Tests de aleatoriedad para secuencias finitas

Verónica Becher

Grupo KAPOW
(Knowledgeable Algorithms for Problems on Words)
Departamento Computación, Facultad de Ciencias Exactas y Naturales, UBA

18 de mayo de 2022

azar - aléatoire - Zufall - rasgelelik - satunnaisuuden - slumpmässighet - randomness - aleatorietà

Todos tenemos una idea intuitiva acerca de lo que es el azar, típicamente relacionada con los "juegos de azar" o con la "suerte"...

En castellano azar y aleatoriedad son sinónimos.
En inglés se dice "random".

La suerte es loca

¿Creerían que se obtienen echando una moneda para cada posición?

11111111111111111111111111111111... ✗
¡Son todos unos!

01001000100001000001000000100000001... ✗
¡Esta secuencia tiene un patrón!

00101001010001101110100010010101111...

Azar es imposibilidad de predecir, es falta de patrón.

La suerte es democrática

Azar es imposibilidad de predecir, es falta de patrón.

Entonces cara y ceca deben ocurrir, a la larga, la misma cantidad de veces

Sino, podríamos aprovecharnos del desvío y bastantes veces podríamos predecir bien.

La suerte es democrática

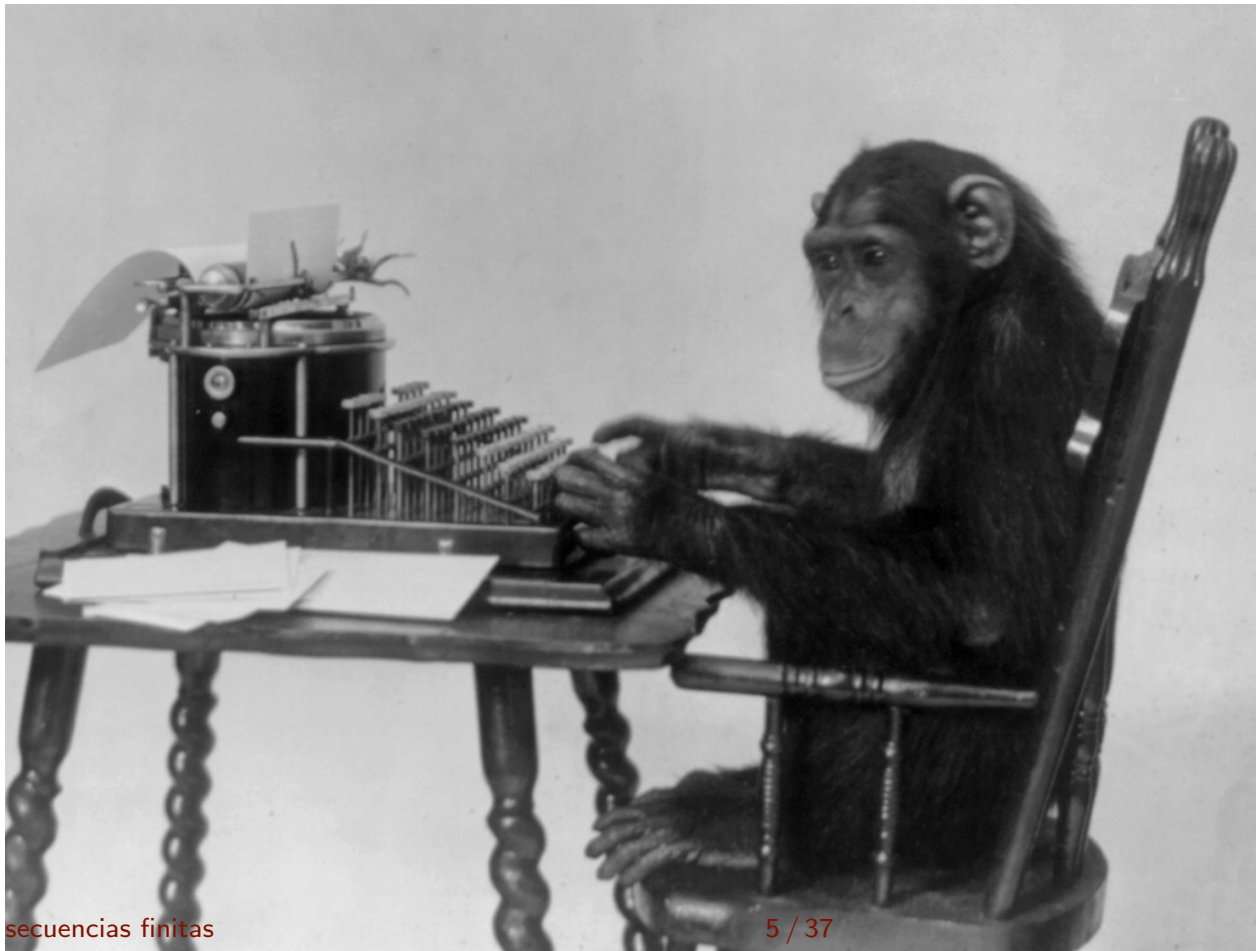
En vez de echar una moneda repartamos cartas.
Si jugamos el suficiente tiempo, y nadie hace trampa, alguna vez me tocarán el ancho de espadas, el ancho de basto y el 7 de espadas.

