una población a través de la información proporcionada por una muestra. Las técnicas a utilizar son fundamentalmente la esti-

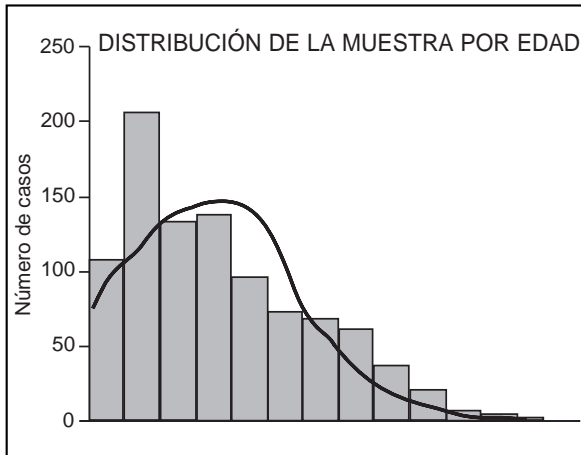


Figura 5. No es extraño que una variable no siga la distribución normal.

mación de parámetros y los contrastes de hipótesis.

Es lógico presuponer que la muestra utilizada ha de ser representativa de la población. Para conseguirlo es preciso que cada individuo de la población tenga la misma probabilidad de salir elegido como integrante de la muestra y que la selección de uno de ellos no condicione la selección de otro (independencia). Estudios de Muestreo definen procedimientos para aproximarse a esta situación ideal, pero ha de advertirse que en los estudios clínicos la selección de la muestra se realiza sencillamente tomando los datos a los que se puede tener acceso. Esta actitud es válida con la consideración de que la muestra debe ser posteriormente revisada por si fuera preciso hacer algunas modificaciones sobre la definición de las características de la población.

La determinación previa del tamaño de la muestra o número de datos que se deben tomar es uno de los primeros pro-

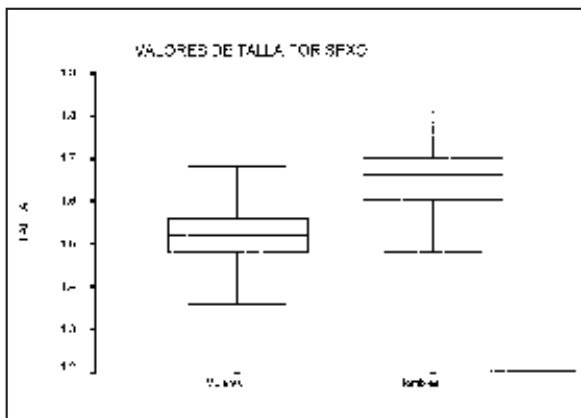


Figura 6. Representación Box - Plot.

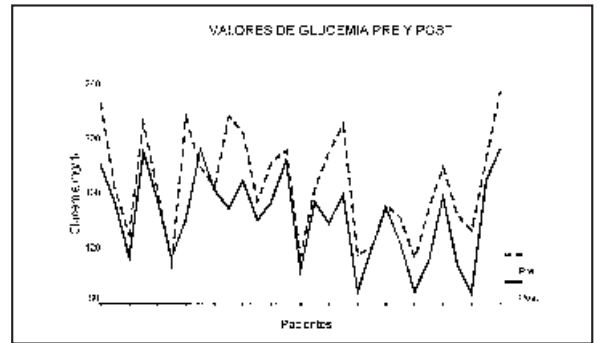


Figura 7. Muestras relacionadas.

blemas que se presenta en el diseño de un estudio estadístico. Su valor depende de la estimación a realizar, de la dispersión de los datos y del error que se pueda admitir, y para cada tipo de estimación existe una expresión o fórmula concreta.

Estimación de parámetros

Consiste en determinar el valor de los parámetros de la población a partir de los estadígrafos de la muestra. La estimación por intervalo consiste en asignar al parámetro desconocido un intervalo de valores a y b entre los cuales está dicho parámetro con una cierta confianza.

