Estadística Descriptiva con R

Antonio Falcó

Lectura 1

Estadística (Wikipedia)

La estadística es una ciencia que estudia la recolección, análisis e interpretación de datos, ya sea para ayudar en la toma de decisiones o para explicar condiciones regulares o irregulares de algún fenómeno o estudio aplicado, de ocurrencia en forma aleatoria o condicional. Sin embargo estadística es más que eso, en otras palabras es el vehículo que permite llevar a cabo el proceso relacionado con la investigación científica.

Estadística (Wikipedia)

La estadística es una ciencia que estudia la recolección, análisis e interpretación de datos, ya sea para ayudar en la toma de decisiones o para explicar condiciones regulares o irregulares de algún fenómeno o estudio aplicado, de ocurrencia en forma aleatoria o condicional. Sin embargo estadística es más que eso, en otras palabras es el vehículo que permite llevar a cabo el proceso relacionado con la investigación científica.

Estadística

La estadística es la ciencia que se ocupa del estudio de los fenómenos aleatorios basado sobre la hipótesis del conocimiento de un número relativamente limitado de observaciones.

Estadística (Wikipedia)

La estadística es una ciencia que estudia la recolección, análisis e interpretación de datos, ya sea para ayudar en la toma de decisiones o para explicar condiciones regulares o irregulares de algún fenómeno o estudio aplicado, de ocurrencia en forma aleatoria o condicional. Sin embargo estadística es más que eso, en otras palabras es el vehículo que permite llevar a cabo el proceso relacionado con la investigación científica.

Estadística

La estadística es la ciencia que se ocupa del estudio de los fenómenos aleatorios basado sobre la hipótesis del conocimiento de un número relativamente limitado de observaciones.

Estadística Matemática

Se centra, fundamentalmente, en el desarrollo de nuevos métodos para la inferencia estadístic (requiere conocimentos avanzados de matemáticas).

Estadística Matemática

Se centra, fundamentalmente, en el desarrollo de nuevos métodos para la inferencia estadístic (requiere conocimentos avanzados de matemáticas).

Estadística Aplicada

Se dedica al empleo de los métodos de la Estadística Matemática sobre áreas específicas: (Economía (Econometría), Psicología (Psicometría) y Ciencias de la Salud (Bioestadística)).

Estadística Matemática

Se centra, fundamentalmente, en el desarrollo de nuevos métodos para la inferencia estadístic (requiere conocimentos avanzados de matemáticas).

Estadística Aplicada

Se dedica al empleo de los métodos de la Estadística Matemática sobre áreas específicas: (Economía (Econometría), Psicología (Psicometría) y Ciencias de la Salud (Bioestadística)).

Bioestadística

Es la rama de la Estadística Aplicada que emples métodos estadísticos para el estudio de problemas biológicos.

Ejemplos de aplicaciones

1. En odontología la recesión gingival es un problema tanto para los pacientes como para los terapeutas. Se realizó un estudio clínico para evaluar y comparar los efectos de un procedimento de regeneración de tejidos en el tratamiento de los defectos de recesión gingival.

Ejemplos de aplicaciones

1. En odontología la recesión gingival es un problema tanto para los pacientes como para los terapeutas. Se realizó un estudio clínico para evaluar y comparar los efectos de un procedimento de regeneración de tejidos en el tratamiento de los defectos de recesión gingival.
2. Dental researchers conducted a study to evaluate relevant variables that may assist in identifying orthodontic patients with signs and symptoms associated with sleep apnea and to estimate the proportion of potential sleep apnea patients whose ages range from 8 to 15 years.

La mayor parte de las investigaciones científicas se llevan a cabo empleando los pasos siguientes:

La mayor parte de las investigaciones científicas se llevan a cabo empleando los pasos siguientes:

Etapas

1. Formulación del problema de investigación.

La mayor parte de las investigaciones científicas se llevan a cabo empleando los pasos siguientes:

Etapas

1. Formulación del problema de investigación.
2. Identificación de las variables clave.

La mayor parte de las investigaciones científicas se llevan a cabo empleando los pasos siguientes:

Etapas

1. Formulación del problema de investigación.
2. Identificación de las variables clave.
3. Diseño del experimento estadístico.

La mayor parte de las investigaciones científicas se llevan a cabo empleando los pasos siguientes:

Etapas

1. Formulación del problema de investigación.
2. Identificación de las variables clave.
3. Diseño del experimento estadístico.
4. Recolección de datos.

La mayor parte de las investigaciones científicas se llevan a cabo empleando los pasos siguientes:

Etapas

1. Formulación del problema de investigación.
2. Identificación de las variables clave.
3. Diseño del experimento estadístico.
4. Recolección de datos.
5. Análisis estadístico de los datos

La mayor parte de las investigaciones científicas se llevan a cabo empleando los pasos siguientes:

Etapas

1. Formulación del problema de investigación.
2. Identificación de las variables clave.
3. Diseño del experimento estadístico.
4. Recolección de datos.
5. Análisis estadístico de los datos
6. Interpretación analítica de los resultados.

Un caso de estudio

Formulación del problema de investigación

Deseamos testar la efeciacia un nuevo dispositivo para medir la presión sanguínea, ya que será distribuido para el uso en establecimientos públicos.

Un caso de estudio

Formulación del problema de investigación

Deseamos testar la efeciacia un nuevo dispositivo para medir la presión sanguínea, ya que será distribuido para el uso en establecimientos públicos.

Identificación de la variables clave

Presión sistólica en sangle (mm Hg).

Diseño del experimento estadístico

Se medirá a cada sujeto su presión sanguínea tanto empleando el dispositivo como manualmente.

Cuestiones

1. ¿Cuántos dispositivos emplearemos en el experimento ?

Diseño del experimento estadístico

Se medirá a cada sujeto su presión sanguínea tanto empleando el dispositivo como manualmente.

Cuestiones

1. ¿Cuántos dispositivos emplearemos en el experimento ?
   - Debido a que no estamos seguros acerca de la comparabilidad de los diferentes dispositivos emplearemos un número de cuatro.

Diseño del experimento estadístico

Se medirá a cada sujeto su presión sanguínea tanto empleando el dispositivo como manualmente.

Cuestiones

1. ¿Cuántos dispositivos emplearemos en el experimento ?
   - Debido a que no estamos seguros acerca de la comparabilidad de los diferentes dispositivos emplearemos un número de cuatro.
2. ¿Cuantas muestras tomaremos por dispositivo?

Diseño del experimento estadístico

Se medirá a cada sujeto su presión sanguínea tanto empleando el dispositivo como manualmente.

Cuestiones

1. ¿Cuántos dispositivos emplearemos en el experimento ?
   - Debido a que no estamos seguros acerca de la comparabilidad de los diferentes dispositivos emplearemos un número de cuatro.
2. ¿Cuantas muestras tomaremos por dispositivo?
   - Empleando fórmulas estadísticas obtenemos un número de 100 muestras.

Diseño del experimento estadístico

Se medirá a cada sujeto su presión sanguínea tanto empleando el dispositivo como manualmente.

Cuestiones

1. ¿Cuántos dispositivos emplearemos en el experimento ?
   - Debido a que no estamos seguros acerca de la comparabilidad de los diferentes dispositivos emplearemos un número de cuatro.
2. ¿Cuantas muestras tomaremos por dispositivo?
   - Empleando fórmulas estadísticas obtenemos un número de 100 muestras.
3. ¿En que orden efectuaremos la toma de medidas?

Diseño del experimento estadístico

Se medirá a cada sujeto su presión sanguínea tanto empleando el dispositivo como manualmente.

Cuestiones

1. ¿Cuántos dispositivos emplearemos en el experimento ?
   - Debido a que no estamos seguros acerca de la comparabilidad de los diferentes dispositivos emplearemos un número de cuatro.
2. ¿Cuantas muestras tomaremos por dispositivo?
   - Empleando fórmulas estadísticas obtenemos un número de 100 muestras.
3. ¿En que orden efectuaremos la toma de medidas?
   - Se realizará de forma aleatoria.

Diseño de un experimento estadístico

Se medirá a cada sujeto su presión sanguínea tanto empleando el dispositivo como manualmente..

Questions

4 ¿Qué datos recogeremos en el cuestionario que pueda influenciar en la comparativa entre ambos métodos?

Diseño de un experimento estadístico

Se medirá a cada sujeto su presión sanguínea tanto empleando el dispositivo como manualmente..

Questions

4 ¿Qué datos recogeremos en el cuestionario que pueda influenciar en la comparativa entre ambos métodos?

- Preguntaremos acerca de la edad, sexo y sobre el historial previo sobre problemas de hipertensión.

Diseño de un experimento estadístico

Se medirá a cada sujeto su presión sanguínea tanto empleando el dispositivo como manualmente..

Questions

4. ¿Qué datos recogeremos en el cuestionario que pueda influenciar en la comparativa entre ambos métodos?
   - Preguntaremos acerca de la edad, sexo y sobre el historial previo sobre problemas de hipertensión.
5. ¿Cómo recolectaremos los datos para facilitar su posterior proceso de computación?

Diseño de un experimento estadístico

Se medirá a cada sujeto su presión sanguínea tanto empleando el dispositivo como manualmente..

Questions

4. ¿Qué datos recogeremos en el cuestionario que pueda influenciar en la comparativa entre ambos métodos?
   - Preguntaremos acerca de la edad, sexo y sobre el historial previo sobre problemas de hipertensión.
5. ¿Cómo recolectaremos los datos para facilitar su posterior proceso de computación?
   - Desarrollaremos una forma de codificación. A cada persona entrevistada se le asignará un número de identificación.

Diseño de un experimento estadístico

Se medirá a cada sujeto su presión sanguínea tanto empleando el dispositivo como manualmente..

Questions

4. ¿Qué datos recogeremos en el cuestionario que pueda influenciar en la comparativa entre ambos métodos?
   - Preguntaremos acerca de la edad, sexo y sobre el historial previo sobre problemas de hipertensión.
5. ¿Cómo recolectaremos los datos para facilitar su posterior proceso de computación?
   - Desarrollaremos una forma de codificación. A cada persona entrevistada se le asignará un número de identificación.
6. Cómo podemos chequear la precisión de los datos numéricos?

Diseño de un experimento estadístico

Se medirá a cada sujeto su presión sanguínea tanto empleando el dispositivo como manualmente..

Questions

4. ¿Qué datos recogeremos en el cuestionario que pueda influenciar en la comparativa entre ambos métodos?
   - Preguntaremos acerca de la edad, sexo y sobre el historial previo sobre problemas de hipertensión.
5. ¿Cómo recolectaremos los datos para facilitar su posterior proceso de computación?
   - Desarrollaremos una forma de codificación. A cada persona entrevistada se le asignará un número de identificación.
6. Cómo podemos chequear la precisión de los datos numéricos?
   - Analizaremos los rangos de cada una de las variables y destacaremos los posibles datos anómalos para un posterior análisis.

Datos

Mean blood pressures and differences between machine and human readings at four locations

		Systolic blood pressure (mm Hg)					
		Machine		Human		Difference	
Location	Number of people	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
A	98	142.5	21.0	142.0	18.1	0.5	11.2
B	84	134.1	22.5	133.6	23.2	0.5	12.1
C	98	147.9	20.3	133.9	18.3	14.0	11.7
D	62	135.4	16.7	128.5	19.0	6.9	13.6

Análisis estadístico de los datos

- Determinar cuando las diferencias en presión sanguínea medida manualmente y mediante el dispositivo en (C,D) fueron *reales* o *debidas al azar*. **Inferencia Estadística**

Análisis estadístico de los datos

- Determinar cuando las diferencias en presión sanguínea medida manualmente y mediante el dispositivo en (C,D) fueron *reales* o *debidas al azar*. **Inferencia Estadística**
- Estudio del error en las estimaciones.

Análisis estadístico de los datos

- Determinar cuando las diferencias en presión sanguínea medida manualmente y mediante el dispositivo en (C,D) fueron *reales* o *debidas al azar*. **Inferencia Estadística**
- Estudio del error en las estimaciones.
- Construcción de un modelo basado en probabilidades.

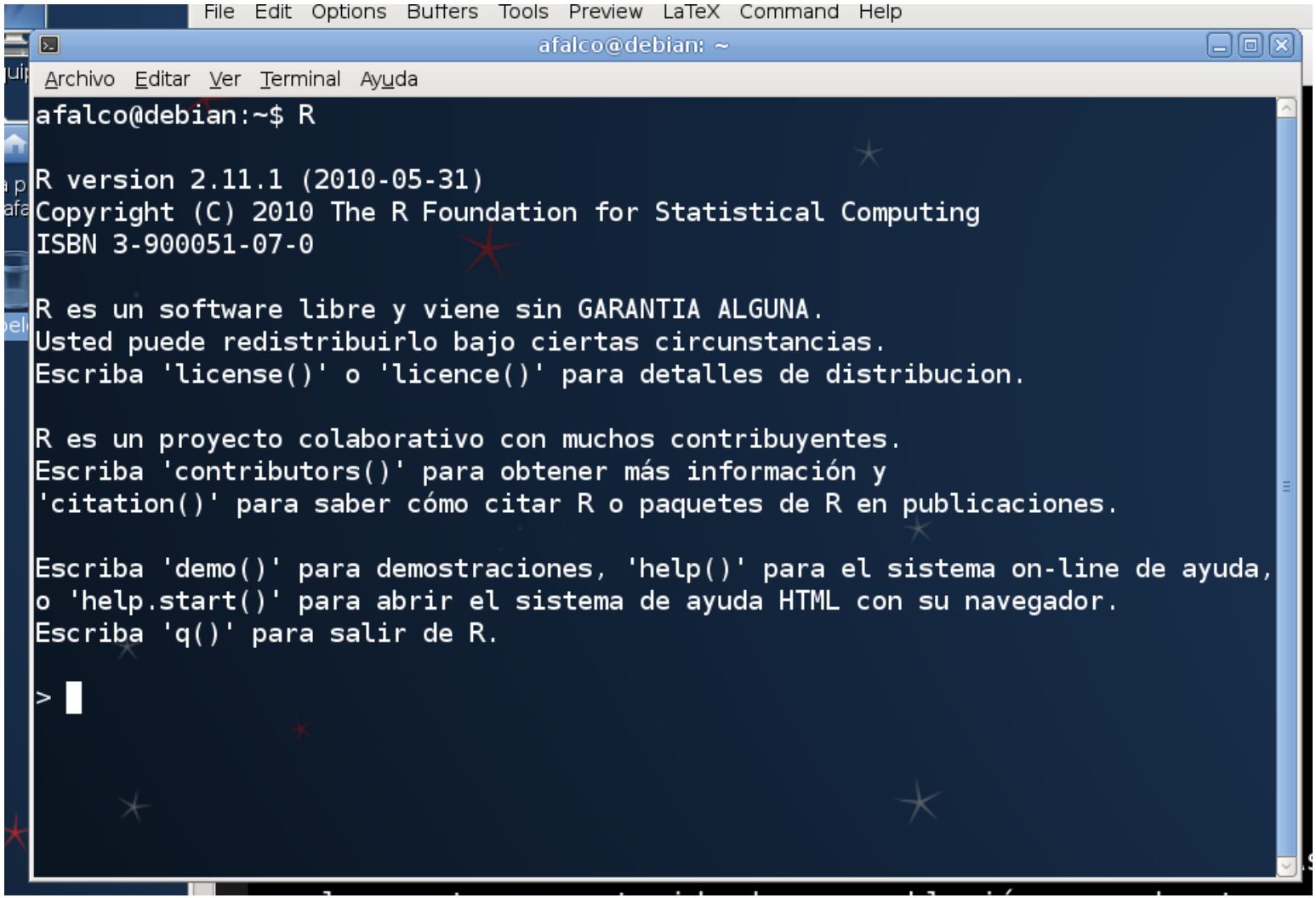
Análisis estadístico de los datos

- Determinar cuando las diferencias en presión sanguínea medida manualmente y mediante el dispositivo en (C,D) fueron *reales* o *debidas al azar*. **Inferencia Estadística**
- Estudio del error en las estimaciones.
- Construcción de un modelo basado en probabilidades.

y finalmente **Interpretar analíticamente los resultados**.

