ANALISIS DE DATOS

Dr. Edgar Acuna

http://academic.uprm.edu/eacuna

UNIVERSIDAD DE PUERTO RICO

RECINTO UNIVERSITARIO DE MAYAGUEZ

2. 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

2.1Organización de datos Cuantitativos Discretos

2.1.1 Tablas de Frecuencias: Los datos cuantitativos discretos se organizan en tablas, llamadas ***Tablas de Distribución de frecuencias. Los*** tipos de frecuencias: son

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.

Tabla de distribucion de frecuencias

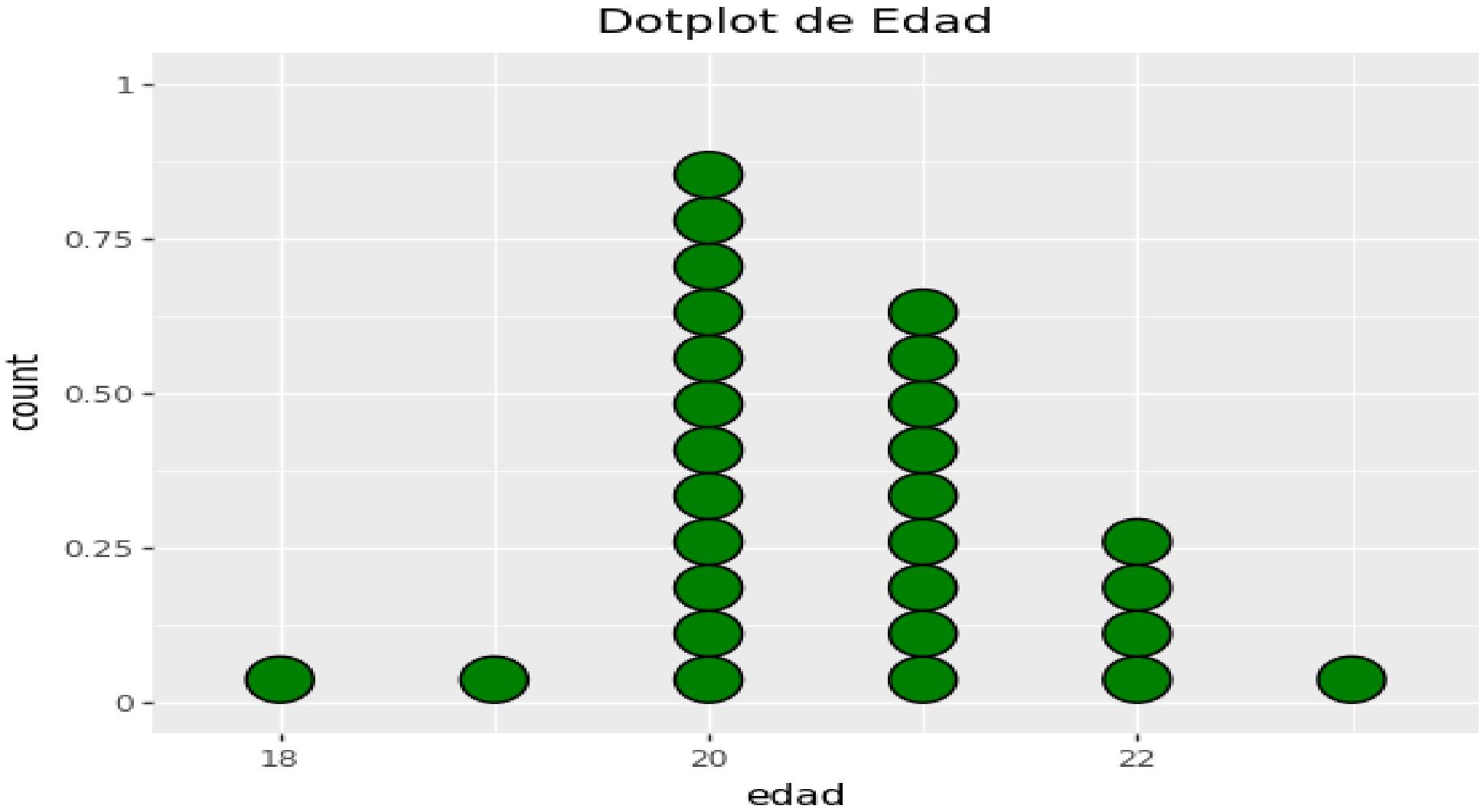
edad	fabs	frelat	fabs.acum	frelat.acum
18	1	3.571429	1	3.571429
19	1	3.571429	2	7.142857
20	12	42.857143	14	50.000000
21	9	32.142857	23	82.142857
22	4	14.285714	27	96.428571
23	1	3.571429	28	100.000000

Laboratorios 1 y 2

Lab 1: Leyendo datos y calculo de frecuencias para datos discretos
Lab 2: Leyendo datos y tabla de freccuencias datos discretos

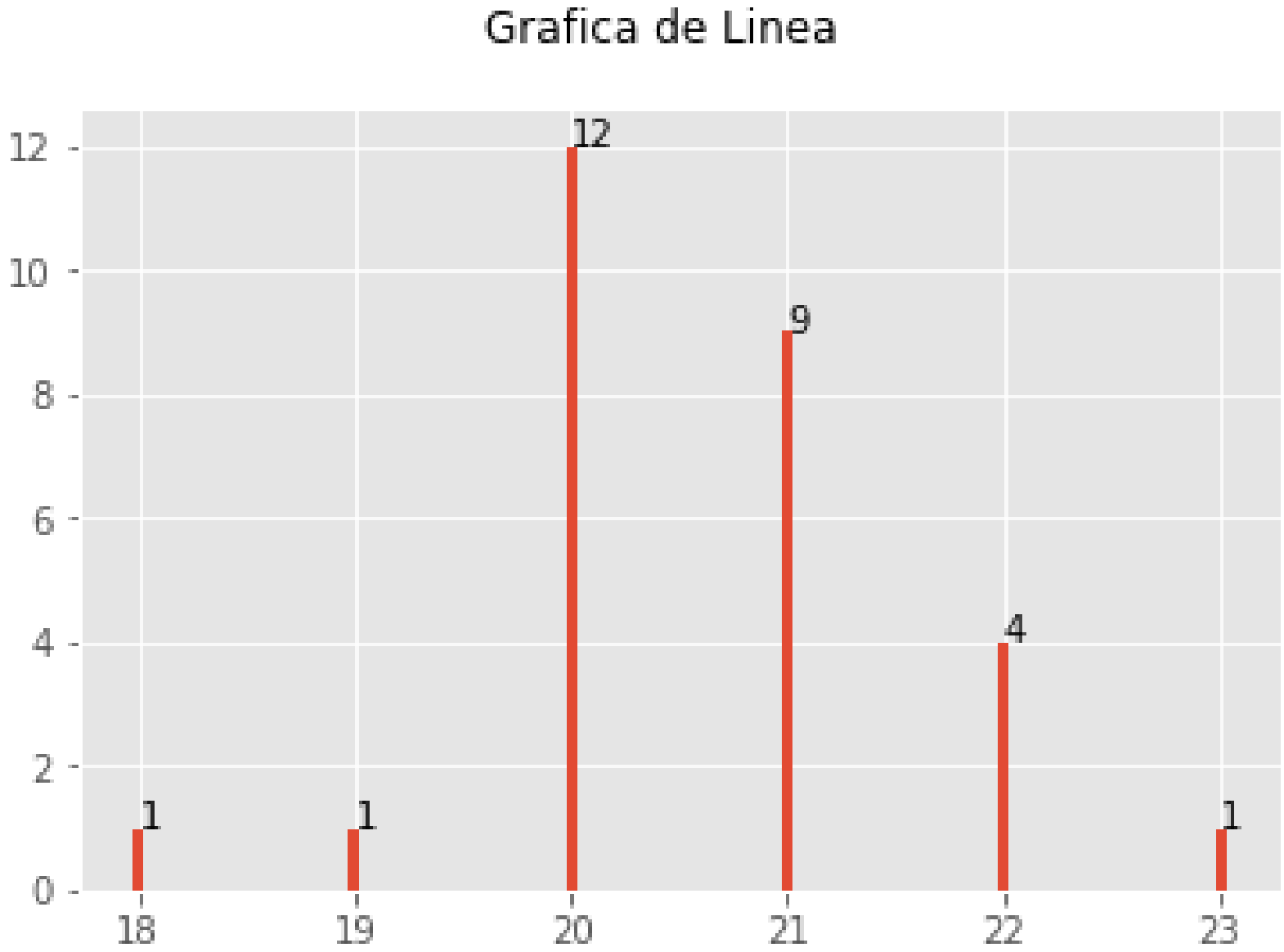
2.1.2 El plot de puntos (“Dotplot”)

