FBioyF - UNR
Area Tecnología en Salud Pública.
Autor: Bioq. L. Eloísa Rodenas.
Año: 2006.

Tema: "Herramientas de Análisis: la estadística descriptiva".

⫽ *Introducción.*

La **Estadística** se utiliza como **tecnología al servicio** de las ciencias, donde la variabilidad y la incertidumbre forman parte de su naturaleza. Es una rama de la matemática y se utiliza para describir, analizar e interpretar ciertas características de un conjunto de individuos llamado población.

🗗 *Pasos en un estudio estadístico.*

1- Plantear hipótesis sobre una *población.*

Población es el conjunto sobre el que estamos interesados en obtener conclusiones, realizar inferencias.

2- Decidir qué datos recoger: diseño de experimentos.

- Qué individuos pertenecerán al estudio: ***muestras*.**

Muestra es un subconjunto de la población, al que tenemos acceso y sobre el que realmente hacemos las observaciones: mediciones.

- Qué datos recoger de los mismos: ***variables*.**

Una ***variable*** es una característica observable que varía entre los diferentes individuos de una población. Es una característica (magnitud, vector o número) que puede ser medida, adoptando diferentes valores en cada uno de los casos de un estudio.

3- Recoger los datos: muestreo.
4- Describir - resumir - los datos obtenidos.
5- Realizar una inferencia sobre la población.
6- Cuantificar la confianza en la inferencia.

- ***La estadística descriptiva***, se dedica a los métodos de recolección, descripción, visualización y resumen de datos originados a partir de los fenómenos en estudio.
- ***La estadística inferencial*** se dedica a la generación de los modelos, inferencias y predicciones asociadas a los fenómenos en cuestión.

Tipos de variables.

En un estudio científico, podemos clasificar las variables

- según la escala de medición.
- según la influencia que asignemos a unas variables sobre otras.

Según la escala de medición.

Variables cualitativas: son las variables que expresan distintas cualidades, características o modalidad. Cada modalidad que se presenta se denomina atributo o categoría y la medición consiste en una clasificación de dichos atributos. Sus valores no se pueden asociar naturalmente a un número: no se pueden hacer operaciones algebraicas con ellos. Las variables cualitativas pueden ser **dicotómicas** cuando sólo pueden tomar dos valores posibles como *sí y no*, *hombre y mujer* o son politómicas cuando pueden adquirir tres o más valores.

Las variables cualitativas pueden ser ordinales y nominales.

1. ***Variable cualitativa ordinal***: la variable puede tomar distintos valores ordenados siguiendo una escala establecida, aunque no es necesario que el intervalo entre mediciones sea uniforme, por ejemplo, *leve, moderado, grave,* grado de satisfacción, intensidad del dolor.

2. ***Variable cualitativa nominal***: en esta variable los valores no pueden ser sometidos a un criterio de orden, como por ejemplo los colores o el lugar de residencia, sexo, grupo sanguíneo, religión, nacionalidad, fumar (Sí/No).

Variables cuantitativas: son las variables que se expresan mediante cantidades numéricas y tiene sentido hacer operaciones algebraicas con ellos.

Las variables cuantitativas además pueden ser:

1. ***Variable discreta****:* es la variable que presenta separaciones o interrupciones en la escala de valores que puede tomar. Estas separaciones o interrupciones indican la ausencia de valores entre los distintos valores específicos que la variable pueda asumir. Toma valores enteros: número de hijos, número de cigarrillos.
2. ***Variable continua:*** es la variable que puede adquirir cualquier valor dentro de un intervalo especificado de valores. Por ejemplo el peso o la altura, que solamente limitado por la precisión del aparato medidor, en teoría permiten que siempre existe un valor entre dos cualesquiera.

⫽ La Estadística Descriptiva.

Se dedica única y exclusivamente al ordenamiento y tratamiento mecánico de la información para su presentación por medio de tablas y de representaciones gráficas, así como de la obtención de algunos parámetros útiles para la explicación de la información.
Este análisis es muy básico, pero fundamental en todo estudio y se lleva a cabo, calculando una serie de medidas de tendencia central, para ver en qué medida los datos se agrupan o dispersan en torno a un valor central.

Presentación ordenada de datos.

Las tablas de frecuencias y las representaciones gráficas son dos maneras equivalentes de presentar la información. Las dos exponen ordenadamente la información recogida en una muestra.

⊟ Tablas de frecuencia.

Exponen la información recogida en la muestra, de forma que no se pierda nada - o poca - de información.

