Representación gráfica en el Análisis de Datos

Pértega Díaz S., Pita Fernández S.
Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Juan Canalejo. A Coruña.
Actualización 02/04/2001.

La realización de los estudios clínico-epidemiológicos implica finalmente emitir unos resultados cuantificables de dicho estudio o experimento. La claridad de dicha presentación es de vital importancia para la comprensión de los resultados y la interpretación de los mismos. A la hora de representar los resultados de un análisis estadístico de un modo adecuado, son varias las publicaciones que podemos consultar[1]. Aunque se aconseja que la presentación de datos numéricos se haga habitualmente por medio de tablas, en ocasiones un diagrama o un gráfico pueden ayudarnos a representar de un modo más eficiente nuestros datos.

En este artículo se abordará la representación gráfica de los resultados de un estudio, constatando su utilidad en el proceso de análisis estadístico y la presentación de datos. Se describirán los distintos tipos de gráficos que podemos utilizar y su correspondencia con las distintas etapas del proceso de análisis.

Análisis descriptivo.

Cuando se dispone de datos de una población, y antes de abordar análisis estadísticos más complejos, un primer paso consiste en presentar esa información de forma que ésta se pueda visualizar de una manera más sistemática y resumida. Los datos que nos interesan dependen, en cada caso, del tipo de variables que estemos manejando[2].

Para variables categóricas[3], como el sexo, estadio TNM, profesión, etc., se quiere conocer la frecuencia y el porcentaje del total de casos que "caen" en cada categoría. Una forma muy sencilla de representar gráficamente estos resultados es mediante diagramas de barras o diagramas de sectores. En los **gráficos de sectores**, también conocidos como diagramas de "tartas", se divide un círculo en tantas porciones como clases tenga la variable, de modo que a cada clase le corresponde un arco de círculo proporcional a su frecuencia absoluta o relativa. Un ejemplo se muestra en la **Figura 1**. Como se puede observar, la información que se debe mostrar en cada sector hace referencia al número de casos dentro de cada categoría y al porcentaje del total que estos representan. Si el número de categorías es excesivamente grande, la imagen proporcionada por el gráfico de sectores no es lo suficientemente clara y por lo tanto la situación ideal es cuando hay alrededor de tres categorías. En este caso se pueden apreciar con claridad dichos subgrupos.

Los **diagramas de barras** son similares a los gráficos de sectores. Se representan tantas barras como categorías tiene la variable, de modo que la altura de cada una de ellas sea proporcional a la frecuencia o porcentaje de casos en cada clase (**Figura 2**). Estos mismos gráficos pueden utilizarse también para describir ***variables numéricas discretas*** que toman pocos valores (número de hijos, número de recidivas, etc.).

Para ***variables numéricas continuas***, tales como la edad, la tensión arterial o el índice de masa corporal, el tipo de gráfico más utilizado es el **histograma**. Para construir un gráfico de este tipo, se divide el rango de valores de la variable en intervalos de igual amplitud, representando sobre cada intervalo un rectángulo que tiene a este segmento como base. El criterio para calcular la altura de cada rectángulo es el de mantener la proporcionalidad entre las frecuencias absolutas (o relativas) de los datos en cada intervalo y el área de los rectángulos. Como ejemplo, la **Tabla I** muestra la distribución de frecuencias de la edad de 100 pacientes, comprendida entre los 18 y 42 años. Si se divide este rango en intervalos de dos años, el primer tramo está comprendido entre los 18 y 19 años, entre los que se encuentra el 4/100=4% del total. Por lo tanto, la primera barra tendrá altura proporcional a 4. Procediendo así sucesivamente, se construye el histograma que se muestra en la **Figura 3**. Uniendo los puntos medios del extremo superior de las barras del histograma, se obtiene una imagen que se llama **polígono de frecuencias**. Dicha figura pretende mostrar, de la forma más simple, en qué rangos se encuentra la mayor parte de los datos. Un ejemplo, utilizando los datos anteriores, se presenta en la **Figura 4**.

Otro modo habitual, y muy útil, de resumir una variable de tipo numérico es utilizando el concepto de percentiles, mediante **diagramas de cajas**[4,5]. La **Figura 5** muestra un gráfico de cajas correspondiente a los datos de la Tabla I. La caja central indica el rango en el que se concentra el 50% central de los datos. Sus extremos son, por lo tanto, el 1er y 3er cuartil de la distribución. La línea central en la caja es la mediana. De este modo, si la variable es simétrica, dicha línea se encontrará en el centro de la caja. Los extremos de los "bigotes" que salen de la caja son los valores que delimitan el 95% central de los datos, aunque en ocasiones coinciden con los valores extremos de la distribución. Se suelen también representar aquellas observaciones que caen fuera de este rango (outliers o valores extremos). Esto resulta especialmente útil para comprobar, gráficamente, posibles errores en nuestros datos. En general, los diagramas de cajas resultan más apropiados para representar ***variables que presenten una gran desviación de la distribución normal***. Como se verá más adelante, resultan además de gran ayuda cuando se dispone de datos en distintos grupos de sujetos.

Por último, y en lo que respecta a la descripción de los datos, suele ser necesario, para posteriores análisis, comprobar la normalidad de alguna de las variables numéricas de las que se dispone. Un diagrama de cajas o un histograma son gráficos sencillos que permiten comprobar, de un modo puramente visual, la simetría y el "apuntamiento" de la distribución de una variable y, por lo tanto, valorar su desviación de la normalidad. Existen otros métodos gráficos específicos para este propósito, como son los **gráficos P-P o Q-Q**. En los primeros, se confrontan las proporciones acumuladas de una variable con las de una distribución normal. Si la variable seleccionada coincide con la distribución de prueba, los puntos se concentran en torno a una línea recta. Los gráficos Q-Q se obtienen de modo análogo, esta vez representando los cuantiles de distribución de la variable respecto a los cuantiles de la distribución normal. En la **Figura 6** se muestra el gráfico P-P correspondientes a los datos de la Tabla I que sugiere, al igual que el correspondiente histograma y el diagrama de cajas, que la distribución de la variable se aleja de la normalidad.

