Repaso de conceptos estadísticos

Sobre las herramientas de software

En primer lugar vamos ha señalar que aún cuando nos parezca como la alternativa más fácil para desarrollar las actividades de esta parte del trabajo, no utilizaremos excel.
Ciertamente una gran cantidad de personas usan esta herramienta a diario, pero lo cierto es que también son muchas las oportunidades en que encontramos errores o incluso no tenemos claro como excel maneja los cálculos estadísticos. Excel es muy bueno para el trabajo de prototipado, pero si tengo que enfrentarme a colecciones de datos muy grandes, la facilidad de uso que tiene la hoja de cálculo se vuelve en contra, respecto a lo que se conoce como error de punto flotante y estabilidad y repetitividad de resultados.

Una buena recomendación es trabajar con add-ons del excel, tales como Crysta Ball de Oracle o XLStat. Otras buenas herramientas podrían ser productos como Minitab, que presenta los datos como si fuesen hojas de cálculo, pero está adaptado y diseñado concretamente para el trabajo estadístico.

Finalmente si su trabajo está relacionado con una inversión muy importante, tal como una obra pública, la infraestructura de un puerto o un hospital, o en el caso de tesis de maestría o doctorado; mi recomendación de inclina por el programa R-Cran. En estos momentos la mayor parte de los trabajos de investigación en el área de estadísticas a nivel mundial y los avances expuestos en los congresos están desarrollados bajo esta plataforma. Como si eso no fuese importante, está bajo licencia GNU y existe un ejercito de Maestrandos y Doctorandos en estadística que aseguran la calidad de los métodos y resultados obtenidos por ellos.

Además les sugiero que como soporte para el curso recurran a lo materiales de estadísticas que han utilizado en la carrera de grado o los cursos que pudiesen haber tomado, Nada como realizar esta parte del aprendizaje usando el lenguaje y conocimiento que ya está adquirido.

Ejercicio 1

Tenemos la necesidad de desarrollar una simulación con un utilitario para establecer el porcentaje de consorcios que han tenido éxito en la disminución de riesgo de quebranto en Argentinna gracias a la generación y adopción de iniciativas de innovación. Se utilizará una escala de 0 a 10 para establecer el impacto o éxito alcanzado . 10 Representa éxito total, 0 representa fracaso. Basados en datos estadísticos del CLADES (Centro Latinoamericano Paa el Desarrollo de Estudios Sociales de Naciones Unidas), se han tomado 250 PYMES (casos testigos) de Colombia y Brasil al azar de un universo de casi dos mil de empresas; en su mayoría se trata de PyMES comerciales y un pequeño sector de ONGs. Según esta muestra se observa que para Brasil el promedio de éxito es 6,5 y el desvío estandard es 3,5 en tanto que para la Colombia esta cifra alcanza una media de 4,5 y sigma 3,2
Represente estos datos con R-Cran y formule algunas conjeturas que expliquen por que esto es así.

R-cran code

```
x=seq(0,10,length=250)
y=dnorm(x,mean=6.5,sd=3.5)
z=dnorm(x,mean=4.5,sd3,2)
plot(x,y,type="l",lwd=2,col="red")
hist(y,main="Histograma de Datos Observados")
plot (density(y) , main="Densidad de datos secuencia")

par(mfrow=c(2,2))
plot(y,x,type="l",lwd=2,col="red")
plot(z,x,type="l",lwd=2,col="blue")
hist(y, main="Histograma Brasil")
boxplot(z, main="Boxplot de Colombia")
```

Probar también este modo

```
z= rnorm(200,70,20)
plot(z,y,type="l",lwd=2,col="red")
hist(y,main="Histograma de Datos Observados")
plot (density(z) , Main="Densidad de datos aleatorios")
```

Actividad Colaborativa:
Ensaye este tipo de análisis con datos de su propia actividad.

Ejercicio 2

Con los datos del ejercicio 1 indicar cuál es la probabilidad de encontrar un caso con índice de éxito en innovación 4 o mayor en cada país.

```
pnorm(4, mean=6.5 , sd=3.5)
```

Gráfica Suplementaria

```
x=seq(0,10,length=250)
y=dnorm(x,mean=6.5,sd=3.5)
plot(x,y,type="l", lwd=2, col="blue")
x=seq(0,4,length=200)
y=dnorm(x,mean=6.5,sd=3.5)
polygon(c(0,x,4),c(0,y,0),col="gray")
```

Ejercicio 3

Ensayos varios
¿Cómo cargar datos desde la línea de comando?

```
 A <- scan()
79.98 80.04 80.02 80.04 80.03 80.03 80.04 79.97
80.05 80.03 80.02 80.00 80.02
 B <- scan()
80.02 79.94 79.98 79.97 79.97 80.03 79.95 79.97
 boxplot(A, B)
```

Actividad colaborativa : pruebe este análisis con sus propios datos.

Ejercicio 4

R-CRAN posee una enorme cantidad de datos que han sido tomados como “set de datos” para medir la performance y exactitud de los métodos y algoritmos propuestos por los propios usuarios. Muchos de estos set de datos, asociados a los paquetes instalados o por instalar son verdaderas bases de datos y pueden ser cargados en el entorno de trabajo con el comando data(). También es posible cargar set de datos en formato CSV (coma separated values) tal como los que provee excel.