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”).


Dotplot de Edad
count
1
0.75
0.50
0.25
0
18
20
22
edad

2.1.3 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.


Grafica de Linea
12
10
8
6
4
2
0
18
19
20
21
22
23
1
1
12
9
4
1

2.2 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:

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

2.2 Organización de datos Cuantitativos Continuos (cont)

Una regla bien usada es que el número de clases debe ser aproximadamente igual a la raíz cuadrada del número de datos. También está la regla de Sturges,en donde el número de clases está dado por 1+3.3*log(número de datos)

Una vez que se determina el número de clases se determina la amplitud de cada clase usando la siguiente fórmula:

Amplitud del intervalo de clase $\approx$ (Mayor Valor-Menor Valor)/numero de intervalos

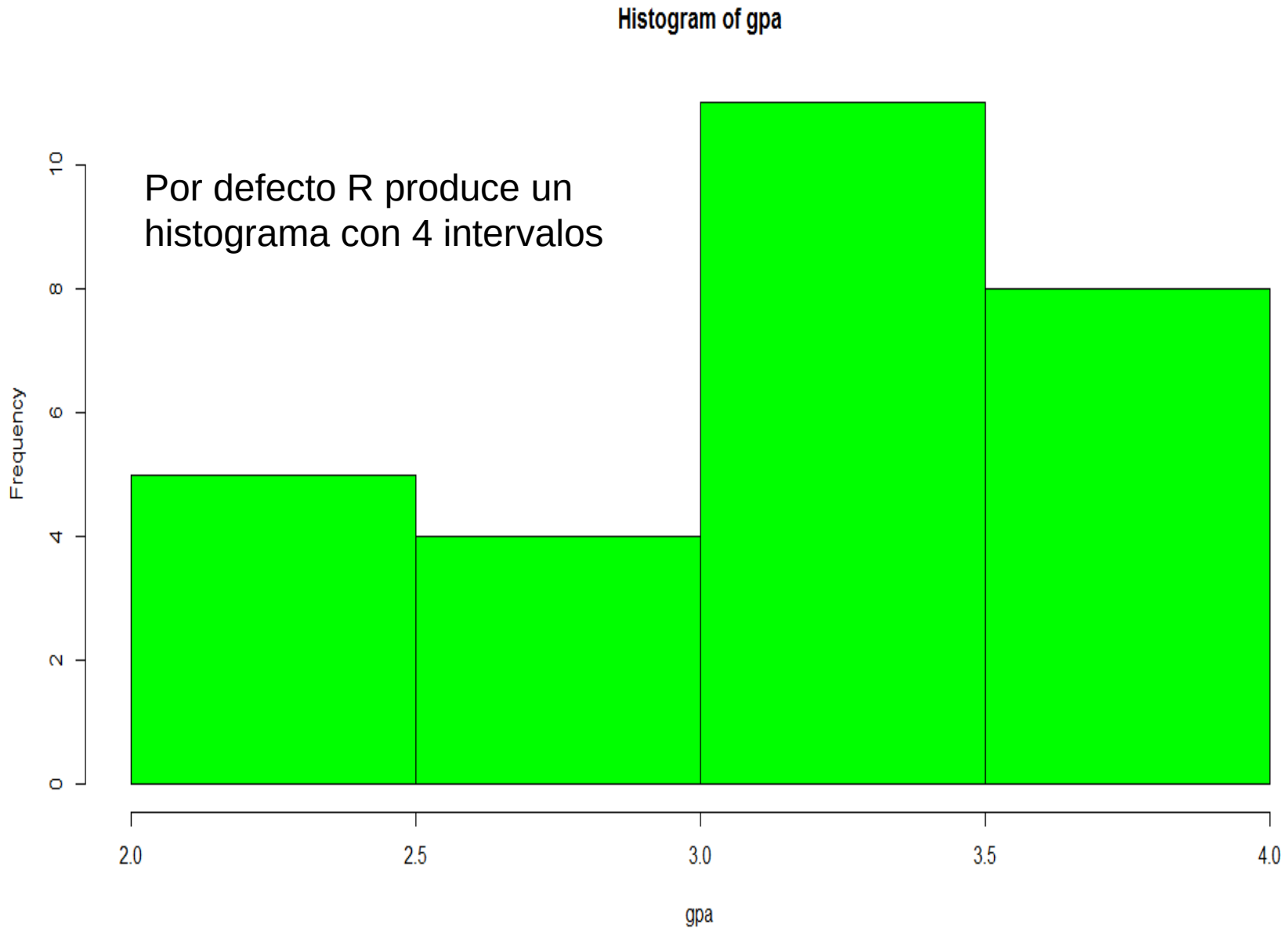
Usualmente la amplitud se redondea a un número cómodo de usar.

