Análisis de los datos con ANALIZAR DATOS de EPI INFO - 2007

Dr. Juan Gagliardi

Generalidades

ANALIZAR DATOS o ANALYSIS es el módulo de Epi Info que permite producir listados, frecuencias, tablas y estadísticas de los archivos de Epi Info o bases realizadas con otros formatos. Con órdenes simples, se pueden seleccionar registros que cumplan determinados criterios, ordenar o listar registros, hacer frecuencias o cruces de variables, hacer operaciones lógicas o matemáticas en un campo, poner los resultados en una nueva variable, y dirigir los resultados a la pantalla, a la impresora o a un archivo específico.

Todas las órdenes que se ejecutan en este módulo generan un archivo con formato HTML con el nombre OUT*.HTML donde el asterisco es un número correlativo que el programa asigna automáticamente en cada sesión de análisis de los datos. El archivo se guarda por defecto en el directorio del programa: C:\Epio_info salvo que se especifique lo contrario (ver más adelante). Este archivo se puede editar para revisar los resultados, copiar las tablas, imprimirlas o editarlas con otros programas para presentaciones, etc.

En esta sección utilizaremos las órdenes más comunes que permiten revisar los datos y realizar una descripción de los mismos.

Información general

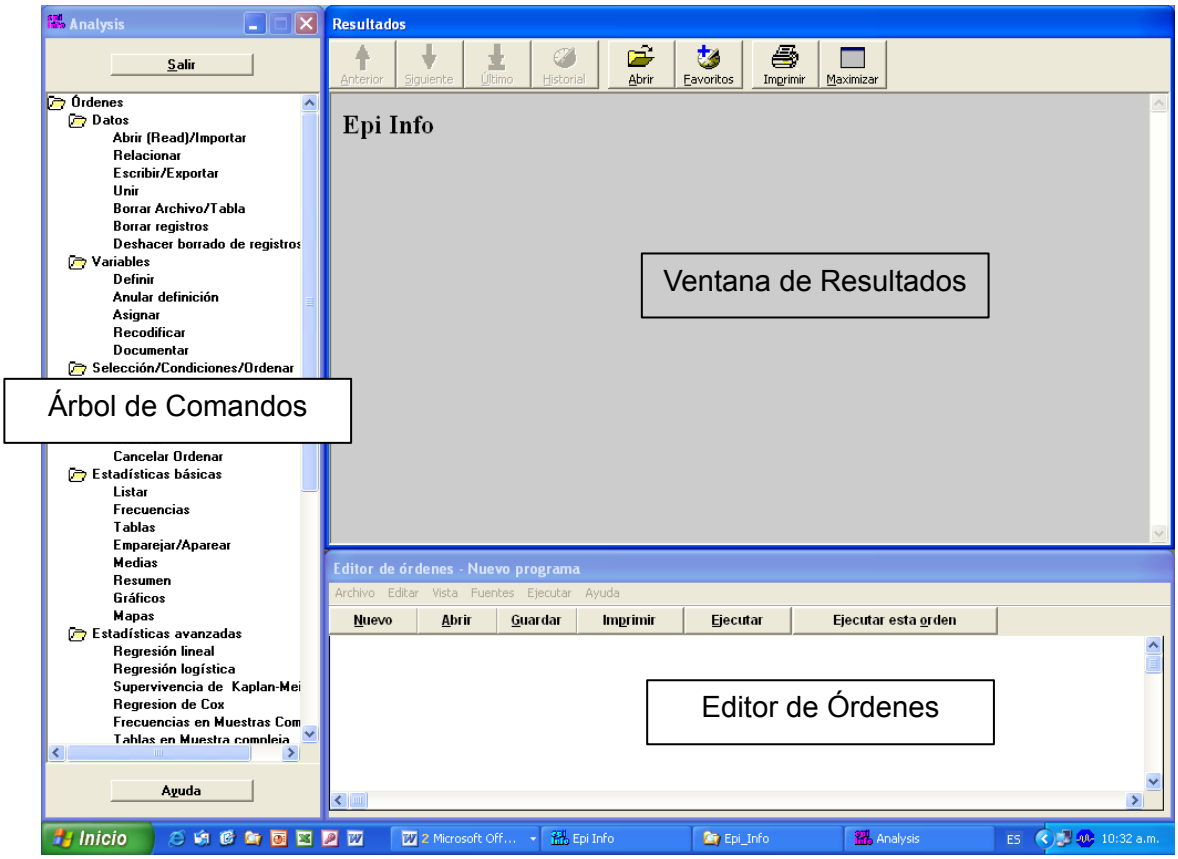
Cuando active ANALIZAR DATOS desde el menú principal, aparecerá una pantalla que contiene 3 ventanas: el Árbol de Comandos, la Ventana de Resultados y el Editor de Órdenes (figura 1).

En la ventana del Árbol de Órdenes o Comandos se encuentran disponibles las diferentes órdenes o comandos, agrupadas en diferentes carpetas.

En la ventana de Resultados, aparecen todos los resultados de las órdenes ejecutadas. Al leer un archivo aparece cuál es la vista que se encuentra activa y cuántos registros tiene esa vista.

En la ventana del Editor de Órdenes se visualizan todas las órdenes que se van generando. También se pueden escribir directamente en este sector las órdenes para se ejecutadas o programas de análisis sistemático de los datos.

En la parte inferior de la ventana del Árbol de Comandos aparece el botón de ayuda.



