

[Artículo de opinión / *Opinion piece*]

## Cinco preguntas para precautelar la honestidad estadística en la sociedad actual.

### *Five questions to safeguard statistical honesty in today's society.*

Juan Carlos Abril<sup>1</sup>  María de las Mercedes Abril<sup>2</sup> <sup>1-2</sup> Instituto de Investigaciones Estadísticas de la Facultad de Ciencias Económicas de la Universidad Nacional de Tucumán\*

\*Correspondencia: jabril@herrera.unt.edu.ar

Fecha de recepción: 3 de abril de 2021 -Fecha de aceptación: 17 de mayo de 2021

### RESUMEN

El presente artículo de opinión tiene el propósito de plantear una estrategia metodológica general para evitar la deshonestidad derivada del engaño o del desconocimiento estadístico. Para ello, presentamos un recorrido por algunos eventos históricos que destacan la importancia de la estadística aplicada. A menudo ocurre que los no expertos tienden a confundir correlación con causalidad. Cualquier relación estadística, por fuerte y sugerente que sea, no puede establecer una conexión causal. Nuestras ideas sobre la causalidad deben provenir de fuera de la estadística; en última instancia, de alguna teoría que se refiera al área de conocimiento de donde provienen los datos. Para identificar estos problemas, basados en la experiencia, proponemos una alternativa metodológica simple, que consiste en responder a cinco preguntas clave sobre el origen, la cantidad, la variabilidad, el contexto y el sentido de los datos. Las reflexiones que se extraen tienen importantes implicaciones epistemológicas para la estadística aplicada.

**Palabras clave:** ciencia, estadística aplicada, deshonestidad estadística.

### ABSTRACT

This study has the purpose of proposing a general methodological strategy to avoid dishonesty derived from deception or statistical ignorance. To do this, we present a tour of some historical events that highlight the importance of applied statistics. It is often the case that non-experts tend to confuse correlation with causality. Any statistical relationship, strong and suggestive as it may appear, cannot establish a causal connection. Our ideas about causality must come from outside of statistics, ultimately from some theory that refers to the area of knowledge from which the data comes. To identify these problems, based on experience, we propose a simple methodological alternative, which consists of answering five key questions about the origin, quantity, variability, context and meaning of the data. The reflections extracted have important epistemological implications for applied statistics.

**Keywords:** science, applied statistics, statistical dishonesty.

## INTRODUCCIÓN

Muchos piensan que la Estadística como ciencia es un campo exclusivo de quienes la estudiaron a profundidad. Pero también hay que tener en cuenta que existe bastante gente que acumula y combina cifras con más o menos suerte y lógica. El método estadístico permite, por su aparente facilidad, todo tipo de abusos; muchas veces sirve para tapar más de un sofisma bajo el disfraz de una precisión engañosa. Con la mejor buena fe del mundo, los espíritus se sorprenden con el atractivo de sus seducciones. Constantemente, periódicos y publicaciones diversas lanzan información repleta de cifras, destinada a determinar cuestiones como la extensión de las tierras cultivadas, el avance de la vegetación, el efecto de las vicisitudes atmosféricas, la consistencia y calidad de los cultivos, la producción industrial, por nombrar algunas. Sin embargo, estas cifras deben aceptarse con mucha cautela. Para el lector, el peligro especial que presenta la Estadística es que sus resultados se expresan de manera precisa, porque las operaciones matemáticas son exactas después de todo.

Los primeros desarrollos estadísticos comenzaron desde el principio de la humanidad, como aparece, incluso, en la misma Biblia. Alexandre Moreau de Jonnés (1829) indicó que ha estado presente desde los tiempos más remotos, inscrito en el Pentateuco con el nombre *arithmi*, que significa números.

La mayor parte del trabajo de un estadístico en la sociedad actual se genera a partir del interés en la estructura conjunta de un par (o quizás más) de variables: podemos describir esto como el problema de las relaciones estadísticas. Existe un campo de interés completamente diferente que se refiere a las relaciones entre variables de una clase estrictamente funcional, como las clásicas en física; este tema es de interés estadístico porque las variables relacionadas funcionalmente están sujetas a errores de observación. A esto lo llamamos «el problema de las relaciones funcionales». Sin embargo, en lo que sigue solo se considerará «el problema de las relaciones estadísticas simples».

