

A large, stylized letter 'R' in a dark grey color, positioned to the left of the main title. It has a 3D effect with blue and grey rectangular blocks forming its structure.

R

Introducción a la estadística básica y gráficos



contáctenos: enei@inei.gob.pe / 433-3127



Pasaje Hernán Velarde 285 Lima.

Entre la cuadra 01 y 02 de la Av. Arequipa.

Correo: enei@inei.gob.pe / campusvirtual@inei.gob.pe

Teléfonos: 433-3127 - 332-4650

Centro Andino de Formación y Capacitación en Estadística

Cursos Especializados en Estadística e Informática

Estadísticos de resumen (1)

Fácilmente se pueden calcular estadísticos sumario tipo media, mediana, desviación, ...

```
> x<-rnorm(50)
> mean(x)
[1] -0.2552258
> sd(x)
[1] 1.209657
> var(x)
[1] 1.463269
> median(x)
[1] -0.3365646

#cuantiles empíricos
> quantile(x)
      0%      25%      50%      75%     100%
-3.4542028 -1.1195259 -0.3365646  0.6758368  2.0094436
> pvec<-seq(0,1,0.1)
> pvec
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
> quantile(x,pvec)
      0%      10%      20%      30%      40%      50%      60%      70%      80%      90%     100%
-3.4542028 -1.7757077 -1.1735237 -0.9867830 -0.6923767 -0.3365646  0.0907579  0.5180732  0.9876644  1.3546966  2.0094436
```

Estadísticos de resumen (2)

```
#exploramos el dataset juul
> library(ISwR)
> data(juul)
> ?juul
> attach(juul)
> mean(igf1)
[1] NA
```

Debemos indicarle que no tenga en cuenta los valores missing:

```
> mean(igf1,na.rm=T)
[1] 340.168
> sd(igf1,na.rm=T)
[1] 171.0356
```

```
#una excepción: la función length
> sum(!is.na(igf1))
[1] 1018
```

```
#directamente, función summary() sobre cualquier dataset
> summary(juul)
```

age	menarche	sex	igf1	tanner	testvol
Min. : 0.170	Min. : 1.000	Min. :1.000	Min. : 25.0	Min. : 1.000	Min. : 1.000
1st Qu.: 9.053	1st Qu.: 1.000	1st Qu.:1.000	1st Qu.:202.3	1st Qu.: 1.000	1st Qu.: 1.000
Median :12.560	Median : 1.000	Median :2.000	Median :313.5	Median : 2.000	Median : 3.000
Mean :15.095	Mean : 1.476	Mean :1.534	Mean :340.2	Mean : 2.640	Mean : 7.896
3rd Qu.:16.855	3rd Qu.: 2.000	3rd Qu.:2.000	3rd Qu.:462.8	3rd Qu.: 5.000	3rd Qu.: 15.000
Max. :83.000	Max. : 2.000	Max. :2.000	Max. :915.0	Max. : 5.000	Max. : 30.000
NA's : 5.000	NA's :635.000	NA's :5.000	NA's :321.0	NA's :240.000	NA's :859.000

Estadísticos de resumen (3)

```
#en el data frame tenemos variables categóricas
```

```
> detach(juul)
> juul$sex<-factor(juul$sex,labels=c("M","F"))
> juul$menarche<-factor(juul$menarche,labels=c("No","Yes"))
> juul$tanner<-factor(juul$tanner,labels=c("I","II","III","IV","V"))
> attach(juul)
> summary(juul)
```

