Universidad Zaragoza

TRABAJO DE FIN DE GRADO

EL PROBLEMA DEL COLECCIONISTA DE CROMOS

Autor:
Javier Morales Hernández

Director:
Francisco Javier López Lorente

UNIVERSIDAD DE ZARAGOZA

FACULTAD DE CIENCIAS

JULIO 2023

Índice

1 Introducción

El problema del coleccionista de cromos es un problema clásico en probabilidad combinatoria. Comencemos describiéndolo: consideremos una persona que colecciona una cantidad finita de distintos cromos, digamos n. Estos cromos se adquieren de uno en uno y la probabilidad de obtener el cromo k es p_k. Como tenemos que obviamente es un espacio de probabilidad completo tenemos que $\sum_{i=1}^{n} p_i = 1$. El objetivo de este problema es estudiar el número de cromos que debemos comprar hasta completar la colección. Si las probabilidades p_k son iguales, nos enfrentamos a un problema relativamente sencillo, mientras que en el caso de que los cromos no sean equiprobables será más complicado pero a su vez más realista.

La historia de este problema comienza en 1708, cuando el problema apareció por primera vez en *De Mensura Sortis (On the Measurement of Chance)* escrito por A. de Moivre. Gracias a Euler y a Lagrange se obtuvieron más resultados en el caso de cromos equiprobables, cuando $p_k = \frac{1}{n}$ para todo k.

En 1954 H. Von Schelling obtuvo por primera vez la distribución del tiempo de espera para completar una colección cuando las probabilidades de cada cromo son distintas y además en 1960 D.J. Newman y L. Shepp calculó la distribución del tiempo de espera para completar 2 colecciones en el caso de cromos equiprobables.

A la hora de realizar un estudio de este problema no nos podemos quedar solo en los casos más simples sino que también hay otras preguntas que plantearnos: ¿Los cromos son equiprobables o no? ¿Queremos completar una o varias colecciones? ¿Se compran de forma individual o en sets de varios cromos?. El estudio de todos estos casos hace que sea un estudio completo del problema, sin embargo, no nos centraremos en todos los casos.

Además, el problema del coleccionista de cromos tiene muchas aplicaciones, especialmente en ingeniería eléctrica, aunque también se puede aplicar en biología o en el ámbito de las telecomunicaciones.

En este trabajo estudiaremos, una colección de cromos equiprobables en el capítulo 2, una colección de cromos no equiprobables en el capítulo 3, varias colecciones de cromos equiprobables en el capítulo 4 y por último las aplicaciones del problema en el capítulo 5.

2 Una colección de cromos equiprobables

2.1 Cálculo de la esperanza y la varianza

En este primer apartado queremos estudiar el caso en el que queremos completar una colección de n cromos equiprobables, es decir, que la probabilidad de conseguir cada cromo cuando compramos uno nuevo es $\frac{1}{n}$. Denotamos al número de cromos que necesitamos comprar para completar la colección como X. Este valor es distinto al número de cromos que hay en la colección ya que al comprarlos podremos conseguir cromos repetidos. Definamos ahora X_i como el número de cromos que debemos comprar para tener una colección con $i-1$ cromos distintos a i, es decir, que la suma de las variables $X_i, i = 1, ..., n$ es:

$$X = X_1 + X_2 + ... + X_n$$

Cuando compramos el primer cromo no tenemos ninguno, podemos asegurar que $X_1 = 1$. Además estamos estudiando el caso en el que todos los cromos tienen la misma probabilidad, que para este enfoque en concreto es $\frac{1}{n}$, por lo que en el momento en el que tenemos i cromos la probabilidad de conseguir uno nuevo es $\frac{n-i}{n}$. Las variables X_i son independientes entre ellas y siguen la ley geométrica con parámetro $\frac{n-i+1}{n}$. Podemos ahora demostrar el siguiente teorema:

Teorema 2.1.1: En el problema del coleccionista de cromos con probabilidades iguales, la esperanza y la varianza del número de cromos que hay que comprar para completar la colección, son respectivamente:

$$E[X] = n\sum_{i=1}^{n} \frac{1}{i}$$

$$Var[X] = n\sum_{i=1}^{n} \frac{i-1}{(n-i+1)^2}$$

Demostración: A partir de la escritura de X: $X = X_1 + X_2 + ... + X_n$ con $X_i \sim Geom(\frac{n-i+1}{n})$. Se tiene:

$$E[X] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n\sum_{i=1}^{n} \frac{1}{i}$$

Por último, recordando que la varianza de una geométrica de parámetro p es $\frac{1-p}{p^2}$, y debido a la independencia de las variables X_i:

$$Var(X) = Var\left(\sum_{i=1}^{n} X_i\right) = Var(X_1 + ... + X_n) = Var(X_1) + ... + Var(X_n) =$$

$$= 0 + \frac{n}{(n-1)^2} + \frac{2n}{(n-2)^2} + ... + \frac{n(n-1)}{2^2} = \sum_{i=1}^{n} \frac{(i-1)n}{(n-i+1)^2}$$

Una vez demostrado que la esperanza y la varianza toman ese valor, otra forma de calcular la esperanza es a través de cadenas de Markov, que nos serán útiles más adelante. Antes de comenzar este nuevo enfoque debemos definir algunos conceptos:

Definición 2.1.1 (Cadena de Markov): Sea E un conjunto contable. Una sucesión $(A_n)_{n\geq 0}$ de variables aleatorias con valores en E se llama cadena de Markov si para todo $n \geq 2$, $i_0, ..., i_n \in E$ se verifica:

$$P\{A_n = i_n | A_0 = i_0, ..., A_{n-1} = i_{i-1}\} = P\{A_n = i_n | A_{n-1} = i_{n-1}\}$$

El conjunto E se llama espacio de estados de la cadena. Si E es finito se dice que la cadena es finita.

