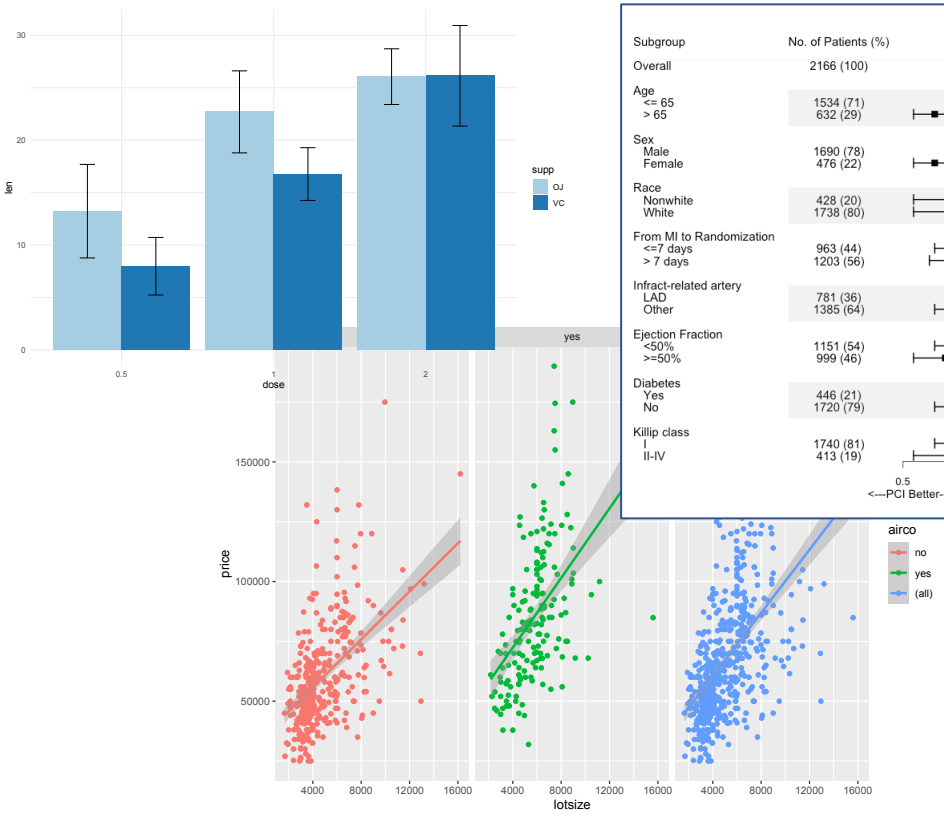
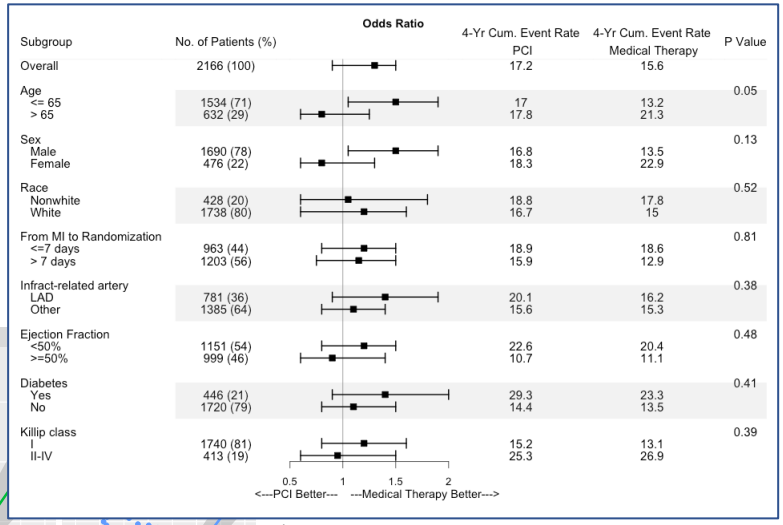
Análisis Estadístico con





Datos del curso

datos.curso1.RData

```
'data.frame':        200 obs. of  11 variables:
 $ ID             : num  137 174 200 23 39 90 40 115 72 27 ...
 $ edad           : num  37 85 29 13 49 12 85 31 39 70 ...
 $ sexo           : chr  "Mujer" "Mujer" "Hombre" "Hombre" ...
 $ estado.civil   : chr  "Casado" "Soltero" "Casado" "Divorciado" ...
 $ nivel.estudios : chr  "Bajo" "Alto" "Bajo" "Alto" ...
 $ peso           : num  59.6 60 79.2 80.8 80.8 ...
 $ altura         : num  151 149 169 171 171 ...
 $ fumador        : chr  "No" "No" "No" "Si" ...
 $ diabetes       : chr  "No" "Si" "Si" "Si" ...
 $ cancer.mama    : chr  "Si" "No" "Si" NA ...
 $ cancer.prostata: chr  NA NA "Si" "Si" ...
```