Un mono y una máquina de escribir

Teorema (Émile Borel 1913)

Si un mono se sienta en una máquina de escribir por siempre jamás escribirá todos los posibles textos, infinitas veces cada uno.

mono escribiendo = secuencia azarosa de símbolos



Sobre el azar

- ¿Hay una definición matemática de azar?
  SI
- ¿Podemos dar ejemplos?
  SI
- ¿Hay grados de azar?
  SI
- ¿Puede una computadora producir secuencias azarosas puras?
  NO
- ¿Qué algoritmos generadores pseudoaleatorios hay ?
  The Art of Computer Programming, Volumen 2, Donald Knuth.
  Ver referencias en esta presentación.
- ¿Cómo podemos convencernos de que una secuencia es azarosa?
  : Tests de aleatoriedad, Hoy!

Estos son los temas de la materia optativa Azar y Autómatas.

Como en un juicio

Una secuencia es aleatoria salvo que haya evidencia en contrario.

Tests de hipótesis - Hipótesis nula

H_0 : La secuencia a testear tiene para cada una de las posiciones la misma probabilidad sobre cada uno de los símbolos del alfabeto y cada posición es independiente de las demás

Los tests de hipótesis permiten refutar H_0. Pero no podemos confirmarla.

La evidencia en contra de H_0 se cuantifica numéricamente con un número real entre 0 y 1 llamado p-valor. Los p-valores muy bajos nos llevan a refutar H_0.

Tests de aleatoriedad

Donald Knuth, en su libro Art of Computer Programming, Volume 2 da una batería de tests para refutar aleatoriedad en secuencias de símbolos sobre alfabetos arbitrarios.

Cada test recibe como parámetro una secuencia y parámetros específicos del test. Arroja un p-valor de resultado.

Casi siempre es así: divide la secuencia en bitstreams (segmentos de igual tamaño) y determina cantidad de categorias. Para cada bitstream

1. computa estadísticas de cada categoría
2. computa test χ^2 (bondad de ajuste Pearson) y consigue un p-valor

Sobre los p-valores obtenidos computa test test Kolmogorov-Smirnov
Consigue el p-valor final. El p-valor bajo de K-S nos lleva a refutar H_0.

Test de Frecuencia de símbolos

Todos los símbolos del alfabeto Σ deberían aparecer con la misma frecuencia $1/|\Sigma|$. Si la secuencia tiene longitud n, cada símbolo debe aparecer en la secuencia aproximadamente $n/|\Sigma|$.

El test consiste en contar la cantidad de apariciones de cada símbolo y realizar un test χ^2 con $|\Sigma| - 1$ grados de libertad.

Sabemos que hay n observaciones. Sabiendo cuantas pertenecen a $k-1$ de las categorías, podemos deducir cuantas pertenecen a la restante.

Test de Frecuencia de Símbolos dentro de Bloque

Sea Σ el alfabeto.
Partir a la secuencia original en bloques de tamaño B.
Sea M es la cantidad de bloques que quedaron.

