6to Congreso Latinoamericano de CIENCIA POLÍTICA

12, 13 y 14 de junio de 2012
FLACSO Sede Ecuador

FLACSO ECUADOR

alacip
ASOCIACIÓN LATINOAMERICANA CIENCIA POLÍTICA

Estadística Adimensional: del dato al gráfico estimado por inducción

Al rescate de un método en crisis de fundamento teórico desde hace dos siglos.

Emilio José Chaves[1]*

RESUMEN

El ensayo narra la experiencia del autor en los cursos de estadística de su época universitaria; plantea luego algunos elementos de la actual crisis profunda de un sector de la teoría estadística tradicional: la inducción (inferencia). Recupera el ordenamiento descendente de datos (Pareto) y lo aplica para construir un modelo estructural de distribuciones univariables contínuas no-negativas. Con datos de predios rurales de Nariño, Colombia-2008, ilustra los componentes básicos de la Estadística Adimensional, una técnica no-paramétrica propuesta en 2009 para separar la media de la estructura distributiva y elaborar la Función de Distribución Acumulativa (FDA) desde los datos y dos premisas indispensables sobre los valores extremos. No usa funciones de densidad de probabilidad (PDFs) y sugiere adoptar por consenso, en su lugar, histogramas adimensionales normatizados para comparar regresiones hechas con métodos distintos sobre datos comunes. La idea puede aplicarse en otros campos de la ciencia y en temas estadísticos especializados, y aún mejor, ser llevada al campo de la enseñanza teórica y práctica para hacerla más sencilla y eficaz en la formación estadística de investigadores jóvenes de la periferia mundial y de otras naciones, luego del debido proceso evaluativo. Resume la parte matemática en forma de gráficos y cuadros. La conclusión reafirma la importancia de la ética, la subjetividad responsable y la autocrítica en el manejo estadístico por parte de los diversos actores desde una visión humanista.

Palabras claves: Crisis de teoría estadística, Ciencias Básicas: Métodos de Investigación, Estadística adimensional, Enseñanza estadística.

Adimentional Statistics: from data to estimated graphs of induction

To the rescue of a method in crisis of its theory fundamentals during last two centuries

Emilio José Chaves*

Summary

The essay tells the author's past experience in his statistics university courses; it treats later some elements of the deep crisis of a sector of traditional statistics theory: induction (or inference). It recovers the descending data ordering (Pareto) which is applied to build a structural model of non-negative univariable continuous distribution functions. Using regional rural data (Nariño, Colombia, 2008), it shows basic components of

[1] * Emilio José Chaves, ingeniero, investigador independiente en Economía y Distribución. Miembro de Comité Editorial de Revista Tendencias Universidad de Nariño-Pasto-Colombia. Correo-electrónico: chavesej@hotmail.com

Adimentional Statistics plus a non-parametric technique proposal (2009) that separates the media from the structural distribution to build the Cummulative Distribution Function (CDF) using only data plus two additional premises on extreme values. It does not need Probability Density Functions (PDF) and suggests their change –through consense- to adimensional normalized histograms in order to compare regressions built with different methods from common data. The idea might be applied to other science fields and specific statistical themes and even more, it might be used in statistics teaching and practice in order to make them simpler and more effective in the education of young periphery researchers and other nations youth -once it fulfills its due evaluation process-. The math part is summarized in graphs and tables. It concludes asserting the importance of ethics, responsible subjectivity and self-criticism in statistical handling by all participants in a humanist perpective.

Key Words: Statistical theory crisis; Basic sciences: Research methods, Adimensional statistics, Statistics teaching.

Introducción

Hace 14 años observé la siguiente discrepancia metodológica en el tema de la distribución de ingresos: mientras el italiano Wilfredo Pareto (1896) , el inglés Leo Chiozza-Money (1905) y el econometrista polonés Oskar Lange (1958) trabajaron el tema en Europa con datos ordenados de mayor a menor (de altos ingresos a bajos), ocurría en Estados Unidos lo contrario: Otto Lorenz (1905), autor de la primera versión de lo que hoy llamamos Curva de Lorenz , (CL), y Paul Samuelson (1962) –en su famoso texto de economía neoclásica- ordenaron los datos de menor a mayor (de bajos ingresos a altos). Esa diferencia tiene efectos en la sencillez y claridad del manejo teórico-matemático y en la calidad de los resultados estimados. Con el paso del tiempo, la investigación derivó en la propuesta de estadística matemática que da el título a este artículo.