Instituto de Salud Carlos III

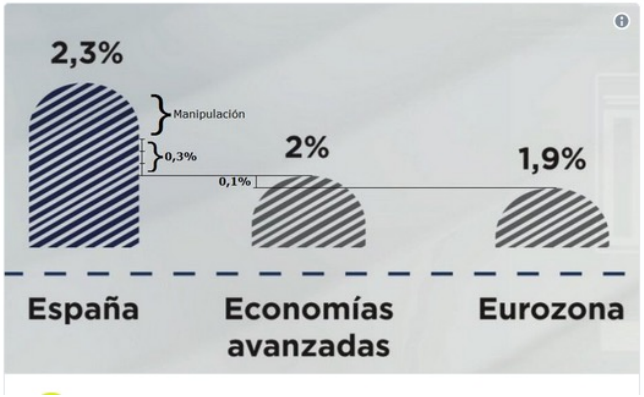
BIODAMA

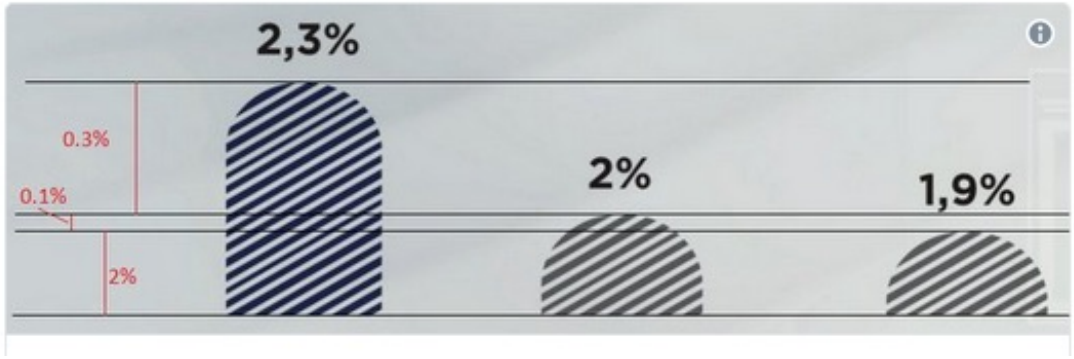
	ID	edad	sexo	estado.civil	nivel.estudios	peso	altura	fumador	diabetes	cancer.mama	cancer.prostata
1	137	37	Mujer	Casado	Bajo	59.58221	150.7163	No	No	Si	<NA>
2	174	85	Mujer	Soltero	Alto	59.95427	149.2075	No	Si	No	<NA>
3	200	29	Hombre	Casado	Bajo	79.20674	168.9795	No	Si	Si	Si
4	23	13	Hombre	Divorciado	Alto	80.78347	171.1568	Si	Si	<NA>	Si
5	39	49	Hombre	Divorciado	Bajo	80.76036	170.5682	Si	Si	<NA>	No
6	90	12	Hombre	Casado	Alto	79.83426	170.8565	No	No	<NA>	Si
7	40	85	Hombre	Casado	Alto	80.69636	168.5586	No	No	<NA>	Si
8	115	31	Mujer	Soltero	Alto	61.28985	150.0667	Si	No	Si	<NA>
9	72	39	Mujer	Divorciado	Bajo	60.70871	150.4091	Si	Si	Si	<NA>
10	27	70	Hombre	Soltero	Medio	78.89874	167.8307	No	Si	<NA>	Si
11	19	24	Mujer	Divorciado	Alto	60.06984	149.5328	No	No	Si	<NA>
12	133	45	Hombre	Soltero	Medio	79.57263	170.7013	Si	Si	<NA>	Si
13	15	42	Mujer	Soltero	Bajo	60.39387	150.7226	Si	No	Si	<NA>
14	44	74	Mujer	Casado	Medio	60.33449	148.2137	No	Si	Si	<NA>
15	179	16	Mujer	Divorciado	Medio	60.01237	151.4407	Si	Si	Si	<NA>
16	148	6	Mujer	Soltero	Bajo	57.93049	149.7306	No	No	Si	<NA>
17	192	31	Mujer	Casado	Bajo	57.99527	148.3326	Si	Si	Si	<NA>
18	186	83	Mujer	Soltero	Bajo	60.44401	150.3675	Si	Si	Si	<NA>
19	18	32	Hombre	Divorciado	Alto	79.39825	168.9841	Si	No	<NA>	No
20	106	67	Hombre	Soltero	Medio	80.07338	170.6679	No	Si	<NA>	Si
21	86	54	Mujer	Divorciado	Medio	60.02079	150.1170	Si	Si	No	<NA>
22	55	18	Hombre	Soltero	Bajo	78.30574	170.9642	No	Si	<NA>	Si
23	20	84	Hombre	Casado	Bajo	80.35391	168.2667	Si	Si	<NA>	No
24	102	40	Mujer	Soltero	Bajo	60.34935	150.3949	Si	No	No	<NA>
25	140	6	Hombre	Divorciado	Bajo	79.47181	170.0953	Si	No	<NA>	Si
26	112	63	Mujer	Soltero	Medio	60.05865	150.2747	No	Si	No	<NA>
27	183	79	Hombre	Soltero	Alto	79.69627	169.4480	Si	Si	<NA>	Si
28	120	27	Mujer	Divorciado	Alto	58.60144	150.2412	No	Si	Si	<NA>
29	117	34	Hombre	Soltero	Alto	80.76927	168.8787	No	No	<NA>	Si
30	130	53	Hombre	Divorciado	Alto	79.05485	170.1690	Si	No	<NA>	Si
31	144	23	Mujer	Divorciado	Alto	59.16448	149.9304	No	No	Si	<NA>
32	41	66	Hombre	Casado	Medio	79.35361	170.4784	Si	No	<NA>	No
33	193	29	Mujer	Casado	Alto	58.42305	149.1501	Si	No	Si	<NA>
34	12	54	Hombre	Soltero	Medio	78.61265	169.0950	Si	Si	<NA>	Si
35	108	32	Mujer	Casado	Bajo	60.56363	150.0661	No	Si	Si	<NA>
36	17	7	Mujer	Soltero	Bajo	58.65925	148.7323	Si	Si	Si	<NA>
37	157	42	Hombre	Soltero	Bajo	79.90425	169.8202	No	Si	<NA>	Si
38	122	68	Mujer	Casado	Alto	61.06287	149.0318	No	No	Si	<NA>
39	25	15	Hombre	Soltero	Alto	81.42743	169.8150	No	Si	<NA>	Si
40	50	6	Hombre	Casado	Bajo	80.68350	168.3866	Si	No	<NA>	Si
41	146	38	Mujer	Soltero	Bajo	58.74115	150.0431	No	Si	Si	<NA>
42	29	28	Hombre	Casado	Medio	80.60408	170.8330	Si	No	<NA>	Si
43	187	27	Mujer	Soltero	Alto	60.28175	149.2774	No	No	Si	<NA>
44	64	7	Mujer	Divorciado	Bajo	60.67000	151.5040	No	No	Si	<NA>
45	185	9	Hombre	Divorciado	Alto	82.13408	169.4244	No	Si	<NA>	Si
46	31	85	Mujer	Divorciado	Medio	59.54810	150.3059	No	No	Si	<NA>
47	21	24	Hombre	Casado	Alto	79.06357	170.3569	Si	No	<NA>	Si
48	78	57	Mujer	Soltero	Medio	58.93866	149.9912	No	Si	Si	<NA>
49	11	30	Hombre	Divorciado	Alto	81.73699	169.1571	Si	Si	<NA>	Si
50	87	84	Hombre	Soltero	Alto	80.20169	170.2267	Si	Si	<NA>	Si
51	116	64	Hombre	Divorciado	Medio	80.71775	167.6929	No	No	<NA>	Si
52	47	81	Mujer	Casado	Bajo	61.21061	149.0885	No	Si	Si	<NA>
53	8	38	Hombre	Casado	Alto	80.21772	170.2392	Si	Si	<NA>	Si
54	94	73	Hombre	Divorciado	Alto	79.47415	168.2748	No	Si	<NA>	Si
55	92	54	Mujer	Casado	Bajo	59.50153	150.7325	No	No	Si	<NA>
56	173	19	Hombre	Soltero	Alto	79.17450	170.0500	Si	No	<NA>	Si
57	175	77	Hombre	Soltero	Medio	79.05946	169.3775	Si	Si	<NA>	Si
58	79	29	Mujer	Soltero	Bajo	59.24916	149.8326	No	Si	Si	<NA>
59	5	81	Hombre	Soltero	Alto	80.11593	168.1224	No	No	<NA>	Si
60	36	20	Hombre	Casado	Alto	80.11153	168.8877	Si	Si	<NA>	Si
61	110	41	Mujer	Casado	Medio	60.35693	148.6183	No	Si	No	<NA>
62	196	44	Hombre	Casado	Alto	79.43454	168.3474	Si	No	<NA>	No

