IX Encuentro de Matemática

y sus Aplicaciones

Análisis Exploratorio

de Datos

Escuela Politécnica Nacional

Departamento de Matemática

Julio 2004

Introducción

Exploratory data analysis is detective work − numerical detective work− or counting detective work − or graphical detective work

Tukey, 1977 (pág 1)

El análisis exploratorio de datos (EDA), según unos, nueva rama de la estadística, según otros, una extensión de la estadística descriptiva, propugna un cambio de actitud y de enfoque metodológico ante el análisis de datos.

El EDA propugna que previo a cualquier análisis estadístico, es necesario un examen cualitativo de los datos, hay que comprender y reflexionar sobre la información que ellos contienen.

La estadística descriptiva clásica se ocupa de describir los datos a través de gráficos y de algunas medidas de tendencia central y de dispersión. El EDA tiene los mismos objetivos pero además pretende detectar anomalías o errores en las distribuciones univariantes de los datos. También intenta descubrir patrones o modelos. Para ello incorpora nuevas técnicas gráficas y busca estadísticos resistentes y robustos basados en las estadísticas de orden y centrados en la mediana.

El EDA potencializa los índices de forma, y la utilización de gráficos, prácticamente, como un índice más, *una gráfica bien realizada puede ser mas informativa que un conjunto de números.*

Estadística descriptiva

Frecuencias e histogramas

Supongamos que se tiene un conjunto de n observaciones , denominado muestra, de una variable aleatoria X o de una población !. Uno de los problemas es conocer la distribución de la muestra. Con este fin se divide a la muestra en rangos o clases continuas de igual longitud, sean los rangos.

Frecuencia relativa

La frecuencia relativa de la clase rj es:

Si la muestra es aleatoria y n es suficientemente grande, se puede hacer la siguiente aproximación

es la probabilidad empírica (suma de frecuencias), mientras que *P* es la probabilidad teórica.

Histograma

El gráfico de las frecuencias relativas; en ordenadas los valores fj y en abscisas las clases rj, se denomina histograma. Es claro que la forma del histograma depende del número de clases, no deben ser muchas ni muy pocas. No existe una regla que determine el número de clases, en general deben ser alrededor de y no menos de 5.

Función de distribución (empírica)

También se define la función de distribución acumulada

donde

Uno de los pilares de la estadística clásica es la convergencia de la distribución empírica hacia la distribución teórica.

Teorema de Glivenko − Cantelli

Ejemplo

Suponga que se ha seleccionado una muestra aleatoria simple de 15 personas y se les ha preguntado su salario mensual en dólares. Los salarios, previamente ordenados, son:

53, 86, 163, 183, 206, 224, 259, 652, 842,

1139, 1433, 2198, 2215, 2410, 4592

Como dividiremos la muestra en 5 clases.

Clase	**ni**	**fj**		**Histograma**
Menos de 620	7	0.47	0.47	/tr>
De 620 a 1755	4	0.27	0.73	>
De 1755 a 2889	3	0.20	0.93	
De 2889 a 4024	0	0	0.93	
Más de 4024	1	0.07	1.00	table> El histograma pone en evidencia una distribución completa−mente asimétrica. El 47% tienen salarios inferiores a $620, mientras que el 7% tienen salarios superiores a 4024. ***Medidas de tendencia central y de dispersión*** **Media** Es el índice clásico de tendencia central. Se define por:

Moda

Es el valor o los valores mas frecuentes.

Su uso es restringido porque pueden existir varias modas o su frecuencia puede ser irrelevante con respecto a la frecuencia de los otros datos, en especial cuando los datos son de tipo continuo. Si en el ejemplo anterior, calculamos la frecuencia de cada uno de los salarios, vemos que todos tiene la misma frecuencia, no existe una moda, pero si consideramos los 5 rangos de salarios, existe uno que es claramente modal, el primero.

Varianza

Es una medida de dispersión, en promedio, mide como se alejan los datos de la media. Su definición es:

Desviación estándar

El problema de la varianza es que sus unidades están elevadas al cuadrado, por ejemplo si calculamos la varianza de los salarios tendríamos *dólares al cuadrado*, lo que no tiene mucho sentido. Por esta razón, se acostumbra calcular la **raíz cuadrada de la varianza**, lo que se denomina desviación estándar.

Coeficiente de variación

Es una medida adimensional de la dispersión. Es la dispersión con respecto a la media, su fórmula es:

Ejemplo. Para los 15 salarios se tiene:

Tanto la desviación estándar como el coeficiente de variación ponen en evidencia que la dispersión es muy grande, la desviación estándar es 1.15 veces la media. Existen salarios muy bajos y salarios muy altos.

