(#397). AJUSTE DE FUNCIONES (I): SPLINES VS MÍNIMOS CUADRADOS

[MONOTEMA] Tras explicar varios de los más empleados métodos de interpolación para buscar aproximarnos a la función que describe los datos empíricos, hemos visto que las splines cúbicas nos ofrecen mucha flexibiliad. Sin embargo, y aunque pueda parecer paradójico, el hecho de buscar una función de interpolación que pase por todos los nodos puede ser contraproducente, porque podemos estar dando demasiada importancia al ruido, cuando lo que buscamos es la señal. Es decir, cuando en los datos hay errores aleatorios (y también algún error sistemático puntual), buscar aproximaciones con funciones que no pasen necesariamente por esos nodos pero que tengan otras propiedades (como la minimización de una variable de error), puede ser más útil.

El método básico para realizar este ajuste es el de mínimos cuadrados, un potente método no paramétrico que minimiza la suma al cuadrado de los errores. Y es lo que vamos a ver en este post, aunque da manera muy simplificada ya que está sobradamente explicado en multitud de textos especializados de prácticamente todas las áreas científicas.

Así, compararemos gráficamente el ajuste por mínimos cuadrados con el método de las splines cúbicas, para estimular la reflexión de los alumnos.

Datos de partida

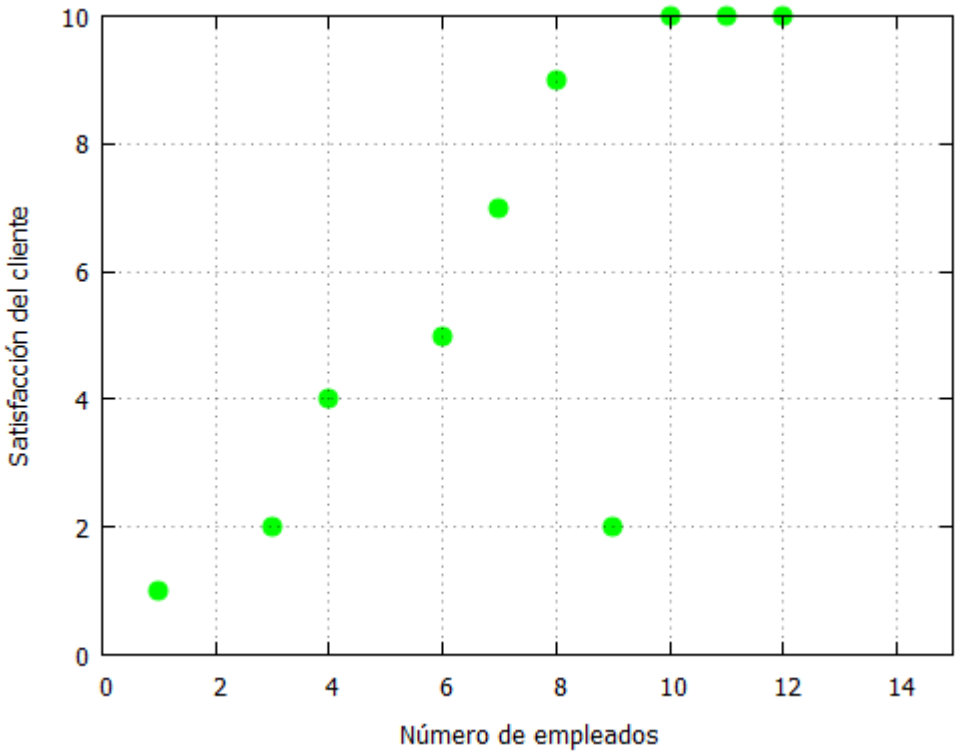
Usaremos los mismos datos que en el método de Lagrange, y en el de todos los ejemplos sobre interpolación:

x	f(x)

1	1
3	2
4	4
6	5
7	7
8	9
9	2
10	10
11	10
12	10

Estos datos relacionan la cantidad de empleados utilizados en un gran supermercado (x) con la saltisfacción del cliente $y = f(x)$. La satisfacción del cliente se mide en una escala de 0 a 10, donde 0 es el valor mínimo y 10 el valor máximo.

```
x_:[1,3,4,6,7,8,9,10,11,12];
fx_:[1,2,4,5,7,9,2,10,10,10];
plot2d([discrete, x_, fx_],
[x,0,15],[y,0,10], [style, points], [color,green],
[xlabel, “Número de empleados”],
[ylabel, “Satisfacción del cliente”], [legend, false]);
```