La aplicación del teorema central del límite permite estimar cualquier parámetro de la población con el uso de un intervalo de confianza, que es un intervalo simétrico en torno a un parámetro (media, mediana, porcentaje, etc.) definido por dos valores entre los cuales se debe encontrar el valor verdadero de la estimación con un nivel de confianza predeterminado. El procedimiento para construirlo toma como valor central el valor del estadígrafo de la muestra y establece límites de variabilidad por encima y por debajo, obtenidos multiplicando el error típico del parámetro por un valor t que se obtiene de tablas estadísticas para una probabilidad prefijada. El valor del error típico es directamente proporcional a la dispersión de la distribución e inversamente proporcional al número de datos de la muestra y se obtiene a partir de una fórmula específica para cada parámetro.

- En nuestro estudio de referencia se obtuvieron ciertos resultados cuya interpretación a nivel de muestra ha sido comentada anteriormente. En este momento se trata de hacer una extrapolación hablando sobre la población que representa, es decir, en individuos en general con una edad superior a los 60 años. Estas conclusiones pueden formularse con sentencias como las siguientes:

Respecto a los valores de colesterol se obtuvieron los valores de media muestral = 231.308 y desviación típica =



38.9191 con un número total de observaciones = 964. La estimación del valor de la media de colesterol en una población con estas características y para una probabilidad del 95% se resuelve en este caso con un intervalo de confianza de la media que está entre 228.848 y 233.769.

Anteriormente constatamos que había 371 individuos en la muestra con hipertensión arterial que constituían un 38,5% del total. Estos resultados de frecuencia absoluta y porcentaje son exactos referidos a la muestra considerada. Si deseamos ampliar estos valores para la población construyendo un intervalo de confianza para el porcentaje podemos concluir que un porcentaje de individuos comprendido entre 35,43% y 41,57% padecen hipertensión arterial con una confianza del 95%.

En la figura 8 puede observarse que para ilustrar los resultados de análisis de varianza de colesterol para las tres categorías de BMI, se utiliza la representación de las medias poblacionales de cada grupo.

Contrastes de hipótesis

Para decidir con objetividad si una hipótesis particular es confirmada por un conjunto de datos, necesitamos un procedimiento que nos lleve a un criterio objetivo para rechazar o aceptar esa hipótesis. Estos procedimientos son los contrastes de hipótesis, cuya aplicación práctica se desarrolla de la forma siguiente:

- Formulación de la hipótesis de nulidad (H_0). Generalmente se formula con la intención expresa de ser rechazada.
- Elección de una prueba estadística para probar H_0 . Existen generalmente varias y se debe seleccionar la óptima según las circunstancias.
- Especificación del nivel de significancia (α) y del tamaño de la muestra (N). Valores comúnmente utilizados para α son 0,05 y 0,01. Hay dos tipos de errores que pueden cometerse al decidir acerca de H_0 : error tipo I o error α es el que se comete al rechazar H_0 siendo verdadera y error tipo II o error β es el que se comete al aceptar H_0 siendo falsa.
- Encuentro (o suposición) de la distribución muestral de la prueba estadística conforme a H_0 . Cumplimiento de hipótesis previas.

– Definición de la región de rechazo. La región de rechazo consiste en un conjunto de valores posibles tan extremos que, cuando H_0 es verdadera, es muy pequeña la probabilidad (α) de que la muestra observada produzca un valor que esté entre ellos. La probabilidad asociada con cualquier valor de la región de rechazo es igual o menor que α .

– Cálculo del valor de la prueba estadística con los datos obtenidos de la(s) muestra(s). Si el valor desciende a la región de rechazo H_0 debe rechazarse y el valor observado es llama-

do "significativo"; si el valor cae fuera de la región de rechazo H_0 no puede rechazarse al nivel de significación escogido.

En la realización de un diseño de investigación es necesario emplear un criterio de selección del contraste más adecuado. Para la resolución de cualquier problema estadístico existen generalmente al menos dos contrastes adecuados. Los contrastes paramétricos proceden de un modelo que obliga a cumplir ciertas condiciones acerca de los parámetros de la población de la que se obtuvo la muestra investigada, mientras que los contrastes no paramétricos no precisan estas condiciones. Lógicamente los primeros son más precisos pero las posibilidades de su aplicación son limitadas. Como norma a seguir se indica que los datos medidos con escalas de intervalo o de proporción deben analizarse por métodos paramétricos si los supuestos del modelo estadístico paramétrico son sostenibles, y que los datos medidos por escalas nominales u ordinales deben analizarse por métodos no paramétricos.

