UNIVERSIDAD DE PUERTO RICO
RECINTO DE RIO PIEDRAS
FACULTAD DE ADMINISTRACION DE EMPRESAS
Instituto de Estadística y Sistemas Computadorizados de Información

MANUAL DE LA ACADEMIA

Estadística Aplicada, usando R

Marzo – 2009

Preparado por:
José Carlos Vega Vilca, Ph.D.
josevega02@yahoo.con

Instituto
de Estadísticas
de Puerto Rico
Estado Libre Asociado de Puerto Rico

INTRODUCCION AL SISTEMA R

R es un lenguaje y entorno de programación para análisis estadístico y gráfico. En un inicio R fue escrito por Robert Gentleman y Ross Ihaka, conocidos como el grupo "R & R" del Departamento de Estadística de la Universidad de Auckland. Actualmente R es el resultado de un esfuerzo colaborativo con contribuciones de todo el mundo.

COMENTARIO

El New York Times publicó una nota recientemente sobre el lenguaje de programación R, destacando el hecho de que se trata de software libre, siendo los analistas de datos los más cautivados por el mismo.

Un creciente número de gente en academias y empresas ha comenzado a utilizarlo dado que el procesamiento de datos se encuentra en la edad de oro, según opina el diario neoyorquino. La operación de procesar datos es utilizada tanto para fijar precios, perfeccionar modelos financieros o encontrar nuevas medicinas, es así que R se utiliza en Pfizer, Merck, Google, el InterContinental Hotels Group, Bank of America o Shell, empresas muy diversas.

¿Y por qué R es tan utilizado? Porque científicos, ingenieros, estadísticos que no son expertos en programación pueden emplearlo rápidamente. El científico investigador de Google Daryl Pregibon expresó que es difícil no sobrevalorar a R dado lo importante que se ha tornado: les permite hacer análisis muy complejos a los estadísticos sin que conozcan en profundidad los sistemas de computación.

The New York Times subraya que grandes empresas como Dell, Hewlett-Packard o IBM hacen mucho dinero al año con la venta de servidores ejecutando GNU/Linux (la competencia libre de Microsoft o Mac OS X), de hecho la mayoría de los sitios Web se basan en el software libre Apache y cada vez hay más confianza en MySQL, la base de dados libre. Por último, el diario estadounidense destaca que los resultados finales de toda esta tecnología abierta y libre son visualizados por millones de personas mediante el navegador Firefox: una cadena libre de software.

http://www.mastermagazine.info/articulo/13495.php
Título del artículo: *R, un lenguaje de programación que seduce*

¿COMO SE INSTALA R?

Google: CRAN R
The **C**omprehensive **R** **A**rchive **N**etwork
Windows
base
Download R 2.8.1 for Windows (34 megabytes)
Run

R, ES LA MEJOR CALCULADORA

Operación aritmética	Solución en R
$3 + 5$	`> 3+5` `[1] 8`
$\frac{3}{4} + \frac{5}{7}$	`> 3/4 + 5/7` `[1] 1.464286`
$2 \times (5 + 7 \times 4)^2$	`> 2*(5+7*4)^2` `[1] 2178`
$1 + 3.3 \times log_{10}35$	`> 1+3.3*log10(35)` `[1] 6.095425`
$\frac{12 - 10}{5/\sqrt{80}}$	`> (12-10)/(5/sqrt(80))` `[1] 3.577709`
$\frac{2^8 + 3^2 - \sqrt{2}}{\sqrt{13}}$	`> (2^8+3^2-sqrt(2))/sqrt(13)` `[1] 73.10554`
$\frac{-b + \sqrt{b^2 - 4ac}}{2a}$	`> (-b+sqrt(b^2-4*a*c))/(2*a)`
$(e^3 - \sqrt[5]{28})^{log_e 41}$	`> (exp(3)-28^(1/5))^log(41)` `[1] 47193.7`

COMANDOS PARA REDONDEAR DATOS

```
> a=110/6
> a
[1] 18.33333

> b=56/3
> b
[1] 18.66667


> ceiling(a)
[1] 19
> ceiling(b)
[1] 19


> floor(a)
[1] 18
> floor(b)
[1] 18

> round(a)
[1] 18
> round(b)
[1] 19

> round(a,1)
[1] 18.3
> round(b,1)
[1] 18.7


> round(a,2)
[1] 18.33
> round(b,2)
[1] 18.67

> round(a,3)
[1] 18.333
> round(b,3)
[1] 18.667
```

COMANDOS DE IMPRESION

```
> print(a)
[1] 18.33333

> print(b)
[1] 18.66667

> print("hola")
[1] "hola"
> print("hoy es miercoles")
[1] "hoy es miercoles"



> cat(a,"\n")
18.33333

> cat(b,"\n")
18.66667

> cat("primer valor=",a,"\n")
primer valor= 18.33333

> cat("segundo valor=",b,"\n")
segundo valor= 18.66667

> cat("primero=",a,"segundo=",b,"\n")
primero= 18.33333 segundo= 18.66667

> list(primero=a,segundo=b)
$primero
[1] 18.33333
$segundo
[1] 18.66667
```

COMANDOS DE CONTROL

```
> x=3
> y=0
> if(x<5) y=4 else y=8
> y
[1] 4
```

```
> x=7
> y=0
> if(x<5) y=4 else y=8
> y
[1] 8


> x=3
> if(x<5) print("menor") else print("mayor")
[1] "menor"

> x=7
> if(x<5) print("menor") else print("mayor")
[1] "mayor"


> for(i in 1:5) print("hola")
[1] "hola"
[1] "hola"
[1] "hola"
[1] "hola"
[1] "hola"

> for(i in 1:5) print(i)
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5

> for(i in 1:5) {a=2*i+5; print(a)}
[1] 7
[1] 9
[1] 11
[1] 13
[1] 15

> n=5
> while(n<10) {print(n); n=n+1}
[1] 5
[1] 6
[1] 7
[1] 8
[1] 9
```

¿COMO HACER UNA FUNCION?

```
hola=function(x) {3*x+1}
> hola(2)
[1] 7
> hola(-4)
[1] -11

> iepr=function(x)
+ { if(x>4) print("mayor") else print("menor")}

> iepr(8)
[1] "mayor"

> iepr(2)
[1] "menor"


> es.par=function(x)
+ {if(x%%2==0) print("numero par") else print("numero impar")}

> es.par(543)
[1] "numero impar"

> es.par(82)
[1] "numero par"


> fahrenheit=function(centigrados)
+ {centigrados*9/5+32}

> fahrenheit(28)
[1] 82.4

> fahrenheit(34)
[1] 93.2

> fahrenheit(37)
[1] 98.6

> fahrenheit(38)
[1] 100.4
```

¿COMO INTRODUCIR DATOS?

```
> datos = c(34,21,29,19,22,28,19,18,38,30)
```

Cálculo del número de datos
```
> NROW(datos)
[1] 10
```

Cálculo de la media
```
> mean(datos)
[1] 25.8
```

Cálculo de la mediana
```
> median(datos)
[1] 25
```

Cálculo de la varianza
```
> var(datos)
[1] 48.84444
```

Cálculo de la desviación estándar
```
> sd(datos)
[1] 6.98888
```

Cálculo de la suma de los datos
```
> sum(datos)
[1] 258
```

Cálculo del cuadrado de cada dato
```
> datos^2
 [1] 1156  441  841  361  484  784  361  324 1444  900
```

Cálculo de la suma de los cuadrados de cada dato
```
> sum(datos^2)
[1] 7096
```

Ordenar los datos de menor a mayor
```
> sort(datos)
 [1] 18 19 19 21 22 28 29 30 34 38
```

Ordenar los datos de mayor a menor
```
> sort(datos,TRUE)
 [1] 38 34 30 29 28 22 21 19 19 18
```

¿COMO LEER DATOS DESDE MS EXCEL 2003?

Se debe instalar la librería *xlsReadWrite*, que lee archivo de datos EXCEL 2003

1) Seleccionar *Packages*

2) Seleccionar *Install package* (*s*)…

3) Seleccionar *Canada* (BC), OK

4) Seleccionar *xlsReadWrite*, OK

5) Escribir: *library(xlsReadWrite)*

6) Escribir: *dat=read.xls("c:/folder/ejemplo1")*

```
> library(xlsReadWrite)
> dat=read.xls("c:/amaquinarrp/acursos/casos-taller/ejemplo1")
> dat
> dat[,1]
> dat[,2]
> dat[,1:2]
> dat[,1:5]
> dat[1:10,]
> dat[1:10,2:4]
```

ORGANIZACION DE DATOS

La base de datos "ejemplo1" contiene datos de las siguientes variables:

1. Razón de preferencia: cualitativa
2. Gastos semanales: cuantitativa continua
3. Ingreso mensual: cuantitativa continua
4. Número de hijos: cuantitativa discreta
5. Forma de pago: cualitativa

ORGANIZACIONES DE DATOS DE LA VARIABLE: "RAZON"

Selección de datos en estudio

```
> razon=dat[,1]
```

Frecuencias absolutas ordenadas alfabeticamente

```
> fabs=table(razon)
> fabs

     Aire   Crédito Guardería    Oferta   Parking
        4         8         5         8        10
```

Ordenamiento por la mayor frecuencia absoluta

```
> fabs=sort(fabs,TRUE)
> fabs

  Parking   Crédito    Oferta Guardería      Aire
       10         8         8         5         4
```

Suma de frecuencias absolutas

```
> n=sum(fabs)
> n
[1] 35
```

Frecuencias relativas

```
> frel=(fabs/n)*100

  Parking   Crédito    Oferta Guardería      Aire
 28.57143  22.85714  22.85714  14.28571  11.42857
```

Frecuencias relativas, con dos decimales

```
> frel=round(frel,2)

  Parking   Crédito    Oferta Guardería      Aire
    28.57     22.86     22.86     14.29     11.43
```

Tabla de frecuencias de la variable “razón”