Definición 2.1.2: Una cadena de Markov $(A_n)_{n\geq 0}$ se llama homogénea si para todo $n \geq 0, i, j \in E$, se verifica $P\{A_{n+1} = j | A_n = i\} = P\{A_1 = j | A_0 = i\}$.

Definición 2.1.3: Sea $(A_n)_{n\geq 0}$ una cadena de Markov homogénea. Se llaman probabilidades de transición de la cadena a $p_{ij} = P\{A_1 = j | A_0 = i\}$ y matriz de transición de la cadena a $\mathbf{P} = (p_{ij})_{i,j\in E}$. Para $n \geq 0$, se llaman probabilidades de transición en n etapas de la cadena a $p_{ij}^{(n)} = P\{A_n = j | A_0 = i\}$ y matriz de transición en n etapas de la cadena a $\mathbf{P}^{(n)} = (p_{ij}^{(n)})_{i,j\in E}$

Ahora que hemos definido estos conceptos veamos cómo la esperanza del número de cromos que debemos comprar tiene el valor del Teorema 2.1.1 pero desde otro enfoque, en este caso, las cadenas de Markov. A diferencia del caso anterior, suponemos que cada cromo llega en una unidad de tiempo, por lo que las variables X_i pueden interpretarse como el número de unidades de tiempo que debemos esperar para pasar de tener $i-1$ cromos distintos a i cromos. Ahora definimos Y_n como el número de cromos que tendremos después de que pasen n unidades de tiempo. La probabilidad de conseguir cualquier tipo de cromo en cualquier momento es $p = \frac{1}{n}$. Cada Y_{n-1} depende de los anteriores, es decir, que el número de cromos después de n unidades de tiempo depende de los cromos que tengamos después de $n-1$ unidades de tiempo o escrito de otra forma:

$$P\{Y_n = i_n | Y_0 = i_0, ..., Y_{n-1} = i_{i-1}\} = P\{Y_n = i_n | Y_{n-1} = i_{n-1}\}$$

por lo tanto las variables Y_n son una cadena de Markov con el espacio de estados $S = \{0, 1, ..., n\}$.
Con la definición de matriz de transición sabemos que $p_{i,j} = P\{X_{n+1} = i | X_n = j\}$. Por lo tanto si $j > i$ la probabilidad es cero y por consiguiente sabemos que la matriz de transición es triangular. Es más si $i + 2 > j$ la probabilidad también es cero por la definición de Y_n ya que si aumenta una unidad el valor de n solo aumenta un cromo. El resto de probabilidades son obvias, por lo que llegamos a la siguiente matriz de transición:

$$\begin{pmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & \frac{1}{n} & \frac{n-1}{n} & 0 & \cdots & 0 \\ 0 & 0 & \frac{2}{n} & \frac{n-2}{n} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \frac{n-1}{n} & \frac{1}{n} \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

Esta cadena de Markov tiene un estado absorbente y nuestro objetivo es conocer el número medio de transiciones hasta llegar al estado deseado partiendo del estado 0. Para ello tenemos:

$$\begin{cases} k_n = 0 \\ k_i = 1 + \sum_{j \neq n} p_{ij} k_j, i \neq n \end{cases}$$

Y resolviendo este sistema llegamos a:

$$k_0 = n \sum_{i=1}^{n} \frac{1}{i}$$

que coincide con la expresión del Teorema 2.1.1

2.2 Función de masa de probabilidad de X

Una vez calculadas la media y la varianza de X podemos preguntarnos por su función de masa de probabilidad. Es decir, la probabilidad de que nos haga falta comprar exactamente K cromos con $K \geq n$ para completar la colección. Para hacer un estudio de la función de masa de probabilidad primero tenemos que definir los números de stirling de segundo tipo.

2.2.1 Definición: Sea $s_{n,k}$ denota el número de particiones de un espacio de n elementos en k bloques. Podemos usar la siguiente fórmula de recurrencia:

$$s_{n,k} = s_{n-1,k-1} + k s_{n-1,k}$$

para $n, k \geq 1$ y con condiciones iniciales $s_{n,0} = s_{0,k} = 0$ y $s_{0,0}$.

