Las Probabilidades
en nuestro mundo II.
Problemas de Bolas

Ulises Cárcamo

Aunque detestados por muchísimos estudiantes y muchos profesores de estadística, los problemas de probabilidades relativos a grupos de esferas son casos particulares de modelos probabilísticos que se aplican a fenómenos tan diversos como el muestreo de aceptación, el control de calidad, problemas de lingüística y el comportamiento de los gases.

INTRODUCCIÓN

En el número 115 de la Revista Universidad EAFIT publiqué un primer artículo sobre probabilidades, en tono jocoso y tratando de mostrar que muchos temas relativos a la probabilidad tienen interpretaciones sencillas; que a partir de elementos no muy complicados se puede llegar a cosas muy interesantes en probabilidad y que debemos usar las herramientas computacionales a nuestro alcance (Hojas electrónicas, paquetes matemáticos, etc.) para facilitar los cálculos. En este artículo trato de continuar con la filosofía del primero.

Un problema bastante odiado

Todos los que hemos estudiado las nociones elementales de las probabilidades hemos enfrentado problemas referentes a bolas o esferas. Un ejemplo típico es el siguiente: "En una urna hay diez bolas de las cuales cuatro son rojas y el resto negras y están todas mezcladas. Si se extraen tres bolas al azar sin reemplazo, es decir, sin volver a depositar en la urna las bolas que se han examinado, ¿cuál es la probabilidad de que dos de ellas sean rojas?"

Son muchísimos los estudiantes y los profesores que les tienen terror o fastidio a esta clase de problemas. Además, cuando a algunos profesores de Estadística les pedimos ayuda para resolver alguno, dicen, con mucha razón, que este tipo de problemas "no son problemas de estadística" e inmediatamente se marchan o muy amablemente nos invitan a salir de su oficina. Esta actitud, en parte justificada, además de conseguir más detractores de este tipo de problemas, oculta a nuestra vista los secretos que ellos guardan.

Por ahora haremos caso omiso de aquellas objeciones y resolveremos el problema planteado.

Una primera solución

Utilicemos algo de sentido común. Si extraemos las bolas una a una, sin reemplazo, tendremos las siguientes posibles sucesiones de color: RRN, RNR y NRR, donde R representa al color rojo y N al color negro. Estas sucesiones representan los únicos "casos favorables" de este problema. La probabilidad del primer evento, RRN, es $\frac{4}{10}\cdot\frac{3}{9}\cdot\frac{6}{8}=\frac{1}{10}$,, ¿sabe el lector atento el porqué? La probabilidad del segundo evento es $\frac{4}{10}\cdot\frac{6}{9}\cdot\frac{3}{8}=\frac{1}{10}$, y la del tercero $\frac{6}{10}\cdot\frac{4}{9}\cdot\frac{3}{8}=\frac{1}{10}$; dado que los

ULISES CÁRCAMO. Licenciado en Educación, Área de Matemáticas, Universidad de Medellín. Máster en Matemáticas Aplicadas, Universidad EAFIT. Profesor, Universidad EAFIT. email: ucarcamo@eafit.edu.co

eventos son mutuamente excluyentes, la probabilidad de encontrar dos esferas rojas en la muestra es la suma de las tres probabilidades, es decir, $\frac{3}{10}$.

Ventajas de este primer método de solución

A simple vista, la resolución del problema por el método anterior no exige el conocimiento de ninguna fórmula en particular, sólo tener nociones intuitivas de la propiedad multiplicativa de las probabilidades y de los eventos mutuamente excluyentes. Por otro lado, el ejemplo nos sirve para ilustrar la noción de probabilidad condicional y la propiedad multiplicativa para tres conjuntos, a saber: $P(A \cap B \cap C) = P(A).\ P(B/A).\ P(C/A \cap B)$, siempre que $P(A)$ y $P(A \cap B)$ sean positivas.

Desventajas de este método

Si pretendemos seguir este mismo procedimiento para resolver problemas más complicados, veremos que la búsqueda de la solución es bastante larga y que, en general, el método es engorroso. Por ejemplo, inténtese encontrar la probabilidad de extraer 5 bolas rojas en una muestra de 15 extraída de una urna donde 20 esferas son rojas y 10 son negras.

Por esta razón intentaremos una segunda solución que sea fácil de generalizar.

Un segundo método (la aparición de la temida Combinatoria)

El problema también puede resolverse de la siguiente manera: Las bolas no se extraen una a una sino simultáneamente, en cuyo caso habrá que averiguar cuántos subconjuntos de tres bolas tiene un conjunto de diez de ellas y entre ellos cuántos tienen dos bolas rojas.

Salta a la vista que el primer problema es un problema clásico de la rama de las matemáticas conocida como Análisis Combinatorio. Esta rama presenta un cierto grado de dificultad mientras nos acostumbramos a pensar de acuerdo con sus lineamientos (¿pero, no es esto común a todas las ramas de la matemática?). Por esto es común escuchar la expresión "Análisis Adivinatorio", para refererirse a esta división de las matemáticas, entre todos aquellos que han sido "golpeados" por la dificultad de sus problemas. Nos propondremos pasar por alto todos estos "comentarios" e ir más allá, para ver las ventajas que la segunda aproximación al problema puede traer.

Algunos resultados importantes del Análisis Combinatorio

Los dos pilares de la combinatoria son los principios fundamentales del conteo:

Primer principio fundamental del conteo: Si una operación debe ejecutarse en k etapas, la primera etapa puede ejecutarse de n_1 maneras, la segunda de n_2 maneras, . . ., la k-ésima de n_k maneras, entonces, el número de posibles maneras en las que es posible realizar la operación es $n_1.n_2. \ldots .n_k$.

