DIFERENCIAS DE VARIABLES BINOMIALES

Por *Gabriel Poveda Ramos*

0. Es bien sabido que en el siglo XVIII el gran matemático Daniel Bernouilli dedujo la fórmula que hoy conocemos con su nombre y que da lugar a la llamada distribución binomial de probabilidad. Dicha fórmula da la probabilidad de que, al extraer aleatoriamente (y con reposición) una muestra de m elementos de un universo estadístico cuyos miembros pueden poseer (o no poseer) una cierta característica o propiedad E, en la muestra aparezca un número determinado r de elementos que poseen esa propiedad. Si la probabilidad de que un elemento al azar del universo tenga la característica E es p, entonces la probabilidad buscada es

$$f_r = \binom{m}{r} p^r (1-p)^{1-r} \qquad (0.01)$$

Este resultado es muy usado en problemas de control de calidad y en muchas otras aplicaciones de la Teoría de Probabilidades a problemas estadísticos. Un ejemplo de estas aplicaciones es el de la efectividad de una droga: Si admitimos que la droga es eficiente en una fracción p de todos los enfermos humanos, entonces la probabilidad de que al tomar m enfermos y aplicarles la droga, resulte efectiva para r de ellos, viene dada por la fórmula (0.01) de Bernouilli.

1. La fórmula (0.01) puede usarse para resolver el problema inverso, de inferencia estadística, que consiste en lo siguiente: Supongamos que no conocemos el valor de la probabilidad p, que se llama la "probabilidad *a priori*", pero se sabe que la característica E (como ser aliviados por la droga) ha aparecido en r miembros de los m elementos que forman una muestra aleatoria simple, extraídos con reposición del universo en cuestión (como ser, por ejemplo, todos los afectados de cierta dolencia), el problema es estimar la probabilidad a priori p.

Según el criterio de máxima verosimilitud, el valor de p es el que haga máxima la expresión f_r, en donde r y m tienen valores fijos y conocidos. Esto implica que debe buscarse p de manera que haga

$$\max_p \; p^r \; (1-p)^{m-r}$$

Esto último equivale, como bien se sabe, a que haga

$$\max_p \; ln \left[p^r \; (1-p)^{m-r} \right]$$

o sea

$$\max \left[r \; ln\, p + (m-r) \; ln \, (1-p) \right] = \max_p S \, (p)$$

Para esto es necesario que $dS\,(p)/dp = 0$, o sea, que:

$$\frac{r}{p} - \frac{m-r}{1-p} = 0$$

$$r \; (1-p) = p \, (m-r)$$

es decir que

$$p = r/m \qquad (1.01)$$

Puede demostrarse que este estimador, dado por la fórmula (1.01) es un estimador centrado (no sesgado) y eficiente, aunque no es usual que aparezca en los textos corrientes de Estadística.

2. Este artículo se refiere al problema análogo, siguiente, que se refiere a dos universos: Una carac-

terística E puede presentarse (o estar ausente) en los miembros de un universo U_1 y de un universo U_2 distinto al primero. De U_1 se extrae una muestra aleatoria simple de m_1 elementos donde E aparece en r_1 de ellos. Y de U_2 se extraen aleatoria y simplemente m_2 elementos, hallando que E está presente en r_2 de ellos. El problema es el de inferir cómo se comparan las probabilidades a priori, p_1, p_2, de E en los dos universos respectivos U_1, U_2.

Este es el caso de un universo U_1 formado por todos los pacientes humanos de cierta dolencia con una droga D_1 y de un universo U_2 de todos los pacientes que pueden tratarse con otra droga D_2. De los primeros se hace una prueba sobre m_1 pacientes, de los cuales r_1 sanan; y de los segundos se hace una prueba sobre m_2 pacientes, de los cuales r_2 responden a la segunda droga. El problema es comparar las probabilidades a priori p_1, p_2 de que las dos drogas respectivas D_1, D_2 sean efectivas en el conjunto de todos los afectados por la dolencia en cuestión.

3. Para responder al problema que hemos planteado en el párrafo anterior, consideramos las variables aleatorias $X_1 = r_1/m_1, X_2 = r_2/m_2$ que representan las proporciones en que la característica E se presenta en las dos muestras que se comparan, tomadas de los dos universos respectivos. Compararemos las dos variables X_1, X_2 estudiando su diferencia $X_1 - X_2 = L$ como variable aleatoria.

La probabilidad de que la variable X_1 valga u (siendo u uno de los números 0, $1/m_1$, $2/m_1$, ..., $(m_1-1)/m_1$, 1) la indicamos como $h_1\,(u)$. Es igual a la probabilidad expresada por la fórmula (0.01) poniendo $r = r_1$, $m = m_1$.

A su vez, la variable X_2 puede tener los valores 0, $1/m_2$, $2/m_2$, ..., $(m_2-1)/m_2$, 1; y la probabilidad de que adopte uno de esos valores (que indicaremos genéricamente como v) es $h_2\,(v)$. Dicha probabilidad coincide con la que da la fórmula (0.01) poniendo $r = r_2$, $m = m_2$.