Volvamos ahora a nuestro problema. Hay n^K formas en las que podría resultar una secuencia de K intentos. El número de formas en que un cromo puede aparecer en el K-ésimo intento, y que sea n-ésimo cromo, es $(n-1)!s_{K-1,n-1}$. Como tenemos n cromos que pueden aparecer por primera vez en la última compra, tenemos $n(n-1)!s_{K-1,n-1} = d!s_{K-1,n-1}$ formas en las que necesitamos exactamente K intentos para tener los n cromos. Y así:

$$P(X = K) = \frac{n!sK-1, n-1}{n^K} = \frac{(n-1)!s_{K-1,n-1}}{n^{K-1}}$$

2.3 Distribución asintótica de $X(n)$

En el apartado anterior hemos visto la distribución del número de cromos necesarios para completar la colección y podría ser interesante ver su comportamiento cuando n tiende a infinito. Para este apartado la variable X suma de variables independientes la reescribiremos como $X(n)$ para una colección de n cromos.

Teorema 2.3.1: Si $X(n)$ es el número de cromos necesarios para completar una colección de n cromos equiprobables, se tiene que $\frac{X_(n)-nlog(n)}{n}$ converge en distribución a una distribución de Gumbel con $\mu = 0$ y $\beta = 1$. Es decir:

$$\lim_{n\to+\infty} P(\frac{X_n - nlog(n)}{n} < x) = e^{-e^{-x}}$$

Demostración: Comencemos denotando $v(n,k)$ el número de cromos que son necesarios para que exactamente tengamos al menos k cromos distintos en una colección de n cromos.
Entendamos ahora la definición de $X(n)$. $X(n)$ como el número de cromos que necesitamos comprar para tener al menos 1 cromo de cada tipo, es decir, los cromos que necesitamos comprar para completar una colección. Lo denotamos así con el objetivo de que nos facilite la expresión en capítulos posteriores.
Por definición $v(n,1) = 1$, $v(n,n) = X(n)$ y podemos reformular las variables X_k como $X_1 = 1$, $X_k = v(n,k) - v(n,k-1)$ para todo $k = 2,3,...,n$. Recordemos que las variables X_k son variables aleatorias independientes con ley geométrica de parámetros $\frac{n}{n-k+1}$.

Tomamos ahora la variable aleatoria:

$$\eta_n = \frac{X(n) - n\sum_{k=1}^{n}\frac{1}{k}}{n} = \frac{\sum_{i=2}^{n} X_i - n\sum_{k=1}^{n}\frac{1}{k}}{n}$$

Antes de seguir, recordemos la definición de función característica:

Definición 2.3.1: Dada una variable aleatoria continua X su función característica es una función $\psi_X : \mathbb{R} \to \mathbb{C}$ definida como

$$\psi_X(t) = E[e^{itX}] = \int_{-\infty}^{+\infty} e^{itx} f_X(x)dx$$

Ahora que la hemos definido, podemos estudiar darle valor a la función característica de la función η_n y es:

$$\psi_n(t) = \frac{1}{\prod_{h=1}^{n} e^{\frac{it}{h}}(1 + \frac{n}{h}(e^{-\frac{it}{n}} - 1))}$$

Si tomamos el límite de cuando n tiende a infinito de esta función característica converge a:

$$\lim_{n\to+\infty} \psi_n(t) = \Gamma(1 - it)e^{-itc}$$

con Γ la función Gamma y con c la constante de Euler. Por la representación integral de la función gamma tenemos:

$$\int_{-\infty}^{+\infty} e^{itx} dF(x) = \Gamma(1-it)$$

con $F(x) = e^{-e^{-x}}$. Así que tenemos que:

$$\lim_{n\to+\infty} P(\frac{v_1(n) - n(1+\frac{1}{2}+...+\frac{1}{n}-c)}{x} < x) = F(x)$$

Por último se sabe que:

$$\sum_{k=1}^{n} \frac{1}{k} = log(n) + c + o(1)$$

Y así obtenemos el resultado esperado y queda acabada la demostración.

Este resultado puede escribirse de forma equivalente usando la distribución exponencial o la chi-cuadrado de dos grados de libertad.

Corolario 2.3.1: En las condiciones del teorema anterior tenemos que $e^{-\frac{X(n)}{n} - log(2n)}$ converge en distribución a una exponencial de parámetro $\frac{1}{2}$ *Demostración:* Es casi inmediato, simplemente debemos destacar que para $t > 0$:

$$P(e^{-(\frac{X(n)}{n} - log(2n))} \leq t) = P(-(\frac{X(n)}{n} - log(2n)) \leq log(t)) =$$

$$= P(\frac{X(n)}{n} - log(n) \geq log(\frac{2}{t})) \to 1 - e^{-e^{-log(\frac{2}{t})}} = 1 - e^{-\frac{t}{2}}$$

Por último solo hace falta comparar las funciones de distribución de una chi-cuadrado de 2 grados de libertad con una exponencial de parámetro $\frac{1}{2}$ para ver que son iguales.

Para finalizar este capítulo estudiemos algunos resultados finales. Definimos las variables W_n como el tiempo que debemos esperar hasta tener $a_n + 1$ cromos con $0 \leq a_n < n$. Estudiemos las distribuciones asintóticas de las variables W_n dependiendo del comportamiento de a_n. Definimos por último $b_n = n - a_n$ y sea μ_n y σ_n^2 la media y la varianza de W_n.

Teorema 2.3.2 : Si $\frac{a_n}{n^{\frac{1}{2}}}$ converge a 0, se tiene que, $W_n - a_n - 1$ converge en probabilidad a 0.