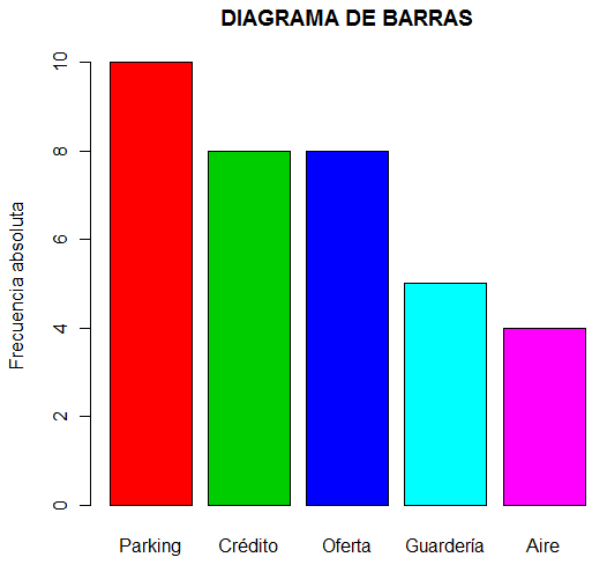
```
> cbind(fabs,frel)
          fabs  frel
Parking     10 28.57
Crédito      8 22.86
Oferta       8 22.86
Guardería    5 14.29
Aire         4 11.43
```

Gráfico circular de la variable “razón”

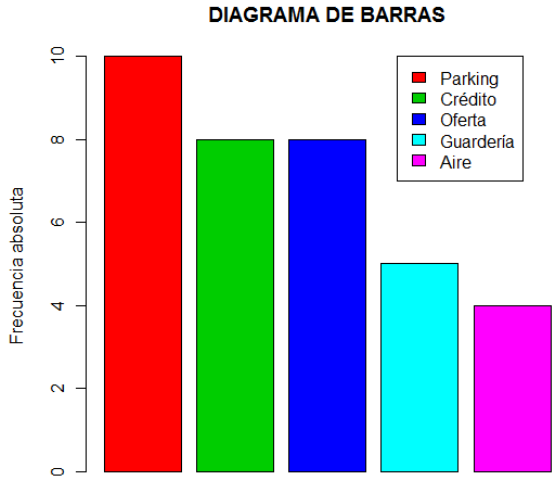
```
> pie(fabs,col=c(2,3,4,5,6),main="GRAFICO CIRCULAR")
```

Diagrama de barras de la variable “razón”

```
barplot(fabs,col=c(2,3,4,5,6),
     main="DIAGRAMA DE BARRAS",
     ylab="Frecuencia absoluta",xlab="  ")
```




```
barplot(fabs,col=c(2,3,4,5,6),names.arg=c(" "),
     main="DIAGRAMA DE BARRAS",
     ylab="Frecuencia absoluta",xlab=" ")
legend(4,10,c("Parking","Crédito",
     "Oferta","Guardería","Aire"),
       fill = c(2,3,4,5,6))
```



Funcion que hace la table de frecuencias

```
ta.frec=function(dato)
{ n=NROW(dato)
  fabs=table(dato)
  fabs=sort(fabs,TRUE)
  frel=(fabs/n)*100
  frel=round(frel,2)
  tabla=cbind(fabs,frel)
  print(tabla)
}
```

```
> ta.frec(razon)
          fabs  frel
Parking     10 28.57
Crédito      8 22.86
Oferta       8 22.86
Guardería    5 14.29
Aire         4 11.43
```

```
ta.frec=function(dato,sorteo)
{ n=NROW(dato)
  fabs=table(dato)
  if(sorteo==1) fabs=sort(fabs,TRUE)
  frel=(fabs/n)*100
  frel=round(frel,2)
  tabla=cbind(fabs,frel)
  print(tabla)
}
```

ORGANIZACIÓN DE DATOS DE LA VARIABLE "HIJOS"

```
> hijos=dat[,4]
> ta.frec(hijos,0)
  fabs  frel
0    3  8.57
1    5 14.29
2   12 34.29
3   11 31.43
4    4 11.43
```

```
fabs=table(hijos)
frel=(fabs/n)*100
```

```
barplot(frel,space=5,col="blue",ylim=c(0,40),ylab="porcentaje"
,main="Gráfico de Lineas: Número de hijos")
abline(h=0)
```

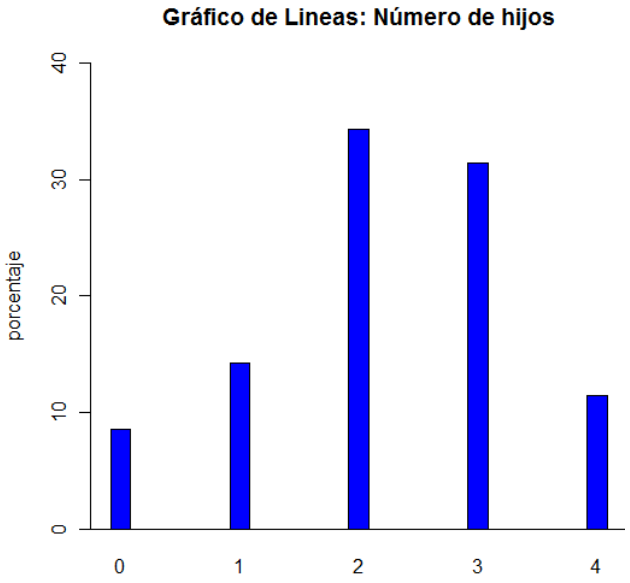


TABLA DE FRECUENCIAS DE LA VARIABLE "GASTOS"

```
gas=dat[,2]
```

Cálculo de TIC

```
tic=function(dato)
{n=NROW(dato)
 rango=max(dato)-min(dato)
 k=1+3.3*log10(n)
 k=round(k)
 tic=rango/k
 list(tic=tic,clases=k)}
```

Limites de clase

```
limites=function(dato,tic,clases)
{mini=min(dato)
 LInf=seq(mini,length=clases,by=tic)
 LSup=seq(LInf[2],length=clases,by=tic)
 marca=(LInf+LSup)/2
 Intervalo=cbind(LInf,LSup,marca)
return(Intervalo)}
```

Marca de clase: Promedio de la clase

```
lim=limites(gas,18.4,6)
marca=lim[,3]
```

Transforma los datos en clases

```
clase=function(dato)
{ n=NROW(dato)
  y=rep(0,n)
  for(i in 1:n)
  { if(dato[i]<48.4)y[i]=1 else {;
    if(dato[i]<66.8)y[i]=2 else {;
    if(dato[i]<85.2)y[i]=3 else {;
    if(dato[i]<103.6)y[i]=4 else {;
    if(dato[i]<122.0)y[i]=5 else y[i]=6}}}}}
  return(y)
}
```

Tabla de frecuencias de GASTOS

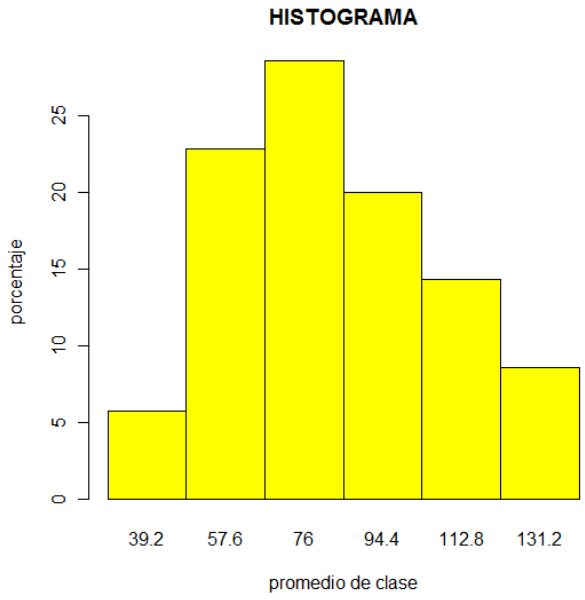
```
tablaf=function(y,lim)
{ n=NROW(y)
  fabs=table(y)
  frel=round((fabs/n)*100,2)
  facum=cumsum(fabs)
  Facum=round((facum/n)*100,2)
  tabla=cbind(lim,fabs,frel,facum,Facum)
  print(tabla)
}
```

```
> tablaf(y,lim)
   LInf  LSup marca fabs  frel facum  Facum
1  30.0  48.4  39.2    2  5.71     2   5.71
2  48.4  66.8  57.6    8 22.86    10  28.57
3  66.8  85.2  76.0   10 28.57    20  57.14
4  85.2 103.6  94.4    7 20.00    27  77.14
5 103.6 122.0 112.8    5 14.29    32  91.43
6 122.0 140.4 131.2    3  8.57    35 100.00
```

HISTOGRAMA DE FRECUENCIAS

```
gas=dat[,2]
n=NROW(gas)
y=clase(gas)
```

```
lim=limites(gas,18.4,6)
marca=lim[,3]
fabs=table(y)
frel=(fabs/n)*100

barplot(frel,space=0,names.arg=marca,
        col="yellow",xlab="promedio de clase",
        ylab="porcentaje",main="HISTOGRAMA")
```



OJIVA DE FRECUENCIAS

```
gas=dat[,2]
y=clase(gas)
lim=limites(gas,18.4,6)
LSup=lim[,2]
fabs=table(y)
Facum=cumsum(fabs)
Frcum=round((Facum/n)*100,2)

barplot(Frcum,space=0,names.arg=LSup,
     col="yellow",xlab="limites superiores",
     ylab="porcentaje acumulado",
     main="Ojiva: Frecuencias acumuladas")
```