Es decir

$$h_1\,(u) = Pbbdd\cdot(X_1 = r_1/m_1 = u) = fr_1 = f_{(u.m_1)}$$

$$h_2\,(v) = Pbbdd.\,(X_2 = r_2/m_2 = v) = fr_2 = f_{(v.m_2)}$$

Es fácil demostrar (aunque casi ningún texto usual de Estadística o de Probabilidades lo hace), que la probabilidad de que, con esta notación, la variable L adopte el valor $L = k$, está dada por la expresión

$$g\,(k) = \sum_u h_1\,(u)\,h_2\,(u-k) \qquad (3.01)$$

en donde u recorre todos sus valores 0, $1/m_1$,.., $(m_1-1)/m_1$, 1. Por esta razón, el factor $h_2\,(u-k)$ sólo adquiere valores no nulos en aquellos valores de u tales que $u-k$ adopta alguno de los valores 0, $1/m_2$, $2/m_2$,.., 1.

Escribiéndolo en forma explícita, se tiene la siguiente expresión (3.02):

$$g\,(k) = \sum_u \binom{m_1}{m_1 u} p_1^{\,m_1 u}\cdot(1-p_1)^{m_1\,(1-u)} \binom{m_2}{m_2\,(u-k)} \cdot p_2^{\,m_2\,(u-k)}\cdot(1-p_2)^{m_2\,(1-u-k)} \qquad (3.02)$$

donde u asume la sucesión de valores $0, 1/m_1$,...., $(m_1-1)/m_1$, 1, al paso que $u-k$ adopta los valores 0, $1/m_2$, $2/m_2$, ... $(m_2-1)/m_2$, 1.

4. Una situación interesante es el caso en que, como resultado de un experimento aleatorio, resulten r_1, r_2 de tales valores que $r_1/m_1 = r_2/m_2$. Este es precisamente el caso en que $L = r_1/m_1 - r_2/m_2 = 0$, o sea, en que $k = 0$. También sucede que $L = 0$ cuando r_1 y r_2 adoptan ambos el valor cero.

La probabilidad *ex-ante* de que esto resulte así vale, según la expresión (3.02), lo siguiente:

$$g(0) = \sum_u \binom{m_1}{m_1 u} p_1^{\,m_1 u}\cdot(1-p_1)^{m_1\,(1-u)}\cdot\binom{m_2}{m_2 u} p_2^{\,m_2 u}\cdot(1-p_2)^{m_2\,(1-u)} \qquad (4.01)$$

El criterio de máxima verosimilitud expresa que, si no se conocen previamente los valores de p_1, p_2 pero sí los de m_1, m_2, la mejor inferencia respecto a las dos probabilidades *a priori* es la que haga

$$\max_{p_1,\,p_2}\; g\,(0)$$

Como bien se sabe, para ello es necesario que se cumplan simultáneamente las dos condiciones:

$$\partial g\,(0)\,/\,\partial p_1 = 0 \qquad (4.02)$$

$$\partial g\,(0)\,/\,\partial p_2 = 0 \qquad (4.03)$$

Tomando la derivada parcial de (4.01) respecto a p_1, y escribiendo, para más comodidad, $m_1 u = r_1$ y $m_2 = a\,m_1$, la ecuación (4.02) se convierte en

$$\sum \binom{m_1}{r_1}\binom{m_1 a}{r_1 a}\; p_2^{\,r_1 a}\cdot(1-p_2)^{(m_1-r_1)a}\cdot$$

$$\left[\; r_1 p_1^{\,r_1-1}\cdot(1-p_1)^{m_1-r_1} - p_1^{\,r_1}\,(m_1-r_1)\cdot (1-p_1)^{m_1-r_1-1}\;\right] = 0 \qquad (4.04)$$

De la misma manera la ecuación (4.03) conduce a la ecuación

$$\sum_{r_1}\binom{m_1}{r_1}\binom{m_1 a}{r_1 a}\; p_1^{\,r_1}\cdot(1-p_1)^{m_1 r_1}\cdot$$

$$\left[\; r_1 a p_2^{\,r_1 a-1}\cdot(1-p_2)^{(m_1-r_1)a} - p_2^{\,r_1 a}\,(m_1-r_1)a\,(1-p_2)^{(m_1-r_1)a-1}\;\right] = 0$$

Un cálculo algebraico sencillo demuestra que la condición necesaria y suficiente para que estas últimas dos condiciones sean compatibles es que se tenga

$$r_1/p_1 - (m_1-r_1)/(1-p_1) = r_1/p_2 - (m_1-r_1)/(1-p_2)$$

para todos los valores de r_1. Esto significa necesariamente, que

$$p_1 = p_2$$

Es decir, que si hemos obtenido $r_1/m_1 = r_2/m_2$ (o sea $k = 0$), mejor inferencia es que las probabilidades *a priori* son iguales.

En el caso de las dos drogas D_1, D_2, si hemos tratado m_1 pacientes (tomados al azar) con la primera droga, y se han aliviado r_1 de ellos; y si, habiendo tratado m_2 pacientes con la segunda, se han aliviado r_2 de ellos, siendo $r_2/m_2 = r_1/m_1$, puede inferirse por el criterio de máxima verosimilitud que las dos drogas tienen una misma eficiencia curativa para todos los pacientes ($p_1 = p_2$) en general. Por lo demás, este es un resultado intuitivamente muy verosímil.

