CAPÍTULO 3

ESTADÍSTICA DESCRIPTIVA

En este capítulo se verán todas las técnicas que se usan para la organización y presentación de datos en tablas y gráficas, así como el cálculo de medidas estadísticas. Se considerarán solamente datos univariados y bivariados.

Ejemplo 3.1 Los siguientes datos provienen de un cuestionario de 10 preguntas que se hizo a 28 estudiantes de una clase de Estadistica Aplicada I en el Recinto Universitario de Mayaguez de la Universidad de Puerto Rico. Un asterisco (*) significa que la pregunta no fue contestada. En lo sucesivo se hará referencia a este conjunto de datos como "**clase97.mtw**"

Row	edad	sexo	escuela	programa	creditos	gpa	familia	hestud	htv
1	21	f	públ	biol	119	3.60	3	35	10
2	18	f	priv	mbio	15	3.60	3	30	10
3	19	f	priv	biot	73	3.61	5	5	7
4	20	f	priv	mbio	*	2.38	3	14	3
5	21	m	públ	pmed	114	3.15	2	25	25
6	20	m	públ	mbio	93	3.17	3	17	6
7	22	m	públ	pmed	120	2.15	5	20	10
8	20	m	priv	pmed	*	3.86	5	15	5
9	20	m	priv	pmed	94	3.19	4	10	2
10	20	f	públ	pmed	130	3.66	6	20	33
11	21	f	priv	mbio	97	3.35	1	15	20
12	20	m	priv	mbio	64	3.17	4	30	2
13	20	f	públ	mbio	*	3.23	2	5	3
14	21	f	públ	mbio	98	3.36	4	15	10
15	21	f	priv	biol	113	2.88	5	15	3
16	21	f	priv	pmed	124	2.80	5	20	10
17	20	f	públ	eagr	*	2.50	4	10	5
18	20	f	priv	mbio	*	3.46	4	18	5
19	22	f	priv	pmed	120	2.74	2	10	15
20	20	f	priv	mbio	95	3.07	3	15	12
21	22	f	priv	biol	125	2.20	3	20	10
22	23	m	públ	eagr	13	2.39	3	10	8
23	21	m	priv	pmed	118	3.05	4	10	10
24	20	f	públ	mbio	118	3.55	5	38	10
25	21	f	públ	mbio	106	3.03	5	36	35
26	20	f	priv	mbio	108	3.61	3	20	10
27	22	f	públ	mbio	130	2.73	5	15	2
28	21	f	priv	pmed	128	3.54	3	18	5

Las variables **edad, familia, hestud y htv** son consideradas como variables cuantitativas discretas. Las variables **créditos y gpa** son consideradas como variables cuantitativas continuas. Las variables **sexo, escuela y programa** son consideradas como variables cualitativas

3.1 Organización de datos Cuantitativos Discretos

3.1.1 Tablas de Frecuencias

Los datos cuantitativos discretos se organizan en tablas, llamadas **Tablas de Distribución de frecuencias.** La primera columna de la tabla contiene los distintos valores que asume la variable ordenados de menor a mayor y las restantes columnas contienen los siguientes tipos de frecuencias.

Frecuencia absoluta: Indica el número de veces que se repite un valor de la variable.

Frecuencia relativa: Indica la proporción con que se repite un valor. Se obtiene dividiendo la frecuencia absoluta entre el tamaño de la muestra. Para una mejor interpretación es más conveniente mutiplicarla por 100 para trabajar con una **Frecuencia relativa porcentual**.

Frecuencia absoluta acumulada: Indica el número de valores que son menores o iguales que el valor dado.

Frecuencia relativa porcentual acumulada: Indica el porcentaje de datos que son menores o iguales que el valor dado.

Para construir una tabla de frecuencias en **MINITAB,** se sigue la secuencia **Stat** ***Tables Tally Individual Variables***. En la ventana de diálogo de ***Tally Individual Variables*** se elige la variable deseada, la cual debe aparecer en la ventanita **Variables.** Se seleccionan todas las opciones de **Display** si se desea una tabla completa con todos los tipos de frecuencias y luego se oprime el botón **OK**. La tabla aparecerá en la ventana **Session.**

En la figura 3.1 se muestra la ventana de diálogo de ***Tally Individual Variables,*** para obtener la tabla de distribución de frecuencias de la variable **familia,** del ejemplo 3.1

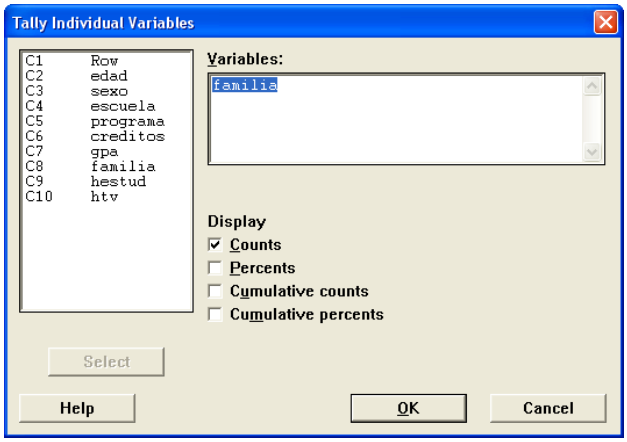


Figura 3. 1 Ventana de diálogo de **Tally Individual Variables** para la tabla de frecuencia de la variable *Familia*

El contenido de la ventana **session** será el siguiente:

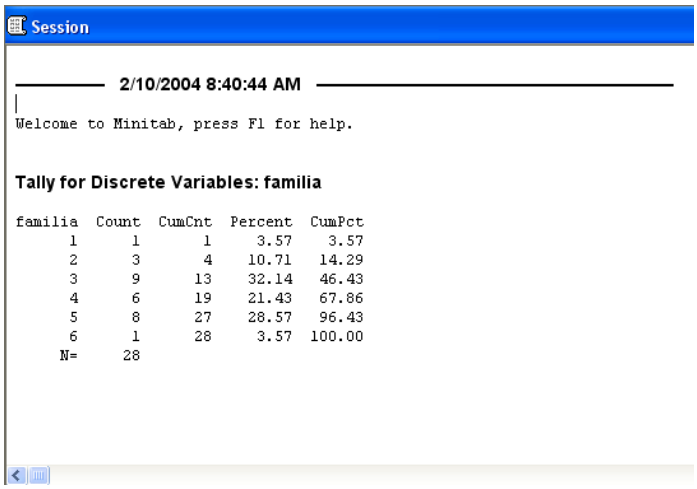


Figura 3.2: Tabla de frecuencia de la variable *Familia*

Interpretación:
Count, representa la frecuencia absoluta. Por ejemplo el tamaño familiar que más predomina es 3.
CumCnt, representa la frecuencia absoluta acumulada.. Por ejemplo 27 de los 28 entrevistados tienen una familia de tamaño menor o igual que 5.
Percent, representa la frecuencia relativa porcentual. Por ejemplo, sólo 3.57 por ciento de las familias de los estudiantes entrevistados son de tamaño 6.
CumPct, representa la frecuencia relativa porcentual acumulada. Por ejemplo, el 94.93% de las familias son de tamaño menor o igual que 5.

3.1.2 El plot de puntos ("Dotplot")

Una vez obtenida la tabla de frecuencia el próximo paso es obtener un gráfica de ella. La gráfica más elemental es el plot de puntos ("Dotplot") que consiste en colocar un punto cada vez que se repite un valor. Esta gráfica permite explorar la simetría y el grado de variabilidad de la distribución de los datos con respecto al centro, el grado de concentración o dispersión de los datos con respecto al valor central y ,también, permite detectar la presencia de valores anormales ("outliers").

En **MINITAB** el plot de puntos se obtiene eligiendo la opción ***Dotplot*** del menú **Graph.** Las ventanas de diálogo para obtener el plot de puntos de la tabla de frecuencias anterior se completará como sigue:

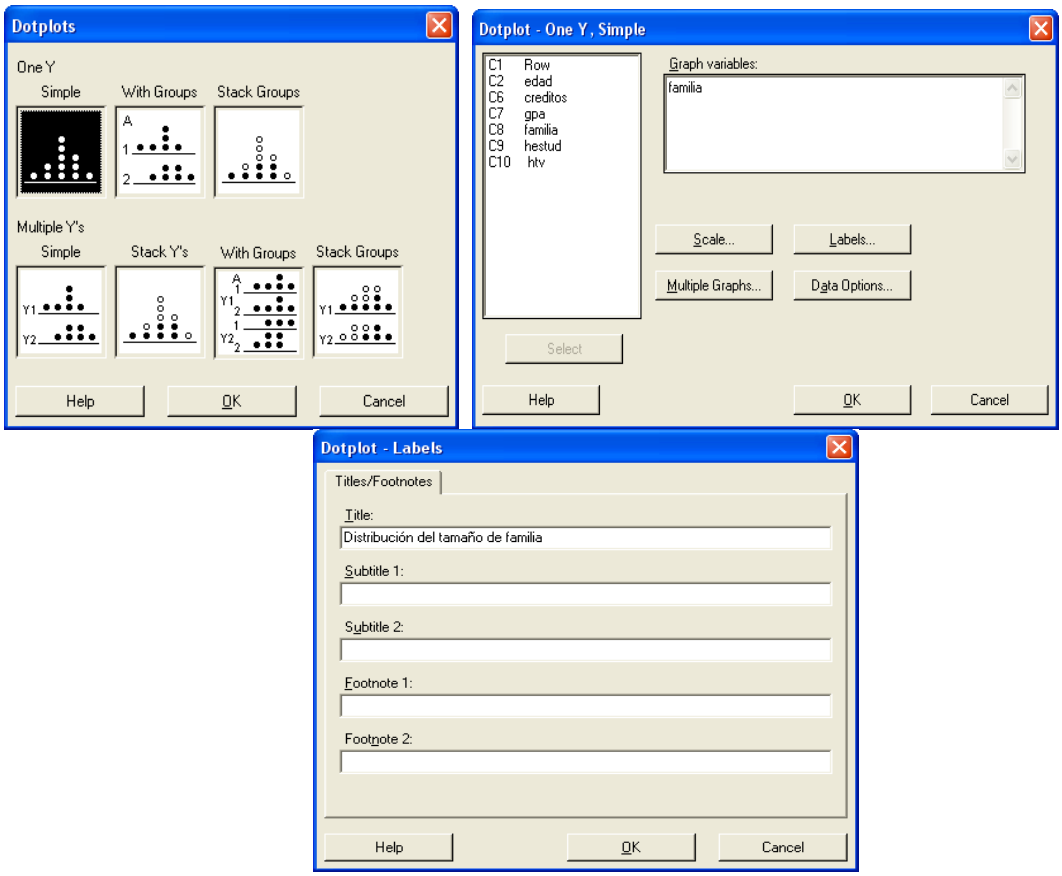


Figura 3.3:Ventanas de diálogo para hallar el dotplot de la variable *Familia*

Obteniéndose la siguiente gráfica:

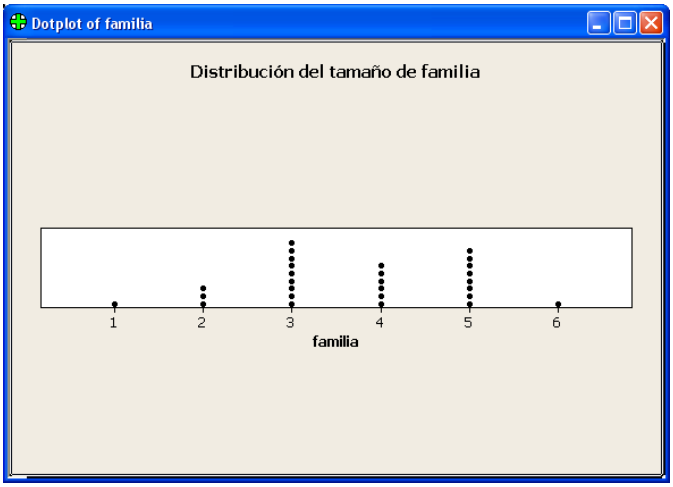


Figura 3.4:Dotplot de la tabla de frecuencia de la variable *Familia*

Interpretación: *La distribución de la variable familia es algo simétrica con respecto al centro. No hay mucha variabilidad y no se observa la presencia de valores anormales.*

También es posible obtener una gráfica de texto del "**Dotplot".** Las gráficas de texto se construyen utilizando caracteres del teclado y no son de alta resolución. Son útiles si se quiere incluir la gráfica como parte de un archivo ASCII o en un correo electrónico a base de texto. Aunque estas gráficas aún están disponibles, ya no aparecen en el menú de **Graph** por defecto. Para añadir la opción de crear gráficas de caracteres al menú de **Graph** utilice la secuencia **Tools > Customize > Commands>Character Graphs** como se muestra a continuación:

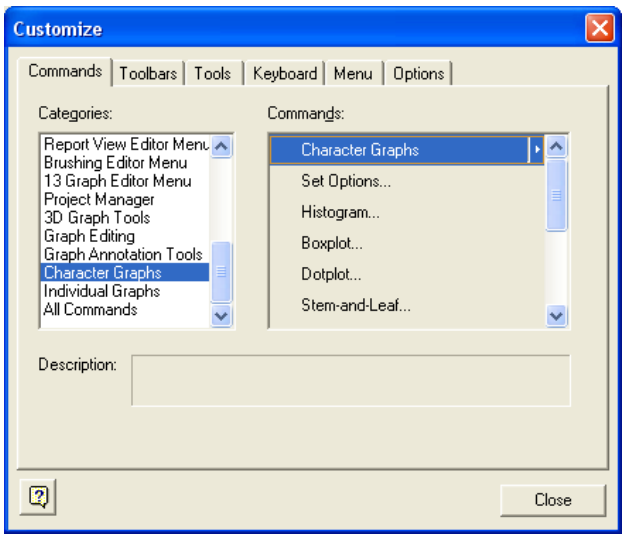


Figura 3.5: Ventana de diálogo para añadir la opción de gráficas de texto

Luego elija el ícono *Character Graphs* de la ventana de **Commands** y arrástrelo hasta el menú de **Graphs** en este caso se sigue la siguiente secuencia **Character Graph ▪ Dotplot** y se obtiene la siguiente salida en la ventana de **Session:**

Dotplot: familia

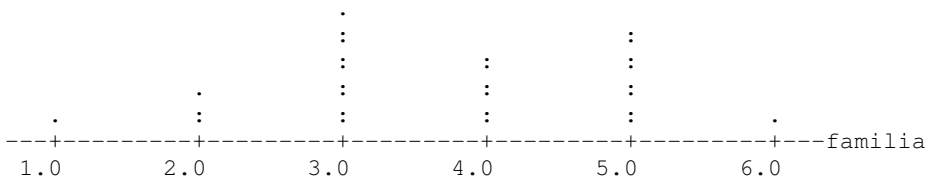


Figura 3.6: Dotplot de la variable *Familia* en modo texto.

3.1.3 Gráfica de Línea

La gráfica de línea es una alternativa a la gráfica de puntos. Por cada valor de la variable se traza una linea vertical de altura proporcional a la frecuencia absoluta del valor de la variable. En **MINITAB** hay una opción directa para obtener esta gráfica la cual será discutida más adelante en la sección 3.2.2.

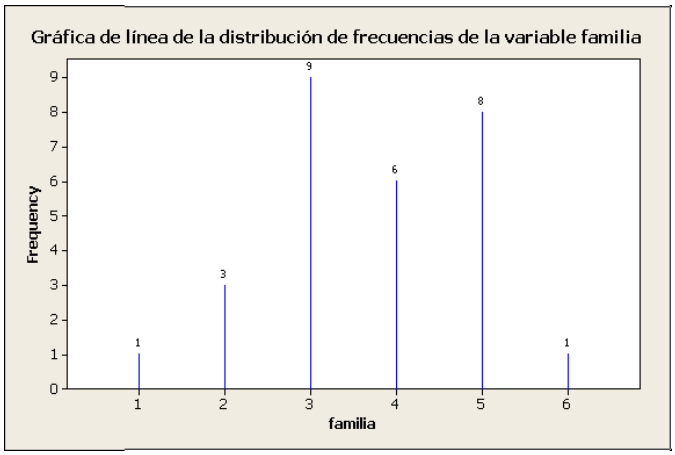


Figura 3.7: Gráfica de línea de la distribución de frecuencias de la variable *familia*

Los números que aparecen en la parte superior de las líneas representan las fecuencias absolutas**.**

Interpretación: *La gráfica tiene algo de simetría, no presenta valores anormales ni tiene mucha variabilidad.*

3.2 Organización de datos Cuantitativos Continuos

Cuando los datos son de una variable continua o de una variable discreta que asume muchos valores distintos, ellos se agrupan en clases que son representadas por intervalos y luego se construye una tabla de frecuencias, cada frecuencia absoluta (relativa porcentual) representa el número (porcentaje) de datos que caen en cada intervalo.

Recomendaciones acerca del número de intervalos de clases:

a) El número de intervalos de clases debe variar entre 5 y 20.
b) Se debe evitar que hayan muchas clases con frecuencia baja o cero, de ocurrir ésto es recomendable reducir el número de clases.
c) A un mayor número de datos le corresponde un mayor número de clases.

Una regla bien usada es que el número de clases debe ser aproximadamente igual a la raíz cuadrada del número de datos. Una vez que se determina el número de clases se determina la amplitud de cada clase usando la siguiente fórmula:

$$\text{Amplitud del intervalo de clase} \approx \frac{\text{Dato mayor - Dato menor}}{\text{número de clases}}.$$

Usualmente la amplitud se redondea a un número cómodo de usar. Si se ha redondeado mucho, entonces el primer intervalo de clase debe empezar un poco antes del valor menor.

MINITAB no tiene una opción para obtener la tabla de frecuencia para datos agrupados, lo único que existe es una opción para obtener la gráfica de la tabla de frecuencias, ésta es llamada ***Histograma*** y puede obtenerse en modo texto o modo gráfico.

3.2.1 Tablas de frecuencias-Histograma en modo texto

La forma de obtener este histograma es eligiendo la opción ***Character Graphs*** del menú **Graph** y luego del submenú que sale se elige ***Histogram***. En la salida aparecerán los puntos medios de los intervalos de clase (llamados también Marcas de clase) y la frecuencia absoluta de cada clase.