R (lenguaje de programación)

Se trata de un proyecto de software libre, resultado de la implementación GNU del premiado lenguaje S. R y S-Plus -versión comercial de S- son, probablemente, los dos lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy populares en el campo de la investigación biomédica, la bioinformática y las matemáticas financieras. A esto contribuye la posibilidad de cargar diferentes bibliotecas o paquetes con finalidades específicas de cálculo o gráfico. R se distribuye bajo la licencia GNU GPL y está disponible para los sistemas operativos Windows, Macintosh, Unix y GNU/Linux.


File Edit Options Buffers Tools Preview LaTeX Command Help
afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
afalco@debian:~$ R
R version 2.11.1 (2010-05-31)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
>

Cuestión

Consideremos una colección de datos de caracter numérico: $x_1, \ldots, x_n$ *donde* x_1 *denota el primer dato muestral y* x_n *denota el n-ésimo dato muestral. Asumendo que la muestra es extraida de una población, que denotaremos por* P*, que podemos inferir o concluir sobre* P *empleando los datos muestrales ?*

Cuestión

Consideremos una colección de datos de caracter numérico: $x_1, \ldots, x_n$ *donde* x_1 *denota el primer dato muestral y* x_n *denota el n-ésimo dato muestral. Asumendo que la muestra es extraida de una población, que denotaremos por* P*, que podemos inferir o concluir sobre* P *empleando los datos muestrales ?*

Antes de responder los datos deben de resumirse empleando las llamadas medidas estadísticas de localización.

Cuestión

Consideremos una colección de datos de caracter numérico: $x_1, \ldots, x_n$ *donde* x_1 *denota el primer dato muestral y* x_n *denota el n-ésimo dato muestral. Asumendo que la muestra es extraida de una población, que denotaremos por* P*, que podemos inferir o concluir sobre* P *empleando los datos muestrales ?*

Antes de responder los datos deben de resumirse empleando las llamadas medidas estadísticas de localización.

Definition

Media Aritmética

$$\overline{x} := \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1}{n} \times \left(x_1 + x_2 + \cdots + x_{n-1} + x_n\right).$$

Example

Consideremos una muestra de datos $\{2, 5, 0, -4\}$. Entonces $n = 4$ y

$$x_1 = 2,\ x_2 = 5,\ x_3 = 0 \text{ and } x_4 = -4.$$

Ahora,

$$\overline{x} = \frac{1}{4} \times (2 + 5 + 0 + (-4)) = \frac{1}{4} \times (2 + 5 + 0 - 4) = \frac{1}{4} \times 3 = \frac{3}{4} = 0.75.$$

Example

Consideremos una muestra de datos $\{2, 5, 0, -4\}$. Entonces $n = 4$ y

$$x_1 = 2,\ x_2 = 5,\ x_3 = 0 \text{ and } x_4 = -4.$$

Ahora,

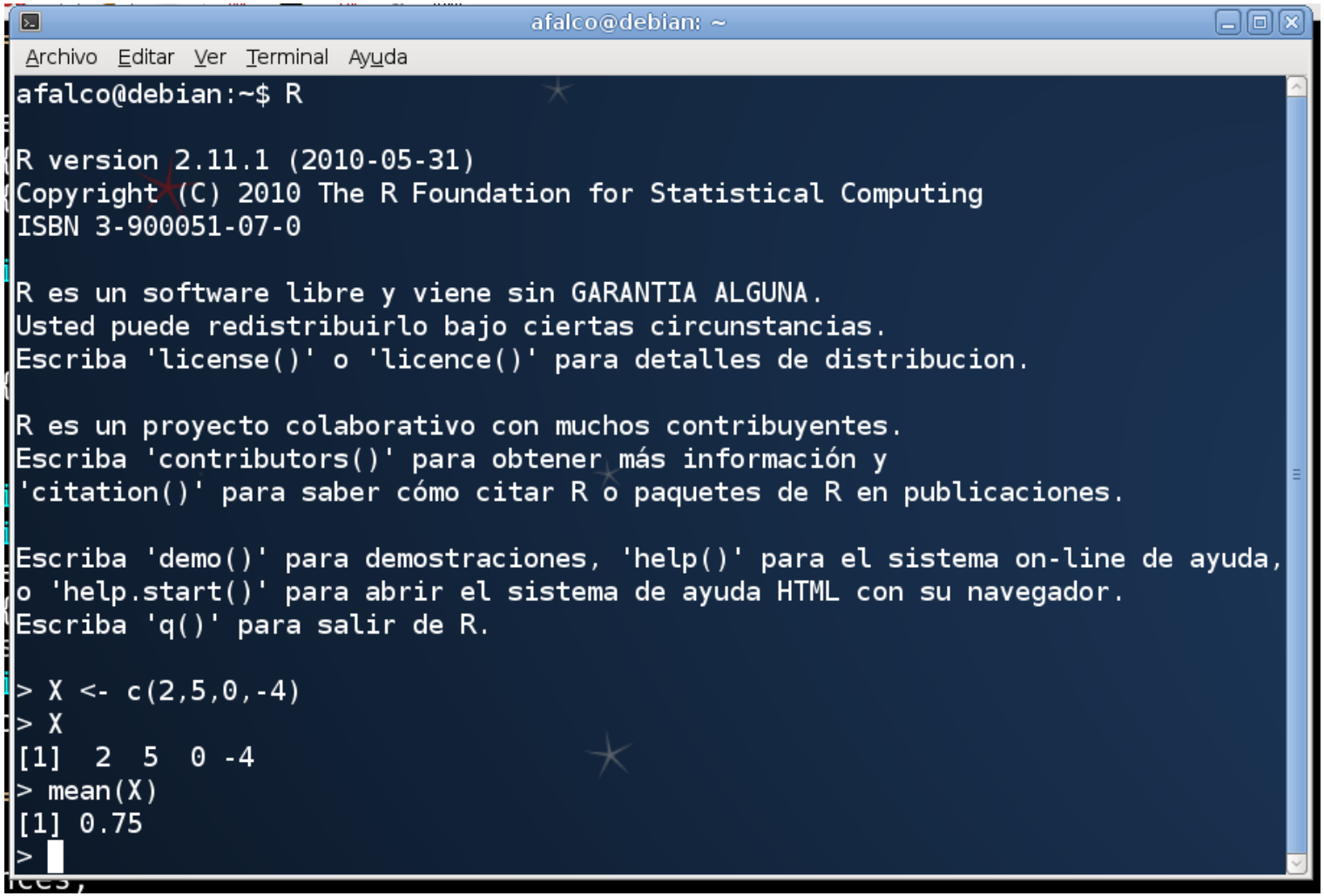
$$\overline{x} = \frac{1}{4} \times (2 + 5 + 0 + (-4)) = \frac{1}{4} \times (2 + 5 + 0 - 4) = \frac{1}{4} \times 3 = \frac{3}{4} = 0.75.$$

Observación

Observése que la medida es independiente de la ordenación de los datos. Si reordenamos la muestra

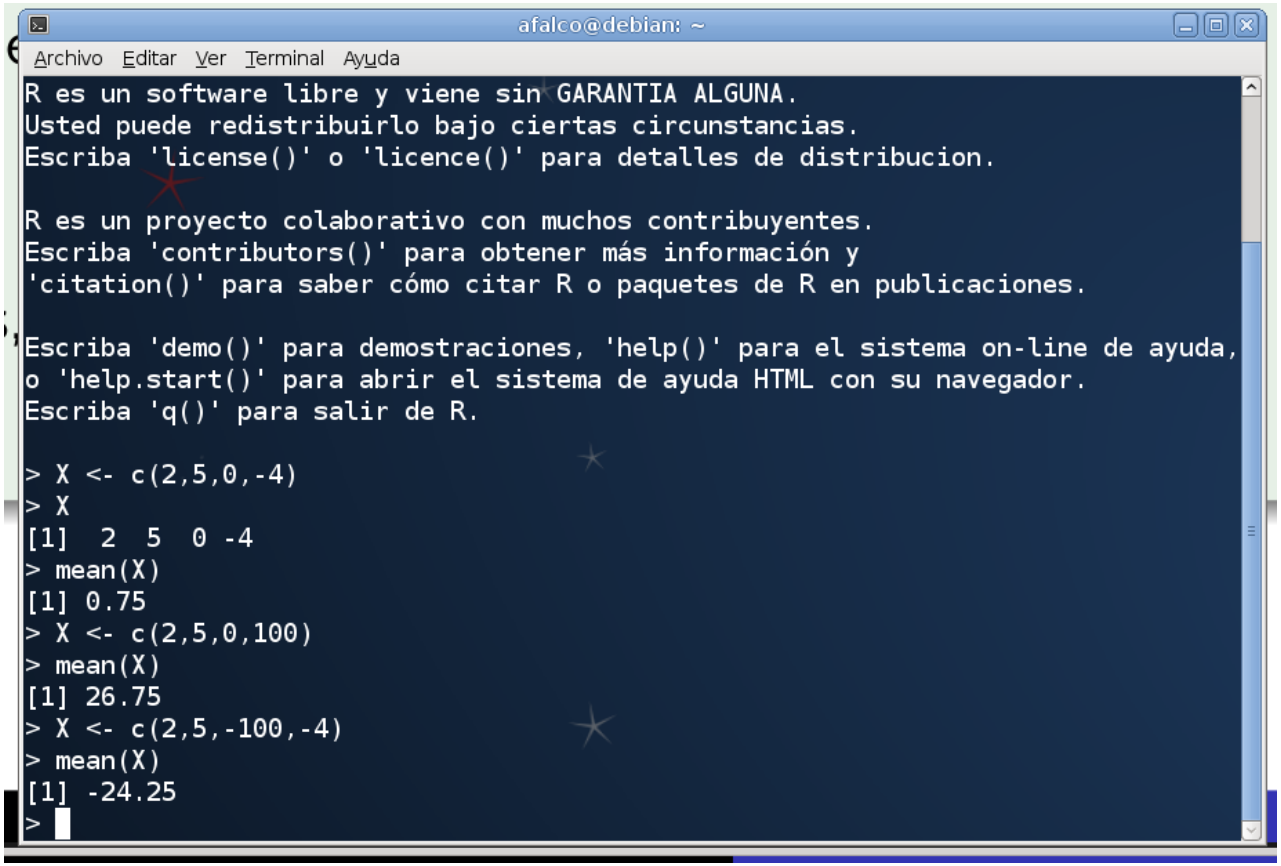
$$x_1 = -4,\ x_2 = 0,\ x_3 = 2 \textit{ and } x_4 = 5.$$

La media aritmética $\overline{x} = 0.75$ *permanece invariante.*


afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
afalco@debian:~$ R
R version 2.11.1 (2010-05-31)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
> X <- c(2,5,0,-4)
> X
[1] 2 5 0 -4
> mean(X)
[1] 0.75
>

Afirmación

La media aritmética es sensible a los valores extremos.