Dentro del campo de las relaciones estadísticas, debe hacerse una distinción adicional útil. Puede que nos interese la interdependencia entre varias de nuestras variables o la dependencia de una o más variables hacia otras. Por otro lado, hay situaciones en las que solo interesa la dependencia. La relación entre la producción de cultivos y la cantidad de lluvia es un ejemplo donde las consideraciones no estadísticas hacen que la cantidad de lluvia “cause” la producción de los cultivos, cuando estamos bastante seguros de que la producción de cultivos no afecta la cantidad de lluvia que cae; por lo tanto, medimos la dependencia de la producción de la cantidad de lluvia caída. Aquí nos referimos a la cantidad de lluvia como «variable de entrada» o «regresor».

No existe una distinción clara en la terminología estadística para las técnicas apropiadas para este tipo de problemas esencialmente diferentes. Por ejemplo, si estamos interesados en la interdependencia de dos variables con los efectos de otras variables eliminadas, usamos el método llamado «correlación parcial», mientras que si estamos interesados en la dependencia de una sola variable de un grupo de otras, usamos el término «correlación

múltiple». Sin embargo, en principio, es cierto que el estudio de la interdependencia conduce a la teoría de la correlación, mientras que el estudio de la dependencia conduce a la teoría de la regresión. La teoría de la correlación fue desarrollada inicialmente a principios del siglo XX por Karl Pearson (1900) y George U. Yule (1907, 1926).

Enfatizamos: cualquier relación estadística, por fuerte y sugerente que sea, no puede establecer una conexión causal; nuestras ideas sobre la causalidad deben provenir de estadísticas externas; en última instancia, de alguna teoría que se refiera al área de conocimiento de la que provienen los datos. Incluso en el ejemplo simple de la producción de cultivos y la cantidad de lluvia presentada anteriormente, no tenemos razones estadísticas para descartar la idea de que la lluvia dependa de la producción de cultivos. Incluso si la lluvia y la producción de cultivos estuvieran en perfecta correspondencia funcional, no deberíamos pensar en revertir la conexión causal “obvia”. Debemos tener cuidado al concluir relaciones causales que podrían no existir. No necesitamos entrar en las consecuencias filosóficas de esto; para nuestros propósitos, solo debemos repetir que las relaciones estadísticas, de cualquier tipo, no pueden implicar lógicamente causalidad.

También debemos señalar que personalidades altamente intelectuales consideraban a la Estadística como una parte esencial de la vida cívica y democrática de cualquier sociedad.

George Bernard Shaw (1856-1950), Premio Nobel de Literatura en 1925 y quien fuera uno de los fundadores de The London School of Economics and Political Science, expuso brillantemente este punto en el prefacio de su libro “El Dilema del Doctor” (1906), donde escribió:

Inclusive estadísticos entrenados frecuentemente fallan en apreciar el grado en el cual la estadística está viciada por supuestos sin constancias hechos por sus intérpretes [...] Es fácil probar que usar sombreros de copa y llevar paraguas agranda el pecho, prolonga la vida, y confiere cierta inmunidad con respecto a las enfermedades [...] Un grado universitario, un baño diario, la posesión de treinta pares de pantalones, el conocimiento de la música de Wagner, un reclinatorio propio en la iglesia, en esencia, cualquier cosa que implique mejores medios y mejor nutrición [...] puede ser tomado estadísticamente como un hechizo mágico que confiere toda clase de privilegios [...] Los matemáticos que llenarían de admiración a Newton pueden, al recolectar y aceptar datos y sacar conclusiones de ellos, caer en errores muy crudos por estos yerros populares como lo he estado describiendo.