Por ejemplo, supongamos que deseamos obtener el histograma de los datos de la variable *gpa*, en el archivo **Clase97.mtw**, agrupando los datos en 5 clases. Primero debemos determinar la amplitud de cada clase, donde $\text{Amplitud} \approx \frac{\text{Dato mayor - Dato menor}}{\text{número de clases}}$. En este caso $\text{Amplitud} \approx \frac{3.86 - 2.15}{5}$ y la primera clase sería: 2.15 - 2.49 con un punto medio igual a 2.32. La ventana de diálogo se completará de la siguiente manera:

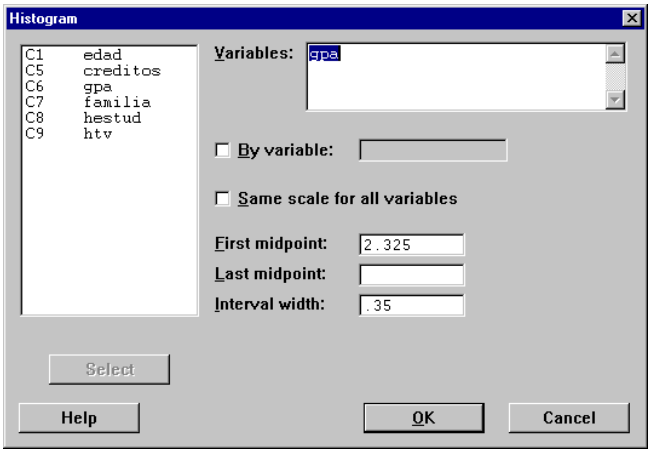


Figura 3.8: Ventana de diálogo para el histograma en modo texto de la variable *gpa*

y en la ventana **session** aparecerá,

```
Histogram
Histogram of gpa   N = 28
Number of observations below the first class = 1

Midpoint        Count
```

```
   2.350          4   ****
   2.700          3   ***
   3.050          8   ********
   3.400          6   ******
   3.750          6   ******
```

donde ***Count*** representa la frecuencia absoluta del intervalo de clase.

3.2.2 Histograma en modo gráfico

Un ***Histograma,*** es la gráfica de la tabla de distribución de frecuencias para datos agrupados, consiste de barras cuyas bases son los intervalos de clases y cuyas alturas son proporcionales a las frecuencias absolutas (o relativas) de los correspondientes intervalos. Un histograma permite ver la forma de la distribucion de los datos, en particular, se puede ver si hay simetría con respecto al centro de la distribución, del grado de dispersión con respecto al centro y permite detectar datos anormales ("outliers") en la muestra. Para obtener un histograma en **MINITAB** se sigue la siguiente secuencia **Graph "Histogram.** Luego, aparece una ventana de diálogo similar a la figura siguiente:

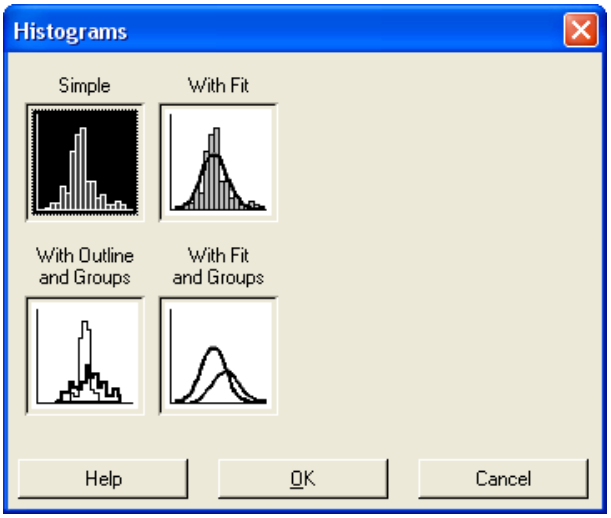


Figura 3.9: Ventana de diálogo para Histograma

En esta ocasión se elije la primera opción y aparece la siguiente ventana:

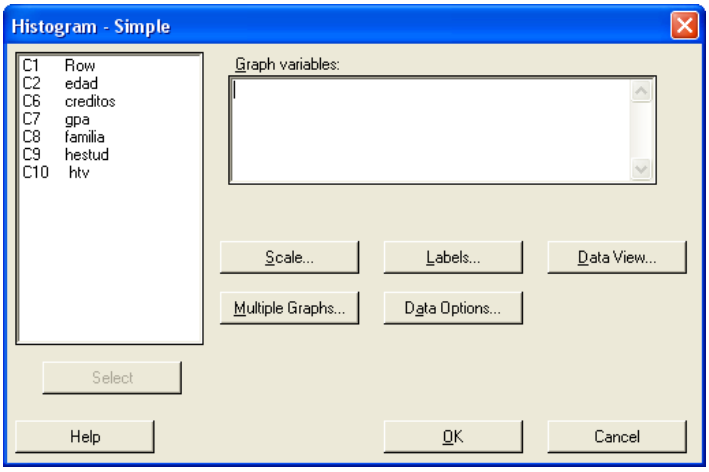


Figura 3.10. Ventana de diálogo para obtener el histograma en modo gráfico de la variable *GPA*.

Graph Variables se escribe la variable cuyo histograma se desea obtener. Si se quiere poner títulos se elige **Labels Titles/Footnotes;** para poner los valores de la frecuencia absoluta encima de cada barra se elige **Labels Data Labels**.

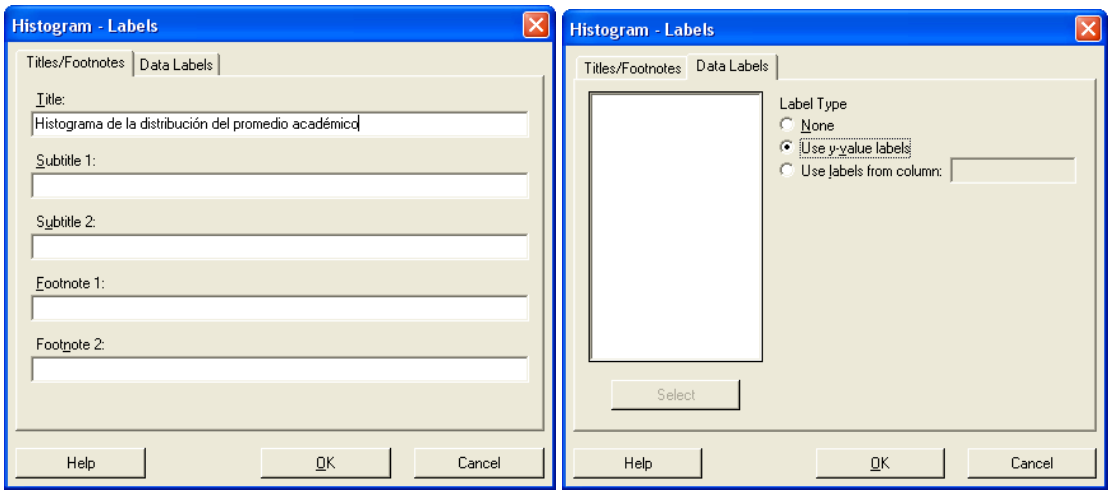


Figura 3.11: Algunas opciones del menú de **Labels**

Oprimiendo **OK** se obtiene el siguiente histograma:

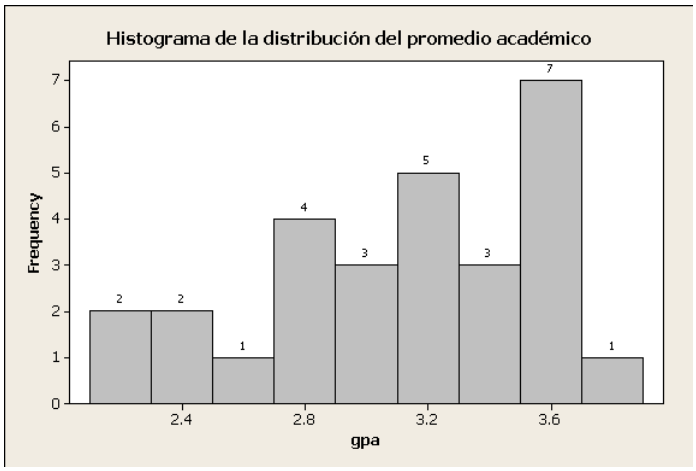


Figura 3.12: El histograma de la variable *gpa*

Interpretación. *El histograma es asimétrico hacia la izquierda. No existe mucha variabilidad, ni hay valores anormales .*

MINITAB elige automáticamente el número de intervalos de clases, si se desea cambiar el número de intervalos de clases, se coloca el cursor en el eje horizontal y se oprime dos veces el botón izquierdo del ratón. Le aparece una ventana de diálogo llamada **Edit Bars.** En esta ventana puede cambiar el color de las barras (***Attributes***) y cambiar el número de intervalos deseado donde aparece ***Binning***. Además se puede entrar los puntos medios de los intrevalos de clase que se desean.

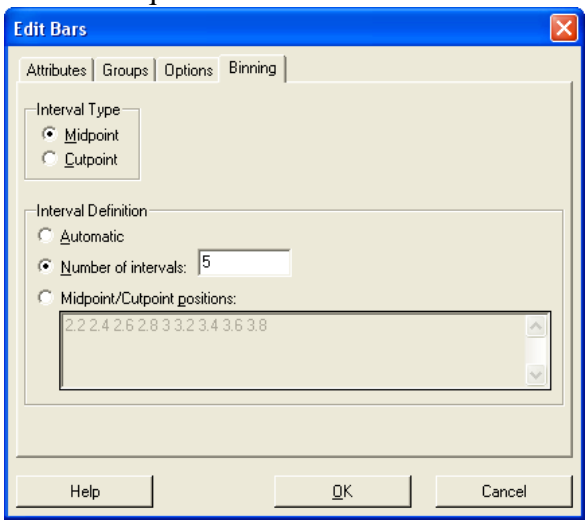


Figura 3.13: Ventana de diálogo para editar un histograma

Para imprimir el Histograma se elige la opción ***Print Graph*** del menú **File.** También es posible obtener el histograma de un conjunto de datos eligiendo la opción **Graph** que aparecen en ciertas ventanas de diálogo, como por ejemplo; cuando se calculan medidas estadísticas básicas.

3.3 Presentación de datos cualitativos

En este caso los datos también se pueden organizar en tablas de frecuencias, pero las frecuencias acumuladas no tienen mucho significado, excepto cuando la variable es ordinal. Para obtener la tabla se sigue la secuencia **STAT ▪ Tables ▪ *Tally*.** Si se desea obtener las frecuencias acumuladas se pueden seleccionar en la ventana Tally.

Por ejemplo, la siguiente sería una tabla de frecuencias para la variable *programa* del Ejemplo 3.1.

programa	Count	Percent
biol	3	10.71
biot	1	3.57
eagr	2	7.14
mbio	13	46.43
pmed	9	32.14
N=	28	

Existen una gran variedad de gráficas para datos cualitativos que se pueden hacer en **MINITAB**. Sólo consideraremos las gráficas de barras y las gráficas circulares ("Pie-Chart").

3.3.1 Gráficas de Barras

Las gráficas de barras pueden ser verticales u horizontales. Las gráficas de barras se obtienen eligiendo la opción **Bar Chart** del menú **Graph.** Aparecerá la ventana de diálogo que se muestra en la primera ventana de la Figura 3.11. Para comenzar, se debe elegir el significado que tendrán las barras y el si se desea una gráfica simple, aglomerado o acumulativa.

Si se desea una gráfica de barras verticales simple, entonces se elige la opción de *Counts of unique variables* como el significado de las barras y simultáneamente la opción *Simple*. Al oprimir *OK*, observará la segunda ventana presentada en Figura 3.14.

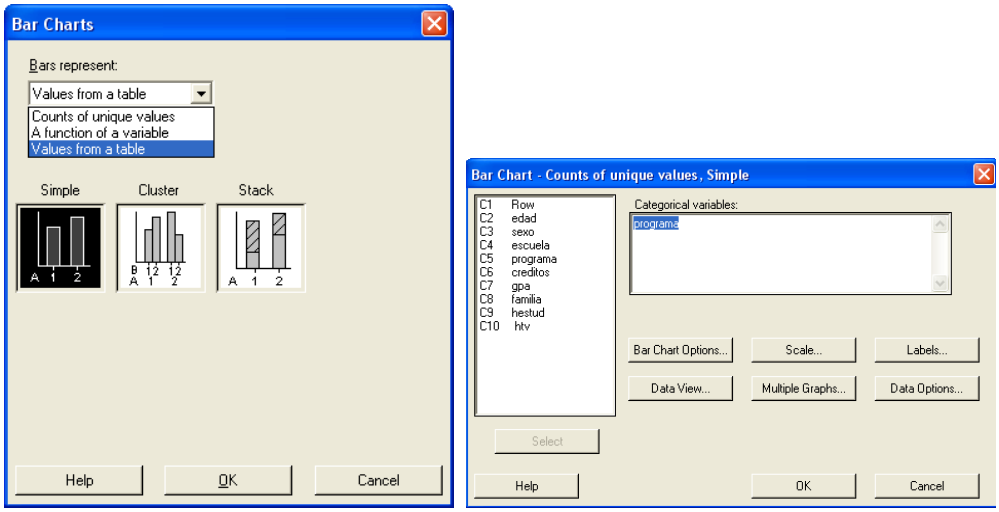


Figura 3.14. Ventanas de diálogo para obtener una gráfica de barras verticales del *programa.*

Ejemplo 3.2. Usando los datos del ejemplo 3.1, hacer una gráfica de barras verticales para representar la distribución de estudiantes por programa.

Se elige las opciones para las primeras dos ventanas de diálogo según se ha descrito en el párrafo anterior. Para colocar el título, en la segunda ventana de diálogo, elija la opción *Labels* y se escribe el título deseado en el renglón titulado *Title*.

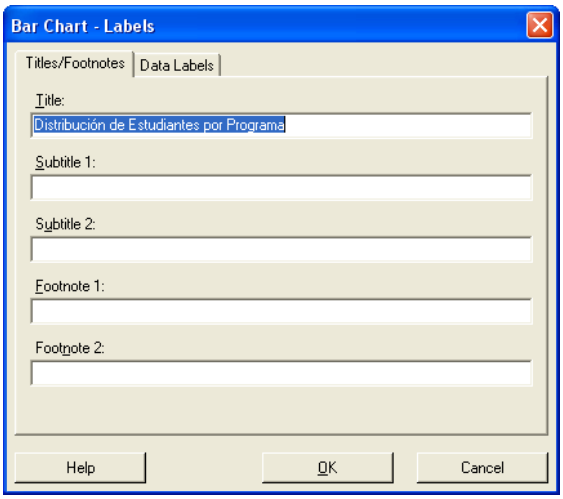


Figura 3.15 Ventana de diálogo para colocar un título a la gráfica de barras.

Al oprimir *OK* dos veces, obtendrá la siguiente gráfica:

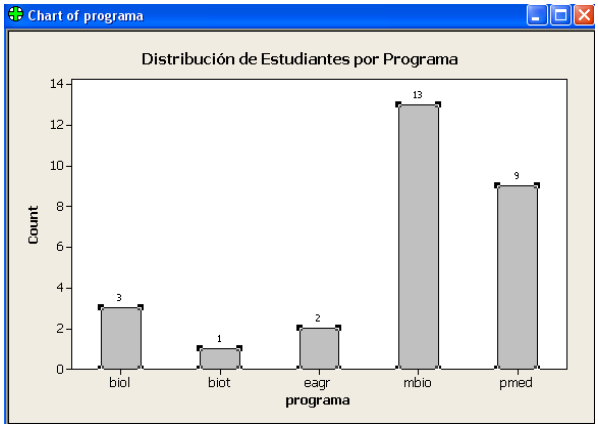


Figura 3.16 Gráfica de Barras verticales de la variable *Programa*

Para hacer una gráfica de barras agrupada, se debe seleccionar *Cluster*, en la primera ventana de diálogo. Luego en la segunda ventana de diálogo, se eligen las variables por las cuales se quiere agrupar. Por ejemplo si deseamos ver la distribución de estudiantes por programa dividido por sexo, elegimos como variable *programa* y luego, *sexo*. Luego de colocar el título, se obtendría la siguiente gráfica:

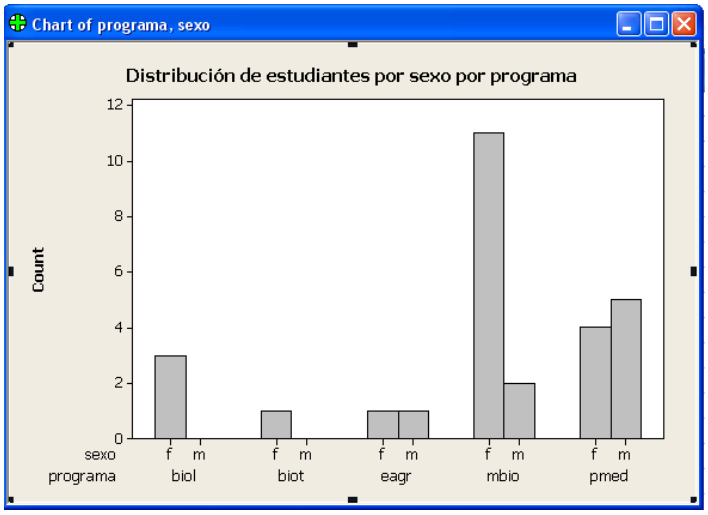


Figura 3.17. Gráfica de barras verticales para la variable programa agrupada por *Sexo*.

3.3.2 Gráficas Circulares

Este tipo de gráfica se usa cuando se quiere tener una idea de la contribución de cada valor de la variable al total. Aunque es usada más para variables cualitativas, también podría usarse para variables cuantitativas discretas siempre que la variable no asuma muchos valores distintos.

Para obtener gráficas circulares se usa la opción ***Pie Chart*** del menú **Graph.** Las ventanas de diálogo de **Pie Chart** que se muestran en la Figura 3.14 son para la variable

programa del Ejemplo 3.1 La gráfica permitirá ver como se distribuyen los estudiantes de la clase según el programa académico.

