Estadística Descriptiva

1. En Estadística Descriptiva se exploran los datos mediante:
   - gráficos
   - medidas de resumen.

   La idea es **visualizar** en forma rápida las principales características del conjunto de datos.

2. Con frecuencia los conjuntos de datos provienen de medir una o más variables en un conjunto de individuos.

3. Se suele comenzar con un análisis individual de cada variable medida y posteriormente se estudian las relaciones entre variables. El orden habitual es primero representaciones gráficas y luego las medidas numéricas o de resumen.

¿Datos?

1. POBLACION: total de sujetos o unidades de análisis de interés en el estudio.
   Estamos interesados en estudiar un fenómeno en una población

2. MUESTRA: cualquier subconjunto de los sujetos o unidades de análisis de la población, en el cual se recolectarán los datos.
   Usamos una muestra para conocer o *estimar* características de la población

3. PARAMETRO: una medida resumen calculada sobre la población: media, varianza, proporción.

4. ESTADISTICO : una medida resumen calculada sobre la muestra.

¿Qué medimos?

- VARIABLE: es una característica que varía de individuo en individuo.
  (edad, peso, altura, género, concentración de colesterol en sangre, club de fútbol preferido, etc.)

- DATOS: son los valores observados de las variables en estudio medidos en la muestra.

¿Qué buscamos?

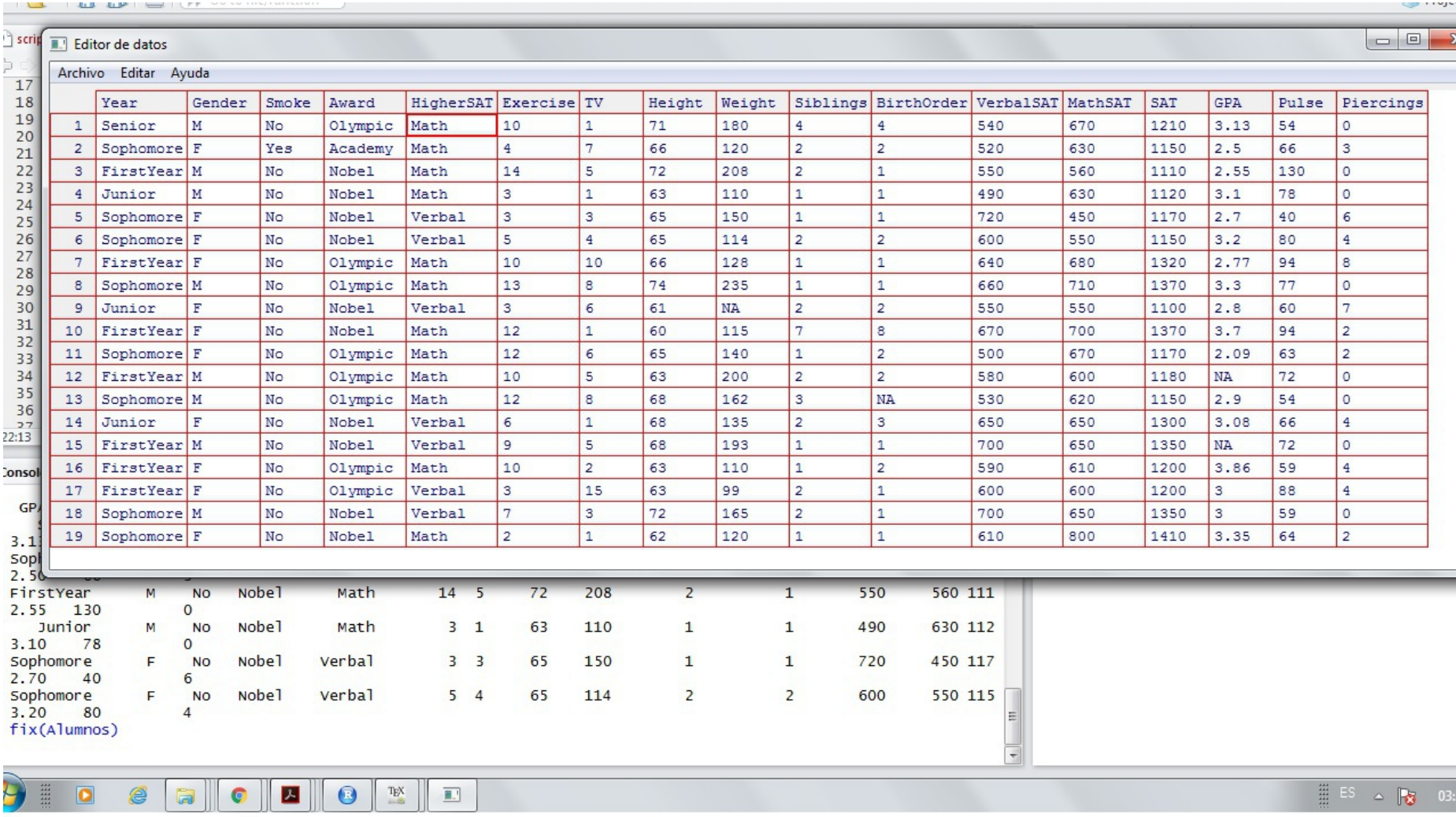
Respecto de una variable, tratamos de responder a preguntas tales como:

- ¿Son los valores medidos casi todos iguales o son muy diferentes entre sí?
- ¿En qué sentido difieren?
- ¿Las mediciones tienden a concentrarse alrededor de un valor?
- ¿Existe algún patrón o tendencia?
- ¿Constituyen un único grupo? ¿Hay varios grupos?
- ¿Difieren algunos pocos datos notablemente del resto?
- ¿Cómo se relaciona esta variable con otra de nuestro interés tomada en la misma muestra?

Veamos un ejemplo

Este archivo corresponde a una encuesta a estudiantes (fuente: Statistics: unlocking the power of the data, $Lock^5$, 2013, Wiley) que se realizó durante muchos años el primer día de clase en una universidad en una materia introductoria de Estadística. En ella se registraron 17 variables y 362 casos.

Archivo: StudentSurvey.txt



Editor de datos

Archivo Editar Ayuda

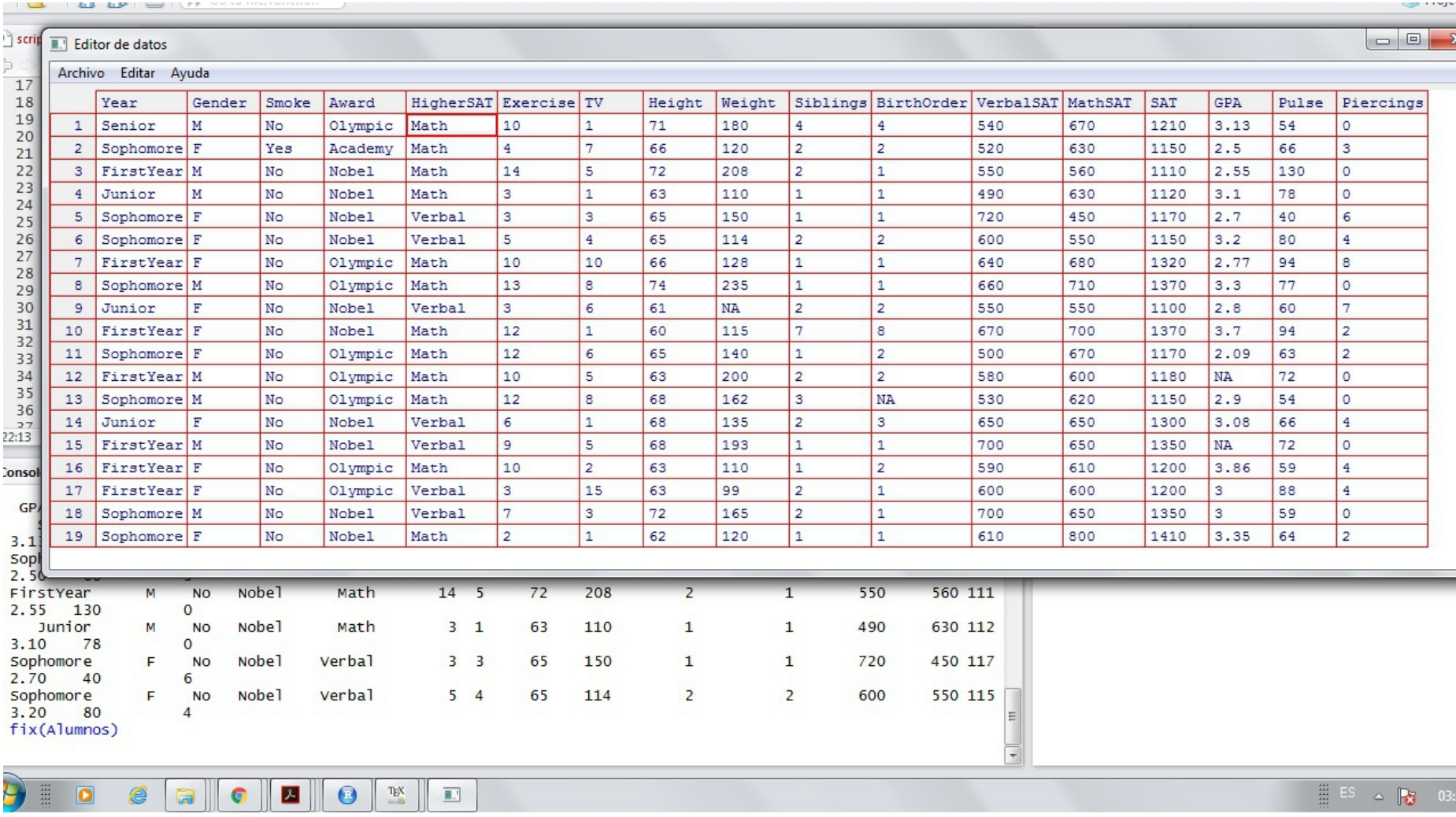
	Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	Weight	Siblings	BirthOrder	VerbalSAT	MathSAT	SAT	GPA	Pulse	Piercings
1	Senior	M	No	Olympic	Math	10	1	71	180	4	4	540	670	1210	3.13	54	0
2	Sophomore	F	Yes	Academy	Math	4	7	66	120	2	2	520	630	1150	2.5	66	3
3	FirstYear	M	No	Nobel	Math	14	5	72	208	2	1	550	560	1110	2.55	130	0
4	Junior	M	No	Nobel	Math	3	1	63	110	1	1	490	630	1120	3.1	78	0
5	Sophomore	F	No	Nobel	Verbal	3	3	65	150	1	1	720	450	1170	2.7	40	6
6	Sophomore	F	No	Nobel	Verbal	5	4	65	114	2	2	600	550	1150	3.2	80	4
7	FirstYear	F	No	Olympic	Math	10	10	66	128	1	1	640	680	1320	2.77	94	8
8	Sophomore	M	No	Olympic	Math	13	8	74	235	1	1	660	710	1370	3.3	77	0
9	Junior	F	No	Nobel	Verbal	3	6	61	NA	2	2	550	550	1100	2.8	60	7
10	FirstYear	F	No	Nobel	Math	12	1	60	115	7	8	670	700	1370	3.7	94	2
11	Sophomore	F	No	Olympic	Math	12	6	65	140	1	2	500	670	1170	2.09	63	2
12	FirstYear	M	No	Olympic	Math	10	5	63	200	2	2	580	600	1180	NA	72	0
13	Sophomore	M	No	Olympic	Math	12	8	68	162	3	NA	530	620	1150	2.9	54	0
14	Junior	F	No	Nobel	Verbal	6	1	68	135	2	3	650	650	1300	3.08	66	4
15	FirstYear	M	No	Nobel	Verbal	9	5	68	193	1	1	700	650	1350	NA	72	0
16	FirstYear	F	No	Olympic	Math	10	2	63	110	1	2	590	610	1200	3.86	59	4
17	FirstYear	F	No	Olympic	Verbal	3	15	63	99	2	1	600	600	1200	3	88	4
18	Sophomore	M	No	Nobel	Verbal	7	3	72	165	2	1	700	650	1350	3	59	0
19	Sophomore	F	No	Nobel	Math	2	1	62	120	1	1	610	800	1410	3.35	64	2

```
FirstYear      M   No   Nobel     Math     14  5    72    208     2        1      550     560 111
2.55   130        0
   Junior      M   No   Nobel     Math      3  1    63    110     1        1      490     630 112
3.10    78        0
Sophomore      F   No   Nobel   Verbal      3  3    65    150     1        1      720     450 117
2.70    40        6
Sophomore      F   No   Nobel   Verbal      5  4    65    114     2        2      600     550 115
3.20    80        4
fix(Alumnos)
```