Comparación de dos o más grupos.

Cuando se quieren comparar las observaciones tomadas en dos o más grupos de individuos una vez más el método estadístico a utilizar, así como los gráficos apropiados para visualizar esa relación, dependen del tipo de variables que estemos manejando.

Cuando se trabaja con ***dos variables cualitativas*** podemos seguir empleando gráficos de barras o de sectores. Podemos querer determinar, por ejemplo, si en una muestra dada, la frecuencia de sujetos que padecen una enfermedad coronaria es más frecuente en aquellos que tienen algún familiar con antecedentes cardiacos. A partir de dicha muestra podemos representar, como se hace en la **Figura 7**, dos grupos de barras: uno para los sujetos con antecedentes cardiacos familiares y otro para los que no tienen este tipo de antecedentes. En cada grupo, se dibujan dos barras representando el porcentaje de pacientes que tienen o no alguna enfermedad coronaria. No se debe olvidar que cuando los tamaños de las dos poblaciones son diferentes, es conveniente utilizar las frecuencias relativas, ya que en otro caso el gráfico podría resultar engañoso.

Por otro lado, la ***comparación de variables continuas*** en dos o más grupos se realiza habitualmente en términos de su valor medio, por medio del test t de Student, análisis de la varianza o métodos no paramétricos equivalentes, y así se ha de reflejar en el tipo de gráfico utilizado. En este caso resulta muy útil un **diagrama de barras de error**, como en la **Figura 8**. En él se compara el índice de masa corporal en una muestra de hombres y mujeres. Para cada grupo, se representa su valor medio, junto con su 95% intervalo de confianza. Conviene recordar que el hecho de que dichos intervalos no se solapen, no implica necesariamente que la diferencia entre ambos grupos pueda ser estadísticamente significativa, pero sí nos puede servir para valorar la magnitud de la misma. Así mismo, para visualizar este tipo de asociaciones, pueden utilizarse dos diagramas de cajas, uno para cada grupo. Estos diagramas son especialmente útiles aquí: no sólo permiten ver si existe o no diferencia entre los grupos, sino que además nos permiten comprobar la normalidad y la variabilidad de cada una de las distribuciones. No olvidemos que las hipótesis de normalidad y homocedasticidad son condiciones necesarias para aplicar algunos de los procedimientos de análisis paramétricos.

Por último, señalar que también en esta situación pueden utilizarse los ya conocidos gráficos de barras, representando aquí como altura de cada barra el valor medio de la variable de interés. Los **gráficos de líneas** pueden resultar también especialmente interesantes, sobre todo cuando interesa estudiar tendencias

a lo largo del tiempo (**Figura 9**). No son más que una serie de puntos conectados entre sí mediante rectas, donde cada punto puede representar distintas cosas según lo que nos interese en cada momento (el valor medio de una variable, porcentaje de casos en una categoría, el valor máximo en cada grupo, etc).

Relación entre dos variables numéricas.

Cuando lo que interesa es estudiar la relación entre ***dos variables continuas***, el método de análisis adecuado es el estudio de la correlación. Los coeficientes de correlación (Pearson, Spearman, etc.) valoran hasta qué punto el valor de una de las variables aumenta o disminuye cuando crece el valor de la otra. Cuando se dispone de todos los datos, un modo sencillo de comprobar, gráficamente, si existe una correlación alta, es mediante **diagramas de dispersión**, donde se confronta, en el eje horizontal, el valor de una variable y en el eje vertical el valor de la otra. Un ejemplo sencillo de variables altamente correlacionados es la relación entre el peso y la talla de un sujeto. Partiendo de una muestra arbitraria, podemos construir el diagrama de dispersión de la **Figura 10**. En él puede observarse claramente como existe una relación directa entre ambas variables, y valorar hasta qué punto dicha relación puede modelizarse por la ecuación de una recta. Este tipo de gráficos son, por lo tanto, especialmente útiles en la etapa de selección de variables cuando se ajusta un modelo de regresión lineal.

Otros gráficos.

Los tipos de gráficos mostrados hasta aquí son los más sencillos que podemos manejar, pero ofrecen grandes posibilidades para la representación de datos y pueden ser utilizados en múltiples situaciones, incluso para representar los resultados obtenidos por métodos de análisis más complicados. Podemos utilizar, por ejemplo, dos diagramas de líneas superpuestos para visualizar los resultados de un análisis de la varianza con dos factores (**Figura 11**). Un diagrama de dispersión es el método adecuado para valorar el resultado de un modelo de regresión logística (**Figura 12**). Existen incluso algunos análisis concretos que están basados completamente en la representación gráfica. En particular, la elaboración de curvas ROC (**Figura 13**) y el cálculo del área bajo la curva constituyen el método más apropiado para valorar la exactitud de una prueba diagnóstica.

Hemos visto, por lo tanto, como la importancia y utilidad que las representaciones gráficas pueden alcanzar en el proceso de análisis de datos. La mayoría de los textos estadísticos y epidemiológicos[4] hacen hincapié en los distintos tipos de gráficos que se pueden crear, como una herramienta imprescindible en la presentación de resultados y el proceso de análisis estadístico. No obstante, es difícil precisar cuándo es más apropiado utilizar un gráfico que una tabla. Más bien podremos considerarlos dos modos distintos pero complementarios de visualizar los mismos datos. La creciente utilización de distintos programas informáticos hace especialmente sencillo la obtención de las mismas. La mayoría de los paquetes estadísticos (SPSS, STATGRAPHICS, S-PLUS, EGRET,...) ofrecen grandes posibilidades en este sentido. Además de los gráficos vistos, es posible elaborar otros gráficos, incluso tridimensionales, permitiendo grandes cambios en su apariencia y facilidad de exportación a otros programas para presentar finalmente los resultados del estudio.