Leer un archivo: READ

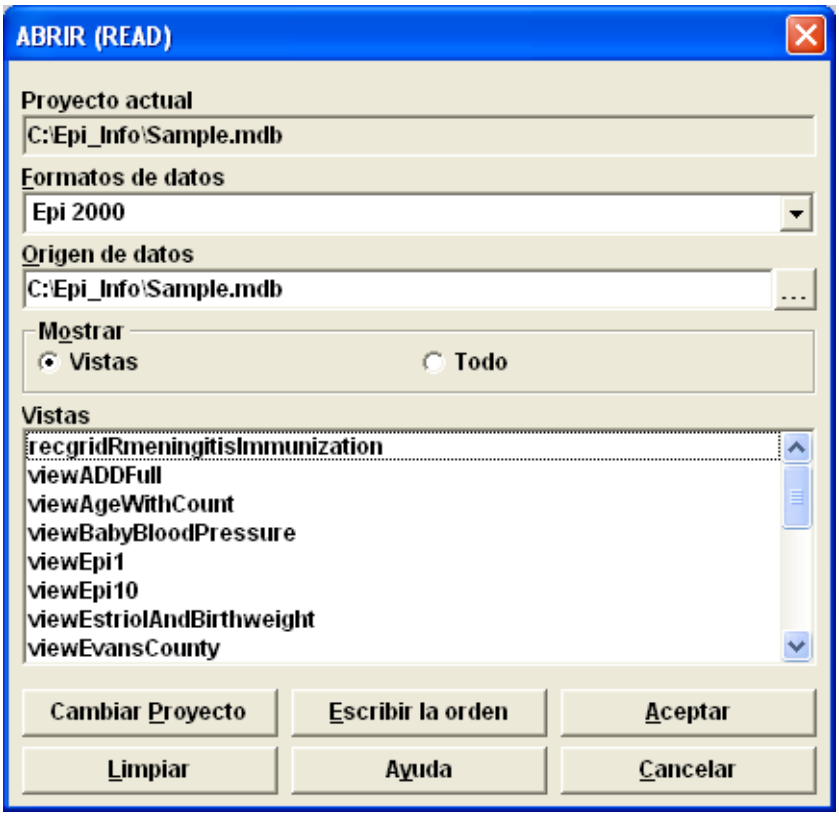
ANALIZAR DATOS debe trabajar con los registros de archivos. Los archivos pueden ser de Epi Info, o tener otros formatos que pueden ser importados directamente con esta órden.

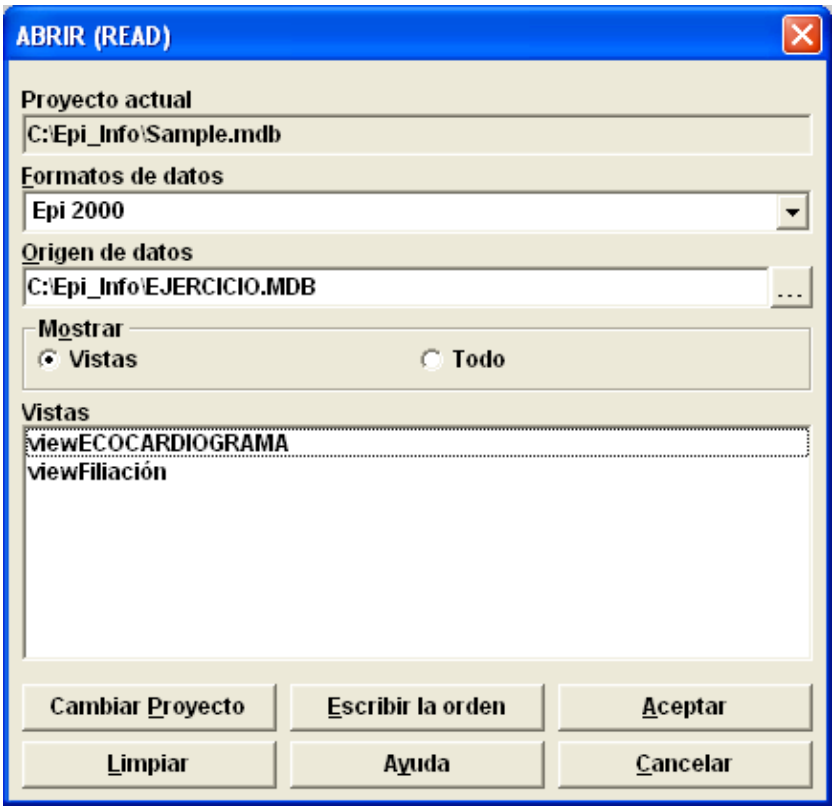
Para abrir el archivo seleccionamos la orden ABRIR (READ)/IMPORTAR de la ventana del Árbol de Comandos.

Se abre así la ventana ABRIR (READ) en la cuál debemos seleccionar el formato de los datos (habitualmente se presenta por default como Epi 2000). Los formatos disponibles son: Access, DBase, Epi2000, Epi6, Excel, FoxPro, HTML, ODBC, Paradox y Text.

En el cuadro de origen de los datos debemos seleccionar la base de datos donde se encuentran almacenados los datos. Presionando el botón  se abre una ventana para poder buscar y seleccionar el archivo.

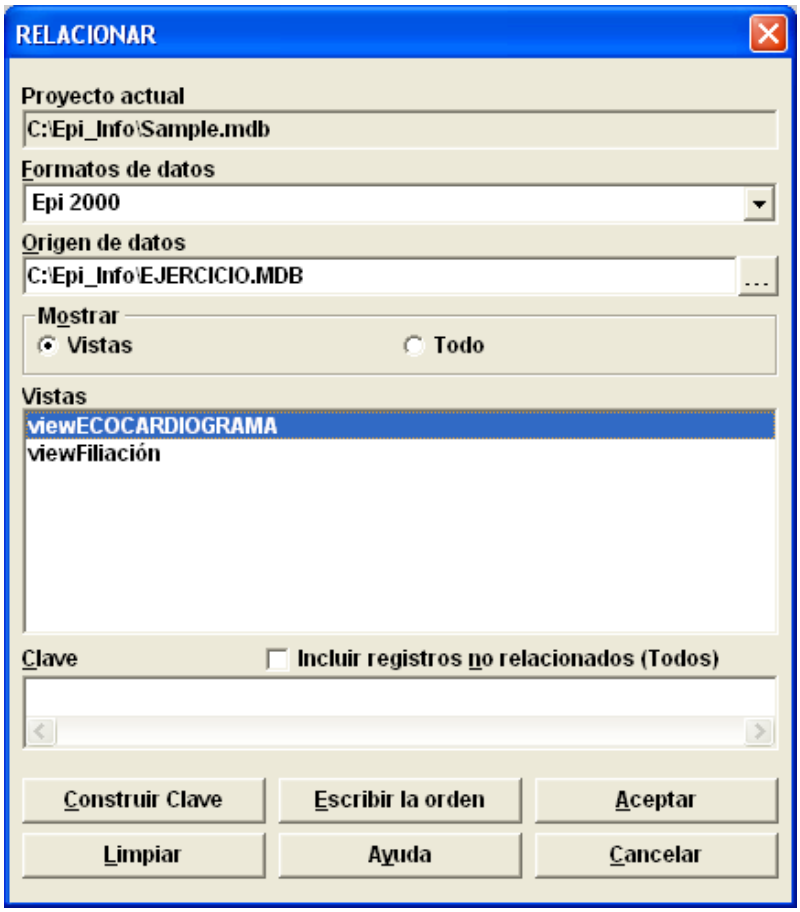
En la ventana inferior aparecerán todas las vistas disponibles, de las cuales debemos seleccionar la que se utilizará para trabajar.