Definition

Si las observaciones muestrales esta ordenadas, la *mediana muestral* viene dada por:

$$\begin{cases} x_{(n+1)/2} & \text{if} \quad n \text{ es impar}, \\ \frac{x_{n/2}+x_{(n/2)+1}}{2} & \text{if} \quad n \text{ es par}, \end{cases}$$

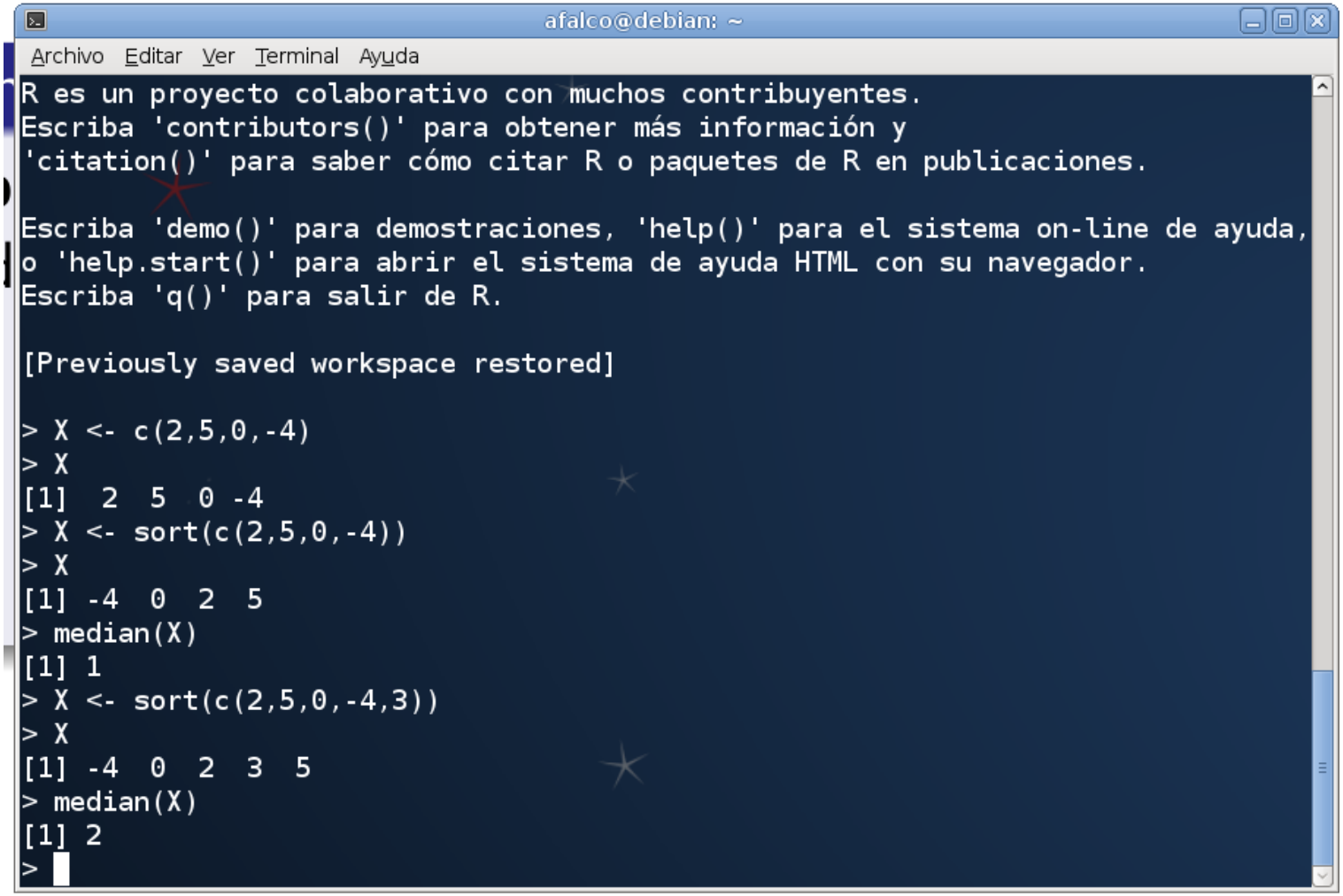
Definition

Si las observaciones muestrales esta ordenadas, la *mediana muestral* viene dada por:

$$\begin{cases} x_{(n+1)/2} & \text{if} \quad n \text{ es impar,} \\ \frac{x_{n/2}+x_{(n/2)+1}}{2} & \text{if} \quad n \text{ es par,} \end{cases}$$

Example

- Para $\{x_1 = -4, x_2 = 0, x_3 = 2\}$ como $n = 3$ calculamos el índice mediante $\frac{n+1}{2} = \frac{3+1}{2} = 2$ y la mediana muestral es $x_2 = 0$.
- Para $\{x_1 = -4, x_2 = 0, x_3 = 2, x_4 = 5\}$ como $n = 4$ calculamos los ínidces $\frac{n}{2} = \frac{4}{2} = 2$ and $\frac{n}{2} + 1 = 3$. En este caso la mediana muestral es $\frac{x_2+x_3}{2} = \frac{0+2}{2} = 1$.


afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
[Previously saved workspace restored]
> X <- c(2,5,0,-4)
> X
[1] 2 5 0 -4
> X <- sort(c(2,5,0,-4))
> X
[1] -4 0 2 5
> median(X)
[1] 1
> X <- sort(c(2,5,0,-4,3))
> X
[1] -4 0 2 3 5
> median(X)
[1] 2
>

Definition

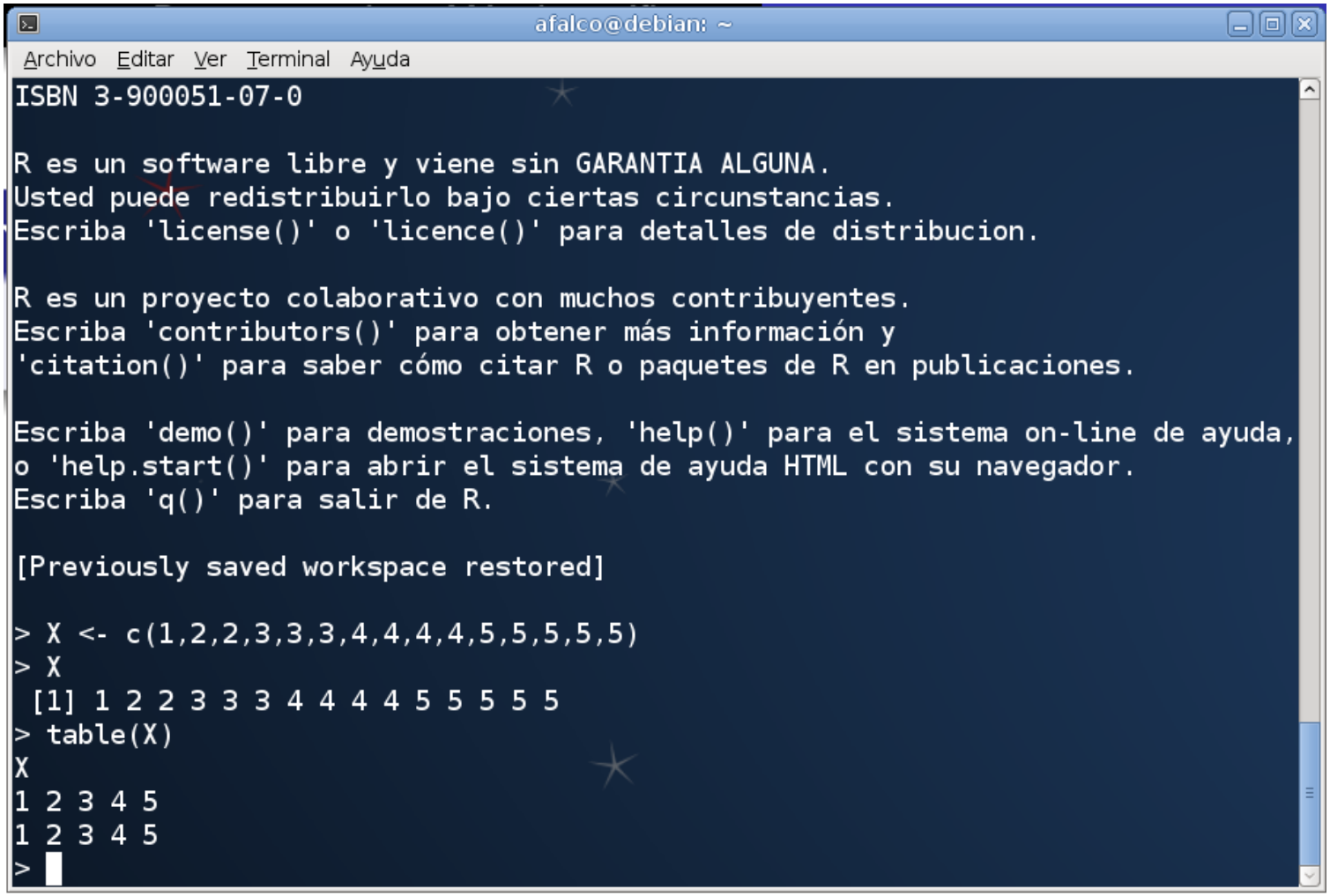
La *moda* es el valor que aparece con mayor frecuencia en la muestra

Example

Sample of time intervals between successive menstrual periods (days) in college-age women

Value	Frequency	Value	Frequency	Value	Frequency
24	5	29	96	34	7
25	10	30	63	35	3
26	28	31	24	36	2
27	64	32	9	37	1
28	185	33	2	38	1

frecuencia es el número de veces que aparece un valor fijado en una muestra dada. Aquí la moda es 28


afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
[Previously saved workspace restored]
> X <- c(1,2,2,3,3,3,4,4,4,4,5,5,5,5,5)
> X
[1] 1 2 2 3 3 3 4 4 4 4 5 5 5 5 5
> table(X)
X
1 2 3 4 5
1 2 3 4 5
>

Notación

A partir de ahora:

$$a \times b = a \cdot b = ab$$

Definition

La *media geométrica* (en realidad la media aritmética en escala logarítmica) se define como:

$$\overline{x}_G = \overline{\log x} = \left(x_1 \cdot x_2 \cdots x_{n-1} \cdot x_n\right)^{1/n}.$$

En la práctica se calcula: $\overline{x}_G = \overline{\log x} = 10^{\left(\frac{1}{n}\sum_{i=1}^{n} \log x_i\right)}$

Definition

La *media geométrica* (en realidad la media aritmética en escala logarítmica) se define como:

$$\overline{x}_G = \overline{\log x} = \left(x_1 \cdot x_2 \cdots x_{n-1} \cdot x_n\right)^{1/n}.$$

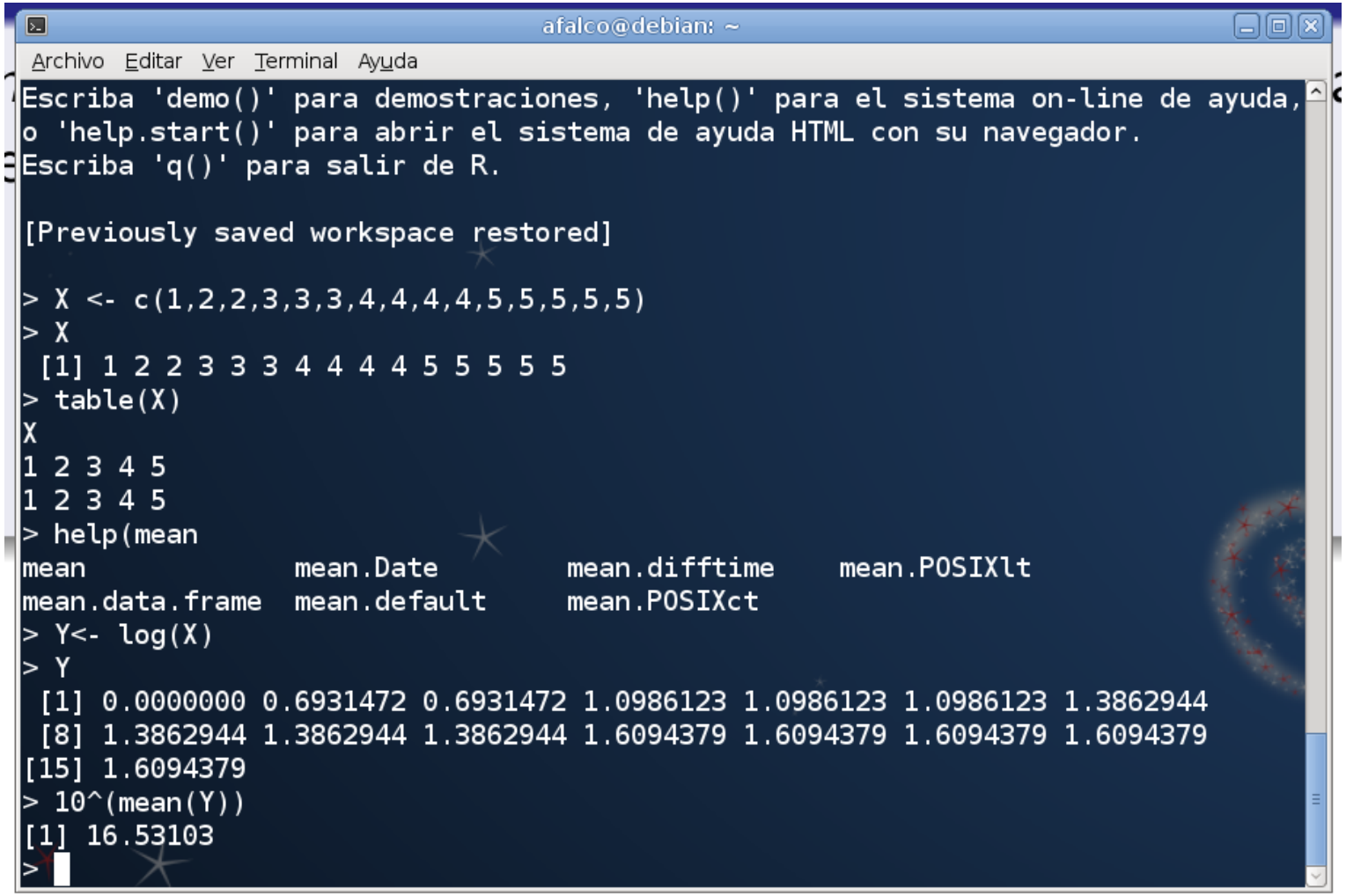
En la práctica se calcula: $\overline{x}_G = \overline{\log x} = 10^{\left(\frac{1}{n}\sum_{i=1}^{n} \log x_i\right)}$

Example

Supongamos $\{x_1 = 10, x_2 = 100, x_3 = 1000\}$ entonces $\{\log x_1 = 1, \log x_2 = 2, \log x_3 = 3\}$. Ahora,

$$\frac{1}{n}\sum_{i=1}^{n} \log x_i = \frac{1}{3}(1+2+3) = \frac{6}{3} = 2.$$

Finalmente, $\overline{x}_G = \overline{\log x} = 10^2 = 10 \cdot 10 = 100$.


afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
[Previously saved workspace restored]
> X <- c(1,2,2,3,3,3,4,4,4,4,5,5,5,5,5)
> X
[1] 1 2 2 3 3 3 4 4 4 4 5 5 5 5 5
> table(X)
X
1 2 3 4 5
1 2 3 4 5
> help(mean
mean mean.Date mean.difftime mean.POSIXlt
mean.data.frame mean.default mean.POSIXct
> Y<- log(X)
> Y
[1] 0.0000000 0.6931472 0.6931472 1.0986123 1.0986123 1.0986123 1.3862944
[8] 1.3862944 1.3862944 1.3862944 1.6094379 1.6094379 1.6094379 1.6094379
[15] 1.6094379
> 10^(mean(Y))
[1] 16.53103
>

Proposición

Supongamos que añadimos una constante c al valor de cada dato muestral, esto es, set $y_i := x_i + c$ para cada $i = 1, 2, \ldots, n$. Entonces

$$\overline{y} = \overline{x} + c$$

Proposición

Supongamos que añadimos una constante c al valor de cada dato muestral, esto es, set $y_i := x_i + c$ para cada $i = 1, 2, \ldots, n$. Entonces

$$\overline{y} = \overline{x} + c$$

Proof.