Segundo principio fundamental del conteo: Si una operación puede realizarse de la forma f_1 o la forma f_2, . . . o la forma f_k y todas las formas son excluyentes, es decir, si se realiza de una de esas formas, no se realiza de las demás formas y al realizar la operación de forma f_1 se puede hacer de n_1 maneras, realizarla de la forma f_2 se pueda hacer de n_2 maneras, . . ., realizarla de la forma f_k se puede hacer de n_k maneras, entonces el número total de maneras en las que se puede realizar la operación es de $n_1+n_2+ \ldots +n_k$ maneras.

A partir de los anteriores principios se pueden deducir los siguientes resultados:

1. El número de permutaciones de n objetos distintos, $P_{n,n}$ es n!

2. El número de permutaciones de r objetos distintos seleccionados de un conjunto de n es $P_{n,r} = n.(n-1).(n-2). \ldots .(n-r+1) = \frac{n!}{(n-r)!}$.

3. El número de subconjuntos (combinaciones) de r elementos extraídos de un conjunto de n elementos distintos es $C_{n,r} = \frac{n!}{r!(n-r)!} = \binom{n}{r}$; este último símbolo se llama símbolo combinatorio o coeficiente binomial y aparece en el Teorema del Binomio del álgebra elemental.

4. Si se tienen dos clases de objetos, la primera con n_1 objetos, la segunda con n_2, y si se quiere formar un conjunto con r_1 elementos de la primera clase y con r_2 elementos de la segunda, entonces el total de maneras posibles de hacer esto

está dado por $\binom{n_1}{r_1}\binom{n_2}{r_2}$; en particular, si $r_1+r_2 = k$, se tendrán $\binom{n_1}{r_1}\binom{n_2}{k-r_1}$ maneras posibles.

5. Más general, si se tienen S clases de objetos con respectivos números de elementos $n_1, n_2, \ldots, n_s$ y se quiere escoger un conjunto que contenga r_1, elementos de la primera clase, r_2 de la segunda, . . .y r_s, de la s-ésima, el número de maneras de hacerlo es $\binom{n_1}{r_1}\binom{n_2}{r_2}\cdots\binom{n_s}{r_s}$; en particular, si $r_1+r_2+\ldots+r_s= k$, el número de maneras puede escribirse como $\binom{n_1}{r_1}\binom{n_2}{k-r_1}\binom{n_3}{k-r_1-r_2}\cdots\binom{n_s}{r_s}$.

6. El número de permutaciones de n objetos, no todos distintos, cuando hay n_1 de una primera clase, n_2 de una segunda clase, . . ., n_k de la k-ésima clase está dado por $\frac{n!}{n_1!.n_2!....n_k!}$.

Naturalmente, estos no son los únicos resultados del Análisis Combinatorio; existen libros completos del tema y revistas especializadas dedicadas a los últimos avances y aplicaciones de esta rama.

¿Hay alguna aplicación práctica de estos resultados?

¡Por supuesto que sí!. Daremos dos ejemplos:

En muestreo

Supóngase que se tiene una población de 20.000, viviendas de entre las cuales se necesita extraer una muestra aleatoria de 1.000 viviendas para realizar una encuesta allí. ¿Cuántas muestras de este tamaño son posibles? Naturalmente, el número de muestras aleatorias, es decir, muestras en las que cada vivienda tiene la misma posibilidad de participar, es igual al número de subconjuntos de 1000 elementos tomados de un conjunto de 20000 es $\binom{20000}{1000}$; este es un número enorme que es aproximadamente $2.4761\text{x}10^{1722}$.

En Computación

Supóngase que se sabe que existe una clave de acceso a un programa ultrasecreto del enemigo, que utiliza 30 de los 225 caracteres del código ASCII. ¿Podría en un tiempo razonable encontrarse esa clave y acceder al programa? Supóngase que el computador empleado puede generar y probar 100 millones de permutaciones de 30 caracteres cada segundo (¿Es éste un supuesto realista?).

Primero calculemos el total de permutaciones de 30 caracteres escogidos de entre 225, que es $\frac{225!}{(225-30)!}$; este número es también muy grande y es aproximadamente igual a 4.85883×10^{69}. Si el computador es capaz de analizar $1\text{x}10^8$ permutaciones por segundo, tardará $4.85883\ \text{x}10^{61}$ segundos en generarlas y probarlas todas, pero esta cantidad de segundos equivale a $1.54072\text{x}10^{62}$ años; por lo tanto, aún con ese computador tan poderoso no podrá resolverse por "fuerza bruta"; hay que olvidarse de problema. ¿Indica esto por qué algunos programas comerciales tienen claves de acceso tan largas?

Usemos los resultados del Análisis Combinatorio para resolver el problema de referencia

Volvamos al primer problema: En número de "casos posibles" es el número de subconjuntos de tres bolas extraídos del conjunto de diez de ellas, $C_{3,10} = \binom{10}{3} = \frac{10!}{3!.7!} = 120$ y el número de "casos favorables" es el número de conjuntos de tres bolas donde hay dos de ellas rojas y es

$$\binom{4}{2}\binom{6}{1} = \frac{4!}{2!.2!}\cdot\frac{6!}{1!.5!} = 36$$ y, por lo tanto, la probabilidad buscada es

$$\frac{\binom{4}{2}\binom{6}{1}}{\binom{10}{3}} = \frac{36}{120} = \frac{3}{10}.$$

Naturalmente, aunque este método requiere más conocimientos, se puede aplicar fácilmente a otros problemas de este tipo y es más fácil de generalizar a otro tipo de situaciones.