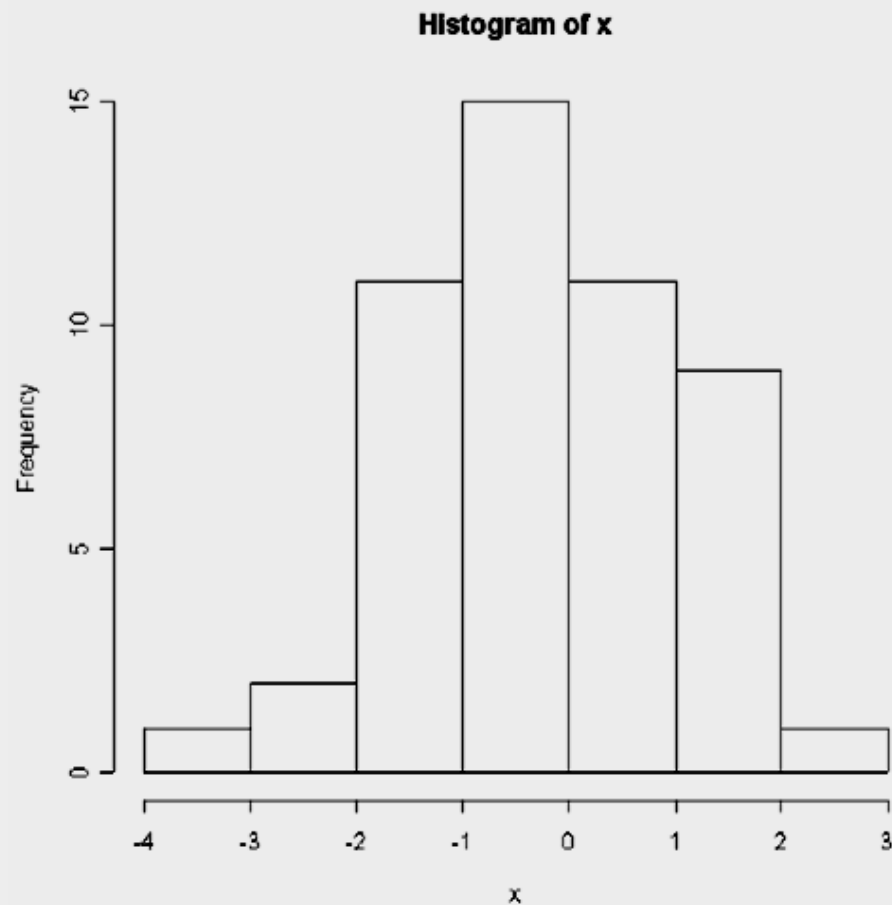
age	menarche	sex	igf1	tanner	testvol
Min. : 0.170	No :369	M :621	Min. : 25.0	I :515	Min. : 1.000
1st Qu.: 9.053	Yes :335	F :713	1st Qu.:202.3	II :103	1st Qu.: 1.000
Median :12.560	NA's:635	NA's: 5	Median :313.5	III : 72	Median : 3.000
Mean :15.095			Mean :340.2	IV : 81	Mean : 7.896
3rd Qu.:16.855			3rd Qu.:462.8	V :328	3rd Qu.: 15.000
Max. :83.000			Max. :915.0	NA's:240	Max. : 30.000
NA's : 5.000			NA's :321.0		NA's :859.000

```
#también podríamos haber utilizado la función transform()
```

```
> juul<-transform(juul,
+ sex=factor(sex,labels=c("M","F")),
+ menarche=factor(menarche,labels=c("No","Yes")),
+ tanner=factor(tanner,labels=c("I","II","III","IV","V"))) )
```

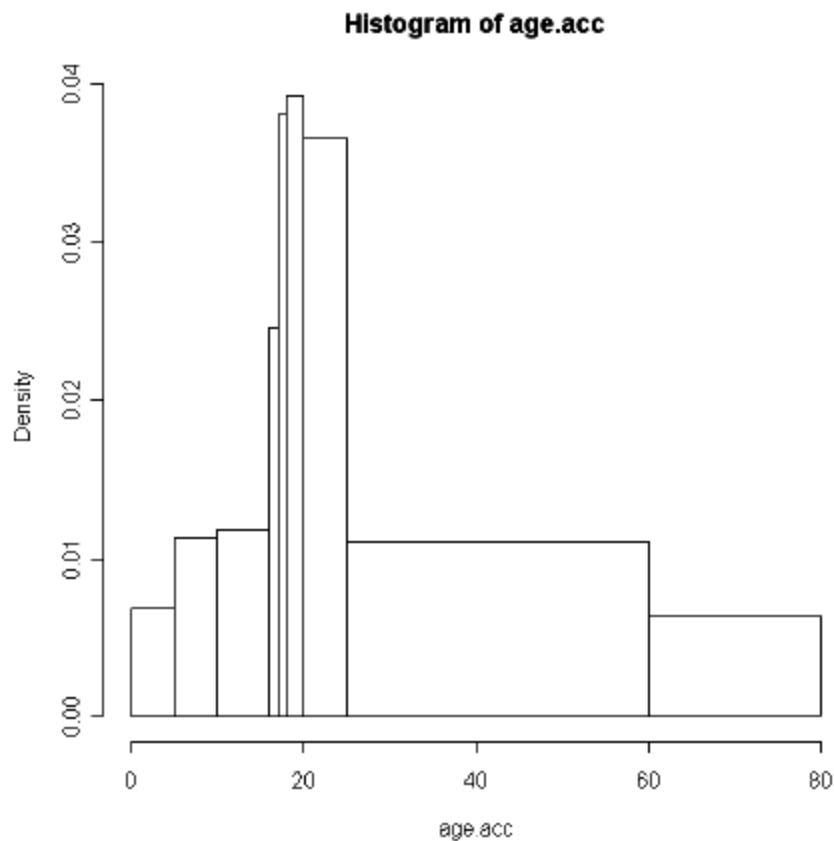
Gráficos para una variable (1)

```
#histogramas. Por defecto R, intenta hacer puntos de corte "adecuados"  
> hist(x)
```