Observése que

$$\begin{aligned} \overline{y} &= \frac{1}{n}(y_1 + \cdots + y_n) = \frac{1}{n}((x_1 + c) + \cdots + (x_n + c)) && (1) \\ &= \frac{1}{n}((x_1 + \cdots + x_n) + (n \cdot c)) = \frac{1}{n}(x_1 + \cdots + x_n) + \frac{1}{n}(n \cdot c) && (2) \\ &= \overline{x} + c && (3) \end{aligned}$$

□

En la práctica

El anterior pricipio es útil ya que algunas veces necesitamos "desplazar" el origen de los datos muestrales.

En la práctica

El anterior pricipio es útil ya que algunas veces necesitamos "desplazar" el origen de los datos muestrales.

Example

Tomemos $c = 28$

Translated sample for the duration between successive menstrual periods in college-age women

Value	Frequency	Value	Frequency	Value	Frequency
−4	5	1	96	6	7
−3	10	2	63	7	3
−2	28	3	24	8	2
−1	64	4	9	9	1
0	185	5	2	10	1

Note: $\bar{y} = [(-4)(5) + (-3)(10) + \ldots + (10)(1)] / 500 = 0.54$

$\bar{x} = \bar{y} + 28 = 0.54 + 28 = 28.54$ days

Proposición

Supongamos que reescalamos la muestra empleando una constante c, esto es, fijemos $y_i := c \cdot x_i$ para cada $i = 1, 2, \ldots, n$. Entonces

$$\overline{y} = c \cdot \overline{x}$$

Proposición

Supongamos que reescalamos la muestra empleando una constante c, esto es, fijemos $y_i := c \cdot x_i$ para cada $i = 1, 2, \ldots, n$. Entonces

$$\overline{y} = c \cdot \overline{x}$$

Proof.

Claramente

$$\overline{y} = \frac{1}{n}(y_1 + \cdots + y_n) = \frac{1}{n}(c \cdot x_1 + \cdots + c \cdot x_n) \tag{4}$$

$$= c \cdot \frac{1}{n}(x_1 + \cdots + x_n) = c \cdot \overline{x}. \tag{5}$$

Example

Supongamos que los datos x_i son los pesos de los recien nacidos expresados en gramos . Si $\overline{x} = 3169.9\, g$. Expresar la media aritmética en onzas donde $1\, oz = 28.35\, g$.

Si

$$1 \cdot oz = 28.35 \cdot g, \text{ then } \frac{1}{28.35} \cdot oz = 1 \cdot g.$$

Entonces,

$$x_i \cdot g = x_i \cdot \frac{1}{28.35} \cdot oz = \frac{1}{28.35} \cdot x_i \cdot oz = y_i \cdot oz.$$

De la última igualdad se obtiene $\frac{1}{28.35} \cdot x_i = y_i$. En consecuencia

$$\overline{y} = \frac{1}{28.35} \cdot \overline{x} = \frac{1}{28.35} \cdot 3169.9 = 111.71.$$

Proposición

Sea $x_1, \ldots, x_n$ *la muestra original de datos e*

$$y_i = c \cdot x_i + d,$$

para $i = 1, 2, \ldots, n$ *donde* c *y* d *son dos constantes dadas, representan una transformación de la muestra original. Entonces*

$$\overline{y} = c \cdot \overline{x} + d.$$

Proposición

Sea $x_1, \ldots, x_n$ la muestra original de datos e

$$y_i = c \cdot x_i + d,$$

para $i = 1, 2, \ldots, n$ donde c y d son dos constantes dadas, representan una transformación de la muestra original. Entonces

$$\overline{y} = c \cdot \overline{x} + d.$$

Example

Recordemos por ejemplo la fórmula de tranformación entre grados Fharenheit y Celsius:

$$[F] = 9.5 \cdot [C] + 32$$

Definition

El *rango* es la diferencia entre el valor mayor y el menor de una muestra dada.

Definition

El *rango* es la diferencia entre el valor mayor y el menor de una muestra dada.

Example

Supongamos una muestra ordenada $\{-3, -2, 1, 1.2, 2.1, 3.3, 5, 7.2\}$. Entonces su rango es

$$7.2 - (-3) = 7.2 + 3 = 10.2.$$

Definition

El *rango* es la diferencia entre el valor mayor y el menor de una muestra dada.

Example

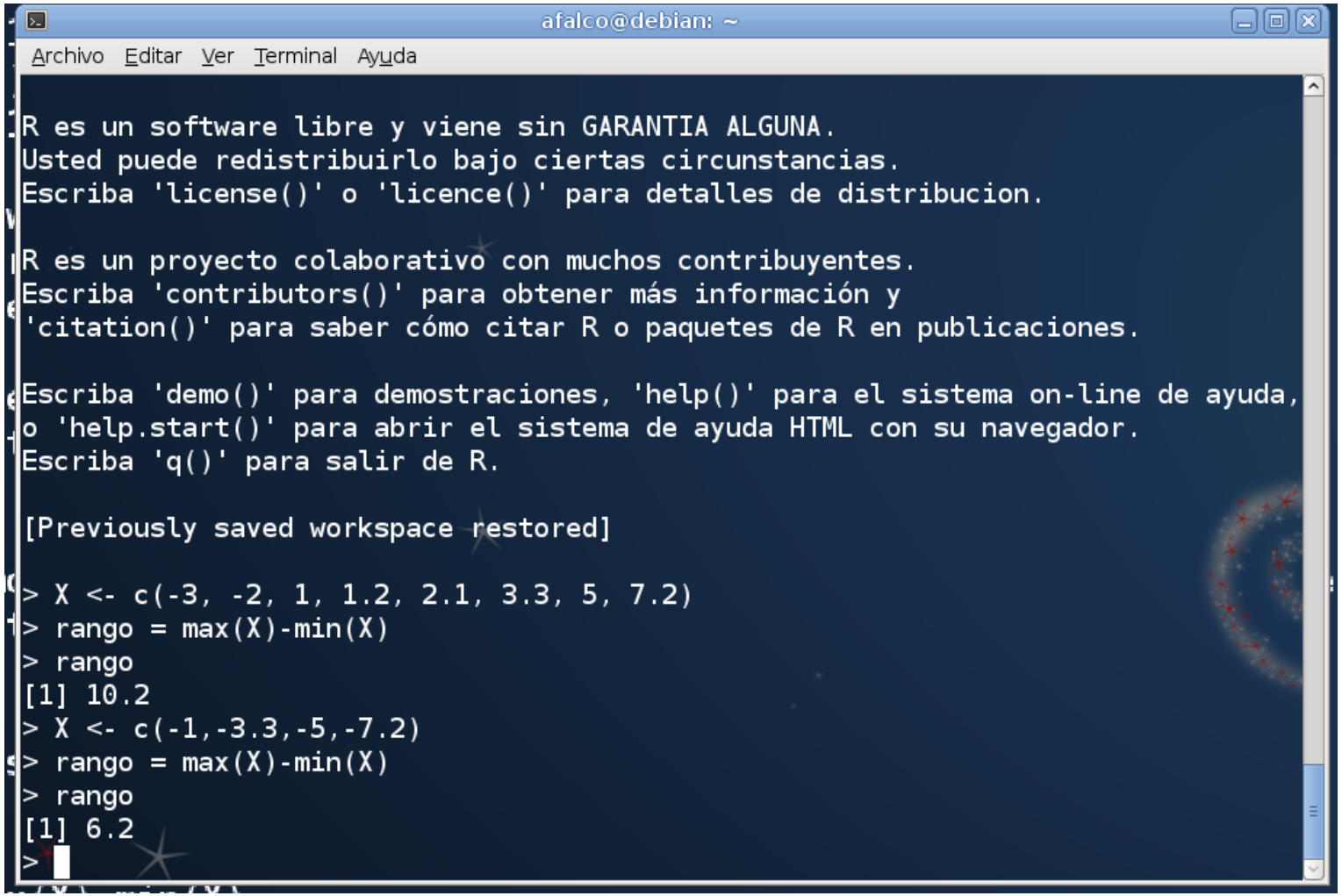
Supongamos una muestra ordenada $\{-3, -2, 1, 1.2, 2.1, 3.3, 5, 7.2\}$. Entonces su rango es

$$7.2 - (-3) = 7.2 + 3 = 10.2.$$

Example

Consideremos la muestra $\{-1, -3.3, -5, -7.2\}$. Entonces su rango es

$$-1 - (-7.2) = -1 + 7.2 = 6.2.$$


afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
[Previously saved workspace restored]
> X <- c(-3, -2, 1, 1.2, 2.1, 3.3, 5, 7.2)
> rango = max(X)-min(X)
> rango
[1] 10.2
> X <- c(-1,-3.3,-5,-7.2)
> rango = max(X)-min(X)
> rango
[1] 6.2
>

Definition

El p-ésimo *percentil* es el valor, denotado por V_p, de forma que el $p\%$ de los datos muestrales son menores o iguales a V_p.

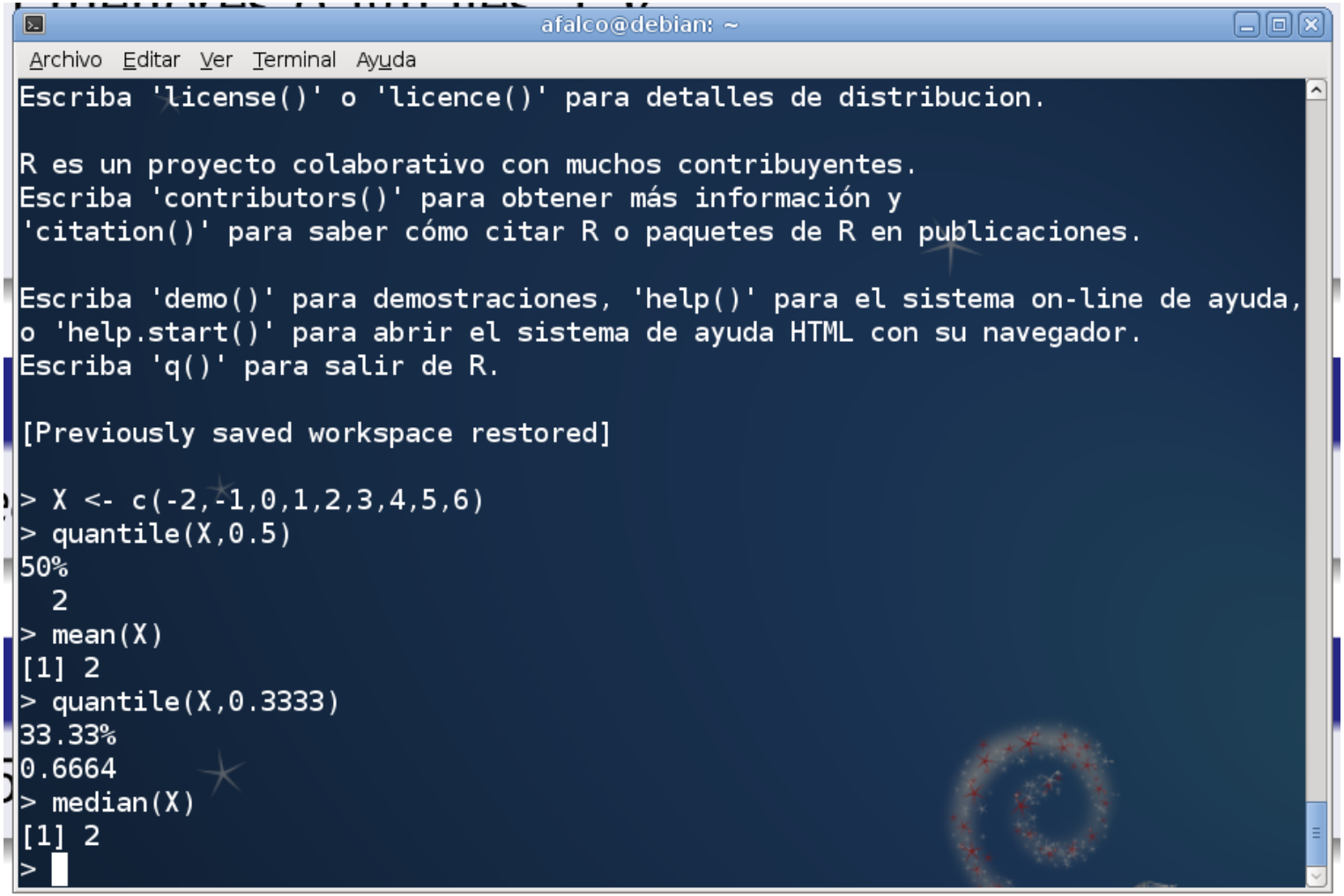
$$\frac{\#\{x_i : x_i \leq V_p\}}{n} = \frac{p}{100}.$$

Observación

El percentil 50 *es la mediana muestral.*

Definition

A los percentiles $0, 25, 50$ y 100 se les suele llamar *cuartiles*


afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
[Previously saved workspace restored]
> X <- c(-2,-1,0,1,2,3,4,5,6)
> quantile(X,0.5)
50%
2
> mean(X)
[1] 2
> quantile(X,0.3333)
33.33%
0.6664
> median(X)
[1] 2
>

La varianza y la desviación estándar

La variabilidad de los datos muestrales se mide para comprender el nivel de dispersión que existe en los datos poblacionales. Los más populares y frecuentemente empleados son la varianza y la desviación típica.

La varianza y la desviación estándar

La variabilidad de los datos muestrales se mide para comprender el nivel de dispersión que existe en los datos poblacionales. Los más populares y frecuentemente empleados son la varianza y la desviación típica.

