Análisis de Datos Multivariados

Leticia Gracia Medrano.
lety@sigma.iimas.unam.mx

07 de agosto del 2014

La definición Estadística de Agresti y Franklin

Estadística es el arte y la ciencia de diseñar estudios y analizar los datos que esos estudios generan. Su fin último es traducir los datos en conocimiento y entendimiento del mundo que nos rodea. En resumen Estadística es el arte y la ciencia de aprender de los datos.

Concepto de medición y de variable

Para cuantificar o clasificar lo que percibimos de un fenómeno aleatorio necesitamos hacer mediciones u observaciones que nos ayudarán a investigar una o varias características de interés sobre el fenómeno.
Para un correcto manejo de nuestras mediciones, las observaciones deben ser registradas tomando en cuenta su *tipo*, para poder saber que operaciones aritméticas podemos hacer con ellas.
Como al medir un fenómeno aleatorio obtenemos diferentes registros llamaremos **variable** al conjunto de posibles resultados que podemos obtener.

Las variables se clasifican de acuerdo a la escala de medición de sus valores.

$$\text{Variables} = \begin{cases} \text{Categóricas} \begin{cases} \text{Ordinales} \\ \text{Nominales} \end{cases} \\ \text{Numéricas} \begin{cases} \text{Continuas} \\ \text{Discretas} \end{cases} \end{cases}$$

Variable Categórica

Cuando el registro de la medición es un elemento de una categoría.

- Ordinales

Cuando el registro de la medición se expresa en grados de intensidad que tienen un orden, pero no se puede determinar el incremento entre los grados.
Con variables de tipo ordinal podemos calcular: la moda, la mediana o los porcentiles de los datos.
Ejemplo: Grados de satisfacción en un servicio "Muy bueno", "Bueno", "Regular" y "Malo".

- Nominales

Cuando las categorías sólo se les da un nombre pero no tienen un orden entre ellas, deben ser mutuamente excluyentes (no hay un elemento que pertenezcan a dos o más categorías a la vez) y exhaustivas (todo elemento pertenece a una categoría). Podemos calcular la(s) moda(s) y la frecuencia de ocurrencia en cada una de las categorías.
Ejemplo: ¿Está de acuerdo con las obras de continuación del segundo piso del Periférico? "Sí" "No".

Variable Numérica

Cuando los registros son valores numéricos

- Discretas

son las variables que toman un número finito o numerable de valores.
Ejemplo: Número de hijos en un matrimonio, número de accidentes.

- Continuas

Toman cualquier valor numérico entero, fraccionario o irracional.
La precisión del registro dependerá del instrumento de medición.
Ejemplo: la estatura de una persona.

Calidad en los datos

Inspección visual. Para detectar si hay datos fuera de los rangos establecidos, conocer el máximo y mínimo de cada variable.
Verificar que las codificaciones sean consistentes en toda la base.
Cuidado con el número de dígitos a usar, puede perderse precisión o al revés desperdiciar espacio.
Tener control sobre los estándares de medición.
Distribución de frecuencias de las variables de mayor interés, ver distribución de la muestra.
Gráficas de dispersión. Identificar grupos u observaciones discrepantes.
Verificar métodos de recolección de los datos para detectar posibles fuentes de sesgo.
Observaciones faltantes. Tratar de rastrearlas, ir a registros originales, razones de su omisión. Definir que se hará con estas observaciones, si se puede usar algún valor de reemplazo o imputación: o seleccionar cuáles si se desechan. Los valores faltantes generan sesgo este tema es de suma importancia
Un grupo de datos de poca calidad no merece un análisis muy

Observaciones Discrepantes

Estas observaciones también son conocidas como aberrantes, discordantes, contaminantes,sorprendentes, en inglés OUTLIER. Puede definírseles de varias formas, una de ellas es decir que es una observación que se encuentra a una distancia ANORMAL de las demás, y entonces hay que definir lo que es una distancia NORMAL, es decir la observación se encuentra fuera de la nube de datos.
Estas observaciones pueden distorsionar la información, también pueden ser una señal de que el modelo de distribución de los datos NO es el adecuado, o reflejar el haber encontrado una situación sorprendente o peculiar. Si la observación causa un impacto en el observador se le llama generalmente **discrepante**.

Una observación **contaminante** será cualquiera que no corresponda a la distribución supuesta, y el problema es que ésta puede no ser percibida por el observador.
Estas observaciones afectan fuertemente al estimador $\bar{X}$ de la media μ, y consecuentemente a los estimadores de $Var(X)$, de las de $Cov(X, Y)$ y de $Corr(X, Y)$.
En análisis de regresión interesa identificar a las observaciones **influyentes,** que son aquellas que al omitirlas del análisis los valores de las $\hat{\beta}$'s varían mucho.
Detectar estas observaciones puede ser una tarea bastante complicada, sobre todo cuando se tienen datos altamente multivariados.
En el caso univariado se les puede detectar muy fácilmente a través de gráficos boxplot o también al verificar si la media de los datos difiere mucho de la mediana.

Datos Faltantes

Datos faltantes completamente al azar

Pueden ser muy variadas las razones por las que existan valores faltantes. Ya sea porque las condiciones climáticas, de seguridad o políticas no permiten recoger la información, porque ese día los instrumentos se descomponen, por que no se encontró a la persona u objeto de la encuesta, aquí se puede pensar que la información se perdió **completamente** al azar (MCAR por su siglas en inglés). Es decir cuando la probabilidad de que X_i sea **no observada** no está relacionada con el valor mismo de x_i o con el de cualquier otra variable.

