Las Probabilidades y sus Aplicaciones a la Estadística

PROF. DGO. ALMENDRAS

(Continuación)

Momentos m_1 m_2 y variancia de $ax + by$.

Si x e y son variables aleatorias independientes, tenemos la relación:

$$
\begin{aligned}
1)\ M_3 (ax + by, t) &= M_1 (ax, t)\ M_2 (by, t) \\
&= M (x, \theta_1)\ M (y, \theta_2) \text{ siendo } at = \theta_1;\ bt = \theta_2.
\end{aligned}
$$

(ver Rev. Mat. N.º 3 pág. 21).

Los momentos m_1 y m_2 se obtendrán de la primera y segunda derivada de (1) con respecto a t y luego haciendo $t = 0$. Estas derivadas son:

$$
\begin{aligned}
2)\ M'_3 (t) &= a\, M'_1 (x, \theta_1)\ M_2 (y, \theta_2) + bM'_2 (y, \theta_2)\ M_1 (x, \theta_1) \\
3)\ M''_3 (t) &= a^2 M''_1 (x, \theta_1)\ M_2 (y, \theta_2) + 2abM'_1 (x, \theta_1)\ M'_2 (y, \theta_2) \\
&\quad + b^2 M''_2 (y, \theta_2)\ M_1 (x, \theta_1)
\end{aligned}
$$

Las derivadas que figuran en el segundo miembro siendo tomadas con respecto a θ_1 y θ_2. Para $t = 0$ los primeros miembros de (2) y (3) se reducen a los momentos m_1 y m_2 de la variable $ax + by$, estos son, pues:

$$
\begin{aligned}
4)\ m_1 (ax+by) &= am_1 (x) + bm_1 (y) = a\bar{x} + b\bar{y} \\
5)\ m_2 (ax+by) &= a^2 m_2 (x) + 2ab\bar{x}\,\bar{y} + b^2 m_2 (y)
\end{aligned}
$$

De la fórmula (8), pág. 19 del Nº 3 de la Rev. de Mat. para $p = 2$ tenemos:

$$
\begin{aligned}
\text{var. } x &= m_2 - 2\, m_1^2 + m_1^2 \\
6)\ \text{var. } x &= m_2 - m_1^2
\end{aligned}
$$

Aplicando esta relación para determinar var. $(ax+by)$, tenemos:

$$
\begin{aligned}
7)\ \text{var. } (ax+by) &= m_2 (ax+by) - m_1^2 (ax+by) \quad \text{de 4) y 5)} \\
&= a^2 [m_2 (x) - \bar{x}^2] + b^2 [m_2 (y) - \bar{y}^2] \\
&= a^2 \text{ var. } x + b^2 \text{ var. } y
\end{aligned}
$$

Esta se puede generalizar a una suma S a_i x_i, siendo las x_i, variables aleatorias independientes.

Ejemplo 1. En una urna U_1 hay 4 bolitas blancas y 2 negras; en otra urna U_2 hay 3 bolitas blancas y 2 negras. En la primera se saca 5 veces sucesivamente una bolita, reponiéndola después de cada tirada y anotando al final de esta experiencia el número x de bolitas negras que han salido; luego en la urna U_2, en la misma forma, se hacen 3 tiradas anotando el número y de bolitas negras que han salido. Admitiéndose un número ilimitado de este par de experiencias se pide el valor medio y variancia de la variable x + y.

Solución. Las 5 tiradas en U_1 podrían ser BBNBN y en la segunda U_2, NBB; en este caso se tendría $x = 2$, $y = 1$; $x+y = 3$.