Pruebe estos comandos

```
help (data)
data(faithful) // Cargar set de datos creados por el profesor H.
                    Faithful

 data()

 help(attach)
 attach(faithful) // Attacher (pegar los datos para utilizarlos)

//Este set de datos es muy utilizado para el análisis de
crecimiento de precios en zonas de vulcanismo activo. (Caso
Iquique)
```

Visualización resumen una colección de datos del set.

```
summary(eruptions)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.600   2.163   4.000   3.488   4.454   5.100

fivenum(eruptions)
[1] 1.6000 2.1585 4.0000 4.4585 5.1000

hist(eruptions)

// Graficos sin outailers
hist(eruptions, seq(1.6,5.2,0.2), prob=TRUE)
//  Agrega líneas de tendencia

lines(density(eruptions, bw=0.9))
lines(density(eruptions, bw=0.1))
```

```
// Agrega puntos en el eje de las abscisas (x)
rug(eruptions)
plot(ecdf(eruptions))

// Si los puntos son muchos conviene eliminarlos
plot(ecdf(eruptions), do.points=FALSE, verticals=TRUE)

plot(eruptions)
```

Revisar ---- No funciona en versiones posteriores a 2.13

```
library() //permite revisar si mclust está instalado.

library(Rcmdr) //para instalar R-commander

library(mclust) // Carga la librería de aglomerado de clusters
FAclusters <- Mclust(faithful)
summary(FAcluster)
plot (FAcluster)

----- fin revisar
```

Ejercicio 5

En algunas ocasiones obtener datos para alimentar un simulador es un proceso difícil, caro o imposible de llevar a adelante. Imagine determinar el promedio y desvío estándar de los flujos de caja de una empresa durante 200 meses.
A pesar de ello es posible construir el set de datos si por otros medios se conocen estos parámetros de la distribución. Este set de datos así creado puede servir para hacer minería de datos.

Generar 200 número aleatorios con media 70 y desvío estandar 20

```
x=seq(0,140,length=200)
y=dnorm(x,mean=70,sd=20)
par(mfrow=c(2,3))
plot(x,y,type="l",lwd=2,col="red")
hist(y,main="Histograma de Datos Observados")
plot (density(y) , main="Densidad de datos secuencia")


z= rnorm(200,70,20)
plot(z,y,type="l",lwd=2,col="red")
hist(y,main="Histograma de Datos Observados")
plot (density(z) , Main="Densidad de datos aleatorios")
```

Ejercicio 6

¿Qué probabilidad tenemos de sacar un evento entre los 200 y que sea igual
o menor que 35

```
pnorm(35, mean=70 , sd=20)
```

```
Gráfica Suplementaria Integradora
```

Ejercico 7

¿Cuál es la probabilidad de hallar una valor mayor (o menor) que 1 sigma, 2 sigma, y 3 sigma?

```
x=seq(-4,4,length=200)
y=dnorm(x)
plot(x,y,type="l", lwd=2, col="blue")
x=seq(-1,1,length=100)
y=dnorm(x)
polygon(c(-1,x,1),c(0,y,0),col="gray")
```

```
Resultados = 68% , 95% y 99,7%
```

Ejercicio 8 Cuantiles

En algunos casos conocemos la curva de probabilidades, e incluso se nos fija en nivel de confianza. ¿Qué valor debería asumir el valor de la variable independiente para tener ese nivel de confianza.

Ejemplo, Calcule el nivel de confianza para el valor de las acciones de una compañía cuyos precios varía con una distribución normal de media 130 y sigma 50. El nivel de confianza debe ser del 90% . Se pretende saber que probabilidad hay de que la acción sobrepase ese valor?

```
qnorm(0.90,mean=130,sd=50)
```

```
Graficar
```

Ejercicio 9

Usando el caso del ejercicio 8 , ¿Qué posibilidad hay de que el valor de la acción esté por debajo de 120?

Ejercicio 10

¿Qué distribución de probabilidad sigue el tiempo de recupero de inversión especulativa en la bolsa? (datos en días, fracción decimal)

```
0.874  0.000  2.469  0.494  3.297 60.459 98.504  0.000  0.001
```

```
33.479  0.001  0.000  0.021  0.000  0.003  0.001 25.733  1.824
74.567  0.001  0.041  0.011  0.056  2.100  0.032
```

Ejercicio 11

Probabilidad condicional

Base teórica, el Teorema de Bayes

Nuestro objetivo es ir avanzando paso a paso, desde conceptos básicos de estadística hasta poder desarrollar las bases del teorema de Bayes, comenzaremos con las probabilidades condicionales (o condicionadas)

Las probabilidades condicionadas se calculan una vez que se ha incorporado información adicional a la situación de partida: Existe un evento incierto y la investigación de campo nos aporta una nueva pista que hace que los cálculos tengan mayor certeza respecto del evento que investigamos.
En los problemas de tránsito es común que tengamos un evento sobre el que nos interesa saber algo y dado ese evento encontramos evidencia que nos aproxima más a la comprensión de lo que ocurre.

Ejemplo: se tira un dado y sabemos que la probabilidad de que salga un 2 es 1/6 (probabilidad a priori). Si incorporamos nueva información (por ejemplo, alguien nos dice que el resultado ha sido un número par) entonces la probabilidad de que el resultado sea el 2 ya no es 1/6.

Para poder deducir la probabilidad de eventos deberemos deducir las reglas de Laplace que nos permiten calcular las probabilidad es de unión o intersección y la de eventos excluyentes.

a) Si lanzamos el dado, ¿Cual es la probabilidad de que salga un número par “0” un menor que 2?