USO DE ALGUNOS CONTRASTES DE SIGNIFICACIÓN ESTADÍSTICA

Pruebas de la bondad de ajuste de la muestra a una distribución determinada

Se trata de contrastes cuya hipótesis nula considera que la muestra obtenida sigue una función de distribución conocida con unos parámetros determinados. Su uso como contraste de normalidad es muy importante en el estudio previo de una muestra, ya que su resultado servirá para definir el tratamiento estadístico posterior de dicha muestra.

Los contrastes correspondientes se basan en la determinación del grado de acuerdo entre la distribución de un conjunto de valores de la muestra (puntajes observados) y la distribución teórica específica. Los más usuales son el contraste χ^2 y la prueba de Kolmogorv-Smirnov. En ambos casos la hipótesis nula H_0 se formula como que la muestra pertenece a la distribución descrita. La valoración del estadístico obtenido lleva a un valor p de probabilidad. En el caso que $p < 0,05$ se rechaza la hipótesis nula concluyendo que la muestra no sigue la distribución teórica.

• En nuestro estudio de referencia utilizando un programa informático para realizar un contraste de normalidad de algunas variables cuantitativas obtenemos un valor de significación p en las variables siguientes: Edad ($p < 0,0001$), Peso ($p = 0,004$), Peso en Hombres ($p = 0,693$), Peso en Mujeres ($p = 0,333$), Talla en Hombres ($p = 0,128$), Talla en Mujeres ($p = 0,128$), Tensión sistólica ($p < 0,0001$), Tensión diastólica ($p < 0,0001$), Colesterol ($p = 0,510$), TSH ($p < 0,0001$) y T4

($p=0,461$). Se observa que debe rechazarse significativamente la hipótesis de normalidad para las variables con $p<0,05$.

Contrastes de independencia de variables cualitativas. Tablas de contingencia

La distribución cruzada de las frecuencias de las diversas categorías de dos variables proporciona la información para valorar la relación de dos variables cualitativas. Es importante señalar si la relación de dependencia encontrada en la distribución muestral es o no significativa a nivel poblacional. Este problema se resuelve por los llamados contrastes de independencia, que partiendo de una hipótesis nula de independencia entre las dos variables obtienen una valoración de la diferencia entre resultados teóricos y reales. La prueba χ^2 es el contraste típico en estas situaciones en las que los datos están formados por frecuencias en categorías discretas (sean nominales u ordinales).

Es interesante destacar el caso particular de las tablas de contingencia 2×2 , en el que los contrastes correspondientes determinarán si los grupos difieren en la proporción correspondiente a las clasificaciones. La hipótesis que usualmente se pone a prueba es que los dos grupos difieren con respecto a alguna característica y, por lo tanto, con respecto a la frecuencia relativa con que los miembros del grupo son encontrados en diferentes categorías. Es habitual corregir el resultado final del contraste χ^2 utilizando la fórmula de Yates. En el caso de que las dos muestras sean pequeñas es sumamente útil el uso de la prueba de probabilidad exacta de Fisher.

- En el estudio de referencia fueron presentadas las relaciones entre dos variables cualitativas en las figuras 3 y 4. La aplicación del contraste χ^2 ofrece en el primer caso un valor de $p=0,035$ que rechaza la hipótesis de independencia entre sexo e hipotiroidismo (significativamente mayor en mujeres). En el segundo caso nos confirma el alto grado de significatividad de valores de BMI con la hipertensión arterial ($p<0,0001$).

Contrastes para comparar dos muestras relacionadas

En realidad se trata de un estudio de una muestra única (diferencia de observaciones) en el que se establece el efecto de un "tratamiento", entendiendo con este concepto una multiforme variedad de condiciones. Las muestras estudiadas se relacionan cuando cada sujeto es su propio control o con parejas de sujetos en las que se asignan los miembros de cada pareja a las dos condiciones. En estos contrastes la hipótesis nula se formula indicando que la media de las diferencias entre las dos poblaciones es cero.

