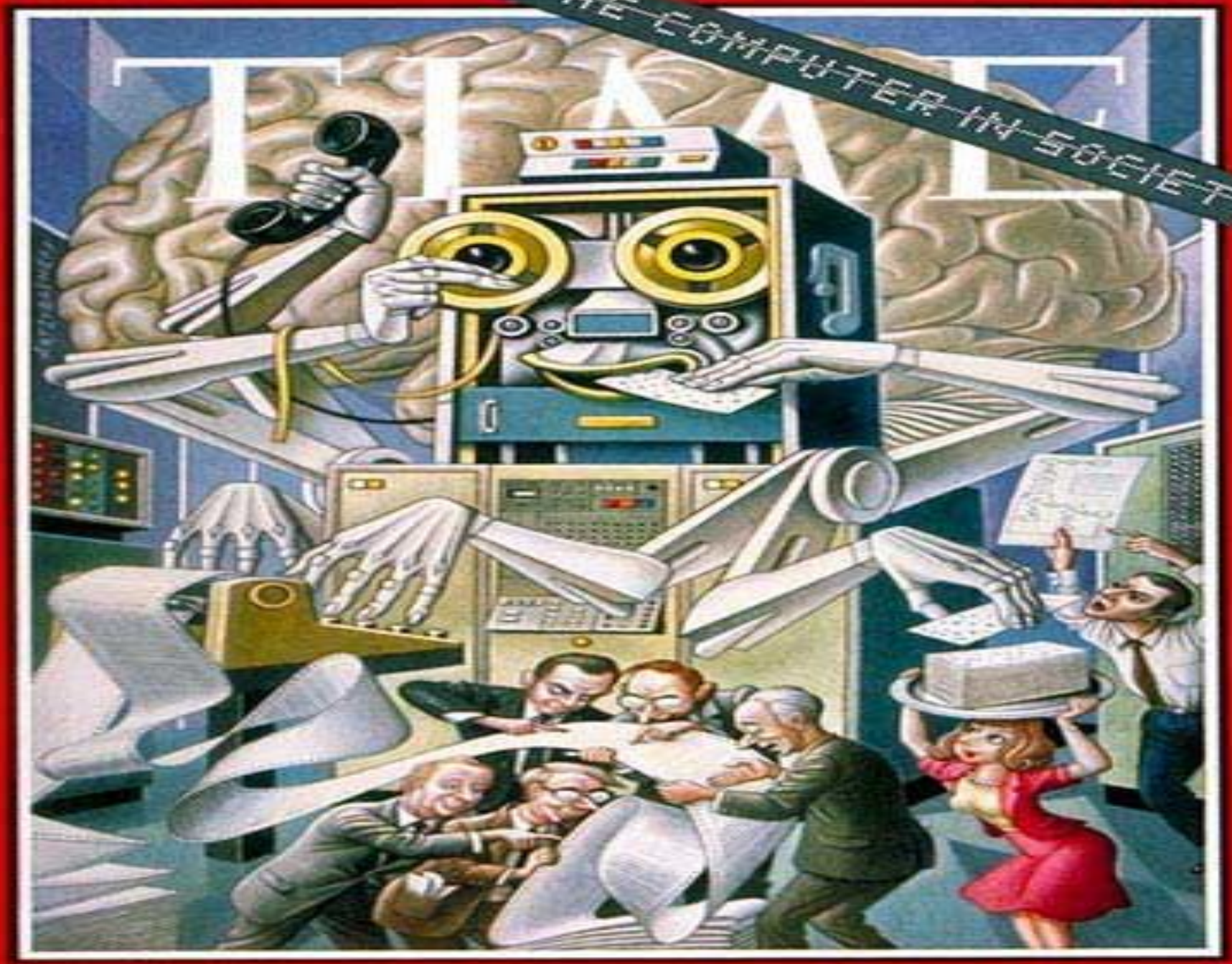
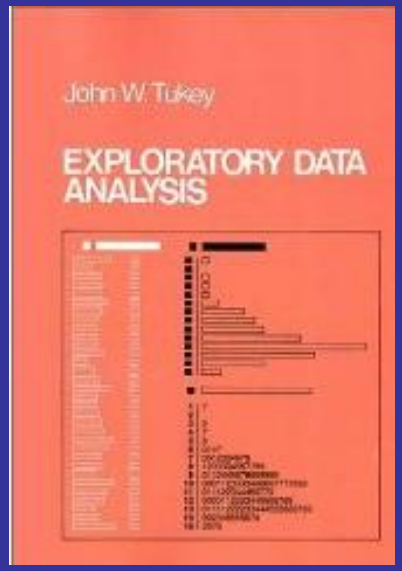
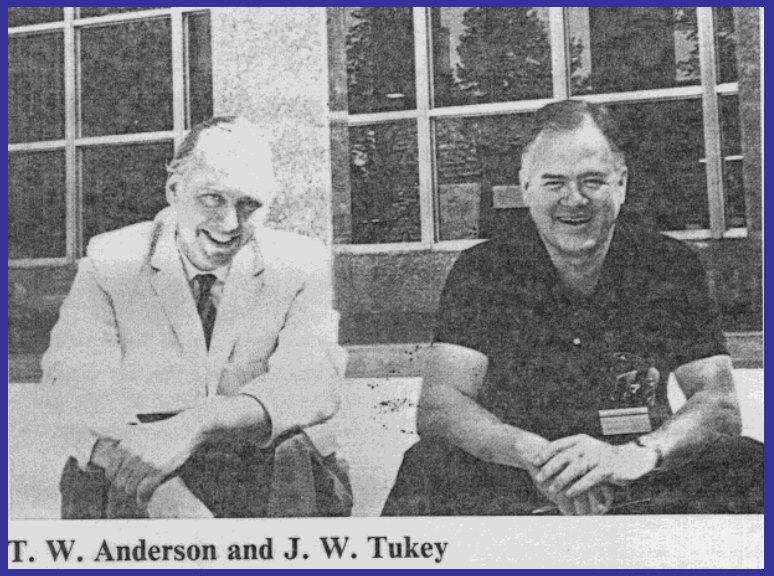

