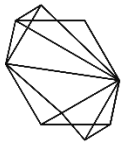


MQECONOMICS

Knowledge network

Técnicas de Análisis Multivariado

Software R

Técnicas de Análisis Multivariado

Índice

Introducción

El análisis multivariante es un conjunto de técnicas que se utilizan sobre un individuo, objeto o muestra (haciendo referencia a unidades de investigación, unidades de muestreo o unidades experimentales) cuando se tienen diversas mediciones (variables). En este caso es importante que las variables estén correlacionadas (Rencher, 2002).

El análisis multivariado siguiendo a Rencher (2002) es utilizado por dos razones: 1) Es tratable matemáticamente (aproximaciones lineales se utilizan en toda la ciencia por la misma razón) y 2) Tienen un comportamiento adecuado en la práctica.

El presente trabajo tiene como propósito analizar las principales técnicas del análisis multivariado utilizando el software R, el cual es un lenguaje y entorno para computación y gráficos estadísticos.

R ofrece una amplia variedad de estadística (lineal y no lineal, pruebas estadísticas clásicas, análisis de series temporales, clasificación, agrupaciones,...) y las técnicas gráfica. R como un conjunto integrado de servicios de softwares incluye[1]:

- *Un manejo eficaz de los datos y la instalación de almacenamiento.*
- *Un conjunto de operadores para los cálculos con matrices en matrices particulares.*
- *Una gran colección, coherente e integrado de herramientas intermedias para el análisis de datos.*
- *Facilidades gráficas para el análisis y visualización de datos ya sea en pantalla o en copia impresa.*
- *Un lenguaje de programación bien desarrollado, simple y efectivo que incluye condicionales, bucles, funciones recursivas definidas por el usuario y facilidades de entrada y salida.*

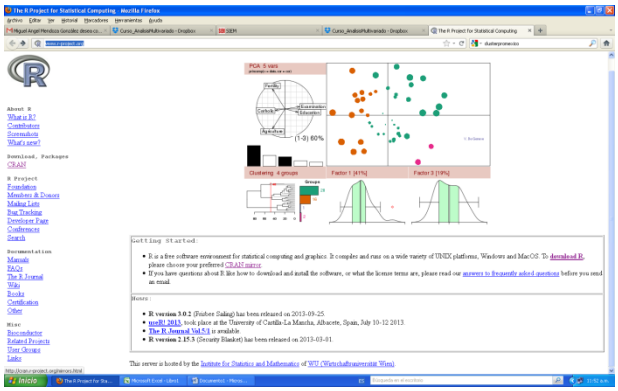
Este documento se elaboró con la información de las lecturas recomendadas y las prácticas de la clase de Métodos y Técnicas de Análisis Regional del Dr. Miguel Ángel Mendoza del Posgrado de Economía de la Universidad Nacional Autónoma de México

[1] R Project, What R? Consultada el 10 de diciembre de 2013 de http://www.r-project.org/

•

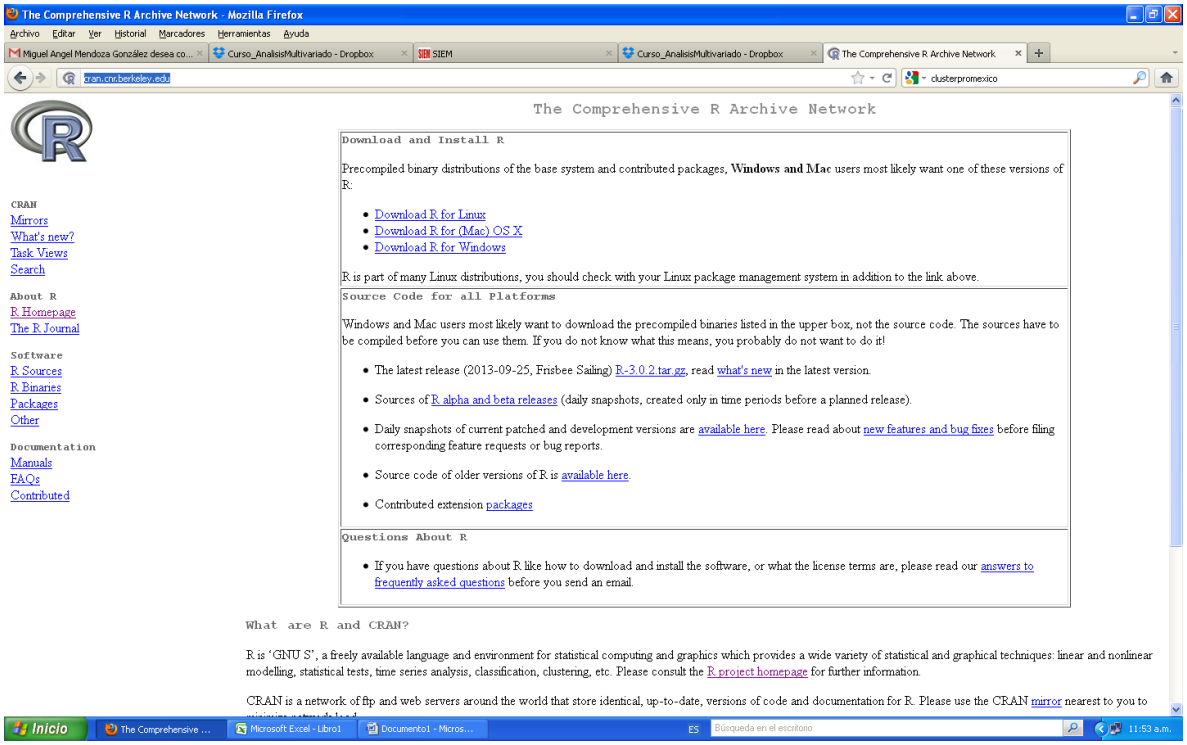
INSTALAR R

http://www.r-project.org/

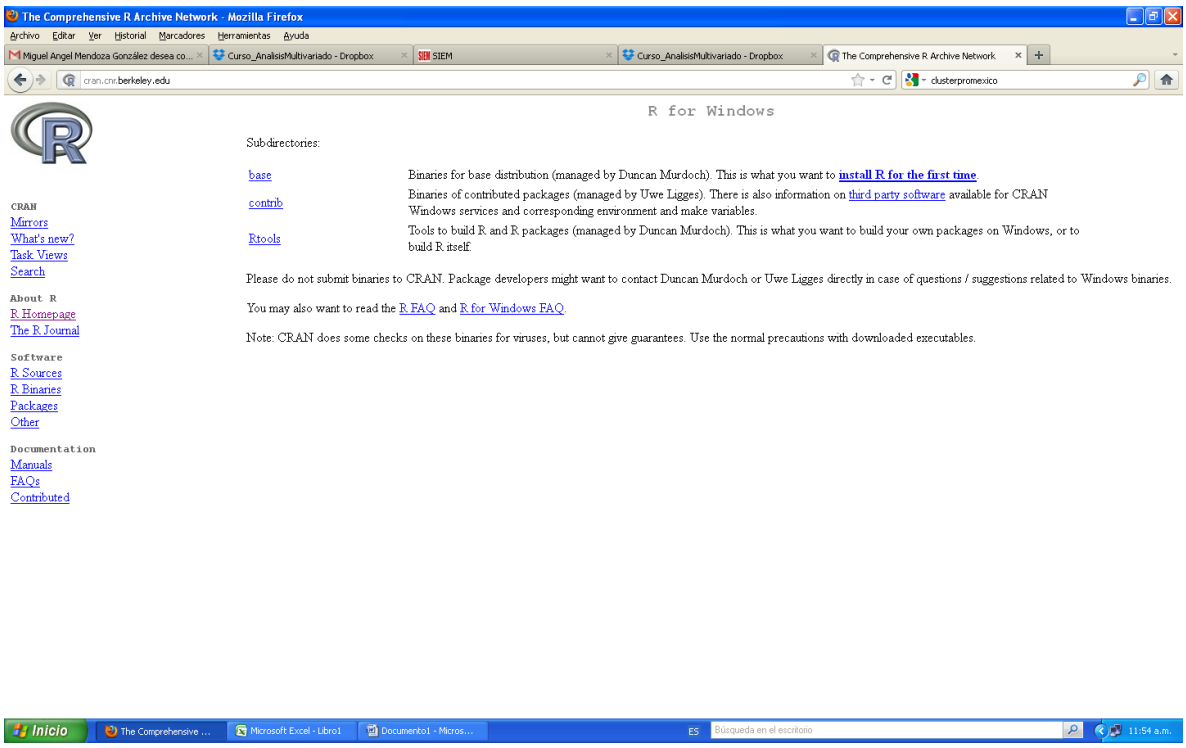


Dowload R

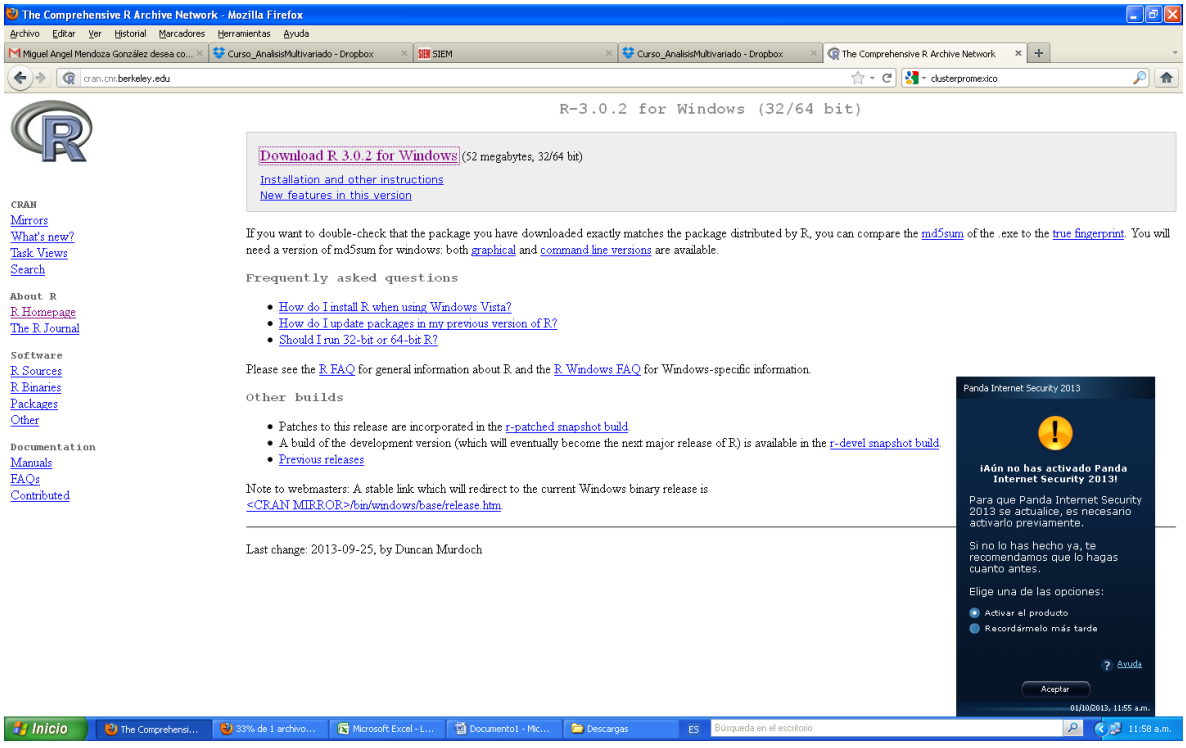
http://cran.cnr.berkeley.edu/



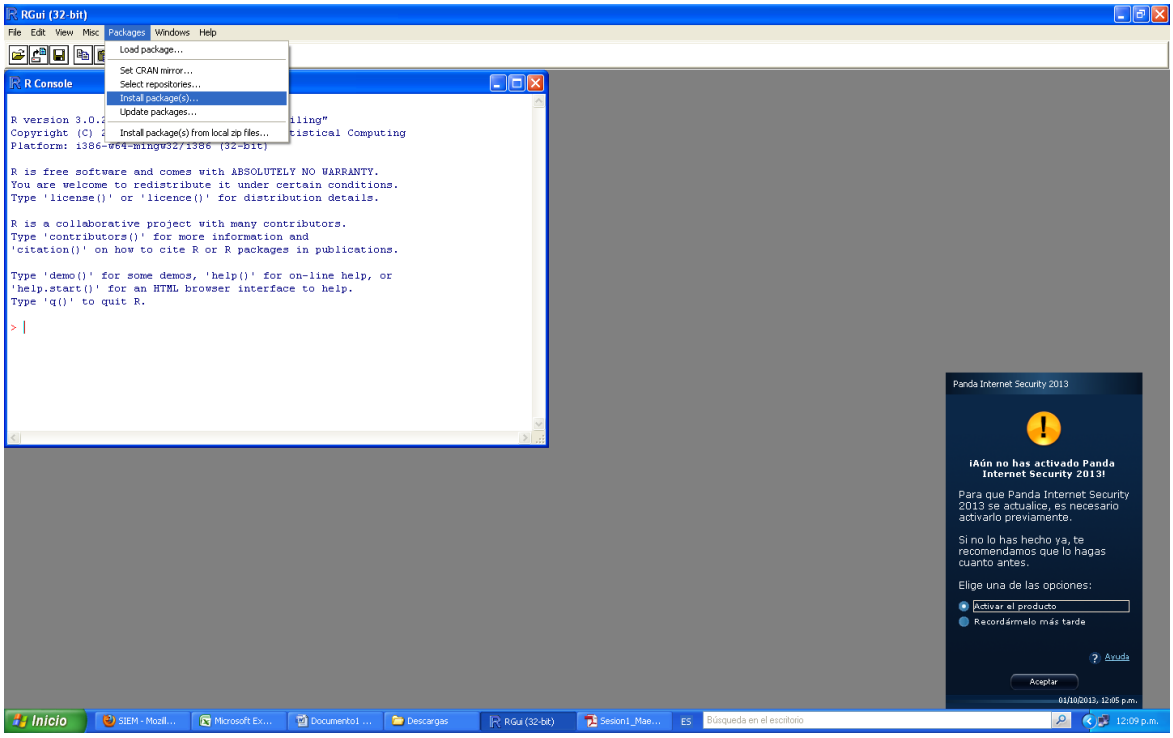
Intall R for fisrt time

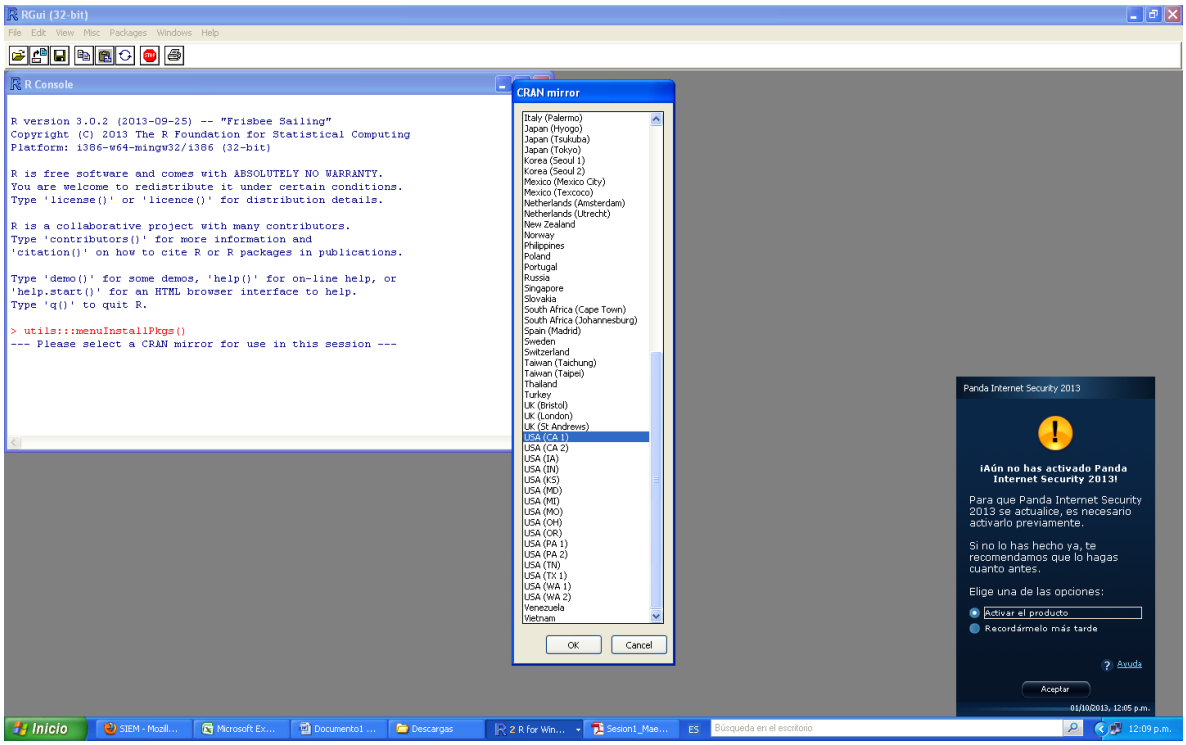


Download R 3.0.2

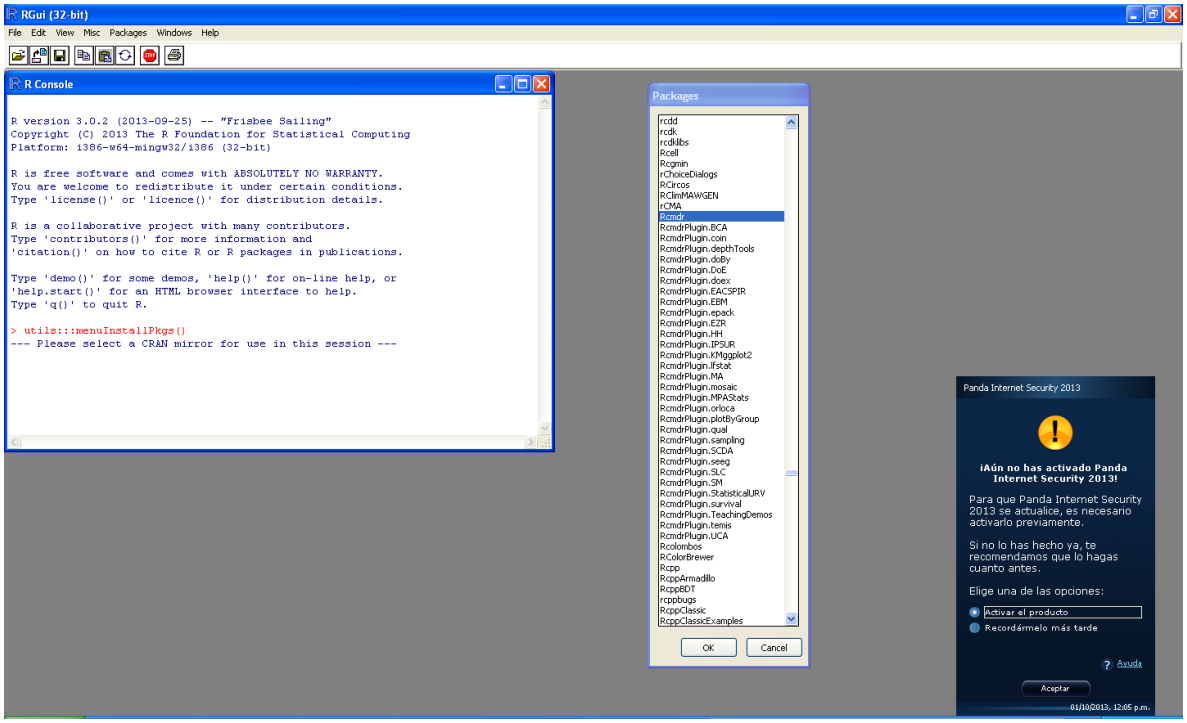


Packages/install packages/CAN 1/CAN 2



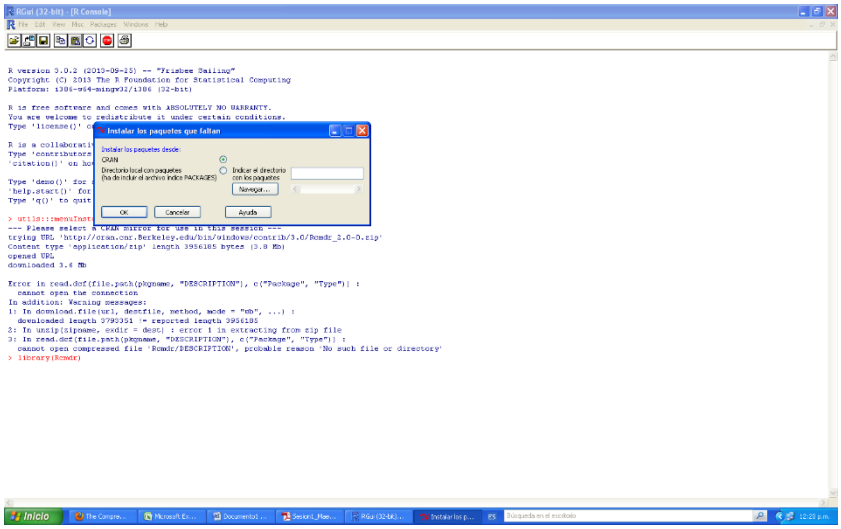


Buscar paquete Rcmdr



Escribir

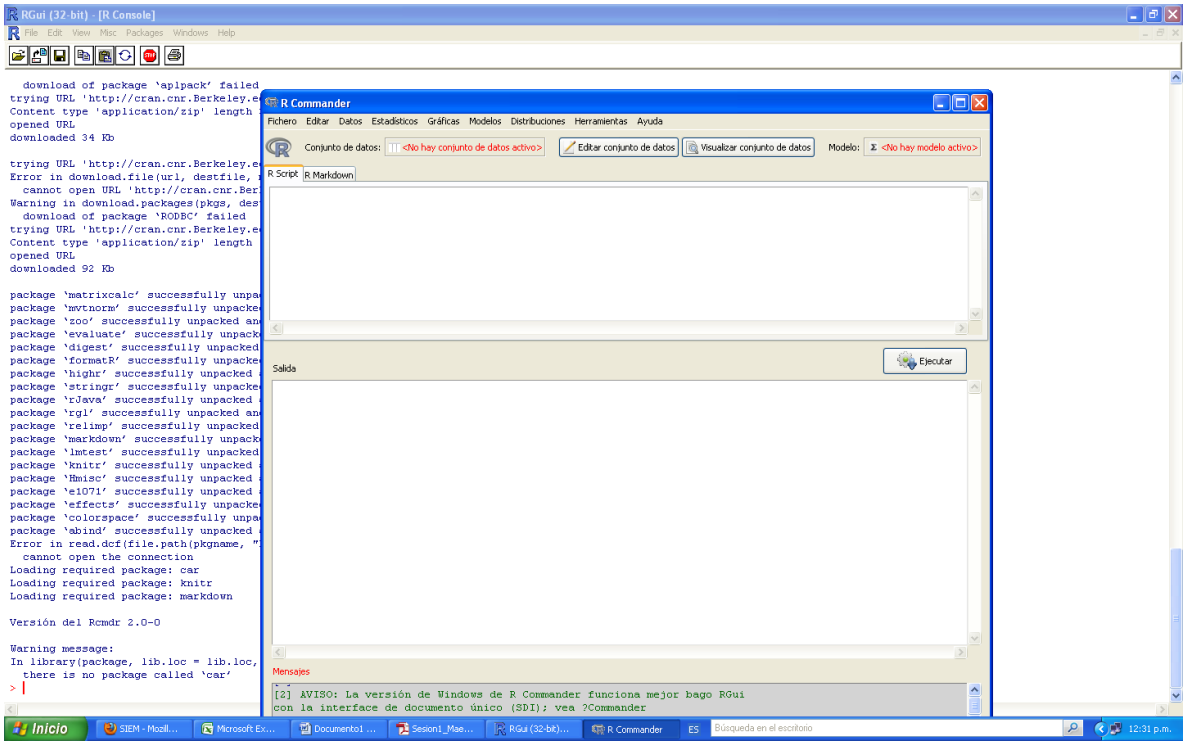
library(Rcmdr)



Aparece error

Poner que si

Se comenzara a cargar y se tarda un poco.