Por ejemplo si las personas con un nivel de ingresos alto tienden a no contestar por miedo a ser sujetos "secuestrables", entonces esa observación no se perdió completamente al azar.

MCAR corresponde a pensar que ese dato se perdió con la misma probabilidad que cualquier otro dato. Si la persona no responde acerca de sus ingresos, de la misma manera que no responde a cuántos hijos tiene, entonces se considera MCAR. En este caso los parámetros pueden estimarse sin sesgo, aunque al perder

Datos faltantes al azar

A diferencia de los datos MCAR, donde la probabilidad de no observar a X_i no depende del valor mismo de x_i o de otras variables. En este caso esa probabilidad no dependerá de x_i **luego de controlar o condicionar con otra variable.**
Por ejemplo, una persona con depresión puede ser que tienda más a no contestar acerca de su ingreso, la gente con depresión a su vez en general tiene menos ingresos, entonces lo que ocurre es que si hay un tasa alta de no respuesta entre las personas con depresión, la media real puede ser menor que la calculada con los datos existentes, es decir sin tomar en cuenta a los datos faltantes. Ahora si entre las personas con depresión la probabilidad de no contestar acerca de su ingreso no está relacionada con su nivel de ingreso, entonces los datos se consideran faltantes al azar, (MAR). **Esto No significa que estos faltantes no produzcan sesgo y que se pueda uno olvidar del problema.**

Datos Faltantes no al azar

Cuando no son MCAR ni MAR entonces se dice que son datos faltantes no al azar (MNAR).
Ejemplo: Si se estudia una cierta enfermedad y las persona que padecen esa enfermedad son las que tienen una mayor probabilidad a no contestar a si la padecen, entonces los datos son faltantes no al azar, MNAR. Claramente el estimador de la proporción que padece esa enfermedad será menor que la proporción que se obtendría con los datos completos. Lo mismo ocurre en el caso de las personas con menor ingreso son las que tienden a no contestar su nivel de ingreso. Esta falta de datos no al azar es un problema, la única manera de obtener un estimador insesgado
Referencia bibliográfica:
http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/M

Tratamiento de datos faltantes

Omisión total

Si los datos son MCAR las estimaciones obtenidas serán insesgadas si no son MCAR serán sesgadas, hay que tener en cuenta que esta pérdida de datos genera pérdida de potencia en las pruebas.

Omisión parcial

Por ejemplo en el cálculo de las correlaciones se usan las observaciones disponibles, pero entonces cada estimación está soportada por diferentes bases de datos. Puede ser el caso que se llegue a una matriz de correlaciones estimada NO definida positiva.
No hay que olvidar que hay que analizar a las observaciones NA y tratar de ver si se comportan (en ciertas variables) como la población total o si difieren.
Otra cosa importante es considerar qué es lo que se tiene perdido. La situación de perder variables explicativas es diferente a perder variables respuesta.

Sustitución

Hot Deck

sustituir el caso por alguno semejante (de dónde sacamos a alguien semejante si ya acabó la encuesta, tener la providencia de guardar un montoncito extra para la sustitución?).

Imputación Simple

- Sustituir los valores faltantes por la media (el estimador de máxima verosimilitud), pero eso tiene consecuencias sobre la estimación de la varianza, porque siempre estaremos sustituyendo con el mismo valor.
- se puede sustituir usando una regresión, pero el problema sigue siendo que se sustituye por una media (esta vez condicionada) SPSS permite sumar una variación aleatoria, se subsana en algo este tipo de problema.
- Se puede usar el Algoritmo EM. En regresión si se conocieran los NA, estimar los parámetros del modelo sería fácil, y si se conocieran los parámetros del modelo de los datos sería sencillo hacer predicciones insesgadas de las observaciones faltantes. Este algoritmo es iterativo y va haciendo ambas

Imputación múltiple

Se generan valores para hacer la imputación basados en los datos existentes. Suponiendo que se estima$y$ usando x, pero esta imputación se hace varias veces, es decir tendremos varios conjuntos de datos completados. Para hacer esto se usan métodos conocidos Markov Chain Monte Carlo.
El programa NORM en la parte llamada *data augmentation* lo hace.
SAS tiene dos procedimientos MI y MIANALYZE.
Schafer, J.L. & Olsden, M. K.. (1998). *Multiple imputation for multivariate missing-data problems:*
A data analyst's perspective. Multivariate Behavioral Research, 33, 545-571.
En R esta el paquete MICE, material con referencia en: Van Buuren, S., Groothuis-Oudshoorn, K. (2011) MICE:
Multivariate Imputation by Chained Equations in R. Journal of Statistical Software.
http://www.stefvanbuuren.nl/publications/MICEinR-Draft.pdf

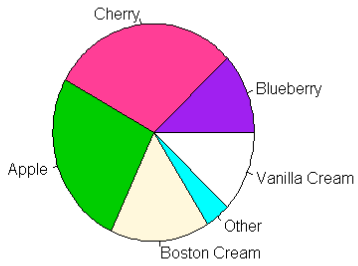
Gráficas datos univariados

- gráfica de barras y de *pie* son solo para datos categóricos, debe haber espacios entre las barras.
- histograma debe tenerse cuidado con los anchos de barras y con los puntos que se consideran en el eje de las x.
- boxplot permite rápidamente ver observaciones discrepantes.
- q-qplot permite ver si dos muestras provienen de la misma distribución.
- tallo y hoja, una versión de los histogramas pero permite ver los datos tal cual.