Algunos índices EDA

Los índices EDA se clasifican en:

- **Localización**: corresponderían a los índices de posición y tendencia central clásicos, indicando los valores límites y promedios de la distribución.
- **Dispersión**: indican el grado de agrupación o disgregación en la distribución. Cuanto menor sea su valor, mas información aportaran los índices de localización.
- **Forma:** evalúan la forma de la distribución de los datos desde ejes verticales (simetría) y desde ejes horizonta−les (curtosis).
- **Gráficos**: mostraran las agrupaciones internas de los valores e indicarán los índices que mejor representan a la distribución.

Indices de localización

Las medidas vistas en estadística descriptiva, son sensibles a los valores extremos, así por ejemplo: si eliminamos el último salario, 4592, se tiene:

Debido a la prioridad que concede el enfoque EDA a la resistencia y a la robustez, sus índices se basan en los percentiles.

Definición. Sea , un percentil de orden es un número real tal que, aproximadamente, 100% de valores xi son inferiores a y 100(1−)% aproximada−mente, son superiores a dicho valor, mas precisamente:

o lo que es equivalente

El percentil divide al conjunto de datos en dos subconjuntos: uno de peso aproximado , a la izquierda de , y otro de peso aproximado (1−), a la derecha de . Para su cálculo se procede como sigue:

- Se ordenan los valores xi de menor a mayor. Escribiremos los valores ordenados.
- Se encuentra el entero menor ( ) del producto , y el entero mayor ()
-

Observación. Si no es entero, entonces y .

Percentiles particulares

- El percentil de orden 0.5 se denomina mediana (Md).
- Los percentiles de órdenes: 0.25, 0.50, y 0.75 se denominan cuartiles: primer cuartil (Q1), segundo cuartil (Q2) y tercer cuartil (Q3) respectivamente. Observe que .
- Los percentiles de órdenes: 0.2, 0.4, 0.6, 0.8 se denominan quintiles.
- Los percentiles de órdenes: 0.1, 0.2, 0.3, ... , 0.9 se denominan deciles.

Los índices EDA de localización son.

Mediana

De acuerdo a la regla dada para el calculo de percentiles

Promedio de cuartiles

Trimedia

Centrimedia o media intercuartílica (MID)

Es el promedio de los valores xi, no repetidos, que se encuentran entre los cuartiles Q1, Q3. Se debe procurar que el número de valores a cada lado de la mediana sea el mismo. Se puede introducir observaciones repetidas para equilibrar los dos costados.

Observaciones

- Si el conjunto de datos es centrado

cualquier diferencia entre estos índices refleja asimetría.

- Los cuatro índices EDA que hemos visto dan cuenta del 50% central de valores, no dependen del 25% de valores inferiores al primer cuartil y del 25% de valores superiores al tercer cuartil, por tanto son resistentes.

Ejemplo. Indices EDA para los salarios

Indice	Q1	Q2 = Md	Q3		TRI	MID
Valor	183	652	2198	1190.5	921.25	679.28

Indices de dispersión

Amplitud intercuartiles

Mediana de desviaciones absolutas

Indices estandarizados

Con el fin de comparar con la ley normal centrada y reducida se estandarizan los dos índices anteriores. Sus estandarizacio−nes se denominan seudo desviaciones estándar.

Los cuartiles de la ley normal centrada y reducida son:

su amplitud intercuartil es .

Las seudo desviaciones estándar son:

Ejemplo. Indices de dispersión y seudo desviaciones estándar para los salarios.

Indice	IQR	MAD	Sd(IQR)	Sd(MAD)
Valor	2015	489	1493.69	724.98

La amplitud intercuartil del lote de salarios es 1494 veces la amplitud intercuartil de la ley normal centrada y reducida. La mediana de desviaciones absolutas de los salarios es 725 veces superior a la correspondiente de la ley normal centrada y reducida.

Indices de forma

Los índices de forma constituyen el principal aporte del EDA

Indice de Yule

- Si H1 = 0, la distribución es simétrica.
- Si H1 > 0, la asimetría es positiva.

La distribución es alargada hacia los valores superiores a la mediana.
- Si H1 < 0, la simetría es negativa. La distribución es alargada hacia los valores inferiores a la mediana.

Indices de simetría de Kelly

La ventaja de H3 sobre H2 es su adimensionalidad. Se interpreta de forma idéntica al índice de Yule.