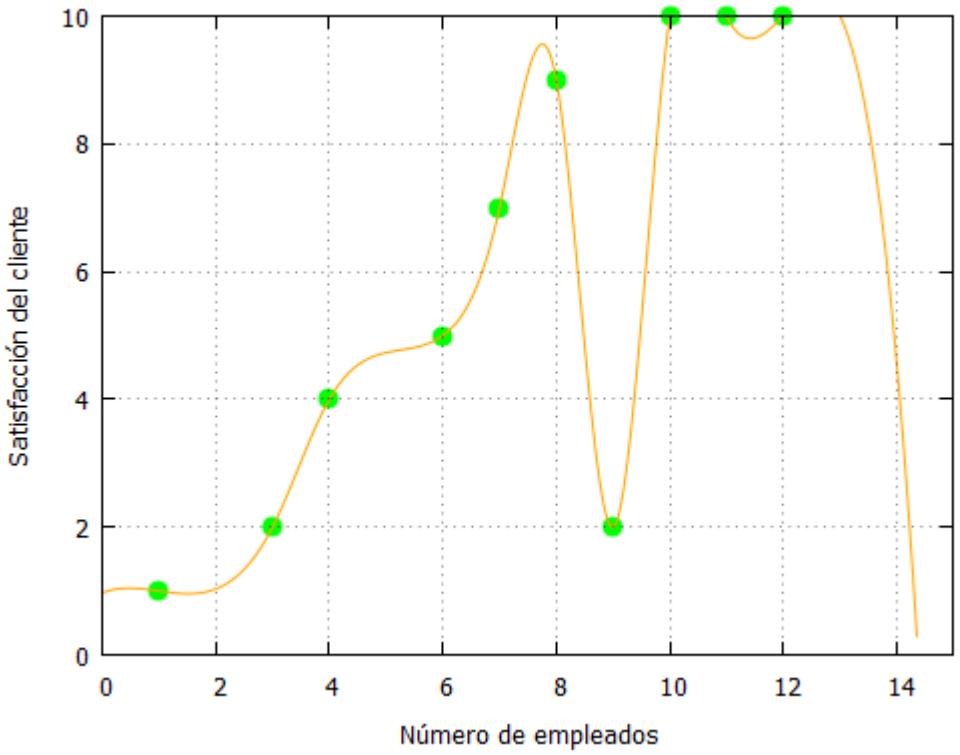
Splines cúbicos

Una vez que ya sabemos programar las splines paso a paso, podemos actuar de forma más directa con Maxima, pidiendo que use todos los datos disponibles:

```
load (interpol);
p:[[1,1],[3,2],[4,4],
[6,5],[7,7],[8,9],[9,2],[10,10],[11,10],[12,10]];
splines: cspline(p);
```

Y lo podemos graficar de la siguiente forma:

```
x_:[1,3,4,6,7,8,9,10,11,12];
fx:[1,2,4,5,7,9,2,10,10,10];
plot2d([[discrete, x_, fx], splines],
[x,0,15],[y,0,10], [style, points, lines], [color,green,
orange],
[xlabel, “Número de empleados”],
[ylabel, “Satisfacción del cliente”], [legend, false]);
```



Vemos que las splines hacen aparentemente un “buen trabajo” de ajuste, pero tenemos ese punto “problemático” [9,2] que hace que la curva de interpolación cambie drásticamente. Sin embargo, en este caso no nos afecta demasiado a nuestro objetivo de interpolar puntos que no están en los datos en el intervalo [1,12]. Tanto el nodo 2 como el nodo 5 se pueden evaluar con la curva de interpolación ya que las splines van de nodo a nodo, por lo que ese comportamiento “raro” en el punto [9,2] no afecta prácticamente a las evaluaciones entre los otros nodos.

Mínimos cuadrados lineales

Este método se basa en la minimización de la función de error:

$$E(\alpha, \beta) = \sum_{i=1}^{n} (f(x_i) - (\beta x_i + \alpha))^2$$

donde $y_i = f(x_i)$ se corresponde con las imágenes de los datos brutos x_i , y α, β son parámetros a estimar.

Minimizar esa función requiere que:

$$\frac{\partial E}{\partial \alpha} = 0$$

$$\frac{\partial E}{\partial \beta} = 0$$

Es importante señalar que se asume que el error tiene de media cero, es decir, es aleatorio (ruido blanco).

