

LA LEY DE BENFORD Y LOS DATOS DEL COVID-19 EN BOLIVIA

BENFORD'S LAW AND COVID-19 DATA IN BOLIVIA

Dindo Valdez Blanco¹

Instituto de Estadística Teórica y Aplicada, Universidad Mayor de San Andrés, La Paz -Bolivia

✉ dvaldez@fcpn.edu.bo

Artículo recibido: 2021-07-30

Artículo aceptado: 2021-09-11

RESUMEN

En la actualidad, con los datos de la pandemia COVID-19, existe la duda en creer que los estados o gobiernos estén informando datos confiables y precisos. En Bolivia, en particular, ocurre lo mismo, en vista que el sistema de salud es precario, se duda de la información reportada hasta el día de hoy, tanto en nuevos casos diarios, casos diarios recuperados y fallecidos, así como en los datos acumulados. Por lo tanto, el objetivo del trabajo de investigación radica en documentar si estos conjuntos de datos informados por el sistema de salud del país siguen la ley de Benford. La metodología del trabajo se basa en el procedimiento de un estudio de investigación de bondad de ajuste pues abarca el uso de dos pruebas de bondad de ajuste denominadas el test Chi cuadrado de Bondad de Ajuste y el test de bondad de ajuste de Kuiper. Los datos recopilados provienen de los reportes diarios del Ministerio de Salud del Estado Plurinacional de Bolivia entre el 1 de abril del 2020 y el 14 de julio del 2021. Para determinar si el primer dígito significativo del número diario de casos confirmados con COVID-19 en Bolivia se adecúa a la ley de probabilidad de Benford se realizan las pruebas de bondad de ajuste Chi cuadrado de Pearson y la prueba de Kuiper, en ambos casos se rechaza la hipótesis que los datos se ajustan a la ley de Benford, la diferencia significativa más grande es con el dígito 1, este hecho sugiere que existe una subestimación en los reportes diarios de casos confirmados.

Palabras clave: *Bondad de ajuste, Prueba Chi Cuadrado, Prueba de Kuiper, Análisis de datos.*

ABSTRACT

At present, with the data from the COVID-19 pandemic, there is doubt in believing that states or governments are reporting reliable and accurate data. In Bolivia, in particular, the same happens, given that the health system is precarious, the information reported to date is doubted, both in new daily cases, daily cases recovered and deaths, as well as in accumulated data. Therefore, the objective of the research work is to document whether these data sets reported by the country's health system follow Benford's law. The work methodology is based on the procedure of a goodness-of-fit research study, since it involves the use of two goodness-of-fit tests called the Chi-square Goodness-of-Fit test and the Kuiper goodness-of-fit test. The data collected comes from the daily reports of the Ministry of Health of the Plurinational State of Bolivia between April 1, 2020 and July 14, 2021. To determine if the first significant digit of the daily number of confirmed COVID-19 cases in Bolivia conforms to Benford's law of probability, Pearson's Chi-square goodness-of-fit tests and Kuiper's test are performed, in both cases the hypothesis that the data conform to Benford's law is rejected, the difference being more significant large is with the digit 1, this fact suggests that there is an underestimation in the daily reports of confirmed cases.

Keywords: *Goodness of Fit, Chi Square Test, Kuiper Test, Data analysis.*

¹ Maestría en Ciencias Estadísticas, Licenciado en Estadística. Profesor de Estadística - Universidad Mayor de San Andrés. ORCID: 0000-0003-0704-0980