Coeficiente de curtosis

o bien empleando octiles

- Si K1 o K2 = 1, la distribución es mesocúrtica.
- Si K1 o K2 > 1, la distribución es leptocúrtica (alargada).
- Si K1 o K2 < 1, la distribución es platicúrtica (plana).

Ejemplo. Indices de simetría y curtosis para los salarios.

Indice	H1	H3	K1	K2
Valor	0.82	0.91	0.61	0.61

Por el tamaño de la muestra: los deciles (extremos) coinciden con los octiles. La distribución tiene una marcada asimetría positiva y es platicúrtica.

Gráficos EDA

Diagrama de puntos

En el gráfico anterior se muestran los salarios repartidos en una recta numérica, este gráfico se denomina diagrama de puntos. Es muy útil para visualizar un conjunto pequeño de datos.

El gráfico muestra la concentración y la dispersión de los mismos. En el caso del ejemplo los salarios se concentran hacia los valores bajos, existe un salario muy alto con respecto al resto.

Diagrama tronco y hojas

Es un diagrama que puede sustituir al histograma. La principal crítica a los histogramas es que los datos se dividen en rangos cuyos extremos pueden no ser representativos de la distribución interna de los datos o no reflejar sus posibles sub−agrupaciones.

El enfoque EDA propone la utilización de representaciones gráficas que potencien la visualización de la información, no solo en lo cualitativo sino en lo cuantitativo, conservando en lo posible los propios valores numéricos.

Los números xi se dividen en dos partes: un tronco formado por el primer dígito o por los dos primeros dígitos, y una hoja por el siguiente dígito. Se desprecian el resto de dígitos.

La parte que define el grupo (el rango en el histograma) es el tronco, éstos se colocan en una columna ordenada a intervalos constantes, desde el valor mas bajo hasta el valor mas alto. Se hallen presentes o no los valores intermedios.

Ejemplo

Para realizar el diagrama tronco y hojas para los salarios, podemos separar los dos primeros salarios, por ser muy pequeños, y suponer que todos los números restantes están formados por cuatro dígitos, a los números de tres dígitos les anteponemos el cero. Si tomamos el primer dígito como tronco y el siguiente como hoja se tiene el gráfico adjunto.

Lo: 53, 86		
Fre.	**Tronco**	**Hojas**
(9)	0	1 1 2 2 2 6 8
6	1	1 4
4	2	1 2 4
1	3	
1	4	5
Unidad = 1000; 1|1 = 1100 − 1199		

- Los dos salarios mas bajos constan en la parte superior acompañados de la palabra Lo = lower.

- En la última fila hemos añadido la unidad, ésta nos indica que son unidades de 1000 y, que si el trono es 1 y la hoja es 1 (1|1) significa que el salario puede ir desde 1100 hasta 1199.

- La primera columna es la frecuencia absoluta acumulada. Las frecuencias se acumulan tanto desde arriba hacia abajo como desde abajo hacia arriba, se encuentran en la clase que contiene la mediana, la misma que se escribe entre paréntesis.

Como se puede ver este gráfico es mucho mas informativo que el histograma y sus clases son menos arbitrarias, prácticamente están determinadas por los valores observados. No obstante el número de clases también puede variar de acuerdo a los mismos criterios de construcción de los histogramas.

En el diagrama del ejemplo se puede ver: la concentración de salarios bajos, al igual que la existencia de un salario muy alto. Además, algo que no se ve en un histograma: hay 2 salarios entre 100 y 199, 3 salarios entre 200 y 299, 1 salario entre 600 y 699, etc.. Hay nueve salarios inferiores a 1000, 2 salarios entre 1100 y 1499, 3 salarios ente 2100 y 2499 , 0 salarios entre 3000 y 3999, 1 salario entre 4500 y 4599.

Si el diagrama anterior no nos satisface, porque concentra mucho los datos, se pueden subdividir los troncos. El tronco 1 se subdividir en dos: 1L para las hojas 0,1,2,3,4 y 1U para las hojas 5,6,7,8,9. como se muestra en el siguiente diagrama.