Una primera generalización

Los resultados que se desprenden de la resolución de problemas matemáticos son, comúnmente, más útiles cuando se generalizan. Esto permite que se abarque una gama más amplia de aplicaciones. El resultado anterior se puede generalizar, en una primera etapa, con el siguiente enunciado:

Supóngase que se dispone N objetos, de los cuales k pertenecen a una clase y los restantes N-k pertenecen a una segunda clase. Si todos se encuentran distribuidos al azar y se escoge un conjunto de n objetos, sin sustitución, entonces la probabilidad de que hayan exactamente x objetos de la primera clase en el conjunto seleccionado es $\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$.

Y, ¿este resultado es también aplicable?

Claro que sí; consideremos una fábrica de cierto tipo de artículos, por ejemplo bombillos, y supongamos que a través de registros estadísticos de la producción, se conoce el porcentaje de defectuosos que produce (bombillos que no encienden, por ejemplo). Dado que los artículos defectuosos se producen no adrede sino por pequeños "accidentes" que podemos considerar aleatorios, (éstos) estarán dispersos entre los demás. Se exceptúan los casos en los que un desperfecto en una de las máquinas produzca defectuosos en serie, y similares. Supongamos además, como es natural, que los artículos se empacan en lotes de n antes de enviarlos a los distribuidores o directamente a los clientes.

¿Cuál es la probabilidad de que haya k defectuosos en un lote de n?

Para concretar, supongamos que la producción total es de N = 10000 bombillos por periodo de producción, el porcentaje de defectuosos es del 1%, n = 400 y k = 4.

Calcular esta probabilidad equivale a calcular $\dfrac{\binom{100}{4}\binom{9900}{396}}{\binom{10000}{400}} \approx 0.200392$. ¿Puede el lector atento explicar el porqué de esos coeficientes binomiales?

En muchos casos una probabilidad de este tamaño puede ser inadmisible; es más, calculemos la probabilidad de que en un lote haya 1,2, 3 ó 4 defectuosos: $\dfrac{\binom{100}{1}\binom{9900}{399}}{\binom{10000}{400}}+\dfrac{\binom{100}{2}\binom{9900}{398}}{\binom{10000}{400}}+\dfrac{\binom{100}{3}\binom{9900}{397}}{\binom{10000}{400}}+\dfrac{\binom{100}{4}\binom{9900}{396}}{\binom{10000}{400}} \approx 0.611382.$

Esta es una probabilidad bastante alta; nuestro lote podría no pasar algunas de las pruebas de control de calidad de los distribuidores o de los consumidores.

¡Ah!, ¿O sea que estas conclusiones y procedimientos se pueden aplicar al control de calidad?

Obviamente, pero antes de esto es necesario hacer una generalización importante: Construir una distribución de probabilidad discreta:

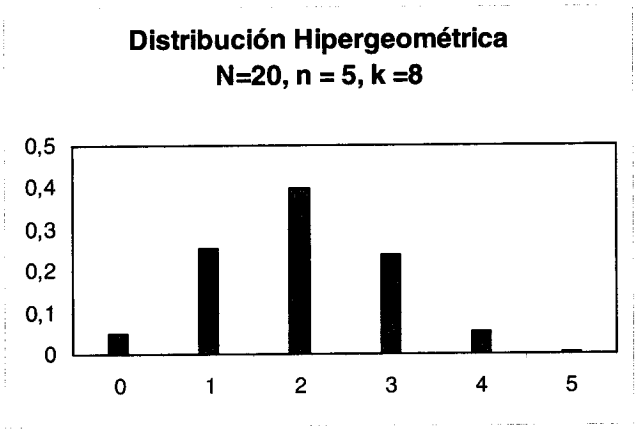
La Distribución Hipergeométrica

Se dice que una variable aleatoria discreta X, tiene distribución hipergeométrica con parámetros N, n y k (enteros positivos), si su función, de masa, de probabilidad está dada por

$$P(X=x)=P(x;N,n,k)=\begin{cases}\dfrac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \text{para } x=0,1,2,\ldots,n; x\le k; n-x\le N-k \\ 0 \text{ en cualquier otro caso}\end{cases}$$

Los tres parámetros le dan bastante versatilidad a la distribución y es casi imposible tener una idea general a cerca de su forma; sin embargo podemos graficarla en algunos casos particulares.

FIGURA No. 1
Media y varianza de una distribución hipergeométrica



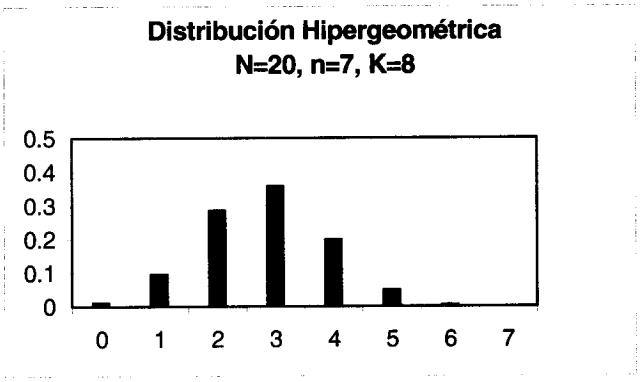
Es posible demostrar que el valor esperado y la varianza de esta distribución están dados, respectivamente, por $E(X)=\frac{nk}{N}$ y $Var(X)=\frac{nk(N-k).(N-n)}{N^2(N-1)}$; esto significa, dentro de nuestro problema, que cada lote tiene, en promedio, $\frac{400*100}{10000}$ = 4 defectuosos; además, la desviación estándar es aproximadamente 1.24.

Los resultados que se desprenden de la resolución de problemas matemáticos son, comúnmente, más útiles cuando se generalizan. Esto permite que se abarque una gama más amplia de aplicaciones.

