Rev. chil. endocrinol. diabetes 2009; 2 (2): 122-126

Variables aleatorias: El caso discreto

Gabriel Cavada Ch[1]

[1]*División de Bioestadística, Escuela de Salud Pública, Universidad de Chile.*

Una variable aleatoria (VA), en general, es una codificación numérica de los posibles resultados que contiene el espacio muestral de un experimento; dicha codificación puede ser arbitraria, pero, si el espacio muestral tiene algún orden jerárquico específico, este mismo orden sugiere la codificación. El empleo de variables aleatorias permite descubrir nuevas propiedades del experimento que se está estudiando. Las variables aleatorias pueden ser discretas o continuas; por ejemplo, el número de hijos que tiene una mujer adulta es una VA discreta pues puede tomar valores 0,1,2…, mientras que la talla de una persona es una VA continua pues puede ser cualquier valor de la recta real: 1.52, 1.76, $\sqrt{2}$, etc.

Variables aleatorias discretas

Si se analiza el experimento de lanzar un dado, para lo cual tenemos:

E: Se lanza un dado
$\Omega = \{1,2,3,4,5,6\}$

el espacio muestral, tiene seis sucesos fundamentales:

A_1: el dado muestra as
A_2: el dado muestra 2
.....
A_6: el dado muestra 6

Sin embargo, estos sucesos pueden ser codificados por la variable X, que tal que: X = 1 si ocurre A_1, X = 2 si ocurre A_2 etc. Hecha esta codificación, podemos hacer una descripción completa del experimento, pues las probabilidades asociadas con cada suceso se pueden representar por una función, que recibe el nombre de FUNCIÓN DE CUANTÍA DE PROBABILIDADES, que para este caso es:

Tabla 1

X	P (X)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
Total	1

Observamos que la suma de las P (X) es igual a 1, pues es la probabilidad del espacio muestral completo.
En general una función es una cuantía si toma sólo valores mayores o iguales 0 para cualquier valor de la variable aleatoria y la suma de los valores funcionales da 1, en símbolos:

$$P(X) \geq 0$$
$$\sum P(X) = 1$$

Caracterización de una VA discreta

Dada una función de cuantía definida sobre el conjunto de valores de la VA : X_1, X_2,…X_n , se define la Esperanza matemática o valor esperado de la VA X a la expresión:

$$E(X) = \mu = X_1 \cdot P(X_1) + X_2 \cdot P(X_2) + ... + X_n \cdot P(X_n) = \sum_{i=1}^{n} X_i \cdot P(X_i)$$

Se debe notar que la esperanza no es más que el promedio ponderado de las X por los valores de probabilidad correspondientes. La interpretación de este número es el promedio de la VA que se obtendría si se repitiera infinitas veces el experimento. Por ejemplo, en el caso del dado, cuya cuantía está en la tabla 1, la esperanza es:

$$E\ (X) = 1*1/6 + 2*1/6 + 3*1/6 + 4*1/6 + 5*1/6 + 6*1/6 = 3,5$$

Si imaginamos que muchísimas personas lanzan un dado honesto y nos reportan que número mostró el dado, el promedio de este reporte será 3,5.
Muchas veces se comete el error de confundir la esperanza con el valor más probable, este ejemplo nos muestra que el valor de la esperanza puede ser un resultado imposible al realizar el experimento: al lanzar un dado no se puede obtener 3,5.

La definición de esperanza puede generalizarse a lo que se conoce como momento de orden "r" de la siguiente forma:

$$E(X^r) = \sum X^r \cdot P(X_i)$$

Naturalmente, la esperanza se consigue poniendo r = 1, la esperanza también se conoce como momento de orden 1.
Evidentemente el momento de orden 2 está dado por la expresión:

$$E(X^2) = \sum X^2 \cdot P(X_i)$$

En el caso del lanzamiento del dado, el momento de orden 2 es:

$$E\ (X^2) = 1^2*1/6 + 2^2*1/6 + 3^2*1/6 + 4^2*1/6 + 5^2*1/6 + 6^2*1/6 = 15,17$$

El momento de orden 2 es imprescindible para definir la varianza de la VA X, es decir el grado de heterogeneidad de los resultados del experimento; en efecto, se define la varianza de la VA X como sigue:

$$Var(X) = E(X^2) - [E(X)]^2$$

Es decir, la varianza de X es el momento de segundo orden menos el cuadrado de momento de primer orden. Y por extensión la desviación estándar de X es:

$$\sigma = \sqrt{Var(X)} = \sqrt{E(X^2) - [E(X)]^2}$$

En nuestro ejemplo del dado se tiene respectivamente:

$$Var\ (X) = 15,17\text{-}3,5^2 = 2,92$$

$$\sigma = 1,\ 71$$

El lector recordará que en condiciones de simetría, la mayoría de los resultados está entre el promedio menos la desviación estándar y el promedio más la desviación estándar. En nuestro caso, la mayoría de los sujetos que han lanzado el dado han obtenido un resultado comprendido entre 3,5 ± 1,71, esto es entre 2 y 5, por ello que es más "romántico" lanzar un dado y obtener como resultado un As o el Seis.

En conclusión, una VA queda caracterizada por su función de cuantía, su esperanza y su varianza.

Rincón de la Bioestadística

Algunas funciones de cuantía de relevancia en medicina

Se conocen como Experimentos de Bernoulli a una secuencia de experimentos que tienen las siguientes características:

1) El experimento tiene sólo dos posibles resultados, que llamaremos éxito y fracaso.
2) Cada vez que se repite el experimento, la probabilidad de aparición del éxito (y de fracaso) se mantiene constante.
3) Cada ensayo es independiente de otro.

Si llamamos p a la probabilidad del éxito en un ensayo, la probabilidad del fracaso es 1-p al que llamaremos q, es decir q = 1-p o bien p + q = 1.

Asociados a los experimentos de Bernoulli, se definen las siguientes distribuciones de probabilidades:

i. La distribución de Bernoulli

En una población que está dicotomizada respecto de un determinado atributo (los elementos que poseen el atributo *versus* el resto de la población), en que la proporción con el atributo es p y q = 1-p la proporción que no lo posee, se realiza el experimento de extraer un elemento y se observa la presencia del atributo, podemos asumir la codificación:

X = 0, si el objeto no tiene el atributo
X = 1, si el objeto tiene el atributo

con lo que se obtiene la siguiente función de cuantía:

X	P (X)
0	p
1	q

o bien:

$$P(X = x) = p^x q^{1-x}, x = 0,1$$

así se tiene que:

$$E(X) = p$$
$$Var(X) = p(1-p) = pq$$

Esta es una cuantía muy trivial de manejar, sin embargo es de interés por la varianza, pues se observa que la variabilidad depende de la probabilidad del éxito; esta varianza se hace máxima cuando p = 0,5, es decir cuando en la población la mitad de los sujetos posee el atributo de interés. Por esta razón cuando no se conoce la prevalencia de una enfermedad y se desea calcular el tamaño de muestra necesario para estimarla, hay que suponer que dicha prevalencia es del 50%.

ii. Distribución Binomial

El modelo binomial, permite calcular la probabilidad de tener k éxitos, cuando se realizaren en n ensayos; si tenemos n intentos la cantidad de éxitos que podríamos obtener va desde 0 a n, es decir X = 0, 1, 2, ..., n . En este contexto:
La cuantía de probabilidades es:

$$P(X) = \binom{n}{x} \cdot q^{n-x} \cdot p^x, X = 0,1,$$

La esperanza y la varianza son:

$$E(X) = n \cdot p$$
$$Var(X) = n \cdot p \cdot q$$

NOTA: Recordemos que:

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \text{ y que } \quad j! = 1 \cdot 2 \cdot 3 \cdots (j-1) \cdot j$$

Ejemplo $3! = 1 \cdot 2 \cdot 3 = 6$

por definición $0! = 1$

Por ejemplo, si se sabe que la prevalencia de diabetes tipo 2 en la población adulta chilena es de un 6%, ¿Cuál es la probabilidad de que en un grupo de 10 personas escogidas al azar haya exactamente un diabético?. Aquí n = 10, p = 0,06 y X = 1, con lo que la probabilidad pedida es:

$$P(X=1) = \binom{10}{1} \cdot 0{,}94^{10-1} \cdot 0{,}06^{1} = 0{,}03438$$

Es decir que en grupos de 10 personas extraídas al azar de la población, encontrar exactamente un diabético tiene probabilidad 3,4%. En grupos de 10 personas esperamos encontrar 10*0,06 = 0,6 diabéticos, es decir menos de uno en promedio.