D={1,2,3,4,5,6} 100% da casos que pueden salir
P(a) = {2,4,6} todos los casos posibles en que el resultado es par.
P(b) = {1} todos los casos en que el resultado es menor que 2

Probabilidad de que sea par “o” menor que 2 = casos exitosos / casos totales ; o sea 4 sobre 6

Esto se llama evento unión y su expresión es la suma de las probabilidades.

$$P(a\,U\,b)=P(a)+P(b)$$

En R el resultado sería

```
Pa= 3/6
Pb=1/6
Pa_or_b= Pa + Pb
```

Pa_or_b

a) Si lanzamos el dado, ¿Cual es la probabilidad de que salga un número par “o” un menor que 2?

D={1,2,3,4,5,6} 100% da casos que pueden salir
P(a) = {2,4,6} todos los casos posibles en que el resultado es par.
P(b) = {1} todos los casos en que el resultado es menor que 2
P or = 4/6

Probabilidad de que sea par “y” mayor que 5 = casos exitosos / casos totales ; o sea 1 sobre 6

Esto se llama evento intersección y señala con el punto como elemento concatenador.

P(A) = {2,4,6} = 3/6
P(b) = {6} = 1/6 Vemos que los dos conjuntos tienen sólo al 6 como elemento común
Se dice que estos eventos no son mutuamente excluyentes.

Las probabilidades condicionadas se calculan aplicando la siguiente fórmula:

$$P(B.A)=\frac{(P(A)\wedge P(b))}{(P(A))}$$

en R la solución sería

PA=3/6 = ½
PB= 1/6
P_a_y_b= (PA * PB) / PA

Probabilidad Condicional

Este es un caso especial , en el que, dependiendo de los eventos sean o no mutuamente excluyentes, nos permitiría deducir los dos casos anteriores.

Este tipo de problema consiste en averiguar la probabilidad de algún evento , dado que algo ha sucedido.

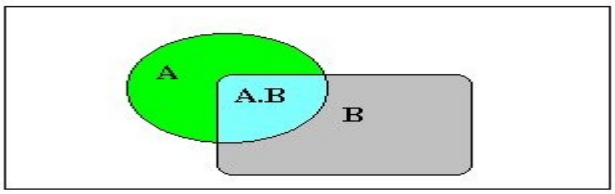
Por ejemplo si tenemos un mazo de 52 cartas, si alguien nos dijese que la carta que salió es un dos, ¿Cuál sería la probabilidad de que esa carta sea 2 de trébol?

Llamaremos evento A a la probabilidad de que salga 2

PA= 4 / 52

Llamaremos B a la probabilidad de que salga el 2 de trébol es

$P(A)\cap P(B)$ PB= 1/52 (son solamente los elementos del espacio muestral que son comunes a ambos conjuntos.



Luego si la ocurrencia del evento está restringida por la ocurrencia de un evento anterior (ejemplo , ya sabemos que es un dos el que salió y queremos saber si es de trébol) se procede a calcular con la siguiente expresión:

$$P(B/A)=\frac{(P(A)\cap P(b))}{(P(A))}$$

Suele denominarse :

P (B/A) como la probabilidad de que se dé el suceso B condicionada a que ya se haya dado el suceso A.

P (B n A) es la probabilidad del suceso simultáneo de A y de B

P (A) es la probabilidad a priori del suceso A

$$P(B/A)=\frac{(P(A)\cap P(b))}{(P(A))}$$

En R-Cran este problema se resuelve así.

```
Pa=4/52
Pb=1/52
Payb= Pa * Pb

Pa_dado_b= (Pa_n_b / Pa)
```

b) Intersección de sucesos: es aquel suceso compuesto por los elementos comunes de los dos o más sucesos que se interseccionan. La probabilidad será igual a la probabilidad de los elementos comunes.

Ejemplo: lanzamos un dado al aire y analizamos dos sucesos: a) que salga número par, y b) que sea mayor que 3. La intersección de estos dos sucesos tiene dos elementos: el 4 y el 6.

Su probabilidad será por tanto:

P(A ^ B) = 2 / 6 = 0,33

Unión de dos o más sucesos: la probabilidad de la unión de dos sucesos es igual a la suma de las probabilidades individuales de los dos sucesos que se unen, menos la probabilidad del suceso intersección

Ejemplo: lanzamos un dado al aire y analizamos dos sucesos: a) que salga número par, y b) que el resultado sea mayor que 3. El suceso unión estaría formado por los siguientes resultados: el 2, el 4, el 5 y el 6.

P(A) = 3 / 6 = 0,50 // Tres número pares (2, 4 y 6) sobre un total de 6

P(B) = 3 / 6 = 0,50 // Tres números del dado son mayores que 3 (el 4, 5 y 6)

P (A ^ B) = 2 / 6 = 0,33 // 4 y 6 aparece en los dos conjuntos

Por lo tanto,

P (A u B) = (0,50 + 0,50) - 0,33 = 0,666

Ejercico 12

Uno de los problemas más importantes que enfrentamos en el financiamineto de la innovación, son los relativos al fraude tecnológico. Los casos que se producen, en algunos paísesdel mundo, son tan importantes que afectan a las estadísticas de riesgo crediticio de algunos clusters. De hecho Naciones Unidad y muchos otros organismos multilaterales u ONGs se refieren a estos como epidemia del Silicon Vally.

Es tan común esta forma de expresar el índice de fracaso tecnológico accarreado por fallas en el financiamiento que amenudo se usan términos de estadística de la saludo al momento de relevar los datos.

