DIPLOMADO
APLICACIÓN DE LA CIENCIA SENSORIAL EN LA INDUSTRIA

Fernando Alonso Velez
Msc. en Estadística

fvelez78@yahoo.es

Septiembre 2017

CONTENIDO GENERAL

CONTENIDO GENERAL

CONTENIDO GENERAL

CONTENIDO GENERAL

www.schoology.com

SIGN UP → STUDENT → CODIGO → DATOS PERSONALES

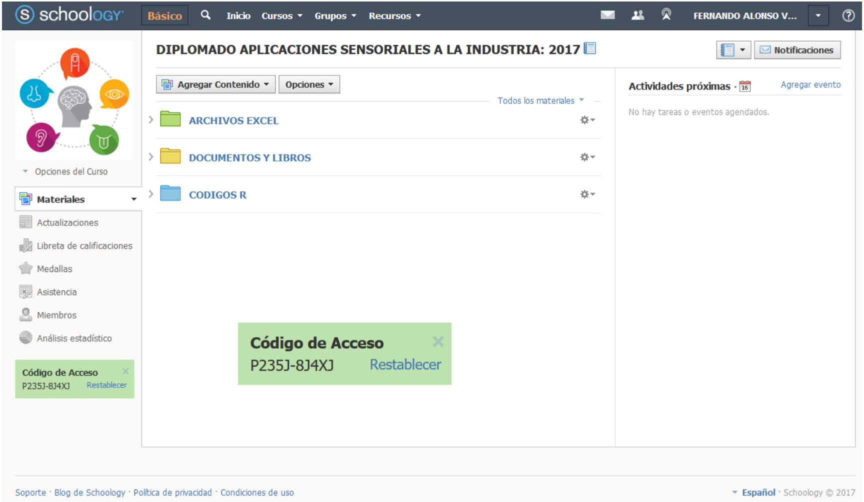


Figura: Punto de encuentro

Sección 1

INTRODUCCIÓN AL ANÁLISIS DE DATOS

Naturaleza de la información en el mundo sensorial

En la evaluación sensorial se consideran variables de diversa naturaleza, las cuales de manera general se pueden agrupar como:

1. Cualitativas
   a. Atributos no ordenados → Nominal (Marca, Color)
   b. Atributos ordenados →Ordinal (Nivel de agrado)
2. Cuantitativas
   a. Discretas → Ordinal (Frecuencia de uso)
   b. Continuas →Intervalo ó Razón (Tiempos de duración)

En resumen:

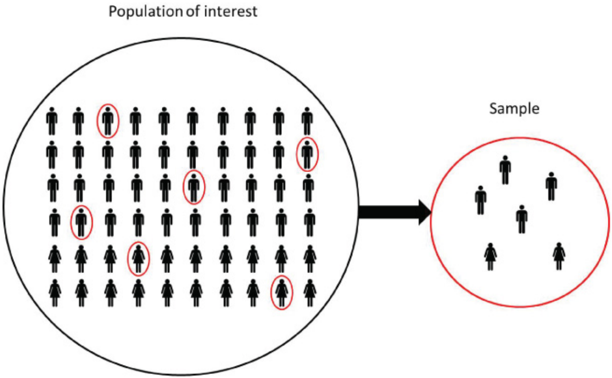
- Nominal: Nombres sin orden
- Ordinal: Nombres con orden o números enteros
- Intervalo: Medidas sin cero unico
- Razón: Medidas con cero unico

Variables Estadísticas

En el lenguaje estadístico se suele emplear algunos términos especificos para referirse a determinados usos de las variables

- Variable respuesta
- Variable Dummy
- Variable de clasificación
- Serie
- Imagenes
- Audio

Población y Muestra



- Población de interés - Muestra Aleatoria
- Población Accesible - Muestra disponible

El role de la percepción

Enfoques a considerar

1. Señal química
2. Respuesta humana

De acuerdo al diseño del experimento se puede obtener información de:

1. Estructura Manifiesta: Son aquellas compuestas por variables que son directamente observables.
   Ejemplo: Características de una manzana fresca.
2. Estructura Latente:Es aquella expresada en términos de variables que no son observables de forma directa.
   Ejemplo: Criterio de compra.

Sección 2

INTRODUCCIÓN A R Y RStudio

Subsección 1

Aspectos generales de R y RStudio

¿Que es R?

Es un lenguaje de programación y un entorno para el cálculo estadístico y elaboración de gráficas básicas avanzadas; es una implementación de código abierto del lenguaje S (S-Plus), desarrollado por los Laboratorios Bell. Escrito inicialmente por Ross Ihaka y Robert Gentleman a mediados de los años 90.
En R se puede realizar análisis hasta con 2 millones de registros y mas de 250.000 variables. Es un programa amplio y flexible de análisis estadístico y gestión de información capaz de trabajar con datos procedentes de distintos formatos proporcionando, desde sencillos gráficos de distribuciones y estadísticos descriptivos, hasta análisis estadísticos complejos que permiten descubrir relaciones de dependencia e interdependencia, establecer clasificaciones de sujetos y variables, predecir comportamientos, etc.

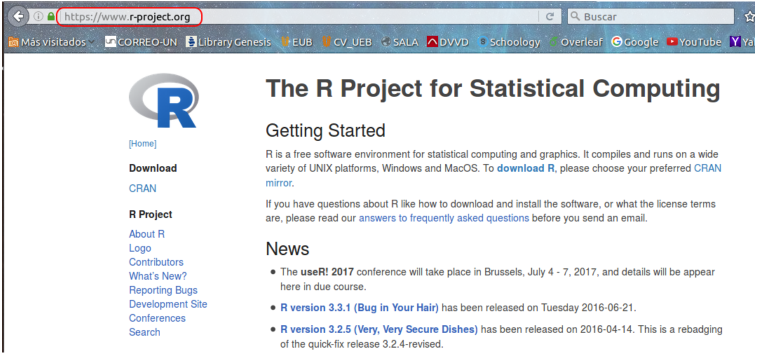
Introducción a R

R es un programa libre bajo licencia GNU GPL(General Public License)

- Ventajas
  1. Libre (Gratis. . .)
  2. Adaptable
  3. Liviano
  4. Gran cantidad de ayuda disponible

- Desventajas
  1. Programar (aunque realmente es una ventaja. . .)
  2. Muchas cosas por aprender...

¿onde se puede descargar?



https://www.r-project.org

Más visitados · CORREO-UN · Library Genesis · EUB · CV_UEB · SALA · DVVD · Schoology · Overleaf · Google · YouTube · Ya

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Development Site
Conferences
Search

The R Project for Statistical Computing

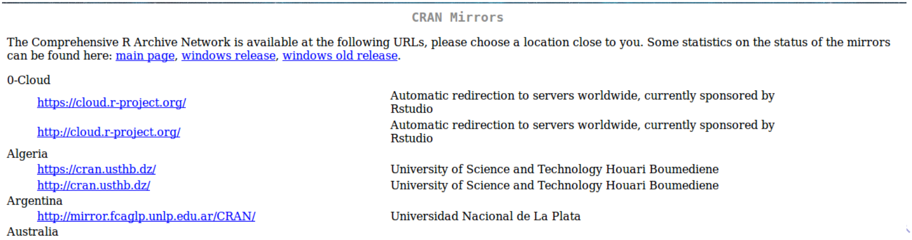
Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- The **useR! 2017** conference will take place in Brussels, July 4 - 7, 2017, and details will be appear here in due course.
- **R version 3.3.1 (Bug in Your Hair)** has been released on Tuesday 2016-06-21.
- **R version 3.2.5 (Very, Very Secure Dishes)** has been released on 2016-04-14. This is a rebadging of the quick-fix release 3.2.4-revised.



CRAN Mirrors

The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the mirrors can be found here: main page, windows release, windows old release.

0-Cloud

https://cloud.r-project.org/	Automatic redirection to servers worldwide, currently sponsored by Rstudio
http://cloud.r-project.org/	Automatic redirection to servers worldwide, currently sponsored by Rstudio

Algeria

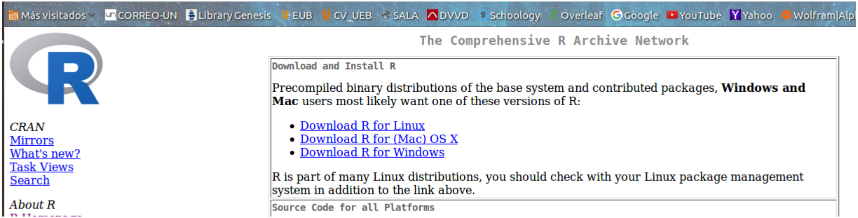
https://cran.usthb.dz/	University of Science and Technology Houari Boumediene
http://cran.usthb.dz/	University of Science and Technology Houari Boumediene

Argentina

http://mirror.fcaglp.unlp.edu.ar/CRAN/	Universidad Nacional de La Plata

Australia

¿Donde se puede descargar?


Más visitados CORREO-UN Library Genesis EUB CV_UEB SALA DVVD Schoology Overleaf Google YouTube Yahoo Wolfram|Alp

The Comprehensive R Archive Network

CRAN
Mirrors
What's new?
Task Views
Search

About R

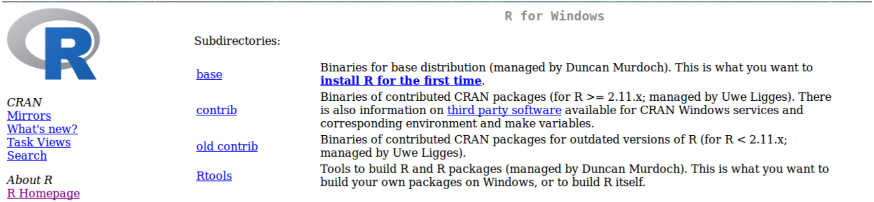
Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

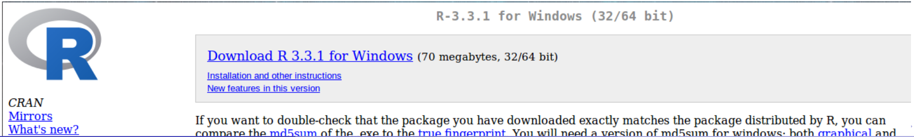

R for Windows

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage

Subdirectories:

base	Binaries for base distribution (managed by Duncan Murdoch). This is what you want to **install R for the first time**.
contrib	Binaries of contributed CRAN packages (for R >= 2.11.x; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
old contrib	Binaries of contributed CRAN packages for outdated versions of R (for R < 2.11.x; managed by Uwe Ligges).
Rtools	Tools to build R and R packages (managed by Duncan Murdoch). This is what you want to build your own packages on Windows, or to build R itself.