5. En cambio, si la droga D_1 alivia a r_1 pacientes entre m_1, al paso que la droga D_2 alivia r_2 pacientes entre m_2; y si designamos con k a la diferencia $r_1/m_1 - r_2/m_2 = k$, la mejor inferencia respecto a p_1, p_2 es la que haga

$$\partial g(k)/\partial p_1 = 0$$

$$\partial g(k)/\partial p_2 = 0$$

de manera simultánea. Escribiendo $g(k)$ en la forma

$$g(k) = \sum_{r_1}\binom{m_1}{r_1}\; p_1^{\,r_1}\cdot(1-p_1)^{m_1-r_1}\binom{m_1 a}{r_1 a - k m_2}\cdot$$

$$p_2^{\,r_1 a - k m_2}\cdot(1-p_2)^{(m_1-r_1)a + k m_2}$$

Tomando las dos derivadas parciales respecto a p_1, p_2 y haciendo las operaciones algebraicas correspondientes, se encuentra que las condiciones de compatibilidad de las ecuaciones es que para todos los valores de r, se tenga

$$m_1 p_1 - r_1 = 0 \qquad (4.06A)$$

y que

$$m_2 p_2 - r_1 a - k m_2 = 0 \quad \text{con } a = m_2/m_1 \qquad (4.06B)$$

cuya solución simultánea única es

$$p_2 = p_1 - k. \qquad (4.07)$$

Podemos pues inferir que si $r_1/m_1 - r_2/m_2 = k$, los poderes curativos de las dos drogas guardan la misma diferencia

$$p_1 - p_2 = k.$$

6. Pero, además de los anteriores resultados, interesa calcular la distribución $g(k)$ de probabilidades de la variable $L = X_1 - X_2$, dada más arriba por las expresiones (3.01) y (3.02).

En primer lugar interesa identificar los valores que pueda adoptar la variable $k = r_1/m_1 - r_2/m_2$. Es evidente que el máximo posible de k se presenta cuando r_1 adopta el valor $r_1 = m_1$, y cuando r_2 toma el valor $r_2 = 0$, que es el mínimo posible. En tal caso $max\ k = 1$. Así mismo, el mínimo valor de k se presenta cuando $r_1 = 0$ y $r_2 = m_2$, en cuyo caso se tiene $min\ k = -1$.

Todos los valores de k se forman haciendo la diferencia $r_1/m_1 - r_2/m_2$, poniendo $r_1 = 0, 1, 2, \ldots$, m_1 y $r_2 = 0, 1, 2, \ldots, m_2$. Examinemos pues, la diferencia de quebrados

$$\frac{r_1}{m_1} - \frac{r_2}{m_2}$$

En la Aritmética elemental se muestra que esta diferencia de quebrados se calcula buscando primero el mínimo común múltiplo de los denominadores:

$$M = m.c.m.\ (m_1, m_2)$$

y sus cuocientes enteros con estos mismos denominadores:

$$P_1 = M/m_1 \quad , \quad P_2 = M/m_2$$

Entonces la diferencia de quebrados que se busca, es

$$k = \frac{r_1}{m_1} - \frac{r_2}{m_2} = \frac{r_1\,(M/m_1) - r_2\,(M/m_2)}{M} =$$

$$\frac{r_1 P_1 - r_2 P_2}{M} = \frac{S}{M} \qquad (6.01)$$

El numerador $r_1 P_1 - r_1 P_2 = S$ adopta valores que son todos números enteros. Tales valores resultan al darle a r_1 los valores 0, 1, 2, ..., m_1, y a r_2 los valores $r_2 = 0, 1, 2, ..., m_2$, y se pueden disponer en una tabla de doble entrada, o matriz rectangular, como los que se muestra en las Tablas 1 y 2. De la fórmula (6.01) resulta que

$$r_1 = m_1 k + (m_1/m_2)\, r_2$$

es decir que

$$r_1 \geqslant m_1 k \qquad (6.02A)$$

o bien que

$$r_1 \geqslant m_1\, S/M \qquad (6.02B)$$

TABLA 1

Valores del numerador $S = r_1 P_1 - r_2 P_2$ con $m_1 = 12, m_2 = 10$

	r_1 : 0	1	2	3	4	5	6	7	8	9	10	11	12
r_2													
0	0	5	10	15	20	25	30	35	40	45	50	55	60
1	−6	−1	4	9	14	19	24	29	34	39	44	49	54
2	−12	−7	−2	3	8	13	18	23	28	33	38	43	48
3	−18	−13	−8	−3	2	7	12	17	22	27	32	37	42
4	−24	−19	−14	−9	−4	1	6	11	16	21	26	31	36
5	−30	−25	−20	−15	−10	−5	0	5	10	15	20	25	30
6	−36	−31	−26	−21	−16	−11	−6	−1	4	9	14	19	24
7	−42	−37	−32	−27	−22	−17	−12	−7	−2	3	8	13	18
8	−48	−43	−38	−33	−28	−23	−18	−13	−8	−3	2	7	12
9	−54	−49	−44	−39	−34	−29	−24	−19	−14	−9	−4	1	6
10	−60	−55	−50	−45	−40	−35	−30	−25	−20	−15	−10	−5	0