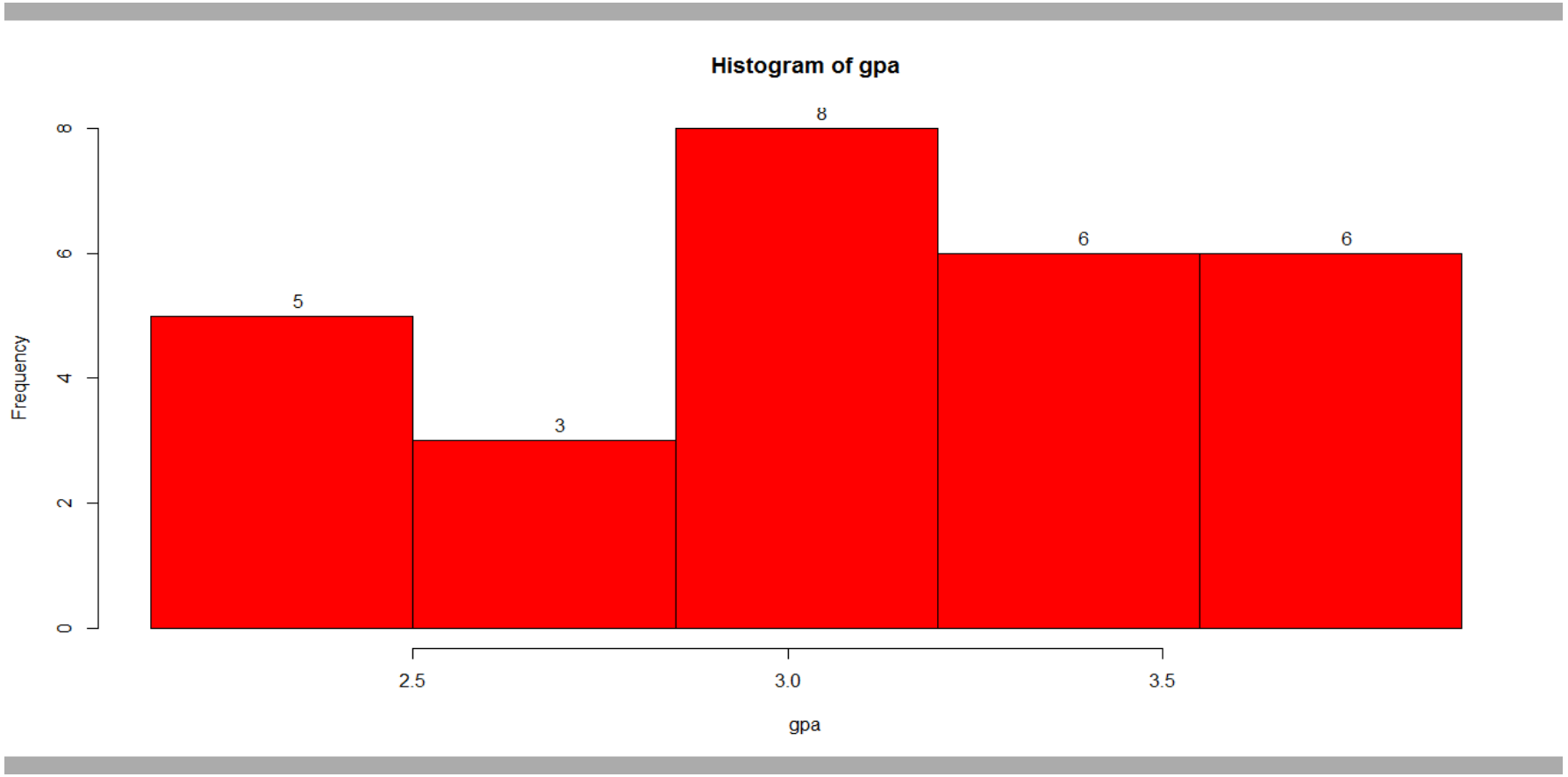
3.2.2 Histograma

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.

Laboratorio 3

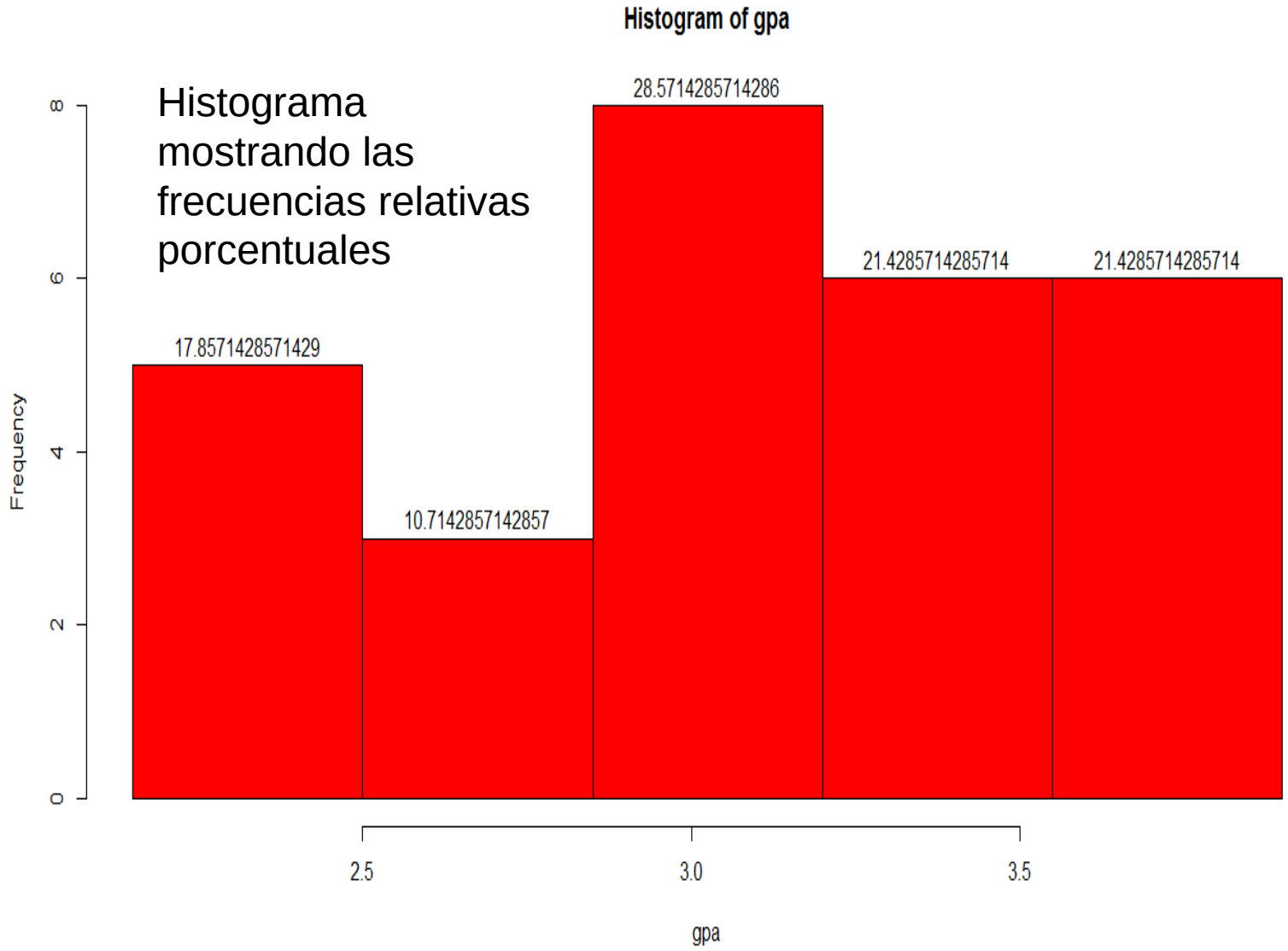
Tabla de frecuencias para datos continuos e histograma
Tabla de frecuencias para datos categoricos

