DISCIPLINAS ESPECIALIZADAS

LA MATEMÁTICA ESTADÍSTICA

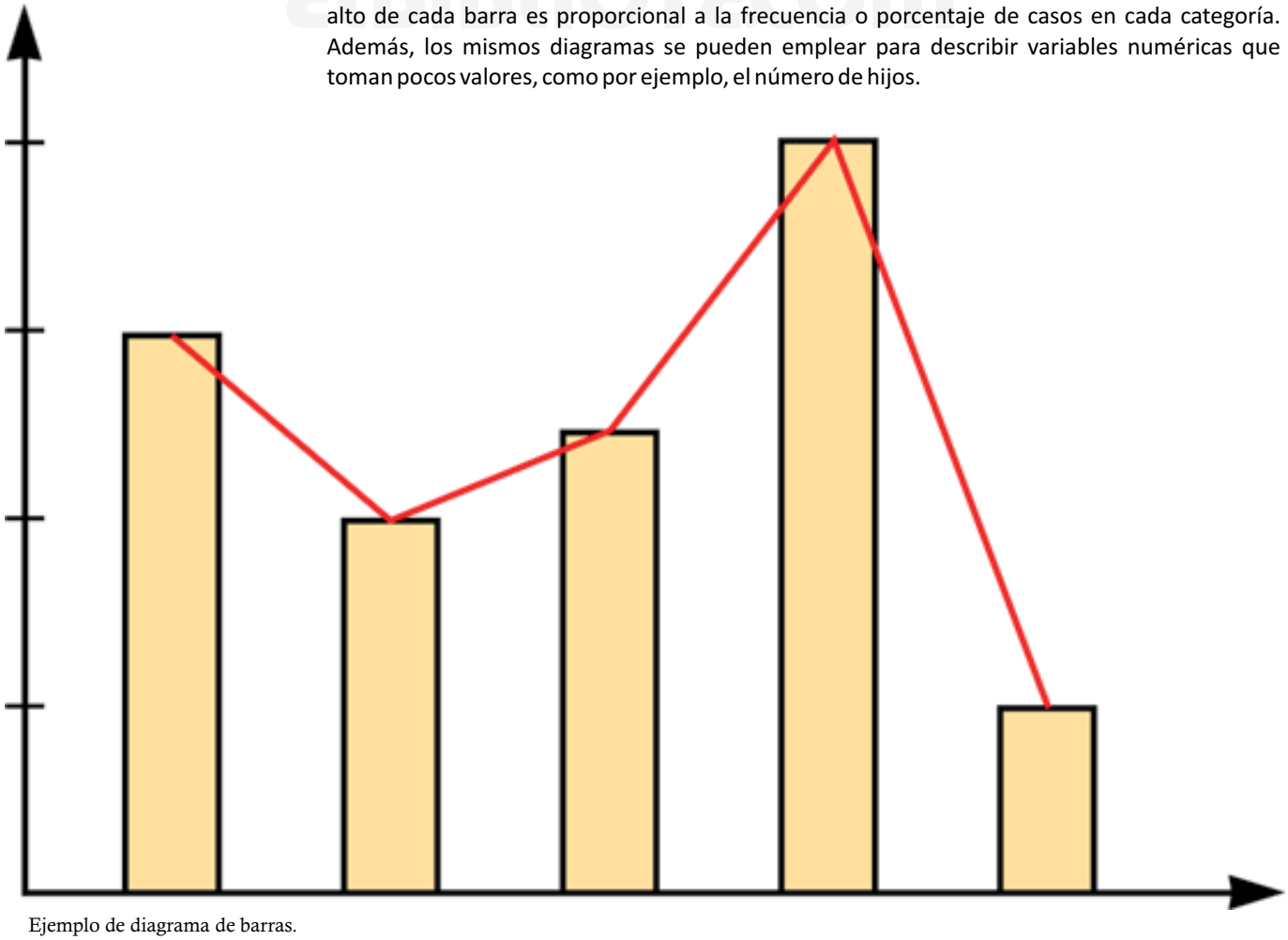
ANÁLISIS DESCRIPTIVO:

Al tener disponibles datos de una población, y a previamente al abordaje de análisis estadísticos de mayor complejidad, lo primero que corresponde hacer es presentar dicha información de una manera tal que la misma pueda ser visualizada sistemática y resumidamente. Los datos que son de interés, para cada caso estar en función del tipo de variables que se manejen.

Entonces, para variables categóricas, tales como el sexo, el estado civil, la profesión, etc., es indispensable conocer la frecuencia y el porcentaje del total de casos que se ubican en cada categoría. Una manera simple de representar gráficamente tales resultados es por medio de la utilización de diagramas de barras o sectoriales.

