¿QUÉ ENSEÑAR?, ¿ESTADÍSTICA DESCRIPTIVA O ANÁLISIS EXPLORATORIO DE DATOS?

Santiago Inzunza Cazares*
sinzunza@uas.uasnet.mx

RESUMEN

En el presente trabajo se discute la pertinencia de un enfoque distinto al enfoque de estadística descriptiva que ha predominado en la enseñanza de la estadística, denominado análisis exploratorio de datos. Se proporcionan fundamentos teóricos y didácticos que hacen factible este enfoque mediante la utilización de herramienta computacional con estudiantes de secundaria y bachillerato, y se ofrecen algunos ejemplos de actividades y la forma como podrían ser implementadas por los profesores en el salón de clases.

INTRODUCCIÓN

Hasta antes de la década de los noventa, los cursos de estadística estaban prácticamente reservados para estudiantes universitarios que cursaban estudios superiores. Sin embargo, ante la relevancia que la estadística está teniendo en la sociedad moderna, -frecuentemente caracterizada como "la sociedad de la información y del conocimiento"-, se han propuesto y realizado reformas a los currículos de matemáticas en diversos países (por ejemplo: NCTM 1989, 2000, en Estados Unidos; AEC 1990, en Australia; MEC 1992, en España) para incorporar contenidos estadísticos en los niveles preuniversitarios. De tal forma, que la estadística y la probabilidad -al igual que otras áreas de las matemáticas, como la geometría, la aritmética y el álgebra-, se han convertido en un área importante del currículo de matemáticas en estos niveles.

En el caso de México, a partir de 1993, con la reforma curricular para la educación secundaria, se incorporan contenidos específicos de estadística bajo el nombre de *presentación y tratamiento de la información* y *probabilidad* en los plantes de estudios de matemáticas. Dichos contenidos se profundizan aún más y se replantea su enseñanza sugiriendo la utilización de herramientas computacionales en la más reciente reforma curricular propuesta para iniciarse en su fase experimental a partir del año 2005 (SEP, 2005).

El nombre que tradicionalmente se daba –incluso todavía persiste en muchos currículos, libros de texto y es utilizado por muchos profesores-, a la parte de la estadística encargada del análisis de datos, es el de *estadística descriptiva*. Como su nombre lo señala, el propósito se centra en la "descripción" de los datos, utilizando para ello procedimientos de agrupamiento a través de tablas o distribuciones de frecuencia y el cálculo de medidas descriptivas, como la tendencia central y la dispersión. Si bien, en estadística descriptiva se utilizan representaciones gráficas, como diagramas de barras y sectores, histogramas y polígonos de frecuencia, el análisis no enfatiza en el uso de ellas en forma sistemática, sino más bien como recurso adicional al cálculo de estadísticos que permiten dar una idea global del comportamiento de los datos.

Actualmente, la denominación de estadística descriptiva ha sido cambiada por la de *análisis exploratorio de datos* (EDA por sus siglas en inglés), en la cual el énfasis se centra en la "exploración" más que en la descripción de los datos. En dicha exploración, las representaciones gráficas ocupan un lugar fundamental. Este cambio de enfoque ha sido derivado del trabajo de John Tukey (1977), quien en la década de los setenta, trabajando para los Laboratorios Bell en los Estados Unidos, desarrolló diversas herramientas para el análisis de datos, las cuales tuvieron como catalizador principal a la tecnología computacional y al software estadístico con capacidades de representación gráfica y tratamiento de datos que en dicha época empezaban a estar disponibles

* Profesor e investigador en la Facultad de Informática-Culiacán y la Escuela Preparatoria Dr. Salvador Allende de la Universidad Autónoma de Sinaloa.

cada vez por un mayor número de personas e instituciones. Podemos decir entonces, que el propósito del análisis exploratorio de datos va mucho más allá de la mera descripción a la que usualmente se limita la estadística descriptiva.