De este modo:

$$\alpha = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i x_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\beta = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Así, podemos programarlo con Maxima usando el siguiente código:

```
x:matrix([1,3,4,6,7,8,9,10,11,12]);
fx:matrix([1,2,4,5,7,9,2,10,10,10]);
x_traspuesta: transpose(x);
fx_traspuesta:transpose(fx);
n: length(x_traspuesta);
datos: zeromatrix(n,1);
datos_xcuad:x_traspuesta.x;
datos_y:fx_traspuesta;
datos_xy:fx_traspuesta.x;
datos_x: x_traspuesta;
print(datos_xcuad,datos_y, datos_xy,datos_x);
suma_xcuad:sum(datos_xcuad[i,i],i,1,n);
suma_y:sum(datos_y[i,1],i,1,n);
suma_xy:sum(datos_xy[i,i],i,1,n);
suma_x:sum(datos_x[i,1],i,1,n);
cuad_suma_x:suma_x^2;
alpha:((suma_xcuad*suma_y)-(suma_xy*suma_x))/((n*suma_xcuad)-cuad_suma_x),
numer;
beta:((n*suma_xy)-(suma_x*suma_y))/((n*suma_xcuad)-cuad_suma_x), numer;
fx_estimada:alpha+beta*x_;
```

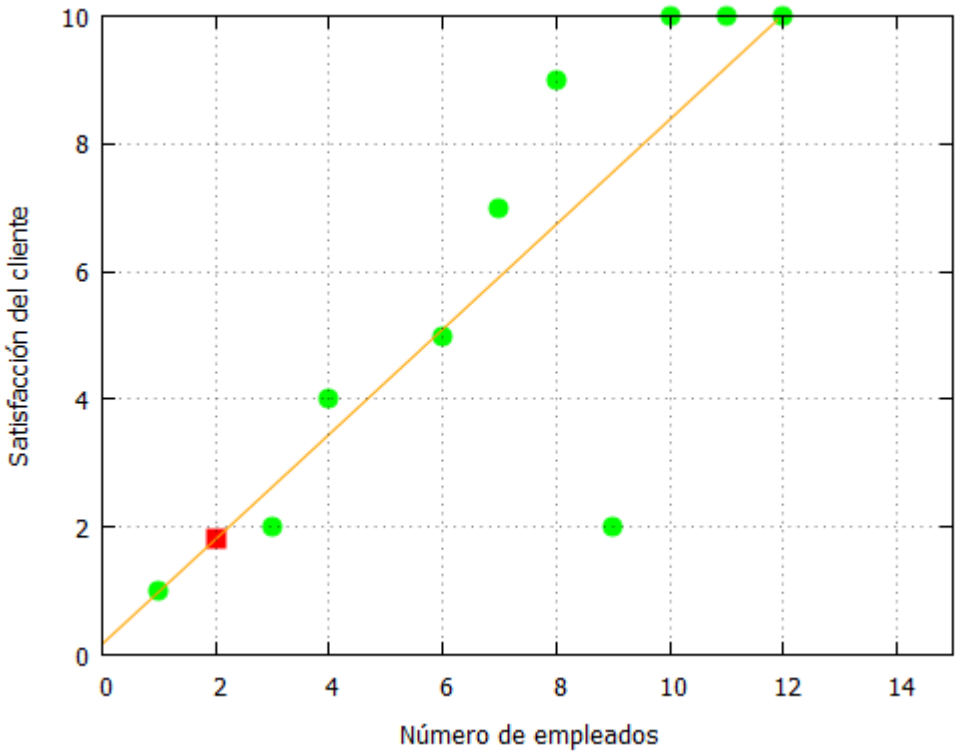
Lo que nos da la siguiente recta de ajuste (la estimación de

la función).

$$f(\hat{x}_i) = \alpha + \beta x_i = 0.1693755346449957 + 0.8212147134302823 x_i$$

Entonces podemos hacer la representación gráfica del ajuste de los datos:

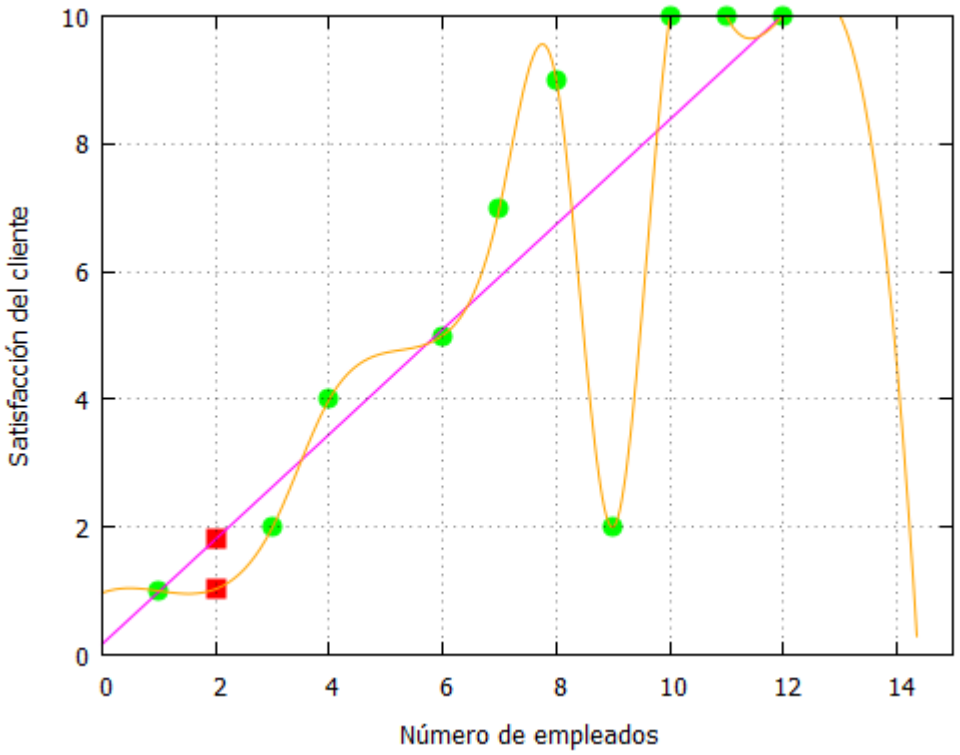
```
kill (all);
x_:[1,3,4,6,7,8,9,10,11,12];
fx:[1,2,4,5,7,9,2,10,10,10];
x_aproximar:2;
fx_estimada:0.8212147134302823*x+0.1693755346449957;
fx_estimada_x_aproximar: ev(fx_estimada, x=x_aproximar),
numer;
plot2d([[discrete, x_, fx], [discrete,
[[x_aproximar,fx_estimada_x_aproximar]]],fx_estimada],
[x,0,15],[y,0,10], [style, points, points, lines],
[color,green, red, orange],
[xlabel, “Número de empleados”],
[ylabel, “Satisfacción del cliente”], [legend, false]);
```



Para 2 empleados la función de ajuste estimada nos dice que la satisfacción será de 1.81.

En próximos posts hablaremos con más detalle de la evaluación de este ajuste, pero ahora simplemente los estudiantes pueden hacer un ejercicio de reflexión ante el visionado de la superposición de las dos funciones computadas, la primera con splines cúbicas y la segunda con el método de los mínimos cuadrados.

```
load (interpol);
p:[[1,1],[3,2],[4,4],
[6,5],[7,7],[8,9],[9,2],[10,10],[11,10],[12,10]];
splines: cspline(p);
x_:[1,3,4,6,7,8,9,10,11,12];
fx:[1,2,4,5,7,9,2,10,10,10];
x_aproximar:2;
x_interpolar:x_aproximar;
fx_estimada:0.8212147134302823*x+0.1693755346449957;
fx_estimada_x_aproximar: ev(fx_estimada, x=x_aproximar), numer;
f_splines: ev(splines,x=x_interpolar), numer;
plot2d([[discrete, x_, fx], [discrete,
[[x_aproximar,fx_estimada_x_aproximar],
[x_interpolar,f_splines]]],fx_estimada,splines],
[x,0,15],[y,0,10], [style, points, points, lines, lines],
[color, green, red, magenta, orange, magenta],
[xlabel, “Número de empleados”],
[ylabel, “Satisfacción del cliente”], [legend, false]);
```



Cuando el número de empleados es 2, la estimación del nivel de satisfacción difiere en función del método empleado. Además, hemos realizado el ajuste de mínimos cuadrados sin tener en cuenta que aparentemente hay una ruptura de la linealidad cuando el número de empleados se incrementa. En posteriores posts, analizaremos con más tranquilidad estos detalles.

Conclusión

Hemos comparado los resultados obtenidos con splines cúbicas con respecto al ajuste de mínimos cuadrados ordinarios, sin entrar en profundidades sobre la idoneidad de ambos análisis. Las predicciones que podemos hacer en base a los datos empíricos divergen, lo que nos debe advertir de lo prudentes que hemos de ser a la hora de analizar los datos.