TIME
THE COMPUTER IN SOCIETY

Introducción al Análisis exploratorio (EDA)
Conceptos Fundamentales

Bibliografía

TUKEY, J.W.(1977).-*Exploratory Data Analysis.* Reading Mass. Addison & Wesley.





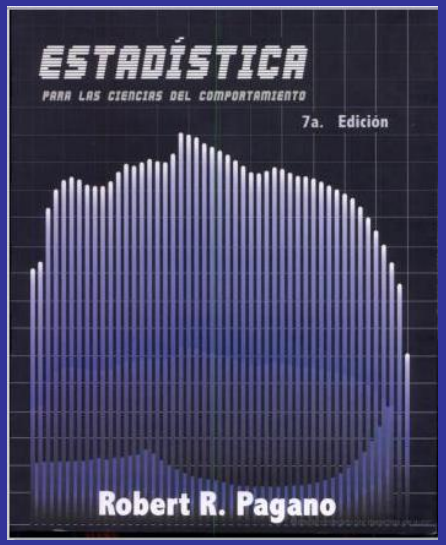
T. W. Anderson and J. W. Tukey

Bibliografía

Hoaglin, D C; Mosteller, F & Tukey, John Wilder (Eds) (1985). *Explorando la Tabla de Datos, Tendencias y Formas*. ISBN 0-471-09776-4.

Hoaglin, D C; Mosteller, F & Tukey, John Wilder (Eds) (1983). *Entendimiento Robusto y Análisis Exploratorio de Datos*. ISBN 0-471-09777-2.

Tukey, John Wilder (1977). *Análisis Exploratorio de datos*. ISBN 0-201-07616-0.



Estadística para ciencias del comportamiento
Escrito por Robert R. Pagano
Edition: 7
Publicado por Cengage Learning Editores, 2006
ISBN 9706865047, 9789706865045
http://books.google.es/books?id=zU1hmIJ4IrcC&pg=PA139&source=gbs_selected_pages&cad=0_1#PPA164,M1

Supuestos paramétricos

- Normalidad multivariable
- Homocedasticidad
- Tamaño muestral n>30
- Linealidad
- Nivel de medición continuo

EDA

1 Perspectiva desarrollada por TUKEY

El Análisis exploratorio de Datos aborda el análisis de datos a partir de conceptos como los de escepticismo, amplitud y flexibilidad en la investigación.

