Fundación J. L. Castaño
Para el desarrollo del Laboratorio clínico

ESTADÍSTICA BÁSICA APLICADA AL LABORATORIO CLÍNICO

Ed Cont Lab Clín; 30: 1 - 8

SEQC

2016-2017

ESTADÍSTICA DESCRIPTIVA.

Inmaculada Pérez de Algaba Fuentes.
UGC Intercentros de Laboratorio. Hospital Regional Universitario de Málaga.

Ana Peña Cabia.
Servicio de Análisis Clínicos. Hospital Virgen de la Luz. Cuenca.

INTRODUCCIÓN

La Estadística es la parte de las Matemáticas que se encarga del estudio de una determinada característica en una población, recogiendo los datos, organizándolos en tablas, representándolos gráficamente y analizándolos para sacar conclusiones de dicha población.

Podemos diferenciar dos situaciones:

- Estadística descriptiva. Realiza el estudio sobre la población completa, o una muestra de ella, limitándose a observar una o varias características y calculando unos parámetros que den información global de toda la población o de la muestra estudiada.
- Estadística inferencial. Realiza el estudio descriptivo sobre un subconjunto de la población llamado muestra y, posteriormente, extiende los resultados obtenidos a dicha población.

A continuación, se describen dos ejemplos para ilustrar estos conceptos:

Ejemplo 1. Cuando van a llegar cualquier tipo de elecciones, como por ejemplo, las generales, es muy frecuente que los medios de comunicación nos adelanten los resultados de encuestas o sondeos en los que se nos indica el resultado final de dichas elecciones con una precisión y con un error determinados (a lo que en medios de comunicación le denominan horquillas). Estos sondeos son realizados por distintas técnicas sobre un grupo (muestra) más o menos numeroso de personas. Es muy difícil, por no decir imposible, entrevistar a todos los españoles con derecho a voto, por tanto, se recurre a entrevistar lo que se conoce como una muestra, cuanto mayor sea esta mayor será la fiabilidad de la encuesta, pero también mayor será el coste del sondeo. El estudio de esta muestra se haría mediante estadística descriptiva, pero lo que nos interesa no es el resultado de este estudio reducido sino el resultado final de las elecciones. El paso de generalizar los resultados de la muestra a toda la población se hace mediante técnicas de Estadística inferencial. La elección de la muestra debe hacerse mediante métodos de muestreo para que el estudio resulte lo más fiable posible.

Ejemplo 2. Supongamos que estamos en un instituto con un número muy elevado de alumnos y alumnas, por ejemplo 500, y queremos hacer un estudio estadístico sobre su altura.

Un método sería pasar clase por clase y medirlos a todos; esto nos podría llevar un tiempo considerable, pero sería la forma más exacta de hacer dicho estudio, aunque es fácil encontrarnos con ausencias y tendríamos que volver varios días y pasar lista para conseguir la estatura de todo el alumnado. Una vez que tengamos todos los datos en nuestro poder los resultados los obtendríamos mediante Estadística descriptiva.

Otra posibilidad podría ser pasar clase por clase, decirle a los alumnos y alumnas que anoten su estatura en un papel y recogerlos todos. También así tendríamos un estudio de Estadística descriptiva, aunque seguramente menos fiable que con el método anterior, pues casi con toda seguridad, es de esperar que algunos alumnos escriban su estatura a cálculo y otros, con ganas de bromas, muy por encima o muy por debajo de la realidad.

Y otra posibilidad sería escoger una muestra, es decir un grupo de por ejemplo 50 personas, hacer el estudio descriptivo sobre ellas y después generalizarlo a todo el instituto con Estadística inferencial. En este caso, comprobaríamos por una parte que cuanto mayor sea la muestra más trabajo tendremos, pero más fiable será el resultado final y por otra, que la elección de la muestra debe hacerse de manera que permita también fiarnos del resultado obtenido.

Conceptos básicos.

Ya hemos hablado de ellos en los ejemplos anteriores, en cualquier estudio estadístico pueden aparecer los conceptos:

- **Individuo**, cada uno de los elementos, personas u objetos que se van a valorar y que contiene cierta información acerca del fenómeno que se desea estudiar.
- **Población**, que es el conjunto formado por todos los elementos a los que les vamos a hacer el estudio.
- **Muestra**, el subconjunto de la población que elegimos para hacer un estudio más reducido. El número de individuos que forman la muestra se le llama tamaño muestral.
- **Parámetro,** son los valores que resumen una determinada información referente a una población.
- **Estadístico,** son los valores que expresan una determinada información referente a una muestra.

