



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Herramientas estadísticas para Big Data

Introducción a la Inferencia Estadística,
Muestreo y Preproceso de datos

Máster **Big Data** Analytics

Departamento de Estadística e
Investigación Operativa Aplicadas
y Calidad

Valencia, Octubre 2017

Elena Vázquez

www.upv.es

bigdata.inf.upv.es



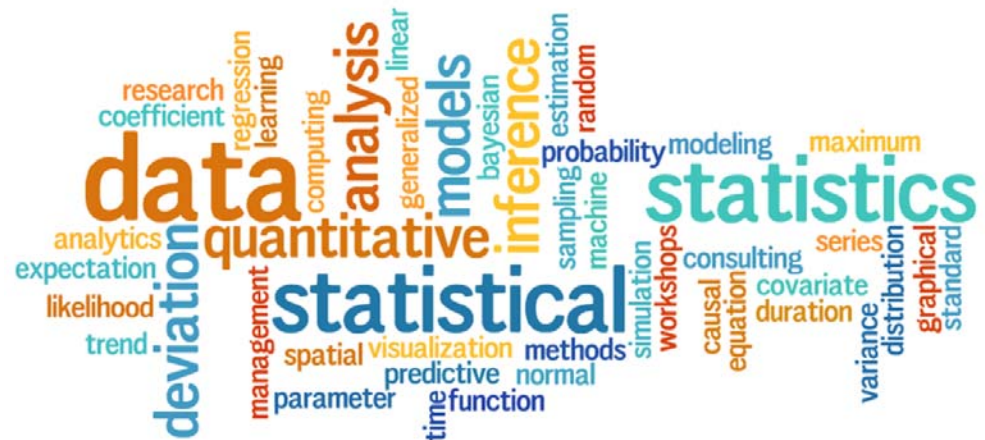
Contenidos

1. Conceptos básicos
2. Probabilidad
3. Variables aleatorias y distribuciones
4. Inferencia en muestras grandes
5. Técnicas de muestreo
6. Preprocesamiento de datos

Glosario

Enlaces de interés

Bibliografía





6 Preprocesamiento de datos

Introducción

Generación y recopilación

1. Limpieza (Data cleaning)

2. Integración

3. Transformación

4. Reducción

Glosario

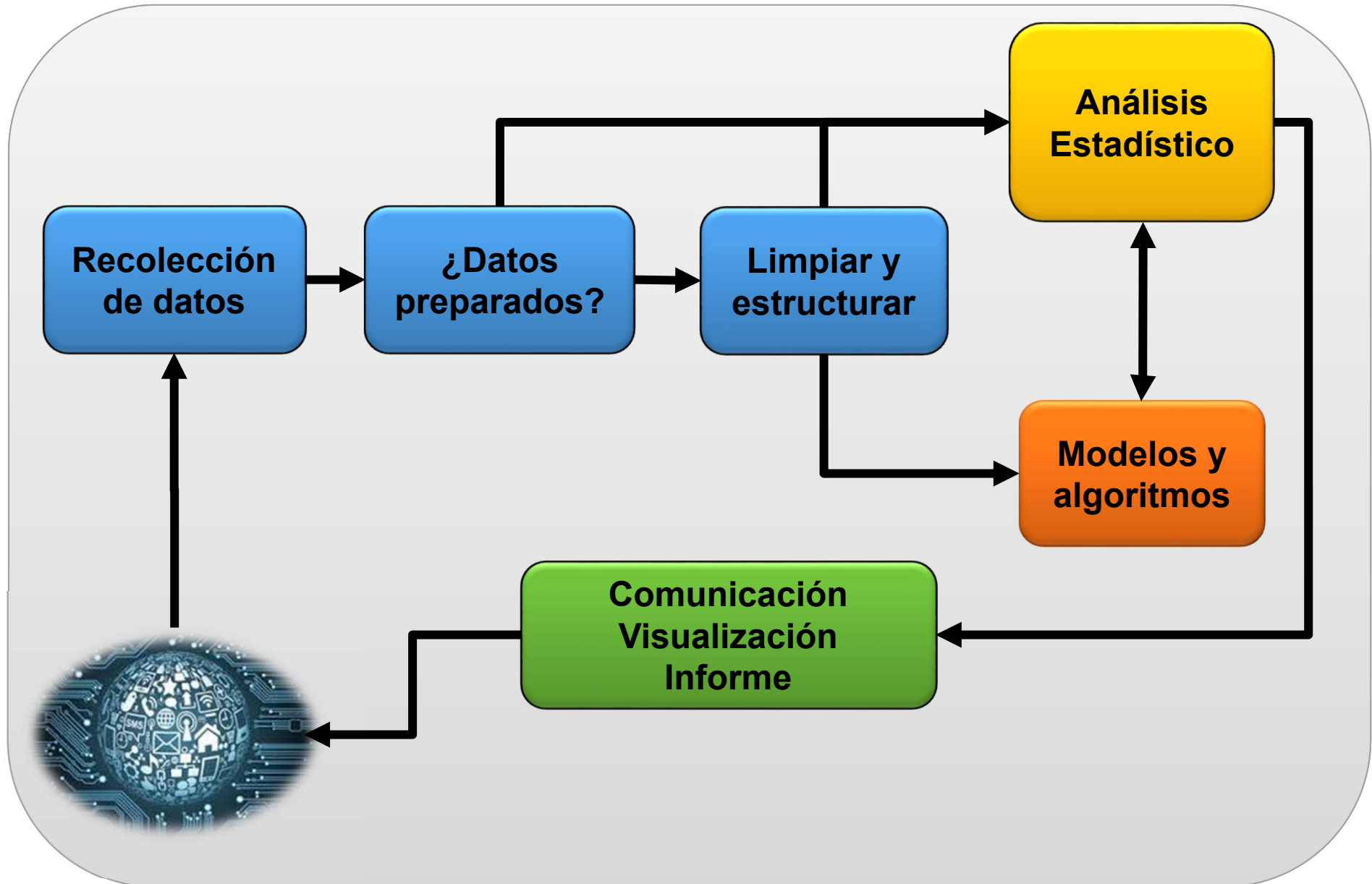


Introducción

- La mayor parte de la teoría estadística se centra en el modelado, predicción e inferencia, asumiendo que los datos son correctos, completos y en perfecto estado para el análisis.
- En la práctica, especialmente en el ámbito del Big Data, los datos crudos o *raw data* necesitan un proceso previo hasta que pueden considerarse *técnicamente correctos*.
- Mediante el **preproceso** se obtienen **datos sin errores, perfectamente codificados y consistentes para poder realizar el análisis estadístico**.
- Los resultados del preproceso pueden afectar significativamente a los resultados del análisis estadístico, por lo que podría considerarse como una fase más de éste.



Introducción



Recopilación o generación

- **Generación:**
 - Se puede simular datos a partir de generación de secuencias aleatorias de una distribución y otros procedimientos más complejos.
 - Encuestas, cuestionarios, DOE
- **Recopilación:**
 - Bases de datos:
 - Públicas
 - Privadas
 - Recolección de datos no estructurados:
 - Web de empresas
 - Otras web (scrapping)
 - Redes sociales
 - Google Analytics

BASES DE DATOS PÚBLICAS ESPAÑA



- INE <http://www.ine.es/>



- CIS <http://www.cis.es/cis/opencms/ES/index.html>



- **Reutilización de la información del sector público**
 - <http://datos.gob.es/acerca-de>



BASES DE DATOS PÚBLICAS INTERNACIONALES



- [Data.gov](#) : La página web Data.gov tiene 210.912 conjuntos de datos que están abiertos y disponibles gratuitamente para su descarga y uso. Muchos de los conjuntos de datos son visibles a través de mapas interactivos.
- [ARDA](#) – Association of Religion Data Archives: Este sitio web tiene también su propia sección de Mapas de GIS donde los usuarios pueden trazar conjuntos de datos religiosos sobre un barrio y / o mapas mundo.
- [Fuentes de datos públicos y abiertos en España y en el mundo](#) : sitio web La Oficina del Censo de EE.UU. tiene la versión más reciente del censo de EE.UU., que se puede descargar gratuitamente. También tienen una galería de visualizaciones de datos donde se pueden encontrar infografías y mapas en los que se han utilizado los conjuntos de datos de la Oficina del Censo.
- [CIA World Factbook](#) : The World Factbook, es elaborado por la Agencia Central de Inteligencia y proporciona información sobre la historia, la gente, gobierno, economía, geografía, comunicaciones, transporte, militar y cuestiones transnacionales de 267 entidades del mundo.
- [Eurostat](#) : La misión de Eurostat es ser el principal proveedor de estadísticas de alta calidad a la Unión Europea y los países candidatos.
- [Observatorio Mundial de la Salud](#) : Esta colección cuenta con más de 50 conjuntos de datos sobre temas prioritarios de salud como la mortalidad y la carga de las enfermedades, los Objetivos de Desarrollo del Milenio (nutrición infantil, salud infantil, la salud materna y reproductiva, la inmunización, el VIH / SIDA, la tuberculosis, el paludismo, las enfermedades olvidadas, agua y saneamiento), las enfermedades no transmisibles y factores de riesgo, enfermedades epidémicas, sistemas de salud, la salud ambiental, la violencia y las lesiones, la equidad, entre otros.



BASES DE DATOS PÚBLICAS INTERNACIONALES



- [Harvard Dataverse Red](#) : Esto es un repositorio para compartir, citando y la preservación de los datos de la investigación, abierta a todos los datos científicos de todas las disciplinas en todo el mundo. Incluye la mayor colección del mundo de los datos de la investigación en ciencias sociales.
- [Fondo Monetario Internacional \(FMI\) de datos](#) : El FMI (Fondo Monetario Internacional) publica una serie de datos de series de tiempo sobre los préstamos del FMI, de cambio y otros indicadores económicos y financieros.
- [NOAA: La Administración Nacional Oceánica y Atmosférica](#) : National Climatic Data Center de la NOAA (CNDC) es el mayor proveedor mundial de datos meteorológicos y climáticos. -Base en tierra, mar, modelo, radar, globo de tiempo, satélite y paleoclimáticos son sólo algunos de los tipos de conjuntos de datos disponibles.
- [Colección de datos del Banco Mundial](#) : El Catálogo de datos proporciona acceso a más de 8.000 descargas indicadores a partir de conjuntos de datos del Banco Mundial, una búsqueda por país, los indicadores, o tema.
- [NYC Open Data](#) : Esta colección cuenta con más de 800 conjuntos de datos que pertenecen a la ciudad de Nueva York, la mayoría de los cuales se puede ver como un mapa interactivo. Incluyen ubicaciones de graffiti, la ubicación de aseos en los parques públicos, lugares wifi, entradas de metro, y mucho más.
- *Global Administrative Areas* (<http://www.gadm.org/country>)





Recopilación: funciones

- Introducción y edición
 - **data.entry()**
 - **de()**
 - **scan()**
 - **edit()**
- Lectura / importación de archivos: diferentes fuentes
 - **load(*.Rdata)**
 - **read.table()**
 -



Importar archivo de datos: ejemplo



```
## Importar dataset  
> datos1<-read.table("EjemploBD.txt", header=T,  
sep="\t", dec=",", na.strings="NA")  
## Guardar dataset como fichero de R  
> save(datos1,file="datos.RData")  
  
> ## Hacer data set "datos1" el objeto predeterminado  
> attach(datos1)
```



Recopilación datos *raw*



- Lectura desde web:
 - Paquetes:
 - XML
 - rvest
 - Rcurl
- Datos redes sociales:
 - API's
 - Facebook
 - Tweeter
 - Youtube
 - ...



Lectura de datos de una web: ejemplo



```
# La dirección es:
# http://analisisydecision.es/manual-curso-
# introduccion-de-r-capitulo-17-analisis-cluster-con-r-
# y-iii/

# Partimos de un archivo de texto delimitado por
# tabuladores con 46 frutas
# y la información que disponemos es:

# Nombre fruta
# Intercambio de hidratos de carbono por gramo
# Kilocalorías
# Proteínas
# Grasas
# (información obtenida de www.diabetesjuvenil.com)
```



Lectura de datos de una web: ejemplo



```
# Leer datos
frutas<-
read.table(url("http://analisisydecision.es/wp-
content/uploads/2009/06/alimentos2.txt"),header=FALSE,
sep="\t")

# Definir nombres
nombres<-
c("nombre","inter_hidratos","kcal","proteinas","grasas
")

# Asignar nombres al dataset
names(frutas)<-nombres

# Editar hoja de datos
edit(frutas)
```



Lectura de datos de una web: ejemplo



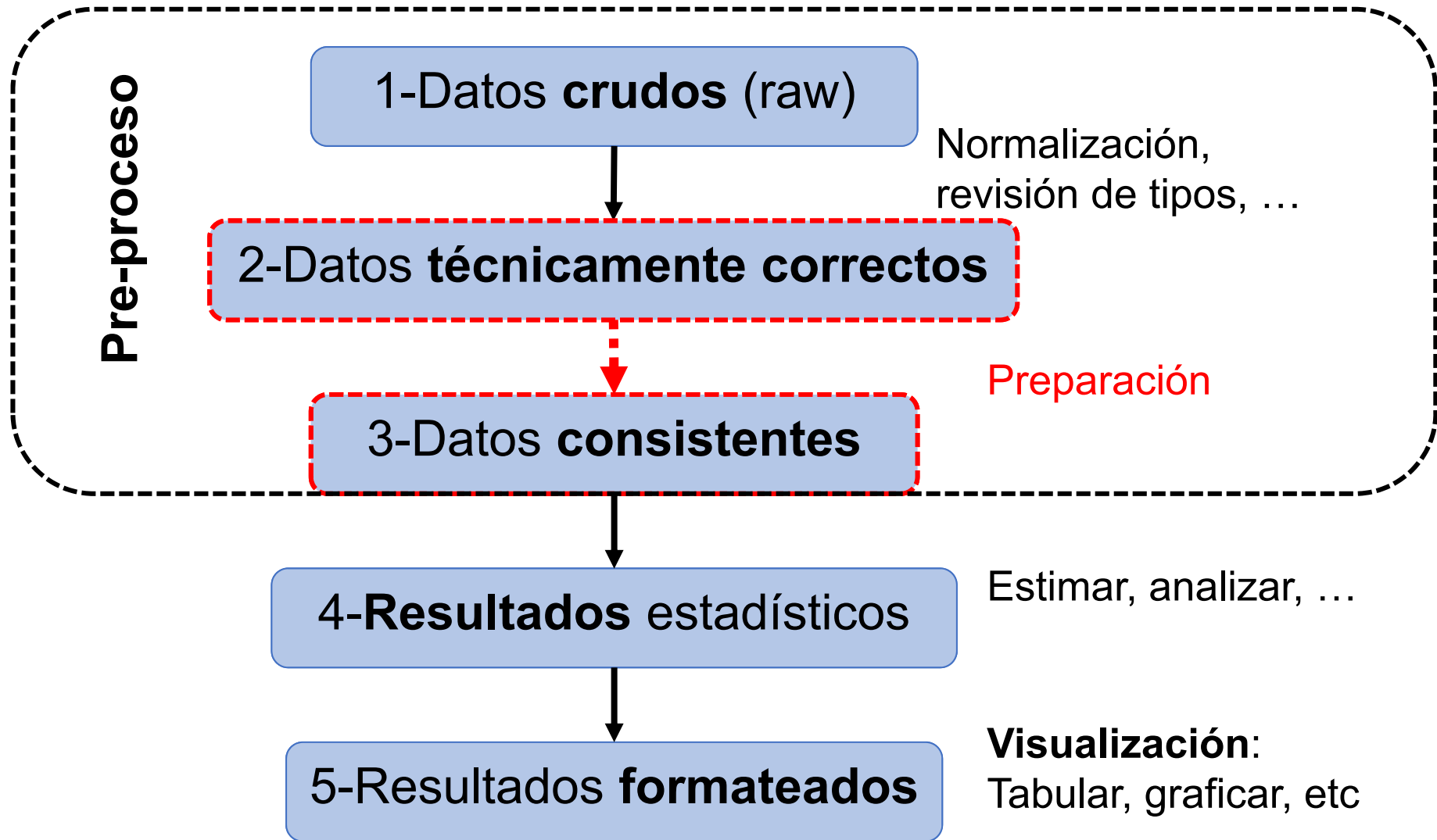
Editor de datos					
Archivo Editar Ayuda					
	nombre	inter_hidratos	kcal	proteinas	grasas
1	AGUACATE	769	1032.9	10	106.2
2	ALBARICOQUE	105	41.8	0.84	0.1
3	ARÁNDANO	145	43.6	0.9	0.3
4	BREVAS	63	41.1	0.75	0.1
5	CAQUIS	63	41	0.45	0.2
6	CEREZAS	74	43.2	0.6	0.4
7	CHIRIMOYAS	50	40.4	1	0.1
8	CIRUELAS	91	40.9	1	0.1
9	DÁTILES	14	39.1	0	0
10	FRAMBUESA	125	48.8	1	0.8
11	FRESA Y FRESÓN	143	49.2	1	0.9
12	GRANADA	133	42.4	1	0.1
13	GROSELLA NEGRA	152	44.3	1	0.2
14	GROSELLA ROJA	227	49.5	3	0.2
15	GUAYABA	149	49.6	1	0.7
16	GUINDAS	74	43.2	1	0.4
17	HIGO CHUMBO	108	45.7	1	0.4
18	HIGOS Y BREVAS	63	41.1	1	0.1
19	KIWI	83	44.8	1	0.4



Fases del preproceso

1. **Limpieza** (*Data cleaning*): a eliminar datos con ruido o incorrectos
2. **Integración**: e integrar diferentes fuentes de datos en un almacén coherente y homogéneo como un *data warehouse* o un *data cube*
3. **Transformación**
4. **Reducción**: reducir el tamaño de los datos mediante la agregación y/o eliminación de características redundantes, o clustering

Data cleaning



NOTA: En este módulo nos centramos en el paso de la fase 2 (**datos técnicamente correctos**) a la fase 3 (**datos consistentes**). Más directamente relacionados con las técnicas estadísticas.

De crudos a técnicamente correctos



Algunas funciones

Description

substr(x, start=n1, stop=n2)

Extract or replace substrings in a character vector.

x <- "abcdef"

substr(x, 2, 4) is "bcd"

substr(x, 2, 4) <- "22222" is "a222ef"

**grep(pattern, x ,
ignore.case=FALSE, fixed=FALSE)**

Search for *pattern* in *x*. If fixed =FALSE then *pattern* is a [regular expression](#). If fixed=TRUE then *pattern* is a text string. Returns matching indices.

grep("A", c("b","A","c"), fixed=TRUE) returns 2

**sub(pattern, replacement,x,
ignore.case =FALSE, fixed=FALSE)**

Find *pattern* in *x* and replace with *replacement* text. If fixed=FALSE then *pattern* is a regular expression.