Función de distribución de la hipergeométrica

Como una herramienta en el cálculo de probabilidades, la función de distribución presta una ayuda muy importante en el momento de las aplicaciones más o menos complejas. En nuestro caso, la función de probabilidad acumulada, como también se le llama, está dada por $F(x)=P(X\leq x)=\sum_{i=0}^{x}\frac{\binom{k}{i}\binom{N-k}{n-i}}{\binom{N}{n}}$; aunque existen tablas para esa función, podemos calcular estas probabilidades en muy poco tiempo con la ayuda de una hoja electrónica.

FIGURA No. 2



Bueno, apliquemos esta distribución.

Aplicaciones a problemas de calidad: muestreo de aceptación.

Problemas de calidad

Volvamos al problema de los bombillos. Si se tiene una producción de 10000 bombillos que se empacan en lotes de 400 y se sabe que el 1% de los 10000 es defectuoso, ¿cuál es la probabilidad de que en un lote haya 1, 2, 3 ó 4 defectuosos?

Si usamos la función de distribución de una distribución hipergeométrica con parámetros N =10000, n = 400, y k = 4, podemos calcular $P(X\leq 4)$ y con la función de masa calculamos $P(X = 0)$; así, la probabilidad buscada es $P(X\leq 4)-P(X=0) = 0.612380$.

El problema del rechazo del lote

Supóngase que el inspector de calidad de uno de los clientes somete los lotes de bombillos que se le venden a la siguiente prueba: Extrae al azar una muestra de 10 bombillos del lote de 400 y los prueba (sin reemplazo). Si ninguno de los diez es defectuoso, acepta el lote, pero si aparecen uno o más defectuosos, el lote se devuelve. ¿Cuál es la probabilidad de que uno de los lotes que tienen 4 defectuosos pase la prueba? ¿Cuál es la probabilidad de que uno cualquiera de sus lotes, fijo, pase la prueba?

Resolver la primera pregunta equivale a encontrar la probabilidad de que haya c=0 defectuosos en una muestra de 10 extraída del lote de 400, o sea P(X=0), cuando x tiene una distribución hipergeométrica con parámetros N=400, n=10 y k=4, y esta probabilidad es $\frac{\binom{4}{0}\binom{390}{10}}{\binom{400}{10}} \approx 0.903338$, que es una probabilidad bastante alta.

Resolver la segunda pregunta es algo más complicado y requiere el concepto de Curvas OC *(Operating Characteristic Curve)*.

Las Curvas OC

Son curvas que representan la probabilidad de aceptación de un lote de tamaño **N** Vs. el porcentaje de defectuosos del lote, cuando se extrae una muestra de tamaño **n** y la condición de aceptación del lote es encontrar como máximo **c** defectuosos. Naturalmente, estas curvas son de gran importancia en el control de calidad. Éstas se construyen calculando las probabilidades de aceptación con base en la distribución hipergeométrica correspondiente al proceso.

Obsérvese cómo la **figura No. 3** muestra que, bajo las condiciones de control, un lote con un 20% de defectuosos tiene una probabilidad aproximada de 0.1 de ser aceptado. Esto pudiera significar que el proceso de control no es suficientemente riguroso. Entonces deberíamos ampliar el número elementos que conforman la muestra del lote y construir otras curvas y elegir el proceso que corresponde a la curva más adecuada.

FIGURA No. 3

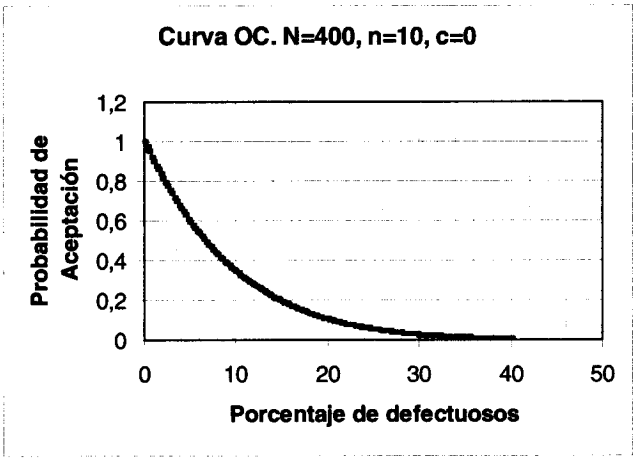
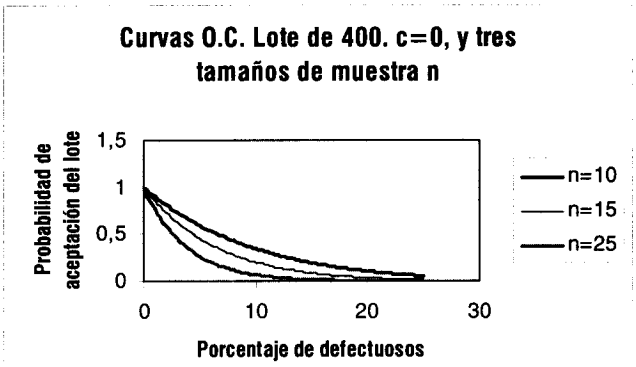


FIGURA No. 4



En la **figura No. 4** se puede observar cómo, al aumentar el tamaño de la muestra, manteniendo el límite de c=0 defectuosos para aceptar el lote, disminuye ostensiblemente la probabilidad de aceptar un lote con un 20% de defectuosos hasta aproximadamente 0.003 con n=25.

Obviamente, a mayor número de artículos examinados, mayor la precisión, pero también mayor el costo del muestreo. Depende de la administración de la empresa compradora fijar las políticas para buscar un equilibrio entre estos dos aspectos más o menos contradictorios. Si se conoce el valor de muestrear según el número de artículos se podrán aplicar técnicas de optimización para buscar el número más adecuado que no exceda un costo fijado de antemano por la empresa.