Figura 1. Ejemplo de gráfico de sectores. Distribución de una muestra de pacientes según el hábito de fumar.

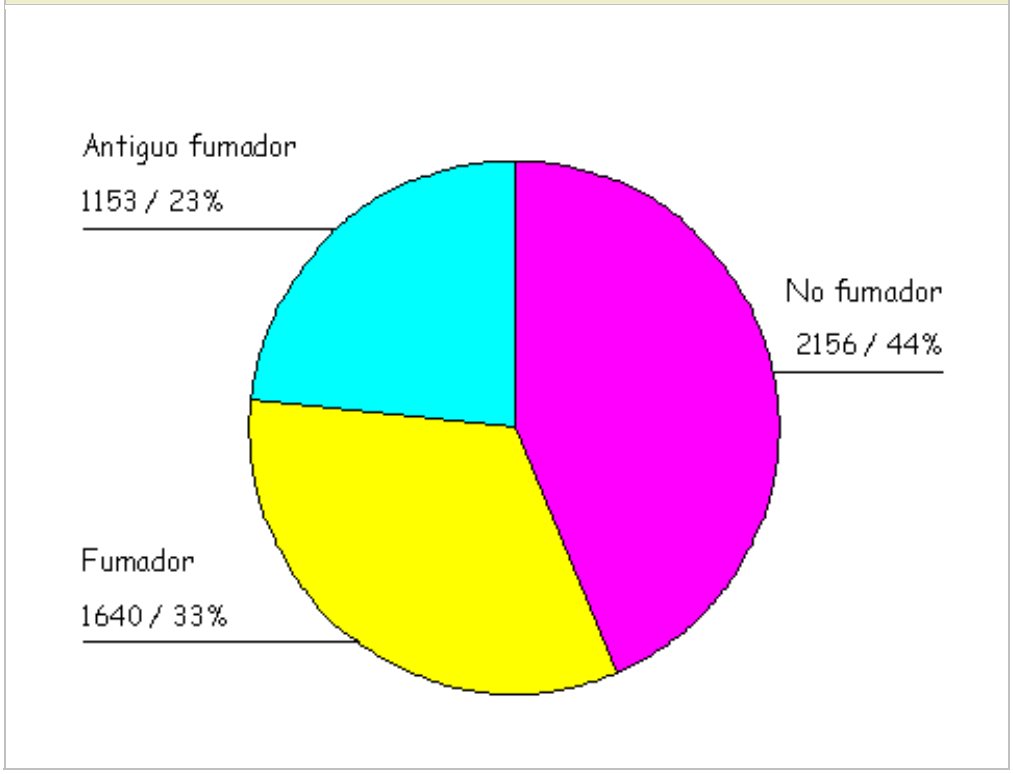


Figura 2. Ejemplo de gráfico de barras. Estadio TNM en el cáncer gástrico.

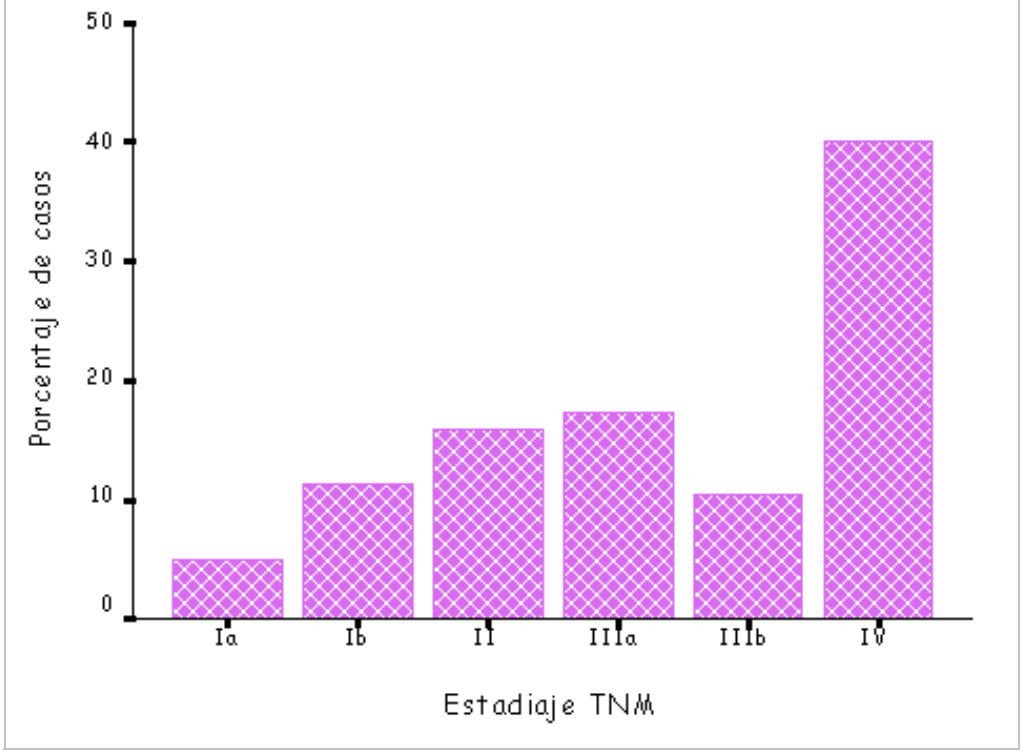


Tabla I. Distribución de frecuencias de la edad en 100 pacientes.

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

Figura 3. Ejemplo de un histograma correspondiente a los datos de la Tabla I.