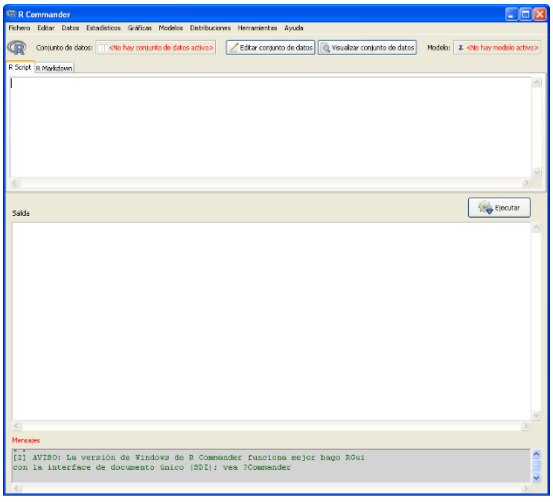
MANEJO DE DATOS

a) Análisis de la base de datos_

Fuente CONEVAL (información por municipio/marginal-)

1. Abrir R
2. Abrir el Rcmdr

```
library(Rcmdr)
```

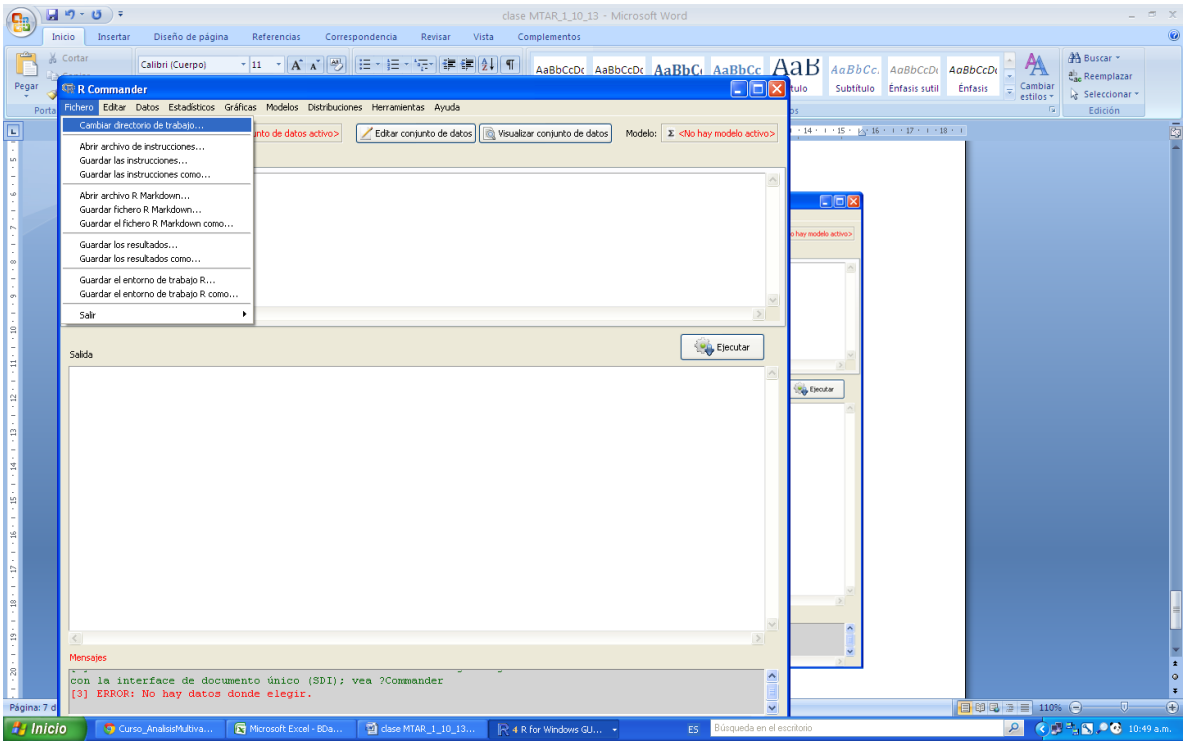


3. Cambiar carpeta para almacenar archivos

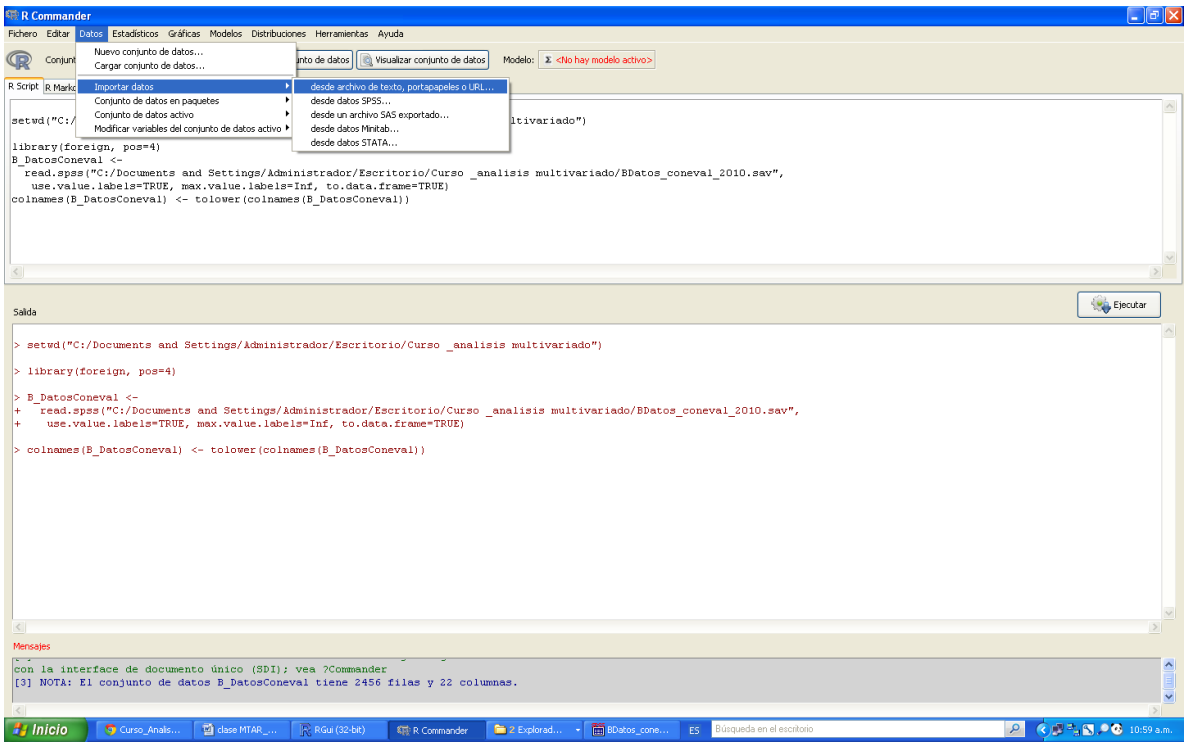
Con las herramientas de Rcomander

o´

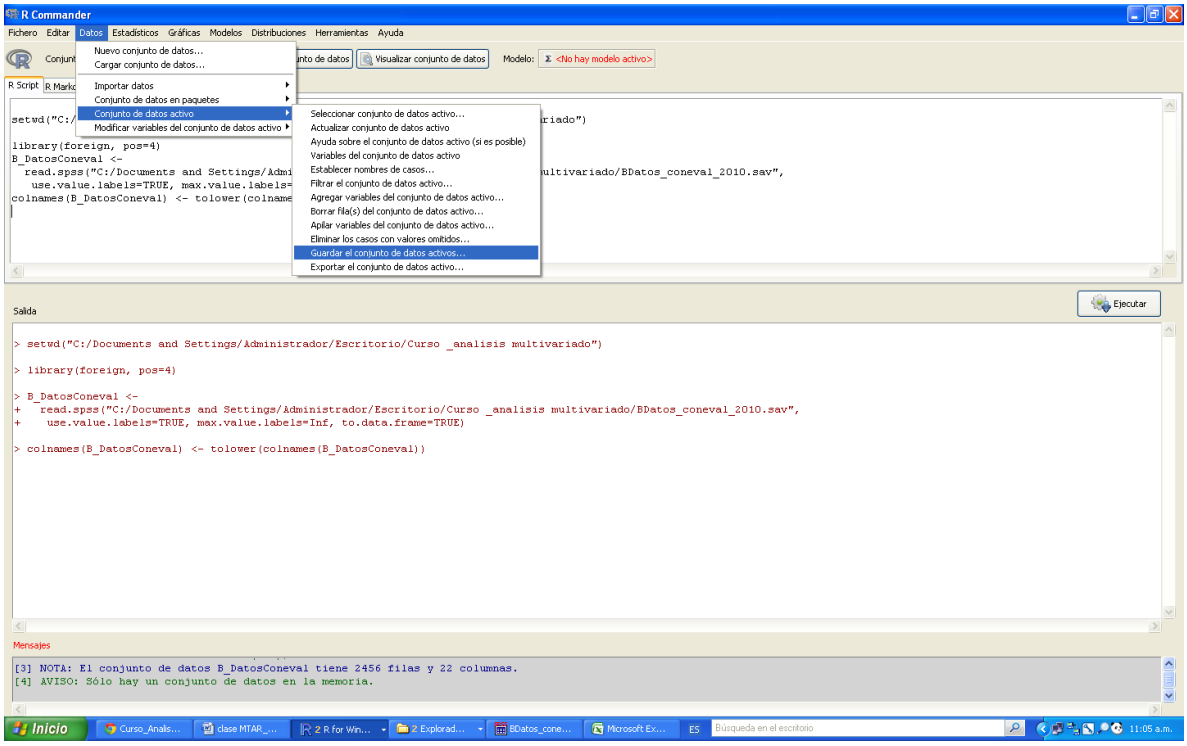
```
setwd("C:/Documents and Settings/Administrador/Escritorio/Curso _analisis multivariado")
```



4. Importar datos desde Excel u otro formato



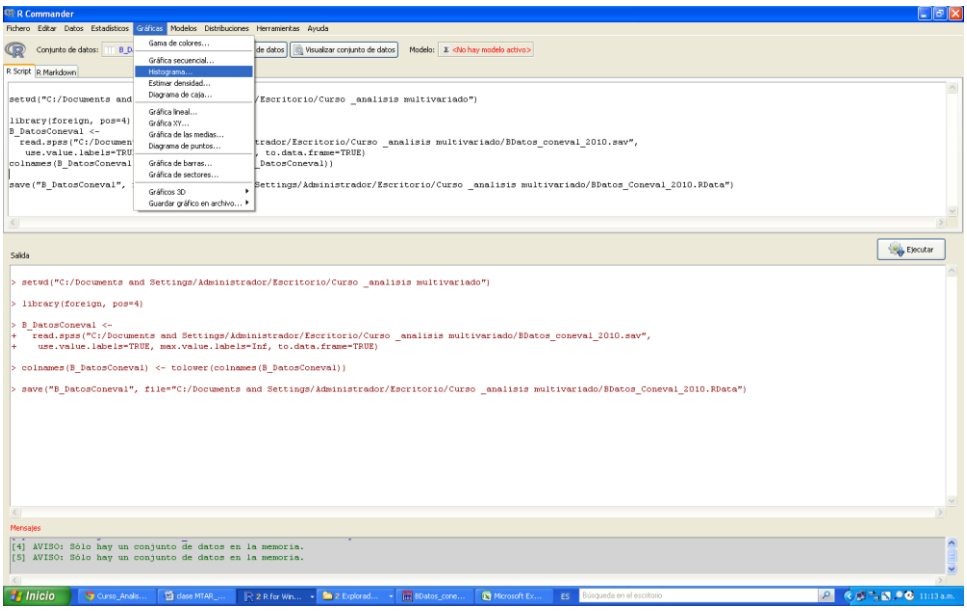
5. Guardar conjunto de datos activos

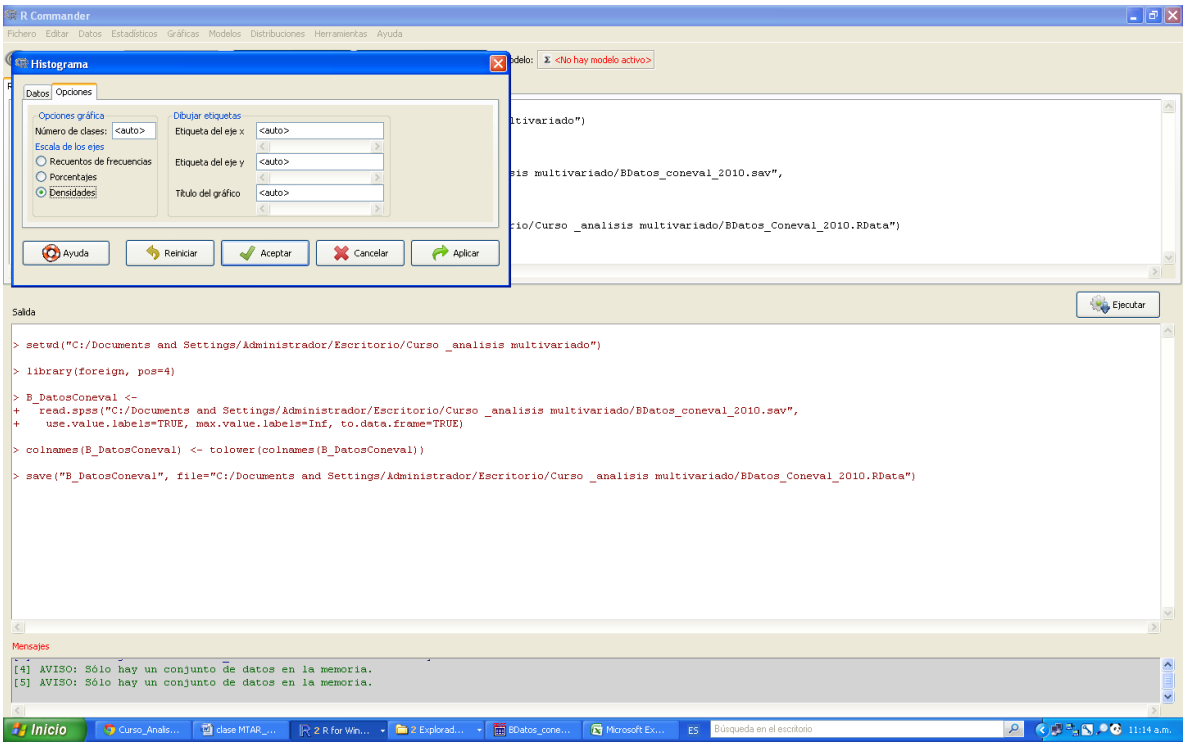


1. Análisis de funciones de distribución normal

Histograma

Un histograma es un gráfico formado por barras verticales construidas sobre una línea recta horizontal delimitada por los intervalos de la variable mostrada. Los intervalos corresponden a los de una tabla de distribución de frecuencias. La altura de cada barra es proporcional al número de observaciones que hay en ese intervalo. El número de observaciones puede indicarse encima de las barras (Newbold et al, 2008).





Variable: **Carencia Alimentaria**

El analisis del histograma nos dice que la variable esta normalmente distribuida, aunque se muestra ligeramente cargada hacia la derecha, por lo que se necesita realizar otra pruena para definir normalidad. Si realizamos una tipologia de muy baja, baja, media, alta y muy alta ordenadamente podriamos afirmar que la mayoria de la poblacion en Guanajuato se encuentra en el rango de baja carencia alimentaria.

Variable: **Grado de Marginación (g_margina_num)**

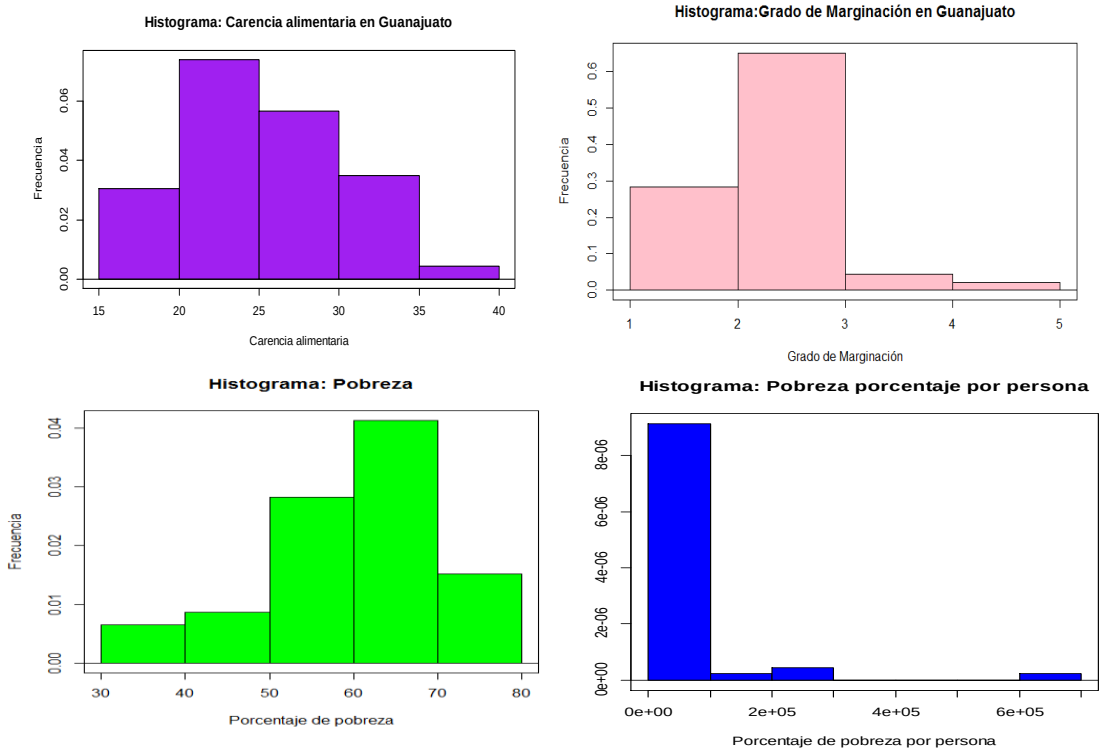
El analisis del histograma nos dice que la variable no esta normalmente distribuida, se muestra ligeramente cargada hacia la izquierda, por lo que se necesita realizar otra prueba para definir normalidad. Si realizamos una tipologia de muy baja, baja, media, alta y muy alta ordenadamente, podemos afirmar que la mayoria de la poblacion en Guanajuato se encuentra en el rango de muy bajo y bajo grado de marginación.

Variable: **Pobreza, porcentaje (pob_por)**

El analisis del histograma nos dice que la variable esta normalmente distribuida, se muestra ligeramente cargada hacia la derecha, por lo que se necesita realizar otra prueba para definir normalidad.

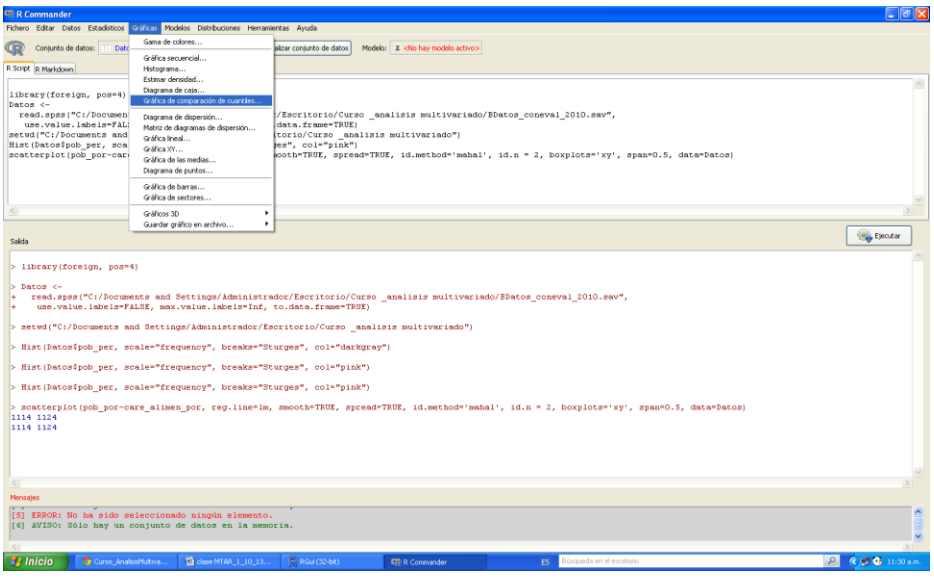
Variable: **Pobreza, porcentaje por persona (pob_per)**

El analisis del histograma nos dice que la variable no esta normalmente distribuida, se muestra ligeramente cargada hacia la izquierda, por lo que se necesita realizar otra prueba para definir normalidad.



Grafica de comparación de cuantiles QQ

Un gráfico Cuantil-Cuantil permite observar cuan cerca está la distribución de un conjunto de datos a alguna distribución ideal ó comparar la distribución de dos conjuntos de datos, es decir compara la distribución de la serie con la distribución normal entre más cerca a la normal más significativa.



Variable: **Carencia Alimentaria (care_alimen_por)**

En el caso de la variable por carencia alimentaria podemos observar que la mayoria de los datos se concentran cerca de la distribucion ideal, con esta prueba se concluye que los datos tiene una distribucion normal.