Cuando tomé el curso de estadística en la universidad estuve a punto de perder la materia. Los capítulos sobre probabilidades discretas eran comprensibles, útiles y bien fundamentados. La confusión empezó cuando aparecieron las campanas “normales” de Gauss (Funciones de Densidad de Probabilidad, FDP), las que remitían al estudiante a unas tablas al final del libro para resolver ejemplos y problemas de tarea; esas tablas arrojaban las cifras claves para solucionar los ejercicios cuando se daban dos parámetros correctos: 1) La media de la distribución y 2) Su desviación estándar. Necesité veinte años más para comprender que las FDP eran unos constructos de tipo platónico, creados por matemáticos brillantes a comienzos del siglo XIX y usados como moldes forzados para ajustar datos reales a la premisa-forma simétrica y centrada de la campana de Gauss. Entonces comprendí que ni el autor del texto, ni el profesor, ni yo, entendíamos bien el tema, y ante eso lo más fácil era repetir la receta de memoria con mucha fé, porque según decían, era un instrumento científico comprobado en la práctica, gracias al cual la ciencia y la técnica habían logrado enormes avances, evidentes hasta en la vida cotidiana. Y lo decían con tanta seguridad, sin el menor atisbo de duda, que uno se sentía amedrentado y se limitaba a callar, memorizar y obedecer.

Al tomar el curso de economía de Samuelson y leer sus anotaciones sobre la Curva de Lorenz intuí que allí anidaba oculta otra explicación clara y diferente sobre la desigualdad. Retomé el asunto en la década de 1990, me interesé en la econometría y encontré en Oskar Lange unos estudios claros y profundos sobre Pareto y la distribución de ingresos. En 1996 realicé un análisis matemático de las distribuciones de Pareto que empleé después para algunas propuestas macro-económicas que hoy necesito mejorar. Ese análisis ya portaba la idea central de ordenar los datos en sentido descendente, la de separar la media de la distribución adimensional tal como lo intuyó Pareto en su época, y la de emplear una función exponencial especial para las curvas de ajuste. Con el tiempo el método fue perfeccionado, desbordó la econometría y se centró en el marco más general de la estadística básica, el análisis datual, y la inferencia no-paramétrica aplicada, un tema esencial en ciencias básicas e investigación moderna.

En los congresos sobre estos temas se habla de la necesidad de integrar la estadística y la probabilidad "bajo un solo paraguas" y de reunir las diferentes tradiciones estadísticas que han hecho aportes desde diferentes ramas de la ciencia al precio de generar una notable dispersión teórica y metodológica. En resumen, necesité veinte años para comprender porqué no pude entender esa parte de la materia y otros veinte para sugerir una contrapropuesta. El objetivo es contribuír a unir el tema de las distribuciones univariables contínuas alrededor de fundamentos más sólidos y sencillos; a facilitar la formación estadística de nuestros jóvenes investigadores del Tercer Mundo y a debatirla con expertos de otras partes del mundo, si lo tienen a bien.

En el proceso hubo varios factores de gran ayuda: 1) El acerbo cultural del mundo occidental y de otras culturas distintas sobre el tema; 2) La aparición de los computadores personales con sus hojas de cálculo; 3) La lectura en Internet de otras visiones del tema por otros autores. 4) El contacto con culturas indígenas, afrodescendientes y campesinas del sur de Colombia y del norte de Ecuador, porque aprendí que ellos hacen a su manera inferencias útiles para su práctica cotidiana, su alimentación, sus medicinas, sus mitos y su relación con el entorno. Por ejemplo, cuando siguen las huellas de los animales, no lo hacen solo para cazar y pescar; también buscan conocer sus costumbres, sus lugares preferidos, sus nichos, sus juegos y su manera de ser.

El artículo resume la propuesta teórica empleando datos de un histograma sobre distribución de tierras rurales de la región donde vivo (Nariño, suroccidente de Colombia, año 2008). Con ellos trabaja y explica el método, las gráficas inferidas, y las ideas que definen la propuesta central. En esencia, pide un cambio radical en el enfoque y la enseñanza de la inferencia. Invoca el principio de la cuchilla de Occam para adoptar por consenso cambios sustanciales en este sector específico de la teoría y la práctica estadística. Dentro de esos cambios solicitados, quizás el más duro y polémico es el de eliminar y/o poner bajo control el uso de las funciones de densidad de probabilidad (PDFs), para concentrarse en una estadística adimensional no-paramétrica basada únicamente en datos y frecuencias reales, que separa la media dimensional de la estructura adimensional de la distribución en su primera fase y las reintegra en el momento de las conclusiones.

Los motivos

En el proceso encontré otros puntos problemáticos en el manejo del tema, poco mencionados por los expertos que pude consultar, los cuales me llevaron a sospechar que este sector de la estadística matemática contenía inconsistencias teóricas y prácticas de modo que, el panorama que yo divisaba era el de una disciplina que avanzaba velozmente por el camino de los análisis multivariables sin disponer aún de un manejo convincente del análisis univariable. A continuación enumero esos puntos críticos que a mi entender son auténticas fisuras de la teoría estadística aún vigente, sin entrar a discutirlos de manera extensa.