En los gráficos sectoriales o de torta, el circulo se divide en una cantidad de proporciones equivalente a la cantidad de clases que posea la variable, por lo tanto, cada clase corresponde a un arco de circulo proporcional a su frecuencia, ya sea esta ultima absoluta o relativa.

Los gráficos de barras son afines a los diagramas de sectores, no obstante, la diferencia radica en que en estos la cantidad de categorías de la variable se representan en barras, de tal forma que el alto de cada barra es proporcional a la frecuencia o porcentaje de casos en cada categoría. Además, los mismos diagramas se pueden emplear para describir variables numéricas que toman pocos valores, como por ejemplo, el número de hijos.



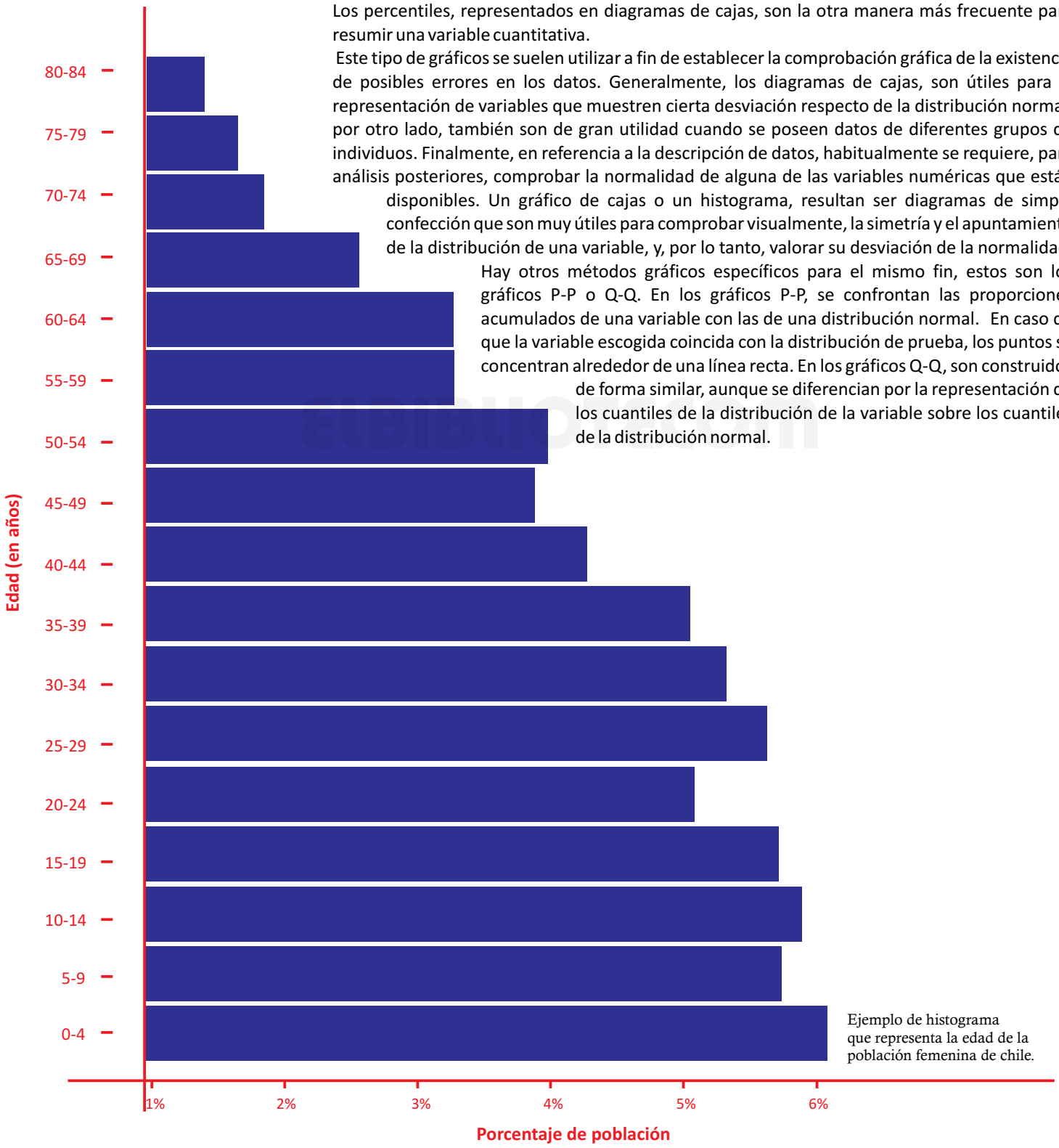
Ejemplo de diagrama de barras.