Variable: **Grado de Marginación (g_margina_num)**

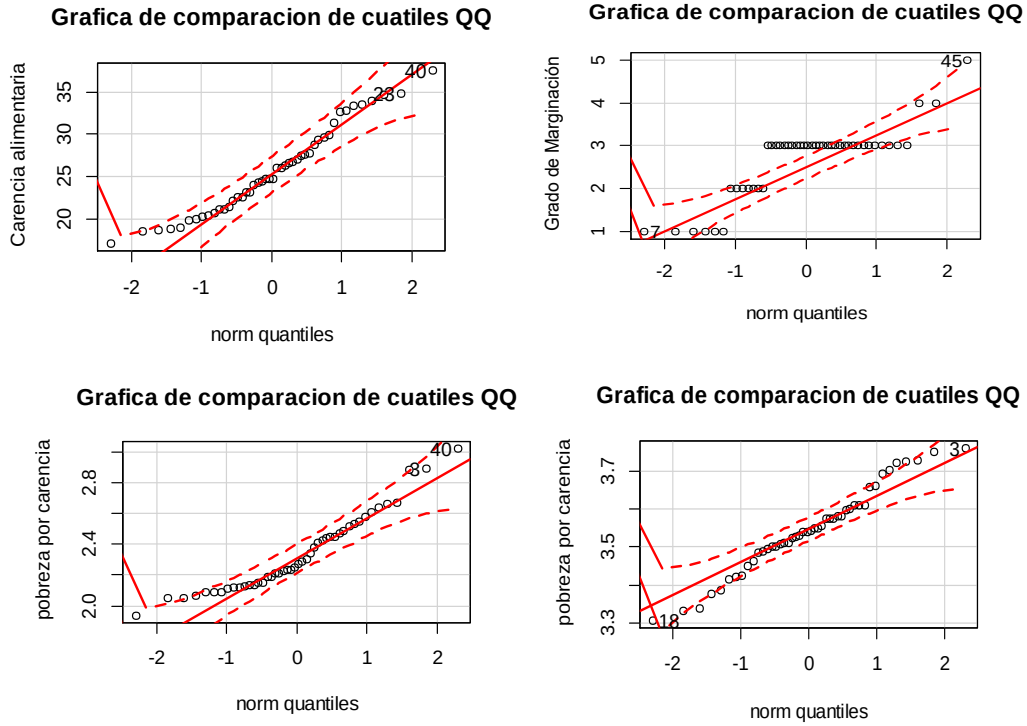
En el caso de la variable de grado de marginacion los datos no presentan una distribucion normal.

Variable: **Pobreza, porcentaje (pob_por)**

En el caso de la variable por pobreza podemos observar que la mayoria de los datos se concentran cerca de la distribucion ideal, con esta prueba se concluye que los datos tiene una distribucion normal.

Variable: **Pobreza, porcentaje por persona (pob_per)**

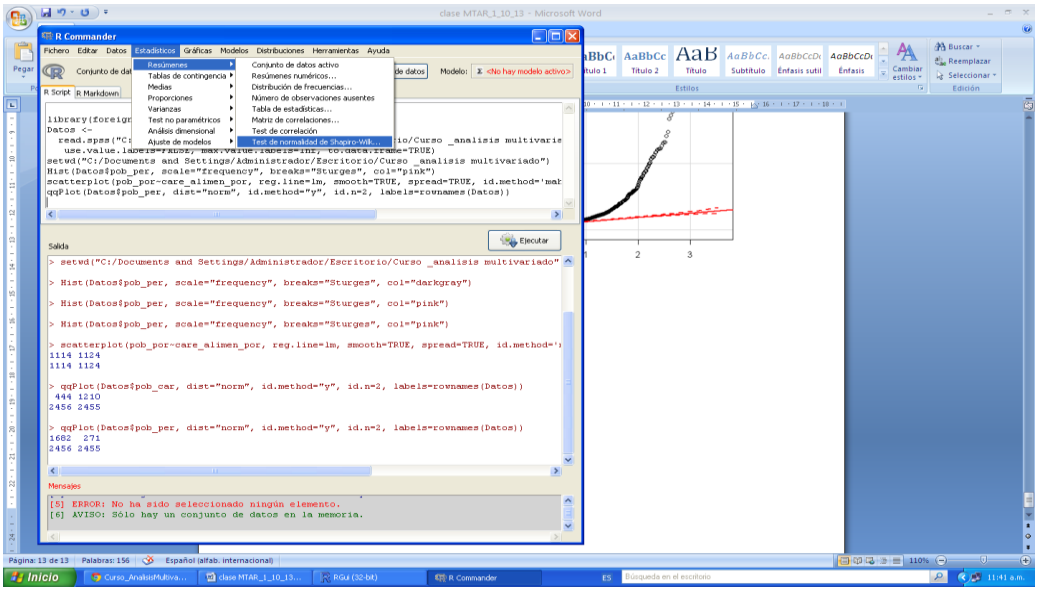
En el caso de la variable pobreza por persona podemos observar que la mayoria de los datos se concentran cerca de la distribucion ideal, con esta prueba se concluye que los datos tiene una distribucion normal.



Prueba Shapiro-Wilk para normalidad

La prueba Shapiro-Wilks se basa en estudiar el ajuste de los datos sobre un gráfico probabilístico en el que cada dato es un punto cuyo valor de abscisas (eje x) indica la probabilidad para un valor determinado de la variable, y el de ordenada (eje y) el valor esperado de probabilidad (Sesión 2). La prueba Shapiro- wilk normality:

- Ho: NORMALMENTE DISTRIBUIDO – Ec < Et y la p-value>0.05
- Ha: NO ES NORMAL - Ec > Et la p-value<0.05



Variable: **Carencia Alimentaria (care_alimen_por)**

Con la prueba de Shapiro-Wilk para normalidad podemos observar que la probabilidad es mayor a 0.05 lo que nos indica, que la muestra se distribuye normalmente.

Variable: **Grado de Marginación (g_margina_num)**

Con la prueba de Shapiro-Wilk para normalidad podemos observar que la probabilidad es menor a 0.05 lo que nos indica, que la muestra no se distribuye normalmente.

Variable: **Pobreza, porcentaje (pob_por)**

Con la prueba de Shapiro-Wilk para normalidad podemos observar que la probabilidad es mayor a 0.05 lo que nos indica, que la muestra se distribuye normalmente.

Variable: **Pobreza, porcentaje por persona (pob_per)**

Podemos observar que la probabilidad es menor a 0.05 lo que nos indica, que la muestra no se distribuye normalmente.

```
Shapiro-Wilk normality test

data:
GTO_CONEVAL$care_alimen_por

W = 0.9604, p-value = 0.1186
```

```
Shapiro-Wilk normality test

data:
GTO_CONEVAL$g_margina_num

W = 0.7672, p-value = 3.906e-07
```

```
Shapiro-Wilk normality test

data:  GTO_CONEVAL$pob_por

W = 0.9553, p-value = 0.0754
```

```
Shapiro-Wilk normality test

data:  GTO_CONEVAL$pob_per

W = 0.4985, p-value = 2.544e-11
```

Covarianza

Si dos variables x y y son medidas sobre cada unidad de investigación, se tiene una variable aleatoria bivariante (x, y). Entonces se dice que *x* y *y* tienden a covariar. La covarianza muestral:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

La covarianza es una medida de dependencia central, entre dos variables aleatorias. Si la covarianza es positiva indica que dos variables aleatorias se mueven en la misma dirección, mientras que si es negativa, indica que las dos variables se mueven en direcciones opuestas (Wooldridge, 2011).

Propiedades:

1. Si X y Y son independientes entonces Sxy = 0
2. Para todas las constantes Cov (a_1X+b_1, a_2Y+b_2)= a_1a_2Cov(X,Y)
3. $|Cov\ (X,Y)| \leq de(X)de(Y)$

Coeficiente de variación (error estándar entre la media)

$$cv = \frac{S}{\bar{x}}$$

Correlación

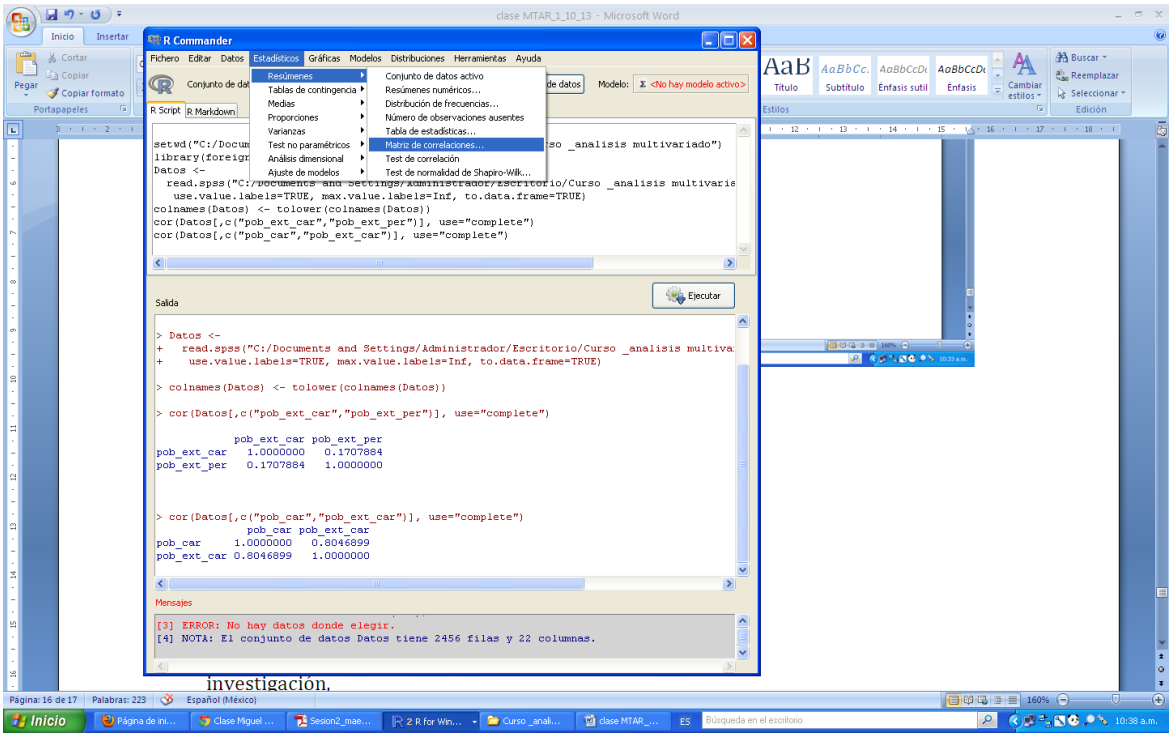
El coeficiente de correlacion se calcula dividiendo la covarianza por el producto de las desviaciones típicas de las dos variables.

$$\frac{cov\ (x,y)}{s_x s_y} = r$$

Ho=$\bar{x} = \bar{\mu}$ (no es estadísticamente significativo) tc<tt p-value>0.05 o (ac>at el error calculado es mayor que el error de tablas) X=0

Hi=x≠0 la media es diferente de cero (es estadísticamente significativo) tc>tt p-value<0.05 (ac>at el error calculado es menor que el error de tablas)

El siguiente ejemplo muestra la correlación entre la pobreza extrema por carencia y la pobreza extrema porcentual. Con este criterio el valor de 0.554 nos indica el grado de correlación alta entre pobreza extrema por carencia y la pobreza extrema porcentual.



investigación,

```
>cor(Datos[,c("pob_ext_per","pob_ext_por")], use="complete")

pob_ext_car pob_ext_per
pob_ext_car  1.0000000  0.5543863
pob_ext_per  0.5543863  1.0000000
```

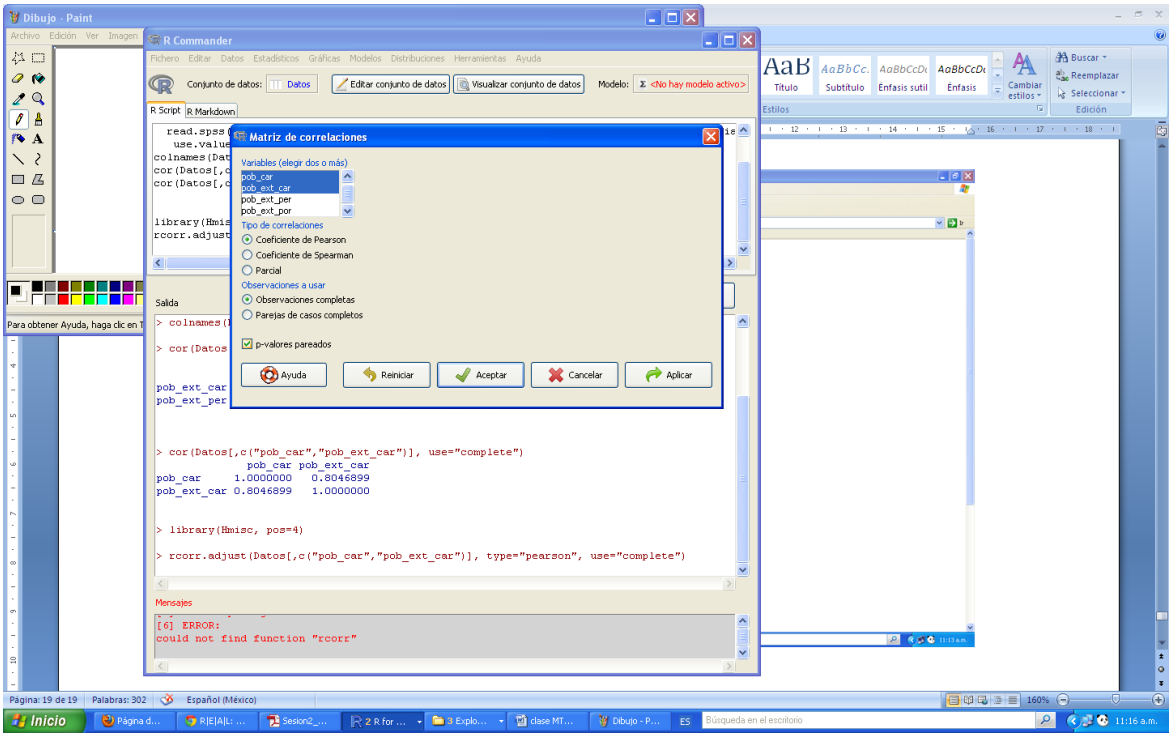
Coeficiente de Pearson (r)

Mide el grado de relación lineal entre dos variables, (dicho proceso consiste en comparar dos distribuciones y una vez que compara, dice que grado de asociación existe).

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - Y)^2}}$$

Supuestos: las variables se distribuyen como una normal, la relación es lineal y homoscedastica**.**

Criterios: el tamaño del efecto se deriva de la propuesta de Cohen; entre .10 y .29 la asociación es pequeña; entre 0.30 y 0.49 es media; y, igual o mayor 0.5 la asociación es considerable.



El siguiente ejemplo podemos observar el grado de asociación entre la variable pobreza extrema por carencia y la pobreza extrema porcentual, la cual nos indica una asociación considerable con un criterio de 0.55. Mientras que el p-value nos indica es menor a 0.05 con lo que aceptamos la hipótesis alternativa, la media es diferente de cero, lo que nos indica que es estadísticamente significativa.

```
> library(Hmisc, pos=4)
> rcorr.adjust(Datos[,c("pob_ext_car","pob_ext_per")], type="pearson")

Pearson correlations:
pob_ext_car pob_ext_per
pob_ext_car      1.0000     0.5544
pob_ext_per      0.5544     1.0000

 Number of observations: 46

 Pairwise two-sided p-values:
         pob_ext_car pob_ext_per
pob_ext_car           0
pob_ext_per  0

 Adjusted p-values (Holm's method)
         pob_ext_car pob_ext_per
pob_ext_car           0
pob_ext_per  0
```

Rango de correlación Sperman (ρ)

Estadístico no-paramétrico que es usado para medir el grado de asociación entre dos variables. No supone la distribución de las variables y es apropiado cuando las variables son medidas con una escala al menos ordinal.
En el siguiente ejemplo podemos ver que existe asociación considerable entre la pobreza por carencia y la pobreza por carencia extrema, con un criterio de 0.64. Mientras que el p-value nos indica es menor a 0.05 con lo que aceptamos la hipótesis alternativa, la media es diferente de cero, lo que nos indica que es estadísticamente significativa.

```
>rcorr.adjust(Datos[,c("pob_car","pob_ext_car")], type="spearman", use="complete")

Spearman correlations:
         pob_car pob_ext_car
pob_car     1.0000     0.6418
pob_ext_car 0.6418     1.0000

 Number of observations: 46

 Pairwise two-sided p-values:
         pob_car pob_ext_car
pob_car          0
pob_ext_car 0

 Adjusted p-values (Holm's method)
         pob_car pob_ext_car
pob_car          0
pob_ext_car 0
```

Rango de correlación de Kendall (τ)

Es una prueba no paramétrica que mide la intensidad de la dependencia entre las dos variables. Si se considera dos muestras, a y b donde el tamaño de la muestra es n y el total de pares con a y b es n(n-1)/2.

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

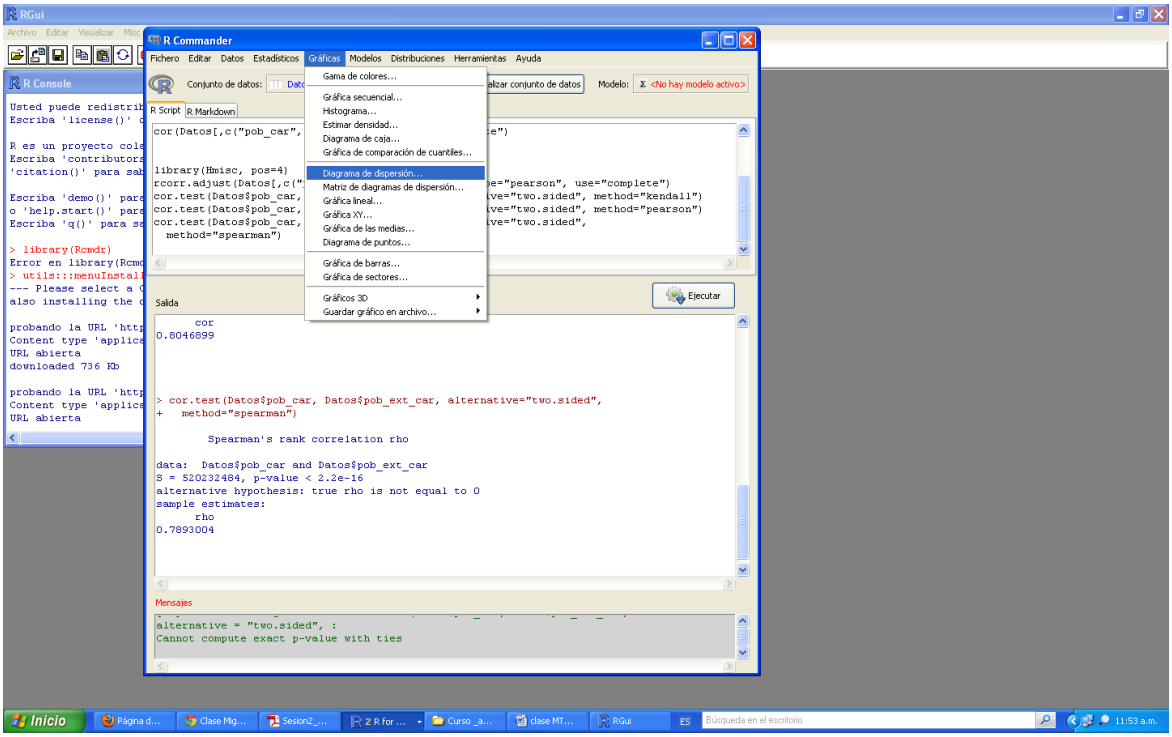
Donde: n_c es el número de concordancias y $n_d$es el número de dis-concordancias.
Concordante: ordenados en el mismo sentido
Discordante: ordenado en sentidos contrarios

Ho: τ=0 no estadísticamente significativa
Ha: $\tau \neq 0$ es estadísticamente significativa

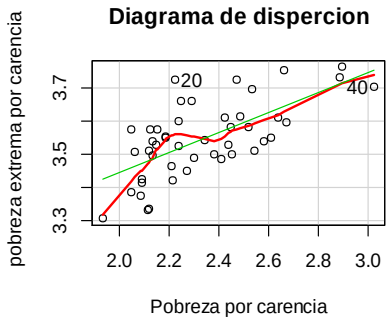
En el siguiente ejemplo la tau es diferente de cero y la probabilidad es menor a 0.05 lo que nos indica que existe una asociación moderada y que es estadísticamente significativa.

```
> cor.test(Datos$pob_car, Datos$pob_ext_car, alternative="two.sided", method="kendall")
        Kendall's rank correlation tau

data:  Datos$pob_car and Datos$pob_ext_car
T = 760, p-value = 1.859e-06
alternative hypothesis: true tau is not equal to 0
sample estimates:
    tau
0.468599
```

Diagramas de dispersión

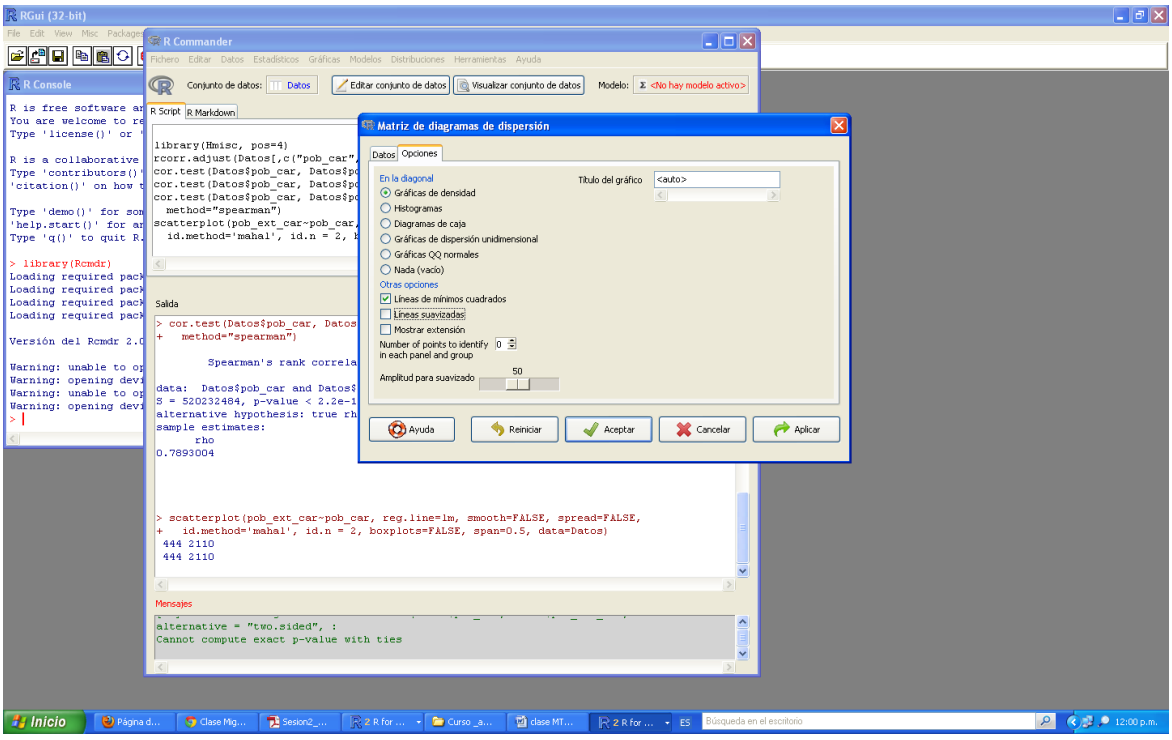


La siguiente grafica de dispersión nos muestra que existe correlación positiva entre las variables de pobreza extrema por carencia y pobreza por carencia.



