

## 4. NOCIONES ELEMENTALES DE ESTADISTICA.

### 4.1. LA CURVA NORMAL Y UTILIZACIÓN DE TABLAS EN LA CURVA NORMAL.

Al iniciar el análisis estadístico de una serie de datos, y después de la etapa de detección y corrección de errores, un primer paso consiste en describir la distribución de las variables estudiadas y, en particular, de los datos numéricos. Además de las medidas descriptivas correspondientes, el comportamiento de estas variables puede explorarse gráficamente de un modo muy simple. Consideremos, como ejemplo, los datos de la Figura 1a, que muestra un **histograma** de la tensión arterial sistólica de una serie de pacientes isquémicos ingresados en una unidad de cuidados intensivos. Para construir este tipo de gráfico, se divide el rango de valores de la variable en intervalos de igual longitud, representando sobre cada intervalo un rectángulo con área proporcional al número de datos en ese rango<sup>1</sup>. Uniendo los puntos medios del extremo superior de las barras, se obtiene el llamado **polígono de frecuencias**. Si se observase una gran cantidad de valores de la variable de interés, se podría construir un histograma en el que las bases de los rectángulos fuesen cada vez más pequeñas, de modo que el polígono de frecuencias tendría una apariencia cada vez más suavizada, tal y como se muestra en la Figura 1b. Esta curva suave "asintótica" representa de modo intuitivo la distribución teórica de la característica observada. Es la llamada función de densidad.

Una de las distribuciones teóricas mejor estudiadas en los textos de bioestadística y más utilizada en la práctica es la **distribución normal**, también llamada **distribución gaussiana**<sup>2,3,4,5</sup>. Su importancia se debe fundamentalmente a la frecuencia con la que distintas variables asociadas a fenómenos naturales y cotidianos siguen, aproximadamente, esta distribución. Caracteres morfológicos (como la talla o el peso), o psicológicos (como el cociente intelectual) son ejemplos de variables de las que frecuentemente se asume que siguen una distribución normal. No obstante, y aunque algunos autores<sup>6,7</sup> han señalado que el comportamiento de muchos parámetros en el campo de la salud puede ser descrito mediante una distribución normal, puede resultar incluso poco frecuente encontrar variables que se ajusten a este tipo de comportamiento.

El uso extendido de la distribución normal en las aplicaciones estadísticas puede explicarse, además, por otras razones. Muchos de los procedimientos estadísticos habitualmente utilizados asumen la normalidad de los datos observados. Aunque muchas de estas técnicas no son demasiado sensibles a desviaciones de la normal y, en general, esta hipótesis puede obviarse cuando se dispone de un

número suficiente de datos, resulta recomendable contrastar siempre si se puede asumir o no una distribución normal. La simple exploración visual de los datos puede sugerir la forma de su distribución. No obstante, existen otras medidas, gráficos de normalidad y contrastes de hipótesis que pueden ayudarnos a decidir, de un modo más riguroso, si la muestra de la que se dispone procede o no de una distribución normal. Cuando los datos no sean normales, podremos o bien transformarlos<sup>8</sup> o emplear otros métodos estadísticos que no exijan este tipo de restricciones (los llamados métodos no paramétricos).

A continuación se describirá la distribución normal, su ecuación matemática y sus propiedades más relevantes, proporcionando algún ejemplo sobre sus aplicaciones a la inferencia estadística. En la sección 3 se describirán los métodos habituales para contrastar la hipótesis de normalidad.

### ▲ La Distribución Normal

La distribución normal fue reconocida por primera vez por el francés Abraham de Moivre (1667-1754). Posteriormente, Carl Friedrich Gauss (1777-1855) elaboró desarrollos más profundos y formuló la ecuación de la curva; de ahí que también se la conozca, más comúnmente, como la "**campana de Gauss**". La distribución de una variable normal está completamente determinada por dos parámetros, su media y su desviación estándar, denotadas generalmente por  $\mu$  y  $\sigma$ . Con esta notación, la densidad de la normal viene dada por la ecuación:

$$\text{Ecuación 1:} \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}; \quad -\infty < x < \infty$$

que determina la curva en forma de campana que tan bien conocemos (Figura 2). Así, se dice que una característica  $X$  sigue una distribución normal de media  $\mu$  y varianza  $\sigma^2$ , y se denota como  $X \approx N(\mu, \sigma)$ , si su función de densidad viene dada por la Ecuación 1.

Al igual que ocurriría con un histograma, en el que el área de cada rectángulo es proporcional al número de datos en el rango de valores correspondiente si, tal y como se muestra en la Figura 2, en el eje horizontal se levantan perpendiculares en dos puntos  $a$  y  $b$ , el área bajo la curva delimitada por esas líneas indica la probabilidad de que la variable de interés,  $X$ , tome un valor cualquiera en ese intervalo. Puesto que la curva alcanza su mayor altura en torno a la media, mientras que sus "ramas" se extienden asintóticamente hacia los

ejes, cuando una variable siga una distribución normal, será mucho más probable observar un dato cercano al valor medio que uno que se encuentre muy alejado de éste.

### Propiedades de la distribución normal:

La distribución normal posee ciertas propiedades importantes que conviene destacar:

- i. Tiene una única moda, que coincide con su media y su mediana.
- ii. La curva normal es asintótica al eje de abscisas. Por ello, cualquier valor entre  $-\infty$  y  $+\infty$  es teóricamente posible. El área total bajo la curva es, por tanto, igual a 1.
- iii. Es simétrica con respecto a su media  $\mu$ . Según esto, para este tipo de variables existe una probabilidad de un 50% de observar un dato mayor que la media, y un 50% de observar un dato menor.
- iv. La distancia entre la línea trazada en la media y el punto de inflexión de la curva es igual a una desviación típica ( $\sigma$ ). Cuanto mayor sea  $\sigma$ , más aplanada será la curva de la densidad.
- v. El área bajo la curva comprendido entre los valores situados aproximadamente a dos desviaciones estándar de la media es igual a 0.95. En concreto, existe un 95% de posibilidades de observar un valor comprendido en el intervalo  $(\mu - 1.96\sigma, \mu + 1.96\sigma)$ .
- vi. La forma de la campana de Gauss depende de los parámetros  $\mu$  y  $\sigma$  (Figura 3). La media indica la posición de la campana, de modo que para diferentes valores de  $\mu$  la gráfica es desplazada a lo largo del eje horizontal. Por otra parte, la desviación estándar determina el grado de apuntamiento de la curva. Cuanto mayor sea el valor de  $\sigma$ , más se dispersarán los datos en torno a la media y la curva será más plana. Un valor pequeño de este parámetro indica, por tanto, una gran probabilidad de obtener datos cercanos al valor medio de la distribución.