Todos los posts relacionados

(#412). LOS JUGADORES MÁS PRODUCTIVOS DE LA NBA 2018/19

(#411). LOS ROOKIES MÁS PRODUCTIVOS DE LA NBA 2018/19

(#410) ¿QUIÉN HA SIDO EL MEJOR ANOTADOR DE LA FASE REGULAR DE LA EUROLIGA 2018/19?

(#409). ALGUNAS CONSIDERACIONES EN LA ESTIMACIÓN DEL RENDIMIENTO POR MINUTO EN BALONCESTO (II)

(#408). ALGUNAS CONSIDERACIONES EN LA ESTIMACIÓN DEL RENDIMIENTO POR MINUTO EN BALONCESTO (I)

(#407). AVANCE DE LOS NOVATOS MÁS PRODUCTIVOS DE LA NBA 2018/19

(#398). AJUSTE DE FUNCIONES (II): CAPACIDAD EXPLICATIVA Y PARTICIÓN DE LA FUNCIÓN

(#397). AJUSTE DE FUNCIONES (I): SPLINES VS MÍNIMOS CUADRADOS

(#396). INTERPOLACIÓN (IV): MÉTODO DE SPLINE CÚBICO

(#395). INTERPOLACIÓN (III): MÉTODO DE DIFERENCIAS DIVIDIDAS

(#394). INTERPOLACIÓN (II): MÉTODO DE NEVILLE

(#393). INTERPOLACIÓN (I): POLINOMIO DE LAGRANGE

(#391). BÚSQUEDA DE SOLUCIONES (V): MÉTODO DE LA POSICIÓN FALSA

(#390). BÚSQUEDA DE SOLUCIONES (IV): MÉTODO DE LA SECANTE

(#389). BÚSQUEDA DE SOLUCIONES (III): METODO DE NEWTON

(#388). BÚSQUEDA DE SOLUCIONES (II): METODO DEL PUNTO FIJO

(#387). BÚSQUEDA DE SOLUCIONES (I): MÉTODO DE LA BISECCIÓN

(#384). MEJOR UN SÓLO ITEM QUE VARIOS PARA MEDIR ACTITUDES

(#381). TABLAS DE TUMORES TOTALES EN EL ESTUDIO DEL NTP SOBRE EFECTOS DE LA RADIACIÓN DE TELÉFONO MÓVIL

(#380). ÍNDICES APROXIMADOS FLEXIBLES EN ECUACIONES ESTRUCTURALES

(377). AVANZANDO EN LA COMPLEJIDAD; DOBLE DEPENDENCIA

(#376). CAMPOS DIRECCIONALES. COMENZANDO A REPRESENTAR LA COMPLEJIDAD

(#373). PREDICCIONES CON BUENAS APROXIMACIONES

(#366). MODELO DE HOMOGENEIDAD EN TABLAS DE 2X2

(#365). BÁSICOS DE ECUACIONES ESTRUCTURALES (V): MATRIZ DE DATOS BRUTOS

(#362). ¿HORMESIS ENTRE EL VOLUMEN DE ENTRENAMIENTO Y LA HIPERTROFIA?

