UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad

www.upv.es bigdata.inf.upv.es

Herramientas estadísticas para Big Data

Introducción a la Inferencia Estadística, Muestreo y Preproceso de datos

Máster Big Data Analytics

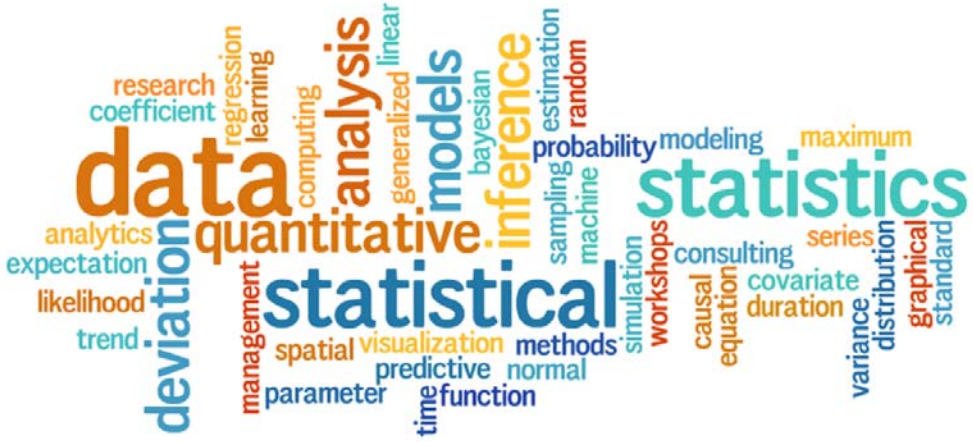
Valencia, Octubre 2017

Elena Vázquez

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



5 Técnicas de muestreo

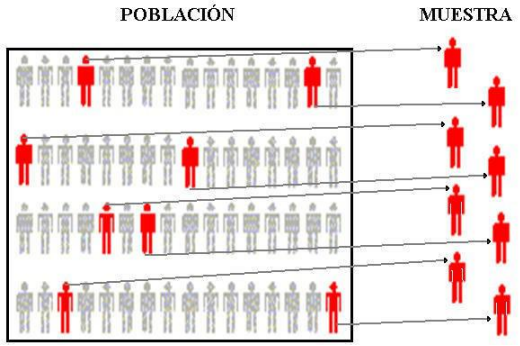
1. Población y muestra
2. ¿Muestreo en Big Data?
3. Técnicas de muestreo
4. Tamaño de la muestra
5. Error muestral
6. Potencia de un contraste
7. Tamaño del efecto

Glosario

Población y muestra

Población

- Conjunto de todos los individuos o entes que constituyen el objeto de un determinado estudio y sobre los que se desea obtener ciertas conclusiones
- Puede ser Finita o Infinita

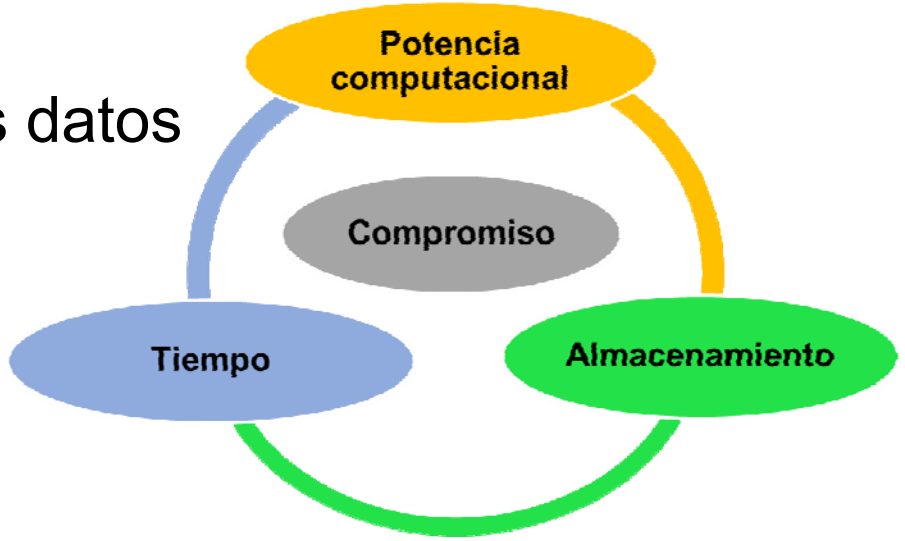


Muestra

- Normalmente no se trabaja con toda la **población**, se analiza sólo una parte, la **muestra**…
  - Porque es imposible si es infinita
  - Porque se destruye a los individuos en el muestreo
  - Por razones técnicas
  - Por razones económicas, …