Las variables registradas son:

1. **Year**: Años en la institución: FirstYear, Sophomore, Junior, o Senior
2. **Gender**: Sexo: F o M
3. **Smoke**: Fuma? No o Yes
4. **Award**: Premio preferido: Academy, Nobel, o Olympic
5. **HigherSAT**: Qué puntaje SAT es más alto? Math o Verbal
6. **Exercise**: Horas semanales de ejercicio
7. **TV**: Horas semanales de TV
8. **Height**: altura (en pulgadas)
9. **Weight**: peso (en libras)
10. **Siblings**: número de hermanos
11. **BirthOrder**: orden de nacimiento, 1 = mayor, 2 = segundo mayor, etc.
12. **VerbalSAT**: puntaje SAT en lengua
13. **MathSAT**:puntaje SAT en matenáticas
14. **SAT**: puntaje SAT combinado de lengua y matemáticas
15. **GPA**: puntaje promedio obtenido en el colegio
16. **Pulse**: pulso cardíaco (latidos por minuto)
17. **Piercings**: número de piercings

Archivo: StudentSurvey.txt



Editor de datos

Archivo Editar Ayuda

	Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	Weight	Siblings	BirthOrder	VerbalSAT	MathSAT	SAT	GPA	Pulse	Piercings
1	Senior	M	No	Olympic	Math	10	1	71	180	4	4	540	670	1210	3.13	54	0
2	Sophomore	F	Yes	Academy	Math	4	7	66	120	2	2	520	630	1150	2.5	66	3
3	FirstYear	M	No	Nobel	Math	14	5	72	208	2	1	550	560	1110	2.55	130	0
4	Junior	M	No	Nobel	Math	3	1	63	110	1	1	490	630	1120	3.1	78	0
5	Sophomore	F	No	Nobel	Verbal	3	3	65	150	1	1	720	450	1170	2.7	40	6
6	Sophomore	F	No	Nobel	Verbal	5	4	65	114	2	2	600	550	1150	3.2	80	4
7	FirstYear	F	No	Olympic	Math	10	10	66	128	1	1	640	680	1320	2.77	94	8
8	Sophomore	M	No	Olympic	Math	13	8	74	235	1	1	660	710	1370	3.3	77	0
9	Junior	F	No	Nobel	Verbal	3	6	61	NA	2	2	550	550	1100	2.8	60	7
10	FirstYear	F	No	Nobel	Math	12	1	60	115	7	8	670	700	1370	3.7	94	2
11	Sophomore	F	No	Olympic	Math	12	6	65	140	1	2	500	670	1170	2.09	63	2
12	FirstYear	M	No	Olympic	Math	10	5	63	200	2	2	580	600	1180	NA	72	0
13	Sophomore	M	No	Olympic	Math	12	8	68	162	3	NA	530	620	1150	2.9	54	0
14	Junior	F	No	Nobel	Verbal	6	1	68	135	2	3	650	650	1300	3.08	66	4
15	FirstYear	M	No	Nobel	Verbal	9	5	68	193	1	1	700	650	1350	NA	72	0
16	FirstYear	F	No	Olympic	Math	10	2	63	110	1	2	590	610	1200	3.86	59	4
17	FirstYear	F	No	Olympic	Verbal	3	15	63	99	2	1	600	600	1200	3	88	4
18	Sophomore	M	No	Nobel	Verbal	7	3	72	165	2	1	700	650	1350	3	59	0
19	Sophomore	F	No	Nobel	Math	2	1	62	120	1	1	610	800	1410	3.35	64	2

```
FirstYear      M   No   Nobel     Math     14  5    72    208     2        1      550     560 111
2.55   130        0
   Junior      M   No   Nobel     Math      3  1    63    110     1        1      490     630 112
3.10    78        0
Sophomore      F   No   Nobel   Verbal      3  3    65    150     1        1      720     450 117
2.70    40        6
Sophomore      F   No   Nobel   Verbal      5  4    65    114     2        2      600     550 115
3.20    80        4
fix(Alumnos)
```

Preguntas posibles en el ejemplo

- ¿Cuál es el porcentaje de hombres y de mujeres?
- ¿Qué porcentaje de los alumnos fuma?
- ¿Quiénes fuman más, hombres o mujeres?
- ¿Cuál es el promedio de horas de ejercicio?
- ¿Hay alumnos con un puntaje de VerbalSAT inusualmente alto o bajo?
- ¿Cuál es el premio más deseado?
- ¿Qué relación hay entre peso y altura? ¿Es la misma en hombres que en mujeres?
- ¿Los alumnos que hacen más ejercicio tienden a tener pulso más bajo?

Tipos de variables o datos

Por lo general, se comienza a trabajar con las variables en forma individual.

Antes que nada debemos identificar el tipo de variable con la que vamos a trabajar.

¿Por qué es importante identificar el tipo de datos?
La naturaleza de una variable determina el tratamiento o análisis estadístico apropiado y válido que podemos aplicar. Por lo general, un método estadístico es específico para un cierto tipo de datos.

Tipos de variables o datos

- Variables cualitativas (factor en R): Describen cualidades o atributos (ej.: género, color del ojos, estado civil, fuma no fuma, severidad de la patología: Ausente/leve/moderado/severo). En Alumnos: por ej.: Gender, Smoke y Award

Tipos de variables o datos

- Variables cualitativas (factor en R): Describen cualidades o atributos (ej.: género, color del ojos, estado civil, fuma no fuma, severidad de la patología: Ausente/leve/moderado/severo). En Alumnos: por ej.: Gender, Smoke y Award

- Variables cuantitativas (numeric en R)
  - Discretas: Toman un cierto número de valores posibles. En general, aparecen por conteo. (ej.: número de miembros del hogar, número de hijos, número de intervenciones quirúrgicas, número de casos notificados de una cierta patología). En Alumnos: por ej.: Siblings, Piercings, Pulse

Tipos de variables o datos

- Variables cualitativas (factor en R): Describen cualidades o atributos (ej.: género, color del ojos, estado civil, fuma no fuma, severidad de la patología: Ausente/leve/moderado/severo). En Alumnos: por ej.: Gender, Smoke y Award

- Variables cuantitativas (numeric en R)
  - Discretas: Toman un cierto número de valores posibles. En general, aparecen por conteo. (ej.: número de miembros del hogar, número de hijos, número de intervenciones quirúrgicas, número de casos notificados de una cierta patología). En Alumnos: por ej.: Siblings, Piercings, Pulse

  - Continuas: Toman valores en un intervalo (ej.: altura, peso, pH, nivel de colesterol en sangre, tiempo hasta que llega un tren). En Alumnos: por ej.: GPA

La distinción más importante es entre datos cuantitativos (numéricos) y categóricos.

Algunas veces esta clasificación no es tan estricta...

Ejemplo: $Edad$ (tiempo de vida) es continua... pero si se la registra en años resulta ser discreta.

Cuando se trabaja con adultos se suele tratar como discreta, sin embargo en menores de 1 año en lo que se registran meses de vida, se la suele tratar como continua.

Los datos numéricos (discretos o continuos) pueden ser transformados en categóricos y ser tratados como tales, por ej.,

$$Edad >= cota \text{ o } Edad < cota.$$

Representación de datos categóricos

Tabla de frecuencia El modo más simple de presentar datos categóricos es por medio de una tabla de **frecuencias** que indica el número observaciones que caen en cada una de las clases de la variable. La **frecuencia relativa** de una clase es el cociente entre la frecuencia y el número total de observaciones.

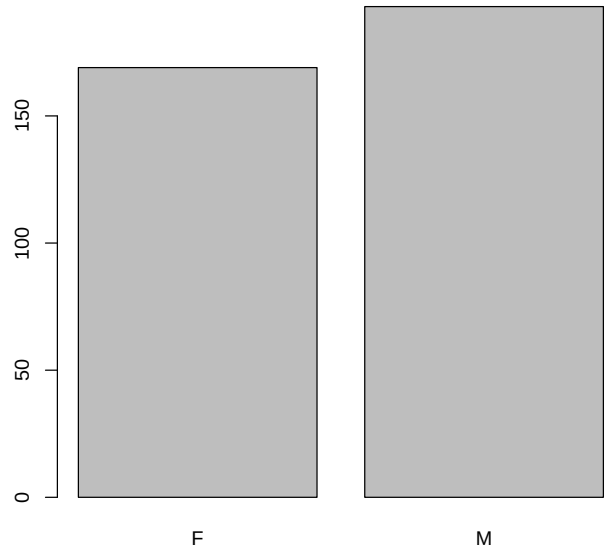
Gráfico de Barras A cada categoría o clase de la variable se le asocia una barra cuya altura representa la frecuencia o la frecuencia relativa de esa clase. Las barras difieren sólo en altura, no en ancho y se representan separadas por un espacio.