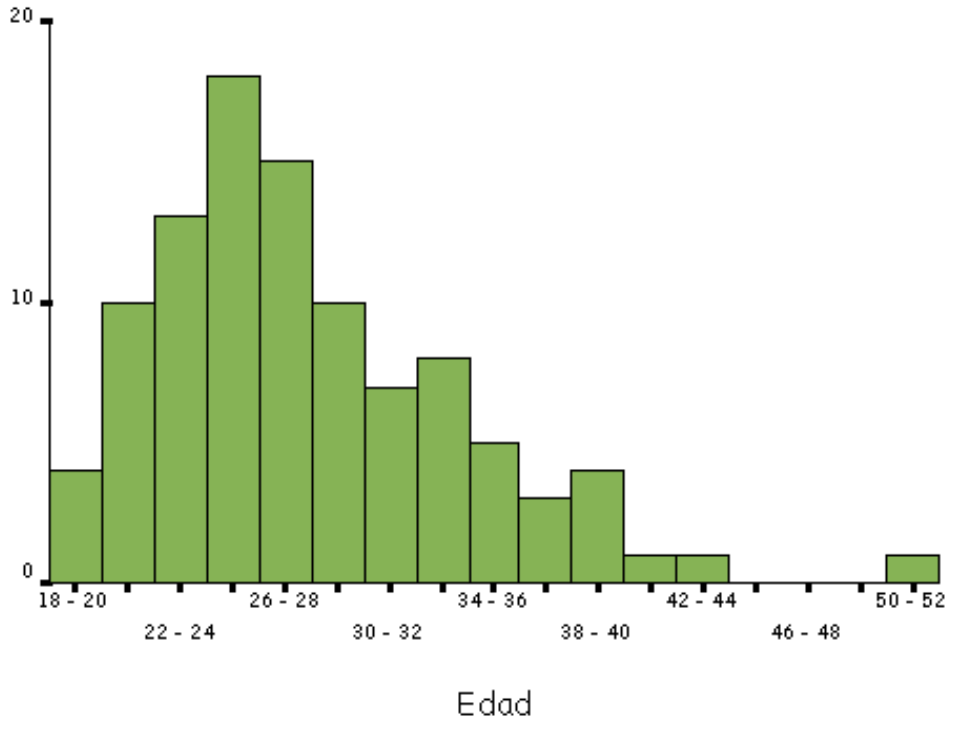


Figura 4. Polígono de frecuencias para los datos de la Tabla I.

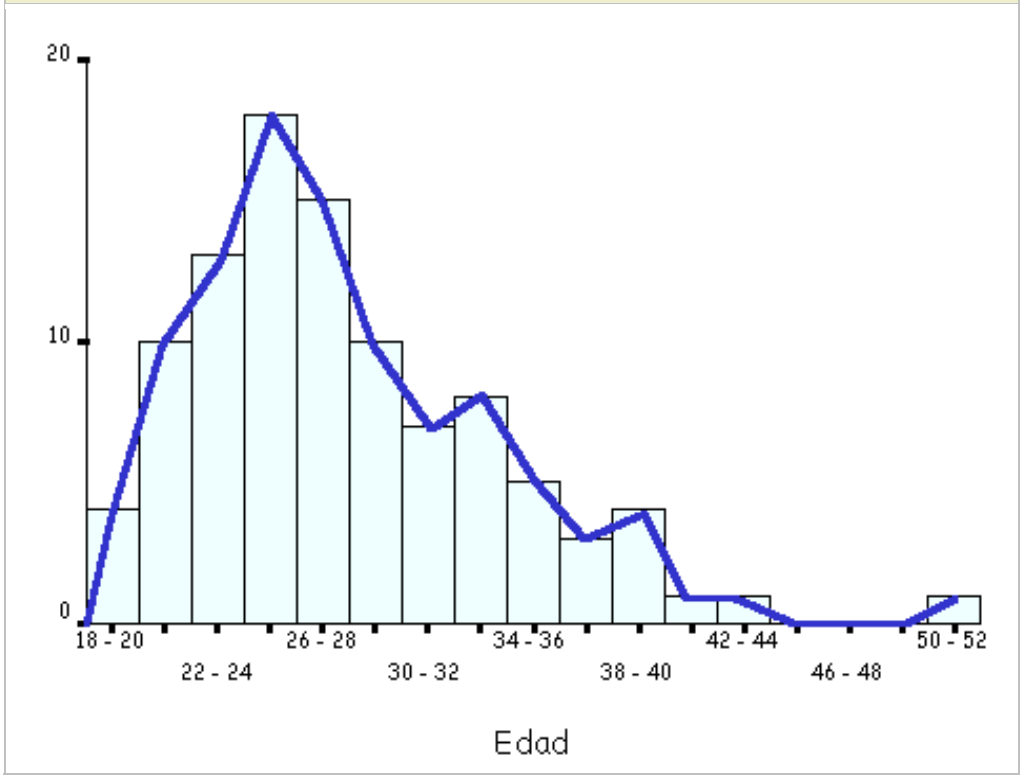


Figura 5. Ejemplo de un diagrama de caja correspondiente a lo datos en la Tabla I.