Supongamos que para $\{x_1, \ldots, x_n\}$ calculamos las desviaciones con respecto la media aritmética

$$x_1 - \overline{x}, \ldots, x_n - \overline{x}.$$

Una medida simple que satisface nuestros objetivos es:

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

Malas noticias

Proposición

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

Proof.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) = \frac{1}{n}\left((x_1 - \overline{x}) + \cdots + (x_n - \overline{x})\right) \tag{6}$$

$$= \frac{1}{n}\left((x_1 + \cdots + x_n) - (n \cdot \overline{x})\right) \tag{7}$$

$$= \frac{1}{n}(x_1 + \cdots + x_n) - \frac{1}{n}(n \cdot \overline{x}) = \overline{x} - \overline{x} = 0. \tag{8}$$

Definition

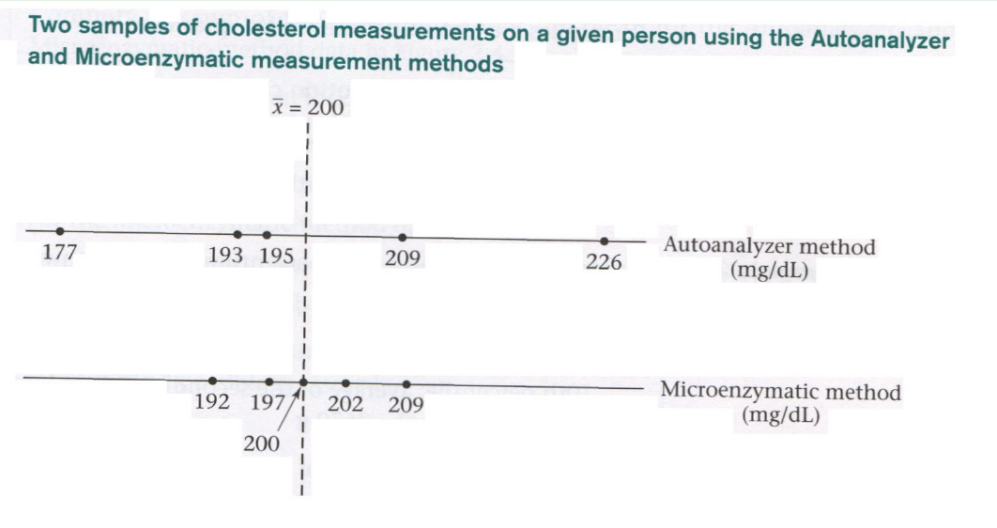
La *varianza* o *varianza* se define como sigue:

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

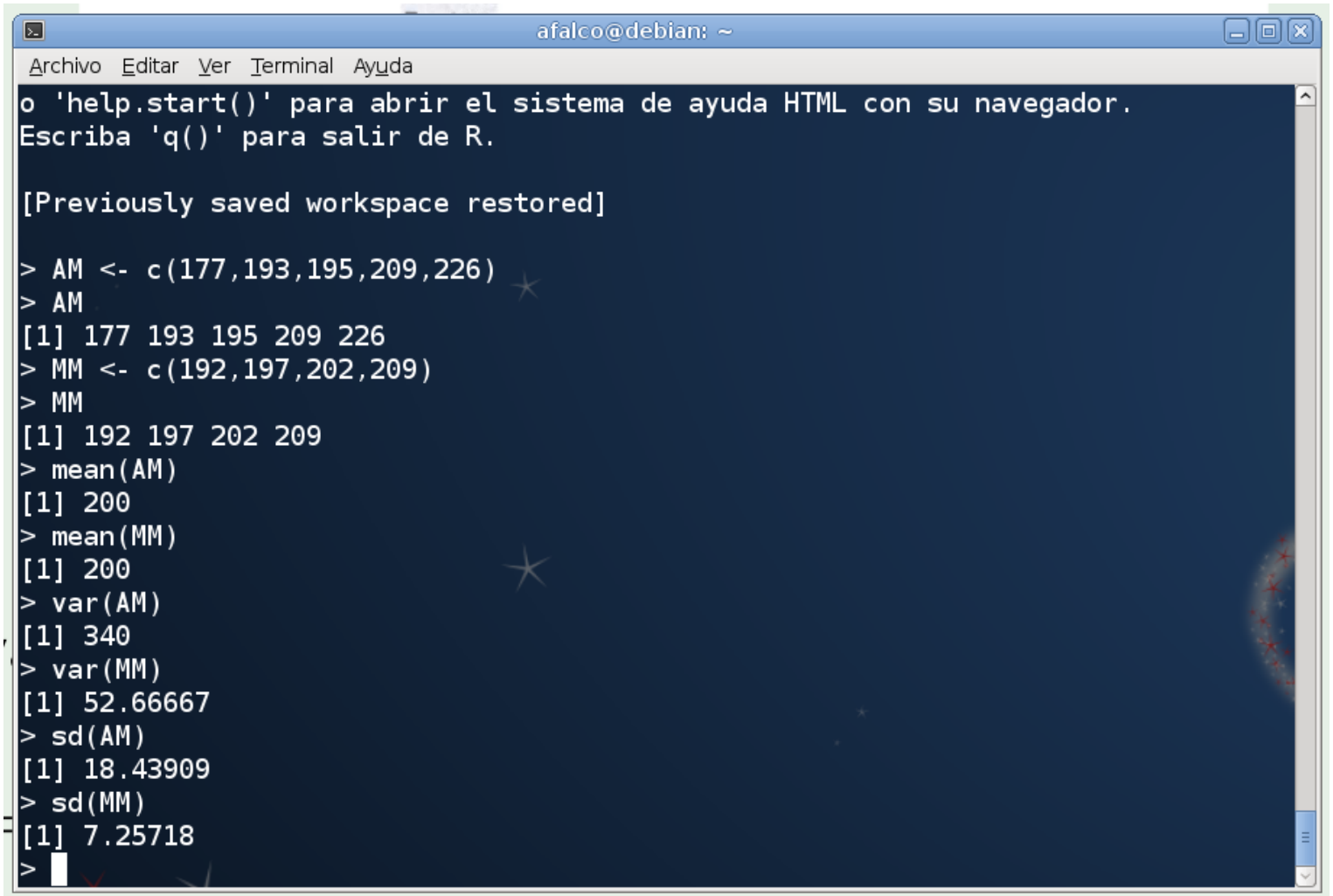
Entonces la *desviación estándar* se define como:

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

Example



Calcular la varianza y la desviación estándar.


afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
[Previously saved workspace restored]
> AM <- c(177,193,195,209,226)
> AM
[1] 177 193 195 209 226
> MM <- c(192,197,202,209)
> MM
[1] 192 197 202 209
> mean(AM)
[1] 200
> mean(MM)
[1] 200
> var(AM)
[1] 340
> var(MM)
[1] 52.66667
> sd(AM)
[1] 18.43909
> sd(MM)
[1] 7.25718
>

Proposición

Supongamos que tenemos dos muestras $x_1, \ldots, x_n$ *e* $y_1, \ldots, y_n$ *donde* $y_i = x_i + c$ *para* $i = 1, 2, \ldots, n$ *para una constante dada* c. *Si las respectivas varianzas muestrales las denotamos por* s_x^2 *y* s_y^2 *entonces se cumple:*

$$s_y^2 = s_x^2.$$

Proposición

Supongamos que tenemos dos muestras $x_1, \ldots, x_n$ *e* $y_1, \ldots, y_n$ *donde* $y_i = x_i + c$ *para* $i = 1, 2, \ldots, n$ *para una constante dada* c*. Si las respectivas varianzas muestrales las denotamos por* s_x^2 *y* s_y^2 *entonces se cumple:*

$$s_y^2 = s_x^2.$$

Proof.

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}((x_i + c) - (\overline{x} + c))^2 \tag{9}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i + c - \overline{x} - c)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s_x^2. \tag{10}$$

□

Proposición

Supongamos que tenemos dos muestras $x_1, \ldots, x_n$ *and* $y_1, \ldots, y_n$ *donde* $y_i = cx_i$ *para* $i = 1, 2, \ldots, n$ *para una constante dada* $c > 0$. *Entonces*

$$s_y^2 = c^2 s_x^2 \text{ y } s_y = cs_x.$$

Proposición

Supongamos que tenemos dos muestras $x_1, \ldots, x_n$ *and* $y_1, \ldots, y_n$ *donde* $y_i = cx_i$ *para* $i = 1, 2, \ldots, n$ *para una constante dada* $c > 0$. *Entonces*

$$s_y^2 = c^2 s_x^2 \text{ y } s_y = cs_x.$$

Proof.

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(cx_i - c\overline{x})^2 \tag{11}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} c^2(x_i - \overline{x})^2 = c^2\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = c^2 s_x^2. \tag{12}$$

□

Example

Si tenemos una muestra de temperaturas Celsius $\{10, 4.5, 7, 6, 2.2\}$. Si conocemos que $s_x = 2.9014$ Calcular la desviación estándar en grados Fahrenheit.

Como $[F] = 9.5 \cdot [C] + 32$, entonces $y_i = 9.5x_i + 32$. Empleando resultados anteriores:

$$s_y^2 = s_{9.5x}^2 = (9.5)^2 s_x^2 \text{ thus } s_y = 9.5 s_x = 9.5 \cdot 2.9014 = 27.563.$$

Observación

Si los datos muestrales se proporcionan en una determinada unidad física (e.g. Celsius), entonces tanto la media aritmética como la desviación estándar vienen dadas en las mismas unidades físicas (e.g. Celsius).

Observación

Si los datos muestrales se proporcionan en una determinada unidad física (e.g. Celsius), entonces tanto la media aritmética como la desviación estándar vienen dadas en las mismas unidades físicas (e.g. Celsius).

Ahora deseamos construir una medida adimensional (sin unidades físicas):

Observación

Si los datos muestrales se proporcionan en una determinada unidad física (e.g. Celsius), entonces tanto la media aritmética como la desviación estándar vienen dadas en las mismas unidades físicas (e.g. Celsius).

Ahora deseamos construir una medida adimensional (sin unidades físicas):

Definition

El *coeficiente de variación (CV)* se define como

$$100\% \cdot \frac{s}{\overline{x}}.$$

Observación

Si los datos muestrales se proporcionan en una determinada unidad física (e.g. Celsius), entonces tanto la media aritmética como la desviación estándar vienen dadas en las mismas unidades físicas (e.g. Celsius).

Ahora deseamos construir una medida adimensional (sin unidades físicas):

Definition

El *coeficiente de variación (CV)* se define como

$$100\% \cdot \frac{s}{\overline{x}}.$$

Observación

If $y_i = cx_i + d$ *then* $\frac{s_y}{\overline{y}} = \frac{s_x}{\overline{x}}$

Example

Consideremos una muestra de temperaturas en Celsius $\{10, 4.5, 7, 6, 2.2\}$. Calculamos $\overline{x} = 5.94$ y $s_x = 2.9014$. Obtener CV

$$\frac{s_x}{\overline{x}} = \frac{2.9014}{5.94} = 0.48845$$

Entonces $CV = 48.845\%$.

Example

Consideremos una muestra de temperaturas en Celsius $\{10, 4.5, 7, 6, 2.2\}$. Calculamos $\overline{x} = 5.94$ y $s_x = 2.9014$. Obtener CV

$$\frac{s_x}{\overline{x}} = \frac{2.9014}{5.94} = 0.48845$$

Entonces $CV = 48.845\%$.

- El CV resulta útil para comparar la variabilidad de muestras diferentes cada una de las cuales tiene una media aritmética diferente

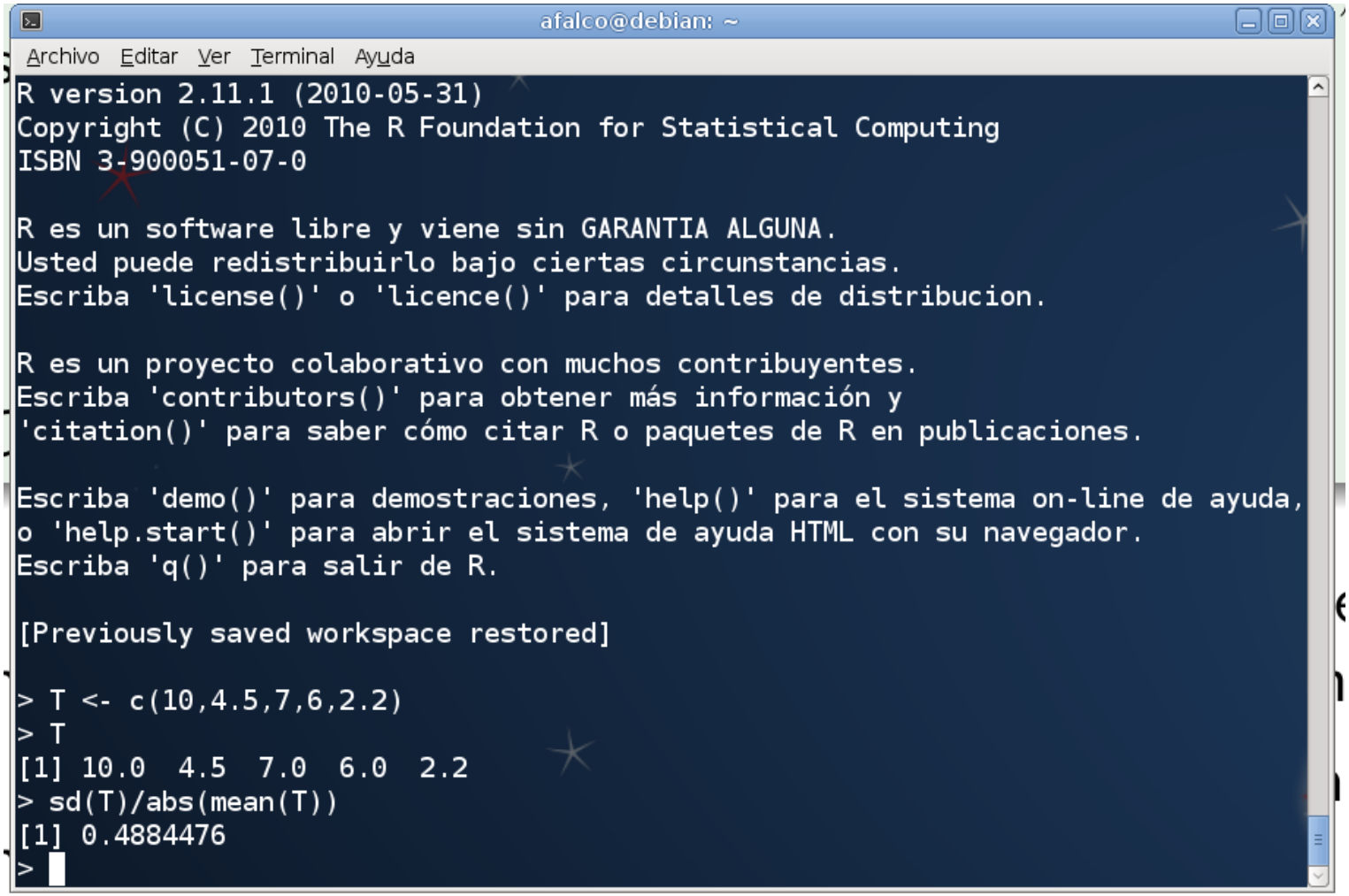
Example

Consideremos una muestra de temperaturas en Celsius $\{10, 4.5, 7, 6, 2.2\}$. Calculamos $\overline{x} = 5.94$ y $s_x = 2.9014$. Obtener CV

$$\frac{s_x}{\overline{x}} = \frac{2.9014}{5.94} = 0.48845$$

Entonces $CV = 48.845\%$.