La solución paramétrica más clásica es ofrecida por el contraste t de Student para datos apareados, y entre las no paramétricas se puede citar la prueba de rangos señalados y pares igualados de Wilcoxon. Como en todos los contrastes los estadísticos obtenidos se valoran con la tabla de distribución correspondiente, y su valoración para $p<0,05$ lleva a la conclusión de que el efecto del "tratamiento" ha sido significativo.

- Utilizaremos ahora los resultados de un estudio realizado en pacientes de la Unidad de Cuidados Intensivos de un Hospital para estudiar los factores de riesgo que puedan determinar la mortalidad de estos pacientes. Una de las características de estos enfermos es que tienen los valores de glucemia muy elevados y es fundamental superar el período inmediato (24 horas) para asegurar el control de la enfermedad.

Se trata de realizar la comparación de los valores de glucemia en el momento del ingreso y los presentados pasadas 24 horas. Después de realizar un contraste de normalidad se acepta la hipótesis de normalidad, por lo que seleccionamos el contraste paramétrico t de Student para realizar la comparación. El valor de $p<0,0001$ nos lleva a rechazar significativamente la hipótesis de igualdad, concluyendo que la acción del tratamiento sobre los valores de glucemia es significativa en las primeras veinticuatro horas (Figura 7).

Contrastes para comparar k muestras independientes

Es frecuente plantearse la duda sobre si varias muestras independientes deben considerarse como procedentes de la misma población. Sobre una base de datos este problema se pone de manifiesto al estudiar los valores de una variable en grupos definidos por otra variable categórica. Los valores de las muestras casi siempre difieren en cierto grado y el problema es determinar si tras las diferencias muestrales observadas hay diferencias entre poblaciones, o si son meramente variaciones al azar que se esperarían entre muestras aleatorias de la misma población.

La técnica paramétrica usual para probar si varias muestras independientes proceden de la misma población es la denominada ANOVA (Análisis de Varianza) y su paralela en el caso no paramétrico es la técnica de Kruskal-Wallis. Ambas contrastan la hipótesis nula que las k muestras independientes se recogieron de la misma población o de k poblaciones idénticas. El resultado de los contrastes indicará en el caso de $p<0,05$ que los distintos grupos son significativamente distintos.

Cuando una prueba total de k muestras permita rechazar la hipótesis de nulidad, se justifica el uso subsecuente de un procedimiento para probar las diferencias significativas entre cualquier par de las muestras (comparaciones dos a dos). En-



tre las soluciones más conocidas en el caso paramétrico se pueden citar los contrastes t de Student, LSD, Duncan, Newman-Keuls, Tukey, Scheffe, Bonferroni, etc. y en el ámbito no paramétrico se resuelve generalmente con el contraste de la U de Mann-Whitney.

- En nuestro estudio de referencia se ha realizado un ANOVA para determinar los valores de colesterol según categorías definidas por BMI (Figura 8). En la tabla 4 puede observarse los resultados estadísticos usuales ofrecidos en un estudio de estas características. La interpretación de estos resultados se inicia con la tabla de Descriptivos que nos ofrece un estudio de la variable colesterol en cada uno de los tres grupos de BMI. La tabla ANOVA tiene un nivel de significación $p=0.017$ que indica que los resultados del colesterol en los grupos definidos por valores de BMI son estadísticamente significativos. El contraste HSD de Tukey es el utilizado para comparaciones dos a dos y los resultados con un * indican la diferencia significativa de valores de colesterol entre Normales y Sobrepeso y Normales y Obesos, pero no entre Sobrepeso y Obesos.

Las técnicas de Análisis de Varianza alcanzan un alto grado de complejidad que hace que deban ser excluidos del presente estudio. Sin embargo se considera de interés presentar en la Figura 9 un sencillo ejemplo para ilustrar el uso de ANOVA en un nivel superior.

RELACIÓN ENTRE VARIABLES

El estudio de la posible relación existente entre variables cuantitativas se resuelve con las técnicas estadísticas de correlación y regresión. El problema se plantea valorando un modelo, de una variable dependiente en función de varias variables independientes. El análisis de correlación determinará el grado de relación del modelo, y el análisis de regresión ofrecerá la medida de esta relación funcional entre variables proporcionando un mecanismo de predicción y pronóstico.