Ahora, si por algún acuerdo entre el productor y el cliente los costos de la devolución del lote corren por cuenta del productor, este debe diseñar estrategias para que sus lotes tengan la máxima probabilidad de ser aceptados por el comprador dentro de los márgenes de los porcentajes de producción de defectuosos y los costos que acarrea el rechazo del lote. La conjunción de estrategias de productor y comprador puede llevar a un problema propio de la Teoría de los Juegos en su versión probabilística.

Un momento, y ¿cómo calculo eso?

Si el lector atento buscó en una tabla de la distribución hipergeométrica, lo más probable es que haya encontrado sólo hasta el valor N=20. Afortunadamente hoy en día podemos hacer estos cálculos mediante las funciones estadísticas y probabilísticas que traen las hojas electrónicas; los cálculos necesarios y las gráficas de la distribución hipergeométrica de

este artículo fueron elaborados con Excel® ;sin embargo, si tratamos de hacer el cálculo de P(X=4, 10000,400, 4) para una distribución hipergeométrica, con una de éstas, con toda seguridad recibiremos un mensaje de error. Esto se debe a la necesidad de calcular factoriales de números muy grandes tales como 10000. Para hacer el cálculo de esa probabilidad se necesita entonces un paquete que permita el cálculo simbólico; el autor usó Derive®, pero si se es un programador, se puede escribir un programa, en cualquier lenguaje que, mediante el uso de listas enlazadas, permita el manejo de números enteros muy grandes.

Entonces, ¿siempre necesitaremos ese tipo de programas?

No necesariamente. Cuando la cantidad de artículos producidos es muy grande y cuando la relación entre el tamaño de la producción y el tamaño de un lote es pequeña, se puede recurrir a la distribución binomial como una buena aproximación para la hipergeométrica.

Aproximación de valores de una hipergeométrica por medio de una binomial

Si consideramos $p=\frac{k}{N}$, la proporción de artículos de la primera clase, entonces la probabilidad P(X=x) se puede escribir como $P(X=x)=P(x;N,n,p)=\frac{\binom{Np}{x}\binom{N-Np}{n-x}}{\binom{N}{n}}$; a partir de acá se puede demostrar que $\lim_{N\to\infty} P(x;N,n,p) = \frac{n!}{(n-x)!.x!}p^x(1-p)^{n-x}$, que es la probabilidad de que X tome el valor x para una variable aleatoria binomial con parámetros n y p.

Obsérvese en las expresiones para el valor esperado y la varianza de la hipergeométrica, que, cuando **N**→∞, E(x)→np y Var(X) →np(1-p). Todo esto significa que, para **N** grande y cuando el tamaño del lote, n, es pequeño en comparación a **N**, si se hace $p=\frac{k}{N}$, la probabilidad P(X=x) de una variable aleatoria hipergeométrica con parámetros N,n y k se puede aproximar con buena exactitud mediante el cálculo de P(Y=x) donde Y es una variable aleatoria binomial con parámetros n y p.

Si aplicamos esto a nuestro problema del lote, P(x;10000, 400,100) se puede aproximar mediante probabilidades de una binomial con parámetros n=400 y p= 0.01. Como las tablas de la binomial no traen un número tal alto de pruebas usamos la hoja electrónica. Usamos Excel 97® y obtuvimos los siguientes resultados para la binomial P(X=0, 400,0.01) = 0.01795055, P(X≤4, 400, 0.01)=0.62883858, con esto hacemos una estimación de la probabilidad de que haya 1, 2, 3 ó 4 defectuosos en el lote de 400 extraído de la población de 10000: P(X=1ó X=2 ó X=3 ó X=4) = 0.62883858-0.01795055= 0.6108880, resultado que comparado al obtenido mediante cálculo simbólico, tiene un error absoluto del orden de 0.002.

¿Hay más aplicaciones?

Sí, hay más aplicaciones posibles.

Ecología y estimación del número de peces en un lago: Captura - Recaptura

Supongamos que se quiere estimar el número N de peces en un lago, ¿cómo hacer esto? Dado que es imposible o impráctico intentar un conteo directo, debemos recurrir a métodos estadísticos. Uno de tales métodos es el siguiente: Primero se recolectan k (un número conocido) peces, mediante una red, se marcan y se devuelven al lago. Se espera un tiempo suficientemente largo como para garantizar que los peces marcados se distribuyeron de manera aleatoria en el lago, pero no demasiado largo de tal manera que el número de peces en el lago no cambie sustancialmente. (Obsérvese que se divide la población en dos partes: los marcados y los sin marcar; la distribución aleatoria de los peces marcados entre el conjunto de peces es lo que hace factible la representación del fenómeno mediante una distribución hipergeométrica). Se toma una muestra aleatoria de n peces y se cuenta el número x de peces marcados en la muestra. Un estimador de N es $\hat{N}=\frac{nk}{x}$; se puede demostrar que éste es un estimador máximo-probable (o máximo verosímil) para N, es decir, un valor para la población que hace máxima la probabilidad del valor observado x.

Por ejemplo, supongamos que se capturaron 30 peces en una laguna pequeña, se marcaron y después se devolvieron a la laguna y que después de un tiempo conveniente se seleccionó una muestra de 50 peces, de los cuales 3 resultaron marcados. Entonces hay aproximadamente $\hat{N}=\frac{50*30}{36}=500$ peces en la laguna.

El principal problema de este estimador es que tiene varianza infinita si el tamaño de la muestra no es mayor que N-k, lo que es natural en la mayoría de los casos. Un par de estimadores que pueden ser más apropiados y que tienen momentos finitos son $\hat{N}_1 = \frac{(n+1)(k+1)}{x+1} - 1$ y $\hat{N}_2 = \frac{(n+2)(k+2)}{x+2}$; utilizando los datos que poseemos se obtiene: $\hat{N}_1 \approx 395$ y $\hat{N}_2 \approx 332$.