Demostración: Comencemos redefiniendo W_n como:

$$W_n = Z_{\frac{n}{n}} + Z_{\frac{n-1}{n}} + ...Z_{\frac{b_n}{n}}$$

Pero si pensamos la definición de estas "nuevas" variables, nos damos cuenta de que $Z_{\frac{n}{n}} = X_1$ definida en el primer apartado. Así sucesivamente reescribimos $Z_{\frac{n-1}{n}} = X_2$, ... ,$Z_{\frac{b_n}{n}} = X_{a_n+1}$, por lo que lo podemos reescribir como:

$$W_n = X_1 + X_2 + ... + X_{a_n+1}$$

Obviamente las variables X_i tendrán las medias y varianzas como las halladas en el primer apartado, pero adaptadas a las definiciones de a_n y b_n y se quedan de la siguiente forma:

$$\mu_n = n \sum_{k=b_n}^{n} \frac{1}{k}$$

$$\sigma_n^2 = n \sum_{k=b_n}^{n} \frac{n-k}{k^2}$$

Notemos que en la varianza sustituyendo k por $n-i+1$ obtenemos la varianza hallada en el primer apartado.

Debido a esto, la función característica de $W_n - a_n$ es:

$$\prod_{k=0}^{a_n} \frac{1-\frac{k}{n}}{1-e^{it\frac{k}{n}}}$$

Usando la aproximación de $1 + z = e^{z+z^2\theta}$ con $|z| \leq \frac{1}{2}$ y $|\theta| \leq 1$ en el numerador y denominador obtenemos lo siguiente:

$$1 + (-\frac{k}{n}) = e^{-\frac{k}{n}+\theta\frac{k^2}{n^2}}$$

$$1 + (-e^{it}\frac{k}{n}) = e^{-e^{it}\frac{k}{n}+\theta(-e^{it}\frac{k}{n})^2}$$

Así, el cociente dentro del producto anterior se puede escribir como:

$$e^{-\frac{k}{n}+\theta(\frac{k}{n})^2+e^{it}\frac{k}{n}-\theta(e^{it}\frac{k}{n})^2}$$

Como es un producto de exponenciales lo podemos reescribir como la exponencial del sumatorio, por lo que llegamos a la siguiente fórmula:

$$exp[\sum_{k=0}^{a_n} \frac{k}{n}(e^{it}-1) + 2\theta\frac{k^2}{n^2}]$$

Recordamos ahora a qué converge el siguiente sumatorio:

$$\sum_{k=0}^{a_n} \frac{k}{n} = \frac{a_n(a_n+1)}{2n} \to \frac{\lambda^2}{2}$$

En este caso en particular $\lambda = 0$, entonces, $\sum_{k=0}^{a_n} \frac{k}{n} \to 0$. Además $\sum_{k=0}^{a_n} \frac{k^2}{n^2} \to 0$. Es decir que si $\lambda = 0$, entonces la función característica de $W_n - a_n - 1$ converge

a 0.

Teorema 2.3.3 : Si $\frac{a_n}{n^{\frac{1}{2}}}$ converge a una constante positiva λ, por lo tanto, $W_n - a_n - 1$ converge en distribución a una distribución de Poisson con media $\frac{\lambda^2}{2}$

Demostración: Razonando exactamente de la misma manera tenemos que bajo esta hipótesis $\lambda > 0$ y entonces la función característica de $W_n - a_n - 1$ converge a $exp[\frac{1}{2}\lambda^2(e^{it} - 1)]$ que es la de Poisson de media $\frac{\lambda^2}{2}$.

Por último enunciaremos el último teorema pero sin demostración ya que es mucho más compleja.

Teorema 2.3.4: Si $\frac{a_n}{n^{\frac{1}{2}}}$ y b_n convergen a infinito, por lo tanto, $\frac{X_n - \mu_n}{\sigma_n}$ converge en distribución a una normal con media 0 y varianza 1.

La demostración de este teorema es muy larga y la dejamos a elección del lector, se puede encontrar en [1]

3 Varias colecciones de cromos equiprobables

3.1 El problema de las n urnas

Consideremos ahora el siguiente caso: en una familia de m hermanos quieren completar m colecciones. Cuando se consigue un cromo nuevo el hermano mayor lo recibe y se lo guarda, cuando conseguimos por segunda vez el cromo se lo guarda el segundo hermano de mayor edad y así sucesivamente. Ahora nos volvemos a plantear las mismas preguntas ¿Cuál es el número esperado de cromos que debemos comprar para completar las m colecciones? ¿Qué distribución sigue esta variable?

El problema del coleccionista de cromos es equivalente al problema de las urnas. Este problema consiste en que tenemos n urnas, que serían los distintos tipos de cromos y vamos lanzando pelotas, que siempre entran en alguna urna pero de forma completamente aleatoria. El acto de lanzar una pelota es equivalente a comprar un cromo, por lo que efectivamente los problemas son equivalentes.

Nos planteamos ahora la siguiente pregunta: ¿Cuántos cromos debemos comprar para completar m colecciones? O equivalentemente, ¿cuántas pelotas debemos lanzar para tener al menos m pelotas en cada una de las n urnas?.

