Esto no es una pipa (sobre el Teorema Central del Límite)[1]

Siempre me resultaron un tanto inconducentes, y quizás sospechosas, las cosas que se definen por la negativa, como los cursos de economía para "no economistas", o los "no docentes". Por eso es que me resulta un tanto incómodo hablar del Teorema Central de Límite (TCL) en base a qué no es el TCL. Pero tengo la fuerte impresión de que una enorme mayoría de alumnos y también profesionales avanzados tienden a entender incorrectamente dicho teorema, y quizás en el peor de los sentidos: no es que no lo entienden sino que entienden correctamente algo que no es el TCL. Y eso es grave, porque no hay yerro más grave que dar la respuesta correcta a la pregunta equivocada. He aquí una larga diatriba sobre esta confusión en relación a este importantísimo resultado de la matemática.

Comienzo declarando que en ningún lugar de esta nota diré formalmente qué es el TCL, ese es justamente el "yeite". Le propongo al lector buscar en un buen libro de estadística o probabilidad, o en Wikipedia. En segundo lugar, no hay tal cosa como "el TCL" sino que TCL se refiere a una suerte de familia de resultados. Para simplificar, nos referiremos al más simple de los casos, correspondientes a una muestra independiente e idénticamente distribuida, con media y varianza bien definidas.

En primer lugar, el TCL no hace referencia a la media muestral *per se*, sino a su versión estandarizada, es decir, habiéndole restado su "centro" (la media poblacional) y dividido por su desvió estándar, que como todos sabemos, y en nuestro caso canónico, es el desvió estándar del error dividido el tamaño de la muestra. Mucha gente piensa que el TCL dice que asintóticamente la media muestral es normal, pero el resultado que "gobierna" la distribución de la media muestral en el límite (cuando el tamaño de la muestra tiende a infinito) no es el TCL sino su hermano, la ley de grandes números (LGN). La LGN dice que la media muestral converge en probabilidad a la media poblacional. Es decir, la LGN provee la distribución límite de la media muestral: es degenerada en un punto (la media

[1] Extracto de Walter Sosa Escudero, 2015, El Lado Oscuro de la Econometria, 2015, Ed. TEMAS, Buenos Aires, en prensa.

poblacional). Los que hayan tomado clase conmigo me han escuchado decir que quizás el principal objetivo de mi curso de posgrado es que distingan el TCL de la LGN. En resumen, dada la LGN, el TCL no puede agregar información a lo que ya sabíamos: en el límite, la distribución de la media muestral es degenerada en un punto. Chocolate por la noticia.

Y entonces, ¿para qué sirve el TCL? Bueno, ahora empiezan las sutilezas y los problemas. Mucha gente cree que el TCL dice que la media muestral es normal para un tamaño de muestra lo suficientemente grande, de hecho Wikipedia pregona, textualmente, que "El teorema del límite central garantiza una distribución normal cuando n es suficientemente grande". El TCL dice que la sucesión de media muestrales estandarizadas (que se forma agregando observaciones de a una), es una sucesión de variables aleatorias que converge en distribución a una variable normal estándar. Eso (su versión formal) es el TCL. Cualquier otra cosa no, como cuando una vez, al preguntar en mi clase que era el TCL, alguien me dijo "es un resultado que dice que todo es normal".

Bien. El TCL como dije, no hace referencia a la media muestral sino a su versión estandarizada. El argumento informal que justifica lo que dice Wikipedia es más o menos el siguiente. Cuando n terminó de tender a infinito, la media muestral estandarizada es exactamente normal. Entonces, un poquito antes de terminar de tender a infinito, la media muestral estandarizada es aproximadamente normal. Fíjense muy bien las palabras que uso; primero digo "exactamente" y luego "aproximadamente". Ahora, la media muestral es una simple transformación lineal de su versión estandarizada. Si agregamos que toda transformación lineal de una variable normal es también normal, y lo juntamos con lo anterior, nos queda lo siguiente: un poquito antes de terminar de tender infinito la media muestral (¡no estandarizada!) es aproximadamente normal. ¿Dónde está la trampa? ¡En que el TCL vale para infinito y no dice nada exacto acerca de qué pasa antes de infinito! El TCL provee una aproximación al comportamiento de la media muestral para muestras "suficientemente grandes". ¿Y que es suficientemente grande? Y esta es LA pregunta. Estrictamente, suficientemente grande es infinito. Ahora, si aceptásemos alguna inexactitud, si bien es imposible conocer la verdadera distribución de la media muestral para un tamaño de muestra finito, el TCL dice que las discrepancias entre ésta y la normal no deberían ser muy grandes para un tamaño de muestra “grande”. ¿Y qué cosa es grande? Bienvenidos a la ciencia. Es una pregunta empírica.