- Frecuencias absolutas: contabilizan el número de individuos de cada modalidad
- Frecuencias relativas o porcentajes: Idem, pero dividido por el total.
- Frecuencias acumuladas: sólo tienen sentido para variables ordinales y numéricas. Muy útiles para calcular cuantiles

⊟ Representaciones gráficas para información univariada.

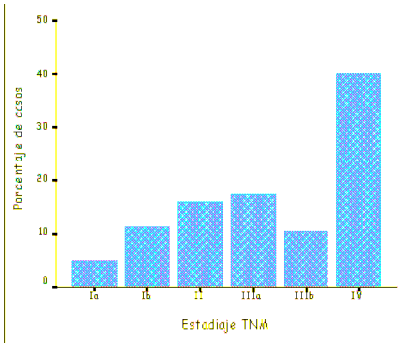
Gran parte de la utilidad que tiene la Estadística Descriptiva es la de proporcionar un medio para informar basado en los datos recopilados: **la transmisión eficiente de la información**. La eficacia con que se pueda realizar tal proceso de información dependerá de la presentación de los datos, siendo la forma gráfica uno de los más rápidos y eficientes.

Existen varios **tipos de gráficas**, o **representaciones gráficas**, utilizándose cada uno de ellos de acuerdo al tipo de información que se está usando y los objetivos que se persiguen al presentar la información.

Diagrama de rectángulos:

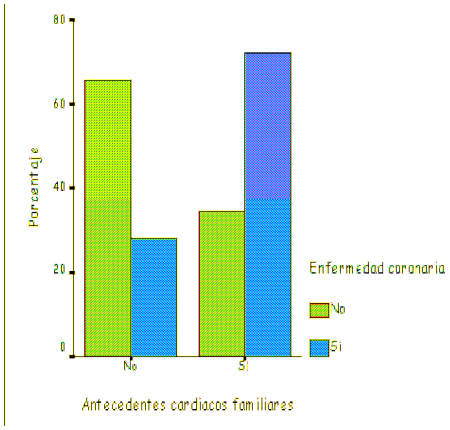
Esta representación gráfica se utiliza para **v*ariables cualitativas nominales*** y consiste en construir tantos rectángulos como modalidades presente el carácter cualitativo en estudio, todos ellos con base de igual amplitud. La altura se toma igual a la frecuencia absoluta o relativa - según la distribución de frecuencias que estemos representando - consiguiendo de esta manera rectángulos con áreas proporcionales a las frecuencias que se quieren representar. 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. Estos mismos gráficos pueden utilizarse también para describir ***variables numéricas discretas*** que toman pocos valores: número de hijos, número de recidivas,etc..

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

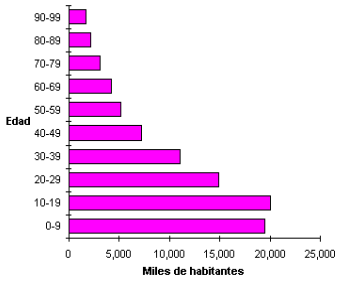


Se pueden representar en la misma gráfica, utilizando las mismas escalas horizontales y verticales, varios datos correspondientes a las mismas variables producto de varias observaciones. Esto produce una gráfica con varias **series**, correspondiendo cada una de ellas a cada observación de la muestra (o población), y teniéndose una gráfica compuesta. Es conveniente que cada serie de datos (u observaciones) sean ilustradas o iluminadas de igual manera entre sí, pero distinta de las demás.

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

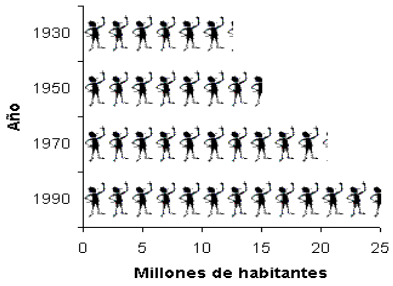


También es posible realizar **gráficas de barras horizontales**, los cuales se parecen mucho a las gráficas de columnas, con la salvedad importante de que la función de los ejes se intercambian y el eje horizontal queda destinado a las frecuencias y el eje vertical a las clases. Es muy común que este tipo de gráficos se utilicen para ilustrar el tamaño de una población dividida en estratos como, por ejemplo, son sus edades.



A este tipo de gráficos en particular se le llama **pirámide de edades** por su forma. Incluso, cuando se compara la población masculina y femenina por estratos de edades, se estila utiliza el lado izquierdo para la población de un sexo y el lado derecho para el otro, el resultado es una "pirámide" casi simétrica (dependerá de la población en particular).