¿OIGA, Y ESOS RESULTADOS TAN DIFERENTES, NO MUESTRAN MUCHA INEXACTITUD?

Naturalmente, la incertidumbre y la inexactitud van de la mano, son algo inevitable en el momento de enfrentar situaciones reales. Además sí se puede, en cierta medida, eliminar algo de incertidumbre pero a mayor costo, utilizando sistemas de muestreo más sofisticados y por tanto más costosos de aplicar. El del ejemplo es apenas un método muy sencillo.

Las discrepancias entre los resultados no deben alarmarnos, más bien éstas indican que el muestreo no es cosa sencilla y mucho menos el muestreo relativo a problemas de Ecología tal como el de poblaciones de especies en vías de extinción o el relativo a problemas sociológicos, tales como la determinación del índice el maltrato infantil o el índice de analfabetismo en una población dada.

Aunque en los medios de comunicación con frecuencia se abusa del muestreo y de las encuestas, no cualquiera puede hacer un buen muestreo y menos interpretarlo correctamente. En algunas carreras de estadística, se estudia muestreo general durante dos semestres y luego se estudia muestreo de encuestas. Debemos ser muy cautelosos con la información que nos llega a través de aquellos.

Aplicaciones a la Lingüística

Johnson[5] cita el artículo de A.S.C. Ross, ***Philológical Probability Problems*** en el "Journal of the Royal Statistical Society", Series B, 12, 19-41. donde se evalúa la cercanía entre dos lenguajes derivados del primitivo lenguaje Indo-Europeo o Proto-Indo-Europeo. De este lenguaje derivaron el sánscrito, persa, griego, latín, hitita, irlandés, antiguo eslavo y el germánico, entre muchos otros. De esa familia de idiomas derivados sobreviven el inglés, español, alemán, griego, ruso, albano, lituano, armenio, persa y el hindi, entre otros.

Naturalmente, la incertidumbre y la inexactitud van de la mano, son algo inevitable en el momento de enfrentar situaciones reales. Además sí se puede, en cierta medida, eliminar algo de incertidumbre pero a mayor costo, utilizando sistemas de muestreo más sofisticados y por tanto más costosos de aplicar.

Esa cercanía entre dos idiomas de esa clase con cierto número de raíces comunes se establece calculando la probabilidad de que los dos lenguajes compartan tales raíces por puro azar; a menor probabilidad, mayor cercanía entre los dos lenguajes; como podríamos esperar, tal cálculo se basa en la distribución hipergeométrica.

Veamos:

En el lenguaje primitivo Indo-Europeo cada palabra se puede expresar en la forma Prefijo-Raiz-Sufijo-Terminación. Con el fin de establecer una medida de la carcanía entre los idiomas derivados ese lenguaje se puede construír una tabla en la que cada columna corresponde a un idioma, y cada fila a una "raíz certificada". Entendiéndose por "raíz certificada" cualquier raíz que es común a por lo menos dos idiomas.

FIGURA No. 5

Raíces	Idioma 1	Idioma 2
Raíz1	*	
Raíz2	*	*
. . .		
Raízn		

En la **figura No. 5** se observa que la raíz número dos es una raíz certificada (pertenece al menos a los idiomas 1 y 2) y esto se señala con un "*".

Consideremos dos lenguajes cualesquiera con R raices comunes. Un criterio que se puede adoptar para evaluar la cercanía de los lenguajes es el siguiente: Si la probabilidad de que por lo menos R raices comunes se hayan presentado por puro azar es pequeña entonces consideramos que los lenguajes están relacionados.

Ahora, si al igual que en la **figura 5,** se señala con asteriscos la pertenencia de una raíz a un lenguaje, calcular la probabilidad de que se presenten **R** raíces comunes es igual a la probabilidad de que hayan **R** filas con un par de asteriscos en cada una . Para concretar, supongase que tenemos **R** raices certificadas, n_A asteriscos en la primera columna y n_B en la segunda. Consideremos dos casos:

Primer Caso: $\mathbf{N \geq n_A + n_B}$

En este caso la probabilidad de obtener **R** pares de asteriscos equivale a la probabilidad de obtener **R** bolas rojas en una muestra de tamaño n_A, extraída sin reemplazamiento, de una urna con **N** bolas, de las cuales n_B son rojas (¿puede explicar el lector atento el por qué?).

Así, la probabilidad que queríamos medir es

$$1-F(R)=1-P(X\leq R)=1-\sum_{i=0}^{R}\frac{\binom{n_B}{i}\binom{N-n_B}{n_A-i}}{\binom{N}{n_A}}$$

Segundo Caso: $\mathbf{N < n_A + N_B}$

En este caso, la probabilidad de obtener R pares de asteriscos equivale a obtener $N+R-n_A-n_B$ bolas rojas en una muestra de tamaño $N-n_A$ extraída sin reemplazo de una urna que contiene **N** bolas, de las cuales $N-n_B$ son rojas. (El lector atento deberá hacer un mayor esfuerzo para verificar la anterior aseveración). La probabilidad que necesitamos calcular es entonces

$$1-F(R)=1-P(X\leq R)=1-\sum_{i=0}^{R}\frac{\binom{N-n_B}{N+i-n_A-n_B}\binom{n_B}{n_B-i}}{\binom{N}{N-n_A}}$$

Naturalmente, existen otros criterios para juzgar la relación entre dos lenguajes provenientes de un tronco común. Escogimos el anterior con el fin de ilustrar la semejanza de estos cálculos con los de un problema relativo a esferas.