If fixed = T then *pattern* is a text string.

sub("\s",".","Hello There") returns "Hello.There"

strsplit(x, split)

Split the elements of character vector *x* at *split*.

strsplit("abc", "") returns 3 element vector "a","b","c"

paste(..., sep="")

Concatenate strings after using *sep* string to separate them.

paste("x",1:3,sep="") returns c("x1","x2" "x3")

paste("x",1:3,sep="M") returns c("xM1","xM2" "xM3")

paste("Today is", date())

toupper(x)

Uppercase

tolower(x)

Lowercase



Preparación de datos

1. Limpieza (*Data cleaning*):

- i. Valores especiales
- ii. Valores perdidos o datos faltantes (missing values)
- iii. Valores anómalos o atípicos (outliers)

2. Integración

- i. Seleccionar datos
- ii. Combinar datos
- iii. Agregar datos

3. Transformación

- i. Normalización de variables cuantitativas
- ii. Transformaciones exponencial y logarítmica
- iii. Discretización de variables cuantitativas en categorías
- iv. Binarización de atributos categóricos

4. Reducción

- i. Análisis de Componentes Principales PCA

1. Limpieza de datos

- El objetivo es obtener **datos consistentes** en los que los valores perdidos, especiales, anómalos y las inconsistencias (obvias) se han detectado, localizado y eliminado, corregido o imputado.
- Básicamente comprenden los procedimientos de detección y tratamiento de:
 - **Valores especiales**
 - **Valores perdidos o datos faltantes (*missing values*)**
 - **Valores anómalos o atípicos (*outliers*)**
 - **Inconsistencias**

Valores especiales

- Como otros lenguajes de programación, R tiene valores especiales que son excepciones a los valores normales de los tipos definidos.
- Los **tipos de datos “normales”** soportados por R son:

numeric #Datos numéricos (aproximación a los números reales, \mathbb{R})

integer #Datos enteros (números enteros, \mathbb{Z})

factor #Datos categóricos (simples clasificaciones, como “género”)

ordered #Datos ordinales (clasificaciones ordenadas, como el nivel de educación)

character #Datos de tipo carácter (*strings*)

raw #Datos binarios





Tipos especiales de datos en R

- **NA (Not Available)** indica que el dato es faltante.
 - Las operaciones básicas funcionan bien con éstos, y devuelven habitualmente otro NA cuando algún dato faltante interviene.
 - No confundir NA con un valor que representa el “0” o “No aplica” y se ha dejado vacío.
- **NULL** Indica que no hay dato, no existe ni ocupa espacio. No es de ninguna clase (*class*)
- **Inf (Infinity)** Técnicamente es un número válido resultado de operaciones como la división por 0.
- **NaN (Not A Number)** Dato que existe (no es NULL) pero su valor no se conoce, no por omisión, sino porque es el resultado de algunas operaciones como 0/ 0, Inf-Inf o Inf/Inf. Técnicamente también es de tipo numérico :-0





Inspección de objetos y archivos

- **Funciones genéricas**

- `dir()`, `list.files()`
- `ls()`, `objects()`
- `head()`, `tail()`
- `str()`, `ls.str()`
- `summary()`
- `typeof()`, `class()`
- ---



Inspección de tipos y clases de objetos



Comprobación de tipos de datos y objetos y coerción

- ***is.***
 - is.numeric()
 - is.integer()
 - is.character()
 - is.logical()
 - is.factor()
 - is.ordered()
 - is.na()
 - is.null()
 - is.nan()
 - is.finite()
 - is.infinite()
- ***Coercion explícita***
 - as.numeric()
 - as.integer()
 - as.carácter()
 - as.logical()
 - as.factor()
 - as.ordered()





Valores especiales

- Las operaciones que implican valores especiales generalmente producen valores especiales.
- Las tesis estadísticas son sobre fenómenos del mundo real y éstas no incluyen nunca valores especiales.

```
#is.finite() determina si los valores son "normales"
```

```
> valores<-c(1, Inf, NA, NaN, NULL)
```

```
> is.finite(valores)
```

```
[1] TRUE FALSE FALSE FALSE
```



Valores perdidos

- También llamados **datos faltantes**, **ausentes** o *missing values*
- Son valores que **no constan** debido a causas tales como errores en la transcripción de los datos o falta de disposición a responder a algunos ítems de una encuesta. También pueden ser *valores especiales*.
- Los datos pueden faltar de manera aleatoria o no aleatoria.
 - Los **datos faltantes aleatorios** pueden perturbar el análisis de datos dado que disminuyen el tamaño de las muestras y en consecuencia la potencia de las pruebas de contraste de hipótesis.
 - Los **datos faltantes no aleatorios** ocasionan, además, disminución de la representatividad de la muestra.

Valores perdidos

- En **R** los datos faltantes se denotan con la etiqueta **NA** (**Not Available**):
 - NA no es, o no debería ser, una categoría por defecto.
 - NA indica que conocemos el tipo de dato, pero no lo tenemos.
- Debe ser el analista quien decida que hacer con ellos.
- **Tratamiento:**
 - Eliminación
 - Imputación

Tratamiento de valores perdidos

- **Eliminación**

- **De casos completos o eliminación por lista (*listwise*):**

Sólo se incluyen en el análisis los casos que presentan observaciones completas en todas las variables. Este método solo debe utilizarse cuando el proceso de recogida de datos es aleatorio, porque en otro caso introduce sesgo. Otro inconveniente es que el tamaño muestral puede reducirse sustancialmente y afectar a la representatividad de la muestra.

- **De casos incompletos o selección por variables (*pairwise*)**

Se mantienen en el *data set* los casos si disponen de datos en las variables que van a ser utilizadas en el análisis. Este procedimiento tiene el inconveniente de generar muestras heterogéneas.

- **Métodos de imputación**

- Los métodos de imputación consisten en estimar los valores ausentes en base a los valores válidos de otras variables y/o casos de la muestra.



Detección de valores perdidos: ejemplo 1

> ## Temperatura recogidas por 3 estaciones meteorológicas

> **load**("datos.Rdata")

Introducción de algún perdido en la hoja de datos **Temper**

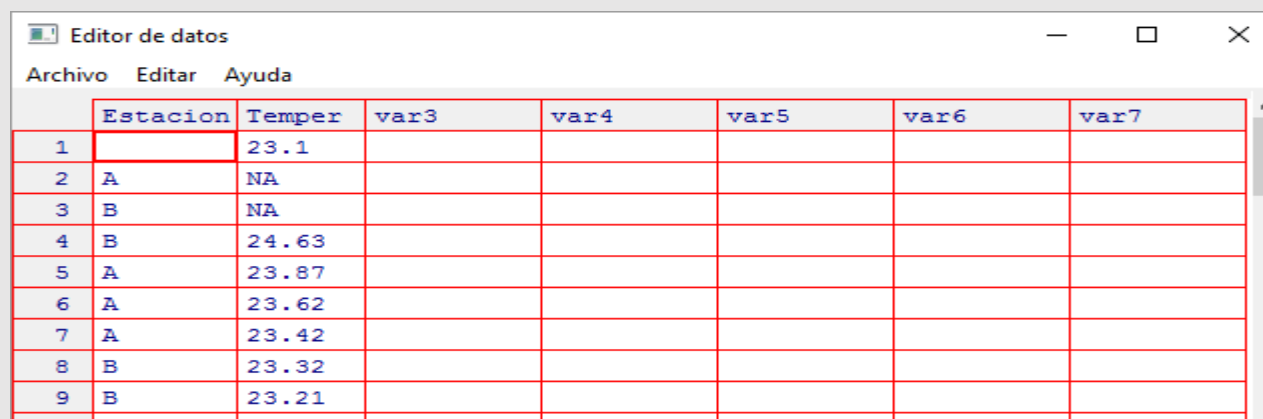
> temper[1,1] <- NA

> temper[2,2] <- NA

> temper[3,2] <- NA

Editar **Temper**

> edit(temper)



	Estacion	Temper	var3	var4	var5	var6	var7
1		23.1					
2	A	NA					
3	B	NA					
4	B	24.63					
5	A	23.87					
6	A	23.62					
7	A	23.42					
8	B	23.32					
9	B	23.21					





Detección de valores perdidos: ejemplo 1

```
> ## ¿El data frame tiene algun dato perdido?
```

```
> anyNA(temper)
```

```
[1] TRUE
```

```
> ## ¿Cuántos faltantes hay en cada variable?
```

```
> summary(temper)
```

Estacion	Temperatura
A :52	Min. :16.11
B :52	1st Qu.:18.55
C :51	Median :19.79
NA's: 1	Mean :19.95
	3rd Qu.:21.21
	Max. :32.00
	NA's :2



Detección de valores perdidos: ejemplo 1

```
> ## ¿Dónde están los datos perdidos y cuántos?
```

```
> which(is.na(temper$Temper))
```

```
[1] 2 3
```

En las posiciones 2 y 3

```
> which(is.na(temper$Estacion))
```

```
[1] 1
```

En la posición 1





Detección de valores perdidos: ejemplo 1

```
> ## ¿Qué filas contienen casos incompletos?
```

```
> which(!complete.cases(temper))
```

```
[1] 1 2 3
```

En los individuos 1, 2 y 3

```
> ## ¿Qué filas contienen casos completos?
```

```
>
```

```
which(complete.cases(temper))
```

```
[1] 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
```

```
.....
```





Acciones ante valores perdidos

na.rm()

Algunas funciones como **mean**, **sum**, **prod**, **quantile**, **sd**, tienen la opción de omitir los valores perdidos. Las funciones de estadística k-dimensional como **cor** y **cov** ofrecen, además, la posibilidad de incluir tratamiento (**listwise**, **pairwise**, etc)

complete.cases ()

Detecta las filas de un data.frame que no contienen valores perdidos

na.omit() na.exclude()

Omite los casos de un data.frame que contienen valores perdidos por columnas o en total.

na.fail()

Sólo trabajará con los datos si no hay ningún datos perdidos.

na.pass()

Trabjará con los datos aunque hayan datos perdidos.



Tratamiento de valores perdidos: ejemplo 2



Cargar datos

```
detach(temper)
```

```
attach(datos1)
```

Insertamos algunos faltantes

```
datos1[c(1,4,5,13,15), "ESTATURA"] <- NA
```

```
datos1[c(2,7), "PESO"] <- NA
```

```
datos1[c(2,4,5,7,11), "GASTO"] <- NA
```

```
View(datos1)
```



Tratamiento de valores perdidos: ejemplo 2



Observad las diferencias

Calculo de la correlacion con eliminación de perdidos por casos

```
cor(datos1[,c("PESO", "ESTATURA", "GASTO")], use="complete.obs")
```

	PESO	ESTATURA	GASTO
PESO	1.00000000	0.73242059	0.07315068
ESTATURA	0.73242059	1.00000000	0.02395298
GASTO	0.07315068	0.02395298	1.00000000

Calculo de la correlacion con eliminación de perdidos por pares.

```
cor(datos1[,c("PESO", "ESTATURA", "GASTO")], use="pairwise.complete.obs")
```

	PESO	ESTATURA	GASTO
PESO	1.00000000	0.73338344	0.08764238
ESTATURA	0.73338344	1.00000000	0.02395298
GASTO	0.08764238	0.02395298	1.00000000





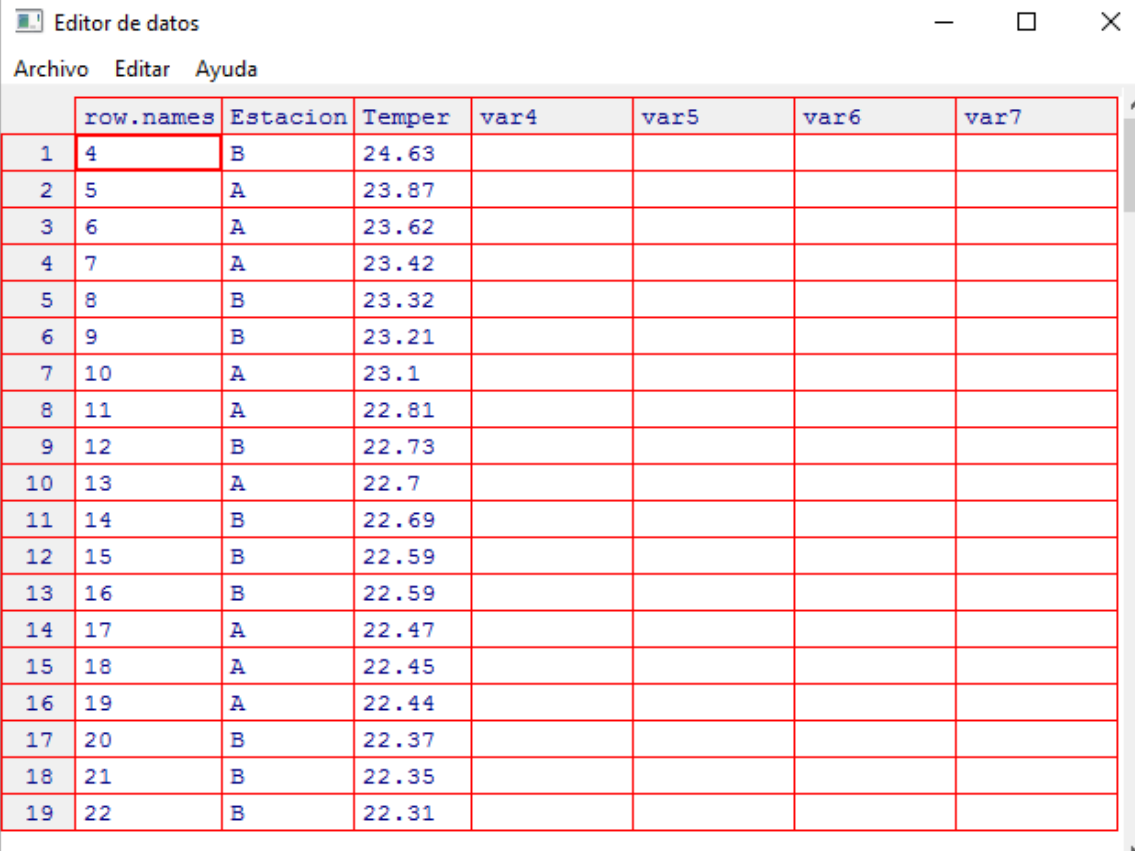
Eliminación de valores perdidos

> ## Eliminamos filas con datos faltantes y creamos un nuevo data frame

> Temper.sin <- **na.exclude**(temper)

> anyNA(Temper.sin)

[1] FALSE



	row.names	Estacion	Temper	var4	var5	var6	var7
1	4	B	24.63				
2	5	A	23.87				
3	6	A	23.62				
4	7	A	23.42				
5	8	B	23.32				
6	9	B	23.21				
7	10	A	23.1				
8	11	A	22.81				
9	12	B	22.73				
10	13	A	22.7				
11	14	B	22.69				
12	15	B	22.59				
13	16	B	22.59				
14	17	A	22.47				
15	18	A	22.45				
16	19	A	22.44				
17	20	B	22.37				
18	21	B	22.35				
19	22	B	22.31				



Imputación de valores perdidos

- Los métodos de imputación consisten en estimar los valores ausentes en base a los valores válidos de otras variables y/o casos de la muestra.
- La estimación se puede hacer a partir de la información del conjunto completo de variables o bien de algunas variables especialmente seleccionadas.
- En este apartado veremos someramente algunas de las técnicas de imputación numéricas más ampliamente utilizadas.
- Aunque , la mayoría de los métodos de imputación se utilizan con variables cuantitativas, hay alguno que puede usarse también con cualitativas.
- Principales **procedimientos**:
 - **Sustitución por la Media u otro parámetro**
 - **Sustitución por constante**
 - **Imputación por regresión**



Sustitución por la Media

- Consiste en sustituir el valor ausente por la **media** de los valores válidos.

$$\hat{x}_i = \bar{x}_i$$

- Este procedimiento plantea inconvenientes como:
 - Dificulta la estimación de la varianza.
 - Distorsiona la verdadera distribución de la variable
 - Distorsiona la correlación entre variables dado que añade valores constantes.
- Una variante en datos muy asimétricos es utilizar la **mediana** en vez de la media.

$$\hat{x}_i = Me_i$$

- Existen otros procedimientos numéricos como el *ratio imputation*



Imputación numérica

```
##Cargar paquete Hmisc
> library(Hmisc)
## Crear vector
> x<-1:5
> x
[1] 1 2 3 4 5
>
> ## Generar un dato anómalo
> x[2]<- NA
> x
[1] 1 NA 3 4 5

## Calcular la media
> mean(x, na.rm = T)
[1] 3.25

> ## Calcular la mediana
> median(x, na.rm = T)
[1] 3.5
```





Imputación numérica

```
## Imputar la media  
> x.imedia<- impute(x, mean)  
> x.imedia
```

1	2	3	4	5
1.00	3.25*	3.00	4.00	5.00

```
## Imputar la mediana  
> x.imedian<- impute(x, median)  
> x.imedian
```

1	2	3	4	5
1.0	3.5*	3.0	4.0	5.0



Sustitución por otras constantes

- Consiste en sustituir los valores ausentes por constantes cuyo valor viene determinado por razones teóricas o relacionadas con la investigación previa.
- Presenta los mismos inconvenientes que la sustitución por la Media, y solo debe ser utilizado si hay razones para suponer que es más adecuado que el método de la media.
- Sin embargo, puede ser utilizado en la imputación sobre variables cualitativas.
- El procedimiento principal es imputación **Hot-Deck**



Imputación numérica

```
## Imputar la constante 100
```

```
> x.iconst<- impute(x, 100)
```

```
> x.iconst
```

1	2	3	4	5
1	100*	3	4	5

```
## Imputar un número aleatorio de entre los de  
la variable
```

```
> x.irnd<- impute(x, "random")
```

```
> x.irnd
```

1	2	3	4	5
1	3*	3	4	5

Consideraciones sobre la imputación

- Aunque en este apartado hemos visto sólo algunos de los métodos más utilizados para variables unidimensionales, existe gran cantidad de información relativa a otros métodos y paquetes.
- No existe un único método que sea válido para todas las situaciones.
- Los métodos de imputación deben aplicarse con gran precaución porque pueden introducir relaciones inexistentes en los datos reales.

Si la muestra es suficientemente grande y los datos faltantes son completamente aleatorios y no siguen ningún patrón, lo óptimo es eliminar casos y/o variables.

Consideraciones sobre la imputación

- Para un conocimiento más profundo de los aspectos teóricos y prácticos sobre los datos faltantes es muy recomendable el capítulo correspondiente [de Hair, J.F. et al., 2005. Multivariate Data Analysis 6th ed., Prentice Hall](#)
- En [De Jonge, E., & van der Loo, M. \(2013\)](#) se puede encontrar más detalles sobre éstas técnicas, paquetes de R y referencias.
- **Paquetes** relacionados:
 - MICE
 - Amelia
 - missForest
 - Hmisc
 - mi

Valores anómalos o atípicos o *outliers*

- Son observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos. Los datos atípicos son ocasionados por:
 - a) Errores de procedimiento en la recogida de datos.
 - b) Acontecimientos extraordinarios.
 - c) Valores extremos.
 - d) Causas no conocidas.
- Los datos anómalos distorsionan los resultados de los análisis.
- Los valores anómalos NO son necesariamente errores, por lo que han de ser detectados, pero no eliminados sistemáticamente.
- La inclusión o no en el análisis es una decisión del analista o investigador.