La prevalencia (probabilidad de ocurrencia) de una enfermedad es calculada por los médicos como:

$$Prevalencia = \frac{(Número\,de\,Casos)}{(Tamaño\,de\,la\,Población\,en\,Riesgo)}$$

Cuando utilizamos el planteamiento de frecuencia relativa para establecer probabilidades, el número que obtenemos como probabilidad adquirirá mayor precisión a medida que aumentan las observaciones

Por ejemplo, el fonarsec estima que en Mendoza hay 8 clusters con iniciativas innovadoras. Se estima que el tiempo para madurar la idea hasta que las ventas superan a los cotos oporativos es de 5 años. En estos clusters hay 200 empresas que se fundan por año. Si la cantidad de quebrantos por año atribuibles a que los fondos llegaron tarde es de 3 empresas por cluster por año, ¿Cuál sería la prevalencia provincial y cuál la prevalencia por cluster?
(La prevalencia siempre debe estar referida a un horizonte de tiempo)

Ejercicio 13

Aplicación de Probabilidad condicional

En un ejemplo de accidentología , la probabilidad de que un vehículo choque en un cruce altamente peligroso fue de 0.6266. Sin embargo, la probabilidad tendría que modificarse si se supiera de antemano cuantas horas condujo el chofer. Esto introduce la idea de probabilidad condicional o, la probabilidad de que A ocurra dado que B ha ocurrido.

Ejemplo:
Suponga que 135 personas que pasaron ese cruce condujeron más de 8 horas seguidas y 60% registró un incidente.
Otras 230 pasaron y habían conducido menos de 8 horas y solamente 6 registraron incidentes.
Luego de este estudio se encuentra un accidente. ¿Qué probabilidad existe que el conductor haya conducido más de 8 horas?

Veamos como resolver esto:

Total de autos que pasaron = 135 + 230 = 365
Probabilidad evento A, 60/100 . Esto implica que 135*0,6 = 81 automóviles del grupo A han tenido incidentes.
Probabilidad evento B = 6 / 230 = 2,6%
Total de automóviles con incidentes = 81 + 6 = 87
Probabilidad de incidente = 87 / 365 (A n B).

La probabilidad de A dado B es:

$$P(A/B)=\frac{(P(A)\cap P(b))}{(P(B))}=\frac{87/365}{135/365}=0{,}6350$$

Si nos hubiesen preguntado ¿Cuál es la probabilidad de que hubiese tenido el accidente y hubiese conducido menos de 8 horas? , la solución sería.

$$P(B/A)=\frac{(P(A)\cap P(b))}{(P(A))}=\frac{87/365}{230/365}=0{,}3782$$

Una tendencia común es pensar que la suma de las probabilidades recién calculadas debería ser la probabilidad total (o sea P=1). Cosa que puede verificarse que no se cumple. En efecto autos que pertenecen a A y a B que no han registrado incidentes.

La relación entre estas probabilidades la aporta la expresión del teorema de Bayes que dice:

$$P(B/A)=P(A/B)\frac{P(B)}{P(A)}$$

Otro ejemplo:

Si A_1, A_2,... , $A_n$son: son sucesos incompatibles 2 a 2.

Y cuya unión es el espacio muestral . $A_1 U A_2 U A_3 ... U A_n = E$

Y B es otro suceso.

Resulta que:

$$p(A_i/B) = \frac{p(A_i) * p(B/A_i)}{p(A_1) + p(B/A_1) + P(A_2) * p(B/A_2) + ... + p(A_n) * p(B/A_n)}$$

Las probabilidades $p(A_1)$ se denominan probabilidades a priori.

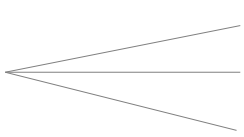
Las probabilidades $p(A_i/B)$ se denominan probabilidades a posteriori.

Las probabilidades $p(B/A_i)$ se denominan verosimilitudes.

Ejercicio 14

Resolución por árboles

El 20% de los empleados de la empresa TENARIS tienen un grado de Ingeniería Industrial y otro 20% son economistas. El 75% de los Ing. Ind. ocupan un puesto directivo y el 50% de los economistas también, mientras que los que no son Ing. Ind. ni son economistas solamente el 20% ocupa un puesto directivo. ¿Cuál es la probabilidad de que un empleado directivo elegido al azar sea Ingeniero Industrial?



0,2	Ing. Ind.	0,75	Directivo
0,2	Economistas	0,5	Directivo
0,6	Otros	0,2	Directivo

$$p(IngInd/directiva)=\frac{0,2*0,75}{0,2*0,75+0,2*0,5+0,6*0,2}=0,405$$

Esta es la expresión general del Teorema de Bayes y nos muestra, como en el caso de la probabilidad condicional, cómo la búsqueda de nueva evidencia nos pone en un terrenos de más certeza respecto al resultado que hallamos.

Veamos un ejemplo con los mismos valores numéricos aplicados al tema de transporte.

Sobre una autopista muy congestionada (con alta prevalencia) se sabe que 20% de los vehículos son remolques porta contenedores, el otro 20% son camiones con carga agrícola a granel.

De la revisión de los videos capturados de estos vehículos se ha visto que el 75% de los camiones con containers tienen un incidente (cuasi accidente) en los 300 km de longitud de la carretera. Esta cifra se reduce al 50% de los camiones de carga agrícola.
Si tomamos la lista de accidentes anual, ¿Qué probabilidad existe de que el accidente corresponda a un camión con container?