Como se deduce de este último apartado, no existe una única distribución normal, sino una familia de distribuciones con una forma común, diferenciadas por los valores de su media y su varianza. De entre todas ellas, la más utilizada es la **distribución normal estándar**, que corresponde a una distribución de media 0 y varianza 1. Así, la expresión que define su densidad se puede obtener de la Ecuación 1, resultando:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right); \quad -\infty < z < \infty$$

Es importante conocer que, a partir de cualquier variable  $X$  que siga una distribución  $N(\mu, \sigma)$ , se puede obtener otra característica  $Z$  con una distribución normal estándar, sin más que efectuar la transformación:

$$\text{Ecuación 2:} \quad Z = \frac{X - \mu}{\sigma}$$

Esta propiedad resulta especialmente interesante en la práctica, ya que para una distribución  $N(0,1)$  existen tablas publicadas (Tabla 1) a partir de las que se puede obtener de modo sencillo la probabilidad de observar un dato menor o igual a un cierto valor  $z$ , y que permitirán resolver preguntas de probabilidad acerca del comportamiento de variables de las que se sabe o se asume que siguen una distribución aproximadamente normal.

Consideremos, por ejemplo, el siguiente problema: supongamos que se sabe que el peso de los sujetos de una determinada población sigue una distribución aproximadamente normal, con una media de 80 Kg y una desviación estándar de 10 Kg. ¿Podremos saber cuál es la probabilidad de que una persona, elegida al azar, tenga un peso superior a 100 Kg?

Denotando por  $X$  a la variable que representa el peso de los individuos en esa población, ésta sigue una distribución  $N(80,10)$ . Si su distribución fuese la de una normal estándar podríamos utilizar la Tabla 1 para calcular la probabilidad que nos interesa. Como éste no es el caso, resultará entonces útil transformar esta característica según la Ecuación 2, y obtener la variable:

$$Z = \frac{X - 80}{10}$$

para poder utilizar dicha tabla. Así, la probabilidad que se desea calcular será:

$$P(X > 100) = P\left(Z > \frac{100 - 80}{10}\right) = P(Z > 2)$$

Como el área total bajo la curva es igual a 1, se puede deducir que:

$$P(Z > 2) = 1 - P(Z \leq 2)$$

Esta última probabilidad puede ser fácilmente obtenida a partir de la Tabla 1, resultando ser  $P(Z \leq 2) = 0.9772$ . Por lo tanto, la probabilidad buscada de que una persona elegida aleatoriamente de esa población tenga un peso mayor de 100 Kg, es de  $1 - 0.9772 = 0.0228$ , es decir, aproximadamente de un 2.3%.

De modo análogo, podemos obtener la probabilidad de que el peso de un sujeto esté entre 60 y 100 Kg:

$$P(60 \leq X \leq 100) = P\left(\frac{60 - 80}{10} \leq Z \leq \frac{100 - 80}{10}\right) = P(-2 \leq Z \leq 2)$$

De la Figura 2, tomando  $a = -2$  y  $b = 2$ , podemos deducir que:

$$P(-2 \leq Z \leq 2) = P(Z \leq 2) - P(Z \leq -2)$$

Por el ejemplo previo, se sabe que  $P(Z \leq 2) = 0.9772$ . Para la segunda probabilidad, sin embargo, encontramos el problema de que las tablas estándar no proporcionan el valor de  $P(Z \leq z)$  para valores negativos de la variable. Sin embargo, haciendo uso de la simetría de la distribución normal, se tiene que:

$$P(Z \leq -2) = P(Z \geq 2) = 1 - P(Z \leq 2) = 1 - 0.9772 = 0.0228$$

Finalmente, la probabilidad buscada de que una persona elegida al azar tenga un peso entre 60 y 100 Kg., es de  $0.9772 - 0.0228 = 0.9544$ , es decir, aproximadamente de un 95%. Resulta interesante comprobar que se obtendría la misma conclusión recurriendo a la propiedad (iii) de la distribución normal.

No obstante, es fácil observar que este tipo de situaciones no corresponde a lo que habitualmente nos encontramos en la práctica. Generalmente no se dispone de información acerca de la distribución teórica de la población, sino que más bien el problema se plantea a la inversa: a partir de una muestra extraída al azar de la población que se desea estudiar, se realizan una serie de mediciones y se desea extrapolar los resultados obtenidos a la población de origen. En un ejemplo similar al anterior, supongamos que se dispone del peso de  $n = 100$  individuos de esa misma población, obteniéndose una media muestral de  $\bar{X} = 75$  Kg, y una desviación estándar muestral  $S = 12$  Kg, querríamos extraer alguna conclusión acerca del valor medio real de ese peso en la población original. La solución a este tipo de cuestiones se basa en un resultado elemental de la teoría estadística, el llamado teorema central del límite. Dicho axioma viene a decirnos que las medias de muestras aleatorias de cualquier variable siguen ellas mismas una distribución normal con igual media que la de la

población y desviación estándar la de la población dividida por  $\sqrt{n}$ . En nuestro caso, podremos entonces considerar la media muestral

$\bar{X} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ , con lo cual, a partir de la propiedad (iii) se conoce que aproximadamente un 95% de los posibles valores de  $\bar{X}$  caerían

dentro del intervalo  $\left(\mu - \frac{1.96\sigma}{\sqrt{n}}; \mu + \frac{1.96\sigma}{\sqrt{n}}\right)$ . Puesto que los valores de  $\mu$  y  $\sigma$  son desconocidos, podríamos pensar en aproximarlos por sus

análogos muestrales, resultando  $\left(78 - \frac{1.96 \times 12}{\sqrt{100}}; 78 + \frac{1.96 \times 12}{\sqrt{100}}\right) = (75.6; 80.3)$

. Estaremos, por lo tanto, un 95% seguros de que el peso medio real en la población de origen oscila entre 75.6 Kg y 80.3 Kg. Aunque la teoría

estadística subyacente es mucho más compleja, en líneas generales éste es el modo de construir un intervalo de confianza para la media de una población.

