ANALISIS DE DATOS

Sixto Ríos

UNIVERSIDAD COMPLUTENSE DE MADRID

Cuando la Estadística se va poblando por todos sus capítulos de modelos de decisión bayesianos, neobayesianos, no bayesianos, intervalos de confianza de otros tantos tipos, etc., parecerá a muchos algo trasnochado el título elegido para estas notas. Pero precísamente hemos escogido este tema porque, como dice Tukey, el análisis de datos es un tema "elemental, importante y despreciado".

Sin duda, la impetuosa corriente, iniciada por Student y Fisher a comienzo de siglo, con el tratamiento de los problemas de inferencia estadística en el marco de un modelo matemático apropiado, ha arrastrado a casi todas las mentes estadísticas nacidas en el medio siglo transcurrido.

Es un hecho reconocido que la idea de considerar un conjunto de datos estadísticos como una muestra aleatoria de una población, real o potencialmente existente, y utilizar los métodos del Cálculo de probabilidades para establecer relaciones de inferencia de las características de la muestra a los parámetros de la población, ha permitido resolver importantes problemas prácticos y teóricos y su desarrollo ha dado lugar a innumerables trabajos y libros.

Pero no todos los problemas en que intervienen datos estadísticos se encuentran en estos modelos de la decisión o la inferencia estadística, que implican

la consideración de una población y de una muestra aleatoria de la misma.

Frecuentemente el estadístico se enfrenta con unos datos a los que no tiene sentido considerar como muestra de ninguna población y deben obtenerse conclusiones basadas sobre los mismos, que no pueden, por tanto, tener el carácter de una inferencia estadística tal como la hemos descrito (*) en párrafos anteriores.

Si comparamos los datos relativos a división en fincas agrícolas de dos provincias de España, p.e., Zamora y Sevilla, ¿podemos decir que estos datos son *indicativos* de una mayor división de la tierra en una provincia que en otra? . Obtener conclusiones sobre situaciones como ésta puede tener interés; pero aquí no se trata de muestras aleatorias y no tiene sentido aplicar la inferencia estadística convencional.

¿Se puede tener una respuesta satisfactoria a este problema utilizando la media y la desviación típica o serán más adecuadas para dar una respuesta comprensible a tal cuestión otras medidas como la mediana y el recorrido intercuartil o bien otras nuevas adecuadamente definidas? .

De un modo general podemos decir que el tratamiento de un problema mediante inferencia formal, es decir, utilizando modelos matemáticos que permitan dar una medida de la incertidumbre de las conclusiones obtenidas mediante alguno de los procedimientos ya clásicos de intervalos de confianza, probabilidades a posteriori, funciones de riesgo, etc., puede no tener sentido por alguna de las siguientes razones (entre otras): a) los datos no permiten conocer los efectos de algunas de las más importantes causas de variación; b) las causas se presentan en una forma claramente no aleatoria; c) no existe un modelo probabilístico apropiado para la situación, pues los conocidos presentan restricciones completamente inadecuadas a la situación real.

En tales situaciones cabrá, sin embargo, hacer *"indicaciones"* como consecuencia del análisis de los datos. Para el investigador el valor principal de los datos reside en lo que *indican* o parecen demostrar.

Ejemplos de indicaciones son por ejemplo:

1) apariencia de similaridad en la composición de dos poblaciones de medidas obtenida a través de sus histogramas por ser ambos de tipo acampanado, o ambos bimodales, etc.;

2) o de comportamiento general de una serie de frecuencia (p.e., parecen decrecer de un modo exponencial, etc),

3) las dispersiones de varias poblaciones presentan una gran estabilidad, mientras los valores centrales son muy distintos.

(*) Tampoco se ha de confundir este aspecto de la Estadística con lo que se suele llamar muestras de poblaciones finitas.

Podemos decir, en general, que una *indicación* es una conclusión comprensible obtenida a través de una cierta elaboración de unos datos estadísticos y que esta elaboración se puede llamar un *indicador*. Un indicador puede ser un histograma o un estadístico (media, varianza, coeficiente de correlación, etc.) o un esquema de cálculo más complicado (análisis complejo de varianza, análisis de regresión múltiple, etc.).

Pero ¿por qué el tratamiento de la "inferencia" con modelos probabilísticos formales ha tenido tan importante desarrollo, mientras el de la "indicación" ha sido tan limitado, hasta el punto que en los libros puede considerarse reducido a los capítulos, más o menos ramplones, de la llamada estadística descriptiva ?

No es fácil contestar a esta pregunta:posiblemente los que constituyen una nueva teoría o técnica comienzan por abrir los caminos más fáciles y muchas veces el peso específico de estos pioneros arrastra a muchos otros investigadores por el camino iniciado, abandonando otros problemas que consideran menos generales o más específicos y difíciles.

En trabajos recientes de J.W. Tukey (*) se describe el análisis de datos de una manera que queremos recoger aquí:"Extensas partes del análisis de datos son inferenciales en el sentido muestra–población, pero esto no es todo el análisis de datos. Extensas partes del análisis de datos son incisivas, permitiendo establecer indicaciones que no se pueden obtener por simple y directo examen de los datos brutos, pero esto no es todo el análisis de datos. Otras partes del análisis de datos se refieren a los métodos para planificar la toma de datos y a la distribución del esfuerzo y otras consideraciones valiosas para la observación, esperimentación y análisis".

Es bien sabido de cuantos aplican la Matemática o la Estadística a los problemas reales la dificultad de ser fieles simultáneamente a la realidad y al modelo, evitando resolver un problema que no es el real sino otro inventado por el matemático, para su más fácil solución o por no ser capaz de crear el método matemático adecuado.

Mi mayor ilusión sería que estas notas valieran para despertar el interés de los estadísticos profesionales por estos aspectos fundamentales del análisis de datos, estimulándoles a la lectura de la memoria de J.W. Tukey, que tantos problemas nuevos suscita y plantea.

(*) The future of data analysis, The Annals of Mathematical Statistics Vol. 33. nº 1.