UNIVERSIDAD
NACIONAL
DE COLOMBIA
Revista de la Facultad de Medicina 1999; 47 (4):232-236

Conceptos básicos de Bioestadística

Constanza Quintero Guzmán, Profesora Asociada, Facultad de Ciencias, Departamento de Matemáticas y Estadística. Aura Nidia Herrera Rojas, Profesora Asistente, Facultad de Ciencias Humanas, Departamento de Psicología. Ricardo Sánchez Pedraza, MD. Profesor Asociado, Centro de Epidemiología Clínica, Facultad de Medicina, Universidad Nacional de Colombia. INCLEN.

SUMMARY

This article is primarily aimed to give an understanding of some basic principles in Biostatistics. Instead of theoretical aspects, the emphasis is firmly on basic and practical applications related with probability, odds ratios and descriptive statistics.

RESUMEN

Este artículo tiene como objetivo aportar elementos para la comprensión de algunos principios básicos de la Bioestadística. En lugar de mostrar los aspectos teóricos, se hace énfasis fundamentalmente en aplicaciones básicas y prácticas en el campo de la probabilidad, los riesgos relativos indirectos y la estadística descriptiva.

INTRODUCCIÓN

En nuestra vida cotidiana permanentemente utilizamos la Estadística. Cuando vemos en la calle una persona con comportamientos no habituales nos causa extrañeza y la catalogamos como rara o fuera de lo normal. En este caso hemos seguido un proceso dentro del cual se han dado cuatro pasos:

1. *Descripción del comportamiento:* "Esa persona se viste descuidadamente, grita, dice cosas raras, es agresiva sin razón aparente".
2. *Generación de una hipótesis:* "Esa persona no es normal".
3. *Prueba de la hipótesis:* "No se comporta como los demás, los que se comportan así son los enfermos mentales, si lo comparo con un enfermo mental esta persona se porta igual".
4. *Conclusión:* "Esta persona es un enfermo mental, definitivamente no es normal".

El anterior proceso se realiza en muchas otras circunstancias: cuando vemos a alguien muy alto, o a una persona que tiene un acento extraño, o un modelo de automóvil nuevo...

Cuando tenemos que analizar una serie de datos de un estudio clínico utilizamos los mismos pasos: Primero los describimos, luego generamos una hipótesis que sometemos a un proceso de prueba y finalmente llegamos a una conclusión.

Detrás de este proceso está el hecho de que no somos iguales, es decir, hay *variabilidad.* En los casos en los que no hay variabilidad no se necesita la Estadística: No nos causa extrañeza el color de un taxi porque prácticamente todos son iguales, no nos llama la atención que un tigre tenga rayas pues casi todos las tienen.

La variabilidad es un fenómeno tan importante que nos permite poder conocer gente nueva, aprender cosas que no sabíamos, extrañarnos, asombrarnos, alegrarnos y entristecernos. Incluso, algunos autores de libros de Estadística la ven como explicación de la infidelidad (1) .

El fenómeno de la variabilidad también es el responsable de que nunca podamos conocer las características de ciertas *poblaciones* por ser estas muy grandes o difíciles de medir. En este caso debemos recurrir al estudio de solo una parte de esa población, lo cual se conoce como *muestra.* El asunto aquí es buscar una muestra que refleje de la mejor manera posible las características de la población de la cual se ha tomado. Obviamente, entre más grande sea la muestra, más se parecerá a la población que pretende representar. En este sentido, una muestra de buena calidad es una *muestra representativa.*

El proceso de describir, generar hipótesis y probarlas y finalmente sacar una conclusión, generalmente se efectúa sobre muestras. El siguiente paso es, mediante un proceso de inferencia, hacer extensivos los hallazgos y conclusiones de la muestra a la población de la cual proviene: esto se ha denominado *inferencia estadística.*

De manera general, las inferencias se hacen mediante dos estrategias (2):

1. Inferencia deductiva: O método hipotético-deductivo. Creamos una teoría a partir de la cual predecimos resultados.
2. Inferencia inductiva: Hacemos varias observaciones, dilucidamos un patrón y proponemos una teoría.

Para Popper la inducción surgida de la acumulación de evidencia no refuerza una teoría. La inducción se basa en nuestra creencia de que lo que aun no hemos observado es igual a lo que ya observamos. Esto quiere decir que no podemos probar una hipótesis con muchas evidencias a favor pero podemos refutarla con una única observación. En este sentido el proceso lógico sería plantear una teoría o conjetura y tratar de demostrar que es falsa. Entre más resista a ser destruida será más fuerte. Las teorías científicas útiles son potencialmente falsificables (3).
Los epidemiólogos hacen inferencias inductivas para generalizar a partir de una serie de observaciones, generan hipótesis y luego usan inferencias deductivas para probar tales hipótesis. Dentro de este proceso tiene un papel fundamental la probabilidad, aspecto sobre el que nos referiremos a continuación.