#### **4.2. ESTUDIO ELEMENTAL DE LA BONDAD DE AJUSTE DE UNA DISTRIBUCIÓN EMPIRÍCA A LA DISTRIBUCIÓN NORMAL.**

La verificación de la hipótesis de normalidad resulta esencial para poder aplicar muchos de los procedimientos estadísticos que habitualmente se manejan. Tal y como ya se apuntaba antes, la simple exploración visual de los datos observados mediante, por ejemplo, un histograma o un diagrama de cajas, podrá ayudarnos a decidir si es razonable o no el considerar que proceden de una característica de distribución normal. Como ejemplo, consideremos los histogramas que se muestran en la Figura 4a, correspondientes a una muestra de 100 mujeres de las que se determinó su peso y edad. Para el caso del peso, la distribución se asemeja bastante a la de una normal. Para la edad, sin embargo, es claramente asimétrica y diferente de la gaussiana.

Resulta obvio que este tipo de estudio no puede llevarnos sino a obtener una opinión meramente subjetiva acerca de la posible distribución de nuestros datos, y que es necesario disponer de otros métodos más rigurosos para contrastar este tipo de hipótesis. En primer lugar, deberemos plantearnos el saber si los datos se distribuyen de una forma simétrica con respecto a su media o presentan algún grado de asimetría, pues es ésta una de las características fundamentales de la distribución de Gauss. Aunque la simetría de la distribución pueda valorarse, de modo simple,

atendiendo a algunas medidas descriptivas de la variable en cuestión<sup>8</sup> (comparando, por ejemplo, los valores de media, mediana y moda), resultará útil disponer de algún índice que nos permita cuantificar cualquier desviación. Si se dispone de una muestra de tamaño  $n$ ,  $\{x_1, \dots, x_n\}$  de una característica  $X$ , se define el **coeficiente de asimetría de Fisher** como:

$$\gamma_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left( \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}}$$

a partir del cual podemos considerar que una distribución es simétrica ( $\gamma = 0$ ), asimétrica hacia la izquierda ( $\gamma < 0$ ) o hacia la derecha ( $\gamma > 0$ ). En segundo lugar, podemos preguntarnos si la curva es más o menos "aplastada", en relación con el grado de apuntamiento de una distribución gaussiana. El **coeficiente de aplastamiento o curtosis de Fisher**, dado por:

$$\gamma_2 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left( \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2} - 3$$

permite clasificar una distribución de frecuencias en mesocúrtica (tan aplanada como una normal,  $\gamma_2 = 0$ ), leptocúrtica (más apuntada que una normal,  $\gamma_2 > 0$ ) o platicúrtica (más aplanada que una normal,  $\gamma_2 < 0$ ).

Siguiendo con los ejemplos anteriores, y tal y como cabía esperar, el coeficiente de asimetría toma un valor mayor para la distribución de la edad ( $\gamma_1 = 0.713$ ) que para el peso observado ( $\gamma_1 = 0.135$ ). En cuanto a los niveles de curtosis, no hay apenas diferencias, siendo de  $-0.320$  para el peso y de  $-0.366$  para la edad.

Los **gráficos de probabilidad normal** constituyen otra importante herramienta gráfica para comprobar si un conjunto de datos puede considerarse o no procedente de una distribución normal. La idea básica consiste en enfrentar, en un mismo gráfico, los datos que han sido observados frente a los datos teóricos que se obtendrían de una distribución gaussiana. Si la distribución de la variable coincide con la normal, los puntos se concentrarán en torno a una línea recta, aunque conviene tener en cuenta que siempre tenderá a observarse mayor variabilidad en los extremos (Figura 4a, datos del peso). En

los **gráficos P-P** se confrontan las proporciones acumuladas de una variable con las de una distribución normal. Los **gráficos Q-Q** se obtienen de modo análogo, esta vez representando los cuantiles respecto a los cuantiles de la distribución normal. Además de permitir valorar la desviación de la normalidad, los gráficos de probabilidad permiten conocer la causa de esa desviación. Una curva en forma de "U" o con alguna curvatura, como en el caso de la edad en la Figura 4b, significa que la distribución es asimétrica con respecto a la gaussiana, mientras que un gráfico en forma de "S" significará que la distribución tiene colas mayores o menores que la normal, esto es, que existen pocas o demasiadas observaciones en las colas de la distribución.

Parece lógico que cada uno de estos métodos se complemente con procedimientos de análisis que cuantifiquen de un modo más exacto las desviaciones de la distribución normal. Existen distintos tests estadísticos que podemos utilizar para este propósito.

#### **4.3. EL TEST DE KOLMOGOROV-SMIRNOV**

Es el más extendido en la práctica. Se basa en la idea de comparar la función de distribución acumulada de los datos observados con la de una distribución normal, midiendo la máxima distancia entre ambas curvas. Como en cualquier test de hipótesis, la hipótesis nula se rechaza cuando el valor del estadístico supera un cierto valor crítico que se obtiene de una tabla de probabilidad. Dado que en la mayoría de los paquetes estadísticos, como el SPSS, aparece programado dicho procedimiento, y proporciona tanto el valor del test como el p-valor correspondiente, no nos detendremos más en explicar su cálculo.