Gráficos de *Pie* y *Dot Chart*

El uso de gráficos circulares o pasteles es bastante común entre personas no profesionales en estadıstica y lamentablemente se ha trivializado tanto que si en muchas de las situaciones donde se usan se suprimieran se ahorrarıan muchas hojas de papel.
Los gráficos de puntos son elegantemente simples y permite numerosas variaciones. La única razón por la cual no se han vuelto populares es que los programas de hojas electrónicas no los elaboren presionando una tecla.

```
> pie(pie.sales) # default colours
> pie(pie.sales, col = c("purple", "violetred1", "green3",
+                        "cornsilk", "cyan", "white"))
```




```
> dotchart(pie.sales)
```

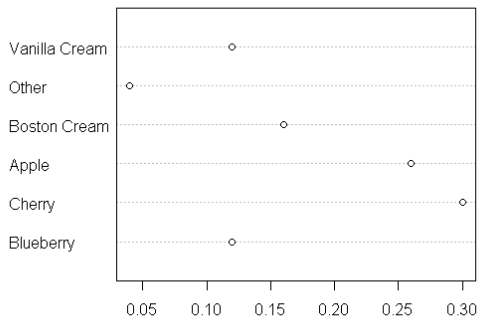
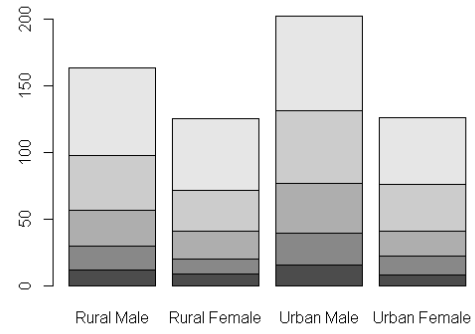


Gráfico de Barras

```
> barplot(VADeaths)
```




```
> barplot(VADeaths, beside = TRUE,
+         col = c("lightblue", "mistyrose", "lightcyan",
+                 "lavender", "cornsilk"),
+         legend = rownames(VADeaths), ylim = c(0, 100))
> title(main = "Death Rates in Virginia", font.main = 4)
```

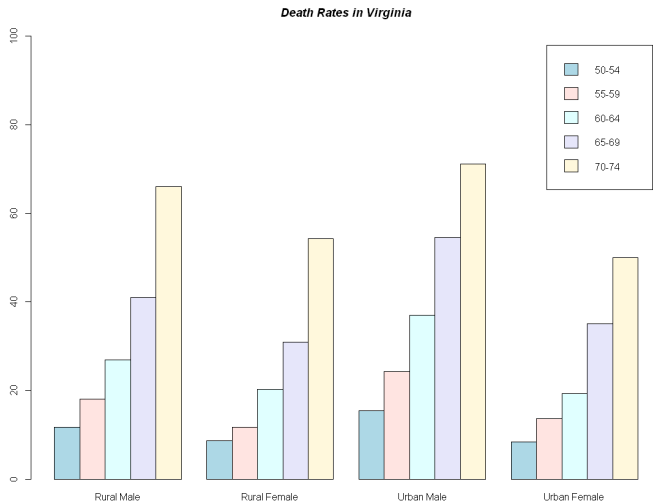


Gráfico de Tallo y Hoja

Este gráfico fue propuesto por Tukey (1977) y a pesar de no ser un gráfico para presentación definitiva se utiliza a la vez que el analista recoge la información ve la distribución de los mismos. Estos gráficos son fáciles ***de realizar a mano*** y se usan como una forma rápida y no pulida de mirar los datos. Qué nos muestra?

1. El centro de la distribución
2. La forma general de la distribución

Simétrica si las porciones a cada lado del centro son imágenes espejos de las otras.

Sesgada a la izquierda Si la cola izquierda (los valores menores) es mucho más larga que los de la derecha (los valores mayores)

Sesgada a la derecha opuesto a la sesgada a la izquierda.

3. Desviaciones marcadas de la forma global de la distribución.

Outliers Observaciones individuales que caen muy por fuera del patrón general de los datos.

gaps Huecos en la distribución

```
> stem(islands)
The decimal point is 3 digit(s) to the right of the |

   0 | 000000000000000000000000000000111111222338
   2 | 07
   4 | 5
   6 | 8
   8 | 4
  10 | 5
  12 |
  14 |
  16 | 0
```

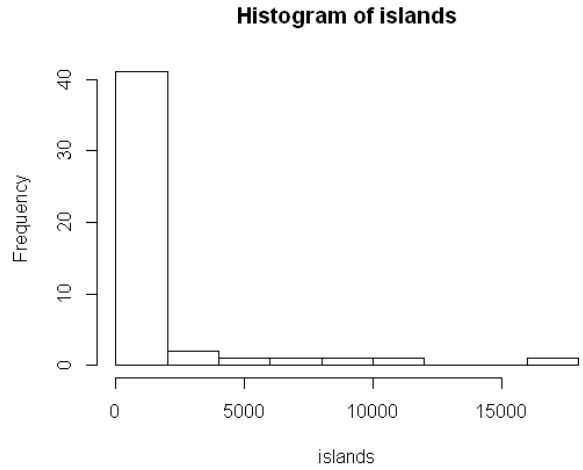
```
> stem(log10(islands))

  The decimal point is at the |

  1 | 1111112222233444
  1 | 5555556666667899999
  2 | 3344
  2 | 59
  3 |
  3 | 5678
  4 | 012
```

```
> as.data.frame(islands)
                islands
Africa            11506
Antarctica         5500
Asia              16988
Australia          2968
Axel Heiberg         16
Baffin              184
Banks                23
Borneo              280
Britain              84
Celebes              73
Celon                25
Cuba                 43
Devon                21
Ellesmere            82
Europe             3745
Greenland           840
Hainan               13
Hispaniola           30
Hokkaido             30
Honshu               89
Iceland              40
Ireland              33
Java                 49
Kyushu               14
Luzon                42
Madagascar          227
Melville             16
Mindanao             36
Moluccas             29
New Britain          15
New Guinea          306
New Zealand (N)      44
New Zealand (S)      58
Newfoundland         43
North America      9390
Novaya Zemlya        32
Prince of Wales      13
Sakhalin             29
South America      6795
Southampton          16
Spitsbergen          15
Sumatra             183
Taiwan               14
Tasmania             26
Tierra del Fuego     19
Timor                13
Vancouver            12
Victoria             82
```