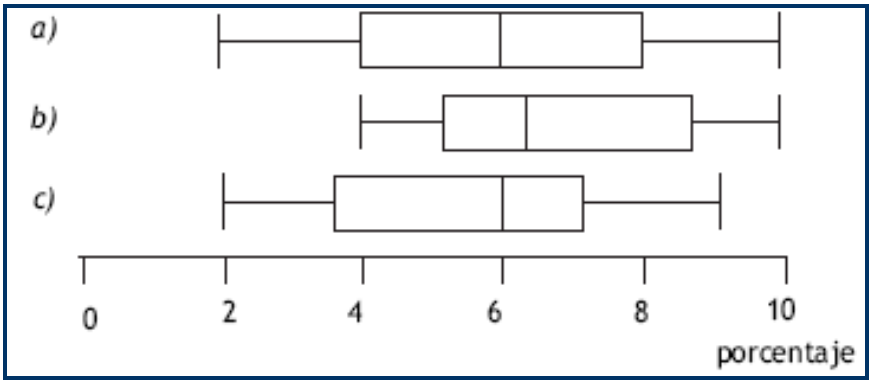
LA NATURALEZA DEL ANÁLISIS EXPLORATORIO DE DATOS

La diferencia entre estadística descriptiva y el análisis exploratorio de datos, no radica solamente en las herramientas que se utilizan, como es la invención de nuevas representaciones gráficas y el uso de computadoras, que incluso, podrían ser utilizadas con un enfoque meramente descriptivo. Más bien, el análisis exploratorio de datos es un conjunto de estrategias para el análisis de datos, cuya esencia es "permitir que los datos hablen y la búsqueda de patrones en los datos, sin una primer consideración sobre si los datos son representativos o no de una población o si pertenecen a una muestra. En muchas situaciones, el análisis exploratorio de los datos puede preceder a una situación de inferencia formal, mientras que en otras, el análisis exploratorio puede sugerir preguntas y conclusiones que se podrían confirmar con un estudio adicional (Moore1992, p.3). De acuerdo con lo anterior, el análisis exploratorio de datos puede ser una herramienta de utilidad en la generación de hipótesis, conjeturas y preguntas de investigación acerca de los fenómenos de donde los datos fueron obtenidos.

Para mostrar lo anterior, tomaremos un ejemplo que se proporciona en los Planes de Estudio de Matemáticas (SEP, 2005, p. 140).

Las siguientes gráficas (diagramas de caja) muestran:

a) La distribución de los puntajes obtenidos por un grupo en un examen.
b) La distribución de los puntajes entre los varones del grupo.
c) La distribución de los puntajes entre las mujeres.



Los diagramas de caja son una de las representaciones gráficas desarrolladas por Tukey (1977) para el análisis de datos. Consiste en partir la distribución de los datos sobre un eje horizontal, en cuarto partes a intervalos de frecuencias iguales (25%, 50%, 70% y 100%), siendo la mediana el valor central (50%).

Para el análisis podemos apoyarnos en las tres etapas del análisis exploratorio de datos que señala Curcio (1989):

1. Ver los datos (un análisis global de forma, centro y dispersión)
2. Ver entre los datos (comparar grupos de datos y variables)
3. Ver más allá de los datos (plantear conjeturas, hipótesis e inferencias).

La construcción del diagrama de puntajes totales puede ubicarse en la primera etapa (ver los datos y realizar un análisis global). El análisis de dicha distribución nos revela que la mediana de los puntajes es igual a 6 y la distribución de las calificaciones del grupo es simétrica, pues las cuatro secciones de la gráfica tienen la

misma longitud, es decir, la dispersión a cada lado de la mediana es la misma. Si el puntaje de 6 se considera aprobatorio, podemos decir que la mitad del grupo acreditó el examen.

El hecho de construir sobre una misma gráfica el diagrama con la totalidad de los datos y los grupos o categorías (hombres y mujeres) en forma separada, se puede ubicar en un segundo nivel de análisis (ver entre los datos), que refleja la intención de comparar ambos grupos para buscar diferencias y similitudes en cuanto a los puntajes obtenidos. El análisis en este nivel revela que en el caso de los hombres (b) la distribución está sesgada a la derecha, pues la mitad de los puntajes mayores se extienden más lejos que la mitad de los puntajes menores, sucediendo exactamente lo contrario en el caso de las mujeres, donde la distribución está sesgada a la izquierda. Se observa además, que la mediana de los puntajes de los hombres es mayor a la de las mujeres y que la dispersión de los puntajes de los hombres es ligeramente menor que la de las mujeres (en el caso de los hombres los puntajes van de 4 a 10, mientras que en las mujeres van de 2 a 9). Teniendo en cuenta lo anterior las tres propiedades de las distribuciones (simetría, centro y dispersión) podemos decir que los hombres obtuvieron mejores puntajes que las mujeres en el examen realizado.