Aun cuando Shaw, en esta ocasión, apoyó una causa cuestionable, su lógica fue válida. En la primera oleada de entusiasmo por las técnicas de correlación, fue fácil ser imprudentes para los seguidores iniciales de Pearson y Yule. No fue sino hasta veinte años después de lo que escribió Shaw que Yule (1926) horrorizó a los estadísticos aduciendo casos de muy altas correlaciones las cuales, obviamente, no tenían causalidad: por ejemplo, la tasa anual de suicidio estaba altamente correlacionada con la membresía a la Iglesia de Inglaterra. La mayoría de estas

correlaciones “sin sentido” operan a través de variaciones concomitantes en el tiempo, y tienen el efecto saludable de mostrar al estadístico que causalidad no puede ser deducida a partir de algunas co-variaciones observadas, no obstante lo cercanas que estén.

Si bien medidas de correlación simple son raramente calculadas como un fin en sí mismas, mucho del análisis multivariado se inicia a partir de un matriz de coeficientes de correlación. De tal manera, el tópic de correlación permanece como uno de importancia central.

Herbert George Wells (1866-1946), escritor inglés, notable novelista y filósofo, famoso por sus novelas de ciencia ficción, de la que es considerado, junto a Julio Verne, uno de sus precursores, dijo: “El pensamiento estadístico llegará a ser tan necesario para ser un ciudadano responsable como el saber leer y escribir”. Por otra parte, está la figura de Francis Galton (1822-1911), polímata, explorador y científico británico con un amplio espectro de intereses (psicología, biología, tecnología, geografía, estadística, meteorología y eugenesia), entre cuyas contribuciones más importantes en estadística están: el uso de la regresión lineal, ser pionero en la utilización de la distribución normal, inventar la máquina Quincunx (instrumento para demostrar la ley del error y la distribución normal), descubrir las propiedades de la distribución normal bivariada y su relación con el análisis de regresión, e introducir en 1888 el concepto de correlación. Este personaje dijo: “Tengo que escribir acerca de un gran tema (las estadísticas), pero siento vivamente mi incapacidad literaria de hacerlo de forma inteligible sin sacrificar su precisión y su veracidad”.

Esto lleva a la necesidad de tener, al interior de la sociedad, una “cultura estadística” que permita manejar la gran cantidad de información disponible. La Estadística como área cuyo contenido técnico debe ser desarrollado por quienes han logrado la formación adecuada para que la información que llega a la sociedad civil sea accesible en su interpretación, veraz en sus apreciaciones y en sus cifras.

Muchas veces, el proceso de mentir por medio de las estadísticas indica que estas operaciones fueron resultado del deseo que tuvo alguien de engañarnos. Sin embargo, un miembro de la Asociación Estadounidense de Estadística (ASA) afirmó que la mayoría de las veces no se trata de fraude, sino de incompetencia. Quizás haya algo de verdad en esa afirmación, pero no creemos que este supuesto (el de la incompetencia) sea menos ofensivo que el de fraude para los expertos en estadística. Posiblemente sea muy importante tener en cuenta que la tergiversación de datos estadísticos y su manipulación para lograr un fin determinado, no siempre son obra de profesionales. Lo que sale lleno de virtudes de la mesa del experto puede ser cambiado, exagerado, simplificado y falsificado cuando lo manipula otra persona.

Quien tiene la culpa es difícil precisarlo. Los gráficos falsos que aparecen en revistas y periódicos son sensacionales en su exageración. Quienes presentan argumentos estadísticos a favor del crecimiento industrial o de la falta de inflación, rara vez entregan a los trabajadores o al público consumidor información que mejore los hechos en sus sectores; en general, los presentan aún peores para ellos. Mientras los errores provengan todos de un lado, no es fácil atribuirlos a la falta de conocimiento o al azar.