Ejercicio 16

Uso de datasets
Como comentamos antes, una de las ventajas de R es la cantidad enorme de Data Sets y tratamientos estadísticos disponibles. Estos procedimientos que casi siemmpre incluyen una colección de datos asociadas se tienen que cargar con el comando:

```
install.packages ("clusters")
install.packages("party")
```

donde cluster y party son los nombres los paquetes que queremos instalar.

Si ya hemos ejecutado estos comando (se necesita hacerlo solo una vez) y queremos usar esos paquetes tendremos que invocarlos con:

```
library(cluster)
library(party)
```

Procederemos a cargar en el área de trabajo un set de datos que hemos generado con excel y lo hemos gravado en el formato CSV (Coma Separated Values)

Usaremos el archivo Herederos.csv

Sistemas_Territoriales";"I_I_Potencial";"Impcato";"Generacion";"Exito_Fracaso"
"1";5.1;3.5;1.4;0.2;"Convocatoria_Acreedores"
"2";4.9;3;1.4;0.2;"Convocatoria_Acreedores"
"3";4.7;3.2;1.3;0.2;"Convocatoria_Acreedores"
"4";4.6;3.1;1.5;0.2;"Convocatoria_Acreedores"
"5";5;3.6;1.4;0.2;"Convocatoria_Acreedores"
….....
….....
"56";5.7;2.8;4.5;1.3;"Fracaso_Tecnológico"
"57";6.3;3.3;4.7;1.6;"Fracaso_Tecnológico"
"58";4.9;2.4;3.3;1;"Fracaso_Tecnológico"
"59";6.6;2.9;4.6;1.3;"Fracaso_Tecnológico"
"60";5.2;2.7;3.9;1.4;"Fracaso_Tecnológico"
"61";5;2;3.5;1;"Fracaso_Tecnológico"
….....
….....
"120";6;2.2;5;1.5;"Cuadruplica_Ventas"
"121";6.9;3.2;5.7;2.3;"Cuadruplica_Ventas"
"122";5.6;2.8;4.9;2;"Cuadruplica_Ventas"
"123";7.7;2.8;6.7;2;"Cuadruplica_Ventas"
"124";6.3;2.7;4.9;1.8;"Cuadruplica_Ventas"
"125";6.7;3.3;5.7;2.1;"Cuadruplica_Ventas"

Para cargar estos datos en el área de trabajo, en primer lugar nos aseguraremos que el archivo esté en una carpeta conocida. Tomaré como ejemplo una carpeta que se llama c:/datasets.
Desde R-cran la carga se realiza así:

```
transit <-
read.table("c:/dataset/Herederos.csv",header=TRUE,sep=";")
```

También puede realizarse esta tarea desde el R-commander.

Para ello debemos asegurarnos que esta instalado en nuestra máquina. Si no lo está hay que ejecutar:

```
install.packages("Rcmdr")
```

```
Cuando ya está instalado se cargar con
```

```
require("Rcmdr")
```

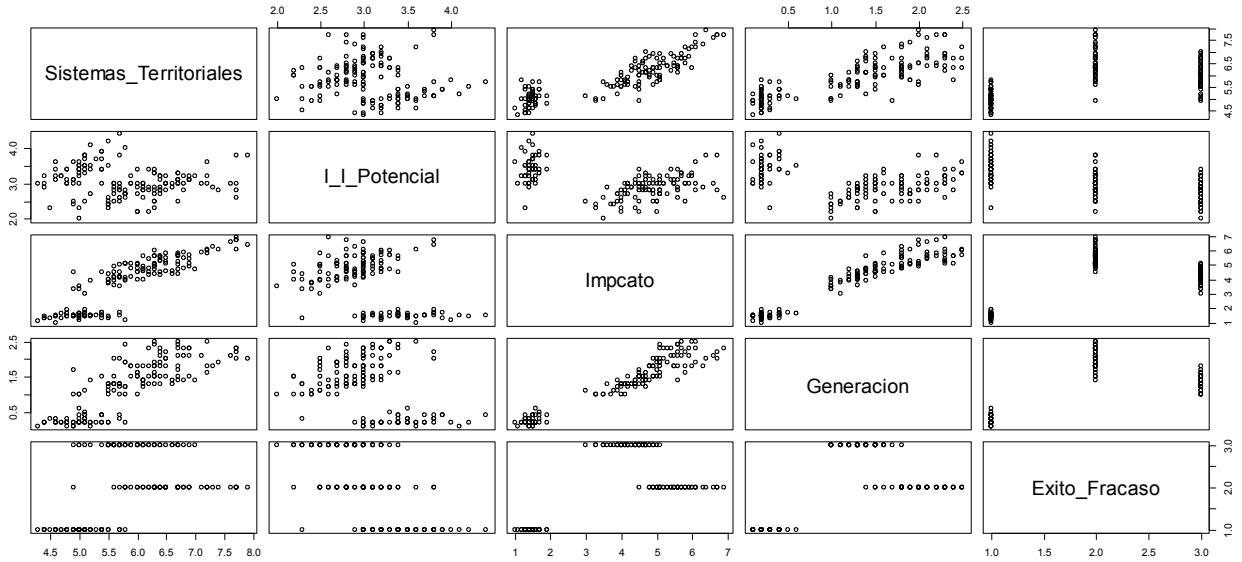
Comenzaremos a examinar el conjunto de datos.

```
plot(transit)
```