En **Chart Raw Data** se coloca la variable de la que se quiere hacer el "pie chart". La ventanita de **Chart values from table** se usa sólo en el caso que en una columna estén las categorías de la variable y en la otra la frecuencia con que se repite cada categoría. En la Figura 3.15 se presenta la gráfica de círculo para la variable *programa*.

Existen formas de modificar la gráfica de círculo para enfatizar ciertas ideas. Por ejemplo, se puede resaltar uno o varios pedazos ("slices") mediante el uso de **Explode slice**. Esta opción se logra seleccionando el pedazo(s) que se quiere(n) explotar. Luego, se oprime el botón izquierdo del ratón y se selecciona **Edit Pie**. La ventana de diálogo que se obtiene se muestra en la Figura 3.16. En esta ventana, se puede también modificar el color del pedazo. Si se selecciona la gráfica completa antes de ir a **Edit Pie**, hay la posibilidad de combinar pedazos que contribuyan con un porcentaje muy bajo al total o de colocar el nombre asociada a cada categoría en la gráfica. En la Figura 3.16 b, se muestra la gráfica de la variable *programa* modificada según se ha descrito anteriormente.

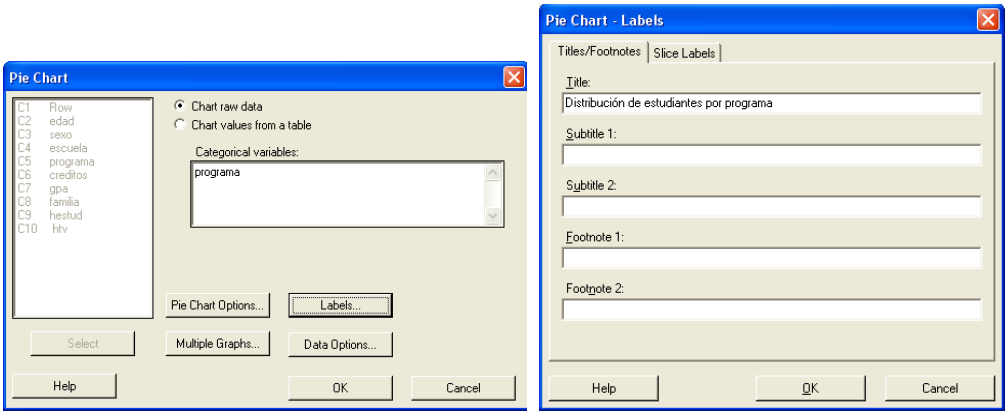


Figura 3.18. Ventanas de diálogo para obtener gráficas circulares

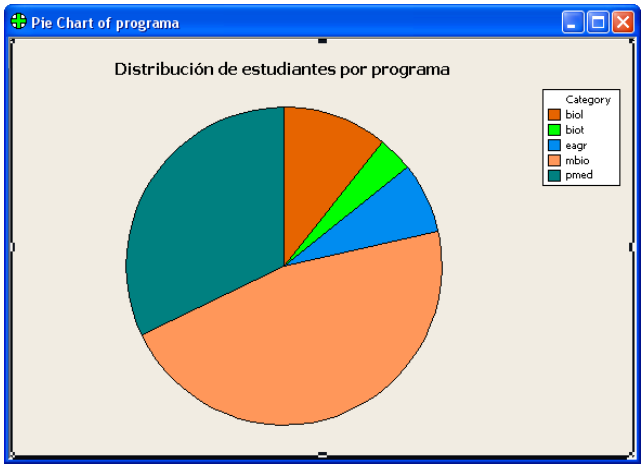
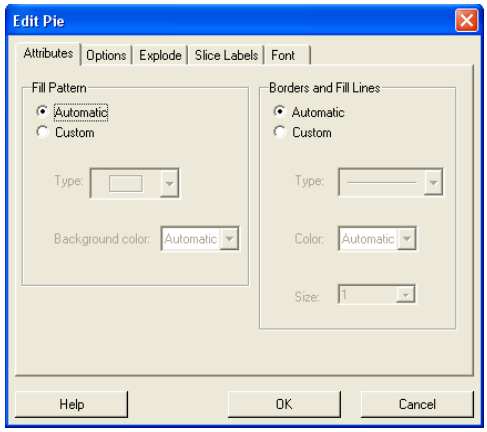


Figura 3.19. Gráfica circular para mostrar la distribución de estudiantes por programa

(a) 

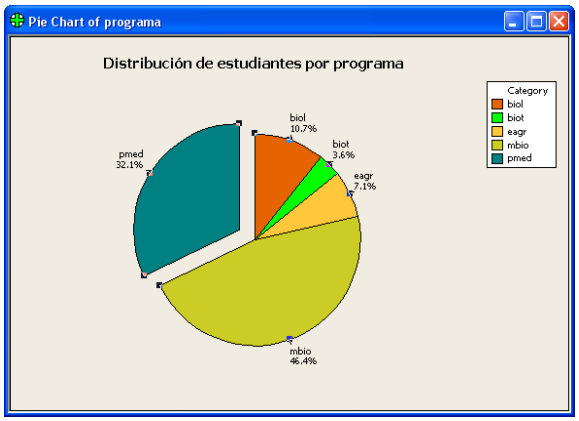
(b) 

Figura 3.20. Ventana de diálogo para modificar la gráfica de la figura 3.19.

Ejemplo 3.3. La siguiente tabla muestra el número de restaurants americanos de comidas rápidas en Puerto Rico a julio de 1997 (Nuevo Día, 31 de Agosto de 1997).

Nombre	**Número**
Burger King	113
McDonald's	97
Taco Maker	63
Kentucky Fried Chicken	58
Pizza Hut	51
Church's	46
Domino's	30
Wendys	24
Taco Bell	22
Ponderosa	21
Little Ceasers	20
Otros	45

Hacer un "Pie-Chart" que muestre qué parte del mercado representa cada franquicia. Enfatizar la franquicia que tiene la mayor parte del mercado y la que tiene la menor parte.

En este caso se elige **Chart values from table**, y en el espacio de **Categorical variable** se coloca el nombre de las columnas que contiene el nombre de los restaurantes y en el rectángulo al lado de **Summary Variables** se coloca la columna que contiene el número de restaurantes de cada tipo. Eligiendo **Labels,** puede indicar el título que tendrá la gráfica y las etiquetas de los pedazos. Al oprimir **ok**, se obtiene la siguiente gráfica:

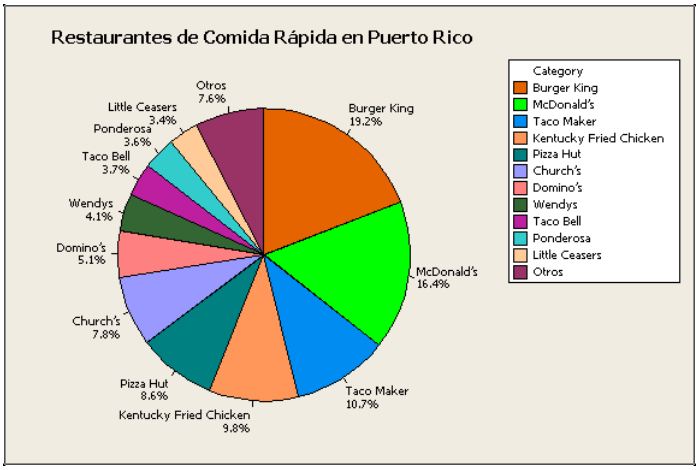


Figura 3.21. Gráfica circular para los datos del ejemplo 3.3

3.4 Gráfica de tallo y hojas ("Stem-and-Leaf ")

La gráfica de tallo y hojas es una gráfica usada para datos cuantitativos. Es la gráfica más básica de un conjunto de técnicas conocido con el nombre de Análisis Exploratorio de Datos (EDA) introducida por John Tukey a mediados de los años 70.
La idea es considerar los primeros dígitos del dato como una rama del tallo ("stem") y el último dígito como una hoja ("leaf") de dicha rama. Las ramas son ordenadas en forma creciente.

Ejemplo 3.4. Los siguientes datos representan pesos de una muestra de 15 varones adultos.
165 178 185 169 152 180 175 189 195 200 183 191 197 208 179
Hacer su gráfica de "Stem-and Leaf".

Solución:
En este caso las ramas la forman los primeros dos dígitos de los datos, y las hojas serán dadas por los últimos dígitos de los datos. Luego el "stem-and leaf " será de la siguiente manera:

15	2
16	59
17	598
18	0935
19	517
20	08

Interpretación: *El uso del "stem-and-leaf" es exactamente igual al del Histograma, la única diferencia está en que del "stem-and-leaf" se pueden recuperar los datos muestrales, pero de un histograma no se puede hacer. En este ejemplo el "stem-and-leaf" es asimétrico a la izquierda, no tiene mucha variabilidad ni "outliers".*

La unidad de la hoja de un "stem-and-leaf" representa la posición del dígito de la hoja en la escala decimal. En el ejemplo anterior el dígito de las hojas está en las unidades luego la unidad de la hoja será 1.0. Si los datos fueran de promedios académicos: 3.17, 3.23, 2.98 entonces, la unidad de la hoja será .01.

Para recuperar los datos de la muestra se juntan las ramas y las hojas del "stem-and-leaf" y se multiplica por la unidad de la hoja.

Hay varias maneras de obtener un "stem-and-leaf" en **MINITAB.** La primera es elegir la opción **stem-and-leaf** del menú **Graph**, la segunda es elegir la opción **Character Graph** del menú **Graph** y luego **stem-and-leaf** del listado que aparece. Finalmente, también se puede elegir la opción **EDA** del menú **Stat** y luego ***Stem-and-Leaf*** del submenú de **EDA.**

La ventana de diálogo para obtener el "stem-and-leaf" de los datos de promedio académico *gpa* del ejemplo 3.1 es como sigue:

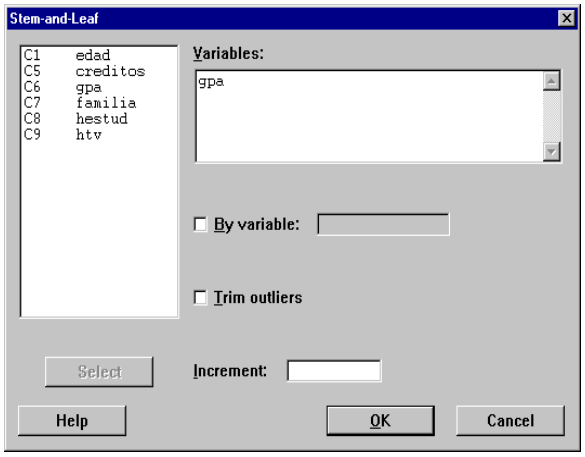


Figura 3.22. Ventana de diálogo para obtener el "stem-and-leaf" de la variable *gpa*

La opción **By variable** se usa cuando se quiere comparar "stem-and-leaf" de dos o más grupos y aqui se escribe la variable que clasifica en grupos.
Si se elige la opción **Trim outliers** en la ventana de diálogo del "stem-and-leaf" se puede detectar los "outliers". La opción **Increment** permite ajustar el número de ramas del **"**stem". En la ventana **session** aparecerá el "stem-and-leaf" de la variable *gpa* que se muestra a continuación.

La unidad de la hoja 0.1 indica la posición de una hoja en la escala decimal. O sea 3 | 6 significa 3.6.

En el ejemplo anterior se han hecho uso de 5 subramas para cada rama principal. Se pueden usar 2 ó 5 subramas por cada rama principal. Si se usa dos subramas, entonces la primera subrama contiene las hojas del 0 al 4 y la segunda las hojas del 5 al 9. En el caso

de 5 subramas, entonces la primera contiene las hojas 0 y 1, la segunda las hojas 2 y 3 y así sucesivamente hasta la quinta que contiene las hojas 8 y 9.

Stem-and-Leaf Display: gpa

```
Stem-and-leaf of gpa  N  = 28
Leaf Unit = 0.10


 1   2  1
 4   2  233
 5   2  5
 7   2  77
 9   2  88
(7)  3  0001111
 12  3  233
 9   3  455
 6   3  66666
 1   3  8
```

Figura 3.23: Ventana de sesión para una gráfica de caracteres de tallo y hoja para la variable *GPA*

Frecuentemente, los programas estadísticos como **MINITAB,** redondean los datos antes de hacer el "stem-and-leaf". Por ejemplo si la muestra contiene los datos,
93 135 178 245 267 342 307, éstos pueden ser redondeados a 90 130 170 240 340 300 y luego el "stem-and-leaf" tendría las ramas 0,1,2 y 3 con unidad de hoja igual a 10.

Ejemplo 3.5 El impuesto por cajetilla de cigarrillos en Puerto Rico es de 83 centavos. Los siquientes datos muestran los impuestos en los 50 estados de los Estados Unidos (Nuevo Dia, 4 de Sept. de 1997)

Estado	**tax**	**Estado**	**tax**
Virg	0.025	DakS	0.330
Kent	0.030	Flor	0.339
CarN	0.050	Nebr	0.340
CarS	0.070	Neva	0.350
Georg	0.120	Iowa	0.360
Wyom	0.120	Mary	0.360
Tenn	0.130	Cali	0.370
Indi	0.155	Maine	0.370
Alab	0.165	Oreg	0.380
Misso	0.170	NewJ	0.400
WestV	0.170	Texas	0.410
Missi	0.180	Wisco	0.440
Mont	0.180	Illin	0.440
Colo	0.200	DakN	0.440
Lousi	0.200	Verm	0.440
NMexi	0.210	Minn	0.480
Oklah	0.230	Conn	0.500
Delaw	0.240	NewY	0.560
Kans	0.240	Ariz	0.580
Ohio	0.240	Hawa	0.600

NHans	0.250	RhodI	0.610
Utah	0.265	WasDC	0.650
Idaho	0.280	Michi	0.750
Alask	0.290	Massa	0.760
Penn	0.310	Washi	0.825
Arka	0.315		

Hacer un "stem-and-leaf" de los datos.

Solución: Usaremos la opción **Trim** de **Stem-and-Leaf** para detectar "outliers".

Stem-and-Leaf Display: tax

```
Stem-and-leaf of tax  N  = 51
Leaf Unit = 0.010


 2   0  23
 4   0  57
 7   1  223
 13  1  567788
 20  2  0013444
 24  2  5689
(5)  3  11334
 22  3  566778
 16  4  014444
 10  4  8
 9   5  0
 8   5  68
 6   6  01
 4   6  5
 3   7
 3   7  56


HI 82
```

Interpretación: *El "stem-and-leaf" indica mucha variabilidad y asimetría hacia la derecha. Además, el estado de Washington representa un "outlier" superior. La unidad de la hoja es .01, o sea 3 | 7 representa 0.37. Se han usado dos subramas por cada rama principal*

3.5 Cálculo de Medidas Estadisticas

Hay dos tipos principales de Medidas Estadísticas: Medidas de Tendencia Central y Medidas de Variabilidad.

Las medidas de tendencia central dan una idea del centro de la distribución de los datos. Las principales medidas de este tipo son la media o promedio aritmético, la mediana, la moda y la media podada.

Las medidas de variabilidad expresan el grado de concentración o dispersión de los datos con respecto al centro de la distribución. Entre las principales medidas de este tipo están la varianza, la desviación estándar, el rango intercuartílico. También hay medidas de posición, como son los cuartiles, deciles y percentiles. Además, una medida de asimetría ("skewness") y una medida de aplanamiento ("kurtosis").

3.5.1 Medidas de Centralidad

La media o promedio se obtiene sumando todos los datos y dividiendo entre el número de datos. Es decir, si $x_1, x_2,\ldots,x_n$, representan las observaciones de una variable X en una muestra de tamaño n, entonces la media de la variable X está dada por:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Ejemplo 3.6. Supongamos que los siguientes datos representan el precio de 9 casas en miles.

74, 82, 107, 92, 125, 130, 118, 140, 153

Hallar el precio promedio de las casas.

Solución:

$$\bar{x} = \frac{74+82+107+92+125+130+118+140+153}{9} = 113.4$$

Es decir que el costo promedio de una casa será 113,400.

La media es afectada por la asimetría de la distribución de los datos y por la presencia de "outliers" como se muestra en el siguiente ejemplo.

Ejemplo 3.7. Supongamos que en el ejemplo anterior se elige adicionalmente una casa cuyo precio es de 500,000.
Luego el promedio será:

$$\bar{x} = \frac{74+82+107+92+125+130+118+140+153+500}{10} = 152.1$$

En este caso la media da una idea errónea del centro de la distribución, la presencia del "outlier" ha afectado la media. Sólo dos de las 10 casas tienen precio promedio mayor de 152,100.

Otras propiedades de la media son:

a) Que el valor de la media debe estar entre el mayor dato y el menor dato.
b) Si a cada dato de la muestra se les suma (o resta) una constante entonces, la media queda sumada (o restada) por dicha constante.
c) Si a cada dato de la muestra se le multiplica (o divide) por una constante entonces, la media queda multiplicada (o dividida) por dicha constante.

Las propiedades b) y c) se usan para hacer cálculos rápidos de la media.

La mediana es un valor que divide a la muestra en dos partes aproximadamente iguales. Es decir, como un 50 por ciento de los datos de la muestra serán menores o iguales que la mediana y el restante 50 por ciento son mayores o iguales que ella.

Para calcular la mediana primero se deben ordenar los datos de menor a mayor. Si el número de datos es impar, entonces la mediana será el valor central. Si el número de datos es par entonces, la mediana se obtiene promediando los dos valores centrales.

Ejemplo 3.8. Calcular la mediana de los datos del Ejemplo 3.6.

Solución:
Ordenando los datos en forma ascendente, se tiene: 74, 82, 92, 107, 118, 125, 130, 140, 153. En este caso el número de datos es impar así que la mediana resulta ser 118 que es el quinto dato ordenado.

A diferencia de la media, la mediana no es afectada por la presencia de valores anormales, como lo muestra el siguiente ejemplo:

Ejemplo 3.9. Calcular la mediana de los datos del Ejemplo 3.7.

Solución:
Ordenando los datos, se tiene:
74, 82, 92, 107, 118, 125, 130, 140, 153, 500.
en este caso el número de datos es par, así que la mediana resulta ser el promedio de los dos valores centrales: $\frac{118+125}{2}$=121.5 y el dato anormal 500 no afecta el valor de la mediana.