El hecho es que, a pesar de su base matemática, la estadística es tanto un arte como una ciencia. Son posibles muchas manipulaciones, e incluso, tergiversaciones dentro de los límites de su jurisdicción. A menudo, el experto en estadística debe elegir entre diferentes métodos y encontrar el que debe utilizar para representar los hechos, lo que sigue siendo un proceso subjetivo. Sin embargo, en situaciones concretas, como es el caso de las estadísticas oficiales, existen métodos que han sido considerados adecuados por una serie de convenciones y que aseguran la uniformidad de los resultados, permitiendo todo tipo de comparaciones. Todos los cambios deben realizarse de acuerdo con metodologías técnica y científicamente aceptadas, y nunca siguiendo los caprichos circunstanciales de algunas personas o sectores.

A continuación sería interesante explicar cómo se deben mirar los datos falsos y desenmascararlos, y –más importante aún– cómo reconocer datos útiles y verdaderos frente a la ola de fraudes.

No toda la información estadística que llega a nuestras manos puede verificarse con la seguridad de un análisis químico; no obstante, podemos tamizar la información a través de cinco preguntas muy simples. Al encontrar las respuestas, evitaremos asumir muchas cosas que no son ciertas.

### **CINCO PREGUNTAS PARA PRECAUTELAR LA HONESTIDAD ESTADÍSTICA EN LA SOCIEDAD ACTUAL**

A continuación, planteamos cinco preguntas que consideramos de vital importancia para reducir la deshonestidad proveniente del desconocimiento de las reglas estadísticas o, en su defecto, del deseo de engañar que puede tener un autor.

#### **Primera pregunta ¿Quién lo dice?**

Lo primero que hay que tener en cuenta es en qué sentido se puede influir en la información. Lo puede hacer el laboratorio que necesita demostrar algo en beneficio de una teoría, una reputación o una tarifa; el periódico cuyo propósito es publicar un buen artículo; o aquellos trabajadores y jefes que discuten un aumento salarial. Puede existir una influencia consciente. El método puede ser una tergiversación directa de un hecho, o una declaración ambigua que igualmente sirve al propósito deseado y cuya culpabilidad no se puede probar. Puede ser una selección de los datos favorables con la eliminación de los desfavorables. Las unidades de medida pueden ser engañosas, como cuando se considera un año favorable para establecer cierta comparación. O cuando se utiliza una medida inapropiada: un promedio en vez de la mediana, que es mucho más informativa.

Ante ello, debemos buscar cuidadosamente la influencia consciente o inconsciente que tiende a cambiar números e información. Los influyentes, a menudo, se filtran en los gráficos y predicciones de muchos economistas y otros usuarios de estadísticas, lo que lleva a resultados curiosos. Por ejemplo, los trastornos en la estructura económica se descartan alegremente y se brinda todo tipo de pruebas estadísticamente justificadas para

demostrar que, de hecho, estamos en una corriente de prosperidad.

Puede que sea necesario realizar una segunda inspección para averiguar quién lo dice.

#### **Segunda pregunta: ¿Cómo lo sabe?**

En una ocasión se realizó un estudio en empresas para saber si los precios de los productos habían subido o no. De 169 empresas que respondieron, 108 dijeron que no habían aumentado los precios y 61 que sí. Dicho así, tenemos que aproximadamente dos tercios de las empresas no están aumentando los precios y solo un tercio lo está haciendo. Y sucede que, en muchos casos, eso es lo que se informa. Pero al profundizar, resultó que el cuestionario se había enviado a 1.200 empresas importantes. Solo 169 respondieron; aproximadamente el 14%. El 86% restante no tenía ningún interés en informar al público si estaba subiendo los precios o no. El reportero acogió con satisfacción los resultados obtenidos, pero en realidad había muy poco de qué presumir. De hecho, de las 1.200 empresas que estaban en la muestra y que debían participar en la encuesta, el 9% dijo que no había subido los precios, el 5% que los había subido y el 86% no dio ninguna respuesta. Los encuestados constituyen una proporción muy pequeña de la muestra y es de esperar un factor de influencia. En estos casos, debemos buscar indicios de que se trata de una muestra sesgada, que ha sido elegida incorrectamente o –como en el caso referido– que se ha seleccionado a sí misma. Preguntémosnos: ¿Es la muestra lo suficientemente grande como para llegar a una conclusión válida?