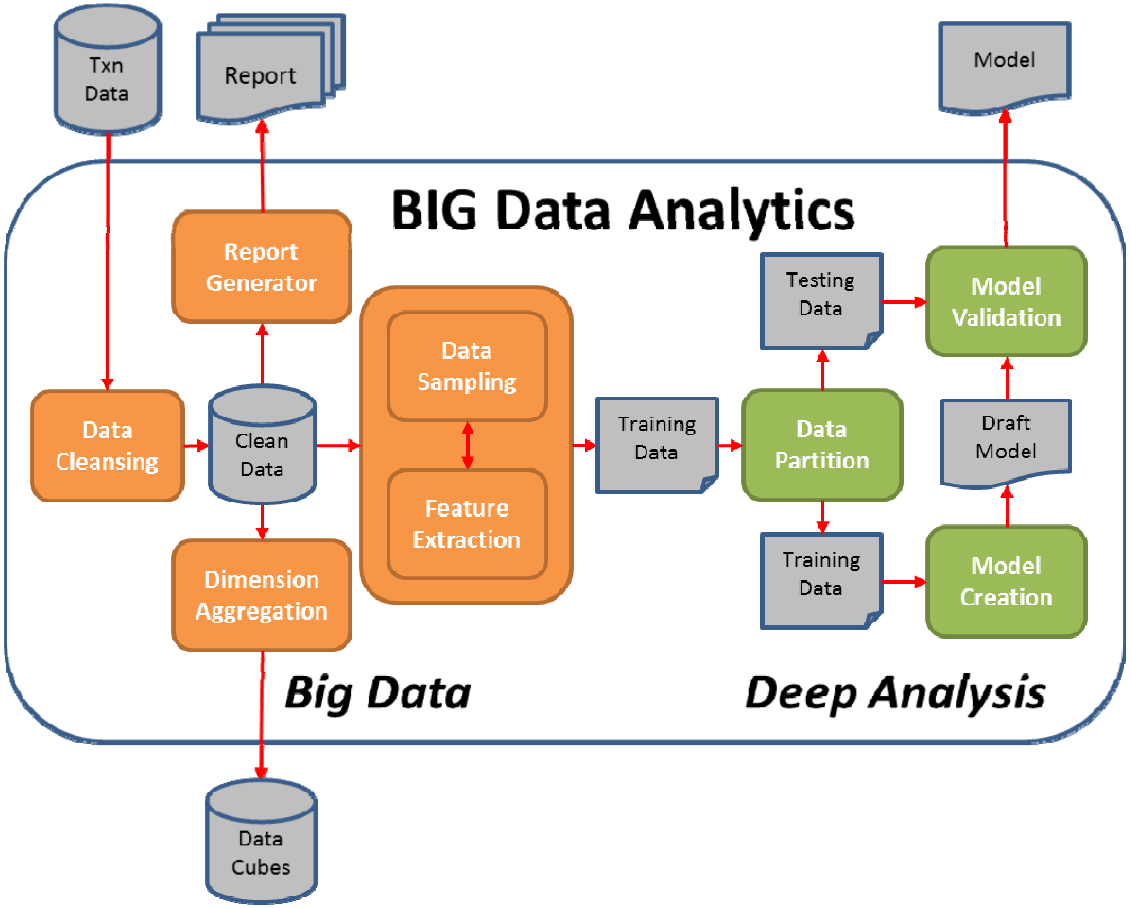
Muestreo en Big Data

- **Aunque se disponga de toda la población** necesitamos el muestreo para:
  - Reducir el tamaño de los datos
  - Equilibrar las clases
  - Facilitar la visualización
  - Realizar un primer estudio exploratorio
  - etc



- En ocasiones los **datos disponibles** ya son la muestra en realidad:
  - Volver a muestrear para aumentar la aleatoriedad