3.2 Cálculo de la esperanza

Denotamos $v_m(n)$ como el número de pelotas que debemos lanzar para tener al menos m pelotas en cada una de las n urnas, por eso en el estudio anterior, como era para una colección, lo hemos denotado como $v_1(n)$. Veamos una aproximación de la media de la variable $v_m(n)$. Comencemos recordando la aproximación del valor de la media que es $\sum_{i=1}^{n} \frac{1}{i} = log(n) + c + \frac{1}{2n} + O(\frac{1}{n^2})$ con c la constante de Euler.

Enunciemos ahora el siguiente teorema:

Teorema 3.2.1: En el problema del coleccionista de cromos con probabilidades iguales y m colecciones, la esperanza del número de cromos que hay que comprar para completar las colecciones es:

$$E[v_m(n)] = nlog(n) + (m-1)nlog(log(n)) + nC_m + O(n)$$

con C_m constante que depende de m, que toma el valor $C_m = c - log(m-1)!$, resultado que no demostraremos debido a su dificultad en cuanto a cómputos. Antes de realizar la demostración del valor de la esperanza debemos enunciar y demostrar un lema:

Lema 3.2.1: Demostrar que el valor de la esperanza es el anterior es equivalente a probar que:

$$\frac{E[v_m(n+1)]}{n+1} - \frac{E[v_m(n)]}{n} = \frac{1}{n+1} + \frac{m-1}{nlog(n)} + \lambda_n$$

con $\sum_n |\lambda_n| < +\infty$.

Demostración: Lo que tenemos que probar es que

$$\frac{E[v_m(n)]}{n} - log(n) - (m-1)log(log(n)) = C_m + o(1)$$

lo que es equivalente a ver que el límite de la primera parte de la igualdad cuando n tiende a infinito es real. Comencemos llamando $\alpha_n = \frac{E[v_m(n)]}{n}$, y tenemos que probar si $\alpha_{n+1} - \alpha_n = \frac{1}{n+1} + \frac{m-1}{nlog(n)} + \lambda_n$, entonces existe $\lim_{n\to\infty}[\alpha_n - log(n) - (m-1)log(log(n))]$ y que pertenece a los reales.
Notemos que $\alpha_n = \alpha_n - \alpha_{n-1} + \alpha_{n-1} - ... + \alpha_2 - \alpha_1$ por lo que α_n se puede expresar como

$$\alpha_n = \sum_{k=2}^{n} \left(\frac{1}{k} + \frac{m-1}{klog(k)} + \lambda_k \right)$$

Por lo que:

$$\alpha_n - log(n) - (m-1)log(log(n)) =$$
$$= \left(\sum_{k=2}^{n} \frac{1}{n} - log(n)) + (m-1) \left(\sum_{k=2}^{n} \frac{1}{klog(k)} - log(log(n) \right) \right) + \sum_{k=2}^{n} \lambda_k$$

Y obviamente si las dos series entre paréntesis convergen a valores reales y el sumatorio de λ_n es una serie absolutamente convergente, entonces existe el límite y es real.

Demostración (Teorema 3.2.1): Como son resultados equivalentes demostremos que:

$$\frac{E[v_m(n+1)]}{n+1} - \frac{E[v_m(n)]}{n} = \frac{1}{n+1} + \frac{m-1}{nlog(n)} + \lambda_n$$

Podemos escribir esta resta como:

$$\frac{E[v_m(n+1)]}{n+1} - \frac{E[v_m(n)]}{n} = \int_0^{+\infty} e^{-t} S_m(t)[1 - e^{-t} S_m(t)]^n dt$$

con $S_m(t) = \sum_{k<m} \frac{t^k}{k!}$. Esta igualdad proviene de $E[v_m(n)] = n \int_0^{+\infty} [1 - (1 - S_m(t)e^{-t})^n] dt$ puede verse la explicación en [2]. Haciendo ahora el cambio de variable $x = 1 - e^{-t} S_m(t)$ llegamos a que $\int_0^1 x^n S_m(t) \frac{(m-1)!}{t^{m-1}} dx$ con obviamente $t = t(x)$. Queremos ahora probar que:

$$\sum_{n=1}^{\infty} \int_0^1 \frac{x^n}{t^k} dx < \infty, k < m$$

$$\int_0^1 \frac{x^n}{t}dx = \frac{1}{nlog(n)} + \alpha n$$

con $\sum \alpha n < \infty$ y así quedaría demostrado el teorema.
Por el cambio de variable es obvio que $t \geq log(\frac{1}{1-x})$ además desarrollando $S_m(t)$ tenemos que $x \leq t^m$. Con estas dos desigualdades podemos ver que la siguiente integral es finita:

$$\int_0^1 \frac{1}{1-x}\frac{dx}{t^k} = \int_0^{\frac{1}{2}} \frac{1}{1-x}\frac{dx}{t^k} + \int_{\frac{1}{2}}^1 \frac{1}{1-x}\frac{dx}{t^k}$$