Probabilidad y odds:

La *probabilidad* de que ocurra un fenómeno se cuantifica entre 0 (seguro no ocurre) y 1 (seguro ocurre). Si la probabilidad de un evento es P y hay N ensayos u oportunidades de que el evento ocurra podemos esperar que el evento ocurra NxP veces *(frecuencia esperada).*
Dos eventos son *mutuamente excluyentes* cuando, si uno ocurre el otro no ocurre: La probabilidad de que uno u otro ocurra es la suma de las probabilidades individuales: En un dado la probabilidad de que salga 3 ó 5 es 1/6 + 1/6 = 2/6. La notación de esta propiedad es:

P(A o B) = P(A) + P(B)

Si dos eventos son independientes, el que ocurra uno de ellos no tiene efecto sobre la ocurrencia del otro. La probabilidad de que ocurran juntos es el producto de las probabilidades individuales. En un dado la probabilidad de que salga un número par menor de 5 es 1/2 x 2/3 = 2/6. La notación de esta propiedad es:

P(A y B) = P(A) x P(B)

Si dos eventos son independientes, la probabilidad conjunta es el producto de las probabilidades individuales. Esta es la base del test de Ji cuadrado.
La probabilidad de que B ocurra dado que A ocurrió se llama probabilidad condicional y se denota:

P(A|B) = P(A y B)/P(B)

Un tipo particular de probabilidad condicional es la Probabilidad Bayesiana. Por ejemplo, si tomamos 100 esquizofrénicos y evaluamos la frecuencia de delirios persecutorios en ese grupo encontrando que 70 los tienen, conocemos la probabilidad de tener delirios persecutorios dado que se tiene esquizofrenia : P(D|E)=0.7
Sin embargo puede ser más importante saber cuál es la probabilidad de tener esquizofrenia dado que se tienen delirios persecutorios [P(E|D)]. Esto se calcula mediante la siguiente fórmula:

$$P(E \mid D) = \frac{P(D \mid E)P(E)}{P(D \mid E)P(E) + P(D \mid \hat{E})P(\hat{E})}$$

El chance (odds) es un concepto familiar en los juegos de azar (4). Cuando se dice que las apuestas están 4 a 1 a favor del boxeador X estamos diciendo que cuatro personas apuestan a que X gana y 1 a que pierde. El odds puede definirse como el cociente entre el número de maneras consideradas favorables o de interés (apostadores a favor del boxeador X) sobre el número de maneras consideradas desfavorables o de no interés (apostadores en contra del boxeador X). Por ejemplo, si tenemos un grupo de 100 personas en las cuales se desea determinar la frecuencia de demencia y se encuentra que 15 de ellos tienen la enfermedad se puede decir que en ese grupo la probabilidad del desenlace es del 15%. También se puede decir que la situación de interés (presencia de demencia) está en 15 personas y la situación de no interés (ausencia de demencia) está en el resto de personas (85). El odds será entonces 15/85. Si interesa analizar un antecedente de dependencia al alcohol como posible factor de riesgo en los pacientes de demencia pueden compararse 2 odds: el odds de demencia en pacientes con dependencia al alcohol (12/30) versus el de pacientes sin dependencia al alcohol (3/55).

	Con demencia	Sin demencia
Con antecedente de alcoholismo	12	30
Sin antecedente de alcoholismo	3	55

Si se divide el odds de demencia en pacientes con dependencia al alcohol sobre el de pacientes sin dependencia al alcohol se tiene una medida de riesgo que dice cuántas veces es mayor el riesgo de tener demencia si se tiene el antecedente de alcoholismo. Esta medida es el OR (5,6), (traducido de diferentes maneras: riesgo relativo indirecto, razón de suertes, razón de chances...). El OR (Odds Ratio) es un cociente entre 2 odss. En el ejemplo anterior tendríamos (12/30)/(3/55)=7.33, lo que quiere decir que el chance de tener demencia es 7.33 veces mayor en los que tienen antecedente de alcoholismo.

Estadística descriptiva:

Para ilustrar los procedimientos básicos de la Estadística descriptiva se tomará un ejemplo consistente en una serie de datos de una muestra de 30 niños, correspondiente a un estudio para averiguar los factores de riesgo de bajo peso al nacer, definido éste como peso menor de 2500g.