Muestreo en Big Data



http://horicky.blogspot.com.es/2012/08/big-data-analytics.html

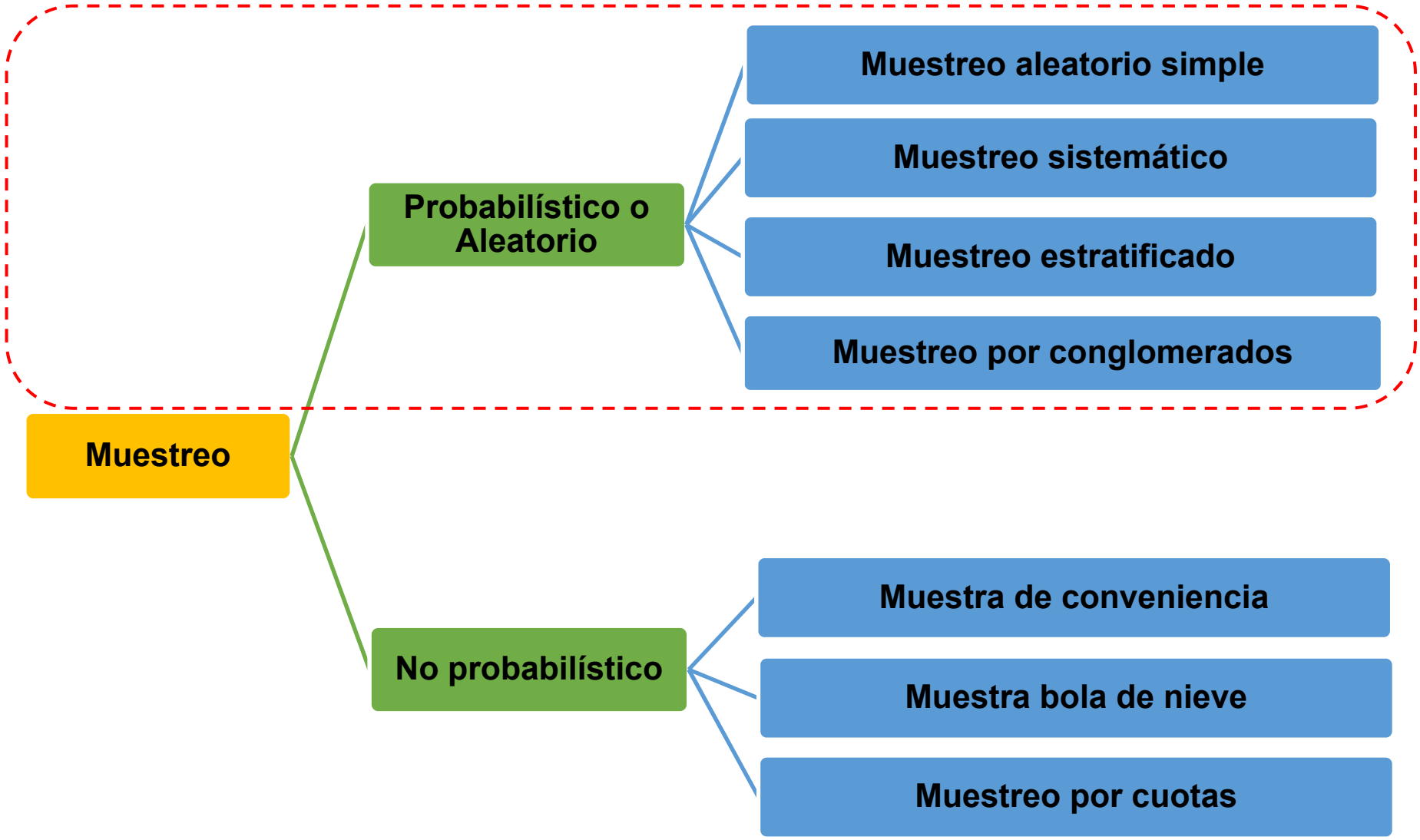
Muestreo en Big Data

- El análisis de datos en Big Data se realiza generalmente en modo *batch* (por ejemplo, una vez al día), por lo que el proceso de datos y el *deep* análisis se llevan a cabo en distintas fases de este proceso por lotes.
- La gran parte de procesamiento de datos (de color naranja) se realiza normalmente utilizando la tecnología Hadoop / PIG / Hive
- La parte del *deep analysis* (de color verde) se hace generalmente en R, SPSS, SAS utilizando una cantidad mucho menor de datos cuidadosamente muestreados que se ajuste a la capacidad de una sola máquina (generalmente menos de dos centenares de miles de registros de datos). Esta parte, por lo general incluye el estudio estadístico de los datos, la visualización de datos, la preparación de datos y los modelo de aprendizaje.

¿Cómo debe ser la muestra?

- La muestra debe ser **representativa de toda la población** y con un **error de muestreo conocido y aceptable**, de modo que permita extraer conclusiones razonablemente válidas sobre la toda la población.

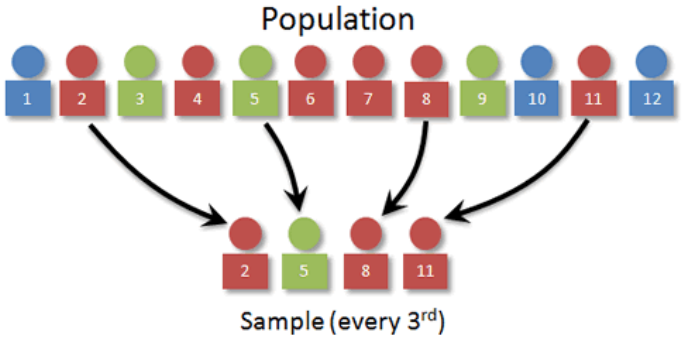
Técnicas de muestreo



En este módulo nos centraremos en el muestreo probabilístico fundamentalmente.

Muestreo sistemático

- Si se tiene que seleccionar una muestra de *n* elementos de una población de tamaño *N*.
- El muestreo sistemático de 1 en *k*, donde *k* = *N*/*n*, se realiza de la siguiente manera:
  - 1) El primer elemento (i) es seleccionado aleatoriamente entre los primeros *k* elementos (arranque)
  - 2) Los próximos elementos son seleccionados cada *k*-elementos (i+k, i+2K, ..i+(n-1)k)
- Solución fácil y práctica
- Puede crear sesgos pues el orden inicial de los datos depende de muchos factores.

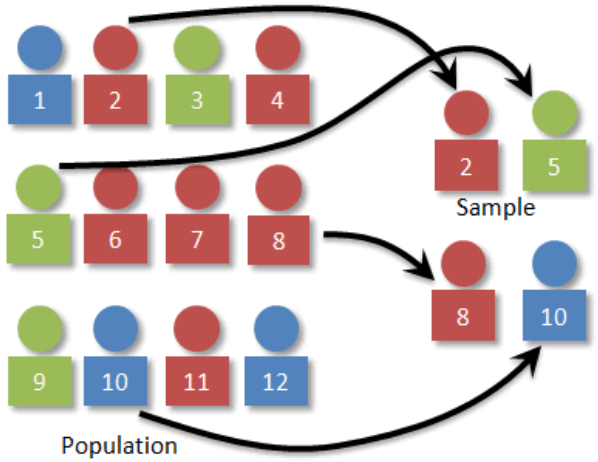


Muestreo sistemático

- Por ejemplo, tenemos una población formada por 100 elementos y queremos extraer una muestra de 25 elementos:
  - en primer lugar debemos establecer el intervalo de selección que será igual a 100/25 = 4
  - A continuación elegimos el elemento de arranque, tomando aleatoriamente un número entre el 1 y el 4 (suponemos que es el 2)
  - a partir de él obtenemos los restantes elementos de la muestra.
  - **Resultado: 2, 6, 10, 14,..., 98**

Muestreo aleatorio simple

- Todos los sujetos tienen la misma probabilidad de ser escogidos (ej. sorteo, lotería)
- Ideal si no hay *clusters o estratos*
- Suele ser una parte de otros tipos de muestreo.
- Puede ser:
  - Con reemplazamiento
  - Sin reemplazamiento
  - Si N→∞ (Big Data) la diferenci inapreciable



Muestreo estratificado

- Si se tiene que seleccionar una muestra de n elementos de una población de tamaño N, la cual está dividida en k **estratos**, mutuamente excluyentes de tamaños N_1, N_2, …, N_k, tal que:

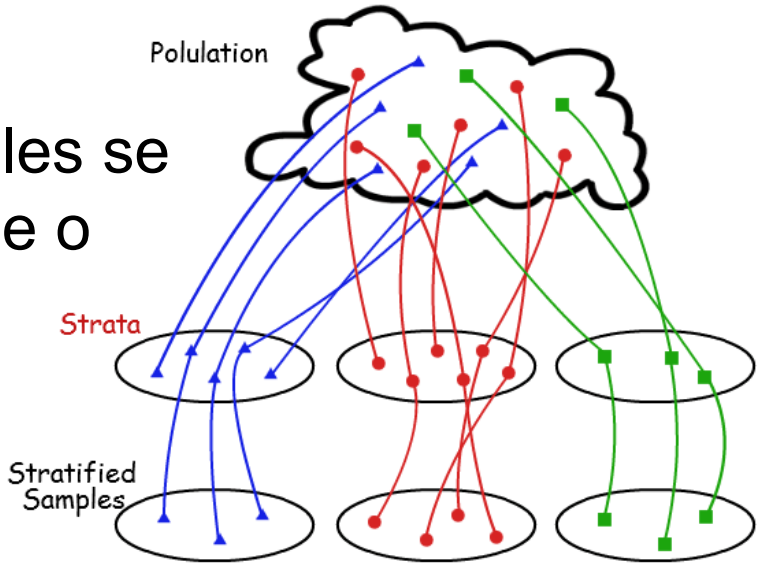
$$N_1 + N_2 + \ldots + N_k = N$$

- El muestreo estratificado consiste en seleccionar una muestra desde cada estrato de tamaños n_1, n_2, …, n_k, tal que:

$$n_1 + n_2 + \ldots + n_k = n$$

- La selección de los sujetos individuales se efectúa por muestreo aleatorio simple o muestreo aleatorio sistemático.

- N_i y n_i son proporcionales



Muestreo estratificado

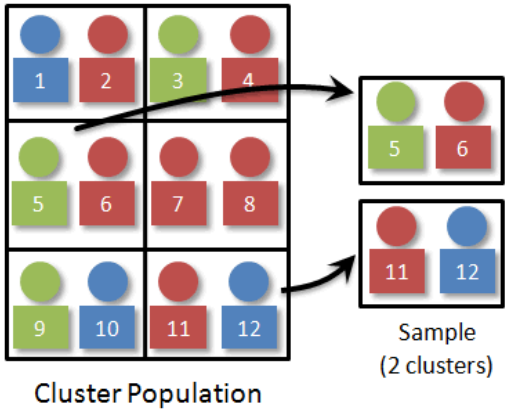
- Por ejemplo, en una fábrica que consta de 600 trabajadores queremos tomar una muestra de 20. Sabemos que hay 200 trabajadores en la sección A, 150 en la B, 150 en la C y 100 en la D.

- $\frac{20}{600} = \frac{x_1}{200}$ $\quad x_1 = 6.6 \approx 7$ trabajadores de A
- $\frac{20}{600} = \frac{x_2}{150}$ $\quad x_2 = 5$ $\quad$ 5 trabajadores de B
- $\frac{20}{600} = \frac{x_3}{150}$ $\quad x_3 = 5$ $\quad$ 5 trabajadores de C
- $\frac{20}{600} = \frac{x_4}{100}$ $\quad x_4 = 3.3 \approx 3$ trabajadores de D

Muestreo por conglomerados

- La población de *N* elementos, está dividida en conglomerados C_1, C_2,… , C_M los cuales forman las **unidades primarias** de muestreo, cada uno de estos conglomerados está constituido por elementos de la población, **unidades finales**:
  - *N* = número de elementos en la población
  - *M* = número de conglomerados en la población
  - *m* = número de conglomerados en la muestra

- El muestreo por conglomerados puede ser realizado en una etapa o en dos etapas.



Muestreo por conglomerados

- **Muestreo por conglomerados de una etapa**
  - Consiste en seleccionar aleatoriamente un cierto número de conglomerados (*m*), de forma que los individuos dentro de cada conglomerado pasan a ser las unidades finales.

- **Muestreo por conglomerados de dos etapas**
  - Consiste en seleccionar aleatoriamente un cierto número de conglomerados (*m*), y dentro de cada conglomerado se realiza un muestreo de las unidades finales.

- La selección de los individuos dentro de cada conglomerado se puede efectuar por muestreo aleatorio simple o muestreo aleatorio sistemático o, incluso, estratificado.

Muestreo por conglomerados

Principales diferencias con el muestreo estratificado:

- Las **etapas**
- Las **unidades de muestreo** (conglomerados o *clusters*) son **agrupaciones naturales**:
  - unidades geográficas o físicas:
    - estados
    - delegaciones
    - distritos
    - etc
  - en base a una organización:
    - escuelas,
    - grado escolar
    - etc

Muestreo por conglomerados

En el muestreo por conglomerados en una y dos etapas se pueden presentar cualquiera de los dos siguientes casos:

Caso 1: Conglomerados de igual tamaño

Cada conglomerado C_1, C_2,… , C_M de la población tiene igual número de unidades primarias.

Sea *u* el número de unidades en cada conglomerado, entonces se cumple que $M = N / u$ y por lo tanto $N = M \times u$

Caso 2: Conglomerados de diferente tamaño

Cada conglomerado C_1, C_2,… , C_M de la población tiene diferente número de unidades primarias. Siendo u_i el número de unidades en el conglomerado C_i para $i = 1, 2, \ldots, M$.

Obtención de una muestra aleatoria simple

```
sample(x, size, replace = FALSE, prob = NULL)


Muestrear índices o código de individuo (números de fila)
# Ejemplo 1 con remplazamiento
x <- 1:100
set.seed(111)
sample(x, size = 10, replace = T)
 [1] 60 72 37 50 98 40  2 97 95  9

# Ejemplo 2 con remplazamiento
set.seed(222)
sample(100, size = 10, replace = T)
 [1] 94  7 50  1 92 96 36 42 58 15
```

Obtención de una muestra aleatoria simple

Muestrear el propio valor de la variable

```
# Ejemplo 3
# Muestra de 10 elementos de una población normal (con
media 4 y desviación típica 1) de tamaño 500
x<-rnorm(500, 4, 1)
> muestra<-sample(x, 10, replace=T)
> muestra
 [1] 5.935148 4.104689 3.515435 3.909264 3.904676
2.726362 4.560953 5.770121 4.739089 3.061024

# Ejemplo 4
load("datos.RData")
sample(datos1$PROVINCIA, size = 4)
[1] Castellon Alicante  Alicante  Teruel
```

Elección de la técnica

- Las muestras obtenidas, y por tanto las conclusiones acerca de la población, pueden variar mucho dependiendo del tipo de muestreo.
- La elección debe fundamentarse en el objetivo o combinación de objetivos del análisis:
  - Maximizar la aleatorización
  - Minimizar costes (tiempo, espacio,…)
  - Entrenar modelos
  - Análisis exploratorio
  - etc
- En cualquier caso es indispensable el conocimiento de la población.
    - Ej. Para realizar el muestreo estratificado es necesario conocer bien cuales son los estratos o grupos de interés.