TABLA 2

Valores del numerador $S = r_1 P_1 - r_2 P_2$ con $m_1 = 8, m_2 = 6$

	r_1 : 0	1	2	3	4	5	6	7	8
r_2									
0	0	3	6	9	12	15	18	21	24
1	−4	−1	2	5	8	11	14	17	20
2	−8	−5	−2	1	4	7	10	13	16
3	−12	−9	−6	−3	0	3	6	9	12
4	−16	−13	−10	−7	−4	−1	2	5	8
5	−20	−17	−14	−11	−8	−5	−2	1	4
6	−24	−21	−18	−15	−12	−9	−6	−3	0

7. En la Tabla 1 está tratado el caso de la diferencia $r_1 P_1 - r_2 P_2$ cuando r_1 va desde cero hasta $m_1 = 12$ y r_1 va desde cero hasta $m_2 = 10$. En esas condiciones, se tiene

$M = m.c.m.\ (10, 12) = m.c.m.\ (5 \times 2, 2 \times 2 \times 3) =$
$5 \times 4 \times 3 = 60$

$P_1 = M/m_1 = 60 \div 12 = 5$

$P_2 = M/m_2 = 60 \div 10 = 6$

$S = r_1 P_1 - r_2 P_2 = 5\, r_1 - 6\, r_2$

El número de entradas o casillas de esta tabla es, por supuesto, $(m_1 + 1) \times (m_2 + 1) = 13 \times 11 = 143$

pero en ella solamente hay 101 valores numéricos distintos, como puede comprobarse contándolos directamente en la tabla. Nótese que en esta situación, el mínimo común múltiplo M se puede expresar como

$$M = 60 = 5 \times 6 + 6 \times 5 = AP_1 + B P_2$$

siendo

$$A = 5, \quad B = 6$$

y que el número N de valores numéricos en la tabla es

$$N = 3AB + A + B = 3 \times 5 \times 6 + 5 + 6 = 101 \qquad (7.01)$$

8. En la Tabla 2 se presentan los valores del numerador $S = r_1 P_1 - r_2 P_2$ cuando $m_1 = 8$ y $m_2 = 6$. En este caso el mínimo común múltiplo M es

$$M = m.c.m.\ (8, 6) = m.c.m.\ (2^3, 2 \times 3) = 2^2 \times 3 = 24$$

y sus cuocientes con m_1, m_2, son

$$P_1 = M/m_1 = 24/8 = 3$$

$$P_2 = M/m_2 = 24/6 = 4$$

Aquí también se tiene la situación de que

$$M = 24 = 4 \times P_1 + 3 \times P_2 = 4 \times 3 + 3 \times 4 = 24$$

de modo que

$$M = A P_1 + B P_2$$

$$A = 4 \quad , \quad B = 3$$

El número de valores distintos en esta Tabla 2 es de $N = 43$ como se obtiene enumerándolos y contándolos directamente. Pero también aquí comprobamos que

$$N = 43 = 3 AB + A + B = 3 \times 4 \times 3 + 4 + 3 = 43 \qquad (8.01)$$

9. En la Tabla 3 se presenta el caso en que $m_1 = 5$, $m_2 = 4$, que son números primos entre sí. En estas condiciones

$$M = m.c.m.\ (m_1, m_2) =$$

$$m.c.m.\ (5, 4) = 5 \times 4 = m_1 \times m_2 = 20$$

$$P_1 = M/m_1 = 20/5 = 4 = m_2$$

$$P_2 = M/m_2 = 20/4 = 5 = m_1$$

En tales condiciónes *no* existen números naturales A, B que permitan escribir $M = AP_1 + B P_2$. El número de valores numéricos distintos que aparecen en la tabla es, más bien,

$$N = 29 = m_1 m_2 + m_1 + m_2 = 5 \times 4 + 5 + 4 = 29 \qquad (9.01)$$

número que puede comprobarse por enumeración en la propia Tabla 3.

TABLA 3

Valores del numerador $S = r_1 P_1 - r_2 P_2$ con $m_1 = 5, m_2 = 4$

r_2 \ r_1:	0	1	2	3	4	5
0	0	4	8	12	16	20
1	−5	−1	3	7	11	15
2	−10	−6	−2	2	6	10
3	−15	−11	−7	−3	1	5
4	−20	−16	−12	−8	−4	0

10. En la Tabla 4 se muestra el caso en que $m_1 = 3, m_2 = 6$, que es el caso en que uno de estos números es múltiplo del otro y cuando

$$M = m.c.m.\ (m_1, m_2) = max\ (m_1, m_2) = m^*$$

$$M = m.c.m.\ (3, 6) = 6$$

$$P_1 = M/m_1 = 6/3 = 2$$

$$P_2 = M/m_2 = 6/6 = 1$$

En este caso sí existen números naturales A, B tales que

$$M = 4 P_1 + B P_2$$

porque, con $A = 2$ y $B = 2$ se obtiene

$$M = 6 = 2 P_1 + 2 P_2 = 2 \times 2 + 2 \times 1 = 6$$

TABLA 4

Valores del numerador $S = r_1 P_1 - r_2 P_2$ con $m_1 = 3, m_2 = 6$

r_2 \ r_1:	0	1	2	3
0	0	2	4	6
1	−1	1	3	5
2	−2	0	2	4
3	−3	−1	1	3
4	−4	−2	0	2
5	−5	−3	−1	1
6	−6	−4	−2	0

