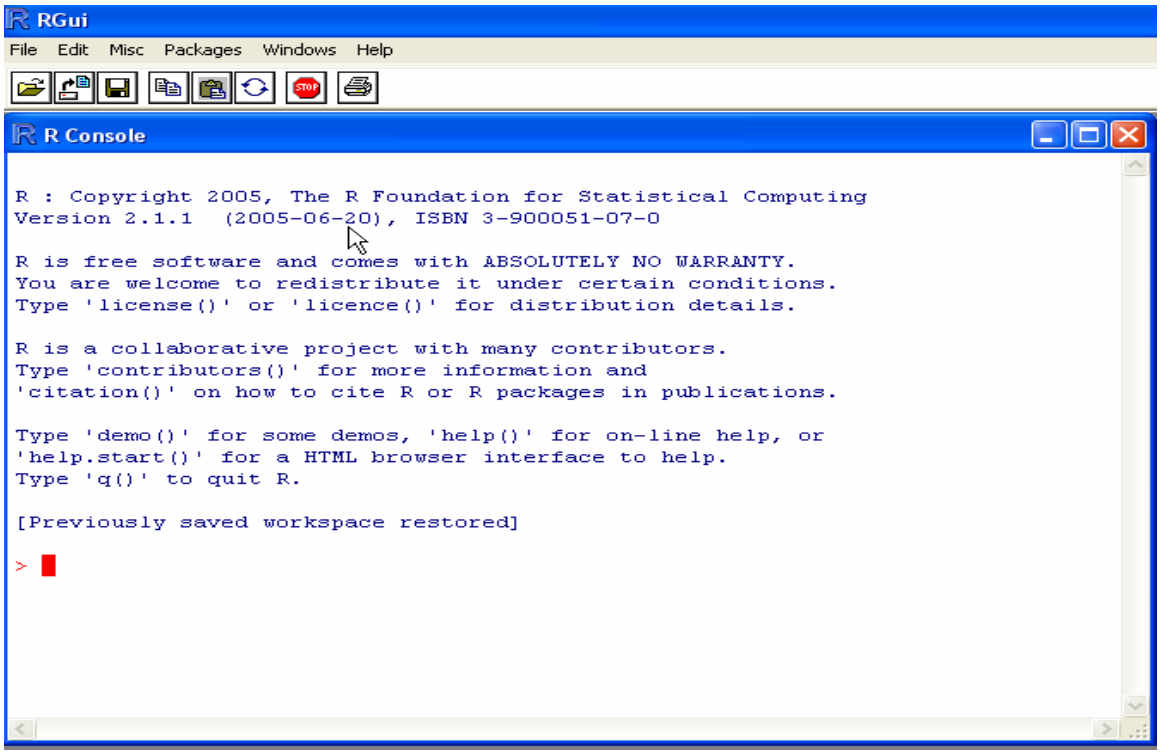
INTRODUCCIÓN AL ANALISIS ESTADISTICO CON R
UNA MIRADA RAPIDA

INTRODUCCION

R es un sistema para análisis estadísticos y gráficos creado por R. Ihaka y R. Gentleman. R tiene una naturaleza doble de programa y lenguaje de programación y es considerado como un dialecto del lenguaje S creado por los Laboratorios AT&T Bell. S está disponible como el programa S-PLUS comercializado por Insightful. Existen diferencias importantes en el diseño de R y S: aquellos interesados en averiguar más sobre este tema pueden leer el artículo publicado por Ihaka & Gentleman (1996) o las Preguntas Más Frecuentes en R, que también se distribuyen con el programa.

R se distribuye gratuitamente bajo los términos de la *GNU General Public Licence*; su desarrollo y distribución son llevados a cabo por varios estadísticos conocidos como el *Grupo Nuclear de Desarrollo de R*. El programa R se puede descargar de **`http://cran.r-project.org`**

Al abrir R aparece la siguiente pantalla de entrada



CAPITULO 1
ESTADISTICA DESCRIPTIVA

Intentemos estudiar algunas opciones básicas del lenguaje R con la aplicación de él, inicialmente al estudio descriptivo de datos. Para ello veamos el siguiente ejemplo

Ejemplo 1
Para conocer el comportamiento de una máquina automática que deposita un líquido en vasos, se seleccionó una muestra de 49 de ellas. Al medir el contenido, en onzas, se obtuvo los siguientes resultados.