Gráfico de Tortas Se representa la frecuencia relativa de cada categoría como una porción de un círculo, en la que el ángulo se corresponde con la frecuencia relativa correspondiente.

Tabla de frecuencia y gráfico de barras: barplot

Gender	
F	M
169	193

Frecuencia



Frecuencia Relativa

Tabla de frecuencia y gráfico de barras: barplot

Gender	
F	M
169	193

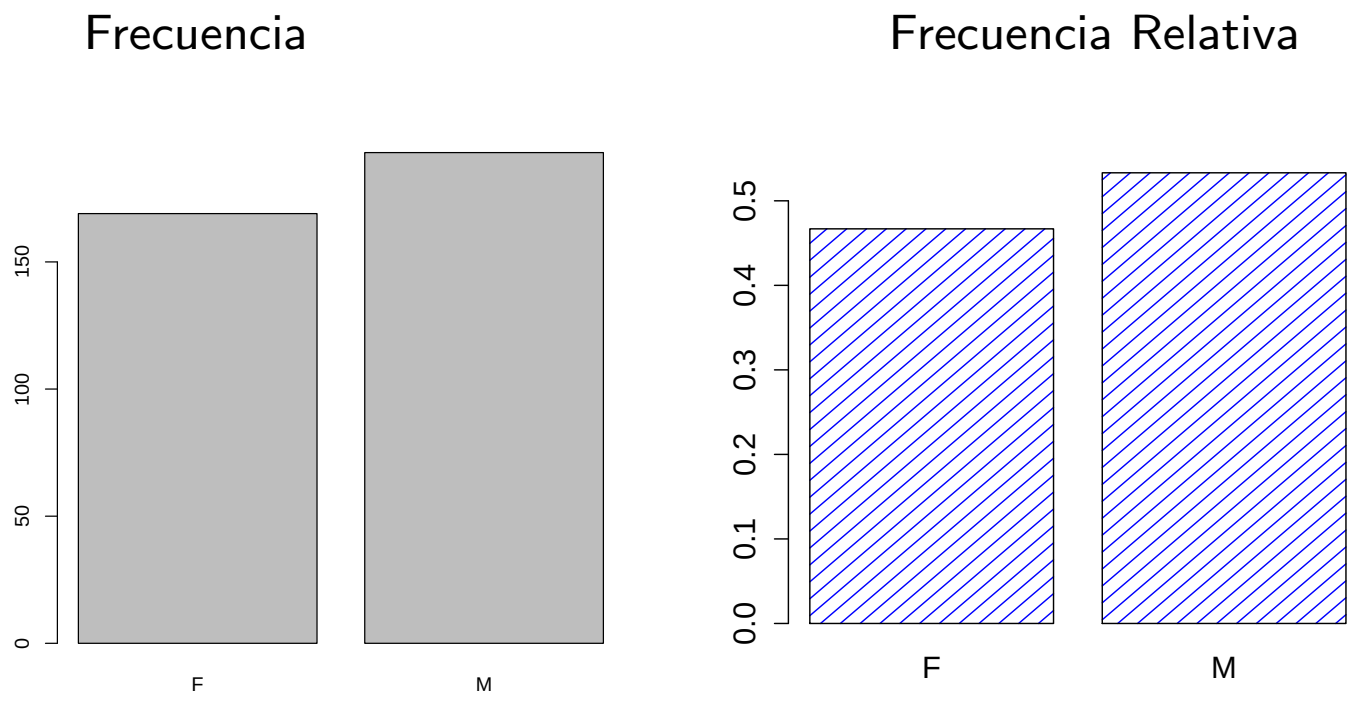


Tabla de frecuencia y gráfico de torta

Smoke	
No	Yes
319	43

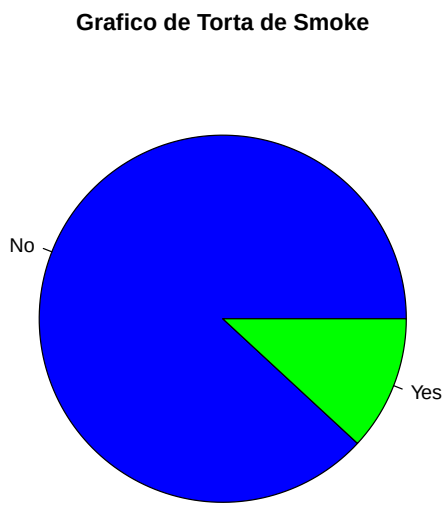
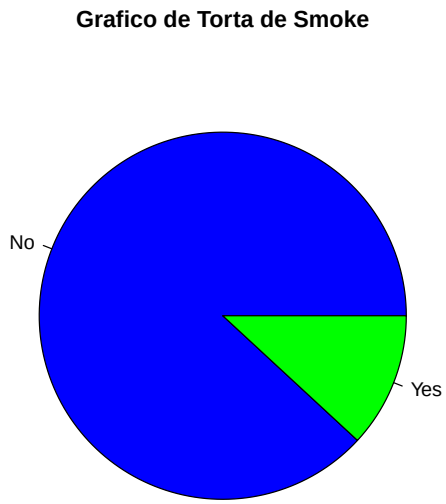


Tabla de frecuencia y gráfico de torta

Smoke	
No	Yes
319	43



Tablas de contingencia

Relación entre variables categóricas

Consideramos cómo se distribuyen los 362 estudiantes entre las categorías que surgen al combinar el género y la condición de fumador.

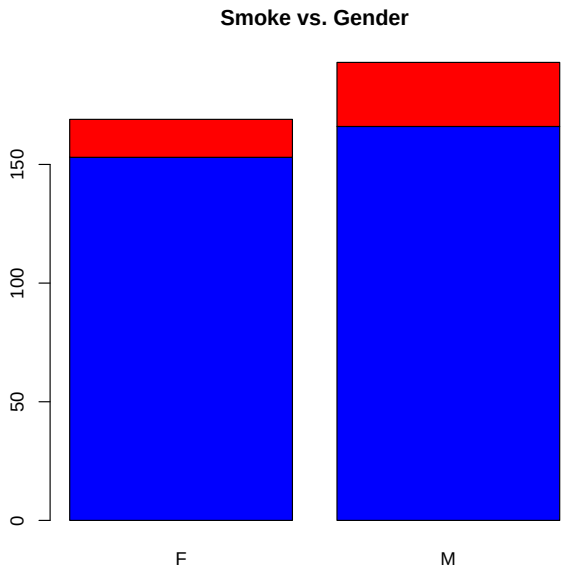
En una tabla de contingencia se clasifican todos los alumnos teniendo en cuenta todas las categorías posibles al cruzar las dos variables.

Esta tabla se puede representar de distintas maneras.