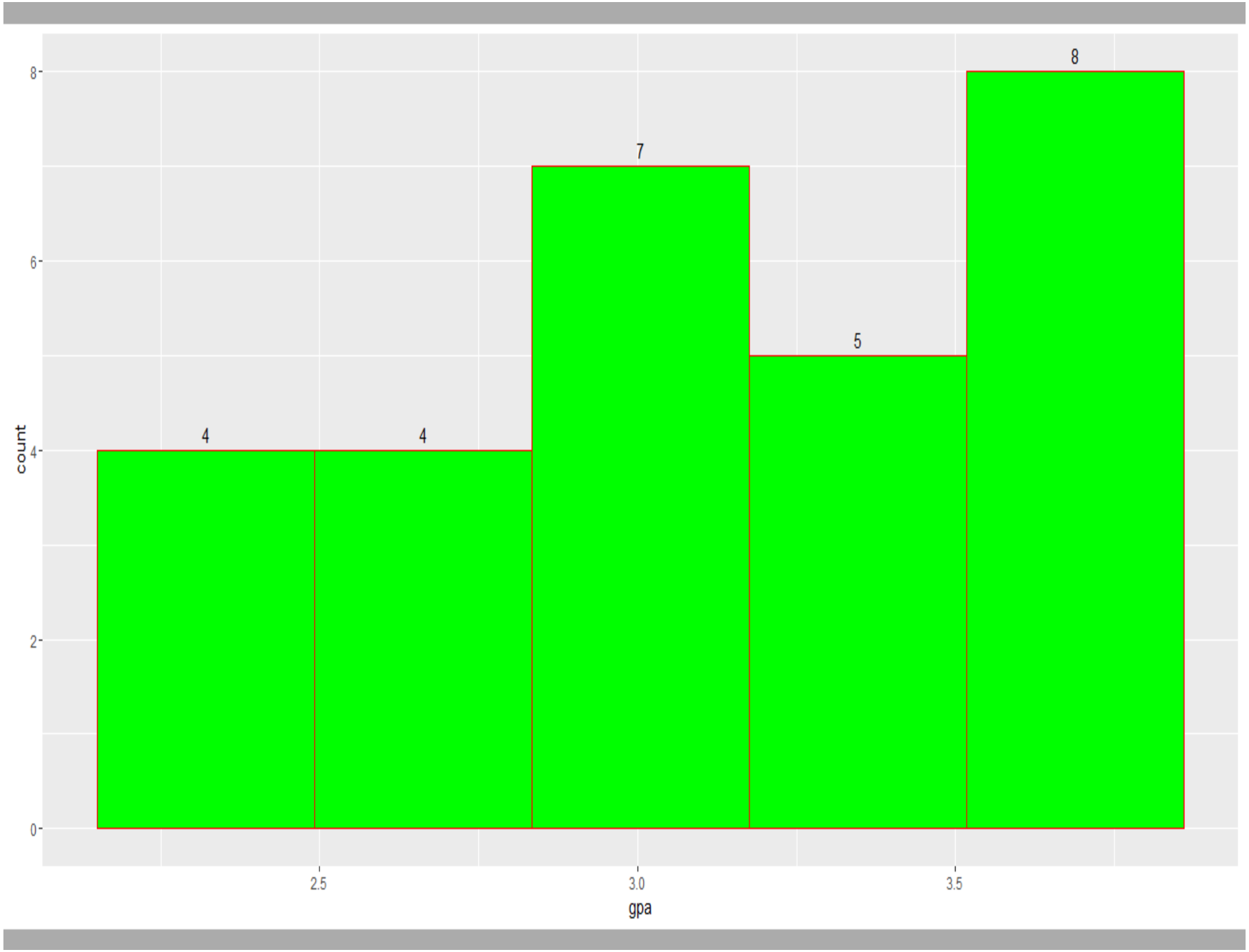

Histogram of gpa
Por defecto R produce un histograma con 4 intervalos
Frequency
0
2
4
6
8
10
2.0
2.5
3.0
3.5
4.0
gpa



Histograma asimetrico hacia la izquierda, la frecuencia es baja para valores bajos de GPA y la frecuencia es alta para valores altos de GPA

Histograma mostrando las frecuencias relativas porcentuales




8
6
4
2
0
count
4
4
7
5
8
2.5
3.0
3.5
gpa

Interpretacion del histograma

Un histograma es simetrico con respecto al centro si los intervalos a la derecha e izquierda del centro tienen similar frecuencia.

Un histograma es asimetrico hacia la derecha si hay mayor concentracion de datos a la izquierda que a la derecha del centro. O sea para valores altos de las variablesle corresponde mayor frecuencia que valores bajos de la variable.

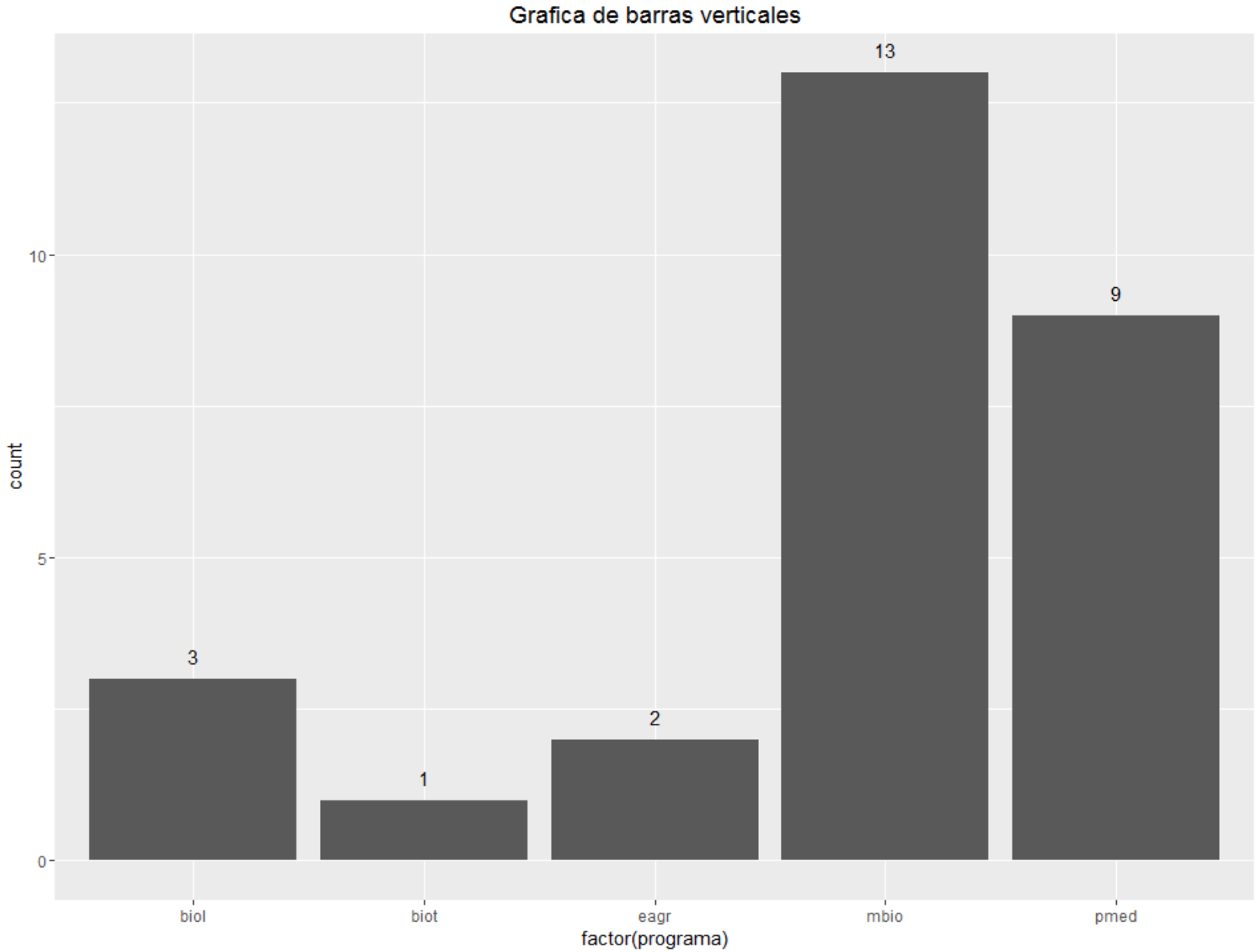
Un histograma es asimetrico a la izquierda si hay mayor concentracion de datos a la derecha que a la izquierda del centro. O sea que a valores bajos de la variable le corresponde mayor frecuencia que a valores altos de la variable.