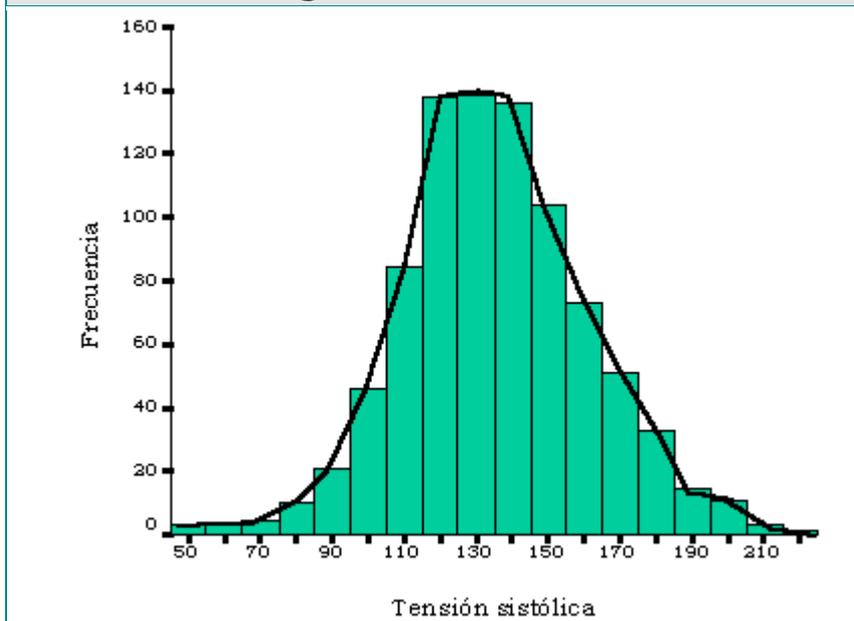
Existen modificaciones de este test, como el de Anderson-Darling que también pueden ser utilizados. Otro procedimiento muy extendido es también el test chi-cuadrado de bondad de ajuste. No obstante, este tipo de procedimientos deben ser utilizados con precaución. Cuando se dispone de un número suficiente de datos, cualquier test será capaz de detectar diferencias pequeñas aún cuando estas no sean relevantes para la mayor parte de los propósitos. El test de Kolmogorov-Smirnov, en este sentido, otorga un peso menor a las observaciones extremas y por lo tanto es menos sensible a las desviaciones que normalmente se producen en estos tramos.

Para acabar, observemos el resultado de aplicar el test de Kolmogorov-Smirnov a los datos de la Figura 4. Para el caso del peso, el valor del estadístico proporcionado por dicho test fue de 0.705, con un p-valor correspondiente de  $p=0.702$  que, al no ser

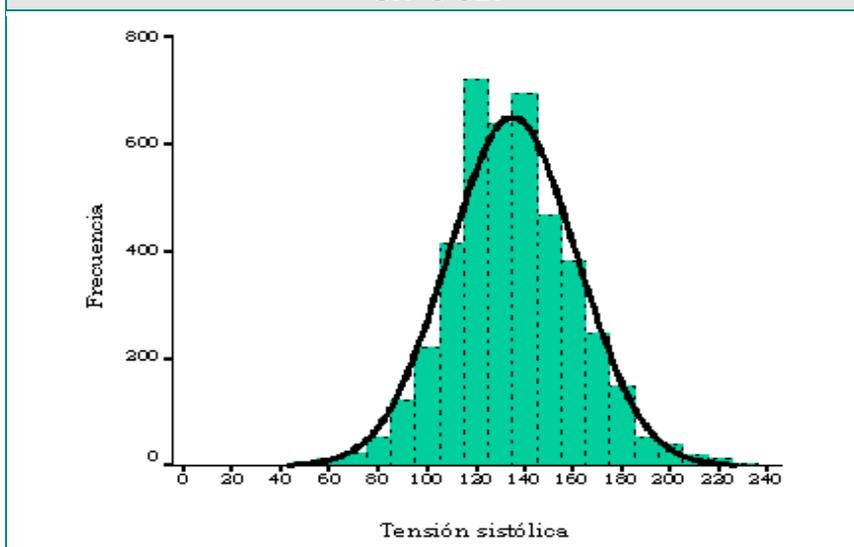
significativo, indica que podemos asumir una distribución normal. Por otra parte, para el caso de la edad, en el que la distribución muestral era mucho más asimétrica, el mismo test proporcionó un valor de 1.498, con  $p=0.022$ , lo que obligaría a rechazar en este caso la hipótesis de una distribución gaussiana.

**Figura 1. Histograma de los valores de tensión arterial sistólica para dos muestras de pacientes isquémicos ingresados en una unidad de cuidados intensivos.**

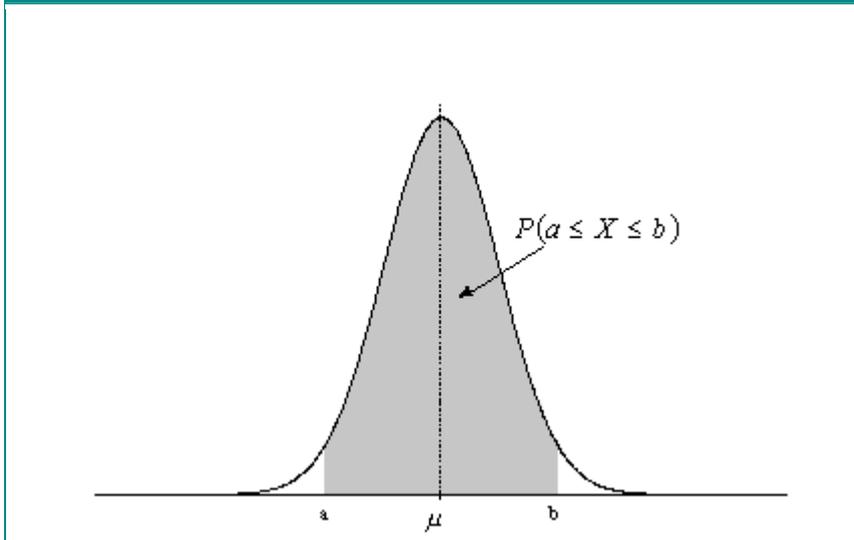
**Figura 1a.- Valores de tensión arterial sistólica en una muestra de 1000 pacientes isquémicos ingresados en UCI.**



