Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 62-76

Tipo de artículo: Artículo original
Temática: Matemática Computacional
Recibido: 06/04/2021 | Aceptado: 28/04/2021

Prueba de bondad de ajuste para la distribución de distancias en secuencias de datos categóricos

Goodness of fit test for distance distribution in categorical data sequences

Niuman Comas Arias [1*] http://orcid.org/0000-0001-6291-3313
Belarmino Catalá González [1] http://orcid.org/0000-0002-6645-6567
Oscar Oro Dosouto [2] http://orcid.org/0000-0002-4339-6186

[1] Universidad de Holguín. Avenida XX Aniversario, vía Guardalavaca, Holguín. ncomasa@uho.edu.cu.
[3] CTE Lidio Ramón Pérez. Felton, Holguín. oscaralejandro.oro@nauta.cu.

*Autor para la correspondencia. (ncomasa@uho.edu.cu)

RESUMEN

El análisis de aleatoriedad en secuencias de datos categóricos es relevante para el estudio de procesos de Markov, fiabilidad de sistemas, big data, generación de números pseudoaletorios y encriptación de datos. Existen diferentes enfoques para el análisis de aleatoriedad implementados en paquetes como la batería de pruebas “Diehard”, el Test U01 y NIST Statistical Test Suite. El presente estudio analiza el comportamiento de secuencias categóricas interpretadas como series cronológicas de tiempo discreto demostrándose que la distribución esperada de las distancias entre eventos de cada categoría corresponde a la distribución

geométrica. La distribución de distancias observadas fue comparada con la teórica mediante prueba de bondad de ajuste basada en el estadístico chi-cuadrado. El algoritmo de la prueba fue implementado como módulo javascript para paquetes estadísticos en plataforma web comprobando su sensibilidad a diversas causas de comportamiento no aleatorio: el carácter periódico de los eventos, agrupamiento en bloques, autocorrelación y los procesos de Markov. La convergencia y robustez de la prueba fueron estudiadas mediante simulación en ordenador detectándose pequeñas desviaciones en la proporción de casos significativos esperados que indican la existencia de sesgos inherentes al criterio de agrupamiento utilizado en la prueba chi-cuadrado.

Palabras clave: Secuencias categóricas; aleatoriedad; prueba de bondad de ajuste.

ABSTRACT

Randomness analysis in categorical sequences is relevant for the study of Markov processes, system realibity, big data, data encryption and evaluation of pseudo-random number generators. Various approaches exist in order to appraise the randomness phenomena, they lead to a variety of tests such as the “Diehard” test battery, the test U01 and the NIST Statistical Test Suite. The behavior of categorical sequences was studied and understood as a discrete time chronological series. It was proved that the geometric distribution is the expected distribution (theoretical distribution) for distances between successes random sequences. The observed distance distribution was compared to the theoretical distribution by goodness of fit test based on chi-square statistic. The test algorithm was implemented as javascript module for web statistical packages checking its sensibility to various no random behavior including the periodical character of successes, blocking, autocorrelation and Markov processes existence. Test convergence and robustness were studied by means of simulation in computer, discovering little deviations in proportion of the significant cases that indicate the existence of inherent biased in chi-square test.

Keywords: Categorical sequences; randomness; goodness of fit test.

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 62-76

Introducción

Las secuencias categóricas están presentes en la naturaleza y en la actividad humana: desde la sucesión de aminoácidos presentes en una proteína (Santoni, 2016), hasta el comportamiento de las fallas en un proceso industrial (Coit, 2018) o la sucesión de eventos reportados en un estudio sociológico o económico (Chou, 2020).

Entre los paradigmas de modelación de secuencias categóricas más utilizados destacan las cadenas o procesos de Markov (Ramaekers, 2019) y las técnicas propias del análisis de fiabilidad de sistemas como los modelos exponenciales y de Weibull para la obtención de valores esperados en variables involucradas en procesos industriales y tecnológicos (Elabatal, 2016) (Beyer, 2018).