7.85 7.86 7.87 7.87 7.88 7.89 7.92 7.94 7.95 7.96 7.97 7.97 7.98
7.99 7.99 8.01 8.03 8.03 8.04 8.05 8.05 8.05 8.05 8.05 8.06 8.06
8.06 8.07 8.07 8.07 8.08 8.09 8.09 8.09 8.10 8.10 8.10 8.11 8.11
8.12 8.16 8.16 8.17 8.19 8.21 8.21 8.22 8.24 8.26

Solución
En File debemos marcar la opción New Scrip y guardar con algún nombre, por ejemplo, vasos R; enseguida usando la opción Open Scrip estamos en condiciones de iniciar la sesión.

El símbolo > llamado **`prompt`** es propio del sistema y nos indica que espera instrucciones.

Ingresamos los datos creando un vector con nombre “vasos” (por ejemplo) usando la función **`c()`**

```
> vasos<-c(7.85,7.86,7.87,7.87,7.88,7.89,7.92,7.94,7.95,7.96,7.97,
+ 7.97,7.98,7.99,7.99,8.01,8.03,8.03,8.04,8.05,8.05,8.05,8.05,
+ 8.05,8.06,8.06,8.06,8.07,8.07,8.07,8.08,8.09,8.09,8.09,8.10,8.10,
+ 8.10,8.11,8.11,8.12,8.16,8.16,8.17,8.19,8.21,8.21,8.22,8.24,8.26)
```

El símbolo + lo coloca el programa cuando cambia de línea

Ejemplo 2
Podemos verificar que la cantidad de datos que hemos ingresado es la correcta, para ello, basta con usar la función **`length`** con el nombre del archivo entre paréntesis:

```
> length(vasos)
[1] 49
```

MEDIDAS DE TENDENCIA CENTRAL
Necesitamos caracterizar la muestra con algunas medidas de tendencia central, las cuales sintetizan, en si mismo, la característica “central” de la muestra. Tenemos:

a) **Media Aritmética** o media de la variable; en este caso la variable es
X = “cantidad en onzas, de un líquido depositado por una máquina automática”.

Denotada $\overline{X}$ es tal que $\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$ donde n es el número de unidades de observación (en nuestro caso el número de vasos seleccionados) y x_i es el valor que toma cada una de ellas.

Ejemplo 3
Con R calculamos la media aritmética como sigue:

```
> mean(vasos)
[1] 8.05
```

Esto significa que, si todos los vasos tuviesen la misma cantidad, esa cantidad común sería 8,05 onzas

b) Una segunda medida de tendencia central es la **mediana**; valor de la variable que supera a no más del 50% de los datos y es superado por no más del 50% de los datos, naturalmente que la variable debe ser al menos ordinal
En R la conseguimos como sigue:

Ejemplo 4

```
> median(vasos)
[1] 8.06
```

Esto significa que el 50% de los vasos tiene un volumen depositado a lo más de 8,06 onzas (o a lo menos de 8,06 onzas)

c) Podemos obtener, además, **percentiles.** Existen 99 percentiles donde, si $p = 1,2,3..,99$ entonces el percentil p denotado P_p es aquel valor de la variable que supera a no mas del p% de los datos y es superado por no más del (100-p)% de los datos. Observamos que el percentil 50 es la mediana, que el percentil 25 es el cuartel 1, que el percentil 75 es el cuartil 3

Ejemplo 5
Si queremos calcular el percentil 10, denotado P_{10} , con R anotamos

```
> quantile(vasos,0.1)
  10%
7.888
```

Esto significa que el 10% de los vasos tiene un contenido a lo más de 7,888 onzas, o lo que es lo mismo, que el 90% de los vasos tiene un contenido de al menos 7,888 onzas

Ejemplo 6
Podríamos estar interesados en el percentil 10 tanto como el percentil 75, al mismo tiempo. Con el lenguaje R procedemos como sigue:

```
> quantile(vasos,c(0.1,0.75))
  10%   75%
7.888 8.100
```