2.3 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.

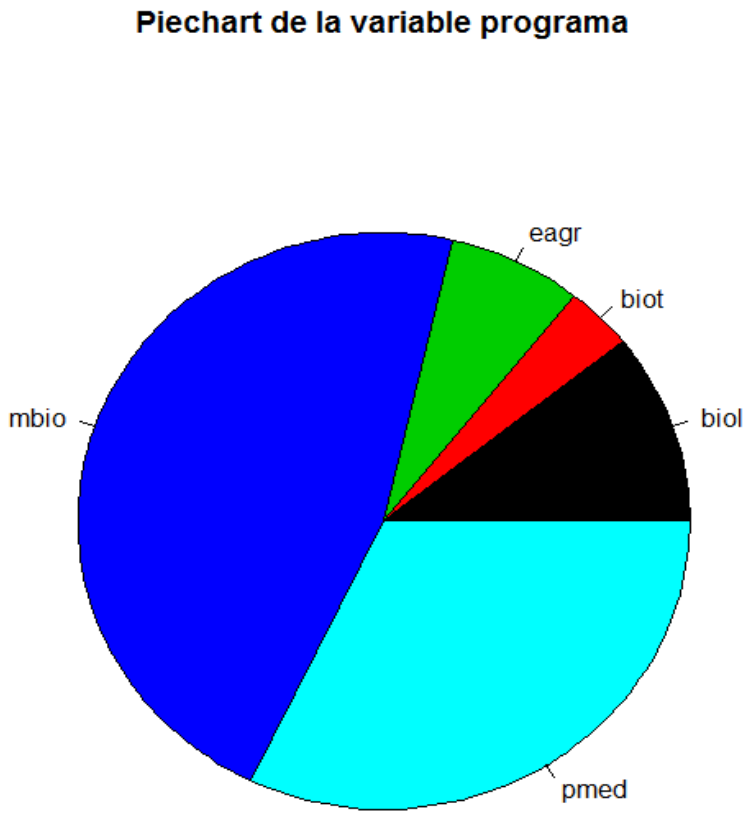
2.3.1 Gráficas de Barras

Las gráficas de barras pueden ser verticales u horizontales.


Grafica de barras verticales
13
9
3
1
2
10
5
0
count
biol
biot
eagr
mbio
pmed
factor(programa)

2.3.2 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.


Piechart de la variable programa
eagr
biot
biol
mbio
pmed

Laboratorio 4

Modulos para Graficas en Python

Hay varias librerias para hacer graficas interactivas en R .
Las principales son:

Seborn (2012)
Ploltnine (2014)
Plotly (2014)

2.4 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. 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").

2.4.1 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,…,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:

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

Medidas de Centralidad (cont.)

La media o promedio se obtiene sumando todos los datos y dividiendo entre el número de datos. Es decir, si x_1, x_2,…,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 2.6. 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.

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

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

Medidas de Centralidad (cont.)

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 2.7. 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.

Medidas de Centralidad (cont.)

Propiedades de la media son:

A) El valor de la media debe estar entre el mayor dato y el menor dato.

B) Si a cada dato de la muestra se le 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.

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.

Medidas de Centralidad (cont.)

Ejemplo 2.8. Calcular la mediana de los datos del Ejemplo 3.6.

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. Asi, la mediana para los datos del ejemplo 3.7, donde hay

un número par de datos, la mediana resulta ser el promedio de los dos valores centrales: =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.

Medidas de Centralidad (cont.)

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 2.10. 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 y 4

Hallar la moda.

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

Ejemplo 2.11. Los siguientes datos representan tipos de sangre de 9 personas

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

Hallar la Moda.

La Moda es el tipo de sangre O.

Medidas de Centralidad (cont.)

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 2.12. Hallar la media podada del 5 por ciento para los datos del Ejemplo 2.7

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

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

.

2.4.2 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.

Otra medida, para determinar el grado de concentración de los datos sería el promedio de las desviaciones con respecto a la media, es decir ,

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

Sin embargo esto siempre daria CERO porque la suma de las desviaciones con respecto a la media es siempre CERO.

Medidas de Variabilidad (cont)

La **varianza** es una medida que da una idea del grado de concentración de los datos con respecto a la media.

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

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

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.

Medidas de Variabilidad (cont)

Ejemplo 2.13. 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.

El **coeficiente de variación** (CV) y que se calcula por $\frac{s}{|\bar{x}|}x100\%$. Si e l CV es mayor que 30% la muestra es muy variable y si CV es menor del 10% 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.

Medidas de Variabilidad (cont)

Ejemplo 2.13. 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.

El **coeficiente de variación** (CV) y que se calcula por $\frac{s}{|\bar{x}|}x100\%$. Si e l CV es mayor que 30% la muestra es muy variable y si CV es menor del 10% 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.

Medidas de Variabilidad (cont)

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". El criterio debe ser usado con precaución, puesto que la media, la varianza y la desviación estándar son afectadas por la presencia de "outliers".

Ejemplo 2.14. Dada la siguiente muestra

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

Determinar si 98 es un "outlier".

Solución:

Como 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".

Laboratorio 5

2.4.3. 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.

Existen varios métodos para calcular los cuartiles pero la manera mas simple es ordenar los datos y considerar Q_1 como 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.