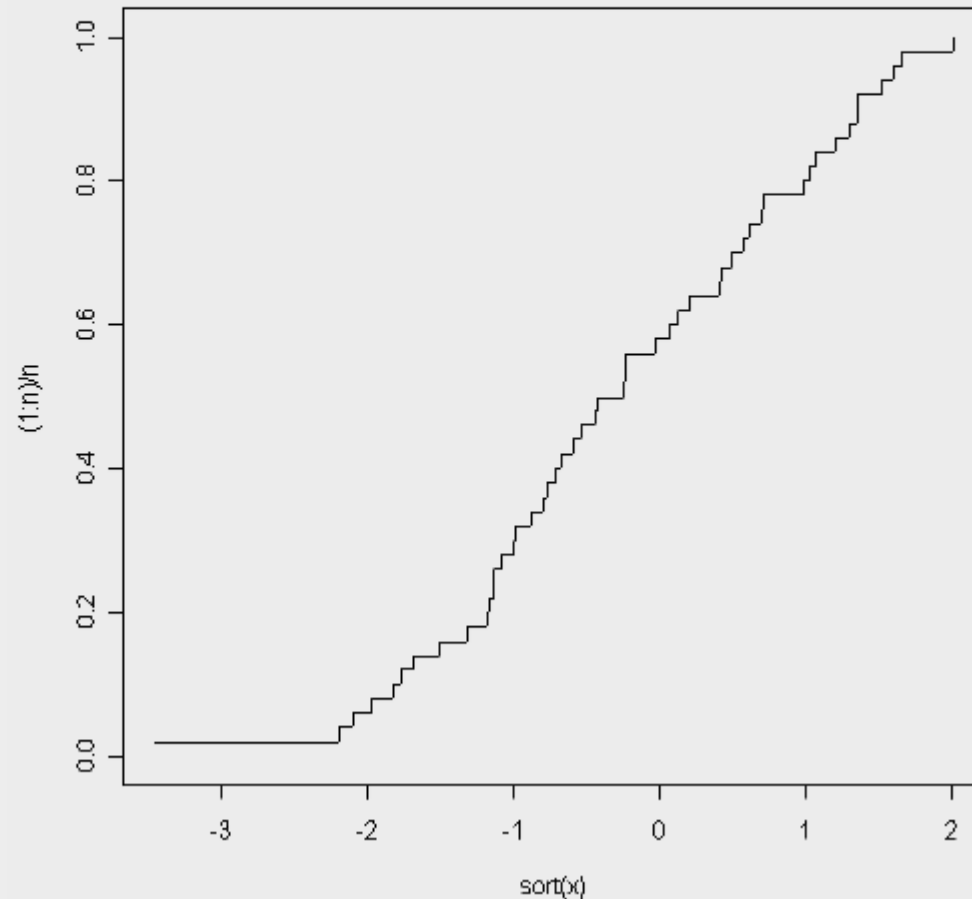
Gráficos para una variable (2)

```
#Ejemplo #accidentes vs edad (0-4,5-9,10-15,16,17,18-19,20-24,25-59,60-79)
> mid.age<-c(2.5,7.5,13,16.5,17.5,19,22.5,44.5,70.5)
> acc.count<-c(28,46,58,20,31,64,149,316,103)
> age.acc<-rep(mid.age,acc.count)
> brk<-c(0,5,10,16,17,18,20,25,60,80)
> hist(age.acc,breaks=brk)
```



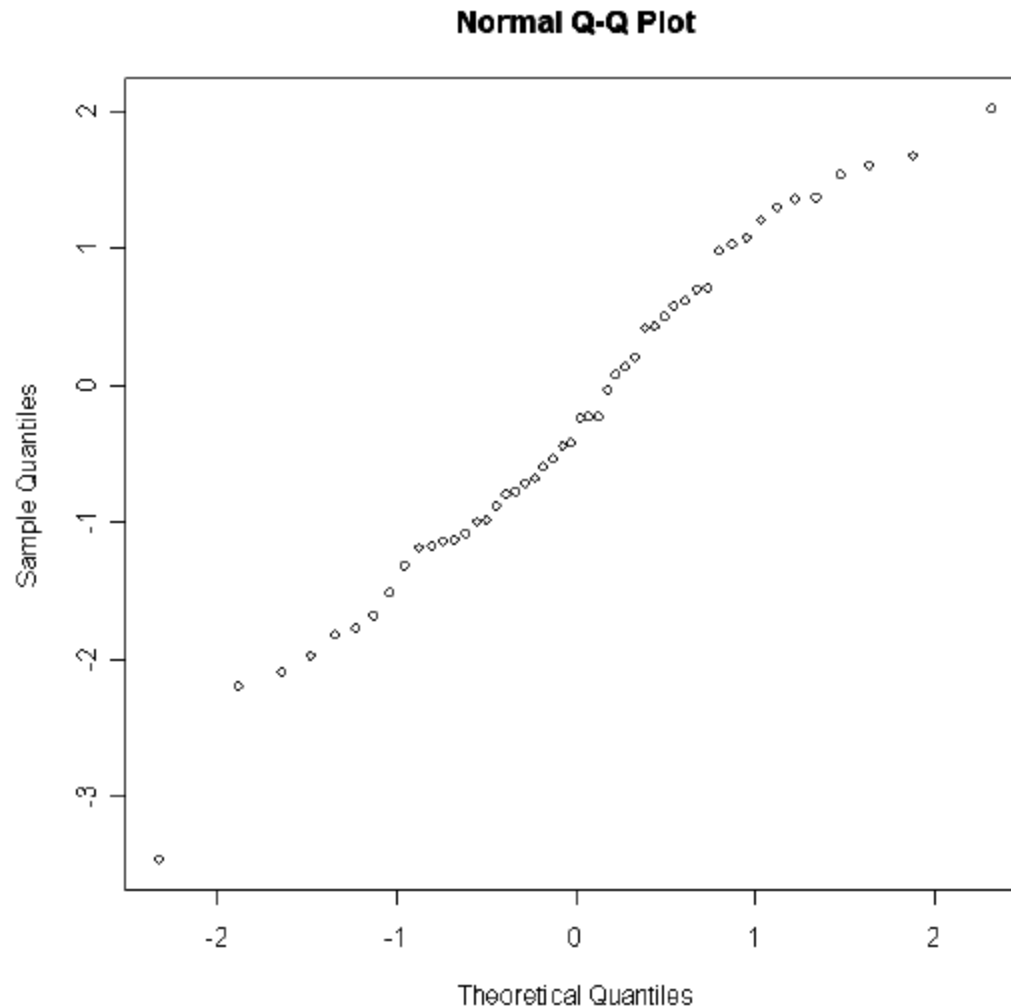
Gráficos para una variable (3)

```
#distribución empírica acumulada  
> n<-length(x)  
> plot(sort(x), (1:n)/n, type="s", ylim=c(0,1))
```