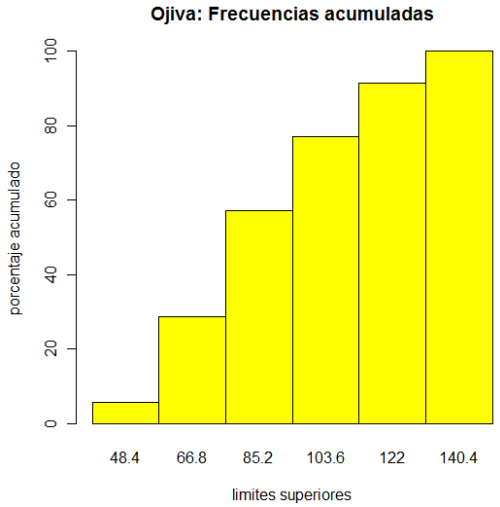
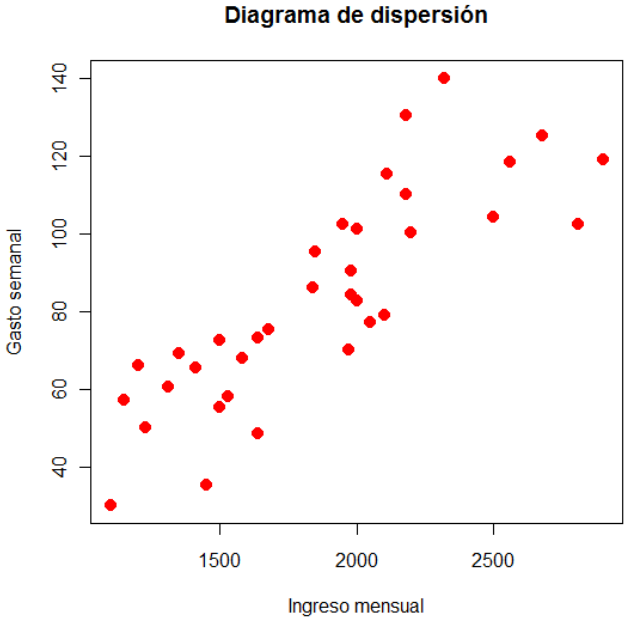


DIAGRAMA DE DISPERSION

```
gas=dat[,2]
ing=dat[,3]
plot(ing,gas)

plot(ing,gas,pch=19,col=2,cex=1.4,
     xlab="Ingreso mensual",ylab="Gasto semanal",
     main="Diagrama de dispersión")
```



TABLAS DE CONTINGENCIA

```
razon=dat[,1]
pago=dat[,5]

table(razon,pago)
table(pago,razon)
```

```
          razon
pago       Aire Crédito Guardería Oferta Parking
  Crédito     1       6         3      3       7
  Efectivo    3       2         2      5       3
```

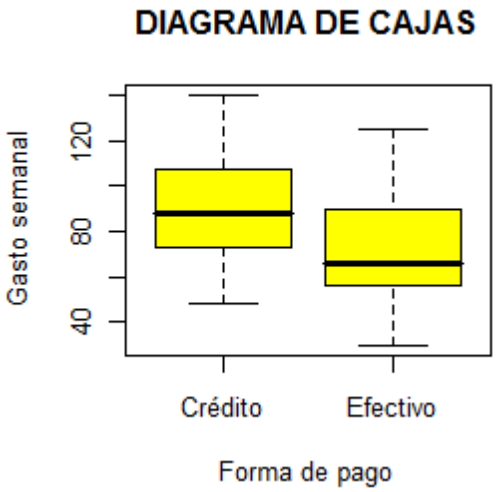
DIAGRAMA DE TALLOS Y HOJAS

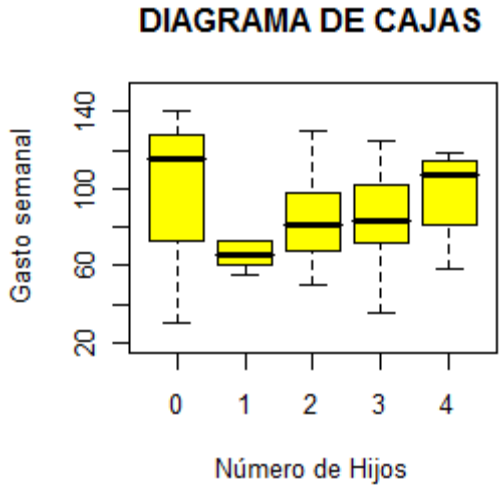
```
gas=dat[,2]
ing=dat[,3]
stem(gas)
stem(ing)
stem(gas,2)
stem(ing,2)
```

```
   3 | 05
   4 | 8
   5 | 0578
   6 | 05689
   7 | 033579
   8 | 346
   9 | 05
  10 | 01224
  11 | 0589
  12 | 5
  13 | 0
  14 | 0
```

DIAGRAMA DE CAJAS

```
razon=dat[,1]
gas=dat[,2]
ing=dat[,3]
hijos=dat[,4]
pago=dat[,5]

boxplot(gas~pago)
boxplot(gas~razon)
boxplot(gas~hijos)

boxplot(gas~pago,col="yellow",
        main="DIAGRAMA DE CAJAS",
        xlab="Forma de pago",
     ylab="Gasto semanal")
```





MEDIDAS DE TENDENCIA CENTRAL

Se continúa trabajando con la base de datos "ejemplo1". Se calcularán las medidas:

1. Media
2. Mediana
3. Moda

```
> library(xlsReadWrite)
> dat=read.xls("c:/amaquinarrp/acursos/casos-taller/ejemplo1")
> dat
```

```
> summary(dat)
       Razón        Gastos          Ingresos       Hijos            Pago
 Aire     : 4   Min.   : 30.00   Min.   :1100   Min.   :0.000   Crédito :20
 Crédito  : 8   1st Qu.: 65.65   1st Qu.:1500   1st Qu.:2.000   Efectivo:15
 Guardería: 5   Median : 79.10   Median :1950   Median :2.000
 Oferta   : 8   Mean   : 83.35   Mean   :1869   Mean   :2.229
 Parking  :10   3rd Qu.:102.20   3rd Qu.:2145   3rd Qu.:3.000
                Max.   :140.00   Max.   :2900   Max.   :4.000
```

```
razon=dat[,1]
gas=dat[,2]
ing=dat[,3]
hijo=dat[,4]
pago=dat[,5]
```

```
> summary(gas)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.00   65.65   79.10   83.35  102.20  140.00
```

```
> summary(ing)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1100    1500    1950    1869    2145    2900
```

```
> summary(razon)
     Aire   Crédito Guardería    Oferta   Parking
        4         8         5         8        10
```

Cálculo de la media y mediana

```
mean(gas)
median(gas)
```

Cálculo de la moda

```
moda=function(arreglo)
{ q=table(arreglo)
  q=sort(q,TRUE)
  return(q[1])}

> moda(razon)
Parking
     10

> moda(hijo)
 2
12
```

MEDIDAS DE POSICION

Cálculo de cuartiles

```
> quantile(gas)
    0%    25%    50%    75%   100%
 30.00  65.65  79.10 102.20 140.00

> quantile(ing)
  0%  25%  50%  75% 100%
1100 1500 1950 2145 2900
```

Cálculo de percentiles

```
> quantile(gas,0.83)
    83%
111.244
> quantile(ing,0.62)
 62%
2000
```

Medidas de tendencia central y de posición para los datos de la variable “gastos”, para clientes con pagos al crédito

```
gasc=dat[dat[,5]=="Crédito",2]

> summary(gasc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  48.40   72.95   88.15   91.30  105.70  140.00
```

Medidas de tendencia central y de posición para los datos de la variable “gastos”, para clientes con pagos en efectivo

```
gase=dat[dat[,5]=="Efectivo",2]

> summary(gase)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.00   56.25   66.00   72.75   89.75  125.10
```

Medidas de tendencia central y de posición para los datos de la variable “ingresos”, para clientes con pagos al crédito

Medidas de tendencia central y de posición para los datos de la variable “ingresos”, para clientes con pagos en efectivo

MEDIDAS DE VARIABILIDAD

Se continúa trabajando con la base de datos "ejemplo1". Se calcularán las medidas:

1. Rango o Amplitud
2. Varianza
3. Desviación estándar
4. Coeficiente de variabilidad
5. Desviación intercuartílica

Cálculo del rango

```
rango=function(arreglo)
{ max(arreglo)-min(arreglo) }
```

Cálculo de la varianza: var

Cálculo de la desviación estándar: sd

Cálculo del coeficiente de variabilidad

```
cv=function(arreglo)
{(sd(arreglo)/mean(arreglo))*100}
```

Cálculo del coeficiente intercuartílico

```
ci=function(arreglo)
{ quantile(arreglo,0.75)-quantile(arreglo,0.25)}
```

EJERCICIOS

1.- ¿Son más variables los gastos de los clientes que pagan al crédito o de los que pagan en efectivo?

2.- ¿Son más variables los ingresos de los clientes que pagan al crédito o de los que pagan en efectivo?

3.- ¿Son más variables los gastos de los clientes que prefieren nuestra tienda por el "parking" o de los que prefieren nuestra tienda por la "oferta"?

4.- ¿Son más variables los ingresos de los clientes que prefieren nuestra tienda por el "parking" o de los que prefieren nuestra tienda por la "oferta"?

5.- ¿Son más variables los gastos de los clientes que tienen 0, 1, 2, 3 ó 4 hijos?

6.- ¿Son más variables los ingresos de los clientes que tienen 0, 1, 2, 3 ó 4 hijos?

7.- Hacer un boxplot de los gastos con respecto a la razón de preferencia

8.- Hacer un boxplot de los ingresos con respecto a la razón de preferencia

PROBABILIDADES

Cálculo del factorial de un número

```
> factorial(5)
[1] 120

> for(i in 5:10){a=factorial(i);print(a)}
```

Cálculo de la combinatoria

```
> choose(5,2)
[1] 10

> choose(8,3)
[1] 56
```

Ejercicios:

1.- Calcular la probabilidad de ganar la LOTO

2.- Calcular la probabilidad de ganar el PEGA 4

3.- En una reunión de 15 personas: 10 mujeres y 5 varones, se va elegir un comité formado por 4 personas. Cuál es la probabilidad de que ese comité esté formado por 2 mujeres y 2 varones.