El histograma es el tipo de gráfico que se utiliza frecuentemente para las variables numéricas continuas, como por ejemplo, edad, tensión arterial, índice de masa corporal, etc. el histograma se construye a partir de las división en rango de valores de la variable en intervalos que tengan la misma amplitud, representando sobre cada intervalo un rectángulo que posea al segmento como base. La altura de rectángulo se calcula manteniendo la proporcionalidad entre las frecuencias absolutas o relativas de los datos en cada intervalo y el área de los rectángulos.

Los percentiles, representados en diagramas de cajas, son la otra manera más frecuente para resumir una variable cuantitativa.

Este tipo de gráficos se suelen utilizar a fin de establecer la comprobación gráfica de la existencia de posibles errores en los datos. Generalmente, los diagramas de cajas, son útiles para la representación de variables que muestren cierta desviación respecto de la distribución normal; por otro lado, también son de gran utilidad cuando se poseen datos de diferentes grupos de individuos. Finalmente, en referencia a la descripción de datos, habitualmente se requiere, para análisis posteriores, comprobar la normalidad de alguna de las variables numéricas que están disponibles. Un gráfico de cajas o un histograma, resultan ser diagramas de simple confección que son muy útiles para comprobar visualmente, la simetría y el apuntamiento de la distribución de una variable, y, por lo tanto, valorar su desviación de la normalidad.

Hay otros métodos gráficos específicos para el mismo fin, estos son los gráficos P-P o Q-Q. En los gráficos P-P, se confrontan las proporciones acumulados de una variable con las de una distribución normal. En caso de que la variable escogida coincida con la distribución de prueba, los puntos se concentran alrededor de una línea recta. En los gráficos Q-Q, son construidos de forma similar, aunque se diferencian por la representación de los cuantiles de la distribución de la variable sobre los cuantiles de la distribución normal.

