DEMOCRACIA EN DIGITAL:

FACEBOOK, COMUNICACIÓN Y POLÍTICA EN COSTA RICA

Ignacio Siles González
Editor

DEMOCRACIA EN DIGITAL:

FACEBOOK, COMUNICACIÓN Y POLÍTICA EN COSTA RICA

Ignacio Siles González
Editor

UNIVERSIDAD DE COSTA RICA
CICOM Centro de Investigación en Comunicación
UCREA Espacio Universitario de Estudios Avanzados

324.972.860.5
S582d Siles González, Ignacio
Democracia en digital: facebook, comunicación y política en Costa Rica / Ignacio Siles González, editor; autores Ignacio Siles González [y otros trece]. – San José, Costa Rica: Universidad de Costa Rica, Centro de Investigación en Comunicación, Espacio Universitario de Estudios Avanzados, 2020.
viii, 305 páginas: ilustraciones (algunas a color). – (Colección tecnología y sociedad)

Autores tomados del colofón.
ISBN 978-9968-919-69-2

1. CAMPAÑA ELECTORAL – COSTA RICA-2018. 2. FACEBOOK (RECURSO ELECTRÓNICO). 3. REDES SOCIALES – COSTA RICA. 4. COMUNICACIÓN EN POLÍTICA. 5. POLÍTICA Y MEDIOS DE COMUNICACIÓN DE MASAS. 6. DEMOCRACIA – COSTA RICA. I. Título. II. Serie.

CIP/3507
CC.SIBDI.UCR

Coordinación editorial: Centro de Investigación en Comunicación (CICOM)
Revisión filológica: La Voz Activa
Diagramación: Nicole Chaves Mora
Diseño de portada: Daniele Lampis
Control de calidad: Grettel Calderón Abarca

Impreso en Lara Segura y Asociados S.A.

Primera edición 2020

Facultad de Ciencias Sociales, Universidad de Costa Rica
Montes de Oca, San José, Costa Rica
2511-6414 www.cicom.ucr.ac.cr

3

La región de probabilidad: un enfoque computacional para la comprensión del consumo de noticias políticas en medios costarricenses

Brayan Rodríguez Delgado

Introducción

Las ciencias sociales no son ajenas a las matemáticas y la estadística, entre todas ellas existe un puente de colaboración en ambos sentidos. Esta conexión, a menudo, se suele extender para abordar una temática específica y para la construcción de nuevos modelos que ayudan a explicar fenómenos sociales observados, lo cual contribuye al desarrollo mutuo entre las ciencias. Estos modelos con el tiempo se convierten en campos completos de conocimiento donde convergen profesionales de distintos ámbitos, quienes trabajan en conjunto, para ampliar la visión general sobre algún acontecer, como por ejemplo se puede citar el uso de la teoría de grafos y matrices en el análisis de redes sociales para representar relaciones entre un grupo de personas u organizaciones (Crane, 2018).

El fenómeno de interés, en el presente estudio, corresponde al comportamiento del consumo de noticias políticas en el contexto costarricense. Respecto a esto, según Siles, Campos y Segura (2018), un grupo de investigadores del Centro de Investigación y Comunicación (CICOM) de la Universidad de Costa Rica, con el apoyo del Laboratorio de Investigación e Innovación Tecnológica (LIIT) de la UNED, había dado seguimiento a la actividad de diez medios de comunicación costarricenses por medio de la red social Facebook. Su monitoreo abarcó las publicaciones entre enero 2016 y diciembre 2017 que poseían el mayor número de interacciones o *engagement* por parte de la persona consumidora final. Mediante un análisis de contenido, determinaron que esta tiende a una mayor preferencia por las noticias de carácter no público. Asimismo, al profundizar más en el estudio, detectaron que la proporción de noticias de índole política no superó cierto umbral o límite con respecto al total de noticias analizadas.

Estas observaciones parecen comprobar la falta de avidez de la ciudadanía por el consumo de noticias de interés público y la preferencia general por otro tipo de

noticias, mencionada por Tristán y Álvarez (2018). Ahora, estas autoras, también hacen referencia a la existencia de periodos tanto de baja como de alta actividad política, donde el número de noticias de esa índole cambia sustancialmente. Dado el criterio de selección de las noticias, basado en la interacción e intereses del usuario, sumado a periodos de actividad política, cabe entonces preguntarse si las proporciones de noticias políticas en Costa Rica realmente no sobrepasan un umbral determinado, además, cómo afectó la presencia de los eventos ocurridos en el 2018, a saber, las elecciones presidenciales, las manifestaciones populares como la huelga del sector público, entre otros.

De esta manera surge la motivación para abordar dicha situación de interés desde una perspectiva matemática, estadística y computacional, con dos objetivos principales, el primero, bajo un nivel de confianza aceptable, comprobar si existe un valor límite para la mayoría de las proporciones de noticias políticas y segundo, encontrar cuál es ese umbral mínimo. Ahora bien, dado que la proporción de noticias políticas es un proceso estocástico, es decir, puede asumir más de un valor a lo largo del tiempo (Beichelt, 2016), resulta conveniente enfocar el cálculo del umbral más como una región de probabilidad que como un valor, de modo que un evento noticioso de corte político pueda ser identificado como dentro o fuera de dicha región, de manera similar a como un químico puede encontrar, dada una probabilidad, en qué orbital va a estar un electrón dentro de un átomo (Albright, Burdett y Whangbo, 2013). Por otra parte, aunque no necesariamente sea un método para detectar datos atípicos como el propuesto por Gbenro (2018), sí podría ser utilizado como un indicador de que un evento ha influido en el comportamiento de las noticias.

Como la región de probabilidad es un concepto abstracto es conveniente iniciar su comprensión desde una perspectiva intuitiva para luego profundizar plenamente en su definición matemática, luego, basados en ella construir un algoritmo que permita calcular dicha región. Una vez diseñado e implementado el algoritmo se expone un ejemplo de aplicación en el contexto costarricense para terminar con una discusión sobre los resultados obtenidos.

La región de probabilidad como modelo de comportamiento de las noticias

La palabra región evoca un lugar delimitado por una frontera, donde los elementos internos tienen ciertas características que los distinguen de aquellos que se encuentran por fuera. Cuando se habla de región de probabilidad, por tanto, se trata una idea similar, pero asociada a un nivel de confianza determinado. A continuación, se exponen tanto el concepto de región de probabilidad como su definición matemática.

Concepto

En algunas ocasiones, cuando solicitamos la dirección a una persona que sabe dónde se encuentra un lugar, pero no sabe explicar la dirección exacta suele usar expresiones como: "cerca de...", "ahí en...", "por ...", por ejemplo, si uno pregunta: "¿dónde puedo encontrar un plato?", puede que reciba la siguiente respuesta: "-¡ahí, en la cocina!", con esta indicación el interlocutor define un espacio físico donde intuye que puede estar el objeto buscado, o más exactamente, donde existe la mayor probabilidad de encontrarlo, a pesar de que a ciencia cierta no lo pueda localizar con precisión, por tanto, no es necesario buscar por toda la casa, sino únicamente en la cocina.

