Si te gusta la estadística, bancate los metámeros

Elio Campitelli

Introducción

En 1973 Frank Anscombe creó cuatro sets de datos que comparten la media y el desvío de cada variable y su coeficiente de correlación, pero que lucen muy distintos cuando se los grafica (Anscombe 1973). Desde entonces, el cuarteto de Anscombe se usa para ilustrar la importancia de visualizar los datos crudos en vez de confiar en los estadísticos. A pesar de eso, no existe mucha investigación sobre el fenómeno general de "datasets disintos con los mismos estadísticos" del cual el cuarteto de Anscombe es sólo un ejemplo. Además usar un conjunto de datos creados hace 50 años para enseñar da la impresión de que es un caso único o extraordinario.

En este artículo propongo el nombre de "metámeros estadísticos" en analogía al concepto de colorimetría y presento el paquete **metamer**, que implementa el algoritmo de Matejka and Fitzmaurice (2017) para la creación automática de metámeros.

Fuente de metamerismo estadístico

El Demonio de Laplace no sabe ni necesita saber estadística. Él puede conocer la posición y velocidad de cada partícula del universo y usar ese conocimiento para predecir su evolución. Pero los seres humanos no podemos analizar más de unos pocos números por vez. Si queremos entender el universo tenemos que resumir grandes cantidad de observaciones en unos pocos números. Necesitamos saber estadística.

La mayoría de los métodos estadísticos buscan reducir grandes cantidades de datos en unos pocos números interpretables y representativos. Esto implica que son funciones continuas (si dos datasets son similares, sus estadísticos deben ser similares) que pasan de un espacio de alta dimensión a uno de dimensión menor. No existen funciones con ambas propiedades que sean biyectivas (Malek et al. 2010), por lo que, para cualquier método estadístico, los mismos pocos números pueden ser producidos por una infinidad de sets de datos distintos. Por ejemplo, se necesitan N momentos estadisticos para caracterizar unívocamente un set de datos de N observaciones[1]. Como colorario, existen infinitos sets de datos de N observaciones que comparten los mismos $n < N$ momentos.

Definición 1 *Sea una función $E : A \to B$ y un $a_0 \in A$. El conjunto $M_{a,E}$ de todos los $a \in A$ tal que $E(a) = E(a_0)$ son los metámeros de a_0 con respecto a E.*

Es decir, toda función no biyectiva tiene metámeros. El metamerismo es una consecuencia inevitable de los métodos estadísticos; no es un bug, es una característica. El Cuarteto de Anscombe es un ejemplo dramático, pero no debe entenderse como aplicable sólo a los momentos estadísticos. Tampoco debe concluirse que visualizar los datos sea la única solución, ya que al proyectar de los datos en un espacio bidimensional, la visualización también sufre de metamerismo.

Cómo crear metámeros

El paquete **metamer** implementa el algoritmo de Matejka and Fitzmaurice (2017) para generar metámeros. Iterativamente perturba un set de datos verificando que se preserve la transformación estadística de interés y, opcionalmente, que minimice una función.

Al ser completamente genérico, permite ilustrar el metamerismo de cualquier transformación. El siguiente ejemplo genera sets de datos que comparten los primeros 3 momentos pero tienen distribuciones substancialmente distintas (sólo algunas se parecen a un sombrero).

[1] Técnicamente unívocamente a menos a menos de una permutación.

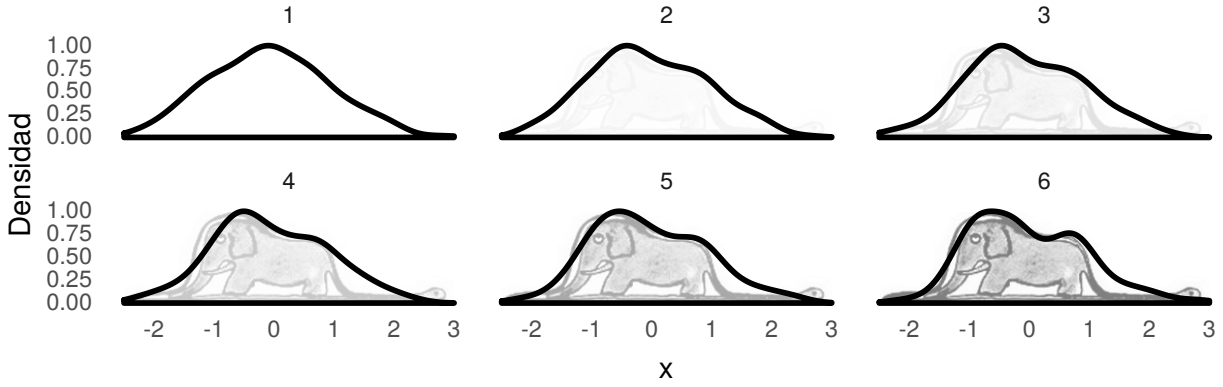

Figura 1: Densidad de probabilidad de 6 metámeros. Todas comparten el mismo promedio, desvío estándard y asimetría.

Bibliografía

Anscombe, F. J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21. https://doi.org/10.2307/2682899.

Malek, Freshteh, Hamed Daneshpajouh, Hamidreza Daneshpajouh, and Johannes Hahn. 2010. "An Interesting Proof of the Nonexistence Continuous Bijection Between Rn and R2 for N != 2." *arXiv:1003.1467 [Math]*, March. http://arxiv.org/abs/1003.1467.

Matejka, Justin, and George Fitzmaurice. 2017. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics Through Simulated Annealing." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 1290–4. Denver, Colorado, USA: ACM Press. https://doi.org/10.1145/3025453.3025912.

Elio Campitelli
Centro de Investigaciones del Mar y la Atmósfera - CONICET
elio.campitelli@cima.fcen.uba.ar