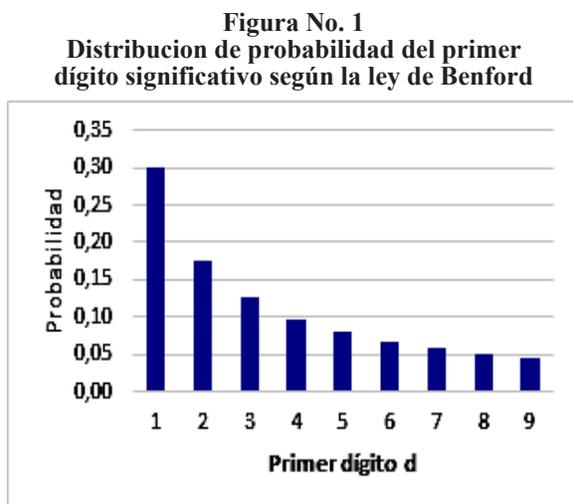
INTRODUCCIÓN

La ley de Benford establece que en muchos conjuntos de datos numéricos que ocurren naturalmente, el primer dígito de los datos no tiene la misma probabilidad de ser 1,..., 9, como cabría esperar, sino que se aproxima bastante por la ley logarítmica:

$$P(d) = \log\left(1 + \frac{1}{d}\right) \quad ; \quad d = 1,2, \dots, 9$$

Fuente: Newcomb (1881)

Donde d es el primer dígito del dato numérico y $P(d)$ es la probabilidad que el dato cuantitativo tenga como primer dígito significativo d . La Figura No. 1 muestra los valores de estas probabilidades.



Fuente: Elaboración propia en base a la distribución de probabilidad logarítmica (Newcomb, 1881)

La ley de Benford fue descubierta por primera vez por Simon Newcomb en su trabajo de 1881 en el *American Journal of Mathematics*. Benford (1938) redescubrió la ley en *Proceedings of the American Philosophical Society* y se le atribuyó el mérito. Descubrió que esta ley logarítmica era bastante precisa en muchas circunstancias; por ejemplo, las cantidades declaradas de impuestos, las longitudes de los ríos, los precios de las

acciones, las constantes universales en física química, el número de habitantes de las grandes ciudades y muchas otras tablas de datos numéricos.

No todos los conjuntos de datos siguen la ley de Benford. Por ejemplo, los números de teléfono de una ciudad determinada no siguen dicha distribución probabilística porque el código de área es el mismo número.

La ley de Benford también se puede utilizar para detectar la manipulación de los estados financieros. Incluso se puede utilizar para detectar fraudes (impuestos, juegos de azar).

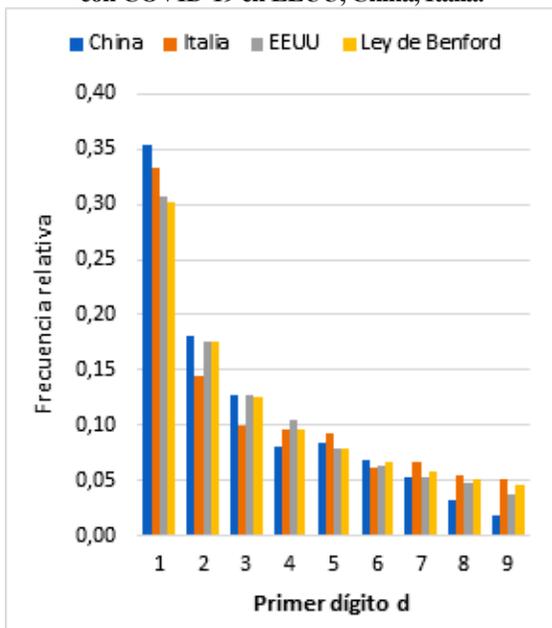
METODOLOGÍA

La metodología del trabajo se basa en el procedimiento de un estudio de investigación de bondad de ajuste pues baraca el uso de dos pruebas de bondad de ajuste denominadas el test Chi-Cuadrado de Bondad de Ajuste y el test de bondad de ajuste de Kuiper. Los datos para el estudio corresponden a los reportes diarios de casos confirmados de COVID-19 en Bolivia realizados por el ministerio de salud y que se encuentran disponibles en el repositorio de datos COVID-19 del Centro de Ciencias e Ingeniería de Sistemas (CSSE, 2121) de la universidad Johns Hopkins Situada en Baltimore, Estados Unidos.