Elección de la técnica

- Se pueden realizar varios muestreos en cascada:
  - Realizar primero un m.a.s. para reducir la dimensionalidad y hacer un primer análisis descriptivo.
  - Realizar muestreo estratificado o conglomerados en función de la información obtenida en el primer muestreo
- En algunos casos el muestreo se utiliza para crear muestras dispares con el fin de generalizar los modelos.
  - Dentro de estas técnicas está el *bootstrapping* usado en la técnica de *bagging* **que se verá en la siguiente parte de este módulo**.

Muestreo no probabilístico

- **Muestra de conveniencia**
  - Es la muestra que está disponible. Útil para estudios descriptivos.
  - Ej. En el estudio de las reclamaciones sobre un producto, si sólo quedan registradas las realizadas a través de la web, son sólo éstas de las que dispongo y podré analizar.
  - Suele ser conveniente muestrear esta muestra (aleatoriedad)
- **Muestra bola de nieve**
  - Los sujetos ya encuestados ayudan a contactar con otros
  - Ej. Redes sociales
- **Muestreo por cuotas**
  - Igual que el muestreo estratificado, pero sin elección aleatoria
  - De cada cuota vamos obteniendo candidatos a formar parte de la muestra de forma no aleatoria y comprobando si el candidato es válidos para el estudio (es decir, si puede formar parte de una de mis cuotas o ya he excedido mi objetivo).

Tamaño de la muestra

¿cuántos datos necesito?

- En el caso de **validación de hipótesis** (contrastes) existen técnicas estadísticas para determinar el número de datos necesarios para tener un nivel de confianza establecido de antemano.
  - Basadas en la distribución del parámetro (Binomial, Normal, …)
  - Basadas en la potencia estadística y/o tamaño del efecto
- En el caso del **entrenamiento de modelos** la respuesta es más compleja. En general lo habitual es llevar a cabo un muestreo progresivo o incremental, en el que la muestra se va haciendo cada vez más grande (**se verá en otra parte de este módulo**.)

El tamaño de la muestra n depende de…

1. El nivel de confianza o riesgo de 1ª especie (α)

El α habitual es 1%, 5% o 10%.

2. La variabilidad estimada de los datos en la población: varianza estimada $\hat{\sigma}^2$

Si los datos no tuvieran variabilidad, nos bastaría conocer el valor de uno solo, pero si hay gran diversidad de datos, hará falta un mayor número de sujetos en la muestra.

3. El error muestral (E)

Es el margen de error que estamos dispuestos a aceptar cuando decimos, por ejemplo, que el 15% de los clientes en la población compra en la primera visita a la web, pero en realidad es un resultado que hemos obtenido a partir de la muestra.

Error muestral

$$IC_m^{nc\%} = \left[\overline{X} - z\alpha/_2 \frac{S}{\sqrt{N}}, \overline{X} + z\alpha/_2 \frac{S}{\sqrt{N}}\right]$$

Error muestral Error muestral

$$IC_P^{nc\%} = \left[\hat{p} - z\alpha/_2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z\alpha/_2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Error muestral Error muestral

Tamaño de la muestra y error muestral

Despejando de las expresiones vistas en Inferencia para la obtención de IC:

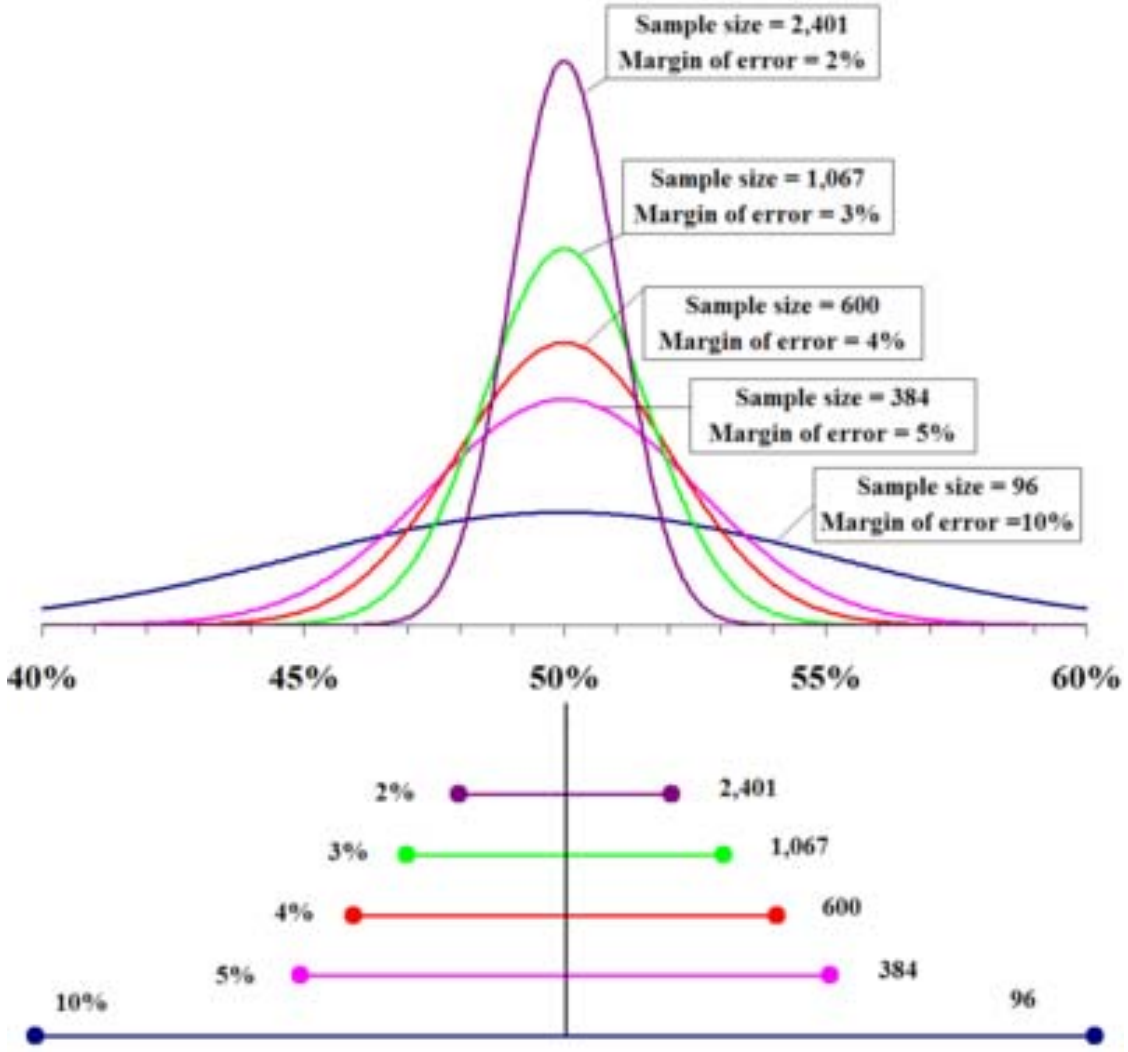
- **Dado un tamaño de la muestra n**, puede obtenerse el error muestral cometido o
- **Dado un error muestral**, puede calcularse el tamaño de la muestra necesario.
- Por ejemplo:
  - **e** y **n** en la estimación de una proporción (población infinita)