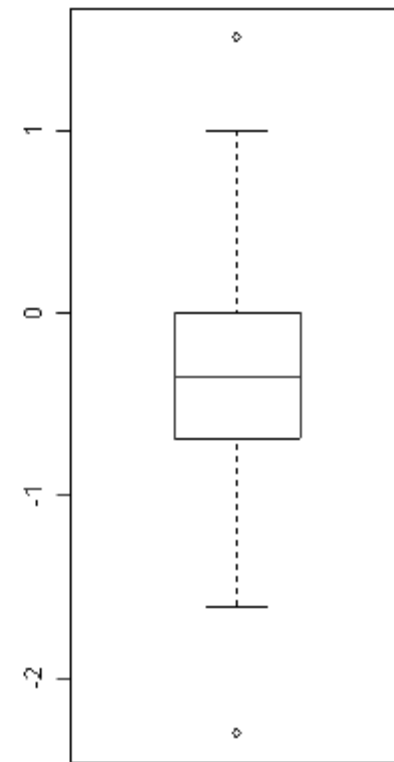
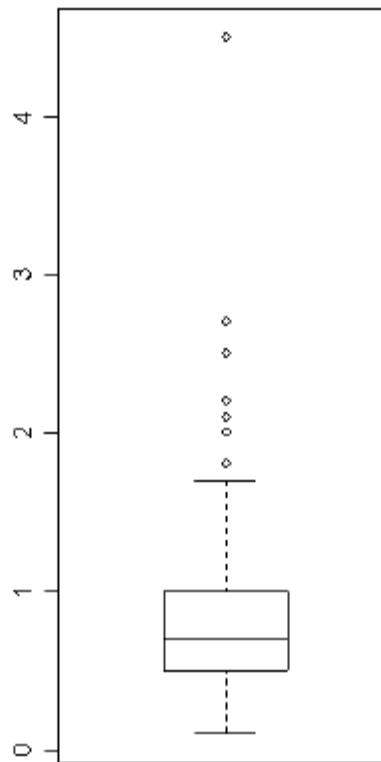
Gráficos para una variable (4)

```
#qqplot  
> qqnorm(x)
```



Gráficos para una variable (5)

```
#Boxplots IgM ( Serum IgM in 298 children aged 6 months to 6 years)
> data(IgM)
> ?IgM
> par(mfrow=c(1,2))
> boxplot(IgM)
> boxplot(log(IgM))
> par(mfrow=c(1,1))
```



Estadísticos de resumen para grupos (1)

```
#Folate concentration in blood cells according to three types of ventilation during anesthesia
```

```
> data(red.cell.folate)
> attach(red.cell.folate)
> ?red.cell.folate
> summary(red.cell.folate)
```

	folate	ventilation
Min.	:206.0	N2O+O2,24h:8
1st Qu.:	249.5	N2O+O2,op :9
Median	:274.0	O2,24h :5
Mean	:283.2	
3rd Qu.:	305.5	
Max.	:392.0	

```
> tapply(folate,ventilation,mean)
```

N2O+O2,24h	N2O+O2,op	O2,24h
316.6250	256.4444	278.0000

```
> #Para tener más de un estadístico resumen por grupo
```

```
> m<-tapply(folate,ventilation,mean)
> s<-tapply(folate,ventilation,sd)
> n<-tapply(folate,ventilation,length)
> cbind(mean=m,std.dev=s,n=n)
```

	mean	std.dev	n
N2O+O2,24h	316.6250	58.71709	8
N2O+O2,op	256.4444	37.12180	9
O2,24h	278.0000	33.75648	5

Estadísticos de resumen para grupos (2)

```
#para el dataset juul  
> tapply(igf1,tanner,mean)  
  I  II III  IV  V  
NA  NA  NA  NA  NA  
> tapply(igf1,tanner,mean,na.rm=T)  
      I      II      III      IV      V  
207.4727 352.6714 483.2222 513.0172 465.3344
```