Ejemplo 7
El lenguaje R provee una función denominada **fivenum,** propuesta por el estadístico John W. Tukey, la cual calcula cinco valores que describen concisamente al conjunto de datos, son: el valor mínimo, los percentiles 25, 50 y 75, y el valor máximo

```
> fivenum(vasos)
[1] 7.85 7.98 8.06 8.10 8.26
```

El valor mínimo es 7,85 ; $P_{25} = Q_1 = 7{,}98$; $P_{50} = Q_2 = Mediana = 8{,}06$
$P_{75} = Q_3 = 8{,}10$; el valor máximo es 8,26

Ejemplo 8
Observe la siguiente instrucción

```
> summary(vasos)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.85    7.98    8.06    8.05    8.10    8.26
```

MEDIDAS DE DISPERSION
Una vez declaradas algunas medidas de tendencia central debemos caracterizar la muestra de vasos con medidas de dispersión
La más básica es la amplitud de la variable o **rango de la variable**, R nos entrega dos números que son los extremos del rango

Ejemplo 9

```
> range(vasos)
[1] 7.85 8.26
```

Ejemplo 10
Otra medida de dispersión es el **recorrido intercuartílico**, se define como $IQR = Q_3 - Q_1$, en el lenguaje R la función es `IQR`

```
> IQR(vasos)
[1] 0.12
```

Esto significa que la mayor diferencia entre dos números dentro del 50% de los valores centrales es 0,12

La **varianza y la desviación estándar** son medidas de dispersión muy usadas.

La varianza de una variable X se denota $V(X)$ o $\sigma^2(X)$ o S_n^2 para el caso en que trabajamos con la población y, para el caso en que trabajamos con una muestra de tamaño n se denota S_{n-1}^2; esta última se define por

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

La desviación estándar es la raíz cuadrada de la varianza, así,

$$S_{n-1} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}}$$

En el lenguaje R conseguimos la varianza y la desviación estándar muestral, respectivamente por **`var y sd`**

Ejemplo 11

```
> var(vasos)
[1] 0.01073333

> sd(vasos)
[1] 0.1036018
```

Ejemplo 12

Si deseamos la varianza poblacional $S_n^2 = \frac{n-1}{n}S_{n-1}^2$, en el lenguaje R la escribimos:

```
> ((length(vasos)-1)/length(vasos))*var(vasos)
[1] 0.01051429
```

Una buena medida de dispersión es el **coeficiente de variabilidad**; es el cuociente entre la desviación estándar y la media, expresada en porcentaje; denotada CV

$$CV(X) = \frac{S}{\overline{X}}100$$

En palabras es el cociente de la desviación estándar y la media expresada en %, esta medida es fácil de interpretar, y en la mayoría de los casos es adecuada para comparar dos conjuntos de datos.

En el lenguaje R se calcula como sigue:

Ejemplo 13

```
> 100*sd(vasos)/mean(vasos)
[1] 34.22613
```

REPRESENTACION GRAFICA

En muchas ocasiones un gráfico es muy eficiente para caracterizar y comunicar mejor a un conjunto de datos.
Algunas de las funciones gráficas más utilizadas en R son: histograma, diagrama de caja, diagrama de barras, diagrama de tortas

1) Histograma

Un diagrama usado preferentemente cuando la variable en estudio es continua o cuando es discreta con muchas observaciones
Es un conjunto de rectángulos contiguos con base la amplitud de cada intervalo o clase y altura el número de observaciones o datos que allí pertenecen, la función es **`hist(x)`**

```
hist(x, nclass=n)
hist(x, breaks=b, ...)
```

Produce un histograma del vector numérico x. El número de clases se calcula habitualmente por defecto, pero puede elegir uno con el argumento **`nclass,`** o bien especificar los puntos de corte con el argumento **`breaks`**. Si está presente el argumento **`probability=TRUE`**, se representan frecuencias relativas en vez de absolutas.

Existe una serie de argumentos que pueden pasarse a las funciones gráficas, entre otros los siguientes:

`type=`

Este argumento controla el tipo de gráfico producido, de acuerdo a las siguientes posibilidades:

`type="p"` Dibuja puntos individuales. Este es el valor predeterminado

`type="l"` Dibuja líneas

`type="b"` Dibuja puntos y líneas que los unen

`type="o"` Dibuja puntos y líneas que los unen, cubriéndolos

`type="h"` Dibuja líneas verticales desde cada punto al eje X

`type="s"`
`type="S"`
Dibuja un gráfico de escalera. En la primera forma, la escalera comienza hacia la derecha, en la segunda, hacia arriba.

