



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA

Departamento de Estadística e  
Investigación Operativa Aplicadas  
y Calidad

[www.upv.es](http://www.upv.es)

[bigdata.inf.upv.es](http://bigdata.inf.upv.es)

# Herramientas estadísticas para Big Data

Introducción a la Inferencia Estadística,  
Muestreo y Preproceso de datos

Máster **Big Data** Analytics

Valencia, Octubre 2017

Mónica Clemente | Ana Debón | Elena Vázquez

# Presentación

Elena Vázquez

@ [evazquez@eio.upv.es](mailto:evazquez@eio.upv.es)

Associated Professor

Department of Applied Statistics and  
Operational Research and Quality

Researcher of Centre for Quality and  
Change Management (CQ)

Secretary of the Department of Statistics

# Software

- R

<http://www.r-project.org/>



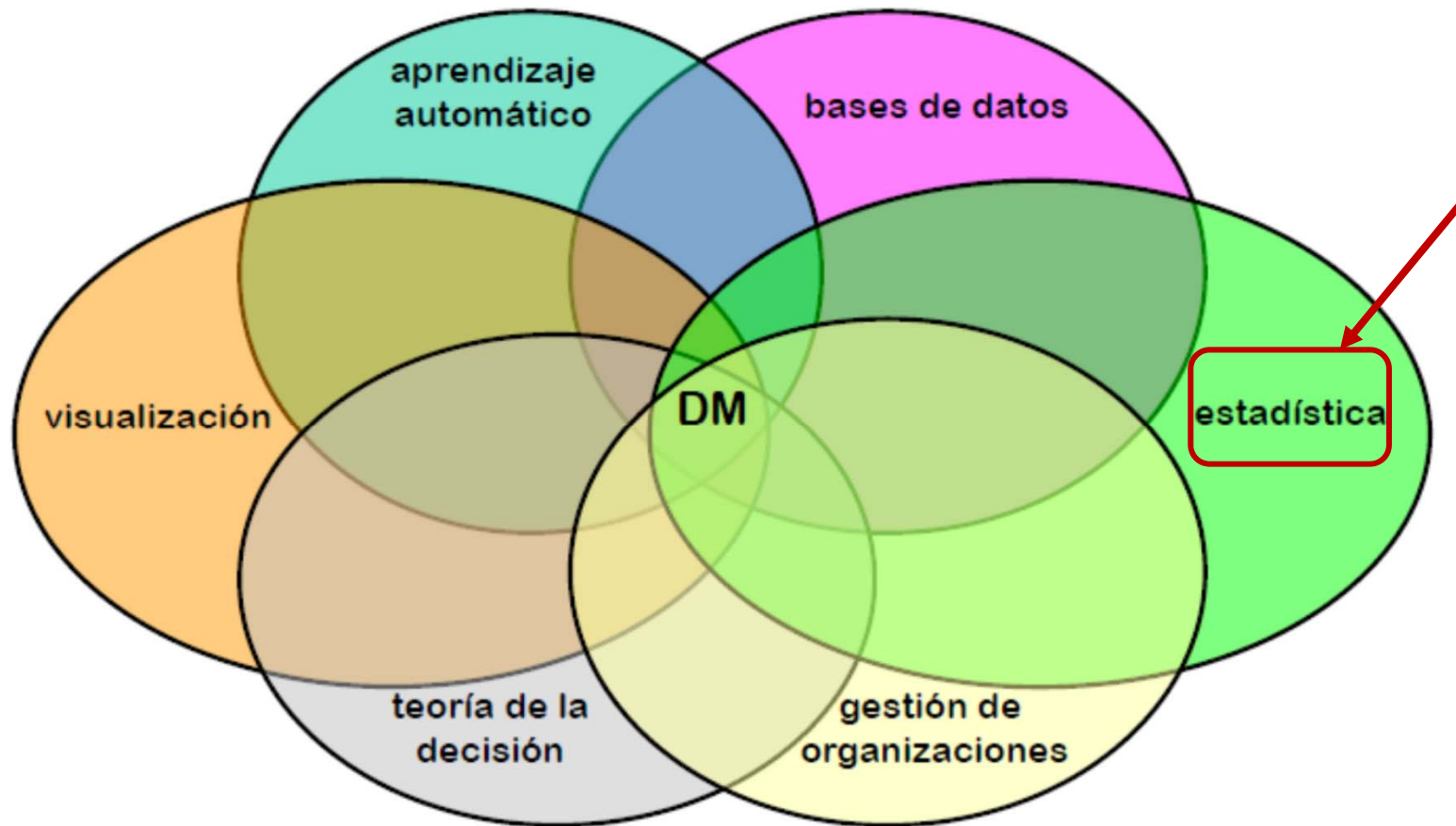
- R Studio

<http://www.rstudio.com/>



Free & Open-Source IDE for R

# ¿Por qué necesitamos la Estadística?



Probar teorías  
cuantitativamente

# ¿Por qué necesitamos la Estadística?

The screenshot shows the Statistics Views website interface. At the top, there's a logo with three colored bars (blue, yellow, red) and the text 'STATISTICS VIEWS'. Below it is a navigation bar with links: HOME, BIOSTATISTICS, BUSINESS & ECONOMICS, ENGINEERING, ENVIRONMENTAL, and ME. The main content area displays the article title 'Statistical Truisms in the Age of Big Data' with a breadcrumb trail: Home → Search Results → Statistical Truisms in the Age of Big Data. The article is by Kirk Borne, dated 19 Jun 2013. The text discusses whether Big Data makes statistics obsolete and examines basic tenets of elementary statistics. A central diagram shows 'Statistics' in a blue circle, surrounded by four colored triangles: green for '(Big) Data Science', orange for 'Decision Sciences', yellow for 'Operations Research', and red for 'Experimental Science'. To the right, there are sections for 'RELATED TOPIC' (Methods - Gen) and 'RELATED PUBLI' (BOOKS, with a link to 'Common Error Statistics (a how to Avoid them), 4th Edition'). Below that, 'RELATED CONT' includes links like 'Embrace Big Data: peril FEATURE' and 'Don't drown in the Big data and city us? JOURNAL AR'.