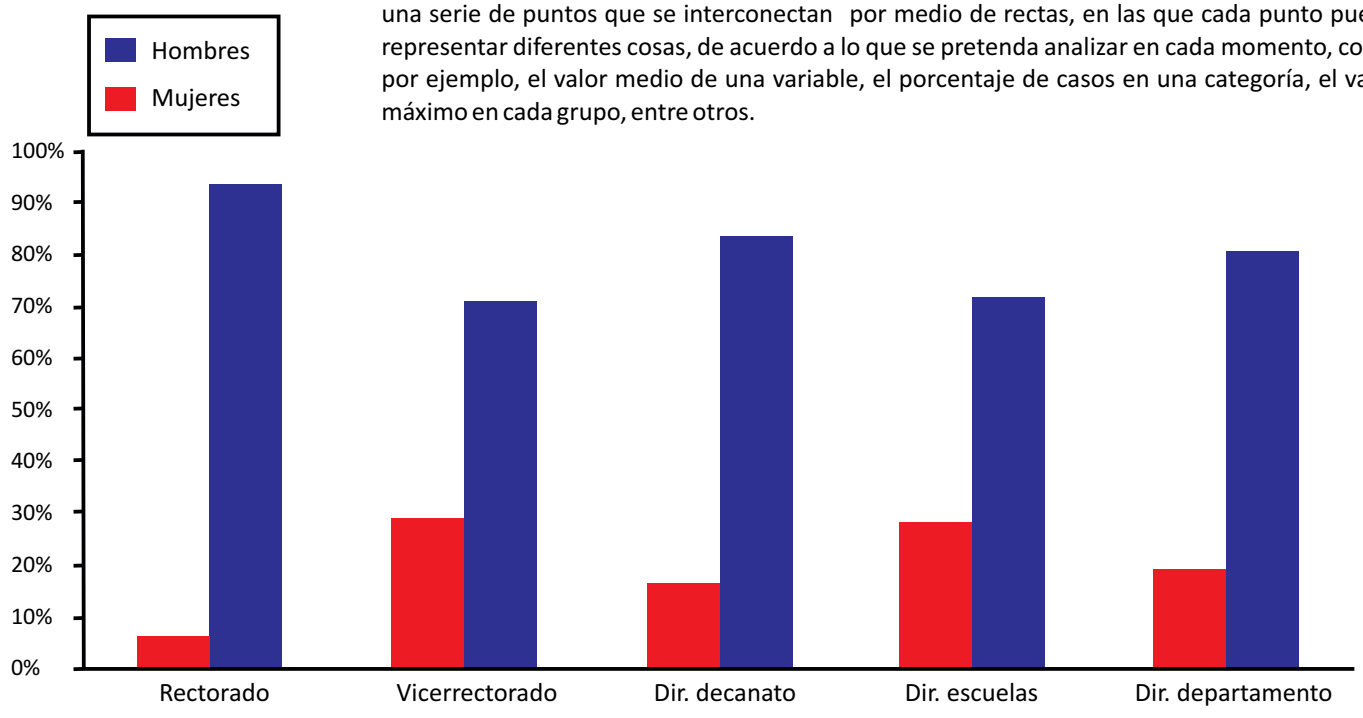


Ejemplo de histograma que representa la edad de la población femenina de chile.

COMPARACIÓN DE DOS O MÁS GRUPOS

Si se pretende efectuar una comparación de las observaciones tomadas en dos o más grupos de sujetos, conviene emplear el método estadístico capaz de ajustarse al tipo de variables manejadas. Al trabajar con dos variables cualitativas es posible continuar la utilización de gráficos de barras o sectoriales, por ejemplo, si desea establecer el hecho de que en una determinada muestra, la frecuencia de individuos que padecen una enfermedad coronaria es mayor en los que tienen antecedente familiares de padecimientos cardiacos. Con esta muestra, es posible representar dos grupos de barras; uno para los sujetos con antecedentes cardiacos de algún familiar y otro para los que no poseen esta clase de antecedentes. Para cada uno de los grupos, se dibujan dos barras que representan el porcentaje de pacientes que tiene o no alguna enfermedad coronaria. Es importante recodar que si los tamaños de ambas poblaciones son distintos, conviene emplear las frecuencias relativas, porque de lo contrario el diagrama podría ser falaz. De otra parte, si se coparan dos o más grupos de variables continuas, dicho procedimiento es efectuado en términos de su valor medio, mediante el test t de Student, análisis de la varianza o métodos no paramétricos equivalentes, a fin de realizar la representación en el tipo de gráfico empleado. Para tal circunstancia, es recomendable la utilización del gráfico de barras; por ejemplo, se puede representar la comparación entre el índice de masa corporal en una muestra de hombres y mujeres, para cada uno de los grupos es representado su valor medio, junto con su intervalo de confianza. No se debe olvidar que aunque los intervalos no se solapen, esto no quiere decir que indispensablemente la diferencia entre los dos grupos pueda ser estadísticamente significativa, sin embargo, si puede ser útil para valorar la magnitud de la misma. Igualmente, es posible visualizar dichas relaciones con dos diagramas de cajas, uno para cada grupo, dichos gráficos o diagramas son importantes en este caso, ya que dan ligar a la visualización de la existencia o inexistencia de diferencia entre los grupos, y también ofrecen la posibilidad de comprobar la normalidad y la variabilidad de cada una de las distribuciones. Recuérdese, que las hipótesis de normalidad y homocedasticidad con condiciones indispensables para la aplicación de alguno de los procedimientos de análisis paramétricos. Para finalizar, es importante indicar que en la misma circunstancia es posible emplear los diagramas de barras, siendo en tal caso la altura de cada barra la representación del valor medio de la variable que se analiza. También son útiles los gráficos de líneas, en especial, cuando se pretende estudiar las tendencias durante el tiempo. Este tipo de gráficos consisten en una serie de puntos que se interconectan por medio de rectas, en las que cada punto puede representar diferentes cosas, de acuerdo a lo que se pretenda analizar en cada momento, como por ejemplo, el valor medio de una variable, el porcentaje de casos en una categoría, el valor máximo en cada grupo, entre otros.

Al trabajar con dos variables cualitativas es posible continuar la utilización de gráficos de barras.