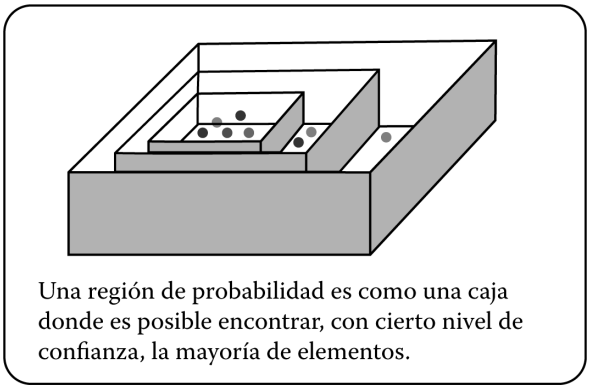

Figura 3.1. Modelo intuitivo de la región de probabilidad
Fuente: Elaboración propia

De aquí, que, entonces, una región de probabilidad puede imaginarse como una especie de caja, en la que se encuentran objetos de diferente naturaleza, la cual al ir haciéndose más pequeña permite ubicar dónde se concentran aquellos que son de mayor interés (ver figura 3.1). Otra manera de conceptualizar la región como una diana para dardos, como el círculo donde haya la mayor parte de blancos hechos.

Aunque en las ciencias naturales el uso de las regiones de probabilidad no es nuevo, ya se mencionó el caso del orbital del electrón (Albright, Burdett y Whangbo, 2013); en las ciencias sociales, el concepto suele estar más asociado a la región de rechazo en el contraste de hipótesis estadísticas (Elorza, 2008; Figueiredo, *et al.*, 2013; Hernández, 2015; Szucs y Ioannidis, 2017; Batanero, López-Martín, Gea y Arteaga, 2018), pero, hasta donde se puede constatar en la literatura consultada, es raramente utilizado como un delimitador o frontera. Por tanto, la presente propuesta puede ser de gran provecho para el estudio de fenómenos con alta aleatoriedad como el caso de las proporciones de noticias de temática política y un pequeño aporte a la teoría del consumo de noticias políticas desde un punto de vista cuantitativo.

Ahora bien, en términos del presente estudio, se define el umbral como el límite de la "caja" más pequeña, dado un nivel de significancia determinado, que contenga a la mayor parte de proporciones calculadas de noticias políticas, es decir, el umbral corresponde, entonces, al mínimo de los máximos para cierto nivel de confianza. Esta definición asegura la existencia de este valor (Khalaf, Kumar y Baladvidhya 2017). Partiendo de esta definición de umbral, se diseña un algoritmo que pueda calcular dicho valor dentro del conjunto de proporciones de noticias políticas.

Construcción de la región de probabilidad

Crear la región de probabilidad consiste en encontrar el mínimo valor del umbral tal que la cantidad de proporciones de noticias políticas debajo de este sea mayor que aquellas que quedan por afuera, lo cual, a su vez requiere asegurar un nivel de significancia relevante (al menos 95 %), para lograrlo es necesario reformular el problema, como se explicará más adelante. También se debe aclarar que, a partir de este momento, se utiliza indistintamente los términos tasa y proporción, puesto que la primera es, en realidad, una proporción calculada a lo largo de un intervalo de tiempo (Gómez, 2014). En las siguientes secciones se explica la definición matemática de la región de probabilidad y el algoritmo utilizado para su construcción.

Definición matemática y criterio de decisión

Primero, considere D como el conjunto de todas las tasas o proporciones de noticias de índole política calculadas en un intervalo de tiempo determinado. Luego, para definir un umbral se divide a D en dos subconjuntos mutuamente excluyentes, sean estos: A y B; para ello se puede encontrar un valor real *m*, tal que A contenga todos los número menores que *m* en D y B los mayores o iguales.

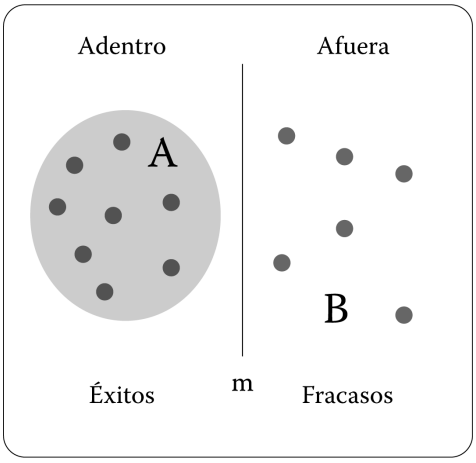

Figura 3.2. División dicotómica del conjunto de proporciones
Fuente: Elaboración propia

Esta división de D permite enfocar el problema de una manera diferente, dado que ahora su naturaleza es completamente dicotómica, como se ilustra en la figura 3.2, o sea, para cada uno de los elementos de D existen solamente dos eventos posibles: la tasa *x* pertenece al subconjunto A (que podemos denominar como éxito), o bien, que la tasa *x* no pertenece (fracaso). En otras palabras, se ha transformado el problema de los datos con una distribución de la cual no se tiene idea alguna a otro de carácter binomial (Beichelt, 2016).

Falta definir el criterio de decisión de la prueba binomial con el fin de ir probando diferentes valores del umbral en forma iterativa, para un nivel de confianza definido

con anterioridad. Considere como hipótesis nula que tanto la proporción de éxitos como la proporción de fracasos son iguales (0,50), por lo cual el valor de *m* debe tener la propiedad de hacer que se rechace dicha hipótesis; sobre esta premisa se construye el algoritmo de búsqueda del umbral. Dada la naturaleza del problema se requiere de aprovechar la potencia de la recursividad y los métodos numéricos.

Algoritmo de búsqueda del máximo límite más pequeño

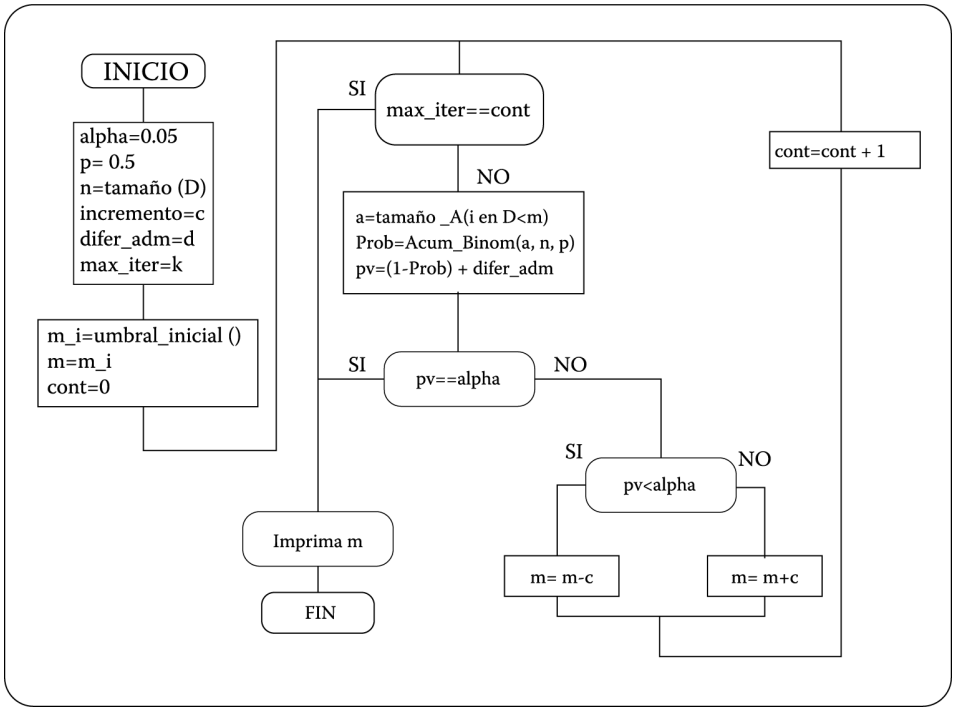

Figura 3.3. Algoritmo simplificado para el cálculo del umbral
Fuente: Elaboración propia