En el presente ejemplo, nos se proporcionan otras variables, como serían las calificaciones en otras pruebas realizadas en distintos momentos, o sobre materias diferentes a al evaluada, para ver si se mantiene el mismo comportamiento, lo que daría mayor profundidad al análisis.

Un tercer nivel de análisis puede consistir en tratar de ir más allá de los datos observados, planteando hipótesis o conjeturas que se podrían poner a prueba mediante un análisis de tipo confirmatorio (por ejemplo, una prueba de hipótesis). Una pregunta para investigar sería: ¿Existen diferencias significativas de género en los puntajes obtenidos en la materia evaluada?

POSIBILIDADES PEDAGÓGICAS DEL EDA

La enseñanza del análisis exploratorio de datos, se encuentra en concordancia con la propuesta de un renombrado grupo de educadores estadísticos que en la década de los noventa fueron encomendados por parte de la MAA (Mathematical American Association) y la ASA (American Statistical Association) para revisar la enseñanza de la estadística en los Estados Unidos. Sus conclusiones fueron que la estadística, en tanto es la ciencia de los datos, debe ser enseñanza como tal, no como un campo de las matemáticas, por lo tanto, su enseñanza debe enfocarse a datos y conceptos, disminuir el uso de fórmulas y procedimientos y automatizar hasta donde sea posible los cálculos y las gráficas.

Batanero, Estepa & Godino (1991), consideran algunas características del análisis exploratorio de datos que lo hacen un tema apropiado de estudio en la enseñanza de secundaria y bachillerato:

1. Posibilidad de generar situaciones de aprendizaje referidas a temas de interés para el alumno.
2. Fuerte apoyo en representaciones gráficas
3. No necesita una teoría matemática compleja
4. Uso de diferentes escalas o reexpresión.

Mientras que Ben-Zvi (2004) señala que la oportunidad pedagógica que permite EDA en la exploración de datos ayudada mediante tecnología, esta en línea con paradigmas educativos actuales como son:

1. La enseñanza y aprendizaje con comprensión
2. Aprendizaje basado en investigación y preguntas.
3. Aprendizaje basado en proyectos

PAQUETES DE SOFTWARE DISPONIBLES PARA EL ANÁLISIS EXPLORATORIO DE DATOS

Existen en la actualidad diversos paquetes computacionales con capacidades gráficas y de procesamiento de datos que pueden utilizarse en la enseñanza del análisis exploratorio de datos. Quizá el más conocido

sea la hoja de cálculo Microsoft Excel, ya que es parte del software que comúnmente traen las computadoras nuevas. Excel tiene esa gran ventaja sobre las demás herramientas, que está disponible prácticamente para cualquier usuario. Sin embargo, dado que no es propiamente un paquete para hacer análisis estadístico, tiene muchas limitaciones en el manejo de las gráficas. Otra opción la constituyen los paquetes estadísticos profesionales, por ejemplo, SPSS u STATGRAPHICS, que disponen de una amplia gama de recursos gráficos para el análisis. Una tercer opción es el software FATHOM, que ha sido utilizado para el análisis de la actividad que se muestra en este trabajo. A diferencia de las otras opciones, FATHOM es un software dinámico que ha sido pensado con propósitos de enseñanza, por lo que es muy flexible y abundante en el uso de representaciones gráficas de los datos.

UN EJEMPLO DE ANÁLISIS EXPLORATORIO DE DATOS

Actividad 1:
La presente actividad se desarrolla a partir de un archivo de datos que contiene diversas variables medidas a 45 pacientes hombres admitidos en un Hospital en la ciudad de Auckland Nueva Zelanda, después de haber sufrido un ataque al corazón. Dichos datos han sido tomados del libro: Chance Encounters: A first course in Data Análisis and Inference (Wild & Seber, 2000) y también se encuentran disponibles en la dirección de internet (http://www.stat.auckland.ac.nz/~wild/ChanceEnc/WSdata/Ch02data/heart.txt). A continuación se detallan los nombres de las variables y su significado.