Para derivar la salida a un gráfico jpg, png, ps , etc tenemos que tipear:

```
 png("c:/Dataset/plot2.png"), plotear y luego ejecutar  dev.off()
```

"png (transit1.png)" y luego volver a generar el ploteo con " plot(transit)". Ver en el explorador de archivos el nuevo fichero transit1.png. Finalmente tipeamos "`graphics.off ()`"

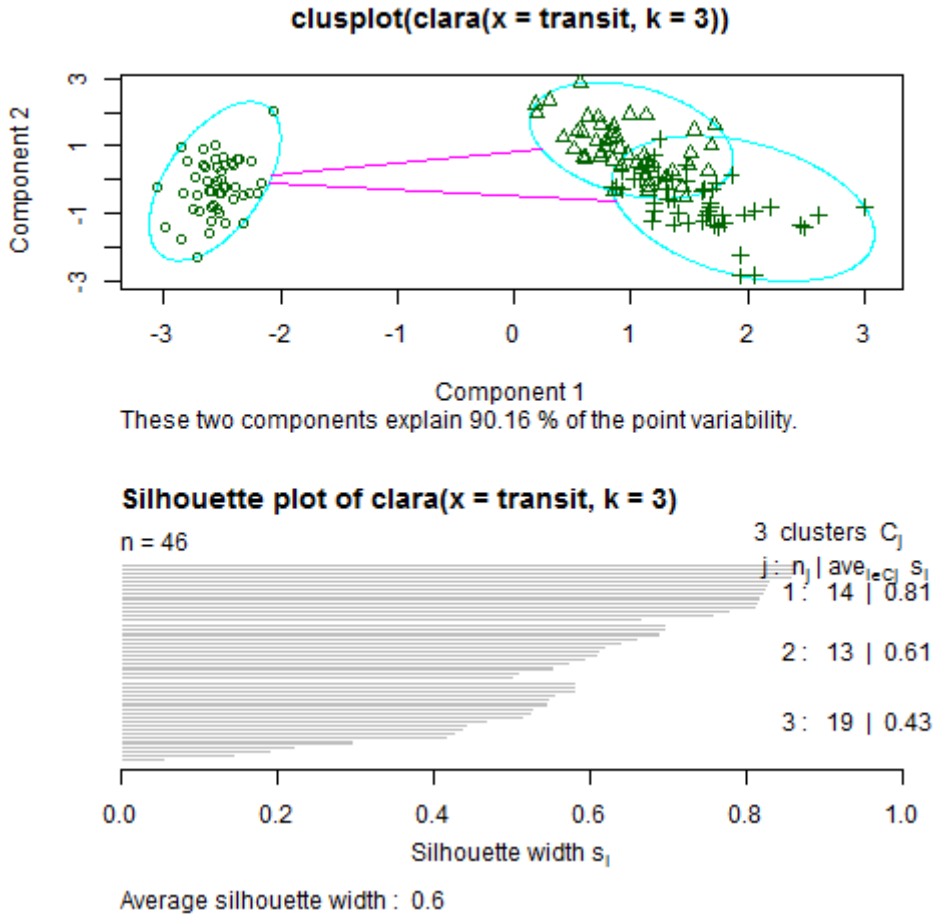


La imagen nos sugiere que existe cierta correlación entre las variables estudiadas. Además podemos inferir que hay ciertos grupos que responden diferente (diferentes correlaciones) e intuimos que estos es consecuencia del tiempo de viaje. En los datos de partida hemos señalado el origen de conductor, pero supondremos que lo desconocemos.

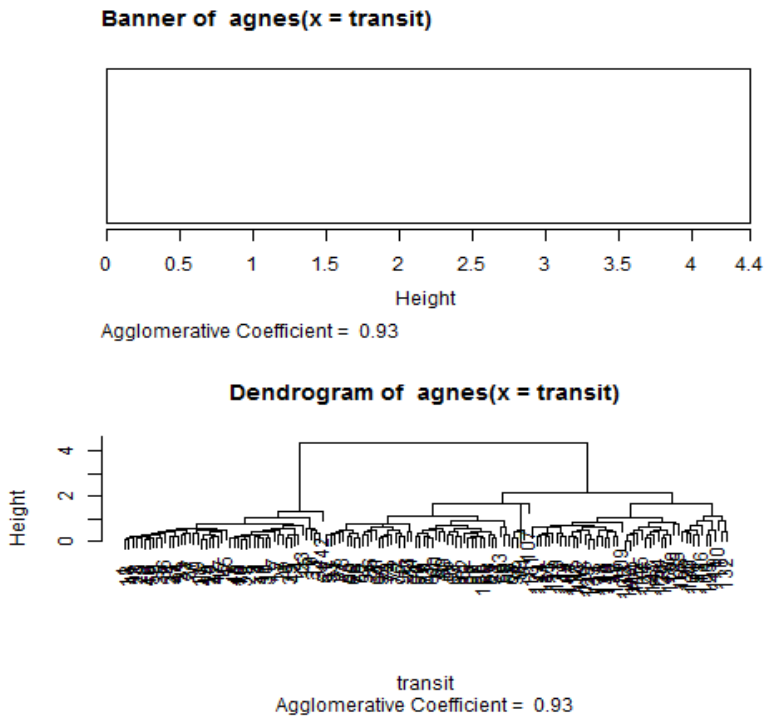
Utilizaremos el paquete cluster para que nos ayude a identificar y clasificar estos aglomerados