(#358). BÁSICOS DE ECUACIONES ESTRUCTURALES (IV): REGRESIÓN LINEAL SIMPLE CON FIABILIDADES DIVERSAS

(#354). BÁSICOS DE ECUACIONES ESTRUCTURALES (III): COVARIANZA Y CORRELACIÓN ENTRE VARIABLES LATENTES

(#353). BÁSICOS DE ECUACIONES ESTRUCTURALES (II): VARIABLES LATENTES Y FIABILIDAD

(#352). TESTOSTERONA Y CONSUMO CONSPICUO EN HOMBRES

(#350). BÁSICOS DE ECUACIONES ESTRUCTURALES (I): COVARIANZAS Y DESVIACIONES SOBRE LA MEDIA

(#346). MODELO NO LINEAL EN LAS VARIABLES PERO LINEAL EN LOS PARÁMETROS CON STATA Y MAXIMA

(#337). EXPLORANDO LA PREDICCIÓN LINEAL EN MARKETING CON MAXIMA

(#336) EXPLORANDO EL DETERMINISMO Y EL AZAR CON C Y MAXIMA

(#335). EXPLORANDO EL COMPORTAMIENTO CAÓTICO CON MAXIMA

(#334). EXPLORANDO LA NO LINEALIDAD CON MAXIMA

(#331). LOS CONSTRUCTOS FORMATIVOS NO REFLEJAN ESTADOS PSICOLÓGICOS

(#323). EL ATAQUE CON MICROONDAS A LA EMBAJADA DE EEUU EN MOSCÚ DURANTE LA GUERRA FRÍA

(#314). REDUCIR EL SESGO EN ESTUDIOS SOBRE EFECTIVIDAD DE VACUNAS

(#313). EL SISTEMA PLANETARIO DE TRAPPIST-1

(#310). MÁS ALLÁ DEL REDUCCIONISMO EN BIOLOGÍA

(#307). LA CRISIS DE LA INVESTIGACIÓN EN PSICOLOGÍA

(#302). CONSTRUCTOS EN PSICOLOGÍA E IMPLICACIONES ESTADÍSTICAS

(#301). CENSURA EN EL LENGUAJE CIENTÍFICO

(#298). ¿CAMBIAR DE ENTRENADOR A MITAD DE TEMPORADA?

(#294). NEUROLOGÍA DE LA MALDAD: PISTAS SOBRE LAS RAÍCES DE LA MALDAD INDIVIDUAL

(#287). MEDIAS VERDADES Y OCULTACIÓN: ACEPTACIÓN DEL ENGAÑO DE CORPORACIONES

(#283). INTERPRETACIÓN FRECUENTISTA DE LA PROBABILIDAD; EL PAPEL DE LOS MODELOS

(#276). LA IRRACIONALIDAD, SCHRÖDINGER, Y EL EFECTO TÚNEL

(#265). DEFENDIENDO EL P-VALOR A TRAVÉS DE SU CORRECTA INTERPRETACIÓN

(#264). MÁS ALLÁ DE LOS ENSAYOS CONTROLADOS ALEATORIZADOS

(#263). CAUSALIDAD EN CASOS INDIVIDUALES: ALGUNAS RECOMENDACIONES PARA SU ESTUDIO

(#260). PROPONEN SER 10 VECES MÁS EXIGENTES CON EL P-VALOR

(#257). ERRORES DE CODIFICACIÓN DE EMBOLIAS PULMONARES AL USAR EL ICD-10

(#252). LA PREGABALINA NO SE ASOCIA A MALFORMACIONES EN EL EMBARAZO

(#243). LOS DEPORTISTAS DE ÉLITE VIVEN MÁS, PERO CON MATICES

(#232). LA OTRA HISTORIA DE NIKE

(#191). LA TRIADA OSCURA, LOS ANTICONCEPTIVOS Y LA PREFERENCIA SEXUAL

(#185). VENTA DE ENTRADAS EN LA NBA Y POTENCIAL DE LOS EQUIPOS

(#163). CRECEN LAS DUDAS SOBRE PLS (PARTIAL LEAST SQUARES)

(#159). ENTENDER LA NATURALEZA DE LOS DATOS Y DE LOS TEST ESTADÍSTICOS

(#128). INDICADORES FORMATIVOS EN ECUACIONES ESTRUCTURALES

(#127). AJUSTE EXACTO EN ECUACIONES ESTRUCTURALES

(#104). CORRUPCIÓN Y DESIGUALDAD ENTRE LOS MUY RICOS

(#100). DESCREÍDOS DE LA EPIDEMIOLOGÍA

(#90). REDES EN PSICOPATOLOGÍA Y EN MARKETING

(#76). EL 85-90% DE LOS NUEVOS MEDICAMENTOS SON INNECESARIOS

(#33). CLUSTERS DE CÁNCER PEDIÁTRICO

(#31). TRÁFICO Y CÁNCER DE PULMÓN

(#26). FOTOVOLTAICAS HIPERGEOMÉTRICAS

(#24). ROMPER EL CRISTAL

(#22). COMPLEJIDAD DE LOS FENÓMENOS SOCIALES

(#20). DECISIONES HEURÍSTICAS

(#17). RADICALES DE LAS VACUNAS

(#13). 24755 CACAS EN EL PARQUE

(#12). ESTADÍSTICAS SIN ESTADÍSTICA

(#10). EL DESCUBRIMIENTO DEL ARGÓN

(#6). ¿VIVIMOS MÁS Y CON MEJOR SALUD?

(#5). ¿ES LA ECONOMÍA UNA CIENCIA?

(#3). HEZONJA Y LAS PREDICCIONES ESTADÍSTICAS

(#1) ¿CÓMO JUGARÁ MARIO HEZONJA SU PRÓXIMO PARTIDO?