Un ejemplo clásico y que da lugar a vergonzantes confusiones es el siguiente. La tabla de la distribución “t” de Student, la que viene en las últimas páginas de los libros de estadística, se detiene misteriosamente en n=30. Siempre pregunto en mi clase ¿Por qué en 30? (y no en 31 o 29). En nuestro contexto, no existe ningún tamaño de muestra finito para el cual podamos garantizar que la media muestral sea normal. Ahora, para n=30, las discrepancias entre la verdadera distribución (que no la sabemos) y la normal son “descartables”. ¿Y por qué para 30 y no 29 o 31? Es una maldita convención, como la que dice que las edades se comunican en años (y no en días), o la temperatura en grados (y no miligrados). Por eso es que la tabla se detiene en 30: porque sí. La idea es que convenimos, socialmente, que más allá de 30 las discrepancias entre los percentiles de la distribución t y los de la normal, si bien difieren, esa diferencia no nos es socialmente relevante, como que yo les diga que tengo 48 años y ustedes sepan que en realidad tengo “algún número comprendido entre 48 y 49”. Algún desprevenido usa este argumento para decir que “más de treinta es una muestra grande”, uno de los disparates más grandes que haya escuchado en mi vida, una clara muestra de inmadurez conceptual, en base a esta ya larguísima discusión.

Entonces, el TCL provee a) un resultado exacto para la media muestral estandarizada, cuando el tamaño de la muestra tendió a infinito, b) una aproximación (¿útil?) para la media muestral, quizás antes de que la muestra haya tendido a infinito. Como diría George Box, a fines prácticos, el TCL es un modelo errado (jamás la muestra es infinita) que solo intenta ser útil. Y la pucha que lo es.

Opino que en matemática pasa algo similar a lo que sucede con las relaciones personales. Uno se gana el derecho de usar apodos una vez que conoce muy bien a las personas. Si me encontrase con el célebre compositor Gustavo Leguizamón (a mi juicio uno de los más relevantes del folclore argentino), le diría “Gustavo” o “Sr. Leguizamón”, y quizás luego de un tiempo de entrar en confianza le diga “Cuchi”, su reconocido apodo. Con la matemática pasa, a mi entender, exactamente lo mismo: uno puede hablar informalmente de los teoremas solo cuando pasó mucho tiempo entendiendo exactamente qué es lo que dicen. Un grueso error de principiantes es hacer resúmenes de la matemática, como si fuese filosofía o historia. El punto es que por definición la matemática ya está resumida. Mi consejo es, con la matemática, hacer exactamente lo contrario. Dejar el resultado intacto e intentar ampliar, agregar ejemplos, preguntas, ejercicios, detalles. Si el TCL ocupa, digamos, una página de tu libro, tus notas tienen que medir tres o cuatro páginas, agregando (y jamás quitando) ejemplos, ideas, extensiones y tus inquietudes.

Como dicen que decía Frank Zappa, a veces pienso que hablar de matemática es como bailar sobre arquitectura. Siendo pesimista, opino que las confusiones sobre el TCL tienen que ver con interpretaciones de cosas que no son el TCL. Y siendo optimista digo que el TCL, uno de los más importantes resultados de la matemática, es un concepto tan útil como complejo de abordar. Y quizás amerite larguísimas discusiones, siempre y cuando no terminemos diciendo "Ok, pero esto no es el TCL", como genialmente decía Magritte de su famosa pipa (busquen en Wikipedia, ¡un poco de cultura!).

Referencias

Billingsley, P., 1995, *Probability and Measure*, 3ra ed, Wiley, Nueva York.

McCabe, B. y Tremayne, A., 1993, *Elements of Modern Asymptotic Theory with Statistical Applications*, Manchester University Press, Manchester.

White, H., 2001, *Asymtotic Theory for Econometricians*, ed. Revisada, Academic Press, San Diego.