Pero el número de valores distintos en la tabla es $N = 13$ como se puede contar sobre ella misma. De manera que en este caso N es distinto de $3\ AB + A + B$, a diferencia de lo que se tenía en (7.01) y en (8.01). En realidad, en estas condiciones el número de valores diferentes que adopta S es

$$N = 2.\ max\ (m_1, m_2) + 1 = 2 \times 6 + 1 = 13 \quad (10.01)$$

11. Los anteriores ejemplos sugieren unas reglas de formación para calcular el número N de los valores k que puede adoptar la diferencia algebraica $L = X_1 - X_2$. Porque cada valor de k corresponde biunívocamente a cada valor distinto de $S = k.\ M = r_1\ P_1 - r_2\ P_2$. Aquí enunciaremos, sin demostrarlas, algunas de estas reglas (que pueden enunciarse como "teoremas" si se prefiere):

a.– Si m_1 *y* m_2 son iguales ($m_1 = m_2 = m$), el número N de valores numéricos que adopta S (o bien k) es

$$N = 2\ m + 1$$

b.– Si m_1, m_2 son primos entre sí (o sea, carecen de factores primos comunes que sean distintos de uno (1), como ocurre en la Tabla 3), el número N de valores numéricos diferentes que toma el numerador S de las fracciones $k = S/M$ es decir, el número de valores diferentes que toma k, es

$$N = m_1\ m_2 + m_1 + m_2 \quad (11.02)$$

La fórmula (9.01) es un ejemplo de esta situación.

c.– Si m_1, m_2 son múltiplos el uno del otro (y por lo tanto *no* son primos entre sí), como en la Tabla 4, el número N de valores numéricos diferentes que toma S es

$$N = 2.\ max\ (m_1, m_2) + 1 = 2\ m^* + 1 \quad (11.03)$$

Un ejemplo de esta situación lo da la fórmula (10.01)

d.– Si m_1, m_2 *no* son primos entre sí (tienen factores comunes distintos de uno), ni uno es múltiplo del otro, pero existen dos números naturales *positivos* A (menor que m_1) y B (menor que m_2) tales que

$$M = m.c.m.\ (m_1, m_2) = A\ (M/m_1) + B\ (M/m_1)$$

o sea que satisfacen la condición

$$A\ /\ m_1 + B\ /\ m_2 = 1$$

entonces el número N de valores de S (o también de valores de L), es

$$N = 3\ AB + A + B$$

Este es el caso presentado en la Tabla 1 y en la Tabla 2.

e.– Otras situaciones entre m_1, m_2 se dejan para la curiosidad del lector que se interese en Aritmética.

12. Para no fatigar al lector no daremos aquí las demostraciones explícitas de todas las fórmulas consignadas en el párrafo anterior. Pero daremos la demostración de la fórmula (11.04) correspondiente al caso (d), por ser una de las más interesantes.

Para fijar ideas, nos remitimos a la Tabla 2. En primer lugar, obsérvese que en el centro geométrico de la tabla está el número cero. En la casilla $r_1 = 4$, $r_2 = 0$ está el número 12 que es 3 x 4 o sea $B\ P_2$. El mismo número 12 está en la casilla $r_1 = 8$, $r_2 = 3$. En los dos extremos de la diagonal principal está el número cero. Y en los dos extremos de la otra diagonal (o diagonal transversal) está el número $24 = M$. Esta configuración de números en la tabla es lo que refleja en ella la relación aritmética

$$M = A\ P_1 + B\ P_2$$

A la derecha de la columna mediana de $r_1 = 4$, hay 4 columnas (en general, A columnas), y a su izquierda hay otras A columnas. Encima de la línea horizontal mediana de $r_2 = 3$ hay 3 columnas (en general, B columnas), y debajo de esa línea hay otras B columnas.

Así pues, la tabla puede descomponerse en los siguientes fragmentos, que no contienen números comunes entre sí

a) La submatriz de la parte superior derecha

15	18	21	24
11	14	17	20
7	10	13	16

que contiene $12 = 4 \times 3$ elementos. En general, AB elementos.

b) La submatriz de la parte inferior izquierda, con números todos negativos, simétrica de la anterior, que contiene también AB elementos.

c) El número cero, que aparece en tres casillas. Es otro elemento más, distinto de los anteriores.

d) Los números en la fila mediana ($r_2 = 3$), a la izquierda de la columna mediana ($r_1 = 4$), que son

$$-12 \qquad -9 \qquad 6 \qquad 3$$

cuyo número es 4 (en general, son A números).

e) Los números

$$-4$$
$$-8$$

de la primera columna ($r_1 = 0$) encima de la fila mediana $r_2 = 3$, y omitiendo el cero (que está en $r_1 = 0$, $r_2 = 0$) el cual ya se ha enumerado. Son pues, otros dos números (en general, $B - 1$ números).

f) Los números

$$-1 \quad 2 \quad 5$$

$$-5 \quad -2 \quad 1$$

que forman una submatriz en la parte superior izquierda de la tabla, cuyo número es $6 = 3 \times 2$ (en general, $(A - 1) \times (B - 1)$ números).

g) Los números

$$3 \qquad 6 \qquad 9 \qquad 12$$

que aparecen en la primera fila, desde $r_1 = 1$ hasta $r_1 = 4$.