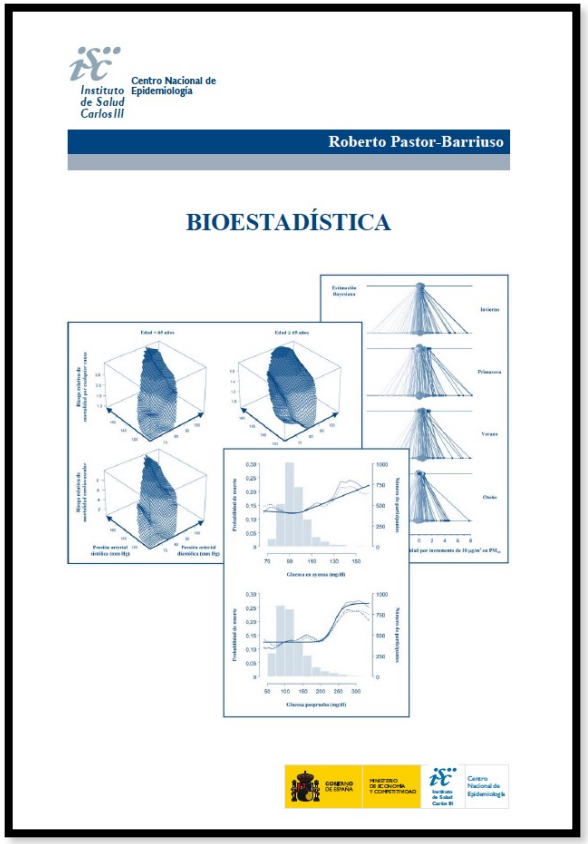
Estadística

Viñeta de Forges del 1 de julio de 2016; "El País"