Medidas de Posición (cont)

Ejemplo 2.15. 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 $Q_1 =$ $= 17.5$

La segunda mitad es: 23, 28, 29,39, 40, 44, luego $Q_3 =$ $= 34$

Una variante que se usa cuando el número de datos es impar es no usar la mediana, para cada mitad. Es decir en el ejemplo b), 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. R usa un proceso de interpolación para calcular los cuartiles

Medidas de Posición (cont)

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

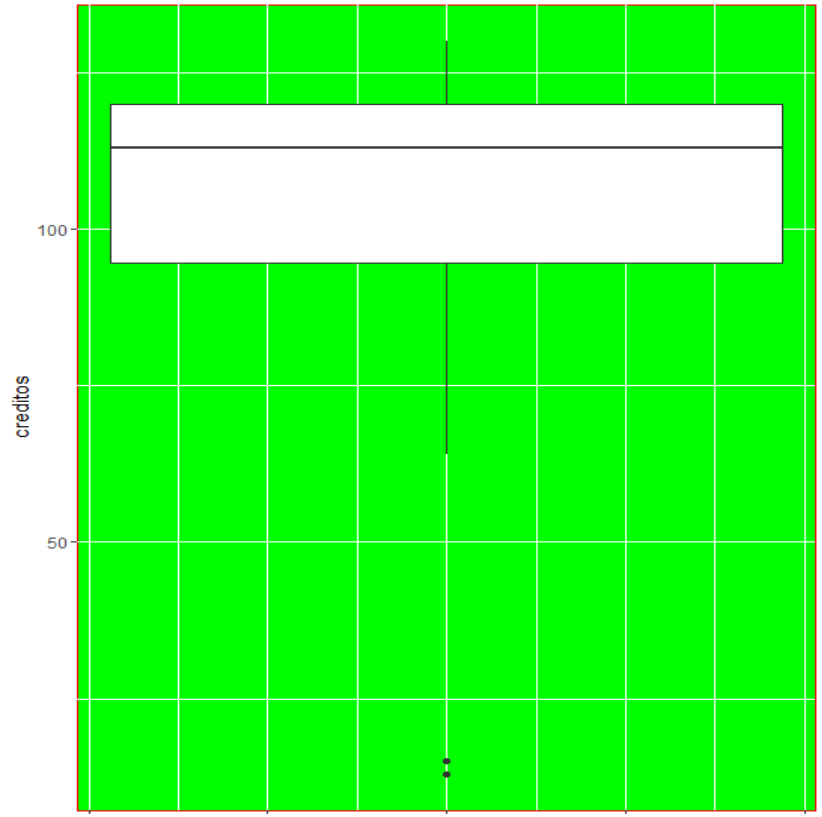
Los Percentiles: Dado un cierto porcentaje 100p, donde p varía entre 0 y 1, el percentil del 100p% es un valor tal que 100p% de los datos caen a la izquierda del percentil. En particular, la mediana y los cuartiles son percentiles. El primer cuartil es el percentil de 25%, la mediana es el percentil del 50% y el tercer cuartil es el percentil del 75%. Por ejemplo el puntaje minimo para que un estudiante este en el tercio superior de su clase es igual al puntaje de 66.67%.

•

2.6 Diagrama de caja (“Boxplot”)

El “**Boxplot”** , 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 **R** hay varias maneras de obtener el “Boxplot” de un conjunto de datos, la primera es usando la funcion la bo***plot*** . Otra manera es usando la funcion geom_boxplot de la librería ggplot2



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.

2.6 Diagrama de caja ("Boxplot") cont.

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**.

Laboratorio 6

2.7 Organización y Presentación de datos Bivariados

2.7.1 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 **R** se usa elige la funcion ***table***

Ejemplo 2.16

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

Ejemplo 2.16 (cont)

Los datos estan disponible en academic.uprm.edu/eacuna/eje316.txt

```
> table(escuela,aprueba)
       aprueba
escuela no si
   priv     3  7
   publ     5  5
```

*Hay 7 estudiantes que son de escuela privada y que aprueban el examen. Un (7/12*100%=58.33% 0de los que aprueban el examen son de escuela privada*

Ejemplo 2.17.

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. Los datos estan disponibles en http://academic.uprm.edu/eacuna/eje2biv.txt.

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

Usar **R** 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 usa el comando xtabs de R

```
> xtabs(conteo~Sexo+Opinion)
        Opinion
Sexo     abst no si
  female   44 31 15
  male     30 20 10
>
```

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.

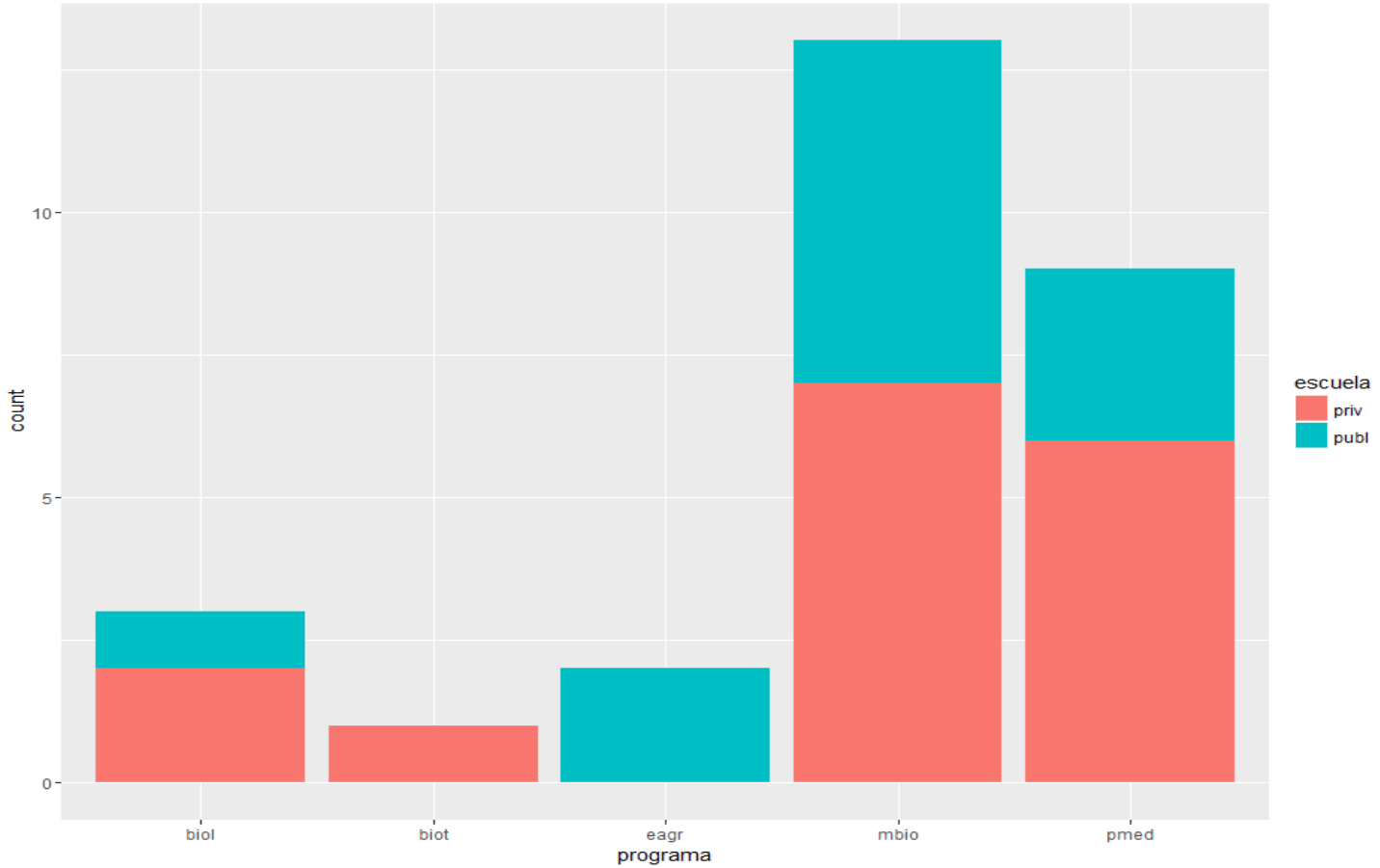
Ejemplo 2.20

Hallar una gráfica de partes componentes para comparar los estudiantes (por programa) según el tipo de escuela de donde proceden, usando datos del ejemplo 2.1.