El lenguaje matemático aquí empleado puede ser mejorado y formalizado con mayor rigor, pero en general puede ser entendido por personas que dispongan de una formación básica en cálculo y elaboración de gráficos. No es mi intención malinterpretar a ningún autor particular, ni a los gremios académicos que los divulgan, aunque sí es mi intención el cuestionar sus técnicas, sus premisas y su escasa autocrítica.

Dentro de la teoría estadística convencional -implícita en textos, aulas, paquetes estadísticos y programas virtuales- hay ciertos elementos teóricos y prácticas que merecen ser cuestionados, o al menos ser usados para reflexionar sobre ellos como base para críticas y autocríticas más profundas:

1) En estadística los datos son siempre discretos y representan una muestra parcial e incompleta del fenómeno estudiado. El analista suele asumir que son representativos del fenómeno general. La estadística suele también usar premisas adicionales que no siempre son declaradas de manera abierta y oportuna; esto es comprensible y se soluciona asumiendo la subjetividad y justificando la necesidad y el efecto posible de las premisas. Aún así, la estadística es muy importante porque permite entender mejor la pluralidad, la variedad y el cambio de los fenómenos, así sea que las respuestas halladas sean normalmente aproximadas, probables y válidas solo para el contexto y momento de recolección de los datos.
2) Los datos del investigador suelen ordenarse en forma de tablas que presentan dos columnas principales: la de valores dimensionales de la variable repartida y la de población asignada. A veces los datos vienen agrupados dentro de un límite inferior y otro superior del valor de la variable. Las mejores tablas suelen informar el promedio de cada grupo y el promedio de la muestra total, también llamado “la media”, pero no es una práctica frecuente.
3) Si los datos presentan valores es porque fueron medidos con alguna técnica, poseen unidades dimensionales apropiadas … vienen en kilogramos, bacterias por litro, dólares, pesos de ingreso mensual per-cápita… tanto para la variable repartida como para la población receptora del reparto.
4) La mayoría de los estudios manejan datos ordenados en orden ascendente, así como algoritmos elaborados para esa premisa. Aquí los datos son ordenados en orden descendente de la variable: de mayor a menor.
5) Es muy frecuente el uso de conjuntos de datos en forma de histogramas dimensionales que suelen llevar casillas cuyo ancho va sobre el eje horizontal para representar los límites de la variable repartida, y cuya altura va en el eje vertical y representa la frecuencia grupal de ocurrencia de cada

casilla. Pero estos histogramas son problemáticos porque una misma muestra ordenada de 20 datos puede dar origen a múltiples histogramas posibles según las preferencias del analista a la hora de agruparlos y graficarlos. Los histogramas requieren unidad y reglas claras por los expertos.

6) Hay muchos casos en los que la única información es una serie de pocos datos (unos veinte) y no se dispone de información de las frecuencias. En estos casos se suele usar el Principio de Laplace (1812), consistente en asumir que la media es el promedio aritmético de los N datos y que cada valor tiene igual frecuencia: f=1/N. Esas dos premisas son muy duras e inexactas ya que al aumentar el tamaño de la muestra cambian sus valores y las frecuencias se hacen más altas hacia la mitad y más bajas en los extremos por lo general, aunque no necesariamente; la regla de Laplace sirve como premisa inicial provisional porque permite hacer una primera estimación tanto de la media como de la desviación estándar que usan las tablas de FDP de los apéndices de los textos.
7) La desviación estándar y la media suelen ser un dúo dimensional. No se puede adelantar el valor de este dúo sin conocer muchos puntos de la distribución. Sería más sensato dejar que los datos hablen por sí mismos, graficarlos y analizarlos en vez de usar el dúo como insumo paramétrico de las tablas de funciones de densidad de probabilidad (PDFs) usados como fuentes de datos.
8) Las PDF son funciones paramétricas (una familia de modelos muy rígidos basada en parámetros) como en el caso de la curva Normal de Gauss (han creado muchas otras PDF con más de dos parámetros y formas muy distintas a la curva "Normal" de Gauss). Ocurre además que dos muestras del mismo fenómeno con 20 datos cada una suelen producir dos parámetros distintos, que conducen a dos curvas PDF diferentes, o sea a dos respuestas distintas e inciertas. Al aumentar el tamaño de las muestras las diferencias disminuyen, pero no necesariamente son representativas. Es posible que una muestra de 10 datos sea más representativa que una de 40, aunque sea menos probable que eso ocurra.
9) Produce dudas ver el uso de las PDFs como fuente de valores estimados para diversos propósitos (p.e: para las inferencias bayesianas, los p-values, y otros fines). Afortunadamente, desde hace décadas en Internet se pueden leer diversas críticas explicadas en lenguaje sencillo de epistemológos de la estadística sobre esos métodos (Cox, Mayo; 2006), (Arévalo; 2000), que recomendaban lo que hoy es una tendencia clara: usar métodos no-paramétricos y no-lineales para desarrollar mejores aproximaciones en estos casos, así como en la inferencia inicial que forma parte de cualquier análisis multivariable.