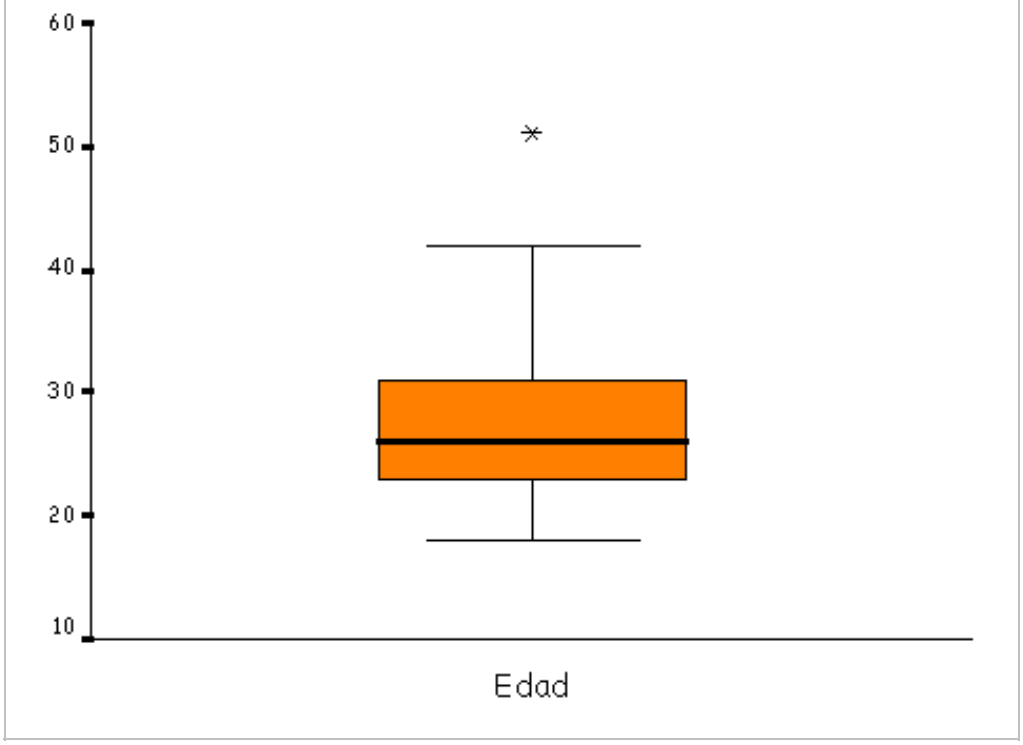


Figura 6. Gráfico P-P de normalidad para los datos de la Tabla I.

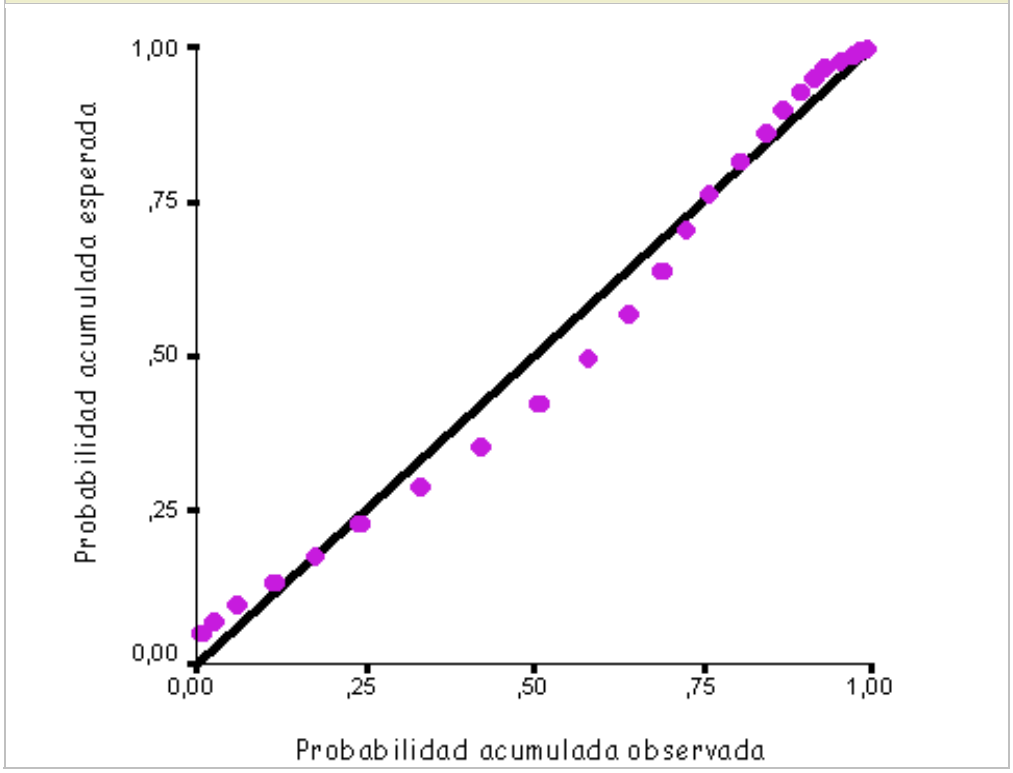


Figura 7. Diagrama de barras agrupadas. Relación entre la presencia de alguna enfermedad coronaria y los antecedentes cardiacos familiares en una muestra.