Cuando la distribución es asimétrica hacia la derecha, la mediana es menor que la media. Si hay asimetría hacia la izquierda entonces la mediana es mayor que la media y cuando hay simetría, ambas son iguales.

La moda es el valor (o valores) que se repite con mayor frecuencia en la muestra. La Moda puede aplicarse tanto a datos cuantitativos como cualitativos.

Ejemplo 3.10. Los siguientes datos representan el número de veces que 11 personas van al cine mensualmente:

3, 4, 4, 5, 0, 2, 1, 5, 4, 5, 4

Hallar la moda.

Solución:
La Moda es 4. O sea que predominan más las personas que asisten 4 veces al mes al cine.

Ejemplo 3.11. Los siguientes datos representan tipos de sangre de 9 personas

A, O, B, O, AB, O, B, O, A

Hallar la Moda.

Solución:
La Moda es el tipo de sangre O.

La media podada es una medida más resistente que la media a la presencia de valores anormales. Para calcular la Media Podada, primero se ordenan los datos en forma creciente y luego se elimina un cierto porcentaje de datos (redondear si no da entero) en cada extremo de la distribución, finalmente se promedian los valores restantes.

Ejemplo 3.12. Hallar la media podada del 5 por ciento para los datos del Ejemplo 3.9.
Solución:
El 5 por ciento de 10 datos es .5 que redondeando a 1 implica que hay que eliminar el mayor (500) y el menor (74) dato. Luego la media podada del 5 por ciento será

$$\frac{82+92+107+118+125+130+140+153}{8} = 118.375 .$$

3.5.2 Medidas de Variabilidad

El rango o amplitud es la diferencia entre el mayor y menor valor de la muestra. Mientras mayor sea el rango existe mayor variabilidad. Lamentablemente el rango es bien sensible a la presencia de "outliers".

La varianza es una medida que da una idea del grado de concentración de los datos con respecto a la media.
De primera intención una medida para determinar el grado de concentración de los datos sería el promedio de las desviaciones con repecto a la media, es decir $\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n}$, pero se

puede mostrar que la suma de las desviaciones es cero, ya que las desviaciones positivas y negativas se compensan, luego la anterior medida de variabilidad sería siempre 0.

La siguiente tabla ilustra lo anteriormente mencionado para un conjunto de datos.

	X	$x-\bar{x}$
	5	-6
	8	-3
	12	1
	17	6
	14	3
	10	-1
Sumas	66	0

La media de la muestra es 11.

Si se cuadran las desviaciones se soluciona este problema y es así que aparece la varianza. La varianza de una muestra de n datos se calcula por:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Se divide por n-1 y no por n, porque se puede demostrar teóricamente que cuando se hace esto s^2 estima más eficientemente a la varianza poblacional

Alternativamente se puede usar la fórmula:

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

Es bastante riesgoso usar solamente el valor de la varianza para concluir que la muestra es muy o poco variable. Su uso es más que todo para comparar la variabilidad de dos o más conjuntos de datos de la misma variable en estudio. Además la varianza tiene el problema de que está expresada en unidades cuadráticas en relación a la medida de los datos tomados.

La desviación estándar es la raíz cuadrada positiva de la varianza y tiene la ventaja que está en las mismas unidades de medida que los datos. Se representa por s.
De por si sola la desviación estándar no permite concluir si la muestra es muy variable o poco variable. Al igual que la varianza es usada principalmente para comparar la variabilidad entre grupos.

Ejemplo 3.13. Las muestras siguientes:

```
muestra1
16    18     25     28     23     42     24     47     38     19
22    34

muestra2
116   118    125    128    123    142    124   147   138     119
122   134
```

tienen medias 28 y 128 respectivamente, e igual desviación estándar s = 10.018. O sea que se puede decir en términos absolutos que tienen igual variabilidad. Sin embargo comparándola con los datos tomados se puede concluir que la muestra 1 es bastante variable, mientras que la muestra 2 es poco variable.

Existe una medida llamada **coeficiente de variación** (CV) y que se calcula por $CV = \frac{s}{\bar{x}} \times 100\%$. Si el CV es mayor que 30% la muestra es muy variable y si CV<30% entonces no existe mucha variabilidad. Para el ejemplo el CV para la muestra 1 es 35.77 y para la muestra 2 es 7.82 concluyéndose que la muestra 1 es bastante variable y la muestra 2 no lo es.

Criterio para detectar "outliers".
Un primer criterio para identificar si un dato es un "outlier" es el siguiente:
Un dato que cae fuera del intervalo $(\bar{x} - 3s, \bar{x} + 3s)$ puede ser considerado un "outlier".

Aún así el criterio no es muy confiable, puesto que la media, la varianza y la desviación estándar son afectadas por la presencia de "outliers".

Ejemplo 3.14. Dada la siguiente muestra

59, 62, 73, 79, 68, 77, 69, 71, 66, 98, 75

Determinar si 98 es un "outlier".

Solución:
Como $\bar{x} = 72.45$ y s=10.43. Se tiene que si un dato cae fuera del intervalo (41.15, 103.75) será considerado un "outlier", 98 cae dentro de dicho intervalo por lo tanto no es "outlier".

3.5.3. Medidas de Posición.

Los Cuartiles: Son valores que dividen a la muestra en 4 partes aproximadamente iguales. El 25% de los datos son menores o iguales que el cuartil inferior o primer cuartil, representado por Q_1. El siguiente 25 % de datos cae entre el cuartil inferior y la mediana, la cual es equivalente al segundo cuartil. El 75 % de los datos son menores o iguales que

el cuartil superior o tercer cuartil, representado por Q_3, y el restante 25% de datos son mayores o iguales que Q_3.

Para calcular los cuartiles simplemente se ordenan los datos y luego Q_1 es la mediana de la primera mitad, o sea aquella que va desde el menor valor hasta la mediana. Similarmente Q_3 es la mediana de la segunda mitad, o sea aquella que va desde la mediana hasta el mayor valor.

Ejemplo 3.15. Calcular los cuartiles de las siguientes muestras:

a) 6, 8, 4, 12, 15, 17, 23, 18, 25, 11

Los datos ordenados serán: 4, 6, 8, 11, 12, 15, 17, 18, 23, 25
La primera mitad es: 4, 6, 8, 11, 12, luego $Q_1 = 8$
La segunda mitad es: 15, 17, 18, 23, 25, luego $Q_3 = 18$

b) 10, 22, 17, 13, 28, 40, 29, 18, 23, 39, 44

Los datos ordenados serán: 10, 13, 17, 18, 22, 23, 28, 29, 39, 40, 44
La primera mitad es: 10, 13, 17, 18, 22, 23, luego $Q_1 = \frac{17+18}{2} = 17.5$
La segunda mitad es: 23, 28, 29,39, 40, 44, luego $Q_3 = \frac{29+39}{2} = 34$

Una variante en este último caso es no usar la mediana. Es decir considerar que la primera mitad es 10, 13, 17, 18, y 22 y la segunda mitad es 28, 29, 39, 40, y 44. Así Q_1 sería 17 y Q_3 sería 39. Existen otros métodos de calcular cuartiles, por ejemplo MINITAB usa un proceso de interpolación para calcularlos.
A la diferencia de Q_3 y Q_1 se le llama **Rango Intercuartílico**, ésta es una medida de variabilidad que puede ser usada en lugar de la desviación estándar, cuando hay "outliers".

Los Deciles: Son valores que dividen a la muestra en 10 partes iguales

Los Percentiles: Dado un cierto porcentaje 100p, donde p varía entre 0 y 1, el percentil del 100p% es un valor tal que 100p% de los datos caen a la izquierda del percentil. En particular, la mediana y los cuartiles son percentiles. El primer cuartil es el percentil de 25%, la mediana es el percentil del 50% y el tercer cuartil es el percentil del 75%.

3.5.4 Cálculo de medidas estadísticas usando MINITAB.

En **MINITAB** se pueden calcular simultáneamente varias medidas estadísticas de centralidad y de variabilidad para un conjunto de datos, para esto se elige la opción

Display Descriptive Statistics del submenú de **Basic Statistics** del menú **STAT.** La ventana de diálogo de ***Display Descriptive Statistics*** para calcular las medidas estadísticas de la variable *gpa* del Ejemplo 3.1 según sexo aparece de la siguiente manera:

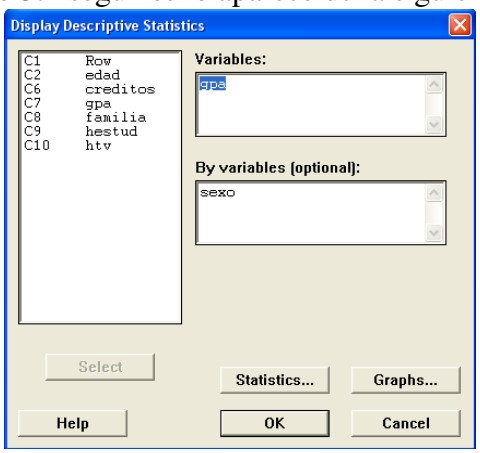


Figura 3.24. Ventana de diálogo para calcular medidas estadísticas de la variable *gpa*, clasificada por *sexo*.

Los resultados aparecerán en la ventana **Session,** como sigue:

Descriptive Statistics: gpa

Variable	sexo	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
gpa	f	20	0	3.145	0.103	0.463	2.200	2.755	3.290	3.588
	m	8	0	3.016	0.187	0.528	2.150	2.555	3.160	3.185

Variable	sexo	Maximum
gpa	f	3.660
	m	3.860

Donde:

N representa el número de datos;

N* representa en número de datos perdidos**,**

Mean, la media muestral;

Median, la Mediana;

Tr Mean, la media podada del 5 por ciento;

StDev, la desviación Estándar;

SE Mean, el error estándar de la Media Muestral, o sea $\frac{s}{\sqrt{n}}$ y los valores restantes representan el **Mínimo,** el **Máximo** y los **cuartiles superior (Q_3) e inferior (Q_1)** de cada variable.

Si se oprime el botón **Graphs** antes de oprimir OK en la ventana de diálogo anterior se obtiene la siguiente ventana de diálogo que permite hacer histogramas, "individual value plot", y "boxplot".

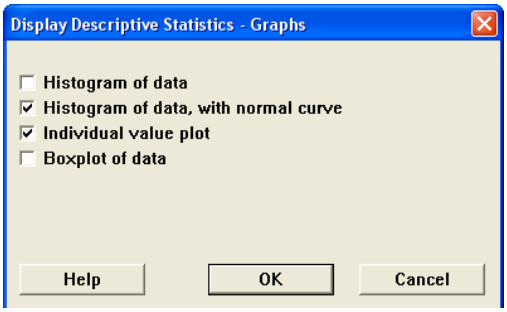


Figura 3.25. Ventana de diálogo de la opción **Graph** de **Display Descriptive Statistics**

Al OK dos veces se obtendrán los siguientes resultados:

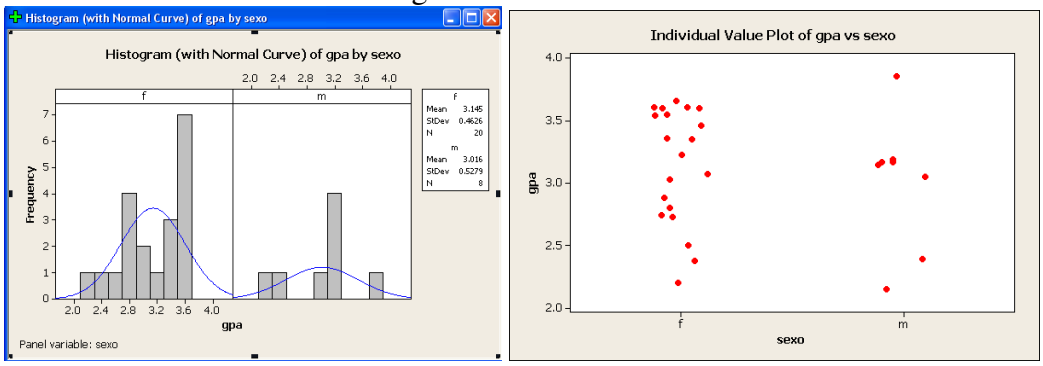


Figura 3.26. Gráficas del Histograma con la curva Normal y un "Individual Value Plot"

También es posible obtener un resumen gráfico del conjunto de datos eligiendo ***Stat-> Basic Statistics -> Graphical Summary***. Los resultados que ofrece Minitab son:

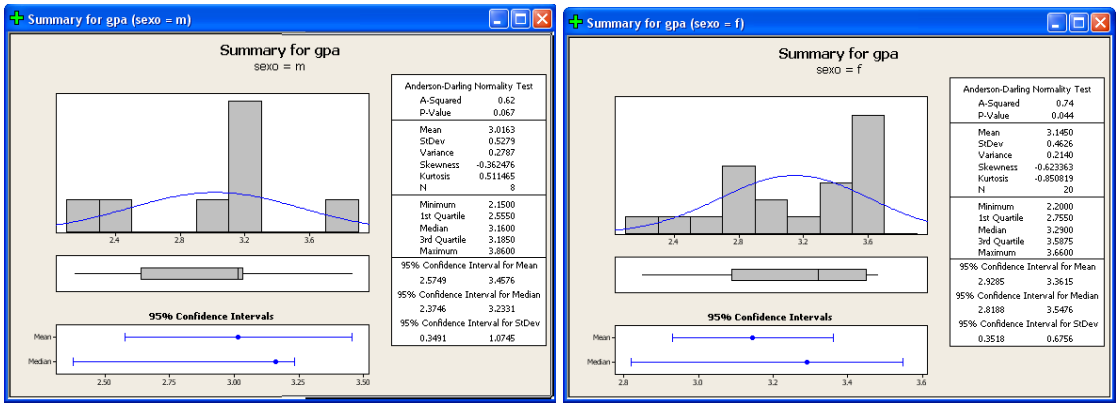


Figura 3.27. Resultados de pedir *Graphical Summary*

Es posible guardar los valores de varias medidas estadísticas en columnas, para esto se elige la opción **Store Descriptive Statistics** del submenú **Basic Statistics.** Al oprimir la opción **Statistics** sale un listado de medidas estadisticas que pueden ser guardadas. Las ventanas de diálogo se muestran a continuación:

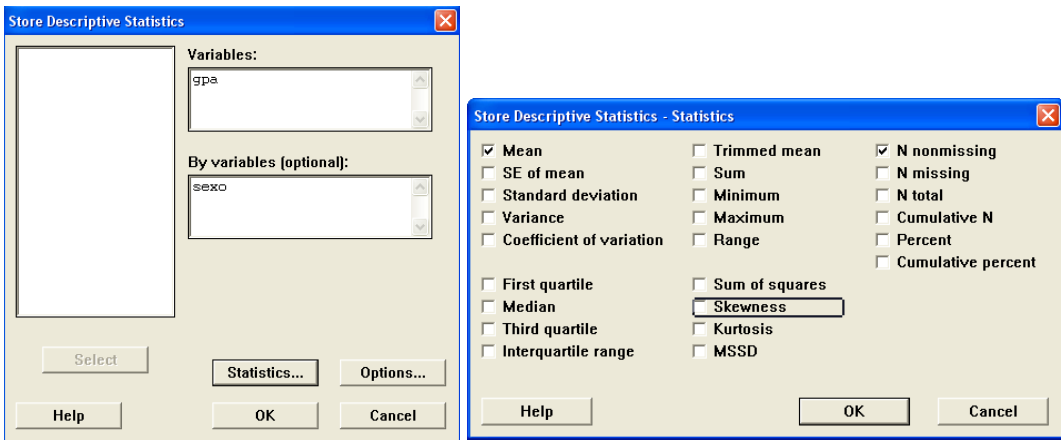


Figura 3.28. Listado de todas las medidas estadísticas que pueden calucularse con MINITAB

Finalmente, también es posible obtener medidas estadísticas, eligiendo la secuencia **CALC "Columns Statistics.**

3.6 El Diagrama de Caja ("Boxplot")

El "**Boxplot"** es una importante gráfica del Análisis Exploratorio de Datos. Al igual que el histograma y el "stem-and-leaf", permite tener una idea visual de la distribución de los datos. O sea, determinar si hay simetría, ver el grado de variabilidad existente y finalmente detectar "outliers". Pero además, el "Boxplot" es bien útil para comparar grupos, es una alternativa gráfica a la prueba estadística t de Student, si se comparan dos grupos o la prueba F del análisis de varianza si se comparan más de dos grupos. Todo lo anterior es posible debido a que se puede hacer múltiples boxplots en una misma gráfica, en cambio los histogramas y "stem-and- leaf" salen en secuencia uno por página.

En **MINITAB** hay varias maneras de obtener el "Boxplot" de un conjunto de datos, la primera es eligiendo la opción ***Boxplot*** del menú **Graph.** En la Figura 3.24 se muestra las ventanas de diálogo para obtener el boxplot de la variable ***creditos*** de los datos del Ejemplo 3.1.

La variable **Y** que aparece debajo de **Graph Variables** es aquella de la cual se desea obtener el "Boxplot", y la variable **X** es usada solo en el caso que se quiera comparar varios grupos usando sus "boxplots". Por ejemplo **X** puede ser: Sexo de la persona, método de Enseñanza, etc.

En **Annotation** se puede poner título, notas al pie, marcar la mediana y también los "outliers"

En **Options** se puede elegir ***Transpose X by Y*** para sacar el boxplot en forma horizontal.

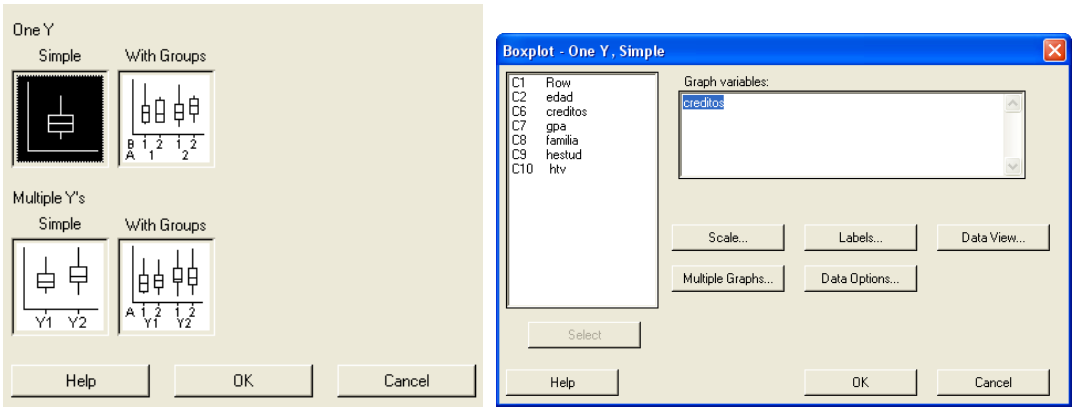


Figura 3.29. Ventanas de diálogo para hallar el Boxplot

El boxplot que se obtiene se muestra a continuación.