Hay otras técnicas que pueden ser cuestionadas, como los análisis de Estadística Descriptiva basados en los datos de intervalos, propios de los histogramas. Entre ellas menciono:

1) a) Asumen que el promedio aritmético de los dos límites de cada casilla es la media grupal y que van centradas respecto a las frecuencias acumulativas, para luego estimar con ellos la media dimensional de toda la distribución; b) A veces corren los límites para graficar particiones equidistantes de la variable repartida en los histogramas sin declarar el efecto sobre las frecuencias y medias grupales; c) Otras veces declaran los

valores muy altos y/o muy bajos como *outsiders* (datos exteriores a los límites aceptables para el analista) para luego mutilarlos del cuerpo datual por conveniencia matemática, ya que manejar valores extremos es complejo y esas son precisamente las regiones más desconocidas de las distribuciones. Lo que deberían hacer es respetar los datos en lugar de mutilarlos o ignorarlos, y buscar explicaciones sobre esos valores extremos.

2) Cuando los datos disponibles se concentran hacia los valores extremos, tanto la media como la desviación estándar resultan afectadas; es preciso acordar unas reglas para su manejo. Conocer que dos distribuciones poseen valores idénticos para ese dúo no nos dice mucho, ya que hay multitud de distribuciones posibles y diferentes para cada dúo. Las cosas se complican aún más cuando se agregan más parámetros de difícil interpretación como los *grados de libertad*, la *entropía* y otros cuantos empleados en modelos multiparamétricos de FDP.

3) Usando la hoja electrónica intenté calcular el área bajo la curva de las FDPs. Siempre obtenía valores diferentes a la unidad. En varios textos de ejemplos observé que recomiendan usar una constante que actuaba como un comodín de cartas de baraja. Recetaban multiplicar la constante por la integral del área y poner a continuación un signo igual y la cifra uno. Por lo tanto, el valor de la constante es el inverso de la integral, y se declaraba resuelto el problema. A mi entender se trata de un recurso inaceptable. Pero hay más: ocurre que cuando se trabaja una distribución adimensional derivada de la curva de Lorenz la integral siempre da la unidad, de modo que intenté usar la hoja electrónica para que estimara la variable distribuída de la curva Normal (de Gauss) para muchos valores de la fracción acumulada, dándole como parámetros la media igual a la unidad, y valores diferentes de la desviación estándar. Cuando esta última era pequeña arrojaba valores positivos y todo parecía correcto, pero si le daba valores altos aparecían valores de la variable negativos, y para completar, el área bajo la curva tampoco arrojaba el valor esperado igual a la unidad. Esa fue la evidencia más fuerte que me llevó a dudar de las PDFs.

Este artículo busca aportar una contrapropuesta usando métodos inferenciales no-paramétricos desde América Latina y desde el Tercer Mundo. Busca responder las mismas preguntas básicas recuperando el empleo del orden datual descendente y el uso de una familia de funciones de regresión nueva, inscrita dentro del contexto de las Curvas de Lorenz y las Funciones Cumulativas de Distribución (CDF) para ese ordenamiento. Se trata de resumir la esencia del modelo ya explicado con más detalle en otros artículos (Chaves; 2009). Trabaja un ejemplo real de carácter local, resume las bases mátemáticas del modelo, las gráficas resultantes de la aplicación, y somete para su evaluación y eventual consenso por parte de la comunidad estadística dos propuestas de gráficos adimensionales obtenidos por regresión deductiva inversa sobre una serie especial de fracciones de población acumuladas, con el fin de construir un histograma unificado que permita comparar los resultados de la regresión con las tablas datuales originales, así como resultados obtenidos a través de métodos diferentes de inferencia cuando se aplican a las mismas series datuales.

Por ahora quedan pendientes otros subtemas para los cuales ya hay propuestas de solución en el marco aquí tratado, tales como la aditividad del modelo, la relectura del Teorema Central del Límite, la inferencia con datos que no permiten estimar directamente la media total ni la media grupal, las condiciones de representatividad de una muestra pequeña, el manejo de datos multivariados, los modelos estadísticos sintéticos, la enseñanza teórica y práctica para nuestros jóvenes, y en general la formulación con más rigor de lo que podríamos llamar el campo de la *estadística adimensional*.