Actualmente y mucho en los medios masivos de comunicación, se utilizan gráficos para ilustrar los datos o los resultados de alguna investigación. Regularmente se utilizan dibujos para representar dicha información, y el tamaño o el número de estos dibujos dentro de una gráfica queda determinado por la frecuencia correspondiente. A este tipo de gráfica se le llama **pictograma.**

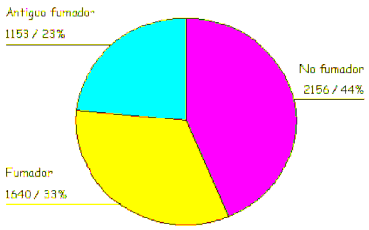


Representa la población de los Estados Unidos y cada hombrecillo representa a dos millones de habitantes.

 # Diagrama de sectores.

Los **diagramas de sectores** son similares a los gráficosde barras.
En los **gráficos de sectores,** 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.

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

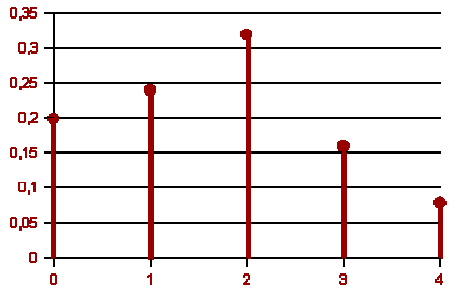


Como se puede observar, la información que se debe mostrar en cada sector hace referencia al número de casos dentro de cada categoría y al porcentaje del total que estos representan. Si el número de categorías es excesivamente grande, la imagen proporcionada por el gráfico de sectores no es lo suficientemente 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.

 # Diagrama de bastones.

Correspondientes a **variables cuantitativas discretas**. Consiste en levantar, para cada valor de la variable, una barra cuya altura sea su frecuencia absoluta o relativa, dependiendo de la distribución de frecuencias que estemos representando.

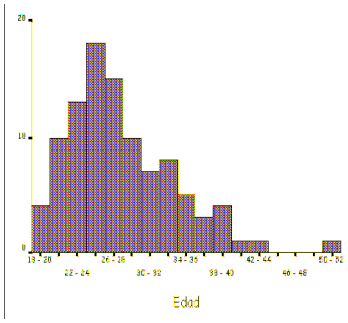
Ejemplo de distribución de frecuencias del nº de hijos



 # El histograma.

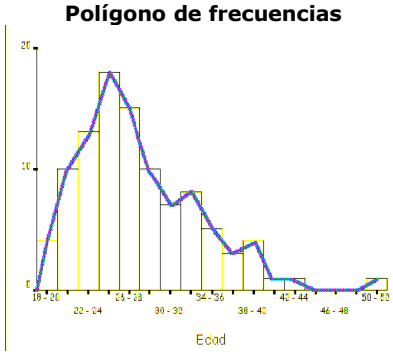
Para ***variables numéricas continuas***, tales como la edad, la tensión arterial o el índice de masa corporal, el tipo de gráfico más utilizado es el **histograma**. Para construir un gráfico de este tipo, se divide el rango de valores de la variable en intervalos de igual amplitud, representando sobre cada intervalo un rectángulo que tiene a este segmento como base.

Distribución de frecuencias de la edad en 100 pacientes.



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 siguiente.

Polígono de frecuencias



#Diagramas de cajas.

Otro modo habitual, y muy útil, de resumir una variable de tipo numérico es utilizando el concepto de percentiles, mediante el **diagramas de cajas.**

En 1977 John Tukey publicó un tipo de gráfico estadístico para resumir información utilizando 5 medidas estadísticas: el valor mínimo, el primer cuartil, la mediana, el tercer cuartil y el valor máximo. Este tipo de gráfico recibe el nombre de *gráfico de caja* (boxplot).

Medidas de posición: valor mínimo y máximo de la variable y la mediana.

- **Mediana (estadística).**

Si tenemos n valores $x_1, x_2, x_3, \ldots, x_n$ habiendo sido ordenados de forma creciente:

Se define ***la mediana como el valor que deja a cada lado (por encima y por debajo) la mitad de los valores de la muestra.***

Valor mediana = (n +1) / 2.

Cuartiles: dividen a la muestra en 4 grupos con frecuencias similares.

- **Cuartil 1 = divide el 25 % de los datos menores del 75 % de los mayores.**

Primer cuartil = Percentil 25 = Cuantil 0,25

- **Cuartil 2 = Mediana**

Segundo cuartil = Percentil 50 = Cuantil 0,5 = mediana

- **Cuartil 3 = divide el 75 % de los mayores del 25 % de los datos menores.**

Tercer cuartil = Percentil 75 = cuantil 0,75

Medidas de dispersión: rango y Rango intercuartílico.