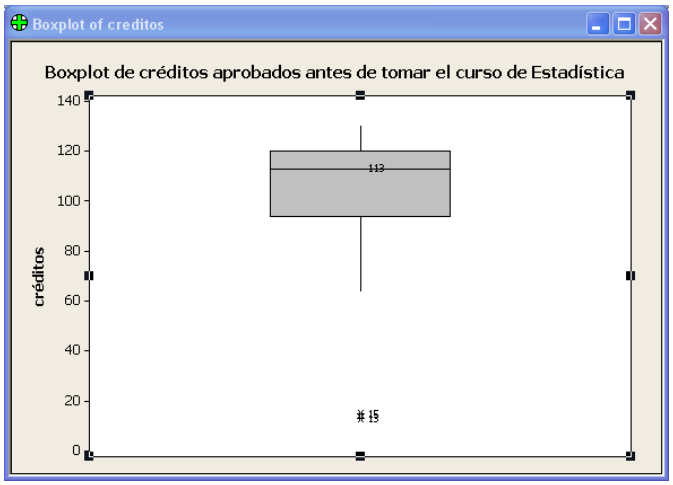


Figura 3.30. "Boxplot" para la variable créditos del Ejemplo 3.1

Interpretación: *La línea central de la caja representa la Mediana y los lados de la caja representan los cuartiles. Si la Mediana está bien al centro de la caja, entonces hay simetria. Si la Mediana está más cerca a Q_3 que a Q_1 entonces la asimetría es hacia la izquierda, de lo contrario la asimetría es hacia la derecha. Si la caja no es muy alargada entonces se dice que no hay mucha* variabilidad.

Si no hay "outliers" entonces las líneas laterales de la caja llegan hasta el valor mínimo por abajo, y hasta el valor máximo por arriba. Cuando hay "outliers" entonces éstos aparecen identificados en la figura y las lineas laterales llegan hasta los valores adyacentes a las fronteras interiores. Si las lineas laterales son bastantes alargadas entonces significa que los extremos de la distribución de los datos se acercan lentamente al eje X.

Las fronteras interiores se calculan como Q_1 - 1.5RIQ y Q_3 + 1.5RIQ respectivamente, donde RIQ = Q_3-Q_1 es el Rango Intercuartílico. Las fronteras exteriores se calculan por Q_1 - 3RIQ y Q_3 + 3RIQ. Si un valor cae más alla de las fronteras exteriores se dice que es **un "outlier" extremo**, en caso contrario el outlier es **moderado.** Un "outlier" moderado se representa por ***** y uno extremo por **0**.

En el "boxplot" de ***créditos*** la mediana es 113, y hay dos "outliers" inferiores 13 y 15. Hay asimetría hacia la izquierda y no hay mucha variabilidad.

Una segunda manera de obtener un "boxplot" es eligiendo la opción **Character Graphs** del menú **Graph** y luego **boxplot** del listado que aparece. En este caso el "boxplot" es de modo texto. Pero aquí se puede notar que los "outliers" son extremos.

```
MTB > BoxPlot 'créditos'.

Boxplot

                                                 -----------
               OO                    ------------I      +  I----
                                                 -----------
              +---------+---------+---------+---------+---------+------
créditos
          0        25        50        75       100       125
```

Otra alternativa de hacer un boxplot en **MINITAB** es elegir la opción **EDA** del menú **Stat** y luego seleccionar **boxplot** del listado que aparece. Aqui el boxplot que resulta es de modo gráfico.

3.7 Organización y Presentación de datos Bivariados

3.7.1 Datos bivariados categóricos.

Para organizar datos de dos variables categóricas o cualitativas se usan tablas de doble entrada. Los valores de una variable van en columnas y los valores de la otra variable van en filas. Para hacer esto en **MINITAB** se elige la opción ***Tables*** del menú **Stat**. y luego la opción ***Cross Tabulation*** del submenú de**Tables**.

Hay dos maneras de usar ***Cross Tabulation*** dependiendo de como se han entrado los datos. Primero, cuando los datos de cada variable están dados en dos columnas distintas. O sea, como si hubiesen sido las contestaciones de un cuestionario.

Ejemplo 3.16. Supongamos que deseamos establecer si hay relación entre las variables tipo de escuela superior y la aprobación de la primera clase de matemáticas que toma el estudiante en la universidad, usando los datos de 20 estudiantes que se muestran abajo:

Est	escuela	aprueba	Est	escuela	aprueba
1	priv	si	11	públ	si
2	priv	no	12	priv	no
3	públ	no	13	públ	no
4	priv	si	14	priv	si
5	públ	si	15	priv	si
6	públ	no	16	públ	no
7	públ	si	17	priv	no
8	priv	si	18	públ	si
9	públ	si	19	públ	no
10	priv	si	20	priv	si

Asumiendo que los datos son entrados en dos columnas: C1: Escuela y C2: aprueba, la ventana de diálogo de ***Cross Tabulation and Chi-Square*** se completerá como aparece en la siguiente figura:

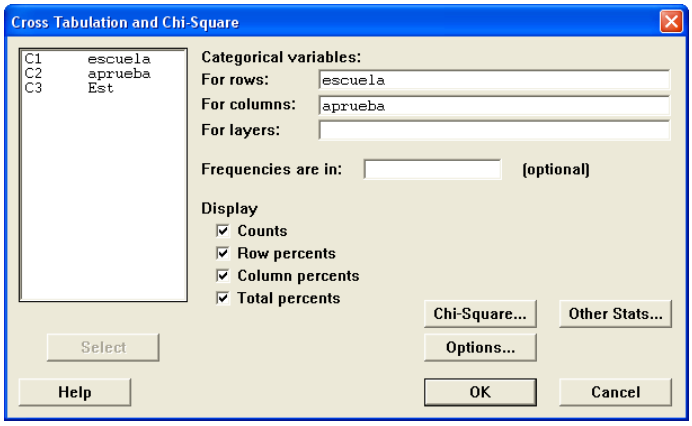


Figura 3.31. Ventana de diálogo para hacer una tabla de contigencia de *escuela* versus *aprueba*

El contenido de la tabla de **session** es el que sigue.

Tabulated statistics: escuela, aprueba

```
Rows: escuela   Columns: aprueba

             no      si     All

priv          3       7      10
             30      70     100
          37.50   58.33   50.00
             15      35      50

públ          5       5      10
             50      50     100
          62.50   41.67   50.00
             25      25      50

All           8      12      20
             40      60     100
         100.00  100.00  100.00
             40      60     100

Cell Contents:      Count
                    % of Row
                    % of Column
                    % of Total
```

Interpretación*: Cada celda contiene 4 valores: La Frecuencia Absoluta, el porcentaje que representa la celda con respecto al total de la fila, el procentaje que representa la celda con respecto al total de la columna, el porcentaje que representa la celda con respecto al total global. Por ejemplo, si cogemos los números de la primera celda, significa que hay 7 estudiantes que son de escuela privada y aprueban el examen. Un 70% de los estudiantes de escuela privada aprueban el examen, 58.33% de los que aprueban el examen son de escuela privada y 35% son estudiantes de escuela pública y aprueban el examen.*

La segunda situación donde ***Cross Tabulation*** es usada, es cuando las frecuencias absolutas de cada celda están totalizados, como en el siguiente ejemplo.

Ejemplo 3.17. Los siguientes datos se han recopilados para tratar de establecer si hay relación entre el Sexo del entrevistado y su opinión con respecto a una ley del Gobierno.

Sexo	**Opinion**	**Conteo**
male	si	10
male	no	20
male	abst	30
female	si	15
female	no	31
female	abst	44

Usar **MINITAB** para construir una tabla de contingencia y responder además las siguientes preguntas:

a) ¿Qué porcentaje de los entrevistados son mujeres que se abstienen de opinar?
b) De los entrevistados varones. ¿Qué porcentaje está en contra de la ley?
c) De los entrevistados que están a favor de la ley. ¿Qué porcentaje son varones?

d) De los que no se abstienen de opinar ¿Qué porcentaje son varones?

Solución:
En este caso se entra la columna c3 (‘conteo’) en la ventanita correspondiente a ***Frequencies are in*** que aparece en la ventana de dialogo de ***Cross Tabulation*** . Los resultados serán como sigue:

Tabulated statistics: Sexo, Opinion

```
Using frequencies in Conteo

Rows: Sexo   Columns: Opinion

             abst      no      si     All

female         44      31      15      90
            48.89   34.44   16.67  100.00
            59.46   60.78   60.00   60.00
            29.33   20.67   10.00   60.00

male           30      20      10      60
            50.00   33.33   16.67  100.00
            40.54   39.22   40.00   40.00
            20.00   13.33    6.67   40.00

All            74      51      25     150
            49.33   34.00   16.67  100.00
           100.00  100.00  100.00  100.00
            49.33   34.00   16.67  100.00

Cell Contents:      Count
                    % of Row
                    % of Column
                    % of Total
```

a) $\frac{44}{150}\times 100 = 29.33\%$

b) $\frac{20}{60}\times 100 = 33.33\%$ (20/60)x100=33.33%

c) $\frac{10}{25}\times 100 = 40.00\%$ (10/25)x100=40.00%

d) $\frac{(10+20)}{(25+51)}\times 100 = \frac{30}{46}\times 100 = 39.00\%$

Cuando se tiene dos variables categóricas se pueden hacer gráficas de barras agrupadas ("bars in clusters") o en partes componentes ("stacked bars") para visualizar la relación entre ellas.

Ejemplo 3.18. Hacer una gráfica de barras agrupadas para mostrar la distribución de los estudiantes por sexo según programa académico para los datos del Ejemplo 3.1.

Para hacer una gráfica de barras agrupadas se debe elegir ***Cluster*** en la ventana de diálago principal. Luego, en la segunda ventana, se eligen las variables que se utilizarán. Como se quiere una gráfica de estudiantes por programa por sexo, se elige primero la variable programa y luego la varible sexo.

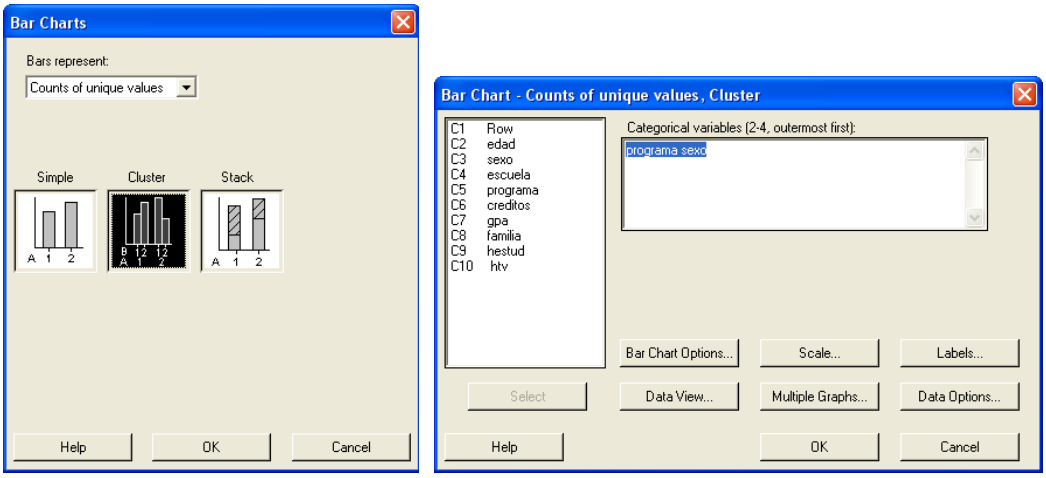


Figura 3.32. Ventana de diálogo para hacer una gráfica de barras agrupadas

Oprimiendo la opción ***Labels***, se puede especificar el título de la gráfica y las etiquetas de las columnas.

Luego, se obtiene la siguiente gráfica de barras agrupadas

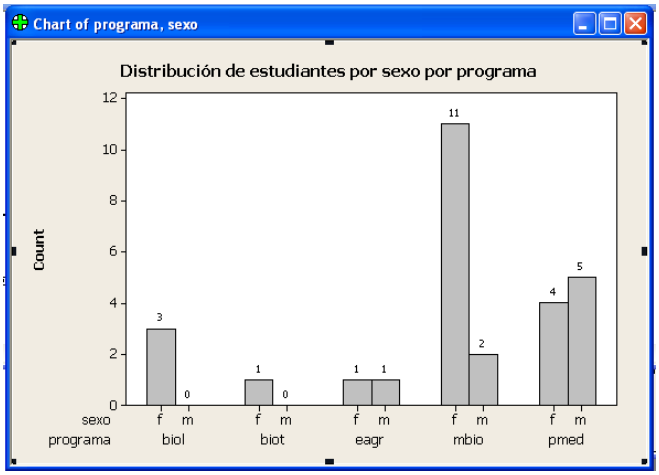


Figura 3.33. Gráfica de barras agrupadas de variable programa según sexo

Ejemplo 3.19. La siguiente tabla muestra el número de estudiantes subgraduados matriculados en el Recinto Universitario de Mayaguez de la Universidad de Puerto Rico en el primer semestre del año académico 96-97.

Facultad	Hombres	Mujeres
Artes y Ciencias	1713	2492
Admistración de Empresas	637	1257
Ingeniería	2885	1720
Agricultura	806	331

Hacer una gráfica de barras agrupadas para comparar el número de estudiantes por sexo en cada facultad.

Solución: Primero que nada hay que entrar los datos en 3 columnas: *Facultad, Sexo* y *cantidad*. Luego se elige Graphs-> Bar Chart. Las opciones de la primera ventana se eligen como se muestra en la Figura 3.34.

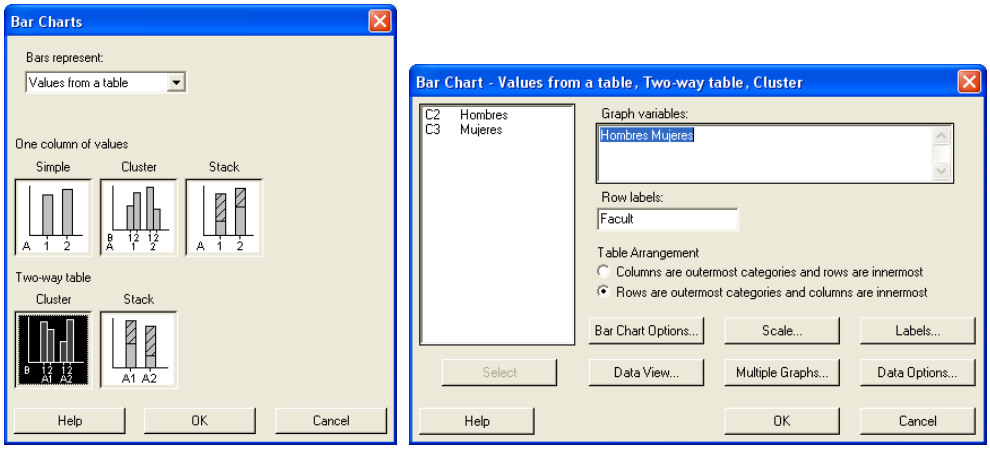


Figura 3.34. Ventana de diálogo para la gráfica de barras agrupadas del Ejemplo 3.19.

Luego de escribir el título deseado en **Labels**, se oprime **ok** para obtener la siguiente gráfica.

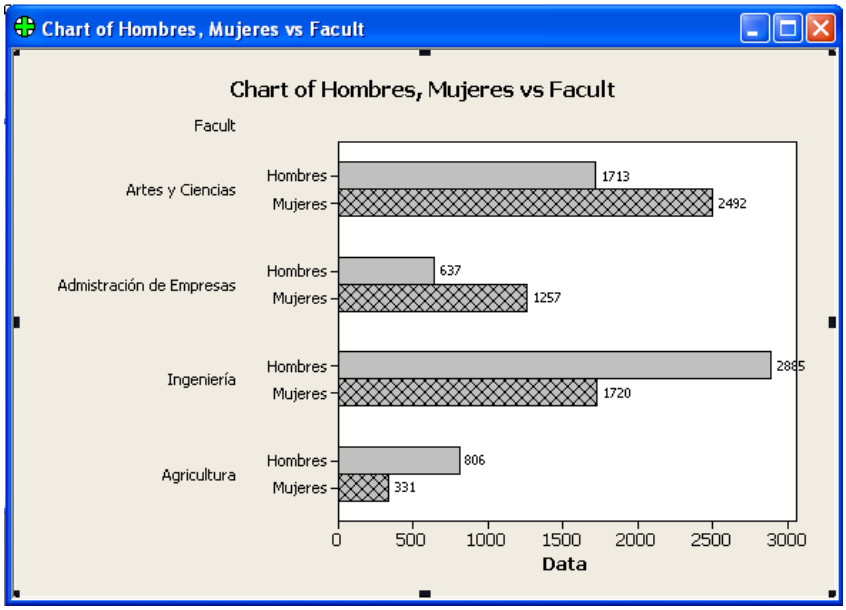


Figura 3.35. Gráficas de barras agrupadas para la variable *Facultad* según *Sexo*.

Ejemplo 3.20. Hallar una gráfica de partes componentes para comparar los estudiantes (por programa) según el tipo de escuela de donde proceden, usando datos del ejemplo 3.1.

Solución: Bajo la opción de **Gráfica -> Bar** Chart, las opciones que se muestran en la figura 3.31.