`xlab=cadena`
`ylab=cadena`
Definen las etiquetas de los ejes x e y, en vez de utilizar las etiquetas predeterminadas, que normalmente son los nombres de los objetos utilizados en la llamada a la función gráfica de nivel alto.

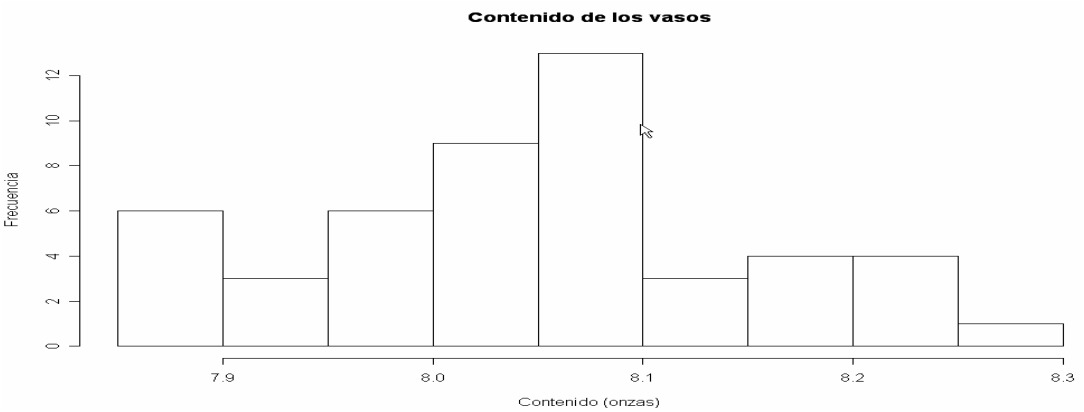
`main=cadena`
Título del gráfico, aparece en la parte superior con tamaño de letra grande.

`sub=cadena`
Subtítulo del gráfico, aparece debajo del eje x con tamaño de letra pequeño.

El histograma del "contenido de los vasos" se obtiene con la instrucción **hist** , debemos observar los parámetros adicionales: **main** para el titulo principal y los rótulos de cada eje dados por **xlab, ylab**

Ejemplo 14

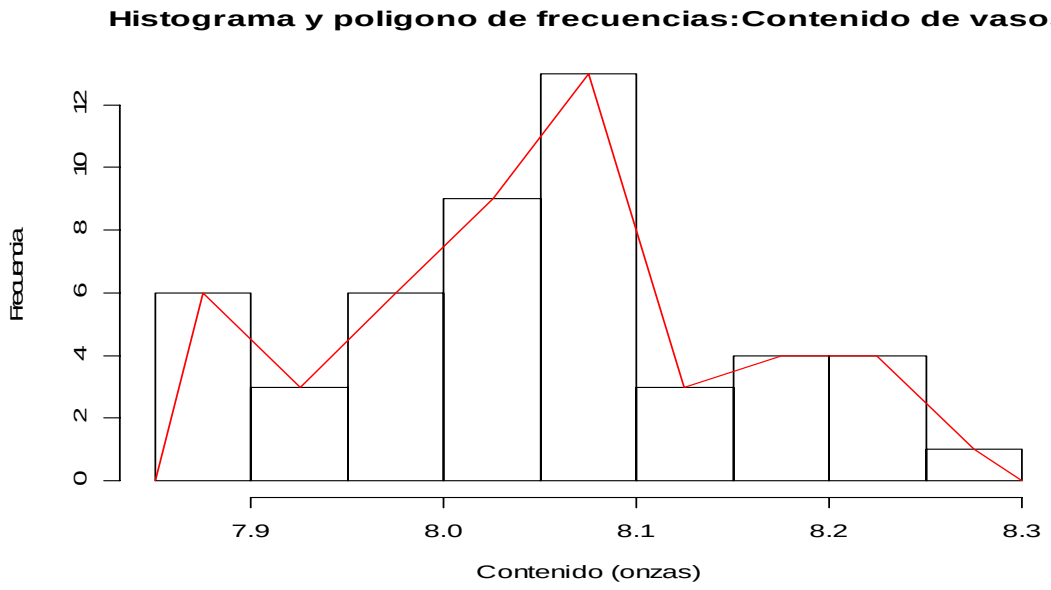
```
>hist(vasos,main="Contenido de los vasos",
      xlab="Contenido (onzas)",ylab="Frecuencia")
```