Detección de valores anómalos

- La herramienta más habitual es el diagrama de caja y bigotes.
- También se puede usar el Papel Probabilístico Normal y de paso verificar la normalidad.
- Existen varios tests, pero no todos se comportan bien con muestras grandes.
- **Tratamiento:**
 1. **Corrección**, si se trata de un error y es posible
 2. **Transformación** de los datos (log, sqrt)
 3. Si sólo necesitamos **descriptivos**, usar **parámetros robustos** (mediana, cuartiles, etc)
 4. **Eliminación** (en último caso)

Detección de valores anómalos



```
## Para determinar e identificar numéricamente los outliers y  
otros parámetros del boxplot  
boxplot.stats(temper$Temperatura, coef=2)
```

\$stats

[1]	16.11	18.54	19.79	21.22	24.63	
	Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo

\$n

[1]	154	Número de datos no NA
-----	-----	-----------------------

\$conf

[1]	19.44878	20.13122	Intervalo que contiene los datos considerados no anómalos (función de coef)
-----	----------	----------	--

\$out

[1]	32	Valores de los datos considerados anómalos. Sólo hay uno.
-----	----	--



Detección gráfica de valores anómalos



- **Funciones**

- `boxplot()`
- `identify()`

- **Datos**

Temperatura recogidas por 3 estaciones meteorológicas

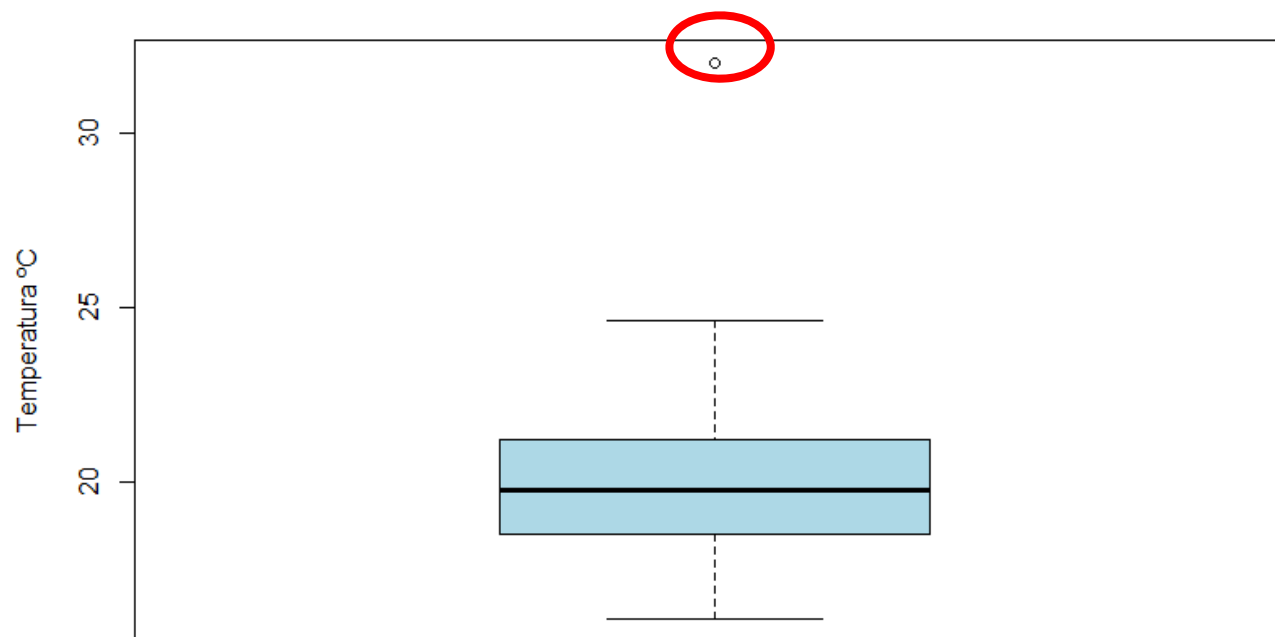
- Var "Temperatura"
- Dataset "temper"
- Archivo "datos.RData"



Detección gráfica de valores anómalos



```
## Temperatura recogidas por 3 estaciones meteorológicas  
## Var "Temperatura". Dataset "temper". Archivo "datos.RData"  
  
#Diagrama de caja (default)  
boxplot(temper$Temperatura, xlab="Todas las estaciones",  
ylab="Temperatura °C", main="¿Hay datos anómalos?",  
col="lightblue")
```



Todas las estaciones

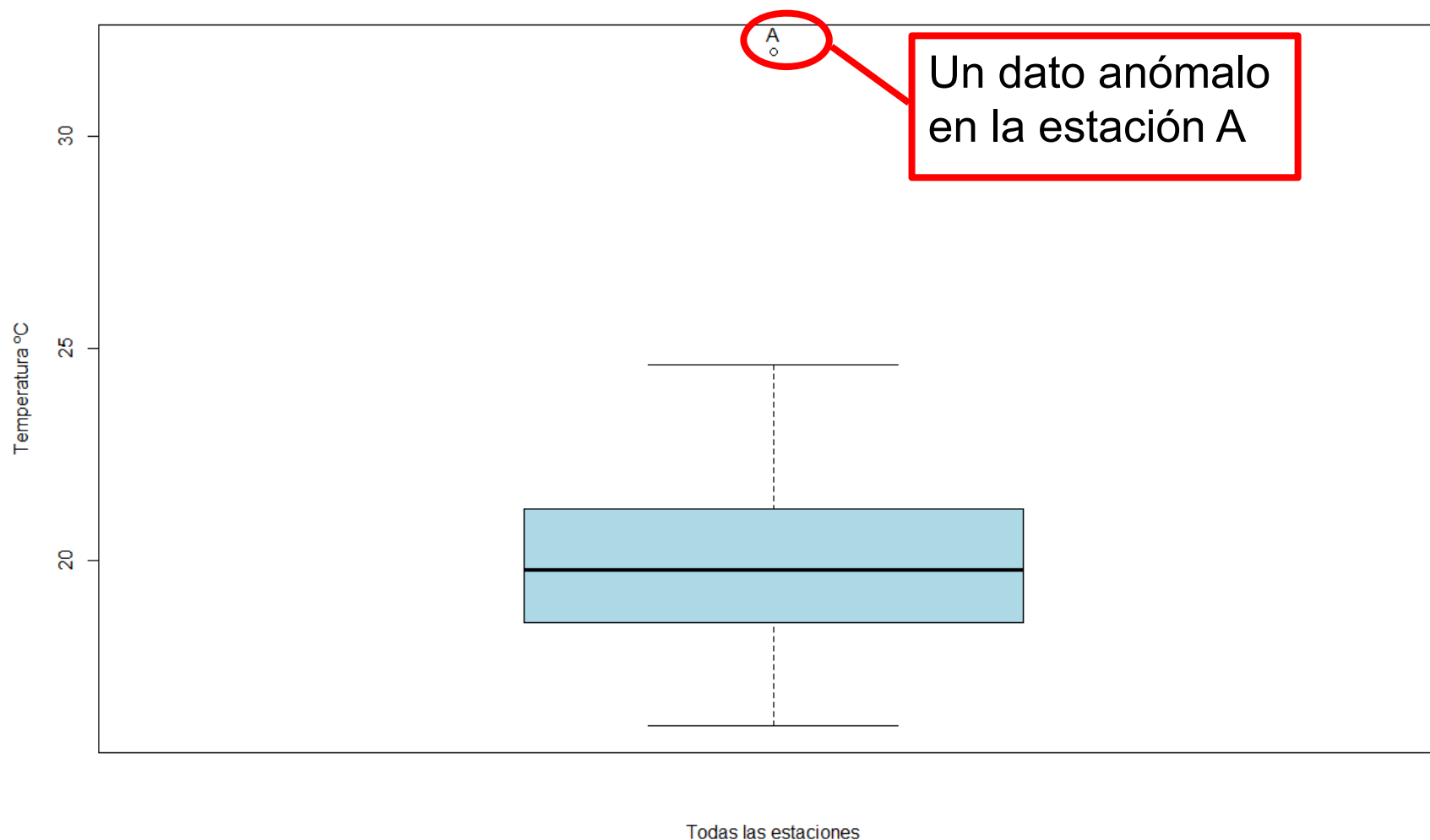




Detección gráfica de valores anómalos

```
## Identificar datos anómalos por estación
```

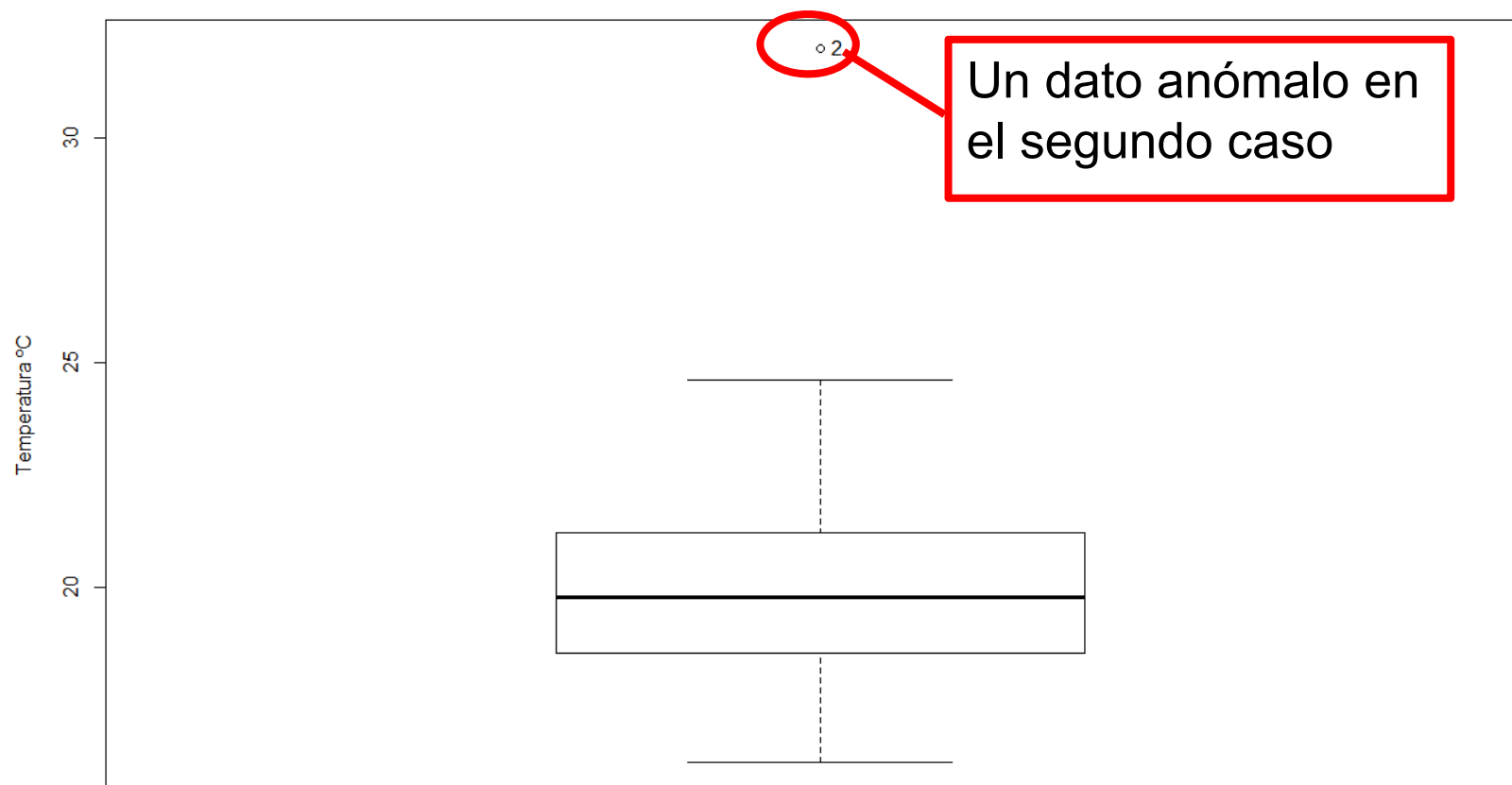
```
identify(rep(1,length(temper$Temperatura)), temper$Temperatura,  
temper$Estacion)
```





Detección gráfica de valores anómalos

```
## Identificar datos anómalos por número de fila  
boxplot(temper$Temperatura, ylab="Temperatura °C")  
identify(rep(1, length(temper$Temperatura)), temper$Temperatura,  
rownmes(temper))
```



Detección gráfica de valores anómalos



– Funciones

- `qqnorm()` : dibuja los datos sobre papel probabilístico normal
- `qqline()` : dibuja los puntos teóricos de probabilidad suponiendo que los datos siguen una distribución normal

– Datos

Temperatura recogidas por 3 estaciones meteorológicas

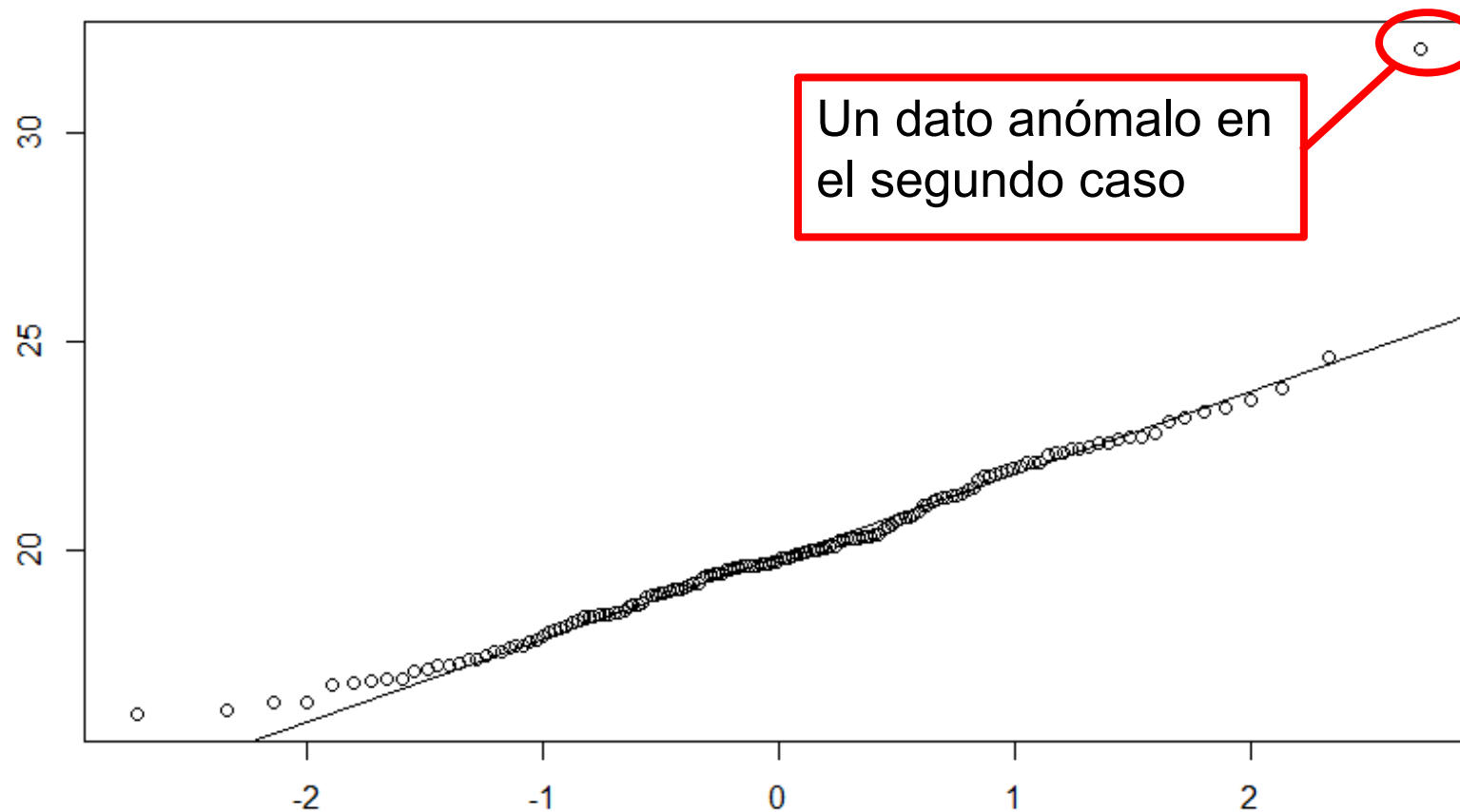
- Var "Temperatura"
- Dataset "temper"
- Archivo "datos.RData"



Detección gráfica de valores anómalos



```
## library(car)  
> qqnorm(Temperatura)  
> qqline(Temperatura)
```



2. Integración

- Seleccionar o filtrar datos
- Ordenar datos
- Combinar datos
- Agregar datos / Añadir nuevas variables
- Paquetes adicionales: **dplyr**

Seleccionar datos

- Selección de un subconjunto de casos del *data frame* a partir de una condición:
 - Por valores de los índices
newdata <- mydata[1:5,]
newdata <- mydata[,6:7]
 - Según valores de variables
newdata <- mydata[**which**(condición1, condición2)]
- La forma más práctica: función **subset()**
 - Permite filtrar por casos y/o por variables

Filtrar según variable



```
> ##### Filtrado de variables
```

```
> ## Creamos un nuevo data frame que contenga solo las  
variables EDAD, ESTATURA y PESO
```

```
> nuevos1<-subset(datos1, select = c("EDAD", "ESTATURA", "PESO"))
```

```
> head(nuevos1)
```

	EDAD	ESTATURA	PESO
1	20	NA	54
2	20	164	NA
3	19	185	70
4	19	NA	63
5	20	NA	63
6	23	159	54

```
>
```



Filtrar según casos



```
> ##### Filtrado de casos
```

```
> nuevos2<-subset(datos1, SEXO=="mujer" & EDAD > 20)
```

```
> head(nuevos2, 10)
```

	SEXO	EDAD	MES	ESTATURA	PESO	PROVINCIA	X	ACCESOS	TIPO	GASTO	COMPRA1	GASTO2	COMPRA2
6	mujer	23	10	159	54	Alicante	5	65	yoga	10.902	no	12.780	si
7	mujer	21	5	160	NA	Alicante	7	45	yoga	NA	no	9.619	si
8	mujer	21	4	155	48	Alicante	3	30	fitness	10.097	no	6.895	si
10	mujer	22	4	172	59	Alicante	7	45	yoga	8.964	no	10.015	si
12	mujer	21	9	160	49	Alicante	4	30		5.856	no	11.422	si
14	mujer	21	9	163	56	Alicante	3	45	fitness	9.617	no	11.842	si
30	mujer	22	6	162	48	Castellon	3	75	aparatos	11.403	no	9.714	si
31	mujer	21	6	170	63	Castellon	7	65	aparatos	10.133	si	10.573	si
34	mujer	21	11	163	54	Castellon	7	10	aparatos	9.154	no	5.362	no
37	mujer	21	4	160	55	Castellon	4	30	yoga	11.639	no	7.436	no



Filtrar según variable y caso



```
> ##### Filtrado de variables y casos
> ## Creamos un nuevo data frame que contenga el SEXO, ESTATURA
y PESO
> ## de las mujeres mayores de 20 años

> nuevos3<-subset(datos1, SEXO=="mujer" & EDAD > 20, select =
c("EDAD", "ESTATURA", "PESO"))

> head(nuevos3, 10)
  EDAD ESTATURA PESO
6    23      159  54
7    21      160 NA
8    21      155  48
10   22      172  59
12   21      160  49
14   21      163  56
30   22      162  48
31   21      170  63
34   21      163  54
37   21      160  55
```



Ordenar datos

- `sort(x, decreasing = FALSE, na.last = NA, ...)`
- `order(x, decreasing = FALSE, na.last = NA, ...)`
- Más práctico usar **order()** en data frames, para ordenar sus índices

Ordenar



```
> ## Ordenar todos los casos de un data frame en función de una  
variable (p.e. EDAD) en orden creciente
```

```
> orden1<-datos1[order(MES, decreasing = F) ,]
```

```
> head(orden1)
```

	SEXO	EDAD	MES	ESTATURA	PESO	PROVINCIA	X	ACCESOS	TIPO	GASTO	COMPRA1	GASTO2	COMPRA2
20	varon	21	1	180	72	Alicante	3	15	aparatos	12.289	no	8.427	si
21	varon	20	1	171	75	Alicante	8	15	yoga	13.391	no	11.128	si
33	mujer	20	1	161	48	Castellon	5	60	pilates	10.464	si	11.458	no
40	mujer	21	1	161	46	Castellon	3	30	fitness	6.184	no	9.756	no
42	varon	20	1	183	76	Castellon	5	15	pilates	8.346	no	5.672	no
45	varon	19	1	179	67	Castellon	7	30	pilates	5.544	no	7.241	no



Ordenar



```
> ## Ordenar todos los casos de un data frame en función  
de una variable (p.e. EDAD) en orden decreciente
```

```
> orden2<-datos1[order(MES, decreasing = T) ,]
```

```
> head(orden2)
```

	SEXO	EDAD	MES	ESTATURA	PESO	PROVINCIA	X	ACCESOS		TIPO	GASTO	COMPRA1	GASTO2	COMPRA2
12	mujer	21	9	160	49	Alicante	4	30			5.856	no	11.422	si
14	mujer	21	9	163	56	Alicante	3	45	fitness		9.617	no	11.842	si
28	varon	21	9	180	90	Alicante	7	15	fitness		10.654	no	9.279	si
46	varon	20	9	171	71	Castellon	3	13	fitness		9.490	no	11.628	no
59	varon	20	9	178	69	Castellon	7	45			11.755	no	12.728	no
68	mujer	23	9	165	55	Teruel	2	35	aparatos		11.068	no	7.247	no



Ordenar



```
> ## Ordenar todos los casos de un data frame en función  
de dos variable (p.e. MES y EDAD) orden creciente
```

```
> orden3<-datos1[order(MES, EDAD) ,]
```

```
> head(orden3)
```