“Sólo examinando los datos podemos encontrar lo que no esperamos”

“Las medidas que resumen una distribución, son eso mismo resúmenes. El análisis debería comenzar sobre los datos, no sobre su sustituto bajo la forma de coeficientes”

EDA

1 Perspectiva exploratoria en el análisis estadístico de datos

El Análisis exploratorio de Datos es una perspectiva alternativa al enfoque confirmatorio tradicional

- La estadística confirmatoria tradicional se centra en los estadísticos de contraste de hipótesis a partir de preguntas como
  "¿Pueden los datos confirmar la hipótesis que relaciona la renta con el nivel de consumo de determinados productos, en la C.V?"
- Una alternativa (EDA), es partir de una pregunta diferente
  "¿que información me ofrecen los datos sobre la relación entre consumo de determinados productos y renta en la C.V?

EDA


TEORÍAS
DEDUCCIÓN LÓGICA
HIPÓTESIS
INTERPRETACIÓN, INSTRUMENTACIÓN, ESCALAJE Y MUESTREO
OBSERVACIONES
MEDICIÓN, RESUMEN DE MUESTRAS Y ESTIMACIÓN DE PARÁMETROS
GENERALIZACIONES EMPÍRICAS
FORMACIÓN DE CONCEPTOS, FORMACIÓN DE PROPOSIOCIONES Y ORDENAMIENTO DE PROPOSICIONES
INFERENCIA LÓGICA
DECISIONES PARA ACEPTAR O RECHAZARHIPÓTESIS
CONTRASTACIÓN DE HIPÓTESIS

EDA

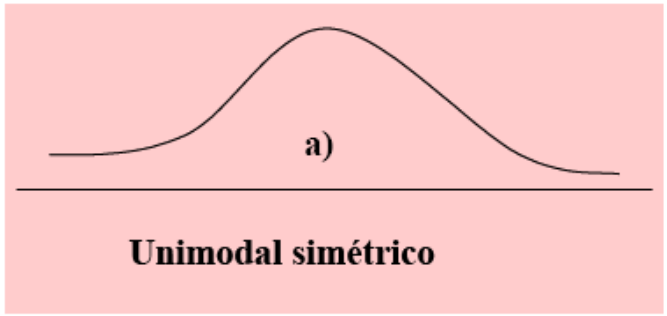
1 Objetivo del EDA

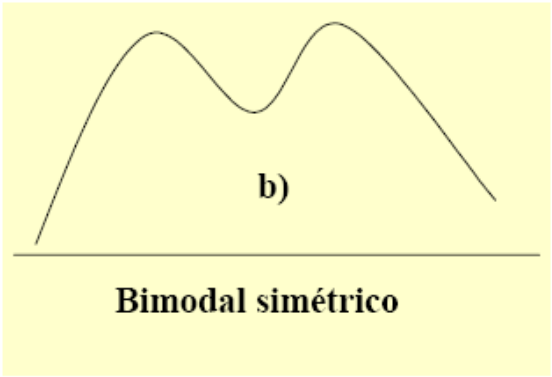
El objetivo en este tipo de análisis es doble:

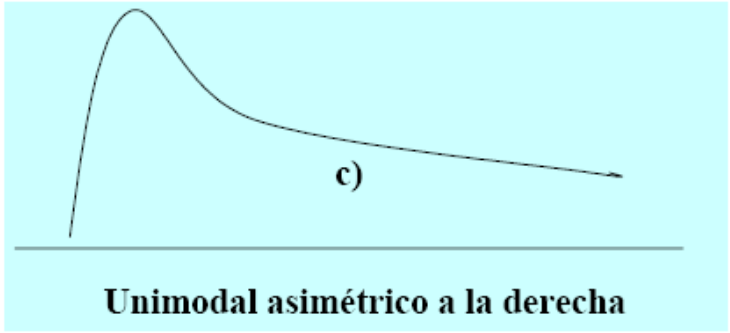
Explorar los datos para descubrir en ellos pautas subyacentes de estructura y relación que de otro modo no se detectarían