Resumen del método

Una primera discusión más detallada de la historia del tema desde Pareto hasta nuestros días, de las premisas y desarrollos puede leerse en (Chaves, E.J.; 2009), donde aparecen las principales referencias, tablas y manejos lógico-matemáticos que sirven de soporte. Por el momento, es más útil explicarlo con uno de los cuadros que componen el afiche (poster) que intenté presentar y publicar sin éxito en los útimos dos años en revistas de dos universidades públicas y otra privada.

Elementos básicos del método desarrollado

Una vez ordenados los datos en orden descendente de la variable distribuída *K* entre la población *q*, tabular un conjunto de *n* cuantiles datuales (x_i , L_i) de la curva de Lorenz, y ejecutar las etapas siguientes:

1. Estimar puntos datuales de función estructural $F(x_i)$ calculando para cada cuantil la función generatriz $F(i) = \ln (L_i) / \ln (x_i)$. Se obtienen *n-1* puntos datuales (x_i , $F(i)$), los cuales se grafican y caben dentro del cuadro de 1*1 típico de la Curva de Lorenz.

2. Con ayuda de la función de regresión de la hoja de cálculo usada, obtener una función de ajuste al vector estructural $F(x)$ para el rango $1/q \leq x \leq 1$. La función $F(x)$ debe ser derivable, guardar su derivada $F'(x)$ para uso posterior. $F(x)$ debe cumplir condiciones lógicas propias del orden descendente (que no abordamos aquí).

3. Calcular la Curva de Lorenz para 100 o más valores de x mediante la expresión: $L(x) = x^{F(x)}$

4. Obtener la Función Adimensional de Distribución Acumulativa $K(x) \geq$ con la expresión: $K \geq(x) = L(x) * [F(x) / x + F'(x) * \ln (x)]$

5. Aplicar las propiedades de las Curvas de Lorenz y otras premisas y criterios, estimar ingresos medios para intervalos grupales, indicadores varios, medir errores promedios entre modelo y datos.

6. Usar indicadores y algoritmos adimensionales para estimar dispersiones, errores promedios, ajustes, indicadores,etc.

En este caso el método está planteado para datos cuantiles de la Curva de Lorenz que estén completos. No es necesario que los datos sean equidistantes. Conviene disponer de datos relativamente bien distribuídos sobre el rango de X y de ser posible que haya cuantiles muy cercanos a los extremos; si esto no ocurre, ayuda mucho asumir con buen criterio un valor mínimo plausible para x=1y otro para X=0. Una vez definida F(x) queda determinada toda la estructura de la distribución. Es recomendable usar el promedio de las desviaciones absolutas porcentuales como medida de la dispersión empleando cien o más puntos equidistantes en lugar de la desviación estándar que es dimensional, endogámica, no comparable y poco informativa.

La función F(x) es una herramienta muy útil para comparar distribuciones adimensionales entre sí porque no depende de la media y sirve para detectar pequeños cambios estadísticos del fenómeno a través del tiempo,

para estudiar otros modelos propuestos para fines parecidos, y para comparar muestras de tamaños distintos y medias diferentes entre sí.

Datos de ejemplo trabajado de inferencia estadística

La Tabla 1 resume datos oficiales de la región sobre la distribución y tamaño de las propiedades rurales.

Histograma datual para alimentar el modelo

Tabla 1. Distribución de Tierra Rural: Departamento de Nariño-2008

Datos de Histograma en Orden Descendente

Límites grupales Hectáreas/Predio	Unidades Predios	Superficie Hectáreas	Número de Propietarios	Hectáreas por predio	Propietarios por predio
2000 y más	88	1.021.527,01	103	11.608,26	1,17
1000 a 2000	31	38.063,92	61	1.227,87	1,97
500 a 1000	65	43.550,92	114	670,01	1,75
200 a 500	226	70.088,77	429	310,13	1,90
100 a 200	522	69.832,60	921	133,78	1,76
50 a 100	1.507	98.956,67	2.265	65,66	1,50
20 a 50	6.070	183.473,75	8.649	30,23	1,42
15 a 20	3.252	55.057,52	4.510	16,93	1,39
10 a 15	7.077	84.040,33	9.792	11,88	1,38
5 a 10	19.882	134.618,35	27.095	6,77	1,36
3 a 5	25.209	92.980,03	33.929	3,69	1,35
1 a 3	78.912	129.347,82	103.030	1,64	1,31
0 a 1	134.726	43.431,85	165.066	0,32	1,23
TOTALES	277.567	2.064.969,54	355.964	7,44	1,28

FUENTE: IGAC (Instituto Geográfico Agustín Codazzi - Gobierno de Colombia)-Enero 2008