GOBIERNO DE ESPAÑA
ESPAÑA SEGUIRÁ A LA CABEZA DEL CRECIMIENTO
PREVISIÓN DE CRECIMIENTO DEL PIB EN 2018
2,3%
2%
1,9%
España
Economías avanzadas
Eurozona
ESPAÑA LÍDERARÁ
en 2017 y 2018 el crecimiento entre los principales países desarrollados
Fuente: : Borrador del Plan Presupuestario 2018 y Fondo Monetario Internacional. Perspectivas Económicas Mundiales (WEO). Octubre 2017


2,3%
Manipulación
0,3%
0,1%
2%
1,9%
España
Economías avanzadas
Eurozona


2,3%
0.3%
0.1%
2%
2%
1,9%



Pastor-Barriuso R. Bioestadística. Madrid. Centro Nacional de Epidemiología, Instituto de Salud Carlos III, 2012, ISBN: 978-84-695-3775-6.

http://gesdoc.isciii.es/gesdoccontroller?action=download&id=03/06/2013-7dd67975c5

Estadística: definiciones

- "La **Estadística** es la rama de las matemáticas aplicadas que permite estudiar fenómenos cuyos resultados son en parte inciertos"*.
- La **Bioestadística** es una rama de la estadística que se ocupa de los problemas planteados dentro de las ciencias de la vida, como la biología, la medicina, entre otros.
- "Al estudiar sistema biológicos, esta incertidumbre se debe al desconocimiento de muchos de los mecanismos fisiológicos y fisiopatológicos, a la incapacidad de medir todos los determinantes de la enfermedad y a los errores de medida que inevitablemente se producen"*
- Dos grandes **grupos de técnicas** en Estadística*:
  - 1. **Descriptiva**s: técnicas necesarias para la organización, presentación y resumen de los datos.
  - 2. **Inferenciales**: técnicas para establecer conclusiones sobre la población a estudio a partir de los resultados obtenidos en una muestra.

**Pastor-Barriuso R. Bioestadística. Madrid. Centro Nacional de Epidemiología, Instituto de Salud Carlos III, 2012, ISBN: 978-84-695-3775-6.*

Tabla 9-1. Validez de los diferentes diseños para la inferencia etiológica.

Validez	Tipos de diseños
La más alta	Ensayo clínico aleatorizado
	Estudio de cohortes prospectivo
	Estudio de cohortes retrospectivo
	Estudio caso-control anidado
	Análisis de series temporales
	Estudios de sección-transversa
	Estudio ecológico
	Análisis de agregaciones
	Estudio de casos
La más baja	Anécdota

Tomado de Künzli y Tager. Environ Health Perspect 1997; 105:1078-83

Método Epidemiológico. Madrid. Escuela Nacional de Sanidad, Instituto de Salud Carlos III, 2009, ISBN: 978–84–95463–56–2.

Estadística: conceptos generales

- **Población**: “conjunto de todos los elementos que cumplen ciertas propiesdades y entre los cuales se desea estudiar un determinado fenómeno”*.

- **Muestra**: subconjunto de la población seleccionado mediante un mecanismo más o menos explícito.

- **Variables**:

  - Definición: propiedades o cualidades que presentan los elementos de una población.

  - Clasificación:

    Cualitativas o atributos (no pueden medirse cuantitativamente):

    Nominales= no pueden ordenarse sus categorías
    Ordinales= se pueden ordenar sus categorías

    Cuantitativas

    Discretas= sólo pueden tomar valores concretos dentro de un intervalo.
    Continuas= pueden tomar cualquier valor dentro de un intervalo.

**Pastor-Barriuso R. Bioestadística. Madrid. Centro Nacional de Epidemiología, Instituto de Salud Carlos III, 2012, ISBN: 978-84-695-3775-6.*

Estadística: conceptos generales

- **Estadístico**: cualquier operación realizada sobre los valores de una variable.

- **Parámetro**: valor de la población sobre el que se desea realizar inferencias a partir de estadísticos obtenidos de la muestras (estimadores).