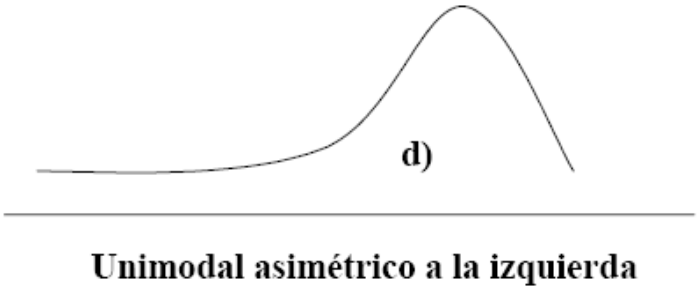
Explorar para observar si se cumplen los supuestos paramétricos y en función de ello elegir los estadísticos más adecuados en cada caso o realizar las modificaciones oportunas

EDA


a)
Unimodal simétrico


b)
Bimodal simétrico


c)
Unimodal asimétrico a la derecha


d)
Unimodal asimétrico a la izquierda

EDA

2 Definición y características principales de una distribución

Una distribución es un conjunto de datos ordenados, cada uno de los cuales representa los valores observados de una característica, a lo largo de un rango de valores entre lo casos de la muestra

EDA

2 Definición y características principales de una distribución

- Localización: hace referencia al anclaje de la distribución en torno a un conjunto de valores
- Dispersión: hace referencia a la variabilidad de los valores, a cuan ancha es la distribución y orienta sobre el numero de valores extremos
- Forma: hace referencia al tipo de distribución, si esta tiende a la normal, si es simétrica, monomodal etc,

EDA

Caracterización de una distribución en el Análisis Univariable

- **LOCALIZACIÓN:** Media / Mediana / Moda

- **DISPERSIÓN:** Desviación típica / Recorrido Intercuartílico

- **FORMA:** Normal, monomodal, simétrica, asimétrica positiva, bimodal asimétrica negativa...

EDA

3 Una nueva terminología

Toda la filosofía que subyace a este nuevo enfoque se refleja tanto en la estructura de sus gráficos como en la terminología que utiliza. Esta, en consonancia con las premisa de esta perspectiva, tiene por objeto el desarrollo de conceptos ilustrativos que reflejen la esencia del EDA. Así conceptos como ***suave***, ***rugoso***, ***robusto***, etc… hacen referencia a otros tantos aspectos fundamentales desde esta orientación.

EDA

SUAVE: Es lo que subyace a lo datos, la estructura simplificada de un conjunto de observaciones. Es la forma general de la distribución o la forma general de la relación. Es la regularidad o el patrón de los datos

ASPERO: Es lo que se aparta de la regularidad, del patrón general de los datos. A este respecto el supuesto *ideológico* que subyace al enfoque tradicional es que estas "rugosidades" o desviaciones respecto a la norma de los datos, son errores de medición

EDA

ROBUSTO: Una medida resistente o robusta es aquella que se ve poco afectada por los cambios en una proporción pequeña de casos, no importa la magnitud de los cambios

Tradicionalmente el análisis confirmatorio se centra en lo Suave aunque el análisis de lo rugoso encierra un gran valor heurístico. En este sentido la estrategia confirmatoria es un sofisma pues impone un modelo a los datos, de tal modo que cuando los analiza obtiene ese mismo modelo, y por tanto suele confirmar la teoría de partida

EDA

ENFOQUE CONFIRMATORIO	ENFOQUE EXPLORATORIO
• Parte de un modelo que intenta imponer a los datos para comprobar las hipótesis de partida	• Parte de los datos para averiguar las pautas de distribución y modelos de relación que subyacen a éstos y a partir de estos resultados generar hipótesis
• Utiliza estadísticos poco robustos como la media y la desviación típica	• Utiliza estadísticos robustos como la mediana y el recorrido intercuartílico o dispersión media o midspeard
• Hace un mayor énfasis en las representaciones numéricas que en las gráficas	• Hace un mayor énfasis en las representaciones gráficas que en las numéricas

EDA

4 Resumen numérico robusto de las características de una distribución

- Localización: Estadística confirmatoria: MEDIA
  Estadística exploratoria: MEDIANA

{ 1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 6 } Media = 3,5 Mediana 3,5

{ 1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 60 } Media = 7,8 Mediana 4

EDA

4 Resumen numérico robusto de las características de una distribución

- Localización: Estadística confirmatoria: MEDIA
  Estadística exploratoria: MEDIANA

