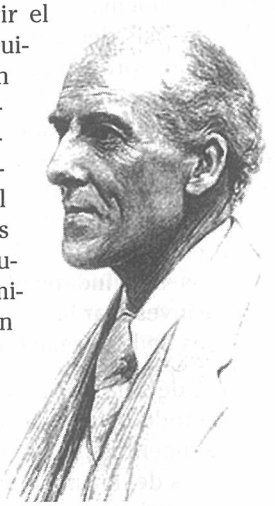
UNA PANORÁMICA DEL ANÁLISIS DE DATOS

Miguel Ángel González Sierra

Por costumbre, se ha considerado la estadística descriptiva y la teórica como dos ramas de la misma con distinta metodología. Años atrás, se trataba de resumir los resultados en términos de "estadística descriptiva" calculando medidas de situación y dispersión, momentos de órdenes más elevados o diversos índices, y también exponer determinadas características de los datos a través de gráficos tales como histogramas, diagramas de barras y gráficas bidimensionales. No se hace referencia al mecanismo estocástico (o distribución de probabilidad) que originó los datos presentados. Las estadísticas descriptivas así efectuadas se utilizan en la comparación de diferentes grupos de datos. Incluso se prescriben diversas reglas para escoger entre estadísticos alternativos, tales como media, mediana y moda, dependiendo de la naturaleza de los datos y de las preguntas que deban ser contestadas. Tales análisis estadísticos son conocidos como análisis descriptivos de datos. En estadística teórica, el objetivo es de nuevo el resultado de los datos, pero con referencia a una determinada familia de distribución de probabilidad (o modelo) subyacente. El resumen estadístico descriptivo, en este caso, depende básicamente del modelo estocástico específico, y las distribuciones de probabilidad correspondientes se utilizan para especificar márgenes de incertidumbre referidos a parámetros desconocidos. Esta metodología se conoce con el nombre de análisis inferencial.

Karl Pearson fue el primero en tratar de cubrir el hueco entre ambos tipos de análisis. Utilizaba la intuición facilitada por el análisis descriptivo basado en los momentos e histogramas para obtener referencias acerca de la familia de distribuciones subyacentes. Para este menester, inventó el primero y quizás el más importante test, consistente en utilizar el estadístico ji-cuadrado, para contrastar la hipótesis de que uno o más datos provenían de una distribución de probabilidad pertenenciente a una determinada familia. Dicho test marcó el comienzo de un nuevo modo de tomar decisiones.



Karl Pearson

Karl Pearson creó un sistema de distribuciones de probabilidad, que se podían generar a partir de sus cuatro primeros momentos. Un hermoso ejemplo de trabajo de investigación fue llevado a cabo por Karl Pearson a través del uso de histogramas y el test ji-cuadrado, consistente en el descubrimiento de que la distribución del tamaño de los tripanosomas hallados en determinados animales, es una mezcla de dos distribuciones normales.

La necesidad de desarrollar métodos generales de estimación surgió a partir de la aplicación del test ji-cuadrado para examinar las hipótesis de que la distribución subyacente pertenece a una determinada familia paramétrica de distribución. Pearson propuso la estimación de parámetros a partir de los momentos, utilizando el test ji-cuadrado basado en la distribución ajustada. Ronald Fisher llevo a cabo dos refinamientos, el primero en términos de obtener un mejor ajuste con los datos obtenidos, a través de la estimación de los parámetros desconocidos, por el método de la máxima verosimilitud y el segundo en el uso correcto del test ji-cuadrado utilizando el concepto de grado de libertad, cuando los parámetros desconocidos son estimados.

Durante las décadas de los años veinte y treinta Fisher introdujo una extraordinaria serie de ideas estadísticas. En un importante escrito de 1922 estableció la bases de la "estadística teórica", analizando los datos por medio de modelos estocásticos, especificados de antemano. Desarrolló una gran variedad de test de hipótesis exactos, para tamaños muestrales pequeños y bajo el supuesto de normalidad, recomendando su uso combinado con la utilización de tablas de ciertos valores críticos, normalmente el 5% y el 1%, cuantiles del estadístico que expresaba los tests. Durante este periodo, bajo la influencia de Fisher, se otorgó gran importancia a los test de significación y gran número de contribuciones acerca de las distribuciones exactas en el muestreo se llevaron a cabo por Hotelling, Bose, Roy y Wilks entre otros. Aunque Fisher aludió a la especificación del modelo, problema considerado en primer lugar por Pearson, como un importante aspecto de la estadística en su artículo de 1922, no prosiguió con el mismo posteriormente. Quizás en el contexto de la investigación biológica que Fisher estaba examinando, en la que manejaba pequeños grupos de datos, no había mucho margen para investigar el problema de la especificación, someter los datos observados a detallados análisis descriptivos en busca de rasgos especiales o empíricamente determinar apropiadas transformaciones de los mismos para ajustarse a un modelo estocástico determinado. Fisher utilizó su propia experiencia para demostrar cómo los datos se verifican para decidir sobre las especificaciones de los modelos. En este punto del desarrollo de la estadística inspirada por las propuestas de Fisher se llevaron a cabo tentativas por parte de otros estadísticos para desarrollar lo que denominamos test de hipótesis no paramétricos, en los que las distribuciones de los estadísticos involucrados para definirlos son independientes del modelo estocástico subyacente de los datos, e investigar la robustez de los tests propuestos por Fisher, frente a desviaciones de la normalidad de la distribución subyacente.