Solución:

Continuación (Ejemplo 2.20)

Solución:

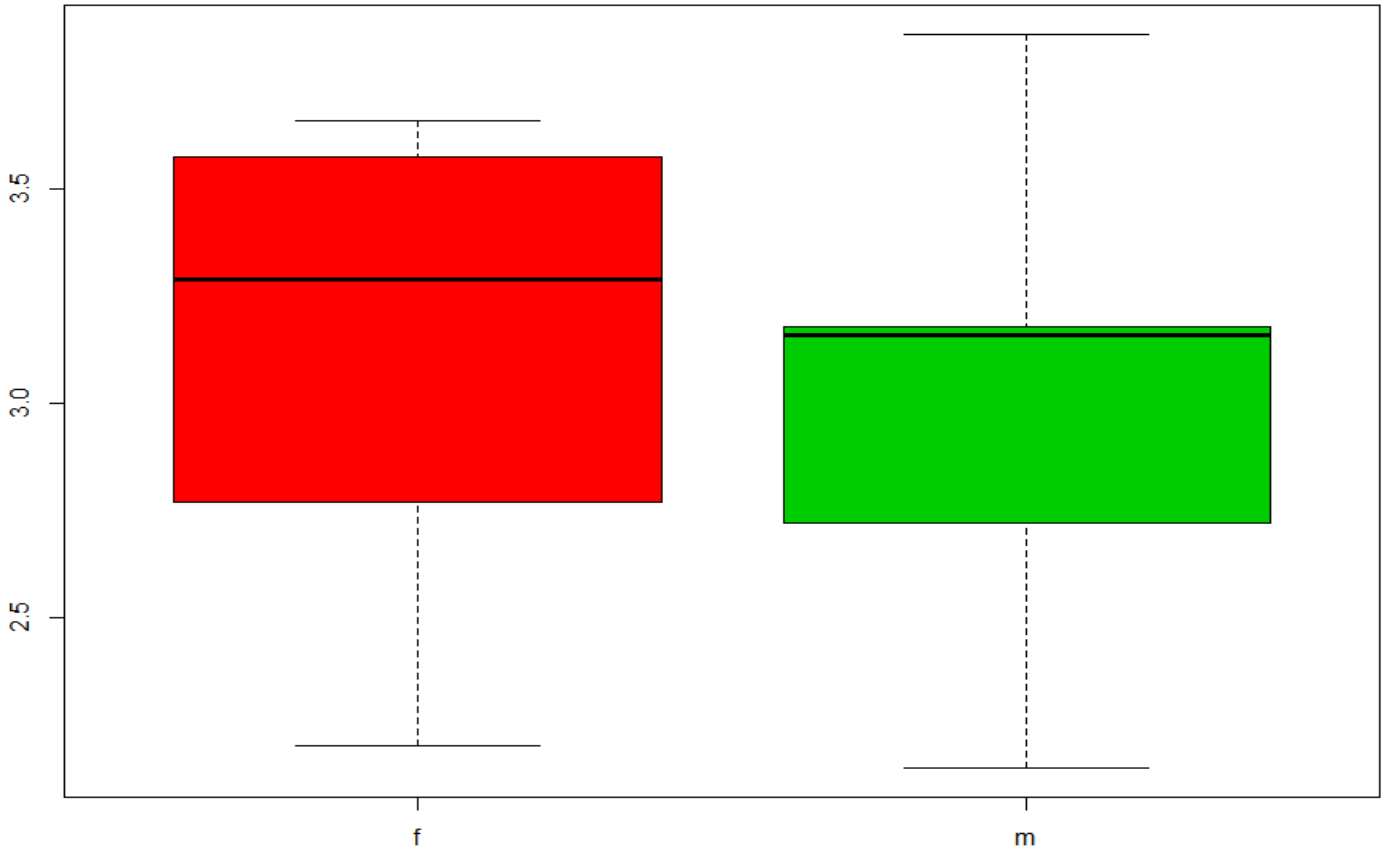


Gráfica de barras en partes componentes para la variable *Programa* según *Escuela*

2.7.2 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* .

La gráfica más adecuada para representar este tipo de información es el "Boxplot".


3.5
3.0
2.5
f
m

2.7.3 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 2.22

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.

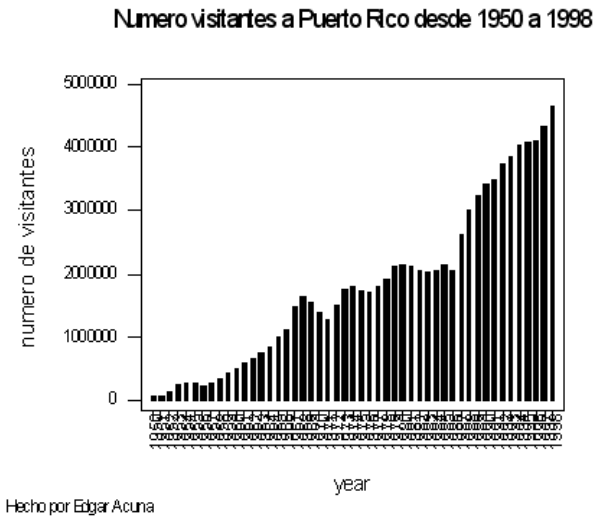


Figura 3.43 Gráfica de barras del número de visitantes a Puerto Rico entre 1950-1998.

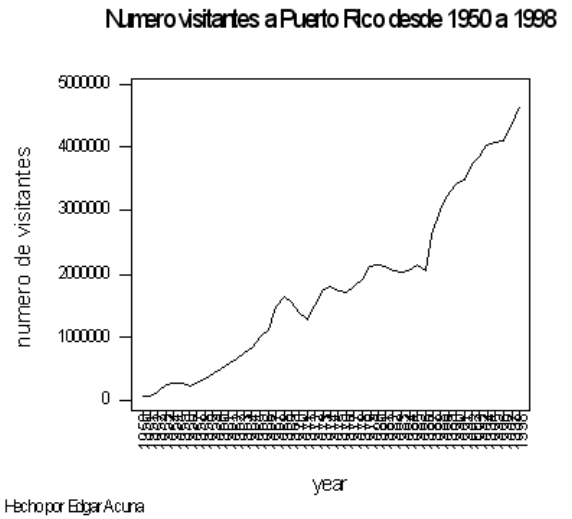


Figura 3.44 Gráfica de barras del número de visitantes a Puerto Rico entre 1950-1998.

Laboratorio 7

2.8 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:

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

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.

Ejemplo 2.23.

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:

Solución: (Ejemplo 2.23.)