El estudio de la aleatoriedad en secuencias categóricas es relevante para la obtención de modelos estadístico-matemáticos y la comprensión de su comportamiento (Traylor, 2017). La aleatoriedad es una propiedad fundamental en el análisis estadístico. Es una de las condiciones inherentes a la teoría de muestreo y de hecho está presente en casi todas las pruebas de hipótesis que se realizan en estadística inferencial. (Shen, 2018)

La aleatoriedad (para secuencias categóricas) puede valorarse bajo diferentes raseros tales como: la igualdad de las frecuencias relativas para cada categoría, las pruebas basadas en rachas, la independencia estadística entre las categorías y la detección de patrones en la secuencia (Koller, 2018).

El campo donde más se ha avanzado en el desarrollo de pruebas de aleatoriedad es el de la evaluación de generadores de números pseudo-aleatorios (Pseudo-Random Number Generator, PRNG) utilizados en encriptación de datos (Martínez, 2018) (Gangyi, 2019). Entre los paquetes de pruebas más conocidos destacan Crypt-XS, la batería de pruebas Diehard “Intransigente” y su desarrollo posterior: el TestU01 (Shen, 2019), NIST Statistical Test Suite (Iwasaki, 2018) y Randomness Testing Toolkit (Obrátil, 2017). Algunas de las pruebas incorporadas son el test de Wald-Wolfowitz, basado en la cantidad de rachas observadas en la secuencia (Doganaksoy, 2015); el test de frecuencias monobit; la prueba de espaciamiento entre cumpleaños; la prueba de los monos, basada en la paradoja de los monos mecanógrafos; el test espectral (aplicación de la transformada discreta de Fourier), la entropía aproximada, y el test de Lempel-Ziv (Obrátil, 2017).

La alternativa que se aborda en el presente trabajo consiste en analizar el comportamiento de las distancias discretas entre eventos de una misma categoría: computar las distancias y juzgar acerca de la aleatoriedad de

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 62-76

la muestra en base a su distribución. Para lograr este objetivo es necesario responder dos preguntas fundamentales: ¿Qué distribución teórica de probabilidades corresponde a la distribución de las distancias entre eventos iguales en una secuencia aleatoria? y ¿Qué pruebas estadísticas pueden aplicarse para determinar la correspondencia entre la distribución teórica y la observada?

Adicionalmente es de interés el estudio minucioso de los resultados que brindan las pruebas aplicadas a fin de evaluar su convergencia y robustez, detectar desviaciones en la frecuencia de errores esperados y comprobar la sensibilidad de las pruebas a diferentes fuentes de comportamiento no aleatorio.

Métodos

Fundamentación teórica de la prueba de las distancias

Sea una secuencia finita de datos categóricos X, compuesta por n eventos y k categorías diferentes. Para cada categoría se define la frecuencia absoluta y frecuencia relativa como:

n_i: frecuencia absoluta: cantidad de eventos de la categoría i,

$r_i = \frac{n_i}{n}$: frecuencia relativa de los eventos pertenecientes a la categoría i. (1)

La secuencia categórica, una vez codificada, puede tener un aspecto como el siguiente:
ABCAABDAEBACABBCDAAABCADE …

Donde cada carácter representa un evento o estado del sistema. Los eventos son mutuamente excluyentes. La posición en que se encuentra cada carácter determina el instante de tiempo en que ocurre cada evento: 1, 2, 3, siendo este una variable discreta autoincremental.

Se define como distancia (d) entre dos eventos sucesivos de la misma categoría ($x_j = x_{j+h}$) la cantidad de pasos necesarios para alcanzar desde x_j desde x_{j+h}.

d = 1, 2, 3,… discreta.

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 62-76

$d(x_j, x_{j+h}) = h;$ (2)

$d(x_j, x_{j+1}) = 1$; la distancia entre dos eventos contiguos es igual a 1.