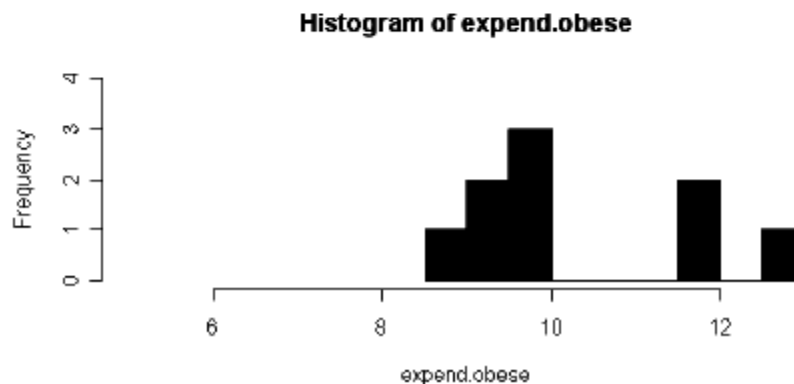
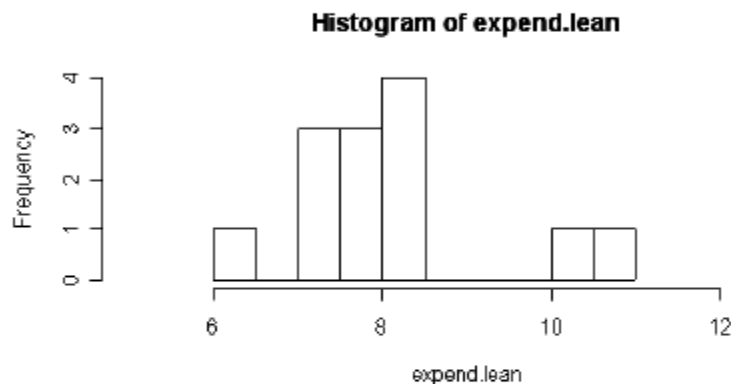

Gráficos para datos agrupados (1)

```
#cargamos el dataset energy
> data(energy)
> attach(energy)
> summary(energy)
```

expend	stature
Min. : 6.130	lean :13
1st Qu.: 7.660	obese: 9
Median : 8.595	
Mean : 8.979	
3rd Qu.: 9.900	
Max. :12.790	

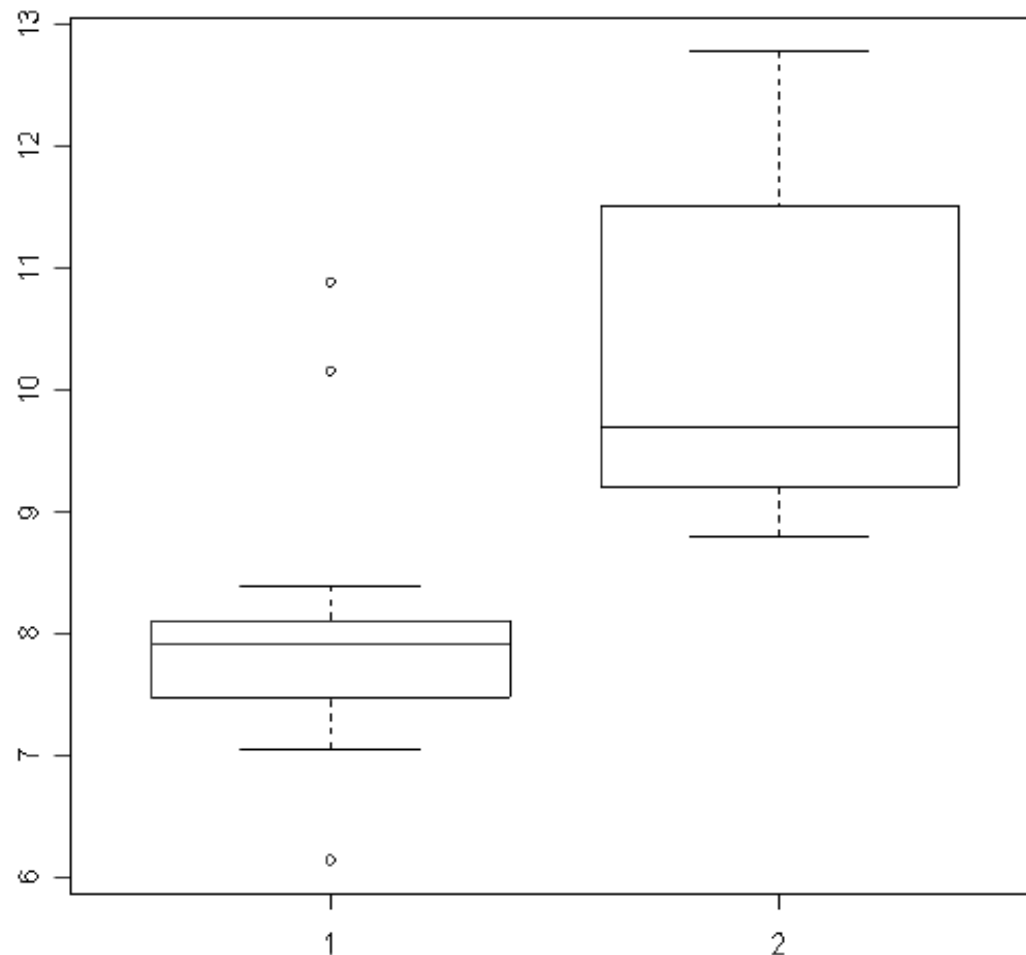
```
> ?energy
```

```
#histogramas para cada grupo de mujeres
> expend.lean<-expend[stature=="lean"]
> expend.obese<-expend[stature=="obese"]
> par(mfrow=c(2,1))
> hist(expend.lean,breaks=10,xlim=c(5,13),ylim=c(0,4),col="white")
> hist(expend.obese,breaks=10,xlim=c(5,13),ylim=c(0,4),col="black")
> par(mfrow=c(1,1))
```