Ejemplo 15

Construyamos un histograma junto con el poligono de frecuencias

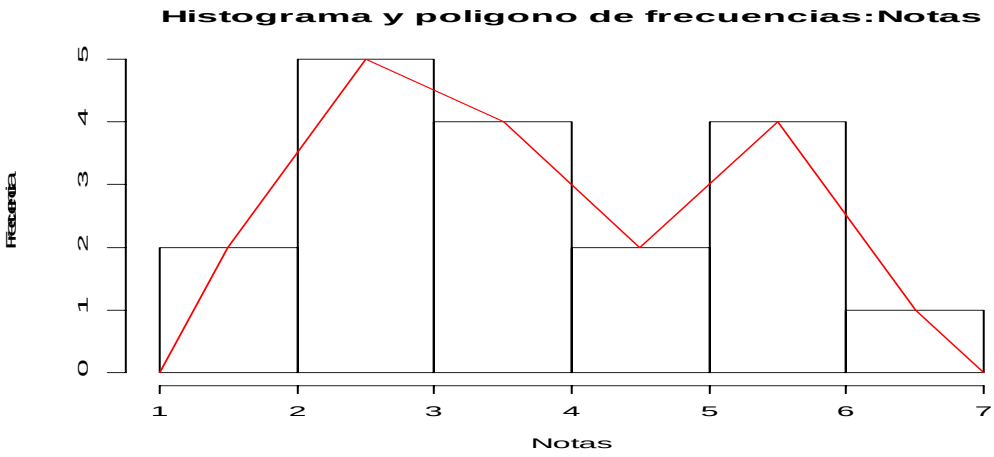
```
> vasos<-c(7.85,7.86,7.87,7.87,7.88,7.89,7.92,7.94,7.95,7.96,7.97,
+ 7.97,7.98,7.99,7.99,8.01,8.03,8.03,8.04,8.05,8.05,8.05,8.05,
+ 8.05,8.06,8.06,8.06,8.07,8.07,8.07,8.08,8.09,8.09,8.09,8.10,8.10,
+ 8.10,8.11,8.11,8.12,8.16,8.16,8.17,8.19,8.21,8.21,8.22,8.24,8.26)
> A<-hist(vasos,main="Histograma y poligono de frecuencias:Contenido de
vasos", xlab="Contenido (onzas)",ylab="Frecuencia")
>lines(c(min(A$breaks),A$mids,max(A$breaks)),c(0,A$counts,0),tipe="l",col
="red"
```



Ejemplo 16

Ahora con otros datos

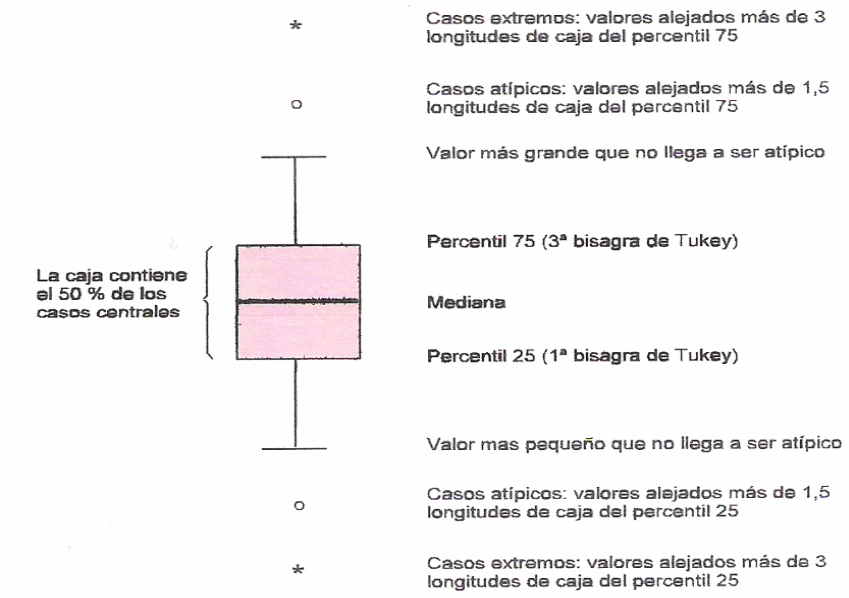
```
> Notas<-c(1.2,2.5,3,4,3,4,3,4,4.5,6,7,4,3,2,6,5.2,5.9,5)
> A<-hist(Notas,main="Histograma y poligono de frecuencias:Notas",
xlab="Notas",ylab="Frecuencia")
>lines(c(min(A$breaks),A$mids,max(A$breaks)),c(0,A$counts,0),tipe="l",col
="red")
```