  - Parámetros poblacionales con letras del alfabeto griego
  - **Estimadores muestrales** con letras de nuestro alfabeto

- **Ej**: Media de colesterol en la población (**μ**) estimada a partir de la Media de colesterol de una muestra de esa población ($\bar{X}$)

  Estadístico: media

**Pastor-Barriuso R. Bioestadística. Madrid. Centro Nacional de Epidemiología, Instituto de Salud Carlos III, 2012, ISBN: 978-84-695-3775-6.*

Estadística descriptiva

Índice

Estadística descriptiva

1. Introducción

Medidas de tendencia central

Medidas de posición

Medidas de dispersión

Representaciones gráficas

Estadística descriptiva

1. Introducción: Medidas de tendencia central

- Informan acerca de cuál es el valor más representativo de una determinada variable.
- Son estimadores que indican alrededor de qué valor se agrupan los datos observados
- Sirven para: resumir los resultados observados y para inferir parámetros poblacionales.
- Principales:

Media aritmética: suma de los valores muestrales dividida por el número de observaciones

Mediana: valor que deja por encima el 50% de los datos de la muestra y por debajo el otro 50%

Media Geométrica:

- Raíz enésima del producto de los valores de una muestra de tamaño n. Consiste en calcular el logaritmo de cada valor muestral, hallar a continuación la media aritmética de los logaritmos y deshacer finalmente la transformación logarítmica.

- Recomendable para variables muy asimétricas, donde un pequeño grupo de observaciones extremas tienen una excesiva influencia sobre la media aritmética.

Estadística descriptiva

1. Introducción: Medidas de posición

- **Cuantiles:** indican la posición relativa de una observación con respecto al resto de la muestra.
- Pueden ser estimados por varios métodos.
- Los más utilizados:

  Percentiles: valores de una variable que dejan un determinado porcentaje de los datos por debajo de ellos.

  Deciles: percentiles 10, 20,…90. (dividen a la muestra en 10 grupos de igual tamaño)
  Quintiles: percentiles 20,40,60 y 80 (dividen a la muestra en 5 grupos de igual tamaño)
  Cuartiles: percentiles 25,50 y 75 (dividen a la muestra en 4 grupos de igual tamaño)
  Terciles: percentiles 33.3 y 66.7 (dividen a la muestra en 3 grupos de igual tamaño)

Estadística descriptiva

1. Introducción: Medidas de dispersión

- Indican el grado de variabilidad de los datos y se complementan con las medidas de tendencia central
- Pueden ser estimados por varios métodos.
- Los más utilizados:

-**Varianza y desviación típica** (influenciadas por los valores extremos)

Varianza muestral: dificil de interprestar ya que sus unidades son las de la variable original al cuadrado.
Desviación típica: raiz cuadrada de la varianza

Rango intercuartílico

Coeficiente de Variación

-**Rango intercuartílico**: Diferencia entre el tercer cuartil y el primer cuartil. Medida de dispersión cuando hay muchos valores extremos. Suele ir asociada a la mediana

-**Coeficiente de Variación**: Cociente entre la desviación típica y la media aritmética expresado como porcentaje. Útil para comparar la variabilidad de distintas variables con distintas medias.

Estadística descriptiva

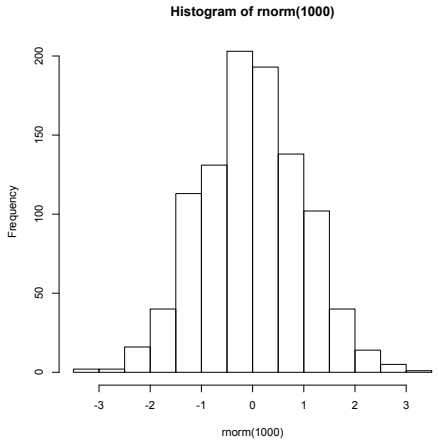
1. Introducción: Representaciones gráficas

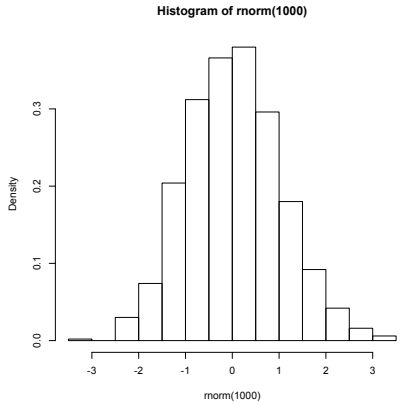
- **Histograma:**

  - Representar la distribución de una variable continua.

  - Los valores de la variable continua se agrupan en categorías
    a) exhaustivas (cubren todo el rango de la variable)
    b) mutuamente excluyentes (no se solapan)

  - Eje x (categorías) Eje y (frecuencias absolutas o relativas) de cada categorías.





Estadística descriptiva

1. Introducción: Representaciones gráficas

- **Grafico de Tallo y hojas (**stem-and-leaf plot)**:**

  - Refleja los datos originales de la muestra y permite visualizar la distribución de frecuencias.

  - Para cada observación de la variable se separa el último dígito significativo (hoja) de los restantes dígitos del valor de la variable (tallo).

  - Luego, todos los posibles tallos se colocan ordenados en la misma columna.