La ley de Benford y los datos del COVID-19

Los conjuntos de datos del COVID-19 reportados en Estados Unidos, China e Italia tienden a ajustarse muy bien a la ley de Benford. La Figura No. 2 muestra el gráfico de la distribución del número de casos diarios confirmados en estos países (Koch y Okamura, 2020).

Figura No. 2
Distribución del primer dígito del número diario de casos confirmados con COVID-19 en EEUU, China, Italia.



Fuente: Elaboración propia en base a los datos disponibles en el repositorio del Centro de Ciencias e Ingeniería de Sistemas (CSSE, 2121).probabilidad logarítmica (Newcomb,1881)

Prueba de bondad de ajuste para la ley de Benford

La prueba de bondad de ajuste más común es la prueba Chi-cuadrado de bondad de ajuste:

$$\chi^2 = \sum_{d=1}^9 \frac{(f_d - e_d)^2}{e_d}$$

Dónde f_d denota la frecuencia observada de los dígitos y e_d es la frecuencia esperada de cada dígito según la ley de Benford.

Prueba de bondad de ajuste de Kuiper

Al aplicar la prueba Chi-cuadrado generalmente no se admite la distribución de Benford, esto ocurre porque la prueba de Chi-cuadrado es una prueba asintótica y tiende a rechazar la significación estadística incluso para pequeñas diferencias. Es por esta razón que es preferible aplicar la prueba de Kuiper definida como:

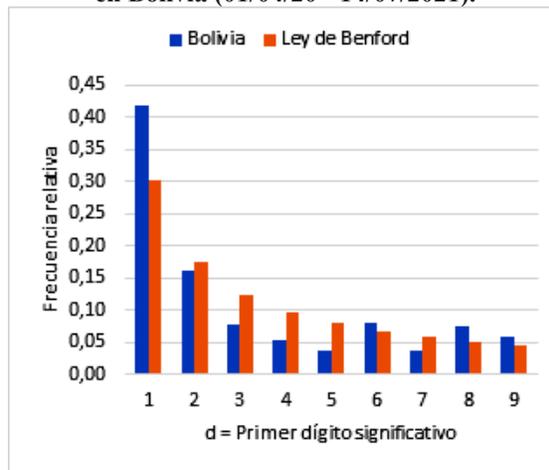
$$T = (D_n^+ + D_n^-) \left[\sqrt{n} + 0.155 + \frac{0.24}{\sqrt{n}} \right]$$

Donde $D_n^+ = \sup(F_d - E_d)$; $D_n^- = \sup(E_d - F_d)$ con F_d y E_d representan las frecuencias acumuladas para el primer dígito de los datos observados y esperados según la ley de Benford.

Evaluación de la ley de Benford y los datos del COVID-19 reportados en Bolivia

Los datos analizados para Bolivia provienen de los reportes diarios del Ministerio de Salud entre el 1 de abril del 2020 al 14 de julio del 2021.

Figura No. 3
Distribución del primer dígito del número diario de casos confirmados con COVID-19 en Bolivia (01/04/20 - 14/07/2021).



Fuente: Elaboración propia en base a los datos disponibles en el repositorio del Centro de Ciencias e Ingeniería de Sistemas (CSSE, 2121).probabilidad logarítmica (Newcomb,1881)

Los datos analizados corresponden al número diario de casos confirmados. El cuadro indica la distribución de frecuencias relativas del primer dígito de los datos observados comparados con la ley de Benford.

Se observa un número elevado para la frecuencia relativa del primer dígito 1 en relación a lo que

se esperaría con la ley de Benford.

El Cuadro No.1 muestra los resultados al aplicar las pruebas de bondad de ajuste Chi Cuadrado y Kuiper.

Cuadro No. 1.
Prueba de bondad de ajuste de los casos diarios confirmados de COVID-19 en Bolivia respecto a la ley de Benford

País	N	Test Chi cuadrado	Test de Kuiper
Bolivia	470	61.177*	3.495*

* La prueba es significativa al 5%

Fuente: Elaboración propia en base a los datos disponibles en el repositorio del Centro de Ciencias e Ingeniería de Sistemas (CSSE,2021).