VARIABLE ALEATORIA

DISTRIBUCION BINOMIAL

Ejemplo:
En una agencia bancaria, el 40% de los clientes tienen certificado bancario. Si se eligen 8 clientes al azar, cuál es la probabilidad de encontrar:
a) Exactamente 6 clientes con certificados bancarios

v.a. X = # de clientes con certificado bancario; $p = 0.40$; $n = 8$

$$P(X=6) = \binom{8}{6} 0.40^6 (1-0.40)^{8-6} = 0.0413$$

```
> dbinom(6,8,0.4)
[1] 0.04128768
```

b) Todos los clientes tienen certificado bancario: P($X = 8$)

```
> dbinom(8,8,0.4)
[1] 0.00065536
```

c) Ningún cliente tenga certificado bancario: P($X = 0$)

```
> dbinom(0,8,0.4)
[1] 0.01679616
```

d) Al menos un cliente tiene certificado bancario: P($X \geq 1$)

```
> 1-dbinom(0,8,0.4)
[1] 0.9832038
```

e) A lo más 6 clientes tienen certificado bancario: P($X \leq 6$)

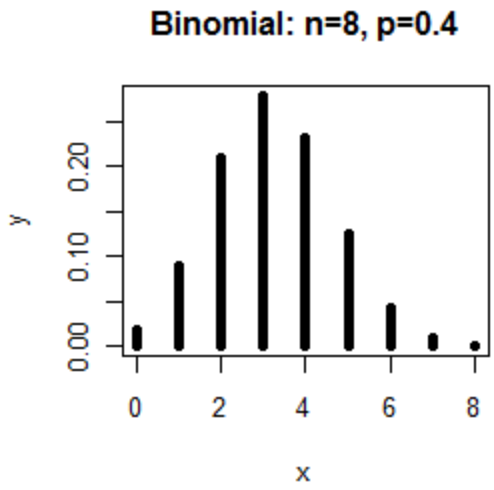
```
> pbinom(6,8,0.4)
[1] 0.9914803
```

e) Al menos cuatro clientes tienen certificado bancario: P(X ≥ 4)

```
> 1-pbinom(3,8,0.4)
[1] 0.4059136
```

f) Graficar la distribución de probabilidades de la variable aleatoria número de clientes con certificado bancario, de un total de 8 clientes. La probabilidad de éxito es 0.40.

```
x=0:8
y=dbinom(x,8,0.4)
plot(x,y,type="h",lwd=5,main="Binomial: n=8, p=0.4")
```



DISTRIBUCION DE POISSON

Ejemplo
En una inmobiliaria se ha determinado que el número promedio de casas vendidas en un día laborable es 1.6 casas/dia. Si el número de casas vendidas es una variable Poisson, calcule la probabilidad de que en un día cualquiera:

a) Se vendan exactamente 4 casas: P($X = 4$)

En este caso t =1 y λ =1.6 ➔ $\mu = \lambda t$ =1.6

$$P(X=4) = \frac{e^{-1.6}\,1.6^4}{4!} = 0.0551312$$

```
 > dpois(4,1.6)
[1] 0.05513121
```

b) No se venda ninguna casa: P($X = 0$)

```
> dpois(0,1.6)
[1] 0.2018965
```

c) Se venda por lo menos una casa: $P(X \geq 1) = 1 - P(X = 0)$

```
> 1-dpois(0,1.6)
[1] 0.7981035
```

d) Se venda entre 2 y 5 casas, inclusive: $P(2 \leq X \leq 5)$

P(X=2) + P(X=3) + P(X=4) + P(X=5)

```
> dpois(2:5,1.6)
[1] 0.25842754 0.13782802 0.05513121 0.01764199

> sum(dpois(2:5,1.6))
[1] 0.4690288
```

e) Cuál es la probabilidad de vender 4 casas en dos días?
En este caso t =2 y λ =1.6 ➔ $\mu = \lambda t = (2)(1.6) = 3.2$

$$P(X = 4) = \frac{e^{-3.2}\, 3.2^4}{4!} = 0.1780928$$

```
> dpois(4,3.2)
[1] 0.1780928
```

f) Cuál es la probabilidad de vender a lo mas 4 casas en dos días?
En este caso t =2 y λ =1.6 ➔ $\mu = \lambda t = (2)(1.6) = 3.2$

P(X≤4) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)

```
> ppois(4,3.2)
[1] 0.7806125
```

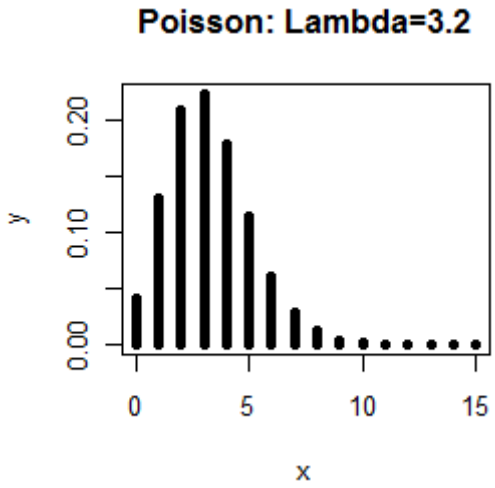
g) Cuál es la probabilidad de vender al menos 4 casas en dos días?
En este caso t =2 y λ =1.6 ➔ $\mu = \lambda t = (2)(1.6) = 3.2$

P(X≥4) = 1- P(X≤3)

```
> 1-ppois(3,3.2)
[1] 0.3974803
```

h) Graficar la distribución de probabilidades de la variable aleatoria número de casas vendidas en dos días si el promedio de ventas es 3.6 casa en dos días.

```
x=0:15
y=dpois(x,3.2)
plot(x,y,type="h",lwd=5,main="Poisson: Lambda=3.2")
```

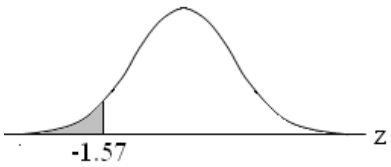


DISTRIBUCION NORMAL ESTANDAR

PROBABILIDADES EN LA DISTRIBUCION NORMAL ESTANDAR

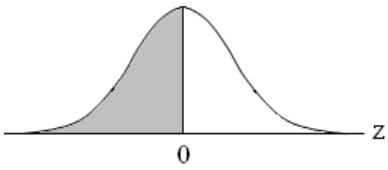
Calcular:

a) P(Z < -1.57) =



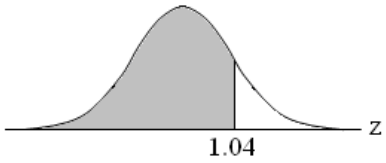

```
pnorm(-1.57)
```

b) P(Z< 0) =



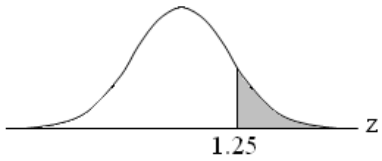

```
pnorm(0)
```

c) P(Z ≤ 1.04) =



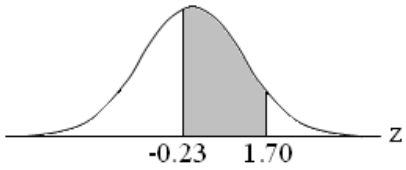

```
pnorm(1.04)
```

d) P(Z≥ 1.25) = 1 – P(Z< 1.25)



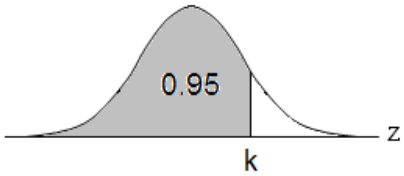

```
1-pnorm(1.25)
```

e) P(-0.23 ≤ Z≤ 1.70) =




```
pnorm(1.70)-pnorm(-0.23)
```

f) Hallar el valor "k", tal que: P(Z< k) = 0.95




```
qnorm(0.95)
```

Ejercicios:

Calcular

1) P(Z> 1.34)
2) P(Z> -2.1)
3) P(Z< -1.24)
4) P(1.1 < Z < 2.2)
5) P(-2 < Z < 1.85)
6) P(-2 < Z < -0.84)

Hallar el valor k, en los siguientes casos

1) $P(Z < k) = 0.37$

2) $P(Z < k) = 0.90$

3) $P(Z > k) = 0.44$

4) $P(0.15 < Z < k) = 0.2$

Ejemplo

En una empresa los pagos mensuales de empleados por trabajar en sobretiempo están distribuidas en forma aproximadamente normal con una media de $200 y una desviación estándar de $20, entonces la probabilidad de que un empleado, seleccionado al azar en esta empresa, tenga un pago mensual por sobretiempo

a) Mayor de 240 dólares, es

$P(X \geq 240) = 1 - P(X < 240)$

$= 1 - \text{pnorm}(240,200,20)$

$= 0.0228$

b) Entre 150 y 250 dólares, es:

$P(150 \leq X \leq 250) = P(X \leq 250) - P(X \leq 150)$

$= \text{pnorm}(250,200,20) - \text{pnorm}(150,200,20)$

$= 0.9876$

Ejercicio

1) Una supervisor ha encontrado que los trabajadores del turno noche, en promedio tardan 10 minutos en realizar una tarea. Si los tiempos requeridos para concluir la tarea están distribuidos en forma aproximadamente normal con una desviación estándar de 3 minutos, encuentre:

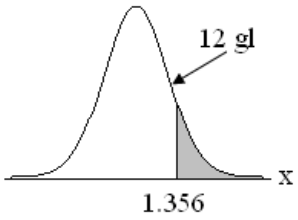
a) La proporción de trabajadores que concluyen la tarea en menos de cuatro minutos.
b) La proporción de trabajadores que requieren más de cinco minutos para concluir la tarea.
c) El supervisor ha determinado que en el turno de la noche el 33% de los trabadores son los mas lentos en completar la tarea. Hallar el tiempo mínimo necesario de un trabajador en completar la tarea para ser considerado dentro del grupo de los más lentos. Resp: 11.32 minutos

DISTRIBUCION t

Ejemplo

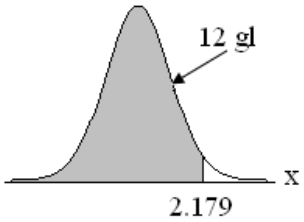
Si $X \sim t_{(12)\,gl}$, calcular:

1) P(X > 1.356) = 0.1



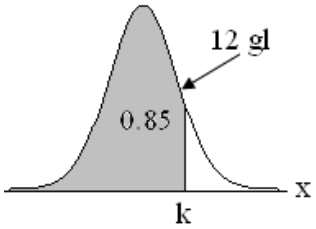

```
1-pt(1.356,12)
```

2) P(X < 2.179) = 0.975




```
pt(2.179,12)
```

3) determinar el k, tal que P(X < k) = 0.85




```
qt(0.85,12)
```

Ejercicios:
Si $X \sim t_{(18)\,gl}$

Calcular la probabilidad:

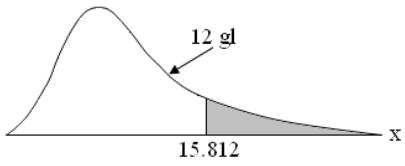
1) P(X > 1.842)
2) P(X < 1.231)
3) P(X < 0.824)
4) P(X > -1.24)
5) P(X < -2.18)
6) P(-1.23 < X < 1.23)

Hallar el valor k en los siguientes casos
7) P(-k < X < k) = 0.95

DISTRIBUCION JI-CUADRADO

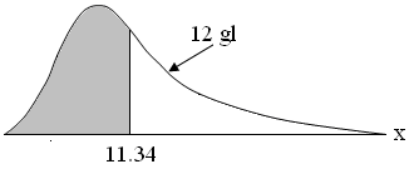
Ejemplo
Si $X \sim \chi^2_{(12)\,gl}$, calcular:

1) P(X > 15.812) = 0.199999



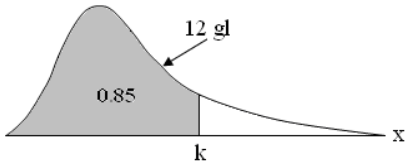

```
1-pchisq(15.812,12)
```

2) P(X < 11.34) = 0.499973




```
pchisq(11.34,12)
```

3) determinar el k, tal que P(X < k) = 0.85




```
qchisq(0.85,12)
```

Ejercicios:
Si $X \sim \chi^2_{(25)\,gl}$

Calcular la probabilidad:

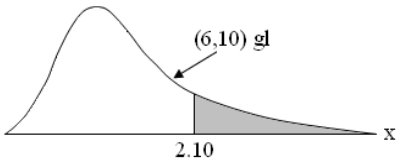
1) P(X > 18.842)
2) P(X < 5.231)
3) P(X < 17.824)
4) P(15.23 < X < 31.23)

Hallar el valor k en los siguientes casos
5) P(5.1 < X < k) = 0.95

DISTRIBUCION F DE SNEDECOR

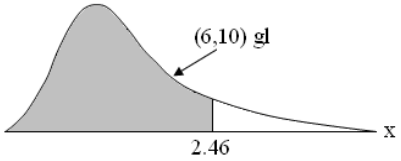
Ejemplo:
Si $X \sim F_{(6,10)\,gl}$, calcular:

1) P(X > 2.10) = 0.1433238



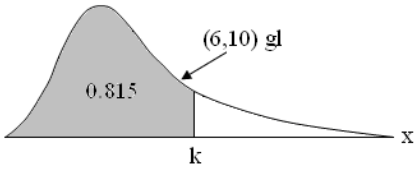

```
1-pf(2.10,6,10)
```

2) P(X < 2.46) = 0.90




```
pf(2.46,6,10)
```

3) determinar el k, tal que P(X < k) = 0.815




```
qf(0.815,6,10)
```

Ejercicios:
Si $X \sim F_{(12,27)gl}$

Calcular la probabilidad:

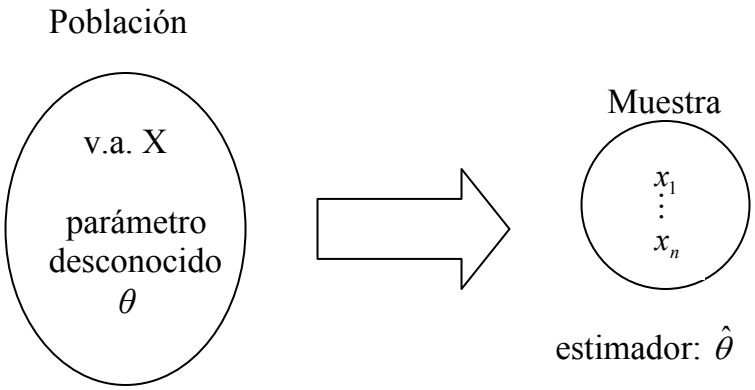
1) P(X > 1.842)
2) P(X < 0.231)
3) P(X < 1.824)
4) P(1.23 < X < 2.23)

Hallar el valor k en los siguientes casos

5) P(0.3 < X < k) = 0.95

ESTADISTICA INFERENCIAL

Se ocupa de los procedimientos que nos permiten analizar y extraer conclusiones de una población a partir de los datos de una muestra aleatoria mediante la teoría de probabilidades y de las distribuciones muestrales.



Estimador: procedimiento de cálculo con los datos muestrales con el objetivo de aproximarse al valor del parámetro.

1) Estimación de Parámetros
- Estimación puntual
- Estimación por intervalo

2) Prueba de Hipótesis

INTERVALO DE CONFIANZA PARA LA MEDIA DE UNA POBLACIÓN

a) Si la varianza σ^2 es conocida (*distribución Z*)

Intervalo de Confianza: $IC(\mu) = \bar{x} \pm Z_0 \frac{\sigma}{\sqrt{n}}$

Ejemplo

Un investigador, interesado en obtener una estimación del nivel promedio diario (μ) de óxido de sulfuro que emite una planta industrial, toma una muestra de 10 días, y calcula la media muestral $\bar{x} = 22$. Suponga que se sabe que la variable de interés presenta una distribución aproximadamente normal con una varianza de 45. Construya un intervalo de confianza del 95% para μ.

Solución:

$$\bar{x} \pm 1.96\, \sigma / \sqrt{n}$$

$$22 \pm 1.96 \sqrt{\frac{45}{10}}$$

$$(17.84, 26.16)$$

Interpretación: El intervalo (17.84, 26.16) brinda un 95% de confianza en contener el verdadero valor de μ

```
icmedia.z=function(n,media,sig2,conf)
{ sig=sqrt(sig2)
  area=(1+conf)/2
  z0=qnorm(area)
  a=media-z0*sig/sqrt(n)
  b=media+z0*sig/sqrt(n)
  print(a)
  print(b)
}

> icmedia.z(10,22,45,0.95)
[1] 17.84229
[1] 26.15771
```

b) Si la varianza σ^2 No es conocida (*distribución t*)

Intervalo de Confianza: $IC(\mu) = \bar{x} \pm t_0 \frac{S}{\sqrt{n}}$

Ejemplo

Una muestra de 30 niños de diez años de edad proporcionó un peso medio y una desviación estándar de 36.5 kg. y 5 kg, respectivamente. Suponiendo una población con distribución normal, encuentre los intervalos de confianza de 90% para la media de la población a partir de la cual se obtuvo la muestra.

Solución: coeficiente de confianza = 90%

$$\bar{x} \pm 1.699\ s/\sqrt{n}$$

$$36.5\ \pm 1.699 \times 5/\sqrt{30}$$

$$(34.94,\ 38.05)$$

```
icmedia.t=function(n,media,sd,conf)
{ area=(1+conf)/2
  t0=qt(area,n-1)
  a=media-t0*sd/sqrt(n)
  b=media+t0*sd/sqrt(n)
  print(a)
  print(b)
}

> icmedia.t(30,36.5,5,0.90)
[1] 34.94892
[1] 38.05108
```

Ejemplo

Hallar un intervalo del 95% de confianza para el promedio de los gastos semanales de todos los clientes de la megatienda VENDO.

```
> gas=dat[,2]
> t.test(gas)

data:  gas
t = 18.1734, df = 34, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 74.02809 92.66906
sample estimates:
mean of x
 83.34857
```

Intervalo de confianza para una proporción: *n* grande

En este caso, la estimación por intervalo para la proporción p de éxitos en cierta población, se obtiene mediante los límites

Intervalo de Confianza: $IC(p) = \hat{p} \pm z_0 \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$

Ejemplo

En una muestra aleatoria de 400 automóviles detenidos en un puesto de revisión, 152 de los conductores llevaban puesto el cinturón de seguridad. Construya el intervalo de confianza del 95% para la proporción real de conductores que llevan puesto el cinturón de seguridad.

Ya que $\hat{p} = \frac{152}{400} = 0.38$ ==> $IC(p) = 0.38 \pm 1.96 \sqrt{\frac{0.38\,(1-0.38)}{400}}$

$$IC(p) = (0.332,\ 0.428)$$

Ejercicio

Hacer un programa R que calcula el intervalo de confianza para el parámetro proporción.

PRUEBA DE HIPOTESIS

Es un método estadístico de comprobación de una hipótesis y es realizado utilizando los valores observados que constituyen la muestra

HIPOTESIS DE INVESTIGACION: es una suposición o reclamo que motiva una investigación. El reclamo pretende describir una característica (parámetro) de la población

HIPOTESIS ESTADISTICA: es una reformulación estadística de una hipótesis de investigación, que refiere al valor de un parámetro.
Se hace uso de dos hipótesis estadísticas complementarias:

- **hipótesis nula**: lo establecido, lo aceptado
- **hipótesis alterna**: el reto, lo nuevo

Pasos necesarios para realizar una prueba de hipótesis

1) Formulación de hipótesis

2) Establecer el nivel de significación: α
Usualmente α = 0.01, 0.02, 0.05, 0.10

3) Determinar la prueba estadística: t, Z, χ^2, F
Establecer las suposiciones de la prueba:
- La muestra fue elegida al azar
- La población de donde se extrae la muestra tiene distribución normal ó las muestras seleccionadas son suficientemente grandes

4) Determinar las regiones de aceptación y rechazo de H_0
Graficar la distribución correspondiente a la prueba elegida en el pto. 3 y representar el valor correspondiente a nivel de significación

5) Realizar el cálculo de la prueba estadística, elegida en el pto. 3

6) Establecer las conclusiones de la prueba

Definición

El *p-value*, es la probabilidad de observar un valor muestral tan extremo o más que el valor observado, si la H_0 es verdadera.

- Si el *p-value* < 0.01, existe una evidencia fuerte en contra de H_0.
- Si 0.01<*p-value*<0.05, existe evidencia moderada en contra de H_0.
- Si el p-value>0.05, existe poca o ninguna evidencia en contra de H_0.