El manejo datual implica tomar datos de la realidad en forma de muestra contextual, parcial y temporal. Siempre hay una etapa en la que ordenamos los datos de alguna manera o tal vez aceptamos la que ya traen. El fin de la inferencia estadística es construir modelos que porten suficiente isomorfismo general respecto a los datos disponibles, y que puedan asumirse como conclusión general aproximada y válida para ellos. Hecho esto, aplicamos la lógica deductiva para derivar conclusiones particulares cuya validez y vigencia dependen de que al replicar el análisis con nuevos datos, no aparezcan divergencias importantes que obliguen a replantear el primer modelo inductivo aceptado como premisa de trabajo. En todo proceso de regresión inferencial hay premisas subjetivas y lo que llamamos objetividad responde al deseo y la voluntad de autocontrolar con responsabilidad ética, con sinceridad, con criterios claros y con transparencia nuestra subjetividad, para evitar los riesgos que conlleva el no hacerlo. En nuestro mundo de alta pluralidad y dinámica de cambio constante necesitamos usar ambas herramientas y conocer sus límites para entender mejor el sistema-mundo y las comunidades que nos acompañan. Todo esto nos ayuda a tomar decisiones mejor informadas y mejor diseñadas para nuestros problemas y proyectos, desde criterios éticos que

favorezcan la vida plena y digna para todos, de modo que si cometemos errores, al menos nos equivoquemos de lado de la vida y de la benevolencia, tal como aconsejó Don Quijote a Sancho Panza en algún pasaje.

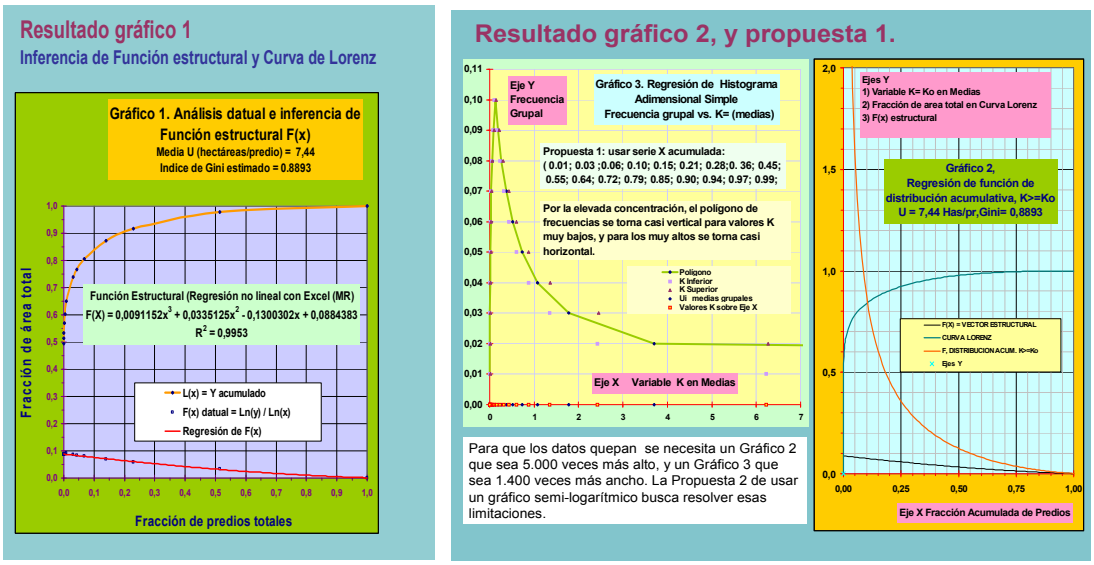
Resultados gráficos y comentarios

En el Gráfico 1 aparece la función estructural obtenida, F(x), la cual ocupa la parte baja del cuadro de uno por uno. En la parte alta está la curva de Lorenz correspondiente al ordenamiento descendente cuyo índice de Gini estimado resultó ser muy alto, cercano a 0.89. Para encontrar la distribución acumulativa de la variable repartida K≥Ko basta derivar la curva de Lorenz respecto a X, usando la fórmula ya mostrada. El Gráfico 2 reune las tres curvas; puede observarse que el valor de la variable K se hace mayor a dos medias para valores menores al 4% de la población, desbordando el espacio asignado en el gráfico.