Lo mismo podría decirse cuando se informa una correlación: ¿Es la muestra lo suficientemente grande como para significar algo? ¿Hay suficientes casos que le den sentido?

#### **Tercera pregunta: ¿Qué está faltando?**

No siempre nos informarán el número de casos. La ausencia de esta figura, particularmente cuando la fuente de esa información es una parte interesada, es suficiente para hacer sospechoso al conjunto. Tampoco debe tomarse demasiado en serio una correlación que ocurre sin una medida de confiabilidad (error probable, desviación estándar). Observemos los promedios cuya variedad no se especifica, en sujetos para los que se puede esperar que la media y la mediana difieran sustancialmente. Tengamos cuidado con los porcentajes cuando no se nos informe del total sobre el que se tomaron.

A veces se mencionan los porcentajes, omitiendo el material numérico básico. Por ejemplo, una empresa informó que el 33,33% de las mujeres de su plantilla de trabajo tenían miembros casados con un miembro del equipo de gestión. Dadas las cifras brutas, el panorama resultó más claro: al momento del informe solo había tres mujeres en la planta de trabajadores y una de ellas se había casado con un gerente.

Si nos enfrentamos a un índice preguntemos qué falta. Quizás sea la base elegida para tergiversar el resultado. En ocasiones falta el factor determinante del cambio; por ejemplo, cuando se reporta un incremento en la

actividad comercial debido a aumentos en las ventas de abril respecto al mismo mes del año anterior, y no se informa que en el año considerado Semana Santa fue en abril y en el año anterior fue en marzo.

#### **Cuarta pregunta: ¿Alguien le dio un cierto giro a la información?**

Cuando compruebe una estadística, busque la posible tendencia que alguien haya introducido en las cifras totales o en las conclusiones. Con frecuencia se da a conocer una cosa en lugar de otra.

Al comprobar un aumento en el número de casos de una cierta enfermedad no significa siempre una mayor virulencia de la misma. Situación que se está produciendo hoy en día con las cifras del COVID 19. La victoria de un candidato en un distrito pequeño no tiene valor para una encuesta pública de mayores alcances. Que un número de lectores exprese preferencia por los artículos de información internacional, no quiere decir que lean estos artículos si se publican.

En cierta oportunidad, en un país se informó que se había producido una vuelta al campo porque en el último censo se habían detectado medio millón más de granjas que en el anterior. Lo que realmente sucedió fue que la definición de granja había cambiado de un censo a otro.

La población de un área de China era de 29 millones de habitantes; Cinco años más tardes alcanzó los 105 millones. Muy poco de este aumento era real. El primer censo era para impuestos y el segundo para la lucha contra el hambre.

#### **Quinta pregunta: ¿Tiene sentido?**

Esta pregunta tiene mucha importancia cuando la aseveración se basa en un supuesto no probado. Existe una fórmula de amenidad en la lectura desarrollada por Rudolf Flesch. Se trata de medir la facilidad de lectura de un pasaje en prosa, por medio de criterios tan simples y objetivos como la longitud de las palabras y de las frases. Como todas las ideas que tienden a reducir a números lo imponderable y sustituyen el juicio por las matemáticas, esta es llamativa para decirlo de alguna manera. Al menos llamó la atención de quienes la usaron, ya que al aplicarla en algunos textos resultó que el “Cuento de la Bella Durmiente” era dos veces más difícil que la “República” de Platón. Aquí tenemos la historia de siempre: las estadísticas se falsean en las propias narices de los usuarios. Se publican solo porque la magia de los números anula el sentido común.

Las extrapolaciones son útiles, particularmente para aquella forma de predicción llamada «previsión de tendencia». Pero al examinar las cifras o los gráficos derivados de las mismas, debe tenerse presente una cosa: la tendencia actual tal vez sea un hecho, pero la tendencia futura solo puede predecirse mediante una hipótesis razonable como la que lleva implícita la sentencia de que “siempre que lo demás siga siendo igual...” y que “las tendencias actuales continúen...”. De un modo u otro nada permanece igual, de lo contrario, la vida sería demasiado monótona.