R-3.3.1 for Windows (32/64 bit)

CRAN
Mirrors
What's new?

Download R 3.3.1 for Windows (70 megabytes, 32/64 bit)

Installation and other instructions
New features in this version

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the md5sum of the .exe to the true fingerprint. You will need a version of md5sum for windows: both graphical and

Instalación de R

1. Ingresamos al https://www.r-project.org/
2. Damos clic en CRAN (ver al costado izquierdo)
3. Escogemos un servidor desde el cual hacer la descarga
4. Se escoge el sistema operativo en el cual se va a instalar
5. Se da clic en el enlace base
6. Clic en Download R 3.3.1 for Windows (70 megabytes, 32/64 bit)
7. Esperar la descarga
8. Siguiente, siguiente, siguiente ...

RStudio

RStudio es un ambiente de desarrollo integrado para R es decir?

- Es un ambiente más amable, funcional y eficiente
- Es menos exigente para el usuario que está iniciando
- Provee ayuda y asistencia en la escritura de comandos

Funcionalidades propias

- R-Script
- Notebooks
- Documentos R-Markdown
- Documentos Sweave (Latex + R)
- Presentaciones R-Markdown
- Shinny Web Apps
- R HTML
- R Presentation

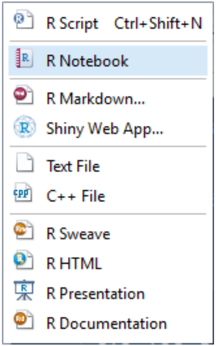


Figura: Menu nuevo documento

INSTALACIÓN DE RStudio

1. Ingresamos a https://www.rstudio.com/
2. Damos clic *Products* → *RStudio* → *Desktop*
3. Clic en **Download RStudio Desktop**
4. Siguiente, siguiente, siguiente...

INSTALACIÓN DE RStudio

1. Ingresamos a https://www.rstudio.com/
2. Damos clic *Products* → *RStudio* → *Desktop*
3. Clic en **Download RStudio Desktop**
4. Siguiente, siguiente, siguiente...

Tools ⟶ Global Options

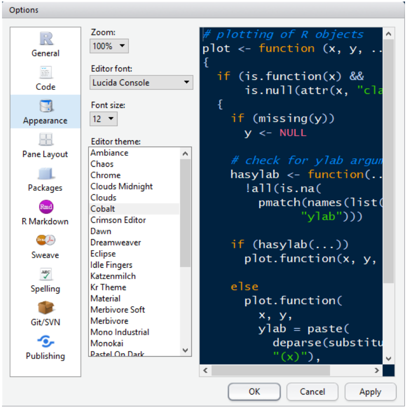


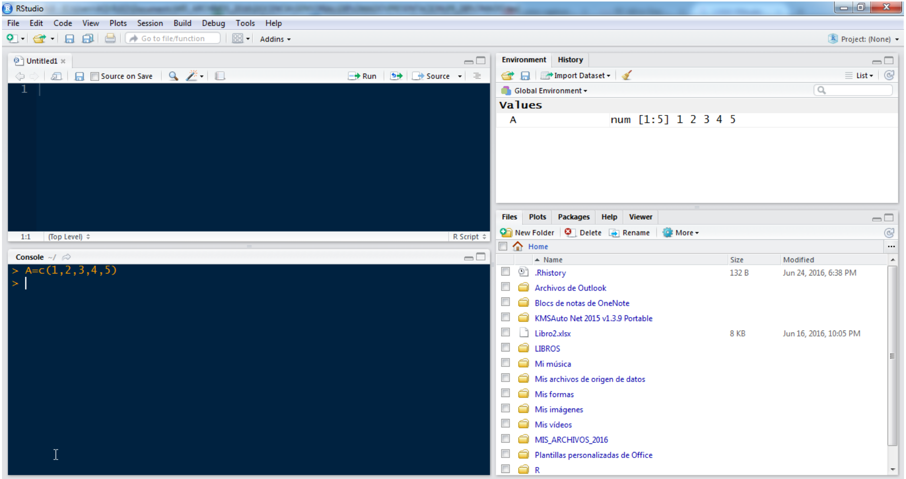
Figura: Opciones Globales

Interface de RStudio

La integración que logra RStudio la logra desde cuatro ventanas

- **Consola**: Ejecución de comandos
- **Scripts**: Desarrollo de programas, R-Markdown, Documentos Latex o Sweave entre otros
- **Navegador**: Permite moverse en directorios, gestionar paquetes, visualizar gráficas, buscar ayuda
- **Historial**: Registro de todo el historial de uso.

Interface de RStudio



Subsección 2

Introducción a R

¿Que puede hacer R?

Ejemplos graficos con R

demo(graphics)

Otras graficas posibles...

demo(persp)

Gestión de paquetes en RStudio

- Cargar un paquete instalado: `library(Nombre paquete)`
- Instalar un paquete:`install.packages("Nombre paquete")`
- Eliminar un paquete:`remove.packages("Nombre paquete")`

Gestión de paquetes en RStudio

- Cargar un paquete instalado: `library(Nombre paquete)`
- Instalar un paquete:`install.packages("Nombre paquete")`
- Eliminar un paquete:`remove.packages("Nombre paquete")`

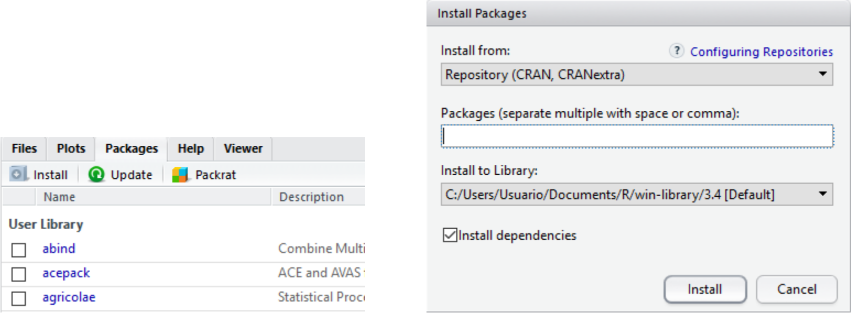


Figura: Paquetes en RStudio

Directorio de trabajo

- Directorio actual de trabajo: `getwd()`
- Declarar directorio de trabajo: `setwd(C:/MIS ARCHIVOS/CARPETA1)`
- Para ver que archivos hay en el directorio de trabajo: `dir()`
- Para ver la variables actuales en el espacio de trabajo: `ls()`

Directorio de trabajo

- Directorio actual de trabajo: `getwd()`
- Declarar directorio de trabajo: `setwd(C:/MIS ARCHIVOS/CARPETA1)`
- Para ver que archivos hay en el directorio de trabajo: `dir()`
- Para ver la variables actuales en el espacio de trabajo: `ls()`

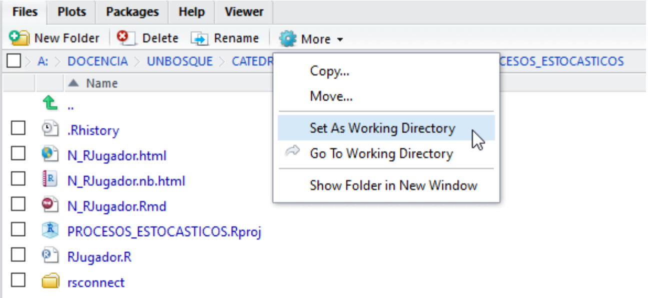


Figura: Navegador de archivos

Principales Objetos en R

La asignación en R se realiza mediante la instrucción `<-` o también `=`
Aunque hay muchos tipos de objetos en R, los más importantes para propósitos generales son:

- Escalares
- Vectores
- Matrices
- Arreglos (Array)
- Dataframe
- Listas

La instrucción `class()` permite conocer el tipo de objeto.

Ejemplo

```
a<-45
class(a)
[1] "numeric"
```

```
b<-4+5
c=b/2
d=sqrt(c)
```

```
e<-log(5)
f=3^2
d=tan(f)
```

Vectores

En R los vectores pueden ser Numéricos, Textuales o Lógicos y se crean con la instrucción `c()` o con la función `scan()`

Vector Numérico

```
A=c(1,3,5,6,9)
```

Textual

```
B<-c(" Cálculo I", " Cálculo II"," Estadística I")
```

Vector Lógico

```
C<-c(FALSE,FALSE,TRUE,TRUE,TRUE)
```

Funciones útiles para trabajo con vectores

Vectores

Usando el vector `A=c(1,3,5,6,9)`

Longitud	`n=length(A)`	`[1] 5`
Tipo	`class(A)`	`[1] "numeric"`
Mínimo	`min(A)`	`[1] 1`
Máximo	`max(A)`	`[1] 9`
Posición de un elemento	`A[4]`	`[1] 6`
Varios elementos	`A[1:3]`	`[1] 1 3 5`
	`A[c(1,3)]`	`[1] 1 5`
Media	`mean(A)`	`[1] 4.8`
Desviación Estóndar	`sd(A)`	`[1] 3.03315`
Varianza	`var(A)`	`[1] 9.2`
Filtros	`A[A<6]`	`[1] 1 3 5`
	`A[A<3 | A>8]`	`[1] 1 9`
	`A[A>=3 & A<8]`	`[1] 3 5 6`

Matrices

Las matrices son un tipo muy importante de objeto en todo lenguaje de programación pues brinda toda la potencia de calculo matricial de gran utilidad en la estadística.
La instrucción para declarar o ingresar una matriz es:

```
matrix(c(),nrow = n,ncol = m)
```

Ejemplo

```
B<-matrix(c(2,4,6,3,5,7),nrow = 2,ncol = 3)
     [,1] [,2] [,3]
[1,]    2    6    5
[2,]    4    3    7
C<-matrix(c(2,4,6,3,5,7),nrow = 2,ncol = 3,byrow=T)
     [,1] [,2] [,3]
[1,]    2    4    6
[2,]    3    5    7
```

Funciones utiles para trabajo con matrices

Matrices

Usando la matriz

```
      [,1] [,2] [,3]
[1,]    2    6    5
[2,]    4    3    7
```