Variables a considerar:

1. IP: Indicador de peso al nacer (0=peso al nacer>2500g, 1=peso al nacer<2500g)

2. Edad: Edad de la madre en años

3. Pesoma: Peso de la madre en libras

4. Raza: Raza de la madre (1=blanca, 2=negra, 3=otra)

5. HT: Historia de hipertensión (1=si,0=no)

6. HF: Hábito de fumar (1=leve, 2=moderado, 3= grave)

7 P: Paridad

8. NS: Nivel socioeconómico (1= bajo, 2=medio)

Los datos se presentan en formato de texto separado por tabulador, donde cada columna corresponde, en el mismo orden, a cada una de las variables mencionadas antes.

1	15	115	3	1	2	1	2
0	18	100	1	1	1	2	2
1	17	100	2	1	3	1	2
0	16	112	1	1	1	2	2
0	24	103	2	0	1	2	2
1	23	110	3	0	1	3	2
1	24	112	3	0	1	2	1
0	22	90	1	0	2	1	1
0	21	127	2	1	1	2	2
1	31	118	2	0	1	1	2
1	20	109	3	0	3	2	2
0	25	118	1	1	3	2	2
1	24	138	1	0	3	2	2
1	25	85	3	0	1	2	1
0	22	115	2	0	1	1	1
1	20	110	1	0	3	3	1
1	17	113	0	1	1	1	1
0	32	121	3	0	2	3	1
1	32	105	1	1	1	2	2
0	31	130	3	0	2	2	2
1	28	120	1	0	3	2	2
0	29	135	1	0	3	1	2
1	26	154	3	0	1	3	2
0	34	170	1	0	1	1	2
0	25	140	1	0	3	2	1
1	34	130	2	0	1	2	2
1	23	124	2	0	1	2	1
0	35	132	1	1	1	3	2
0	25	140	1	0	3	3	2
1	26	154	3	0	1	1	2

El anterior formato de presentación de datos tiene la ventaja de que puede ser leído fácilmente por cualquier programa estadístico o por cualquier hoja electrónica.

Es muy importante considerar que hay diferencia entre datos cualitativos y datos cuantitativos (7). Los cualitativos son aquellos que no son caracterizados por valores numéricos, y en general describen la cualidad de una persona o cosa; en este estudio son: Indicador de peso al nacer, raza, historia de hipertensión y hábito de fumar. La asignación de un valor numérico a este tipo de datos es artificial y solo se usa para permitir su procesamiento en los procedimientos o programas estadísticos. Si las categorías siguen algún orden se habla de variables ordinales. En estos casos sacar un promedio no tiene sentido.

Los datos cuantitativos son aquellos que poseen valor numérico. De manera general puede decirse que son aquellos en los cuales sacar un promedio tiene sentido.

Es importante describir y resumir la información obtenida sin pérdida de sus características esenciales. Un método es elaborar tablas de frecuencia: En el caso de variables nominales y numéricas discretas se reparten los datos en sus categorías y se cuenta el número de observaciones en cada clase (frecuencia absoluta); posteriormente se calcula la frecuencia relativa dividiendo la frecuencia absoluta por el número total de observaciones. Para las variables continuas cada clase es un intervalo, escogido de tal manera que cada observación pertenezca a una y sólo una clase.

Se ilustra este concepto para la variable continua *edad* y para la variable nominal *raza*.

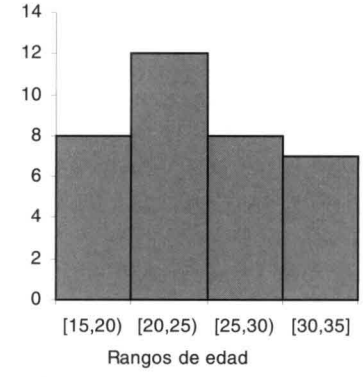
EDAD:

Clase	Frecuencia absoluta	Frecuencia relativa (%)
[15,20)	8	22.9
[20,25)	12	34.3
[25,30)	8	22.9
[30,35]	7	20.0

RAZA:

Clase	Frecuencia absoluta	Frecuencia relativa (%)
1	15	42.9
2	8	22.9
3	12	34.2

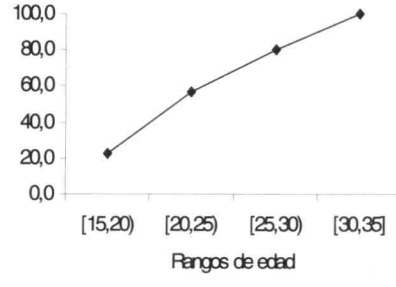
Estas distribuciones de frecuencia pueden representarse gráficamente por medio de un histograma. En el caso de la edad la representación gráfica es la siguiente:



La frecuencia acumulada representa la suma de frecuencias desde la clase inferior hasta la actual. La tabla siguiente representa la frecuencia acumulada de la variable edad.