El algoritmo, como se puede observar en la figura 3.3, se inicializa fijando el tamaño del conjunto completo de tasas calculadas (designado con *n*); un incremento *c*, o valor que irá disminuyendo o aumentando el umbral *m*, y, dado que los resultados de las iteraciones pueden resultar con algunas diferencias en decimales, se define un valor de control denominado diferencia admisible, según un nivel de precisión deseado. También se fija la proporción *p* de prueba del criterio binomial (la hipótesis nula).

Se inicia el algoritmo con un valor del umbral o límite: *m_i*, el cual será al principio *m*, con dicho valor se calculará el tamaño del conjunto A, denotado con *a,* en otras palabras, el número de tasas que son menores a *m*.

La primera iteración calculará la probabilidad acumulada binomial (*Acum_ Binom*) con los parámetros p (proporción a comparar: 0,50), *a* y *n*. La probabilidad acumulada se utiliza para el cálculo de la región de rechazo, al restar a la unidad dicha probabilidad, el resultado será nuestro valor *pv* o p-valor.

Si *pv* más la diferencia admisible es igual al nivel de significancia o *alpha* (0,05), se habrá localizado el umbral deseado: la frontera de la región de probabilidad que se busca. Por tanto, el algoritmo se detiene. Si ocurre lo contrario, pueden existir dos casos, el primero es que el valor *pv* sea muy inferior en magnitud al *alpha*. Esto implica que existe otro umbral *m* más pequeño que el anterior, dado esto, se le resta el incremento y se realiza todo el procedimiento nuevamente hasta que *pv* sea igual al *alpha*, o bien se cumpla un número máximo de iteraciones. El segundo caso posible es que *pv* sea mucho mayor al *alpha*, eso implica que el umbral en realidad es más alto del que se calculó inicialmente, por lo que, en lugar de restar el incremento, se suma y se ejecuta recursivamente hasta que se alcance la igualdad del *pv* con *alpha.*

Ahora, el algoritmo requiere de un primer valor del umbral para iniciar la búsqueda, aunque en teoría podría ser cualquier valor del conjunto, su promedio, mediana o alguno de los cuartiles. Una metodología que nos asegura el nivel de confianza mínimo es la desigualdad de Chebyshev (Nishiyama, 2018).

Según el teorema, para una variable aleatoria Y, con función de probabilidad *p(y)*, media y varianza finita $\sigma^{\wedge 2}$ se cumple que, para un k real, *k>0:*

$$P(|(Y-\mu)| < k\sigma) \leq 1-\frac{1}{k^2} \qquad (1)$$

Esto se puede expresar de la siguiente manera:

$$P(\mu - k\sigma < Y < \mu + k\sigma) \leq 1 - \frac{1}{k^2} \quad (2)$$

En particular, tomando a k=μ/σ se tiene que:

$$P(0 < Y < 2\mu) \leq 1 - \frac{\sigma^2}{\mu^2} \quad (3)$$

Es decir que sin importar la distribución de probabilidades que sigan los datos, se puede acotar los valores a

$$m = 2\mu \quad (4)$$

con una probabilidad de

$$1 - \frac{\sigma^2}{\mu^2} \quad (5)$$

por lo tanto, se tiene un límite inicial, el cual se puede aproximar con la media muestral

$$m = 2\overline{x} \quad (6)$$

En resumen, lo que se busca es una manera de acotar la región, dado un nivel de confianza (en este caso 95 %), de modo que se contenga en su interior la mayor proporción posible de valores.

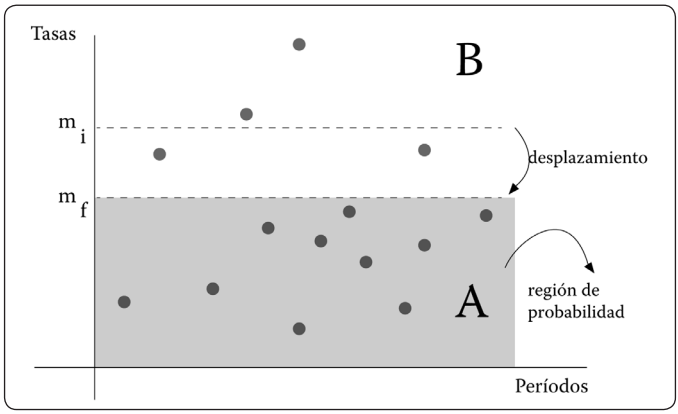

Figura 3.4. Delimitación de la región de probabilidad dada por el algoritmo
Fuente: Elaboración propia

En la figura 3.4, se muestra el funcionamiento del algoritmo y, como con un umbral inicial, se va delimitando la región de manera más precisa con cada iteración, dejando por fuera de ella los elementos del conjunto B, los cuales pueden ser clasificados como eventos de alta actividad política. Los elementos dentro de la región, el conjunto A, pueden ser clasificados como "regulares".

Una vez definido el algoritmo de la región, se procedió a implementar el concepto en dos medios costarricenses. Para ello, se desarrolló un mecanismo de cálculo de las proporciones de noticias de índole política, esto se detalla en la siguiente sección.

Casos de estudio: *Nacion.com* y *CRHoy.com*

Como el primer paso de la implementación consiste en calcular las proporciones de noticias políticas, se requirió poseer una base de datos de noticias con el fin de extraer la información necesaria para realizar el cálculo preliminar de las tasas y, de esta manera, comparar el procedimiento computacional con el realizado por los expertos del CICOM. Se seleccionaron dos medios noticiosos *Nacion.com* (LN) y *CRHoy.com* (CRH) cuya codificación por parte de los investigadores iniciales se encontraba más avanzada.

Descripción de la base datos

A partir de las noticias codificadas bajo el criterio de los expertos mencionados anteriormente, se construyó una base de datos secundaria con los atributos necesarios para el conteo de noticias de interés y de esta manera realizar el cálculo de las proporciones. Los atributos de los registros seleccionados para el presente estudio incluyeron nombre del medio, tema específico, tipo de noticias (público, no público), tema general (disponibles a partir de febrero del 2018) y fecha de publicación. Del total de 7 900 noticias pertenecientes a la base de datos, 2 403 corresponden a noticias de CRH y 2 754 a LN.

Es importante aclarar que la porción de noticias codificadas por los investigadores del CICOM no corresponde a una muestra aleatoria, ni a todas las noticias monitoreadas, por lo que no pueden realizarse inferencias al total de la población. Sin embargo, la cantidad de datos de la muestra permite corroborar la eficacia del algoritmo propuesto.