En la década de los años veinte y treinta también se produjeron avances en la metodología de la recolección sistemática de datos a través del diseño de experimentos introducidos por Fisher, lo cual permite que los datos sean analizados de una manera específica a través del análisis de la varianza, interpretándose de forma que tengan significado; el diseño dictaba el análisis y el análisis revelaba el diseño.

Mientras que gran parte de las primeras etapas de la investigación estadística fue motivada por problemas originados por la biología, tenían lugar

desarrollos paralelos a pequeña escala en el uso de la estadística en la producción industrial. Shewhart (1931) introdujo el proceso gráfico simple a través de cuadros de control para detectar cambios en los procesos de producción, lo cual es probablemente la primera contribución a la detección de valores atípicos o de puntos de cambio, en un valor de una magnitud estudiada.

Gran parte de la metodología propuesta por Ronald Fisher estaba basada en la intuición, y una teoría sistemática de la inferencia estadística no pudo ser por aquel entonces desarrollada. Esta fue suministrada por J. Neyman y E. S. Pearson en 1928, formulando una teoría general de la decisión. Fisher mantenía que su metodología era más apropiada en la inferencia estadística mientras concedía a las ideas de Neyman y Wald mayor relevancia en las aplicaciones tecnológicas, aun cuando Wald afirmaba la validez universal para sus teorías. Wald también introdujo métodos secuenciales de aplicación en inspección de muestras, que tuvieron también aplicaciones en biología, según encontró Fisher.

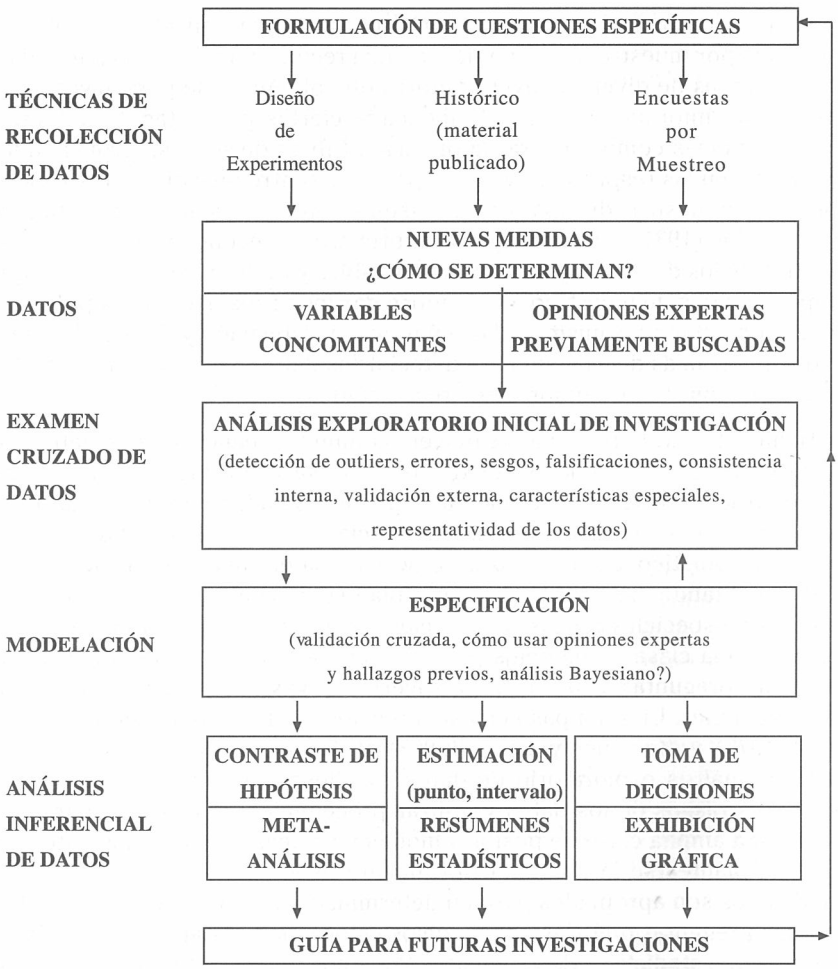
En la década de los cuarenta hubo un gran auge en el desarrollo de encuestas por muestreo, que implicaban una recolección de gran cantidad de datos a través de diversos investigadores que obtenían, de personas escogidas al azar, información obtenida mediante ciertas preguntas. En tal situación, problemas como asegurar la precisión (libres de sesgos, errores en los registros y en las respuestas) y la comparación (entre investigadores y métodos de encuesta) de los datos, asumieron la máxima importancia. Mahalanobis (1931, 1934) fue quizás el primero en reconocer que tales errores en trabajos de investigación eran inevitables y podrían ser mas serios que simples errores de muestreo, debiéndose dar los pasos necesarios para controlar y detectar a los mismos al diseñar una investigación y desarrollar apropiados programas de examen para detectar los datos atípicos (outliers) y los valores inconsistentes en los datos recolectados.