Son cuatro números (en general, son A números).

h) Los números

$$8$$

$$4$$

que aparecen en la mitad superior de la columna mediana ($r_1 = 4$), omitiendo el cero (0) y el doce (12) que ya están enumerados. Son dos números (en general, son $B - 1$ números).

i) Todas las demás casillas de la tabla contienen números que ya están incluidos en alguno de los fragmentos anteriormente descritos.

En consecuencia, el número de valores distintos que contiene la tabla, es, sumando los cardinales indicados atrás,

$$N = 2AB + 1 + A + (B-1) + (A-1)(B-1) + A + (B-1) = 3AB + A + B$$

que es lo que se trataba de demostrar.

Las otras fórmulas, (11.01), (11.02) y (11.03) son más sencillas de demostrar.

13. Volviendo a la distribución de probabilidades de L, dada por

$$g(k) = \sum_u h_1(u)\, h_2(u-k)$$

esta distribución se refiere a los N valores diferentes que puede adoptar el número $k = S/M$, donde $M = m.c.m.(m_1, m_2)$ y $S = r_1 (M/m_1) - r_2 (M/r_2)$ permitiendo a r_1 recorrer la sucesión $1, 2, \ldots, m_1$ y permitiendo a r_2 que recorra la sucesión $1, 2, \ldots m_2$. De todas maneras, los valores de k pertenecen a la sucesión -1, $(-M + 1)/M$, $(-M + 2)/M$,... $-1/M$, 0, $+1/M$, $2/M$,..., $(M-2)/M$, $(M-1)/M$, 1, e incluye necesariamente a $k = -1$, a $k = 0$ y a $k = +1$.

14. Con lo observado hasta aquí, podemos construir el siguiente algoritmo para calcular la distribución de probabilidades de la diferencia $X_1 - X_2 = L$; donde $X_1 = r_1/m_1$ y $X_2 = r_2/m_2$; donde r_1 tiene distribución binomial desde 0 hasta m_1, con probabilidad *a priori* p_1; y donde r_2 es también binomial, desde cero hasta m_2, con probabilidad *a priori* p_2.

a) Anotar m_1, m_2, p_1, p_2

b) Calcular $M = m.c.m.(m_1, m_2)$

c) Calcular: $P_1 = M/m_1$ y $P_2 = M/m_2$

d) ¿Es $m_1 = m_2$?

d.1.- Sí. Ponga $m = m_1 = m_2$. Calcule $N = 2m + 1$

d.2.- No. ¿Son m_1, m_2 primos entre sí?

d.2.A.- Sí. Calcule $N = m_1 m_2 + m_1 + m_2$

d.2.A.- No. ¿Son m_1, m_2 múltiplos el uno del otro?

d.2.A.I. Sí. Calcule $m^* = max(m_1, m_2)$ y calcule $N = 2m^* + 1$

d.2.A.II. No. ¿Hay dos números positivos A m_1 y B m_2 tales que $M = AP_1 + BP_2$?

d.2.A.II.1.- Sí. Calcule $N = 3AB + A + B$

d.2.A.II.2.- No. Siga al paso (e)

e) Forme la tabla T de las $(m_1 + 1) \times (m_2 + 1)$ diferencias

$$r_1/m_1 - r_2/m_2$$

f) Cuente el número N de valores numéricos distintos que forman esta tabla. Si es del caso, confróntelo con el que ya se haya calculado en el paso (d), en la parte pertinente.

Cerciórese de que $N \leqslant 2M + 1$.

g) Anote $k = -1$ como primer valor para k. Le corresponde el valor $S = -M$, que está colocado en el extremo inferior izquierdo de la tabla T, donde $r_1 = 0$ y $r_2 = m_2$.

h) Calcule $S = kM$. Escriba k como $k = S/M$.

i) Calcule el valor correspondiente de la expresión $Sm_1/M = k\, m_1 = R_1$.

j) En la tabla T inspeccione las columnas que tienen valores de r_1 enteros y que sean iguales o mayores que R_1 ($r_1 \geqslant R_1$) leyéndolos en orden creciente, de derecha a izquierda. Identifique y enumere ordenadamente aquellas donde aparezca la S que calculamos en el paso (h). Note que puede no haber columnas que contengan S. En tal caso, pase al paso (ñ).

k) En cada una de las columnas así señaladas, encabezadas por los valores indicados de r_1, anote el valor de r_2 de la fila donde aparezca el valor de S ya calculado. Dicho valor de r_2 es único para cada r_1, y puede también calcularse por la fórmula.

$$r_2 = (r_1 P_1 - kM)/P_2$$

l) Anote todas las parejas (r_1, r_2) así formadas, que corresponden a cada valor de R_1 (es decir, al valor de k prescrito en (g) y en (h) ya calculado en el paso (i)).

m) Formar para cada una de estas parejas el producto

$$\binom{m_1}{r_1} p_1^{r_1} (1-p_1)^{m_1 - r_1} \cdot \binom{m_2}{r_2} p_2^{r_2} (1-p_2)^{m_2 - r_2}$$

n) Sumar estos productos para el conjunto de las parejas obtenidas en el paso (l), que corresponden a un mismo valor de k (o, si se quiere, a un mismo valor de S). Esta suma es el valor de $g(k)$. Anótelo.