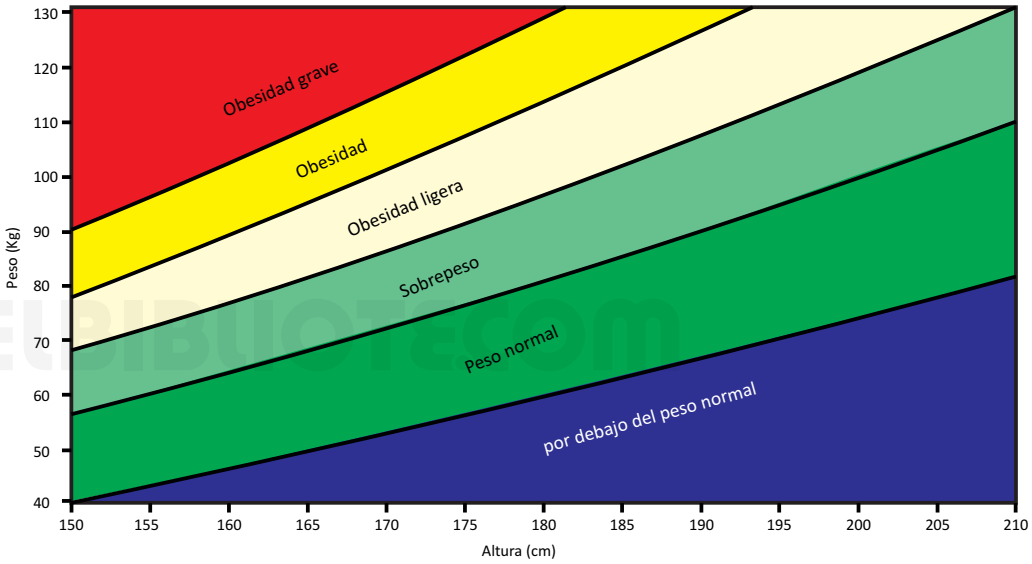


Ejemplo de gráfico con dos grupos.

RELACIÓN ENTRE DOS VARIABLES NUMÉRICAS

El estudio de correlación constituye el método de análisis apropiado para estudiar la relación entre dos variables numéricas; los coeficientes de correlación determinan el valor del aumento o disminución de una de las variables al crecer el valor de la otra variable. En el momento en que están disponibles todos los datos, es posible comprobar grafica y sencillamente la existencia de una alta correlación a través de diagramas de dispersión, en los que se confronta, en el eje horizontal, el valor de una variable y en el eje vertical, el valor de la otra. Se puede mencionar como un ejemplo fácil de variables altamente correlacionadas, la relación existente entre peso y talla de un individuo. A partir de una muestra seleccionada arbitrariamente, es posible confeccionar el grafico de dispersión, donde se puede observar claramente la relación directa que existe entre ambas variables, y valorar hasta qué punto esta relación se puede modelar por la ecuación de una recta. En consecuencia, estos diagramas son de gran utilidad en la fase de selección de variables al momento de ajustar un modelo de regresión lineal.

Relación existente entre peso y talla de un individuo.