La distancia para el primer elemento de la categoría i se define como la suma de los pasos desde el inicio de la secuencia hasta el primer $x_j = K_i$ más los pasos desde el último $x_j = K_i$ hasta el final de la secuencia.
El promedio (media aritmética) de las distancias para cada categoría es:

$$\bar{d}_i = \frac{1}{r_i} \quad (3)$$

Tomando la frecuencia relativa para cada clase como estimador de probabilidad con que ocurre el evento K_i pueden definirse las probabilidades de obtener las distancias al próximo evento como sigue:

$P(1) = r_i$ probabilidad de que la distancia al próximo evento de la clase i sea d = 1.

$P(2) = r_i(1 - r_i)$ ídem, d = 2. Resulta de aplicar la regla de multiplicación para dos eventos sucesivos: que no ocurra el evento x_i en un primer momento $(1 - r_i)$ y sí ocurra en el segundo momento consecutivo de la secuencia (r_i).

$P(3) = r_i(1 - r_i)^2$ ídem, d = 3.

$P(d) = r_i(1 - r_i)^{d-1}$ generalización. (4)

La ecuación (4) se corresponde con la distribución geométrica, con la diferencia que aquí el argumento de la función, comienza en el valor d = 1 y no en cero como generalmente se presenta esta distribución: $P(x) = r\,(1 - r)^x$ (NIST, 2018). Entonces puede afirmarse que, para una distribución aleatoria de los datos a lo largo de la secuencia, las distancias entre los eventos de una misma clase siguen una distribución geométrica tal como se describe en (4).

La esperanza matemática (media aritmética) y la varianza teórica esperada para esta distribución se calculan según:

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 62-76

$$E(d) = \frac{1}{r} \quad (5)$$

$$V(d) = \frac{1-r}{r^2} \quad (6)$$

Tomando las probabilidades teóricas calculadas mediante (4) como frecuencias relativas esperadas para cada distancia (lo que es válido para n lo suficientemente grande) puede estimarse la frecuencia absoluta esperada para cada distancia como el producto:

$$f_{esp}(d) = n_i \cdot p(d) \quad (7)$$

Al computar en tabla de frecuencias absolutas las distancias observadas (para cada clase independientemente) en la secuencia de datos (muestra), las diferencias cuadráticas entre frecuencias observadas y esperadas se calculan como:

$$\chi^2 = \sum \frac{(f_{esp} - f_{obs})^2}{f_{esp}} \quad (8)$$

Siendo m la cantidad de filas en la tabla de frecuencias, el estadístico χ^2 distribuye Chi-cuadrado con m - 2 grados de libertad puesto que se ha estimado un parámetro, r_i, para calcular las frecuencias esperadas. McClave & Sincich, 2018).

El p-valor se determina como en una prueba unilateral de cola derecha. La expresión: = 1 – CHISQ.DIST(χ^2, m – 1, TRUE) devuelve el resultado necesario tal y como se implementa la función Chi-cuadrado en tabuladores electrónicos tipo Microsoft Excel.

Comoquiera que las distancias observadas alcanzan un valor máximo, quedarían sin computar las distancias no observadas después de esta distancia máxima (d_{max}), para las cuales existe una probabilidad marginal. En este caso la tabla de frecuencias debe tener en cuenta una fila adicional donde $f_{obs} = 0$, el estadístico chi-cuadrado para esta última fila sería:

$$\chi_{m+1}^2 = \sum \frac{(f_{esp(r)} - 0)^2}{f_{esp(r)}} = \sum f_{esp(r)}$$

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 62-76

(10) (10)

Donde $\Sigma f_{esp(r)}$ corresponde a la suma de frecuencias esperadas (remanentes) para las distancias mayores que d_{max}, la que puede calcularse según:

$$\sum f_{esp(r)} = n_i - \sum f_{esp}(d \leq d_{max}) \quad (11)$$

De manera que el efecto de las distancias ausentes resulta en adicionar la $\Sigma f_{esp(r)}$ al cómputo de chi-cuadrado obtenido en (6).