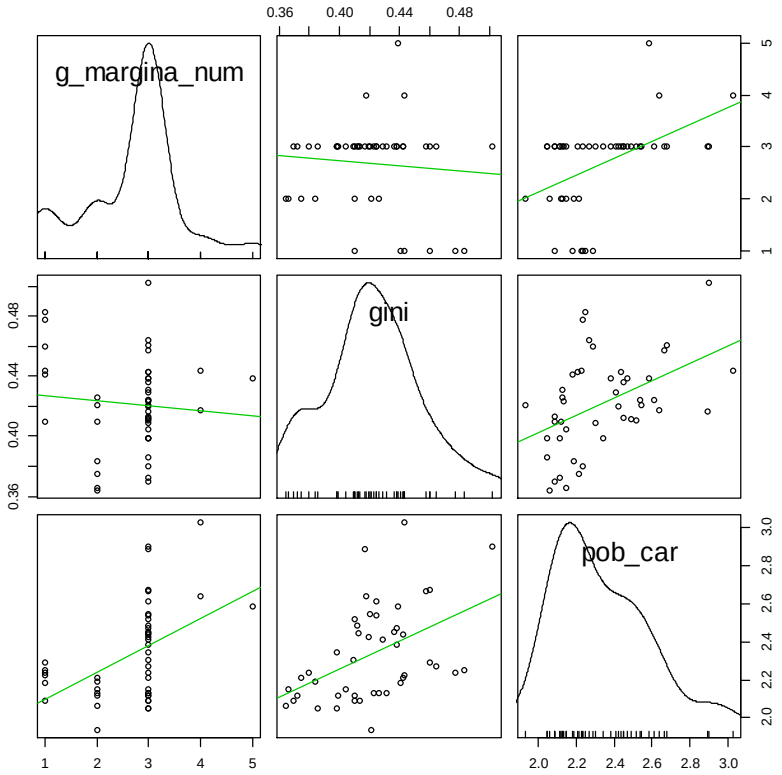

```
>scatterplot(pob_ext_car~pob_car, reg.line=lm, smooth=FALSE, spread=FALSE,
id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=Datos)
20 40
20 40
```

Matriz de diagramas de dispersión (Grafica de densidad)



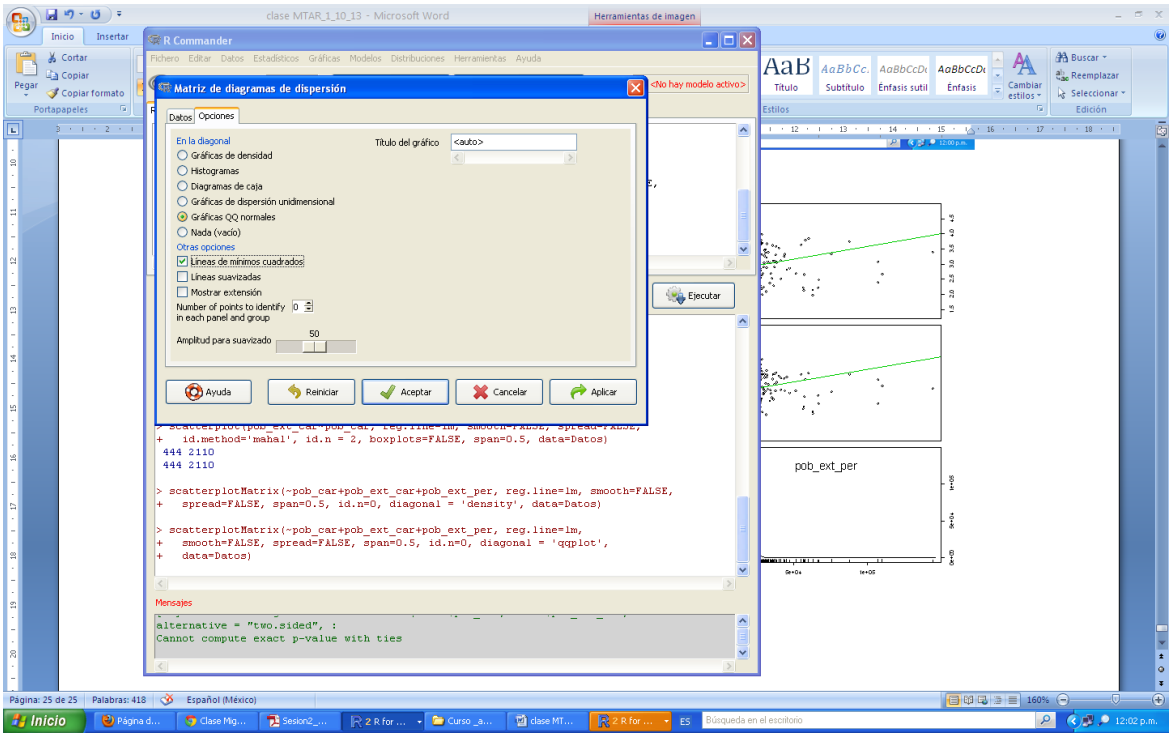
Las matrices de dispersión pueden variarse para presentar información diversa de modo gráfico sobre un mismo conjunto de datos multivariados. En el siguiente grafico se muestra en la diagonal los gráficos de densidad de cada variable, en la primera fila segunda columna se muestra una relación entre la el índice de Gini y el grado de marginación, a mayor concentración poblacional menor grado de marginación,

mientras la tercera columna de la primera fila nos muestra una correlación positiva, esto indica que el grado de marginación más alto corresponde a una pobreza por carencia más alto.

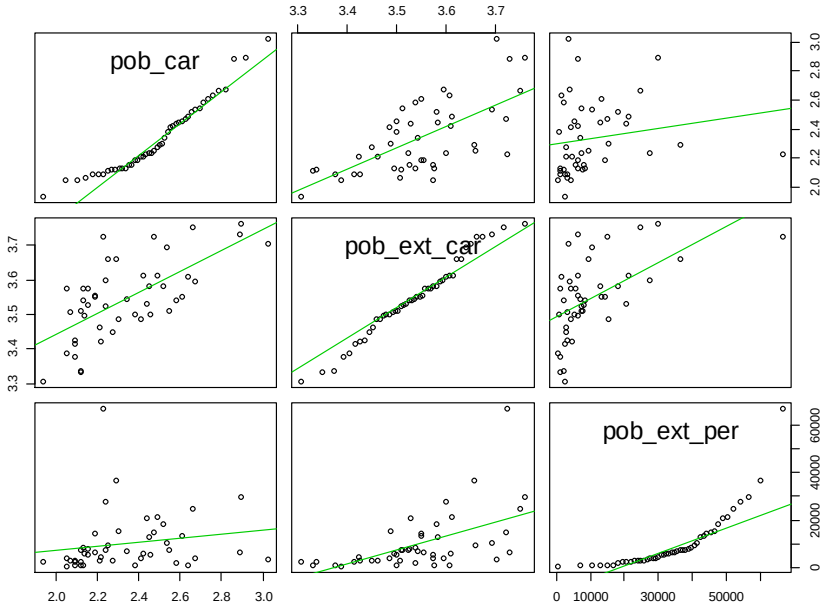


Matriz de diagramas de dispersión (Grafica QQ normales)

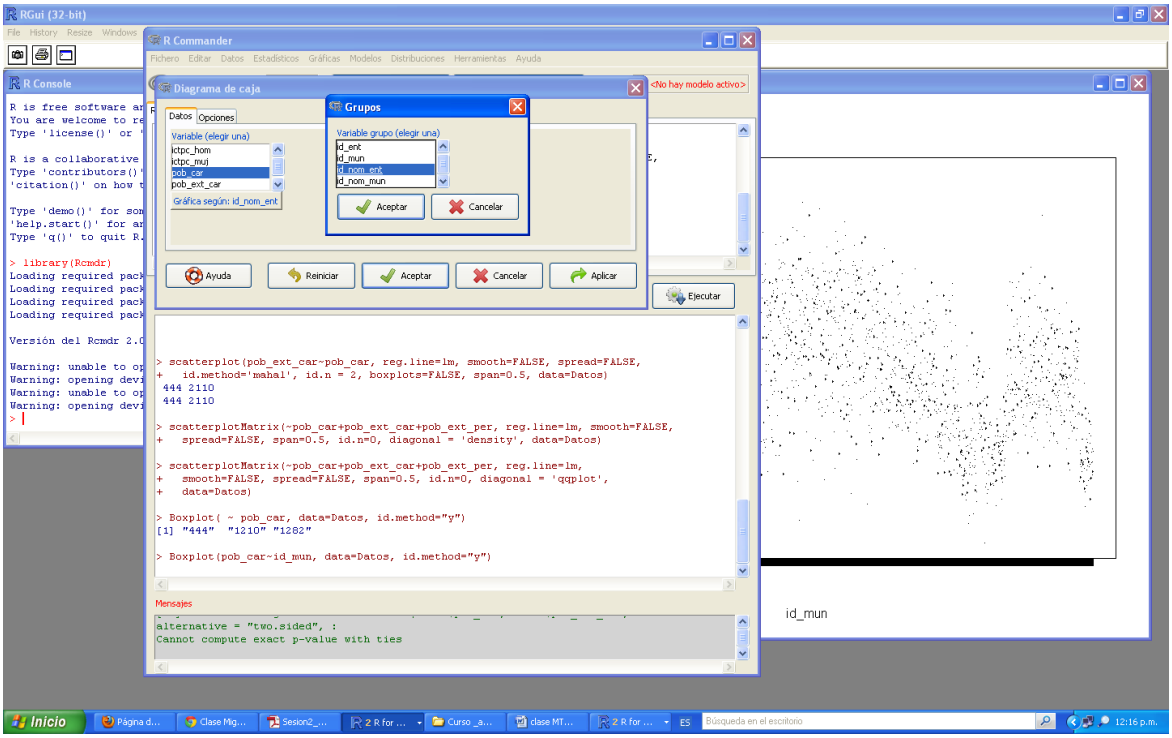
En la siguiente grafica en la diagonal se presenta la gráfica cuantil - cuantil normales (QQ), la cual Un gráfico Cuantil-Cuantil permite observar cuan cerca está la distribución de un conjunto de datos a alguna distribución ideal ó comparar la distribución de dos conjuntos de datos, es decir compara la distribución de la serie con la distribución normal entre más cerca a la normal más significativa.



En el siguiente ejemplo podemos observar que las variables de población por carencia y población extrema por carencia se ajustan a la recta normal lo cual nos indicaría que la correlación es significativa, mientras que la variable de pobreza extrema per no es se ajusta a la normal lo que pone en duda la significancia de la correlación.



Gráficas: Boxplot

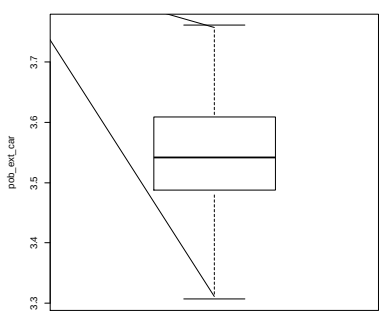


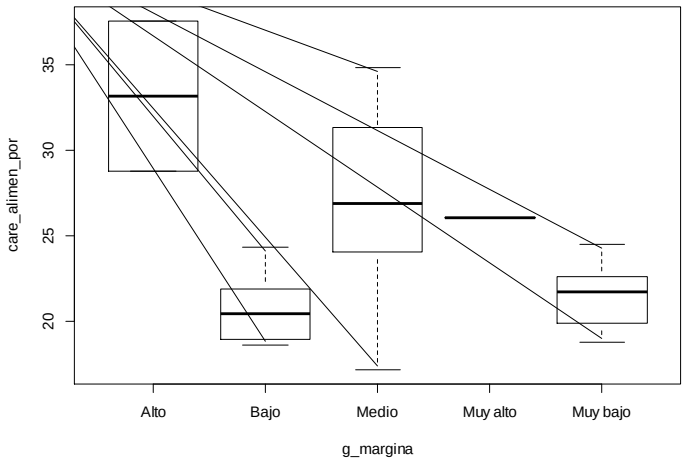
Pobreza extrema por carencias

Los diagramas representan las variables que presentan una gran desviación de la distribución normal. Al centro de la caja se muestra la media, la línea más alta representa el máximo y la línea más baja representa un mínimo, los extremos es por tanto el primer y tercer cuartil.

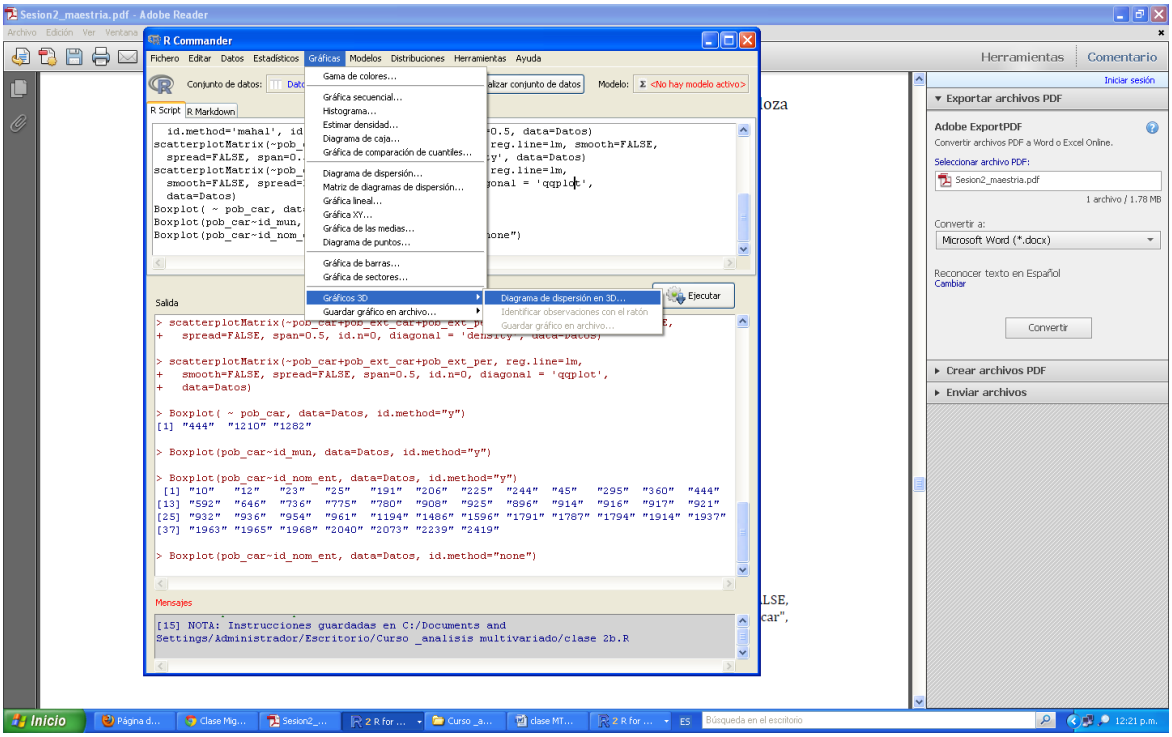
Pobreza por carencias alimentaria condicionada por grado de marginación.

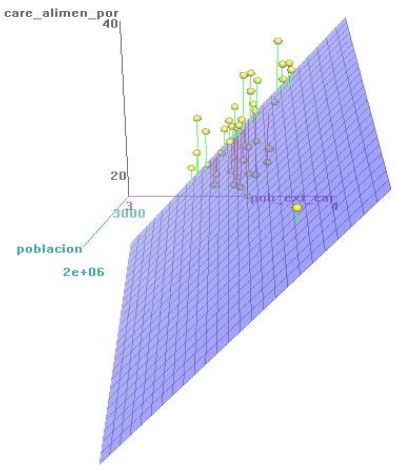
Los diagramas representan las variables que presentan una gran desviación de la distribución normal. Al centro de la caja se muestra la media, la línea más alta representa el máximo y la línea más baja representa un mínimo, los extremos es por tanto el primer y tercer cuartil. Con el análisis comparativo podemos observar que a un grado de marginación media la carencia alimentaria es más dispersa, mientras que a un grado de marginación alta la carencia alimentaria la media es más alta y concentrada.

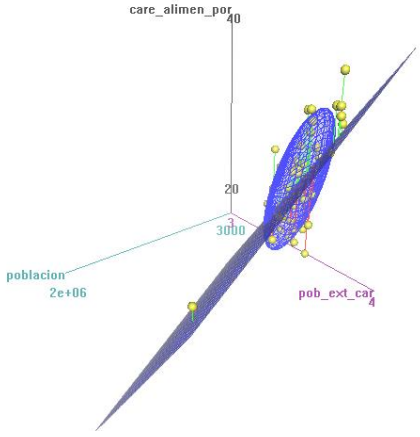




Gráficas: 3D de Pobreza extrema por carencias

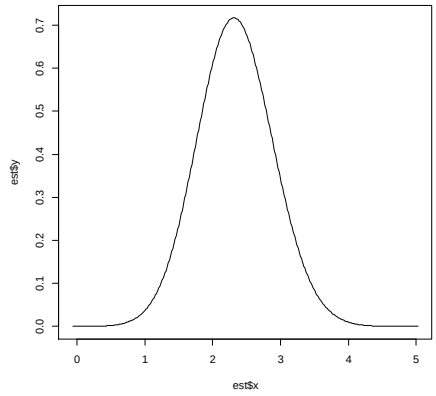






Función Kernell

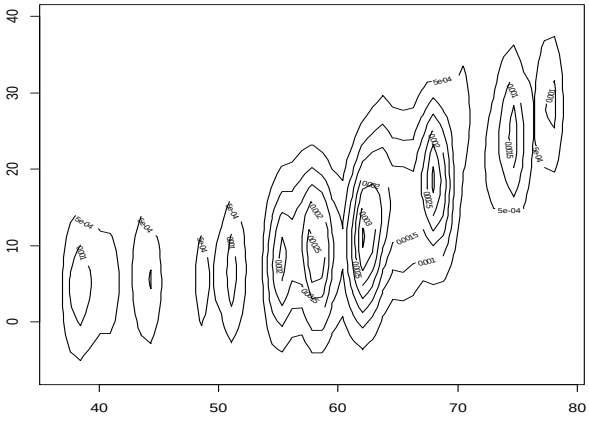
Se mide la densidad, Puede existir dos o más modas, eso nos indica que existen dos o más grupos de concentración. Si analizamos el crecimiento podemos observar que existen diferentes Clubs de convergencias si existen diferentes modas.




```
library(KernSmooth,pos=4)
x<-Datos$pob_car
est<-bkde(x, bandwidth=0.5)
 plot(est,type="l")
```

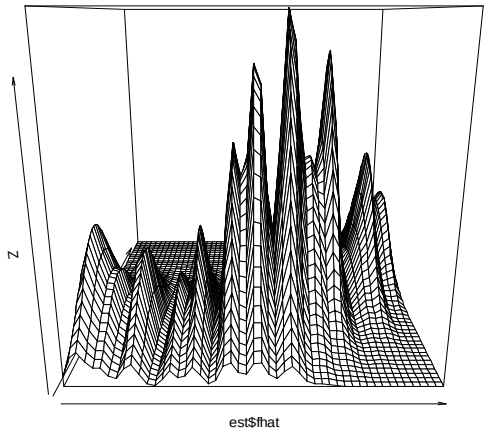
Funciones Kernell contorno

Se mira la función de densidad desde arriba y cada contorno nos muestras diferentes modas en el siguiente ejemplo



Código complementario al anterior

```
x<-cbind(Datos$pob_por,Datos$pob_ext_por)

est<-bkde2D(x,bandwidth=c(0.7,7))

contour(est$x1,est$x2, est$fhat)e
```

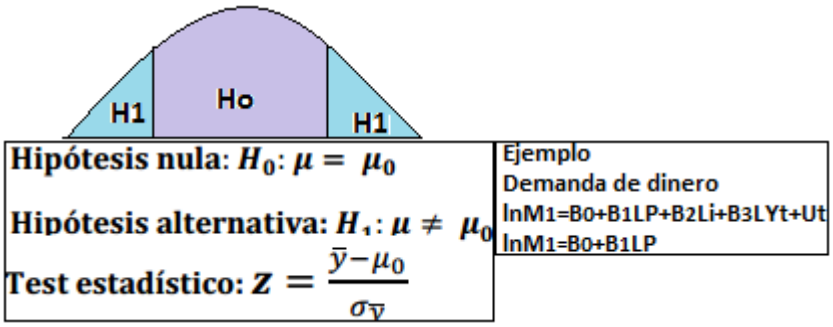



```
x<-cbind(Datos$pob_por,Datos$pob_ext_por)
est<-bkde2D(x,bandwidth=c(0.7,7))
contour(est$x1,est$x2, est$fhat)
persp(est$fhat)
```

2. Comparación de Medias

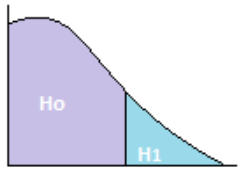
Prueba de hipótesis sobre μ con $\sum$ conocida y es una variable aleatoria que distribuye como N ($\mu, s2$)

1. Univariada



2. Análisis multivalente

Si el análisis es multivariado, la principal ventaja es que analizar la causalidad de dos o más variables y no solo una sola variable.