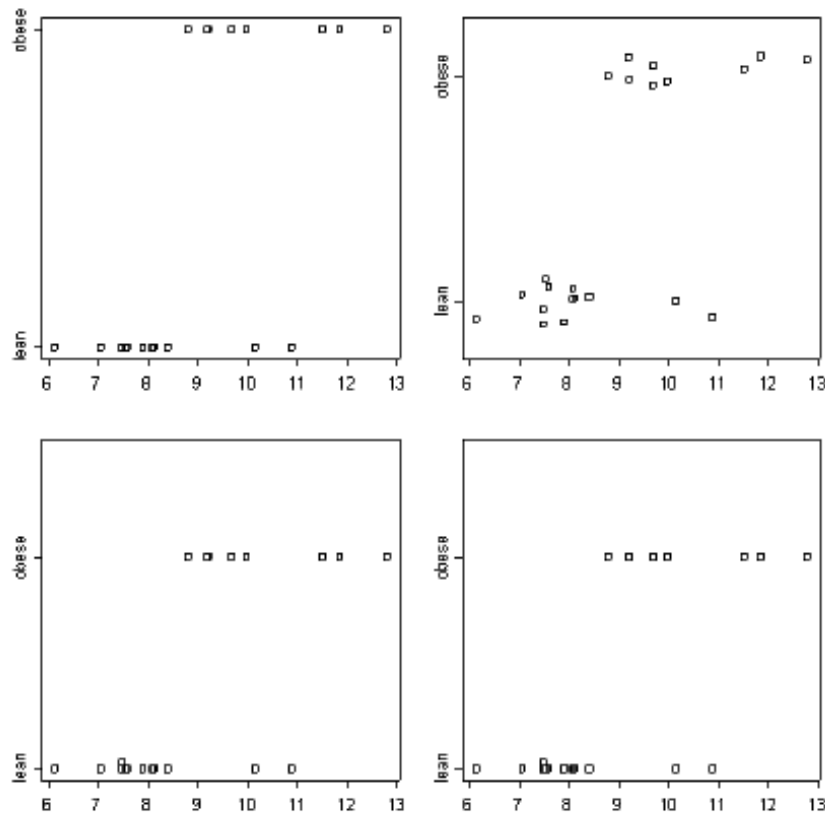
Gráficos para datos agrupados (2)

```
#boxplots para cada grupo  
> boxplot(expend~stature)  
> boxplot(expend.lean, expend.obese)
```



Gráficos para datos agrupados (3)

```
#con muestras tan pequeñas, los boxplots pueden resultar engañosos
#gráficos de los datos originales, punto a punto
> opar<-par(mfrow=c(2,2),mex=0.8,mar=c(3,3,2,1)+0.1)
> stripchart(expend~stature)
> stripchart(expend~stature,method="jitter")
> stripchart(expend~stature,method="stack")
> stripchart(expend~stature,method="stack",jitter=0.03)
> par(opar)
```



Tablas (1)

```
#Una tabla debe estar en un objeto tipo matriz
#Ejemplo mujeres consumo cafeína vs estado civil
> caff.marital<-matrix(c(652,1537,598,242,36,46,38,21,218,327,106,67),nrow=3,byrow=T)
> caff.marital
```

	[,1]	[,2]	[,3]	[,4]
[1,]	652	1537	598	242
[2,]	36	46	38	21
[3,]	218	327	106	67

```
> colnames(caff.marital)<-c("0","1-150","151-300",>300")
> rownames(caff.marital)<-c("Married","Prev.married","Single")
> caff.marital
```

	0	1-150	151-300	>300
Married	652	1537	598	242
Prev.married	36	46	38	21
Single	218	327	106	67

```
#también podemos crearla a partir de variables categóricas de un dataset
table(sex)
```

```
sex
  M   F
621 713
```

```
> table(sex,menarche)
```

```
      menarche
sex No  Yes
  M   0   0
  F 369 335
```

```
> table(menarche,tanner)
```

```
      tanner
menarche I   II  III IV  V
  No  221  43  32  14   2
  Yes   1   1   5  26 202
```


Tablas (2)

```
#podemos transponer las tablas
```

```
> t(caff.marital)
```

```
      Married Prev.married Single
0           652           36    218
1-150       1537           46    327
151-300      598           38    106
>300         242           21     67
```

```
#para calcular las frecuencias marginales, perfiles fila, .-> prop.table(tanner.sex,1)
```

```
> tanner.sex<-table(tanner,sex)
```

```
> tanner.sex
```

```
      sex
tanner M  F
  I   291 224
  II   55  48
  III  34  38
  IV   41  40
  V   124 204
```

```
> margin.table(tanner.sex,1)
```

```
tanner
  I  II III  IV  V
515 103  72  81 328
```

```
> margin.table(tanner.sex,2)
```

```
sex
  M  F
545 554
```

```
      sex
tanner M      F
  I   0.5650485 0.4349515
  II  0.5339806 0.4660194
  III 0.4722222 0.5277778
  IV  0.5061728 0.4938272
  V   0.3780488 0.6219512
```

```
> prop.table(tanner.sex,1)*100
```