	SEXO	EDAD	MES	ESTATURA	PESO	PROVINCIA	X	ACCESOS	TIPO	GASTO	COMPRA1	GASTO2	COMPRA2
45	varon	19	1	179	67	Castellon	7	30	pilates	5.544	no	7.241	no
21	varon	20	1	171	75	Alicante	8	15	yoga	13.391	no	11.128	si
33	mujer	20	1	161	48	Castellon	5	60	pilates	10.464	si	11.458	no
42	varon	20	1	183	76	Castellon	5	15	pilates	8.346	no	5.672	no
61	mujer	20	1	168	61	Teruel	7	7	aparatos	7.942	no	8.970	no
20	varon	21	1	180	72	Alicante	3	15	aparatos	12.289	no	8.427	si



Ordenar



```
> ##### Ordenar todos los casos de un data frame en
función de una variable (p.e. EDAD) orden creciente,

> ## dejando al final los valores perdidos
> orden4<-datos1[order(MES, EDAD, na.last = T) ,]

> tail(orden4)
```

	SEXO	EDAD	MES	ESTATURA	PESO	PROVINCIA	X	ACCESOS	TIPO	GASTO	COMPRA1	GASTO2	COMPRA2
130	varon	22	9	174	80	Valencia	7	20	yoga	10.104	no	9.474	no
68	mujer	23	9	165	55	Teruel	2	35	aparatos	11.068	no	7.247	no
128	varon	23	9	176	74	Valencia	6	10	yoga	12.057	no	7.354	no
3	varon	19	<NA>	185	70		5	30	pilates	8.541	no	10.426	no
1	mujer	20	<NA>	NA	54		3	30	pilates	6.124	si	7.385	no
2	mujer	20	<NA>	164	NA		4	10	pilates	NA	si	10.980	no



Agregar datos

- Añadir casos
- Añadir nuevas variables
 - merge()
 - rbind()
 - cbind()

Agregar nuevas variables



```
> ## Simulamos una nueva variable que contiene la precipitación
media en l/m2
> x<-rnorm(length(Temperatura), 300, 50)
>
> ## Añadimos la nueva variable al data frame
> temper$Precipitacion<-x
>
> head(temper)
```

	Estacion	Temperatura	Precipitacion
1	A	NA	278.5127
2	A	32.00	197.1572
3	B	NA	298.3071
4	B	22.73	292.6677
5	<NA>	23.42	348.4969
6	C	19.56	293.0860





Agregar nuevos casos

```
> ## Creamos dos grupos de datos a partir de datos1
> muestra1<-datos1[1:5,]
> muestra2<-datos1[6:10,]
>
> ## Combinamos los casos de ambos data frames
> muestra<-rbind(muestra1, muestra2)

> head(muestra)
```

	SEXO	EDAD	MES	ESTATURA	PESO	PROVINCIA	X	ACCESOS	TIPO	GASTO	COMPRA1	GASTO2	COMPRA2
1	mujer	20	<NA>	NA	54		3	30	pilates	6.124	si	7.385	no
2	mujer	20	<NA>	164	NA		4	10	pilates	NA	si	10.980	no
3	varon	19	<NA>	185	70		5	30	pilates	8.541	no	10.426	no
4	varon	19	12	NA	63		8	20	fitness	NA	no	12.320	no
5	varon	20	6	NA	63		5	12	pilates	NA	no	9.394	si
6	mujer	23	10	159	54	Alicante	5	65	yoga	10.902	no	12.780	si



3. Transformación

- Para aplicar la técnica adecuada estadística, necesitamos que las variables tengan el tipo y la distribución de frecuencias correctas según el caso.
- Esto nos puede llevar a necesitar transformar la variable para que sea manejable.
- Las transformaciones necesarias más comunes son:
 - Normalización de variables cuantitativas (escalado)
 - Transformaciones exponencial y logarítmica
 - Discretización de variables cuantitativas en categorías

Normalización de variables cuantitativas

- La normalización, estandarización o tipificación de una variable consiste en la siguiente transformación lineal:

$$Z = \frac{X - \mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma}$$

- La variable tipificada expresa el número de desviaciones típicas que cada observación dista de la media (adimensional).
- Por ello, se puede comparar la posición relativa de los datos de diferentes distribuciones, siendo posible comparar variables medidas sobre diferentes escalas o magnitudes.
- Es sencillo programar el cálculo de la normalización en R, pero éste dispone de funciones como **scale()**

Normalización de variables cuantitativas



```
> # Estandarizar las variables ESTATURA y PESO y añadirlas al
dataset datos1

> datos1$est.norm <- scale(ESTATURA, center = T, scale = T)

> datos1$pes.norm <- scale(PESO, center = T, scale = T)

> summary(datos1[,c("ESTATURA", "PESO", "est.norm",
"pes.norm")])
```

ESTATURA	PESO	est.norm.V1	pes.norm.V1
Min. :152.0	Min. :45.00	Min. :-2.295127	Min. :-2.0336382
1st Qu.:165.2	1st Qu.:58.00	1st Qu.: -0.846464	1st Qu.: -0.8015532
Median :174.0	Median :67.00	Median : 0.110201	Median : 0.0514287
Mean :173.0	Mean :66.46	Mean : 0.000000	Mean : 0.0000000
3rd Qu.:179.0	3rd Qu.:74.00	3rd Qu.: 0.656866	3rd Qu.: 0.7148591
Max. :198.0	Max. :90.00	Max. : 2.734195	Max. : 2.2312714
NA's :5	NA's :2	NA's :5	NA's :2



Transformaciones exponencial y logarítmica

- Son transformaciones no lineales que se usan generalmente para **corregir la asimetría o kurtosis** de la distribución de frecuencias de una variable.
- Si se tienen distribuciones de frecuencias con **asimetría negativa** (frecuencias altas hacia el lado derecho de la distribución), es conveniente aplicar una transformación que comprima la escala para valores pequeños y la expande para valores altos:

$$Y = X^2$$

- Para distribuciones con **asimetría positiva** se usan las transformaciones que compriman los valores altos y expanden los pequeños:

$$Y = \sqrt{X}$$

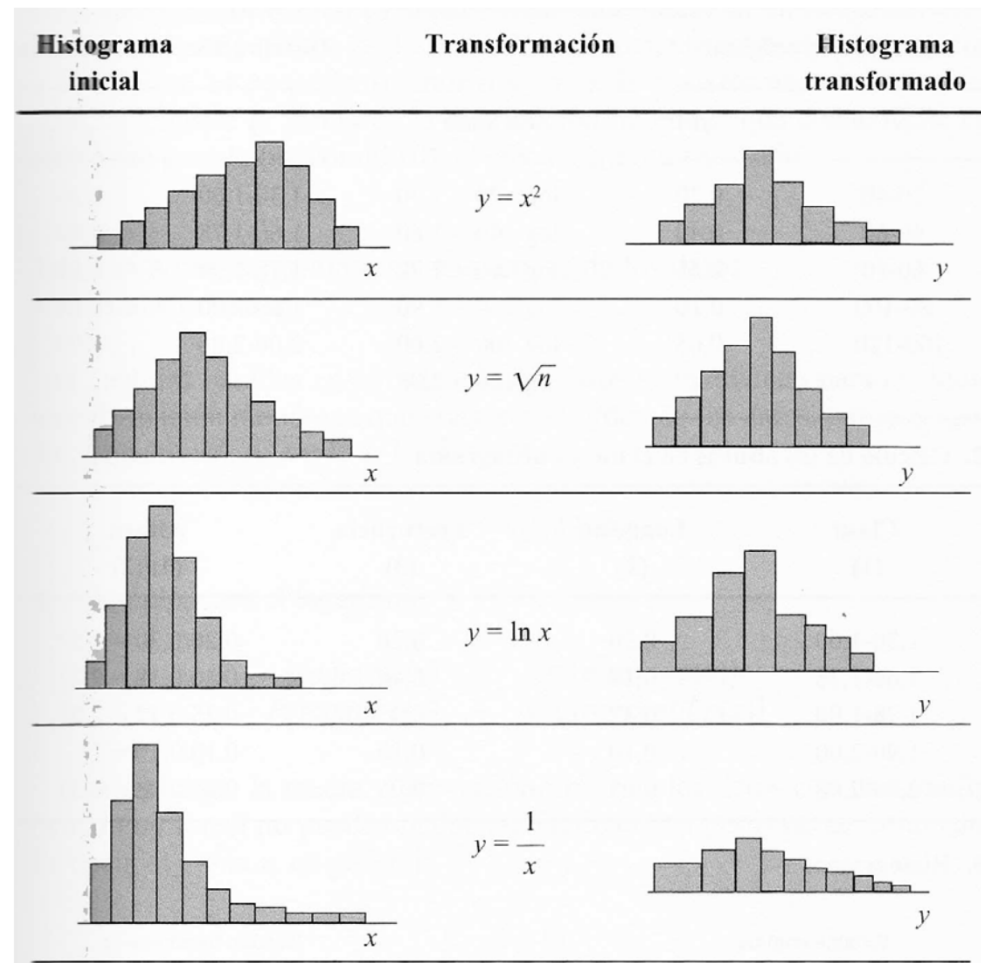
$$Y = \ln(X)$$

$$Y = 1/X$$

El efecto de estas transformaciones está en orden creciente: menos efecto \sqrt{x} , más $\ln(x)$ y más aún $1/x$.

Transformaciones exponencial y logarítmica

- La transformación más utilizada es la del logaritmo. Muchas distribuciones de datos económicos, o de consumos se convierten en simétricas al tomar la transformación logaritmo.



Discretización de variables cuantitativas



```
> table(ESTATURA)
ESTATURA
152 155 156 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182
183 185 186 190 191
  1   2   1   2   1   5   2   5   4   2   7   1   2   5   2   5   2   7   4  11  12   3   1   5   4   8   1   2
4   7   2   1   1
192 195 196 198
  1   1   1   1
```



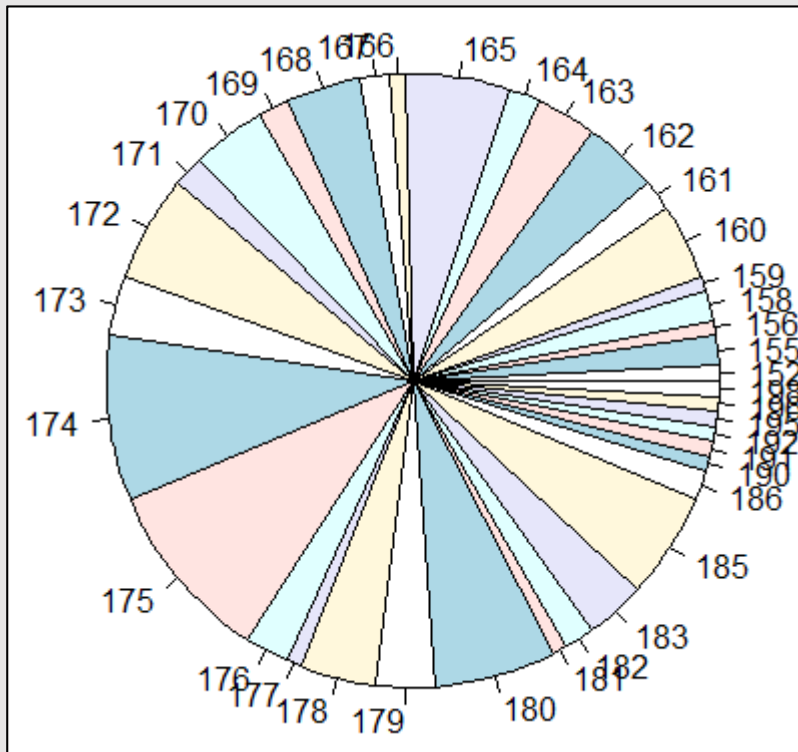
```
> ## Discretizar con 5 tramos de estatura
> est.disc <- cut(ESTATURA, breaks = 5)
>
> ## Tabular la nueva variable
> table(est.disc)
est.disc
(152,161] (161,170] (170,180] (180,189] (189,198]
          14         33         49         24         6
```



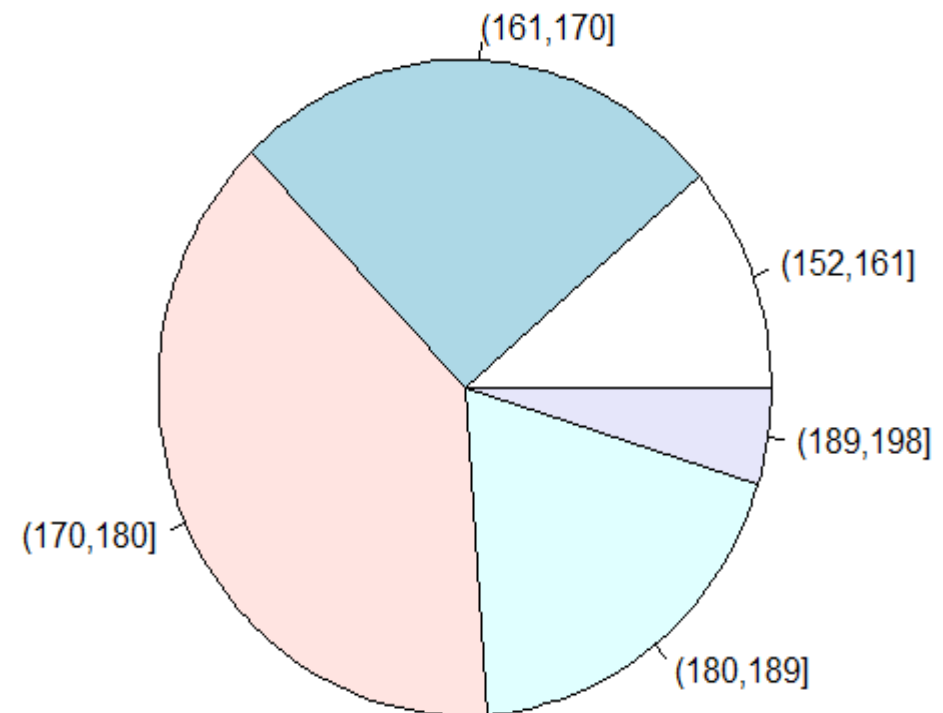
Discretización de variables cuantitativas



```
> pie(table(ESTATURA))
```



```
> pie(table(est.disc))
```



4. Reducción

- Cuando se dispone de un *data set* con muchas variables, puede ser necesario reducir el *data set* a un menor número de variables o de casos perdiendo la menor cantidad de información posible.
- Las técnicas estadísticas en estas situaciones más habituales son:
 - Análisis de Componentes Principales (descriptivo)
 - Análisis Factorial (inferencia)
 - Análisis de Correspondencias (descriptivo)

Análisis de Componentes Principales PCA

Objetivo: condensar la información contenida en J variables originales (correlacionadas) en un número reducido de nuevas variables (incorrelacionadas) llamadas **componentes** (F) definidas como combinaciones lineales de las J variables primitivas ($F \ll J$)

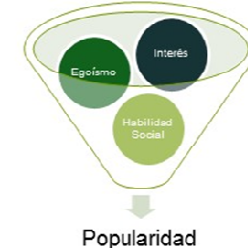
- Situaciones complejas que no pueden resolverse actuando sobre una única variable aleatoria (necesitan muchas variables).
 - imagen de un producto en el mercado
 - satisfacción de los clientes con un producto o servicio
 - actitud de compra de los clientes frente a un producto....
- Características que no pueden medirse directamente:
 - características de un vino
 - popularidad de una persona...



Análisis de Componentes Principales PCA

- **Ejemplo:** en el contexto de alguna investigación sobre la popularidad de los alumnos de la UPV¹ se han medido aspectos como:
 - habilidades sociales personales (**HSOC**)
 - egoísmo (**EGOIS**)
 - capacidad de despertar interés en otros (**INTER**)
 - proporción de tiempo en una conversación, que una persona es capaz de hablar sobre la otra (**CONVO**)
 - proporción de tiempo en una conversación, que una persona es capaz de hablar sobre ellos mismos (**CONVA**)
 - la predisposición a mentir (**MENT**)

Variables observadas



¹Ejemplo extraído de Andy Field, Jeremy Miles, and Z.F., 2012. *Discovering statistics using R*, SAGE Publications.

¿Qué nos dicen las correlaciones?

R	CONVO	HSOC	INTER	CONVA	EGOIS	MENT
CONVO	1,000	Dimensión 1				
HSOC	0,772	1,000				
INTER	0,646	0,879	1,000			
CONVA	0,074	-0,120	0,054	1,000	Dimensión 2	
EGOIS	-0,131	0,031	-0,101	0,441	1,000	
MENT	0,068	0,012	0,110	0,361	0,277	1,000

- Se observan dos “grupos de correlación”: dos grupos de **variables** que correlacionan entre ellas, pero no con el resto.
- Se podría decir que cada grupo puede medir una misma dimensión subyacente común o **componente**

¿Qué nos dicen las correlaciones?

- Si conseguimos de algún modo obtener estas componentes (variables no observables) a partir de las variables originales, habremos resumido los datos a un número menor de variables que explican la máxima cantidad posible de la “información” común de nuestros datos observados inicialmente.
- En el ejemplo, las componentes podrían interpretarse como:
 - *sociabilidad general*
 - *consideración*

Representación matemática

- Las K componentes pueden expresarse como una combinación lineal de las J variables observadas ($K < J$):

$$C_k = b_{1k} X_1 + b_{2k} X_2 + \cdots + b_{Jk} X_J$$

Los coeficientes b_{jk} son los llamados **pesos** o **cargas**:

- Dan una idea del **peso** o importancia relativa que las variables observadas tienen sobre cada componente.
- Se representan como **matriz de pesos** (**A**):

$$A = \begin{pmatrix} b_{11} & \cdots & b_{K1} \\ \vdots & \ddots & \vdots \\ b_{1J} & \cdots & b_{KJ} \end{pmatrix}$$

Representación matemática

Ejemplo: Las variables observadas pueden expresarse como una combinación lineal de las componentes *Sociabilidad General* (C1) y *Consideración* (C2):

$$C1 = b_{11} CONVO + b_{21} HSOC + b_{31} INTER + b_{41} CONVA + b_{51} EGOIS + b_{61} MENT$$

$$C2 = b_{12} CONVO + b_{22} HSOC + b_{32} INTER + b_{42} CONVA + b_{52} EGOIS + b_{62} MENT$$

$$C1 = 0,87 CONVO + 0,96 HSOC + 0,92 INTER - 0,01 EGOIS + 0,09 MENT$$

$$C2 = 0,01 CONVO - 0,03 HSOC + 0,04 INTER + 0,82 CONVA + 0,75 EGOIS + 0,7 MENT$$

$$A = \begin{pmatrix} 0,87 & 0,01 \\ 0,96 & -0,03 \\ 0,92 & 0,04 \\ 0,00 & 0,82 \\ -0,01 & 0,75 \\ 0,09 & 0,70 \end{pmatrix}$$

Representación matemática

$$C1 = 0,87 CONVO + 0,96 HSOC + 0,92 INTER - 0,01 EGOIS + 0,09 MENT$$

$$C2 = 0,01 CONVO - 0,03 HSOC + 0,04 INTER + 0,82 CONVA + 0,75 EGOIS + 0,7 MENT$$

Se observa que para la componente de sociabilidad (C1), los pesos son altos para las habilidades sociales (HSOC), la capacidad de hablar sobre otros (CONVO) y la de despertar interés en otras personas (INTER) y muy bajos (cercanas a 0) para el egoísmo (EGOIS), la capacidad de mentir (MENT) y la proporción de tiempo hablando de uno mismo (CONVA). En el caso de la componente *consideración* (C2) ocurre al revés.