Ho:tc<tt a>0.05
H_1:tc>tt a<0.05

Ejemplo
Demanda de dinero
lnM1=B0+B1LP+B2Li+B3LYt+Ut
lnM1=B0+B1LP

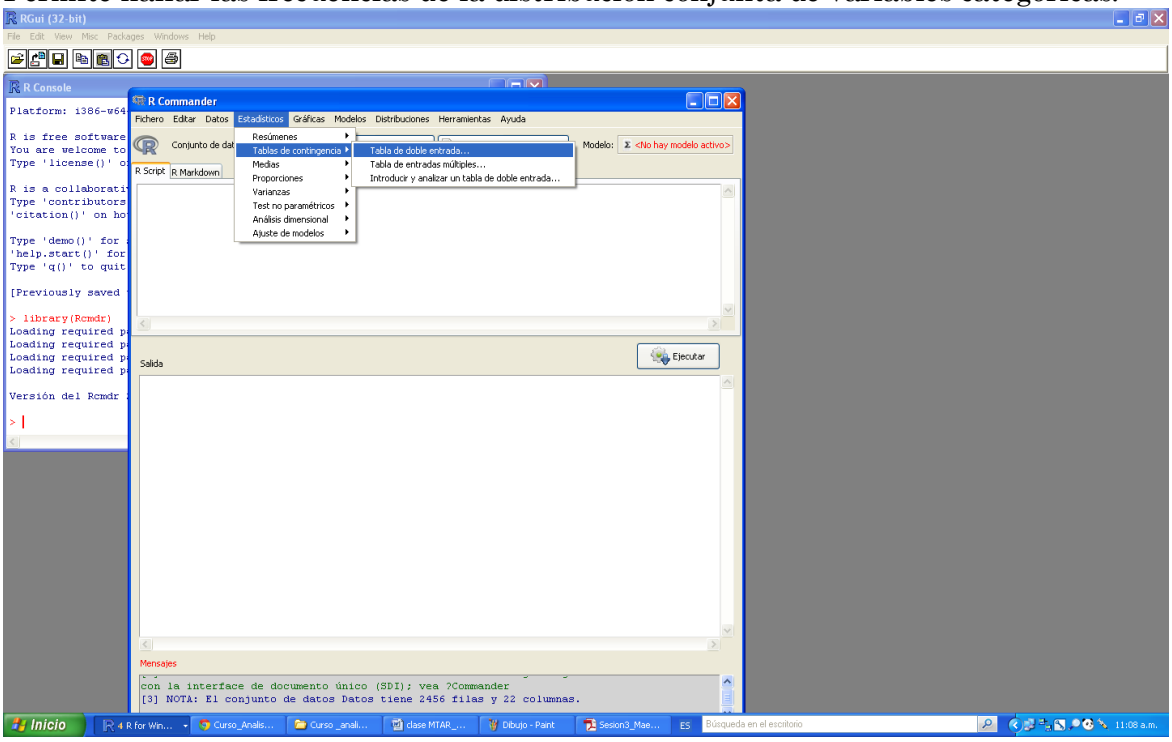
Hipótesis nula: $H_0: \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \mu_{0p} \end{pmatrix}$

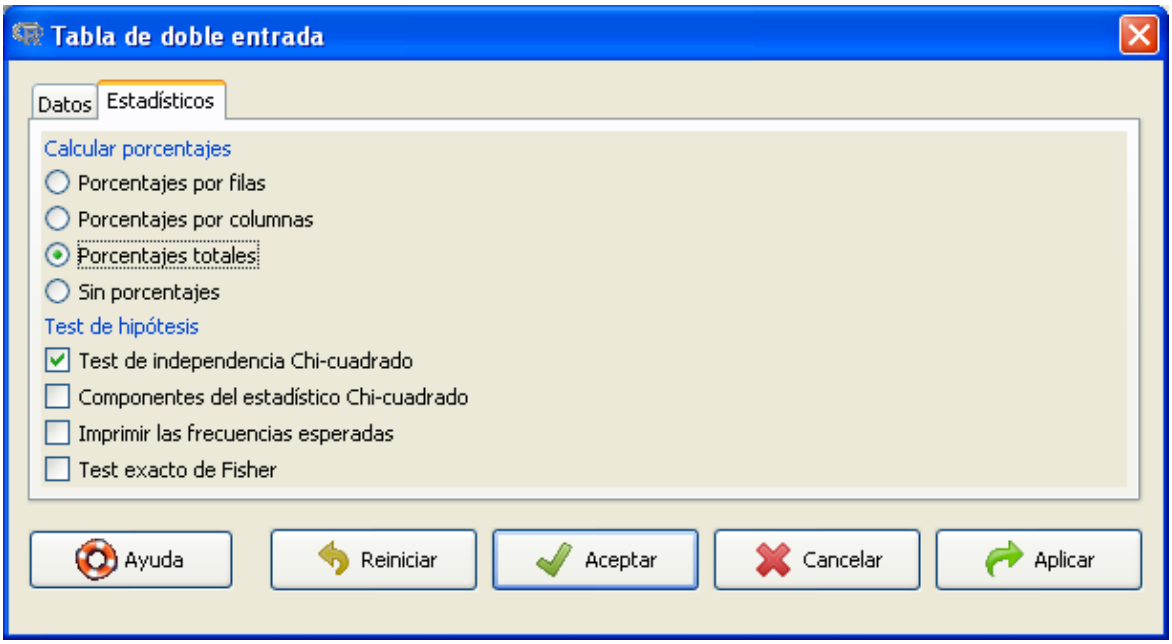
Hipótesis alternativa $H_1: \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \neq \begin{pmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \mu_{0p} \end{pmatrix}$

Test estadístico: $z^2 = n(\bar{y} - \mu_0)' \sum^{-1} (\bar{y} - \mu_0)$

3. Tablas de contingencia

Permite hallar las frecuencias de la distribución conjunta de variables categóricas.





a. Tabla de doble entrada : Variable de fila y variable de columna

Con las siguientes graficas podemos concluir que el 65.8% de la población indígena total tiene un grado de marginación medio en el estado de Guanajuato. Mientras que el 4.3 y 2.2 por ciento de la población indígena tienen un grado de marginación alto y muy alto respectivamente y el 15% y 13% de la población indígena total tiene un grado de marginación bajo y muy bajo respectivamente.

	g_marginal				
CDI	Alto	Bajo	Medio	Muy alto	Muy bajo
Mpio. con población indígena dispersa	2	7	29	1	5
Mpio. con presencia indígena	0	0	1	0	1

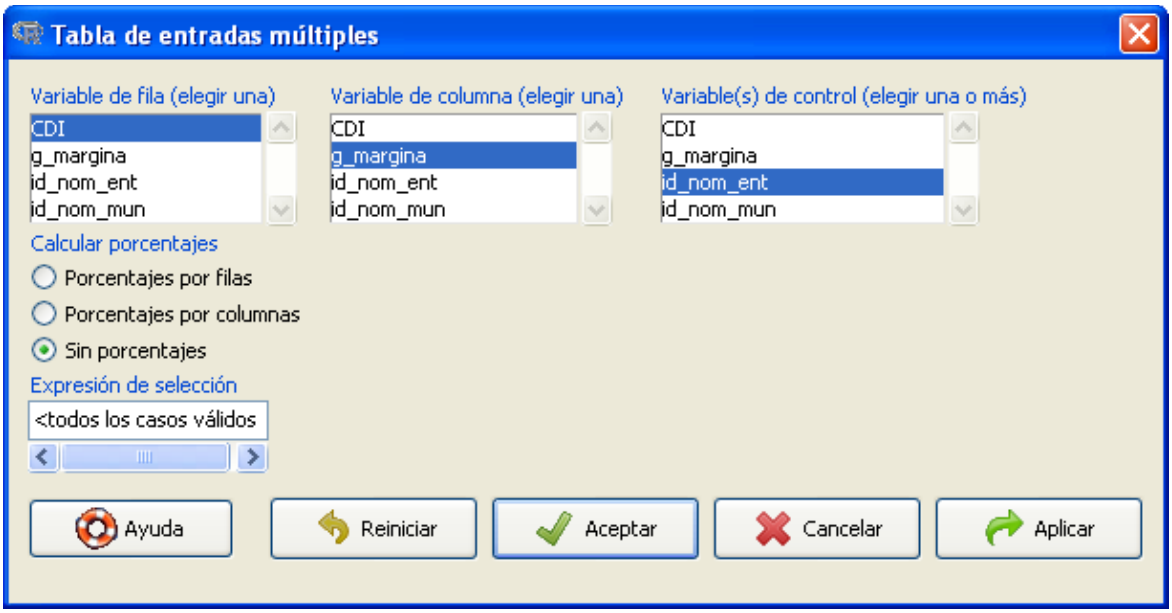
```
> library(abind, pos=4)
> .Table <- xtabs(~CDI+g_margina, data=Datos)
> .Table
```

	Alto	Bajo	Medio	Muy alto	Muy bajo	Total
Mpio. con población indígena dispersa	4.3	15.2	63	2.2	10.9	95.7
Mpio. con presencia indígena	0	0	2.2	0	2.2	4.3
Total	4.3	15.2	65.2	2.2	13	100

```
> totPercents(.Table)
# Percentage of Total
```

```
> .Test <- chisq.test(.Table, correct=FALSE)
> .Test
Pearson's Chi-squared test
data:  .Table
X-squared = 2.7182, df = 4, p-value = 0.606
> remove(.Test)
> remove(.Table)
```

b. Tabla de contingencia (entrada múltiples)

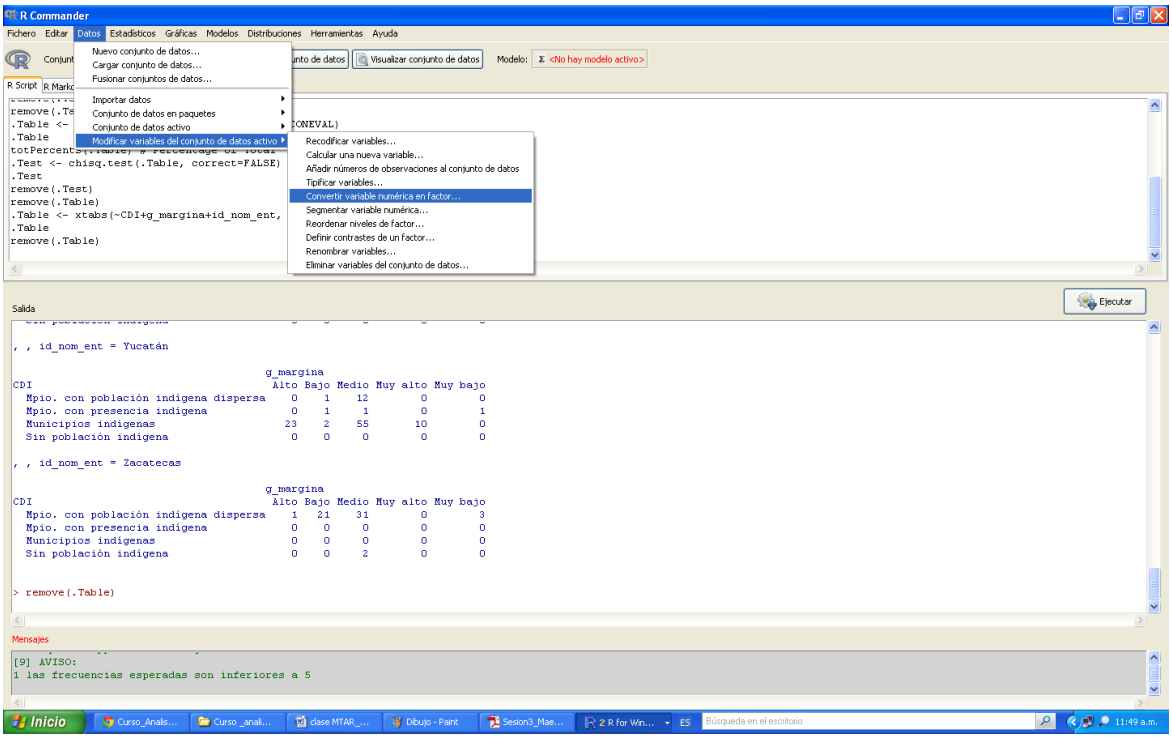


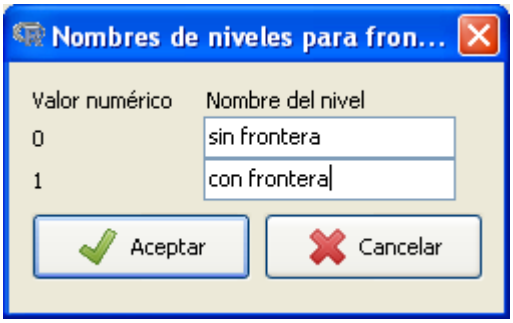
Permite hallar las frecuencias de la distribución conjunta de más de dos variables categóricas. La siguiente tabla nos muestra que existen 29 municipios de población indígena con grados de marginación media, 2 Municipios de población indígena con grados de marginación alta y 1 con población indígena con grado de marginación muy alto, por otro lado existe 7 y 5 municipios con población bajo y muy bajo grado de marginación respectivamente.

	g_margina				
CDI	Alto	Bajo	Medio	Muy alto	Muy bajo
Mpio. con población indígena dispersa	2	7	29	1	5
Mpio. con presencia indígena	0	0	1	0	1

```
> .Table <- xtabs(~CDI+g_margina+id_nom_ent, data=Datos)
> .Table
, , id_nom_ent = Guanajuato
> remove(.Table)
```

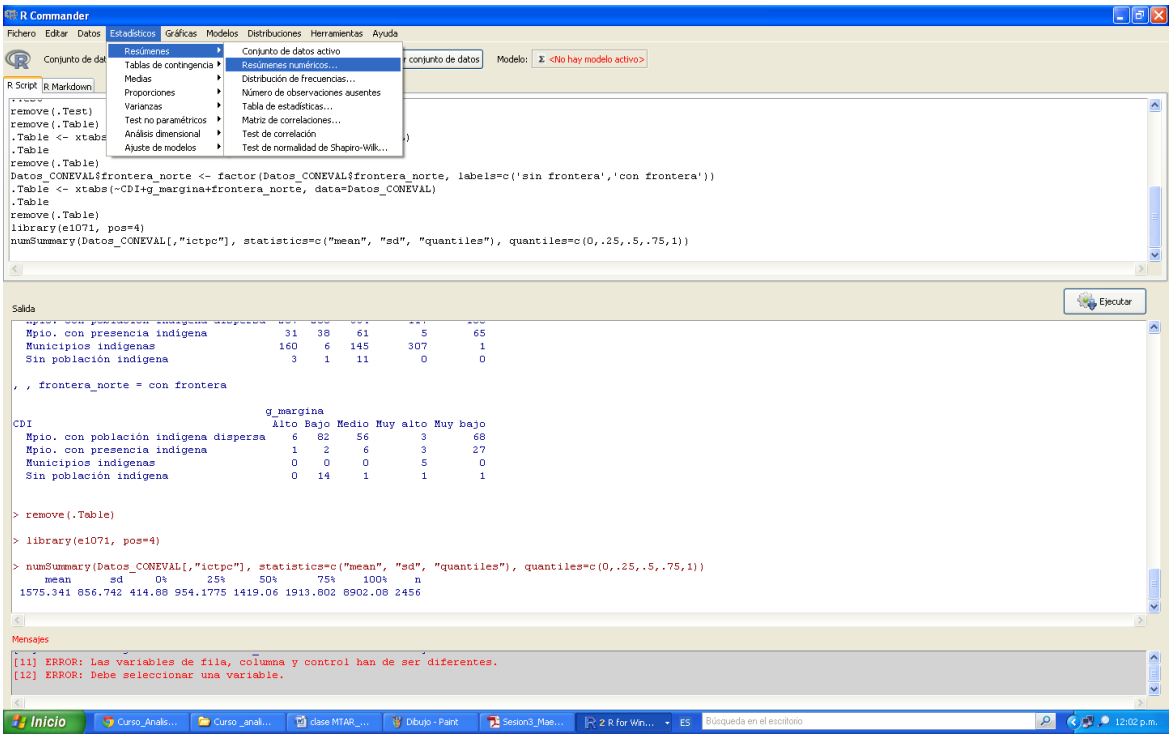
4. Convertir una variable cuantitativa en cualitativa






```
> Datos$frontera_norte <- factor(Datos$frontera_norte, labels=c('con frontera'))
> .Table <- xtabs(~CDI+g_margina+frontera_norte, data=Datos)
> .Table
, , frontera_norte = sin frontera
> remove(.Table)
```

5. Principales estadísticos



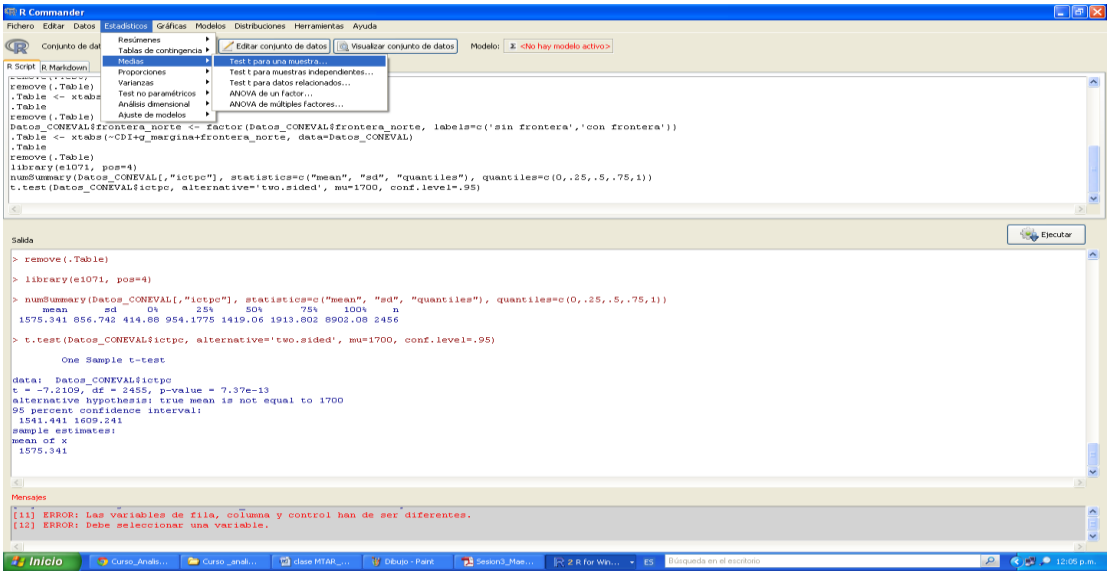
mean	sd	IQR	0%	25%	50%	75%	100%	n
1779.985	587.7988	509.6975	931.21	1421.287	1686.785	1930.985	3450.89	46

```
> library(e1071, pos=4)
> numSummary(Datos[,"ictpc"], statistics=c("mean", "sd", "IQR", "quantiles"),
quantiles=c(0,.25,.5,.75,1))
```

6. Test t para una muestra

Sirve para hacer contrastes de hipótesis para la media de una variable cuantitativa con distribución normal.

- Tipo de contraste –bilateral(igualdad), unilateral (menor o mayor)
- Valor de la media bajo la hipótesis nula
- Nivel de confianza para el intervalo de estimación



El ingreso mensual promedio en el estado de Guanajuato por municipios es de 1573 a precios de 2010.

1. Ho= μ = 1,800 t.test(Datos$ICTPC, alternative=' two.sided', mu=1573, conf.level=.95)
2. Ho= μ > 1,800 t.test(Datos$ICTPC, alternative='greater', mu=1573, conf.level=.95)
3. Ho= μ < 1,800 t.test(Datos$ICTPC, alternative='less', mu=1573, conf.level=.95)

Para la siguiente hipótesis a un nivel de confianza de .95% no se rechaza la hipótesis nula, por lo que la media es igual a 1800.

Ho= μ = 1,800
Ha= $\mu \neq$ 1,800

```
data:  Datos$ictpc
t = -0.2309, df = 45, p-value = 0.8184
alternative hypothesis: true mean is not equal to 1800
95 percent confidence interval:
 1605.430 1954.539
sample estimates:
mean of x
 1779.985
```

No se rechaza la hipótesis nula por lo que la media es mayor a 1800.

Ho= μ > 1,800
Ha= μ <1,800

```
> t.test(Datos$ictpc, alternative='less', mu=1800, conf.level=.95)
        One Sample t-test
data:  Datos$ictpc
t = -0.2309, df = 45, p-value = 0.4092
alternative hypothesis: true mean is less than 1800
95 percent confidence interval:
    -Inf 1925.534
sample estimates:
mean of x
 1779.985
```

Para la siguiente hipótesis a un nivel de confianza de .95% no se rechaza la hipótesis nula, por lo que la media no es mayor que 1800.