- El CV resulta útil para comparar la variabilidad de muestras diferentes cada una de las cuales tiene una media aritmética diferente
- El CV es tambien útil para comparar la reproducibilidad de diferentes variables.


afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
R version 2.11.1 (2010-05-31)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
[Previously saved workspace restored]
> T <- c(10,4.5,7,6,2.2)
> T
[1] 10.0 4.5 7.0 6.0 2.2
> sd(T)/abs(mean(T))
[1] 0.4884476
>

Definition

Una *distribución de frecuencias* is una visión ordenada de cada una de los valores de un conjunto de dadtos junto a sus *frecuencias*, esto es, la cantidad de veces que cada valor aparece en el conjunto de datos. Además, el porcentaje que toma cada valor particular se suele tambien dar de forma usual.

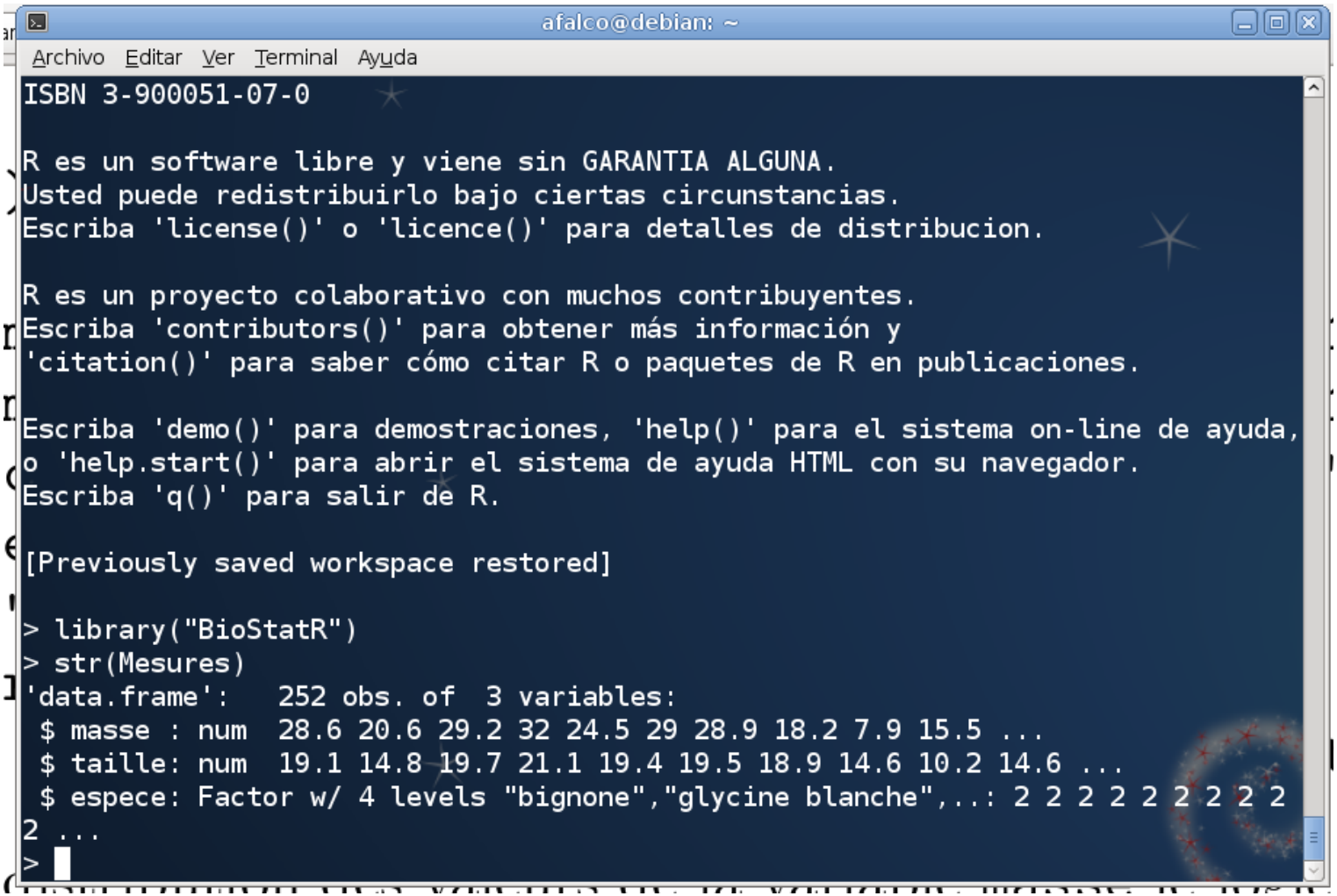
General layout of grouped data

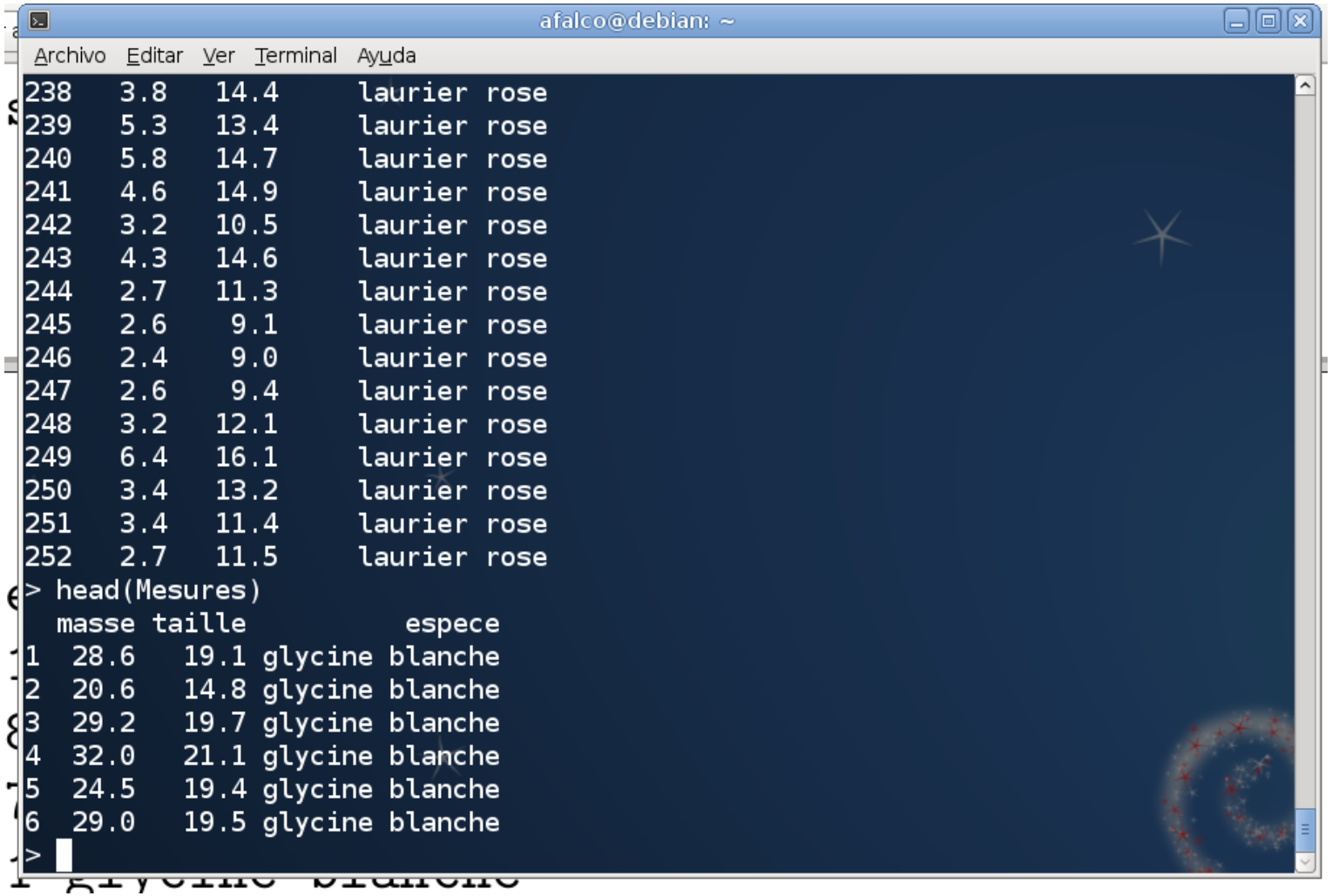
Group interval	Frequency
$y_1 \leq x < y_2$	f_1
$y_2 \leq x < y_3$	f_2
.	.
.	.
.	.
$y_i \leq x < y_{i+1}$	f_i
.	.
.	.
.	.
$y_k \leq x < y_{k+1}$	f_k

Grouped frequency distribution of birthweight (oz) from 100 consecutive deliveries

Group interval	Frequency
$29.5 \le x < 69.5$	5
$69.5 \le x < 89.5$	10
$89.5 \le x < 99.5$	11
$99.5 \le x < 109.5$	19
$109.5 \le x < 119.5$	17
$119.5 \le x < 129.5$	20
$129.5 \le x < 139.5$	12
$139.5 \le x < 169.5$	6
	100

Note: If birthweight can only be measured to an accuracy of 0.1 oz, then a possible alternate representation of the group intervals in Table 2.10 could be 29.5–69.4, 69.5–89.4, to 139.5–169.5.


afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
[Previously saved workspace restored]
> library("BioStatR")
> str(Mesures)
'data.frame': 252 obs. of 3 variables:
$ masse : num 28.6 20.6 29.2 32 24.5 29 28.9 18.2 7.9 15.5 ...
$ taille: num 19.1 14.8 19.7 21.1 19.4 19.5 18.9 14.6 10.2 14.6 ...
$ espece: Factor w/ 4 levels "bignone","glycine blanche",..: 2 2 2 2 2 2 2 2 2
2 ...
>


afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
238 3.8 14.4 laurier rose
239 5.3 13.4 laurier rose
240 5.8 14.7 laurier rose
241 4.6 14.9 laurier rose
242 3.2 10.5 laurier rose
243 4.3 14.6 laurier rose
244 2.7 11.3 laurier rose
245 2.6 9.1 laurier rose
246 2.4 9.0 laurier rose
247 2.6 9.4 laurier rose
248 3.2 12.1 laurier rose
249 6.4 16.1 laurier rose
250 3.4 13.2 laurier rose
251 3.4 11.4 laurier rose
252 2.7 11.5 laurier rose
> head(Mesures)
masse taille espece
1 28.6 19.1 glycine blanche
2 20.6 14.8 glycine blanche
3 29.2 19.7 glycine blanche
4 32.0 21.1 glycine blanche
5 24.5 19.4 glycine blanche
6 29.0 19.5 glycine blanche
>

Gráficas de Barras

Un *gráfico de barras* puede construirse como sigue.

Gráficas de Barras

Un *gráfico de barras* puede construirse como sigue.

1. Los datos se dividen en grupos.

Gráficas de Barras

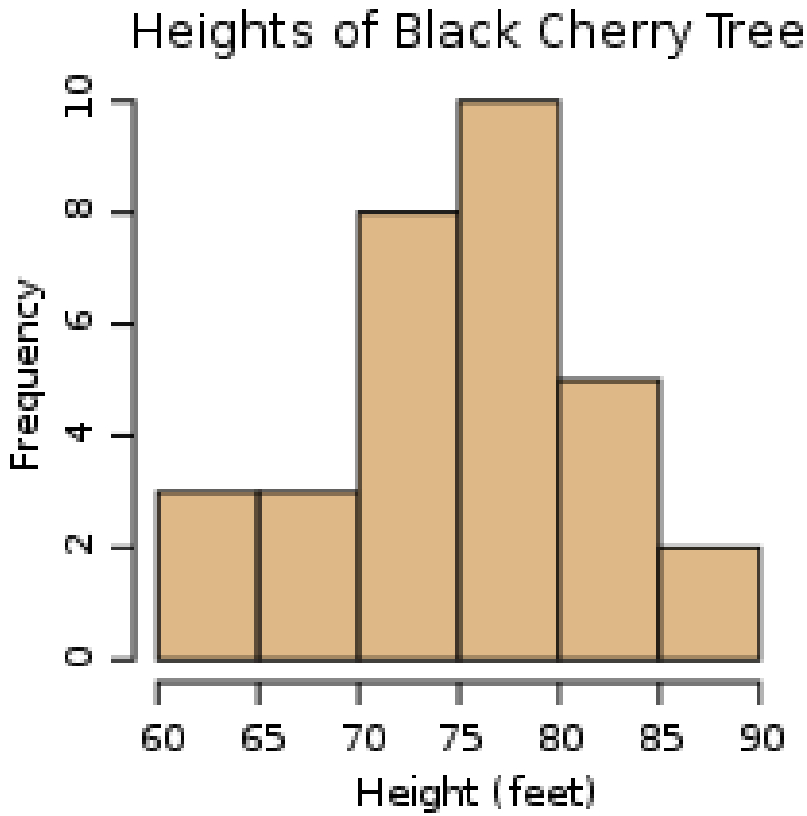
Un *gráfico de barras* puede construirse como sigue.