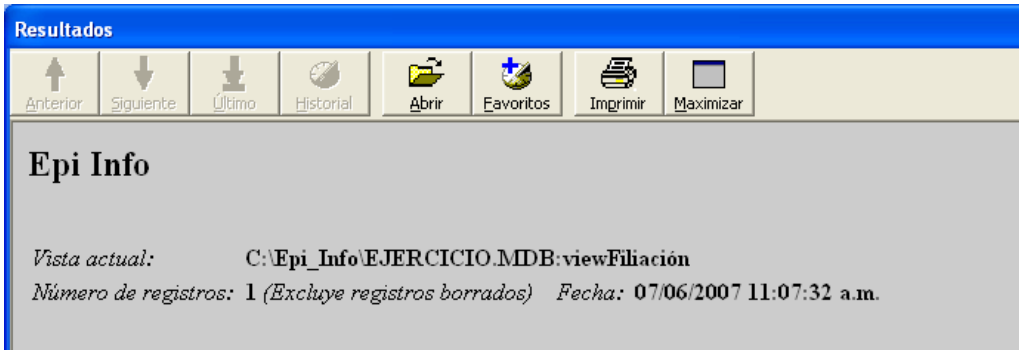


Si queremos relacionar las tablas para poder analizar datos de 2 o más tablas del mismo proyecto, debemos seleccionar la orden RELATE de la ventana del Árbol de Comandos. Aparece una ventana denominada RELACIONAR donde debemos seleccionar la vista a relacionar y construir la

clave que permitirá unir las tablas por un campo común (por ejemplo Nº de Historia Clínica).



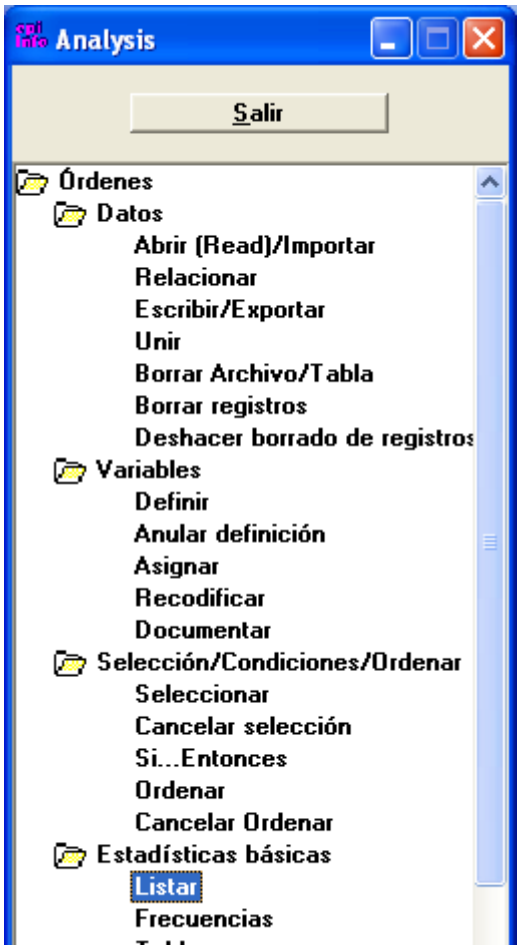
Una vez leído el archivo, en la ventana de resultados se muestra cuál es la vista activa del proyecto y cuantos registros tiene esa tabla. A partir de allí se puede comenzar el análisis de los datos.



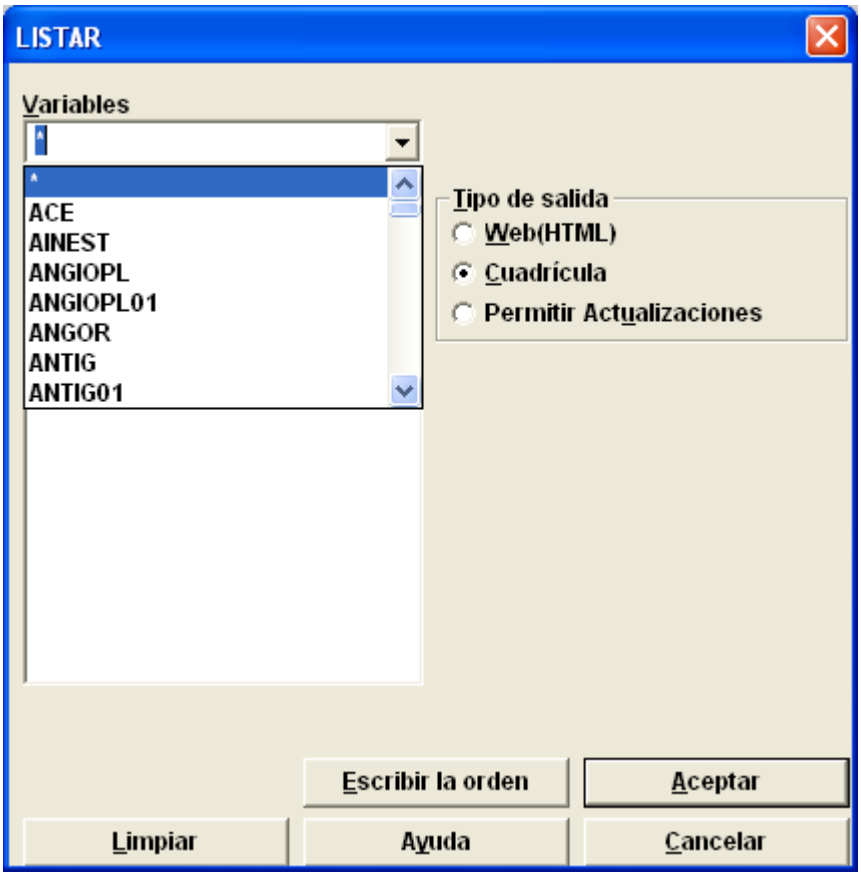
Para el análisis inicial de los datos es necesario producir listados, establecer frecuencias para la descripción de datos cualitativos y medias para describir datos cuantitativos.