Y queda así demostrado el primer punto. Con el desarrollo del logaritmo y la primera desigualdad llegamos a que $t \geq x^r log(r)$ y, por lo tanto, $\int_0^1 \frac{x^n}{t}dx \leq \frac{1}{(n-r+1)log(r)}$. Tomando un valor $u \geq 1$ definimos $a = 1 - S_m(u)e^{-u}$ tenemos que:

$$\int_0^1 \frac{x^n}{t}dx \geq \frac{1}{u}\int_0^a x^n dx \geq \frac{1}{(n+1)u} - \frac{S_m(u)e^{-u}}{u}$$

Y llegamos a que:

$$\int_0^1 \frac{x^n}{t}dx \geq \frac{1}{(n+1)u} - u^{m-2}e^{1-u}$$

Tomando ahora $r = nlog(n)$ y $u = log(n) + mlog(log(n))$ obtenemos que:

$$\frac{1}{nlog(n)} - \frac{Clog(log(n))}{nlog^2(n)} \leq \int_0^1 \frac{x^n}{t}dx \leq \frac{1}{nlog(n)} + \frac{Clog(log(n))}{nlog^2(n)}$$

y así queda acabada la demostración si $\sum \frac{log(log(n))}{nlog^2(n)}$.

Al igual que en el caso equiprobable, veamos ahora el enfoque de las cadenas de Markov para $m = 2$. Sean $Y_n, n > 0$ las variables definidas en el anterior enfoque de cadena de Markov. Sin embargo, ahora tomamos el espacio de estados $S' = \{(i,j) : i,j \in \{0,1,...,n\}, i \geq j\}$ con $|S| = \frac{(n+1)(n+2)}{2}$, el estado (i,j) define el momento en el que tenemos i cromos distintos de la primera colección y j de la segunda colección, por eso es obvio que $i \geq j$. Entonces las probabilidades de transición vienen dadas como $(0,0) \to (1,0)$ con probabilidad 1, $(i,j) \to (i+1,j)$ con probabilidad $\frac{n-i}{n}$, $(i,j) \to (i,j+1)$ con probabilidad $\frac{i-j}{n}$ y $(i,j) \to (i,j)$ con probabilidad $\frac{j}{n}$. Por último tenemos que obviamente $(n,n) \to (n,n)$ con probabilidad 1. Ademas en el caso de que $i = n$ tenemos que las probabilidades de transición son $(n,j) \to (n,j+1)$ con probabilidad $\frac{n-j}{n}$ y $(n,j) \to (n,j)$ con probabilidad $\frac{j}{n}$.

Para conocer el tiempo de espera para completar las colecciones debemos calcular el valor de $k_{(0,0)}^{(n,n)}$ que se hace resolviendo el sistema lineal como en el caso visto anteriormente.

3.3 Distribución asintótica de $v_m(n)$

Del valor de la esperanza podemos sacar n factor común y a partir de ahí podemos estudiar la convergencia asintótica de la distribución de probabilidad, es decir, la convergencia en distribución.

En el apartado anterior hemos calculado una expresión asintótica para la esperanza del número de cromos necesarios para completar m colecciones. El siguiente teorema, que enunciamos sin demostrar da la convergencia en distribución de $v_m(n)$. Notar que para $m = 1$ recuperamos el Teorema 2.3.1.

Teorema 3.2.1: Si $v_m(n)$ es el número de cromos necesarios para completar m colecciones de n cromos, entonces se cumple la siguiente fórmula, para todo x real:

$$\lim_{n\to+\infty} P(\frac{v_m(n)}{n} < log(n) + (m-1)log(log(n)) + x) = exp(-\frac{e^{-x}}{(m-1)!})$$

De todas formas en caso de que se quiera ver una demostración completa del Teorema 3.2.1 se puede leer en [3]

4 Una colección de cromos no equiprobables

4.1 Introducción al problema

Una vez estudiado el caso de que todos los cromos tengan la misma probabilidad, estudiemos un caso algo más realista, una colección donde no todos los cromos tienen la misma probabilidad. Normalmente en las colecciones de cromos existen distintas categorías como común, raro, ...

Definimos así p_i como la probabilidad de conseguir el cromo i-ésimo cada vez que compramos uno. Al tratarse de un espacio de probabilidad completo obviamente tenemos que $\sum_{i=1}^{n} p_i = 1$ con $p_i > 0$.

Para este nuevo capítulo definimos los T_i como el número de cromos que debemos comprar hasta conseguir el cromo de tipo i-ésimo. Cada vez que compramos un cromo tenemos una posibilidad de p_i de conseguir el cromo i-ésimo y una posibilidad de $1 - p_i$ de conseguir cualquier otro, es decir, de no conseguirlo. Con este argumento deducimos que las variables T_i siguen una ley geométrica de parametro p_i.

Definimos ahora la variable X como el número total de cromos que debemos comprar para completar la colección. A diferencia del anterior capítulo por definición, X no va a ser la suma de todos los anteriores T_i sino que seguirá la siguiente fórmula:

$$X = \max(T_1, T_2, ..., T_n)$$

4.2 Cálculo de la esperanza

Como el valor que nos interesa estudiar es la esperanza de la variable X y esta se define con la función de máximo, lo primero que se nos ocurre es usar la identidad de mínimo-máximo.