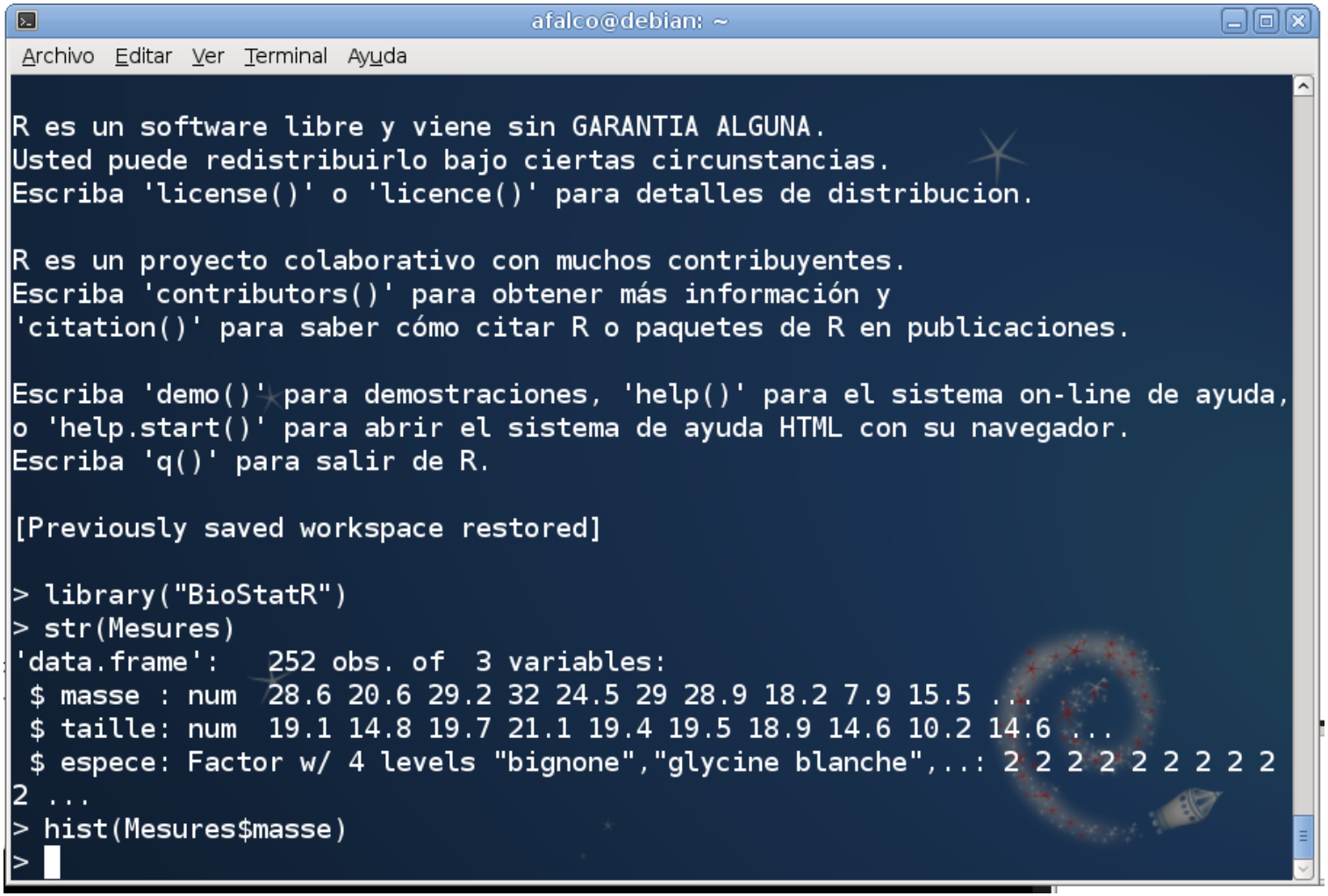
1. Los datos se dividen en grupos.
2. Para cada grupo construimos un rectángulo con una base de anchura constante y una altura proporcional a la frecuencia del grupo.

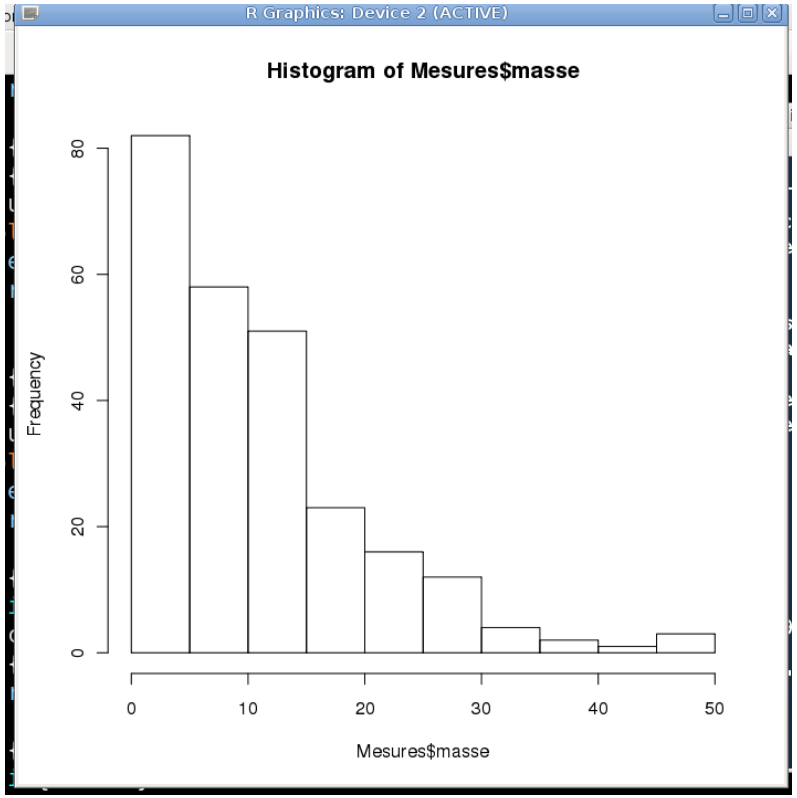
Gráficas de Barras

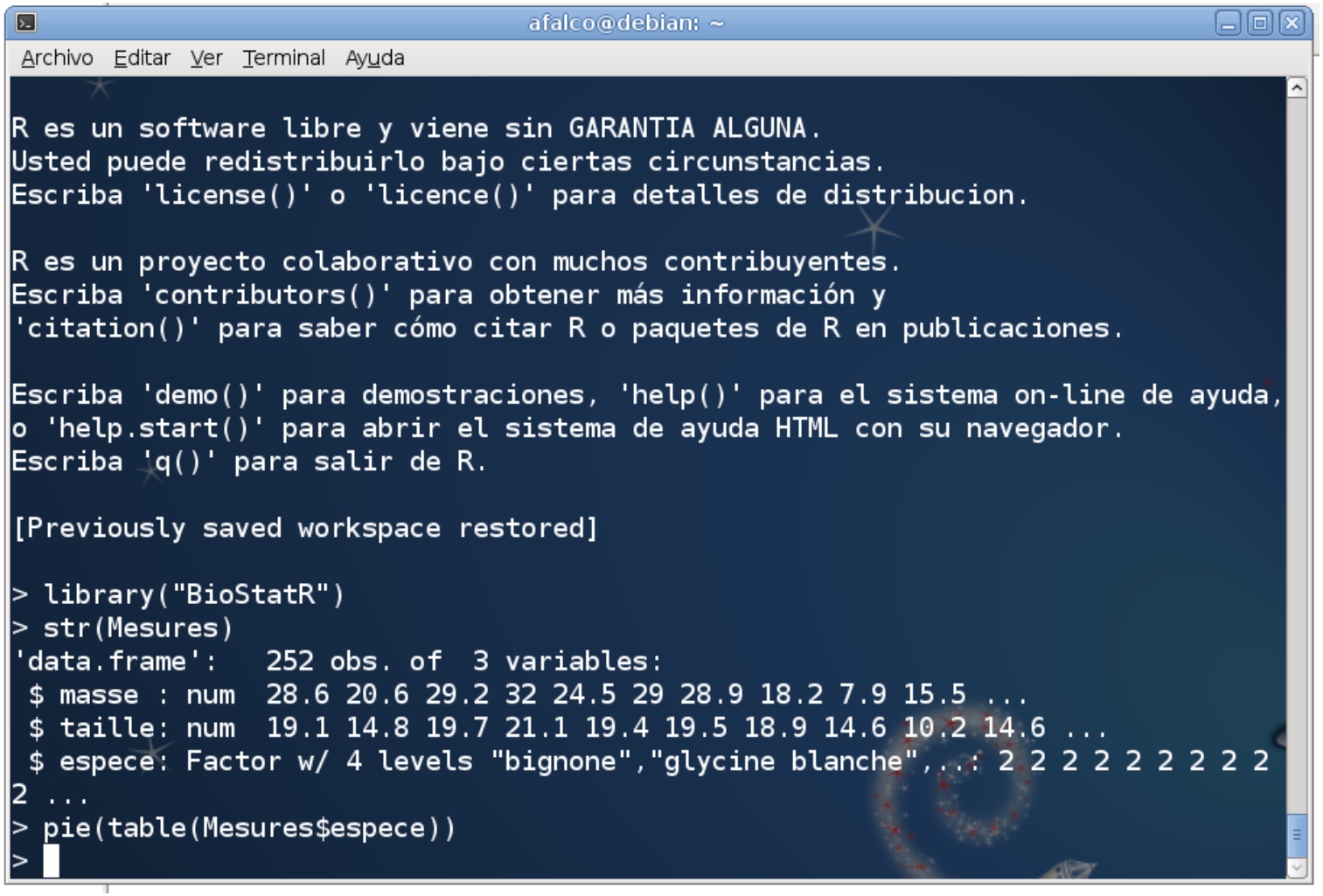
Un *gráfico de barras* puede construirse como sigue.

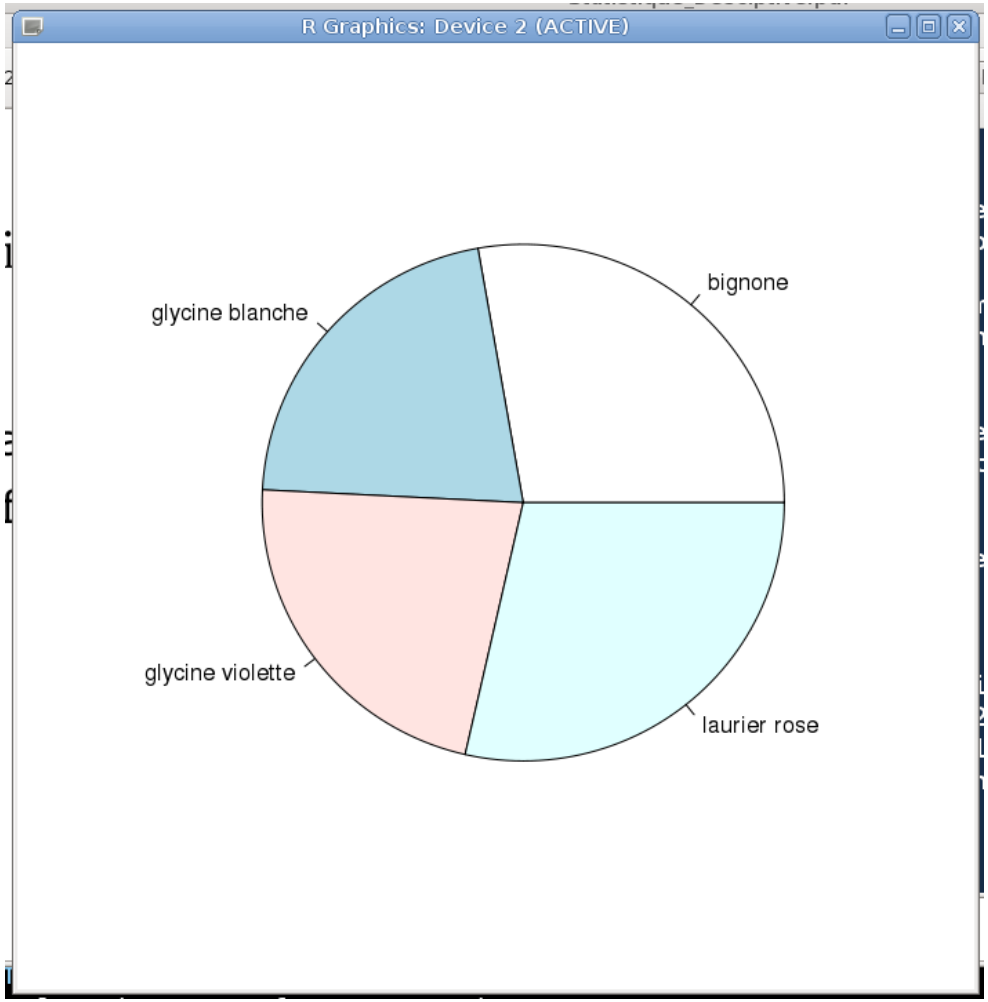
1. Los datos se dividen en grupos.
2. Para cada grupo construimos un rectángulo con una base de anchura constante y una altura proporcional a la frecuencia del grupo.
3. Los rectángulos nos osn generalmete contiguos y estan igualemente espaciados unos de otros.