ID:	Número de identificación del paciente.
IZQUIERDO:	Porcentaje de sangre en el ventrículo izquierdo del corazón expulsado en un latido.
VOLSYS:	Volumen sistólico final (Una medida del tamaño del corazón)
VOLDIAS:	Volumen final diastólico (Otra medida del tamaño del corazón)
BLOQUEO:	Porcentaje del miocardio del ventrículo izquierdo con arterias que están totalmente bloqueadas.
ESTRECHO:	Porcentaje de arterias que están muy estrechas pero no bloqueadas
TIEMPO:	Tiempo en meses desde que el paciente fue admitido hasta el resultado.
RESULTADO:	0 = vivo, 1 = muerte cardiaca súbita, 2 = muerte dentro de 30 días de haber sufrido el ataque, 3 = muerte por falla en el corazón, 4 = muerte durante o después de una cirugía coronaria, 5 = muerte no cardiaca
EDAD:	Edad en años del paciente
FUMA:	Si el paciente fuma o no 1 = si, 2 = no
BETA:	Si el paciente estuvo tomando drogas (beta-bloqueadores); 1 = si, 2 = no
COLEST:	Colesterol en la sangre
CIRUGÍA:	Si el paciente tuvo cirugía 0 = no cirugía, 1 = cirugía como parte del tratamiento, 2 = cirugía por síntomas dentro de un año, 3 = cirugía por síntomas dentro de 1 a 5 años, 4 = cirugía por síntomas después de 5 años.

Tabla 1. Datos de 45 pacientes que ingresaron a un hospital después de haber sufrido un ataque al corazón.

INFARTO_CORAZON

	ID	IZQUIERDO	VOLSYS	VOLDIAS	BLOQUEO	ESTRECHO	TIEMPO	RESULTADO	EDAD	FUMADOR	BETA	COLEST	CIRUGIA
28	283	58	71	167	27	0	138	0	45	SI	NO	46	0
29	210	42	92	159	0	0	139	0	57	SI	SI	58	0
30	397	68	50	156	0	100	138	0	51	SI	NO		0
31	211	43	146	259	47	33	3	1	56	SI	SI	70	0
32	398	67	43	130	0	70	138	0	49	SI	SI		3
33	284	52	70	146	0	23	137	0	47	NO	SI		0
34	399	63	73	195	27	0	136	0	36	NO	NO	61	0
35	285	54	62	133	33	23	137	0	38	SI	SI		0
36	71	37	93	148	47	0	137	0	59	SI	SI		0
37	286	51	65	133	43	7	136	0	54	SI	SI		0
38	212	42	95	163	40	10	109	3	57	SI	SI		4
39	400	66	49	144	10	50	65	1	52	SI	SI	55	0
40	287	54	66	145	7	40	136	0	47	SI	SI	62	0
41	81	39	144	237	13	87	136	0	39	SI	SI	56	3
42	813	63	52	141	0	47	43	3	48	SI	SI		0
43	68	30	219	314	33	45	76	1	53	NO	SI		0
44	288	59	39	94	0	0	135	0	47	NO	SI	63	0
45	407	67	39	117	0	73	53	1	57	SI	SI	62	2

Dada las limitaciones de espacio, solo nos limitaremos a la exploración de algunas variables.

La metodología del análisis que sugerimos puede empezar considerando variables en forma independiente (ver los datos, en el sentido de Curcio, 1989), para posteriormente pasar al análisis entre variables (ver entre los datos), y plantear algunas preguntas e hipótesis que sugieren los datos (ver más allá de los datos). Es conveniente utilizar diferentes tipos de gráficas, ya que cada una de ellas puede aportar información distinta y que puede ser de importancia para el conocer el comportamiento de los datos. Finalmente podemos añadir resúmenes numéricos de aspectos concretos de los datos.