Hemos discutido brevemente que era común la creencia de que había dos ramas de la estadística, la estadística descriptiva y la estadística deductiva, así como la necesidad experimentada por los estadísticos de depurar los datos de posibles defectos que pueden viciar las inferencias obtenidas del análisis estadístico. Lo que quizás se necesitaba era una aproximación integrada, facilitando métodos para un apropiado conocimiento de los datos, sus defectos y especiales rasgos, y por selección de un apropiado modelo estocástico o una clase de modelos para el análisis de datos con el objetivo de contestar preguntas especificas y sugerir nuevas preguntas para futuras investigaciones. Un gran paso en esa dirección fue llevado a cabo por Tukey (1962, 1977) y Mosteller y Tukey (1968) al desarrollar lo que es conocido como el análisis exploratorio de datos. La filosofía básica es entender los especiales rasgos de los datos y utilizar procedimientos robustos para acomodar una amplia clase de posibles modelos estocásticos para los datos. En lugar de plantearse la pregunta típicamente fisheriana de qué compendios estadísticos son apropiados para un determinado modelo estocástico, Tukey propuso preguntar qué clase de modelos estocásticos es apropiado dado un compendio estadístico determinado. Podemos hacer referencia a lo que

Chatfied (1985) describe como análisis de datos inicial, lo que parece ser un amplio análisis descriptivo de datos basando sus conclusiones en el sentido común y la experiencia, con un mínimo uso de la metodología estadística tradicional.

Un esquema de análisis de datos está en el cuadro adjunto. Éste se puede aplicar al análisis de grandes grupos de datos y parece combinar la estadística descriptiva de K. Pearson, la inferencial de Fisher, el análisis exploratorio de datos de Tukey y la preocupación de Mahalanobis sobre el muestreo sin errores.

ANÁLISIS ESTADÍSTICO DE DATOS



En dicho cuadro, los datos son usados para representar el grupo entero de medidas registradas u observaciones, y como han sido estas obtenidas, a través de un experimento, de encuestas o de registros históricos, así como los procedimientos operacionales envueltos en el registro de las observaciones, y cualquier información previa (incluyendo opiniones de expertos) sobre la naturaleza de los datos o del modelo estocástico subyacente en los datos.

Se entiende por examen cruzado (o validación) de datos a cualquier estudio exploratorio o inicial efectuado para entender la naturaleza de los datos, detectar errores de medición, errores de registro y datos anómalos, para probar la validez de la información previa y examinar si los datos son genuinos o falsos. El estudio inicial también intenta probar la validez de un modelo específico, seleccionar un modelo estocástico más apropiado o bien una clase de modelos estocásticos para posteriores análisis de datos.

El análisis inferencial de datos significa el cuerpo entero de métodos estadísticos tales como la estimación, la predicción, el contraste de hipótesis y la toma de decisiones, basándonos en un modelo estocástico específico, a partir de datos observados y con propósitos determinados. El propósito del análisis de datos debería ser extraer toda la información obtenible de los datos y no solamente reducirse a contestar cuestiones específicas. Los datos a menudo contienen valiosa información indicando nuevas líneas de investigación y avances en el diseño de futuros experimentos o planes de muestreo para la recolección de datos. Se puede proponer el principio fundamental del análisis de datos en forma de ecuación fundamental:

Análisis de datos = Respuesta a Cuestiones Concretas + Suministro de Información para Nuevas Líneas de Investigación

Las secuencias del análisis de datos indicadas en el cuadro no deberían ser observadas como distintas categorías con diferentes metodologías. Sólo muestra lo que deberíamos hacer para empezar cuando tratamos con datos y en qué forma el resultado final se expresa y usa en aplicaciones prácticas.

Bibliografía

Fisher R. A.: "On the mathematical foundations of theorical statistics" *Philos. Trans. Roy. Soc.* 222 (1922) pp. 309-368.

Shewhart, W, A.: *Economic Control of Quality of manufactured Product.* D. Van Nostrand, Nueva York, 1931.

Tukey, J.: *Exploratory Data Analysis.* Addison Wesley, 1977.