  - Finalmente, para cada valor de la variable, se coloca su hoja a la derecha del tallo

  - Las hojas de un mismo tallo suelen colocarse en orden creciente.

```
 [1] 37.81471 37.88094 38.14626 38.24091 38.37333 38.39332 38.51006 38.52017
 [9] 38.56449 38.56957 38.58035 38.62006 38.62867 38.73480 38.75368 38.76241
[17] 38.79192 38.81605 38.94836 38.95155 39.03235 39.04506 39.09779 39.12784
[25] 39.16968 39.17134 39.23746 39.31244 39.32239 39.32513 39.34844 39.35411
[33] 39.40083 39.40369 39.51879 39.54382 39.56961 39.58365 39.59936 39.62634
[41] 39.63632 39.66544 39.67546 39.69879 39.74352 39.74622 39.76177 39.77734
[49] 39.78150 39.80485 39.80852 39.81575 39.89824 39.92205 39.92604 39.97118
[57] 40.01875 40.06954 40.08935 40.14817 40.15883 40.18493 40.23253 40.29099
[65] 40.29455 40.34012 40.36209 40.38092 40.38979 40.44024 40.48298 40.50582
[73] 40.53290 40.56274 40.59283 40.59496 40.65523 40.65976 40.71289 40.71660
[81] 40.73569 40.74139 40.75578 40.78634 40.83447 40.89293 40.92552 40.96857
[89] 40.98744 41.06638 41.08655 41.10178 41.15535 41.17271 41.21613 41.22703
[97] 41.36795 41.52259 42.13777 42.22052
```

```
The decimal point is at the |

37 | 89
38 | 1244
38 | 5566666788889
39 | 000112223333444
39 | 556666677777888888999
40 | 001112223334444
40 | 55566677777788899
41 | 0011122224
41 | 5
42 | 12
```

Estadística descriptiva

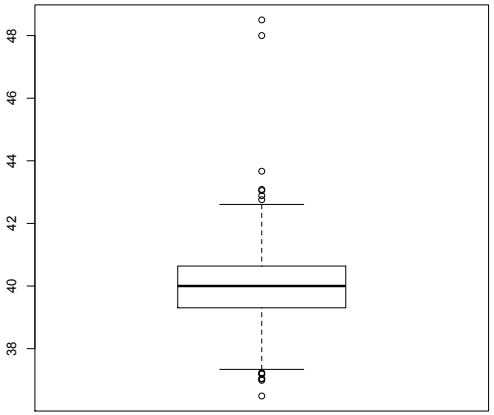
1. Introducción: Representaciones gráficas

- **Diagrama de caja:**

  - Permite evaluar la tendencia central, la dispersión y la simetría de la distribución de una variable, así como identificar valores extremos.

  - Percentiles 25 y 75
  - Altura de la caja: rango intercuartílico
  - Línea central: mediana
  - Líneas verticales: 1.5 veces el rango intercuartílico.
  - Valores extremos: aquellos distanciados 1.5 (circulo) y 3 (asterisco) veces el rango intercuartílico desde el límite de la caja.

upper whisker = min(max(x), Q_3 + 1.5 * IQR)
lower whisker = max(min(x), Q_1 – 1.5 * IQR)



Estadística descriptiva

2. Univariable cuantitativa

Estadísticos básicos

- Media: **mean**()
- Mediana: **median**()
- Media geométrica: **geometric.mean**() [psych]; exp(mean(log( )))
- Cuantiles: **quantile**()
- Deciles: **quantile**(x,prob=seq(0,1,1/10))
- Cuartiles: **quantile**(); **quantile**(x,prob=seq(0,1,1/4))
- CV: (**sd**(x)/**mean**(x))*100
-Varianza: **var**()
-Desviación estándar: **sd**()
-Rango intercuartílico: **IQR**()
-**range**()
-**summary**()

Estadística descriptiva

2. Univariable cuantitativa: gráficos

- Histogramas

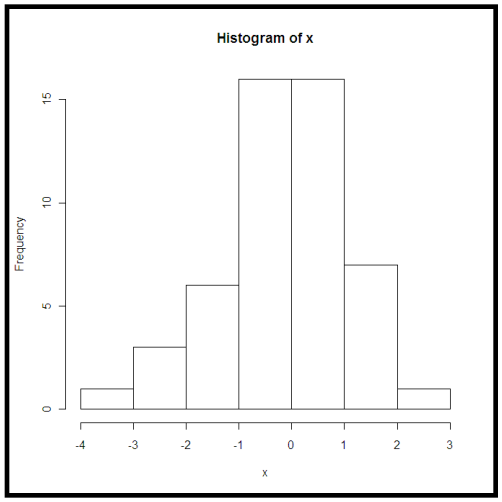
  hist()