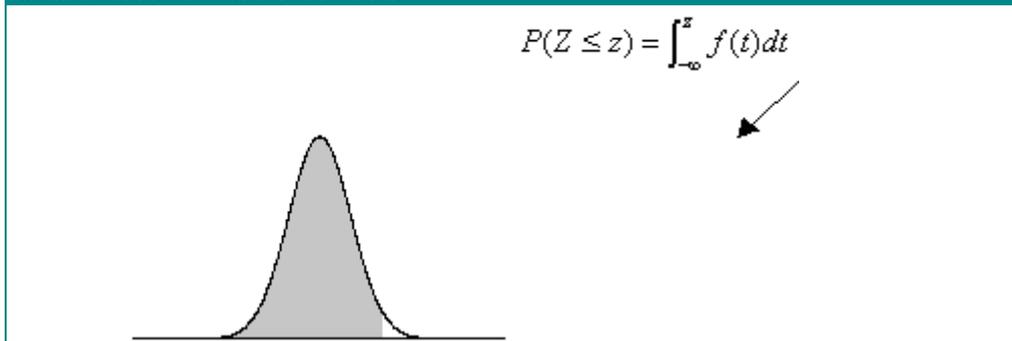
**Figura 1b.- Valores de tensión arterial sistólica de una muestra de 5000 pacientes ingresados en UCI.**



**Figura 2. Gráfica de una distribución normal y significado del área bajo la curva.**



**Tabla 1. Áreas bajo la curva normal estándar. Los valores de la tabla que no se muestran en negrita representan la probabilidad de observar un valor menor o igual a  $z$ . La cifra entera y el primer decimal de  $z$  se buscan en la primera columna, y el segundo decimal en la cabecera de la tabla.**



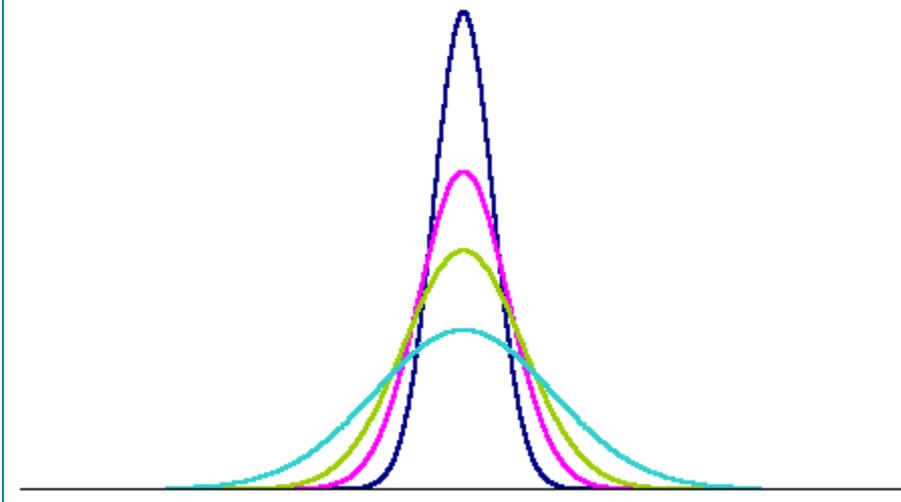
		<b>Segunda cifra decimal del valor de <math>z</math></b>									
<b><math>z</math></b>	<b>0.00</b>	<b>.01</b>	<b>.02</b>	<b>.03</b>	<b>.04</b>	<b>.05</b>	<b>.06</b>	<b>.07</b>	<b>.08</b>	<b>.09</b>	
<b>0.0</b>	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359	
<b>0.1</b>	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753	
<b>0.2</b>	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141	
<b>0.3</b>	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517	
<b>0.4</b>	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879	
<b>0.5</b>	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224	
<b>0.6</b>	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549	
<b>0.7</b>	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852	
<b>0.8</b>	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133	
<b>0.9</b>	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389	
<b>1.0</b>	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621	
<b>1.1</b>	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830	
<b>1.2</b>	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015	
<b>1.3</b>	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177	
<b>1.4</b>	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319	

<b>1.5</b>	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
<b>1.6</b>	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
<b>1.7</b>	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
<b>1.8</b>	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
<b>1.9</b>	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
<b>2.0</b>	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
<b>2.1</b>	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
<b>2.2</b>	.9861	.9864	.9868	.9871	.9875	.4878	.9881	.9884	.9887	.9890
<b>2.3</b>	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
<b>2.4</b>	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
<b>2.5</b>	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
<b>2.6</b>	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
<b>2.7</b>	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
<b>2.8</b>	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
<b>2.9</b>	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
<b>3.0</b>	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
<b>3.1</b>	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
<b>3.2</b>	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
<b>3.3</b>	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
<b>3.4</b>	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998