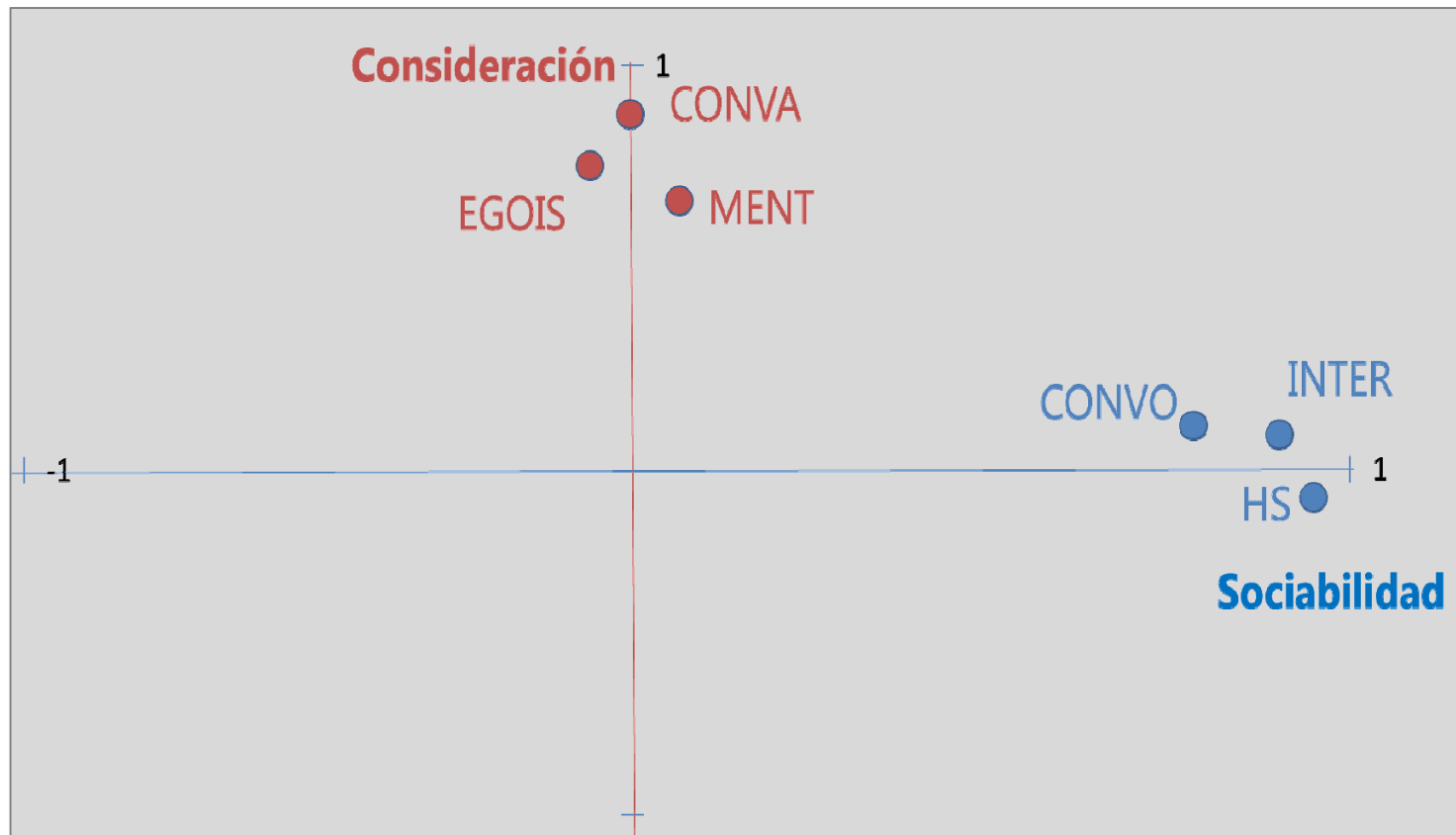
Representación matemática

$$C_k = b_{1k} X_1 + b_{2k} X_2 + \cdots + b_{Jk} X_J$$

- C1, C2 ... y Ck son nuevas variables aleatorias que resumen la información de las variables observadas y que podrían utilizarse como sustitutas de las variables originales en posteriores análisis estadísticos.
- Estas nuevas variables o factores tienen diferentes valores para los distintos individuos de la muestra y se denominan **puntuaciones factoriales**.

Representación gráfica

Las componentes pueden considerarse como ejes de clasificación a lo largo de los cuales pueden representarse las variables observadas utilizando los pesos como coordenadas:



Fases del PCA



1 Identificación y extracción

¿Cuántas dimensiones subyacentes hay?



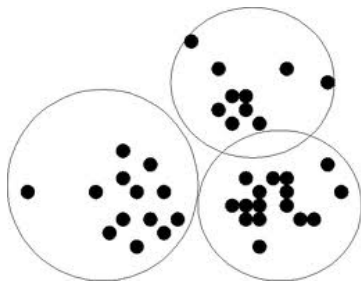
¿1 componente
o ninguna?



¿9 componentes o
ninguno?



¿3 componentes?



identificar los grupos de variables que
tienen “algo” en común entre ellas
(**correlación**), pero no con el resto

Técnicas

¿Cuántas componentes vamos a extraer?

Existen varios criterios y como no hay unanimidad y algunos de ellos son complementarios, tendremos que tenerlos todos en cuenta.

Características de las componentes

- No observables directamente
- Incorrelacionadas entre sí
- La primera componente principal explica tanto de la variabilidad de los datos (varianza) como sea posible
- Cada componente principal obtenida a partir de la primera debe explicar tanto de la variabilidad restante como sea posible
- Los vectores que definen las **K** componentes principales son los **vectores propios** asociados a los **K mayores valores propios** $\lambda_1, \lambda_2, \dots, \lambda_K$ de la **matriz de correlación** R^1 de las J variables primitivas

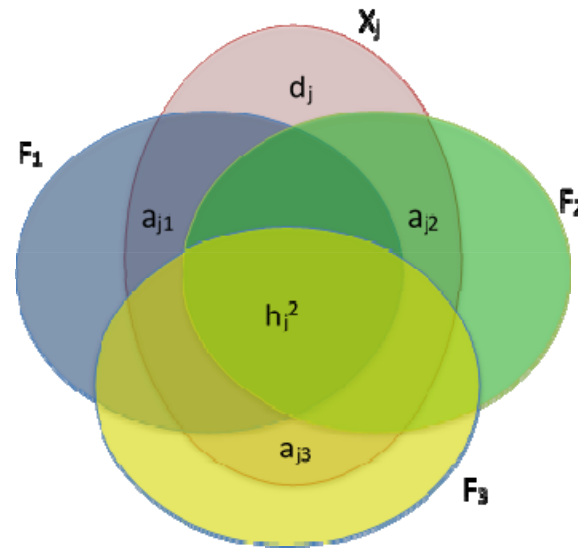
¹ En realidad es la matriz de varianzas-covarianzas, que si las variables están tipificadas coincide con R

Varianza compartida

- Cada variable observada (*egoísmo*, etc.) tiene una determinada varianza, pues no todos los alumnos tienen el mismo nivel de *egoísmo*, por ejemplo.
- A esta varianza se le denomina **varianza total**. Si las variables están tipificadas, la varianza puede ser **1** como máximo.
- Si dos variables están correlacionadas positivamente, como las habilidades sociales y el interés que un alumno despierta en sus semejantes, dichas variables compartirán parte de su varianza total. A esa parte compartida se le denomina **varianza común**.
- El resto es la varianza particular (no compartida) de cada variable o **varianza única**.

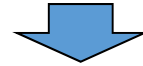
Comunalidad y unicidad

- La proporción de la varianza total de una variable que es común al resto de variables es la **comunalidad** (h^2) y el resto (hasta 1) la **unicidad**.
- Dependiendo del número de componentes (K) que se incluyan en el modelo y de la “importancia” de cada una de ellas (valor de las cargas), mejor o peor estará explicada cada variable observada X_i .

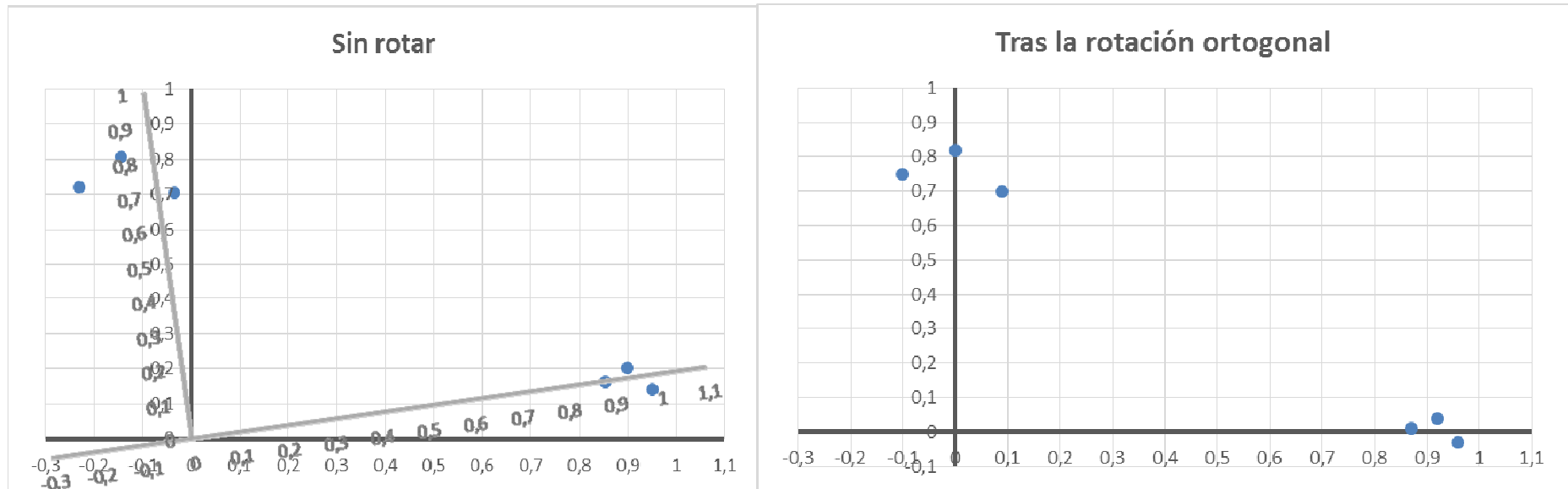


2 Interpretación

En la mayoría de las veces, en los casos reales, los grupos de variables correlacionadas no aparecen tan claros



Rotación



3 Utilización



Determinar el modelo o solución factorial final

Precisar:

- cómo se denomina cada componente
- qué representa
- qué variables se asocian con cada una de ellas

Utilizar la solución factorial final

- Obtener los valores de las componentes para cada individuo de la muestra (**puntuaciones factoriales**)
- Utilizar las **puntuaciones factoriales** en análisis posteriores

Supuestos



Supuestos

Asumir ciertas hipótesis y condiciones en nuestros datos. Muchas de estas hipótesis son comunes a otras técnicas como la de regresión, ANOVA, etc. (linealidad, normalidad,...)

Adecuación muestral

Garantizar la obtención de una solución factorial para un conjunto de variables para la que las componentes encontradas tengan algún significado y/o sean importantes.



Paquetes y funciones



Paquetes

`corpcor`

`Psych`

`GPArotation`

Funciones

- `cor()`
- `det()`
- `cortest.bartlett()`
- `kmo()`
- `principal()`
- `plot()`
- `factor.residuals()`
- `residual.stats()`
- `print.psych()`
- `cor.plot()`



Ejemplo: Cuestionario “R anxiety” RAQ¹

Item	Descripción
1	La Estadística me hace llorar
2	Mis compañeros piensan que soy un inepto por no apañarmelas con el R
3	La Desviación Típica me excita
4	Sueño con que Pearson me ataca con coeficientes de correlación
5	No entiendo la Estadística
6	No tengo mucha experiencia con los ordenadores
7	Los ordenadores me odian
8	Nunca he sido bueno con las matemáticas
9	Mis amigos son mejores estadísticos que yo
10	Los ordenadores sólo son útiles para jugar
11	Las matemáticas no me fueron bien en el colegio
12	La gente trata de convencerme de que R hará más fácil la Estadística, pero no es así
13	Me preocupa causar algún daño irreparable por mi incompetencia con los ordenadores
14	Los ordenadores tienen mentes propias y funcionan mal deliberadamente cuando yo los uso
15	Los ordenadores van a por mi
16	Me "hago encima" sólo con oír la palabra Inferencia
17	Entro en coma cuando veo una ecuación
18	R siempre se cuelga cuando trato de usarlo
19	Todos me miran cuando uso R
20	No puedo dormir pensando en valores propios
21	Me despierto bajo las sábanas pensando que estoy atrapado bajo una distribución normal
22	Mis amigos son mejores que yo con R
23	Si soy bueno en estadística mis amigos pensarán que soy un empollón

¹ **R Anxiety Questionnaire**. Ejemplo extraído de Andy Field, Jeremy Miles, and Z.F., 2012. *Discovering statistics using R*, SAGE Publications. Los datos con las respuestas al cuestionario están en PoliformaT en el **fichero RAQ.RData**

Observación de R

Primer paso, como parte de un análisis descriptivo más amplio, explorar la matriz de correlaciones R para evitar que:

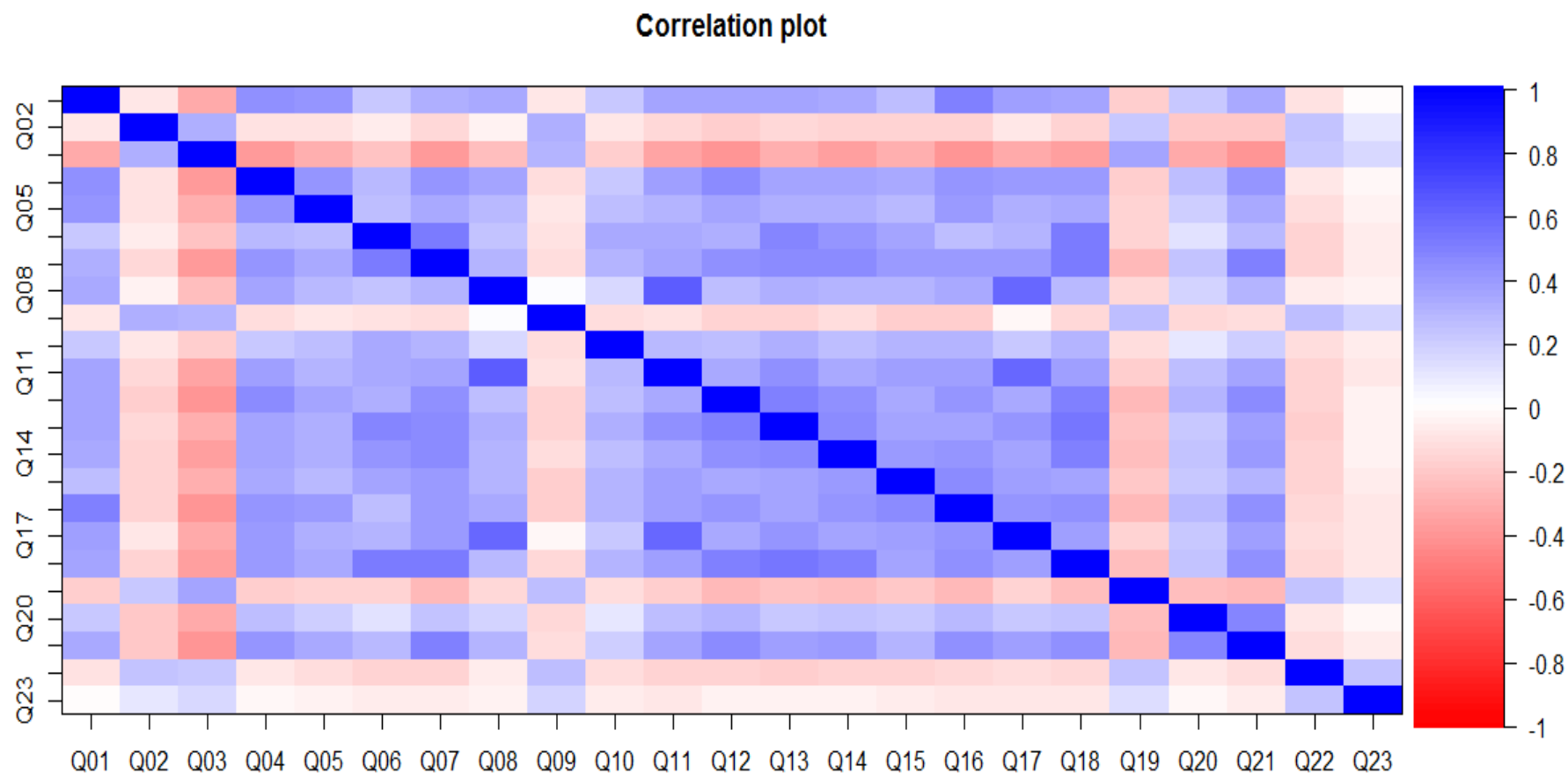
- las variables correlacionen poco:
 - $r < 0,3$
- varias variables correlacionan mucho:
 - $r > 0.8$ (multicolinealidad)
- algunas variables correlacionan perfectamente:
 - $r = 1$ (singularidad)

Eliminar variables que puedan causar problemas

Observación de R gráficamente



```
> cor.plot(raqR)
```



Adecuación muestral

Regla general para considerar adecuada la aplicación del ACP							
Coeficientes matriz de correlaciones	Todos los valores entre 0,3 y 0,9						
Determinante matriz de correlaciones	Debe ser mayor que 0,00001						
Indicador de multicolinealidad							
Coeficientes matriz anti-imagen (en valor absoluto)	Elementos de la diagonal principal: mayoritariamente superiores a 0,5, preferiblemente cercanos a 1						
Medida de adecuación	Resto de elementos: mayoritariamente del orden o inferiores a 0,3						
Test de Esfericidad de Bartlett							
Comprobar si R es la matriz I	El test debe resultar significativo (p-valor < 0,05)						
Índice KMO (Kaiser–Meyer–Olkin)							
Comprobar si las correlaciones no son demasiado altas	<table> <tr> <td>KMO < 0,5</td><td>No adecuado hacer un AF</td></tr> <tr> <td>0,5 < KMO < 0,6</td><td>Aceptable hacer un AF</td></tr> <tr> <td>KMO > 0,7</td><td>Sí adecuado hacer un AF</td></tr> </table>	KMO < 0,5	No adecuado hacer un AF	0,5 < KMO < 0,6	Aceptable hacer un AF	KMO > 0,7	Sí adecuado hacer un AF
KMO < 0,5	No adecuado hacer un AF						
0,5 < KMO < 0,6	Aceptable hacer un AF						
KMO > 0,7	Sí adecuado hacer un AF						



Observación de R



```
> # Obtener la matriz de correlación R
> raqR<-cor(raqDatos)
> round(raqR, 2)
```

	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08
Q01	1.00	-0.10	-0.34	0.44	0.40	0.22	0.31	0.33
Q02	-0.10	1.00	0.32	-0.11	-0.12	-0.07	-0.16	-0.05
Q03	-0.34	0.32	1.00	-0.38	-0.31	-0.23	-0.38	-0.26
Q04	0.44	-0.11	-0.38	1.00	0.40	0.28	0.41	0.35
Q05	0.40	-0.12	-0.31	0.40	1.00	0.26	0.34	0.27
Q06	0.22	-0.07	-0.23	0.28	0.26	1.00	0.51	0.22
Q07	0.31	-0.16	-0.38	0.41	0.34	0.51	1.00	0.30
Q08	0.33	-0.05	-0.26	0.35	0.27	0.22	0.30	1.00
Q09	-0.09	0.31	0.30	-0.12	-0.10	-0.11	-0.13	0.02
Q10	0.21	-0.08	-0.19	0.22	0.26	0.32	0.28	0.16
Q11	0.36	-0.14	-0.35	0.37	0.30	0.33	0.34	0.63
Q12	0.35	-0.19	-0.41	0.44	0.35	0.31	0.42	0.25
Q13	0.35	-0.14	-0.32	0.34	0.30	0.47	0.44	0.31
Q14	0.34	-0.16	-0.37	0.35	0.32	0.40	0.44	0.28
Q15	0.25	-0.16	-0.31	0.33	0.26	0.36	0.39	0.30
Q16	0.50	-0.17	-0.42	0.42	0.39	0.24	0.39	0.32
Q17	0.37	-0.09	-0.33	0.38	0.31	0.28	0.39	0.59
Q18	0.35	-0.16	-0.38	0.38	0.32	0.51	0.50	0.28
Q19	-0.19	0.20	0.34	-0.19	-0.17	-0.17	-0.27	-0.16
Q20	0.21	-0.20	-0.32	0.24	0.20	0.10	0.22	0.18
Q21	0.33	-0.20	-0.42	0.41	0.33	0.27	0.48	0.30
Q22	-0.10	0.23	0.20	-0.10	-0.13	-0.17	-0.17	-0.08

```
> # Obtener el determinante de la matriz de correlación R
> det(raqR)
```

```
[1] 0.0005271037 > 0,00001 → No hay multicolinealidad
```