ñ) Pase a $k' = k + 1/M$. Esto significa que S pasa a $S' = S + 1$.

o) Repita todos los pasos desde (h) hasta (n), para k'. Así obtiene

$$g(k')$$

p) Repita la operación (ñ) (y las sucesivas de (h) a (n), que le corresponden) hasta llegar a $k = +1$, o sea a $S = +M$.

q) Calcule $g(k)$

r) Haga el listado de $g(-1), .., g(0), .., g(+1)$ ya encontrados.

s) Pare.

De esa manera se ha examinado toda la secuencia de $2M+1$ números $-1, (-M+1)/M, ..., -2/M, -1/M, 0, +1/M, 2/M, .., (M-1)/M, +1$ y se han identificado aquellos N números de la secuencia anterior que son valores de $k = r_1/m_1 - r_2/m_2$. Además, se han calculado y listado los N valores de la distribución de probabilidad $g(k)$.

15. El algoritmo que dejamos escrito nos permite calcular los valores numéricos de la distribución de probabilidades $g(k)$. Además, dándole a k los N valores numéricos $k_1, k_2, ..., k_N$ que se le han identificado, también puede calcularse numéricamente el valor medio (o esperanza) de k, realizando las operaciones implícitas en la fórmula bien conocida.

$$E[k] = \bar{k} = \sum_{i=1}^{N} k_i \cdot g(k_i) \qquad (15.01)$$

Dado que $k = r_1/m_1 - r_2/m_2$, y recordando un teorema bien conocido en Estadística, deberá resultar que, además, $\bar{k}$ resulta igual a

$$\bar{k} = p_1 - p_2 \qquad (15.02)$$

Porque

$$\bar{r}_1 = m_1 p_1 \quad , \quad \bar{r}_2 = m_2 p_2$$

Podemos también calcular la varianza de k mediante fórmula conocida

$$V[k] = \sum_i g(k_i) \cdot (k_i - \bar{k})^2 \qquad (15.03)$$

Una identidad bien conocida en Estadística, y válida en nuestro caso, permite escribir

$$V[k] = V[r_1/m_1 - r_2/m_2] = (1/m_1^2) V[r_1] + (1/m_2^2) V[r_2] =$$

$$(1/m_1^2) m_1 p_1 (1-p_1) + (1/m_2^2) m_2 p_2 (1-p_2) =$$

$$p_1 (1-p_1)/m_1 + p_2 (1-p_2)/m_2 \qquad (15.04)$$

Este resultado (15.04) debe coincidir con el resultado numérico (15.03).

16. Por último, señalaremos una forma de escribir la distribución $g(k)$ que también puede usarse en otros casos en que se manejan variables aleatorias binomialmente distribuidas.

Comenzamos por observar que la expresión (3.02) puede también escribirse como

$$g(k) = \sum_{r_1, r_2} \binom{m_1}{r_1} p_1^{r_1} (1-p_1)^{m_1 - r_1} \binom{m_2}{r_2} p_2^{r_2} (1-p_2)^{m_2 - r_2} \qquad (16.01)$$

en donde r_1, r_2 son, para cada valor de k, las parejas identificadas en el paso (1) del algoritmo descrito en el párrafo 14, y entre las cuales se cumple la condición

$$r_1/m_1 - r_2/m_2 - k = 0$$

siendo $r_1 = 0, 1, 2, ..., m_1$ y $r_2 = 0, 1, 2, ..., m_2$. Además k adopta luego los valores $k_1, k_2, ..., k_N$

(que forman una sub-sucesión de la sucesión $k_1 = -1$, $(-M+1)/M, \ldots, 1/M, 0, +1/M, +2/M, \ldots, (M-1)/M, +1. = k_N$).

Los coeficientes binomiales que aparecen en la fórmula (16.01) se pueden también escribir

$$\binom{m_1}{r_1} = \frac{m!}{r!\,(m_1 - r_1)!} = \frac{\Gamma(m+1)}{\Gamma(r+1)\,.\,\Gamma(m-r+1)}$$

$$= \frac{m\,\Gamma(m)}{r\,\Gamma(r)\,.\,(m-r)\,\Gamma(m-r)} \quad \frac{m}{r(m-r)} \quad \frac{\Gamma(m)}{\Gamma(r)\,.\,\Gamma(m-r)}$$

$$\frac{m}{r(m-r)}\, B(r, m-r)$$

donde B (r, m) es la función llamada "beta" u "euleriana de primera especie".

$$B(r, s) = \int_0^1 x^{r-1} (1-x)^{s-1}\, dx$$

y que mantiene una identidad muy conocida

$$B(r,s) = \frac{\Gamma(r+s)}{\Gamma(r)\,.\,\Gamma(s)}$$

con la función "gama".

$$\Gamma(t) = \int_0^\infty e^{-x}\, .\, x^{t-1}\, .\, dx$$

llamada "euleriana de segunda especie".