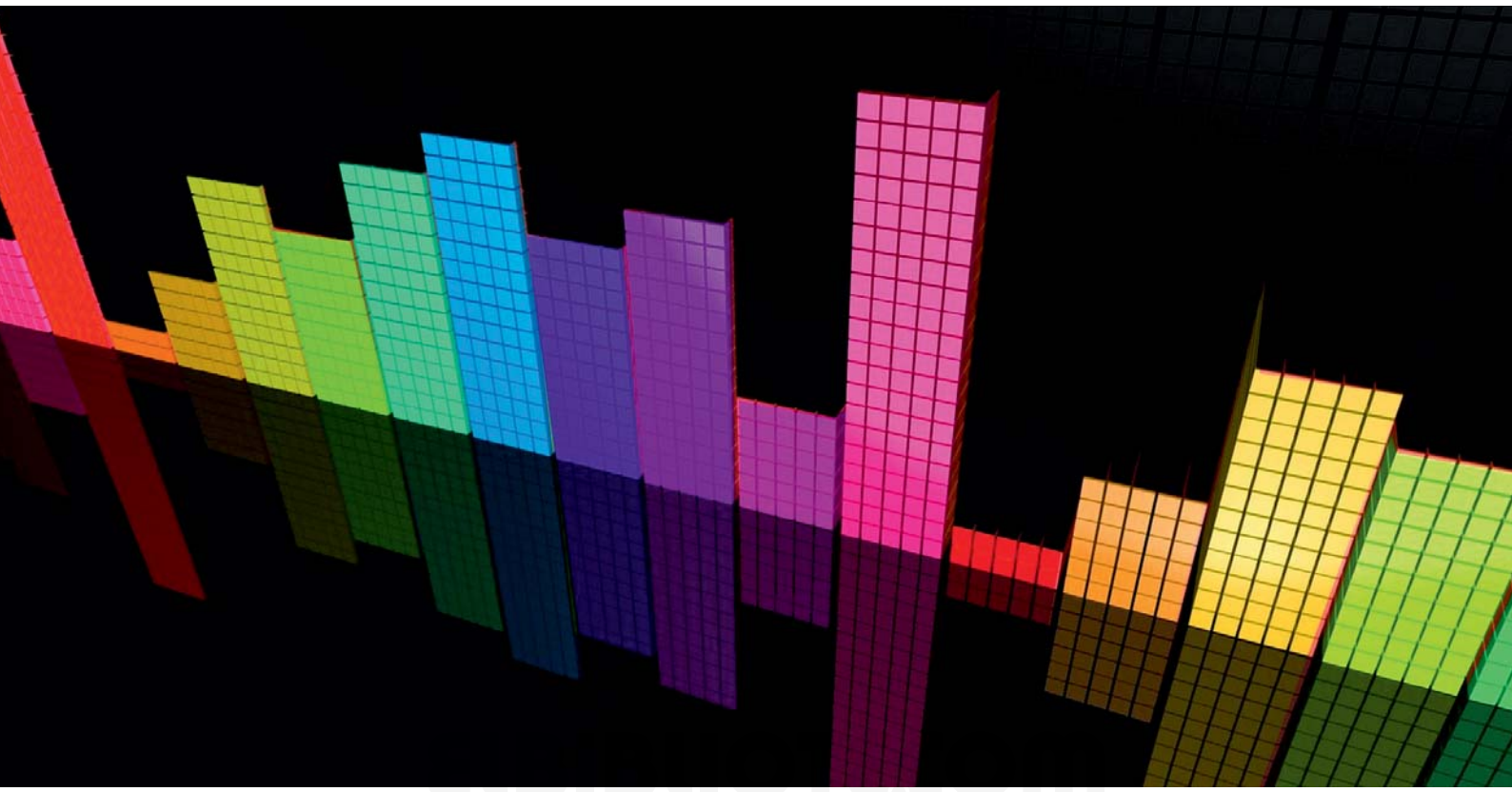


OTROS GRAFICOS

En general, los textos estadísticos y epidemiológicos, enfatizan en las diferentes clases de gráficos que se pueden crear, constituyendo como herramienta indispensable en la presentación de resultados y el proceso de estadístico.

Los gráficos explicados hasta el momento son los de más fácil manejo, sin embargo, tienen la capacidad de brindar enormes posibilidades para la representación de datos y es posible emplearlos en gran cantidad de situaciones, de hecho si se pretende representar los resultados obtenidos a través de métodos de análisis más complicados. Un ejemplo de esos últimos, son dos gráficos de líneas superpuestos para la visualización de los resultados de un análisis de varianza con dos factores. Incluso, existen análisis concretos, basados por completo en la representación grafica. Un caso particular al respecto es el de la confección de curvas ROC y el cálculo del área bajo la curva, ambos el método más adecuado para la valoración de la exactitud de una prueba diagnóstica.

Por lo tanto, a partir de todo lo explicado hasta aquí, es posible comprender la relevancia y utilidad que las representaciones diagramales poseen en el proceso de análisis de datos. En general, los textos estadísticos y epidemiológicos, enfatizan en las diferentes clases de gráficos que se pueden crear, constituyendo como herramienta indispensable en la presentación de resultados y el proceso de estadístico. Sin embargo, resulta difícil saber cuándo es más conveniente un gráfico en lugar de una tabla; entonces, lo que sí se puede hacer más fácilmente es considerarlos como dos formas distintas aunque complementarias para ver los mismo datos. Debido al aumento en el empleo de sistemas informáticos, se ha ido haciendo más fácil la consecución de los mismos; la mayor parte de los paquetes estadísticos brindan enormes posibilidades al respecto. También se pueden construir otros gráficos, en algunos casos tridimensionales, lo que permite obtener grandes transformaciones en su apariencia y facilita la exportación a otros programas que presentan finalmente los resultados del estudio.



ESTADÍSTICA UNIDIMENSIONAL

VARIABLE ALEATORIA Y VARIABLE ESTADÍSTICA

En un experimento aleatorio, los resultados que se obtengan pueden ser eventos dependientes del azar, dando lugar a una variable cuyos valores probablemente se podrán repetir, dichas variables se denominan variables aleatorias, las cuales se constituyen como discretas únicamente cuando tienen la capacidad de tomar un número infinito de valores. En cambio, si se toman muestras de un experimento realizado, los resultados reales reciben la denominación de variable estadística.

Los términos de variable aleatoria y probabilidad, se constituyen como conceptos teóricos obtenidos a partir de una abstracción realizada respecto de los términos de variable estadista y frecuencia, los cuales son considerados luego de haber realizado el experimento, a diferencia de los primeros que son considerados antes de la experimentación.

MEDIDAS DE CENTRALIZACIÓN

Las medidas de centralización representan un conjunto de datos y en general se sitúan en el centro del conjunto de datos, los cuales estarán ordenados de acuerdo a su magnitud.