Producir listados: La orden LIST

Si requiere un listado de los registros del archivo, seleccione de la ventana del árbol de comandos la orden LISTAR



Se abre una ventana en la que debemos seleccionar la o las variables a listar (si seleccionamos el asterisco se listarán todas las variables). Al mismo tiempo podemos seleccionar el tipo de salida, ya se en formato html, grilla o cuadrícula y si se permitirán actualizaciones, es decir si sobre el listado se pueden introducir modificaciones sobre los datos de la tabla. Este último punto puede ser necesario si al revisar una tabla se encuentran datos discordantes o mal cargados como por ejemplo una edad de 630 años en lugar de 63.



La orden LIST también permite, marcando el botón correpondiente, listar todas las variables excepto las seleccionadas.

Los listados aparecerán la ventana de Resultados de la siguiente forma:

Resultados

HCLINICA	EDAD	SEXO	INGRESO	MOTCONS	INFARTO	ANGOR	AINEST	ANGIOPL0
80207002	86	F	06/01/2000	DISNEA DE	No	Missing	No	No
3003284700	80	F	12/01/2000	DERIVADO I	No	Missing	No	No
8740001	63	M	24/02/2000	DERIVADO I	Yes	Missing	No	No
7997501	86	M	26/02/2000	BAJO VOLUI	Yes	Missing	No	No
1,0256E+13	45	M	26/02/2000	ANGINA DE	No	Missing	Yes	Yes
61963402	94	F	28/02/2000	PRECORDIA	No	Missing	No	No
117764900	56	M	28/02/2000	DERIVADO I	No	Missing	No	No
1,00392E+1	80	M	29/02/2000	DISNEA DE	No	Missing	No	No
4731204	46	M	29/02/2000	ANGINA DE	No	Missing	No	Yes
3003466300	64	M	01/03/2000	DERIVADO I	No	Missing	No	No
1,01166E+1	83	F	01/03/2000	ANGINA DE	No	Missing	No	Yes
66273102	70	F	01/03/2000	DERIVADO I	Yes	Missing	No	No
6981201	59	M	01/03/2000	PRECORDIA	No	Missing	Yes	Yes
8740001	63	M	02/03/2000	DISNEA DE	Yes	Missing	No	No
4751102	68	F	03/03/2000	DERIVADO I	No	Missing	No	No
2072102	62	F	05/03/2000	ANGINA DE	No	Missing	Yes	Yes
81727702	83	F	07/03/2000	DERIVADO I	No	Missing	No	No
200170790	54	M	07/03/2000	DERIVADO I	No	Missing	No	No
81735401	88	M	08/03/2000	PRECORDIA	Yes	Missing	No	No
3000889700	70	F	09/03/2000	DERIVADO I	No	Missing	No	No
106062	90	F	10/03/2000	PRECORDIA	No	Missing	Yes	No
117665300	71	M	12/03/2000	DERIVADO I	No	Missing	No	No
121536701	62	F	04/03/2000	BAJO VOLUI	No	Missing	No	No
61430901	83	M	07/03/2000	DISNEA DE	Yes	Missing	No	No
1,01553E+1	76	M	07/03/2000	MAREOS	Yes	Missing	No	No
62796302	74	F	08/03/2000	DERIVADO I	No	Missing	No	No
1519075005	76	F	08/03/2000	TAQUIARRI	No	Missing	No	No

Frecuencias: La orden FREQ

La orden FRECUENCIAS (FREQ), se utiliza para la descripción de datos cualitativos. Contará cada categoría para una variable especificada y dará los resultados absolutos y frecuencias relativas para cada una.

Por ejemplo para describir el SEXO, elegimos la orden FRECUENCIAS y se abre una pantalla de FRCUENCIAS (FREQ si utilizamos el idioma inglés). Allí se despliega el listado de variables y se selecciona la variable sexo y se presiona el botón ACEPTAR.

El resultado se informa de la siguiente manera indicando la frecuencia relativa de F y M. Se presenta primero el número de casos, seguido del porcentaje relativo y el porcentaje acumulado, y luego el intervalo de confianza del 95% de cada uno de los porcentajes.

SEXO	Frecuencia	Porcentaje	Porcentaje acumulado	
F	2504	46,5%	46,5%	
M	2882	53,5%	100,0%	
Total	5386	100,0%	100,0%	

		Int. Conf. 95 %
F	45,2 %	47,8 %
M	52,2 %	54,8 %

Si el campo es numérico, se presentará la frecuencia relativa de cada uno de los números así como el intervalo de intervalo de confianza del 95% de cada uno de los porcentajes.

EDAD

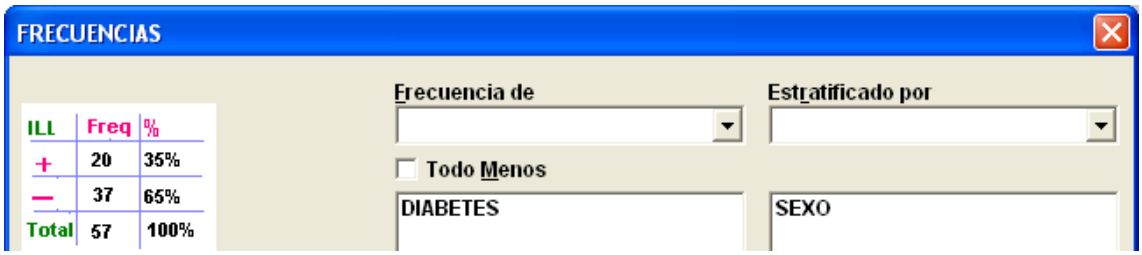
EDAD	Frecuencia	Porcentaje	Porcentaje acumulado	
0	1	0,0%	0,0%	
2	2	0,0%	0,1%	
3	1	0,0%	0,1%	
…..	……	……..	…….	
53	79	1,5%	18,2%	
54	90	1,7%	19,9%	
55	88	1,6%	21,5%	
56	66	1,2%	22,8%	
57	73	1,4%	24,1%	
…..	……..	……..	…….	
97	13	0,2%	99,7%	
98	11	0,2%	99,9%	
99	4	0,1%	100,0%	
Total	5403	100,0%	100,0%	