EDAD Rango	Frec. absoluta acumulada	Frec. relativa (%) acumulada
[15,20)	8	22.9
[20,25)	20	57.1
[25,30)	28	80.0
[30,35]	35	100.0

Una gráfica de frecuencias relativas acumuladas proporciona información visual de los valores acumulados:



Con frecuencia se observan dos características a la vez y se quiere información sobre la frecuencia en una categoría determinada de una variable y de la otra simultáneamente. Esta presentación simultánea y cruzada de características se realiza por medio de las Tablas de Contingencia. Como ilustración se presenta la tabla de contingencia para las variables raza y edad.

EDAD					
Raza	[15,20)	[20,25)	[25,30)	[30,35]	**Total**
1	3	4	5	3	**15**
2	2	4	0	2	**8**
3	3	4	3	2	**12**
Total	**8**	**12**	**8**	7	**35**

Frecuentemente es preferible disponer de medidas que resuman los datos. En el caso de datos cualitativos las medidas de resumen empleadas son la moda y los porcentajes o proporciones. En el caso de datos ordinales también se utilizan como medidas de resumen la moda y la mediana. En el caso de datos cuantitativos se utilizan la media, la moda y la mediana (8).

En general, estas medidas de resumen permiten tener una idea de la localización de los datos. Esta medición debe acompañarse de otra medida que refleje la variabilidad de los datos (Medidas *de dispersión*):

Localización Central de los datos (Medidas de tendencia central) (9)*:*

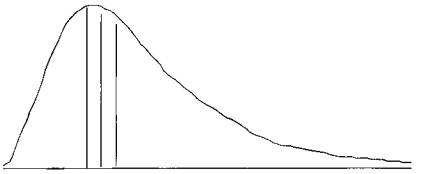
Este término se refiere a la media, a la mediana o a la moda, medidas apropiadas para análisis descriptivos de datos:

La moda: La moda de una distribución es, en variables cualitativas, el valor que aparece el mayor número de veces, es decir el que tiene mayor frecuencia. Si la variable es cuantitativa, es el punto medio de la clase en la cual aparece el mayor número de observaciones.

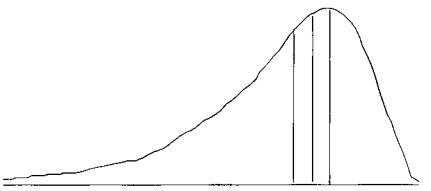
La mediana: Después de ordenar los datos de manera ascendente o descendente, la mediana es el valor central. Si el número de valores en el conjunto de datos N es impar, la mediana es el valor que aparece en el lugar (N +1)/2. Cuando N es par hay dos valores centrales y la mediana es el promedio de éstos.

La media: Es la más común entre las medidas de localización central. Se define como el promedio aritmético, es decir la suma de los N valores de la variable dividida por N. La media se usa para datos numéricos. La media es sensible a los valores extremos. Si se tiene una tabla de frecuencias, se puede estimar la media por un promedio ponderado que se obtiene multiplicando el punto medio de cada intervalo por el número de observaciones en ese intervalo.

En una distribución simétrica la moda, la mediana y la media son iguales.
Si la media es mayor que la mediana la distribución será sesgada a la derecha (positivamente sesgada):



En una distribución sesgada a izquierda , se tiene que la mediana es mayor que la media:



Medidas de Dispersión:

Las medidas de localización central no son suficientes para resumir los datos ya que no tienen en cuenta la variabilidad de estos. Por lo tanto se requiere una medida que indique la variabilidad de los datos. Una buena medida de dispersión o variabilidad debe ser independiente de la localización central de los datos y debe considerar todas las observaciones. Dentro de las medidas de dispersión tenemos:

El Rango: Es la diferencia entre los valores máximo y mínimo de la variable. Es independiente de la localización central pero considera sólo dos valores del conjunto de datos. Además un valor extremo altera el rango considerablemente.