Como criterio de agrupamiento se propone utilizar la expresión para calcular la suma mínima de frecuencias esperada para cada clase de la tabla de frecuencias:

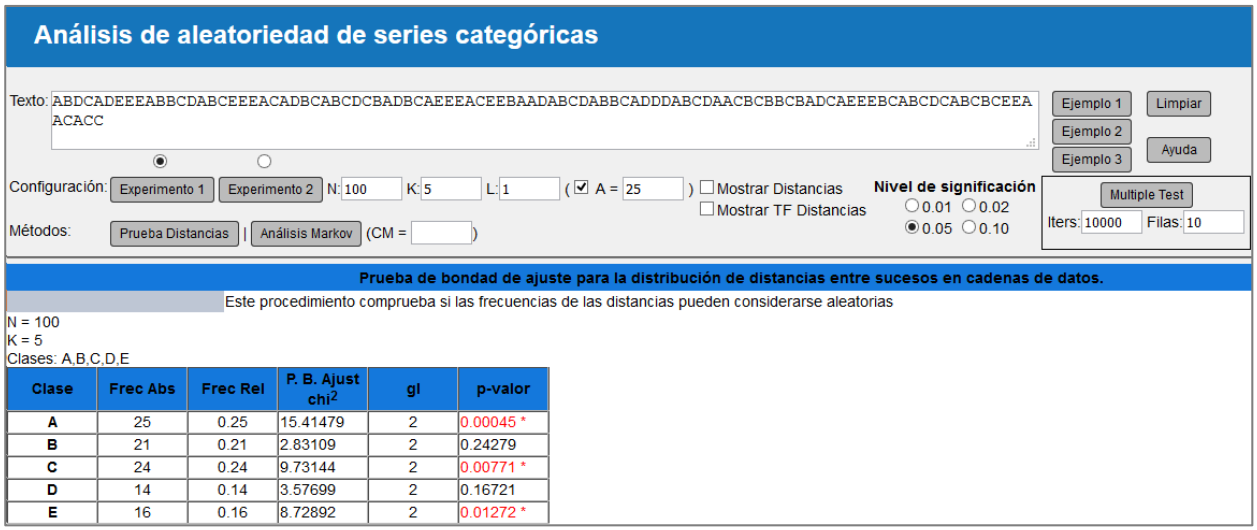
$$\sum f_{\text{esperada mínima}} \geq \ln(n_i) \quad (12)$$

Agregando la condición adicional de cota mínima absoluta para $\sum f_{\text{esperada mínima}} \geq 5$ evitar un uso indiscriminado de la corrección de Yates (Corder & Foreman, 2016).

Resultados y discusión

La prueba de las distancias ha sido implementada como módulo javascript para paquetes estadísticos de plataforma web el cual puede trasladarse a otros lenguajes con relativa facilidad. La interfaz gráfica de usuario fue elaborada en lenguaje HTML y contiene un panel para la entrada de datos y selección de opciones y un panel de salida de resultados. Incluye secuencias de ejemplo prediseñadas y permite generar secuencias pseudo-aleatorias bajo diferentes condiciones definidas por el usuario. Adicionalmente puede realizar simulaciones múltiples de las pruebas estadísticas incorporadas con el objetivo de evaluar su comportamiento.

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 62-76



Clase	Frec Abs	Frec Rel	P. B. Ajust chi^2	gl	p-valor
A	25	0.25	15.41479	2	0.00045 *
B	21	0.21	2.83109	2	0.24279
C	24	0.24	9.73144	2	0.00771 *
D	14	0.14	3.57699	2	0.16721
E	16	0.16	8.72892	2	0.01272 *

Fig. 1 – Interfaz gráfica del módulo para análisis de aleatoriedad de series categóricas.

El siguiente ejemplo didáctico muestra los resultados obtenidos al aplicar la prueba de las distancias a una secuencia previamente diseñada para ilustrar diferentes clases de comportamiento de los datos.