2) Grafico de caja

El gráfico de caja (box-plot) es la forma gráfica de los cinco números (fivenum). Dicho gráfico adopta la siguiente forma

Figura 11.4. Destalles de un diagrama de caja.

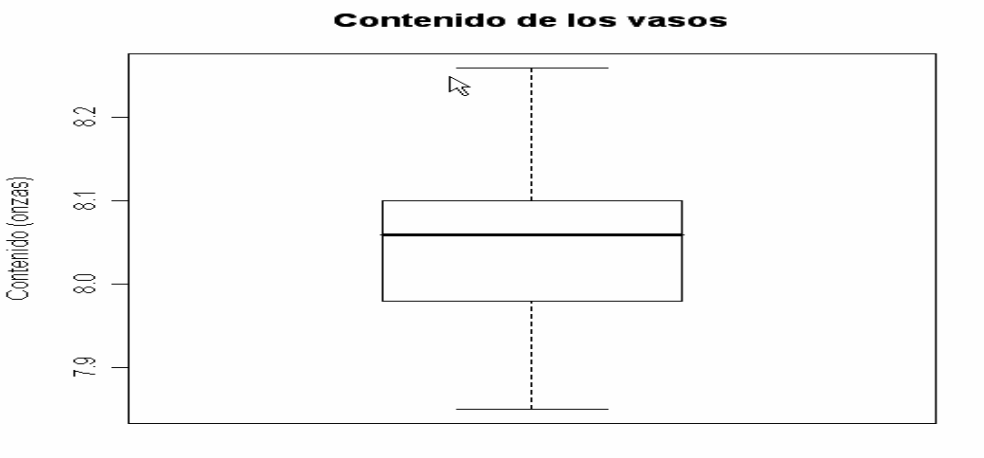


En el lenguaje R lo obtenemos de la siguiente forma

Ejemplo 17

```
> boxplot(vasos,main="Contenido de los vasos",ylab="Contenido (onzas)")
```

El gráfico que se obtiene es



3) Grafico de rama y hoja

Otra opción gráfica es el gráfico de rama y hoja (stem and leaf plot), en lenguaje R es

Ejemplo 18

```
> stem(vasos)
```

Nos entrega

```
  The decimal point is 1 digit(s) to the left of the |

  78 | 567789
  79 | 24
  79 | 5677899
  80 | 1334
  80 | 555556667778999
  81 | 000112
  81 | 6679
  82 | 1124
  82 | 6
```

Los números que se muestran a la izquierda del carácter | son los dígitos mas significativos, y como advierte la leyenda anterior al gráfico el punto decimal está ubicado un dígito a la derecha del carácter | , en otras palabras la primera línea 78 | 567789 se lee como el primer valor 7,85 (por el 78 | 5), luego hay 7,86, un 7,87,otro 7,87, un 7,88 y finalmente un 78,9

Existen muchas otras posibilidades para el manejo de las funciones gráficas, por ejemplo el gráfico de tortas, se obtiene por **piechart(datos)**, en todo caso, en http://www.r-project.org/ hay un completísimo manual.