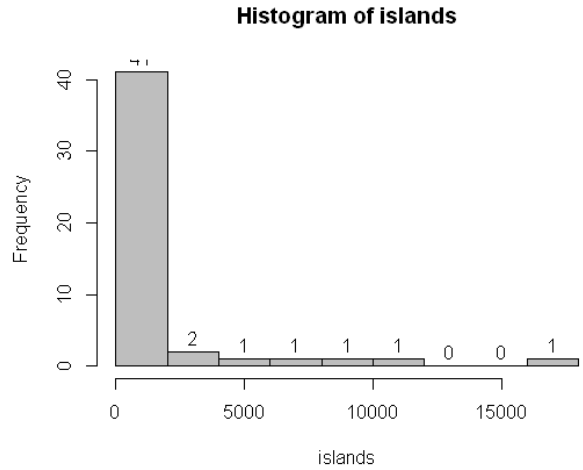
Histograma

El histograma es el gráfico estadístico por excelencia. El histograma de un conjunto de datos es un gráfico de barras que representan las frecuencias con que aparecen las mediciones agrupadas en ciertos rangos o intervalos. Para uno construir un histograma se debe dividir la recta real en intervalos o clases (algunos recomiendan que sean de igual longitud) y luego contar cuantas observaciones caen en cada intervalo.
formula de Sturges para determinar el numero de barras.
Regla de Sturges: $k = 1 + log_2(n)$
Scott (1992), basado en la distribuci´on normal recomienda el siguiente número de barras para el histograma Regla de Scott:$k = (2n)^{1/3}$

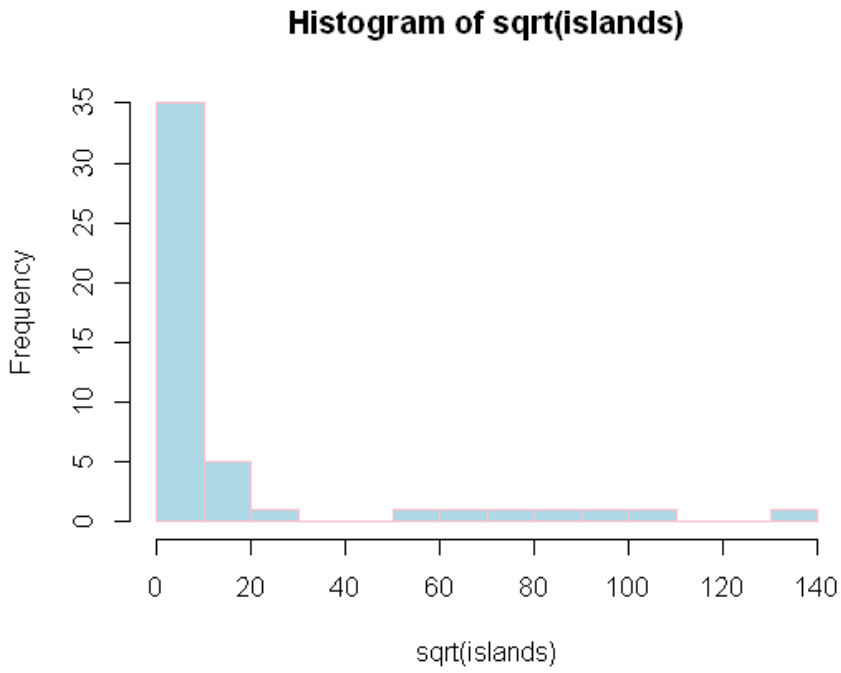
```
> hist(islands)
```




```
> utils::str(hist(islands, col="gray", labels = TRUE))
List of 7
 $ breaks     : num [1:10] 0 2000 4000 6000 8000 10000 12000 14000
16000 18000
 $ counts     : int [1:9] 41 2 1 1 1 1 0 0 1
 $ intensities: num [1:9] 4.27e-04 2.08e-05 1.04e-05 1.04e-05 1.04e-05
...
 $ density    : num [1:9] 4.27e-04 2.08e-05 1.04e-05 1.04e-05 1.04e-05
...
 $ mids       : num [1:9] 1000 3000 5000 7000 9000 11000 13000 15000
17000
 $ xname      : chr "islands"
 $ equidist   : logi TRUE
 - attr(*, "class")= chr "histogram"
```



```
> hist(sqrt(islands), breaks = 12, col="lightblue", border="pink")
```