Análisis de correlación

El análisis de correlación se realiza cuantificando el grado de la relación entre variables en un valor único llamado coeficiente de correlación. Existen diferentes coeficientes de correlación paramétricos (Pearson) y no paramétricos (Spearman, Kendall, C, etc) que toman un valor positivo o negativo de valor absoluto comprendido entre 0 y 1. El valor absoluto del coeficiente indicará un nivel de relación mejor según se acerque al valor 1, de tal forma que el caso en que $r=1$ indicará que la relación lineal es perfecta. Respecto al signo, valores positivos indican que las variables crecen simultáneamente, y

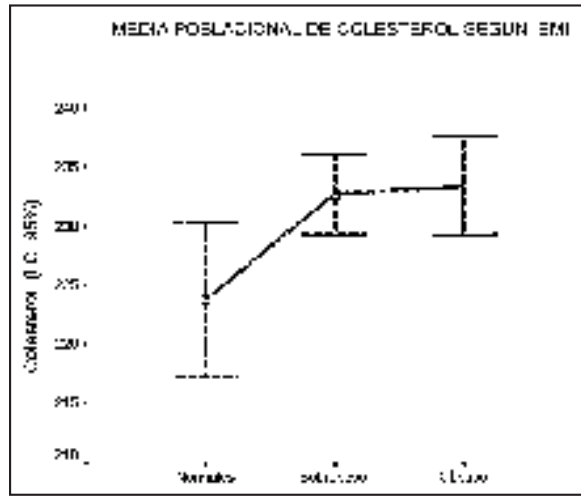


Figura 8. Anova de clasificación simple.

los valores negativos indican un sentido inverso en los grados de crecimiento.

Es interesante determinar el valor a partir del cual se puede considerar significativa la correlación entre dos variables. Nubes de dispersión diseminadas, pero con una ligera tendencia creciente o decreciente, proporcionarán coeficientes de correlación pequeños pero significativos, porque traducen esa ligera tendencia. Para el coeficiente de Pearson existe un contraste estadístico muy sencillo que resuelve esta cuestión, en el que la hipótesis nula se plantea que r sea igual a 0.

Análisis de regresión

El análisis de regresión comprende las técnicas estadísticas para determinar una fórmula que servirá para obtener los valores de una variable dependiente en función de n variables independientes. Elegido el modelo teórico el problema se concreta con la estimación de los coeficientes contenidos en dicho modelo, utilizando datos experimentales de una muestra.

El modelo lineal se resuelve con exactitud y rapidez con el uso del cálculo matricial utilizando el método de mínimos cuadrados de Gauss (regresión mínimo cuadrática). Los resultados se concretan con los valores de los coeficientes de la ecuación de regresión, pero también deben comprobarse la tabla de análisis de la varianza y el análisis de residuos para juzgar la bondad del ajuste.

El procedimiento de regresión paso a paso es un modelo general de regresión múltiple con la particularidad que selecciona las variables independientes más representativas. El proceso se realiza valorando los coeficientes de correlación calibrando la situación del modelo tras la inclusión de una nueva variable.

TABLA 4. Presentación de resultados de análisis de varianza

Descriptivos								
							Intervalo de confianza para la media al 95%	
			N	Media	Desviación típica	Error típico	Límite inferior	Límite Superior
Colesterol	BMI	Normales	174	223.71	43.69	3.31	217.17	230.24
		Sobrepeso	450	232.69	36.23	1.71	229.33	236.04
		Obesos	340	233.37	39.41	2.14	229.17	237.58
		Total	964	231.31	38.92	1.25	228.85	233.77
ANOVA								
			Suma de cuadrados	gl	Media cuadrática	F	Sig.	
Colesterol	Inter-grupos		12359.063	2	6179.532	4.106	.017	
	Intra-grupos		1446290.4	961	1504.985			
	Total		1458649.5	963				
Comparaciones múltiples								
Variable dependiente: COLESTEROL								
HSD de Tukey								
							Intervalo de confianza al 95%	
(I) BMI	(J) BMI	Diferencia de medias (I-J)		Error típico	Sig.	Límite inferior	Límite Superior	
Normales	Sobrepeso	-8.98		3.463	.026	-17.10	-.86	
	Obesos	-9.67*		3.616	.021	-18.14	-1.19	
Sobrepeso	Normales	8.98*		3.463	.026	.86	17.10	
	Obesos	-.69		2.788	.967	-7.22	5.85	
Obesos	Normales	9.67*		3.616	.021	1.19	18.14	
	Sobrepeso	.69		2.788	.967	-5.85	7.22	

* La diferencia entre las medias es significativa al nivel .05.