- **Rango = V máximo – Valor mínimo.**
- **Rango intercuartílico (RIC) = C3 – C1.**

Un gráfico de este tipo consiste en un rectángulo – ***caja*** - donde los lados más largos muestran el recorrido intercuartílico (RIC). Este rectángulo está dividido por un segmento vertical que indica donde se posiciona la mediana y por lo tanto su relación

con los cuartiles primero y tercero (recordemos que el segundo cuartil coincide con la mediana).

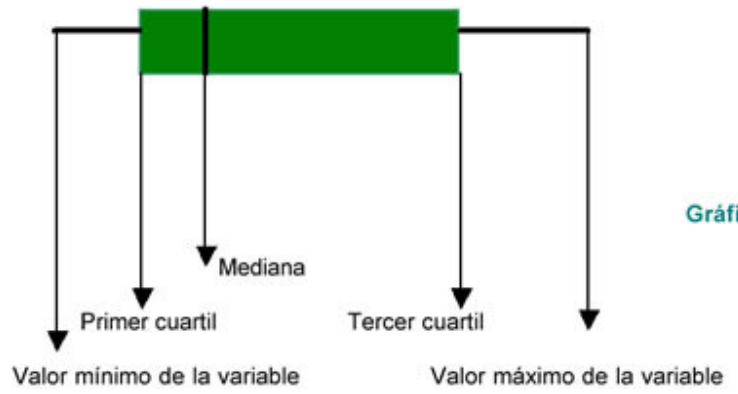


Gráfico 1

A diferencia de otros métodos de presentación de datos, los gráficos de caja muestran los valores atípicos de la variable.

Llamaremos **valores atípicos de la variable** a aquellos que están tan apartados del cuerpo principal de los datos que bien pueden representar los efectos de causas extrañas, como algún error de medición o registro. Su eliminación no se justifica, ya que el propósito del gráfico de caja consiste en brindarnos un mayor conocimiento de la forma en que se distribuyen los datos.

Tukey introduce un criterio para fijar los extremos de los bigotes. Para esto calcula 4 barreras, dos interiores y dos exteriores:

Barrera interior inferior = Primer cuartil – 1,5 . RIC

Barrera interior superior = Tercer cuartil + 1,5 . RIC

Barrera exterior inferior = Primer cuartil – 3 . RIC

Barrera exterior superior = Tercer cuartil + 3 . RIC

Si se consideran **los valores de la variable comprendidos entre las dos barreras interiores,** el valor mínimo de la variable y el valor máximo son los extremos de los bigotes.

Si existen valores de la variable comprendidos entre las barreras interiores y exteriores se consideran valores atípicos y se indican con *. Si existieren valores fuera de las barreras exteriores se consideran valores todavía más atípicos.

Este tipo de gráfico nos proporciona información con respecto a la simetría o asimetría de la distribución. Se utilizan los siguientes criterios: si la mediana está en el centro de la caja o cerca de él, constituye un indicio de simetría de los datos, si la mediana está considerablemente más cerca del primer cuartil indica que los datos son positivamente asimétricos y si está más cerca del tercer cuartil, señala que los datos son negativamente asimétricos. Asimismo, la longitud relativa de los bigotes se puede emplear como un indicio de su asimetría.

En resumen:

- la caja central indica el rango en el que se concentra el 50% central de los datos.

- sus extremos son el 1er y 3er cuartil de la distribución.
- la línea central en la caja es la mediana. 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***.

Bibliografía.

Apuntes y vídeos de Bioestadística. Disponible en URL:http://campusvirtual.uma.es/est_fisio/apuntes/#1ciclo. Consultado 18/09/06.

Estadística Descriptiva. Disponible en URL: http://www.uaq.mx/matematicas/estadisticas/xu3.html. Consultado 18/09/06.

Apuntes de clase. Estadística. Disponible en URL: http://www.liccom.edu.uy/bedelia/cursos/metodos/materiales.html. Consultado 18/09/06.

Estadística descriptiva con Minitab. Disponible en URL: http://www.uoc.edu/in3/e-math/docs/Estad_Descriptiva.pdf#search=%22estad%C3%ADstica%20descriptiva%2Brepresentaciones%20gr%C3%A1ficas%22. Consultado 18/09/06.

Los Gráficos de Caja: Un Recurso Innovador. Disponible en URL: http://www.rieoei.org/experiencias93.htm. Consultado 18/09/06.

Representación gráfica en el Análisis de Datos. http://www.fisterra.com/mbe/investiga/graficos/graficos.htm. Consultado 18/09/06.