

# **3. ESTADÍSTICA DESCRIPTIVA**

**Dr. Edgar Acuna**

**<http://math.uprm.edu/~edgar>**

**UNIVERSIDAD DE PUERTO RICO  
RECINTO UNIVERSITARIO DE MAYAGUEZ**

# ESTADÍSTICA DESCRIPTIVA

En este capítulo se verán 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

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	publ	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

# 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*. 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.

## 3.1.2 El plot de puntos (“Dotplot”)

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 permite detectar la presencia de valores anormales (“outliers”).

En **MINITAB** el plot de puntos se obtiene eligiendo la opción *Dotplot* del menú **Graph**.

## 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 línea vertical de altura proporcional a la frecuencia absoluta del valor de la variable.

### 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.

## 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.

## 3.2.2 Histograma en modo gráfico

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.

## 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.

### 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**. 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*.

## 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**.

## 3.4 Gráfica de tallo y hojas (“Stem-and-Leaf”)

Es una gráfica usada para datos cuantitativos.

**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.

*continuación: ...*

## Ejemplo 3.4.

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”.*

# 3.5 Cálculo de Medidas Estadísticas

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. Aparte 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  $x_1, x_2, \dots, x_n$ , representan las observaciones de una variable  $X$  en una muestra de tamaño  $n$ , entonces la media de la variable  $X$  está dada por:

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

**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.

## 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.

**La varianza** es una medida que da una idea del grado de concentración de los datos con respecto a la media. Para determinar el grado de concentración de los datos sería el promedio de las desviaciones con respecto a la media, es decir ,

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

**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  $s$ .

## 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 Q1. 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 Q3, y el restante 25% de datos son mayores o iguales que Q3.

**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**.

## 3.6 El Diagrama de Caja (“Boxplot”)

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” .

En **MINITAB** hay varias maneras de obtener el “Boxplot” de un conjunto de datos, una de ellas es eligiendo la opción *Boxplot* del menú **Graph**. Otra manera es obtener un “boxplot” es eligiendo la opción **Character Graphs** del menú **Graph** y luego **boxplot** del listado que aparece.

# 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.

La segunda situación donde *Cross Tabulation* es usada, es cuando las frecuencias absolutas de cada celda están totalizados

# 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?

De los entrevistados que están a favor de la ley. ¿Qué porcentaje son varones?

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:

```
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
```

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

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

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

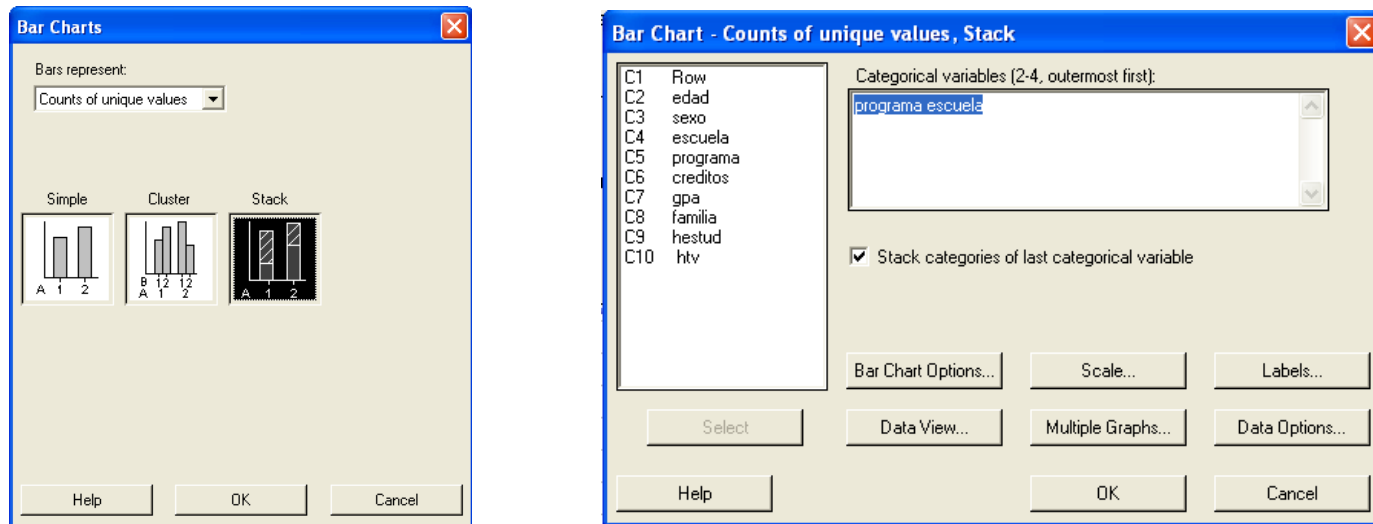
$$d) \frac{(10+20)}{(25+51)} \times 100 = \frac{30}{46} \times 100 = 39.00\%$$