**Figura 3. Ejemplos de distribuciones normales con diferentes parámetros.**

(a) Distribuciones normales con distinta desviación estándar e igual media



(b) Distribuciones normales con diferentes medias e igual desviación estándar

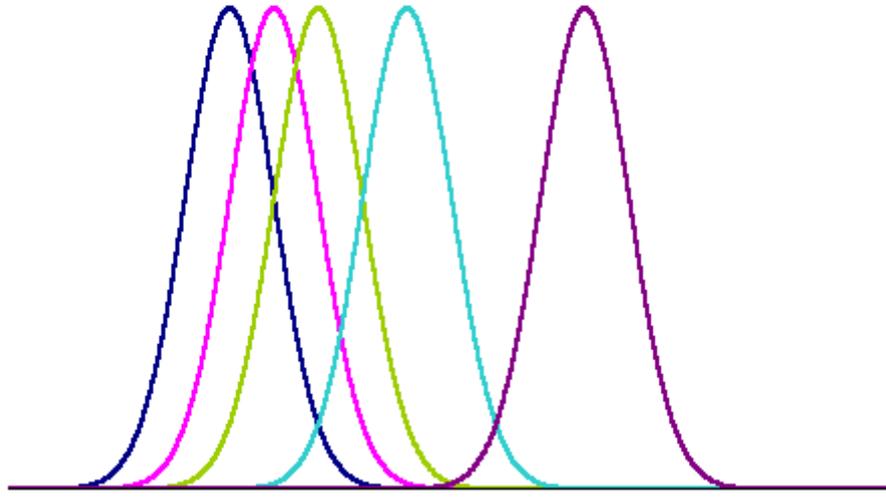
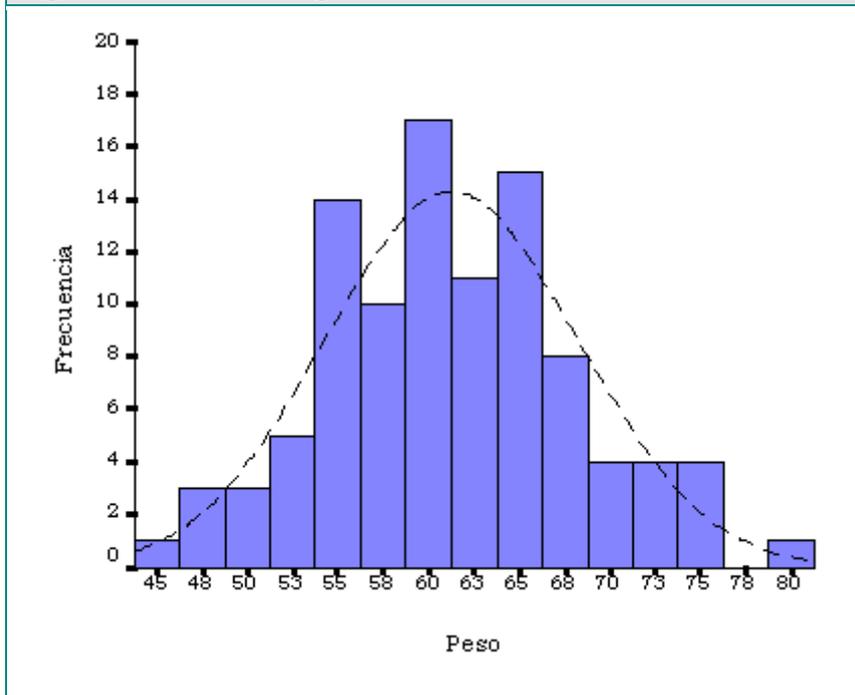
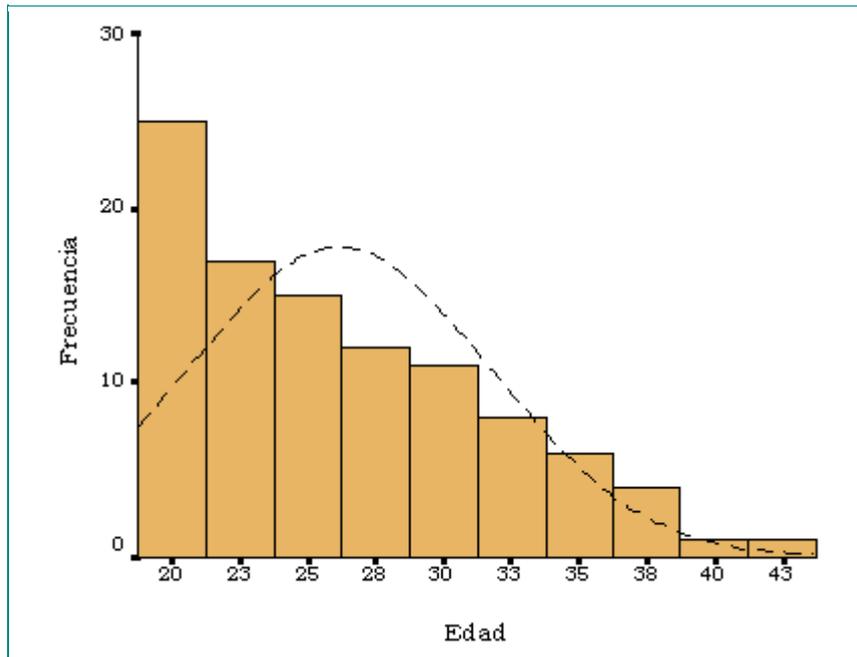


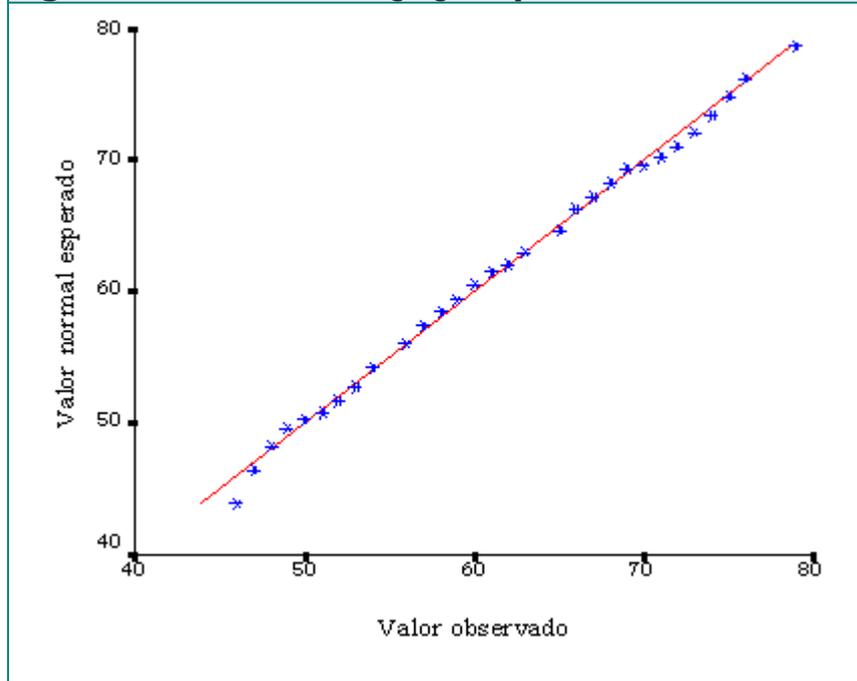
Figura 4. Histogramas y gráficos de probabilidad normal de los valores de peso y edad en dos muestras de pacientes.