Emulación del criterio de cálculo de las proporciones y su fórmula

El cálculo de las tasas se realizó emulando el criterio de discriminación utilizado por los investigadores iniciales. Este consistió en delimitar dentro de un rango de fechas (fecha inicial y fecha final) el número de noticias de cada uno de los medios estudiados (LN y CRH), con el fin de contabilizar el total de noticias dentro de ese rango. Luego se procedió a enumerar aquellas que contienen como tema general política y de tipo público. Cuando no era posible tener tema general, se utilizaron las mismas listas de tema específico de los investigadores como filtro en la obtención de la tasa del periodo estudiado.

Una vez calculado el número de noticias de interés público y temática general política, junto con el total de noticias del periodo estudiado, se utilizó la siguiente fórmula para obtener la tasa:

$$C_p = \frac{K_p}{N} * 100 \qquad (7)$$

Con C_p siendo la tasa o proporción de noticias público-político del periodo según el medio (LN y CRH), K_p el número de noticias que son de interés público y su tema general es política y, por último, N el número total de noticias publicadas en el periodo de estudio.

Obtenidas las proporciones, en los mismos intervalos de tiempo que utilizaron los expertos, se procedió a validar el algoritmo, mediante la comparación de los resultados con los datos obtenidos por aquellos. Se realizó una prueba de t-Student (Sánchez, 2015) para probar que no existen diferencias significativas, por medio del siguiente estadístico:

$$t = \frac{|media\ cálculo\ computacional - media\ cálculo\ manual|}{\frac{desviación\ estándar\ cálculo\ computacional}{\sqrt{N}}} \qquad (8)$$

Cuadro 3.1. Medidas estadísticas de las proporciones calculadas computacionalmente en contraste con las obtenidas por los investigadores

Medio	CRH Comput.	CRH Invest.	LN Comput.	LN Invest.
Promedio	25,22	27,45	31,23	33,92
Desv. estándar	14,98	17,39	19,51	20,44
Coefic. variación	59,39 %	63,35 %	62,47 %	60,25 %

Fuente: Elaboración propia

Una vez calculadas las medias y desviaciones estándar correspondientes a cada medio, como puede verse en el cuadro 3.1, se aplicaron las pruebas correspondientes para verificar la presencia de diferencias significativas entre las medias de los datos calculados por el programa y aquellos obtenidos inicialmente por los investigadores del CICOM.

Con un 95 % de confianza, no se encontró evidencia de diferencias significativas entre las medias de los dos grupos de datos tanto para el medio CRH (estadístico t: 0,5568, p-value: 0,5870), como para LN (estadístico: 0,5151, p-value: 0,6150).

Para constatar los resultados anteriores se procedió a graficar los datos de los investigadores del CICOM y los obtenidos computacionalmente. Como se puede observar en la figura 3.5, el cálculo para CRH resultó muy similar, salvo en unos momentos muy puntuales, como enero y abril del 2018, pero casi idéntico a partir de julio del 2018.

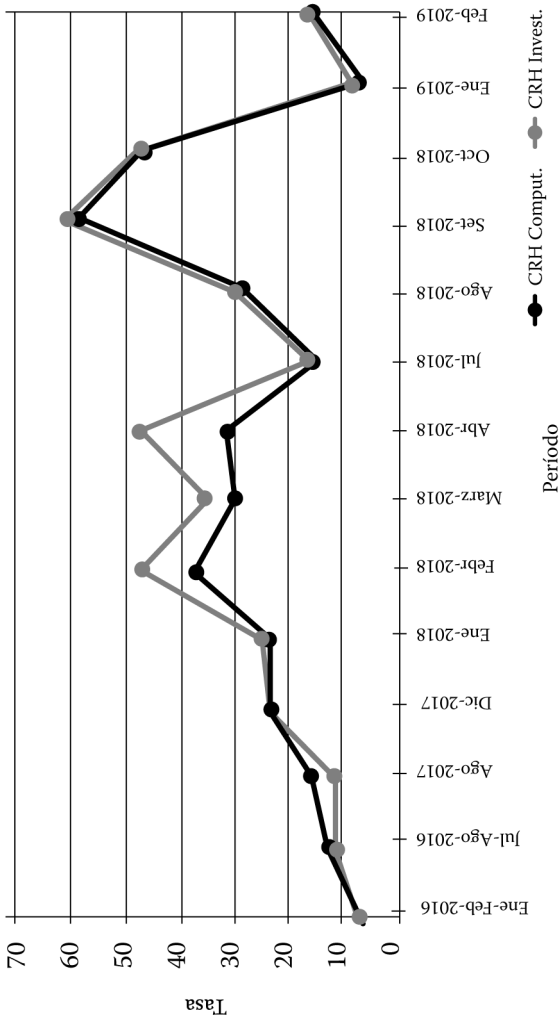

Figura 3.5. Proporciones computacionales y de los investigadores para CRH
Fuente: Elaboración propia, con datos propios y de los investigadores del CICOM

Igualmente, en la figura 3.6, las proporciones obtenidas para LN, no difieren mucho en comparación con la de los expertos. Además, se puede notar que aquellos momentos donde sí existe alguna diferencia esta es menos acentuada en LN que en CRH. También se puede ver que las tasas en ambos medios tienen el mismo comportamiento en lo que se refiere a la los intervalos de variación, siendo los picos más altos los meses febrero-abril del 2018 y setiembre-octubre 2018, elecciones y huelga respectivamente; sin embargo, este hecho no es suficiente prueba para afirmar ni descartar la existencia de un umbral.

Al aplicar pruebas tradicionales como el test de Grubbs, recomendada por Zmuk (2017), no se encontraron evidencias, con un nivel de significancia de 0.05, de que los periodos con tasas altas fueran acontecimientos atípicos (CRH p-value: 0,1172; LN p-value: 0,6636). Esto puede ser debido a que los datos en general no siguen una distribución normal o bien al tamaño de la muestra (Gbenro, 2018). Este hecho plantea entonces la necesidad de enfocar el problema de una manera diferente, es decir la región de probabilidad.

Ahora, una vez validado el procedimiento para las tasas de noticias políticas, se procede a calcular los valores límites de la región. Como lenguaje de programación, se seleccionó Python por contener librerías de procesamiento de datos muy eficientes como *pandas* y *numpy*; para aplicar las pruebas estadísticas de validación, se utilizó la librería *scipy.stats.*