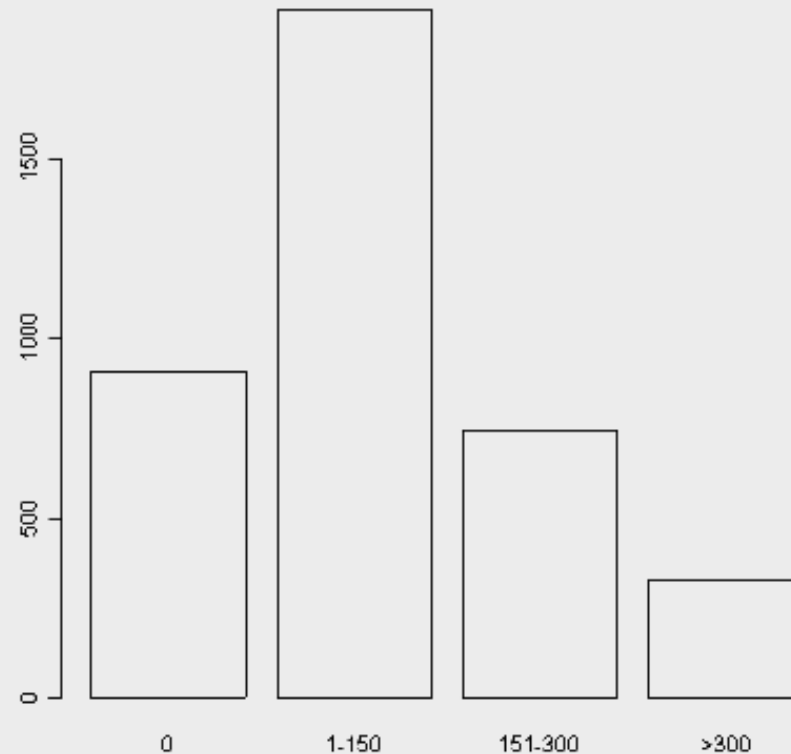
```
      sex
tanner M      F
  I   56.50485 43.49515
  II  53.39806 46.60194
  III 47.22222 52.77778
  IV  50.61728 49.38272
  V   37.80488 62.19512
```

```
> tanner.sex/sum(tanner.sex)
```

```
      sex
tanner M      F
  I   0.26478617 0.20382166
  II  0.05004550 0.04367607
  III 0.03093722 0.03457689
  IV  0.03730664 0.03639672
  V   0.11282985 0.18562329
```

Gráficos para tablas (1)

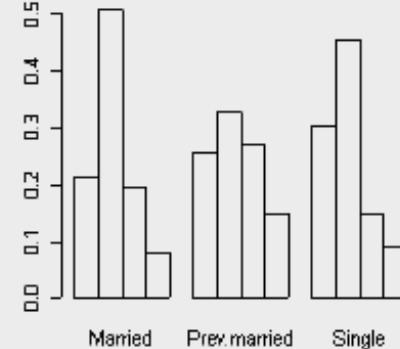
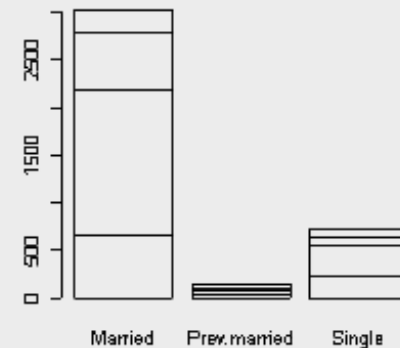
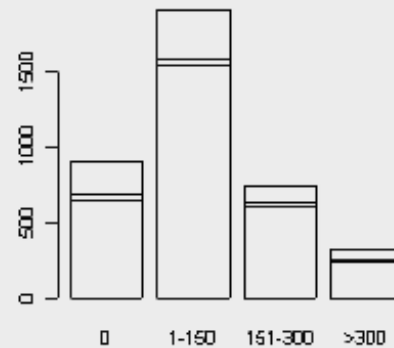
```
#diagrama de barras  
> total.caff<-margin.table(caff.marital,2)  
> total.caff  
      0    1-150 151-300    >300  
906    1910    742    330  
> barplot(total.caff,col="white")
```



Gráficos para tablas (2)