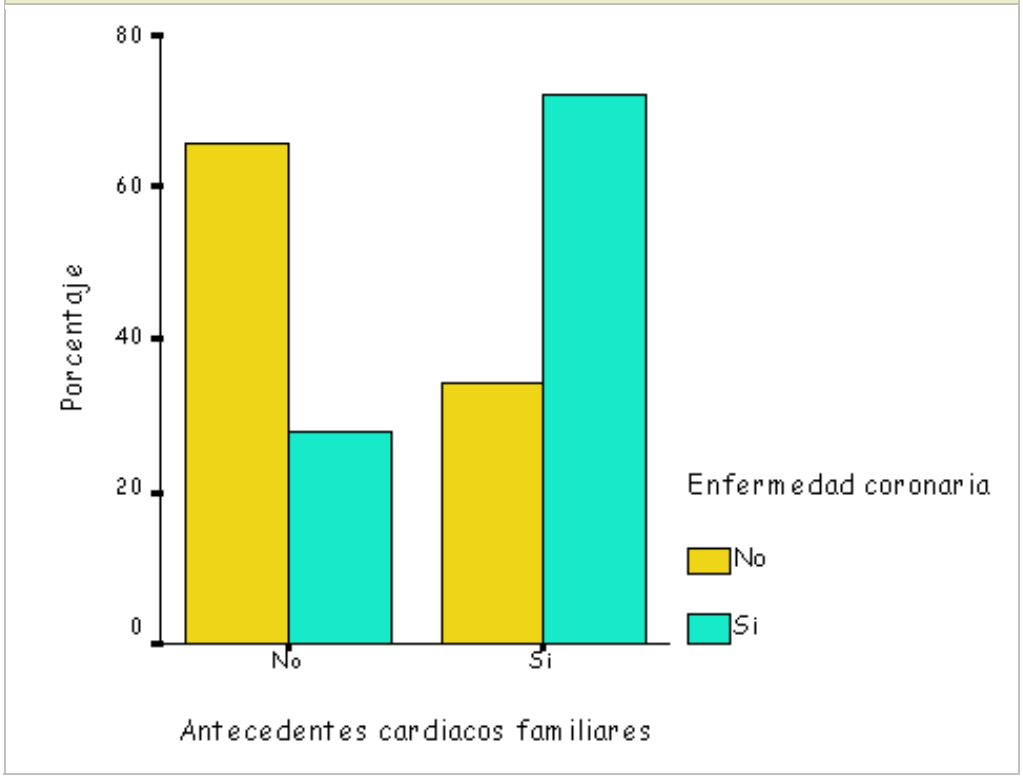


Figura 8. Barras de error. Variación en el índice de masa corporal según el sexo.

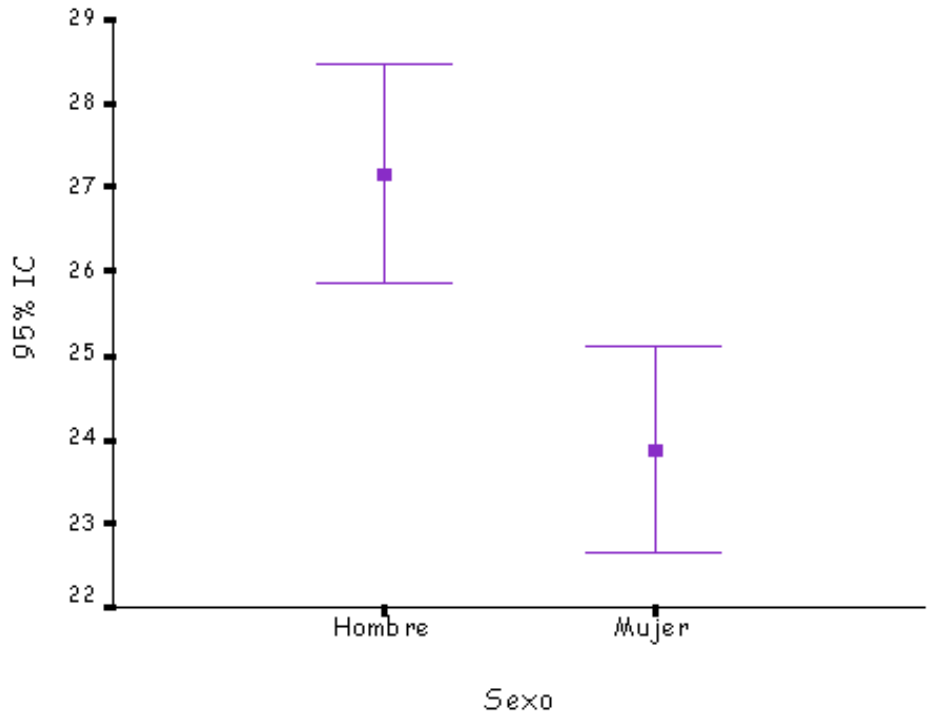


Figura 9. Gráfico de líneas. Número de pacientes trasplantados renales en el Complexo Hospitalario "Juan Canalejo" durante el periodo 1981-1997.

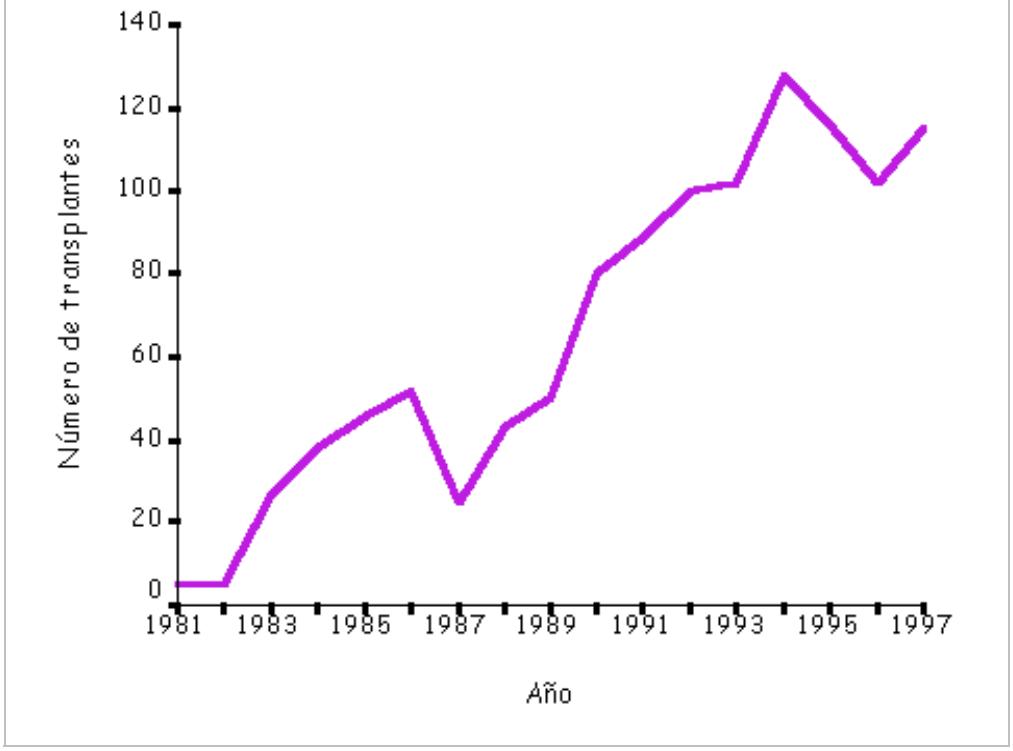


Figura 10. Diagrama de dispersión entre la talla y el peso de una muestra de individuos.