`library (cluster)` , luego utilizaremos los métodos de clasificación de clusters llamados Clara, Meliza y Agnes

```
par(mfrow=c(2,1), pch=14)
plot(clara(transit,3))
```



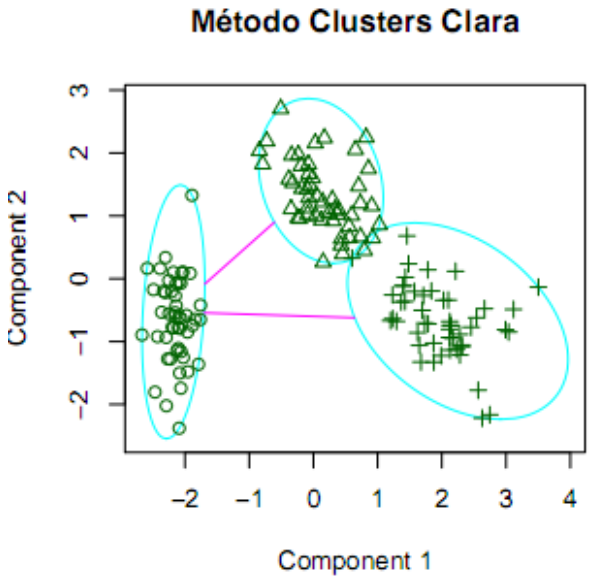

```
plot(agnes(transit))
```

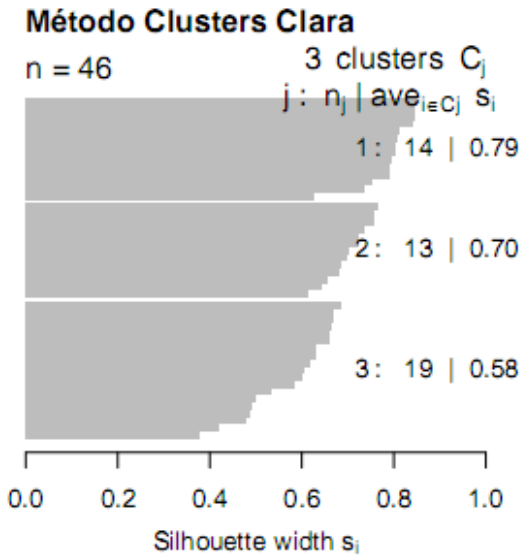


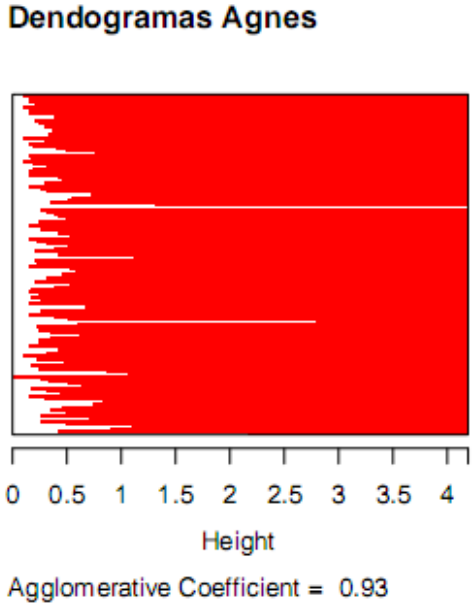
Intentaremos poner todos los gráficos en una misma figura.

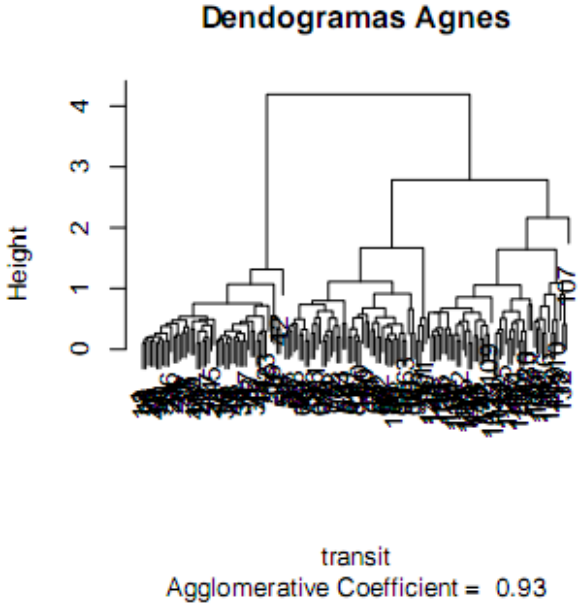
```
png(file="clusters.png",width=1024,height=768,res=150)
par(mfrow=c(2,2), pch=10)
plot(clara(transit,3), main="Método Clusters Clara")
plot(agnes(transit), main="Dendogramas Agnes")
graphics.off ()

jpg (file="clusters2.jpg")
par(mfrow=c(2,2), pch=14)
plot(clara(transit,3), main="Método Clusters Clara")
plot(agnes(transit), main="Dendogramas Agnes")
graphics.off ()
```

Existen en el paquete clusters además de los métodos vistos otros tantos llamados , Fanny, Mona, Daisy, Diana, Pam que puede ser de mucha utilidad, pero dejamos aquí la reseña para que luego lo evalúen ustedes.