Se desea analizar la distribución de distancias en la secuencia:

ABDCADEEEABBCDABCEEEACADBCABCDCBADBCAEEEACEEBAADABCDABBCADDDABC DAACBCBBCBADCAEEEBCABCDCABCBCEEAACACC.

Las hipótesis nula y alternativa se plantean como sigue:

H_0: Las distribuciones de frecuencias de distancias observadas y esperadas son iguales, indicando que los datos se distribuyen de manera aleatoria.

H_1: Las distribuciones de frecuencias de distancias observadas y esperadas difieren significativamente, por lo tanto, los datos no están distribuidos de manera aleatoria.

Análisis de las distancias:

n = 100

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 62-76

k = 5, xi = {A, B, C, D, E}

Tabla 1 - Resultados de la prueba de distribución de las distancias para cada categoría.

Categoría	Frec. Abs.	Frec. Rel.	χ^2	gl	p-valor
A	25	0.25	15.4148	2	0.00045 *
B	21	0.21	2.8311	2	0.24279
C	24	0.24	9.7314	2	0.00771 *
D	14	0.14	3.5770	2	0.16721
E	16	0.16	8.7289	2	0.01272 *

* El asterisco indica significación estadística.

Las distribuciones de distancias de las categorías A, C y E presentan diferencias significativas respecto al comportamiento esperado en condiciones de aleatoriedad.

El p-valor obtenido para la categoría A es significativamente menor que 0.01, esto indica que la probabilidad de que distribución de distancias de los eventos A sea aleatoria es muy pequeña, inferior al uno por ciento. La inspección detallada de los datos muestra que para la categoría A existe un marcado predominio de la distancia $d_{A\text{-}A} = 4$ (8 ítems), la cual coincide con la media aritmética de las distancias entre eventos A. Esto hace sospechar que A presenta una fuerte tendencia a ocurrir cada 4 eventos de manera periódica.

En el caso de la categoría E predomina la distancia $d_{E\text{-}E} = 1$ (10 ítems), indicando la tendencia de E a ocurrir en rachas de varias apariciones consecutivas. Obsérvese una regularidad en C: de 24 apariciones en la secuencia, en 14 ocasiones C es inmediatamente precedido por el evento B, lo cual es un indicio de que la secuencia tiene características de un proceso de Markov, al menos para la interacción B → C.

Para comprobar la propiedad markoviana en la secuencia anterior se confeccionó una tabla de contingencia para la relación causa-efecto entre todos los eventos tomados uno a continuación del otro, esta fue sometida a una prueba chi-cuadrado de independencia estadística utilizando el paquete estadístico Statgraphics Centurión (2020). Se contrasta la hipótesis nula: los valores observados para la variable Efecto (evento

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 62-76

posterior) son independientes de la variable Causa (evento anterior) versus la hipótesis alternativa: existe relación entre la Causa y el Efecto. La tabla 2 muestra los resultados obtenidos.

Tabla 2 - Relación causa-efecto entre eventos sucesivos.

	Efecto					
Causa	A	B	C	D	E	Total
A	3 *	9	5	6	2	25
B	3	3	14 *	1	0	21
C	10	5	1 *	5	3	24
D	5	2	4	2	1	14
E	4	2	0 *	0	10 *	16
Total	25	21	24	14	16	100

* Los asteriscos indican significación estadística.

$Chi^2 = 66.895$ Coef. Contingencia = 0.63311

p-valor = $3.5582 \cdot 10^{-8}$ * Coef. V de Cramer = 0.40895

La frecuencia con que el evento A actúa como causa de sí mismo es significativamente pequeña, lo cual es coherente con el carácter periódico de A descrito anteriormente. Por el contrario, se observa una fuerte autocorrelación para el evento E como causa de sí mismo; lo que se corresponde con la tendencia observada a la aparición de rachas E consecutivos. Se revela el efecto causa-efecto para el binomio B → C, haciéndose patente la no independencia estadística entre las clases. El p-valor obtenido para la prueba de independencia estadística significativamente inferior a 001, por lo que debe rechazarse la hipótesis nula con nivel de confianza superior al 99%.