El caso de regresión simple es el caso más sencillo de análisis de regresión lineal. Se contemplan exclusivamente dos variables y la relación matemática que se trata de determinar es $y = a + bx$ (Línea recta). Los resultados del método se concretan en la obtención del coeficiente de correlación y la determinación de los valores a y b de la fórmula que servirá para predecir en términos de probabilidad los valores de y en función de x . La representación gráfica de los puntos experimentales mostrará una situación intermedia entre estas dos posibilidades:

- La nube de puntos estará totalmente contenida en una franja estrecha debido a que los conjuntos de ordenadas relativos a cada valor de x son poco dispersos y la línea de regresión se ajustará muy bien a ella.

- Los puntos son muy dispersos, la franja no es estrecha y la línea se separa bastante de muchos puntos de la nube. El ajuste es poco preciso y el error cometido al sustituir las 'y' observa-

das por las 'y' estimadas es grande, por lo que el modelo puede carecer de significado. El coeficiente de correlación es bajo.

La función exponencial de ecuación $y = Ae^{bx}$ es muy importante pues aparece de forma espontánea en la naturaleza y en multitud de fenómenos biológicos. Esta función se considera lineizable porque se convierte en una línea recta mediante un cambio de variable $z = \log y$, por lo que el ajuste de funciones de este tipo queda reducido a un análisis de regresión simple realizando este cambio de variable.

La regresión logística es un método similar al de regresión lineal, con la particularidad que la variable dependiente no es un valor cuantitativo sino que es una variable dicotómica (sólo puede tomar dos valores) y la variable predictora puede ser categórica o cuantitativa. Generalmente la variable dependiente toma los valores 1 ó 0 y mide la presencia o ausencia de alguna característica de interés en los sujetos de es-

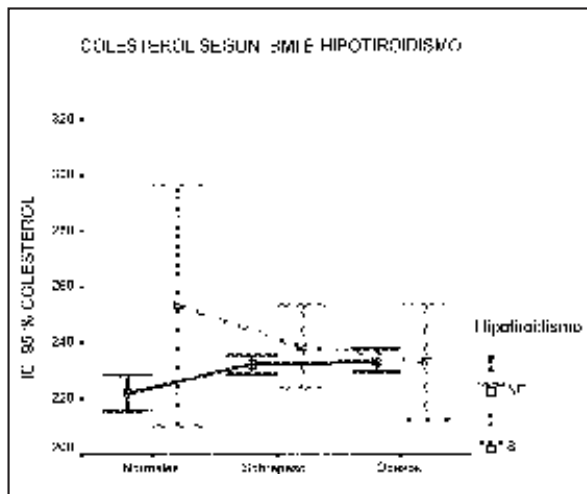


Figura 9. Análisis de Varianza de dos factores.

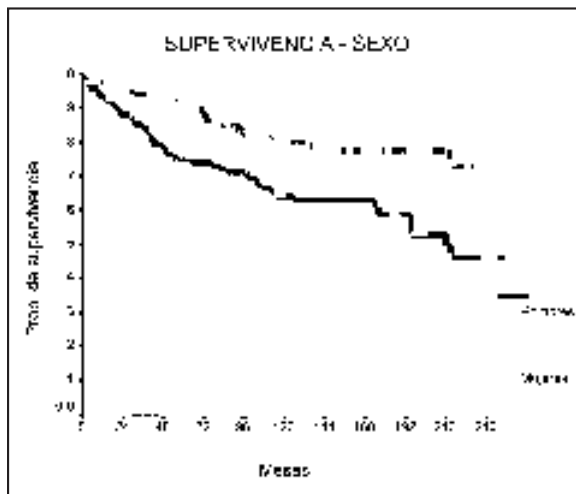


Figura 11. Curvas de Supervivencia.

tudio, por lo que los resultados del ajuste nos ofrecerá una expresión para cuantificar la probabilidad de que un individuo presente la característica de interés.