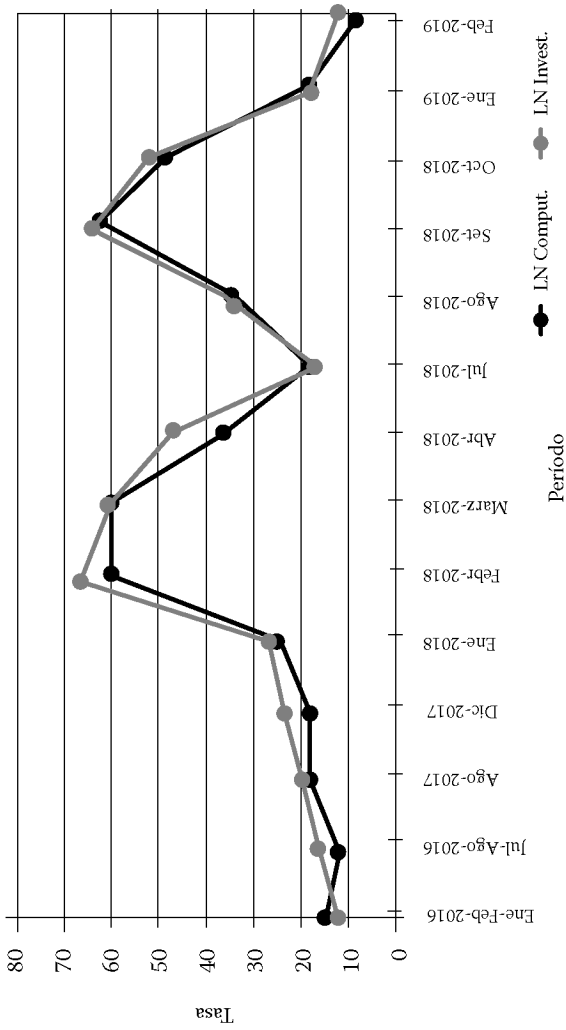

Figura 3.6. Proporciones computacionales y de los investigadores para LN
Fuente: Elaboración propia, con datos propios y de los investigadores del CICOM

Resultados de la aplicación del algoritmo de la región de probabilidad

Las tasas de noticias políticas pueden variar en función del denominador. Es importante hacer el cálculo de la región de probabilidad en diferentes intervalos de tiempo. De esta forma se pueden ubicar eventos influyentes con una mayor precisión, los cuales podrían no ser advertidos cuando se toman periodos más grandes por el problema del denominador. Para una mayor conveniencia en la comprensión del fenómeno se han seleccionado tres periodos: mensual, quincenal y semanal.

Cálculo del umbral mes a mes

Para el periodo mensual, a diferencia de los investigadores del CICOM, se dividieron enero y febrero, así como julio y agosto del 2016, con lo cual se obtuvo una mayor cantidad de datos que permitió mejorar el cálculo del umbral. También se incluyó el mes de mayo 2017, el cual no fue tomado en cuenta dentro de las observaciones de dichos investigadores por cuanto las noticias codificadas se encuentran en un periodo muy corto (una semana); no obstante, la tasa calculada en dicho mes no es despreciable en términos del presente estudio.

Al incluir la información adicional, se obtuvieron dos conjuntos de 17 tasas, uno para LN y otro de CRH. Luego se aplicaron la pruebas Anderson-Darling y Shapiro Wilk para verificar si corresponden a datos normales; pero, según los resultados, LN no sigue la distribución normal, (p-value: Anderson-Darling: 0,01688; Shapiro-Wilk: 0,02134) mientras que CRH sí lo hace (p-value: Anderson-Darling: 0,2854; Shapiro-Wilk: 0,1894). La falta de normalidad en los datos de LN no permite realizar ninguna comparación entre las dos nuevas medias a partir de métodos estadísticos cuyo supuesto sea la normalidad, como el ANOVA (Rodríguez, *et al.* 2018). Si bien, existen métodos no paramétricos como Kruskal-Wallis (Hamada, 2018), tales pruebas indicarían si los grupos de datos siguen una misma distribución, pero no ofrecerían información adicional sobre el umbral buscado.

Ahora, con respecto al estudio de la variabilidad, con los datos adicionales, se obtuvo en CRH un promedio de 24,36 con desviación estándar de 15,20 y en LN la media aritmética fue de 29,06, con una desviación estándar de 18,73. Se puede notar que existe una alta variabilidad respecto a la media dado que los datos de ambas muestras supera el 60 % según el coeficiente de variabilidad (LN: 64,51; CRH: 62,42), esto indica que los valores se encuentran muy dispersos respecto a la media.

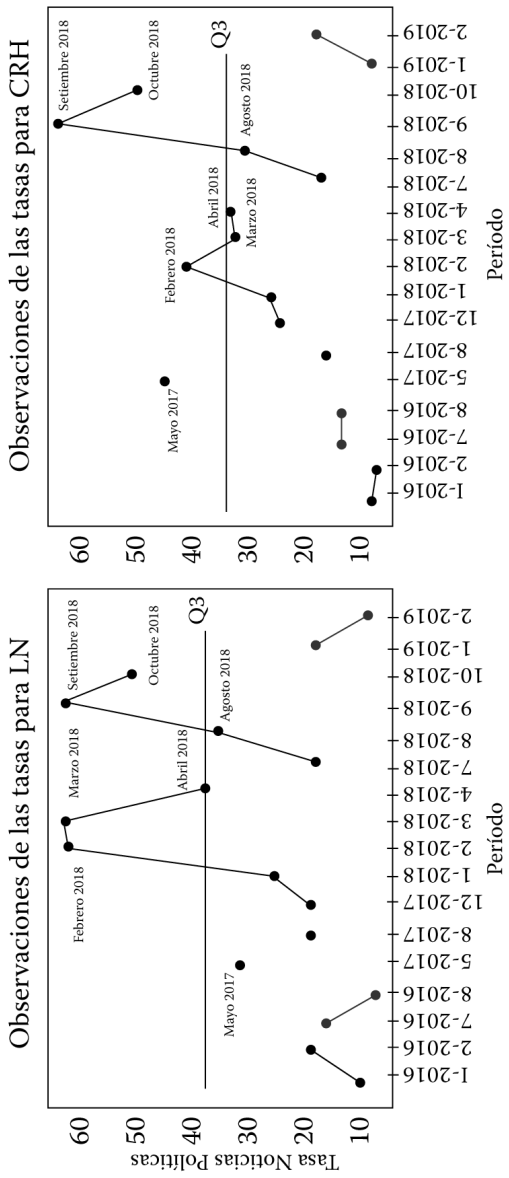

Figura 3.7. *Cálculo de las tasas incluyendo separación de meses para LN y CRH*
Fuente: Elaboración propia

Un hecho remarcable es que el 75 % de los datos (Q3) de CRH resulta menor o igual a 31,13 mientras que para LN posee un tercer cuartil de 36,86, como se puede observar en la figura 3.7. No obstante, no se puede aseverar que estos valores corresponden al mínimo umbral que maximiza el número de proporciones debajo de él. Por otro lado, se puede recalcar el contraste con cálculos de las muestras originales, ya que la inclusión de más datos incrementó la variabilidad, por lo que cabe pensar entonces que dada la dispersión de los datos respecto a la media aritmética, mencionada anteriormente, el promedio no podría ser tampoco el umbral de la región buscada.

Mediante el algoritmo de la región de probabilidad, se calculó, para el conjunto de datos mensuales de LN un umbral o frontera de 37,21. Repitiendo el mismo ejercicio para CRH el resultado fue 31,18; esto significa que, con un 95% de confianza, más de la mitad de las tasas de noticias político-públicas en cada medio estaría por debajo esos valores. Un aspecto importante que destacar es que el umbral calculado para LN es superior al de CRH, lo que parece indicar, al menos en proporción, que existe una mayor cantidad de noticias de corte político por parte del primer medio en comparación con el segundo.