Tablas de Contingencia

	Gender	
Smoke	F	M
No	153	166
Yes	16	27
Total	169	193

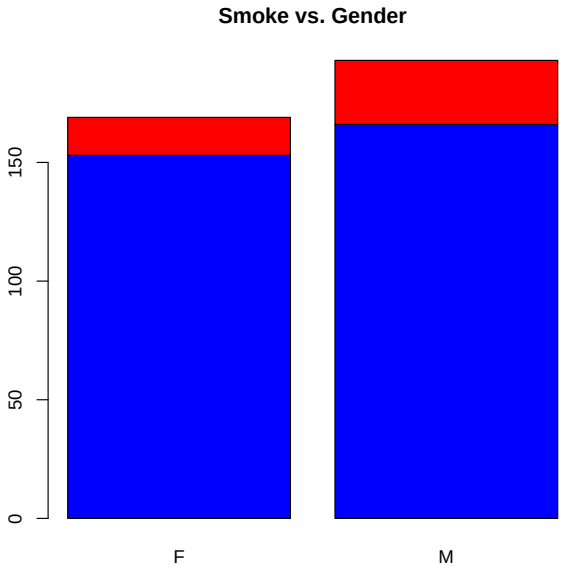
Frecuencia



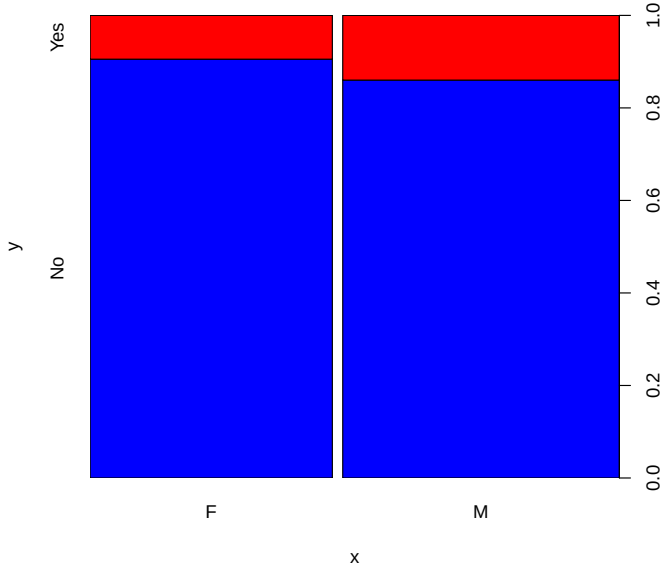
Tablas de Contingencia

	Gender	
Smoke	F	M
No	153	166
Yes	16	27
Total	169	193

Frecuencia

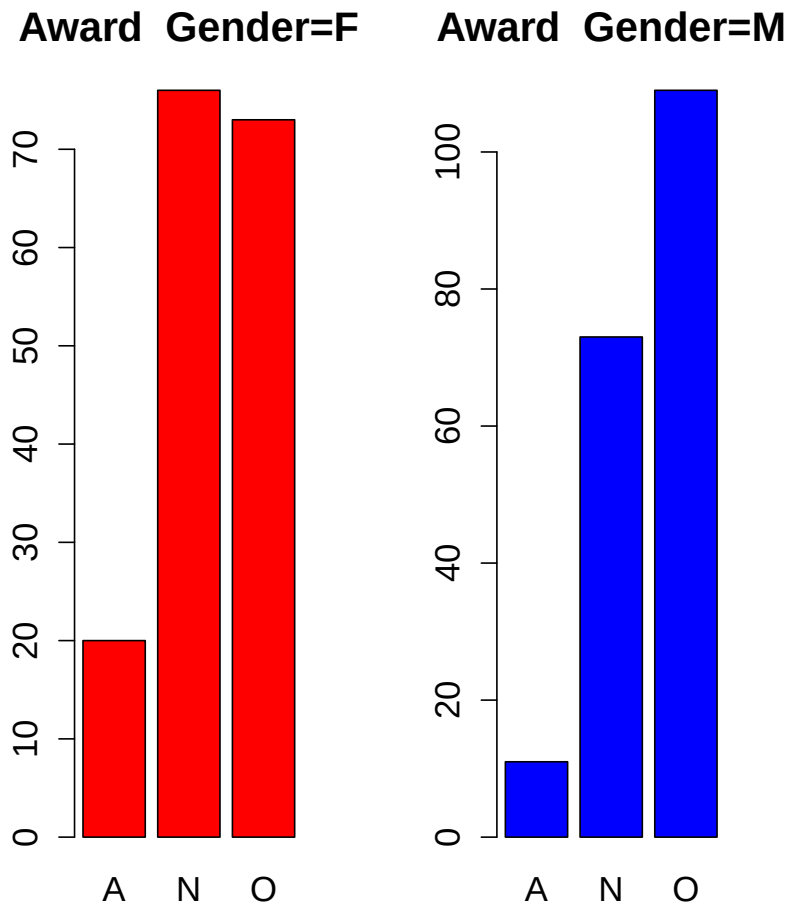


Frecuencia Relativa



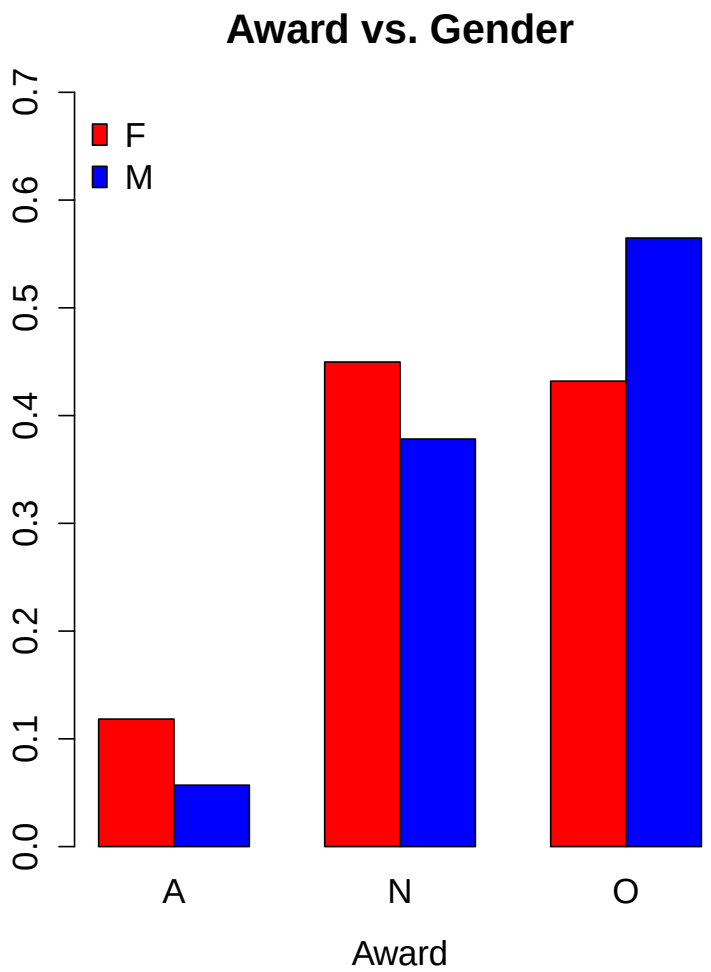
Award vs. Gender

	Award			
Gender	Academy	Nobel	Olympic	Total
F	20	76	73	169
M	11	73	109	193



Award vs. Gender

	Award			
Gender	Academy	Nobel	Olympic	Total
F	20	76	73	169
M	11	73	109	193



Gráficos para variables cuantitativas

El histograma es el más conocido de los gráficos para resumir un conjunto de datos cuantitativos o numéricos.

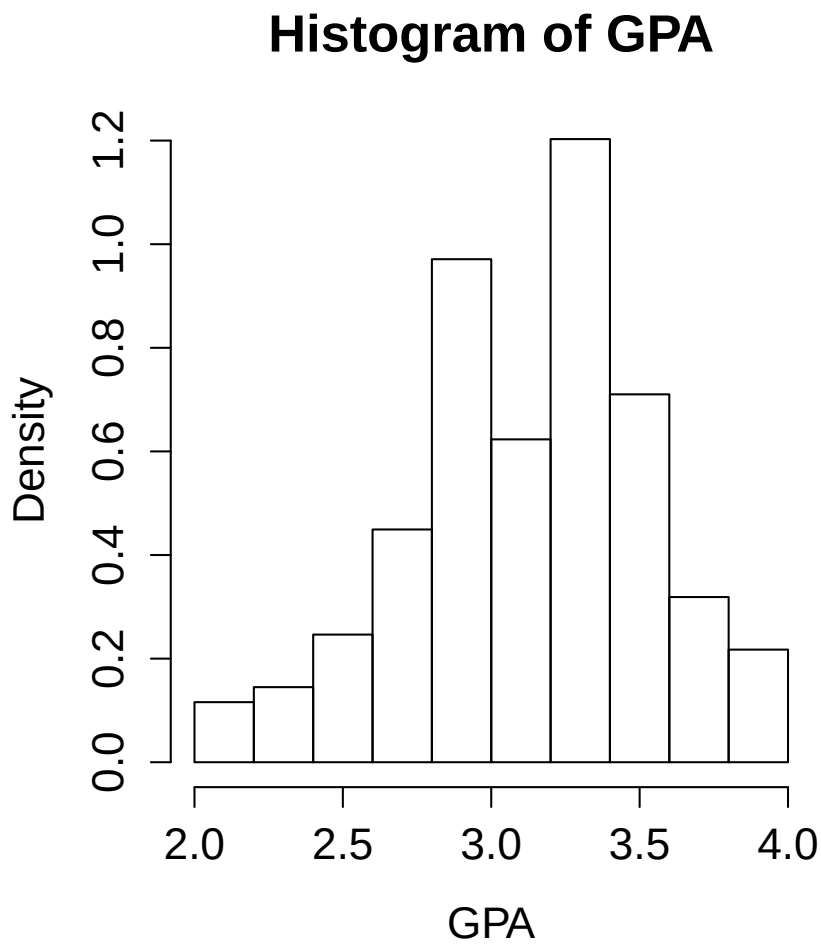
1. Para construir un histograma es necesario previamente construir una tabla de frecuencias: dividimos el rango de los n datos en intervalos o clases, que son excluyentes y exhaustivas.
2. Contamos la cantidad de datos en cada intervalo o clase i, es decir la frecuencia, f_i y calculamos la frecuencia relativa:

$$fr_i = f_i/n$$

3. Graficamos el histograma en un par de ejes coordenados representando en las abscisas los intervalos y sobre cada uno de ellos un rectángulo cuya área es la frecuencia relativa de dicho intervalo.

Comencemos con **GPA**, que es continua.

Histograma de GPA

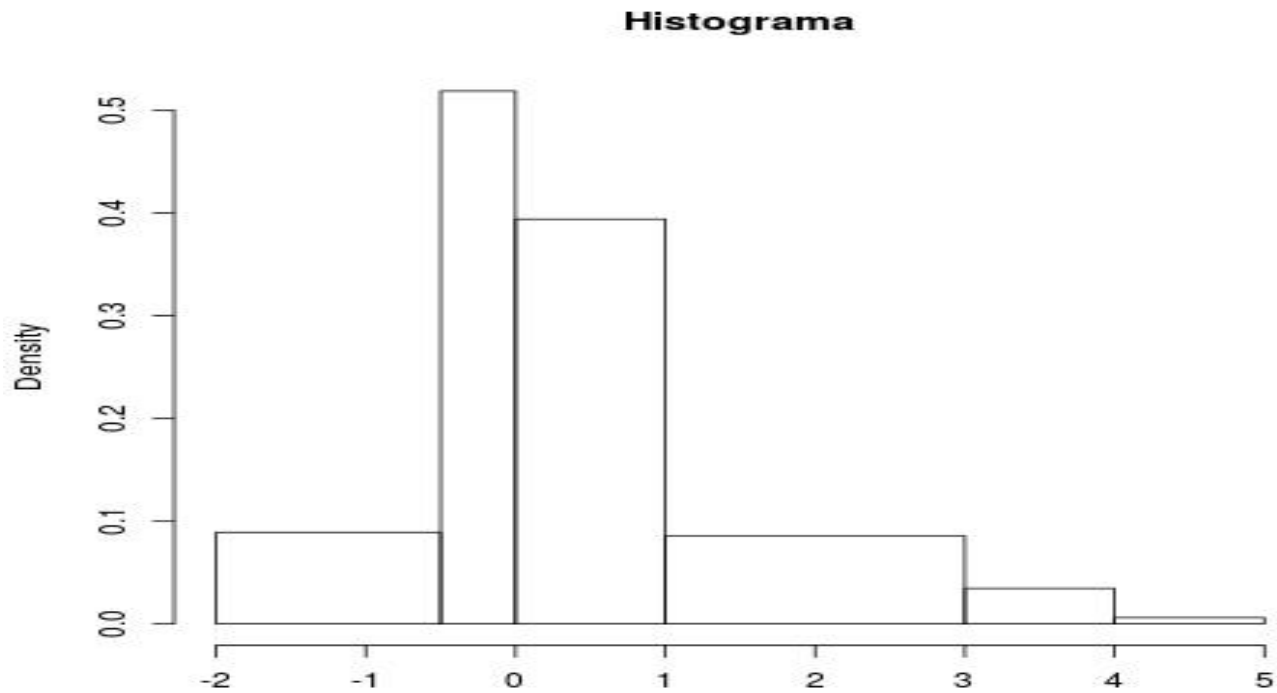


En general, los intervalos se toman de igual longitud y de esa manera la altura es proporcional a la frecuencia relativa, pero esto no es necesariamente siempre así.

Al usar frecuencias relativas, se facilita la comparación de distintos histogramas.