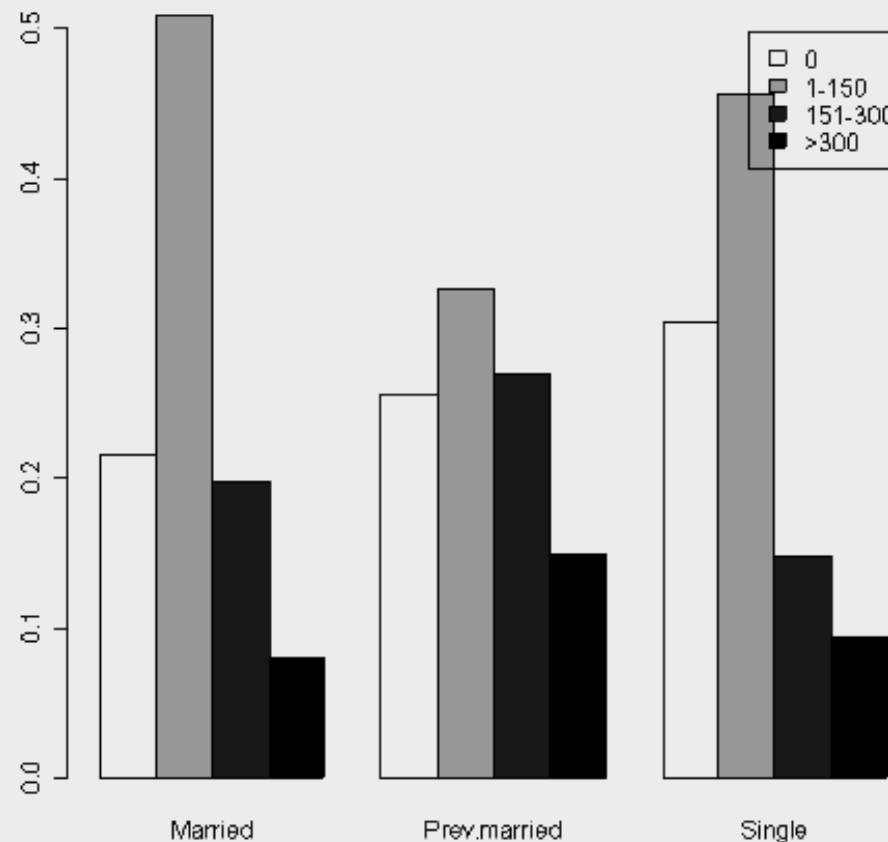
#diagramas de barras para una tabla de contingencia

```
> par(mfrow=c(2,2))  
> barplot(caff.marital,col="white")  
> barplot(t(caff.marital),col="white")  
> barplot(t(caff.marital),col="white",beside=T)  
> barplot(prop.table(t(caff.marital),2),col="white",beside=T)  
> par(mfrow=c(1,1))
```



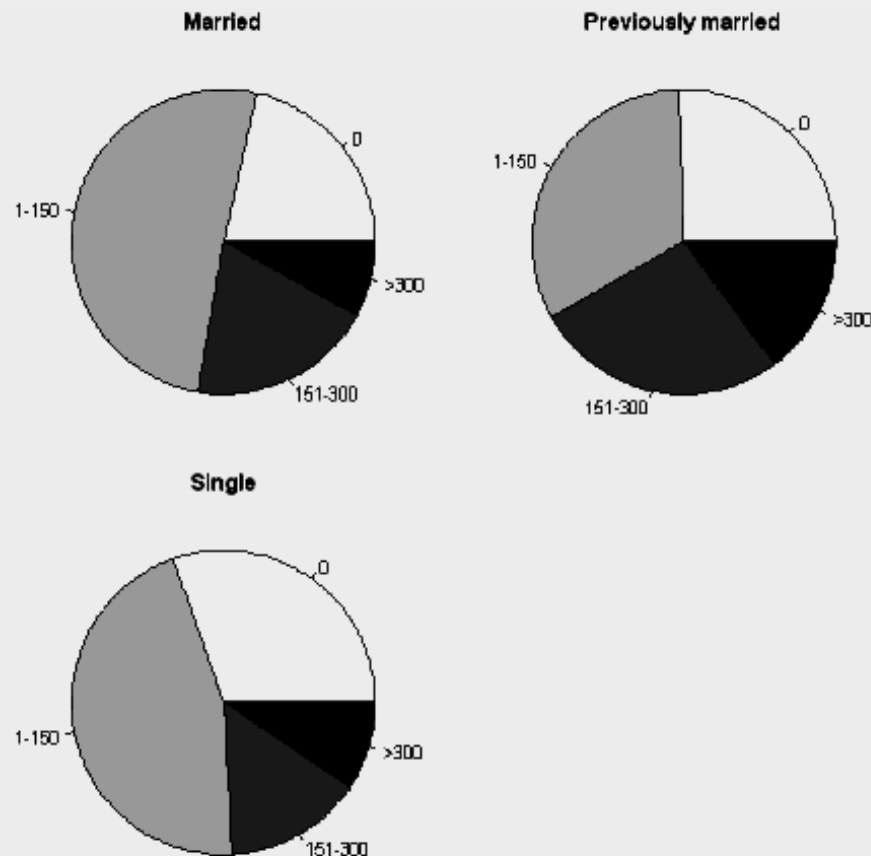
Gráficos para tablas (3)

```
#otro diagrama de barras para una tabla de contingencia  
> barplot(prop.table(t(caff.marital),2),beside=T,  
+         legend.text=colnames(caff.marital),  
+         col=c("white","grey80","grey50","black"))
```



Gráficos para tablas (4)

```
#diagrama de sectores para una tabla de contingencia  
> opar<-par(mfrow=c(2,2),mex=0.8,mar=c(1,1,2,1))  
> slices<-c("white","grey80","grey50","black")  
> pie(caff.marital["Married",],main="Married",col=slices)  
> pie(caff.marital["Prev.married",],main="Previously married",col=slices)  
> pie(caff.marital["Single",],main="Single",col=slices)  
> par(opar)
```



Comunicación constante con la Escuela del INEI

Correo de la Escuela del INEI
enei@inei.gob.pe

Área de Educación Virtual
campusvirtual@inei.gob.pe