La variable x puede tomar los valores: 0, 1, 2, 3, 4 y 5. La variable y puede tomar los valores: 0, 1, 2 y 3. Vemos que estas variables son independientes y que la ley de distribución de cada una de ellas es binomial (Rev. de Mat. Nº 2, págs. 24 y 25). Los datos que tenemos para la resolución del problema son:

$$n_1 = 5 \; ; p_1 = 1/3 \; ; q_1 = 2/3$$
$$n_2 = 3 \; ; p_2 = 2/5 \; ; q_2 = 3/5$$

Luego tenemos:

$$\begin{aligned} m_1 (y + y) &= n_1 p_1 + n_2 p_2 \\ &= 5/3 + 6/5 = 43/15 = 2{,}865 \\ \text{var.} (x + y) &= n_1 p_1 q_1 + n_2 p_2 q_2 \\ &= 10/9 + 18/25 = 412/225 \\ &\doteq 1{,}618 \end{aligned}$$

Desviación standard. La raíz cuadrada de la variancia es un número muy importante en toda la teoría estadística y se le designará por σ

En el ejemplo anterior tenemos

$$\text{var.}(x + y) = 1{,}618$$

luego: $\sigma_{x+y} = \sqrt{1{,}618} = 1{,}27$

Ejemplo 2. En una urna que contenía 18 bolitas casi idénticas, una persona hizo 200 extracciones de 5 bolitas en cada tirada y anotaba el número de bolitas (1) que salían. El resultado de las experiencias está contenido en el cuadro que se acompaña. Se pide averiguar la probabilidad de salida de una bolita azul, valor medio del número de bolitas azules que sale en cada experiencia y la variancia de este número. Sabiendo que de las 18 bolitas 5 eran azules, se pide calcular teóricamente: probabilidad de salida de una bolita, azul, valor medio del número de bolitas azules que salen en cada experiencia y la variancia de dicha variable.

(1) azules.

Solución. Los datos de la experiencia están en las dos primeras columnas. Las otras dos columnas fx y fx^2 sirven para el cálculo del valor medio y del momento m_2.

Nº de azules por cada experiencia x	frecuencia f	fx	fx^2
0	27	0	0
1	80	80	80
2	74	148	296
3	17	51	153
4	2	8	32
5	0	0	0
	200	287	661

La probabilidad de salida de una bolita azul es:

$$p' = \frac{287}{1000} = 0{,}287$$

porque de 200 tiradas de 5 bolitas que equivale a 1000 tiradas de una solamente resultaron 287 azules.

El momento m_1 es $m'_1 = \dfrac{661}{200} = 3{,}305$

$$\text{var.x} = 3{,}305 - m'^2_1$$

Pero $m'_1 = 287/200 = 1{,}435$, por consiguiente,

$$\text{var.x} = 3{,}305 - 1{,}435^2$$

$$\delta_x' = 1{,}246.$$

Sabiendo que de 18 bolitas 5 eran azules, la probabilidad de salida de una bolita azul al sacar una cualquiera al azar será $5/18 = 0{,}277 = p$. Esta probabilidad es algo menor que la que resultó de la experiencia.

La variancia teórica es:

$$\frac{N-n}{N-1}\, n\, p'q' = \frac{5}{18} \cdot \frac{13}{18} \cdot \frac{13}{17} = \frac{4225}{5508} = 0{,}77$$

$$\sigma_x = 0{,}877$$

Concepto de muestra al azar. Dada una población de individuos, por ejemplo, un conjunto de seres humanos, podemos clasificarlos considerando aspectos cualitativos o cuantitativos. Los primeros se llaman atributos y los segundos variables cuantitativas o parámetros. Son atributos de una población el sexo, el grado de instrucción, el idioma, etc.; son parámetros de una población la estatura media, el salario medio, la edad.

En muchos problemas de estadística se presentan sólo dos aspectos que podemos dominar A y B de modo que cada individuo tiene o uno o el otro, pero no ambos. En una población de nacimientos durante un año los atributos A y B serían los sexos masculino y femenino; en la población de los que han rendido bachillerato en la temporada enero, los atributos A y B podrían ser: aprobados o reprobados; hombres y mujeres, etc.

Los que poseen el atributo A es una fracción p del total y los que poseen el atributo B será la fracción q = 1 — p.