Ho= μ < 1,800
Ha= μ > 1,800

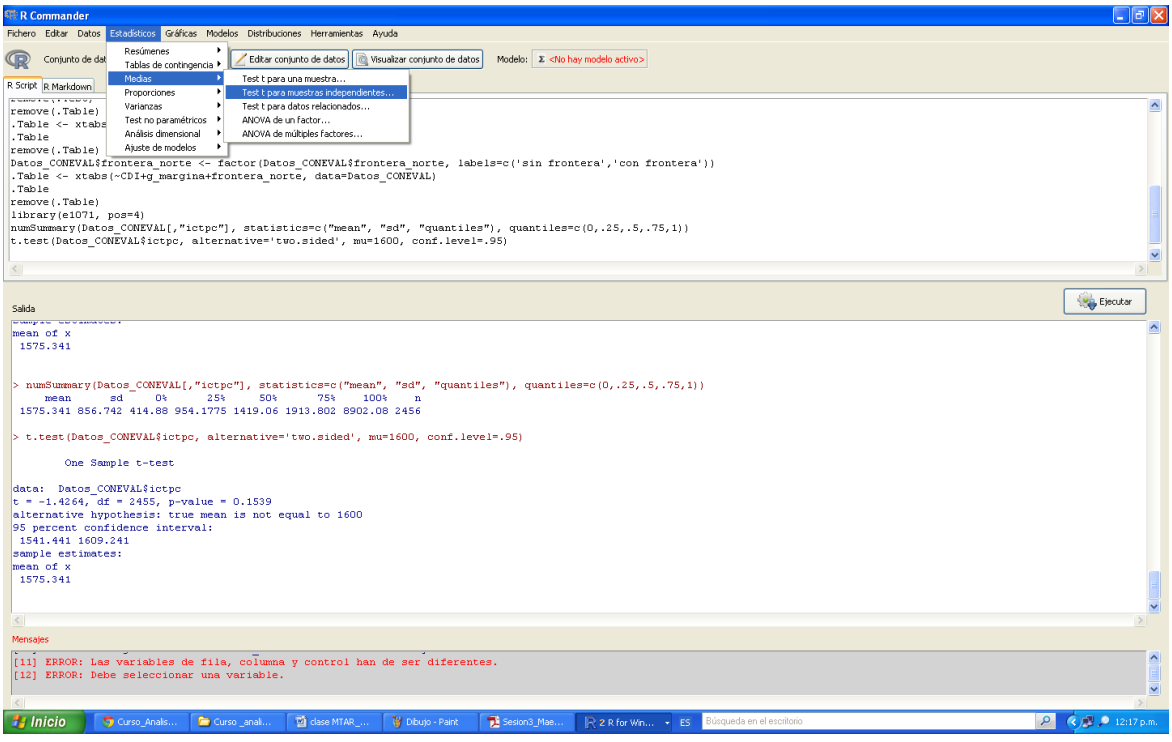
```
> t.test(Datos$ictpc, alternative='greater', mu=1800, conf.level=.95)
        One Sample t-test
data:  Datos$ictpc
t = -0.2309, df = 45, p-value = 0.5908
alternative hypothesis: true mean is greater than 1800
95 percent confidence interval:
 1634.435      Inf
sample estimates:
mean of x
 1779.985
```

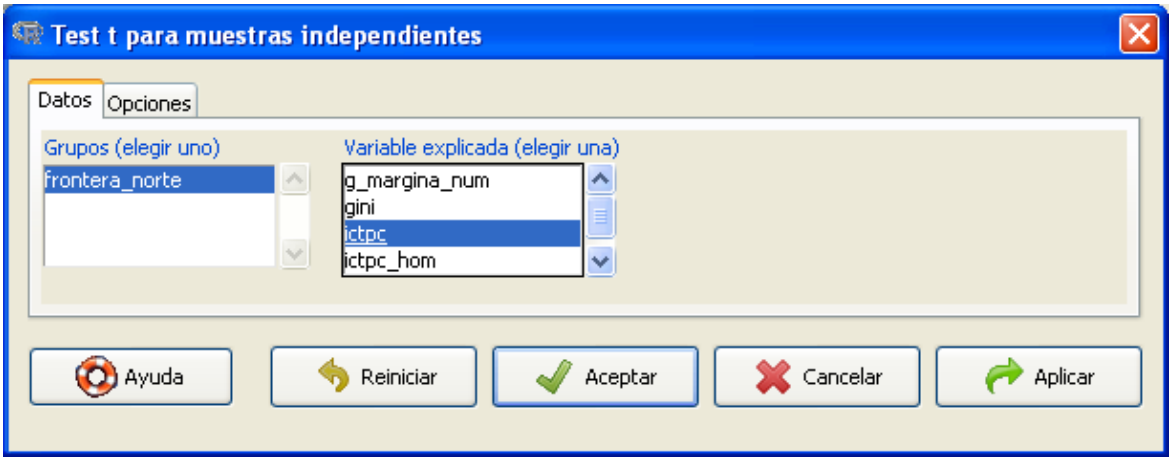
Test t para muestras independientes.

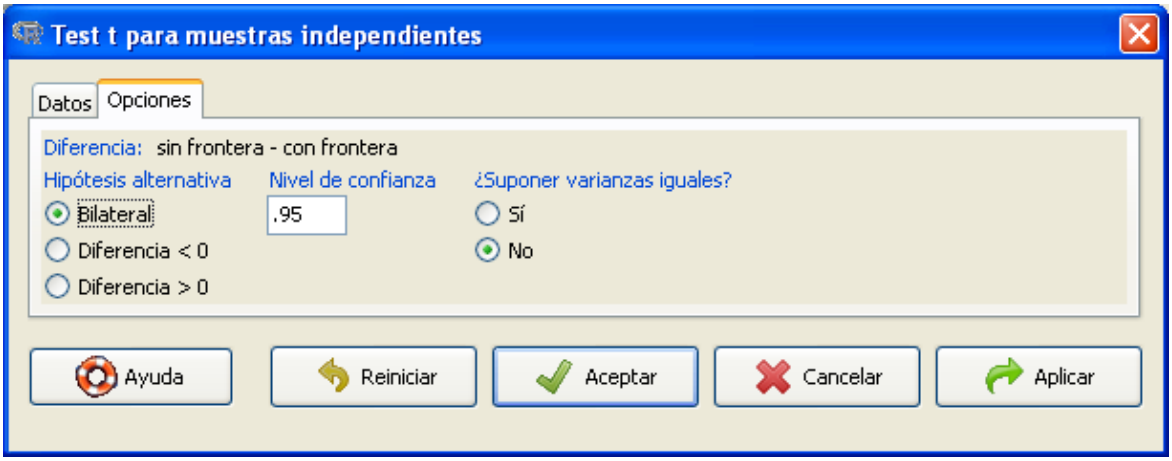
Realiza un contraste de hipótesis de igualdad de medias entre dos muestras. Se necesita por un lado una variable cuantitativa y, por otro, una variable dicotómica (cualitativa o factor, con sólo dos modalidades) que indique los dos grupos (por ejemplo, frontera norte).

Para la hipótesis siguiente

Ho: mpid=mppi
Ha: mpid≠mppi







Los ingresos de la población indígena dispersa y los municipios con presencia indígena no difieren significativamente para el estado de Guanajuato.

```
Welch Two Sample t-test
data:  ictpc by CDI
t = -0.9181, df = 1.02, p-value = 0.5248
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10948.833   9408.773
sample estimates:
mean in group Mpio. con población indígena dispersa       mean in group Mpio. con presencia indígena
                                          1746.505                                          2516.535
```

Si asumimos varianzas iguales se concluye lo mismo, es decir que los ingresos de la población indígena dispersa y los municipios con presencia indígena no difiere significativamente para el estado de Guanajuato.

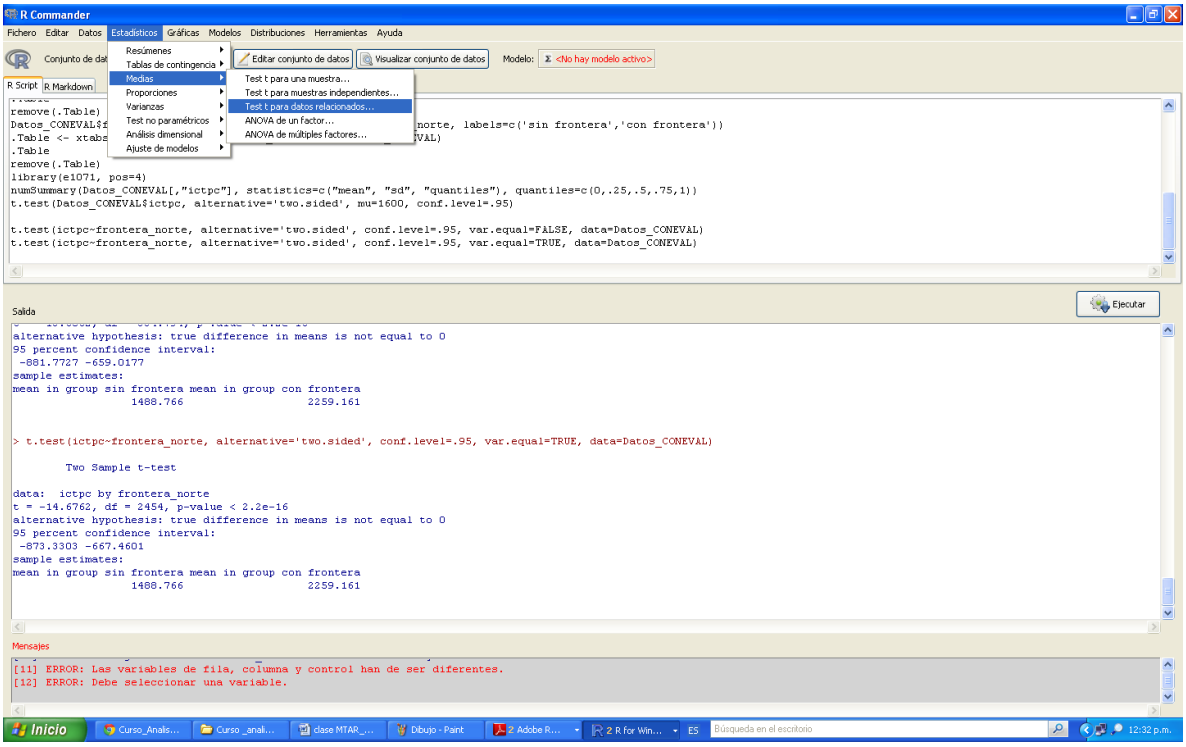
```
Two Sample t-test
data:  ictpc by CDI
t = -1.8609, df = 44, p-value = 0.06946
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1603.99965    63.93965
sample estimates:
mean in group Mpio. con población indígena dispersa       mean in group Mpio. con presencia indígena
                                          1746.505                                          2516.535
```

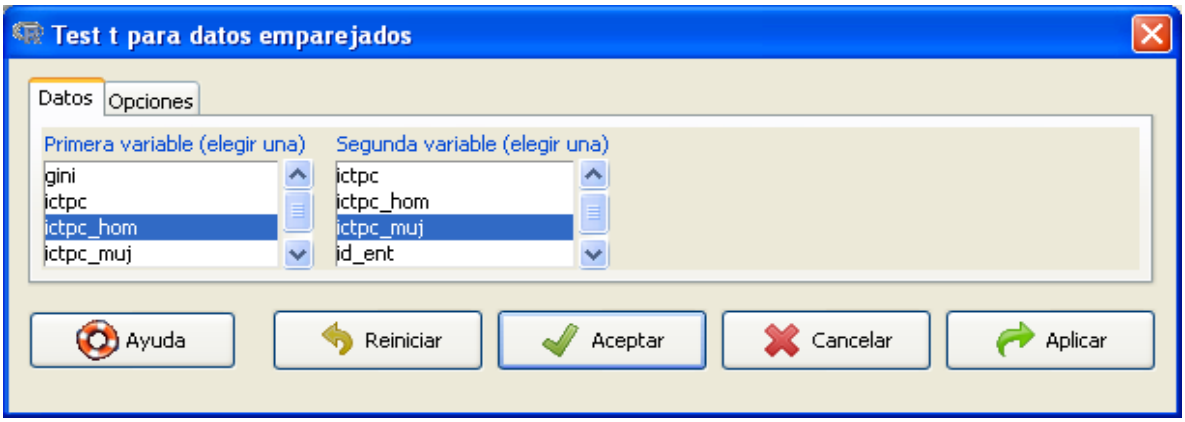
Test t para datos emparejados

Se utiliza para contrastar dos variables en que los datos se suponen relacionados, como al comparar dos variables observadas sobre los mismos individuos.

Ho:mh=mm

Ha: mh≠mm



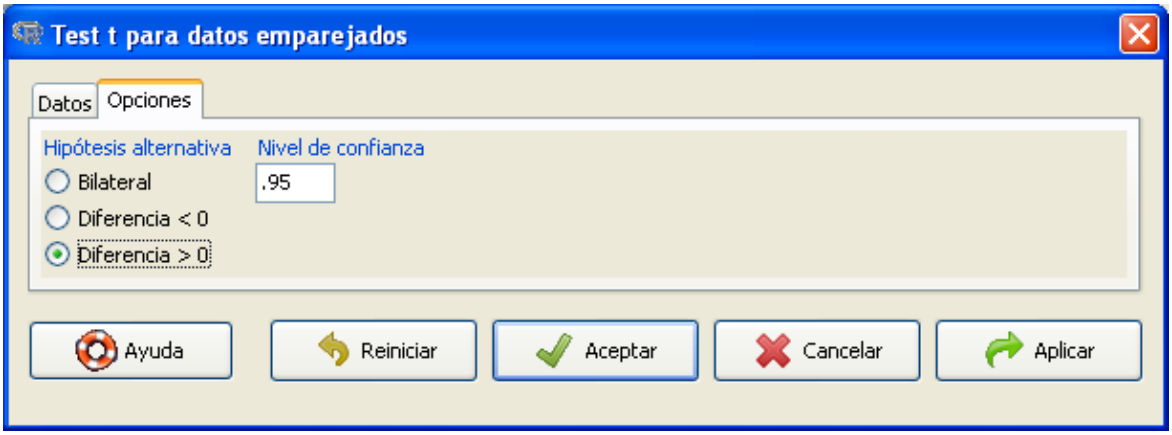


Los ingresos de los hombres y mujeres de los municipios del estado de Guanajuato difieren estadísticamente, se rechaza la hipótesis nula.

```
> t.test(Datos$ictpc_hom, Datos$ictpc_muj, alternative='two.sided', conf.level=.95, paired=TRUE)
        Paired t-test
data:  Datos$ictpc_hom and Datos$ictpc_muj
t = 4.6826, df = 45, p-value = 2.626e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 19.41070 48.71202
sample estimates:
mean of the differences
        34.06136
```

Para saber si los ingresos de los hombres son mayores que los de las mujeres, se realiza la siguiente hipótesis:

Ho:mh=mm
Ha:mh>mm



Con base en la siguiente información podemos afirmar que los ingresos de los hombres son mayores a los de la mujeres, al aceptar la hipotesis alternativa a un nivel de confianza del 95%.

```
>t.test(Datos$ictpc_hom, Datos$ictpc_muj, alternative='greater', conf.level=.95, paired=TRUE)
        Paired t-test
data:  Datos$ictpc_hom and Datos$ictpc_muj
t = 4.6826, df = 45, p-value = 1.313e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 21.84515      Inf
sample estimates:
mean of the differences
        34.06136
```

Análisis ANOVA Guanajuato

- **Análisis de Varianza de un Factor**

ANOVA de un factor permite contrastar la igualdad de medias en varias muestras independientes, mediante un análisis de varianzas

Ho: α_1=0
Ha: μ_1= μ_2=μ_3
Ho: α_1≠0:
Ha: μ_1≠μ_(2;) μ_1≠μ_3; μ_2≠μ_3; los ingresos son diferentes en términos de ingresos promedios
Ho: Si Fc < Ft zona de hipótesis nula (las medias son iguales); la probabilidad asociada es mayor a 0.05
Ha: Si Fc > Ft Hipótesis alternativas (se acepta que por lo menos hay dos medias diferentes); si la probabilidad asociada es menor a 0.05

En el siguiente ejemplo, la probabilidad es menor a 0.05, lo que nos indica que no se rechaza la hipótesis nula, por lo que las medias son iguales. En el ingreso mensual influye el nivel de marginación del municipio.

	Df	Sum Sq	Mean Sq	F value Pr(>F)
g_margina	4	1.35E+09	337988897	1841 <2e-16 ***
Residuals	2451	4.50E+08	183611	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

```
> library(multcomp, pos=4)
> library(abind, pos=4)
> AnovaModel.1 <- aov(ictpc ~ g_margina, data=coneval2010gto)
> summary(AnovaModel.1)
```

	mean	sd	data:n
Alto	1007.3791	222.4519	408
Bajo	2141.3016	442.1751	401
Medio	1484.0722	334.6874	944
Muy alto	752.7635	168.4615	441
Muy bajo	3306.9938	945.9848	262

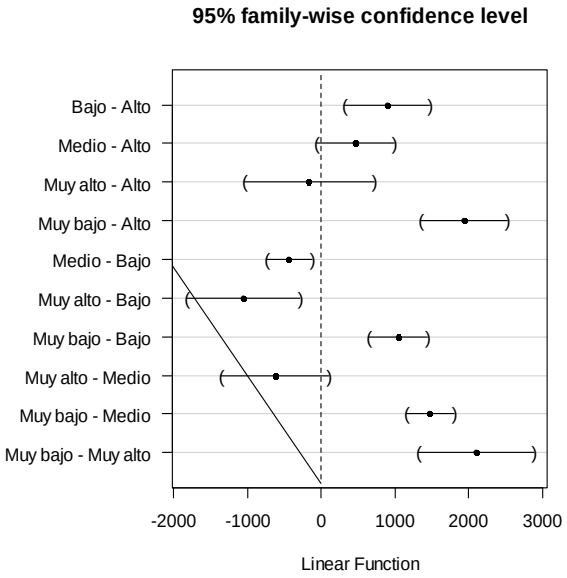
```
> numSummary(coneval2010gto$ictpc , groups=coneval2010gto$g_margina, statistics=c("mean", "sd"))
```

Método de Turkey (1953) conocida como diferencia significativa de Turkey. Todas las comparaciones son referidas a una misma diferencia mínima.

Simultaneous Tests for General Linear Hypotheses				
Multiple Comparisons of Means: Tukey Contrasts				
Fit: aov(formula = ictpc ~ g_margina, data = coneval2010gto)				
Linear Hypotheses:				
	Estimate	Std. Error	t value	Pr(>|t|)
Bajo - Alto == 0	890.9	206.4	4.317	< 0.001 ***
Medio - Alto == 0	458	188	2.436	0.11236
Muy alto - Alto == 0	-166.2	315.3	-0.527	0.98171
Muy bajo - Alto == 0	1931.6	210.2	9.19	< 0.001 ***
Medio - Bajo == 0	-433	108.1	-4.007	0.00194 **
Muy alto - Bajo == 0	-1057.1	275.2	-3.841	0.00319 **
Muy bajo - Bajo == 0	1040.7	143.2	7.267	< 0.001 ***
Muy alto - Medio == 0	-624.1	261.7	-2.385	0.12463

Simultaneous Confidence Intervals				
Multiple Comparisons of Means: Tukey Contrasts				
Fit: aov(formula = ictpc ~ g_margina, data = Datosgtoconeval)				
Quantile = 2.7956				
95% family-wise confidence level				
Linear Hypotheses:				
Estimate	lwr		upr	
Bajo - Alto == 0	890.9207		313.916	1467.9255
Medio - Alto == 0	457.9507		-67.6084	983.5097
Muy alto - Alto == 0	-166.195		-1047.5843	715.1943
Muy bajo - Alto == 0	1931.6483		1344.0555	2519.2412
Medio - Bajo == 0	-432.97		-735.0441	-130.896
Muy alto - Bajo == 0	-1057.1157		-1826.4554	-287.7761
Muy bajo - Bajo == 0	1040.7276		640.3506	1441.1047
Muy alto - Medio == 0	-624.1457		-1355.6929	107.4015
Muy bajo - Medio == 0	1473.6977		1151.8598	1795.5355
Muy bajo - Muy alto == 0	2097.8433		1320.531	2875.1556
Muy bajo - Medio == 0	1473.7	115.1	12.801	< 0.001 ***
Muy bajo - Muy alto == 0	2097.8	278.1	7.545	< 0.001 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
(Adjusted p values reported -- single-step method)				

```
> .Pairs <- glht(AnovaModel.2, linfct = mcp(g_margina = "Tukey"))
> summary(.Pairs) # pairwise tests
> confint(.Pairs) # confidence intervals
```




```
> cld(.Pairs) # compact letter display
  Alto    Bajo   Medio Muy alto Muy bajo
   "a"     "b"     "a"     "a"      "c"
```

- **Test F para 2 varianzas**

Para comparar medias de muestras independientes, depende de la Homoscedasticidad (igualdad de varianzas) en las dos muestras.

Ho: Homoscedasticas (varianzas iguales). Razón de varianzas = 1

Ha: Heteroscedasticas (varianzas diferentes). Razón de varianzas es diferente de 1

Los resultados a continuación consideran todos los municipios y todos los estados para poder aplicar los resultados de la frontera norte. En este caso la probabilidad es menor a 0.05 y la ratio de la varianza es deferente de 1, esto nos indica que existe heteroscedasticidad.

```
> coneval2010gto$frontera_norte <- factor (coneval2010gto$frontera_norte,
+   labels=c('sin frontera','con frontera'))
> tapply(coneval2010gto$ictpc, coneval2010$frontera_norte,  var, na.rm=TRUE)
sin frontera con frontera
   659112.1     801387.0
> var.test(ictpc ~ frontera_norte, alternative='two.sided', conf.level=.95, data=coneval2010gto)
            F test to compare two variances
data:  ictpc by frontera_norte
F = 0.8225, num df = 2179, denom df = 275, p-value = 0.02525
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6843215 0.9763135
sample estimates:
ratio of variances
         0.8224642
```

Test de Bartlett

Para efectos de análisis y nuevamente como Guanajuato no tiene frontera se utilizó el total nacional, En este caso la probabilidad es menor a 0.05 la hipótesis Ho (Hipótesis nula) se rechazan esto nos muestra que las varianzas no son iguales.

```
> tapply(coneval2010$g_margina_num, coneval2010gto$frontera_norte, var, na.rm=TRUE)
sin frontera con frontera
   1.333358    1.068920
> bartlett.test(g_margina_num ~ frontera_norte, data=coneval2010gto)
        Bartlett test of homogeneity of variances
data:  g_margina_num by frontera_norte
Bartlett's K-squared = 5.6278, df = 1, p-value = 0.01768
```

Test de Levene.