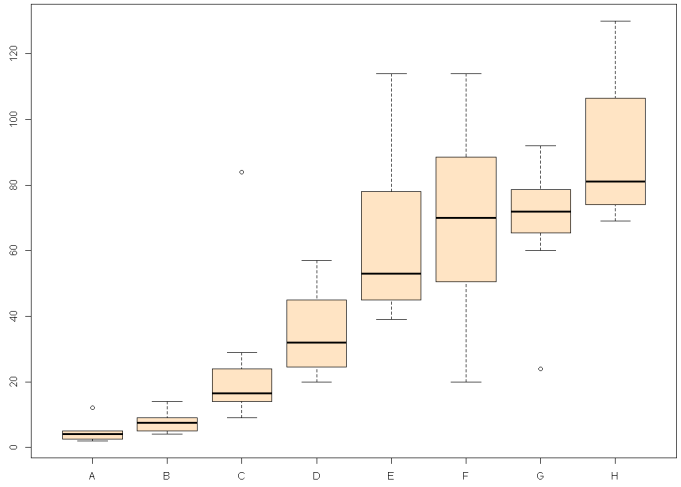


Boxplot o Caja de Tukey

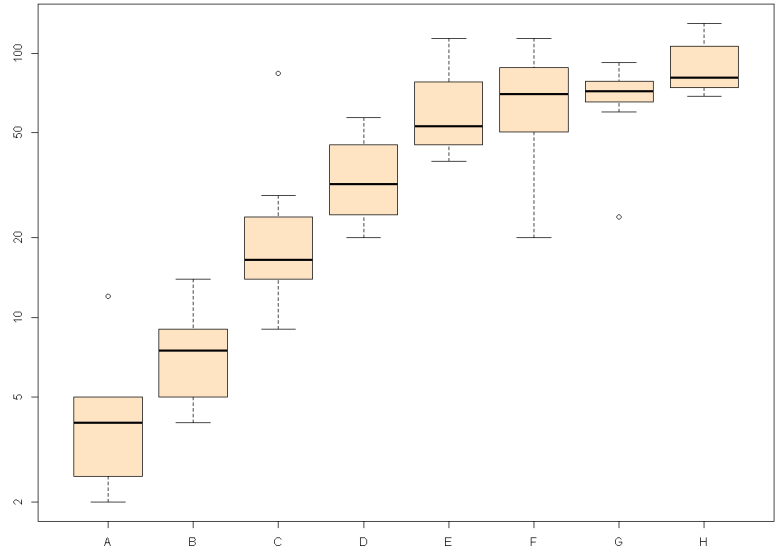
Realizado por Tukey (1977). Es un gráfico simple, ya que se realiza básicamente con cinco números.
Permite comparar diversos conjuntos de datos simultáneamente.
Este gráfico contiene un rectángulo, usualmente orientado con el sistema de coordenadas tal que el eje vertical tiene la misma escala del conjunto de datos. La parte superior y la inferior del rectángulo coinciden con el tercer cuartil y el primer cuartil de los datos. Esta caja se divide con una linea horizontal a nivel de la mediana. Se define un "paso" como 1.5 veces el rango intercuartil, y una linea vertical (un bigote) se extiende desde la mitad de la parte superior de la caja hasta la mayor observación de los datos si se encuentran dentro de un paso. Igual se hace en la parte inferior de la caja Las observaciones que caigan más allá de estas líneas son dibujadas individualmente. La definición de los cuartiles puede variar y otras definiciones de el paso son planteadas por otros autores.

La localización esta representada en la línea que corta la caja y representa la mediana (que esta dentro de la caja), la dispersión esta dada por la altura de la caja, como por la distancia entre los extremos de los bigotes. El sesgo se observa en la desviación que exista entre la linea de la mediana con relación al centro de la caja, y también la relación entre las longitudes de los bigotes. Las colas se pueden apreciar por la longitud de los bigotes con relación a la altura de la caja, y también por las observaciones que se marcan explícitamente.

```
> boxplot(decrease ~ treatment, data = OrchardSprays, col = "bisque")
```




```
> boxplot(decrease ~ treatment, data = OrchardSprays,
+         log = "y", col = "bisque")
```



¿Qué es un cuantil?

Son puntos tomados a intervalos regulares de la función acumulativa de distribución de una variable aleatoria. Dividir al conjunto de los datos ordenados en **q** conjuntos del mismo tamaño, es el objetivo de los q-cuantiles. Los cuantiles son las fronteras entre los conjuntos.

Cuantiles más comunes

El 2-cuantil, parte en dos partes iguales y es la mediana.

Los 3-cuantiles o terciles,

Los 4-cuantiles o cuartiles,

los 10-cuantiles o deciles,

los 100-cuantiles o porcentiles.

El k-ésimo q cuantil satisface lo siguiente:

$Pr(X < x) \leq k/q$. y $Pr(X \leq x) \geq k/q$

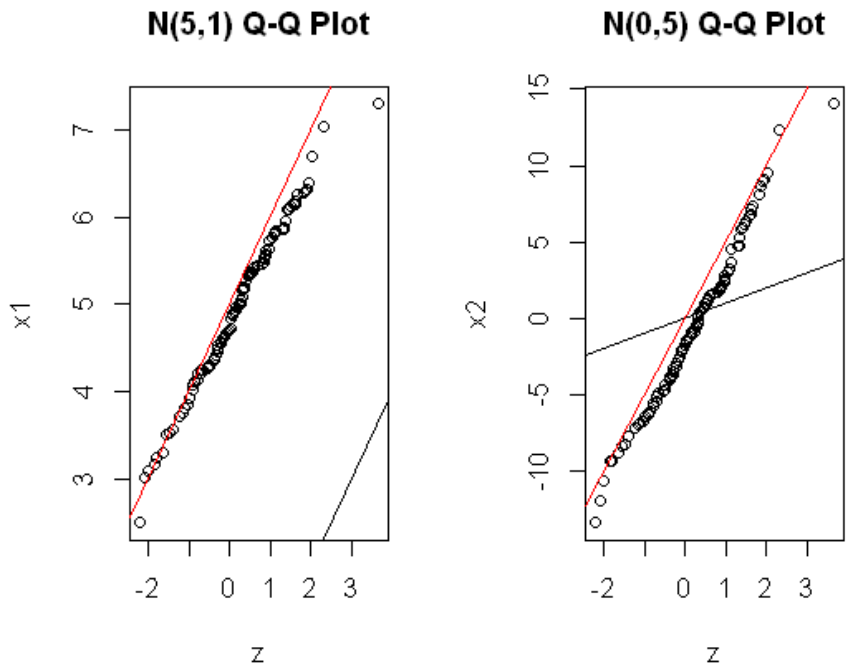
Para un conjunto tamaño N

puede calcularse como $I_p = N * (k/q)$, si es un entero se elige la observación que ocupe esa posición ordenada y ¿si no es un entero???, se redondea, o se toma una cierta interpolación entre las dos observaciones.

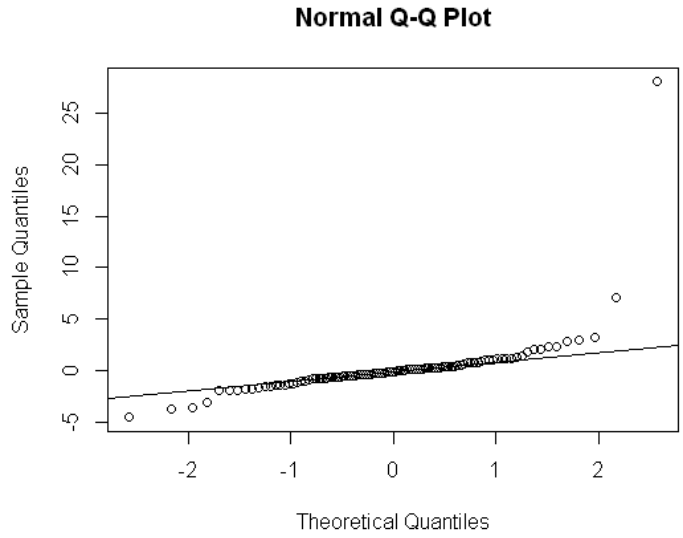
QQplot

Sirve para determinar si dos conjuntos de datos provienen de poblaciones con la misma distribución.
Se grafican los cuantiles del primer conjunto contra los cuantiles del segundo conjunto. Se dibuja también una recta de 45 grados de pendiente(es decir $y = x$). Si las observaciones provienen de la misma distribución, caerán aproximadamente sobre la recta. Entre más se separan de la recta, más alejadas serán sus distribuciones. Si caen sobre una recta con pendiente de 45 grados pero con distinta ordenada al origen, tendrán un traslado en el parámetro de localización, si varía la pendiente variará en la desviación estándar.
Los conjuntos pueden ser de distinto tamaño(se hacen corresponder los cuantiles del conjunto más grande con los valores ordenados del más pequeño, y los cuantiles intermedios se interpolan).
Una **gráfica de probabilidad** es semejante a una qqplot solo que se sustituyen al segundo conjunto de datos por los cuantiles de la distribución teórica a probar.

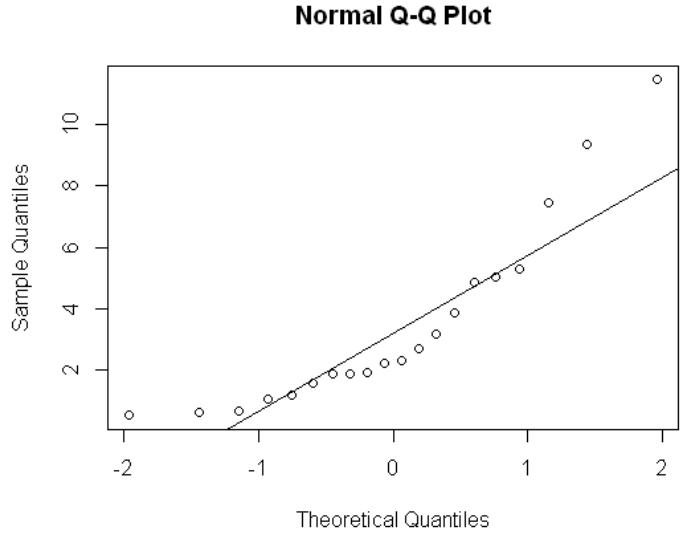
```
> x1<-rnorm(100,5,1)
> z<-rnorm(100)
> x2<-rnorm(100,0,5)
> z<-rnorm(100)
> x2<-rnorm(110,0,5)
> par(mfrow=c(1,2))
> qqplot(z,x1,main="N(5,1) Q-Q Plot")### variando la media
> abline(0,1)
> abline(5,1,col=2)
> qqplot(z,x2,main="N(0,5) Q-Q Plot")#### variando la desviacion
estandar
> abline(0,1)
> abline(0,5,col=2)
```



```
x <- rt(100, df=3)

# normal fit
qqnorm(x); qqline(x)
```