Nos interesa también proponer un histograma estandarizado que permita comparar los resultados de las regresiones inferenciales sobre una base común. Hay una serie especial de frecuencias grupales que tiene la virtud de generar frecuencias grupales y acumuladas de población idénticas para el ordenamiento ascendente y descendente al mismo tiempo. La serie de frecuencias grupales es la siguiente:
0.01-0.02-0.03-0.04-0.05-0.06-0.07-0.08-0.09-0.10-0.09-0.08-0.07-0.06-0.05-0.04-0.03-0.02-0.01

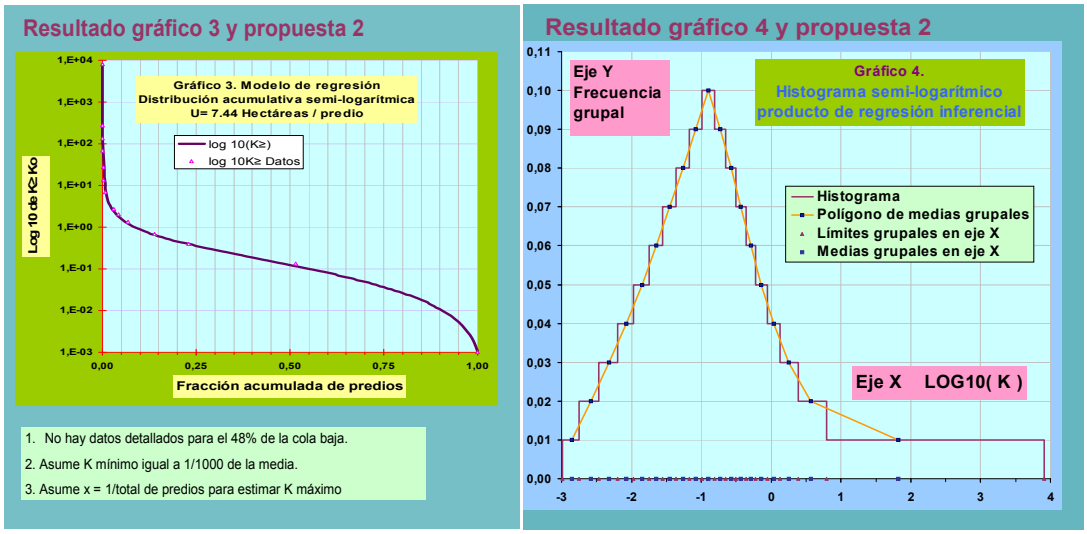
La suma total de ellas da la unidad, y las frecuencias acumuladas siempre dan estos valores:
0.01-0.03-0.06-0.10-0.15-0.21-0.28-0.36-0.45-0.55-0.64-0.72-0.79-0.85-0.90-0.94-0.97-0.99-1.00

La Propuesta 1 sugiere usar esas series para construir el histograma estandarizado del Gráfico 3,



donde las casillas tienen ancho diferentes, las medias grupales rara vez pasan por el centro de las casillas, y podemos constatar que el Gráfico 3 tampoco nos alcanza para mostrar la distribución integral, a pesar de que en el eje horizontal la variable adimensional K tiene un rango que va de cero a 7 medias.

En este punto con casos de muy alta desigualdad, viene la propuesta 2, que consiste en usar gráficos semilogarítmicos para la variable distribuída K, tanto en la distribución acumulativa, como en el histograma, combinado con la serie que conforma la propuesta-1, el Gráfico 3 y el Gráfico 4 ofrecen así un resultado integral, adimensional y de lectura más nítida y fácil. Esto nos permite resolver el problema de la multiplicidad de histogramas posibles para series datuales iguales, el cálculo de las medias grupales, y un método unificado para análisis comparativos, sea que se hagan con ordenamientos datuales ascendentes o descendentes.



Conclusiones y comentario final

Habitamos una pequeña partícula que se desplaza, gira y cambia dentro de un contexto cósmico que hace otro tanto. Durante el día los paisajes que contempla la mitad de la humanidad dependen del instante, del sitio desde donde los miramos, de la luz que salió del sol ocho minutos antes y de nuestra manera de mirar e interpretar lo que vemos. Al mismo tiempo, la otra mitad recibe destellos en la noche con antiguas noticias de eventos remotos del pasado luego de viajar infinidad de años por el espacio a la velocidad de la luz. El pasado lejano y el reciente coexisten con el instante presente en el que diseñamos el futuro. También se dan juntos el infinito grande –por ejemplo, cuando definimos la selva amazónica como ecosistema contextual-, con el infinito pequeño -cuando levantamos una laja al borde del camino y descubrimos bajo ella complejos ecosistemas poblados de multitud de seres vivos diminutos que interactúan a otros ritmos en medio de la oscuridad y la humedad-. En ambos casos se dan contrastes enormes, tal como en el ejemplo presentado sobre datos de distribución de tierras de hace dos años en mi región, que son distintos de otras regiones y épocas en su tamaño medio y en su estructura distributiva.