Variables

El método científico está basado en la observación de un hecho determinado, de un evento o de una circunstancia, los registros de estas observaciones son las variables del estudio.

Una variable es una característica observable que se desea estudiar en una muestra de individuos, pudiendo tomar diferentes valores. Nos podemos encontrar con diferentes tipos de variables:

- **Cualitativas** o nominales (atributos): son aquellas que no pueden medirse numéri-

camente. Los diferentes valores que pueden tomar se conocen como modalidades o categorías. A su vez, pueden ser:

- o Binarias o dicotómicas: cuando existen únicamente dos categorías. Es el caso del género o sexo (masculino y femenino, hombre y mujer)
- o No dicotómicas: cuando existen más de dos categorías para definir a la variable. Por ejemplo, cuando hablamos de raza o profesión.

- **Ordinales:** son aquellas variables que van a ser medidas utilizando una escala ordinal. Tenemos como ejemplos las jerarquías de autoridad, la intensidad del dolor o el estadio de un tumor.
- **Cuantitativas:** son aquellas que se pueden medir numéricamente, es decir, que se pueden cuantificar. Toman valores con significado matemático, y se acompañan de unidades de medida. Estas, a su vez, pueden ser:
  - o Discretas: son aquellas en las cuales entre dos valores consecutivos no se encuentra ningún otro valor (ejemplos: número de hijos, número de camas de un hospital). Estas variables van a tomar valores que podemos suponer que son siempre enteros.
  - o Continuas o proporcionales: son aquellas en las que, teóricamente, entre dos valores consecutivos pueden encontrarse infinitos valores (ejemplos: glucemia, la edad, el peso…).

Una variable predictora o independiente es aquella que medimos para un estudio, mientras que la variable resultado o dependiente, es aquella cuyo valor depende de los resultados de la variable predictora.

Estadística descriptiva

La descripción de las variables suele ser un resumen representativo de estas, lo que es especialmente importante cuando los registros de la observación son muy grandes.

La descripción de las variables de un estudio la haremos según el tipo de variable que tratemos. Normalmente las variables cualitativas se describirán calculando la frecuencia con que se presenta cada categoría de la variable. Esta frecuencia puede ser absoluta, relativa o acumulativa, estos datos pueden describirse mediante tablas o gráficos. Mientras que las variables cuantitativas suelen describirse mediante lo que se conoce como medidas de tendencia central (media, mediana) o medidas de dispersión (desviación estándar, varianza, coeficiente de variación), estos datos pueden mostrarse también en tablas o mediante gráficos.

Descripción de variables cuantitativas:

Medidas de tendencia central

La **moda** es el valor con una mayor frecuencia en una distribución de datos.

La **media aritmética** es el valor promedio de los datos de una distribución, o lo que es lo mismo, la suma de los valores divido por el número de observaciones. En una distribución

normal, la media aritmética y la mediana son iguales, la media es un estadístico en el que influyen mucho la homogeneidad de los datos y sobre todo los valores aberrantes.

$$\bar{x} = \frac{\sum x_i}{n}$$

La **mediana** es el valor central de una distribución de valores ordenada y coincide con el percentil 50 de la distribución.

Medidas de posición

Los **percentiles** nos dan una idea del número de observaciones que hay por debajo o por encima del percentil, así el percentil 25, representa el número de valores de una distribución que se encuentran por debajo del 25 % de los valores, el percentil 50 representa el número de datos que hay por debajo del 50 % de los datos de la distribución, que coincide con la mediana, el percentil 0 sería el valor mínimo de la distribución mientras que el percentil 100 representaría el valor máximo de la distribución. Se conoce como rango interuartílico (IQR), el número de datos comprendidos entre el percentil 25 y el 75, es decir el 50 % central.

Los cuartiles dividen la distribución en cuatro partes y se corresponde con los percentiles 25 %, 50 % y 75 %. Se conoce como deciles a los valores que dividen la distribución en diez partes, correspondientes a los percentiles 10 %, 20 %, 30 %…, es decir el primer decil se corresponde con el percentil 10 %, el sexto decil con el percentil 60 % y así sucesivamente. Por último los quintiles dividen la distribución en 5 partes, lo que corresponde a los percentiles 20 %, 40 %, 60 % y 80 %. Los percentiles que ya hemos descrito dividen la distribución en 100 partes.