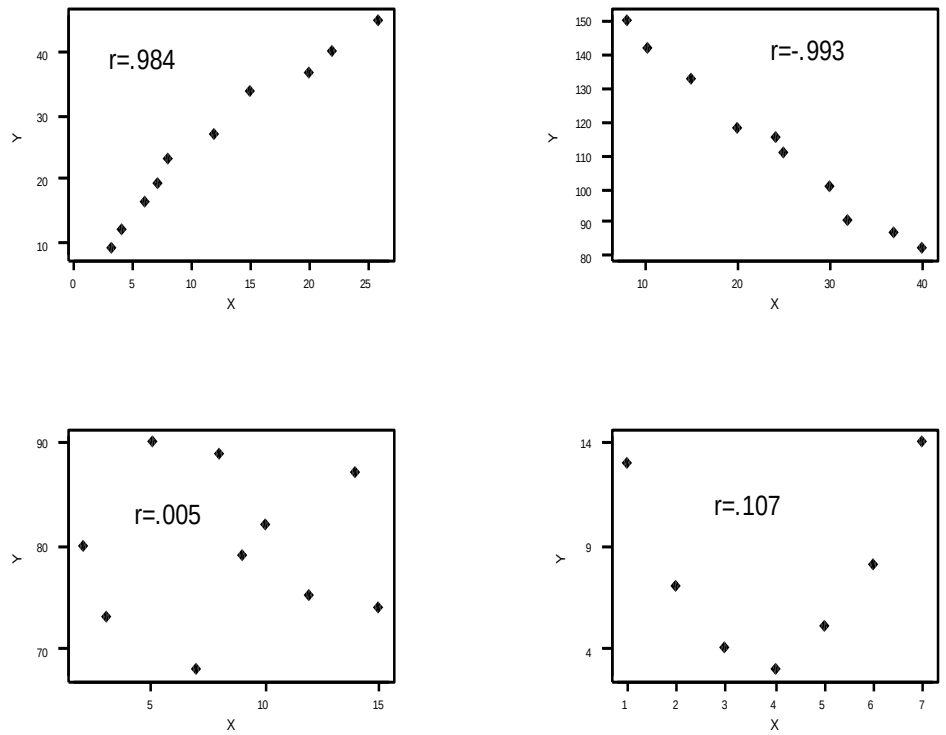
Interpretación:

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.

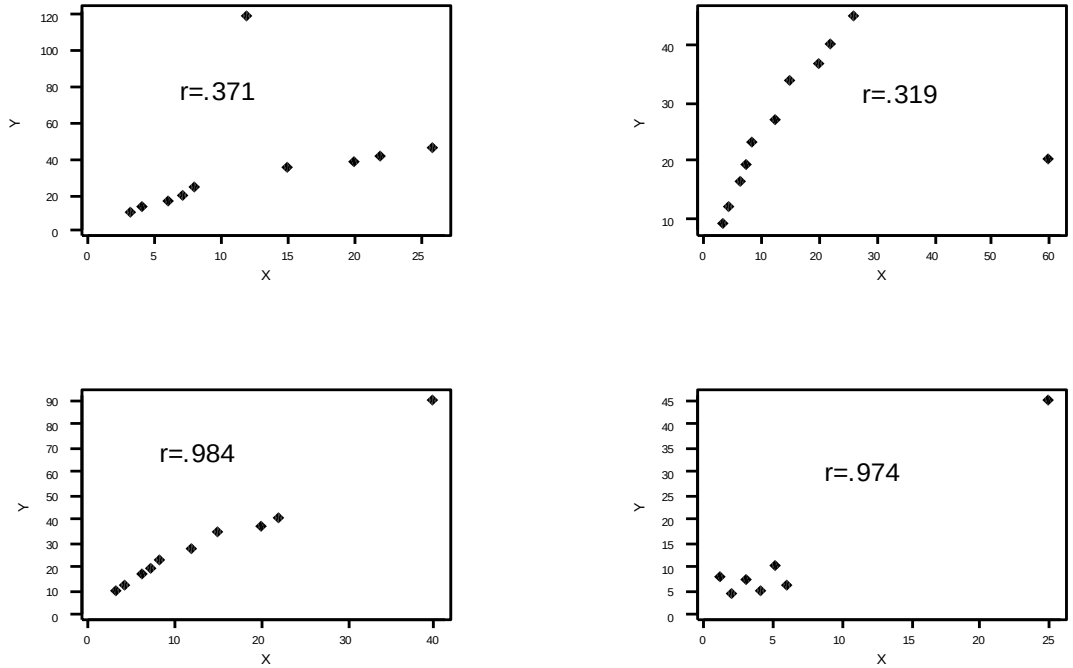
Row	years	ventas	Sxx	Syy	Sxy	r
1	3	9	590.1	1385.6	889.4	0.983593
2	4	12				
3	6	16				
4	7	19				
5	8	23				
6	12	27				
7	15	34				
8	20	37				
9	22	40				
10	26	45				

En **R**, el coeficiente de correlación se puede obtener usando la función **cor**

Coeficiente de Correlacion para diversos plots



Efecto de valores anormales en el valor de la correlacion



2.9 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:

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

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

2.8 Una introducción a Regresión Lineal (cont)

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* $\hat{\alpha}$ *.*

El coeficiente de detreminacion R^2

Es una medida de la bondad de ajuste de la linea a los datos. Varia entre 0 y 100% y mientras mas ce acerca a 100% mejor es el ajuste.

Se puede calcular elevando la correlacion al cuadrado y multiplicando por 100

Ejemplo 2.25.

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
7	652	2.72

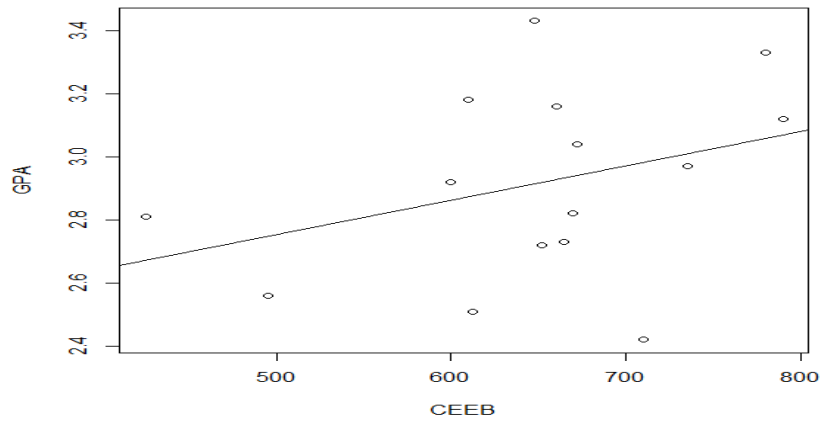
Est	CEEB	GPA
8	660	3.16
9	665	2.73
10	670	2.82
11	720	3.04
12	710	2.42
13	735	2.97
14	780	3.33
15	790	3.12

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

Los datos estan disponibles en academic.uprm.edu/eacuna/eje1reg.txt

Solución (Ejemplo 2.25.)

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 esta dada por

> print(lm(GPA~CEEB))

Call:

lm(formula = GPA ~ CEEB)

Coefficients:

(Intercept)	CEEB
2.209878	0.001087

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.

Laboratorio 8