Lo: 53, 86		
Frec.	Tronco	Hoja
7	0L	1 1 2 2 2

(2)	0U	6 8
6	1L	1 4
4	1U	
4	2L	1 2 4
1	2U	
1	3L	
1	3U	
1	4L	
1	4U	5
Unidad = 1000; 1U|1 = 1100 − 1199		

Si se quiere desagregar mucho mas cada tallo (original) se subdivide en 5 partes. 1z, 1t, 1f, 1s y 1e para las hojas, {0,1}, {2,3}, {4,5},{6,7},{8,9}, respectivamente. El nuevo diagrama se presenta en la página siguiente. En él se incluye una fila para la observación masa alta y que ahora se visualiza muy alejada del resto.

Diagrama de caja

Es una presentación visual que describe al mismo tiempo varias características importantes de un conjunto de datos, tales como: el centro, la dispersión, la asimetría y la identificación de observaciones que se alejan de forma poco usual del resto de datos.

El diagrama de caja se basa en los cuartiles y en los valores extremos (xmin y xmax). Su presentación puede ser vertical u horizontal. Se colocan a escala los cuartiles Q1, Q2, Q3. Se realizan pequeños trazos que indican su posición y se forma una caja con ellos, así (Ver gráfico en la página siguiente.)

Lo: 53, 86		
Frec	**Tronco**	**Hoja**
4	0z	1 1
7	0t	2 2 2
7	0f	
(1)	0s	6
7	0e	8
6	1z	1

5	1t	
5	1f	4
4	1s	
4	1e	
4	2z	1
3	2t	2
2	2f	4
1	2s	
1	2e	
Hi : 4592		
Unidad = 1000; 1z|1 = 1100 − 1199		

Diagrama tronco − hoja

Diagrama de caja

A derecha e izquierda se trazan rayas cuya longitud máxima es 1.5*IQR*, a condición de que dicha longitud no exceda la posición de los valores extremos. Las observaciones cuyos valores superan estos límites se marcan individualmente, mediante cualquier símbolo que represente a los puntos. Las observaciones que se encuentran entre 1.5*IQR* y 3*IQR* (a cualquiera de los lados) se denominan **observaciones atípicas**, las que superan ese rango son **observaciones atípicas extremas.** En el diagrama de caja anterior hay una observación atípica y ninguna observación atípica extrema.

Los diagramas de caja son especialmente útiles cuando se quiere comparar varias muestras.

Parejas de variables

Introducción

Supongamos que se observa una pareja de variables (X,Y). X e Y son dos medidas que se observan sobre un mismo individuo.

Por ejemplo:

- X es la calificación de álgebra, Y es la calificación de Educación Física de un estudiante.

- X es la potencia de un vehículo, Y es su velocidad máxima.
- X es el ingreso de un hogar, Y es su gasto en consumo.
- X es la masa monetaria mensual, Y es la tasa de inflación mensual de un mismo país.

Supongamos que se dispone de una muestra {(x1,y1), (x2,y2),..., (xn,yn)}. El estudio de la pareja (X,Y) a partir de la muestra se lo puede realizar desde diferentes ángulos:

- Distribución de probabilidad conjunta.
- Descripción de los individuos a partir de los valores (xi,yi)
- Búsqueda de una relación funcional entre las variables.

Distribuciones de probabilidad asociadas

Distribución conjunta

Se divide en rangos, de acuerdo a los criterios antes indicados, tanto los valores {x1, x2, ...,xn} como los valores {y1, y2, ...,yn}.

Sean y los rangos, nij el número de observaciones . nij es la frecuencia absoluta de la clase cruzada $R_i^X \quad R_j^Y$. La frecuencia relativa se define por y se interpreta como una probabilidad:

El conjunto de todas estas frecuencias se denomina *distribución conjunta de {X,Y}*. Estas frecuencias pueden ser visualizadas en un histograma tridimensional, pero su representación suele ser poco útil.

Distribuciones marginales

Se puede calcular la distribución de cada una de las variables, éstas se denominan distribuciones marginales.

La distribución marginal de X es

La distribución marginal de Y se define de manera similar.

Distribuciones condicionales

También se puede calcular la distribución de X cundo Y toma un valor particular, lo que se denomina probabilidad condicional de X dado Y.

La probabilidad condicional de dado se define por

Es la probabilidad de sabiendo que . Igualmente se define la probabilidad condicional de dado

Para comparar las distribuciones condicionales se puede trazar en un solo gráfico sus histogramas. También se pueden calcular los diferentes índices antes estudiados a mas de sus diagramas de caja.

Descripción de los individuos

Para describir los individuos se puede recurrir a un gráfico de los puntos (xi,yi) en un plano cartesiano. Es de particular interés cuando los puntos forman grupos o una estructura particular, como en las siguientes figuras.