MEDIANA:

La mediana es el valor de la variable estadística, la cual divide en dos partes iguales a los individuos de una población, supuestos ordenados crecientemente. De manera general, se trata del valor en el cual la función de distribución F (x), toma el valor medio, aunque al definirla de este modo no es única, por lo que se toma la media aritmética de los valores de mediana, o de no existir, se toma como mediana el valor de la población más aproximado a esa mediana ideal.

La transformación de karhunen-loeve es una técnica utilizada frecuentemente en las ciencias sociales y naturales, al intentar resumir un amplio grupo de variables en un nuevo conjunto, más pequeño, manteniendo una porción significativa de la información original.

TRANSFORMACIÓN DE KARHUNEN-LOEVE

La Transformación de Karhunen-Loeve, también conocida como Transformación de Hotelling o Análisis de Componentes Principales, es una técnica cuyo origen está habitualmente asociado a un artículo publicado por Karl Pearson en el año 1901, aunque su primer desarrollo teórico data de 1933 en un artículo de Hotelling.

Se trata de una técnica utilizada frecuentemente en las ciencias sociales y naturales, al intentar resumir un amplio grupo de variables en un nuevo conjunto, más pequeño, manteniendo una porción significativa de la información original. Se trata de determinar la cantidad de dimensiones que se encuentran en un conjunto de datos y buscar los coeficientes que especifican la posición de los ejes que apuntan en las direcciones de máxima variabilidad de los datos. Se origina en la redundancia que existe gran cantidad de veces entre diferentes variables, dicha redundancia son los datos, no la información. Entonces, lo que se intenta:

- Procurar facilidad del estudio de las relaciones que hay entre las variables.

- Procurar que el análisis de dispersión de las observaciones sea más sencillo, mediante la puesta en evidencia de probables agrupamientos, detectando las variables responsables de la dispersión.

Distancia del PH y volumen de LBM1, LBM4, LBM7, LBM9, LBM13 y control con respecto al valor de referencia.

AISLADO	VOLUMEN AISLADO	PH	DISTANCIA EUCLIDIANA
9	475	4,66	43,417
13	426,6	4,65	56,707
7	315,0	5,60	156,529
4	378,3	4,58	99,908
1	371,6	4,71	105,900
Control	375	5,58	102,866

Análisis estadísticos de Hotelling para verificar Ho: mu=mu0 (los valores medios de volumen y pH para cada aislado son iguales a los valores de referencia de volumen y pH: 490 cm3; 4,5)

FORMULACIÓN DESCRIPTIVA

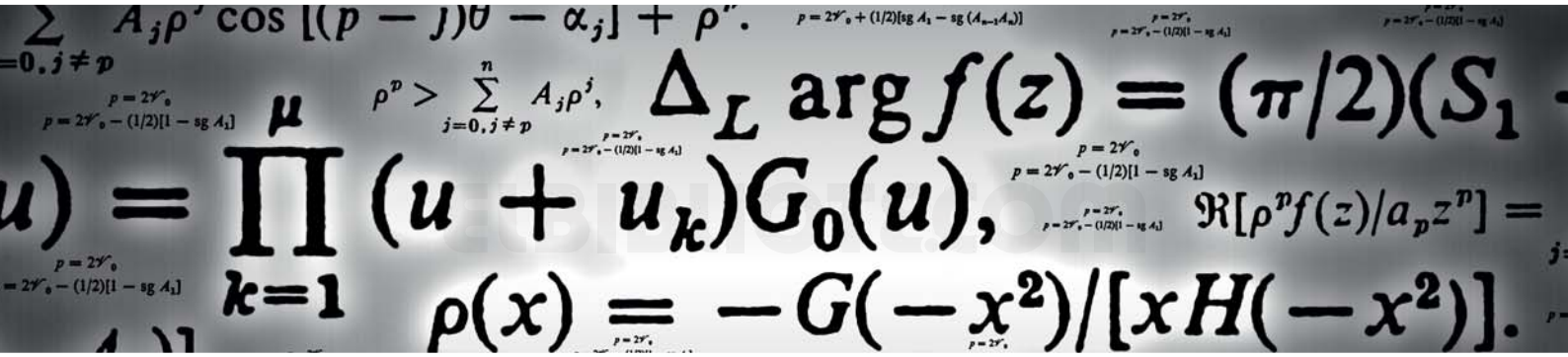
La formulación descriptiva es un sistema multivariante, en el que la forma de la elipse n-dimensional está definida por la matriz de varianza-covarianza calculada para las n variables. Mientras que la varianza es proporcional a la dispersión de puntos en la dirección paralela al eje de esa variable, por su parte, la covarianza define la forma de esa elipse.. En caso de que las variables no posean dimensiones pasibles de comparación, las varianzas tampoco podrán ser comparadas, razón por la cual, conviene recurrir a la matriz de correlación, debido a que el coeficiente de correlación es la covarianza de la mediad para valores estandarizados o normalizados, en consecuencia, la diagonal principal es todo unos. Por lo tanto, es posible emplear la matriz de varianza-covarianza, o la de correlación.