Problemas generalizados de bolas y la distribución hipergeométrica multivariada

Ya sin miedo a los problemas de bolas, consideremos el siguiente: Una urna contiene N bolas repartidas al azar. De las N bolas N_1 son de un primer color, N_2 de un segundo color y N_n son del n-ésimo color, de tal manera que $N_1+N_2+\ldots+N_n = N$. Se extraen m bolas sin reposición, $m \leq n$ y se considera la siguiente variable aleatoria: X_i es el número de bolas de color i que fueron extraídas. De la variable aleatoria n-dimensional (o vector aleatorio de n dimensiones) $\vec{X}=(X_1,X_2,\ldots,X_n)$ se dice que sigue una distribución hipergeométrica multivariada con parámetros N, m , p_1, p_2, p_3, . . ., p_n, donde $p_i=\frac{N_i}{N}$, para i = 1, 2, . . ., n. Es la proporción de bolas de color i antes de hacer la primera extracción.

Se puede demostrar fácilmente que la función, de masa, de probabilidad de esta función está dada por:

$$f(\vec{X})=P(X_1=k_1, X_2=K_2,\ldots,X_n=K_n)=\frac{\binom{N_1}{k_1}\binom{N_2}{k_2}\cdots\binom{N_n}{k_n}}{\binom{N}{m}}=$$

$$\frac{\binom{Np_1}{k_1}\binom{Np_2}{k_2}\cdots\binom{Np_n}{k_n}}{\binom{N}{m}},\quad \text{donde}\quad K_i = 0,1,2,\ldots,m,\ \text{con}\ k_1+k_2+\ldots+k_n=m.$$

¡Sí, si ya sé que me va a preguntar!, le tengo una aplicación

La junta directiva de la asociación de ciudadanos del sector E de una ciudad está formada por 30 miembros de esa comunidad que asocia a igual número de barrios, un representante por cada barrio. En ese sector, 3 barrios son de estrato 5, 7 barrios de estrato 4, 7 de estrato 3 y el resto de estrato 2. Se elige una comisión de 4 personas de la directiva, al azar, para formar una comisión para asistir a una sesión del concejo de la ciudad. ¿Cuál es la probabilidad de que en la comisión haya un representante de cada estrato?

Utilizando la distribución hipergeométrica multivariada, definimos: X_i = número de miembros del comité que pertenecen al estrato i, i = 1, 2, 3,4. Calculamos entonces la probabilidad

$$P(X_1=1, X_2=1, X_3=1, X_4=1)=\frac{\binom{3}{1}\binom{5}{1}\binom{9}{1}\binom{13}{1}}{\binom{30}{4}}= 0.06403941;$$

ésta es una probabilidad bastante baja.

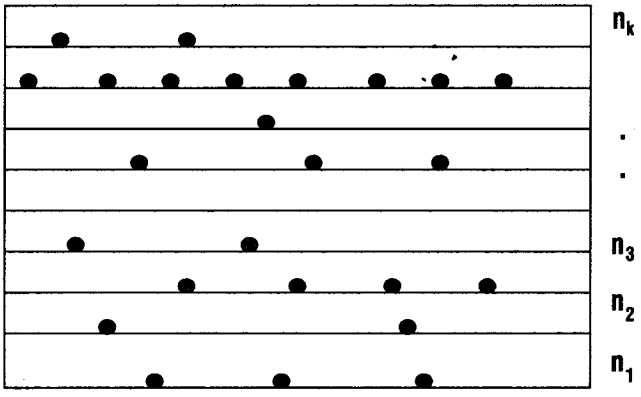
Naturalmente que ésto se puede extender, por ejemplo, a problemas de calidad de la producción cuando lo "defectuoso" se puede clasificar en varios grados, por ejemplo "levemente defectuoso", "moderadamente defectuoso", "altamente defectuoso" e "inservible", y se conocen los porcentajes de producción de cada tipo de defecto.

¿Hay otra clase de problemas que lleven a ese tipo de generalizaciones?

Claro que sí; consideremos el siguiente problema:

Distribución de n esferas en n estados: Se tienen **N** bolas del mismo tamaño y color pero numeradas, es decir, distinguibles, y se tiene una alacena con k entrepaños; en cada entrepaño se puede copiar un número fijo de bolas: n_1 en el primero, n_2 en el segundo . . ., n_k en el k-ésimo, de tal manera que $n_1+n_2+ \ldots +n_k = \mathbf{N}$ ¿de cuántas maneras es posible hacer ésto?

Figura No. 6



Solucionemos el problema de la siguiente manera: Podemos pensar que realizamos un proceso en k etapas; en la primera etapa escogemos el número de elementos por ubicar en el primer entrepaño, y esto se puede hacer de $C_{N,n1} = \binom{N}{n_1}$ maneras; en la segunda etapa, de entre los $N-n_1$ elementos que nos quedan seleccionamos los n_2 elementos para el segundo entrepaño y esto se puede realizar de $C_{N-n1,n2} = \binom{N-n_1}{n_2}$ maneras, . . . ; en la k-ésima etapa "escogemos" los elementos del k-ésimo entrepaño y eso se puede hacer de $C_{nk,nk}$ maneras. Aplicando el primer principio fundamental del conteo, encontramos que el proceso completo se puede realizar en $C_{N,n1} \cdot C_{N,n1,n2} \cdot \ldots \cdot C_{nk,nk,} =$

$$\frac{N!}{n_1!.(N-n_1)!}\cdot\frac{(N-n_1)!}{n_2!.(N-n_1-n_2)!}\cdots\frac{(N-n_1-n_2-\ldots n_{k-1})!}{.n_k!0!}=\frac{N!}{n_1!.n_2!\ldots.n_k!}$$

maneras, lo que equivale a encontrar todas las particiones posibles de un conjunto de N elementos en k subconjuntos con números de elementos respectivos $n_1, n_2, \ldots, n_k$.