	Int. Conf. 95 %	
…	…..	…..
56	1,0 %	1,6 %
57	1,1 %	1,7 %

58	1,3 %	2,0 %
59	1,1 %	1,7 %
60	1,2 %	1,9 %
61	1,3 %	2,0 %

En la ventana de Frecuencias se pueden seleccionar varias variables para analizarlas una a continuación de otra sin necesidad de escribir las órdenes cada vez.

También es posible estratificar el análisis por ejemplo por sexo o grupo de pacientes, seleccionado la variable a describir y la variable por la cual se va a estratificar.

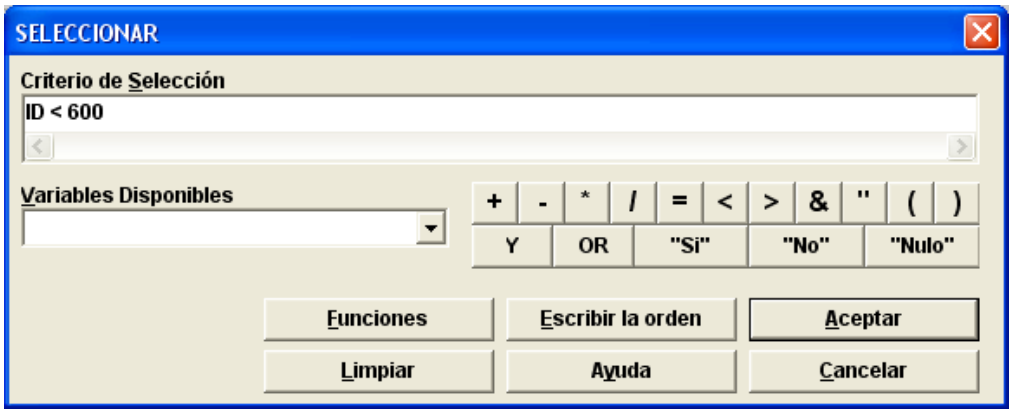


En este caso el resultado será una descripción de la variable "Diabetes" para cada sexo pero no se brindarán estadísticas que comparen ambos resultados.

Ejercicio con el Intervalo de Confianza

A los fines de ejercitación, seleccionaremos los primeros 600 pacientes ingresados en la base de datos, es decir cuyo numero de identificación (ID) sea menor de 600.

Utilizaremos la orden SELECT, y en el listado de "Variables Disponibles" elegimos ID (número de identificación de los pacientes, que en este caso saltea muchos números) indicando que sea menor de 600. Repetiremos luego el análisis de la frecuencia del sexo con la orden FREQ SEXO. ¿Qué pasó con el intervalo de confianza del sexo? ¿Es más o menos amplio que en toda la población? ¿Por qué?



El resultado se verá de la siguiente forma:

SEXO	Frecuencia	Porcentaje	Porcentaje acumulado	
F	264	45,7%	45,7%	
M	314	54,3%	100,0%	
Total	578	100,0%	100,0%	

	Int. Conf. 95 %	
F	41,6 %	49,8 %
M	50,2 %	58,4 %

Si comparamos los porcentajes veremos que son similares a los resultados obtenidos con la población total. Sin embargo los intervalos de confianza son más amplios, indicando una menor seguridad respecto del valor real de la población.

El *concepto de intervalo de confianza* de cualquier valor es una medida de la seguridad que tenemos de que el valor real, si repetimos experimentos o encuestas, será similar al que afirmamos. Esta confianza surge en forma directa del número de pacientes que se incluyeron en el estudio.

Para volver a trabajar con toda la base repetimos en el editor de órdenes la orden SELECT sin indicar ningún criterio (deselecciona todo criterio que se haya establecido) o seleccionamos Cancelar Selección del Árbol de Ordenes.

La orden MEANS (MEDIAS)

La orden MEDIAS ofrece una tabla de datos continuos u ordinales y realiza las estadísticas adecuadas.

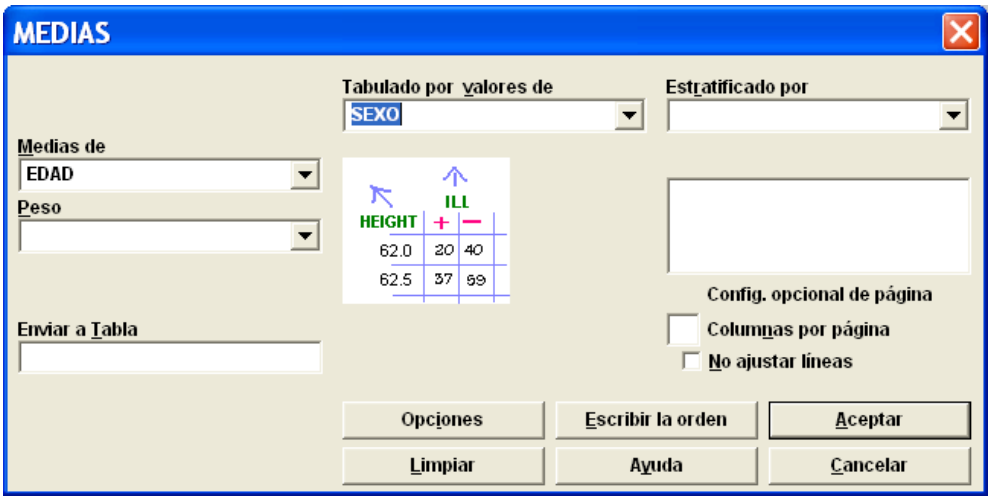
Habitualmente la orden MEANS requiere dos ítems de información: la variable que contiene los datos que van a ser analizados y la variable que indica qué grupos se han de diferenciar. La orden que se escribirá en el Editor de Órdenes es:

`MEANS` [variable numérica a analizar] [variable de agrupación]

Al ejecutar la orden se presentará una tabla de frecuencias con la cantidad de individuos para cada valor. Si prefiere que no aparezca la tabla de datos se debe quitar el tilde en "Mostrar Tablas" de las preferencias de Análisis (ver más adelante).