Más crucial que el inevitable desfase entre los tiempos fácticos, datuales y analíticos puede ser el hecho de que la teoría estadística fue colocada sobre un pedestal como criterio máximo de rigor científico sin el debido

control por parte de los filósofos y epistemólogos especializados en lógica y matemática-estadística, a quienes se silenció por mucho tiempo en nombre de la productividad y el pragmatismo, lo que contribuyó a su dispersión y crisis actual (Méndez, E.; 2000). Numerosas experiencias fallidas y excesos encontrados en investigaciones y aplicaciones supuestamente controladas en su parte estadística han abierto cajas de Pandora que ya afectan a la misma ciencia. Pienso que la ciencia ha progresado más por la tenacidad, la intuición, la creatividad, la autocrítica y las virtudes personales y colectivas de investigadores y grupos de investigación, que por los aportes de la teoría y la práctica estadística. De hecho, ya lleva más de un siglo contaminada por la presencia de las PDFs, entidades innecesarias a la luz de este y otros estudios, que a mi parecer, hoy solo sirven para que profesores que no entienden lo que enseñan, torturen y desmotiven a estudiantes que no están en condiciones de cuestionar tanta sabiduría convencional pero que podrían ser grandes investigadores si les enseñamos a confiar en sí mismos para recoger los datos de su interés, para procesarlos por sí mismos, y para interpretar los resultados con criterios propios sanos, con la convicción de que aparte de la estadística los investigadores siempre disponen de muchos otros recursos que les aportan indicios y criterios valiosos para detectar la novedad, el avance y la incoherencia en la marcha de sus proyectos.

Los inmensos retos del mundo actual requieren jóvenes investigadores muy creativos y dotados de criterios éticos al servicio de la vida digna de la humanidad, que se sientan corresponsables y cocreadores de la marcha del planeta, que se atrevan a desobedecer órdenes absurdas y/o infames y a decir sin temor a los funcionarios del poder y del saber .. "el emperador está desnudo".

Conclusiones y opiniones:

1. Una vez encontradas la media dimensional y la función estructural F(x) la distribución queda determinada. Al escoger F(x) el analista ejerce una subjetividad responsable y auto-controlada que afecta el resultado y define el valor mínimo de la variable repartida.

2. F(x) es una herramienta valiosa para el análisis datual. Cuando su valor es constante nos encontramos ante la distribución de Pareto.

3. Con F(x) podemos detectar cambios pequeños en la evolución de la estructura distributiva en el tiempo, analizar distribuciones paramétricas, estudiar grupos muestrales, revisar otras metodologías, diseñar distribuciones sintéticas, incursionar en otros campos de la estadística como el análisis multivariado.

4. Podemos redefinir el término *probabilidad* dentro del contexto del histograma como una certeza aproximada de que entre los límites de cada casilla tienen lugar eventos discretos cuyo valor medio y cuya frecuencia *como grupo* son cercanos a los pronosticados en el histograma, aunque cada evento individual sea impredecible.

5. El futuro de la estadística parece orientarse hacia los métodos no-paramétricos que son más aptos para entender un mundo siempre cambiante en su variedad y pluralidad.

6. Este texto es portador de una propuesta implícita de cambio en la enseñanza estadística para nuestros jóvenes, sin PDFs.

Bibliografía

Cowell, Frank A. **Measuring Inequality**, Mayo, 2000. 3a. Ed. Oxford University Press, U.K.. Disponible en: http://sticerd.lse.ac.uk/research/frankweb/MeasuringInequality/index.html (Junio 15/2009)

Chaves, Emilio J. **Curvas Funcionales de Lorenz, Análisis Datual e Inferencias**. Tendencias. Volumen X No. 2 - Segundo Semestre 2009. Revista de la Facultad de Ciencias Económicas y Administrativas. Universidad de Nariño, Pasto, Colombia. Disponible en: www.udenar.edu.co/general/tendencias/contenidos/Vol10.2/**CHAVEZ**.pdf

Mayo, Deborah G.; Cox, D. R. **Frequentist statistics as a theory of inductive inference**. Viriginia Tech and Nuffield College, Oxford. IMS Lecture Notes–Monograph Series-2nd Lehmann Symposium – Optimality Vol. 49 (2006) 77–97. TM Institute of Mathematical Statistics, 2006.
Disponible en:
http://www.error06.econ.vt.edu/MayoCox.pdf

Méndez, Evaristo. **El Desarrollo de la ciencia: un enfoque epistemológico.**
Espacio Abierto, Asociación Venezolana de Sociología. Vol. 9-No. 4/ISSN 1315-0006 octubre-diciembre 2000/ pp.505-534.
Disponible en:
http://redalyc.uaemex.mx/pdf/122/12290403.pdf

Rivadulla, Andrés. **Inducción, deducción y decisión en las teorías estadísticas de la inferencia científica.** Revista de Filosofía, 3a. épocam vol. VI, núm.9. págs. 3-13. 1993. Editorial Complutense, Madrid.
Disponible en:
http://revistas.ucm.es/fsl/00348244/articulos/RESF9393120003A.PDF