Prueba de hipótesis acerca de la media

σ^2 es conocido $\quad \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \Rightarrow Z_{calculado} = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

σ^2 no es conocido $\quad \dfrac{\bar{x}-\mu}{S/\sqrt{n}} \sim t\ (n-1)g.l. \Rightarrow t_{calculado} = \dfrac{\bar{x}-\mu}{S/\sqrt{n}}$

Prueba de hipótesis acerca de una proporción

$$\frac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}} \sim N(0,1) \Rightarrow Z_{calculado} = \frac{\hat{p}-k}{\sqrt{\dfrac{k(1-k)}{n}}}$$

Ejercicios:

1) El fabricante de llantas radiales con cinturón de acero X-15 para camiones señala que el millaje medio que la llanta recorre antes de que se desgasten las cuerdas es de 60000 millas, con desviación estándar de 5000 millas. Una compañía compró 48 llantas y encontró que el millaje medio para sus camiones es de 59500 millas. ¿Se puede afirmar que el verdadero millaje medio de las llantas es menor de lo que afirma el fabricante?

2) Una compañía analiza una nueva técnica para armar un carro de golf; la técnica actual requiere 42.3 minutos, en promedio. El tiempo medio de montaje de una muestra aleatoria de 24 carros, con la nueva técnica, fue de 40.6 minutos y la desviación estándar de 2.7 minutos. ¿Se puede afirmar que el tiempo de montaje con la nueva técnica es más rápida?

3) Por mucho tiempo, se ha afirmado que el 60% de los jóvenes de una ciudad, son fumadores. Actualmente un investigador social dice que esta proporción ha disminuido, debido a una campaña de educación en salud. Para probar esta afirmación se hizo un estudio que consistió de una muestra aleatoria de 350 jóvenes de esa ciudad y se encontró que 210 fuman

4) Se afirma que el saldo bancario de los habitantes de una ciudad es mayor de 400 dólares. Para probar esta afirmación se seleccionó una muestra de 120 habitantes; los datos del estudio están en el archivo: ***"ejemplo2.xls"***

```
dat=read.xls("c:/CASOS-TALLER/ejemplo2")
sal=dat[,1]
t.test(sal,mu=400,a="g")
```

Prueba de hipótesis acerca de diferencia de medias: muestras independientes

Varianzas poblacionales: son conocidas ➔ Prueba Z

Se considera que los sueldos de trabajadores de la construcción en dos ciudades A y B, son variables con distribución normal, con desviaciones estándar de 4 y 6 dólares, respectivamente. ¿Se puede afirmar que el promedio de sueldos de los trabajadores de la ciudad B es mayor que el promedio de sueldos en la ciudad A?. Use los datos del archivo "***hipótesis1.xls***".

$$\frac{(\bar{x}_B - \bar{x}_A) - (\mu_B - \mu_A)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \sim N(0,1)$$

1.- Formular las hipótesis

2.- Programa R, que hace los cálculos

```
ztest=function(datoA,datoB,sigmaA,sigmaB)
{nA=NROW(datoA)
 nB=NROW(datoB)
 mediaA=mean(datoA)
 mediaB=mean(datoB)
 zcal=(mediaB-mediaA)/sqrt((sigmaA^2/nA+sigmaB^2/nB))
 pvalor=1-pnorm(zcal)
 list(Zcalculado=zcal,PVALOR=pvalor)
}
```

3.- Conclusión

Varianzas poblacionales: **no** son conocidas ➔ Prueba T

En un estudio reciente se comparó el tiempo (minutos) que pasan juntas las parejas: las parejas en que sólo trabaja uno de los cónyuges versus las parejas en que ambos trabajan. ¿Se puede concluir que en promedio las parejas en que sólo trabaja uno de los cónyuges pasan más tiempo, juntos viendo TV?. Use los datos del archivo "***hipótesis2.xls***".

```
dat=read.xls("c:/CASOS-TALLER/hipotesis2")
uno=dat[,1]
dos=dat[1:35,2]
```

1.- Formular las hipótesis de homogeneidad de varianzas

2.- Evaluación de la homogeneidad de varianzas

```
> var.test(dos,uno)

        F test to compare two variances

data:  dos and uno
F = 1.4084, num df = 34, denom df = 41, p-value = 0.2936
alternative hypothesis: true ratio of variances is not equal
to 1
```

3.- Formular las hipótesis de diferencia de medias

4.- Evaluación de la diferencia de medias

```
> t.test(uno,dos,var.equal=TRUE,a="g")

        Two Sample t-test

data:  uno and dos
t = 2.2971, df = 75, p-value = 0.01220
alternative hypothesis: true difference in means is greater
than 0
```

5.- Conclusión

Prueba de hipótesis de dos muestras: muestras dependientes

La gerencia de una cadena de mueblerías, diseño un plan de incentivos para sus agentes de ventas. Para evaluar este plan innovador, se seleccionó a 30 vendedores, al azar, y se registraron sus ingresos "antes" y "después" de aplicar el plan. ¿Se puede afirmar que hubo un aumento significativo en el ingreso semanal del vendedor?. Usar los datos del archivo "***hipótesis3.xls***".

```
dat=read.xls("c:/CASOS-TALLER/hipotesis3")
antes=dat[,2]
despues=dat[,3]
```

1.- Formular las hipótesis

2.- Evaluación de la hipótesis

```
> t.test(despues,antes,paired=TRUE,a="g")

        Paired t-test

data:  despues and antes
t = 4.1146, df = 29, p-value = 0.0001464
alternative hypothesis: true difference in means is greater
than 0
```

3.- Conclusión

Prueba de hipótesis en tablas de contingencia

Prueba de diferencia de más de dos proporciones

En un estudio se obtuvo una muestra de tres grupos de personas: se preguntó a 100 hombres, 130 mujeres y 90 niños, si les agradaba o no el sabor de una nueva pasta dental. Los resultados fueron los siguientes:

Las hipótesis son:

H_0: La proporción de "gusto por la nueva pasta dental" es la misma en los tres grupos de personas

H_1: Al menos en uno de los grupos la proporción es diferente.

Valores observados

	Hombres	Mujeres	Niños	
Les gustó el sabor	60	67	49	**176**
No les gustó el sabor	40	63	41	**144**
Total	**100**	**130**	**90**	**320**

1.- Formular las hipótesis

2.- Evaluación de la hipótesis

```
> a=matrix(c(60,40,67,63,49,41),nc=3)
> chisq.test(a)
```

3.- Conclusión

Prueba de homogeneidad de poblaciones

	Hombres	Mujeres	Niños	
Les gustó el sabor	52	56	45	**153**
Les resulta indiferente	15	23	11	**49**
No les gustó el sabor	33	51	34	**118**
Total	**100**	**130**	**90**	**320**

1.- Formular las hipótesis

2.- Evaluación de la hipótesis

3.- Conclusión

Prueba de independencia de variables

Se quiere investigar si existe en realidad una relación entre el "*desempeño en el programa de capacitación*" de la compañía y el "*éxito final en el trabajo*".

Desde una muestra de 400 empleados sacados de los grandes archivos de una compañía, se obtuvo los siguientes resultados:

Desempeño en el programa de capacitación

Éxito en el trabajo (clasificación de la empresa)	Inferior a lo normal	En el nivel normal	Superior a lo normal	Total
Deficiente	23	60	29	**112**
Normal	28	79	60	**167**
Muy bueno	9	49	63	**121**
Total	**60**	**188**	**152**	**400**

1.- Formular las hipótesis

2.- Evaluación de la hipótesis

3.- Conclusión

ANALISIS DE REGRESION y CORRELACION

El estudio de las relaciones entre dos o más variables se puede llevar a cabo desde dos puntos de vista:

Análisis de Regresión ⇨ Estudio de la relación funcional existente entre las variables

Análisis de Correlación ⇨ Estudio del grado de asociación existente entre las variables

ANALISIS DE REGRESION LINEAL

El objetivo de este análisis es estimar y analizar una ecuación o modelo, que describa la relación funcional existente entre las variables:

$$Y = f(X_1, X_2, \ldots, X_p)$$

variable dependiente

variables independientes

COEFICIENTE DE CORRELACION LINEAL

Es una medida de asociación lineal entre dos variables aleatorias. Para una muestra de divariada de n-datos: $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, el coeficiente de correlación muestral es definido por la siguiente fórmula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{SP(x,y)}{\sqrt{SC(x)\,SC(y)}}$$

Propiedades de r

1) $-1 \le r \le 1$
2) No depende de las unidades de las variables en estudio.
3) El signo de r es el mismo que b_1

Ejemplo 1
Se consideran los datos mensuales de producción y costos de operación de una empresa británica de transporte de pasajeros por carretera durante los años 1949-52
X: producción, miles de millas recorridos por los vehículos, en un mes
Y: costo de operación, en miles de dólares por mes.
Usar los datos del archivo: *"**regresion1.xls**"*

```
library(xlsReadWrite)
dat=read.xls("c:/CASOS-TALLER/regresion1")
dat=dat[1:33,1:3]
costo=dat[,2]
produ=dat[,3]
```

Gráfico de las variables "costo" y "producción"

```
plot(produ,costo,pch=19)
```

Modelo de regresión lineal

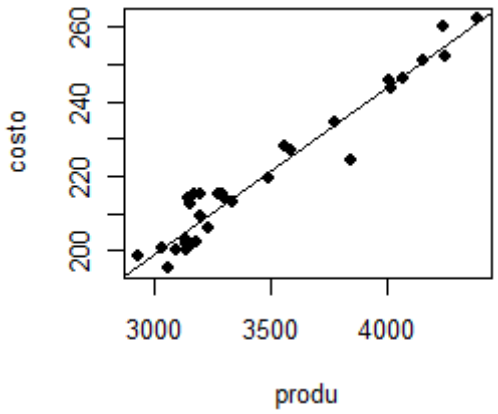
```
regre=lm(costo~produ)
```

```
> regre

Call:
lm(formula = costo ~ produ)

Coefficients:
(Intercept)        produ
   64.96328      0.04467

> summary(regre)

Call:
lm(formula = costo ~ produ)

Residuals:
      Min        1Q    Median        3Q       Max
-12.28613  -3.17076   0.06495   2.73430   8.58943

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.963277   6.635974    9.79 5.31e-11 ***
produ        0.044673   0.001909   23.40  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.626 on 31 degrees of freedom
Multiple R-squared: 0.9464,     Adjusted R-squared: 0.9447
F-statistic: 547.7 on 1 and 31 DF,  p-value: < 2.2e-16
```

Gráfico del modelo de regresión estimado

```
plot(produ,costo,pch=19)
abline(regre)
```



Pronóstico del "costo", cuando la "producción" es 3500 y 4000 miles de millas

```
> new=data.frame(produ=c(3500,4000))
> predict(regre,new)
       1        2
221.3186 243.6551
```

La línea de regresión estimada:

COSTOS = 64.963 + 0.04467 PRODUCCION

$b_0 = 64.963$ Cuando NO hay producción en un mes determinado, el costo de operación en promedio es 64,963 dólares.