## **CONCLUSIONES**

Con esto hemos querido dar un breve recorrido por algunas de las tergiversaciones más frecuentes en Estadística y presentar un conjunto de preguntas básicas que debemos hacernos para iniciar un proceso de aceptación de la información estadística que puede llegar a nuestras manos. Creo importante remarcar que el libro “Cómo mentir con Estadísticas” de Darrell Huff (1954) sigue siendo hoy en día una obra muy esclarecedora en estas circunstancias.

Es bien conocida la enorme importancia que tienen los estudios estadísticos cuando se trata de examinar y comparar el proceso evolutivo de cualquier manifestación de la actividad humana a través del tiempo y del espacio.

Sin la Estadística sería del todo imposible dar una explicación satisfactoria a una gran cantidad de los hechos humanos, así como sin ella sería absolutamente imposible preverlos en todos sus alcances, orientar el criterio con que se ha de juzgarlos, tomar a tiempo las medidas más conducentes para evitarlos o apresurarlos y para conservarlos inmutados o bien modificarlos, descubrir con rapidez y certidumbre cuál es el factor que prevalece en tal o cual sentido sobre el resultado final en cada caso.

Puesto que la vida individual y colectiva de los seres humanos –tanto en su faz constructiva, productiva y creadora como en la de destrucción, de consumo y de desgaste– es una lucha ininterrumpida entre elementos afines y antagónicos, simples y complejos, cercanos y lejanos, que tienden a sobreponerse los unos a los otros de acuerdo a la ley de que el elemento más fuerte, más sagaz, más activo y más previsor aventaja al más débil y más despreocupado, prescindir del ponderado estudio de la Estadística equivale tanto para los individuos como para las colectividades de diferente orden –desde las pequeñas sociedades comerciales hasta las supremas entidades estatales que son representativas de los intereses generales de toda una ciudad o una nación entera– equivale, decimos, a encontrarse en el campo de la batalla eterna e ineludible con los ojos vendados y a ser, por lo tanto, una fácil presa de los contendientes más avezados que poseen información y análisis estadístico veraz y de alta calidad.

Para terminar, debemos decir que sin Estadística sería completamente imposible dar una explicación satisfactoria a un gran número de hechos humanos. Sin ella, sería absolutamente imposible preverlos en todos sus alcances, orientar los criterios con los que deben ser juzgados y tomar las medidas más propicias en el tiempo.

## **BIBLIOGRAFÍA**

- Fisher, R. (1921). On the mathematical foundations of theoretical statistics. *Phil. Trans.*, a, 222, 309
- Galton, F. (1909). *Memories of my life*. P. Dutton and company.
- Huff, D. (1954). *Cómo mentir con estadísticas*. Norton Company Inc.
- Moreau de Jonnés, A. (1829). *Recherches statistiques et économiques sur les pâturages de l'europ*. impr. D'éverat.
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical magazine 5th series*, 50, 157-175.
- Popper, K. (1968). *The logic of scientific discovery*. Hutchinson.

- Popper, K. (1969). *Conjectures and refutations*. Routledges and Kegan Paul.
- Popper, K. y Miller, D. (1983). A proof of the impossibility of inductive probability. *Nature*, 302, 687-698.
- Shaw, G. B. (1906). *The doctor's dilemma*. R. & R. Clarke, Limited.
- Stigler, S. (1986). *The history of statistics: the measurement of uncertainty before 1900*. Harvard University Press.
- Stuart, A. y Ord, J. (1987). *Kendall's advance theory of statistics* (6th ed.). vol. 1. Charles Griffin and Company Limited.
- Yule, G. (1907). On the theory of correlation for any number of variables, treated by a new system of notation. *Proceedings of the royal society a: mathematical, physical and engineering science*, 79, 182–193.
- Yule, G. (1926). Why do we sometimes get nonsense-correlations between time series? *J. R. Statist. Soc.*, 89, 1-6.