Histograma: Area=Frecuencia Relativa

int.:	$[-2,-0.5)$	$[-0.5,0)$	$[0,1)$	$[1,3)$	$[3,4)$	$[4,5)$
frec.:	174	337	512	223	45	8
fre. rel.:	$174/n$	$337/n$	$512/n$	$223/n$	$45/n$	$8/n$



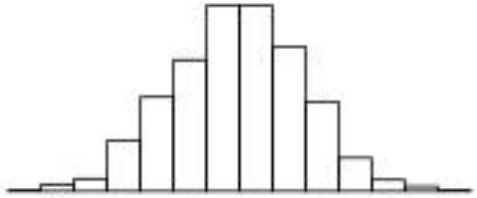
¿En qué difieren un gráfico de barras y un histograma?

El gráfico de barras representa el porcentaje en la altura de la barra. Mientras que en un histograma el porcentaje se representa en el área de la barra.

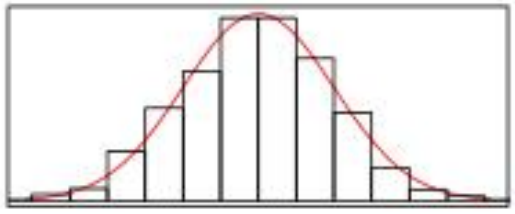
En el gráfico de barras, las barras se representan separadas para indicar que no hay continuidad entre las categorías. En un histograma barras adyacentes deben estar en contacto indicando que la variable es continua o discreta.

Algunos Histogramas

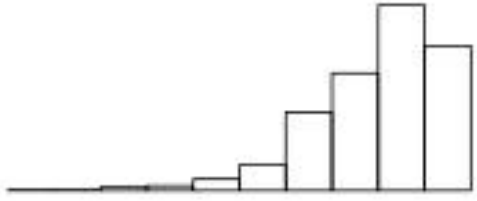
Acampanado



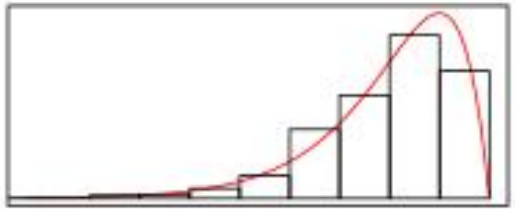
Acampanado



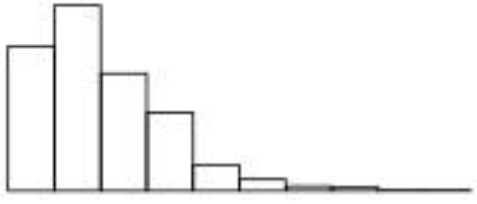
Colas pesada a Iquierda



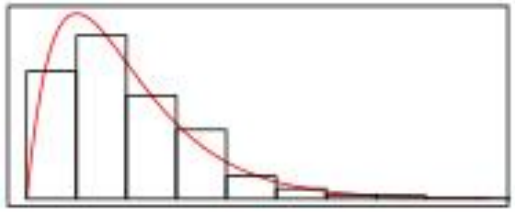
Colas pesada a Iquierda



Colas pesada a Derecha



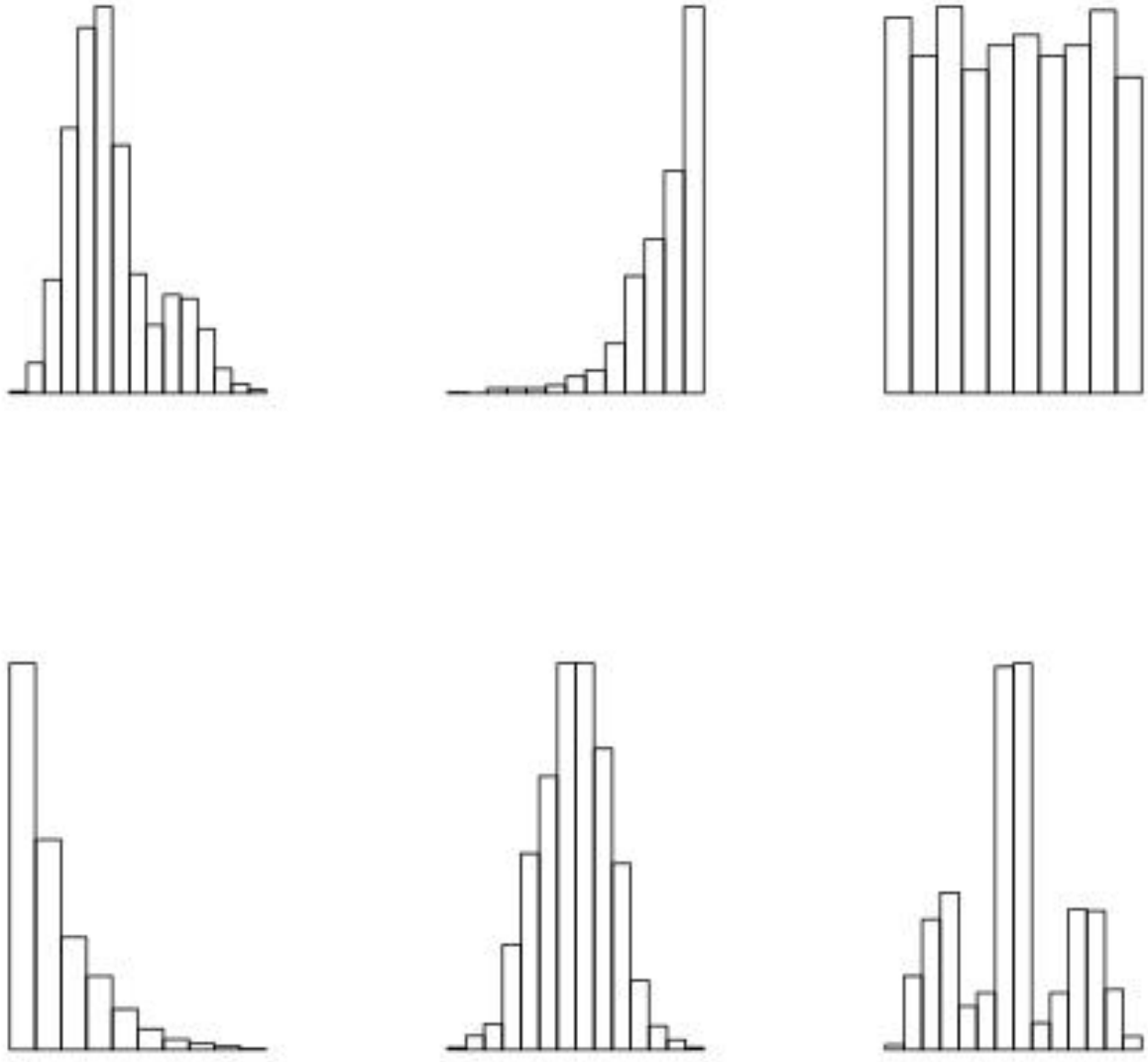
Colas pesada a Derecha



¿Qué miramos en un histograma?

1. Rango de variación de los datos (Mínimo y Máximo).
2. Intervalos más frecuentes
3. ¿La distribución es unimodal o hay más de una moda (pico)?
4. ¿La distribución es simétrica?
5. Si es asimétrica, ¿la asimetría es a derecha o a izquierda?
6. ¿En torno a qué valor están aproximadamente centrados los datos?
7. ¿Cuán dispersos en torno a este centro están los datos ?
8. ¿Hay datos atípico en relación a la mayoría de los datos?

Más Histogramas



Variables cuantitativas: Medidas de resumen

Resumimos la información de los datos mediante medidas de fácil interpretación que reflejen sus características más relevantes.
Las medidas de resumen son útiles para comparar conjuntos de datos y para presentar los resultados de un estudio.

Podemos agruparlas en:

Medidas de posición o localización: describen un valor alrededor del cual se encuentran las observaciones.

Medidas de dispersión o escala: expresan la variabilidad presente en un conjunto de datos.

Medidas de Posición

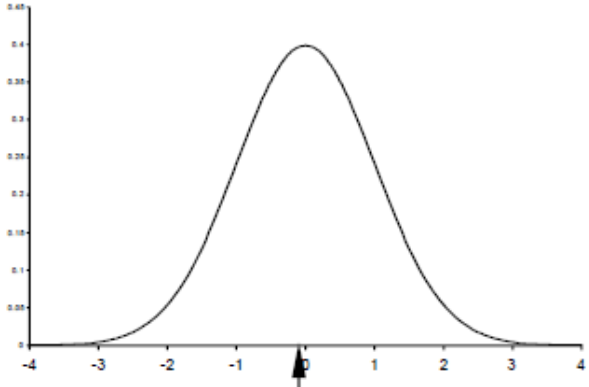
Medidas de Posición o Centrado

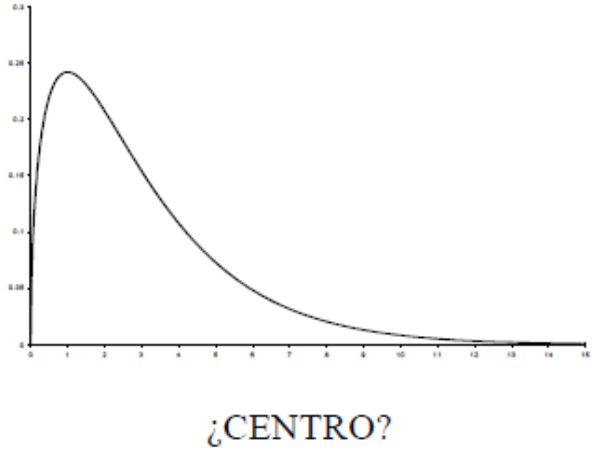
¿Cuál es el valor central o que mejor representa a los datos?

Buscamos un valor típico que represente a los datos.

Si la distribución es simétrica diferentes medidas darán resultados similares y hay un claro centro.

Si es asimétrica no existe un centro evidente y diferentes criterios para resumir los datos pueden diferir considerablemente.





Medidas de resumen

- n datos: $x_1, x_2, \ldots, x_n$

Medidas de Posición

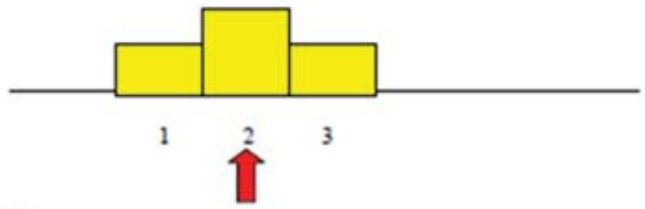
Media o Promedio Muestral

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

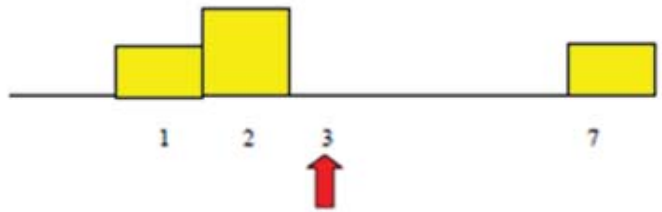
Media o Promedio Muestral

Es el punto de equilibrio del conjunto de datos.