La cantidad esperada de cada símbolo en cada bloque es $B/|\Sigma|$
El test consiste en contar la cantidad de apariciones de cada símbolo en cada bloque y realizar un test χ^2 con grados de libertad $M(|\Sigma| - 1)$.

Test χ^2 (Test de bondad de ajuste de Pearson)

Supongamos k categorías (mutuamente exclusivas) y probabilidad p_i, $i = 1, .., k$, de que una observación sea de la i-ésima categoría.
Supongamos n observationes.
Contamos x_i, para observaciones de categoría $i = 1, \ldots, k$.
Las cantidad esperada en categoría i es $m_i = np_i$, $i = 1, .., k$,

$$\sum_{i=1}^{k} m_i = n \sum_{i=1}^{k} p_i = n = \sum_{i=1}^{k} x_i$$

Se calcula

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

Se definen los grados de libertad ℓ. Se busca X en la Tabla la distribución χ^2 para ℓ grados de libertad y se obtiene obtiene un p-valor (el p-valor es la probabilidad de hallar un valor así o peor suponiendo que la hipótesis nula es correcta).

En el reglón correspondiente a los grados de libertad encontrar las dos celdas cuyos valores encierran el valor calculado χ^2.

TABLA 3-Distribución Chi Cuadrado χ^2

P = Probabilidad de encontrar un valor mayor o igual que el chi cuadrado tabulado, ν = Grados de Libertad

v/p	0,001	0,0025	0,005	0,01	0,025	0,05	0,1	0,15	0,2	0,25	0,3
1	10,8274	9,1404	7,8794	6,6349	5,0239	3,8415	2,7055	2,0722	1,6424	1,3233	1,0742
2	13,8150	11,9827	10,5965	9,2104	7,3778	5,9915	4,6052	3,7942	3,2189	2,7726	2,4079
3	16,2660	14,3202	12,8381	11,3449	9,3484	7,8147	6,2514	5,3170	4,6416	4,1083	3,6649
4	18,4662	16,4238	14,8602	13,2767	11,1433	9,4877	7,7794	6,7449	5,9886	5,3853	4,8784
5	20,5147	18,3854	16,7496	15,0863	12,8325	11,0705	9,2363	8,1152	7,2893	6,6257	6,0644
6	22,4575	20,2491	18,5475	16,8119	14,4494	12,5916	10,6446	9,4461	8,5581	7,8408	7,2311
7	24,3213	22,0402	20,2777	18,4753	16,0128	14,0671	12,0170	10,7479	9,8032	9,0371	8,3834
8	26,1239	23,7742	21,9549	20,0902	17,5345	15,5073	13,3616	12,0271	11,0301	10,2189	9,5245
9	27,8767	25,4625	23,5893	21,6660	19,0228	16,9190	14,6837	13,2880	12,2421	11,3887	10,6564
10	29,5879	27,1119	25,1881	23,2093	20,4832	18,3070	15,9872	14,5339	13,4420	12,5489	11,7807
11	31,2635	28,7291	26,7569	24,7250	21,9200	19,6752	17,2750	15,7671	14,6314	13,7007	12,8987
12	32,9092	30,3182	28,2997	26,2170	23,3367	21,0261	18,5493	16,9893	15,8120	14,8454	14,0111
13	34,5274	31,8830	29,8193	27,6882	24,7356	22,3620	19,8119	18,2020	16,9848	15,9839	15,1187
14	36,1239	33,4262	31,3194	29,1412	26,1189	23,6848	21,0641	19,4062	18,1508	17,1169	16,2221
15	37,6978	34,9494	32,8015	30,5780	27,4884	24,9958	22,3071	20,6030	19,3107	18,2451	17,3217
16	39,2518	36,4555	34,2671	31,9999	28,8453	26,2962	23,5418	21,7931	20,4651	19,3689	18,4179
17	40,7911	37,9462	35,7184	33,4087	30,1910	27,5871	24,7690	22,9770	21,6146	20,4887	19,5110

Test de Kolmogorov-Smirnov

El test de Kolmogorov-Smirnov evalúa si un conjunto de números reales generados de forma independiente están distribuidos uniformemente en el intervalo [0,1].