Medidas de dispersión

Las más habituales son la varianza, la desviación estándar (DS) y el coeficiente de variación (CV %)

La **varianza** es el cociente entre el sumatorio de la diferencia al cuadrado de los valores medidos menos la media, entre el numero de datos.

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x - \bar{x})^2$$

La **DS** es la raíz cuadrada de la varianza

$$\sigma = \sqrt{\sigma_n^2}$$

Coeficiente de variación es el cociente entre la DS y la media y multiplicado por 100.

$$CV\% = \frac{\sigma}{\bar{x}} \times 100$$

Descripción de variables cualitativas

Las variables cualitativas se suelen agrupar mediante frecuencia, o lo que es lo mismo, número de individuos de una muestra que comparten la misma característica. Esta frecuencia puede ser absoluta, relativa (% de individuos que presentan una característica) o acumulativa.

Un ejemplo sería la siguiente tabla

G Tumor	Frecuencia	Frecuencia relativa	Frecuencia acumulativa
IV	25	24,8%	24,8%
III	40	39,6%	64,4%
II	26	25,7%	90,1%
I	10	9,9%	100,0%

Tablas

Frecuentemente las variables se suelen describir en los trabajos científicos mediante tablas o gráficos. Tanto tablas como gráficos siempre deben ser autoexplicativos, o lo que es lo mismo, que no necesitan de una explicación añadida en el texto del documento. Por tanto las tablas o los gráficos deben contener la mayor información posible de forma clara y resumida. En la tabla anterior sabemos que 10 individuos tienen un tumor en grado 1, que representa el 9,9 de la muestra, y que el 100 % de la muestra tiene un tumor con grado 1 o más.

En el caso de las variables cuantitativas en las tablas se suelen describir los datos estadísticos de tendencia central y de dispersión. Normalmente se detalla la media y la DS, el IQR, los percentiles 0 (mínimo), 25, 50 (mediana), 75 y 100 (máximo), así como el número de datos, y a veces el número de datos faltante.

mean	Sd	IQR	0%	25%	50%	75%	100%	n
51	13,87	19	18	42	51	61	81	113

Esta tabla podemos ver que corresponde a la edad de los participantes en un estudio. Sabemos que la media de los participantes es de 51 años con una edad mínima de 18 y máxima de 81, la mediana coincide con la media, la DS es de 13,7 años, el 50 % están entre 42 y 61 años, y han participado 113 individuos en el estudio.

Las tablas pueden hacerse con más de una variable, por ejemplo en el caso anterior podríamos hacer una tabla con dos líneas, la primera sería el descriptivo de los participantes que presentan la condición de estudio y los participantes del grupo control. Es muy típica la tabla de los estudios de seguimiento en la que se estudia una condición en una población. En este caso se suele hacer una tabla para describir las muestras que presentan la condición de estudio y de las que no la presentan, la idea de esta tabla es mostrar qué casos y controles

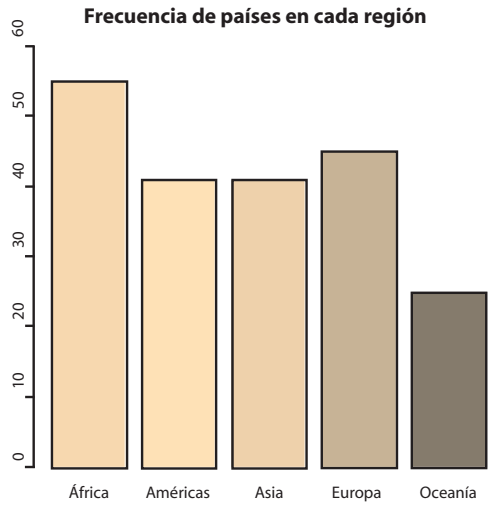
son iguales en todo excepto en la condición de estudio, de forma que cualquier resultado obtenido será debido a la condición estudiada y no a otra variable que pueda confundir.