$$e = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

  - **e** y **n** en la estimación de una media

$$e = z_{\alpha/2}\frac{S}{\sqrt{N}} = z_{\alpha/2}SE$$

Tamaño de la muestra y error muestral



n para estimar una proporción

- **Población finita**. Conocemos el tamaño de la población N

$$n = \frac{N}{1 + \frac{e^2(n-1)}{z^2_{\propto/2}\, p(1-p)}}$$

- **Población infinita**. N>30.000

$$n = \frac{z^2_{\propto/2}\, p(1-p)}{e^2}$$

N = Total de la población (si se conoce)

$z^2_{\propto/2}$ = valor de z / P(Z ≥ |z|) =α

p = proporción esperada de éxitos

e = error muestral o precisión

Obtención de n para para estimar p

```
### Sea X la variable definida como "n° de clientes
que compran en la primera visita a la web" de una
determinada tienda on line.

Estimar n para la proporción (P) de clientes que
compran en la primera visita a la web de dicha tienda.


p=0,1

P(1-p)=1/4

E=5%

Alfa=5%
```

n para estimar la media

- **Población finita**. Conocemos el tamaño de la población N

$$n = \frac{N z^2_{\propto/2}\, S^2}{\mathrm{e}^2(N-1) + z^2_{\propto/2}\, S^2}$$

- **Población infinita**. N>30.000

$$n = \frac{z^2_{\propto/2}\, S^2}{e^2}$$

N = Total de la población (si se conoce)

$z^2_{\propto/2}$ = valor de z / P(Z ≥ |z|) =α

S^2= varianza o su estimación

e = error muestral o precisión

Obtención de n para estimar m, dado un e

```
### Estimación del gasto medio por cliente en una
tienda on line (X).


> ## Nivel de significación y precisión
> alfa<-0.05
> z<-qnorm(alfa/2, lower.tail=F)
> e<-0.05 # Error de 5%
>
> ## Tamaño de la muestra, despejando n de las
expresiones anteriores
> n<-ceiling(z^2*S2.gasto/e^2)
> n
[1] 5486
```

n para un contraste de proporciones

$$n = \frac{z_{\propto}\sqrt{2p(1-p} + z_{\beta}\sqrt{p_1(1-p_1) + p_2(1-p_2)}}{p_1 - p_2}$$

n debe obtenerse para cada una de las muestras

$z_{\propto}$ = valor de z / P(Z ≥ |z|) =α

z_{β} = valor de z / P(Z ≥ |z|) =1-β

p_1 = proporción en el grupo de control o tratamiento habitual

p_2 = proporción en el grupo de del nuevo tratamiento

p = media de p_1 y p_2

n para un contraste de medias

$$n = \frac{2\,(z_{\propto} + z_{\beta})^2\, S^2}{d^2}$$

n debe obtenerse para cada una de las muestras

$z_{\propto}$ = valor de z / P(Z ≥ |z|) =α

z_{β} = valor de z / P(Z ≥ |z|) =1-β

S^2= varianza del grupo de control o su estimación

d = valor mínimo de la diferencia que se desea detectar

Tamaño de la muestra para un contraste

En la comparación de medias o proporciones es necesario conocer:

- La magnitud de la diferencia a detectar, de forma que ésta tenga interés práctico relevante.
- Tener una idea aproximada de los parámetros de la variable que se estudia (referencias, estudios previos).
- Nivel de significación (α)
- Potencia estadística (1 - β)
- Determinar si la hipótesis va a ser unilateral o bilateral.
  - La hipótesis bilateral es una hipótesis más conservadora y disminuye el riesgo de cometer un error de tipo I

Importancia del tamaño de la muestra

1) Obtener 2 muestras aleatorias de tamaño 10 de una distribución normal con sigma=2:

- media 10
- media 11

Obtener el intervalo de confianza para la comparación de las 2 muestras aleatorias independientes.