Para averiguar en corto tiempo y costo relativamente pequeño, ciertos aspectos de una población, se toman de ellas muestras satisfaciendo condiciones determinadas científicamente. Entonces los parámetros de la muestra son representativos de los correspondientes de la población entera.

Una encuesta dirigida a una población entera mediante formularios puede ser menos significativa que una muestra bien planeada, aun cuando ésta contenga mayor número de preguntas, debido, principalmente, a que la primera contiene numerosas no-respuestas y los datos aprovechables no serán representativos.

Entre los tipos de muestras más usada en estadística se mencionan las tomadas al azar.

Para comprender el concepto de muestra al azar imaginaremos que deseamos conocer la estatura media de los concriptos de la guarnición de Santiago tomándola de una muestra de 200 individuos. Se deberá enumerar a todos los concriptos de la guarnición, luego se colocarán en una urna igual número de bolitas idénticas en cuya superficie supondremos que llevan inscritos los números desde uno hasta el número total de concriptos, de modo que se pueda hacer corresponder cada bolita a un determinado concripto. Removiendo lo más posible la urna se sacarán por sorteo 200 bolitas que determinará el grupo de concriptos cuya estatura media será un valor aproximado a la de la población completa.

Como en la práctica es difícil hacer un sorteo mediante bolitas colocadas en una urna, se han construído tablas de dígitos tomados al azar que permiten, con gran rapidez, obtener una muestra al azar. Una parte de estas tablas es la que sigue y que enseñaremos a manejar:

Tabla de 400 dígitos tomados al azar

54463	22662	65905	70639
15389	85205	18850	39226
85941	40756	82414	02015
61149	69440	11286	88218
05219	81619	10651	67079
41417	98326	87719	92294
28357	94070	20652	35774
17783	00015	10806	83091
40950	84820	29881	85966
82995	64157	66164	41180
96754	17676	55659	44105
34357	88040	53364	71726
06318	37403	49927	57715
62111	52820	07243	79931
47534	09243	67879	00544
98614	75993	84460	62846
24856	03648	44898	09351
96887	12479	80621	66223
90801	21472	42815	77408
55165	77312	83666	36028

Parte de la tabla que figura en "Métodos de Estadística" por G. Snedecor.

Supongamos que de una población de 500 individuos se desea sacar una muestra al azar de 10 individuos. Se supondrá enumerados los individuos de 1 hasta 500, luego, considerando los números de tres cifras que se pueden formar con las tres primeras cifras de la izquierda en cada grupo de 5 dígitos, empezando con la columna de la izquierda y de arriba hacia abajo se comenzarán a marcar los números menores que 500 hasta completar 10 de estos números. Si algún número se repite se anotará una sola vez. Así en el ejemplo que consideramos deberemos tomar los números 544, 153, 859, 611 ... pero debemos marcar sólo los menores que 500. Estos son: 153, 052, 414, 283, 177, 409, 343, 063, 475, 248. Los números de tres cifras se pueden tomar de la derecha de cada grupo de 5 dígitos y comenzar a leer verticalmente u horizontalmente. Así, tomando las tres cifras de la derecha se puede obtener la muestra formada por los siguientes números 463, 389, 149, 219, 417, 357, 318, 111, 165, 205. Los conscriptos que llevan los números de la primera o segunda muestra forman dos muestras al azar de 10 individuos.

Veamos otro ejemplo. En el cuadro siguiente se dan las longitudes de Trypanosomas en micrones y las frecuencias respectivas y se pide sacar una muestra al azar de 20 individuos y hacer el cuadro de frecuencia de la muestra:

Longitud en micrones	Frecuencia simple	Frecuencia acumulada
15	7	7
16	31	38
17	148	186
18	230	416
19	326	742
20	252	994
21	237	1231
22	184	1415
23	143	1558
24	115	1673
25	130	1803
26	110	1913
27	127	2040
28	133	2173
29	113	2286
30	96	2382
31	54	2436
32	44	2480
33	11	2491
34	7	2498
35	2	2500