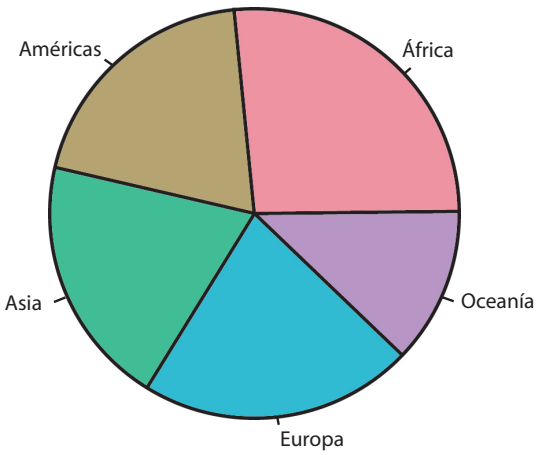
Gráficos

Otra forma de describir las variables es mediante gráficos, que como ya hemos comentado deben ser autoexplicativos, por lo que es importante describir los elementos del gráfico con leyendas adecuadas, los títulos de los ejes adecuados y una pequeña explicación si es necesaria a pie de gráfico.

Al igual que en el caso de las tablas, las variables cualitativas y cuantitativas requieren un tipo de gráfico característico.

En el caso de variables cualitativas los gráficos más habituales son el gráfico de barras y el de sectores, en el que suelen representarse las frecuencias de las variables cualitativas, si el número de categorías de una variable es muy grande, no se debe emplear el gráfico de sectores ya que su interpretación se vuelve muy confusa. A continuación, se muestra un gráfico de barras y uno de sectores:





En el caso de las variables cuantitativas se suelen describir mediante el histograma o los gráficos de cajas.

El histograma es el gráfico estadístico por excelencia. El histograma de un conjunto de datos es un gráfico de barras que representan las frecuencias con que aparecen las mediciones agrupadas en ciertos rangos o intervalos. La idea de agrupar datos en forma de histogramas se conoce desde 1662 con el trabajo de Graunt.

Realizar histogramas de esta manera tiene las siguientes ventajas:

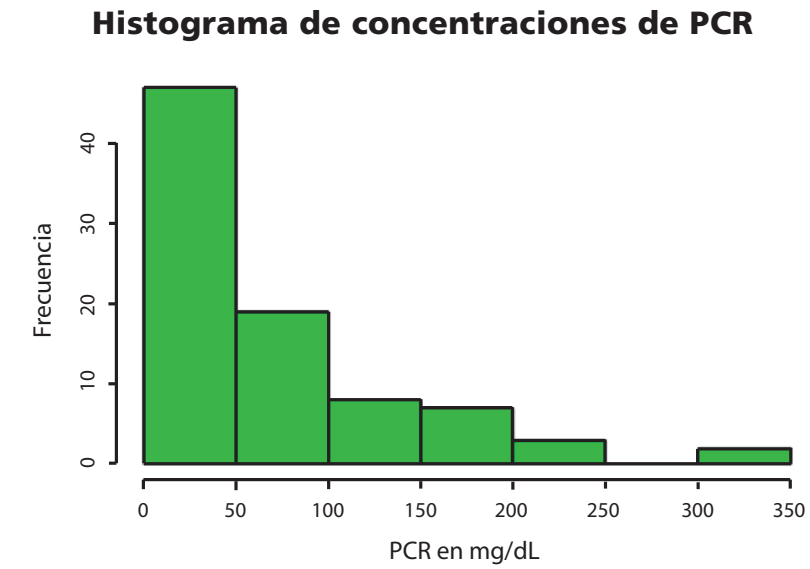
1. Es útil para apreciar la forma de la distribución de los datos, si se escoge adecuadamente el número de clases y su amplitud.
2. Se puede presentar como un gráfico definitivo en un informe.