X´s: 1, 2, 2, 3



X´s: 1, 2, 2, 7



Es una medida muy sensible a la presencia de datos anómalos (outliers).

Mediana y Media $\alpha-$podada Muestrales

- n datos: $x_1, x_2, \ldots, x_n$
- datos ordenados
  $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(k-1)} \leq x_{(k)} \leq x_{(k+1)} \leq \ldots \leq x_{(n)}$

Mediana Muestral

$$\tilde{x} = \begin{cases} x_{(k+1)} & \text{si } n = 2k+1 \\ \\ \frac{x_{(k)} + x_{(k+1)}}{2} & \text{si } n = 2k \end{cases}$$

Mediana y Media $\alpha-$podada Muestrales

- n datos: $x_1, x_2, \ldots, x_n$
- datos ordenados
  $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(k-1)} \leq x_{(k)} \leq x_{(k+1)} \leq \ldots \leq x_{(n)}$

Mediana Muestral

$$\tilde{x} = \begin{cases} x_{(k+1)} & \text{si } n = 2k+1 \\ \frac{x_{(k)} + x_{(k+1)}}{2} & \text{si } n = 2k \end{cases}$$

Media $\alpha-$podada Muestral

$$\overline{x}_\alpha = \frac{x_{([n\alpha])} + \ldots + x_{(n-[n\alpha])}}{(n - 2[n\alpha])}$$

Ejemplo de Juguete

- $n = 10$ datos ordenados: $2, 5, 8, 10, 14, 17, 21, 25, 28, 40$

Media Muestral:

$$\overline{x} = \frac{170}{10} = 17$$

Mediana Muestral: $10 = 2 \times 5 \Rightarrow k = 5$

$$\tilde{x} = \frac{x_{(5)} + x_{(6)}}{2} = 15.5$$

Media 10%podada Muestral

$$\overline{x}_{10} = \frac{x_{(2)} + \ldots + x_{(9)}}{8} = 16$$

Ejemplo de Juguete

- $n = 10$ datos ordenados: $2, 5, 8, 10, 14, 17, 21, 25, 28, 40$

Media Muestral:

$$\overline{x} = \frac{170}{10} = 17$$

Mediana Muestral: $10 = 2 \times 5 \Rightarrow k = 5$

$$\tilde{x} = \frac{x_{(5)} + x_{(6)}}{2} = 15.5$$

Media 10%podada Muestral

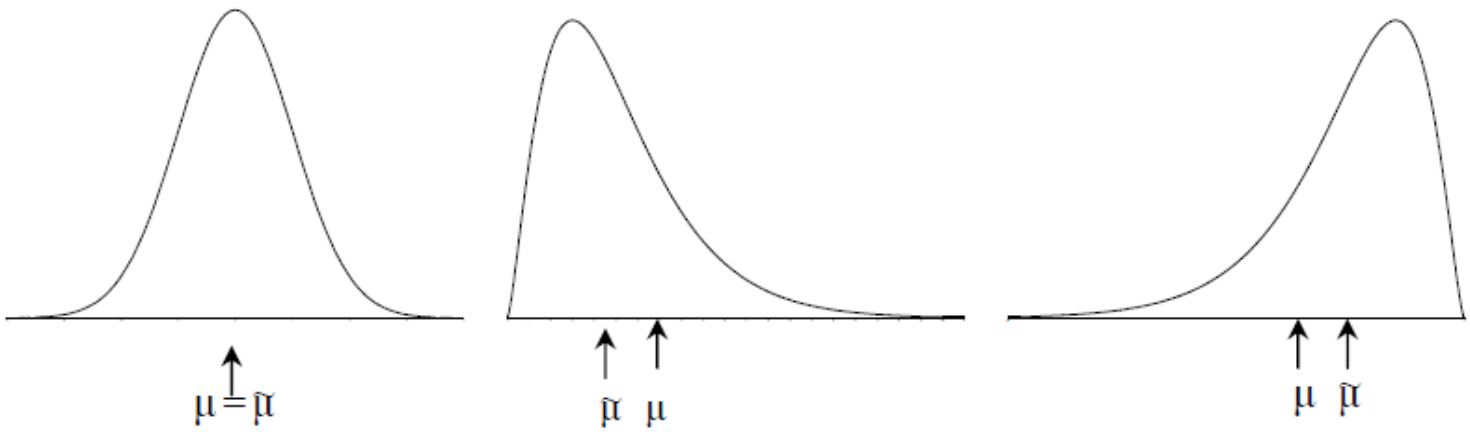
$$\overline{x}_{10} = \frac{x_{(2)} + \ldots + x_{(9)}}{8} = 16$$

¿Qué pasa si reemplazamos a 40 por 400?

$$\overline{x} = 53 \ , \ \tilde{x} = 15.5, \ \overline{x}_{10} = 16$$

Medidas de Posición en la Población

Medidas de Posición o Centrado



Medidas de resumen: Percentiles

- n datos: $x_1, x_2, \ldots, x_n$
- datos ordenados $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ - en R: sort(datos)

El $\alpha \times 100\%$ percentil de la muestra es el valor debajo del cual se halla el $\alpha \times 100\%$ de los datos ordenados.

Hay distintas formas de calcularlos. Una de ellas es la siguiente:

- Ordenamos los n datos de menor a mayor.
- Hallamos el dato que ocupa la posición $\alpha(n+1)$ en la muestra ordenada. Si este valor no es entero, pueden interpolarse los datos que están en las dos posiciones adyacentes.

Percentiles Muestrales

Ejemplo: Tenemos 19 datos que ordenados son

1 1 2 2 3 4 4 5 5 6 7 7 8 8 9 9 10 10 11

Percentil	Posición	Valor	
10%	0.10 (19+1) = **2**	1	
25%	0.25 (19+1) = **5**	3	**Cuartil Inferior**
50%	0.50 (19+1) = **10**	6	**Mediana**
75%	0.75(19+1) = **15**	9	**Cuartil Superior**
95%	0.95(19+1) = **19**	11	

- Notemos que el percentil 50% o segundo cuartil coincide con la mediana. Denotaremos Q_1 al primer cuartil (25%) y Q_3 al tercer cuartil (75%).
- Los cuartiles y la mediana dividen a la muestra en cuatro partes igualmente pobladas: 25% de la muestra en cada una de ellas.
- Entre Q_1 y Q_3 se halla el 50% central de los datos.

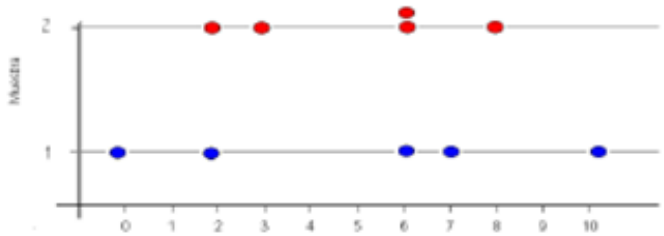
Medidas de Dispersión

Medidas de Dispersión o Variabilidad:

¿Cuán dispersos están los datos? ¿Cuán cercanos son los datos al valor típico?

Supongamos que tenemos datos $x_1, x_2, \ldots\ldots, x_n$

X's: 0 2 6 7 10
Y's: 2 3 6 6 8



$$\overline{X} = \overline{Y} = 5$$
$$\tilde{X} = \tilde{Y} = 6$$

¿Cómo medir la diferencia que se observa entre ambas muestras?

Medidas de Dispersión

Rango Muestral

$$rango = x_{(n)} - x_{(1)}$$

Varianza Muestral

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

Desvío Standard Muestral

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

Mediana de Desvíos Absolutos

$$\mathsf{mad} = 1.4826\ \mathsf{med}_{1 \leq i \leq n}(|x_i - \tilde{x}|)$$

Distancia Intercuartil

$$d_I = Q_3 - Q_1$$

Medidas de Dispersión

Bajo normalidad, tenemos que

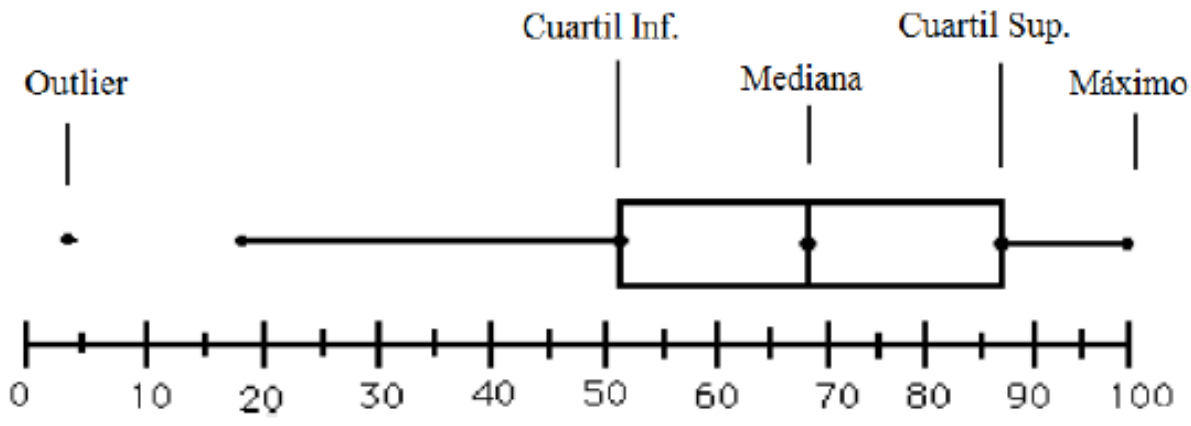
$$\mathsf{s} = \mathsf{mad} = d_I/1.349$$

Números de resumen

Resultan muy útiles para describir la muestra las siguientes medidas conocidos como Números de resumen:

1. Mínimo
2. Q_1 : Cuartil Inferior
3. Mediana
4. Q_3 : Cuartil Superior
5. Máximo

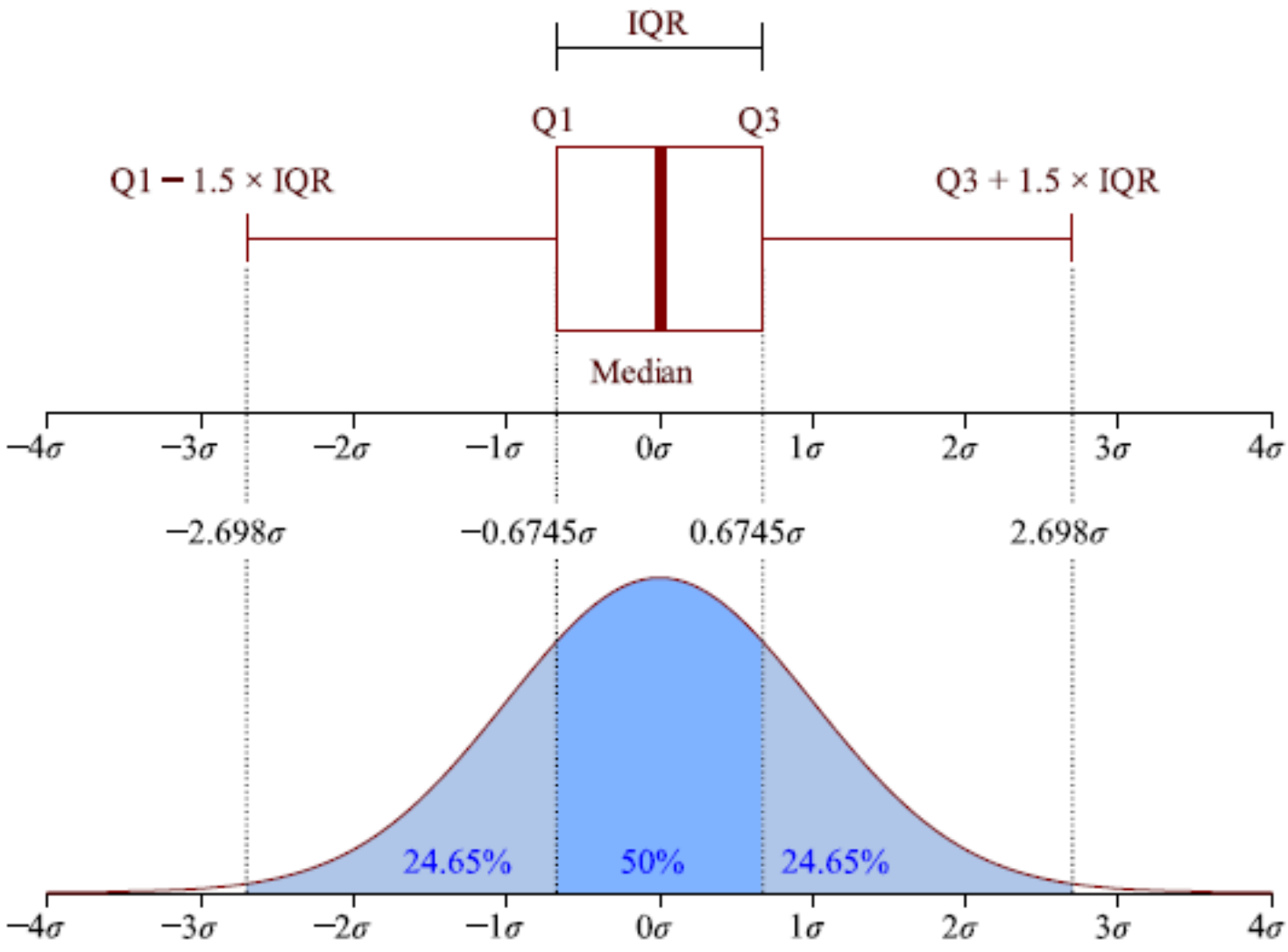
Boxplot


Outlier
Cuartil Inf.
Mediana
Cuartil Sup.
Máximo
0
10
20
30
40
50
60
70
80
90
100

Boxplot

1. Representamos una escala vertical u horizontal.
2. Dibujamos una caja cuyos extremos son los cuartiles y dentro de ella un segmento que corresponde a la mediana.
3. A partir de cada extremo dibujamos un segmento hasta el dato más alejado que está a lo sumo 1.5 d_I del extremo de la caja. Estos segmentos se llaman bigotes.
4. Marcamos con * a aquellos datos que están a más de $1.5d_I$ de cada extremo de la caja.

Datos normales

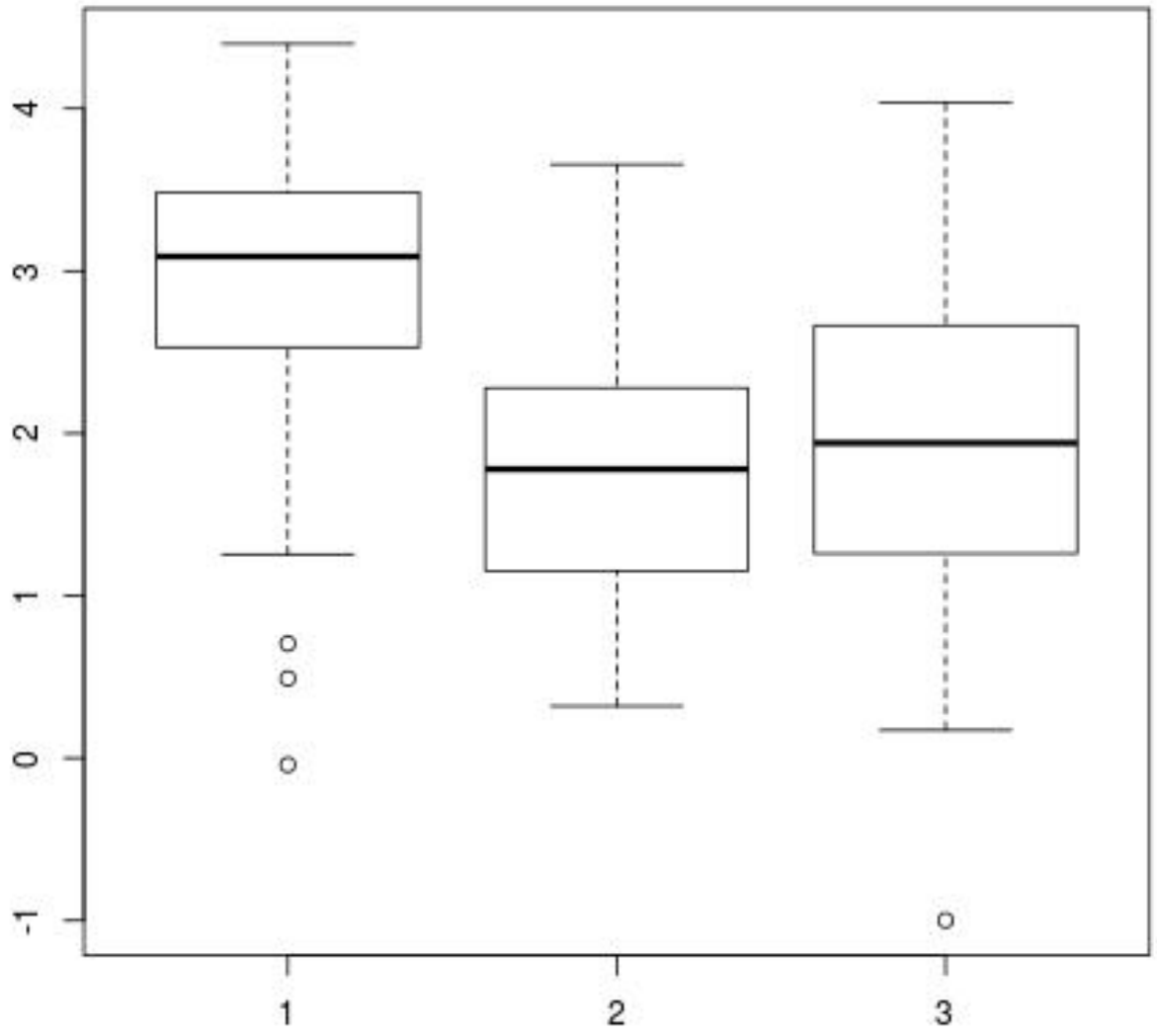


¿Qué vemos en un box-plot?

1. Posición
2. Dispersión
3. Asimetría
4. Longitud de las colas
5. Puntos anómalos (outliers)

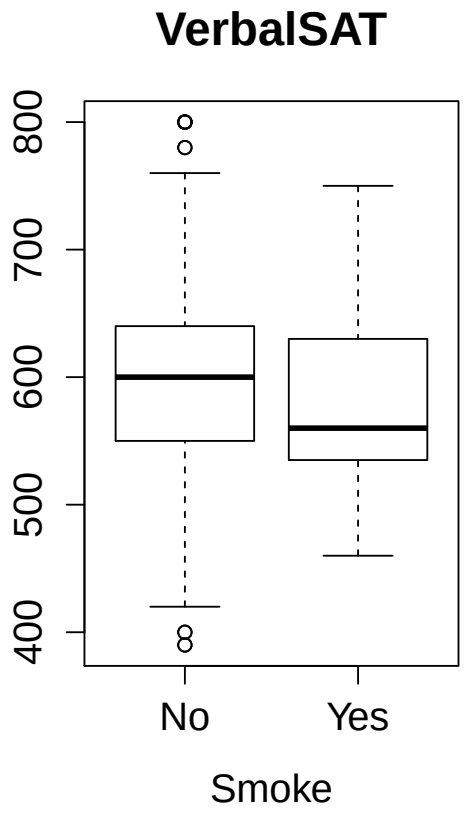
Los boxplots son muy útiles para comparar varios conjuntos de datos, pues nos dan una rápida impresión visual de sus características.