Método Clusters Clara
Component 2
Component 1
These two components explain 85.3 % of the


Método Clusters Clara
n = 46
3 clusters Cj
j : nj | avei∈Cj si
1 : 14 | 0.79
2 : 13 | 0.70
3 : 19 | 0.58
Silhouette width si
Average silhouette width : 0.68


Dendogramas Agnes
Height
Agglomerative Coefficient = 0.93


Dendogramas Agnes
Height
transit
Agglomerative Coefficient = 0.93

Ejercicio 19

Entrenamiento de árboles de decisión.

Esta técnica utiliza un set de datos representativos de una situación y utilizando recursivamente el teoréma de Bayes puede armar un pronosticador o clasificador de datos. Es una técnica parecida a la de clustering, pero más refinada, pues no se basa en reglas sino en parendizaje del set de datos usado como entrenamento. En el paquete "party" existen dos funciones ctree () que se utiliza para entrenar y predict() que se usa para pronosticar o generar la regla de decición que debemos usar.

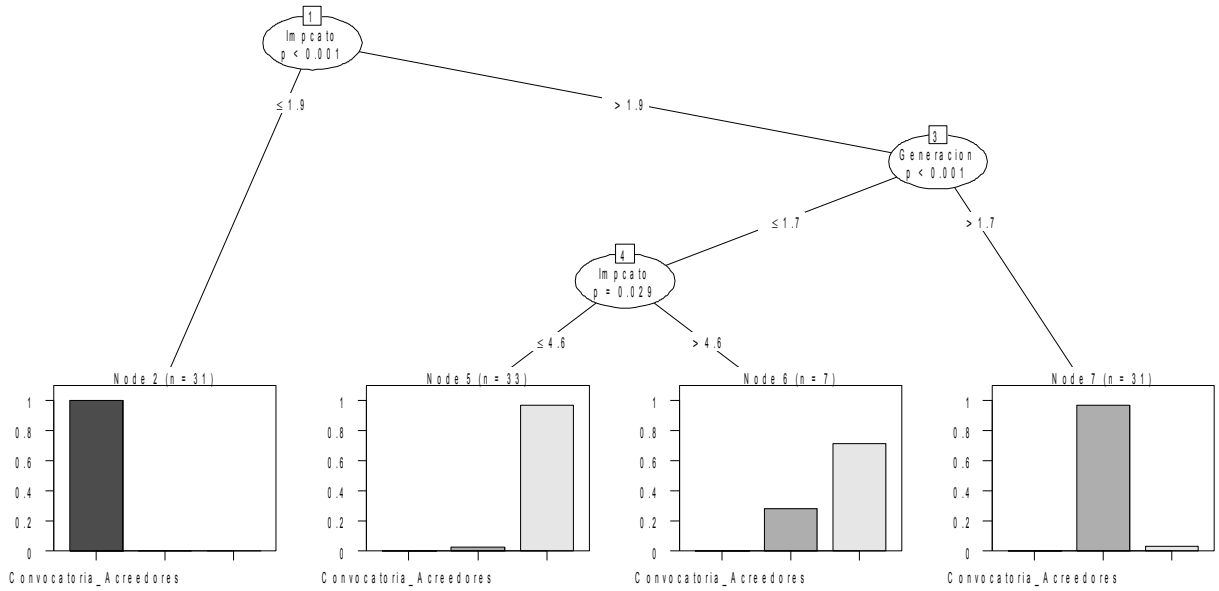
```
transit <- read.table("Herederos.csv",header=TRUE,sep=";")
attach (transit)
str(transit)    // describe al objeto transit

ind <- sample(2, nrow(transit), replace=TRUE, prob=c(0.7,
0.3))  // toma una muestra

ind  // nos imprime la muestra tomada.

trainData <- transit [ind==1,]    // genero un set de
entrenamiento
testData <- transit [ind==2,]    // genero un set de datos de
prueba

library(party)   // recargo la librería de particionado aleatorio
myFormula <- Exito_Fracaso ~  Sistemas_Territoriales +
I_I_Potencial + Impcato + Generacion
transit_ctree <- ctree(myFormula, data=trainData)    // creo el
motor de entrenamiento
# Verificar las prediciones
table(predict(transit_ctree), trainData$Exito_Fracaso)
```

```
                                  Convocatoria_Acreedores Cuadruplica_Ventas
Fracaso_Tecnológico
  Convocatoria_Acreedores                              31                  0                  0
  Cuadruplica_Ventas                                    0                 30                  1
  Fracaso_Tecnológico                                   0                  3                 37
```

```
print(transit_ctree)
```

```
 Conditional inference tree with 4 terminal nodes

Response:  Exito_Fracaso
Inputs:  Sistemas_Territoriales, I_I_Potencial, Impcato, Generacion
Number of observations:  102

1) Impcato <= 1.9; criterion = 1, statistic = 94.788
  2)*  weights = 31
1) Impcato > 1.9
  3) Generacion <= 1.7; criterion = 1, statistic = 47.384
    4) Impcato <= 4.6; criterion = 0.971, statistic = 7.165
      5)*  weights = 33
    4) Impcato > 4.6
      6)*  weights = 7
  3) Generacion > 1.7
    7)*  weights = 31
```

```
plot(transit_ctree)
```


1
Impcato
p < 0.001
≤ 1.9
> 1.9
3
Generacion
p < 0.001
≤ 1.7
> 1.7
4
Impcato
p = 0.029
≤ 4.6
> 4.6
Node 2 (n = 31)
Node 5 (n = 33)
Node 6 (n = 7)
Node 7 (n = 31)
1
0.8
0.6
0.4
0.2
0
Convocatoria_Acreedores