Dimensión	`dim(B)`	`[1] 2 3`
Tipo	`class(B)`	`[1] "matrix"`
Posición de un elemento	`B[2,3]`	`[1] 7`
Fila	`B[1,]`	`[1] 2 6 5`
Columna	`B[,2]`	`[1] 6 3`
Transpuesta	`t(B)`	
Borra columna	`C=B[,1:2]`	
Determinante	`det(C)`	
Inversa	`solve(C)`	

Dataframe

Este es el elemento que con mayor frecuencia se emplea para almacenar datos estadísticos. En esencia es un arreglo bidimensional en el cual como es común se emplean

Filas $\rightarrow$ Registros

Columnas $\rightarrow$ Variables

Ejemplo

```
ID<-c(1:10)
GENERO<-c("H","M","M","M","H","H","M","M","H","H")
EDAD<-c(18,21,19,20,19,21,20,19,21,18)
IMC<-c(23.1,24.2,25.3,27.1,19.0,23.0,22.8,23.0,21.8,24.3)

DT<-data.frame(ID,GENERO,EDAD,IMC)

rownames(DT)<-c("Jhon","Luz","Olga","Sara","Luis",
"Diego","Ana","Ivon","Juan","Carlos")
```

Dataframe

Ejemplo

DT

ID	GENERO	EDAD	IMC
Jhon	1	H	18 23.1
Luz	2	M	21 24.2
Olga	3	M	19 25.3
Sara	4	M	20 27.1
Luis	5	H	19 19.0
Diego	6	H	21 23.0
Ana	7	M	20 22.8
Ivon	8	M	19 23.0
Juan	9	H	21 21.8
Carlos	10	H	18 24.3

head(DT)

ID	GENERO	EDAD	IMC
Jhon	1	H	18 23.1
Luz	2	M	21 24.2
Olga	3	M	19 25.3
Sara	4	M	20 27.1

tail(DT)

ID	GENERO	EDAD	IMC
Ana	7	M	20 22.8
Ivon	8	M	19 23.0
Juan	9	H	21 21.8
Carlos	10	H	18 24.3

Dataframe

Ejemplo

```
str(DT)
'data.frame': 10 obs. of  4 variables:
$ ID    : int  1 2 3 4 5 6 7 8 9 10
$ GENERO: Factor w/ 2 levels "H","M": 1 2 2 2 1 1 2 2 1 1
$ EDAD  : num  18 21 19 20 19 21 20 19 21 18
$ IMC   : num  23.1 24.2 25.3 27.1 19 23 22.8 23 21.8 24.3
```

Para iniciar las principales funciones útiles para el manejo de Dataframe son:

- `data.frame`
- `head`
- `tail`
- `rownames`
- `colnames`
- `str`
- `view`
- `dim`

Importación de Archivos de datos

Existen diversas formas de importar datos a R una de las cuales se describe aquí. A partir de una hoja de Excel, se guarda como archivo ‘csv‘ en el directorio de trabajo y luego se importa mediante la instrucción

```
DT<-read.table(file.choose(),header=T,sep=",",dec=".")
```

Es muy importante especificar los elementos

`header=TRUE` Reconocer nombres de columna

`sep=","` El separador entre columnas(, ; : tab espacio)

`dec="."` Separador de cifras decimales (. o ,)

Filtrando un dataframe

Ejemplo

```
DT1=DT[DT$PAIS==1,]
str(DT1)
'data.frame': 200 obs. of  5 variables:
$ TIPO : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 ...
$ PAIS : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 ...
$ BOLSA: int  1 2 3 4 5 6 7 8 9 10 ...
$ ARETE: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 ...
$ PESO : num  9.09 9.06 9.62 9.21 9.32 9.22 9.17 9.1  ...

TIPO PAIS BOLSA ARETE PESO
1    1    1     1     1 9.09
2    1    1     2     1 9.06
3    1    1     3     1 9.62
4    1    1     4     1 9.21
```

Factores

En general se puede redefinir el tipo de variable de una columna en un dataframe con el fin de preparar la base de datos para próximos análisis
Algunas conversiones importantes son:

- `as.numeric(A)` Convierte un elemento en vector numérico
- `as.factor(A)` Convierte un vector en factor
- `as.matrix(A)` Convierte un arreglo en matriz
- `as.data.frame(A)` Convierte un arreglo en data.frame

En el ejemplo se tendría

Ejemplo

```
DT$PAIS=as.factor(DT$PAIS)
```

Practica 1

En grupos de dos personas van a realizar las siguientes actividades

1. Importe el archivo `BASE ATB.csv`
2. Revísela y asegúrese que todas las variables sean tipo numéricas o enteros, excepto `MST` que es un factor.
3. Seleccione 5 atributos y obtenga un resumen de medidas descriptivas para cada muestra (`MST`).
4. Determine si los datos de esos 5 atributos se distribuyen como una normal.
5. Realice un histograma de frecuencias para una variable de las cinco escogidas.

Sección 3

ANÁLISIS UNIVARIADO DE DATOS EN SENSORIAL

Estadística Univariada

En términos generales describir todas las herramientas estadísticas univariadas que pueden ser empleadas resulta una amplia tarea, pero sin duda unas de las principales áreas que se suelen emplear son:

- E. Descriptiva
- E. Inferencial
  - Pruebas de hipótesis
  - Diseño y análisis de experimentos
  - Modelos estadísticos (Regresión)
  - Análisis bayesiano
  - Modelos longitudinales
- Muestreo

Subsección 1

Inferencia Estadística

Inferencia Estadística

Es el procedimiento mediante el cual se pueden sacar conclusiones acerca de la población, partiendo de la información contenida en una muestra extraída de una población.

En general hay dos tipos de inferencia estadística

- Estimación
- Prueba de Hipótesis

Inferencia Estadística

Es el procedimiento mediante el cual se pueden sacar conclusiones acerca de la población, partiendo de la información contenida en una muestra extraída de una población.
En general hay dos tipos de inferencia estadística

- Estimación
- Prueba de Hipótesis

Sobre que elementos se desarrolla la estimación o las pruebas de hipótesis?

- Medias
- Proporciones
- Varianzas
- Coeficientes de regresión
- Correlaciones
- ...

Estimación

Es el proceso mediante el cual intentamos determinar el valor de un parámetro de la población, a partir de la información de una muestra.

Estimación: es el valor numérico que asignamos a un parámetro.
Estimador: es la formula utilizada para hacer la estimación.

Ejemplo

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{X} = 3{,}25$$

Estimación

Es el proceso mediante el cual intentamos determinar el valor de un parámetro de la población, a partir de la información de una muestra.

Estimación: es el valor numérico que asignamos a un parámetro.
Estimador: es la formula utilizada para hacer la estimación.

Ejemplo

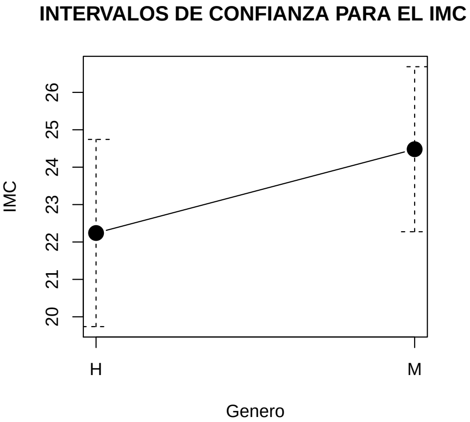
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{X} = 3{,}25$$

Hay dos tipos de estimaciones para cualquier parámetro, estos son:

- Estimación Puntual
- Estimación por Intervalo

Uso de la estimación por intervalo

Los IC se emplean mayoritariamente para variables continuas o también llamadas variables métricas.



El nivel de confianza denota la probabilidad de contener el valor verdadero.

Hipótesis

- Hipótesis: Afirmación no probada acerca de un problema o situación especifica.
- Hipótesis científica: es una proposición aceptable que ha sido formulada a través de la recolección de información y datos, aunque no esté confirmada, sirve para responder de forma alternativa a un problema con base científica.
- Hipótesis estadística:En el contexto estadístico una Hipótesis se entiende como una afirmación acerca de un conjunto de parámetros que describen una característica de una o más poblaciones.

Una Prueba de Hipótesis entonces debe entenderse como un procedimiento mediante el cual se determina el grado de verdad más probable de tal afirmación, en función de la información disponible en la muestra.

Error Tipo I y II en Pruebas de Hipótesis

	ACEPTA H_0	RECHAZA H_0
H_0 Verdadera	Buena decisión	Error Tipo I
H_0 Falsa	Error Tipo II	Buena decisión

Error Tipo I y II en Pruebas de Hipótesis

	ACEPTA H_0	RECHAZA H_0
H_0 Verdadera	Buena decisión	Error Tipo I
H_0 Falsa	Error Tipo II	Buena decisión

En general se emplea la notación siguiente:

- α: Error Tipo I
- β: Error Tipo II

Toda prueba de hipótesis se diseña con el objeto de controlar el error tipo I y se deja libre el tipo II.

En el caso de los intervalos de confianza el término $\left(1-\dfrac{\alpha}{2}\right)\times 100\,\%$ denota precisamente el *Nivel de confianza del intervalo*

TIPOS DE PRUEBAS

Para varios casos de inferencia estadística existen fundamentalmente dos tipos de pruebas que en general se pueden describir como:

- Paramétricas: Estas hacen consideraciones acerca de la distribución de la población.
  - Prueba t (una y dos medias)
  - ANOVA
  - Regresión
  - ...

TIPOS DE PRUEBAS

Para varios casos de inferencia estadística existen fundamentalmente dos tipos de pruebas que en general se pueden describir como:

- Paramétricas: Estas hacen consideraciones acerca de la distribución de la población.
  - Prueba t (una y dos medias)
  - ANOVA
  - Regresión
  - ...
- No Paramétricas: Generalmente no asume distribuciones para la población.
  - Prueba de Wilcoxon, Rangos Signados
  - Kruskal-Wallis
  - Friedman,
  - ...

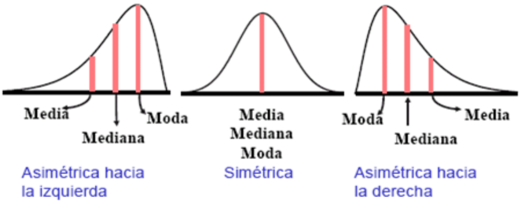
NOTA IMPORTANTE:En general estas pruebas tiendes a coincidir en la conclusión, sin embargo, en ocasiones se presentan discrepancias cuando los datos presentan tendencia hacia un lado de la escala de medida.