2) Obtener 2 muestras aleatorias de tamaño 1000 de una distribución normal con sigma=2:

- media 10
- media 11

Obtener el intervalo de confianza para la comparación de las 2 muestras aleatorias independientes.

¿Qué se observa?

Importancia del tamaño de la muestra

```
# Muestra pequeña
set.seed(1010)
muestra1.10 <- rnorm(10, 10, 2)
muestra2.10 <- rnorm(10, 11, 2)
test.10 <- t.test(muestra1.10, muestra2.10,
alternative = "two.sided", conf.level = 0.95)

test.10$conf.int

cat("amplitud IC (n=10)", test.10$conf.int[2]-
test.10$conf.int[1])

[1] -1.826977  2.087928
amplitud IC (n=10) 3.914906
```

Importancia del tamaño de la muestra

```
# Muestra grande
set.seed(1010)
muestra1.1000 <- rnorm(1000, 10, 2)
muestra2.1000 <- rnorm(1000, 11, 2)

test.1000 <- t.test(muestra1.1000, muestra2.1000,
alternative = "two.sided", conf.level = 0.95)

test.1000$conf.int

cat("amplitud IC (n=1000)", test.1000$conf.int[2]-
test.1000$conf.int[1])
[1] -1.2059580 -0.8563001
amplitud IC (n=1000) 0.3496579
```

Potencia de un test

- La potencia de un contraste se asocia con su capacidad para determinar con precisión, es decir con un margen de incertidumbre reducido, la magnitud del efecto que se está estudiando.
- Un test será tanto más potente cuanto menor sea, para un nivel de confianza determinado, la amplitud del Intervalo de Confianza correspondiente.

La **potencia** es la probabilidad ($1-\beta$) de detectar un efecto (rechazar H0) cuando en realidad sí existe (H0 es falsa)

Es aconsejable usar valores de β inferiores a 0,2 → $1-\beta \geq 0{,}8$ (Cohen, 1988, 1992)

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences. 2nd edition*. Hillsdale, NJ: Erlbaum.
Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155. doi:http://dx.doi.org/10.1037/0033-2909.112.1.155

La potencia depende…

1. Del **tamaño del efecto** (**d**)**.** Cuanto más grande es el efecto, más fácil es identificarlo.
2. Del **nivel de significación** (α). Lo estrictos que seamos al decidir cuándo un efecto es significativo.

   Si α disminuye → β aumenta → 1- β disminuye
3. Del **tamaño de la muestra** (**n**). A mayores valores de n, menor es el error muestral y mayor la potencia.

Conocidos **d**, α y **n**, podemos obtener la potencia de un contraste.

Y, de mayor interés práctico resulta: dados **d**, α y la **potencia** (**1-** β), podemos obtener el tamaño muestral (**n**).

Potencia de un contraste

function	power calculations for
pwr.2p.test	two proportions (equal n)
pwr.2p2n.test	two proportions (unequal n)
pwr.anova.test	balanced one way ANOVA
pwr.chisq.test	chi-square test
pwr.f2.test	general linear model
pwr.p.test	proportion (one sample)
pwr.r.test	correlation
pwr.t.test	t-tests (one sample, 2 sample, paired)
pwr.t2n.test	t-test (two samples with unequal n)

Paquete **pwr**

Potencia de un contraste

- Ejemplo de llamada:

pwr.t.test(n = , d = , sig.level = , power = , type = c("two.sample", "one.sample", "paired"))

- Para cada una de las anteriores funciones, especificando 3 de los 4 parámetros **d**, α, (**1-** β) y **n**, se obtiene el cuarto.
- El nivel de significación por defecto es 0,05, Para calcular el valor α utilizar "sig.level=NULL"
- Determinar el tamaño del efecto no es sencillo. La sugerencia de Cohen (1992) es útil, pero no debe prevalecer sobre la experiencia en el problema que se está tratando.

Obtención de n en función de la potencia

```
### Cargar paquete pwr
library(pwr)
### Obtener n para test sobre la media de una población
n=NULL ## esto es lo que queremos obtener
efecto<-0.6
alfa<-0.05
potencia<-0.8
tipo<-"one.sample"
res<-pwr.t.test(n, efecto, alfa, potencia, tipo)
# Entero inmediatamente superior
ceiling(res$n)

[1] 24 # Tamaño de la muestra
```

Obtención de d en función de n

```
### Obtener el tamaño del efecto para un n dado en un
test sobre la media de una población

n = observaciones disponibles en la variable GASTO del
dataset datos1

efecto ?

alfa = 0.05

potencia = 0.8
```