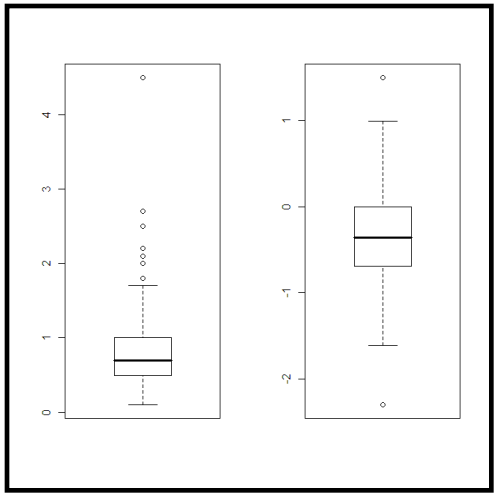
- Boxplots

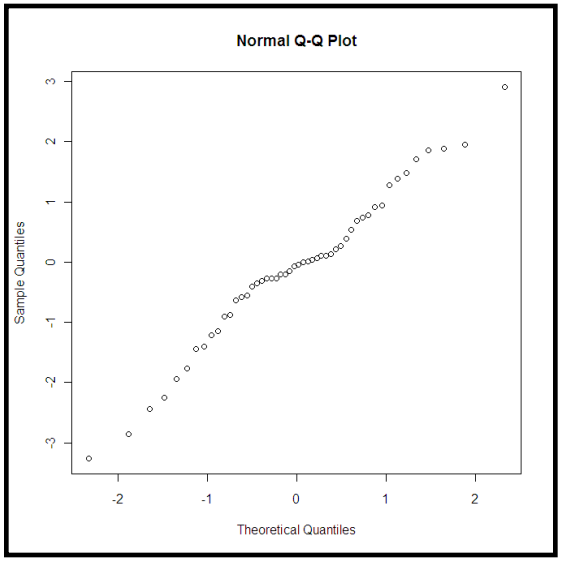
  boxplot()

- Q-Q plots

  qqnorm()







Estadística descriptiva

3. Datos agrupados cuantitativa

Cuando se trabaja con datos agrupados, solemos querer estadísticos agrupados por grupos.

Ej. Tabla de medias y desviaciones estándar. Para ello usamos **tapply**

tapply(x,grupo.variable,statistics,na.rm=T)

Estadística descriptiva

3. Datos agrupados cuantitativa: gráficos

- Histogramas

  par(mfrow=c())

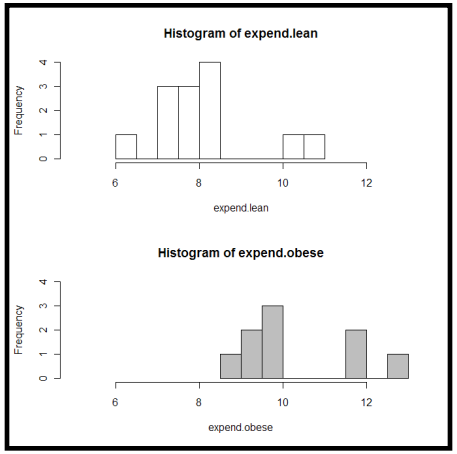
  hist()

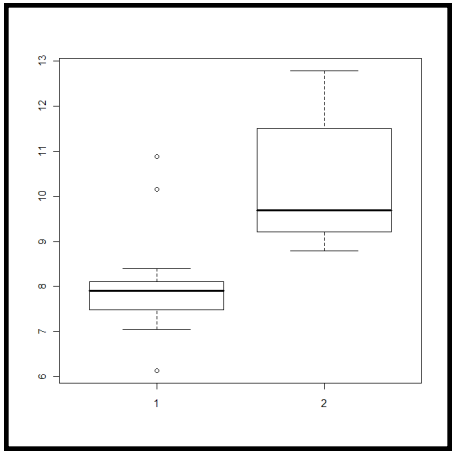
- Boxplots (diagrama de caja)

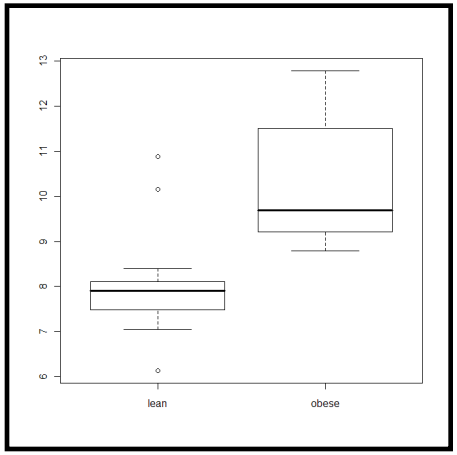
  boxplot(A~B)

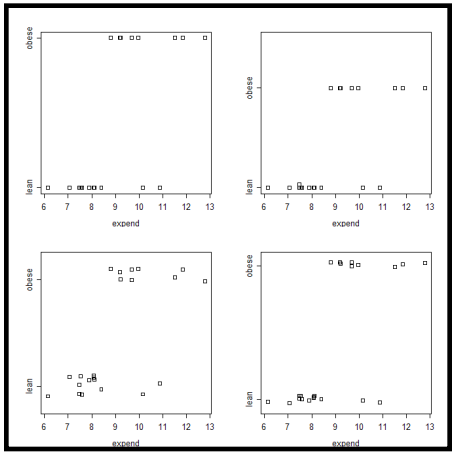
- Stripcharts

  stripchart()