La orden MEANS EDAD SEXO comparará las edades de las mujeres ("F") con las de los hombres ("M").

Para relizarla se selecciona del Árbol de Comandos "MEDIAS" o "Means" de la versión en inglés. Se abre una ventana donde debemos seleccionar la variable a analizar en el cuadro "Medias de:" y la variable de agrupación en la casilla "Tabulado por valores de:"



Se producirán los siguientes resultados:

SEXO

EDAD	F	M	TOTAL
43	2	10	12
44	3	12	15
45	1	13	14

Etc.	....	....	.....
87	3	2	5
89	3	0	3
91	0	1	1
TOTAL	321	699	1020

Estadísticas descriptivas de la variable de cruce

	Obs	Total	Media	Varianza	Desviación típica
F	321	21104,0000	65,7445	120,9658	10,9984
M	699	42522,0000	60,8326	131,5780	11,4707

	Mínimo	25%	Median	75%	Maximum	Mode
F	40,0000	57,0000	66,0000	74,0000	89,0000	68,0000
M	31,0000	53,0000	61,0000	69,0000	91,0000	60,0000

ANOVA, test paramétrico para comparación de medias

(Para datos normalmente distribuidos)

Variación	SC	gl	MS	Estadístico F
Entre grupos	5307,4562	1	5307,4562	41,3862
Intra grupo	130550,4693	1018	128,2421	
Total	135857,9255	1019		

Estadístico T = 6,4332

Valor p= 0,0000

Test de Bartlett para igualdad de Varianzas poblacionales

Chi cuadrado de Bartlett=	0,7667	df= 1	valor p=0,3813

Un valor de p pequeños (menor de 0,05) sugiere que las varianzas no son homogeneas y que el test ANOVA no es apropiado.

Test de dos muestras de Mann-Whitney/Wilconxon (Test de Kruskal-Wallis para dos grupos)

H de Kruskal-Wallis (equivalente a Chi cuadrado)=	37,3173
Grados de libertad=	1
Valor p=	0,0000

La tabla nos indica que existen 2 mujeres de 43 años, que de los 15 pacientes de 44 años, 12 son hombres y 3 son mujeres, y así sucesivamente.

La media de edad en las mujeres fue de 65,7 años y en los hombres 60,8. El valor de p por el test de ANOVA (en este caso equivalente al Test de T de Student) es <0.00001, por lo que podemos concluir que la diferencia de edad en los dos grupos es altamente significativa.

También la orden MEANS puede utilizarse sólo con una variable cuantitativa para describir los datos. La orden en el editor de órdenes es:

`MEANS` [variable numérica a analizar]

De esta manera la orden:

```
MEANS EDAD
```

producirá una tabla similar a la de la orden FREQ con los siguientes resultados:

EDAD	Frecuencia	Porcentaje	Porcentaje acumulado
31	1	0,1%	0,1%
34	2	0,2%	0,3%
36	4	0,4%	0,7%
38	7	0,7%	1,4%
39	3	0,3%	1,7%
40	7	0,7%	2,4%
Etc	.....	.....	.....
61	28	2,7%	45,7%
62	23	2,3%	47,9%
63	41	4,0%	52,0%
64	35	3,4%	55,4%
86	6	0,6%	99,1%
87	5	0,5%	99,6%
89	3	0,3%	99,9%

91	1	0,1%	100,0%
Total	1020	100,0%	100,0%

Observaciones	Total	Media	Varianza	Desviación típica
1020	63626,0000	62,3784	133,3248	11,5466

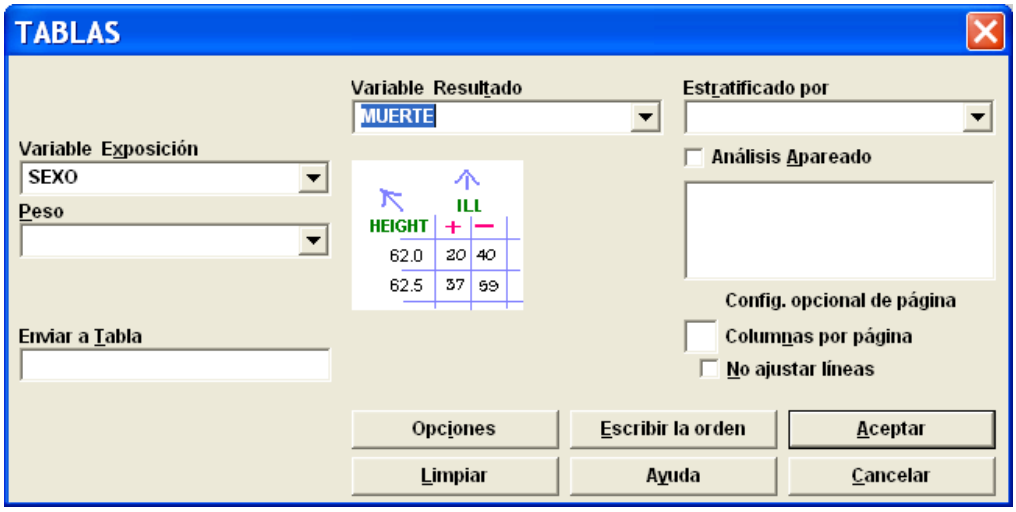
Mínimo	25%	Mediana	75%	Máximo	Moda
31,0000	54,0000	63,0000	71,0000	91,0000	63,0000

La media de la edad en esta muestra de 1020 pacientes es 62,4 años y el desvío estándar es 11.5 años. El rango de edad es de 31 a 91 años. La mediana y la moda son de 63 años.