DATOS CATEGORICOS

Los datos con los que trabajamos pueden ser, en general, clasificados en: categóricos, discretos y continuos.
Para el caso de los datos categóricos, la variable tiene respuestas que no son numéricas y naturalmente no podemos calcular medidas de tendencia central y de dispersión; lo que corresponde es resumir los datos en tablas y en gráficos

Ejemplo19

10 personas han contraído una enfermedad y deseamos conocer el comportamiento de sobre vivencia a la enfermedad después de un determinado tratamiento; las respuestas son Si o No y los datos obtenidos son:

Si, Si, No, Si, Si, No, No, Si, Si, No

Ingresamos los datos con el comando **c()**y le asignamos el nombre x , a continuación construimos una **tabla** con el comando **table,** tenemos:

```
> x<-c("Si","Si", "No", "Si", "Si", "No", "No", "Si", "Si", "No")
> x              # presenta los valores de x
 [1] "Si" "Si" "No" "Si" "Si" "No" "No" "Si" "Si" "No"

> table(x)
x
No Si
4  6
```

La tabla obtenida nos entrega la frecuencia absoluta para cada respuesta de la variable

Gráficos de barras

El gráfico de barras muestra barras con área proporcional a su frecuencia absoluta o relativa presente en la tabla

Ejemplo 20

Supongamos que se encuesta a 25 personas en relación a su preferencia sobre 4 marcas de un determinado producto, las marcas son A, B, C, D y los resultados son:

A, A, B, B, C, D, C, A, A, B, C, D, D, A, B, C, C, D, C, D, A, A, B, C, C

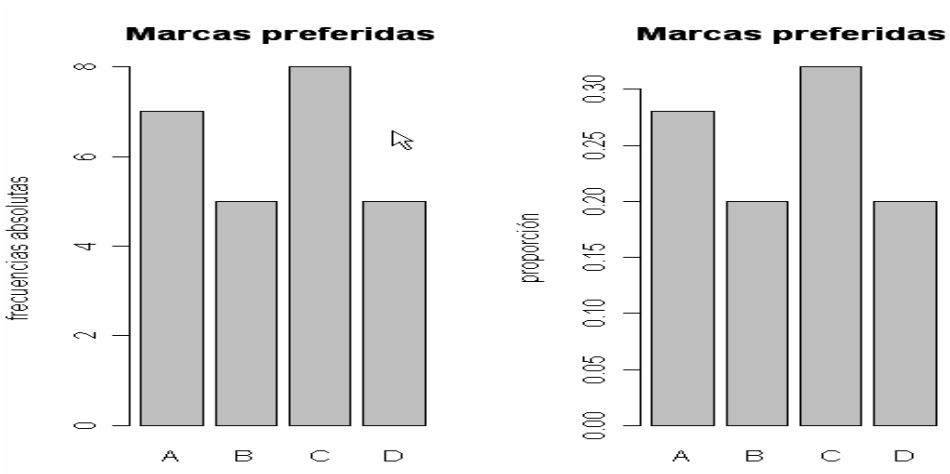
Ingresamos los datos con la función **c()**

```
 x<-
c("A","A","B","B","C","D","C","A","A","B","C","D","D","A","B","C","C",
"D","C","D","A","A","B","C","C")

> table(x)              # entrega las frecuencias absolutas
x
A B C D
7 5 8 5

> table(x)/length(x)    # entrega la proporción
x
   A    B    C    D
0.28 0.20 0.32 0.20
```

Podemos realizar un **diagrama de barras**, **`barplot`**, considerando las frecuencias absolutas y las proporciones y lo podemos hacer en un único grafico como sigue:

Ejemplo 21

```
par(mfrow=c(1,2))# dos graficos en una fila y dos columnas
barplot(table(x), main="Marcas preferidas",ylab="frecuencias absolutas")
barplot(table(x)/length(x), main="Marcas preferidas",ylab="proporción")
```



2) Diagrama de torta

Los mismos datos pueden graficarse con un diagrama de torta usando la función **`pie`**.La función es similar a **`barplot`** pero con algunos aditivos

Tablas con intervalos

Podemos construir tablas estadísticas que den cuenta del número de observaciones que caen en determinados intervalos en los que el rango de la variable se ha dividido. Para hacer esto, R usa las funciones **`cut() y table()`**

Ejemplo 22

Supongamos que el salario anual, en miles de unidades monetarias que tienen 10 trabajadores son: 12 3 60 25.5 32 28 12 18 6 52 .