Estadística descriptiva

4. Univariable cualitativa

- Para datos categóricos
- Generar tablas

  table()

- Tablas marginales y frecuancias relativas

  margin.table(**table**())

  prop.table(**table**())

```
> table(menarche,tanner)
        tanner
menarche   I  II III  IV   V
     No  221  43  32  14   2
     Yes   1   1   5  26 202
```

```
> 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
```

Estadística descriptiva

4. Univariable cualitativa: gráficos

- Gráficos de barras

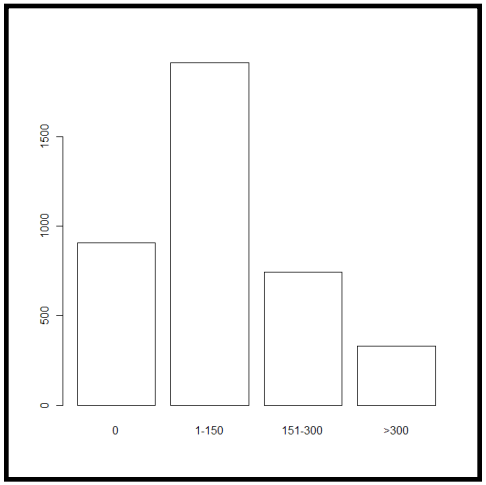
  barplot()

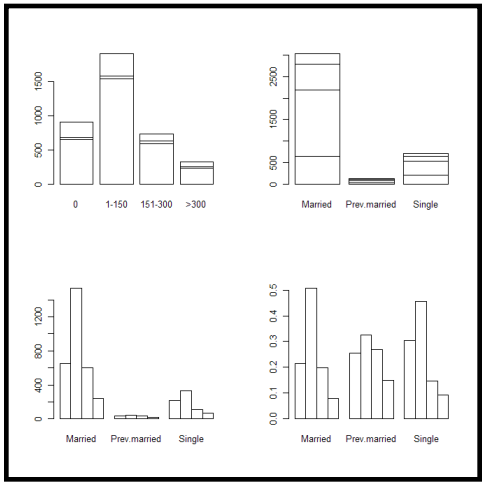
- Dotcharts

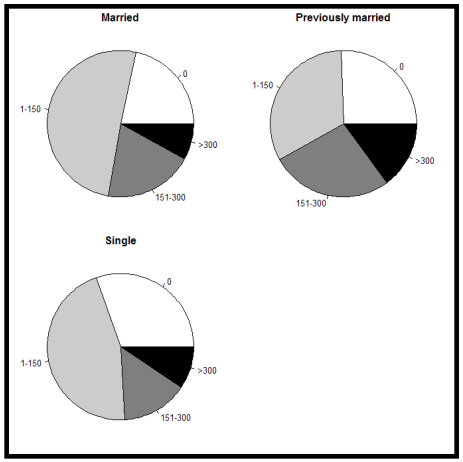
  dotchart()

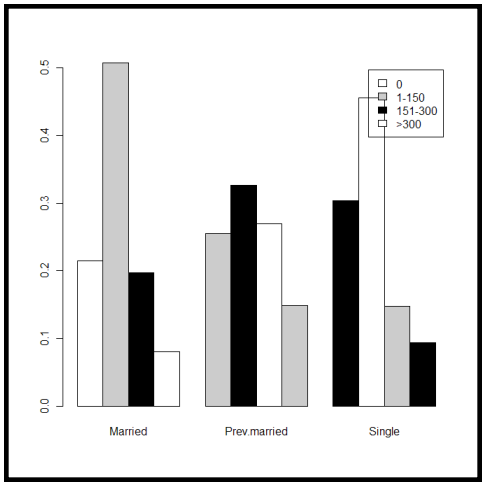
- Piecharts (diagramas de sectores)

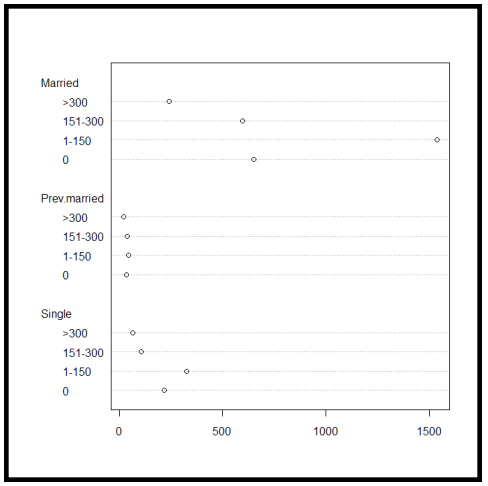
  pie()











Estadística descriptiva

5. Datos agrupados cualitativa

- Generar tablas

  table(x,y)

  gmodels::**CrossTable**

  crosstable::**crosstable**

EJERCICIOS

"ejercicios.estadistica.descriptiva.pdf"