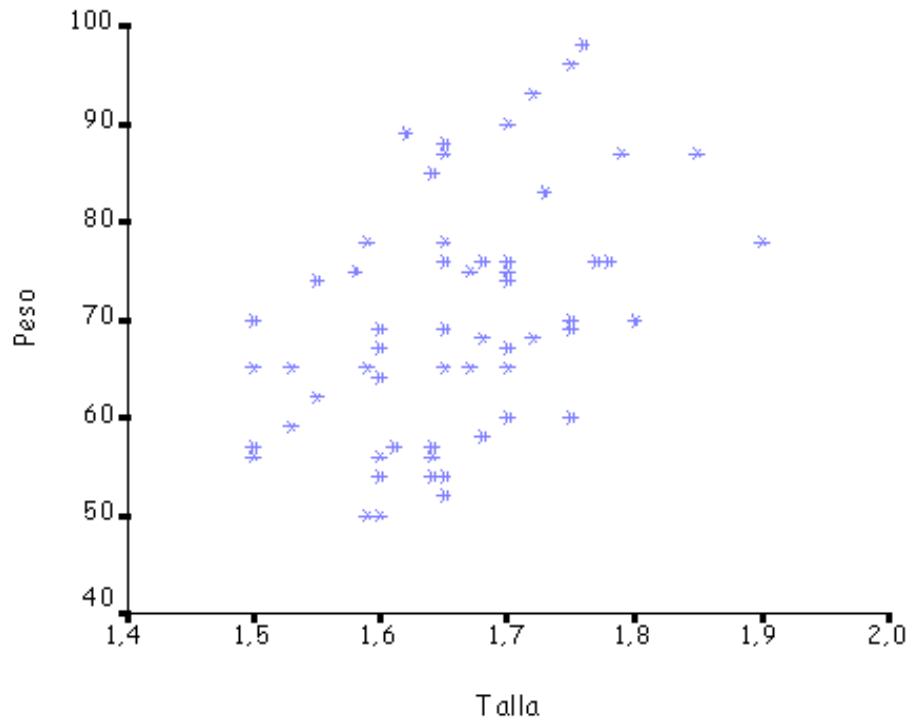


Figura 11. Dos diagramas de líneas superpuestos. Variación en el peso medio de una muestra de recién nacidos según el control ginecológico del embarazo y el hábito de fumar de la madre.

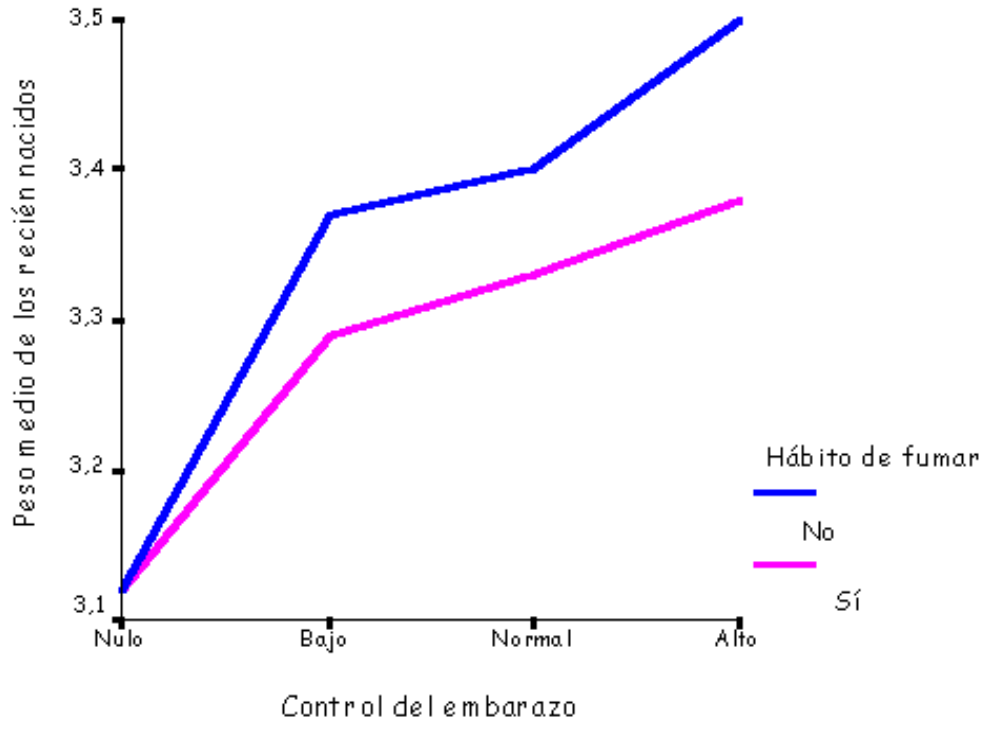


Figura 12. Diagrama de dispersión (regresión logística). Probabilidad de padecer cirrosis hepática, según un modelo de regresión logística ajustando por el % de protrombina y el presentar o no hepatomegalia.