Estos valores confirman la utilidad del algoritmo para calcular la mínima cota superior de la región, el cual, al tomarse como valor de referencia, abre una posibilidad para la detección de momentos de alta actividad en los medios en la temática de política pública. Estos periodos de gran actividad en ambos conjuntos de datos coincidieron precisamente con las elecciones presidenciales y la huelga del sector público en 2018, como puede verse en la figura 3.8. También se puede notar que la huelga y las elecciones tuvieron casi misma importancia dentro de LN. Sin embargo, en CRH, existe una diferencia muy notable en el trato dado a dichos eventos. Este fenómeno motivó a calcular las tasas en periodos más cortos: quincenalmente y cada siete días con el fin de estudiar el comportamiento con mayor detalle, como se verá en las siguientes secciones.

LN:
cot=40.114165, pv=0.049042
cot=39.532800, pv=0.049042
cot=38.951435, pv=0.049042
cot=38.370071, pv=0.049042
cot=37.788706, pv=0.049042
cot=37.207341, pv=0.049042
cot=36.625976, pv=0.143463

Mínima cota admisible a un alfa de 0.05 es 37.207341

CRH:
cot=33.613553, pv=0.049042
cot=33.126400, pv=0.049042
cot=32.639247, pv=0.049042
cot=32.152094, pv=0.049042
cot=31.664941, pv=0.049042
cot=31.177788, pv=0.049042
cot=30.690635, pv=0.143463

Mínima cota admisible a un alfa de 0.05 es 31.177788

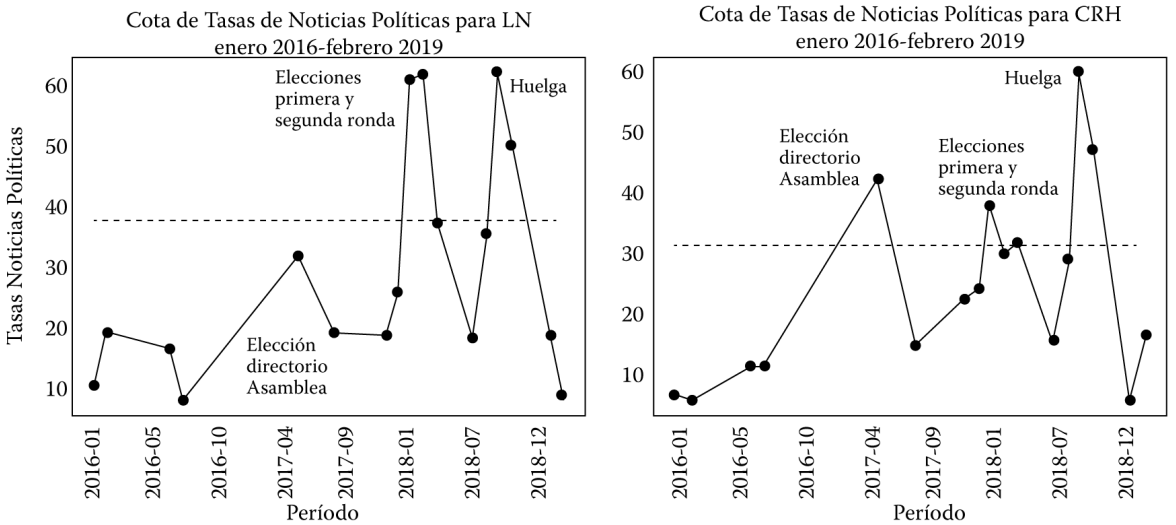

Figura 3.8. Cálculo del umbral mínimo para las tasas mensuales
Fuente: Elaboración propia

Cálculo cada quince días

Las tasas de noticias político-públicas cada quince días, ofrecen un total de 33 observaciones para cada medio; no obstante, cuando se aplicaron las pruebas de normalidad, se constató que las tasas de LN no siguen una distribución normal (Anderson Darling, p-value: 0,007; Shapiro-Wilk p-value: 0,01107). Respecto a los datos de CRH, existe una discrepancia en ambas pruebas aplicadas (Anderson Darling, p-value: 0,1354; Shapiro-Wilk p-value: 0,02407), por lo que simplemente se considera con sospechas de no seguir una distribución normal.

Cuadro 3.2. Medidas estadísticas de las tasas de noticias político público cada quince días

Medio	CRH	LN
Promedio	21,18	28,62
Desv. estándar	15,56	19,79
Coefic. variación	73,47 %	69,15 %

Fuente: Elaboración propia

Igualmente, como se muestra en el cuadro 3.2, la variabilidad de los datos supera el 60 %, por lo que nuevamente la media no es un buen indicador del comportamiento. Se tiene que el 75 % de los datos se encuentran por debajo de 26,8 (CRH) y 42,42 (LN), por lo que se procede a calcular el umbral mínimo para la división de periodos.

El umbral para la LN es de 31,48 y para CRH de 25,83 (ver figura 3.9); el cálculo quincenal permite ver con mayor detalle el comportamiento de los medios ante algunos eventos de carácter político. Se puede notar que ambos medios aumentan la proporción de noticias político-públicas en la segunda semana del mes de enero de 2018, cuya cumbre corresponde a la semana de la primera ronda. Sin embargo, lo que podría llamarse la "cobertura" realizada por LN, parece en proporción ser más amplia que la realizada por CRH, a juzgar por la forma de la curva de los gráficos.

También se pueden notar "fosas" o depresiones pronunciadas, entre periodos de relativa calma y periodos de gran importancia, sin importar si los cálculos son

mensuales o quincenales, como por ejemplo en diciembre 2017, antes de la campaña y las elecciones, además en julio 2018, principalmente antes del fallo de la Sala IV a favor del matrimonio igualitario, junto con las protestas en Nicaragua (agosto, 2018) y sobre todo antes de la huelga del sector público (setiembre-octubre, 2018).

Cálculo cada siete días

Para este tercer ejercicio de cálculo se obtuvieron 62 observaciones por cada medio. Las pruebas realizadas en las muestras confirman que tanto la LN como CRH poseen distribuciones no normales (LN, Anderson-Darling p-value: 0,002, Shapiro Wilk p-value: 0,002; CRH Anderson-Darling p-value: 0,003, Shapiro Wilk p-value: 0,004).

El umbral para LN es de 31,47 mientras que el de CRH es 25,83; la diferencia de estos valores con los respectivos promedios de las muestras es notable (LN media: 23,78, ±19,01; CRH media 29,64, ±21,55). Los coeficientes de variación superan el 70 % en ambos casos, es decir, hay una alta variabilidad en ambos conjuntos de datos. Por otra parte, el 75 % de los datos para LN se encuentra por debajo de 42,88, mientras que para CRH está por debajo de 34,62. Ni la media aritmética ni el tercer cuartil representan el mínimo valor posible para la región de probabilidad, como se ha anotado anteriormente, esto muestra la ventaja que ofrece el algoritmo de la región de probabilidad.

Los cálculos realizados cada siete días proporcionan una visión más detallada del comportamiento para ambos medios. Tanto la figura 3.10, correspondiente a LN, como la figura 3.11 de CRH, muestran que los dos medios, aparentemente, no dan la misma importancia a los acontecimientos políticos. Respecto a la proporción de noticias, para LN en la primera ronda fue de 85,45, mientras que para CRH, llegó a 59,52. Las diferencias entre las proporciones durante la campaña para la segunda ronda de elecciones, de marzo a abril, son más acentuadas para LN, pues se mantuvieron en un rango de 54 a 70, contrario a CRH, que en este periodo no superaron el valor de 40, excepto la primera semana de abril (53,33). En ambos medios la primera ronda tuvo mayor importancia que la segunda; sin embargo, se puede observar que LN publicó, en proporción, mayor número de noticias políticas para la segunda ronda que CRH.