Primera etapa: Ver los datos

Podemos empezar graficando por ejemplo, la edad de los pacientes, sin son fumadores o no, los niveles de colesterol, buscando que los estudiantes consideren si dichas variables influyen en el riesgo de sufrir un infarto al corazón. Algunas preguntas que podrían plantearse a este nivel serían: ¿quiénes sufren más infartos, los fumadores o los no fumadores?, ¿Cuál es el intervalo de edad más frecuente de los pacientes infartados? ¿Los niveles de colesterol de los pacientes están dentro de los límites aceptables o están por encima de ellos?

En esta etapa se puede discutir el tipo de variable y la gráfica más conveniente para representarla. Por ejemplo, en la figura 1 se opta por una histograma y por un diagrama de caja, porque se trata de una variable continua, mientras que en la figura 2 se decide por una diagrama de barras y por un diagrama de barras acumuladas (Ribbon Chart), pues se trata de una variable categórica. Con ello se busca que los alumnos sean conscientes que el tipo de gráfica está en función del tipo de datos que quieren representar. Para un mismo tipo de variable, se puede discutir incluso las ventajas de una gráfica sobre la otra. Por ejemplo, el diagrama de caja nos proporciona el valor de cinco estadísticos muy importantes (los percentiles 25%, 50%, 75% y 100% y la mediana), mientras que el histograma nos muestra la frecuencia de cada edad, información que no proporciona el diagrama de cajas.