Para informar el resultado se usa un p-valor. Valores muy bajos indican ausencia de distribución uniforme.

Test de Frecuencia de Palabras Alineadas dentro de Bloque

Partimos la secuencia original en bloques de tamaño B y fijamos un tamaño de palabra m.

Contamos cantidad de ocurrencias de cada palabra de longitud m, SIN SOLAPAMIENTO, dentro de cada bloque.
La cantidad esperada de ocurrencias de cada palabra en cada bloque es $B/(m|\Sigma|^m)$.

Realizamos un test χ^2 con $M \cdot |\Sigma|^m$ categorías y grados de libertad $M(|\Sigma|^m - 1)$.

Test de Racha mas larga dentro de bloque

(Longest Run Within Block)

Una racha es una secuencia maximal de símbolos idénticos consecutivos.

Elegir un símbolo, por ejemplo 0.

Fijamos longitud n. Dividimos la secuencia de entrada en bloques de longitud n.

Dentro de cada uno de estos bloques se computa la longitud de la racha más larga de dicho símbolo.

El test determina si las longitudes encontradas corresponden con las esperado por aleatoriedad.

Test χ^2 con $n = 1$ categorias, el 1 adicional es por rachas de longitud 0.

Test de Racha mas larga dentro de bloque

Elegimos un símbolo t.

Sea $D(n,k)$ la cantidad de cadenas de longitud n cuya racha más larga de ts tiene longitud k, y la secuencia no comienza y termina con t.

Para simplificar, asumimos $n \geq k+2$.

Dada una secuencia que tiene una racha más larga de longitud k, consideramos tres segmentos consecutivos.

$$\underbrace{\cdots j}_{j \neq t} \quad \underbrace{t \cdots t}_{\text{longitud } k} \quad \underbrace{i \cdots}_{i \neq t}$$

El primero no empieza ni termina en t y cuya racha de ts es menor que k.
El segundo es exactamente la racha de ts de longitud k.
El tercero no empieza ni termina en t y su racha de ts es menor o igual que k.
El primer y/o el tercer segmento podrían tener un solo símbolo.
Por ejemplo, con $t = 0$ y $\Sigma = \{0,1,2,3\}$:

$$\underbrace{1}_{\text{primer bloque}} \quad \underbrace{0000}_{k=4} \qquad \underbrace{12000312}_{\text{la racha de } 0 \text{ tiene longitud } < k}$$

Test de Racha mas larga dentro de bloque

$D(n,k)$ es la cantidad de cadenas de longitud n cuya racha más larga de ts tiene longitud k,y la secuencia no comienza y termina con t.
Para todo m,

$$D(n,0) = (|\Sigma| - 1)^n$$
$$D(n,n-2) = (|\Sigma| - 1)^2$$

Para todo k,

$$D(1,k) = (|\Sigma| - 1)$$
$$D(2,k) = (|\Sigma| - 1)^2$$

Para $n \geq 3$ y $0 < k < n-2$,

$$D(n,k) = \sum_{j=2}^{n-k-1} \sum_{s=0}^{k-1} D(j-1,s) \sum_{u=0}^{k} D(n-j+1-k,u)$$

donde j indexa la primera aparición de la racha, s indexa la racha del primer segmento, y t indexa la racha del tercer segmento.

Test de Racha mas larga dentro de bloque

Sea $C(n,k)$ la cantidad de cadenas de longitud n cuya racha de ts tiene longitud k. Luego

$$C(n,k) = D(n+2,k)/(|\Sigma|-1)^2$$

A partir de $C(n,k)$ determinamos la probabilidad p_k de que en una secuencia de n símbolos haya una racha de k ts, para $k = 1, 2, \ldots n-2$.