Análisis de la convergencia y robustez de la prueba

La comprobación de estas características fue realizada mediante la herramienta Múltiple Test con generación de datos aleatorios. Para secuencias generadas aleatoriamente es de esperar que la proporción de "falsos positivos" (Error Tipo I) sea aproximadamente igual al nivel de significación utilizado en la prueba (alfa). Para detectar la convergencia de la prueba se computó la desviación estándar para la proporción de casos significativos ("falsos positivos") a medida que se incrementa la cantidad de iteraciones.

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 62-76

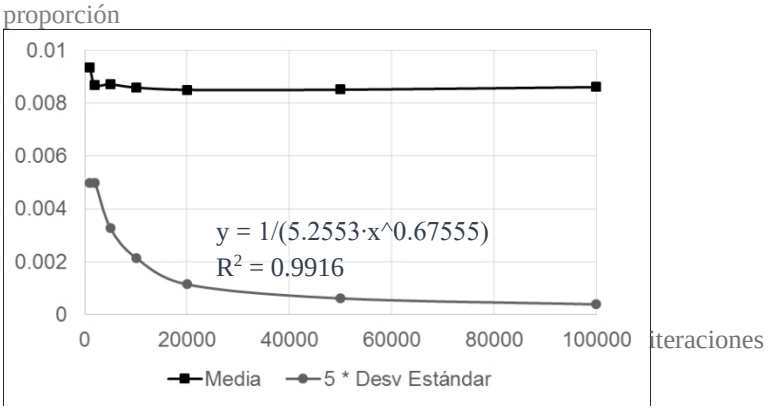


Fig. 2 - Convergencia de la proporción de casos significativos al incrementarse el número de iteraciones. Simulación para n = 1000, k = 5. Significación: α = 0.01.

La curva superior en la figura 2 representa el promedio de las proporciones y la inferior las desviaciones estándar correspondientes a cada bloque de iteraciones. La desviación estándar de la proporción de casos significativos es inversamente proporcional a la cantidad de iteraciones.

La figura 3 analiza la robustez de la prueba. Para $n_i < 100$ existe comportamiento errático en la proporción de casos estadísticamente afectándose la confiabilidad del test chi-cuadrado.

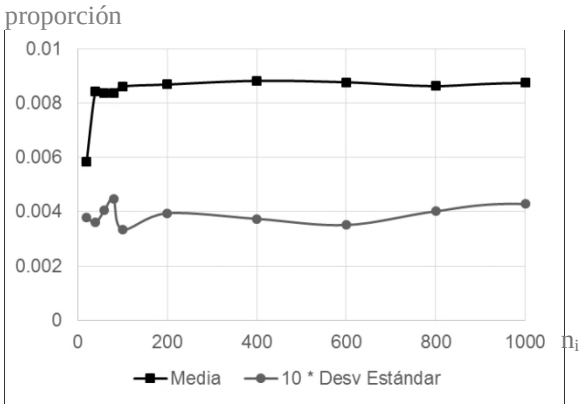


Fig. 3 - Comportamiento de la proporción de casos significativos según el tamaño individual de la muestra (n_i) para muestras con k = 5 categorías equiprobables (frecuencia relativa $r_i = 0.20$). (α = 0.01)

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 62-76

Para de $n_i > 100$ la tendencia la proporción de casos con significación estadística se comporta de manera estable, con fluctuaciones alrededor de 0.00873< 0.01, sin aproximación asintótica. La desviación estándar de la proporción de "falsos positivos" es significativamente inferior (1:20) al valor de la proporción, lo cual es un indicador de la robustez de la prueba. Los resultados acusan la existencia de un sesgo inherente a la prueba en el sentido de disminuir la probabilidad de cometer errores de tipo I (rechazar la hipótesis nula sobre la aleatoriedad de la secuencia cuando en realidad esta sí se cumple),