Distribuciones sin ninguna restricción

Ahora, si se permite que los n_i, $i = 1,2, \ldots k$ varíen en el rango desde 0 hasta N, sin ningún tipo de exclusión, el número total de configuraciones de este tipo será N^k; dentro de estas configuraciones están aquellas en las que todas las esferas están en un mismo entrepaño.

Distribuciones con probabilidades asignadas a cada estado

Supóngase ahora que las bolas se ubican al azar en los anaqueles de tal manera que se cumpla:

1. La probabilidad de que una esfera cualquiera se encuentre en el entrepaño j es cierta cantidad q_j, fija, para $j = 1, 2, \ldots k$.

2. Las posibles ubicaciones de las N bolas son mutuamente independientes.

Bajo estos supuestos, la probabilidad de que las bolas aparezcan en una cualquiera de las configuraciones (fija) con k_j bolas en el entrepaño j ($j= 1, 2, \ldots, k$) es $q_1^{n_1}.q_2^{n_2}\ldots q_k^{n_k}$ y la probabilidad de que aparezcan en al menos una de ellas es

$$P_1=\frac{N!q_1^{n_1}.q_2^{n_2}\ldots q_k^{n_k}}{n_1!.n_2!\ldots n_k!}=N!.\prod_{j=1}^{k}\frac{q_j^{n_j}}{n_j!}.$$

Supongamos ahora que lo que realmente nos interesa para distinguir una configuración de otra es el número de bolas en cada anaquel, sin tener en cuenta cuáles bolas están en cuáles anaqueles, entonces las permutaciones que tienen igual número de esferas en cada uno de los estantes son equivalentes y las n esferas tendrán N! particiones equivalentes y, en este caso, la probabilidad de que las esferas estén en esa configuración es

$$P_1=\frac{q_1^{n_1}.q_2^{n_2}\ldots q_k^{n_k}}{n_1!.n_2!\ldots n_k!}=\prod_{j=1}^{k}\frac{q_j^{n_j}}{n_j!}.$$

Una aplicación a la Mecánica Estadística

Esta última probabilidad se conoce como estadística de Maxwell-Boltzmann y es de suprema importancia en mecánica Estadística Clásica, la rama de la física clásica que aplica los principios físicos al estudio de los sistemas aislados formados por muchísimas partículas; por ejemplo, gases confinados en un recipiente (un centímetro cúbico de un gas en condiciones normales tiene unas $3x10^{19}$ moléculas).

Acá no hablamos de estantes, anaqueles o entrepaños sino de estados de energía y no hablamos de esferas sino de partículas tales como moléculas. Además, las probabilidades g_i se llaman *probabilidades intrínsecas* de cada estado.

En esta ciencia es muy importante obtener la partición (o configuración) de máxima probabilidad porque corresponde a la configuración de equilibrio del sistema. En general se considera que, en cualquier instante, el sistema presenta leves desviaciones a partir de la configuración de equilibrio.

Teniendo en cuenta consideraciones físicas tales como la conservación de la energía y el hecho de que el número de partículas se considera fijo, se encuentra que el estimador máximo-probable (o máximo verosímil) del número de partículas en el estado i es $\hat{n}_1 = g_i e^{\alpha - \beta E_i}$, donde α y β dependen de las condiciones físicas del sistema y E_i es la energía del estado i . Haciendo $Z = \sum_i g_i e^{\alpha - \beta E_i}$ se obtiene $\hat{n}_i = \frac{N}{Z} g_i e^{- \beta E_i}$.

A esta expresión se le llama Ley de Distribución de *Maxwell-Boltzmann* y es especial para el estudio de los gases desde el punto de vista de la Mecánica Estadística clásica.

O sea que, a partir de esos problemas de bolas ¿se llega a problemas de física?

Claro que sí; además, de manera semejante a la anterior, pero teniendo en cuenta los postulados de la Mecánica Cuántica (la parte de la física que fundamenta el estudio de las moléculas, los átomos y las partículas subatómicas), se deducen las distribuciones de Bose-Einstein y de Fermi-Dirac, esenciales en la Mecánica Estadística Cuántica.

En resumen, problemas aparentemente sin uso son, muchas veces, un bosquejo elemental de procedimientos más o menos complejos de gran aplicabilidad. En vez de huirles o rechazarlos de manera arbitraria, deberíamos indagar sobre su naturaleza, sobre aquello que ocultan. No olvidemos que también en matemáticas "Las apariencias engañan".

En resumen, problemas aparentemente sin uso son, muchas veces, un bosquejo elemental de procedimientos más o menos complejos de gran aplicabilidad. En vez de huirles o rechazarlos de manera arbitraria, deberíamos indagar sobre su naturaleza, sobre aquello que ocultan. No olvidemos que también en matemáticas "*Las apariencias engañan*".

REFERENCIAS

Alonso, Marcelo y Finn, Edward. (1976.). Física. Volumen III: Fundamentos Cuánticos y Estadísticos. Edición revisada. México. Addison-Wesley Iberoamericana.

Cárcamo C., Ulises. (1999). Las probabilidades en nuestro mundo. En: Revista Universidad EAFIT, No. 115, pp. 57-70.

Fernández-Abascal, H y otros. (1994). Cálculo de probabilidades y Estadística. Barcelona. Ariel.

Grant, Eugene L. y Leavenwoth. (1994). Richard S. Control Estadístico de Calidad. México: CECSA.

Johnson, Norman L. y Kotz, Samuel. (1970). Discrete Distributions. New York. John Wiley & Sons.