$b_1 = 0.04467$ Cuando la producción se incrementa en mil millas-vehículo recorrido por mes, el costo de operación en promedio se incrementa en 44.67 dólares.

Ejemplo 2
Se consideran los datos de 69 pacientes de los que se conoce su edad y una medición de su tensión sistólica. Si estamos interesados en estudiar la variación en la tensión sistólica en función de la edad del individuo, deberemos considerar como variable respuesta la tensión y como variable predictora la edad.
X: edad
Y: tensión sistólica
Usar los datos del archivo: "**regresión2.xls**"

```
library(xlsReadWrite)
dat=read.xls("c:/CASOS-TALLER/regresion2")
dat=dat[1:69,1:3]
tens=dat[,2]
edad=dat[,3]

regre=lm(tens~edad)
plot(edad,tens,pch=19)
abline(regre)
summary(regre)
```

Ejemplo 3
En 1962 el economista norteamericano Arthur Okun planteó un modelo macroeconómico para explicar las variaciones en la tasa de desempleo. Según este modelo, que se conoce hoy en día como la "ley de Okun," existe una relación lineal entre el cambio en la tasa de desempleo y la tasa de crecimiento del Producto Interno Bruto (PIB) real. Se consideran los datos sobre desempleo y crecimiento económico en los Estados Unidos durante el período 1966-95.
Usar los datos del archivo: "**regresión3.xls**"

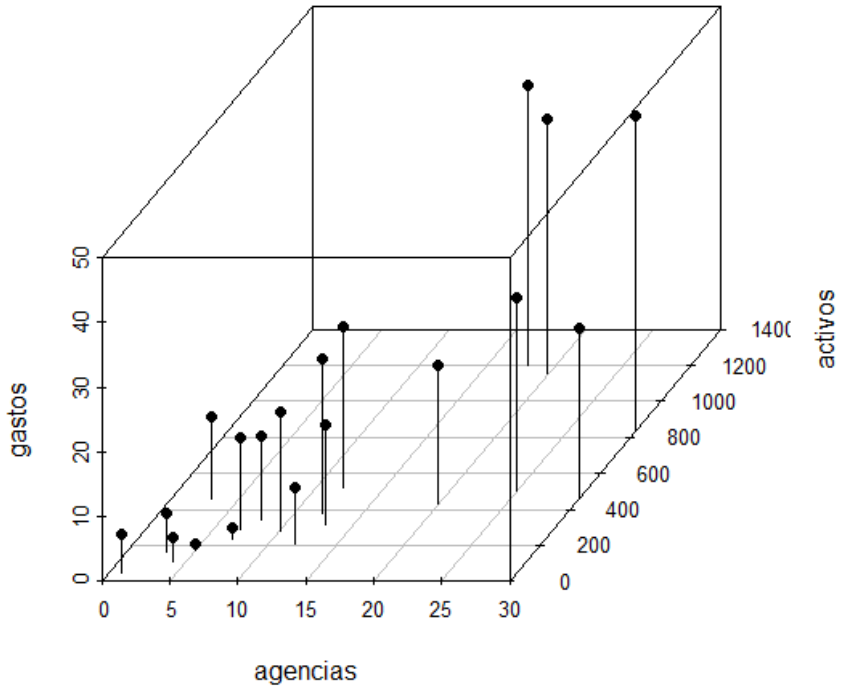
a) Use estos datos para estimar el modelo de Okun, y explique el significado de los coeficientes obtenidos.

b) En este problema, el punto donde la recta intersecta al eje X tiene un significado económico interesante. Determine este punto para este caso, y explique su significado en términos del modelo de Okun.

```
library(xlsReadWrite)
dat=read.xls("c:/CASOS-TALLER/regresion3")
des=dat[,2]
pbi=dat[,3]
regre=lm(des~pbi)
plot(pbi,des,pch=19)
abline(regre)
summary(regre)
```

Ejemplo 4
Se consideran los datos de un estudio estadístico de los costos administrativos en los bancos comerciales en Guatemala.
Y: Gastos Generales y de Administración, miles de dólares.
X1: Total de activos del banco, miles de dólares.
X2: Número de agencias del banco
Usar los datos del archivo: "**regresión4.xls**"

```
library(xlsReadWrite)
dat=read.xls("c:/CASOS-TALLER/regresion4")
gastos=dat[,2]
activos=dat[,3]
agencias=dat[,4]

regre=lm(gastos~activos+agencias)
summary(regre)

library(scatterplot3d)
sss=cbind(agencias,activos,gastos)
scatterplot3d(sss,type="h",pch=16,angle=50)
```



MUESTREO

Cuando se desea obtener información de los miembros de una población; es decir cuando se desea conocer los parámetros de una población, la primera alternativa es realizar un censo. Hay varias razones por las que a menudo se prefiere un muestreo a un censo.

VENTAJAS DEL METODO DE MUESTREO

Costo reducido.- Si los datos se obtienen únicamente de una pequeña fracción del total, los gastos son menores que los que se realizarían en un censo.

Mayor rapidez.- Los datos pueden ser recolectados y resumidos más rápidamente con una muestra que con un censo.

Mayor exactitud.- Si el volumen de trabajo es reducido se puede emplear personal capacitado al cual se le puede someter a entrenamiento intensivo

Cuidado de la población.- En estudios destructivos, conserva los elementos de la población; como por ejemplo, el estudio del tiempo de duración de baterías.

MUESTREO PROBABILISTICO

Todos los individuos tienen probabilidad conocida de ser elegidos.
Todas la posibles muestras de tamaño *n* tienen probabilidad conocida de ser elegidas.
Sólo estos métodos nos aseguran *representatividad* de la muestra.
Los tipos de muestreo probabilístico son:

1. Muestreo Aleatorio Simple
2. Muestreo Aleatorio Sistemático
3. Muestreo Aleatorio Estratificado
4. Muestreo Aleatorio por Conglomerados

MUESTREO NO PROBABILISTICO

Aplicado cuando el muestreo probabilístico resulta excesivamente costoso
Todos los individuos **no** tienen la misma probabilidad de ser elegidos.
No se tiene la certeza de que muestra extraída sea representativa
No se puede hacer generalizaciones.

SELECCIÓN ALEATORIA

Una muestra tiene *selección aleatoria* cuando el proceso de selección de unidades se hace por sorteo, ya que de esta manera todas las unidades tienen la misma probabilidad de ser seleccionadas.

Uso de función R: `sample`

Ejercicios

1.- Seleccionar aleatoriamente 5 elementos, de un total de 20

2.- Seleccionar aleatoriamente 6 elementos de un total de 46

3.- Seleccionar aleatoriamente 80 elementos de un total de 5000

MUESTREO ALEATORIO SIMPLE

Si se tiene que seleccionar una muestra de *n* elementos de una población de tamaño *N*. El muestreo aleatorio simple es aquel en el que cada muestra posible de tamaño *n* tienen la misma probabilidad de ser seleccionada.

Estimación de la media poblacional: μ

Sean $x_1, x_2, \cdots, x_n$ los valores observados de una muestra de tamaño *n*, tomada de una población de tamaño *N*.

1) Estimación puntual de la media: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

2) Estimación de la varianza de la media muestral: $var(\overline{x}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$

3) Estimación del error estándar de la media muestral: $se(\overline{x}) = \sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)}$

4) Estimación por intervalos de la media: $\overline{x} \pm z_0 \times se(\overline{x})$

Estimación del total de la poblacional: *X*

Sean $x_1, x_2, \cdots, x_n$ los valores observados de una muestra de tamaño *n*, tomada de una población de tamaño *N*.

1) Estimación puntual del total: $\hat{X} = N\,\overline{x}$

2) Estimación por intervalos del total: $N\,\overline{x} \pm z_0 \times N\,se(\overline{x})$

Estimación de la proporción poblacional: P

Sean $x_1, x_2, \cdots, x_n$ los valores observados ("1" y "0") de una muestra de tamaño n, tomada de una población de tamaño N.

1) Estimación puntual de la proporción: $\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$

2) Estimación de varianza de la proporción muestral: $var(\hat{p}) = \frac{\hat{p}\,\hat{q}}{n-1}\left(\frac{N-n}{N}\right)$

3) Estimación del error estándar de la proporción muestral: $se(\hat{p}) = \sqrt{var(\hat{p})}$

4) Estimación por intervalos de la media: $\hat{p} \pm z_0 \times se(\hat{p})$

Ejemplo1

Una empresa tiene 189 contables. En una muestra aleatoria de 50 de ellos, el número medio de horas trabajadas en sobretiempo en una semana fue de 9.7 horas con una desviación estándar de 6.2 horas. Halle un intervalo del 95% de confianza para el número medio de horas trabajadas en sobretiempo en una semana.

```
icmedia=function(n,N,media,s,conf)
{ varm=(s^2/n)*(N-n)/N
  sdm=sqrt(varm)
  area=(1+conf)/2
  z0=qnorm(area)
  a=media-z0*sdm
  b=media+z0*sdm
  cat("Linf=",a,"Lsup=",b,"\n")
}
```

Ejemplo2

Un auditor, examinando un total de 840 facturas pendientes de cobro, de una empresa, tomó una muestra aleatoria de 120 facturas. Usando los datos del archivo "**muestreo1.xls**", mediante muestreo aleatorio simple.

a) Hallar un intervalo del 95% de confianza para estimar la cantidad total de cobros pendientes

```
library(xlsReadWrite)
dat=read.xls("c:/CASOS-TALLER/muestreo1")

### muestra #####
m=sample(840,120)
datos=dat[m,]
media=mean(datos)
s=sd(datos)
```

icmedia(120,840,media,s,0.95)

b) Hallar un intervalo del 95% de confianza para estimar la proporción de facturas por cobrar con menos de 100 dólares

```
y=rep(0,120)
for(i in 1:120)
{ if(datos[i]<100) y[i]=1 }


icp=function(n,N,y,conf)
{ p=mean(y)
  q=1-p
  varp=(p*q/(n-1))*(N-n)/N
  sdp=sqrt(varp)
  area=(1+conf)/2
  z0=qnorm(area)
  a=p-z0*sdp
  b=p+z0*sdp
  cat("Linf=",a,"Lsup=",b,"\n")
}
```

MUESTREO SISTEMATICO de 1 en *k*

Si se tiene que seleccionar una muestra de n elementos de una población de tamaño N. El muestreo sistemático de 1 en k, donde $k = N/n$, se realiza de la siguiente manera:
1) El primer elemento es seleccionado aleatoriamente entre los primeros k elementos
2) Los próximos elementos son seleccionados cada k-elementos.