FORMULACIÓN MATEMÁTICA

Notación Y Ordenamiento De Los Datos:
Cuando se trata de una serie de datos, como por ejemplo una imagen, corresponde tomar uno u otro tipo de ordenación, siempre que el mismo sea consistente con la totalidad de las imágenes que participen de la transformación.
El primer paso es encontrar la matriz de correlación de las variables, la que se calcula de distintas formas, a saber:

Con los datos normalizados o normalización de datos: se trata de encontrar la matriz de varianza-covarianza para los datos normalizados.

Con los datos originales: la fórmula de cálculo del coeficiente de correlación lineal es aplicada entre dos variables, es decir, coeficiente de correlación de Pearson, el coeficiente de correlación entre las variables Xa y Xb se denota rab, donde Sxa y Sxb son las desviaciones típicas de las variables Xa y Xb respectivamente, y Sxab es la covarianza muestral. La matriz de correlación se forma mediante el ordenamiento de los diferentes coeficientes de correlación existentes en una matriz de filas y columnas. Con los datos normalizados o normalización de datos: se trata de encontrar la matriz de varianza-covarianza para los datos normalizados.



En primer lugar se deben calcular las estadísticas esenciales de cada variables Xa, su medida y desviación estándar. A partir de dichos datos, las variables se pueden estandarizar, es decir, cambiar ese conjunto de datos por otro, con media cero y desviación estándar uno. Entonces, se pasa de la variable Xa a la variable Za, y así sucesivamente. Partiendo de las variables estandarizadas Z2, Z2, Z3, etc., Zp, se calculan sus varianzas, las que obviamente dan como resultado uno; y las covarianzas entre las variable.

Estos datos se ordenan en forma de matriz, con filas y columnas que representan variables, entones, mediante la relación existente entre la matriz de varianza-covarianza y la matriz de correlación se obtiene la matriz de correlación.

Otra manera de calcular la matriz de correlación a través de las variables estandarizadas es ordenando en primer lugar las variables estandarizas como una matriz, por ejemplo en filas.
El segundo paso para la notación y ordenamiento de los datos cosiste en calcular los valores y vectores propios de la matriz de correlación que se ha calculado. Los valores propios son las raíces de un polinomio, para cada valor propio se obtiene una ecuación distinta, de la cual además se obtendrá un vector propio diferente y relacionado a su respectivo. Los vectores propios asociados a dichos valores propios, serán calculados mediante la sustitución de los valores propios.

Componentes principales:
Los coeficientes de la transformación que se debe efectuar para pasar de las variables originales a las nuevas variables o componentes principales son las coordenadas de los vectores propios que se han encontrado. Los valores propios indican el orden en que corresponde ubicar a esos vectores propios; el valor propio mayor indica que su vector propio asociado apunta en la dirección del primer componente principal; el segundo valor propio efectúa la misma acción con su vector propio, este indica que apunta en la dirección de máxima variabilidad ortogonal, y así continuamente.

Coordenadas:
El último paso es calcular las coordenadas de las variables originales en la nueva base de variables componentes principales. Tales coordenadas serán las que permitirán diferenciar unas variables de otras.

LA ESTADÍSTICA DE MEDICINA

Los gráficos estadísticos son aquellas imágenes en las que mediante la combinación del sombreado, los colores, los puntos, las líneas, los números, texto y coordenadas, es posible presentar datos numéricos. Los gráficos o diagramas sirven para dos cosas esenciales, en primer lugar son un sustituto de las tablas y en segundo lugar, son instrumentos fundamentales para el análisis de datos, llegando a ser el medio más efectivo describir, resumir y analizar la información. En definitiva, colaboran en la comprensión y comunicación de la evidencia arrojada por los datos en referencia a una hipótesis de estudio, entonces, un gráfico científico es útil a fin de representar la realidad, no para generar nuevas realidades existentes fuera de su propia imagen.

Un gráfico científico es útil a fin de representar la realidad, no para generar nuevas realidades existentes fuera de su propia imagen