Boxplot Paralelos

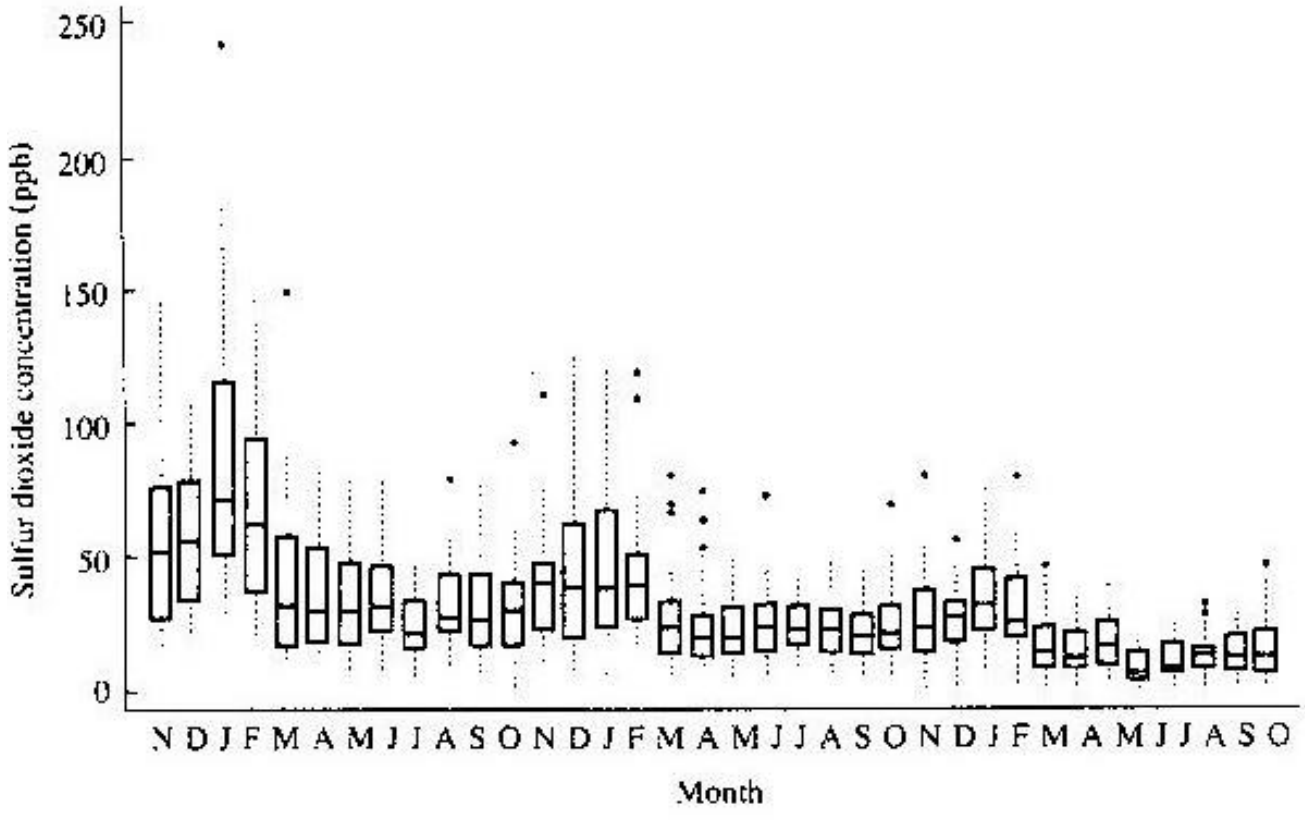


VerbalSAT

Realizamos dos bxp paralelos para VerbalSAT clasificando a los estudiantes de acuerdo con su condición de fumador.



Los datos graficados corresponden a las máximas concentraciones diarias (en partes por mil millones) de dióxido de azufre en Bayonne desde noviembre de 1969 hasta octubre de 1972, agrupadas por mes. Los boxplots se realizaron en base a los 36 grupos (meses) de 30 mediciones cada uno.



Bagplots

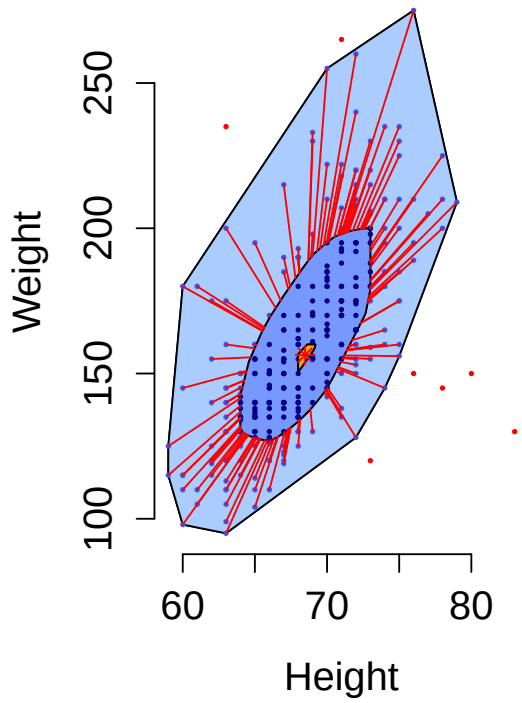
El boxplot tal como lo conocemos está fuertemente ligado a la distribución normal.

Existen otras opciones, tal como el adjbox del paquete robustbase, introducido por Vandervieren y Hubert (2004), diseñado para distribuciones fuertemente asimétricas (library(robustbase)).

Existen extensiones del boxplot a tipos de datos con estructura más complejos, tales como datos bivariados o datos funcionales.
En el paquete alplack el comando bagplot(), que generaliza el boxplot usando el concepto de profundidad a datos bivariados y que permite visualizar cómo varían conjuntamente dos variables e identificar outliers en el espacio bidimensional.

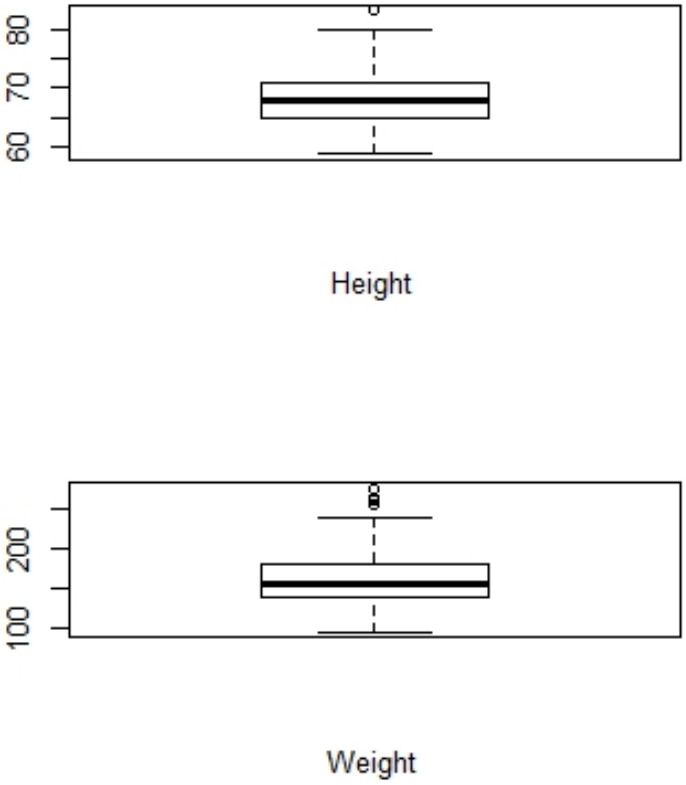
Bagplots (TP 28)

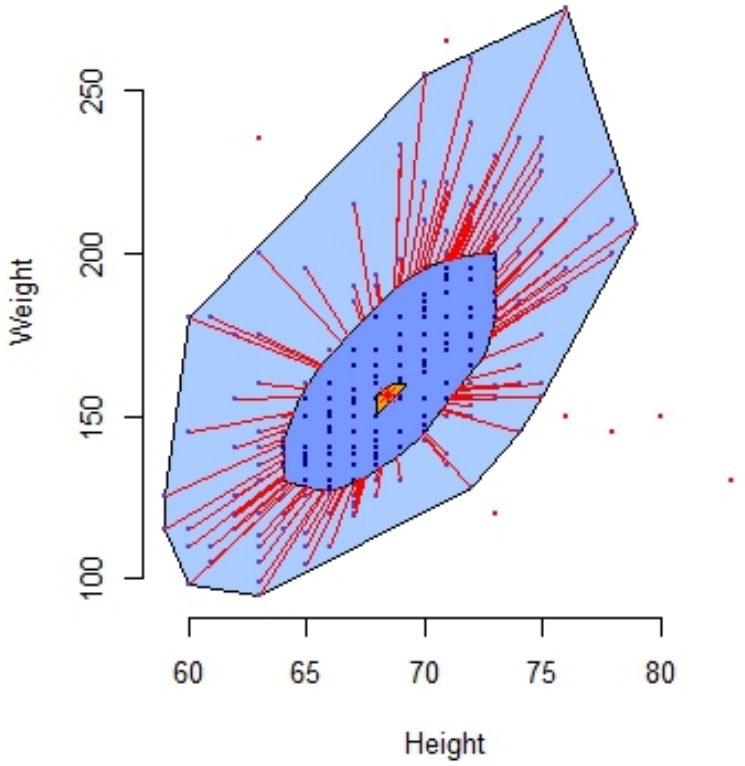
Graficamos el bagplot de Height y Weight.



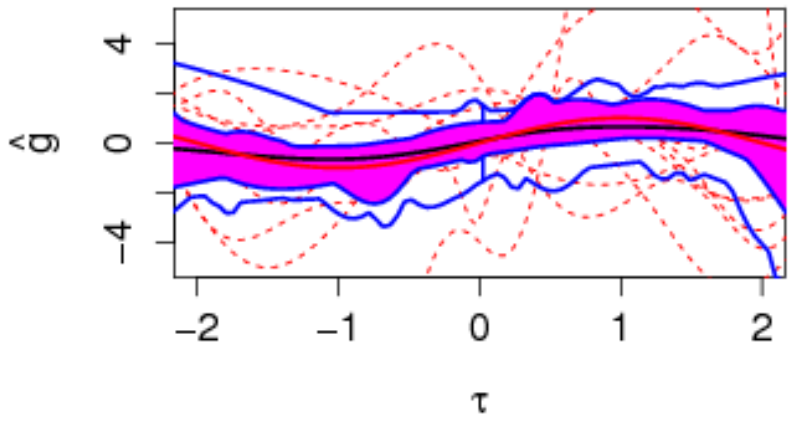
Boxplots y Bagplots

Graficamos los boxplots de Height y Weight y su bagplot.





Boxplots para datos funcionales



Referencias

- Lock, R., Lock, P., Frazer Lock , K., Lock Morgan, E. and Lock, D. (2013). Statistics: unlocking the power of the data, Wiley.
- Paradis, E. (2003). R para principiantes. Disponible en $http://cran.r-project.org/doc/contrib/rdebuts\_es.pdf$
- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.
- Vasishth, A. (2014). An introduction to statistical data analysis (Summer 2014) Lecture note. Disponible en $http://www.ling.uni-potsdam.de/\sim vasishth/$
- .... y muchos más