Figura 3.9. Cálculo del umbral mínimo para las tasas cada quince días
Fuente: Elaboración propia

LN:
Mínima cota admisible a un alfa de 0.05 es 31.417374

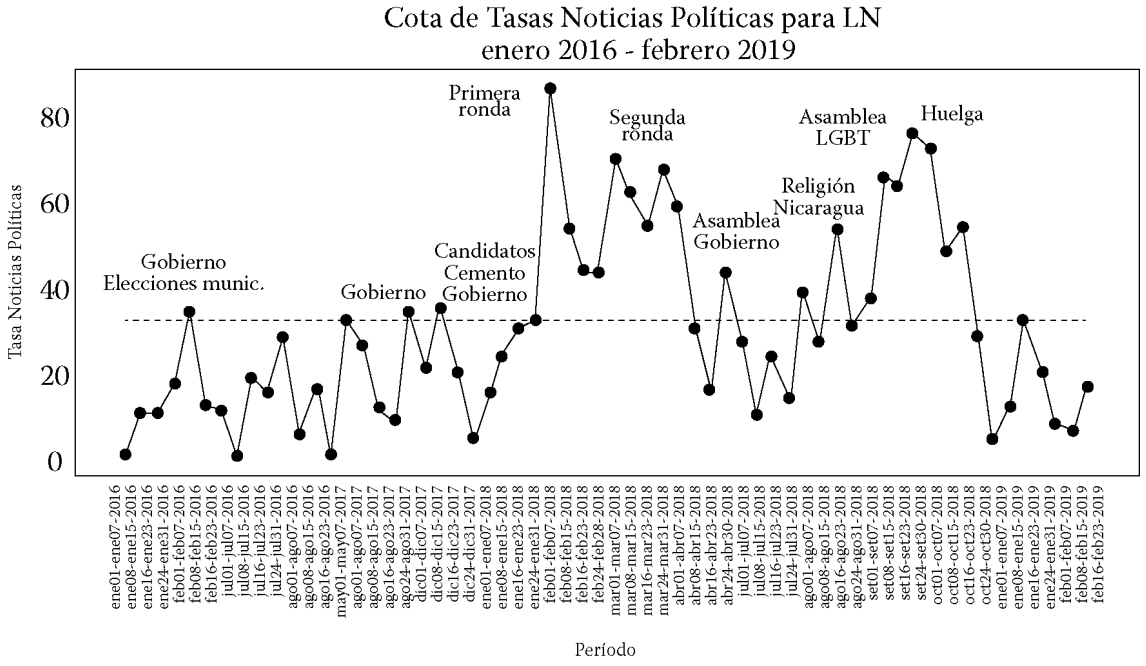

Figura 3.10. Cálculo del umbral mínimo para las tasas cada siete días LN
Fuente: Elaboración propia

Las elecciones presidenciales del 2018 sí influyeron en el aumento de noticias político-públicas en ambos medios, tanto para la primera como para la segunda media; no obstante, no tuvo la misma importancia para los medios noticiosos estudiados.

Otro de los eventos identificados que superaron el umbral se tiene en el periodo preelecciones, específicamente, en la primera semana de febrero de 2016, donde la proporción para LN superó el umbral calculado, con temas relativos a Gobierno y elecciones municipales, mientras que a CRH apenas le interesó tal acontecimiento (no superó el umbral en dicho medio). También, para diciembre de 2017, se identificó el caso de "Cemento chino". Este apenas repercutió en la LN, como se puede notar al compararlo con el límite de la región; el mismo evento en CRH no hizo mella alguna (ver figura 3.11).

Seguido de las elecciones viene un periodo de relativa calma en ambos medios, con un número mayor de "fosas", salvo por un "pico" relativamente importante que corresponde principalmente a las protestas en Nicaragua (julio-agosto, 2018) y en menor grado a la orden de la Sala IV de aprobar el matrimonio igualitario ratificando la resolución de la Corte IDH ocurrido en enero, además de temas varios referentes al Gobierno y a la Asamblea Legislativa.

Otro aspecto importante que considerar es la huelga del sector público de setiembre-octubre 2018. Como se puede observar en la figura 3.11, en CRH este evento superó en "importancia mediática" a las elecciones presidenciales tanto en primera y segunda ronda, mientras que en LN solamente superó a la segunda ronda. El seguimiento dado a la huelga por CRH es mucho mayor en comparación con el realizado por LN según la curva entre setiembre y octubre 2018, pero para CRH aparentemente perdió importancia, dado un descenso más abrupto en la proporción de noticias hacia el final de ese periodo.

La forma de las curvas en los meses de diciembre tanto 2017 como 2018 sugieren una relativa calma en dichos periodos, con una respectiva disminución de la tasa de noticias políticas. Aunque no se puede afirmar nada concluyente hasta contar con la codificación completa de noticias en tales periodos, es promisorio para el estudio de actividad política y su relación con las publicaciones de los medios noticiosos.

CRH:
Mínima cota admisible a un alfa de 0.05 es 25.683794

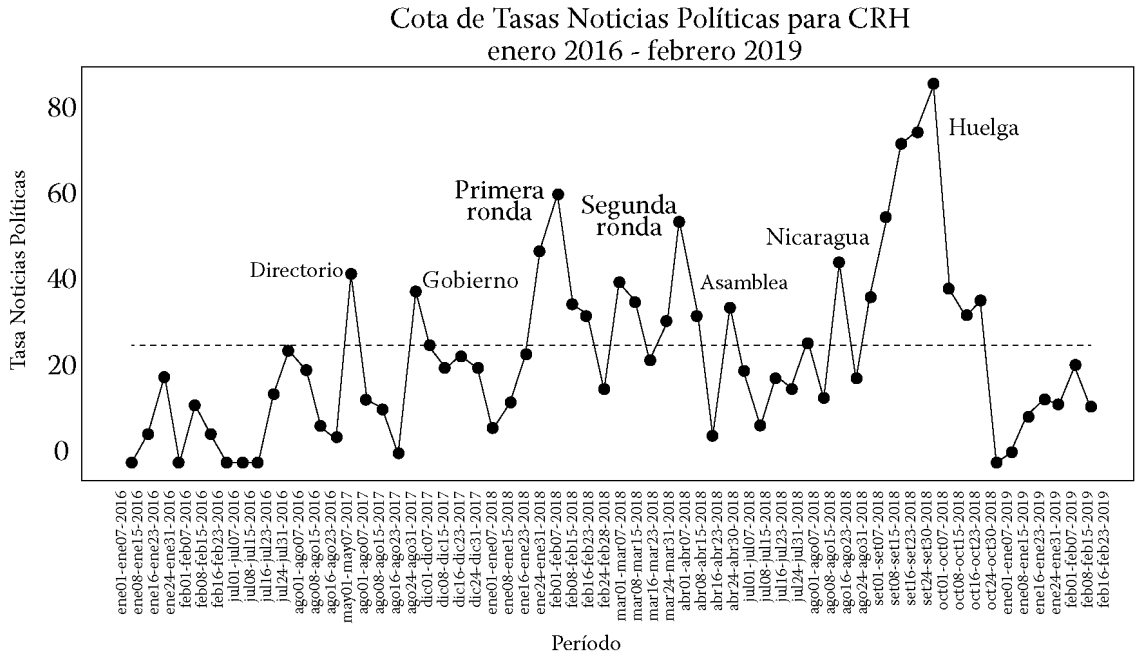

Figura 3.11. Cálculo del umbral mínimo para las tasas cada siete días CRH
Fuente: Elaboración propia

Consideraciones finales

Considerando los resultados obtenidos, existe un valor mínimo en las tasas de noticias de índole política calculadas, tanto para LN como CRH, con una confianza del 95 %, correspondiente a la frontera de la región de probabilidad. La proporción de valores dentro de la región es superior a aquella que se encuentran por fuera, lo que es un gran indicio para corroborar la hipótesis de Siles, Campos y Segura (2018) y la noción de "cuota" sugerida en el capítulo 2 de este libro.