Por lo tanto, la fórmula (16.01) puede escribirse

$$g(k) = \sum_{r_1}^{m_1} \frac{m_1}{r_1\,(m_1 - r_1)}\, B(r_1, m_1 - r_1)\,.$$

$$p_1^{r_1}\,(1-p_1)^{m_1 - r_1}\,.\, \frac{m_2}{r_2\,(m_2 - r_2)}\,.$$

$$B(r_2, m_2 - r_2)\; p_2^{r_2}\; (1-p_2)^{m_2 - r_2} \qquad (16.02)$$

en donde:

a) cada valor de k, r_1 es la variable que cumple las dos condiciones

$$r_1 \in (0, 1, 2, 3, \ldots, m_1)$$

$$r_1 \geqslant k\, m_1$$

b) a cada valor de k y de r_1 le corresponde el valor único r_2 dado por las condiciones

$$r_2 = m_2\,(r_1/m_1 - k) \in (0, 1, 2, \ldots, m_2)$$

c) la variable k recorre la sucesión $k_1, k_2, \ldots, k_N$ que se ha formado en las etapas (g) y (ñ) del algoritmo del párrafo 14.

17. Si suponemos que $p_1 = p_2 = p$, la expresión (16.02) toma la forma

$$g(k) = \sum_{r_1} \binom{m_1}{r_1} \binom{m_2}{m_2\,(r_1/m_1 - k)}\,.$$

$$p^{\,r_1\,(1\,+m_2/m_1)\,-m_2 k}\qquad (1-p)^{\,m_1\,+m_2\,(1+k)\,-r_1\,(1+m_2/m_1)} \qquad (17.01)$$

con las advertencias (a), (b), (c) del párrafo 16. Numéricamente pueden calcularse los N valores de $g(k)$ con $k = -1, k_2, k_3, \ldots, k_N = 1$

18. Volviendo al caso concreto de la comparación de las drogas, la distribución $g(k)$ permite comparar la una contra la otra en aquellos casos en que se conoce la eficacia p_1 de la droga D_1 pero no hay certidumbre sobre la eficacia p_2 de la droga D_2. En este caso, se siguen los siguientes pasos:

a) Se formula la hipótesis nula H_0 de que $p_2 = p_1 = p$.

b) Se construye la tabla T; se extraen los N valores distintos de S que contiene, y se dividen por M para obtener los N valores distintos de $k = S/M$.

c) Para cada valor de k se identifican los valores de r_1 que deben entrar en la sumación de la fórmula (17.01), como en el paso (j) del algoritmo.

d) Se valora la fórmula (17.01) sumando sobre los valores de r_1 correspondientes a cada valor de k.

e) Se obtiene así la distribución

$$g_i = g(k_i)\ , \quad k_i = k_1, k_2, \ldots, k_N$$

que tiene dos colas: en el lado de los valores más bajos de k (k_1, k_2 ...), y en el lado de los valores más altos de k (..., k_{N-2}, k_{N-1}, k_N)

f) Se toma una muestra de m_1 pacientes y se tratan con la droga D_1. Se cuentan los r_1 de ellos en que la droga es eficaz.

g) Se toma una muestra de m_2 pacientes. Se tratan con D_2. Se cuenta el número r_2 de los que resultan aliviados.

h) Se mide la diferencia k de $r_1/m_1 - r_2/m_2 = k$.

j) Se establece un nivel α de significación para la prueba. Usualmente se toma 1% ó 5% según que se quiera ser más o menos exigente.

k) Se identifican los primeros valores k_1, k_2,.. de la cola izquierda, cuya suma de probabilidades sea más próxima a $\alpha/2$

$$g_1 + g_2 + \ldots + g_{N_1} \cong \alpha / 2$$

l) Se hace lo mismo en el extremo de la cola superior:

$$g_{N_2} + g_{N_2+1} + \ldots + g_N \cong \alpha / 2$$

m) Si el valor medido de k pertenece a la sucesión de la cola izquierda

$$k_1, k_2, \ldots, k_{N_1}$$

se rechaza la hipótesis H_0 y se formula una nueva hipótesis nula H'_0 de que $p_2 > p_1$

n) Si el valor medido de k pertenece a la sucesión de la cola derecha

$$k_{N_2}, k_{N_3}, \ldots, k_N$$

se rechaza la hipótesis H_0 y se formula una nueva hipótesis H''_0 de que $p_2 < p_1$

o) Si el valor medido de k cae en el centro de la distribución, es decir en uno de los valores

$$k_{N_1}, k_{N_1+1}, \ldots, k_{N_2-2}, k_{N_2-1}$$

no se rechaza la hipótesis H_0 sino que se toman nuevas muestras hasta que sus extensiones reunidas (m_1, m_2) hagan que la desviación típica (Ver fórmula 15.04) sea suficientemente pequeña respecto a p. Por ejemplo, que sea

$$\sqrt{p(1-p)\left(\frac{1}{m_1} + \frac{1}{m_2}\right)} < 0.01\, p$$

si es que se usa este criterio de validez para aceptar la hipótesis H_0.

19. Es evidente que el procedimiento que se ha expuesto aquí sirve no solamente para comparar la eficacia de dos drogas, sino para comparar muchas otras parejas comparables de datos. Ejemplos de tales comparaciones son:

a) Las resistencias mecánicas de dos marcas de barras de metal.

b) Las velocidades de dos corredores.

c) Los contenidos de un cierto mineral en dos yacimientos distintos.

d) Las durabilidades de dos marcas de buriles para metal.

f) Las eficiencias de dos procedimientos de fabricación; y muchas otras.