Test de Bartlett

```
cortest.bartlett(raqR, dim(raqDatos) [1])
```

```
$chisq
```

```
[1] 19334.49 (estadístico1)
```

```
$p.value (p-valor)
```

```
[1] 0
```

p-valor < 0.05 (alfa). No podemos aceptar la H0 de que R=I ($r < 0.3$) → nuestros datos son adecuados para el AF.

```
$df
```

```
[1] 253 (grados de libertad)
```

¹Estadístico χ^2

Índice KMO



```
> raq.kmo <- kmo(raqDatos)
```

```
$overall
```

```
[1] 0.9302245 KMO > 0.7 Buena adecuación
```

```
$report
```

```
[1] "The KMO test yields a degree of common variance marvelous."
```

```
$individual KMO para cada variable > 0.5 Buena adecuación
```

Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10	Q11	Q12
0.9297610	0.8747754	0.9510378	0.9553403	0.9600892	0.8913314	0.9416800	0.8713055	0.8337295	0.9486858	0.9059338	0.9548324
Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	
0.9482270	0.9671722	0.9404402	0.9336439	0.9306205	0.9479508	0.9407021	0.8890514	0.9293369	0.8784508	0.7663994	

```
$AIS Matriz anti-imagen de R
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	0.627153118	-0.014264007	0.032684206	-0.103441650	-1.037583e-01	0.012088782	0.013059119	-0.0277420250	-0.011472364
[2,]	-0.014264007	0.811755252	-0.109450940	-0.028631346	7.727771e-03	-0.036374795	0.010812929	-0.0210898651	-0.153262421

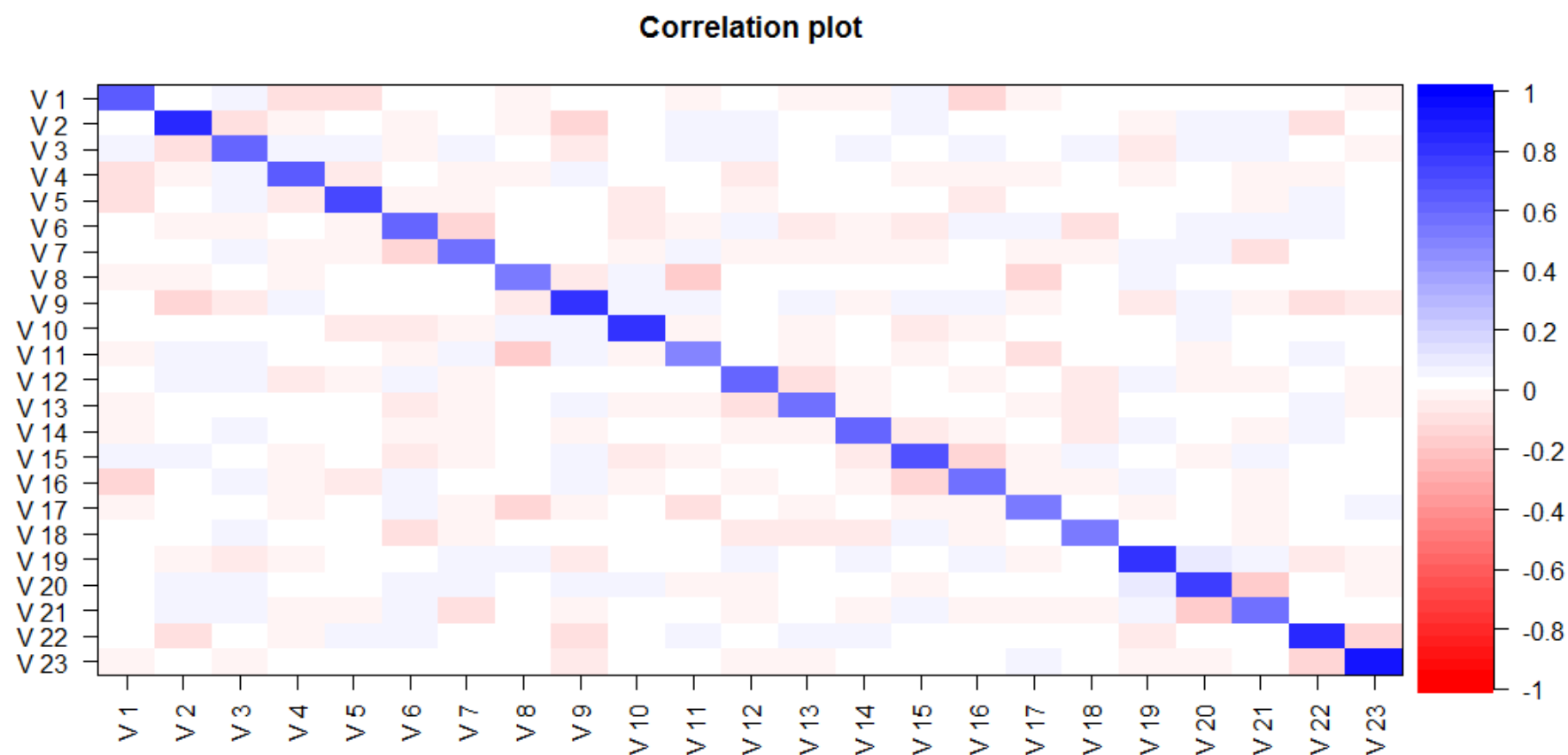
.....



Observación de anti-imagen



```
> cor.plot(rag.kmo$AIS)
```



Modelo 1: 23 componentes



Inicialmente extraemos tantas Componentes como variables (23), obteniendo un primer modelo de componentes principales (mcp1)

```
mcp1<- principal(raqR, nfactors = ncomp, rotate = "none")
```

Matriz R
o data frame

Nº de
componentes
a extraer

Si solución rotada
y de qué tipo

```
# O bien principal(raqDatos, nfactors = ncomp, rotate = "none")
```

Principal Components Analysis

```
Call: principal(r = raqR, nfactors = ncomp, rotate = "none")
```

Modelo 1: 23 componentes



Standardized loadings (pattern matrix) based upon correlation matrix

Matriz de pesos A (b_{jk}) \$loadings

\$communality
Comunalidad → Unicidad

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22	PC23	h2	u2	com
Q01	0.59	0.18	-0.22	0.12	-0.40	-0.11	-0.22	-0.08	0.01	-0.10	0.11	-0.12	0.30	-0.25	0.18	0.12	-0.05	-0.17	0.16	-0.01	-0.21	0.05	0.01	1	0.0e+00	6.0
Q02	-0.30	0.55	0.15	0.01	-0.03	-0.38	0.19	-0.39	0.01	-0.12	0.30	0.27	-0.02	0.01	-0.24	-0.05	-0.08	0.00	0.01	-0.02	-0.02	0.03	0.02	1	-2.9e-15	6.1
Q03	-0.63	0.29	0.21	-0.07	0.02	0.00	0.01	-0.05	0.20	0.10	0.15	0.03	0.10	0.13	0.40	-0.06	0.43	0.08	0.09	0.05	0.01	0.00	0.05	1	-8.9e-16	4.4
Q04	0.63	0.14	-0.15	0.15	-0.20	-0.12	-0.06	0.11	-0.11	-0.01	-0.03	0.34	-0.32	-0.17	0.12	0.31	0.19	0.05	-0.21	0.04	0.09	-0.02	0.02	1	3.3e-16	4.9
Q05	0.56	0.10	-0.07	0.14	-0.42	-0.17	-0.06	0.11	0.24	0.09	-0.30	0.16	0.12	0.48	-0.07	-0.08	-0.04	0.01	-0.04	0.00	-0.02	0.02	0.01	1	-1.1e-15	5.2
Q06	0.56	0.10	0.57	-0.05	0.17	0.01	0.00	0.05	0.00	0.00	-0.13	0.20	0.24	-0.03	0.08	0.20	-0.14	0.05	0.09	-0.07	0.04	-0.32	-0.11	1	-6.7e-16	4.4
Q07	0.69	0.04	0.25	0.10	0.17	-0.08	0.05	0.03	-0.08	0.13	-0.27	0.20	0.04	-0.22	0.00	-0.23	0.03	-0.15	0.20	0.16	0.14	0.24	0.09	1	-2.2e-16	4.1
Q08	0.55	0.40	-0.32	-0.42	0.15	0.10	-0.07	-0.04	0.01	-0.05	-0.09	0.03	-0.01	0.04	-0.04	0.03	0.10	0.07	0.12	-0.15	0.06	0.16	-0.36	1	-4.4e-16	5.7
Q09	-0.28	0.63	-0.01	0.10	0.17	-0.27	-0.01	-0.03	0.16	0.32	-0.22	-0.37	-0.17	-0.07	0.12	0.11	-0.19	-0.02	-0.08	-0.03	0.04	-0.01	0.03	1	-2.2e-16	5.0
Q10	0.44	0.03	0.36	-0.10	-0.34	0.22	0.44	-0.03	0.37	-0.22	-0.11	-0.21	-0.17	-0.15	-0.07	0.03	0.07	-0.01	0.00	0.04	-0.03	0.02	-0.04	1	-4.4e-16	7.7
Q11	0.65	0.25	-0.21	-0.40	0.13	0.18	-0.01	0.03	0.10	-0.14	0.00	0.03	0.02	0.03	-0.02	0.07	-0.05	0.07	0.07	-0.18	0.06	0.00	0.41	1	-4.4e-16	4.1
Q12	0.67	-0.05	0.05	0.25	0.04	-0.08	-0.14	0.08	0.01	-0.11	0.19	-0.07	-0.45	0.17	0.09	-0.10	-0.08	0.04	0.36	0.00	-0.04	-0.10	-0.02	1	-1.1e-15	3.8
Q13	0.67	0.08	0.28	-0.01	0.13	0.03	-0.21	0.05	0.08	-0.22	0.24	-0.08	0.01	0.12	0.14	-0.11	-0.06	-0.32	-0.30	-0.06	0.16	0.08	-0.05	1	-8.9e-16	4.2
Q14	0.66	0.02	0.20	0.14	0.08	-0.03	-0.10	-0.06	-0.14	0.16	0.08	-0.29	0.07	0.14	-0.37	0.25	0.34	-0.09	0.06	0.02	0.03	-0.01	0.05	1	1.1e-15	4.3
Q15	0.59	0.01	0.12	-0.11	-0.07	0.29	0.32	-0.12	-0.27	0.41	0.15	0.09	-0.09	0.16	0.16	0.06	-0.12	-0.10	-0.04	-0.07	-0.19	0.10	0.00	1	-2.2e-16	5.6
Q16	0.68	0.01	-0.14	0.08	-0.32	0.00	0.12	-0.14	-0.19	0.15	0.16	-0.19	0.12	-0.08	0.06	-0.22	-0.03	0.22	-0.02	-0.04	0.35	-0.12	-0.01	1	-1.6e-15	4.0
Q17	0.64	0.33	-0.21	-0.34	0.10	0.05	-0.02	0.03	-0.04	0.02	0.01	-0.03	-0.01	-0.01	-0.05	-0.18	0.04	-0.04	-0.10	0.42	-0.15	-0.23	-0.01	1	-8.9e-16	4.3
Q18	0.70	0.03	0.30	0.13	0.15	-0.09	-0.10	0.06	-0.06	-0.12	0.05	-0.11	0.09	0.00	0.03	-0.01	-0.06	0.45	-0.15	0.08	-0.18	0.23	0.01	1	-6.7e-16	3.4
Q19	-0.43	0.39	0.10	-0.01	-0.15	0.07	0.05	0.68	0.02	0.16	0.29	0.04	0.06	-0.09	-0.16	-0.03	-0.06	0.01	0.05	-0.02	0.02	0.04	-0.02	1	-2.2e-16	3.5
Q20	0.44	-0.21	-0.40	0.30	0.33	-0.01	0.34	0.03	0.33	0.02	0.21	0.04	0.17	0.07	0.05	0.22	-0.09	0.00	0.04	0.18	0.10	0.06	-0.04	1	-4.4e-16	8.7
Q21	0.66	-0.06	-0.19	0.28	0.24	-0.15	0.18	0.10	0.12	0.08	-0.02	0.04	0.03	-0.15	-0.04	-0.27	0.20	-0.03	-0.11	-0.31	-0.20	-0.13	-0.01	1	-1.6e-15	4.6
Q22	-0.30	0.47	-0.12	0.38	0.07	0.12	0.31	0.12	-0.41	-0.39	-0.19	-0.10	0.08	0.15	0.09	0.01	0.04	-0.06	0.02	0.00	0.01	-0.01	0.01	1	0.0e+00	7.2
Q23	-0.14	0.37	-0.02	0.51	0.02	0.62	-0.28	-0.22	0.18	0.08	0.00	0.13	-0.01	-0.07	-0.12	-0.06	-0.03	0.05	-0.03	0.01	-0.01	-0.02	0.00	1	-2.2e-16	4.2

Comunalidad h^2 = proporción de la varianza de una variable que son capaces de explicar las componentes extraídas (cuanto más cerca de 1, mejor).

Unicidad u^2 = 1 - Comunalidad



Extracción

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22	PC23
SS loadings	7.29	1.74	1.32	1.23	0.99	0.90	0.81	0.78	0.75	0.72	0.68	0.67	0.61	0.58	0.55	0.52	0.51	0.46	0.42	0.41	0.38	0.36	0.33
Proportion Var	0.32	0.08	0.06	0.05	0.04	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01
Cumulative Var	0.32	0.39	0.45	0.50	0.55	0.59	0.62	0.65	0.69	0.72	0.75	0.78	0.80	0.83	0.85	0.88	0.90	0.92	0.94	0.95	0.97	0.99	1.00
Proportion Explained	0.32	0.08	0.06	0.05	0.04	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01
Cumulative Proportion	0.32	0.39	0.45	0.50	0.55	0.59	0.62	0.65	0.69	0.72	0.75	0.78	0.80	0.83	0.85	0.88	0.90	0.92	0.94	0.95	0.97	0.99	1.00

Mean item complexity = 5

Test of the hypothesis that 23 components are sufficient.

The root mean square of the residuals (RMSR) is 0

Fit based upon off diagonal values = 1

SS Loadings (Sum of Squared Loadings) `$values` = Valores propios de R = parte de la varianza total de los datos explicada por cada componente

Proportion var = la proporción de los SS Loadings con respecto a la varianza total (nº de componentes) = 23 (e.g. $7.29/23 = 0,32$)

Cumalitive var = la "Proportion var" de cada componente acumulada

Root mean square of the residuals (RMSR) = Raíz del cuadrado medio residual

Fit based upon off diagonal values `$fit.off` = Indicador del ajuste.
Valores > 0,95 → Buen ajuste

¿Cuántos componentes extraemos?

- **Regla de Kaiser:** el número de valores propios¹ de $R > 1$
- **Criterio del porcentaje de la varianza:** número de componentes mínimo necesario para que el porcentaje acumulado de la varianza de los datos observados explicado por las componentes alcance un nivel satisfactorio (75%, 80%)
- **Criterio de Castell:** representación gráfica del tamaño de los valores propios. En esta representación denominada **Gráfico de Sedimentación** o **Scree Plot** el eje de abscisas representa los diferentes factores extraídos y el de ordenadas los valores propios. Las componentes con varianzas altas suelen diferenciarse de las que poseen varianzas bajas, esta diferenciación se manifiesta en el gráfico como un punto de inflexión. Se recomienda conservar los factores situados antes de este punto de inflexión.

¹Valor propio o autovalor de R representa la parte de la varianza total de los datos que explica una componente o factor. Es una medida de la importancia de un cada factor

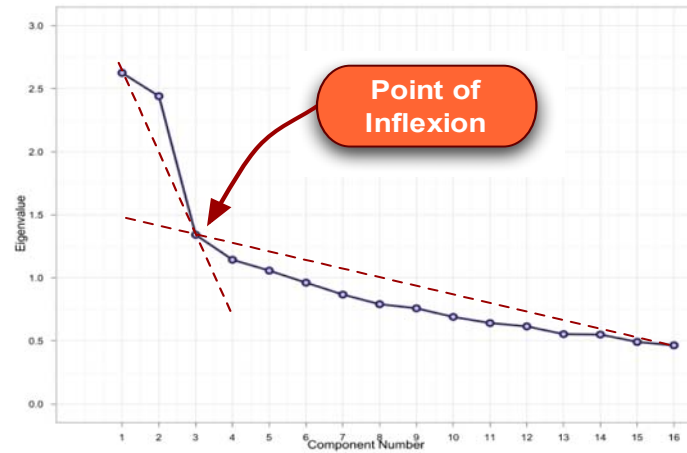
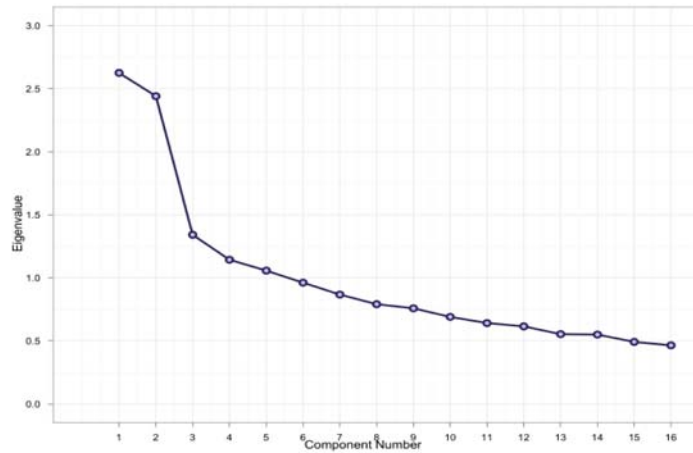
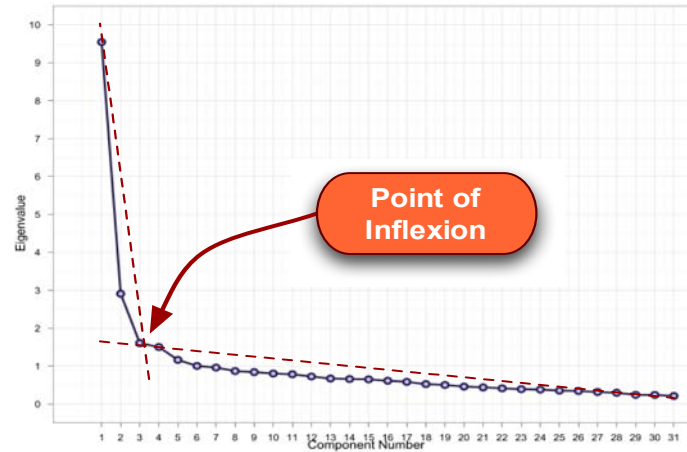
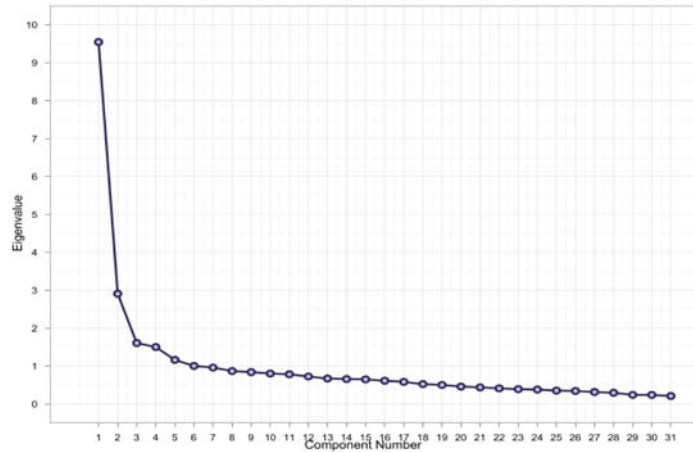
¿Qué criterio uso?

- La **Regla de Kaiser**:
 - Menos de 30 variables
 - Comunalidades tras la extracción $> 0,7$
 - Tamaño de la muestra > 250 y comunalidad media $\geq 0,6$

Big Data: $N \rightarrow \infty$

- El **Criterio de Castell**:
 - Bueno para muestras de tamaño > 200

Screepplot

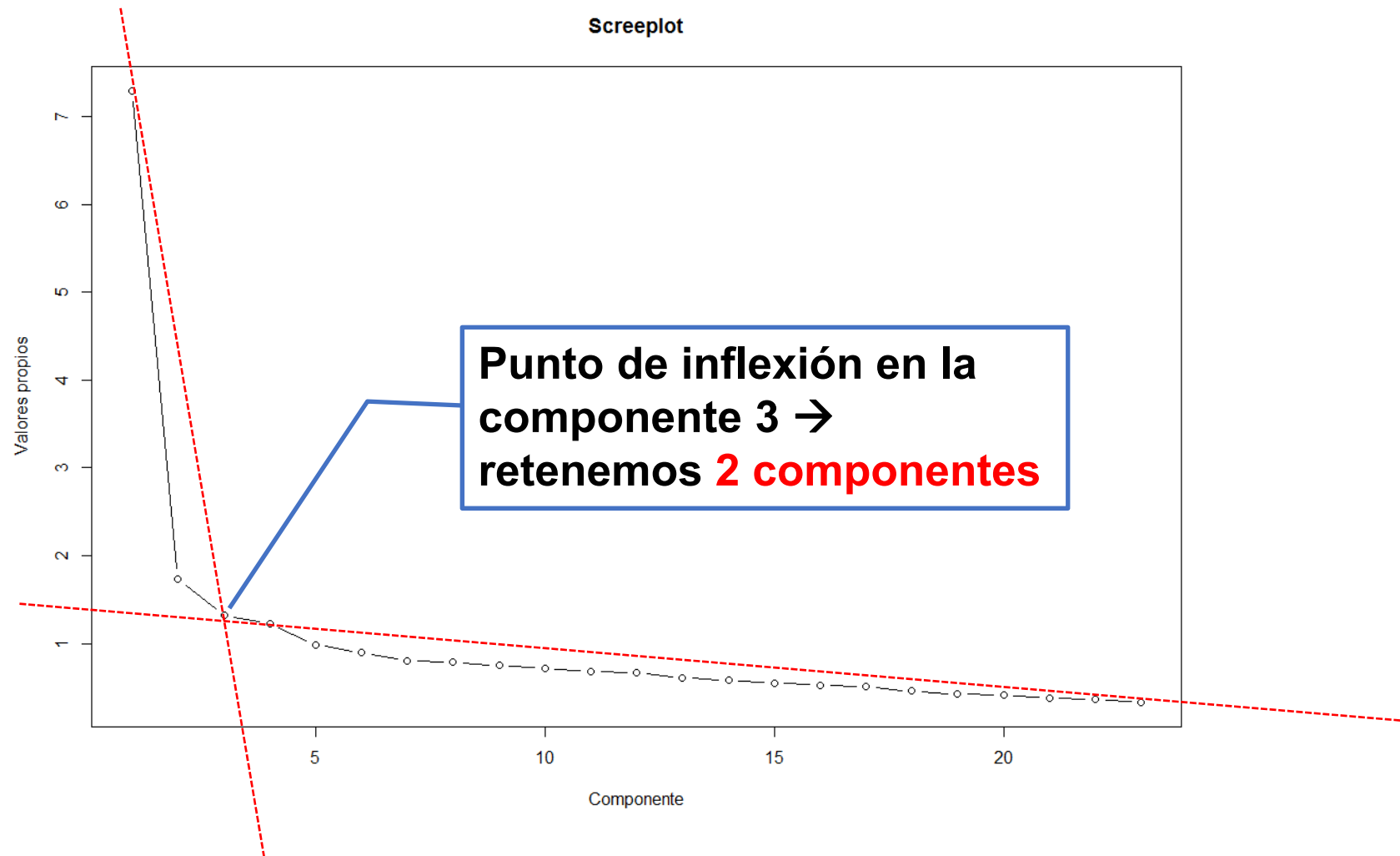


Ejemplos de gráficos de sedimentación que probablemente muestran la existencia de 2 componentes.

Screeplot

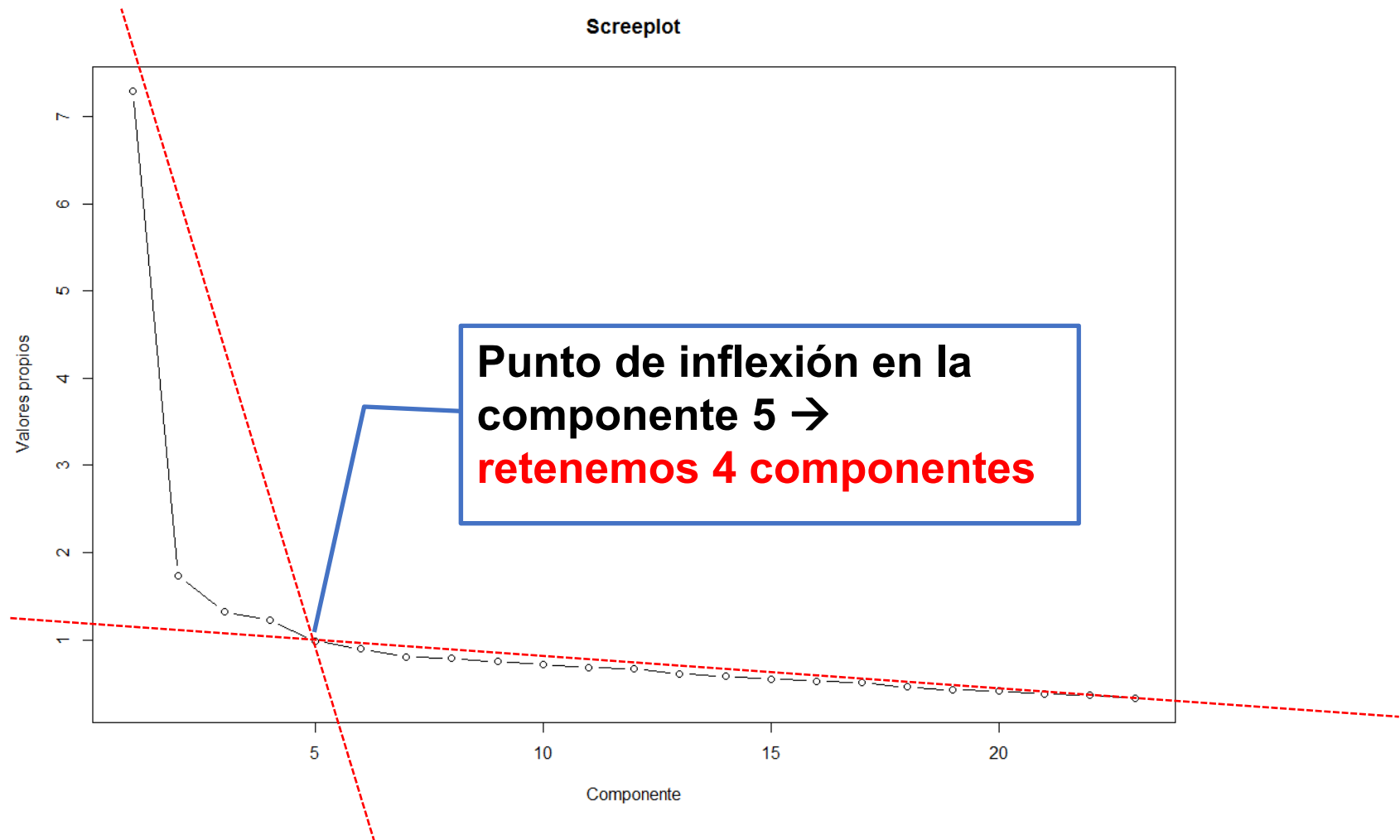