En el gráfico 1 hay 4 grupos mas o menos definidos. Para concretar las ideas supongamos que X es la calificación de álgebra y Y es la calificación en deportes. Leyendo en le sentido de las manecillas de un reloj: encontramos un grupo que tiene calificaciones alteas en las dos materias, el siguiente grupo tiene calificaciones satisfactorias en álgebra pero deficientes en deportes, el tercer grupo tiene calificaciones bajas en ambas materias, el grupo último tiene calificaciones bajas en álgebra paro satisfactorias en deportes.

El gráfico 2 muestra claramente una tendencia lineal sugiere que existe una relación lineal entre las variables .

En el gráfico 3 es difícil visualizar grupos o una relación de tipo funcional. No se

puede decir mucho sobre la relación entre las variables o las características de los individuos.

Búsqueda de una relación funcional

El método clásico se basa en la regresión lineal o mas generalmente en los modelos lineales generalizados. Aquí presentaremos un método alternativo, pero antes estudiaremos el coeficiente de correlación.

Coeficiente de correlación

Si las variables que se observan son cuantitativas, es decir si los valores observados (xi,yi) son valores numé−ricos se puede calcular la **covarianza** que se define por:

y la **correlación** que se define por:

La ventaja de la correlación sobre la covarianza es su adimensionalidad, a mas de los siguientes resultados:

Teorema. Para todo conjunto de observaciones numéricas {(x1,y1), (x2,y2),..., (xn,yn)} se tiene que:

Teorema. Para todo conjunto de observaciones numéricas {(x1,y1), (x2,y2),..., (xn,yn)}. si y solo si existen constantes a, b, tales que

Teorema. Si las variables X, Y son independientes, su correlación es nula.

Los teoremas anteriores permiten interpretar y comprender el coeficiente de correlación. Un coeficiente de correlación cercano a uno, en valor absoluto, sugiere la existencia de una relación lineal entre las observaciones. Una correlación cercana a cero puede ser causada por la independencia de las variables aleatorias o por una relación no lineal entre las observaciones, por ejemplo una relación cuadrática.

Consideremos el conjunto (−2,4), (−1,1), (0,0), (1,1), (2,4). Es claro que su

correlación es nula, en efecto:

lo que implica .

No obstante la relación entre X e Y es cuadrática, como se aprecia en el siguiente gráfico.

Recta de regresión de mínimos cuadrados

La ecuación de la recta que pasa por los puntos (x1,y1), (x2,y2) es , donde b es la pendiente, y está dada por:

Observe que si x = x1, entonces y = y1; y si x = x2, y = y2. Si tenemos n puntos para cada pareja podemos obtener una recta, así tendríamos n(n−1)/2 rectas.

El problema es: encontrar una recta que de alguna manera sea la mas próxima a todos los puntos.

El método de mínimos cuadrados propone encontrar una recta y = a + bx que minimice la suma de residuos al cuadrado

Los estimadores de mínimos cuadrados son:

La ecuación de la recta es: . La recta pasa por el punto y tiene pendiente .

El problema es que la recta puede estar determinada por pocos puntos y no reflejar la verdadera relación entre la mayoría de puntos.

Ejemplo. Suponga que se han realizado 10 observaciones de una pareja (X; Y).

En el primer gráfico se muestra la nube de los 10 puntos con la recta de mínimos cuadrados y la ecuación de la recta: **y = 4.8 + 3.7x**. Es evidente que hay un punto alejado del resto y es muy influyente.

En el segundo gráfico se ha eliminado el último punto. La ecuación de la nueva recta es: **y = 2.6 + 1.7x** que es muy distinta a la anterior.

Línea resistente o línea de Tukey

La línea resistente está ligada a un estadístico resistente, la mediana.

Cálculo de los coeficientes a, b

- Se divide a X en tercios, de acuerdo a los siguientes criterios:

Tercios	Si n = 3K	Si n = 3K+1	Si n = 3K+2
Inferior	K	K	K + 1
Medio	K	K + 1	K
Superior	K	K	K + 1

S varios puntos tienen el mismo valor, se asignan al mismo tercio, buscando siempre el equilibrio.

- En cada tercio se calcula la mediana de los {xi} y la me−diana de los {yi}. Sean: (xinf, yinf), (xmed, ymed), (xsup,ysup) las parejas de medianas de cada tercio.
- La pendiente de la recta resistente es:

la intersección con el eje Y es:

Ejemplo. Línea resistente para los datos del ejemplo anterior.

	Tercio inferior	Tercio medio