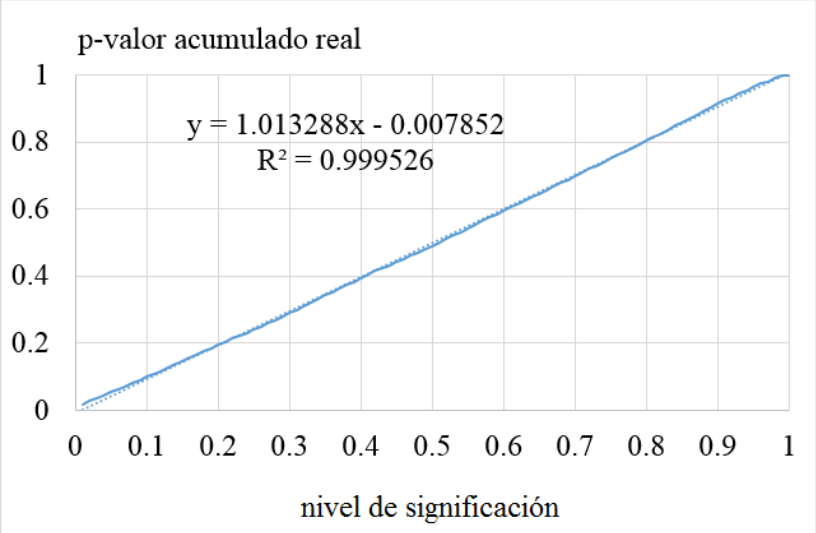


Fig. 4 - Comportamiento de la proporción de casos significativos respecto al nivel de significación de la prueba de bondad de ajuste para la distribución de distancias.

El gráfico representado en la figura 4 corresponde al comportamiento observado para eventos que ocurren con una frecuencia relativa r = 0.20 y tamaño de muestra n = 200. El comportamiento esperado indica una proporción del 1 % de casos significativos para un nivel de significación $\alpha = 0.01$. Para un nivel de significación $\alpha = 0.05$ se espera un 5% de casos significativos y así sucesivamente para $\alpha = 0.10$, $\alpha = 0.20$, hasta cubrir el intervalo (0;1). Se observan ligeras desviaciones respecto a la línea recta esperada (línea de puntos), Para las muestras estudiadas la desviación máxima resultó del 1.67 % de casos significativos observados en diferencia con la proporción de casos significativos esperados.

Conclusiones

La frecuencia esperada para las distancias entre eventos de una misma clase en secuencias de datos categóricos sigue una distribución geométrica con parámetro r determinado por la frecuencia relativa de la clase en la muestra.

La prueba de bondad de ajuste para las distancias permite detectar desviaciones de la aleatoriedad respecto a la distribución longitudinal de los datos en la secuencia.

Se comprobó la sensibilidad de la prueba de aleatoriedad a comportamientos de tipo periódico, tendencia al agrupamiento de los eventos en bloques, existencia de autocorrelación y asociación entre las categorías (propiedad Markoviana).

La convergencia y robustez de la prueba son estudiadas mediante la simulación en ordenador detectándose desviaciones máximas del 1.67 % en la proporción casos significativos respecto a lo esperado que indica la existencia de sesgos inherentes al criterio de agrupamiento utilizado en la prueba chi-cuadrado.

La prueba es apropiada para su aplicación al estudio de aleatoriedad en el comportamiento de las variables involucradas en procesos y sistemas tecnológicos, el análisis de series cronológicas de naturaleza categórica, detección de procesos de Markov, análisis de textos y Big Data.

Referencias

Beyer, B; Murphy, N. R; Rensin, D. K; Kawahara, K; Thorne, S. The Site Reliability Workbook. O'Reilly, 2018. ISBN 978-1-492-02950-2.

Elbatal, I; Mansour, M; Ahsanullah, M. The Additive Weibull-Geometric Distribution: Theory and Applications. Journal of Statistical Theory and Applications, 15 (2), 2016.