```
plot(mcp1$values, type = "b", ylab = "Valores propios", xlab =  
"Componente", main = "Screeplot")
```



Screeplot

- Existe ambigüedad, podemos probar varios modelos y comprobar su ajuste



Modelo 2: 4 componentes



```
mcp2<- principal(raqR, nfactors = 4, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation
matrix
```

Matriz de pesos A (b_{jk})

	Comunalidad				Unicidad		
	PC1	PC2	PC3	PC4	h2	u2	com
Q01	0.59	0.18	-0.22	0.12	0.43	0.57	1.6
Q02	-0.30	0.55	0.15	0.01	0.41	0.59	1.7
Q03	-0.63	0.29	0.21	-0.07	0.53	0.47	1.7
Q04	0.63	0.14	-0.15	0.15	0.47	0.53	1.3
Q05	0.56	0.10	-0.07	0.14	0.34	0.66	1.2
Q06	0.56	0.10	0.57	-0.05	0.65	0.35	2.1
Q07	0.69	0.04	0.25	0.10	0.55	0.45	1.3
Q08	0.55	0.40	-0.32	-0.42	0.74	0.26	3.5
Q09	-0.28	0.63	-0.01	0.10	0.48	0.52	1.5
Q10	0.44	0.03	0.36	-0.10	0.33	0.67	2.1
Q11	0.65	0.25	-0.21	-0.40	0.69	0.31	2.2
Q12	0.67	-0.05	0.05	0.25	0.51	0.49	1.3
Q13	0.67	0.08	0.28	-0.01	0.54	0.46	1.4
Q14	0.66	0.02	0.20	0.14	0.49	0.51	1.3
Q15	0.59	0.01	0.12	-0.11	0.38	0.62	1.2
Q16	0.68	0.01	-0.14	0.08	0.49	0.51	1.1
Q17	0.64	0.33	-0.21	-0.34	0.68	0.32	2.4
Q18	0.70	0.03	0.30	0.13	0.60	0.40	1.4
Q19	-0.43	0.39	0.10	-0.01	0.34	0.66	2.1
Q20	0.44	-0.21	-0.40	0.30	0.48	0.52	3.2
Q21	0.66	-0.06	-0.19	0.28	0.55	0.45	1.6
Q22	-0.30	0.47	-0.12	0.38	0.46	0.54	2.8
Q23	-0.14	0.37	-0.02	0.51	0.41	0.59	2.0



Extracción

	PC1	PC2	PC3	PC4
SS loadings	7.29	1.74	1.32	1.23
Proportion Var	0.32	0.08	0.06	0.05
Cumulative Var	0.32	0.39	0.45	0.50
Proportion Explained	0.63	0.15	0.11	0.11
Cumulative Proportion	0.63	0.78	0.89	1.00

The root mean square of the residuals (RMSR) is 0.06

Fit based upon off diagonal values = 0.96

La proporción de varianza explicada por parte de cada componente es:

PC1 → 32% PC2 → 8% PC3 → 6% PC4 → 5%

Cada nueva componente contemplada explica menor cantidad de la variabilidad contenida en las variables originales.

En total, considerando las 4 componentes principales, se explica un 50% de la varianza total.

El índice de ajuste es bastante bueno (> 0.95)

Modelo 2: 2 componentes



```
mcp2<- principal(raqR, nfactors = 2, rotate = "none")  
Standardized loadings (pattern matrix) based upon correlation  
matrix
```

Matriz de pesos A (b_{jk})

	PC1	PC2	Comunalidad h2	Unicidad u2	com
Q01	0.59	0.18	0.37	0.63	1.2
Q02	-0.30	0.55	0.39	0.61	1.6
Q03	-0.63	0.29	0.48	0.52	1.4
Q04	0.63	0.14	0.42	0.58	1.1
Q05	0.56	0.10	0.32	0.68	1.1
Q06	0.56	0.10	0.33	0.67	1.1
Q07	0.69	0.04	0.47	0.53	1.0
Q08	0.55	0.40	0.46	0.54	1.8
Q09	-0.28	0.63	0.47	0.53	1.4
Q10	0.44	0.03	0.19	0.81	1.0
Q11	0.65	0.25	0.49	0.51	1.3
Q12	0.67	-0.05	0.45	0.55	1.0
Q13	0.67	0.08	0.46	0.54	1.0
Q14	0.66	0.02	0.43	0.57	1.0
Q15	0.59	0.01	0.35	0.65	1.0
Q16	0.68	0.01	0.46	0.54	1.0
Q17	0.64	0.33	0.52	0.48	1.5
Q18	0.70	0.03	0.49	0.51	1.0
Q19	-0.43	0.39	0.33	0.67	2.0
Q20	0.44	-0.21	0.23	0.77	1.4
Q21	0.66	-0.06	0.44	0.56	1.0
Q22	-0.30	0.47	0.31	0.69	1.7
Q23	-0.14	0.37	0.16	0.84	1.3



Extracción

	PC1	PC2
SS loadings	7.29	1.74
Proportion Var	0.32	0.08
Cumulative Var	0.32	0.39
Proportion Explained	0.81	0.19
Cumulative Proportion	0.81	1.00

Mean item complexity = 1.3

Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is 0.06

Fit based upon off diagonal values = 0.95

En el nuevo modelo, en total, considerando las 2 componentes principales, se explica un 39% de la varianza total, menor que el modelo de 4 componentes, obviamente.

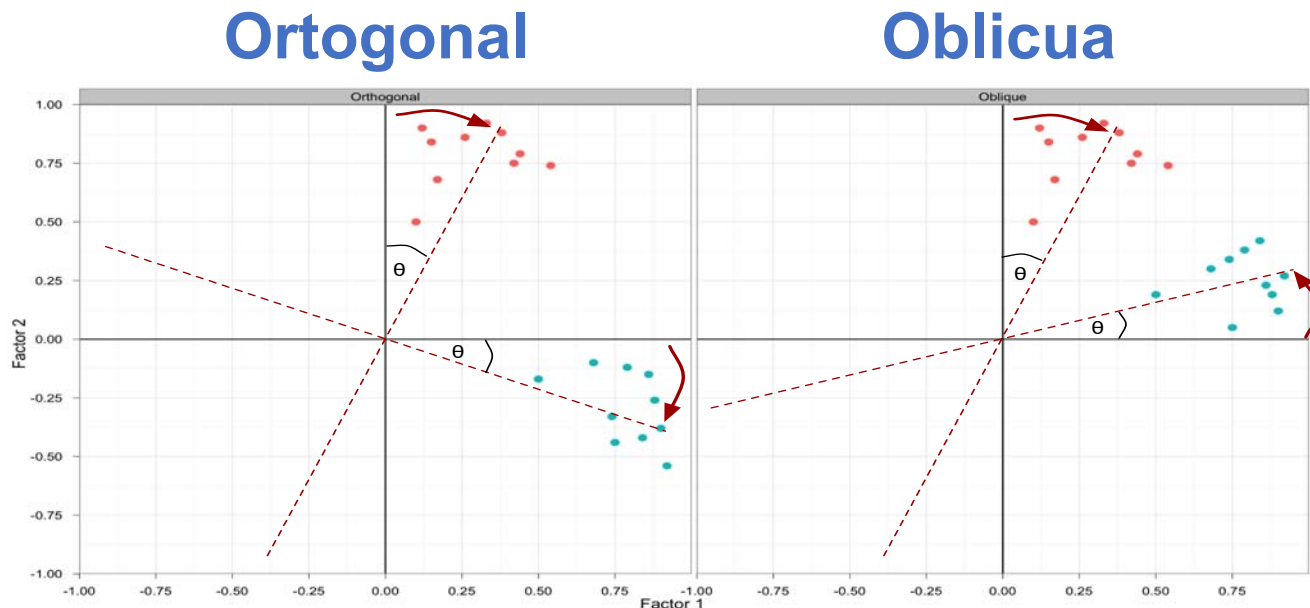
Además, el índice de ajuste ha empeorado.

En principio nos quedaríamos con el modelo 2 de 4 componentes.



Rotación

- Para mejorar la interpretación de las componentes es posible maximizar la carga de una variable sobre uno de los factores y minimizarla en el resto.
- Esto se conoce como rotación.
- Hay dos tipos:
 - **Ortogonal** (las componentes permanecen incorrelacionadas)
 - **Oblicua** (las componentes pueden correlacionar)



Rotación

a) Rotación ortogonal

- **Varimax**: es el método más habitual. Se obtiene maximizando la varianza de los factores, de este modo se consigue que las variables tengan un peso elevado en un solo factor y pequeño en el resto.
- **Quartimax**: este método persigue que muchas variables tengan pesos grandes en el mismo factor. Es sólo aconsejable si se cree que hay un solo factor principal que explica toda la varianza.
- **Equamax**: este método es un híbrido de los dos anteriores. El inconveniente es que posee un comportamiento algo errático.

b) Rotación oblicua. Dentro de este tipo de rotación también encontramos diferentes métodos como el **Direct**, **oblimin** o el **Promax**.

Rotación

- Para un **primer análisis exploratorio se aconseja usar la rotación *varimax***, pues es, en general, una buena estrategia y simplifica mucho la interpretación de los factores.
- Teóricamente, la elección más adecuada del método de rotación dependerá en gran parte de si podemos considerar que los factores pueden estar o no correlacionados. **Si admitimos que los factores pueden tener cierto grado de correlación entre ellos**, lo cual no es raro en el ámbito de las ciencias sociales, por ejemplo, deberíamos **elegir la rotación oblicua (preferiblemente *oblimin*)**.

Rotación Modelo 2: 4 componentes



```
> # Rotacion ortogonal varimax del modelo 2 (4 componentes)
> ncomp <- 4
> mcp2.RVari <- principal(raqR, n factors = 4, rotate = "varimax")

> # Obtención matriz de residuos
> resid.2.Rvar <- factor.residuals(raqR, mcp2.RVari$loadings)

> Índice de ajuste del modelo ¿> 0,9?
> mcp2.RVari$fit.off
[1] 0.9645252 > 0,9 > → Buen ajuste
```

Ajuste y comparación de modelos. Residuos

- Si el modelo de K componentes puede resumir los datos originales, la matriz R reproducida a partir de éste deberá ser parecida a la matriz R de las variables primitivas.
- La diferencia entre ambas matrices produce una matriz de residuos que deberán ser pequeños (de orden 0)
- Un índice del tamaño medio de estos residuos es el ***Fit based upon off diagonal values (\$fit.off)*** de la función **factor.residuals**

Obtención de los residuos y estadísticas mediante las funciones

```
> factor.residuals(R, cargas)
> residual.stats(residuos) #Script de Andy Field
```

Regla general para considerar un buen ajuste	
Fit based upon off diagonal values	> 0,9 o > 0,95
Valor de los residuos (absoluto)	Menos del 50% deben ser > 0,05



Comparación de modelos

```
residual.stats(residuos)  
$fit.off
```

Mejor ajuste

```
> # Modelo 2: 4 componentes sin rotar  
Proportion of absolute residuals > 0.05 = 0.3596838  
> mcp2$fit.off "Fit based upon off diagonal values"  
[1] 0.9645252 > 0,9 > → Buen ajuste  
35,97% (< 50%) residuos en valor absoluto > 0,05 → Buen ajuste
```

```
> # Modelo 3: 2 componentes sin rotar  
Proportion of absolute residuals > 0.05 = 0.4505929  
> mcp3$fit.off "Fit based upon off diagonal values"  
[1] 0.9533968 > 0,9 > → Buen ajuste  
45,06% (< 50%) residuos en valor absoluto > 0,05 → Buen ajuste
```



Interpretación del modelo

- Para facilitar la interpretación del pódolo podemos usar la función:

```
print.psych(mcp2, cut = 0.3, sort = TRUE)
```

modelo

Valor mínimo de los pesos de la matriz de cargas a visualizar

Ordenar la matriz de cargas

Interpretación



Standardized loadings (pattern matrix) based upon correlation matrix

	item	PC3	PC1	PC4	PC2	h2	u2	com
	Q06	0.80				0.65	0.35	1.0
	Q18	0.68	0.33			0.60	0.40	1.5
	Q13	0.65				0.54	0.46	1.6
	Q07	0.64	0.33			0.55	0.45	1.7
	Q14	0.58	0.36			0.49	0.51	1.8
	Q10	0.55				0.33	0.67	1.2
	Q15	0.46				0.38	0.62	2.6
	Q20		0.68			0.48	0.52	1.1
	Q21		0.66			0.55	0.45	1.5
	Q03		-0.57		0.37	0.53	0.47	2.3
	Q12	0.47	0.52			0.51	0.49	2.1
	Q04	0.32	0.52	0.31		0.47	0.53	2.4
	Q16	0.33	0.51	0.31		0.49	0.51	2.6
	Q01		0.50	0.36		0.43	0.57	2.4
	Q05	0.32	0.43			0.34	0.66	2.5
	Q08			0.83		0.74	0.26	1.1
	Q17			0.75		0.68	0.32	1.5
	Q11			0.75		0.69	0.31	1.5
	Q09				0.65	0.48	0.52	1.3
	Q22				0.65	0.46	0.54	1.2
	Q23				0.59	0.41	0.59	1.4
	Q02		-0.34		0.54	0.41	0.59	1.7
	Q19		-0.37		0.43	0.34	0.66	2.2



Interpretación del modelo

- **La información contenida en las 2571 observaciones de cada una de las 23 variables (59133 datos) las hemos podido resumir en 4 variables sólo.**
- Estas nuevas 4 variables llamadas **componentes principales** podrían representar:
 - **CP1 : miedo al manejo del ordenador**
 - **CP2: miedo a la Estadística**
 - **CP3: miedo a las matemáticas**
 - **CP4: miedo al uso del R**



Puntuaciones factoriales

- Son los valores de las componentes (nuevas variables) para cada individuo.
- Podemos operar con ellas como cualquier variable y que resumen la información de todas las variables originales.