El umbral o frontera de la región de probabilidad funciona también como una herramienta visual para determinar los periodos con una mayor actividad política. Igualmente permite comparar cuál temática fue la que impulsó más el aumento de la tasa de noticias político-públicas en un instante determinado. Sin embargo, cabe aclarar que no es prudente, en este momento, considerar a la región de probabilidad como un detector de datos atípicos, lo cual requiere un estudio comparativo entre periodos de igual longitud de medias y varianzas.

Por otro lado, basado en el comportamiento de noticias detectado, es importante notar que, entre pico y pico, existen depresiones, es decir, la tasa de noticias político-públicas cae por debajo del umbral. Podría, entonces, especularse la existencia de una saturación por parte del consumidor sobre el evento.

Ahora, la existencia de tales decensos entre picos o "cumbres" sí representa una evidencia de periodos de alta actividad política seguidos de baja intensidad. Este resultado concuerda con lo comentado por Tristán y Álvarez (2018). Es importante considerar aquí el hecho que aún existen datos pendientes de codificar, lo cual pueden contribuir a la mejora del cálculo del umbral para delimitar tales periodos.

Respecto a los eventos políticos en general, algunos parecen repercutir más que otros. La huelga del sector público, por ejemplo, repercutió más en el comportamiento de las proporciones que el caso del "cementazo". Tales diferencias se pueden notar al trabajar la región de probabilidad con intervalos de tiempo cortos y esto puede deberse a las diferencias en el denominador que poseen las tasas a lo largo del tiempo. A pesar de ello, el modelo resulta útil para encontrar un punto de referencia con el cual hacer comparaciones entre medios y eventos.

Al estudiar las curvas de los gráficos de evolución de las tasas de noticias políticas, el patrón sugiere que los eventos son tratados con grados de importancia diferente por parte de los medios noticiosos analizados. Para obtener un panorama

más amplio de este fenómeno, sería recomendable realizar el experimento a futuro con todos los medios recolectados.

Queda pendiente un estudio sobre la composición temática de las proporciones de noticias políticas, para determinar si la oferta solamente se concentra en un pequeño grupo de temas; este estudio deberá considerar la composición tanto de aquellas tasas que superen el umbral como de aquellas por debajo de la misma.

En conclusión, el enfoque de la región de probabilidad, dada la naturaleza aleatoria de las tasas de noticias políticas, es una alternativa novedosa para describir su comportamiento. Proporciona un punto de referencia de comparación entre periodos y también entre medios noticiosos, asegurando un nivel de confianza aceptable. Permite realizar el cálculo aprovechando el potencial de la tecnología actual y no requiere ninguna suposición respecto a la distribución de los datos, lo cual ofrece una ventaja frente a aquellos métodos estadísticos que requieren comprobación de los supuestos antes de realizar las pruebas. Con este pequeño aporte al estudio del comportamiento de noticias políticas se espera abrir las puertas a nuevas investigaciones y colaboración entre profesionales de distintas ramas del saber, con miras a la comprensión de los fenómenos sociales en nuestro país.

Referencias bibliográficas

Albright, T., Burdett, J. y Whangbo, M. (2013). *Orbital Interactions in Chemistry* (2nd ed.). New Jersey: Wiley.

Batanero, C., López-Martín, M., Gea, M. y Arteaga, P. (2018). Conocimiento del contraste de hipótesis por futuros profesores de Educación Secundaria y Bachillerato. *Publicaciones, 48*(II): 73-95.

Beichelt, F. (2016). *Applied probability and stochastic processes* (2nd ed.). Sudáfrica: CRC Press.

Crane, H. (2018). *Probabilistic Foundation of Statistical Network Analysis* (1st ed.). Boca Raton: CRC Press.

Elorza, H. (2008). *Estadística para las ciencias sociales, del comportamiento y de la salud* (3ª ed.). Ciudad de México: CENGAGE Learning.

Figueiredo, D., Paranhos, R., da Rocha, E., Batista, M., Da Silva, J., Santos, M. y Marino, J. (2013). When is statistical significance is not significant? *Brazilian Political Science Review, 7*(1), 2013: 31-55.

Gbenro, N. (2018). Using Extreme Value Theory to test for Outliers. Décimo tercera jornada de metodología estadística del Insee, 12-14 de junio de 2018, París, Francia.

Gómez, M. (2014). *Elementos de estadística descriptiva* (4ª ed.). San José: EUNED.

Hamada, C. (2018). Statistical analysis for toxicity studies. *Journal of Toxicologic Pathology. 31*(2018), 15-22.

Hernández, O. (2015). *Elementos de probabilidades e inferencia estadística para ciencias sociales* (2ª ed.). San José: Editorial UCR.

Khalaf, A., Kumar, C. y Baladvidhya, S. (2017). Real Analysis of Real Numbers -Cantor and Dedekind real number structuring. *IOSR Journal of Mathematics (IOSR-JM), 13*(5), Ver. I1 (Sep. - Oct. 2017), 32-40.

Nishiyama, Y. (2018). *Improved Chebyshev inequality: new probability bounds with known supremum of PDF.* Metodologías Estadísticas, Universidad de Cornell. arXiv:1801.10770v2 [stat.ME]

Rodríguez, K, Altamirano, K., Adden, K., Cabraca, V., Mora, L. y Briones, J. (2018). Evaluación de la tensión elástica de papel elaborado a partir de desechos de raquis de palma africana y bagazo de caña. *Ingeniería, 28*(1), 29-40.

Siles, I., Campos, P. y Segura, A. (2018). Sitios costarricenses de noticias en Facebook: ¿Qué "likean", comentan y comparten sus usuarios? *Revista de Ciencias Sociales, 160*(II), 37-55.

Sánchez, R. (2015). t-Student. Usos y abusos. *Revista Mexicana de Cardiología, 26*(1), 59-61.

Szucs, D. y Ioannidis, J. (2017). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Frontiers in human neuroscience, 11*(390), 1-21.

Tristán, L. y Alvarez, M. (2018). "¿Brecha de las noticias?". Una comparación de la oferta y el consumo de contenidos en Nacion.com y CRHoy.com. *Revista de Ciencias Sociales, 160*(II), 57-74.

Zmuk, B. (2017). Speeding problem detection in business surveys: benefits of statistical outlier detection methods. *Croatian Operational Research Review, 8*(2017), 33-59.