El contraste de Levene es para el análisis de varianza Homoscedasticidad entre las muestras independientes centrando en la media.

```
> leveneTest(Datosgtoconeval$ictpc, Datosgtoconeval$g_margina, center=median)
Levene's Test for Homogeneity of Variance (center = median)

  Df   F value      Pr(>F)
group  4  3.3479    0.01842 *
       41
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA de más de un factor

Permite contrastar la igualdad de medias en varias muestras independientes, para cada uno de los factores y la misma cantidad de individuos.

Factor 1 **Ho** = $\mu_{1,1} = \mu_{1,2} = \mu_{1,3} = \cdots, \mu_{1,k}$ $\boldsymbol{donde\ k = 1, 2, 2, \ldots, muestras}$

Factor 2 **Ho** = $\mu_{2,1} = \mu_{2,2} = \mu_{2,3} = \cdots, \mu_{2,k}$ $\boldsymbol{donde\ k = 1, 2, 2, \ldots, muestras}$

.

.

.

Factor n **Ho** = $\mu_{n,1} = \mu_{n,2} = \mu_{n,3} = \cdots, \mu_{n,k}$ $\boldsymbol{donde\ k = 1, 2, 2, \ldots, muestras}$

Los resultados del análisis ANOVA para probar si los grupos de la frontera norte son importantes, nos indican que los municipios que definen la frontera norte y lo que definen los grados de marginación son importantes para diferencial el ingreso, debido a que se rechaza la hipótesis nula, lo que indica que existen diferentes medias.

Anova Table (Type II tests)				
Response: ictpc				
	Sum Sq	Df	F value	Pr(>F)
frontera_norte	842545	1	4.6279	0.0315534 *
g_margina	1207398154	4	1657.9951	< 2.2e-16 ***
frontera_norte:g_margina	3877207 4	5.3242	0.0002873 ***	
Residuals	445311317	2446		

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**				

```
> AnovaModel.5 <- (lm(ictpc ~ frontera_norte*g_margina, data=coneval2010gto))
> Anova(AnovaModel.5)
```

	g_margina				
frontera_norte	Alto	Bajo	Medio	Muy alto	Muy bajo
sin frontera	1003.418	2164.037	1478.21	751.0379	3391.19
con frontera	1234.271	2071.008	1566.054	814.4567	3161.405

```
> tapply(coneval2010gto$ictpc, list(frontera_norte=coneval2010gto$frontera_norte,
g_margina=coneval2010gto$g_margina), mean, na.rm=TRUE) # means
```

	g_margina				
frontera_norte	Alto	Bajo	Medio	Muy alto	Muy bajo
sin frontera	215.2847	455.2709	340.0371	169.2392	1027.1097
con frontera	453.6292	392.8573	235.0997	128.348	769.7335

```
> tapply(coneval2010gto$ictpc, list(frontera_norte=coneval2010gto$frontera_norte,
g_margina=coneval2010gto$g_margina), sd, na.rm=TRUE) # std. deviations
```

	g_margina				
frontera_norte	Alto	Bajo	Medio	Muy alto	Muy bajo
sin frontera	401	303	881	429	166
con frontera	7	98	63	12	96

```
> tapply(coneval2010gto$ictpc, list(frontera_norte=coneval2010gto$frontera_norte,
g_margina=coneval2010gto$g_margina), function(x) sum(!is.na(x))) # counts
```

Análisis de componentes principales

Este análisis nos permite construir índices, lo cual nos va a permitir generar una metodología. **El primer componente principal es la combinación lineal con varianza máxima, esencialmente estamos buscando una dimensión en la que las observaciones son máximamente separados o extendidas**. **El segundo componente principal es la combinación lineal con varianza máxima en una dirección ortogonal al primer componente principal, y así sucesivamente.** En general, los componentes principales definen diferentes dimensiones de los definidos por las funciones discriminantes o variables canónicas.

En algunas aplicaciones, los componentes principales son un fin en sí mismos y pueden ser susceptibles de interpretación. Más a menudo que se obtienen para su uso como entrada para otro análisis. Por ejemplo, **dos situaciones en regresión donde los componentes principales pueden ser útiles son (**1) si el número de variables independientes es grande en relación con el número de observaciones, una prueba puede ser ineficaz o incluso imposible, y (2) si las variables son independientes altamente correlacionada, las estimaciones de los coeficientes de represión pueden ser inestables. **En tales casos, las variables independientes se pueden reducir a un número menor de componentes principales que darán una mejor prueba o estimaciones más estables de los coeficientes de regresión (Rencher, 2002).**

- Metodología para la construcción de índices

Etapas para la construcción de índices:

1. Construcción de n indicadores
2. Estandarización de los indicadores por el método Gaussiano o de normalización: menos la media entre la desviación estándar. Valores positivos (negativo) indica una posición arriba (abajo) de la media; (normalizar las series) positivo por arriba de la media, valor negativo por debajo de la media.

$$z = \frac{x_i - \bar{x_i}}{\sigma x_i}$$

3. Aplicación del método de componentes principales para reducir la escala de n indicadores a un solo índice;
4. Estandarización de índice por el método Gaussiano o de normalización
5. Estratificación del índice por un método de clasificación con el de Dalenius Hodges, k-medias, etc.

- Método de estandarización de las variables

Para la construcción de Índice se identifican como IDij a las n variables para municipios del país; donde i idéntica a la variable en los j municipios. Para eliminar los efectos de escala entre las IDij se estandarizan mediante el método Gaussiano; esto es, la variable menos su media entre la desviación estándar:

$$Z_{ij} = \frac{ID_{ij} - \overline{ID_i}}{ds_i}$$

Donde:

Zij= Indicador estandarizado para la variable i para la unidad de observación j por municipios.
IDij =Variable
$\overline{IDij}$ = Media de la variable
dsi= Desviación estándar de la variable IDij

- Método de componentes principales

El método de componentes principales se utiliza para combinar la n indicadores estandarizados ZIJ para obtener el índice. La técnica consiste en transformar el espacio de los vectores Zij como j entradas dependiendo si los datos para los municipios, para entonces encontrar

Esto se puede escribir con vectores característicos, porque eso implica que son linealmente independientes, es decir no están correlacionados.

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a11} \dots & \boldsymbol{a1n} \\ \boldsymbol{am1} \dots & \boldsymbol{amn} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\lambda 1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\lambda 2} \end{bmatrix}$$

Componente 1

$$\begin{bmatrix} \boldsymbol{ID1} \\ \boldsymbol{ID2} \\ \boldsymbol{ID3} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda 1} / \sum \boldsymbol{Xi} \\ \boldsymbol{\lambda 2} / \sum \boldsymbol{Xi} \\ \boldsymbol{\lambda 3} / \sum \boldsymbol{Xi} \end{bmatrix}$$

Aplicación del Índice de Marginación para el Estado de Guanajuato

```
# Estandarizar la matriz, la cual se llama matriz Z
Z<-scale(conapo2010gto[,c("ind1","ind2","ind3","ind4","ind5","ind6","ind7","ind8","ind9")])
Z
```

	ind1	ind2	ind3	ind4	ind5	ind6	ind7	ind8	ind9
1	0.06	0.54	0.10	-0.38	-0.23	0.77	-0.48	0.34	1.04
2	-0.19	-0.34	-0.73	-0.55	-0.45	-1.23	-1.00	-0.65	0.20
3	0.11	-0.25	0.78	0.34	0.38	0.33	0.60	-0.08	-0.83
4	-0.03	0.05	-0.14	-0.18	-0.47	0.30	-0.07	-0.07	-0.07
5	-0.35	-0.69	-0.58	-0.41	-0.01	0.23	-0.36	0.19	-0.37
6	2.22	1.49	1.70	2.72	4.67	0.62	0.96	1.58	0.99
7	-1.64	-2.18	-1.06	-0.62	-0.42	-1.59	0.51	-1.46	-1.54
8	0.44	1.04	0.34	0.10	-0.29	-0.07	-0.65	0.15	0.65
9	0.24	-0.08	0.58	0.18	0.67	1.18	2.39	-0.23	-0.15

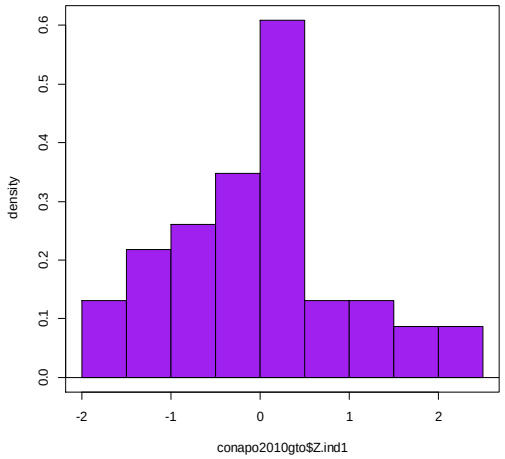
10	0.22	0.57	-0.08	-0.04	-0.51	-0.53	-0.95	1.58	0.82
11	-0.98	-1.07	-0.80	-0.63	-0.47	-0.55	1.14	-1.08	-0.91
12	0.06	0.03	-0.47	-0.19	-0.52	-0.19	-0.05	-0.37	0.76
13	1.25	0.47	0.64	0.06	-0.64	1.42	-0.46	0.74	0.62
14	0.08	0.21	0.73	0.52	-0.11	0.91	1.58	0.06	0.24
15	-1.70	-2.12	-0.55	-0.27	0.43	-1.02	-0.88	-1.02	-1.90
16	-0.63	0.42	-0.65	-0.53	-0.63	-0.02	-1.16	0.54	1.03
17	-1.42	-1.88	-0.85	-0.70	-0.47	-0.81	-0.45	-1.30	-1.49
18	-0.58	-0.74	-0.85	-0.58	-0.68	-0.48	-0.75	-0.55	0.09
19	1.56	1.36	1.49	0.18	0.62	0.16	-0.86	1.00	2.00
20	-1.73	-1.98	-1.08	-0.69	0.08	-1.25	-0.78	-1.92	-2.12
21	-1.11	-0.77	-0.97	-0.86	-0.60	-2.24	-1.32	-1.76	-1.08
22	0.49	1.24	0.46	1.01	0.13	0.45	-0.54	0.49	1.09
23	-0.01	0.50	0.20	-0.22	-0.07	-0.09	-0.25	0.31	0.30
24	0.40	0.58	-0.41	-0.59	-0.60	-0.57	-1.20	1.58	0.62
25	-1.22	-0.42	-0.98	-0.54	-0.57	1.06	-0.87	-0.83	-1.90
26	-0.07	0.71	-0.31	-0.54	-0.49	0.87	0.08	0.16	0.69
27	-1.13	-1.69	-0.86	-0.78	-0.45	-1.54	-0.81	-1.11	-1.21
28	0.10	0.19	-0.76	-0.56	-0.44	-0.84	0.09	-0.50	0.47
29	0.88	1.20	2.10	0.13	-0.38	0.94	0.78	0.85	1.27
30	0.46	0.65	1.39	0.35	1.03	1.38	0.23	0.36	0.53
31	-0.92	-0.69	-0.61	-0.75	-0.43	-0.27	-0.85	-0.81	-1.77
32	-0.43	-0.92	-0.20	-0.22	-0.51	0.43	0.18	0.35	-1.19
33	0.33	0.05	0.49	0.30	0.06	1.12	1.77	-0.28	0.05
34	1.90	0.77	0.74	2.85	0.08	0.19	-0.13	1.58	0.33
35	0.18	0.28	-0.30	-0.34	-0.30	0.74	1.87	-0.77	0.85
36	1.28	1.29	-0.96	-0.83	-0.68	-1.72	-0.71	1.58	0.83
37	-0.76	-0.80	-0.55	-0.49	0.40	0.93	-0.38	-0.18	-0.66
38	-0.32	0.03	-0.61	-0.03	-0.46	-0.87	-0.88	-0.40	0.92
39	0.55	0.97	-0.65	-0.48	-0.49	-0.91	0.20	0.28	0.37
40	1.27	0.68	2.66	2.41	0.40	2.71	2.27	1.58	0.33
41	-1.02	-0.15	-0.98	-0.80	-0.60	-1.09	-1.34	-1.72	-0.98
42	-0.35	0.13	-0.26	-0.63	-0.52	-0.15	-0.25	-0.26	0.26
43	0.72	-0.05	1.83	1.11	1.75	0.88	1.43	1.58	-0.09
44	-0.98	-1.37	-1.02	-0.55	-0.21	-0.21	1.11	-1.03	-0.62
45	2.48	1.73	2.39	3.37	3.26	1.03	1.43	1.58	1.34
46	0.30	0.98	-0.33	-0.67	-0.25	-0.39	-0.19	-0.13	0.22
attr(,"scaled:center")									
	ind1	ind2	ind3	ind4	ind5	ind6	ind7	ind8	ind9
11.6	30.90	11.39	2.71	6.36	40.07	4.73	58.49	53.55	
attr(,"scaled:scale")									
	ind1	ind2	ind3	ind4	ind5	ind6	ind7	ind8	ind9
	3.76	6.60	9.60	2.80	8.88	6.63	2.42	26.20	13.00

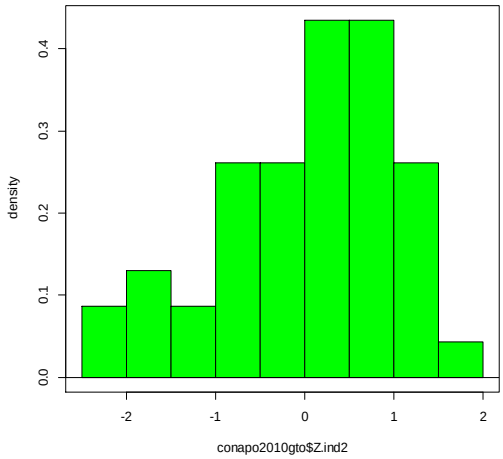
#Incorporar las variables a la base

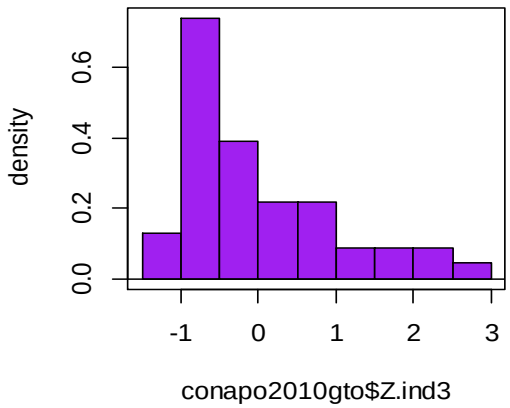
```
conapo2010gto$Z.ind1 <-Z[,1]
conapo2010gto$Z.ind2 <-Z[,2]
conapo2010gto$Z.ind3 <-Z[,3]
conapo2010gto$Z.ind4 <-Z[,4]
conapo2010gto$Z.ind5 <-Z[,5]
conapo2010gto$Z.ind6 <-Z[,6]
conapo2010gto$Z.ind7 <-Z[,7]
conapo2010gto$Z.ind8 <-Z[,8]
conapo2010gto$Z.ind9 <-Z[,9]
```

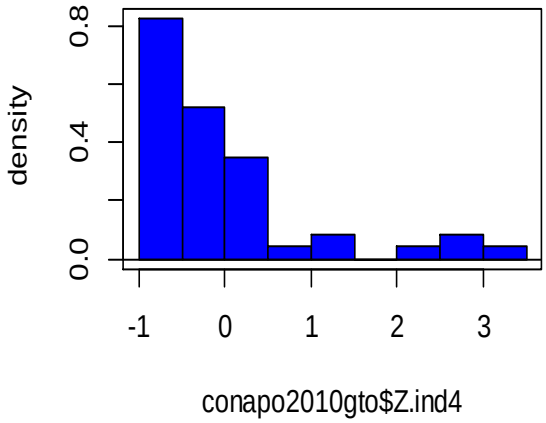
 # Histograma de cada indicador

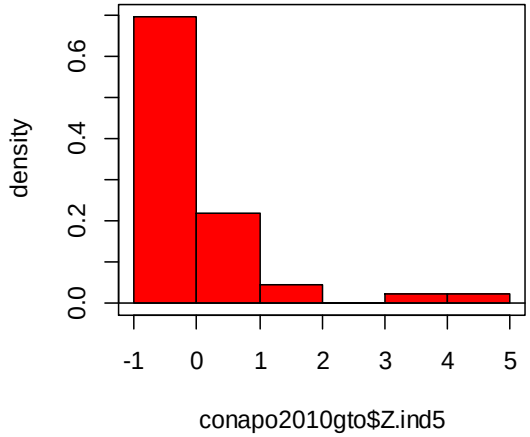
```
Hist(conapo2010gto$Z.ind1, scale="density",  breaks="Sturges", col="purple")
```

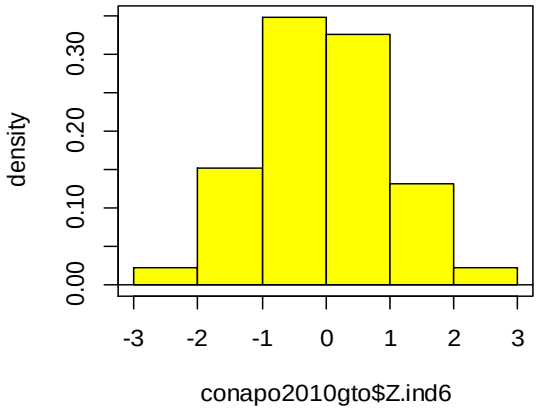


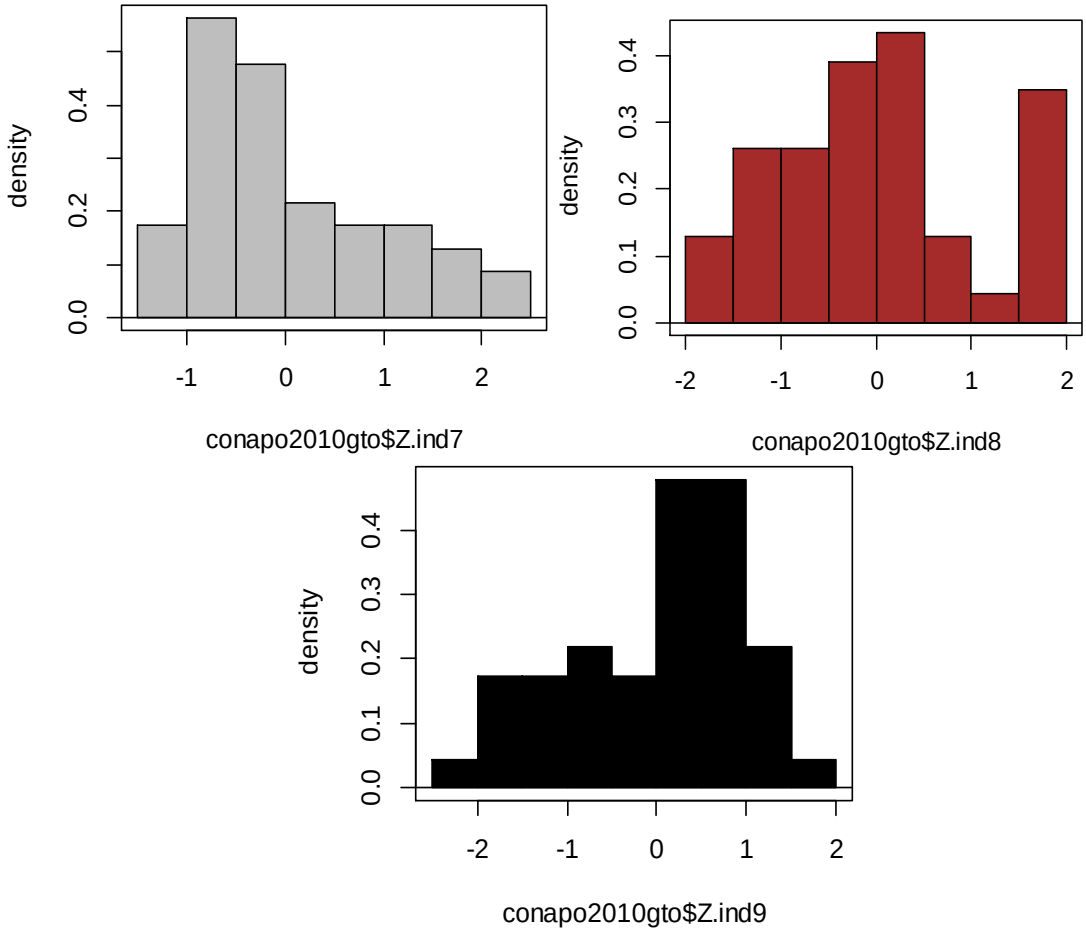












Análisis de correlaciones entre indicadores

```
Hist(conapo2010gto$Z.ind1, scale="density",  breaks="Sturges", col="purple")
```

Pearson correlations:									
	ind1	ind2	ind3	ind4	ind5	ind6	ind7	ind8	ind9
ind1	1.000	0.867	0.770	0.747	0.541	0.494	0.353	0.854	0.779
ind2	0.867	1.000	0.599	0.494	0.314	0.426	0.143	0.761	0.865
ind3	0.770	0.599	1.000	0.809	0.646	0.722	0.549	0.691	0.516
ind4	0.747	0.494	0.809	1.000	0.749	0.547	0.484	0.633	0.386
ind5	0.541	0.314	0.646	0.749	1.000	0.366	0.419	0.418	0.220
ind6	0.494	0.426	0.722	0.547	0.366	1.000	0.596	0.496	0.342
ind7	0.353	0.143	0.549	0.484	0.419	0.596	1.000	0.204	0.141
ind8	0.854	0.761	0.691	0.633	0.418	0.496	0.204	1.000	0.698
ind9	0.779	0.865	0.516	0.386	0.220	0.342	0.141	0.698	1.000
Number of observations: 46									
Pairwise two-sided p-values:									
	ind1	ind2	ind3	ind4	ind5	ind6	ind7	ind8	ind9
ind1		<.0001	<.0001	<.0001	0.0001	0.0005	0.0162	<.0001	<.0001
ind2	<.0001		<.0001	0.0005	0.0333	0.0032	0.3428	<.0001	<.0001

ind3	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	0.0002
ind4	<.0001	0.0005	<.0001		<.0001	<.0001	0.0007	<.0001	0.0081
ind5	0.0001	0.0333	<.0001	<.0001		0.0123	0.0037	0.0039	0.1423
ind6	0.0005	0.0032	<.0001	<.0001	0.0123		<.0001	0.0005	0.0202
ind7	0.0162	0.3428	<.0001	0.0007	0.0037	<.0001		0.1736	0.3517
ind8	<.0001	<.0001	<.0001	<.0001	0.0039	0.0005	0.1736		<.0001
ind9	<.0001	<.0001	0.0002	0.0081	0.1423	0.0202	0.3517	<.0001	

Adjusted p-values (Holm's method)

	ind1	ind2	ind3	ind4	ind5	ind6	ind7	ind8	ind9
ind1		<.0001	<.0001	<.0001	0.0019	0.0074	0.1134	<.0001	<.0001
ind2	<.0001		0.0002	0.0074	0.1667	0.0383	0.6857	<.0001	<.0001
ind3	<.0001	0.0002		<.0001	<.0001	<.0001	0.0015	<.0001	0.0041
ind4	<.0001	0.0074	<.0001		<.0001	0.0016	0.0085	<.0001	0.0728
ind5	0.0019	0.1667	<.0001	<.0001		0.0984	0.0412	0.0412	0.569
ind6	0.0074	0.0383	<.0001	0.0016	0.0984		0.0003	0.0074	0.1212
ind7	0.1134	0.6857	0.0015	0.0085	0.0412	0.0003		0.569	0.6857
ind8	<.0001	<.0001	<.0001	<.0001	0.0412	0.0074	0.569		<.0001
ind9	<.0001	<.0001	0.0041	0.0728	0.569	0.1212	0.6857	<.0001	

III. Aplicación del método de componentes principales para reducir la escala de *n* indicadores a un solo índice;

Instrucciones desde Rcmdr

