Si te gusta la estadística, bancate los metámeros

Elio Campitelli

Palabras clave: metámeros - estadística descriptiva - algoritmos genéticos - cuarteto de anscombe

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 sumarios. Sin embargo, no existe mucha investigación sobre el fenómeno general de "sets de datos disintos con iguales 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.

Fundamentos 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 representar grandes cantidades de datos con unos pocos números interpretables, lo cual implica una reducción de la dimensionalidad. Intuitivamente, parecería que no se puede representar N números con M números menor que N, aunque esta propiedad debe demostrarse para cada método estadístico. Se pueden usar las identidades de Newton y los polinomios simétricos elementales para demostrar que se necesitan N momentos para describir unívocamente una muestra de tamaño N[1]. Como colorario, existen infinitos sets de datos de N observaciones que comparten los mismos $n < N$ momentos.

Voy a llamar "metámeros" de un determinado set de datos al conjunto datos que comparten el valor de una transformación estadística. El nombre surge por analogía al metamerismo en colorimetría; el fenómeno por el cual nuestros ojos perciben el mismo color a partir de espectros distintos al reducirlos a los mismos 3 números (rojo, verde y azul).

Es decir, toda transformación estadística no inyectiva tiene metámeros. El Cuarteto de Anscombe es un ejemplo dramático, pero no debe entenderse como aplicable sólo a los momentos estadísticos. Ninguna transformación estadística representable como una función contínua es inyectiva si reduce la dimensión del problema (Malek et al. 2010). Como en general se busca que datos similares tengan transformaciones similares, el metamerismo es una consecuencia inevitable de los métodos estadísticos. No es un bug, es una característica.

Tampoco debe concluirse que visualizar los datos sea la única solución. Al proyectar los datos en un espacio bidimensional se pierde información y, como la silueta de un sombrero que puede confundirse con la de una boa digiriendo un elefante, los gráficos también sufren de metamerismo.

Cómo crear metámeros

El paquete **metamer** implementa el algoritmo de Matejka and Fitzmaurice (2017) para generar metámeros. Perturba un set de datos iterativamente, verificando que se preserve la transformación estadística de interés y,

[1] Técnicamente unívocamente a menos de una permutación.

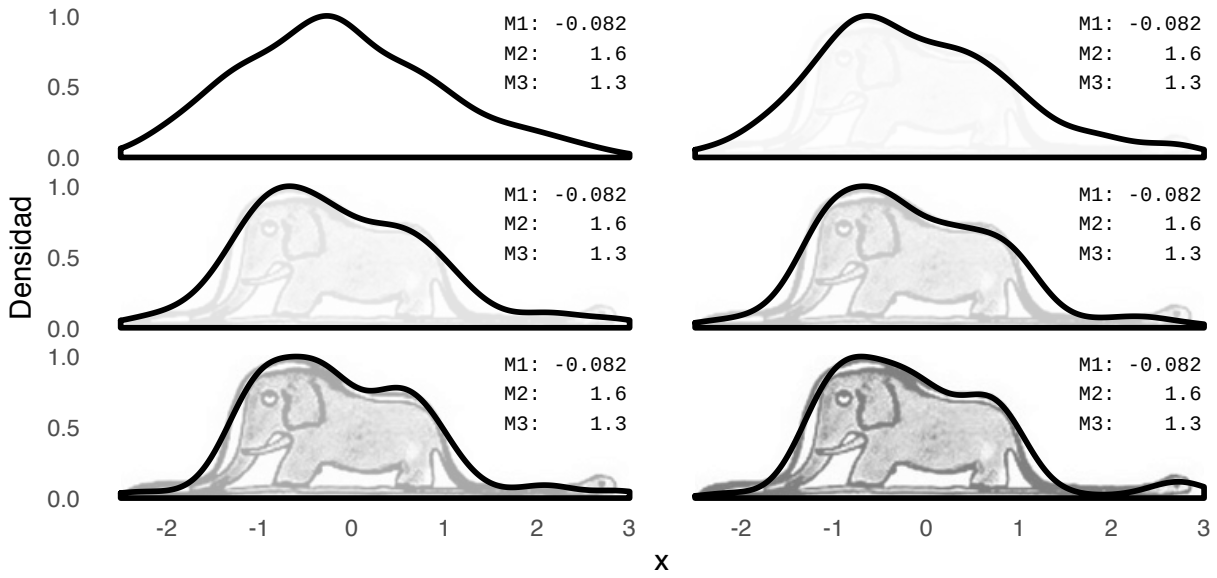

Figura 1: Densidad de probabilidad de 6 metámeros. Todas comparten los primeros tres momentos no centrados hasta 2 cifras significativas (M1, M2 y M3).

opcionalmente, que se minimice una función.

Al ser completamente genérico, permite ilustrar el metamerismo de cualquier transformación. El siguiente código genera sets de datos que comparten los primeros 3 momentos al menos con 2 cifras significativas pero cuyas distribuciones son substancialmente distintas (Figura 1).

```
library(metamer)
metameros <- metamerize(data.frame(x = rt(200, 5)), # dataset inicial
                        preserve = moments_n(1:3),  # función a preservar
                        minimize = elefante,        # función a minimizar
                        annealing = FALSE,          # todos los metámeros minimizan `minimize`
                        N = 250000)                 # número de iteraciones
```

Referencias

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 Rnˆ and R2ˆ 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 - UBA/CONICET
elio.campitelli@cima.fcen.uba.ar