Medias

Estas pruebas aplican a: Variables continuas

En el caso de las media las hipótesis se orientan a los siguientes casos:

1. Una muestra
2. Dos muestras
   1. Independientes
   2. Relacionadas
3. Tres o más muestras



Una Media

La prueba paramétrica corresponde a la prueba t y la no paramétrica a la de Wilcoxon (en su versión de una muestra).

Hipótesis:

$$H_0 : \mu = \mu_0 \qquad vs. \qquad \begin{cases} H_a : \mu > \mu_0 \\ H_a : \mu \neq \mu_0 \\ H_a : \mu < \mu_0 \end{cases}$$

donde μ_0 es un valor fijo y conocido frente al cual se desea hacer la comparación.

En Rcmdr...

Menú → Estadísticos → Medias→Test t para una muestra

Menu → Estadísticos → Test no paramétricos → Test de Wilcoxon para una muestra

Implementación en R (Prueba t)

Instrucción

```
with(DT, (t.test(PESO, alternative='two.sided', mu=8,
conf.level=.95)))
```

Salida

```
One Sample t-test

data:  PESO
t = -6.8476, df = 395, p-value = 2.886e-11
alternative hypothesis: true mean is not equal to 8
95 percent confidence interval:
7.091569 7.496846
sample estimates:
mean of x
7.294207
```

Implementación en R (Prueba de Wilcoxon)

Instrucción

```
with(DT, median(PESO, na.rm=TRUE))
with(DT, mean(PESO, na.rm=TRUE))
with(DT, wilcox.test(PESO, alternative='two.sided',
mu=8))
```

Salida

```
Wilcoxon signed rank test with continuity correction

data:  PESO
V = 19701, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 8
```

Dos Medias

En el caso de dos muestras se debe considerar si estás son:

1. Independientes: Cuando las observaciones o datos corresponden a muestras diferentes.
   Ejemplo: Grupos de usuarios diferentes que participan en una prueba.

2. Relacionadas: Cuando las observaciones provienen de las mismas unidades muestrales y se han tomado en dos momentos diferentes (También llamadas muestras pareadas).
   Ejemplo: Un grupo de jueces sensoriales cuantifican la intensidad olfativa de una fragancia en dos momentos diferentes luego de ser aplicada en un portador.

Dos Medias

Muestras Independientes

La prueba paramétrica corresponde a la prueba t y la no paramétrica a la de Wilcoxon (en su versión para dos muestras).

Hipótesis:

$$H_0 : \mu_1 = \mu_2 \qquad vs. \qquad \begin{cases} H_a : \mu_1 > \mu_2 \\ H_a : \mu_1 \neq \mu_2 \\ H_a : \mu_1 < \mu_2 \end{cases}$$

En Rcmdr...

Menú → Estadísticos → Medias→Test t para muestras independientes

Menú → Estadísticos → Test no paramétricos → Test de Wilcoxon para dos muestras

Implementación en R (Prueba t)

Instrucción

```
t.test(PESO TIPO, alternative='two.sided',
conf.level=.95, var.equal=FALSE, Rcmdr+ data=DT)
```

Salida

```
Welch Two Sample t-test

data:  PESO by TIPO
t = 191.73, df = 296.93, p-value < 2.2e-16
alternative hypothesis: true difference in means
is not equal to 0 95 percent confidence interval:
4.033423 4.117082
sample estimates:
mean in group Corto mean in group Largo
9.331833            5.256581
```

Implementación en R (Prueba de Wilcoxon)

Instrucción

```
with(DT, tapply(PESO, TIPO, median, na.rm=TRUE))
wilcox.test(PESO  TIPO, alternative="two.sided",
data=DT)
```

Salida

```
Corto  Largo
9.4100 5.2495

Wilcoxon rank sum test with continuity correction

data:  PESO by TIPO
W = 39204, p-value < 2.2e-16
alternative hypothesis:
true location shift is not equal to 0
```

Dos Medias

Muestras Relacionadas

La prueba paramétrica corresponde a la prueba t y la no paramétrica a la de Wilcoxon para muestras pareadas.

Hipótesis:

$$H_0 : \mu_1 - \mu_2 = 0 \qquad vs. \qquad \begin{cases} H_a : \mu_1 - \mu_2 > 0 \\ H_a : \mu_1 - \mu_2 \neq 0 \\ H_a : \mu_1 - \mu_2 < 0 \end{cases}$$

En Rcmdr...

Menú → Estadísticos → Medias→Test t para datos relacionados

Menú → Estadísticos → Test no paramétricos → Test de Wilcoxon para dos muestras

Practica 1: Prueba de hipótesis para medias

BASE:Aretes.csv

1. Para cada tipo de arete se le pidió al proveedor que tuvieran un peso especifico donde el liviano debe tener un peso medio de 5.3 gramos y el pesado debe tener 9.1 gramos. ¿Cuál es la prueba de hipótesis correspondiente?
2. Cada par de aretes se marca con un código en cada grupo y se dividen en dos grupos que son enviados a dos países. ¿Presentan las muestras que van a los dos países el mismo valor medio?
3. Realice un gráfico para cada caso.

Una Proporción

Estas pruebas aplican a: Variables discretas ordinales y nominales

La prueba paramétrica corresponde a la prueba Z.

Hipótesis:

$$H_0 : \pi = \pi_0 \qquad vs. \qquad \begin{cases} H_a : \pi > \pi_0 \\ H_a : \pi \neq \pi_0 \\ H_a : \pi < \pi_0 \end{cases}$$

donde π_0 es un valor fijo y conocido frente al cual se desea hacer la comparación.

En Rcmdr...

Menú → Estadísticos → Proporciones→Test de proporciones para una muestra

Dos Proporciones

La prueba paramétrica corresponde a la prueba Z.

Hipótesis:

$$H_0 : \pi_1 = \pi_2 \qquad vs. \qquad \begin{cases} H_a : \pi_1 > \pi_2 \\ H_a : \pi_1 \neq \pi_2 \\ H_a : \pi_1 < \pi_2 \end{cases}$$

En Rcmdr...

Menú → Estadísticos → Proporciones→Test de proporciones para dos muestras

Practica 2: Prueba de hipótesis para proporciones

1. Identifique las variables SABOR, CALIDAD, CALIDAD y MARCA.
2. Recodificar cada una de las variables
3. Convertir la variable en factor. En Rcmdr ir a DATOS; Modificar variables del conjunto de datos activo; Convertir variable numérica en factor
4. Hacer la prueba de hipótesis para una proporción. Estadísticos; Proporciones; Test de proporciones para una muestra

Subsección 2

Pruebas Discriminantes

Seis Pruebas básicas

Entre las diversas pruebas sensoriales que se pueden realizar, las seis pruebas básicas son:

PRUEBA	No. MUESTRAS	RESPUESTA
2-AFC	2	Más Fuerte o Débil
3-AFC	3	Más Fuerte o Débil
Duo-Trio	3	Par igual
Triangular	3	Muestra diferente
A - No A	1	Igual a referente
Igual-Diferente	2	Diferencia o no

Las cuatro primeras corresponden a **Métodos de selección forzada**, pues los participantes deben hallar una selección aún si no la pueden detectar claramente.

Métodos de Selección Forzada

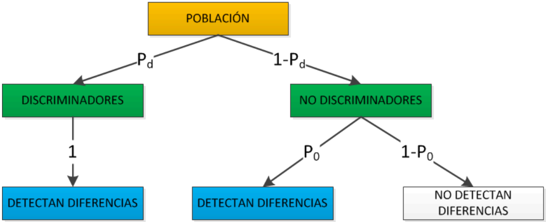
Estos métodos se pueden emplear con dos finalidades

1. Pruebas de diferencia
2. Pruebas de Preferencia

En general las seis pruebas de discriminación básicas son procedimientos basados en pruebas de hipótesis para proporciones las cuales para muestras pequeñas emplean la Distribución Binomial a fin de establecer el resultado de la prueba.

$$X \sim B(n,p) \qquad P[X = i] = \binom{n}{i} p^i (1-p)^{n-i}$$

Modelo para pruebas de diferencia



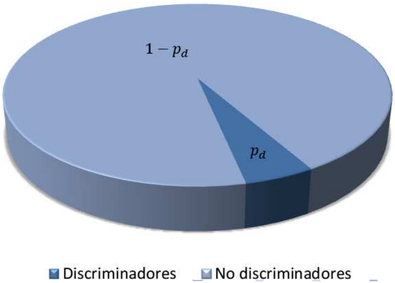
$$p_c = p_d + p_0(1 - p_d) \qquad (1)$$

donde:

p_c: Prob. de respuesta correcta

p_0: Prob. de adivinar

p_d: Proporción de discriminadores



Modelo para pruebas de diferencia

En una prueba de diferencia entonces los pares de hipótesis a probar serón:

$$H_0 : p_c = p_0$$
$$H_a : p_c \neq p_0$$

La proporción de aciertos es diferente de la probabilidad de adivinar

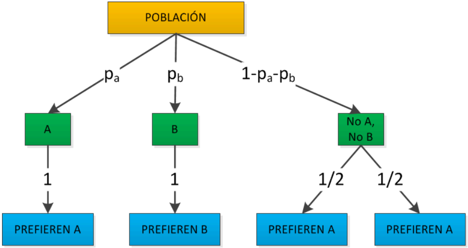
$$H_0 : p_c = p_0$$
$$H_a : p_c > p_0$$

La proporción de aciertos es mayor que la probabilidad de adivinar

La probabilidad de adivinar en cada una de la pruebas es:

PRUEBA	p_0
2-AFC	$1/2$
Duo-Trio	$1/2$
3-AFC	$1/3$
Triangular	$1/3$

Modelo para pruebas de preferencia



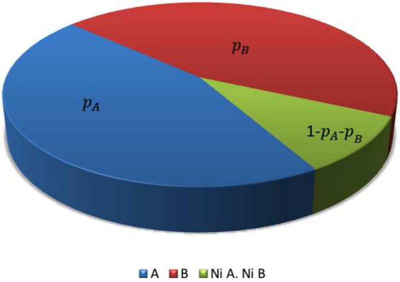
$$P_A = p_a + \frac{1 - p_a - p_b}{2} \qquad (2)$$

donde:

p_a: Prop. que prefiere A

p_b: Prop. que prefiere B

P_A: Prop. que escoge A



Estimación de p_d

En pruebas discriminativas el tamaño de muestra se establece a partir de fijar los valores α, β y p_d por lo cual es indispensable establecer una estimación de tal valor.

$$\hat{p}_d = \frac{\hat{p}_c - p_0}{1 - p_0} \tag{3}$$

donde $\hat{p}_c$ es la proporción observada de aciertos en el panel. Con lo cual un intervalo de confianza del $95\,\%$ para p_d será:

$$\hat{p}_d \pm 1{,}96\sqrt{V(\hat{p}_d)} \tag{4}$$

Donde $V(\hat{p}_d)$ es la varianza de la estimación y está dada por:

$$V(\hat{p}_d) = \frac{1}{(1 - p_0)^2}\frac{\hat{p}_d(1 - \hat{p}_d)}{N} \tag{5}$$

Ejemplo...