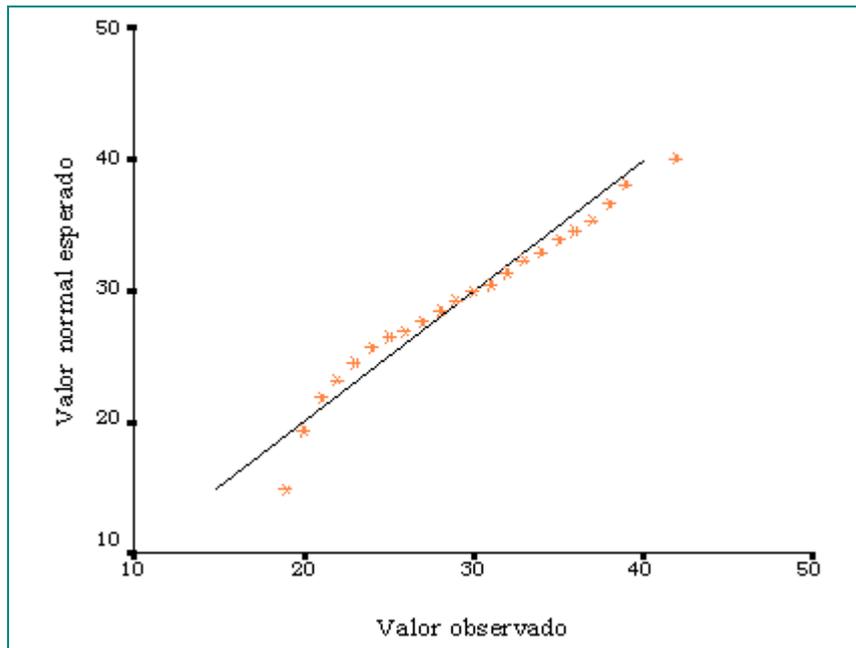
### Figura 4a.- Histogramas





**Figura 4b.- Gráficos Q-Q de probabilidad.**





### 4.3.1. PRUBA DE BONDAD DE LA CHI CUADRADO

La metodología de la prueba es la siguiente:

1. Se colocan los  $n$  datos históricos en una tabla de frecuencia de  $m=n^{1/2}$  intervalos. Se obtiene la frecuencia observada en cada intervalo  $i$  ( $FO_i$ ). Se calcula la media y la varianza de los datos.
2. Se propone una distribución de probabilidad de acuerdo con la forma de la tabla de frecuencias obtenida en el paso 1.
3. Con la distribución propuesta, se calcula la frecuencia esperada para cada uno de los intervalos ( $FE_i$ ) mediante la integración de la distribución propuesta y su posterior multiplicación por el número total de datos.
4. Se calcula el estimador:

$$C = \frac{(FE_1 - FO_1)^2}{FE_1} + \frac{(FE_2 - FO_2)^2}{FE_2} + \frac{(FE_3 - FO_3)^2}{FE_3} + \dots + \frac{(FE_m - FO_m)^2}{FE_m}$$

5. Si el estimador  $C$  es menor o igual al valor correspondiente  $C$  con  $m-k-1$  grados de libertad ( $K =$  número de parámetros estimados de la distribución) y a un nivel de confianza, entonces no se puede rechazar la hipótesis de que la información histórica sigue la distribución propuesta en el punto 2.

**(A)**

## PRUBA DE BONDAD DE AJUSTE DE KOLMOGOROV.

Si el objetivo es encontrar el tipo de distribución de probabilidad de una serie de datos, es posible utilizar la prueba de bondad de ajuste de Kolmogorov-Smirnov.

La metodología es la siguiente:

1. Se colocan los  $n$  datos históricos en una tabla de frecuencias con  $m = n^{(1/2)}$  intervalos. Para cada intervalo se obtendrá la frecuencia observada  $i$  ( $FO_i$ ).
2. Se divide la frecuencia observada de cada intervalo por el número total de datos. A este resultado se le llama la probabilidad observada  $i$  ( $PO_i$ ).
3. Se calcula la probabilidad acumulada observada de cada intervalo ( $PAO_i$ ) del paso 2.
4. Se propone una distribución de probabilidad de acuerdo con la forma de la tabla de frecuencias obtenida en 1. (elaborar histograma)
5. Con la distribución propuesta se calcula la probabilidad esperada para cada uno de los intervalos ( $PE_i$ ) utilizando la distribución propuesta.
6. Se calcula la probabilidad acumulada esperada ( $PAE_i$ ) para cada intervalo de clase.
7. Se calcula la diferencia absoluta entre  $PAO_i$  y  $PAE_i$  para cada intervalo y se selecciona la máxima diferencia, llamándola  $DM$ .
8. El estimado  $DM$  se compara con un valor límite proporcionado en tabla, con  $n$  datos y a un nivel de confiabilidad de  $1 - \alpha$ . Si el estimador  $DM$  es menor o igual al valor límite de la tabla, entonces no se puede rechazar que la información histórica sigue la distribución propuesta en el paso 4.

#### 4.4. ESCALAS Y PUNTUACIONES TIPIFICADAS BASADAS EN LA PUNTUACIÓN (Z, T, S Y CEEB).

##### PUNTUACIÓN Z

Es la puntuación típica o **resultado estándar individual** más utilizada. Gracias a la puntuación z podemos comparar los resultados de un individuo (o de varios) en distintas pruebas en las que las distribuciones de los resultados de los otros miembros de la población sean distintas.

##### Medida típica Z

Si se divide una medida diferencial entre la desviación típica del grupo, se obtiene la *medida o puntuación típica Z*:

$$Z = \frac{X - \bar{X}}{s} = \frac{x}{s}$$

Esta medida es muy importante, pues permite comparar dos colectivos distintos, debido a la siguiente propiedad:

**La media de las puntuaciones Z siempre es cero y su desviación típica siempre es 1**

Como consecuencia de lo anterior, se cumple una curiosidad:

$$\sum Z^2 = N$$

De esta forma, mediante Z, las medidas obtenidas por cualquier sujeto en variables diferentes, siempre tendrán media 0 y desviación 1, con lo que Z mide **el verdadero nivel** dentro de cada grupo, al haber eliminado los parámetros de centro y dispersión. Es como si dos conjuntos los redujéramos a la misma escala para poderlos comparar.