Definición 4.2.1: La identidad de mínimo-máximo es una relación entre el máximo de un conjunto S de n elementos y el mínimo de los $2^n - 1$ subespacios no vacíos de S. Tomando $S = \{x_1, x_2, ..., x_n\}$ la identidad dice:

$$\max(x_1, x_2, ..., x_n) = \sum_{i=1}^{n} x_i - \sum_{i<j} \min(x_i, x_j) + \sum_{i<j<k} \min(x_i, x_j, x_k) - ...$$

$$... + (-1)^{n+1} \min(x_1, x_2, ..., x_n)$$

Ahora que sabemos la existencia de esta identidad solo nos hace falta saber el valor de las funciones mínimo para las variables T_i. Pensemos en las definiciones de estas variables, son el número de cromos que debemos comprar hasta conseguir el cromo i-ésimo por lo que el mínimo entre dos variables es el número de cromos que debemos comprar hasta conseguir uno de los dos tipos de cromos, por lo que

la distribución del mínimo de las dos variables es una geométrica con parámetro $p_i + p_j$. Continuemos calculando el $\min(T_i, T_j, T_k)$, que deducimos que también sigue una ley geométrica con parámetro $p_i + p_j + p_k$ y así sucesivamente hasta que llegamos a que $\min(T_1, T_2, ..., T_n)$ es una geométrica con parámetro $\sum_{i=1}^{n} p_i$. Ahora que ya tenemos todos los valores calculados podemos desarrollar la esperanza de X:

$$E[X] = E[\max_{i=1,2,..,n} X_i] = \sum_{i=1}^{n} E[T_i] - \sum_{i<j} E[\min(T_i, T_j)] +$$

$$+ \sum_{i<j<k} E[\min(T_i, T_j, T_k)] - ... + (-1)^{n+1} E(\min(T_1, T_2, ..., T_n)) =$$

$$= \sum_i \frac{1}{p_i} - \sum_{i<j} \frac{1}{p_i + p_j} + ... + (-1)^{n+1} \frac{1}{p_1 + ... + p_n}$$

Como $\int_0^{+\infty} e^{-px} dx = \frac{1}{p}$ la expresión anterior se puede escribir como:

$$E[X] = \int_0^{+\infty} (\sum_i e^{-p_i x} - \sum_{i<j} e^{-(p_i+p_j)x} ... + (-1)^{n+1} e^{-(p_1+...+p_n)x}) dx$$

y utilizando la igualdad:

$$1 - \prod_{i=1}^{n} (1 - e^{-p_i x}) = \sum_i e^{p_i x} - \sum_{i<j} e^{-(p_i+p_j)x} + ... + (-1)^{n+1} e^{-(p_1+...+p_n)x}$$

concluimos que el valor de la esperanza es:

$$E[X] = \int_0^{+\infty} (1 - \prod_{i=1}^{n} (1 - e^{-p_i x})) dx$$

5 Aplicaciones

5.1 Problema del rastreo de la IP

Esta primera aplicación se encuentra en el ámbito de las telecomunicaciones. Es comúnmente sabido que los ataques DoS (Denial of Service) son los ataques de seguridad más complicados en el ámbito de la ciberseguridad, y es aún más complicado saber de dónde, o quién provoca este ataque. Este problema, el de determinar quién provoca el ataque, se conoce como el problema del rastreo de la IP. Se propuso una prometedora solución para este problema, llamada el PPM (probabilistic packet marking), o en español, marcado probabilístico de paquetes. La idea de esta solución es escoger de forma probabilística información parcial de cada paquete de la ruta del atacante. A pesar de que cada paquete representa solo información parcial de la ruta de ataque, una víctima puede construir la ruta completa combinando la información obtenida de un número modesto de los paquetes escogidos.

En un esquema PPM, la cantidad de paquetes que la víctima debe recibir para reconstruir la ruta del ataque es equivalente a la cantidad de cromos que necesitamos comprar para completar un set de cromos en el problema del coleccionista de cromos. A la cantidad de cromos que necesitamos comprar para completar el set que queremos, en el problema del rastreo de la IP se denomina coste de detección, por lo tanto, analizar la eficencia del esquema PPM se reduce a resolver nuestro problema. En particular, la proporción de falsos negativos de un esquema PPM está dada por la función de supervivencia del coste de detección, por lo tanto, es muy importante calcular esta función a la hora de evaluar la eficiencia de estos esquemas.

El objetivo a la hora de aplicar los resultados de nuestro problema es calcular la función de supervivencia del coste de detección con el menor tiempo de cómputo posible pero con una precisión suficiente para poder asegurar su eficacia en la práctica. Por eso mismo, es buena idea estudiar los límites superiores e inferiores de la función de distribución complementaria.

Más concretamente estudiemos que el ataque DoS proviene de un atacante con una sola fuente y que entre el atacante y la víctima existen n enrutadores, es decir, n pasos. Denominamos como enlace i al enlace entre el enrutador $i-1$ e i. Cada paquete consta de 64 bits de información supongamos que tienen dos paquetes de 32 bits. Cuando un enrutador marca un paquete, escribe su dirección IP en uno de los campos, que llamaremos campo fuente. El siguiente enrutador, escribe su dirección IP en el mismo campo, pero lo denominará campo de destino. De esta forma conseguimos que no halla dos direcciones IP en el mismo paquete de información y se evitan malos entendidos. La probabilidad de que el enlace i marque un paquete recibido por la víctima es p_i, donde $p_i = p(1-p)^{i-1}$ con p la probabilidad de que un enrutador decida marcar el paquete.