Cruce de variables: TABLAS

La orden TABLAS ("TABLES" en la versión en inglés) se utiliza para el análisis de variables cualitativas. Contará los registros que cumplan los mismos criterios para dos variables.

Seleccionando del árbol de comandos la orden TABLAS se abre una ventana donde se deben seleccionar las variables a cruzar, por ejemplo sexo y muerte.



Se producirá el siguiente resultado:

MUERTE

SEXO	+	-	TOTAL
F	7	316	323

M	13	686	699
TOTAL	20	1002	1022

Análisis de tabla simple

	Point Estimación	95% Intervalo de Confianza L. Inferior	L. Superior	
PARAMETROS: Basados en OR				
Odds Ratio (producto cruzado)	1,1689	0,4619	2,9582	(T)
Odds Ratio (EMV-MLE)	1,1688	0,4336	2,9423	(M)
		0,3909	3,1881	(F)
PARAMETROS: Basados en el riesgo				
Razón de Riesgos (RR)	1,1653	0,4694	2,8930	(T)
Diferencia de Riesgos (DR)	0,3074	-1,5701	2,1848	(T)

(T=Series Taylor;C=Cornfield;M=P-Media;F=Fisher)

TEST ESTADÍSTICOS	Chi cuadrado	p de 1 cola	p de 2 colas
Chi-square - uncorrected	0,1088		0,7415252447
Chi-square - Mantel-Haenszel	0,1087		0,7416472136
Chi-square - corrected (Yates)	0,0076		0,9306925167
P-media exacta		0,3656456392	
Test exacto de Fisher		0,4539666885	

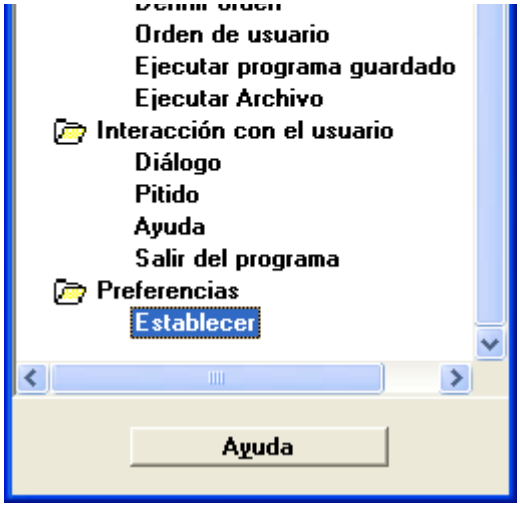
Observe que la interpretación de la razón de riesgos (riesgo relativo) depende de la orientación de la tabla y no todos los riesgos relativos tendrán explicación. El riesgo relativo es la razón de tasas del primer factor en la primera línea comparada con la tasa del primer factor en la segunda línea o (a/a+b)/(c/c+d). La interpretación depende del valor del riesgo relativo. Si los factores de riesgo están en el lado izquierdo de la tabla y la enfermedad arriba, el riesgo relativo representa en este caso el riesgo de morir de los sujetos con el primer factor (las mujeres) en relación a aquellos con el segundo factor. Como se indica en las observaciones, el riesgo relativo deberá ignorarse en los estudios de casos controles al no poder interpretarse los resultados (Ver texto: "Análisis de datos cualitativos").

Valores Nulos

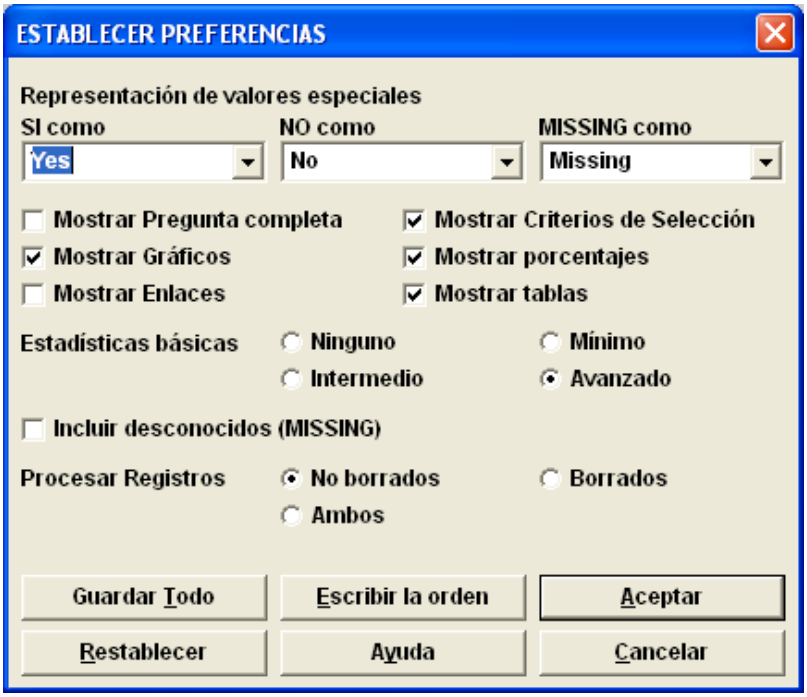
Los valores nulos en Epi Info se escriben como "blancos" en el campo correspondiente. Durante el proceso de introducción de datos, si se pulsa <Enter> en un campo en vez de escribir un valor, se considera un valor nulo ("missing value").

En los procedimientos de análisis de los datos, como con MEANS, TABLES y FREQ se ignorarán los valores nulos si así se estableció en las preferencias del módulo de análisis.

Para ello se debe seleccionar del Árbol de Comandos la orden ESTABLECER de la carpeta PREFERENCIAS:



Se abre una ventana donde se pueden seleccionar todas las preferencias para el análisis de datos:



Allí se puede seleccionar como se representarán los valores especiales, si se mostrarán los criterios de selección, los porcentajes, las tablas de datos.

También la complejidad del análisis estadísticos que se va a realizar y si se van a incluir los valores desconocidos ("Missing").