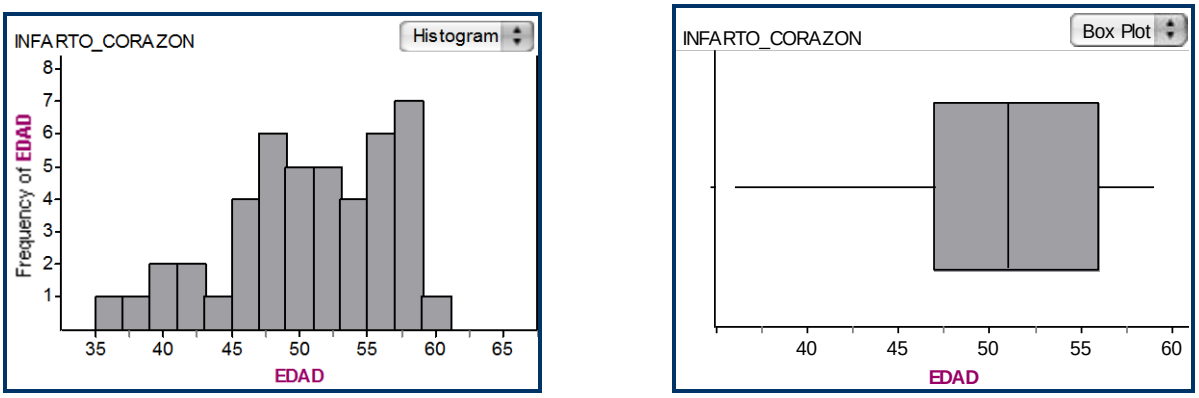


Fig. 1: Distribución de la edad de los pacientes que sufrieron ataque al corazón

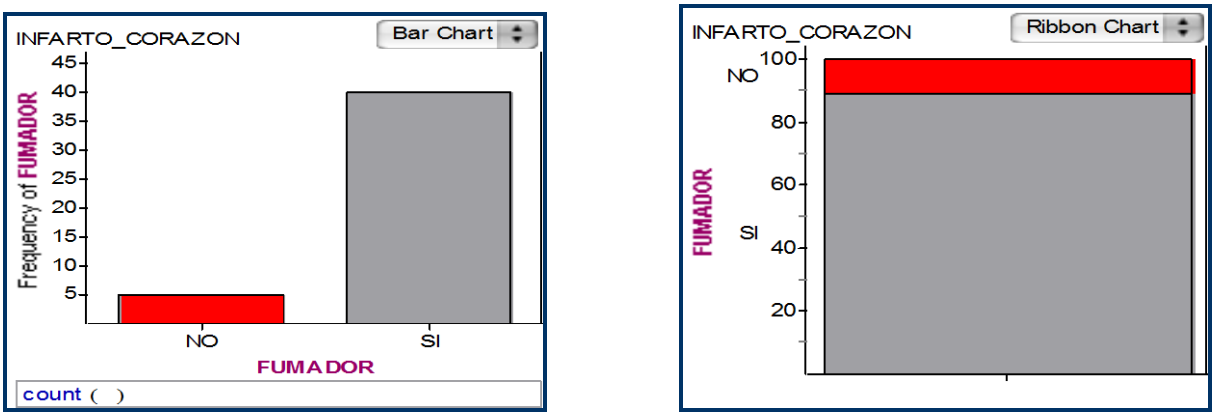


Fig. 2: Distribución de los pacientes infartados según su condición de fumador

Segunda etapa: Ver entre los datos

En esta fase se puede explorar la relación que existe en algunas variables. Para ello se pueden graficar sobre un mismo sistema de ejes coordenados dos variables a la vez (una sobre el eje de las abcisas y otra sobre el eje de las ordenadas). Por ejemplo, supongamos que nos interesa explorar la relación entre la variable edad y la variable resultado del paciente. Una gráfica de apropiada puede ser un diagrama de puntos como la que se muestra en la figura 3.

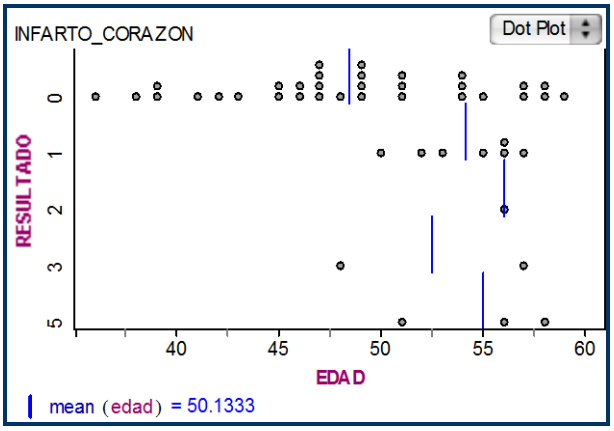


Fig. 3: Gráfica de puntos (Dot Plots) con los promedios de edad por categoría de resultado.

Obsérvese que la variable edad es numérica discreta y la variable resultado es categórica con cinco niveles de clasificación, por ello el software proporciona cinco gráficas de puntos en una (una para cada nivel del resultado).

Lo importante aquí es comprar las cinco categorías del resultado (0, 1, 2, 3, 4 y 5) respecto a la variable edad. Podemos observar que la categoría 0 (sobrevivientes) tiene a los pacientes más jóvenes, así lo muestra el comportamiento de la gráfica y sus medias por cada categoría. El grupo 0 tiene el promedio de edad menor de todos los grupos. Obsérvese que en este caso estamos recurriendo al cálculo de medidas descriptivas para complementar el análisis de datos.

Otra exploración interesante podría ser entre las variables colesterol y edad, como se muestra en el siguiente diagrama de puntos.

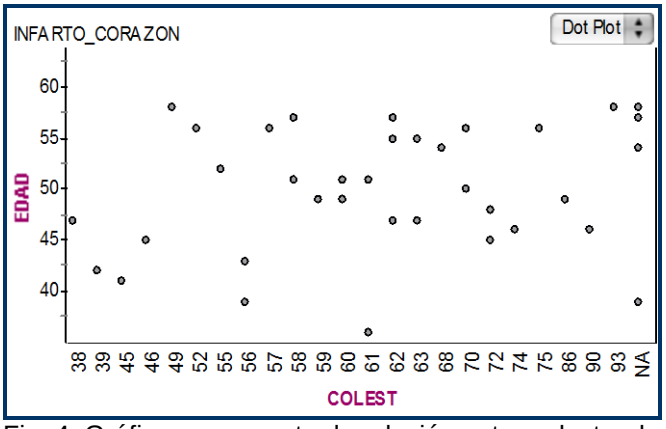


Fig. 4: Gráfica que muestra la relación entre colesterol y edad

La gráfica muestra que no hay un patrón predecible entre ambas variables, es decir no podemos decir por ejemplo que los pacientes de mayor edad presentaron mayor nivel de colesterol, puesto que como hay pacientes mayores con un nivel alto de colesterol también lo hay pacientes jóvenes. En suma, no existe una tendencia predecible.

Finalmente podríamos verificar la relación entre el porcentaje de arterias estrechas pero no bloqueadas (estrecho) con la variable fumador. Para ello recurrimos a un diagrama de cajas simultáneo como se muestra en la siguiente gráfica.