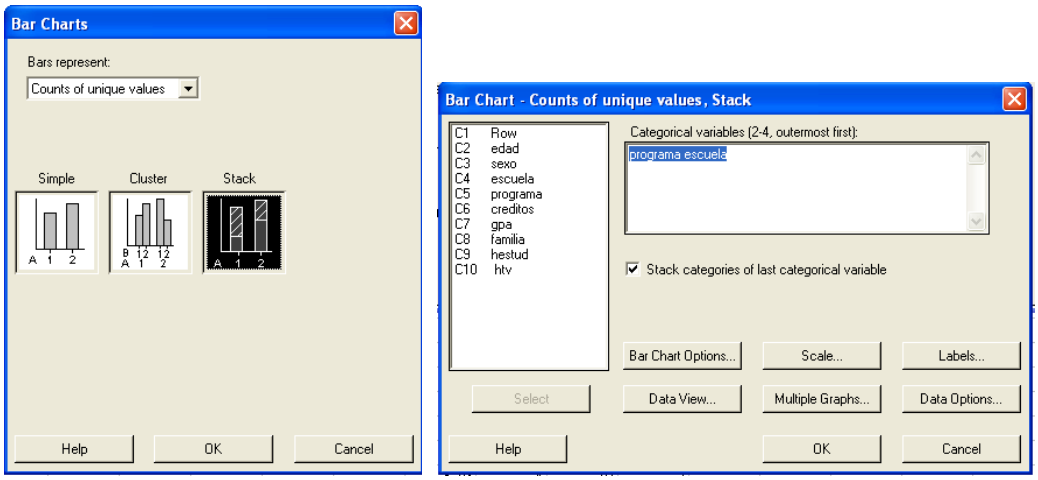


Figura 3.36: Ventanas de diálogo para una gráfica de partes componentes

Luego, en la ventana de **Scale -> Axes and Ticks** elija la opción "Transpose value and category scales" y en la ventana de **Labels** coloque el título de la gráfica y los valores correspondientes a las barras. La gráfica resultante se muestra en la Figura 3.37.

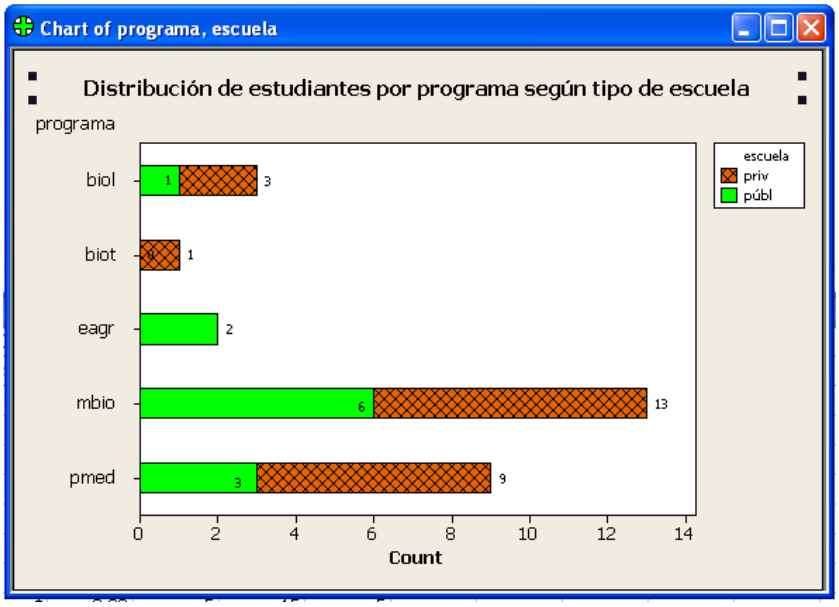


Figura 3.37. Gráfica de barras en partes componentes para la variable *Programa* según *Escuela*

Algunas veces ocurre que una variable cuantitativa es convertida en categórica agrupándola en clases o grupos. Por ejemplo, la edad puede ser convertida en cualitativa si se consideran grupos de edades. Similarmente, años de educación pueden ser convertida en cualitativa si se consideran niveles de educación.

Ejemplo 3.21. La siguiente gráfica muestra la distribución de la población en Puerto Rico según grupos de edades y por sexo.

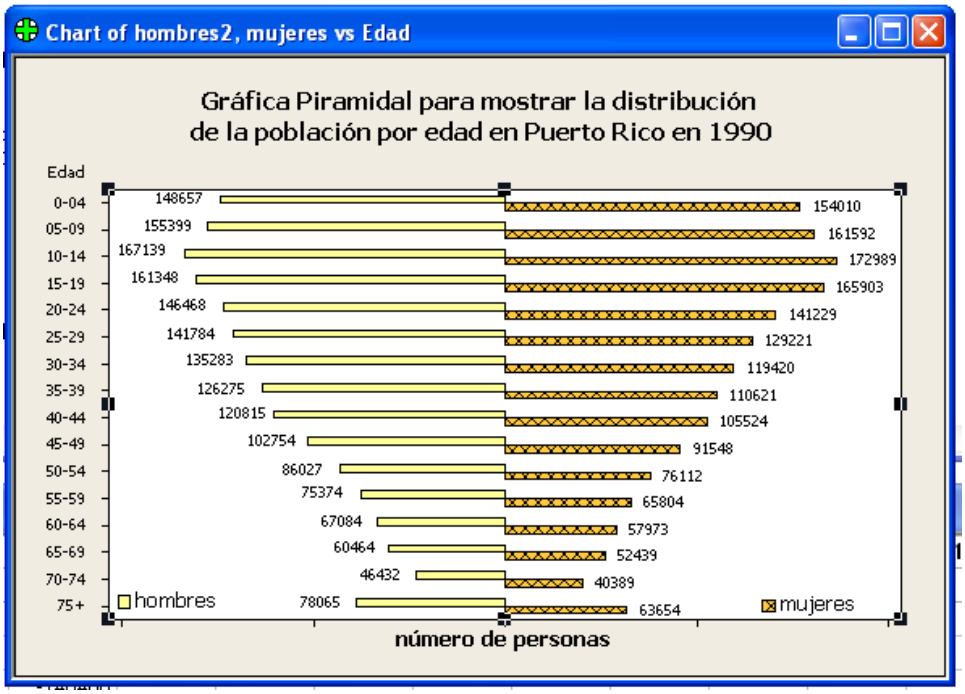


Figura 3.38: Distribución de la población por grupo de edades en Puerto Rico

3.7.2 Datos que contienen una variable cualitativa y otra cuantitativa

Un ejemplo de un conjunto de dos variables en el cual una variable es cualitativa y la otra cuantitativa puede el conjunto compuesto por *método de enseñanza* (cualitativa) y *nota obtenida por el estudiante* (cuantitativa). Otro ejemplo sería, el conjunto compuesto por la variable cualitativa *profesión de una persona* y la variable cuantitativa *salario anual*.

La forma estándar de presentar los datos es en columnas donde cada columna representa un valor de la variable cualitativa y los valores dentro de cada columna representan valores de la variable cuantitativa. En general el objetivo es comparar los valores de la variable cualitativa según los valores de la variable cuantitativa, esto se lleva a cabo con una técnica llamada *análisis de varianza* (ver capítulo 10).

La gráfica más adecuada para representar este tipo de información es el "Boxplot". La gráfica de la Figura 3.39 muestra los “boxplots” de los promedios académicos de los estudiantes varones y mujeres del Ejemplo 3.1.

Interpretación: *De la gráfica se puede ver que en promedio las mujeres tienen mejor promedio académico (GPA) que los hombres, y que la distribución de sus GPA es ligeramente más variable. Además no hay "outliers".*

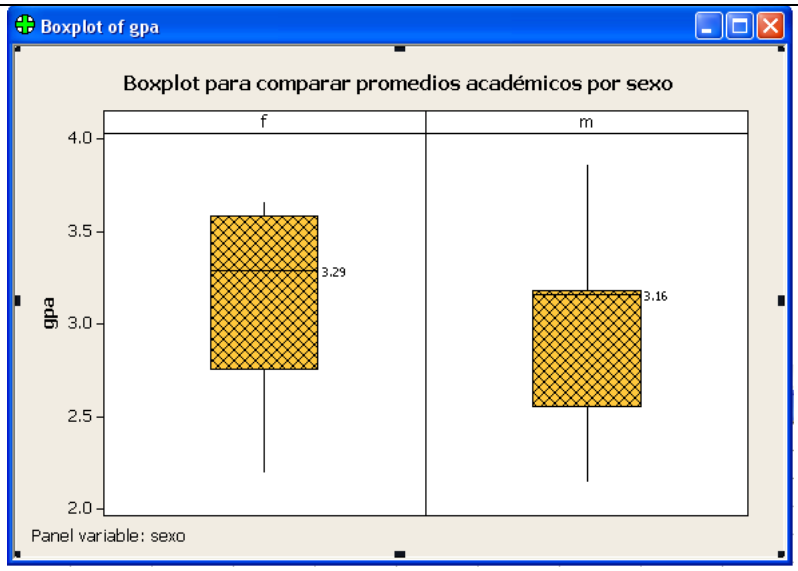


Figura 3.39: Boxplot para comparar los promedios de hombres y mujeres

3.7.3 Datos Bivariados Continuos

Si se quiere representar la relación entre dos variables cuantitativas entonces se usa un diagrama de dispersión ("Scatterplot"). Para obtener un diagrama de dispersión entre dos variables X e Y se usa la opción ***Scatterplots*** del menú **Graph.** La ventana de diálogo para hacer el diagrama de dispersión del promedio académico (*gpa*) versus el tamaño de la familia usando los datos del Ejemplo 3.1 es la siguiente:

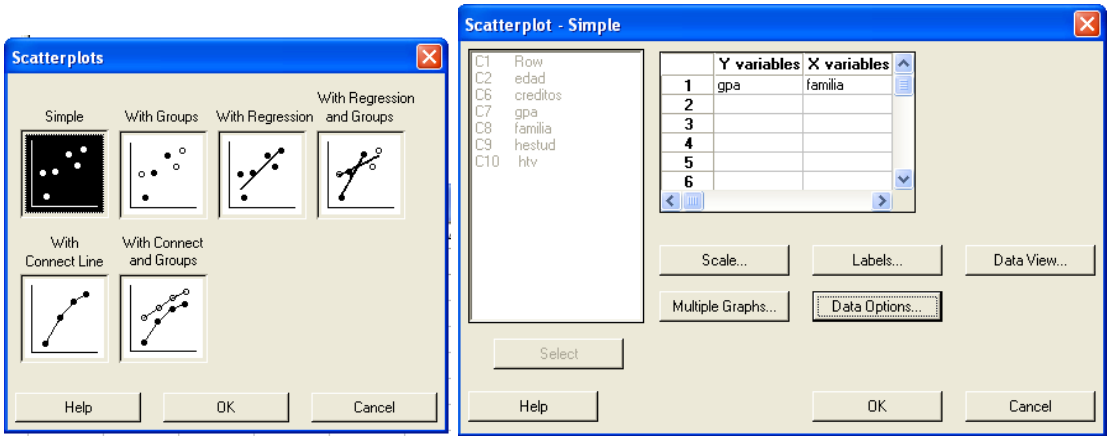


Figura 3.40: Ventanas de diálogo para obtener el plot de gpa versus familia.

La gráfica se muestra en la siguiente figura, donde además cada punto es marcado con el

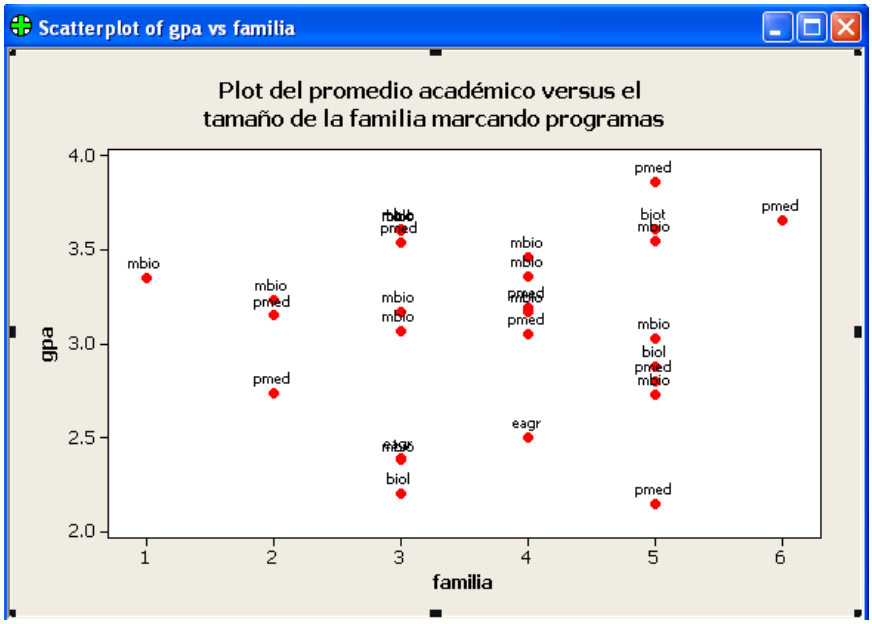
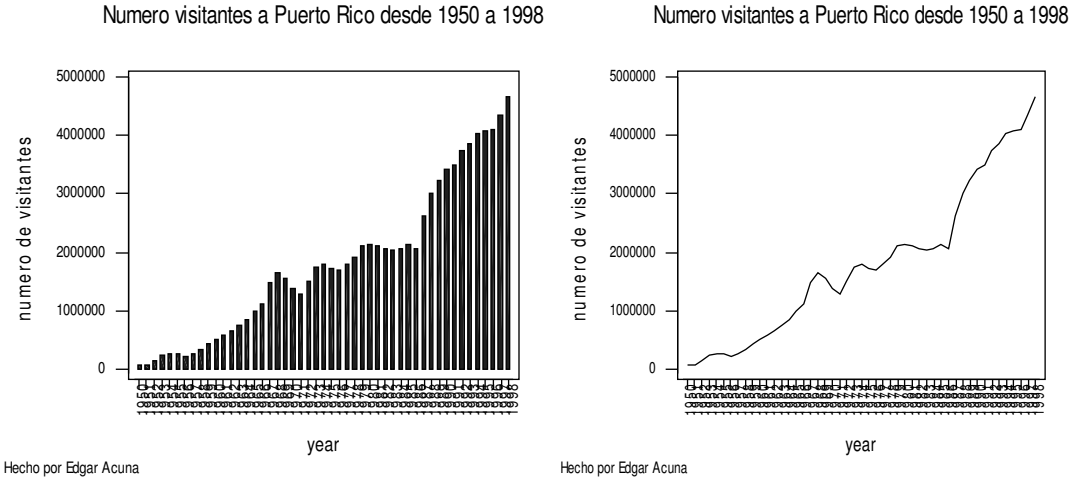


Figura 3.41: Plot de gpa versus familia marcando cada punto con el programa del estudiante

programa al cual pertenece el estudiante, ésto se consigue eligiendo la opción ***Labels -> Data Labels*** y luego entrando la variable *programa* en la ventanita correspondiente a ***Use labels from column.*** Para cambiar de símbolo, colores y tamaños a los puntos del plot, oprima el botón de la izquierda del ratón dos veces seguidos sobre cualquiera de los símbolos para abrir la opción ***Edit Attributes***.

Ejemplo 3.22. Es bien frecuente tener datos de una variable para un período de tiempo (dias, meses o años), estos tipos de datos son llamados series cronológicas o series temporales. Para este tipo de datos se pueden hacer gráficos de barras (aunque éstas son



inadecuadas si el período de tiempo es muy grande) y gráficas lineales. Las siguientes gráficas se refieren al número de visitantes a Puerto Rico desde 1950 hasta 1998.

Figura 3.42 Gráfica de barras del número de visitantes a Puerto Rico entre 1950-1998.

Figura 3.43 Gráfica de barras del número de visitantes a Puerto Rico entre 1950-1998.

3.8 El Coeficiente de Correlación

El coeficiente de correlación lineal, llamado también coeficiente de correlación de Pearson, se representa por **r** y es una medida que representa el grado de asociación entre dos variables cuantitativas X e Y. Se calcula por

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Donde:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \quad , \quad S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \quad \text{y} \quad S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

S_{xx} es llamada la Suma de Cuadrados corregida de X, S_{yy} es la Suma de Cuadrados Corregida de Y, y S_{xy} es la Suma de Productos de X e Y. Tanto S_{xx} como S_{yy} no pueden ser negativas, S_{xy} si puede ser positiva o negativa.

La correlación varia entre -1 y 1. Un valor de r cercano a 0 indica una relación lineal muy pobre entre las variables. Un valor cercano a 1 indica que hay una buena relación lineal entre la variable y además al aumentar una de ellas la otra también aumenta. Un valor cercano a –1 indica una buena relación lineal pero al aumentar el valor de una de las variables la otra disminuye.

En términos generales un valor de correlación mayor que 0.75 ó menor que -0.75 indica una buena relación lineal entre las variables. Aunque el tipo de datos que se está usando influye en el momento de decidir si la correlación es suficientemente alta. Si los datos provienen de un área donde se exige mucha precisión, como en ingeniería o medicina entonces la correlación debe estar lo más cerca posible a 1 ó –1, en áreas como economía o en ciencias sociales una correlación de 0.6 en valor absoluto pudiera ser considerada aceptable. Pero si hay un consenso general que una correlación entre -0.3 y 0.3 es indicativo de una relación lineal bastante pobre entre las dos variables.

Ejemplo 3.23. El dueño de una empresa que vende carros desea determinar si hay relación lineal entre los años de experiencia de sus vendedores y la cantidad de carros que venden. Los siguientes datos representan los años de experiencia (X) y las unidades de carros vendidas al año (Y), de 10 vendedores de la empresa.

X(años)	3	4	6	7	8	12	15	20	22	26
Y(ventas)	9	12	16	19	23	27	34	37	40	45

Haciendo uso de la calculadora de **MINITAB.** Se obtienen los siguientes resultados

```
Row   years   ventas   Sxx      Syy      Sxy          r

  1       3        9   590.1   1385.6   889.4   0.983593
  2       4       12
  3       6       16
  4       7       19
  5       8       23
  6      12       27
  7      15       34
  8      20       37
  9      22       40
 10      26       45
```

Interpretación*:*
Existe una buena relación lineal entre los años de experiencia y las unidades que vende el vendedor. Además mientras más experiencia tiene el vendedor más carros venderá. Se puede usar los años de experiencia para predecir las unidades que venderá anualmente a través de una linea recta.