```
principal(raqR, nfactors = 4, rotate = "varimax", scores = T)
```



Referencias adicionales

- Quick R
<http://www.statmethods.net/advstats/factor.html>
- R Bloggers <http://www.r-bloggers.com/computing-and-visualizing-pca-in-r/>
- Hair, J.F. et al., 2005. *Multivariate Data Analysis* 6th ed., Prentice Hall.
- Peña, D., 2002. *Análisis de datos multivariantes*, McGraw-Hill.

Verificación de hipótesis

- Depende de la técnica a utilizar, pero, en general, hay tres hipótesis básicas que en la mayoría de las técnicas de inferencia clásica deben asumirse:
 - Normalidad
 - Homocedasticidad (igualdad de varianzas)
 - Independencia
- En muchos de los modelos estadísticos (Regresión, ANOVA, AF, etc) otra hipótesis fundamental es la:
 - Linealidad

Normalidad de los datos

- Tests estadísticos formales:
 - Test G1-2, test de Kolmogorov, etc.
 - Exigen muchos datos en general.
 - Poco útiles en la práctica: **“respuesta correcta a una pregunta equivocada”**

Pregunta equivocada: ¿proceden los datos de una distribución normal teórica?

Pregunta relevante: ¿es la pauta de variabilidad constatada en los datos lo suficientemente parecida a la postulada por el modelo de la distribución normal, como para que se uso conduzca a resultados “válidos” o “útiles” en la práctica?



Análisis descriptivo de los datos

Análisis descriptivo de los datos

- Primera fase de cualquier estudio estadístico
- Forma parte del preproceso
- Permite:
 - Detectar la no normalidad de los datos
 - Estimar los parámetros poblacionales (parámetros muestrales)
 - Detectar valores anómalos
 - Detectar características y peculiaridades de la muestra...

Nota sobre el incumplimiento de hipótesis

Homocedasticidad

- Si $\sigma^2_1 \neq \sigma^2_2 \rightarrow$ el test de comparación de medias visto, y la correspondiente fórmula para el intervalo de confianza, tienen sólo carácter aproximado.
- Sin embargo, dicho test es bastante “robusto” frente al incumplimiento de esta hipótesis de homocedasticidad (igualdad de varianzas), especialmente si el número de observaciones en ambas muestras es parecido.
- En la práctica puede ser razonable seguir utilizando, con carácter aproximado los procedimientos de comparación de medias expuestos en esta presentación, aunque existan diferencias entre las varianzas poblacionales, y en cualquier caso se puede recurrir a pruebas que lo tengan en cuenta.

Normalidad

- Los tests son más sensibles frente a su incumplimiento, aunque para muestras grandes, puede admitirse (Teorema Central del Límite).

Papel Probabilístico Normal (PPN o QQPLOT)

- **Herramienta** extremadamente práctica de análisis estadístico para el estudio de distintos tipos de distribuciones.
- Trataremos sólo del papel probabilístico para distribuciones normales: **Papel probabilístico normal (PPN)**.

Objetivo: “determinar” si la v.a con la que se está trabajando sigue una distribución normal.



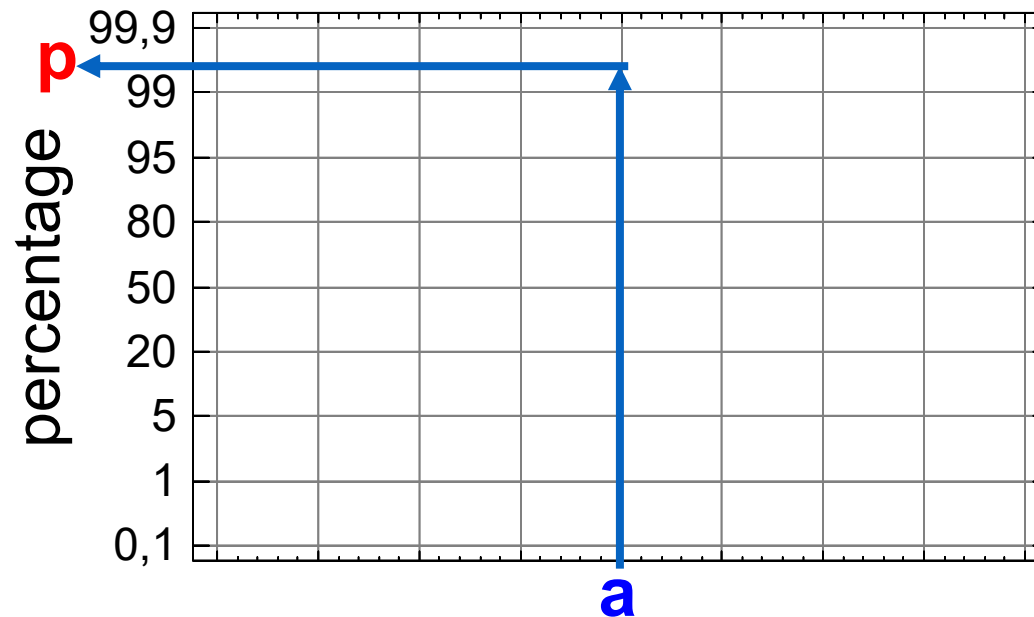
“determinar” si los datos de la muestra proceden de una población normal

NOTA: no se trata de un test de inferencia estadística. El PPN es una herramienta descriptiva que permite ver si el **modelo normal** se ajusta lo suficientemente bien a la realidad de los datos observados (muestra) como para utilizarlo.

Construcción del PPN

Se parte de una plantilla de PPN (o se dibuja), dónde se hace corresponder un punto a cada observación.

Normal Probability Plot

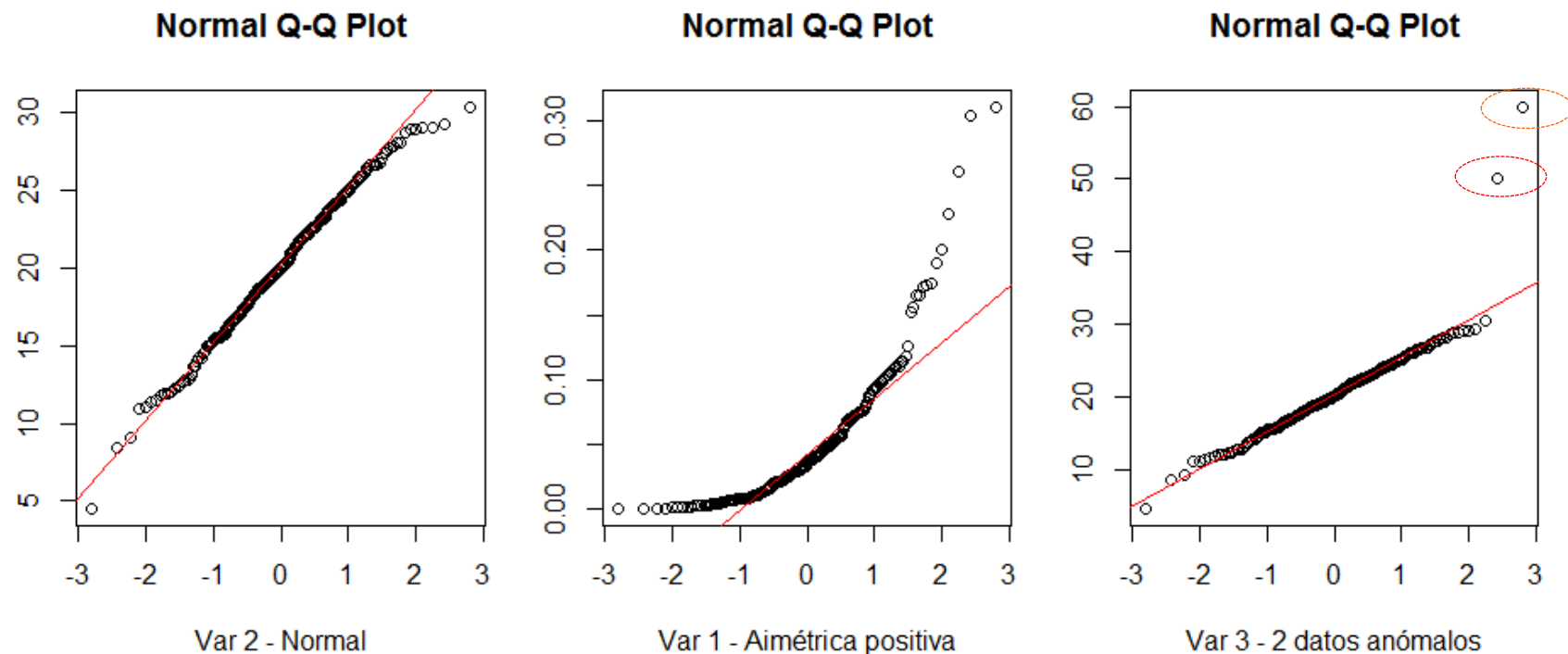


- La abscisa del punto es el valor observado (a)
- La ordenada del punto es el porcentaje de valores en la muestra que son menores o iguales que el considerado (p)

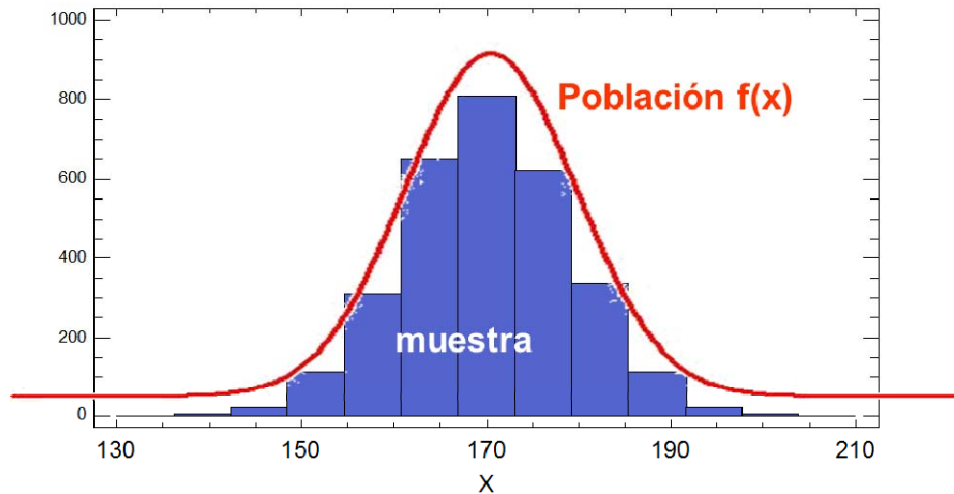
$$p = P(X \leq x)(\%) = \frac{i - \frac{1}{2}}{n} \cdot 100$$

Aplicación del PPN

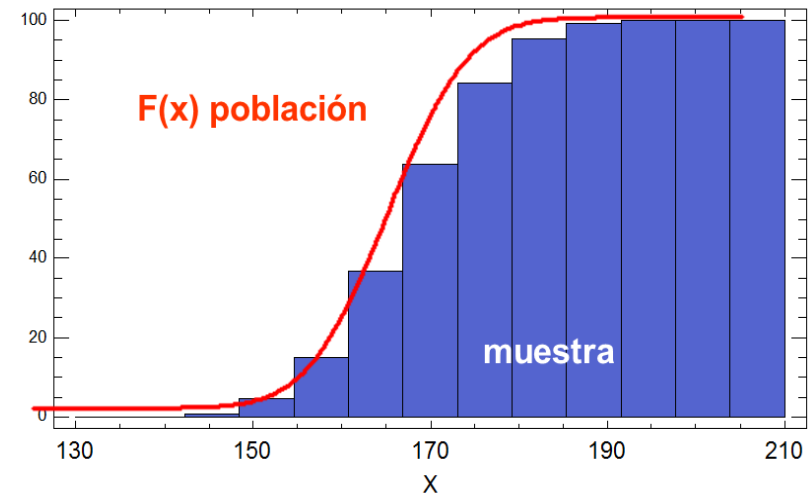
Si al dibujar los valores de la muestra sobre PPN, los puntos se alinean formando una recta, entonces, podemos decir que la v.a. sigue una distribución normal



Fundamento del PPN

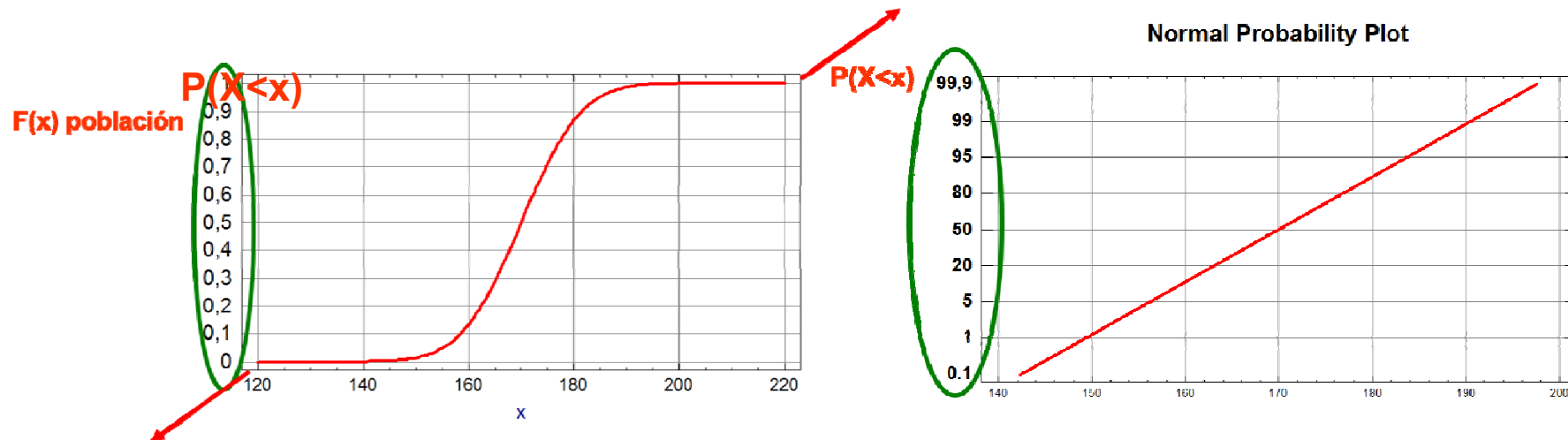


Si los datos proceden de una población normal, el histograma de las frecuencias absolutas tiene forma de campana de Gauss.



Si los datos proceden de una población normal, el histograma de las frecuencias relativas acumuladas tiene la forma de la figura de arriba.

Fundamento del PPN



Si se modifica la escala vertical, de forma que los valores de $P(X \leq x)$ de una normal tipificada $N(0,1)$ tengan forma de recta, el procedimiento es más sencillo.

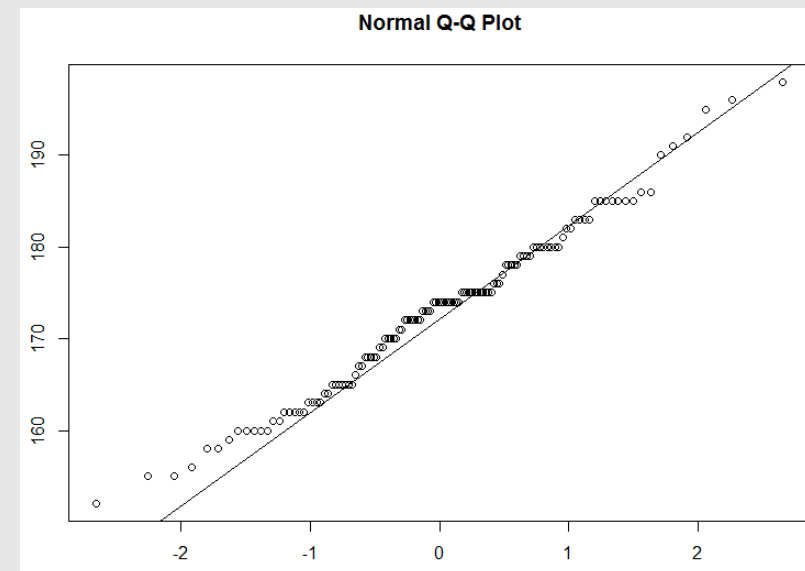
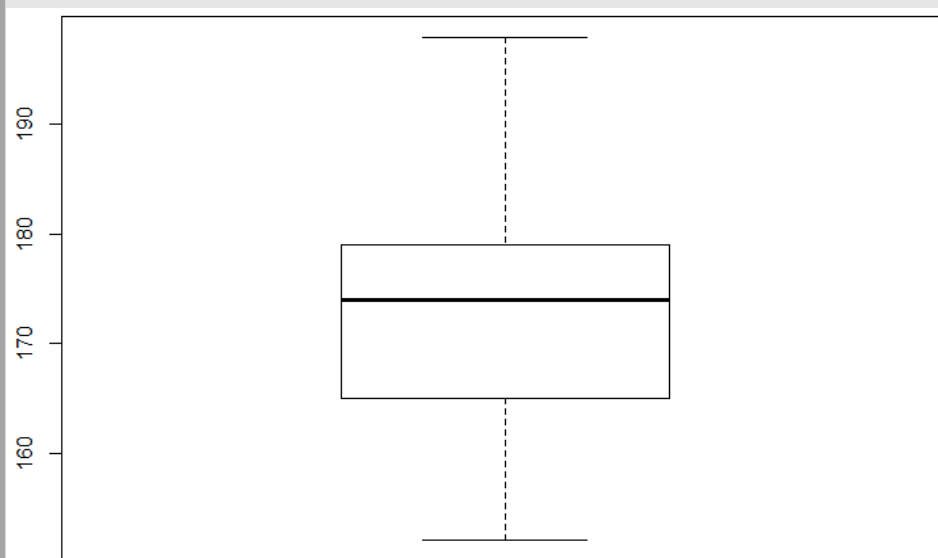
Después de cambiar la escala vertical la forma de $P(X \leq x)$ de una normal tipificada $N(0,1)$ tiene forma de recta.

Ejemplo



```
> summary(ESTATURA)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
152.0	165.2	174.0	173.0	179.0	198.0	5



Ejemplo análisis descriptivo: conclusiones

- A la vista de los parámetros muestrales (media, mediana y coeficientes de asimetría y curtosis) y de las representaciones gráficas estudiadas →
 - No se muestran indicios claros de NO normalidad
 - Tampoco se han apreciado anomalías en los datos



Se puede asumir el modelo normal como adecuado

Guardar *data frame*

- Una vez con los datos técnicamente correctos y consistentes se han de almacenar:
 - Local / remoto
 - Exportar a distintos formatos:
 - R
 - txt, csv
 - Excel, SPSS, ...
 - ...
- **save(), write.table(), ...**

Glosario



Análisis de Componentes Principales
Binarización
Datos consistentes
Datos crudos
Datos Técnicamente correctos
Discretización
Hipótesis
Integración
Limpieza
Normalización
Reducción
Transformación
Valores atípicos o anómalos
Valores especiales
Valores perdidos o faltantes



Herramientas Estadísticas para Big Data
Introducción a la Inferencia Estadística, Muestreo y Preproceso de datos

6- Preproceso de datos



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

www.upv.es

E. Vázquez
Dto. De Estadística e Investigación Operativa, Aplicadas y Calidad