Propiedad de la media $\Sigma (X_i - \text{Media}) = 0$

EDA

4 Resumen numérico robusto de las características de una distribución

- Dispersión: Estadística confirmatoria: Desviación típica
  Estadística exploratoria: Dispersión media
  o (recorrido intercuartílico)

{ 1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 6 } S = 1,46 IQR = 2,5

{ 1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 60 } S = 15,7 IQR = 2,5

EDA

4 Resumen numérico robusto de las características de una distribución

- Dispersión: Estadística confirmatoria: DESVIACIÓN TÍPICA
  Estadística exploratoria: DISPERSION MEDIA

Propiedad de la Desviación típica

$$S = \sqrt{\frac{\Sigma (X_i - \text{Media})^2}{N}}$$

EDA

LOCALIZACIÓN, DISPERSIÓN Y FORMA DESDE LA ÓPTICA ROBUSTA

ASIMÉTRICA NEGATIVA

ei	Q1	Medn	Q3	es
9,5	40,1	45,05	48,6	50,4
30,60	4,95		3,55	1,80
	35,5	8,5	5,35	

DISTRIBUCIÓN NORMAL [1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 6]

1	2,5	3,5	4,5	6
1,5	1		1	1,5
	2,5	2	2,5	

Medn - ei = es - Medn
Medn - Q1 = Q3 - Medn
Q1 - ei = es - Q3

EDA

	Perspectiva confirmatoria	Perspectiva exploratoria
Localización	MEDIA	Mediana
Dispersión	DESVIACION TIPICA	Dispersión media o IQR
Forma	ASIMETRÍA Y CURTOSIS	Resumen numérico robusto y representaciones gráficas STEM and LEAF BOX PLOT

EDA

Aspectos a observar desde EDA Univariable

- ASIMETRÍAS / CASOS ATÍPICOS Y EXTREMOS
- DISCONTINUIDADES / MULTIMODALIDAD

EDA

5 Aspectos a observar desde EDA Univariable
DISCONTINUIDADES / MULTIMODALIDAD

- Forma : Estadística confirmatoria: HISTOGRAMA DE FRECUENCIAS
  Estadística exploratoria: STEM and LEAF (Tallos y hojas)

STEM and LEAF: es una combinación de una distribución de frecuencias y un histograma. Es un gráfico espacialmente diseñado para detectar problemas de multimodalidad, es decir, discontinuidades y problemas en el centro de la distribución.

EDA

Valor	Frecuencia
0	1
2	1
3	1
6	1
8	1
10	1
11	1
16	1
23	1
24	1
25	1
28	1
30	1
31	4
32	1
33	3
34	4
35	7
36	3
37	4
38	4
39	9
40	12

```
0 * | 023
0 . | 68
1* | 01
1 . | 6
2 * | 34
2 . | 58
3* | 0111123334444
3 . | 555555566677778888999999999
4* | 000000000000
```


EDA

CONDENSADO	EXTENSO
0 02368	0 *0
1 016	0 t23
2 3458	0 f
3 011112333444455555556667777888899999999	0 s6
4 000000000000	0.8
	1 *01
	1 t
	1 f
	1 s6
	1 .
	2 *
	2 t3
	2 f45
	2 s
	2 .8
	3 *01111
	3 t2333
	3 f44445555555
	3 s6667777
	3 .8888999999999
	4 *000000000000

EDA

Valor	Frecuencia
0	1
2	1
3	1
6	1
8	1
10	1
11	1
16	1
23	1
24	1
25	1
28	1
30	1
31	4
32	1
33	3
34	4
35	7
36	3
37	4
38	4
39	9
40	12

Frequency		Stem & Leaf
8,00	Extremes	(0), (2), (3), (6), (8), (10), (11), (16)
2,00	2*	34
2,00	2 .	58
13,00	3*	0111123334444
27,00	3 .	555555566677778888999999999
12,00	4*	000000000000

Stem width: 10,00
Each leaf: 1 case(s)

EDA

Back to back stem and leaf display. The left side shows the 1998 TD data and the right side shows the 2000 TD data.

1998		2000
11	4	
	3	7
332	3	233
8865	2	889
44331110	2	001112223
987776665	1	56888899
321	1	22444
7	0	69