Se puede demostrar que una cota inferior es:

$$P(X > k) \geq \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} (1 - i\mathbf{p})^k$$

con $\mathbf{p} = \frac{1}{n} \sum_{i=1}^{n} p_i$ y una cota superior:

$$P(X > k) \leq \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} (1 - ip_{min})^k$$

con $p_{min} = \min\{p_1, ..., p_n$

Estudiemos ahora la proporción de falsos negativos de un ataque de fuente única. Definimos $p(k : l) = (p_1(l), ..., p_1(l), ..., p_n(l), ..., p_n(l))$ cada uno k veces con $p_i(l) = \frac{p(1-p)^{i-1}}{l}$. Definimos $P_{fn}(n)$ como la probabilidad de que no se pueda encontrar el camino hasta el atacante cuando la víctima recibe los n paquetes. Es obvio que el número de paquetes necesarios para obtener el camino del atacante es equivalente al coste de detección del problema del coleccionista de cromos, cuando tenemos que $p(1 : 1)$ y así obtenemos:

$$p_{fn}(n) = P(X(p(1 : 1)) > n)$$

5.2 Método de test aleatorio

Una secuencia aleatoria es una secuencia infinita de dígitos binarios que aparece aleatoria a cualquier algoritmo. Esta definición puede extenderse a cualquier conjunto finito de carácteres. Estas secuencias son muy importantes y tienen dos propósitos comunes. Uno de estos usos es que en la mayoría de algoritmos de cifrados necesitan una base de datos aleatorios, en la que escogemos las secuencias aleatorias. Un ejemplo muy conocido donde si se pierde la aleatoriedad se pierde la seguridad es el sistema, o los basados en el sistema RSA.

El otro uso de los números aleatorios es los generadores de números aleatorios (RNG), herramientas básicas del modelado estocástico. En la actualidad, existen muchos conjuntos de pruebas para evaluar la aleatoriedad de las secuencias binarias de bits como el de los conjuntos de pruebas NIST entre otras. Debido a que hay tantas pruebas para saber si una secuencia es aleatoria o no, normalmente el resultado de la prueba es parte de la aleatoriedad.

En la secuencia aleatoria propuesta por el problema del coleccionista de cromos vemos la longitud de la cadena en 10 caracteres, entre el 0 y el 9, es decir la cantidad de números que debemos escoger hasta obtener un número de cada tipo. De esta forma podemos conseguir una cadena de números que son completamente aleatorios. Es fácil ver la relación con el problema del coleccionista de cromos, ya que si tenemos una colección de 10 cromos X, el

número de cromos que necesitamos comprar para completar una colección, será justo la longitud de la primera cadena de números. Usando la función de masa de probabilidad con los números de Stirling podemos ver qué valores son los más probables y cuáles no.

Los valores de los cálculos son muy complejos pero a la vez muy interensantes de ver, todos estos se pueden ver en [4]

5.3 Otras aplicaciones

Además de estas dos aplicaciones el problema del coleccionista de cromos tiene muchas más aplicaciones, entre ellas tenemos las siguientes:

- Detección de todas las restricciones necesarias en un problema de optimización con restricciones.

- Determinar la clausura convexa en un conjunto de puntos $S \in \mathbb{R}^n$.

- Pruebas con cultivos biológicos para la contanimación.

- Desarrollo de procesos estocásticos con aplicaciones en redes "peer-to-peer".

Bibliografía

[1] Baum, LeonardE Patrick Billingsley: *Asymptotic distributions for the coupon collector's problemAsymptotic distributions for the coupon collector's problem.* The Annals of Mathematical Statistics, 36(6):1835–1839, 1965.

[2] Newman, DonaldJ: *The double dixie cup problemThe double dixie cup problem.* The American Mathematical Monthly, 67(1):58–61, 1960.

[3] Erdős, Paul Alfréd Rényi: *On a classical problem of probability theoryOn a classical problem of probability theory.* Magyar Tud. Akad. Mat. Kutató Int. Közl, 6(1):215–220, 1961.

[4] Zhang, Qinglong, Zongbin Liu, Quanwei Cai Ji Xiang: *Tst: A new randomness test method based on coupon collector's problemTST: A New Randomness Test Method Based on Coupon Collector's Problem. International Conference on Security and Privacy in Communication Systems*, 362–373. Springer, 2014.

[5] Galvin, David: *Basic Discrete Mathematics.* Departamento de matemáticas, Universidad de Notre Dame, 2013.

[6] Har-Peledx, Sariel: *Class notes for randomized algorithmsClass notes for randomized algorithms*, 2005.

[7] Shioda, Shigeo: *Some upper and lower bounds on the coupon collector problemSome upper and lower bounds on the coupon collector problem.* Journal of Computational and Applied Mathematics, 200(1):154–167, 2007.

[8] Ferrante, Marco Monica Saltalamacchia: *The coupon collector's problemThe coupon collector's problem.* Materials matemàtics, 0001–35, 2014.