- En nuestro estudio de referencia obtenemos entre otros los siguientes valores del coeficiente de correlación de Pearson para estudiar la relación lineal de:

Peso con: Talla ($r=0,49$), IMC1 ($r=0,83$), Colesterol ($r=0,19$), TSH ($r=-0,05$) y T4 ($r=-0,03$).

Talla con: IMC1 ($r=-0,05$), Colesterol ($r=-0,01$), TSH ($r=-0,09$) y T4 ($r=-0,12$).

IMC1 con: Colesterol ($r=0,24$), TSH ($r=0,00$) y T4 ($r=0,05$).

Colesterol con: TSH ($r=0,03$) y T4 ($r=0,10$).

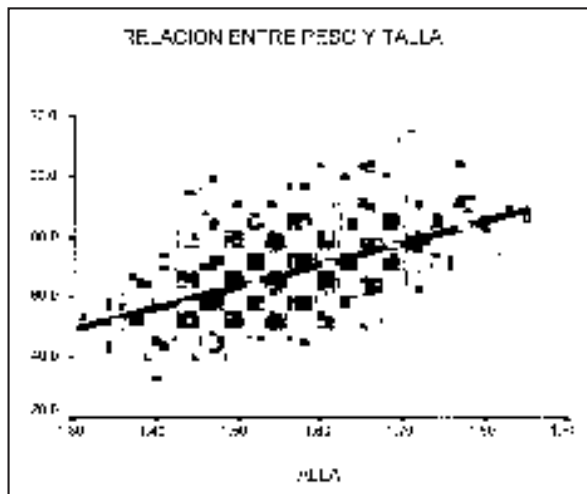


Figura 10. Regresión Lineal.

TSH con: T4 ($r=-0,21$).

- Utilizando los datos de nuestro estudio de referencia, vamos a presentar un estudio de relación en el grupo de hombres entre las dos variables Peso en función de Talla.

Sus resultados son evidentes y por lo tanto de muy fácil comprensión. Según el modelo $\text{Peso} = a + b * \text{Talla}$ el coeficiente de correlación es 0,50 que es muy significativo ($p < 0,0001$). Los valores de los coeficientes son: $a = -61,79$ y $b = 82,85$, y su representación gráfica aparece en la figura 10.

- Para ilustrar el uso de Regresión Múltiple seleccionamos un estudio para determinar un índice de actividad en la enfermedad de Crohn. Se trata de definir unos parámetros clínicos, morfológicos y/o bioquímicos que permitan conocer el grado de actividad y/o gravedad del proceso intestinal.

Los datos recogen el contenido de las variables: TEMP (Temperatura del cuerpo), HB (Valor de Hemoglobina), PLAQ (Número de Plaquetas), LEUC (Número de Leucocitos), VSG (Velocidad de Sedimentación), ALBU (Valor de Albúmina), FE (Hierro sérico), OROSO (Orosomucoide) y VD (Valoración del paciente: 1-Muy Bien, 2-Bien/Regular, 3-Mal, 4-Muy Mal).

Se trata de diseñar un modelo de regresión múltiple que permita obtener un índice de actividad (VD) a partir de una fórmula que relacione las variables más significativas del grupo descrito.

Utilizando el método de regresión paso a paso se seleccionan las variables que influyen significativamente en el modelo que son: TEMP, PLAQ, ALBU y OROSO.

El modelo de regresión ajustado es:

$$VD = -22,55 + 0,71 * \text{TEMP} + 0,0018 * \text{PLAQ} - 0,673 * \text{ALBU}$$