Heights of Black Cherry Trees
Frequency
0
2
4
6
8
10
60
65
70
75
80
85
90
Height (feet)


afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
[Previously saved workspace restored]
> library("BioStatR")
> str(Mesures)
'data.frame': 252 obs. of 3 variables:
$ masse : num 28.6 20.6 29.2 32 24.5 29 28.9 18.2 7.9 15.5 ...
$ taille: num 19.1 14.8 19.7 21.1 19.4 19.5 18.9 14.6 10.2 14.6 ...
$ espece: Factor w/ 4 levels "bignone","glycine blanche",..: 2 2 2 2 2 2 2 2 2 2 ...
> hist(Mesures$masse)
>


R Graphics: Device 2 (ACTIVE)
Histogram of Mesures$masse
Frequency
0
20
40
60
80
0
10
20
30
40
50
Mesures$masse


afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
[Previously saved workspace restored]
> library("BioStatR")
> str(Mesures)
'data.frame': 252 obs. of 3 variables:
$ masse : num 28.6 20.6 29.2 32 24.5 29 28.9 18.2 7.9 15.5 ...
$ taille: num 19.1 14.8 19.7 21.1 19.4 19.5 18.9 14.6 10.2 14.6 ...
$ espece: Factor w/ 4 levels "bignone","glycine blanche",..: 2 2 2 2 2 2 2 2 2
2 ...
> pie(table(Mesures$espece))
>


R Graphics: Device 2 (ACTIVE)
bignone
glycine blanche
glycine violette
laurier rose

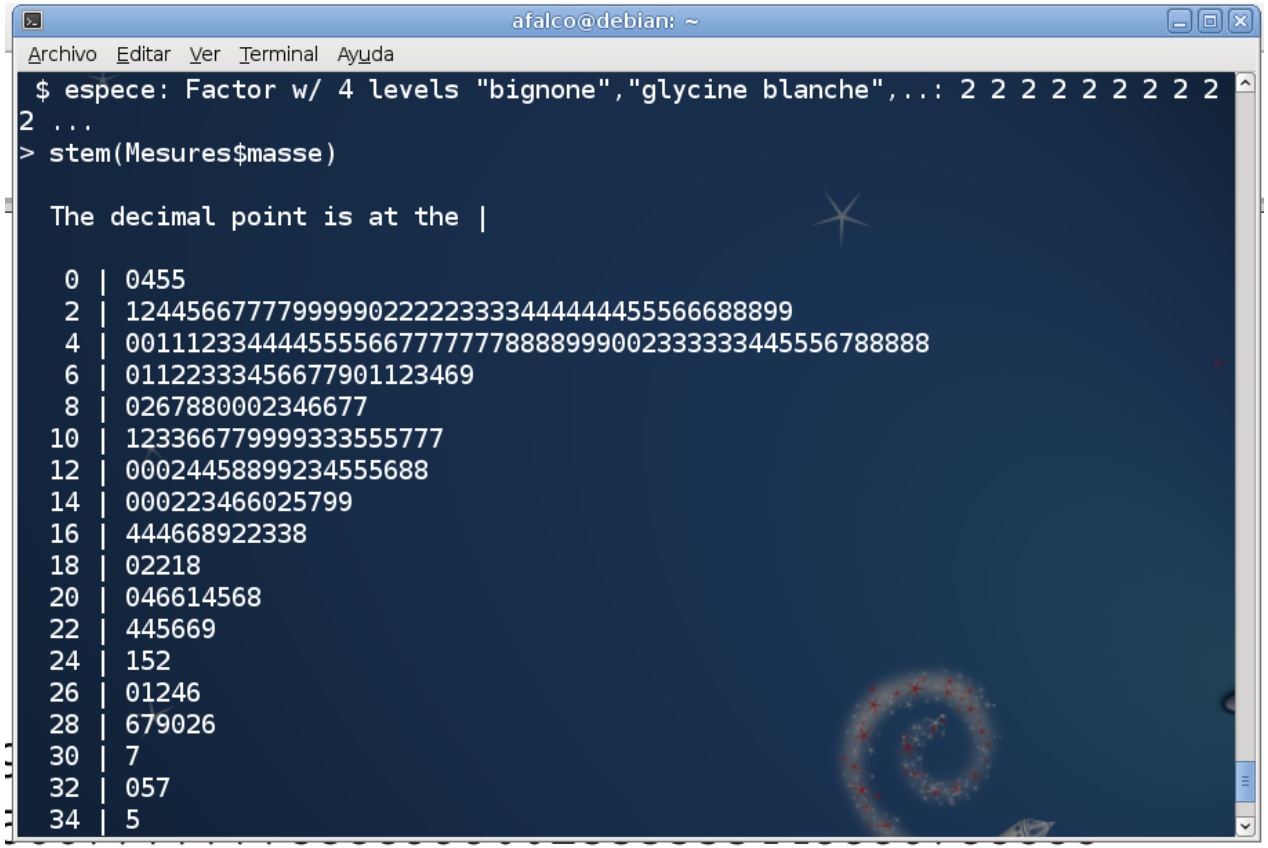
Grafo de hoja y tallos

Escribir los datos en forma ascendente

44, 46, 47, 49, 63, 64, 66, 68, 68, 72, 72, 75, 76, 81, 84, 88, 106.

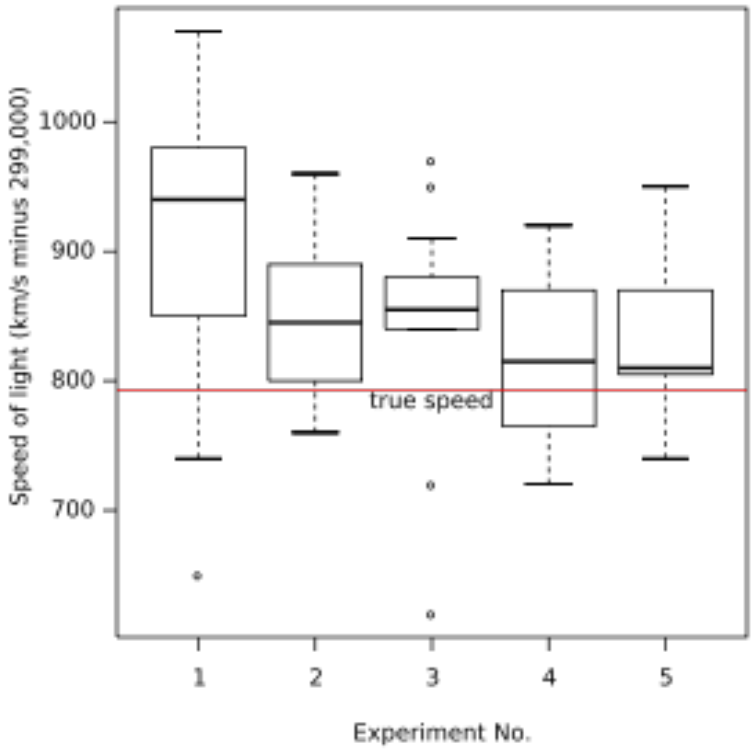
```
  4 | 4 6 7 9
  5 |
  6 | 3 4 6 8 8
  7 | 2 2 5 6
  8 | 1 4 8
  9 |
 10 | 6
key: 6|3=63
leaf unit: 1.0
stem unit: 10.0
```

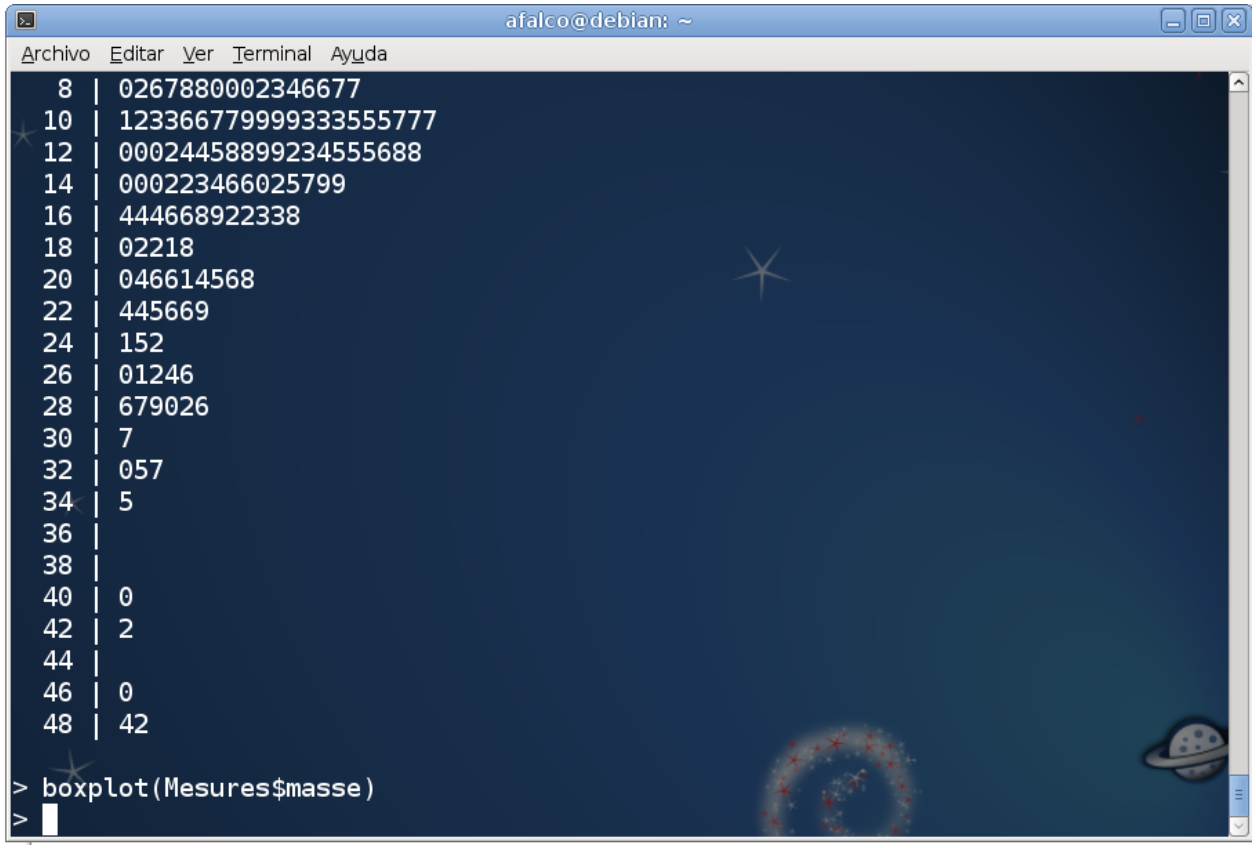
Tablas de hoja y tallos en las estaciones de metro japonesas

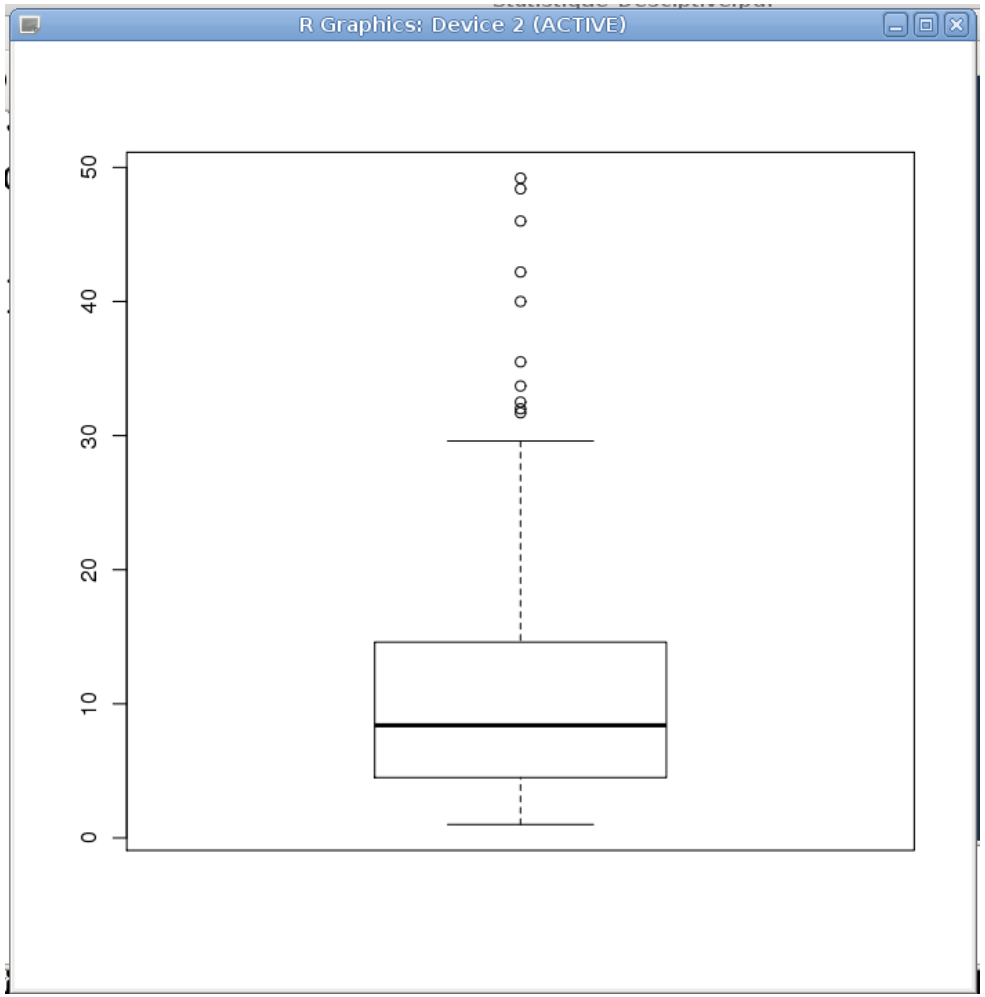

afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
$ espece: Factor w/ 4 levels "bignone","glycine blanche",..: 2 2 2 2 2 2 2 2 2
2 ...
> stem(Mesures$masse)
The decimal point is at the |
0 | 0455
2 | 12445667777999990222223333444444455566688899
4 | 00111233444455556677777778888999002333333445556788888
6 | 01122333456677901123469
8 | 0267880002346677
10 | 123366779999333555777
12 | 00024458899234555688
14 | 000223466025799
16 | 444668922338
18 | 02218
20 | 046614568
22 | 445669
24 | 152
26 | 01246
28 | 679026
30 | 7
32 | 057
34 | 5

Definition

En Estadística descriptiva, u *box plot o boxplot* (tambien conocido como box-and-whisker) es una forma de representación gráfica de datos conveniente cuendo se han de comparar grupos empleando los cuartiles y la mayor de las observaciones. Ademas se puede indicar que valores son anómalos (outliers).


Speed of light (km/s minus 299,000)
1000
900
800
700
true speed
1
2
3
4
5
Experiment No.


afalco@debian: ~
Archivo Editar Ver Terminal Ayuda
8 | 0267880002346677
10 | 123366779999333555777
12 | 00024458899234555688
14 | 000223466025799
16 | 444668922338
18 | 02218
20 | 046614568
22 | 445669
24 | 152
26 | 01246
28 | 679026
30 | 7
32 | 057
34 | 5
36 |
38 |
40 | 0
42 | 2
44 |
46 | 0
48 | 42
> boxplot(Mesures$masse)
>


R Graphics: Device 2 (ACTIVE)
0
10
20
30
40
50

Definition

An *outlying value* (0) is a value x such that either

1. $x >$ upper quartile $+ 1.5 \times$ (upper quartile $-$ lower quartile) or
2. $x <$ lower quartile $- 1.5 \times$ (upper quartile $-$ lower quartile).

Definition

An *extreme outlying value* $*$ is a value x such that either

1. $x >$ upper quartile $+ 3.0 \times$ (upper quartile $-$ lower quartile) or
2. $x <$ lower quartile $- 3.0 \times$ (upper quartile $-$ lower quartile).