En una prueba con 100 consumidores se le presenta a cada participante dos productos lácteos A y B, siendo el primero, el producto que actualmente esta en el mercado y B corresponde al mismo producto con un cambio importante de una materia prima. A cada participante se le pide probar los productos y determinar si son iguales o diferentes. Luego de la prueba se encuentra que 62 personas detectan la diferencia. ¿Cuál es la proporción de discriminadores p_d para ese producto en la población?

Estimación puntual de aciertos: $p_c = 0{,}62$. Probabilidad de adivinar: $p_0 = 0{,}5$

$$\hat{p}_d = \frac{\hat{p}_c - p_0}{1 - p_0} \qquad \hat{p}_d = \frac{0{,}62 - 0{,}5}{1 - 0{,}5} \qquad \hat{p}_d = 0{,}24$$

$$V(\hat{p}_d) = \frac{1}{(1-p_0)^2}\frac{\hat{p}_d(1-\hat{p}_d)}{N} \qquad V(\hat{p}_d) = \frac{1}{(0{,}5)^2}\frac{0{,}24(0{,}76)}{100} \qquad V(\hat{p}_d) = 0{,}097$$

De donde

$$\hat{p}_d \pm 1{,}96\sqrt{V(\hat{p}_d)} \qquad \boxed{(0{,}05, 0{,}43)}$$

Ejemplo...

En una prueba triangular se registro 22 respuestas correctas de un total de 52 evaluaciones. ¿Cual es el intervalo de confianza del $95\,\%$ para la proporción de discriminadores?

Tamaños de Muestra para una prueba triangular

α		β 0.50	0.40	0.30	0.20	0.10	0.05	0.01	0.001
	p_d=50%								
0.40		3	3	3	6	8	9	15	26
0.30		3	3	3	7	8	11	19	30
0.20		4	6	7	7	12	16	25	36
0.10		7	8	8	12	15	20	30	43
0.05		7	9	11	16	20	23	35	48
0.01		13	15	19	25	30	35	47	62
0.001		22	26	30	36	43	48	62	81
	p_d=40%								
0.40		3	3	6	6	9	15	26	41
0.30		3	3	7	8	11	19	30	47
0.20		6	7	7	12	17	25	36	55
0.10		8	10	15	17	25	30	46	67
0.05		11	15	16	23	30	40	57	79
0.01		21	26	30	35	47	56	76	102
0.001		36	39	48	55	68	76	102	130
	p_d=30%								
0.40		3	6	6	9	15	26	44	73
0.30		3	8	8	16	22	30	53	84
0.20		7	12	17	20	28	39	64	97
0.10		15	15	20	30	43	54	81	119
0.05		16	23	30	40	53	66	98	136
0.01		33	40	52	62	82	97	131	181
0.001		61	69	81	93	120	138	181	233

Figura: Tomado de Sensory Evaluation Techniques

Pruebas de Similaridad

No inferioridad

$$H_0 : p_a - p_b \leq -\Delta_0 \quad vs \quad H_a : p_a - p_b > -\Delta_0$$

La proporción p_a no es inferior a p_b

No Superioridad

$$H_0 : p_a - p_b \geq \Delta_0 \quad vs \quad H_a : p_a - p_b < \Delta_0$$

La proporción p_a es mayor a p_b

Subsección 3

Diseño y Análisis de Experimentos

Conceptos de Diseño Experimental

- Factor: Son variables controlables de interés para el investigador.
- Nivel: Cada uno de los valores que toma un factor (Alto, medio, bajo).
- Variable respuesta: Es lo que se mide o cuenta y es el objetivo de todo el diseño.
- Tratamiento: Son las combinaciones de todos los niveles de los factores.
- Modelo: Es la expresión matem?tica que describe te?ricamente la variable respuesta en función de los factores.

Conceptos de Diseño Experimental

- Bloque: Son variables categóricas que se introducen al diseño con el fin de organizar el material experimental en grupos más homogéneos entre si.
- Unidad Experimental: Es cada individuo, animal u objeto que se somete a un tratamiento.
- Observación: Cada valor de la variable respuesta en cada unidad experimental.
- Unidad Observacional: Es el elemento especifico sobre el cual se realiza la medición en la unidad experimental.
- Replica: Corresponde al número de unidades experimentales que se asignan a cada tratamiento.

Ejemplo

En el desarrollo de una mascara para pestañas (pestañina) se desea mejorar la distribución del producto cuando las usuarias lo aplican, para lo cual se proponen siete tipos diferentes de aplicadores los cuales se probaron con dos formulaciones comercialmente exitosas de la compañía. La evaluación se realiza con 70 voluntarias y 3 jueces, así cada usuaria se aplica las dos mascaras con el mismo cepillo y las jueces evalúan de manera independiente las dos muestras.
Tras la aplicación del producto, cada juez evalúa de manera independiente si el producto está homogeneamente distribuido usando una escala de siete puntos.

Ejemplo

Factor:	Tipo de aplicador
Niveles:	Los siete tipos de aplicador
Bloque:	Las dos tipos de mascara empleada
Tratamientos:	Los siete tipos de aplicador
Unidades Experimental:	Usuaria
Unidad Observacional:	Cada ojo
V. Respuesta:	Valoración de las jueces
No. de replicas:	10

DISEÑOS DE UN FACTOR

Cuando se tiene un solo factor de interés el análisis estadístico tiene como principal objetivo determinar si existen diferencias significativas entre las medias de los tratamientos que para el caso son los mismos niveles del factor.
Esto de expresa en la siguiente hipótesis estadística:

$$\boxed{H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \qquad vs. \qquad H_a : \mu_i \neq \mu_j \qquad \forall i \neq j} \tag{6}$$

Está hipótesis se prueba mediante la técnica de **Análisis de Varianza**.

DISEÑOS DE UN FACTOR

Cuando se tiene un solo factor de interés el análisis estadístico tiene como principal objetivo determinar si existen diferencias significativas entre las medias de los tratamientos que para el caso son los mismos niveles del factor.
Esto de expresa en la siguiente hipótesis estadística:

$$\boxed{H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \qquad vs. \qquad H_a : \mu_i \neq \mu_j \qquad \forall i \neq j} \tag{6}$$

Está hipótesis se prueba mediante la técnica de **Análisis de Varianza**.

IMPORTANTE

Notese que el Análisis de Varianza solo le permite determinar su las medias son iguales o si existe al menos un par de medias diferentes. Pero no le establece cuales pares de tratamientos son diferentes.

Estructura de Datos

Tratamientos	Observaciones				Medias
1	y_{11}	y_{12}	$\cdots$	y_{1r}	$\overline{y}_1$
2	y_{21}	y_{22}	$\cdots$	y_{21}	$\overline{y}_2$
$\vdots$	$\vdots$		$\ddots$	$\vdots$	$\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kr}	$\overline{y}_k$

Modelo estadístico

$$\boxed{y_{ij} = \mu_i + \epsilon_{ij}} \tag{7}$$

donde $i = 1, \ldots, k \quad j = 1, \ldots, r$ y además:

μ_i: Media del *i-esimo* tratamiento

ϵ_{ij}: Error experimental

Tabla ANOVA

El análisis de varianza para un diseño de un factor tiene la siguiente estructura:

Fuente	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrados Medios**	$\mathbf{F_c}$
$Tratamientos$	$k-1$	$SCTr$	$CMTr$	$\frac{CMTr}{CME}$
$Error$	$n-k$	SCE	CME	
$Total$	$n-1$	SCT		

Adicionalmente siempre se reporta el valor P correspondiente para decidir respecto al nivel de significancia α.

Pruebas de Comparación Múltiple

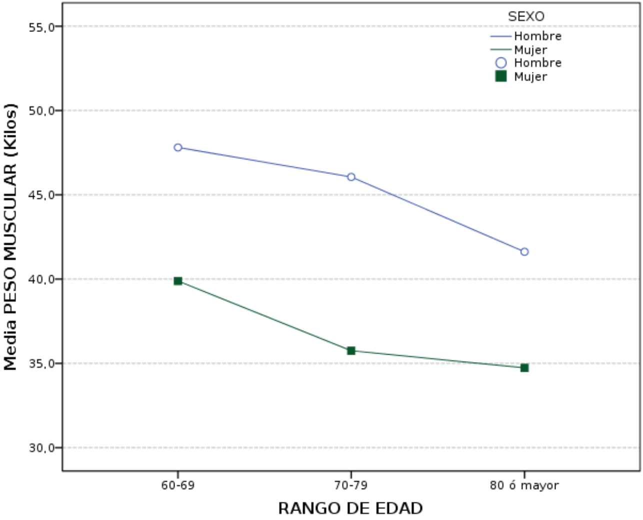
Cuando del análisis de varianza se concluye que existen diferencias estadísticamente significativas entre las medias de los tratamientos es necesario establecer que pares de ellos presentan diferencia y cuales no. Este es el propósito de las *Pruebas de comparación múltiple*.