BU + 0,006*OROSO

ANÁLISIS DE SUPERVIVENCIA

Su denominación procede de la importancia de su uso para determinar probabilidad de muerte a lo largo del tiempo en pacientes con una enfermedad determinada. Bajo un punto de vista estadístico las técnicas que resuelven el problema quedan englobadas dentro del grupo de variables del tipo " tiempo hasta que ocurre un suceso". Bajo un punto de vista práctico su uso no debe restringirse exclusivamente al uso de la muerte, sino que puede aplicarse para estudiar cualquier otro tipo de circunstancias, como recidiva, aparición de ciertos síntomas, etc. Es importante realizar estudios de supervivencia paralelos en poblaciones similares con alguna característica diferente determinada por la clase de tratamiento, estadio de la enfermedad, edad, etc., para determinar la influencia de estos factores en la supervivencia de los pacientes. Su resolución se presenta con la determinación de probabilidades y la presentación gráfica de curvas de supervivencia, que indican el porcentaje de individuos que sobreviven a una fecha determinada o los que mueren en un período de tiempo dado. Para finalizar el estudio es importante el uso de un contraste estadístico para valorar si la diferencia entre poblaciones es o no significativa.

El método de Kaplan Meier es el procedimiento más conocido para el cálculo de la probabilidad. Realiza el trata-

miento de la información según las normas siguientes:

- La población definida suele estar integrada generalmente por individuos con una enfermedad determinada.

- El dato generado por cada individuo es el período de tiempo transcurrido desde el momento de origen (diagnóstico, inicio del tratamiento, etc) hasta el momento de fin (muerte, recidiva, curación, etc.).

- Existen individuos que no aportan información sobre el tiempo exacto de vida porque no han llegado al momento de fin (datos censurados), pero incluyen una información importante ya que se sabe que su tiempo de vida es superior al tiempo actualmente contabilizado. Este grupo está formado por los individuos que estaban vivos en la última revisión y los que se ha perdido el contacto. Es importante indicar que los individuos señalados como perdidos no deben superar en número el 5% del total.

El resultado del análisis queda ilustrado con las denominadas curvas de supervivencia (Figura 11) que representan la variable tiempo en el eje horizontal y en el eje vertical miden la probabilidad que un individuo del grupo considerado sobreviva a un instante de tiempo dado. El test de Log Rank es el contraste más utilizado para conocer si existen diferencias significativas entre las curvas de supervivencia de dos o más grupos. En este contraste la hipótesis nula supone que la supervivencia de grupos es igual y el procedimiento se resuelve comparando las frecuencias observadas con las frecuencias es-

BIBLIOGRAFÍA

- Armitage P, Berry G. *Statistical Methods in Medical Research*, Oxford: Blackwell Scientific Publications; 1987.
- Domenech i Massons JM. *Bioestadística. Métodos estadísticos para investigadores*. Barcelona: Edición Herder; 1980.
- Hazard Munro B. *Statistical Methods for Health Care Research*. 3.ª ed. Nueva York: Lippincott; 1997.
- Lee Elisa T. *Statistical Methods for Survival Data Analysis*. Belmont: Lifetime Learning Publications; 1984.
- Martín Andrés A, Luna del Castillo JD. *Bioestadística para las Ciencias de la Salud*. 4ª ed. Madrid: Ediciones Norma; 1994.
- Milton JS, Tsokos JQ. *Estadística para Biología y Ciencias de la Salud*. Nueva York: McGraw-Hill; 1987.
- Pérez de Vargas A y Abaira V. *Bioestadística*. Madrid: Ed. C. de E. Ramón Areces; 1996.
- Plasencia A, Porta Serra M. La calidad de la información clínica (II): Significación estadística. *Med Clin* 1988;90:122-6.
- Porta Serra M, Álvarez-Dardet C, Bolívar F, Plasencia A y Velilla E. La calidad de la información clínica (I): Validez. *Med Clin* 1987;89:741-7.
- Porta Serra M, Plasencia A y Sanz F. La calidad de la información clínica (III): ¿Estadísticamente significativo o clínicamente importante? *Med Clin* 1988;90:463-8.
- Sackett DL, Haynes RB, Tugwell P. *Epidemiología Clínica*. Madrid: Ediciones Díaz de Santos SA; 1989.
- Siegel S. *Estadística no paramétrica aplicada a las ciencias de la conducta*. 4ª ed. Ed. Mexico: Trillas; 1978.
- Sokal RR y Rohlf FJ. *Biometría. Principios y Métodos Estadísticos en la Investigación Biológica*. Madrid: H.Blume; 1979.
- SPSS v 7.5 para Windows 98. Copyright 1997 by SPSS Inc. Chicago Illinois 60611.