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 3.20

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 **Graphs** -> **Bar Chart**, las opciones que se muestran en la figura 3.37.



**Figura 3.37:** Ventanas de diálogo para una gráfica de partes componentes

# Continuación (Ejemplo 3.20)

## Solución:

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 3.38.

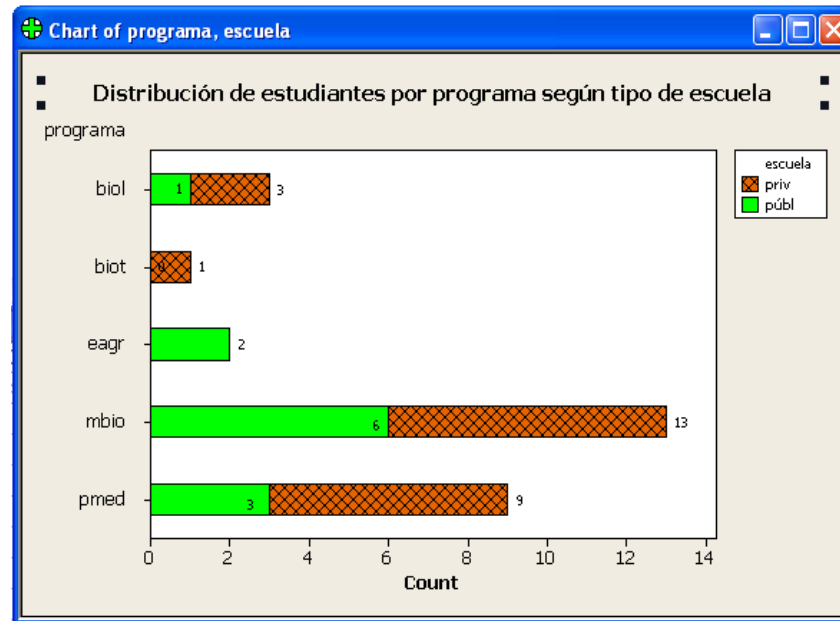


Figura 3.38. Gráfica de barras en partes componentes para la variable *Programa* según *Escuela*

## 3.7.2 Conjunto de datos que contienen una variable cualitativa y otra cuantitativa

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".

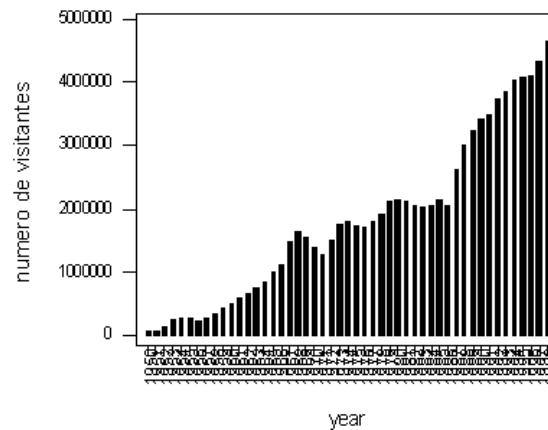
## 3.7.3 Datos Bivariados Continuos

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**.