Esta distribución juega un papel importante cuando se quiere muestrear la población para detectar sujetos con enfermedades de muy escasa prevalencia. Por ejemplo, si una rara enfermedad tiene una prevalencia de 0,3%, es decir la probabilidad de detectar un sujeto enfermo es de 0,003, ¿Cuántos sujetos hay que muestrear para detectar a lo **menos uno** con la enfermedad con probabilidad de encuentro del 99%? Es decir, ¿de qué tamaño debe ser el grupo de sujetos para que tengamos un 99% de certeza que encontraremos a lo menos un enfermo?

Si suponemos que el grupo de sujetos tiene tamaño n, y X es el número de sujetos enfermos en dicho grupo, el problema se plantea de la siguiente forma:

$$P(X \geq 1) = 0{,}9$$

Notar que el suceso complementario es que X=0, por lo tanto el problema se puede plantear como:

$$P(X=0) = 0{,}01$$

Es decir la probabilidad de no encontrar sujetos enfermos en el grupo será sólo de un 1%; si el requerimiento lo ponemos en la función de cuantía de la distribución binomial, se tiene que n es la incógnita, X = 0 y p = 0,03 (prevalencia de la enfermedad) con lo que:

$$P(X=0) = \binom{n}{0} \cdot 0.,7^{n-0} \cdot 0{,}03 = 0{,}97^{n} = 0{,}01$$

$$n \ln(0{,}97) = \ln(0{,}01)$$

$$n = \frac{\ln(0{,}01)}{\ln(0{,}97)} = 151{,}2 \approx 153$$

Es decir en un grupo de 153 sujetos hay un 99% de certeza de encontrar a lo menos un enfermo.

iii. Distribución de Poisson

La distribución de Poisson es de suma importancia, pues modela sucesos de "rara" ocurrencia; por ejemplo, la cantidad de temblores que ocurren en un año en un determinado lugar o la cantidad de ovulaciones de una adolescente insulino resistente en determinado período de tiempo.

Dado un fenómeno de rara ocurrencia por alguna unidad de medida, es posible, por la experiencia acumulada, establecer una tasa de ocurrencia que llamaremos l (una incidencia). En estas condiciones la variable X es la cantidad de veces que aparece el fenómeno en un período, así X puede tomar valores enteros desde 0 al infinito, con lo que:

$$P(X) = \frac{\lambda^{x}}{X!} \cdot e^{-\lambda}, X = 0,1,2,3..$$

$$e = 2{,}71828...$$

Rincón de la Bioestadística

La esperanza y la varianza de esta distribución son, respectivamente:

$$E(X) = \lambda$$
$$Var(X) = \lambda$$

Notar que la esperanza y la varianza son iguales a la incidencia del fenómeno que estamos estudiando.
Por ejemplo, si en la historia de un paciente diabético no controlado se han producido 5 hiperglicemias en 10 años de seguimiento. ¿Cuál es la probabilidad de que en el próximo año calendario el paciente no presente una hiperglicemia? Aquí la tasa de incidencia anual del evento de hiperglicemia es:

$$\lambda = 5/10 = 0{,}5 \left[\frac{hiperglicemias}{año}\right]$$

La cantidad de hiperglicemias que se podrían presentar en un año son X=0,1,2,… en teoría se podrían presentar hasta infinitos eventos. Así la cuantía de probabilidades para el número de eventos anuales es:

$$P(X) = \frac{0{,}5^x}{X!} \cdot e^{-0{,}5}$$

Así, la probabilidad de que este paciente no presente el evento en un año calendario es:

$$P(X = 0) = \frac{0{,}5^0}{0!} \cdot e^{-0{,}5} = 0{,}6065$$

o bien 60,7%; obviamente, la probabilidad de que a lo menos presente un evento es 39,3%.

La Figura 1 muestra las frecuencias teóricas que tiene el número de eventos en este paciente:

Figura 1.