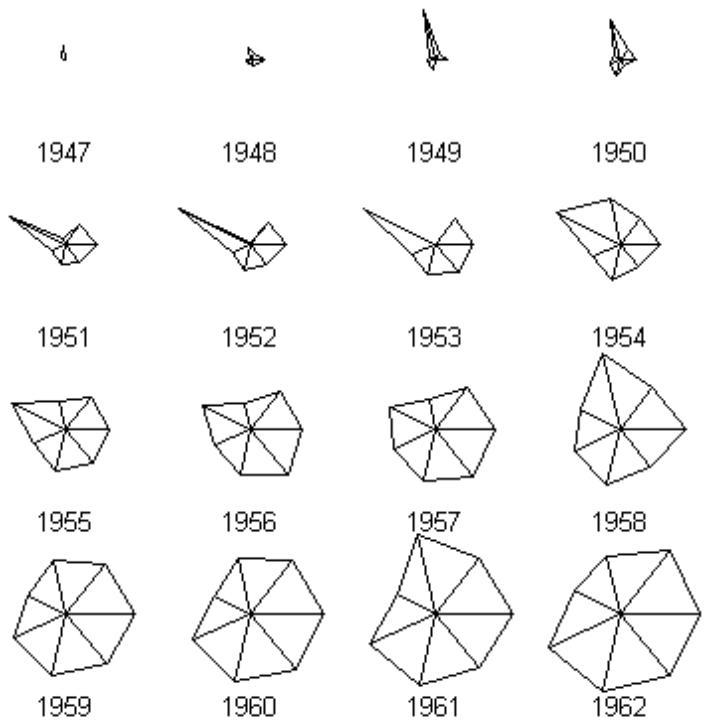

```
> x<-rchisq(20,3)
> qqnorm(x); qqline(x)
```



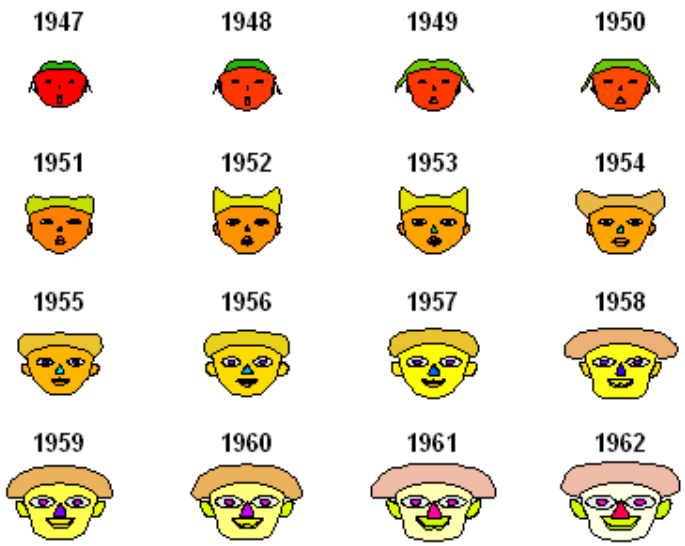
Gráficas datos multivariados

- Estrellas. Convienen cuando no se tienen muchos atributos, pues con más de 10 o 12 aristas las confundimos en su forma.
- Caritas, debidas a Chernov, dado que el ojo humano esta muy entrenado para reconocer rostros humanos. A cada elemento de la cara: pelo, ancho cara, largo nariz, tamaño de ojos se le asocia una característica.

```
> stars(longley)
```



```
> faces(longley)
effect of variables:
 modified item      Var
 "height of face   " "GNP.deflator"
 "width of face    " "GNP"
 "structure of face" "Unemployed"
 "height of mouth  " "Armed.Forces"
 "width of mouth   " "Population"
 "smiling          " "Year"
 "height of eyes   " "Employed"
 "width of eyes    " "GNP.deflator"
 "height of hair   " "GNP"
 "width of hair    " "Unemployed"
 "style of hair    " "Armed.Forces"
 "height of nose   " "Population"
 "width of nose    " "Year"
 "width of ear     " "Employed"
 "height of ear    " "GNP.deflator"
```



Curvas de Andrews

A cada individuo se le asigna una curva de la siguiente manera: $t \in [-\pi, \pi]$ Si p es impar

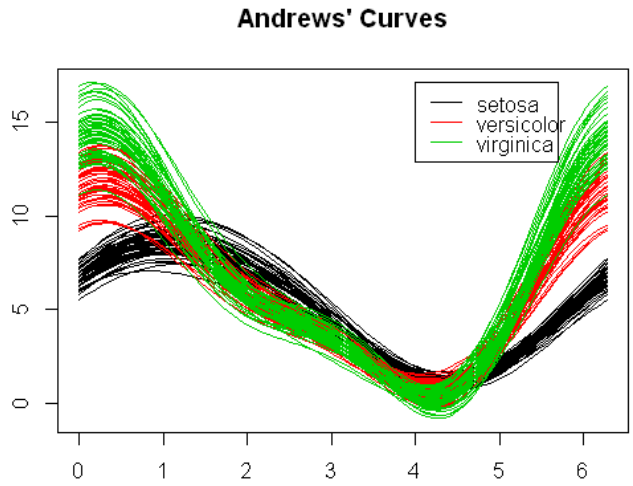
$$f_i(t) = \frac{X_{i1}}{\sqrt{2}} + X_{i2}\sin(t) + X_{i3}\cos(t) + \ldots + X_{ip}\cos(\frac{(p-1)}{2}t)$$

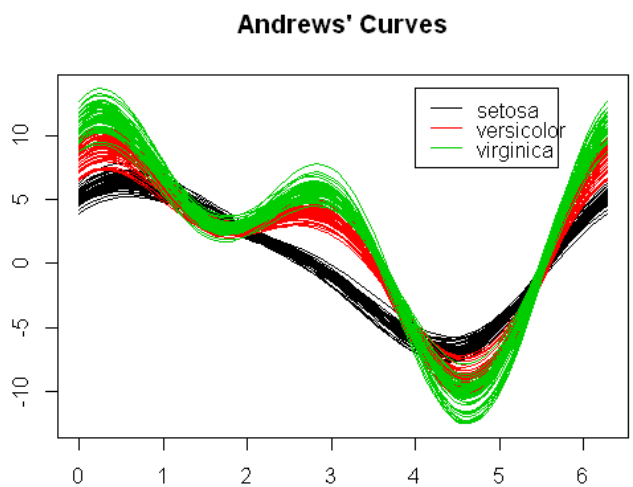
Si p es par

$$f_i(t) = \frac{X_{i1}}{\sqrt{2}} + X_{i2}\sin(t) + X_{i3}\cos(t) + \ldots + X_{ip}\sin(\frac{p}{2}t)$$