En ambos casos se rechaza que los datos del número diario de casos confirmados con COVID-19 en Bolivia correspondan con la ley de Benford, esto es inquietante, en vista que en países como Estados Unidos si se corresponde con dicha ley de probabilidad.

DISCUSIÓN

En el trabajo realizado por Koch y Okamura (2020) se concluye que los datos diarios de casos confirmados por COVID-19 durante la gestión 2020 reportados por el CSSE de la Universidad Johns Hopkins respecto a los países de Italia, Estados Unidos y China se adecúan a la Ley de Benford al realizar las pruebas de bondad de ajuste Chi cuadrado y Kuiper, sin embargo, en el análisis realizado con los datos reportados en Bolivia por el mismo centro de monitoreo de la Universidad

Johns Hopkins, se concluye que los datos no corresponden a la distribución de probabilidad de Benford. Este hecho puede tener diversos motivos que es necesario indagar en futuras investigaciones.

CONCLUSIÓN

Los datos analizados en Bolivia con respecto a la ley de Benford a primera vista no se corresponden. Esto no debería ser una sorpresa, en vista de la precariedad del sistema de salud del Estado Boliviano. Una cosa a tener en cuenta es que las frecuencias de los dígitos no disminuyen estrictamente a medida que el dígito aumenta para todas estas categorías, por ejemplo, un 6 es más común que un 5 o 4.

Al aplicar las pruebas de bondad de ajuste Chi-cuadrado y Kuiper la diferencia más significativa radica en el dígito 1, esto sugiere que los reportes del Ministerio de Salud pueden tener algún indicio de fraude, sin embargo, a medida que se reportan más y más datos de COVID-19, es posible que los datos se ajusten de mejor manera a la ley de Benford para ver si están reportando datos precisos y completos. Aunque es posible que por el sistema de información que se tiene y los recursos para realizar pruebas a tantas personas, los datos en Bolivia no sean los verdaderos y se está subestimando el reporte diario de casos nuevos confirmados de COVID-19.

REFERENCIAS BIBLIOGRÁFICAS

- Balsari, S.; Buckee, C.; Khanna, T. Which (2020) «COVID-19 Data Can You Trust?» Harvard Business Review, 2020.
- Benford, Frank (1938). «The Law of Anomalous Numbers». American Philosophical Society 78 (4): pp. 551-572.
- Cho, T.W.; Gaines, B.J. (2007) «Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance». Am. Stat. 61, pp. 218–223.
- Coronavirus Resource Center (2021). «COVID-19 Global Cases by the Center

- for Systems Science and Engineering (CSSE) at Johns Hopkins University». <https://coronavirus.jhu.edu/map.html>.
- Durtschi, C.; Hillison, W.; Pacini, C. (2004) «The Effective Use of Benford's law to assist in Detecting Fraud in Accounting Data». *JFAR*, 5, pp. 17–34.
- Goodman, Q. (2016) «The promises and pitfalls of Benford's law». *Significance*, 13, pp. 38–41.
- Koch, C.; Okamura, K. (2020) «Benford's Law and COVID-19». https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3586413.
- Kuiper, N. H. (1960). «Tests concerning random points on a circle». *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Series A*. 63: pp. 38–47.
- Lee, K.B.; Han, S.; Jeong, Y. (2020) «COVID-19 flattening the curve, and Benford's law». *Phys. A* 2020.
- Newcomb, Simon (1881). «Note on the Frequency of Use of the Different Digits in Natural Numbers». *American Journal of Mathematics* 4 (1): pp. 39-40.
- Roukema, B.F. (2014) «A first-digit anomaly in the 2009 Iranian presidential election». *J. Appl. Stat.* 41, pp. 164–199.
- World Health Organization. (2020) «Coronavirus Disease (COVID-19) Outbreak». WHO: Geneva, Switzerland, <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>.