Entre las múltiples pruebas existentes algunas importantes son:

- Diferencias mínimas significativas (LSD)
- Tukey
- Bonferroni
- Scheffe
- Duncan
- Dunnett (Comparación con un control)

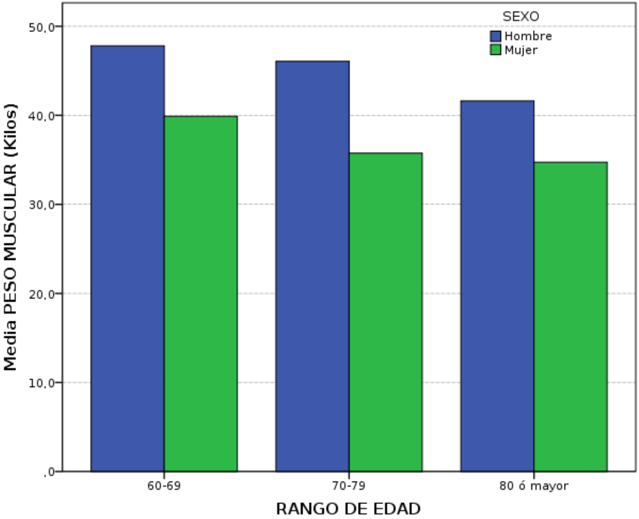
Representación Gráfica

Diagrama de Lineas



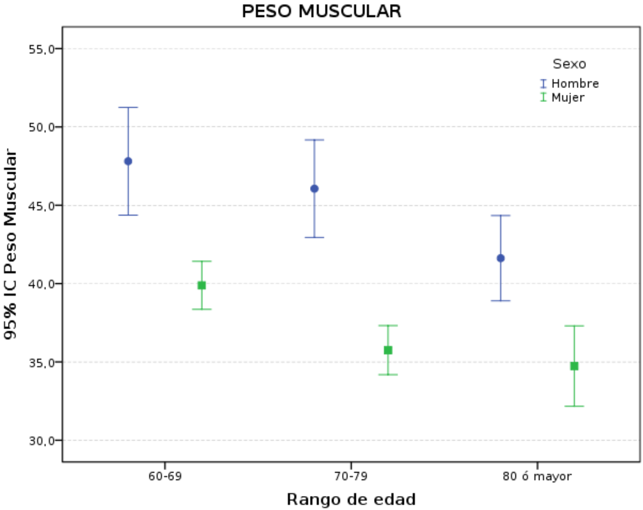
Representación Gráfica

Diagrama de Barras



Representación Gráfica

Diagrama de Barras de Error



Supuestos

El análisis de varianza hace los siguientes supuestos acerca del diseño, que deben ser validados para garantizar la confiabilidad de los resultados, estos son:

1. Igualdad de varianza en los tratamientos (*Homocedastisidad*)
   Prueba de igualdad de varianzas de Bartlett o Levene
2. Independencia de las observaciones
   Diagramas de dispersión, Prueba de Durbin-Watson
3. Las observaciones se distribuyen de acuerdo a una normal
   Prueba de Normalidad

Supuestos

El análisis de varianza hace los siguientes supuestos acerca del diseño, que deben ser validados para garantizar la confiabilidad de los resultados, estos son:

1. Igualdad de varianza en los tratamientos (*Homocedastisidad*)
   Prueba de igualdad de varianzas de Bartlett o Levene
2. Independencia de las observaciones
   Diagramas de dispersión, Prueba de Durbin-Watson
3. Las observaciones se distribuyen de acuerdo a una normal
   Prueba de Normalidad

¿Y si los supuestos no se cumple?

En general depende de cuales sean los supuestos no satisfechos, pero algunas formas de mejorar los resultados implica transformar la variable respuesta:

- $y \to \log y$
- $y \to \sqrt{y}$
- $y \to y^{\lambda}$

Implementación en R

En el archivo BASE1F.xls se presentan los datos de una prueba comparativa de duración de 4 fragancias luego de 6 horas de aplicada en piel sobre dos portadores.

```
head(BASE)
ID FR  INT
1  1  3 3.29
2  2  3 2.62
3  3  3 2.25
4  4  3 2.54

str(BASE)
'data.frame': 150 obs. of  3 variables:
$ ID  : num  1 2 3 4 5 6 7 8 9 10 ...
$ FR  : num  3 3 3 3 3 3 3 3 3 3 ...
$ INT : num  3.29 2.62 2.25 2.54 2.24 2.25 1.29 ...
```

DISEÑOS DE DOS FACTORES

Para el caso de dos factores de interés el análisis estadístico tiene como principal objetivo determinar si existen diferencias significativas entre las medias de los tratamientos que corresponden a las combinaciones de los niveles de ambos factores.
Si A y B son los factores en estudio con n y m niveles respectivamente, entonces el modelo correspondiente es:

$$\boxed{y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}} \tag{8}$$

donde $i = 1, \ldots, n \quad j = 1, \ldots, m$, $k = 1, \ldots, r$.

Con ello la principal hipótesis estadística de interés es:

$$\boxed{H_0 : \mu_{ij} = \mu_{st} \qquad vs. \qquad H_a : \mu_{ij} \neq \mu_{st}} \tag{9}$$

donde $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ se llama *Media de celda*.

Estructura de Datos

De manera similar por cada factor se puede proba las hipótesis:

$$H_0 : \alpha_i = \alpha_l \quad vs. \quad H_a : \alpha_i \neq \alpha_l$$
$$H_0 : \beta_j = \beta_t \quad vs. \quad H_a : \beta_j \neq \beta_t$$

	FACTOR B						
FACTOR A	1			$\cdots$	m		
1	y_{111}	$\cdots$	y_{11r}	$\cdots$	y_{1m1}	$\cdots$	y_{1mr}
2	y_{211}	$\cdots$	y_{211}		y_{2m1}	$\cdots$	y_{2mr}
$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$
n	y_{n11}	$\cdots$	y_{n1r}	$\cdots$	y_{nm1}	$\cdots$	y_{nmr}

Tabla ANOVA

El análisis de varianza para este diseño de dos factor tiene la siguiente estructura:

Fuente	**gl**	**SC**	**CM**	$\mathbf{F_c}$
$Factor A$	$n-1$	SCA	CMA	$\frac{CMA}{CME}$
$Factor B$	$m-1$	SCB	CMB	$\frac{CMB}{CME}$
$Interaccion$	$(n-1)(m-1)$	$SCAB$	$CMAB$	$\frac{CMAB}{CME}$
$Error$	$nm(r-1)$	SCE	CME	
$Total$	$nmr-1$	SCT		

Las pruebas de comparación múltiple aplican igualmente para contrastar los tratamientos y los niveles de cada factor si se detecta diferencia estadística.

Implementación en R

En el archivo BASE2F.xls se presentan los datos de una prueba clínica para valorar el stress de los pacientes. Se usan tres formas de controlar el stress y se considera el sexo como otro factor.

```
head(BASE2F)
Treatment Gender StressReduction
1   medical      F              1
2   medical      F              1
3   medical      F              1
4   medical      F              1
5   medical      F              2
6   medical      F              2
> str(BASE2F)
'data.frame': 60 obs. of  3 variables:
$ Treatment  : Factor w/ 3 levels "medical","mental",..: 1 1
$ Gender     : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 ...
$ StressReduction: num  1 1 1 1 2 2 3 3 3 3 ...
```

ANÁLISIS CONJUNTO

Esta es una técnica que permite la evaluación de varios perfiles de un producto con el fin de establecer las mejores combinaciones de los niveles de los factores en estudio en relación a la valoración del producto.

ANÁLISIS CONJUNTO

Esta es una técnica que permite la evaluación de varios perfiles de un producto con el fin de establecer las mejores combinaciones de los niveles de los factores en estudio en relación a la valoración del producto.

Este análisis permite:

- Identificar los factores mas importantes.
- Cuantificar cuales niveles son mejor valorados en cada factor.
- Obtener un puntaje para cada combinación (Perfil) evaluada.
- Clasificar los individuos participantes en grupos con valoraciones similares.

ANÁLISIS CONJUNTO

El análisis conjoint (análisis conjunto) es una herramienta que surge de la investigación de mercados a fin de obtener él o los mejores perfiles de atributos que puedan lograr aumentar el impacto favorable en el consumidor o cliente final.

Postulado

La utilidad (o beneficio sobre el desempeño comercial del producto) de una combinación de atributos puede ser descompuesta en contribuciones especificas de cada atributo y posiblemente de las interacciones entre ellos.

ANÁLISIS CONJUNTO

El análisis conjoint (análisis conjunto) es una herramienta que surge de la investigación de mercados a fin de obtener él o los mejores perfiles de atributos que puedan lograr aumentar el impacto favorable en el consumidor o cliente final.

Postulado

La utilidad (o beneficio sobre el desempeño comercial del producto) de una combinación de atributos puede ser descompuesta en contribuciones especificas de cada atributo y posiblemente de las interacciones entre ellos.

Típicamente es una prueba orientada al consumidor final en la cual cada participante puede valorar los perfiles mediante:

- Escala de preferencia (Liker)
- Ordenamiento

ANÁLISIS CONJUNTO

Guía de diseño

Una guía muy general para el diseño, preparación, desarrollo y análisis se presenta a continuación:

1. Definición del Problema
   - Definición del Problema
   - Selección de atributos y niveles
2. Diseño de perfiles
   - Preparación del diseño ortogonal, perfiles, tarjetas
   - Administración de la muestra
3. Desarrollo del análisis
   - Estimación de la funciones de preferencia (partworth)
   - Importancia de atributos
4. Uso de resultados
   - Segmentación
   - Selección de mejores perfiles
   - Determinación de precios
   - Estimación de canibalización
5. Reporte

Análisis Estadístico

En el Análisis Conjoint (AC) clásico, se emplea un análisis basado en un Modelo Lineal en el cual los factores (atributos) son la variables explicativas, las cuales se introducen como variables dummy y la variable respuesta es la *Utilidad*. Para tres atributos el modelo más sencillo resulta ser:

$$U = \beta_0 + U_1 + U_2 + U_3 + \epsilon_i$$

Análisis Estadístico

En el Análisis Conjoint (AC) clásico, se emplea un análisis basado en un Modelo Lineal en el cual los factores (atributos) son la variables explicativas, las cuales se introducen como variables dummy y la variable respuesta es la *Utilidad*. Para tres atributos el modelo más sencillo resulta ser:

$$U = \beta_0 + U_1 + U_2 + U_3 + \epsilon_i$$

donde los términos U_i corresponde a la función de utilidad para el atributo i (partworths).