Desviación estándar y varianza: Para hallar la varianza, cuya notación es S^2 ó σ^2 ,se calcula para cada observación x su desviación con respecto a la media, se eleva este resultado al cuadrado, se suma sobre todas las observaciones y esta suma se divide por N número de observaciones:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

La varianza es una buena medida de variabilidad ya que es independiente de la localización central y en ella intervienen todas las observaciones. Como la varianza se expresa en unidades al cuadrado, es conveniente definir otra medida de variabilidad: la desviación estándar, como la raíz de la varianza. Esta última es más conveniente para describir la variabilidad, ya que se expresa en las mismas unidades que los datos originales.

En nuestro estudio se tiene:

Variables	σ^2	**Moda**	**Mediana**	**Rango**	σ
Edad	27.98	24.0	24.0	19.0	5.29
Pesoma	119.2	115.0	115.0	85.0	10.92
Paridad	1.91	2.00	2.00	2.00	1.38

Hasta aquí se han discutido medidas apropiadas para resumir observaciones sobre una característica. Sin embargo, en un estudio es útil conocer las relaciones entre dos o más características. A continuación se hará la discusión para examinar la relación entre dos características numéricas (edad y peso

de la madre) y entre dos características ordinales (nivel socioeconómico e historia de hipertensión).

Para estimar la relación entre dos características numéricas se usa el coeficiente de correlación de Pearson, dado por

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{N\sigma_x\sigma_y}$$

Este coeficiente toma valores entre -1 y 1, el -1 describe una relación negativa perfecta y el 1 describe una relación positiva perfecta. El valor del coeficiente es independiente de las unidades de medida y está influenciado por valores extremos de la característica. Es importante notar que correlación no implica causalidad (10).

El coeficiente de correlación entre edad y peso de la madre es 0.254, lo cual indica que hay pobre correlación entre estas variables por ser éste muy alejado de 1 y de -1.

Para estimar la relación entre dos características ordinales se usa el coeficiente de correlación de Spearman, el cual considera el rango de las observaciones después de ser ordenadas, como si fueran los valores reales de las observaciones. Este coeficiente toma valores entre -1 y 1; -1 y 1 indican perfecta correlación entre los rangos de los valores y no entre los valores mismos.

El coeficiente de correlación entre nivel socioeconómico e historia de hipertensión es 0.1324; por lo tanto puede decirse que no existe correlación entre las dos características.

Para comparar la variabilidad de una característica cuantitativa en grupos de una característica nominal, es conveniente usar el coeficiente de variación que estandariza la variación pues equivale a la variación relativa al tamaño de la media. Su fórmula es:

$$CV = \frac{\sigma}{\mu} 100$$

Los conceptos hasta aquí ilustrados comprenden los elementos básicos para iniciar el proceso de formulación y prueba de hipótesis. No es posible plantear adecuadamente el método estadístico para probar una hipótesis si antes no se ha hecho una adecuada descripción de los datos, partiendo de una clara definición de las variables que se están manejando.

En una entrega posterior se presentarán los principales métodos estadísticos utilizados en la literatura médica, haciendo énfasis en los aspectos prácticos relativos a este tipo de aplicaciones.

REFERENCIAS BIBLIOGRÁFICAS

1. **Norman GR, Streiner DL:** Bioestadística. Barcelona: Mosby Doyma Libros: 1996.
2. **Wassertheil-Smoller S.** Biostatistics and Epidemiology. A primer for health proffesionals, 2nd ed. New York: Springer Verlag:1995;2-6.
3. **Buck C.** Popper's Philosophy for Epidemiologists. International Journal of Epidemiology 1975; 4:159-167.
4. **Ahlbom A.** Biostatistics for Epidemiologists. Boca Raton: Lewis Publishers:1993;76-78.
5. **Selvin S.** Statistical Analysis of Epidemiologic Data. 2nd ed. New York: Oxford University Press: 1996;93-94.
6. **Dunn G, Everitt B.** Clinical Biostatistics. An Introduction to Evidence-Based Medicine.New York: Edward Arnold:1995;12-20.
7. **Dawson-Saunders B, Trapp RG.** Bioestadística Médica. México: Manual Moderno:1993;24-26.
8. **Rosner B.** Fundamentals of Biostatistics, 4th ed. Belmont: Duxbury Press:1995;5-29.
9. **Daniel WW.** Bioestadística. Base para el análisis de las ciencias de la salud, 3ª ed. México: Noriega Limusa:1991;34-40.
10. **Altman DG.** Practical Statistics for Medical Research. London: Chapman& Hill:1991;277-298.

Nota: Tomado del libro "Estrategias de Investigación Medica Clínica "E Ardila , R Sanchez, J Echeverry Eds. (en prensa).