Chou, E.; Mcvey, C.; Hsieh, Y.; Enriquez, S.; Hsieh, F. Extreme-K Categorical Samples Problem. Arxiv: 2007.15039, 2020.

Coit, D.; Zio, E. The Evolution of System Reliability Optimization. Elsevier, 2019.
Corder, G. W.; Foreman, D. I. *Nonparametric Statistics For Non-Statisticians: A Step-By-Step Approach*. New York: Wiley, 2016.

Doganaksoy, A; Sulak, F; Uguz, M; Seker, O; Akcengiz, Z. New Statistical Randomness Tests Based On Length of Runs. *Mathematical Problems in Engineering*, 2015.
Gangyi, H; Jin, P; Weili, P. A Novel Algorithm for Generating Pseudo-Random Number. International Journal of Computational Intelligence Systems, 12(2), 2019, Disponible En Https://Www.Atlantis-Press.Com/Article/125909667.Pdf
Iwasaki, A. Diagonalizing Method Among Test Items Included In Nist Randomness Test Tool. Fukuoka Institute of Technology, 2018.
Koller, Z. *Measuring Loss and Reordering With Few Bits*. Master Thesis, Zurich: Swiss Federal Institute of Technology, 2018.
Martinez, A; Solís, A; Díaz-Hernández, R; Et Al. Testing Randomness in Quantum Mechanics. *Entropy*, 2018.
Mcclave, J. T.; Sincich, T. *Statistics*. Boston: Pearson Education, Inc, 2018.

Nist/Sematech. *E-Handbook Of Statistical Methods*. 2018, Disponible En Http://Www.Itl.Nist.Gov/Div898/Handbook/.

Obrátil, L. *The Automated Testing Of Randomness with Multiple Statistical Batteries*. Master's Thesis, Brno: Masaryk University, 2017.
Santoni, D; Felici, G; Vergni, D. Natural vs. Random Protein Sequences: Discovering Combinatorics Properties on Amino Acid Words. *Journal of Theoretical Biology*, Volume 391, 2016.
Shen, A. Making Randomness Tests More Robust. Hal Archive, 2018, Disponible En Https://Hal.Archives-Ouvertes.Fr/Hal-01707610
Shen, A. *Randomness Tests: Theory and Practice*. Reporte Preliminar, 2019, Disponible En Http://Www.Lirmm.Fr/~Ashen/Racaf/2019-Preliminary-Report.Pdf.
Shen, A. *Making Randomness Tests More Robust*, 2018, Disponible En Https://Hal.Archives-Ouvertes.Fr/Hal-01707610
Statpoint Technologies, Inc. *Statgraphics Centurion 18.* Warrenton, Va: Statpoint Technologies, Inc., 2020, Disponible En Https://Www.Statgraphics.Com/Download18

Traylor, R; Hatchcock, J. Vertical Dependency in Sequences of Categorical Random Variables. Academic Advances of The Cto, 1 (2), 2017.

Conflicto de interés

Los autores autorizan la distribución y uso de su artículo.

Contribuciones de los autores

1. Conceptualización: Niuman Comas Arias.
2. Curación de datos: Belarmino Catalá González
3. Análisis formal: Niuman Comas Arias
4. Investigación: Belarmino Catalá González
5. Metodología: Belarmino Catalá González
6. Administración del proyecto: Niuman Comas Arias
7. Recursos: Oscar Oro Dosouto
8. Software: Niuman Comas Arias
9. Supervisión: Belarmino Catalá González
10. Validación: Oscar Oro Dosouto
11. Visualización: Niuman Comas Arias
12. Redacción – borrador original: Oscar Oro Dosouto
13. Redacción – revisión y edición: Niuman Comas Arias

Financiación

El trabajo se enmarca en las líneas de investigación del Departamento de Matemática Aplicada de la Facultad de Informática y Matemática de la Universidad de Holguín en colaboración con el Departamento de Energética de la Facultad de Ingeniería Mecánica y el Grupo de Planificación de Mantenimientos de la Central Termoeléctrica Lidio Ramón Pérez de Felton.