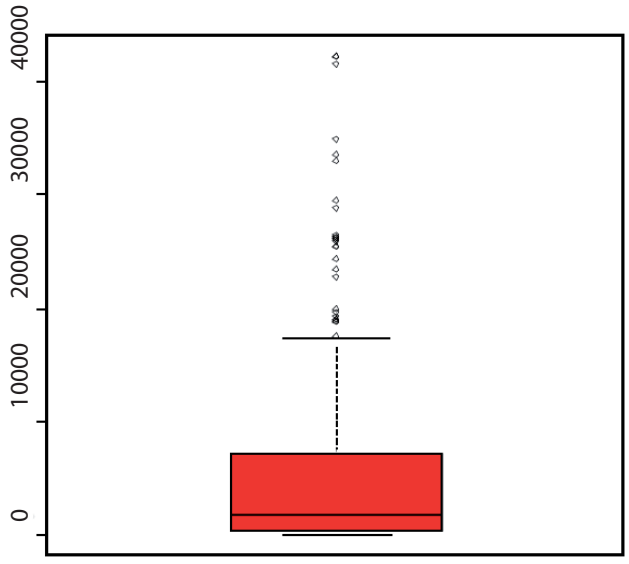
3. Se puede utilizar para comparar dos o más muestras o poblaciones.
4. Se puede refinar para crear gráficos más especializados, por ejemplo, la pirámide poblacional.

Y las siguientes desventajas:

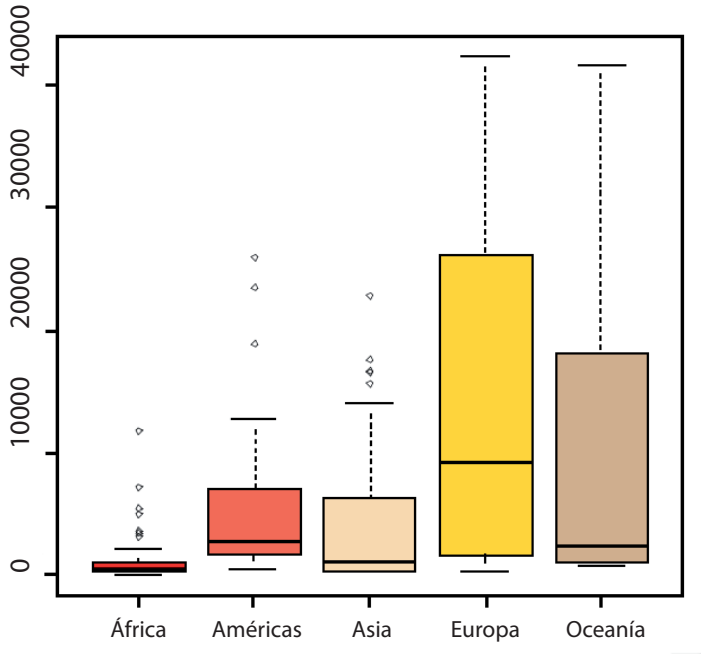
1. Las observaciones individuales se pierden.
2. La selección del número de clases y su amplitud que adecuadamente representen la distribución puede ser complicado. Un histograma con muy pocas clases agrupa demasiadas observaciones y uno con muchas deja muy pocas en cada clase. Ninguno de los dos extremos es adecuado.

Un ejemplo de histograma sería el siguiente:





El otro tipo de gráfico que podíamos construir para variables numéricas ya hemos dicho que es el gráfico de cajas. Este ha sido un aporte fundamental realizado por Tukey (1977). Es un gráfico simple, pero poderoso. Se observa de una forma clara la distribución de los datos y sus principales características. Permite comparar diversos conjuntos de datos simultáneamente. Como herramienta visual se puede utilizar para ilustrar los datos, para estudiar simetría, para estudiar las colas, y supuestos sobre la distribución, también se puede usar para comparar diferentes poblaciones.



Este gráfico contiene un rectángulo, usualmente orientado con el sistema de coordenadas tal que el eje vertical tiene la misma escala del conjunto de datos. La parte superior y la inferior del rectángulo coinciden con el tercer cuartil y el primer cuartil de los datos. Esta caja se divide con una línea horizontal a nivel de la mediana. Se define un "paso" como 1,5 veces el rango intercuartil, y una línea vertical (un bigote) se extiende desde la mitad de la parte superior de la caja hasta la mayor observación de los datos si se encuentran dentro de un paso. Igual se hace en la parte inferior de la caja. Las observaciones que caigan más allá de estas líneas son dibujadas individualmente. La definición de los cuartiles puede variar y otras definiciones del paso son planteadas por otros autores (Frigge et al., 1989).

Existen otros tipos de gráficos que iremos estudiando a medida que avance el curso, será el caso de curvas ROC, curvas de supervivencia y curvas de regresión.

ISSN 1887-6463 – Octubre 2016 (recibido para publicación Mayo 2016).