Según la **desigualdad de Tchebychev**,

$$Pr(|x_i - \bar{x}| \leq ks) \geq 1 - \frac{1}{k^2}$$

las medidas Z comprendidas entre -2 y 2 suponen como mínimo el 75% de los datos. Así, puntuaciones superiores a 2 o inferiores a -2 son extraordinarias, en el sentido de que lo probable es lo contrario.

De igual forma, entre -3 y 3 están contenidos al menos el 89% de los datos. Puntuaciones más alejadas que 3 y -3 se consideran improbables.

## **PUNTUACIÓN T**

### **Escala T**

La puntuación T se obtiene multiplicando Z por 10 y después sumando 50:

$$T = 10Z + 50$$

Lo normal es que una puntuación T oscile entre 20 y 80 puntos. Las medidas más extremas son improbables.

### **Cociente intelectual**

Independientemente de su definición, el **cociente intelectual** (y escalas similares) está construido con la fórmula

$$CI = 15Z + 100$$

Por tanto, los cocientes “normales” estarán entre 55 y 145

## **LA PUNTUACIÓN S**

La escala S tiene como fórmula  $S = 50 + 20Z$

La escala D equivale a  $D = 5 + 2Z$

## **PUNTUACIONES CEEB**

El **College Board** es una organización sin fines de lucro examen bordo en el Estados Unidos que se formó en 1900 como el **College Entrance Examination Board (CEEB)**. Que gestiona diversas pruebas estandarizadas que en general atienden a las instituciones de evaluación de desempeño de los estudiantes en el tercer o cuarto año de escuela secundaria de planificación en continuar sus estudios en una etapa posterior a la secundaria.

El College Board mantiene un registro numerado de los países, universitarias, colegios, programas de becas, centros de prueba, y las escuelas secundarias. En los Estados Unidos, además de la Escuela de la Junta de uso interno de este registro es prestado por otras instituciones como medio de identificación inequívoca, por lo que un estudiante podría dar su departamento de orientación, no sólo un colegio de nombre y dirección, sino también su CEEB código, a fin de garantizar que su transcripción es enviado correctamente. Existe un conjunto similar de ACT códigos para colegios y becas [1], centros de prueba [2], y las escuelas secundarias [3] Sin embargo, estos códigos son menos utilizadas fuera de ACT, Inc.

### **Las pruebas del College Board**

#### **Sábado**

La Prueba de Razonamiento del Sábado es una prueba de admisión universitaria en los Estados Unidos. El SAT es administrado por el College Board de corporación en los Estados Unidos y es desarrollada, publicada, y anotó por el Educational Testing Service (ETS). SAT Subject Tests se dice para medir el rendimiento de los alumnos en áreas específicas, tales como las matemáticas, la ciencia, y historia.

#### **PSAT / NMSQT**

PSAT / NMSQT significa preliminares SAT / National Merit Scholarship Qualifying Test. Es una prueba que ofrece de primera mano para la práctica de prueba de Razonamiento del SAT. También funciona como una prueba de calificación para la National Merit Scholarship Corporation programas de becas.

### **Colegio Nivel Examen del Programa**

Colegio Nivel Examen Programa provee a los estudiantes de cualquier edad con la oportunidad de demostrar el logro de nivel universitario a través de un programa de exámenes en los cursos de pregrado universitario. Los críticos del Advanced Placement Program acusación de que los cursos y los exámenes se centran en la amplitud de la cobertura de contenido en lugar de profundidad, y que los estudiantes aprenden casi exclusivamente acerca de los EE.UU. o Europa occidental. Adelantado los exámenes de ubicación, a lo largo de los

exámenes del Bachillerato Internacional, a menudo avanzada de pie para recibir mejores puntuaciones en los exámenes de AP o IB. Hay 2.900 colegios que otorgan el crédito y / o avanzados de pie.

### **Advanced Placement Program**

El Colegio de la Junta de Advanced Placement Program es un amplio programa que ofrece la escuela secundaria los estudiantes la oportunidad de participar en lo que describen como las clases de nivel universitario, según se informa, la ampliación de sus horizontes intelectuales y su preparación para la universidad el trabajo. También desempeña una parte importante en el proceso de admisión universitaria, que muestra tanto la capacidad intelectual del estudiante y genuino interés en el aprendizaje. El programa permite a los estudiantes a ganar muchos créditos universitarios de alto rendimiento en los exámenes de la AP, tanto en la misma forma que el CLEP. Concesión de créditos sin embargo, se encuentra aún en la discreción de la universidad.

### **CSS / Ayuda Financiera PERFIL**

El College Board ofrece también la CSS / Ayuda Financiera PERFIL, una solicitud de ayuda económica de servicios que muchas instituciones la aplicación de la familia en la determinación de la contribución y la asistencia financiera a los paquetes.

## BIBLIOGRAFÍA

Pértega Díaz S, Pita Fernández S. Representación gráfica en el análisis de datos. *Cad Aten Primaria* 2001; 8: 112-117.

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

Daniel WW. *Bioestadística. Base para el análisis de las ciencias de la salud*. Mexico: Limusa; 1995.

Elston RC, Johnson WD. *Essentials of Biostatistics*. Philadelphia: F.A. Davis Company; 1987.

Altman DG, Bland JM. Statistics notes: The normal distribution. *BMJ* 1995; 310: 298-298.

Bisquera, R., *Metodología de la investigación educativa*, Madrid, La Muralla, 2005.

Etxeberría, J.; Tejedor, F. J., *Análisis descriptivo de datos en educación*, Madrid, la Muralla, 2005.

Fernández Díaz, M. J., García Ramos, J.M., *225 problemas de estadística aplicada a las Ciencias Sociales*, Madrid, Síntesis. 1991.

Fernández Díaz, M. J., García Ramos, J. M., *Resolución de problemas de la estadística aplicada a las Ciencias Sociales*. Madrid, Síntesis, 1991.

Latorre, A., del Rincón, D., y Arnal J., *Bases Metodológicas de la Investigación Educativa*, Barcelona, GR92. 1996.