# Ejemplo 3.22

Es bien frecuente tener datos de una variable para un período de tiempo (días, 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.

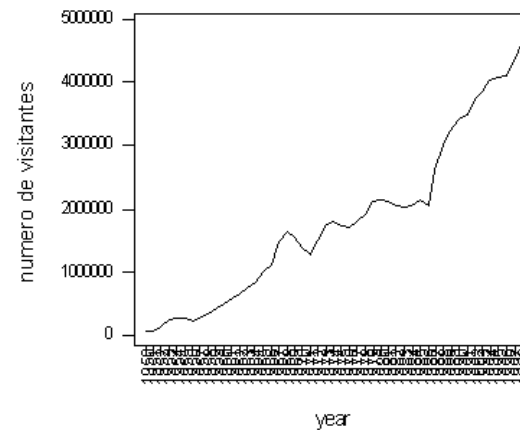
Numero visitantes a Puerto Rico desde 1950 a 1998



Hecho por Edgar Acuna

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

Numero visitantes a Puerto Rico desde 1950 a 1998



Hecho por Edgar Acuna

Figura 3.44 Gráfica de barras del número de visitantes a Puerto Rico entre 1950-1998.

# 3.8 El Coeficiente de Correlación

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:

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

$S_{xx}$  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 varía entre -1 y 1.

# 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

## Solución:

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

# Solución: (Ejemplo 3.23.)

## *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 línea recta.*

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				

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

# 3.9 Una introducción a Regresión Lineal.

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 línea 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.

# 3.9 Una introducción a Regresión Lineal.

## *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 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<sup>2</sup> y la línea de regresión. Para esto, se sigue la secuencia **Stat ▶ Regression ▶ Fitted line Plot**

# Ejemplo 3.25.

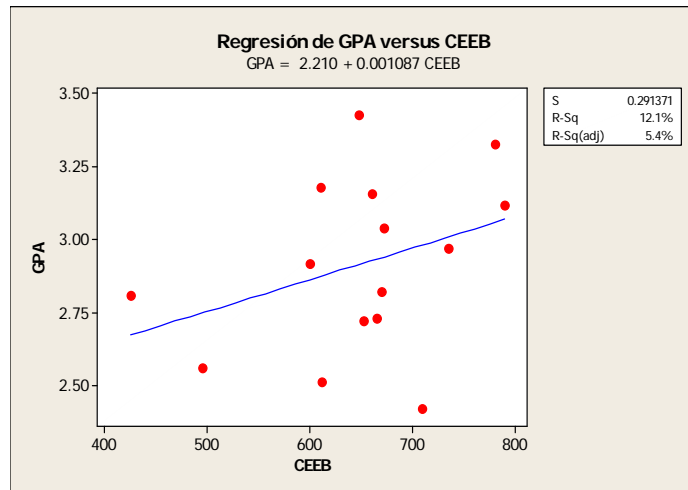
Supongamos que se desea establecer una relación entre la nota que un estudiante obtiene en la parte de aprovechamiento matemático de ingreso (CEEB) y el Promedio académico al final de su primer año de universidad (GPA). Se toma una muestra de 15 estudiantes y se obtiene los siguientes datos:

Est	CEEB	GPA	Est	CEEB	GPA
1	425	2.81	8	660	3.16
2	495	2.56	9	665	2.73
3	600	2.92	10	670	2.82
4	610	3.18	11	720	3.04
5	612	2.51	12	710	2.42
6	648	3.43	13	735	2.97
7	652	2.72	14	780	3.33
			15	790	3.12

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 (Ejemplo 3.25.)

La variable independiente es CEEB y la variable dependiente es GPA. La gráfica es:



**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 línea.

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.

# 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.