Queremos usar los intervalos $(0,10]$, $(10,20]$, $(22,30]$, $(30,40]$, $(40,50]$

Para usar la función **cut** debemos especificar los puntos de corte de los intervalos, en nuestro caso, 0, 10, 20, 30, 40, 50.

```
> sal<-c(12,3,60,25.5,32,28,12,18,6,52)#entramos los datos con nombre sal
> cortes<-cut(sal,breaks=c(0,20,30,40,60))# declaramos los cortes
> table(cortes)
> cortes # vemos los valores
 (0,20] (20,30] (30,40] (40,60]
      5       2       1       2
> levels(cortes)<-c("bajo","mediano","alto","superior")# cambiamos labels
> table(cortes)
cortes
    bajo mediano alto superior
     5      2      1     2
```

Ejercicios

Para los siguientes problemas entregue un informe estadístico (no importa que sea completo, extendido, latoso) e interprete

1) Los siguientes valores corresponden a los sueldos mensuales (miles de u.m) de un grupo de 80 profesionales de una empresa

158	264	173	112	239	248	187	139	90	132
227	98	62	147	175	261	128	286	176	237
268	227	180	205	110	209	155	194	167	107
191	152	229	266	204	214	192	216	169	190
185	230	246	201	162	180	77	135	235	145
144	296	194	170	208	243	225	246	184	181
83	219	123	223	133	118	193	20	257	318
259	105	159	275	181	179	94	241	201	285

2) Una empresa metalmecánica recibió un pedido urgente del mayor número de piezas 210020 que pudiese entregar diariamente durante un periodo de 6 semanas. Los expedientes de la empresa ofrecen las siguientes entregas diarias:

22	65	65	57	55	50	65
77	73	30	62	54	48	65
79	60	63	45	51	68	79
83	33	41	49	28	56	61
65	75	55	75	39	87	45
51	67	65	59	25	35	53

3) Se realiza un estudio para ayudar a comprender el efecto de fumar en los patrones del sueño. La variable considerada es X = tiempo en minutos que se tarda en quedarse dormido. Las muestras de fumadores y no fumadores producen las siguientes observaciones sobre X

No fumadores						Fumadores					
17,2	19,7	18,1	15,1	18,3	17,6	15,1	20,5	17,7	21,3	16,0	24,8
16,2	19,9	19,8	23,6	24,9	20,1	16,9	21,2	18,1	22,1	15,9	25,2
19,8	22,6	20,0	24,1	25,0	21,4	22,8	22,4	19,4	25,2	18,3	25,0
21,2	18,9	22,1	20,6	23,3	20,2	25,8	24,1	15,0	24,1	21,6	16,3
21,1	16,9	23,0	20,1	17,5	21,3	24,3	25,7	15,2	18,0	23,8	17,9
21,8	22,1	21,1	20,5	20,4	20,7	23,2	25,1	16,1	17,2	24,9	19,9
19,5	18,8	19,2	22,4	19,3	17,4	15,8	15,3	19,9	23,3	23,0	25,1

4) Los siguientes datos representan el número de automóviles vendidos por un distribuidor durante 8 semanas de 6 días hábiles cada una.

13	19	22	14	13	16	19	21
23	11	27	25	17	17	14	20
23	17	26	20	24	15	20	21
23	17	29	17	19	14	20	20
10	23	18	25	16	23	19	20
23	18	18	24	23	20	19	26

5) Las siguientes son las notas finales de la asignatura de Probabilidades y Estadística de un curso de 45 alumnos.

5.6	3.6	5.0	5.2	5.0	5.7	2.2	5.6	1.5
4.6	3.8	4.2	4.6	3.0	6.5	6.5	2.8	2.9
4.2	1.0	3.9	5.4	5.0	3.6	3.0	5.6	3.6
6.5	5.8	5.4	3.6	3.8	6.0	3.0	2.5	5.1
1.6	3.1	5.4	5.6	6.2	3.7	4.8	3.0	4.2

6) Genere 50 números aleatorios de un anormal estándar con la función rnorm(Si a tales números los denomina A entonces la instrucción es **`A<-rnorm(n)`** donde n indica la cantidad de números) y analice

7) Genere cuatro series de números aleatorios desde una normal estándar con n=10, n= 50, n= 100, n= 200 y realice un único gráfico que contenga el histograma para cada serie. Nomínelas Histograma A B,C,D u otro nombre que usted estime conveniente usando la opción **`main="nombre"`**