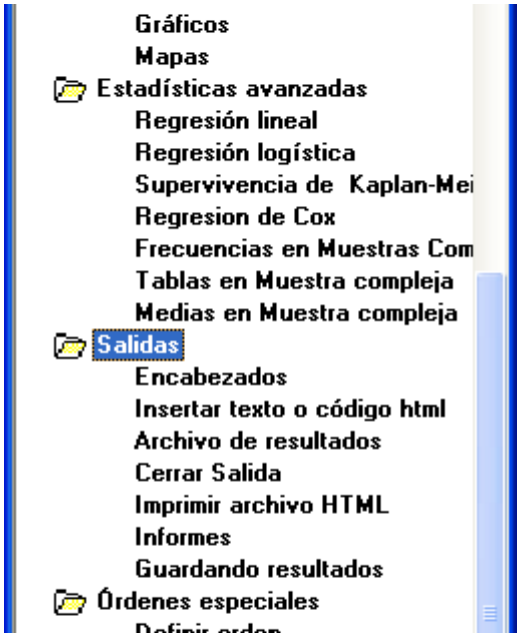
Si se borra un registro, EpiInfo no lo elimina totalmente de la base sino que lo marca como tal. También se puede seleccionar si esos registros se incluyen o no en el análisis. En general no se incluyen.

Si ha utilizado otro código para los valores nulos, como 99, asegúrese de seleccionar sólo los valores no-nulos antes de usar la orden MEANS. Esto se puede hacer usando la orden: SELECT EDAD<>99, por ejemplo. Tenga especial cuidado si los datos han sido importados de otro sistema en el que la codificación de valores nulos puede ser diferente.

SALIDAS DE LOS RESULTADOS.

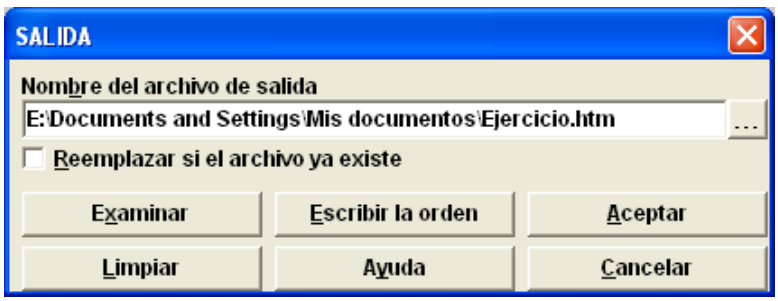
Todos los resultados que se obtengan al procesar los datos pueden enviarse a una archivo para luego editarlos o imprimirlos.

La carpeta SALIDAS del Árbol de Comandos permite seleccionar una serie de puntos para dar formato a esa salida.



Así se puede seleccionar un encabezado para todas las salidas de resultados, insertar textos, generar informes o enviar los resultados a un archivo distinto del que utiliza EpiInfo por defecto.

De esta forma si se selecciona el ítem "Archivo de Resultados" se abre una ventana donde se debe escribir el nombre del archivo donde se guardarán los datos y su ubicación en el directorio de la computadora:



EDITOR DE ÓRDENES

En la parte inferior de la pantalla se encuentra el editor de órdenes. Allí van quedando escritas las órdenes que fuimos generando a través de los botones. Al familiarizarse con las órdenes puede luego escribirlas en forma directa e indicar que corra ese comando haciendo click sobre el botón "Ejecutar esta orden" o "Run this Command" de la versión en inglés.

En ocasiones es necesario realizar operaciones bastante complejas para ser escritas sobre la marcha o que deben ser utilizadas repetidamente durante el análisis de los datos.

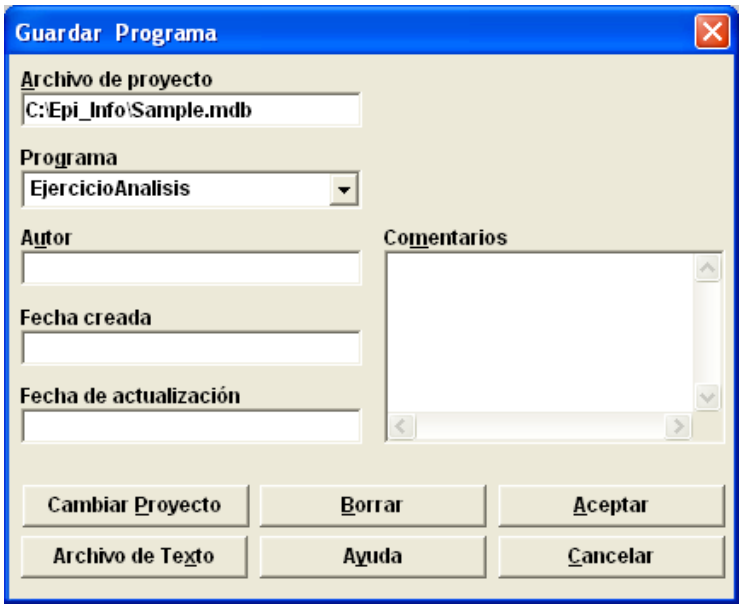
A veces necesitamos hacer un informe periódico de datos o simplemente queremos evaluar la marcha del estudio que estamos llevando a cabo con análisis repetidos tras la incorporación de determinado número preestablecido de pacientes.

En ambos casos, el programa Epi Info permite escribir "programas" que pueden almacenar distintos tipos de órdenes de análisis como crear variables transitorias, visualizar o editar datos, ordenar registros por orden numérico o alfanumérico, seleccionar datos para incluir o excluir registros del análisis, asignar condiciones a las operaciones (como IF - THEN), manejar fechas e intervalos de tiempo, etc.

También es muy útil utilizar un programa para hacer las manipulaciones básicas de un archivo, como leer la base con READ, seleccionar registros con SELECT, recodificar variables creando por ejemplo grupos etarios con RECODE o crear nuevas variables con DEFINE y asignarle valores con IF-THEN para luego hacer al análisis ya sea con ordenes guardadas o utlizando el árbol de ordenes.

Las órdenes almacenadas en el programa podrán ser utilizadas cada vez que "corra" el programa con la orden EJECUTAR (RUN).

Para guardar una serie de órdenes ya utilizadas se presiona el botón guardar para asignar un nombre al programa y elegir el directorio y proyecto donde será guardado.



Con el botón abrir se puede recuperar el archivo para luego hacerlo correr completamente con la orden "Ejecutar".