<http://www.statisticsviews.com/details/feature/4911381/Statistical-Truisms-in-the-Age-of-Big-Data.html>

# Vamos a hablar de ...

- No es una materia sobre R, R es sólo la herramienta, aunque se presuponen conocimientos básicos (**Taller de R**)
- Muchos conceptos se volverán a ver o a ampliar en asignaturas posteriores, como **Business Intelligence**
- Las cuestiones relacionadas con la visualización de datos y resultados se ampliarán en la materia **Visualización de datos**

# Contenidos Viernes 6 y Sábado 7

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

Bibliografía básica

Bibliografía adicional

Enlaces de interés





# 1 Conceptos básicos

1. Introducción
2. Variables, observaciones y casos
3. Tipos de variables
4. Conceptos básicos
5. Análisis descriptivo







## 2 Probabilidad y variables aleatorias

1. Sucesos
2. Probabilidad: concepto y propiedades
3. Teorema de Bayes





## 3 Probabilidad y variables aleatorias

1. Distribuciones de probabilidad
2. Esperanza Matemática
3. La Distribución Normal
4. La Distribución Binomial
5. Teorema Central del Límite
6. Aproximaciones normales





## 4 Inferencia en muestras grandes

1. Introducción
2. Modelos estadísticos
3. Muestreo y distribuciones en el muestreo
4. Estimación puntual
5. Intervalos de confianza
6. Test o contrastes de hipótesis





## 5 Técnicas de muestreo

1. Población y muestra
2. ¿Muestreo en Big Data?
3. Técnicas de muestreo
4. Tamaño de la muestra
5. Error muestral y potencia de un contraste





## 6 Preprocesamiento de datos

1. Limpieza (Data cleaning)
2. Integración
3. Transformación
4. Reducción

Glosario



# Bibliografía básica

**Charte, F., 2014. *Análisis exploratorio y visualización de datos con R*.** Attribution-NonCommercial 3.0 Unported (CC BY-NC 3.0) :  
<http://www.fcharte.com/libros/ExploraVisualizaConR-Fcharte.pdf>

De Jonge, E. & van der Loo, M., 2013. *An introduction to data cleaning with R*, Statistics Netherlands. Available at: [http://cran.r-project.org/doc/contrib/de\\_Jonge+van\\_der\\_Loo-Introduction\\_to\\_data\\_cleaning\\_with\\_R.pdf](http://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf).

**Field, A., Miles, J, & Field, Z., 2012. *Discovering statistics using R*,** SAGE Publications.

Gómez, A. A., 2008. *Estadística básica con R y R-Commander*. Servicio Publicaciones UCA

González, A. & González, S., 2000. *Introducción a R*. R Development Core Team. Available at: <https://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf>

**Sáez Castillo, A.J., 2010. *Métodos Estadísticos con R y R Commander*,** Jaén: Universidad de Jaén. Available at: <http://cran.r-project.org/doc/contrib/Saez-Castillo-RRCmdrv21.pdf>.

Sorribas, A., 2013. *Introducción al análisis gráfico con ggplot2*. Available at: <http://web.udl.es/Biomath/Bioestadistica/R/Grafics/ggplot2.pdf>

# Bibliografía adicional

Diez, D.M., Barr, C.D. & Cetinkaya, M., 2010. *OpenIntro : Statistics Preliminary Edition*,

**Hair, J.F. et al., 2005. Multivariate Data Analysis** 6th ed., Prentice Hall.

Hernández, J., Ramírez, M. J. and Ferri, C. , 2004. *Introducción a la Minería de Datos*, volume 17. Prentice.

Sanchez De Rivera, D.P., 1993. *Estadística modelos y metodos 1. fundamentos*, Alianza.

Scheaffer, R., Mendenhall, W. & Ott, L., 2007. *Elementos de muestreo*, Editorial Paraninfo.

Peña, D., 2002. *Análisis de datos multivariantes*, McGraw-Hill

# Enlaces de interés

- **Quick-R** <http://www.statmethods.net/>
- **R-bloggers** <https://www.r-bloggers.com/>
- **Stackoverflow** <http://stackoverflow.com/>
- **Stackoverflow** en español <http://es.stackoverflow.com/>
- **R for Data Science** <http://r4ds.had.co.nz/data-visualisation.html#the-layered-grammar-of-graphics>
- <https://www.coursera.org/course/statistics>
- [http://spark.rstudio.com/minebocek/dist\\_calc/](http://spark.rstudio.com/minebocek/dist_calc/)





# Paquetes de R *library*

- e1070
- MASS
- pwr
- (Hmisc)
- car
- psych
- corpcor
- GPArotation
- (nFactors)
- (FactoMineR)
- XML
- rvest
- Rcurl
- ggplot2



# Gracias por vuestra atención



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA

[www.upv.es](http://www.upv.es)