```
PC <- princomp(~ind1+ind2+ind3+ind4+ind5+ind6+ind7+ind8+ind9,cor=TRUE, data=conapo2010gto)
unclass(loadings(PC))
```

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8	Comp. 9
ind1	-0.399	0.186	-0.089	0.141	0.207	-0.079	-0.194	-0.244	0.795
ind2	-0.344	0.406	0.087	0.129	-0.257	-0.122	-0.382	-0.494	-0.471
ind3	-0.386	-0.180	0.018	-0.205	-0.054	-0.314	0.743	-0.346	-0.055
ind4	-0.359	-0.234	-0.338	-0.115	0.281	-0.550	-0.307	0.406	-0.221
ind5	-0.281	-0.332	-0.608	0.147	-0.480	0.434	-0.019	0.005	0.007
ind6	-0.298	-0.274	0.542	-0.518	-0.369	0.107	-0.270	0.166	0.159
ind7	-0.220	-0.515	0.418	0.622	0.281	0.150	-0.033	-0.060	-0.138
ind8	-0.362	0.243	-0.046	-0.320	0.537	0.595	0.092	0.075	-0.222
ind9	-0.311	0.452	0.174	0.360	-0.275	-0.033	0.296	0.613	0.022

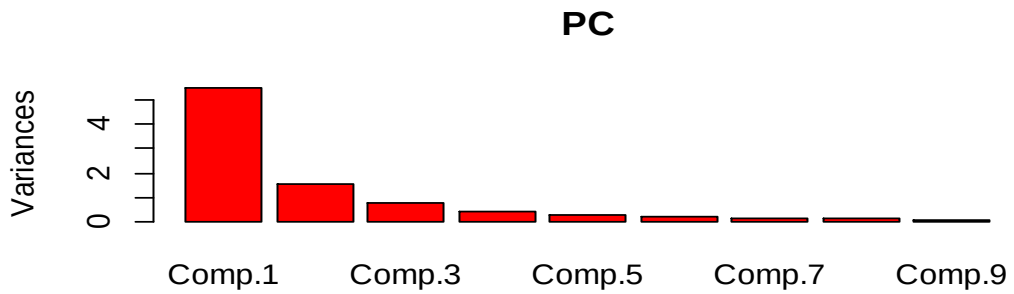
 # proportions of variance

```
summary(PC)
```

Importance of components:									
	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8	Comp. 9
Standard deviation	2.3427	1.2407	0.8866	0.6459	0.5129	0.4327	0.3816	0.3508	0.2245
Proportion of variance	0.6098	0.171	0.0873	0.0464	0.0292	0.0208	0.0162	0.0137	0.0056
Cumulative Proportion	0.6098	0.7808	0.8682	0.9145	0.9437	0.9646	0.9807	0.9944	1

#Gráfica de componentes

```
screeplot(PC, col="red")
```



Hacer lo mismo pero con valores estandarizados

Instrucciones desde Rcmdr

```
PC <-princomp(~Z.ind1+Z.ind2+Z.ind3+Z.ind4+Z.nd5+Z.ind6+Z.ind7+Z.ind8+Z.ind9,cor=TRUE,
data=BDatos)
unclass(loadings(PC))
```

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8	Comp. 9
ind1	-0.399	0.186	-0.089	0.141	0.207	-0.079	-0.194	-0.244	0.795
ind2	-0.344	0.406	0.087	0.129	-0.257	-0.122	-0.382	-0.494	-0.471
ind3	-0.386	-0.180	0.018	-0.205	-0.054	-0.314	0.743	-0.346	-0.055
ind4	-0.359	-0.234	-0.338	-0.115	0.281	-0.550	-0.307	0.406	-0.221
ind5	-0.281	-0.332	-0.608	0.147	-0.480	0.434	-0.019	0.005	0.007
ind6	-0.298	-0.274	0.542	-0.518	-0.369	0.107	-0.270	0.166	0.159
ind7	-0.220	-0.515	0.418	0.622	0.281	0.150	-0.033	-0.060	-0.138
ind8	-0.362	0.243	-0.046	-0.320	0.537	0.595	0.092	0.075	-0.222
ind9	-0.311	0.452	0.174	0.360	-0.275	-0.033	0.296	0.613	0.022

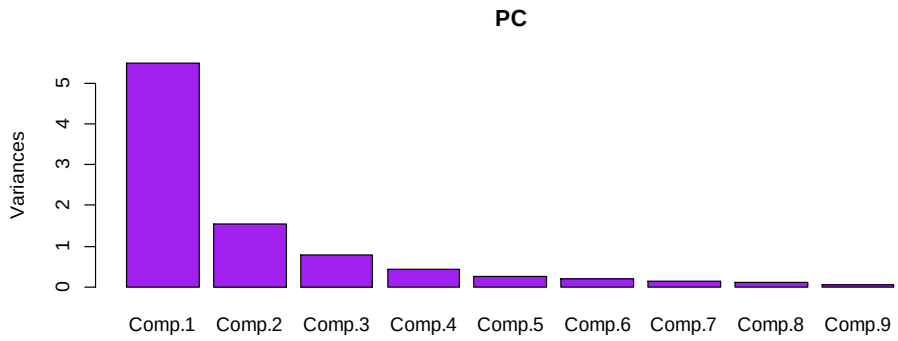
 # proportions of variance

```
summary(PC)
```

Importance of components:									
	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8	Comp. 9
Standard deviation	2.3427	1.2407	0.8866	0.6459	0.5129	0.4327	0.3816	0.3508	0.2245
Proportion of variance	0.6098	0.171	0.0873	0.0464	0.0292	0.0208	0.0162	0.0137	0.0056
Cumulative Proportion	0.6098	0.7808	0.8682	0.9145	0.9437	0.9646	0.9807	0.9944	1

#Grafica de varianzas

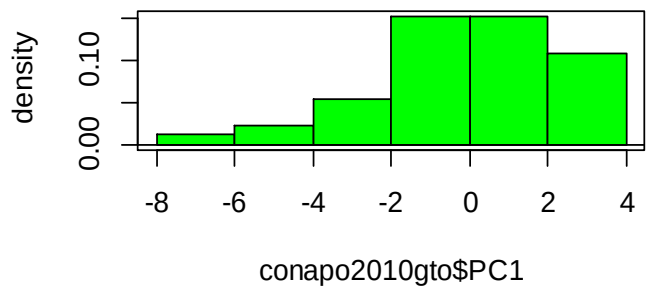
```
screeplot (PC, col="purple")
```



construir el índice con un componente con diferentes combinaciones de los componentes

```
conapo2010gto$PC1 <- PC$score[,1]
```

```
Hist(conapo2010gto$PC1, scale="density", breaks="Sturges", col="green")
```



```
Z <-scale(conapo2010gto[,c("PC1")])
```

Para cambiar el sentido de la curva para que los valores más altos signifique la peor situación.

```
conapo2010gto$Z.PC1 <- -Z[,1]
```

Comprobar que son iguales con el coeficiente de correlación

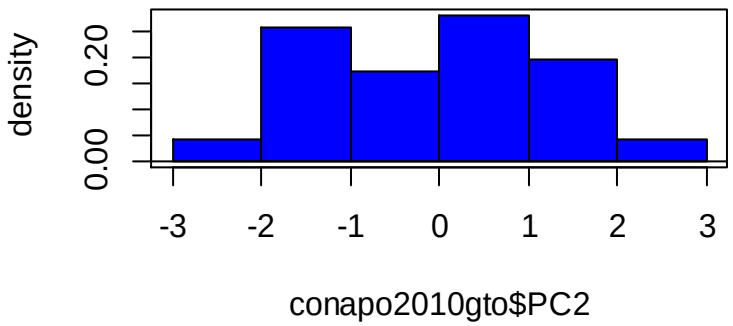
```
rcorr.adjust(conapo2010gto[,c("Z.PC1","ind_marg")],type="pearson")
Pearson correlations:
         Z.PC1 ind_marg
Z.PC1    1.0000   0.9944
ind_marg 0.9944   1.0000
 Number of observations: 46
 Pairwise two-sided p-values:
        Z.PC1 ind_marg
Z.PC1           0
ind_marg  0
 Adjusted p-values (Holm's method)
        Z.PC1 ind_marg
Z.PC1           0
ind_marg  0
```

Con el segundo componente:

construir el índice con un componente con diferentes combinaciones de los componentes

```
> conapo2010gto$PC2 <- PC$score[,2] # contruir el indice con dos componente
```

```
Hist(conapo2010gto$PC2, scale="density", breaks="Sturges", col="blue")
```




```
> Z <-scale(conapo2010gto[,c("PC2")]) # de conapo2010gto elija la variable de PC2 estandarizada y lo guarde en conapo2010gto
> Z
```

Análisis Factorial

Lo que se pretende con el análisis factorial (análisis de Componentes Principales o de Factores Comunes) es simplificar la información que nos da una matriz de correlaciones para hacerla más fácilmente interpretable. Mediante el análisis factorial se analiza la varianza común a todas las variables.

Se pretende encontrar una respuesta a esta pregunta: ¿Por qué unas variables se relacionan más entre sí y menos con otras…? La respuesta hipotética es porque existen otras variables, otras dimensiones o factores que explican por qué unos ítems se relacionan más con unos que con otros. Se trata en definitiva de un análisis de la estructura subyacente a una serie de variables (Morales, 2013).

Estandarizar la matriz, la cual se llama matriz Z

```
Z<-scale(conapo2010gto[,c("ind1","ind2","ind3","ind4","ind5","ind6","ind7","ind8","ind9")])
Z
```

#Incorporar las variables a la base

```
conapo2010gto$Z.ind1 <-Z[,1]
conapo2010gto$Z.ind2 <-Z[,2]
conapo2010gto$Z.ind3 <-Z[,3]
conapo2010gto$Z.ind4 <-Z[,4]
conapo2010gto$Z.ind5 <-Z[,5]
conapo2010gto$Z.ind6 <-Z[,6]
conapo2010gto$Z.ind7 <-Z[,7]
conapo2010gto$Z.ind8 <-Z[,8]
conapo2010gto$Z.ind9 <-Z[,9]
```

I. Aplicación del análisis factorial para reducir la escala de n indicadores a un solo índice;

#Análisis factorial sin rotación, puntuación factorial y un factor

```
FA <-factanal(~Z.ind1+Z.ind2+Z.ind3+Z.ind4+Z.ind5+Z.ind6+Z.ind7+Z.ind8+Z.ind9,factors=1,
rotation="none",scores="none", data=conapo2010gto)
FA
```

Se acepta la hipótesis alternativa pvalue es menor a 0.05 lo que nos indica que el factor es suficiente

```
Call:
factanal(x = ~Z.ind1 + Z.ind2 + Z.ind3 + Z.ind4 + Z.ind5 + Z.ind6 +     Z.ind7 + Z.ind8 + Z.ind9, factors
= 1, data = conapo2010gto,     scores = "none", rotation = "none")

Uniquenesses:
Z.ind1 Z.ind2 Z.ind3 Z.ind4 Z.ind5 Z.ind6 Z.ind7 Z.ind8 Z.ind9
 0.025  0.241  0.378  0.434  0.701  0.724  0.870  0.253  0.383

Loadings:
      Factor1
Z.ind1 0.988
Z.ind2 0.871
```

```
Z.ind3 0.789
Z.ind4 0.753
Z.ind5 0.547
Z.ind6 0.526
Z.ind7 0.360
Z.ind8 0.864
Z.ind9 0.785
                Factor1
SS loadings      4.992
Proportion Var   0.555

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 116.25 on 27 degrees of freedom.
The p-value is 4.8e-13
```

#Análisis factorial sin rotación, puntuación factorial y dos factor

```
FA <-factanal(~Z.ind1+Z.ind2+Z.ind3+Z.ind4+Z.ind5+Z.ind6+Z.ind7+Z.ind8+Z.ind9,factors=2,
rotation="none",scores="none", data=conapo2010gto)
FA
```

Se acepta la hipótesis alternativa pvalue es menor a 0.05 lo que nos indica que el factor es suficiente.

```
Call:
factanal(x = ~Z.ind1 + Z.ind2 + Z.ind3 + Z.ind4 + Z.ind5 + Z.ind6 +     Z.ind7 + Z.ind8 + Z.ind9, factors
= 2, data = conapo2010gto,     scores = "none", rotation = "none")

Uniquenesses:
Z.ind1 Z.ind2 Z.ind3 Z.ind4 Z.ind5 Z.ind6 Z.ind7 Z.ind8 Z.ind9
 0.059  0.086  0.197  0.134  0.390  0.593  0.641  0.259  0.185

Loadings:
      Factor1 Factor2
Z.ind1  0.970
Z.ind2  0.888  -0.354
Z.ind3  0.816   0.371
Z.ind4  0.769   0.525
Z.ind5  0.562   0.542
Z.ind6  0.565   0.297
Z.ind7  0.368   0.474
Z.ind8  0.860
Z.ind9  0.801  -0.417

                Factor1 Factor2
SS loadings      5.136   1.321
Proportion Var   0.571   0.147
Cumulative Var   0.571   0.717

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 34.62 on 19 degrees of freedom.
The p-value is 0.0155
```

#Análisis factorial con rotación ortogonal varimax, puntuación factorial Bartlett y dos factores

```
FA <-factanal(~Z.ind1+Z.ind2+Z.ind3+Z.ind4+Z.ind5+Z.ind6+Z.ind7+Z.ind8+Z.ind9,factors=2,
rotation="varimax",scores="Bartlett", data=onapo2010gto)
FA
```

```
Call:
factanal(x = ~Z.ind1 + Z.ind2 + Z.ind3 + Z.ind4 + Z.ind5 + Z.ind6 +     Z.ind7 + Z.ind8 + Z.ind9, factors
= 2, data = conapo2010gto,     scores = "Bartlett", rotation = "varimax")

Uniquenesses:
Z.ind1 Z.ind2 Z.ind3 Z.ind4 Z.ind5 Z.ind6 Z.ind7 Z.ind8 Z.ind9
 0.059  0.086  0.197  0.134  0.390  0.593  0.641  0.259  0.185

Loadings:
      Factor1 Factor2
Z.ind1 0.815   0.526
Z.ind2 0.935   0.200
Z.ind3 0.472   0.762
Z.ind4 0.347   0.864
Z.ind5 0.165   0.763
Z.ind6 0.304   0.561
Z.ind7         0.598
Z.ind8 0.742   0.438
Z.ind9 0.898

               Factor1 Factor2
SS loadings      3.357   3.100
Proportion Var   0.373   0.344
Cumulative Var   0.373   0.717

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 34.62 on 19 degrees of freedom.
The p-value is 0.0155
```

#Análisis factorial con rotación oblicua Promax, puntuación factorial Regresión y dos factores

```
FA <-factanal(~Z.ind1+Z.ind2+Z.ind3+Z.ind4+Z.ind5+Z.ind6+Z.ind7+Z.ind8+Z.ind9,factors=2,
rotation="promax",scores="regression", data=conapo2010gto)
FA
```

```
Call:
factanal(x = ~Z.ind1 + Z.ind2 + Z.ind3 + Z.ind4 + Z.ind5 + Z.ind6 +     Z.ind7 + Z.ind8 + Z.ind9, factors
= 2, data = conapo2010gto,     scores = "regression", rotation = "promax")

Uniquenesses:
Z.ind1 Z.ind2 Z.ind3 Z.ind4 Z.ind5 Z.ind6 Z.ind7 Z.ind8 Z.ind9
 0.059  0.086  0.197  0.134  0.390  0.593  0.641  0.259  0.185

Loadings:
      Factor1 Factor2
Z.ind1  0.379   0.706
Z.ind2          0.977
Z.ind3  0.755   0.218
Z.ind4  0.911
Z.ind5  0.846  -0.130
Z.ind6  0.569   0.111
Z.ind7  0.689  -0.201
Z.ind8  0.296   0.659
Z.ind9 -0.147   0.977

               Factor1 Factor2
SS loadings      3.169   2.960
```

```
Proportion Var   0.352   0.329
Cumulative Var   0.352   0.681

Factor Correlations:
       Factor1 Factor2
Factor1   1.000   0.561
Factor2   0.561   1.000

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 34.62 on 19 degrees of freedom.
The p-value is 0.0155
```

calcular las puntuación factoriales

```
conapo2010gto$F1 <-FA$scores[,1]

conapo2010gto$F2 <-FA$scores[,2]
```

#Estandarización de factores

```
Z <-scale(conapo2010gto[,c("F1","F2")])
conapo2010gto$Z.F1 <-Z[,1]
conapo2010gto$Z.F2 <-Z[,2]
```

	F1	F2		F1	F2
1	-0.75803447	0.87488633	24	-1.13536579	1.25555414
2	-0.73426637	0.25406	25	-0.15387395	-0.81801564
3	0.94121218	-0.75949258	26	-0.95357447	0.92952136
4	-0.23117464	0.13859783	27	-0.0031919	-1.22562279
5	-0.04355013	-0.37620746	28	-0.89406465	0.67330073
6	2.43221857	-0.02242135	29	-0.04392468	1.04268105
7	0.4387611	-1.93762609	30	0.5463392	0.16555041
8	-0.58671135	0.99254954	31	-0.11284149	-0.80800396
9	0.84925716	-0.52944652	32	0.4700198	-0.90338822
10	-0.74568961	0.98837147	33	0.59782243	-0.26071431
11	0.02286541	-0.95404519	34	1.64512263	0.05800619
12	-0.58616183	0.48456166	35	-0.43956863	0.48517751
13	-0.09757143	0.82731259	36	-1.78831539	2.21073815
14	0.57516501	-0.21260936	37	0.14091008	-0.74021147
15	0.91630207	-2.23948025	38	-0.73901584	0.47206312
16	-1.28786827	0.93452801	39	-1.11058779	1.26476707
17	0.26983293	-1.61783175	40	2.42635128	-0.64024612
18	-0.5740177	-0.11993913	41	-0.87537945	-0.13601033
19	-0.42008339	1.6737171	42	-0.75040855	0.41567879

20	0.4331277	-2.01167523	**43**	1.87193108	-0.82442461
21	-0.68456702	-0.51525213	**44**	0.14164371	-1.08505898
22	-0.09751043	0.86193746	**45**	2.6289428	0.10984522
23	-0.43381216	0.51134993	**46**	-1.06669378	1.11296784
attr(,"scaled:center")					
	F1	F2			
	2.42E-17	-1.71E-17			
attr(,"scaled:scale")					
	F1	F2			
	1.132008	1.14847			

Métodos para segmentar, estratificar o formar grupos

Distancia Euclidiana[2]

La distancia entre dos puntos es la longitud de la trayectoria de conectarlos. En el plano, la distancia entre los puntos (x_1, y_1)y (x_2, y_2)viene dada por la teorema de Pitágoras.

$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$.

En euclidiano de tres el espacio, la distancia entre los puntos (x_1, y_1, z_1)y (x_2, y_2, z_2)es

$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$.

En general, la distancia entre los puntos $\mathbf{x}$y $\mathbf{y}$en un espacio euclidiano $\mathbb{R}^n$ está dada por

$$d = |\mathbf{x} - \mathbf{y}| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} .$$

Estratificación de Índices por el método de Dalenius---Hodges

El método de Dalenius---Hodges se utiliza para forman estratos, con la condición de que la varianza sea mínima entre cada uno de los grupos. Con n igual al número de observaciones, L igual al número de estratos, el procedimiento de estratificación consiste en:

1. Ordenar las observaciones de manera ascendente
2. Agrupar las observaciones en J clases, donde J=min (L*10,n).
3. Calcular los límites para cada clase

#Segmentos Equidistantes

```
conapo2010gto$Seg1_indice <-bin.var(conapo2010gto$ind_marg,bins=5,method='intervals',labels=c('Muy Bajo','Bajo','Medio','Alto','Muy Alto'))
```

[2] Gray, A. "La idea intuitiva de distancia sobre una superficie." § 15.1 de *Modern Geometría Diferencial de Curvas y Superficies con Mathematica , 2 ª ed.* Boca Raton, FL: CRC Press, pp 341-345, 1997.

#Segmentos de igual cantidad

```
conapo2010gto$Seg2_indice <-bin.var(conapo2010gto $ind_marg, bins=5, method='proportions',
labels=c('Muy Bajo','Bajo','Medio','Alto','Muy Alto'))
```

#Segmentos naturales (mediante agrupación por K-medias)

```
conapo2010gto$Seg3_indice <-bin.var(conapo2010gto$ind_marg, bins=5, method='natural', labels=c('Muy
Bajo','Bajo','Medio','Alto','Muy Alto'))
```

#Análisis de las distribuciones

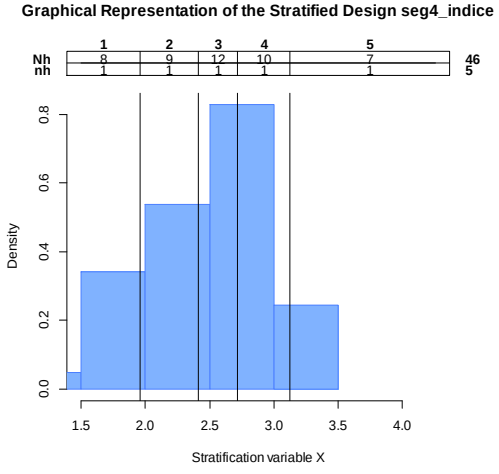
```
Seg1_indice <-bin.var(conapo2010gto$ind_marg, bins=5, method='intervals', labels=c('Muy
Bajo','Bajo','Medio','Alto','Muy Alto'))

Table <-table(Seg1_indice)
Table # counts for Seg1_indice
round(100*Table/sum(Table), 2) # percentages for Seg1_indice
remove(Table)
```

> Table <-table(Seg1_indice)				
> Table # counts for Seg1_indice				
Seg1_indice				
Muy Bajo	Bajo	Medio	Alto	Muy Alto
10	9	18	6	3
> round(100Table/sum(Table), 2) # percentages for Seg1_indice				
Seg1_indice				
Muy Bajo	Bajo	Medio	Alto	Muy Alto
21.74	19.57	39.13	13.04	6.52

Método de Dalenius---Hodges

Cargar el paquete stratification library(stratification)

```
library(stratification)
seg4_indice <-strata.cumrootf (x= conapo2010gto$ind_marg+3, nclass = 50, CV = 0.05, Ls = 5, alloc = c(0.5,
0,0.5))
seg4_indice
plot(seg4_indice, drop=5)
```



Bibliografía

1. Rencher A. C. (2002). Methods of Multivariate Analysis (2da ed.). Canada: John Wiley & Sons, Inc

2. Wooldrige J.M. (2011). Introduccion a la econometria. Un enfoque moderno (4ta ed.). Mexico: Cengage Learning.

3. Newbold P., Carlson W. y Thorne B. (2008). Estadística para administración y economía (6ta ed.). Madrid: PEARSON. Prentice Hall.

Presentaciones de clase

1. Sesión 2, 3, 4 y 5. Maestría. Métodos y Técnicas de Análisis Regional y Urbano. Mtro. Miguel Ángel Mendoza, Semestre 2014-1, Faculta de Economía, UNAM

Páginas de Internet

1. R Project, What R? Consultada el 10 de diciembre de 2013 de http://www.r-project.org/