Ejemplo1
Desde una población de $N = 12$ hogares, se selecciona una muestra de 4 hogares para investigar acerca de la variable "número de personas que viven en el hogar"

hogares	1	2	3	4	5	6	7	8	9	10	11	12
#personas	4	3	5	6	3	4	3	4	7	5	2	1

1) Usando el muestreo aleatorio simple, seleccionar los hogares

2) Usando el muestreo sistemático de 1 en 3, seleccionar los hogares.

```
dat=c(4,3,5,6,3,4,3,4,7,5,2,1)
```

Las posibles muestras:

```
muestra1=seq(1,12,by=3)
muestra2=seq(2,12,by=3)
muestra3=seq(3,12,by=3)
```

Los datos de las posibles muestras

```
dat[muestra1]
dat[muestra2]
dat[muestra3]
```

Ejemplo2
Un auditor, examinando un total de 840 facturas pendientes de cobro, de una empresa, tomó una muestra aleatoria de 120 facturas. Usando los datos del archivo "**muestreo1.xls**", mediante muestreo sistemático de 1 en 7

1) Hallar un intervalo del 95% de confianza para estimar la cantidad total de cobros pendientes

```
library(xlsReadWrite)
dat=read.xls("c:/Users/Princess/Documents/PAPA/CASOS-
   TALLER/muestreo1")

### muestra #####
k=sample(7,1)
m=seq(k,840,by=7)
datos=dat[m,]
media=mean(datos)
s=sd(datos)

icmedia(120,840,media,s,0.95)
```

2) Hallar un intervalo del 95% de confianza para estimar la proporción de facturas por cobrar con menos de 100 dólares

```
k=sample(7,1)
m=seq(k,840,by=7)
datos=dat[m,]
y=rep(0,120)
 for(i in 1:120){if(datos[i]<100) y[i]=1}
p=mean(y)

icp(120,840,y,0.95)
```

MUESTREO ESTRATIFICADO

Si se tiene que seleccionar una muestra de n elementos de una población de tamaño N, la cual está dividida en k estratos, mutuamente excluyentes de tamaños $N_1, N_2, \ldots, N_k$, tal que:

$$N_1 + N_2 + \cdots + N_k = N$$

El muestreo estratificado consiste en seleccionar una muestra desde cada estrato de tamaños $n_1, n_2, \ldots, n_k$, tal que

$$n_1 + n_2 + \cdots + n_k = n$$

Estimación de la media poblacional: μ

Sean $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_k$ y $s_1^2, s_2^2, \cdots, s_k^2$ las medias y las varianzas muestrales desde cada estrato

1) Estimación puntual de la media: $\bar{x}_{str} = \frac{1}{N} \sum_{i=1}^{k} N_i \bar{x}_i$

2) Estimación de la varianza de la media muestral:

$$var(\bar{x}_{str}) = \frac{N_1^2\ var(\bar{x}_1) + N_2^2\ var(\bar{x}_2) + \cdots + N_k^2\ var(\bar{x}_k)}{N^2}$$

Donde: $var(\bar{x}_i) = \frac{s_i^2}{n_i}\left(\frac{N_i - n_i}{N_i}\right) \qquad i = 1, 2, \cdots, k$

3) Estimación del error estándar de la media muestral: $se(\bar{x}_{str}) = \sqrt{var(\bar{x}_{st})}$

4) Estimación por intervalos de la media: $\bar{x}_{str} \pm z_0 \times se(\bar{x}_{str})$

Estimación del total de la poblacional: X

Sean $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_k$ y $s_1^2, s_2^2, \cdots, s_k^2$ las medias y las varianzas muestrales desde cada estrato

1) Estimación puntual del total: $\hat{X} = N\,\overline{x}_{str}$

2) Estimación por intervalos del total: $N\,\overline{x}_{str} \pm z_0 \times N\,se(\overline{x}_{str})$

Estimación de la proporción poblacional: P
Sean $\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_k$ las proporciones muestrales desde cada estrato

1) Estimación puntual de la proporción: $\hat{p}_{str} = \frac{1}{N}\sum_{i=1}^{k} N_i\,\hat{p}_i$

2) Estimación de varianza de la proporción muestral:

$$var(\hat{p}_{str}) = \frac{N_1^2\,var(\hat{p}_1) + N_2^2\,var(\hat{p}_2) + \cdots + N_k^2\,var(\hat{p}_k)}{N^2}$$

Donde: $var(\hat{p}_i) = \frac{\hat{p}_i\,\hat{q}_i}{n_i - 1}\left(\frac{N_i - n_i}{N_i}\right) \quad i = 1, 2, \cdots, k$

3) Estimación del error estándar de la proporción muestral: $se(\hat{p}_{str}) = \sqrt{var(\hat{p}_{str})}$

4) Estimación por intervalos de la media: $\hat{p}_{str} \pm z_0 \times se(\hat{p}_{str})$

Ejemplo1:
Una pequeña ciudad contiene un total de 1800 hogares. La ciudad está dividida en tres distritos que contienen 820, 540 y 440 hogares, respectivamente. Una muestra aleatoria estratificada de 310 hogares contiene 120, 100 y 90 hogares, respectivamente de estos tres distritos. Se pide a los miembros de la muestra que calculen su factura total de electricidad consumida en los meses de invierno. Las respectivas medias muestrales son \$290, \$352 y \$427, y las respectivas desviaciones típicas muestrales son \$47, \$61 y \$93.

Distritos	N_i	n_i	promedio	desviación típica
1	820	120	290	47
2	540	100	352	61
3	440	90	427	93

1) Hallar un intervalo del 95% de confianza para estimar la media de la factura total de electricidad consumida en los meses de invierno.
2) Hallar un intervalo del 95% de confianza para estimar la cantidad total de electricidad consumida en los meses de invierno.

```
icmedia=function(dato,conf)
{ N=sum(dato[,1])
  m.str=crossprod(dato[,1],dato[,3])/N
  a1=(dato[,4]^2/dato[,2])*(dato[,1]-dato[,2])/dato[,1]
  a2=dato[,1]^2
  v.str=crossprod(a1,a2)/N^2
  sd.str=sqrt(v.str)
  area=(1+conf)/2
  z0=qnorm(area)
  a=m.str-z0*sd.str
  b=m.str+z0*sd.str
  cat("Linf=",a,"Lsup=",b,"\n")
}
```

Ejemplo2:
En una ciudad que tiene tres distritos se quiere conocer la proporción de hogares con alguna persona profesional. Se toman muestras aleatorias de esos hogares en cada uno de los tres distritos y se obtienen los resultados que muestra la tabla

Distritos	N_i	n_i	Hogares con Profesionales	Proporción
1	1200	180	80	0.4444
2	1350	190	50	0.2632
3	1050	140	45	0.3214

```
icprop=function(dato,conf)
{ N=sum(dato[,1])
  p=dato[,3]/dato[,2]
  q=1-p
  p.str=crossprod(dato[,1],p)/N
  a1=(p*q/(dato[,2]-1))*(dato[,1]-dato[,2])/dato[,1]
  a2=dato[,1]^2
  v.pstr=crossprod(a1,a2)/N^2
  sd.pstr=sqrt(v.pstr)
  area=(1+conf)/2
  z0=qnorm(area)
  a=p.str-z0*sd.pstr
  b=p.str+z0*sd.pstr
  cat("Linf=",a,"Lsup=",b,"\n")
}
```

Ejemplo3:
Una empresa tiene tres divisiones y los auditores están intentando estimar la cantidad total en facturas pendientes de cobro de la empresa. Hay un total de 870 facturas y en cada división hay 250, 300 y 320 facturas respectivamente. Una muestra aleatoria estratificada de 195 facturas contiene 60, 65 y 70 facturas tomadas desde las tres divisiones respectivamente. Usar los datos del archivo "**muestra2.xls**"

```
library(xlsReadWrite)
dat=read.xls("c:/CASOS-TALLER/muestreo2")

div1=dat[dat[,2]==1,1]
div2=dat[dat[,2]==2,1]
div3=dat[dat[,2]==3,1]

m1=sample(250,60)
m2=sample(300,65)
m3=sample(320,70)

dat1=div1[m1]
dat2=div2[m2]
dat3=div3[m3]

media1=mean(dat1) ; desv1=sd(dat1)
media2=mean(dat2) ; desv2=sd(dat2)
media3=mean(dat3) ; desv3=sd(dat3)
```

Completar el cuadro

Divisiones	N_i	n_i	promedio	desviación típica
1	250	60		
2	300	65		
3	320	70		

Intervalo de confianza para la media

Intervalo de confianza para el total

BIBLIOGRAFIA

Berenson, M. L., Levine, D. M., Krehbiel, T. C. (2008) "Basic Business Statistics", Eleventh Edition, Pearson Prentice Hall.

Black, K., (2008) "Business Statistics", 5th Edition, Wiley.

Cochran, W. G., (1977) "Sampling Techniques", Thirds Edition, Wiley, Ney York.

Levy P. S., Lemeshow S. (1999), "Sampling of Populations, Methods and Applications", Thirds Edition, John Wiley & Sons, Inc.

Lind, D., Marchal, W. G., Wathen, S. A. (2008) "Estadística Aplicada a los negocios y a la Economía", Decimotercera Edición, McGraw-Hill, Mexico D. F.

Newbold, P., Carlson, W., Thorne, B. (2008) "Estadística para Administración y Economía", Sexta Edición, Pearson Educación, S. A. Madrid, España.