Obtención de d en función de n

```
> pwr.t.test(n = length(GASTO), d=NULL, sig.level =
0.05, power = 0.8, type = "one.sample")

     One-sample t test power calculation

              n = 131
              d = 0.2466023 (d < 0,5 puede detectar efectos pequeño)
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

¿Qué tamaño de la muestra necesitaríamos para, con los mismos α y potencia, detectar un tamaño del efecto de 0,1, por ejemplo?

Tamaño del efecto

Uno de los problemas de los tests de hipótesis tradicionales es que el hecho de que un efecto resulte significativo no da información sobre la importancia del mismo. En los enfoques más modernos aparece el concepto de tamaño del efecto.

El **tamaño del efecto** es una medida objetiva (y generalmente estandarizada) de la magnitud del efecto observado.

El tamaño del efecto:

- Es comparable entre diferentes estudios y diferentes escalas de medida.
- No depende tanto del tamaño de la muestra.
- Permite evaluar objetivamente la importancia de un efecto.

Importancia del tamaño de la muestra

1) Obtener 2 muestras aleatorias de tamaño 10 de una distribución normal con sigma=2:

- media 10
- media 11

Obtener el intervalo de confianza para la comparación de las 2 muestras aleatorias independientes.

2) Obtener 2 muestras aleatorias de tamaño 1000 de una distribución normal con sigma=2:

- media 10
- media 11

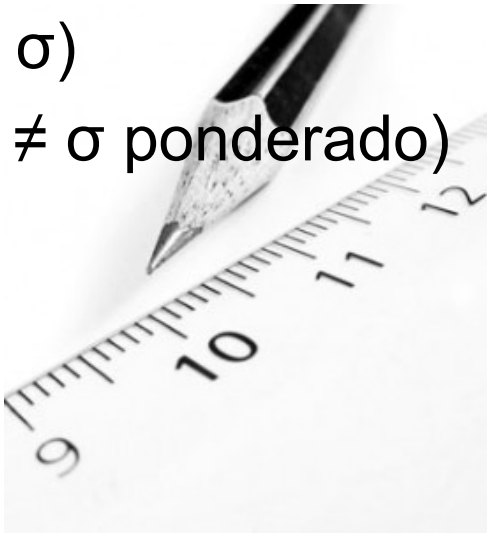
Obtener el intervalo de confianza para la comparación de las 2 muestras aleatorias independientes.

¿Qué se observa?

Medidas del tamaño del efecto

- Un problema en las muestras grandes (**Big Data**) es que un efecto se puede detectar aunque éste sea poco importante.

- Se pueden usar distintas medidas para el tamaño del efecto:
  - ***d*** d**e Cohen** (comparación medias)
  - ***h*** d**e Cohen** (comparación proporciones)
  - ***r*** de **Pearson** (comparación de efectos)
  - **Δ** de **Glass** (comparación de medias ≠ σ)
  - ***g*** de **Hedges** (comparación de medias ≠ σ ponderado)
  - **Odd ratio** / risk rates



d de Cohen

- Para la comparación de medias
- Es la diferencia entre las medias de los tratamientos comparados estandarizada por la desviación típica.

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

- Si no podemos asumir igualdad de varianzas, se utiliza como estimación de σ, si no se conocen, la del grupo de control (**Δ** de **Glass**), o bien una media ponderada de las varianzas de las dos muestras (**g** de **Hedges**).

$$\hat{d} = \frac{\bar{X}_1 - \bar{X}_2}{S_p}$$

$$S_p = \sqrt{\frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

h de Cohen

- Para la comparación de proporciones
- Es la diferencia entre las proporciones de los tratamientos comparados transformada mediante arcoseno.

$$h = 2(arcsen\sqrt{p_1}) - arcsen\sqrt{p_2}$$

- Donde p_1 y p_2 son las proporciones de las muestras que se comparan

r de Pearson

- La r de Pearson, es le **coeficiente de correlación** entre dos variables:

$$r_{XY} = \frac{Cov_{XY}}{S_X S_Y} = \frac{S_{XY}}{S_X S_Y}$$

- Es una medida bastante buena e intuitiva del tamaño del efecto, aunque ésta sí depende del tamaño de la muestra, al contrario que la d de Cohen.
- Dado que r es una medida de la fuerza de la relación entre dos variable, de algún modo describe el tamaño del posible efecto de una sobre otra.

d de Cohen vs r de Pearson

– Una de las ventajas de la r de Pearson es que sólo toma valores entre -1 y 1, por lo que la magnitud del efecto puede determinarse fácil y objetivamente.

– La ventaja de la d de Cohen es que no depende del tamaño muestral y es fiable en muestras pequeñas o análisis con muestras de tamaños muy distintos.

Magnitud del tamaño del efecto

d de Cohen	
0,2	Pequeño
0,5	Mediano
0,8	Grande

r de Pearson	
0,1	Pequeño (10%)
0,3	Mediano (30%)
0,5	Grande (50%)

Estos valores deben considerarse en el contexto del análisis o investigación.

Obtención de d para la comparación de medias

```
### d de Cohen PARA LA COMPARACIÓN DE MEDIAS
#########################################
Calcular la d de Cohen del ejemplo de "Importancia del
tamaño de la muestra", en el caso de la muestra de
tamaño 10 y en la de 1000 y haz un contraste en cada
caso.
```

El tamaño del efecto puede considerarse pequeño.

¿Es este resultado coherente con el del contraste?

Obtención de d para la comparación de medias

```
### d de Cohen PARA LA COMPARACIÓN DE MEDIAS
#########################################################
### Ejemplo de comparación del gasto total por cliente
en una tienda on line (X) entre hombres y mujeres
### Variables "SEXO" y "GASTO". Data set "datos1".
Archivo "datos.RData"

> d ## d de Cohen
[1] -0.1194692
```

El tamaño del efecto puede considerarse pequeño.

¿Es este resultado coherente con el del contraste?

Consideraciones sobre el muestreo

1. No solo es necesario que el tamaño de la muestra sea suficiente, sino también que la **muestra sea representativa** de la población que tratamos de describir
2. Una **muestra de gran tamaño no garantiza** que el margen de error sea pequeño, pues puede estar sesgada hacia segmentos de la población representados en exceso o poco representados
3. Si la población a estudiar es demasiado grande es recomendable **segmentarla en estratos** y valorar en cuáles de ellos pueden obtenerse muestras representativas, facilitando así una interpretación de los resultados más precisa
4. En general, **el margen de error en cada estrato** suele ser superior al margen de error de toda la muestra en conjunto. Es recomendable ser consciente de esta diferencia de precisión en la interpretación de resultados.

Más información

- El paquete ***sampling*** de R también proporciona otros métodos de muestreo

- Ver ***Random*** para la generación de números aleatorios
  - http://127.0.0.1:30245/help/library/base/help/RNG

Glosario

Conglomerado
Error Muestral
Estrato
Muestra
Muestra de conveniencia
Muestreo aleatorio o probabilístico
Muestreo aleatorio simple
Muestreo estratificado
Muestreo no aleatorio
Muestreo por conglomerados
Muestreo por cuotas
Muestreo por etapas
Muestreo sistemático
Población
Potencia de un test
Tamaño efecto
Tamaño muestra

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

5- Muestreo

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

www.upv.es

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