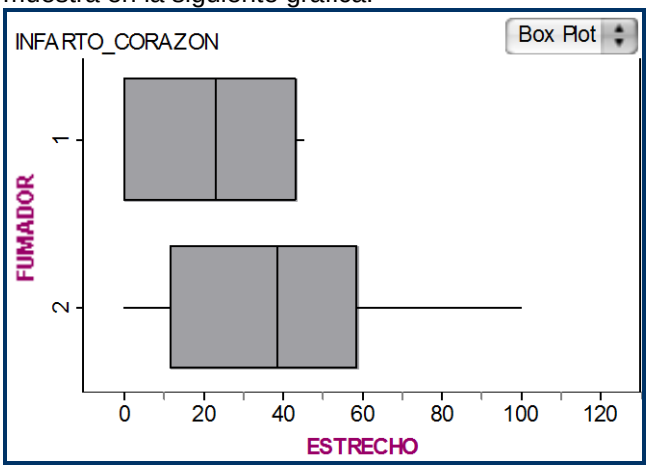


Fig. 5: Relación entre paciente fumador y porcentaje de arterias estrechas

Obsérvese que la mediana de los fumadores es mayor que la mediana de los no fumadores, es decir, los pacientes fumadores tuvieron un mayor porcentaje de arterias estrechas que los no fumadores.

CONCLUSIONES

El análisis exploratorio de datos facilita el vínculo de la estadística con otras áreas del conocimiento, situación que se promueve en muchos currículos actuales. Permite trabajar con datos reales que pueden ser obtenidos de diversas áreas como la geografía, biología, química u otras, a través de experimentos, bases de datos en internet, revistas u periódicos, o bien diseñar sus propias encuestas sobre asuntos de interés. Ello facilita el aprendizaje activo y colaborativo y le da a la estadística el rango de una ciencia de los datos en contexto, como lo señala Moore (1992).

Por otra parte, el análisis exploratorio de datos con ayuda de la herramienta computacional hace posible un análisis de múltiples variables, buscando establecer relaciones entre ellas y conjeturar algunas hipótesis sobre su relación y el efecto que pueden tener en el fenómeno estudiado. A diferencia de la estadística descriptiva que con frecuencia se limita al análisis de una sola variable.

En suma, consideramos que el análisis exploratorio de datos tiene mucho potencial para que los estudiantes aprendan estadística en contexto y mejoren su razonamiento estadístico.

REFERENCIAS

AEC (1990). A National statement on Mathematics for Australian schools*. Australian Education Council*. Currículum Corporation, Carlton Victoria, Australia.

Batanero, C.; Estepa, A. & Godino, J. (1991). Análisis exploratorio de datos: sus posibilidades en la enseñanza secundaria. *Revista Suma,* N° 9, pp. 25-31.

Ben-Zvi, D. (2004). Reasoning about Data Analysis. En D. Ben-Zvi & J. Garfield (Eds.). *The challenge of developing statistical literacy, reasoning and thinking*. Pp. 121-145. Kluwer Academic Publishers. Drdrecht Netherlands.

Curcio, F. R. (1989). *Developing Graph Comprehension.* NCTM, Reston, VA.

MEC (1992). *Matemática. Secundaria Obligatoria.* Madrid: Ministerio de Educación y Ciencia.

Moore, D. S. (1992). *What is Statistics?* En D. C. Hoaglin & D. S. Moore. *Perspectives on Contemporary Statistics*. Mathematical Association of America. MAA Notes 21. pp.1-17.

NCTM (1989). *Curriculum and evaluation standars for school Mathematics.* Reston, VA: National Council of Teachers of Mathematics.

NCTM (2000). *Principles and standards for school mathematics*. National Council Teachers of Mathematics. Reston, VA: National Council of Teachers of Mathematics.

SEP (2005). *Educación Secundaria. Matemáticas. Programas de Estudio*. Secretaría de Educación Pública. México.

Tukey, J. (1977). *Exploratory Data Analysis*, Reading , MA: Addison-Wesley.

Wild, Ch. J. & Seber, G. A. (2000). *Chance Encounters: A first course in Data Análisis and Inference.* John Wiley Sons.