En **MINITAB**, el coeficiente de correlación se puede obtener eligiendo la opción ***correlation*** del submenú **Basic Statistics** del menú **Stat.**

Ejemplo 3.24. La siguiente salida muestra la correlación entre el tamaño de la familia del estudiante y su promedio académico *gpa* del Ejemplo 3.1.

Correlations (Pearson)

```
Correlation of gpa and familia = 0.061
```

Interpretación:
La correlación de .061 indica una muy pobre relación lineal entre las variables familia y gpa. No tiene sentido predecir el promedio académico del estudiante usando el tamaño de su familia a través de una linea recta.

La Figura 3.36, muestra cuatro diagramas de dispersión y sus respectivas correlaciones. Notar que en los dos últimos plots la correlación es cercana a cero, pero en el primero de ellos no parece haber ningún tipo de relación entre las variables, en tanto que en el segundo no hay relación lineal pero si existe una relación cuadrática.

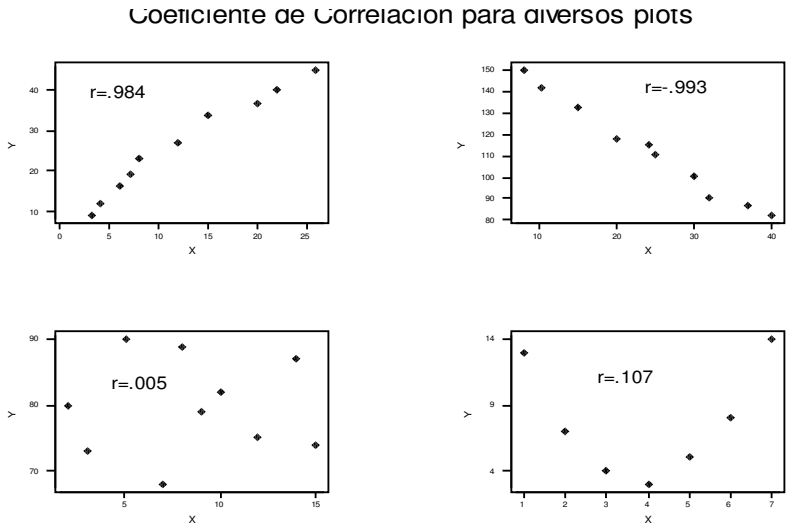


Figura 3.44: Valor de la correlación para diversos plots.

El valor de correlación es afectado por la presencia de valores anormales, en la siguiente gráfica se puede ver el efecto de los valores anormales en el valor de la correlación para 4 diferentes relaciones.

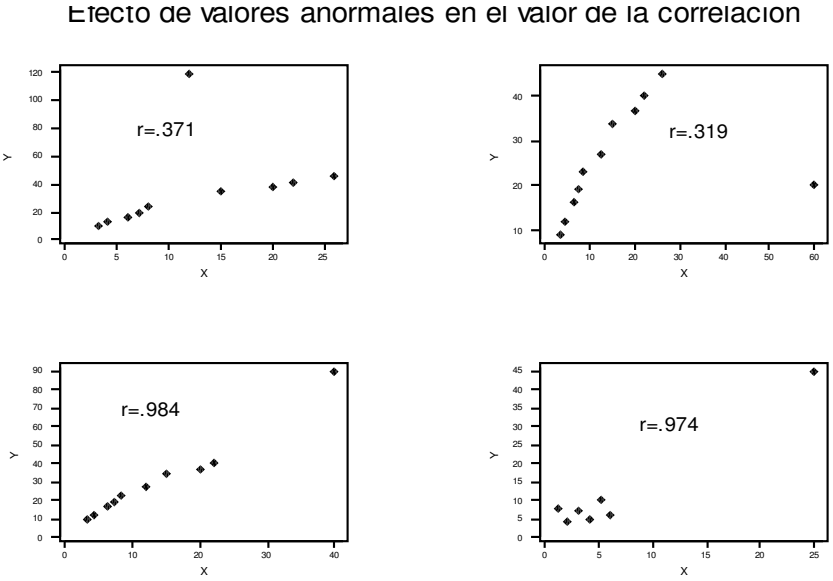


Figura 3.45: Efectos de valores anormales en la correlación

Interpretación de la figura 3.45: *En el primer caso existe un valor bastante anormal en la dirección vertical que hace que la correlación sea bastante bajo a pesar de que los otros valores parecen estar bastante alineados. En el segundo caso, existe un valor bastante alejado horizontalmente de la mayor parte de los datos y que hace que la correlación sea relativamente baja a pesar de que los otros valores muestran una alta asociación lineal.*

En el tercer caso hay, una observación bastante alejado en ambas direcciones sin embargo no tiene ningun efecto en la correlación.

En el cuarto caso, hay un valor bastante alejado en ambas direcciones y las restantes observaciones están poco asociadas, pero el valor anormal hace que el valor de la correlación sea bastante alto.

El cuadrado del coeficiente de correlación expresado en porcentaje es llamado el **Coeficiente de Determinación (R^2).** Un R^2 mayor de 70% indica una buena asociación lineal entre las variables X e Y.

3.9 Una introducción a Regresión Lineal.

Si se ha determinado que la correlación lineal entre las variables Y y X es aceptable entonces el próximo paso es determinar la línea que representa la tendencia de la relación entre las dos variables cuantitativas, ésta es llamada la ***linea de regresión estimada.*** La variable Y es considerada como la ***variable dependiente*** o ***de respuesta*** y la variable X es considerada la ***variable independiente o predictora.*** La ecuación de la línea de regresión es

$$\hat{Y} = \hat{\alpha} + \hat{\beta}\,X,$$

donde: $\hat{\alpha}$ es el intercepto con el eje Y, y $\hat{\beta}$ es la pendiente de la linea de regresión. Ambos son llamados los coeficientes de la línea de regresión.

Los estimadores $\hat{\alpha}$ y $\hat{\beta}$ son hallados usando el método de mínimos cuadrados, que consiste en minimizar la suma de los errores cuadráticos de las observaciones con respecto a la línea. Las fórmulas de cálculo son:

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} \quad \text{y} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

donde $\bar{x}$ es la media de los valores de la variable X y $\bar{y}$ es la media de los valores de Y.

Interpretación de los coeficientes de regresión:

La pendiente $\hat{\beta}$ se interpreta como el cambio promedio en la variable de respuesta Y cuando la variable predictora X se incrementa en una unidad adicional.
El intercepto $\hat{\alpha}$ indica el valor promedio de la variable de respuesta Y cuando la variable predictora X vale 0. Si hay suficiente evidencia de que X no puede ser 0 entonces no tendría sentido la interpretación de $\hat{\alpha}$

En **MINITAB**, es posible obtener simultáneamente, el "scatterplot", el coeficiente R^2 y la línea de regresión. Para esto, se sigue la secuencia **Stat ▪ Regression ▪Fitted line Plot** como se muestra en Figura 3.46:

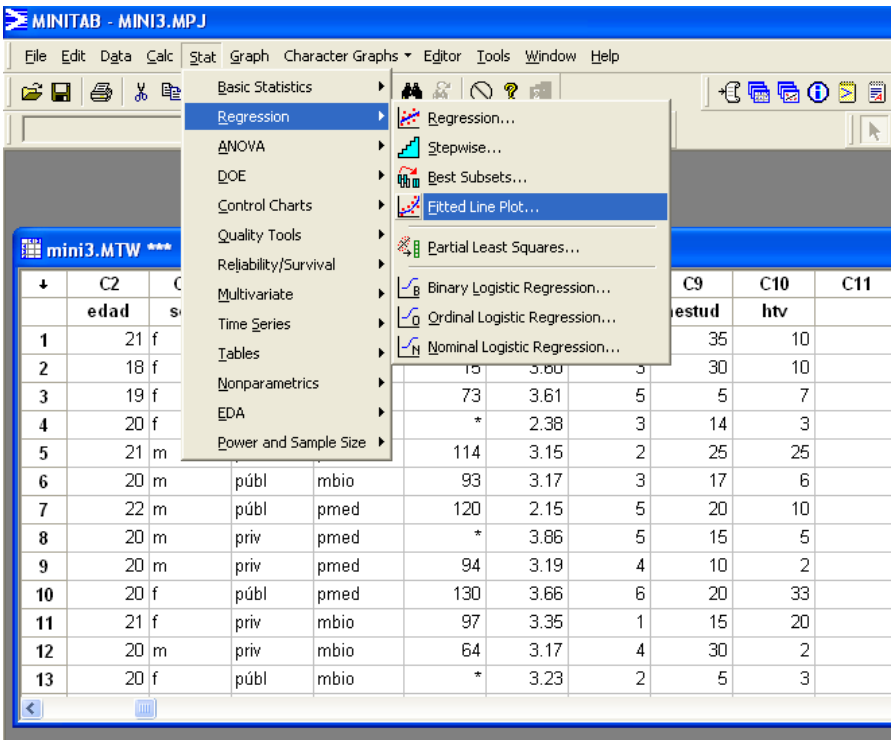


Figura 3.46: Las opciones del menú **regression**

Ejemplo 3.25. Supongamos que se desea establecer una relación entre la nota que un estudiante obtiene en la parte de aprovechamiento matemático de ingreso (CEEB) y el Promedio académico al final de su primer año de universidad (GPA). Se toma una muestra de 15 estudiantes y se obtiene los siguientes datos:

Est	CEEB	GPA
1	425	2.81
2	495	2.56
3	600	2.92
4	610	3.18
5	612	2.51
6	648	3.43
8	660	3.16
9	665	2.73
10	670	2.82
11	720	3.04
12	710	2.42
13	735	2.97
14	780	3.33
15	790	3.12

```
7     652    2.72
```

Obtener el diagrama de dispersión de los datos, la ecuación de la línea de regresión y trazar la línea encima del diagrama de dispersión.

Solución: Primero hay que notar que la variable independiente es CEEB y la variable dependiente esGPA. Luego, la ventana de diálogo para la opción ***Fitted line Plot*** lucirá como sigue:

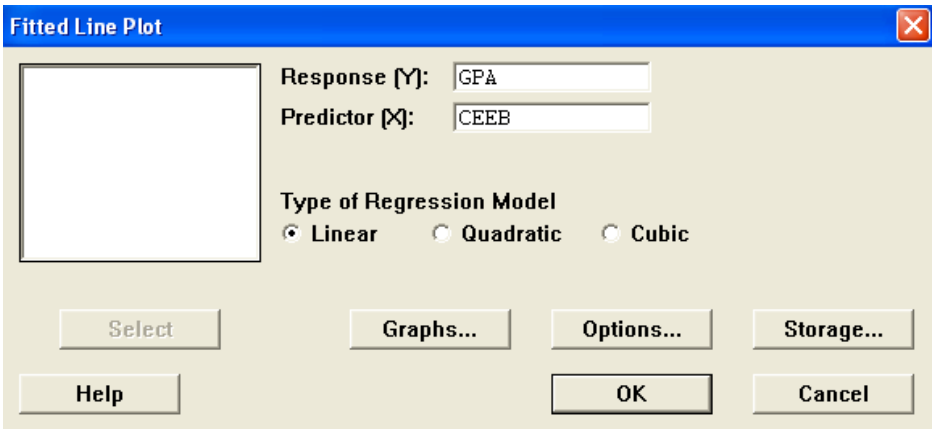


Figura 3.47: Ventana de diálogo para obtener el diagrama de dispersión y la linea de regresión de *gpa* versus *familia*

y la gráfica aparecerá como

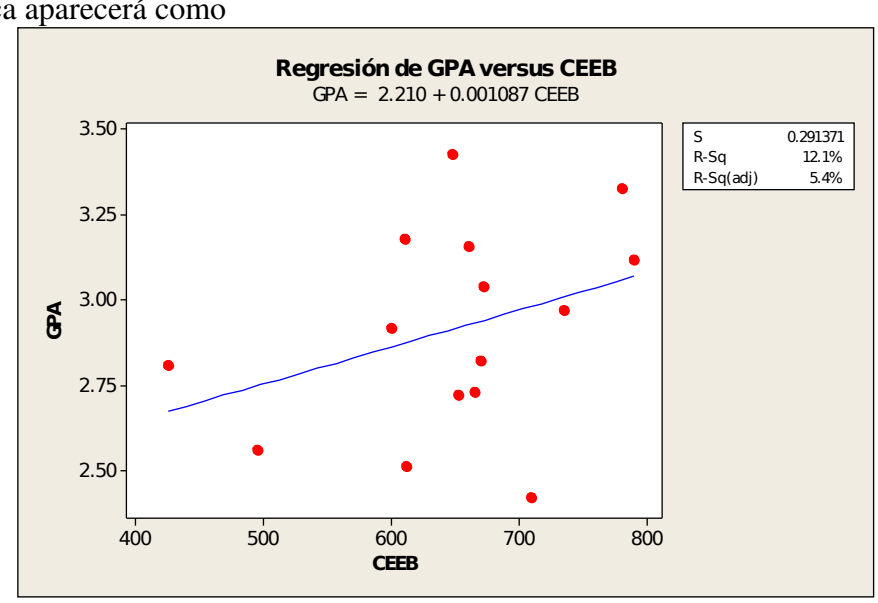


Figura 3.48: Diagrama de puntos y linea de regresión de gpa versus familia

Interpretación: *El coeficiente de determinación es .121 y como la pendiente de la línea de regresión es positiva resulta ser que la correlación es .11, esto indica una pobre relación lineal entre las variables CEEB y GPA. O sea que es poco confiable predecir GPA basado en el CEEB usando una linea.*

La ecuación de la línea de regresión aparecerá en la ventana **session**

Regression

```
The regression equation is
y = 2.21 + 0.00109 x

Predictor       Coef       StDev          T        P
Constant      2.2099      0.5319       4.15    0.001
x          0.0010872   0.0008122       1.34    0.204

S = 0.2914      R-Sq = 12.1%     R-Sq(adj) = 5.4%
```

Interpretación: *La pendiente 0.00109 indica que por cada punto adicional en el College Board el promedio del estudiante subiría en promedio en 0.00109, o se podría decir que por cada 100 puntos más en el College Board el promedio académico del estudiante subiría en .109. Por otro lado, si consideramos que es imposible que un estudiante sea admitido sin tomar el College Board, podemos decir que no tiene sentido interpretar el intercepto.*

El uso de los botones **Options** y **Storage** y de otros aspectos de regresión serán discutidos más detalladamente en el capítulo 8 de este texto.

Predicción

Uno de los mayores usos de la línea de regresión es la predicción del valor de la variable dependiente dado un valor de la variable predictora. Esto se puede hacer fácilmente sustituyendo el valor dado de X en la ecuación.

Por ejemplo, supongamos que deseamos predecir el promedio académico de un estudiante que ha obtenido 600 puntos en la parte matemática del examen de ingreso. Sustituyendo x =600 en la ecuación de la línea de regresión se obtiene Y=2.21+.00109*600=2.21+.654=2.864. Es decir que se espera que el estudiante tenga un promedio académico de 2.86.

MINITAB también tiene una opción que permite hacer predicciones pero, esto será tratado en el capítulo 9 del texto.

EJERCICIOS

1. La siguiente tabla representa el crecimiento poblacional y vehicular de Puerto Rico desde 1950.

Año	Población	Vehículos
1950	2,200,000	57,120
1960	2,345,000	172,077
1970	2,710,000	478,340
1980	3,182,328	1,129,312
1990	3,522,037	1,582,061
1996	3,782,862	2,168,697

Hacer una gráfica que represente la información dada.

2. La siguiente tabla representa los porcentajes de familias americanas en diversos niveles de ingreso en 1969 y 1994.

Ingreso	year 1969	year 1994
Less 10,000	7.9	8.7
10,000 - 14,999	6.7	6.9
15,000 - 24,999	15.8	15.0
25,000 - 34,999	19.1	14.3
35,000 - 49,999	24.7	18.0
50,000 - 74,999	17.8	19.9
75,000 - 99,999	5.0	8.8
100,000 and over	2.9	8.4

a) Hacer una gráfica de barras que permita comparar como han cambiado los porcentajes de familias a varios niveles de ingreso de 1969 a 1994. Comentar la gráfica.
b) Hacer un pie-chart para ver la distribución de personas por nivel de ingreso en los dos años.

3. La siguiente tabla muestra los casos reportados y las muertes por SIDA en Puerto Rico desde 1992 hasta 1996.

Número de casos	tipo	año
2386	reportados	92
1633	muertos	92
2619	reportados	93
1647	muertos	93
2253	reportados	94
1211	muertos	94
1903	reportados	95
800	muertos	95
1152	reportados	96
259	muertos	96

Hacer una gráfica de Barras agrupadas para representar la información.

4. Hacer un"Pie Chart" para representar la siguiente información

Casos de SIDA en Puerto Rico desde 1992

Región	Casos
Aguadilla	600
Mayaguez	930
Arecibo	1199
Ponce	3602
Bayamón	3220
San Juan	2334
Caguas	2352
Fajardo	608

5. Los siguientes datos representan tiempos de sobrevivencia (en dias) de 30 pacientes aquejados de cáncer

42 45 51 46 340 81 243 63 155 151 37 138 245 377 537 455 776
163 20 1234 201 2970 456 1235 1581 40 3808 1804 719 365

a) Calcular la media, la mediana y la desviación estándar. Comentar sus resultados.
b) Hacer el histograma de los datos y comentar la gráfica.
c) Hacer el "stem-and-leaf".
d) Hacer el "Boxplot" de los datos y comentar la gráfica.

6. Elegir la mejor contestación en cada una de las siguientes preguntas

I. ¿Cuál de las siguientes afirmaciones es FALSA?

a) Una variable es cualitativa si los valores que asume expresan atributos o categorias.
b) Tipo de sangre es una variable cualtitativa.
c) La Mediana puede usarse cuando los datos son cualitativos.
d) Un gráfico de barras se usa cuando los datos son cualitativos.