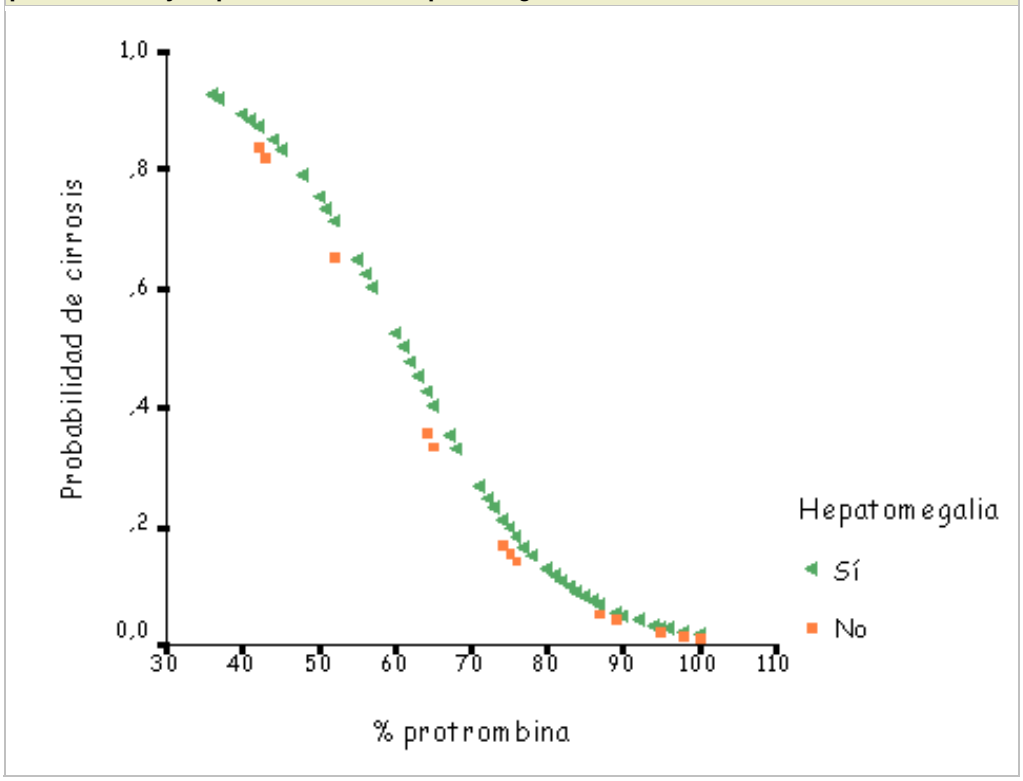
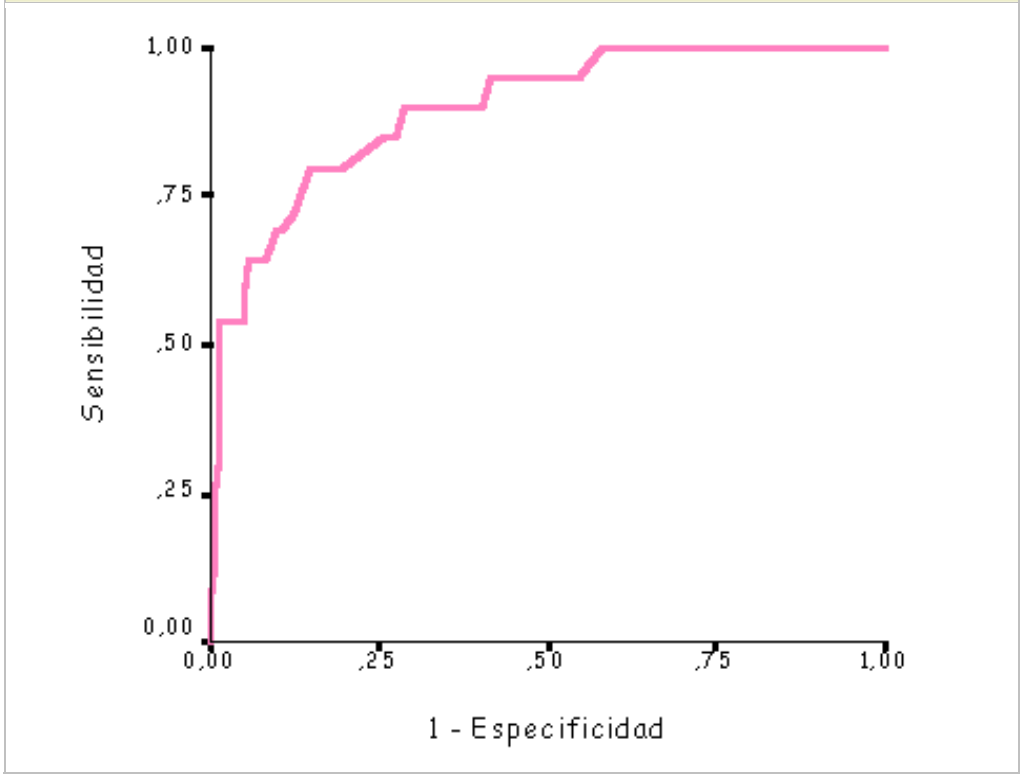


Figura 13. Curva ROC para el porcentaje de protrombina en la predicción de cirrosis.



Bibliografía

1. Lang TA, Secic M. How to report statistics in medicine. Annotated Guidelines for authors, Editors, and reviewers. Philadelphia: Port City Press; 1997.

2. Altman DG, Bland JM. Statistics Notes: Presentation of numerical data. BMJ 1996; 312: 572. [Medline] [texto completo]

3. Singer PA, Feinstein AR. Graphical display of categorical data. J Clin Epidemiol 1993; 46(3): 231-6. [Medline]

4. Simpson RJ, Johnson TA, Amara IA. The box-plot: an exploratory analysis for biomedical publications. Am Heart J 1988; 116 (6 Part 1): 1663-5. [Medline]

5. Williamson DF, Parker RA, Kendrick JS. The box plot: a simple visual method to interpret data. Ann Intern Med 1989; 110 (11): 916-21. [Medline]

6.Altman DA. Practical statistics for medical research. 1th ed., repr. 1997. London: Chapman & Hall; 1997.