Test de Racha mas larga dentro de bloque

Fijemos M la máxima longitud de racha de ts que vamos a considerar.
Las probabilidades teóricas p_k para $k = 1, 2, \ldots M-1$ y una última categoría para las rachas de longitud mayor o igual que M.
Esta es la distribución teórica esperada para bloques de longitud $n = 400$ en el alfabeto que contiene los símbolos del 0 al 9:

k	p_k
1	$4.977414122938492e-19$
2	0.025339385657741895
3	0.6727353262392227
4	$0.2668025235392436 6$
5	0.03156393698492928
6	0.003203288734464395
7	0.0003200694497359544
8	$3.193140065308969 4e-05$
9	$3.185094068318768e-06$
10	$3.176999412983973e-07$

Test de Poker

El test de Poker divide la secuencia de entrada en bloques consecutivos de igual tamaño. Los llamamos *manos* y los analizamos como si se trataran de una mano de cartas.

Por ejemplo, consideremos la siguiente secuencia sobre $\Sigma = \{0, 1, 2, 3\}$ y manos de 5 cartas:

$$\underbrace{1 \quad 1 \quad 1 \quad 1 \quad 2}_{poker} \quad \underbrace{3 \quad 3 \quad 3 \quad 0 \quad 0}_{full}$$

Para generalizar más fácilmente el tamaño de la mano Knuth propone una variación del test donde en lugar de buscar patrones típicos del poker, sólo cuenta la cantidad de símbolos distintos en cada mano.

Volviendo al ejemplo:

$$\underbrace{1 \quad 1 \quad 1 \quad 1 \quad 2}_{\text{2 distintos}} \quad \underbrace{3 \quad 3 \quad 3 \quad 0 \quad 0}_{\text{2 distintos}}$$

Contamos la cantidad de símbolos distintos que hay en cada mano. Determinamos una categoría por cada cantidad de símbolos distintos posibles. Hacemos un χ^2 contado cuántas manos pertenecen a cada categoría.

Test de Poker - Números de Stirling de Segundo Orden

El número de Stirling de segundo orden n, k

$$\left\{ {n \atop k} \right\} = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (k-i)^n$$

es la cantidad de maneras de hacer una partición de un conjunto de n elementos en k subconjuntos disjuntos no vacías.

Test de Poker

La probabilidad de que una mano de tamaño n tenga k símbolos distintos, entonces, es:

$$p_k = \frac{|\Sigma|(|\Sigma|-1)\cdot\ldots\cdot(|\Sigma|-k+1)}{|\Sigma|^k}\left\{{n \atop k}\right\}$$

p_k suele ser muy pequeña para valores muy chicos de k

$$\underbrace{p_1 \quad p_2} \quad p_3 \cdots p_k$$

Entonces agrupamos algunas categorías juntas para que no queden vacías.

$$p_1 + p_2 \quad p_3 \cdots p_k$$

Luego realizamos un test χ^2 comparando la cantidad esperada con los valores observados.

Test de colisión

Es similar a contar colisiones en una tabla de hashing y comparar con la cantidad esperada.

Supongamos u urnas y n bolas, donde u es mucho mayor que n.

Esperamos que la mayoría de las bolas caigan en urnas vacías.
Si una bola cae en una urna ocupada, hay una colisión.

El test cuenta la cantidad de colisiones y se fija que no sean ni demasiadas ni muy pocas.

Test de colisión

Consideramos una secuencia de entrada y fijemos un valor entero m. Cadenas de m símbolos de la secuencia de entrada, sin superposición, hacen el rol de las bolas.

Cantidad de urnas es $|\Sigma|^m$.

Por ejemplo, sea la siguiente secuencia sobre $\Sigma = \{0, 1, 2, 3\}$:

$$0 \quad 2 \quad 3 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 0 \quad 3 \quad 1 \quad 0$$

Si $m = 4$,

$$\underbrace{0 \quad 2 \quad 3 \quad 1} \quad \underbrace{2 \quad 3 \quad 1 \quad 2} \quad \underbrace{0 \quad 3 \quad 1 \quad 0}$$

Test de colisión

Ahora tenemos n/m bolas (cadenas de m símbolos)

La probabilidad de que ocurran c colisiones es
la probabilidad de que solamente se ocupen $n/m - c$ urnas :

$$p_c = \frac{|\Sigma|^m(|\Sigma|^m - 1) \cdot \ldots \cdot (|\Sigma|^m - \frac{n}{m} + c + 1)}{(|\Sigma|^m)^{\frac{n}{m}}} \left\{ \begin{matrix} \frac{n}{m} \\ \frac{n}{m} - c \end{matrix} \right\}$$

Utilizando estas probabilidades se calcula una tabla utilizando puntos percentuales (por ejemplo, 0.01, 0.05, 0.25, 0.50, 0.75, 0.95, 0.99, 1.00)

colisiones $\leq$	101	108	119	126	134	145	153
con probabilidad	$.009$	$.043$	$.244$	$.476$	$.742$	$.946$	$.989$

Test de colisión

Se divide la entrada en bloques de tamaño B de palabras de longitud m.