II. ¿Cuál de las siguientes afirmaciones es CIERTA?

a) La muestra al azar es aquella que hace que la media de la muestra sea igual a la media poblacional.
b) La varianza de una muestra siempre es mayor que la varianza poblacional porque en la primera se divide por n-1.
c) En la fórmula de la varianza de la muestra se divide por n-1 porque excluyendo un dato se obtiene un mejor estimado de la varianza poblacional.
d) Una muestra al azar hace que la media muestral sea un estimado bastante confiable de la media poblacional.

III. ¿Cuál de los siguientes enunciados es CIERTO?

a) La media es una mejor medida que la mediana cuando todos los datos son pequeños.

b) La mediana es afectada por la presencia de outliers.
c) La varianza es afectada por la presencia de outliers.
d) La media es mejor medida que la mediana cuando la muestra es asimetrica a la derecha.

IV. Un histograma es asimétrico hacia la derecha.

a) Si todos los datos son positivos.
b) Si para valores bajos de la variable la frecuencia es alta, y para valores grandes la frecuencia es baja.
c) Si para valores bajos de la variable la frecuencia es baja, y para valores grandes la frecuencia es alta.
d) Si la media de los datos es positivo.

V. ¿Cuál de las siguientes afirmaciones con respecto a la amplitud de clase es FALSA?

a) La amplitud es igual al rango o alcance dividido entre el número de clases.
b) La amplitud es igual a la diferencia de dos marcas de clases consecutivas.
c) La amplitud de una clase es CERO si su frecuencia absoluta es CERO.
d) La amplitud es igual a la diferencia de dos limites inferiores de clases consecutivas.

VI. ¿Cuál de las siguientes afirmaciones es CIERTA?

a) El stem-and-leaf es una mejor gráfica que el histograma cuando existen outliers.
b) El stem-and-leaf sólo se usa para valores positivos.
c) El stem-and-leaf es una mejor gráfica que el histograma cuando los datos son solamente números enteros.
d) El stem-and-leaf permite recuperar los datos de la muestra lo cual no se puede hacer con el histograma.

7. Dado el siguiente stem-and-leaf

2	34578
3	459
4	21
5	0

Si, la unidad de la hoja=.01.
¿Cuál de los siguientes enunciados es FALSO?

a) 5 | 0 representa 0.50.
b) La muestra tiene 11 datos.
c) La muestra es asimétrica a la izquierda .
d) La mediana es 0.34.

8. ¿Cuál de los siguientes enunciados es FALSO?

a) El rango intercuartílico es una medida de variabilidad.
b) Si la desviación estandar es grande no se puede concluir que la muestra tenga mucha variabilidad.

c) Un dato es considerado un outlier si es un número positivo bien grande.
d) Un dato es considerado un outlier extremo si cae fuera del intervalo (Q1-3RIQ , Q3+ 3RIQ).

9. ¿Cuál de los siguientes no es un método de Muestreo?

a) Sistemático b) Estocástico c) Estratíficado d) Por Conglomerados.

10. ¿Cuál de las siguientes afirmaciones es CIERTA?

a) El parámetro es un valor que varía con la muestra tomada.
b) El valor estadístico por lo general permanece constante.
c) Una muestra al azar es aquella que hace que la media muestral sea un estimador confiable de la media poblacional.
d) Un Censo es un listado de todos los elementos de una muestra.

11. ¿Cuál de las siguientes afirmaciones es FALSA?

a) Una variable es cuantitativa discreta si los valores que asume resultan de hacer conteos.
b) La opinión que expresa una persona es una variable cualitativa.
c) La Media puede usarse cuando los datos son cualitativos.
d) Un gráfico de barras se usa cuando los datos son cualitativos.

12. ¿Cuál de los siguientes NO es una gráfica para datos cualitativos?

a) Pie- Chart b) Gráficas de barras agrupadas c) El dotplot d) Ninguna de las anteriores

13. ¿Cuál de las siguientes No es una acción que se puede hacer al elegir el botón **Annotation** de las ventana **Histogram?**

a) Poner título a la gráfica.
b) Poner notas al pie de la gráfica.
c) Indicar cuántos datos hay en cada intervalo de clase.
d) Poner nombre a los ejes coordenados.

14. La opción del menú TABLES que se usa para obtener una tabla de frecuencias en MINITAB es:

a) FREQUENCIES b) TABLE c) COUNT d) TALLY

15. Los resultados de ejecutar los comandos en MINITAB eligiendo las opciones del menú aparecen en la ventana

a) PROJECT b) WORKSHEET c) SESSION d) FILE

16. ¿Cuál de los siguientes enunciados es CIERTO?

a) La mediana es siempre un dato de la muestra tomada.
b) El "dotplot" es una gráfica para distribución de datos discretos.
c) El tiempo de espera para que un estudiante escoja sus secciones en la matricula es una variable cuantitativa discreta.
d) Si el tamaño de la muestra es n entonces la mediana es n/2.

17. Marcar con una C si es cierto y una F si es falso en cada uno de los siguientes enunciados.

a) La mediana es siempre un dato de la muestra tomada.
b) El parámetro es un valor que caracteriza a la muestra.
c) El número de carros que pasan por una estación de peaje entre las 7am y 9am es una variable cualitativa continua.
d) Las gráficas Circulares y de Barras se usan para presentar datos cualitativos.
e) Si el tamaño de la muestra es n, entonces la mediana es (n+1)/2.
f) El comando PRINT en MINITAB se usa para imprimir los resultados en el papel .
g) Un Censo es un listado de todos los elementos de la población.

18. Los siguientes datos representan el número de asesinatos reportados durante 15 fines de semana en una ciudad:

4 5 0 5 3 2 1 4 3 2 4 4 1 12 5

a) ¿Cuál es el número promedio de asesinatos durante los fines de semana?
b) ¿Cuál es el número más frecuente de asesinatos en los fines de semana?
c) ¿Piensa Ud. que 12 es un valor anormal? Justifique su contestación.

19. La siguiente tabla muestra la distribución de frecuencias de una muestra de los tiempos (en minutos) que tienen que esperar las personas para ser atendidos en un Banco:

Intervalos de clases	Frec. Abs f	Frec. Rel. Porcentual	Frec. Abs. Acumul.	Frec. Rel. Porc. Acum..
1.0 - 4.9	3			
5.0 - 8.9	10			
9.0 - 12.9	14			
13.0 - 16.9	25			
17.0 - 20.9	17			
21.0 - 24.9	9			
25.0 - 28.9	2			

a) ¿Cuál es la amplitud de cada clase?
b) ¿Cuál es la marca de clase (midpoint) de la tercera clase?
c) ¿Cuál es el tamaño de la muestra?
d) Cálcular las frecuencias relativas porcentuales y las frecuencias acumuladas.
e) Hacer el histograma y comentar acerca de su forma.

20. Una muestra tiene el siguiente BOXPLOT

Poner una X al lado de las afirmaciones que son CIERTAS

a) La muestra es asimétrica hacia la izquierda.
b) El dato menor es 6.
c) Existe mucha variabilidad.
d) La media de la muestra es 10.
e) El * representa un valor mayor que 18.
f) La frontera exterior superior es 25.
g) El valor adyacente inferior es 6.
h) El valor mayor es 16.

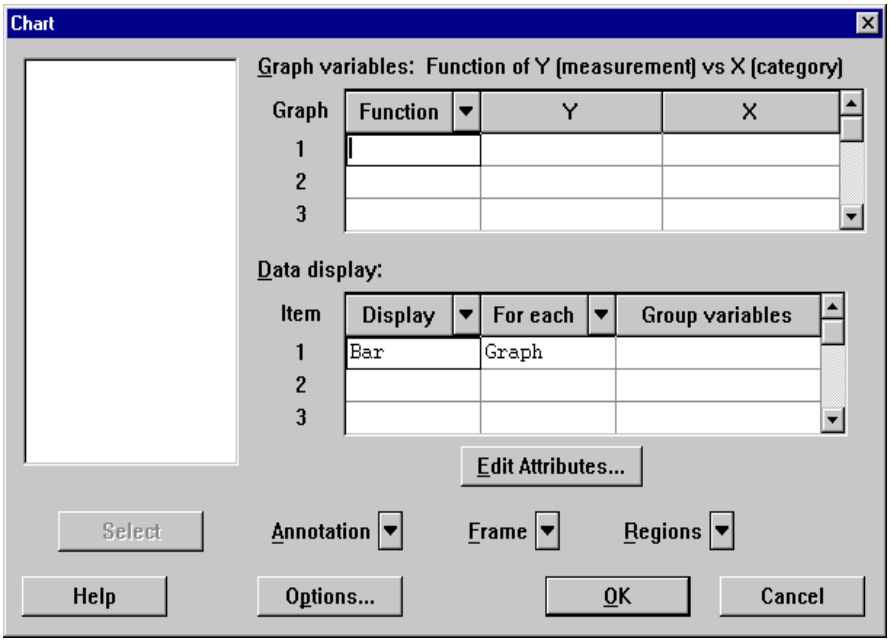
21. Los siguientes datos representan la tasa de criminalidad por cada 100000 habitantes en cada estado de los Estados Unidos.

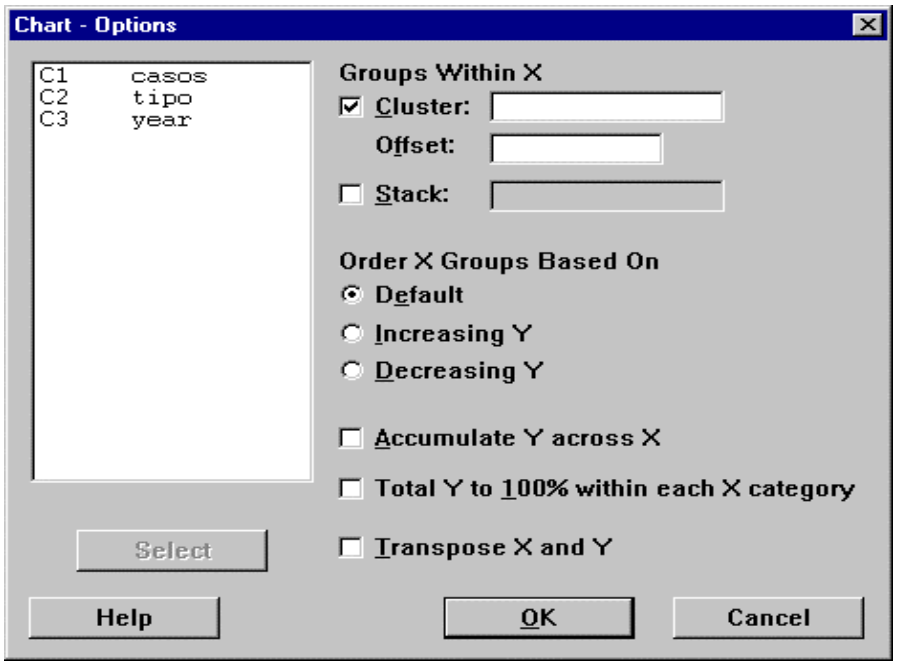
STATE	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto
Alabama	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
Alaska	10.8	51.6	96.8	284	1331.7	3369.8	753.3
Arizona	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
Arkansas	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
California	11.5	49.4	287	358	2139.4	3499.8	663.5
Colorado	6.3	42	170.7	292.9	1935.2	3903.2	477.1
Connecticut	4.2	16.8	129.5	131.8	1346	2620.7	593.2
Delaware	6	24.9	157	194.2	1682.6	3678.4	467
Florida	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4
Georgia	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9
Hawaii	7.2	25.5	128	64.1	1911.5	3920.4	489.4
Idaho	5.5	19.4	39.6	172.5	1050.8	2599.6	237.6
Illinois	9.9	21.8	211.3	209	1085	2828.5	528.6
Indiana	7.4	26.5	123.2	153.5	1086.2	2498.7	377.4
Iowa	2.3	10.6	41.2	89.8	812.5	2685.1	219.9
Kansas	6.6	22	100.7	180.5	1270.4	2739.3	244.3
Kentucky	10.1	19.1	81.1	123.3	872.2	1662.1	245.4
Louisiana	15.5	30.9	142.9	335.5	1165.5	2469.9	337.7
Maine	2.4	13.5	38.7	170	1253.1	2350.7	246.9
Maryland	8	34.8	292.1	358.9	1400	3177.7	428.5
Massachusetts	3.1	20.8	169.1	231.6	1532.2	2311.3	1140.1
Michigan	9.3	38.9	261.9	274.6	1522.7	3159	545.5
Minnesota	2.7	19.5	85.9	85.8	1134.7	2559.3	343.1
Mississippi	14.3	19.6	65.7	189.1	915.6	1239.9	144.4
Missouri	9.6	28.3	189	233.5	1318.3	2424.2	378.4
Montana	5.4	16.7	39.2	156.8	804.9	2773.2	309.2
Nebraska	3.9	18.1	64.7	112.7	760	2316.1	249.1
Nevada	15.8	49.1	323.1	355	2453.1	4212.6	559.2
New Hampshire	3.2	10.7	23.2	76	1041.7	2343.9	293.4
New Jersey	5.6	21	180.4	185.1	1435.8	2774.5	511.5
New Mexico	8.8	39.1	109.6	343.4	1418.7	3008.6	259.5
New York	10.7	29.4	472.6	319.1	1728	2782	745.8
North Carolina	10.6	17	61.3	318.3	1154.1	2037.8	192.1

North Dakota	0.9	9	13.3	43.8	446.1	1843	144.7
Ohio	7.8	27.3	190.5	181.1	1216	2696.8	400.4
Oklahoma	8.6	29.2	73.8	205	1288.2	2228.1	326.8
Oregon	4.9	39.9	124.1	286.9	1636.4	3506.1	388.9
Pennsylvania	5.6	19	130.3	128	877.5	1624.1	333.2
Rhode Island	3.6	10.5	86.5	201	1489.5	2844.1	791.4
South Carolina	11.9	33	105.9	485.3	1613.6	2342.4	245.1
South Dakota	2	13.5	17.9	155.7	570.5	1704.4	147.5
Tennessee	10.1	29.7	145.8	203.9	1259.7	1776.5	314
Texas	13.3	33.8	152.4	208.2	1603.1	2988.7	397.6
Utah	3.5	20.3	68.8	147.3	1171.6	3004.6	334.5
Vermont	1.4	15.9	30.8	101.2	1348.2	2201	265.2
Virginia	9	23.3	92.1	165.7	986.2	2521.2	226.7
Washington	4.3	39.6	106.2	224.8	1605.6	3386.9	360.3
West Virginia	6	13.2	42.2	90.9	597.4	1341.7	163.3
Wisconsin	2.8	12.9	52.2	63.7	846.9	2614.2	220.7
Wyoming	5.4	21.9	39.7	173.9	811.6	2772.2	282

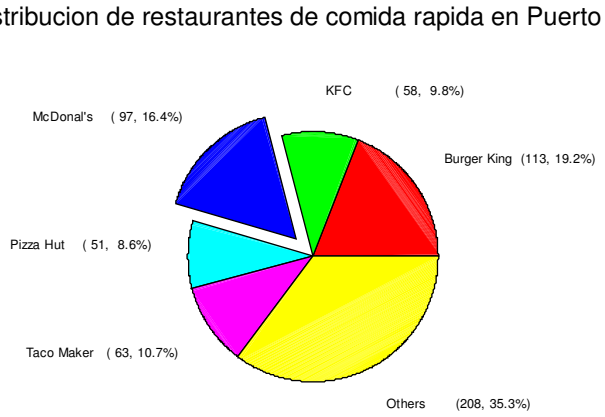
a) Hacer un histograma con 7 clases de la variable robo de auto. Comentar la gráfica.
b) Escoger cualquiera de las otras variables y hacer lo siguiente:
   i) Hacer un stem-and-leaf. Comentar su gráfica.
   ii) Hacer un boxplot. Comentar su gráfica.

22. Suponga que una Worksheet de MINITAB tiene 3 columnas: La primera es llamada **Casos** y contiene la cantidad de casos de SIDA reportados en Puerto Rico desde 1992 hasta 1996, la segunda columna llamada **Tipo** indica si son nuevos casos de SIDA en el año o si son casos de muertes por Sida, la tercera columna llamada **year**, contiene los años de la ocurrencia de los casos. Se desea hacer una gráfica de barras agrupadas. Indicar cómo se debe llenar la columna **Y**, la columna **X** y la columna **Group variables** de la ventana Chart y la ventana **Chart-Options** , las cuales se muestran en las siguientes figuras.





23. Comentar la siguiente gráfica.



24. Los siguientes datos representan la duración en horas de un cierto tipo de baterias

0.4 1.5 0 0.9 0.8 1.2 1.1 1.4 2.3 1.3 2.2 1.6 2.1 1.2 2.4 1.9 2.9 1.7

a) Hacer el "stem-and-leaf" de los datos, usando subramas si es necesario. Indicar la unidad de la hoja y comentar la forma de la gráfica.

b) ¿Cuál es el tiempo promedio de la duración de las baterias?
c) ¿Cuál es el tiempo más frecuente de duración de las baterias?
d) Hallar la mediana de los tiempos de duración.
e) Hallar la media podada del 10% de los tiempos de duración.

25. En un país se eligen 10 pueblos al azar y se anota el ingreso personal promedio de los habitantes (en miles) y la tasa de divorcio (por cada 1000 personas). Usar la siguiente tabla de datos para responder las siguientes preguntas.

Obs	Ingreso X	Divorcio Y	X^2	Y^2	XY
1	7.7	7.2	59.29	51.84	55.44
2	10.9	3.3	118.81	10.89	35.97
3	10.1	2.9	102.01	8.41	29.29
4	9.3	3.7	86.49	13.69	34.41
5	9.9	4.4	98.01	19.36	43.56
6	9.2	4.1	84.64	16.81	37.72
7	6.5	6.9	42.25	47.61	44.85
8	10.0	3.4	100.00	11.56	34.00
9	9.4	3.0	88.36	9.00	28.20
10	8.7	3.2	75.69	10.24	27.84
Sumas	91.7	42.1	855.55	199.41	371.28

a) Hacer un plot de los datos.
b) Hallar el coeficiente de correlación r e interpretarlo.
c) Hallar la línea de regresión estimada e interpretar las constantes $\hat{\alpha}$ y $\hat{\beta}$.
d) Trazar la línea de regresión sobre el plot de la parte a).
e) Hallar la tasa de divorcio estimada si el ingreso es de 11,000.