Dependiendo de la naturaleza de los atributos a trabajar, los partworths U_i pueden ser:

- Los atributos mismos cuando estos son valores en escala continua, esto es $U_i = \beta_i X_i$.
- Agregaciones de variables dummy correspondientes a los niveles del atributo cuando está en escala ordinal o nominal.

$$U_1 = \beta_{11} x_{11} + \beta_{12} x_{12} \qquad \text{Atributo de tres niveles}$$

Análisis Estadístico

Siendo muy común que los atributos estén en escala ordinal o nominal consideremos dos atributos,

$$\mathbf{X_1} \text{ (Alto, Medio, Bajo) y } \mathbf{X_2} \text{ (A, B)}$$

de tal manera que el modelo para este caso será:

$$U = \beta_0 + U_1 + U_2 + \epsilon_i$$

Análisis Estadístico

Siendo muy común que los atributos estén en escala ordinal o nominal consideremos dos atributos,

$$\mathbf{X_1} \text{ (Alto, Medio, Bajo) y } \mathbf{X_2} \text{ (A, B)}$$

de tal manera que el modelo para este caso será:

$$U = \beta_0 + U_1 + U_2 + \epsilon_i$$

donde

$$U_1 = \beta_{11}x_{11} + \beta_{12}x_{12}$$
$$U_2 = \beta_{21}x_{21}$$

x_{11}	x_{12}	x_{21}	Perfil
1	0	1	Alto - A
1	0	0	Alto - B
0	1	1	Medio - A
0	1	0	Medio - B
0	0	1	Bajo - A
0	0	0	Bajo - B

$$U = \beta_0 + \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{21}x_{21} + \epsilon_i$$

Ejemplo: Inmuebles

FACTOR	No. de Niveles	Niveles
Habitaciones	2	2 y 3
Baños	2	1 y 2
Garaje	2	1 y 2

Diseño: Factorial Completo 2^3. Total perfiles: 8
Modo de Evaluación: Ordenamiento

Ejemplo: Inmuebles

FACTOR	No. de Niveles	Niveles
Habitaciones	2	2 y 3
Baños	2	1 y 2
Garaje	2	1 y 2

Diseño: Factorial Completo 2^3. Total perfiles: 8
Modo de Evaluación: Ordenamiento

No.	HABITACIÓN	BAÑO	GARAJE
1	2	1	1
2	2	1	2
3	2	2	1
4	2	2	2
5	3	1	1
6	3	1	2
7	3	2	1
8	3	2	2

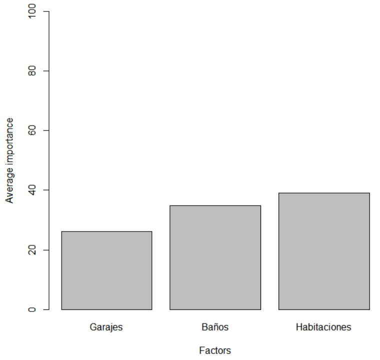
Implementación en R

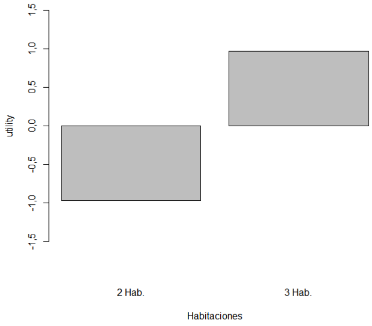
```
Residuals:
Min       1Q  Median      3Q     Max
-5,3409 -1,4242  0,3409  1,4242  5,3409

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)             4,50000    0,07958  56,548   <2e-16 ***
factor(x$Garajes)1      0,11742    0,07958   1,476    0,141
factor(x$Baños)1       -0,99242    0,07958 -12,471   <2e-16 ***
factor(x$Habitaciones)1 -0,96591    0,07958 -12,138   <2e-16 ***
---
Signif. codes:  0 ?***? 0,001 ?**? 0,01 ?*? 0,05 ?.? 0,1 ? ? 1

Residual standard error: 1,829 on 524 degrees of freedom
Multiple R-squared:  0,3679,Adjusted R-squared:  0,3643
F-statistic: 101,7 on 3 and 524 DF,  p-value: < 2,2e-16
```

Implementación en R

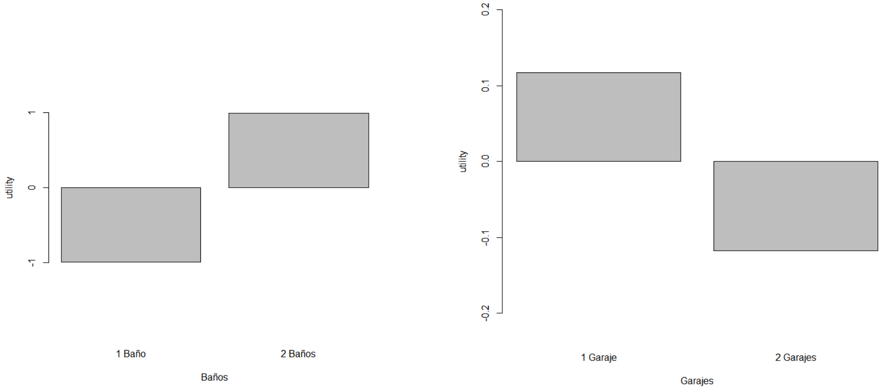
```
[1] "Part worths (utilities) of levels
(model parameters for whole sample):"
levnms    utls
1 intercept     4,5
2  1 Garaje  0,1174
3 2 Garajes -0,1174
4     1 Baño -0,9924
5   2 Baños  0,9924
6    2 Hab. -0,9659
7    3 Hab.  0,9659
[1] "Average importance of factors (attributes):"
[1] 26,07 34,91 39,03
[1] Sum of average importance:  100,01
```

Implementación en R - Gráficas





Implementación en R - Gráficas



DISEÑO PARA MEDIDAS REPETIDAS

Un diseño en el cual varios individuos son medidos o evaluados varias veces en el tiempo en varios tratamientos se denomina *estudio longitudinal* o *estudio panel*. Este tipo de análisis permite establecer:

- Evolución o cambio de la variable respuesta en el tiempo
- Interacción de los factores con el tiempo
- Efecto de los tratamientos

En el caso de un factor con s niveles, t medidas en el tiempo, n_h individuos en el grupo h, se tiene el modelo:

$$\boxed{y_{hij} = \mu + \alpha_h + \beta_j + (\alpha\beta)_{hj} + \pi_{i(h)} + \epsilon_{hij}} \tag{10}$$

donde $i = 1, \ldots, n_h$, $j = 1, \ldots, t$ $h = 1, \ldots, s$.

DISEÑO PARA MEDIDAS REPETIDAS

Con ello las hipótesis estadísticas de interés son:

Grupo (Tratamientos)

$$\boxed{H_0 : \alpha_h = \alpha_k \qquad vs. \qquad H_a : \alpha_h \neq \alpha_k} \tag{11}$$

Tiempo

$$\boxed{H_0 : \beta_j = \beta_l \qquad vs. \qquad H_a : \beta_j \neq \beta_l} \tag{12}$$

Interacción Grupo $\times$ Tiempo

$$\boxed{H_0 : \mu_{h \cdot j} = \mu_{k \cdot l} \qquad vs. \qquad H_a : \mu_{h \cdot j} \neq \mu_{k \cdot l}} \tag{13}$$

donde $\mu_{h \cdot j} = \mu + \alpha_h + \beta_j + (\alpha\beta)_{hj}$ con $h \neq k$ y $j \neq l$.

Supuestos del Modelo

Este tipo de análisis de varianza considera varios supuestos como son:

- Los errores se distribuyen de acuerdo a una normal

$$\epsilon_{hij} \sim N(0, \sigma_{\epsilon}^2)$$

- Los individuos se consideran como un factor aleatorio cuya variabilidad se reúne en el término $\pi_{i(h)}$ el cual se asume cumple

$$\pi_{i(h)} \sim N(0, \sigma_{\pi}^2)$$

- Los errores y el efecto de los individuos no están correlacionados

$$cov(\pi_{i(h)}, \epsilon_{hij}) = 0$$

- La homogeneidad de varianzas en los tiempos de medida este modelo se denomina Esfericidad.

Estructura de Datos

Típicamente los datos para un ANOVA de medidas repetidas se disponen de la siguiente manera:

Sujeto	Tratamiento	Y_1	$\ldots$	Y_t
1	1	y_{111}	$\cdots$	y_{11t}
2	1	y_{121}	$\cdots$	y_{12t}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n_1	1	y_{1n_11}	$\cdots$	y_{1n_1t}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
1	s	y_{s11}	$\cdots$	y_{s1t}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n_s	s	y_{sn_s1}	$\cdots$	y_{sn_st}

Tabla ANOVA

El análisis de varianza para este diseño de un factor con medidas tiene la siguiente estructura:

Fuente	**gl**	**SC**	**CM**	$\mathbf{F_c}$
$Tratamientos$	$s-1$	SCG	CMG	$\frac{CMG}{CME}$
$Individios(Trat.)$	$n-s$	SCI_G	CMI_G	
$Tiempo$	$t-1$	SCT	CMT	$\frac{CMT}{CME}$
$Trat. \times Tiempo$	$(s-1)(t-1)$	$SCGT$	$CMGT$	$\frac{CMGT}{CME}$
$Error$	$(n-s)(t-1)$	SCE	CME	

Las pruebas de comparación múltiple aplican igualmente para contrastar los tratamientos.

Sección 4

ANÁLISIS MULTIVARIADO DE DATOS EN SENSORIAL

Naturaleza multivariada en Sensorial

La descripción estadística simultanea de cualquier situación mediante más de una variable corresponde en general a un análisis multivariado.
En análisis sensorial esto se presenta con mucha frecuencia pues un producto se suele describir con múltiples atributos, los cuales estón típicamente correlacionados entre si.

Algunas técnicas multivariadas son:

- Análisis de componentes principales
- Análisis de correspondencias simples y múltiples
- Análisis cluster
- Análisis discriminante
- Escalamiento Multidimensional
- Regresión múltiple
- ...

Elementos Básicos de Análisis Multivariado

- Vector de medias `M=sapply(DT,mean)`

$$\mu = (\mu_1, \mu_2, \mu_3)$$

- Matriz de varianzas-covarianzas `cv=cov(DT)`

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}$$

- Matriz de correlaciones `cr=cor(DT)`

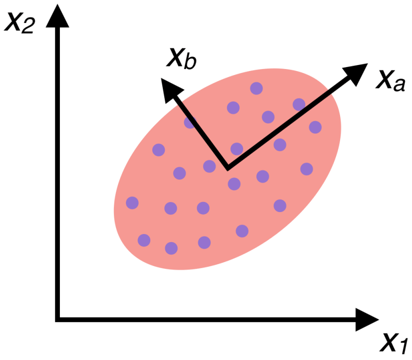
$$\rho = \begin{pmatrix} \rho_1 & \rho_{12} & \rho_{13} \\ \rho_{12} & \rho_2 & \rho_{23} \\ \rho_{13} & \rho_{23} & \rho_3 \end{pmatrix}$$

Subsección 1

Análisis de Componentes Principales

Análisis de Componentes Principales ACP

El ACP es una técnica de reducción de dimensión que permite crear nuevas variables sintéticas que logren reunir la mayor cantidad de la información original (*Varianza*) a fin de lograr representaciones gráficas y análisis estadísticos más sencillos.