Estas tres gráficas no son únicas, pues según ordenemos las variables darán origen a estrellas, curvas o caras distintas.

```
andrews.curves(iris[,c(4,2,1,3)], iris[,5], title="Iris Data")
```

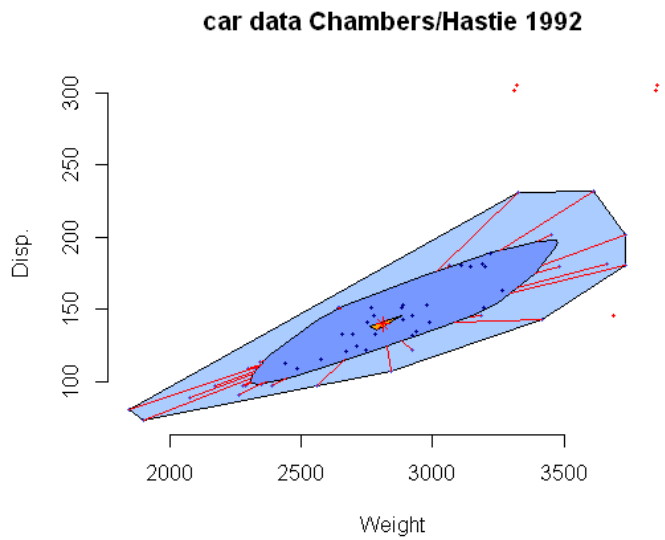




Bagplot

Parecida a un *boxplot* pero en dos dimensiones.

```
> cardata
      Weight Disp.
 [1,]   2560    97
 [2,]   2345   114
 [3,]   1845    81
 [4,]   2260    91
 [5,]   2440   113
 [6,]   2285    97
 [7,]   2275    97
 [8,]   2350    98
 [9,]   2295   109
[10,]   1900    73
...
[59,]   3185   146
[60,]   3690   146
> bagplot(cardata,factor=3,show.baghull=TRUE,
+         show.loophull=TRUE,precision=1,dkmethod=2)
> title("car data Chambers/Hastie 1992")
```



Gráfica de paralelas

Se usan sobre todo cuando hay varias mediciones para un solo individuo.

```
parallel(~iris[,1:4],col=as.numeric(iris$Species),main="Parallelplot IRIS")
```

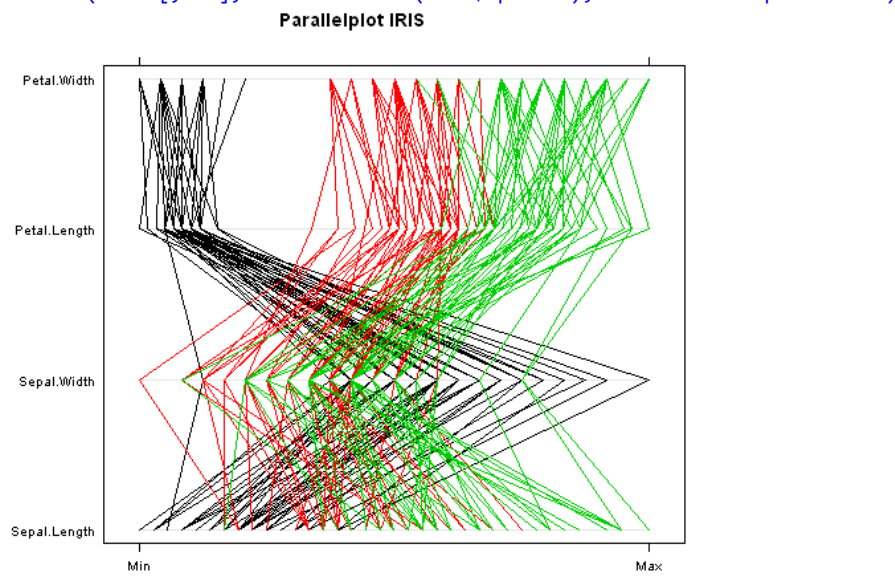


Gráfico series de tiempo múltiples

```
> USeconomic
        log(M1) log(GNP)          rs          rl
1954 Q1 6.111246 7.249073 0.010800000 0.02613333
1954 Q2 6.115892 7.245084 0.008133333 0.02523333
1954 Q3 6.129268 7.257003 0.008700000 0.02490000
1954 Q4 6.141177 7.271565 0.010366667 0.02566667
1955 Q1 6.151881 7.292746 0.012600000 0.02746667
1955 Q2 6.159307 7.303641 0.015133333 0.02816667
1955 Q3 6.162472 7.316880 0.018633333 0.02926667
1955 Q4 6.161840 7.325610 0.023466667 0.02890000
1956 Q1 6.164157 7.323633 0.023800000 0.02886667
...
1987 Q1 6.448731 8.236606 0.055333333 0.07636667
1987 Q2 6.453310 8.248791 0.057333333 0.08576667
1987 Q3 6.445879 8.259795 0.060333333 0.09083333
1987 Q4 6.446513 8.274612 0.060033333 0.09240000
```