Para cada a bloque contamos su cantidad de colisiones.
Determinamos la cantidad de categorías, rangos de cantidad de colisiones.
Utilizamos tabla de arriba para las probabilidades asociadas a las categorías para un test χ^2.
Esta foma de hacer el test arroja UN p-valor.

También es posible subdividir la entrada en grandes segmentos de la misma lingitud, y a cada uno de ellos correr el test recién descripto, y luego hacer un K-S.

Test del Album de Figuritas

Debemos contar cuántos símbolos de la secuencia mirar hasta conseguir que hayan ocurrido todos los símbolos del alfabeto Σ al menos una vez (coompletar el album).

Supongamos que tenemos la siguiente secuencia sobre $\Sigma = \{0, 1, 2, 3\}$:

$$0 \quad 0 \quad 3 \quad 1 \quad 0 \quad 2 \quad 2 \quad 2 \quad 3 \quad 0 \quad 1$$

Completamos dos álbumes:

$$\underbrace{0 \quad 0 \quad 3 \quad 1 \quad 0 \quad 2}_{\text{Primer album}} \quad \underbrace{2 \quad 2 \quad 3 \quad 0 \quad 1}_{\text{Segundo album}}$$

Test del Album de Figuritas

El test cuenta la longitud de las secuencias que completan el album.

Contamos la cantidad estas longitudes $|\Sigma|, \cdots, t-1$ y de longitud $\geq t$.
(Si el valor de t no es provisto, damos t que no deje categorías vacías)

Dado t, la probabilidad de que una secuencia tenga longitud r es:

- Para r tal que $|\Sigma| \leq r < t$:

$$p_r = \frac{|\Sigma|!}{|\Sigma|^r} \left\{ \begin{matrix} r-1 \\ |\Sigma|-1 \end{matrix} \right\}$$

- Para $r = t$:

$$p_r = 1 - \frac{|\Sigma|!}{|\Sigma|^{t-1}} \left\{ \begin{matrix} t-1 \\ |\Sigma| \end{matrix} \right\}$$

Test del Album de Figuritas

Luego realizamos un test χ^2 con $t - |\Sigma| + 1$ categorías:
una categoría por cada longitud entre $|\Sigma|$ y $t - 1$ y
una categoría $\geq t$ y $t - |\Sigma|$ grados de libertad

Test sobre numeros reales y sobre numeros enteros

Hay test diseñados para secuencias finitas de números reales (o racionales) entre 0 y 1.
Podemos transformar nuestra secuencia de símbolos del alfabetu Σ a una secuencia de racionales y aplicar el test.

Para esto, dada la secuencia original, y un valor m netero, dividimos la secuencia en bloques de tamaño m e interpretamos cada bloque de tamaño m como la parte decimal de un número de la forma $0, x$

Por ejemplo, para $\Sigma = \{0, 1, 2, 3\}$ y $m = 3$,

$$1 \quad 1 \quad 2 \quad 1 \quad 3 \quad 1$$

En este caso, obtendríamos esta nueva secuencia:

$$0.112 \quad 0.131$$

Test de Huecos (gap test)

Fijamos un intervalo $[\alpha, \beta]$, llamamos hueco a los elementos consecutivos que están fuera del intervalo. El test de huecos examina la longitud de huecos.