Se supone los Trypanosomas enumerados de 1 a 2500 de modo que uno que tenga el número 1156, por ejemplo, según se ve en la columna de frecuencias acumuladas, tendrá una longitud de 21 micrones. Tomando los números de cuatro cifras de la derecha y de arriba hacia abajo y menores que 2500 elegiremos 20. Estos son: 1538, 0521, 1778, 0631, 2485, 2266, 0001, 1767, 0924, 0364. Los datos relativos a esta muestra son los del cuadro siguiente:

Longitud en micrones	Frecuencia simple
15	1
18	1
19	2
20	1
23	1
25	2
29	1
33	1
	10

Intervalo de confianza. Cuando se toma una muestra al azar de una cierta población ya sea con el fin de obtener la proporción de individuos que poseen un determinado atributo o de obtener un número característico de ella, como ser una estatura media, un salario medio, etc., es de fundamental importancia saber qué aproximación puede proporcionar dicha muestra con respecto a los valores verdaderos y cual es la probabilidad de obtener una aproximación que quede dentro de un intervalo dado. La teoría matemática de las muestras, que analizaremos en próximos artículos, nos enseña que si p es la proporción (frecuencia relativa) de los individuos que poseen un atributo E y σ_p la desviación standard (típica) o si x es el promedio de la muestra, y σ_x la desviación standard, entonces hay una probabilidad de 0,95 o de 95% que una proporción p′ o un promedio x′ quede comprendido entre $p-2\sigma_p$, $p+2\sigma_p$ y $\bar{x}-2\sigma_x$, $\bar{x}+2\sigma_{\bar{x}}$ respectivamente, es decir:

$$p-2\sigma_p \leq p' < p + 2\sigma_p \; ; \bar{x}-2\sigma_x \leq x' \leq \bar{x}+2\sigma_x, \text{ con probabilidad } 0.95.$$

De donde se deduce que:

$$p'-2\sigma_p \leq p \leq p'+2\sigma_p \text{ y } x-2\sigma_x \leq \bar{x} \leq x+2\sigma_x \text{ con probabilidad de } 95\%$$

En la práctica las desviaciones típicas se calculan aproximadamente mediante los datos tomados de la misma muestra.

Ejemplo 3. Se tomó el color de 150 automóviles que pasaron sucesivamente por Avenida Beaucheff, resultando 35 de color negro. Se pregunta ¿qué número de automóviles de color negro se podrían contar al pasar 1000 automóviles, con seguridad de 95%?

Solución. Se supone que la muestra de 150 automóviles es al azar y para el cálculo σ_p admitiremos que la distribución de los automóviles negros sigue la ley binomial, siendo $p'=35/150=7/30$; $q'=23/30$; luego:

$$\sqrt{np'q'} = 5,17$$

$$2\sigma_p = 10,3$$

Luego el intervalo de seguridad de 95% será (35-10,3 ;35 +10,3)=(25;45). Este intervalo expresado en % del tamaño de la muestra será (16,6;30), por consiguiente en 1000 automóviles el intervalo de seguridad de 95% será (166 ;300) o sea " de 1000 automóviles tomados al azar hay una seguridad de 95% de contar un número de automóviles negros comprendido entre 166 y 300.

Ejemplo 4. De acuerdo con la ley de la herencia de Mendel, el cruce de dos clases de frejoles produce del color amarillo y verde en la razón 3:1. En un experimento resultaron 176 amarillos y 48 verdes. Se pregunta si este resultado está de acuerdo con la teoría. (Mathematical Statistics por Paul G. Hoel).

Solución. Los datos son: $n = 224; p=3/4; q=1/4$.

$$\sigma_p = \sqrt{npq} = 6,5$$

$$2\sigma_p = 13 \text{ ; } np = \bar{x} = 224 \times 3/4 = 168.$$

El intervalo de seguridad de 95% es: (155;181) y, vemos que el resultado obtenido de 176 amarillos queda dentro del intervalo anterior y está de acuerdo con la teoría.