El conjunto de variables originales tiene la caracterástica de ser correlacionadas, en tanto que las componentes principales son independientes.

Naturaleza de las variables

El Análisis de Componentes Principales (ACP) es en esencia una técnica matemática que en principio no hace ningún tipo se supuesto acerca de las variables, las cuales en general son de tipo cuantitativo, típicamente en escalas continuas.
De acuerdo al interés de las variables en el ACP estas pueden ser:

- Activas: Son las variables a partir de las cuales se construye todo al análisis y cuya dimensión se desea reducir.

- Suplementarias:Son variables cuantitativas o cualitativas que no aportan información en la construcción de las componentes pero sobre las cuales si hay interés de visualizar en los planos factoriales a fin de establecer asociaciones.

Aunque menos frecuente, algunos individuos también puede incluirse como Suplementarios con una finalidad similar al caso de las variables.

Construcción de las Componentes Principales

Las direcciones en el espacio en las cuales se presenta las mayor variación de los datos se denominan *Componentes Principales* y están dados por una combinación de todas las variables originales así:

$$
\begin{aligned}
F_1 &= \alpha_{11}X_1 + \alpha_{12}X_2 + \ldots + \alpha_{1p}X_p \\
F_2 &= \alpha_{21}X_1 + \alpha_{22}X_2 + \ldots + \alpha_{2p}X_p \\
&\vdots \\
F_k &= \alpha_{k1}X_1 + \alpha_{k2}X_2 + \ldots + \alpha_{kp}X_p
\end{aligned}
$$

Donde los coeficientes α_{ij} describen la contribución porcentual de la variable j en el factor i y se denomina *Cargas factoriales* o simplemente *Contribuciones*.

Elementos importantes del ACP

Tipicamente, las salidas de un ACP comprenden:

- Valores propios:Brindan la información de cuanta varianza esta explicada en cada componente.
- Contribuciones: Describen la contribución porcentual de la variable j en el factor i.
- Matriz de correlaciones: Cuantifica el grado de asociación entre las variables analizadas.
- Coordenadas: Tanto para variables como para individuos se puede obtener sus coordenadas en cada componente, de tal manera que se pueden representar gráficamente.
- Gráficas factoriales: A partir de las dos primeras componentes se pueden hacer representaciones bidimensionales.

Implementación en R

Un paquete util para la realización de Análisis Multivariados es factominer, en el cual se tiene una base de datos de vinos llamada wine.

```
library(FactoMineR)
data(wine)
str(wine)
'data.frame': 21 obs. of  31 variables:
$ Label: Factor w/ 3 levels "Saumur","Bourgueuil",..: 1 1 ...
$ Soil : Factor w/ 4 levels "Reference","Env1",..: 2 2  ...
$ Odor.Intensity.before.shaking: num  3.07 2.96 2.86 ...
$ Aroma.quality.before.shaking : num  3 2.82 2.93 2.59 ...
$ Fruity.before.shaking        : num  2.71 2.38 2.56 2.42 ...
.
.
$ Harmony                      : num  3.14 2.96 3.14 2.04 ...
$ Overall.quality              : num  3.39 3.21 3.54 2.46 ...
$ Typical                      : num  3.25 3.04 3.18 2.25 ...
```

Implementación en R

De las 29 variables numéricas vamos a extraer las variables correspondientes al gusto.

```
DT_Sabor=wine[,c(1,2,21,22,23,24,25,26,27,28,29)]
names(DT_Sabor)=c('Marca','Suelo','Int.Inicial','Acidez',
'Astringencia','Alcohol','Balance','Suavidad','Amargor',
'Intensidad','Armonia')
str(DT_Sabor)
```

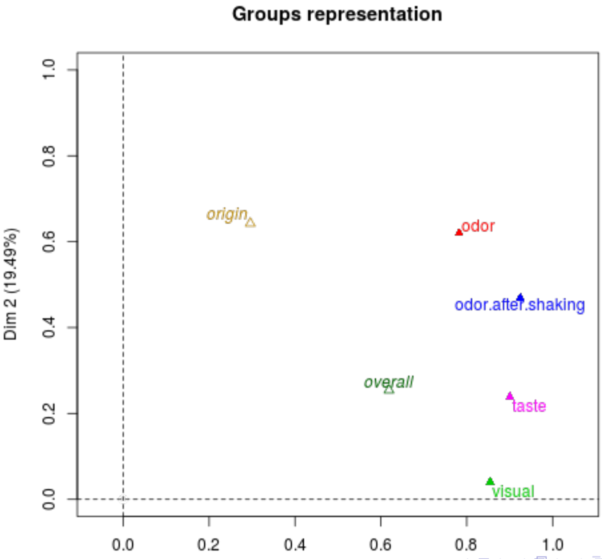
```
data.frame': 21 obs. of  11 variables:
$ Marca       : Factor w/ 3 levels "Saumur","Bourgueuil",..
$ Suelo       : Factor w/ 4 levels "Reference","Env1",..
$ Int.Inicial : num  2.96 3.04 3.22 2.7 3.46 ...
$ Acidez      : num  2.11 2.11 2.18 3.18 2.57 ...
$ Astringencia: num  2.43 2.18 2.25 2.19 2.54 ...
$ Alcohol     : num  2.5 2.65 2.64 2.5 2.79 ...
$ Balance     : num  3.25 2.93 3.32 2.33 3.46 ...
```

Implementación en R

```
sw_pca=PCA(DT_Sabor,ncp=5,quali.sup=c(1,2),graph=T)

Eigenvalues               Dim.1   Dim.2   Dim.3   Dim.4   Dim.5
Variance                  5.642   1.791   0.674   0.351   0.313
% of var.                62.689  19.900   7.490   3.895   3.475
Cumulative % of var.     62.689  82.589  90.079  93.973  97.448

$contrib         Dim.1      Dim.2      Dim.3
Int.Inicial  15.565865  0.3861023  2.0953619
Acidez        1.148341 26.6354370 67.3482364
Astringencia 10.491302 10.3152025  4.0961623
Alcohol      10.562720  8.1309186  0.9823138
Balance      12.575911 10.0642661  4.0839002
Suavidad     14.353862  8.1594278  1.6300898
Amargor       2.517722 32.4497430 17.7635511
Intensidad   16.547731  0.7569382  1.6848031
Armonia      16.236547  3.1019645  0.3155813
```

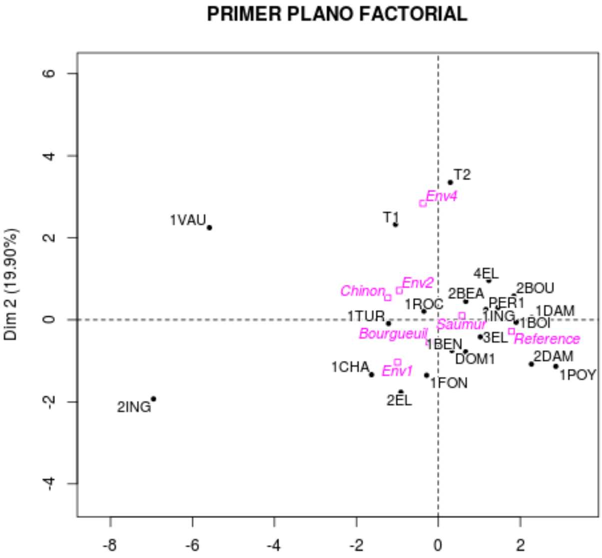
Representación Gráfica

Variables



Representación Gráfica

Individuos

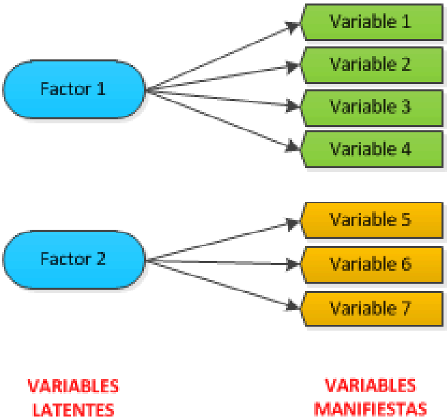


Subsección 2

Análisis Factorial Multiple

Análisis de Factorial Múltiple

El AFM es una técnica de reducción de dimensión que mediante la construcción de nuevas variables denominadas factores permite obtener identificar la estructura latente de las variables.



Análisis de Factorial Múltiple

Objetivo del AFM: Describir la relación de covariación entre múltiples variables en términos de pocas variables no observables llamadas *Factores*.
El AFM puede ser de dos tipos dependiendo la finalidad con la cual se realice:

- Exploratorio: En este caso no se tiene información acerca de la estructura del problema y se quiere precisamente establecer que variables latentes (factores) están presentes.

- Confirmatorio: Bajo una estructura o agrupamiento de variables manifiestas en factores, el AFM provee una confirmación o negación de la misma.

Modelo en el AFM

A diferencia del ACP, el AFM considera un modelo estadístico de describe las variable de estudio, este es:

$$\boxed{\mathbf{X} = \Lambda\mathbf{F} + \epsilon} \tag{14}$$

donde:

- $\mathbf{X}$ es la matriz de variables observadas.
- Λ es la matriz de cargas factoriales.
- $\mathbf{F}$ es el vector de factores.
- ϵ es el vector de errores.

$$\begin{aligned} X_1 &= \lambda_{11}f_1 + \lambda_{12}f_2 + \ldots + \lambda_{p1}f_k + \epsilon_1 \\ &\vdots \\ X_p &= \lambda_{p1}f_1 + \lambda_{p2}f_2 + \ldots + \lambda_{pk}f_k + \epsilon_p \end{aligned}$$

Implementación en R

```
w_mfa = MFA(wine, group=c(2,5,3,10,9,2),
type=c("n",rep("s",5)),ncp=5,
name.group=c("origin","odor","visual",
"odor.after.shaking","taste","overall"),
num.group.sup=c(1,6))
```

DIPLOMADO
APLICACIÓN DE LA CIENCIA SENSORIAL EN LA INDUSTRIA

Fernando Alonso Velez
Msc. en Estadística

fvelez78@yahoo.es

Septiembre 2017