Supongamos que tenemos el intervalo $[\alpha, \beta]$, con $\alpha = \frac{1}{3}$ y $\beta = \frac{2}{3}$

Sea la siguiente secuencia de números reales:

$$0.1 \quad 0.3 \quad 0.2 \quad 0.4 \quad 0.5 \quad 0.1 \quad 0.3 \quad 0.9 \quad 0.4 \quad 0.7$$

Buscamos las longitudes de subsecuencias maximales de números que estén fuera del intervalo. En nuestro ejemplo:

$$\underbrace{0.1 \quad 0.3 \quad 0.2}_{3} \quad 0.4 \quad 0.5 \quad \underbrace{0.1 \quad 0.3 \quad 0.9}_{3} \quad 0.4 \quad \underbrace{0.7}_{1}$$

Test de Huecos

$$\underbrace{0.1 \quad 0.3 \quad 0.2}_{3} \quad 0.4 \quad 0.5 \quad \underbrace{0.1 \quad 0.3 \quad 0.9}_{3} \quad 0.4 \quad \underbrace{0.7}_{1}$$

Contamos huecos de longitudes $0, .., t-1$ y de longitud $\geq t$.

Nota: Si el valor de t no es provisto se busca uno tal que no queden categorías vacías hacia el final

En el ejemplo, si $t = 3$ tendríamos

tamaño hueco	0	1	2	≥ 3
cantidad observada	0	1	0	2

Test de Huecos

Si $p = \beta - \alpha$, la probabilidad de que la longitud de un hueco sea i es:

- Para i tal que $0 \leq i < t$:

$$p_i = p(1-p)^i$$

- Para $i = t$:

$$p_i = (1-p)^i$$

Con estas probabilidades se puede obtener la cantidad esperada de huecos (multiplicando por la cantidad de huecos observados).

Luego se realiza un test χ^2 (hay $t+1$ categorías, una por cada longitud menor a t y una por longitud $\geq t$) con t grados de libertad.

Valores críticos en cada test

1. Cantidad de símbolos de la secuencia de entrada
2. Cantidad de bitstream y su tamaño
3. Cantidad de categorías
4. Grados de libertad (cantidad de categorías independientes)
5. Las categorías no deben estar vacías

¿Siempre ceros y unos?

¿Qué pasa si las secuencias de entrada en secuencias son de un del alfabeto con más de dos símbolos ?
¿Está bien o mal transformar las secuencias de entrada a dos símbolos y utilizar baterías ya existentes?

Mal

Hay secuencias no aleatorias en alfabeto de tres símbolos, que pasan los tests de aleatoriedad al pasar a dos símbolos.
Por ejemplo la expansión en base 3 de todos los números reales del conjunto ternario de Cantor no tiene el dígito 1. Sin embargo, la mayoria de ellos son normales en base 2. (resultado de Cassels 1959).
Hay ejemplos concreto en Tesis de Licenciatura de Donatucci.

Baterías existentes

- NIST, alfabeto $\Sigma = \{0, 1\}$
- DieHard, alfabeto $\Sigma = \{0, 1\}$
- TestU01 , alfabeto $\Sigma = \{0, 1\}$ y números reales.
- Batería Tesis de Licenciatura Nicolás Donatucci, 2022. alfabeto Σ arbitrario. En Python 3, `https://github.com/NDonatucci/donut_tests`


Nicolás Donatucci.
Tests de aleatoriedad para alfabetos arbitrarios. Tesis de Licenciatura en Ciencias de la Computación, FCEyN, UBA, 2021.


Pierre L'ecuyer and Richard Simard.
TestU01: A C library for empirical testing of random number generators. ACM Transations omn Mathematical Software 33(4): 22, 2007.


Donald Knuth
The Art of Computer Programming Volume II, 2nd Edition *Addison-Wesley*, 1981


Onur Koçak and Fatih Sulak and Ali Doganaksoy and Muhiddin Uğuz
Modifications of Knuth randomness tests for integer and binary sequences *Commun. Fac. Sci. Univ. Ank. Ser. A1 Math. Stat. vol 67, number 2*, 64-81, 2018


George Marsaglia
The Marsaglia Random Number CDROM including the Diehard Battery of Tests of Randomness. *Florida State University*, 1995


National Institute of Standards and Technology,
Technology Administration, US Department of Commerce A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. *Special Publication 800-22 - Revision 1a*, 2010


Harald Niederreiter, Arne Winterhof
Applied Number Theory, Springer, 2015.

Burns tenía mil monos con mil máquinas de escribir...