Informes automatizados de estadística univariante y bivariante

III Jornadas de Usuarios de R

Tania Iglesias, Patricia Díaz, Alexandra González

Unidad de Consultoría Estadística
Servicios Científico-Técnicos
Universidad de Oviedo

Madrid, 18 de noviembre de 2011

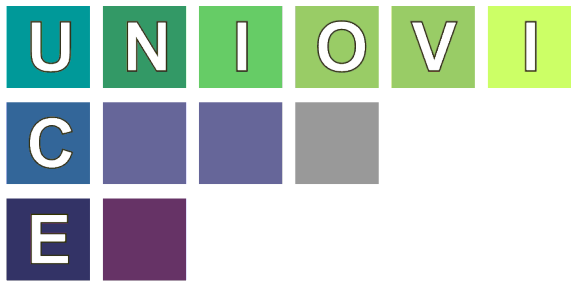
Índice

1. Introducción

2. Análisis estadístico
   - Análisis Univariante
     - Variables continuas
     - Variables nominales
   - Análisis Bivariante
     - Análisis bivariante nominal nominal
     - Análisis bivariante nominal continua
     - Análisis bivariante continua continua

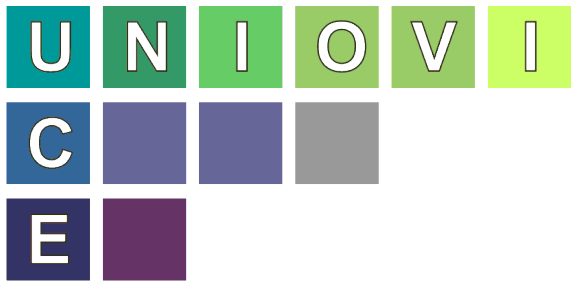
Primeros pasos...

1. Petición por parte del usuario
2. Reunión conjunta para establecer objetivos
3. Realización del análisis estadístico
4. Entrega de informe de resultados



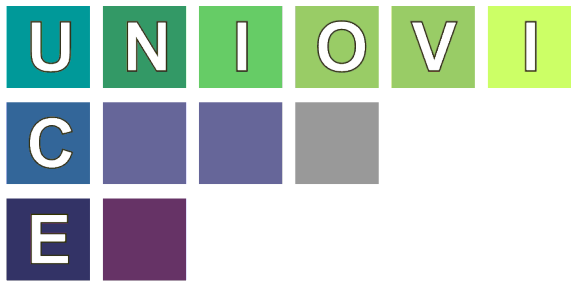
Primeros pasos...

1. Petición por parte del usuario
2. Reunión conjunta para establecer objetivos
3. Realización del análisis estadístico
4. Entrega de informe de resultados



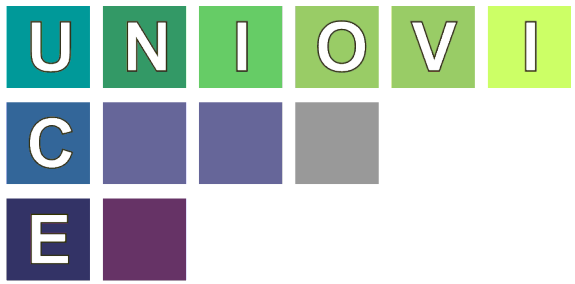
Primeros pasos...

1. Petición por parte del usuario
2. Reunión conjunta para establecer objetivos
3. Realización del análisis estadístico
4. Entrega de informe de resultados



Primeros pasos...

1. Petición por parte del usuario
2. Reunión conjunta para establecer objetivos
3. Realización del análisis estadístico
4. Entrega de informe de resultados



Primeros pasos...

Pasos previos

1. Análisis exploratorio o descriptivo de datos
2. Representaciones gráficas
3. Estudio de relaciones entre variables

Necesidad y realización de informes

- Automatización mediante la definición de funciones en R
- Utilización del paquete Sweave: integración con LaTeX

Primeros pasos...

Pasos previos

1. Análisis exploratorio o descriptivo de datos
2. Representaciones gráficas
3. Estudio de relaciones entre variables

Necesidad y realización de informes

- Automatización mediante la definición de funciones en R
- Utilización del paquete Sweave: integración con LaTeX

Definición de funciones en R

Estadística Univariante

- Descripción de variables continuas
- Descripción de variables categóricas o nominales

Estadística Bivariante

- Relación entre variables nominales
- Relación entre variables continuas
- Relación entre una variable continua y otra nominal

Estadística Univariante

Variables continuas

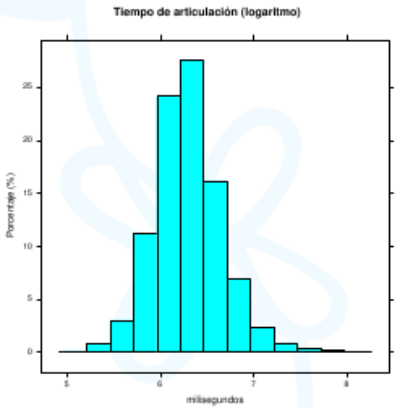
```
univariantecontinualatexparagraph <- function
  ( x
  , labelx = deparse(substitute(x))
  , namex = deparse(substitute(x))
  , printoutliers = TRUE
  , printtable = TRUE
  , printgraph = TRUE
  , printtestdistribution = TRUE
  , printblock = TRUE
  , h0mu = mean(x, na.rm=TRUE )
  , latexwidthblock = 0.485
  , dirgraph = "Dibujos"
  , latexwidthgraph = 0.485
  , significancelevelalpha = 0.05
  , ... )
```

Análisis univariante variable numérica

- Descriptivos
  - Media, mediana, cuartiles
  - Número de casos válidos
  - Número de casos perdidos
  - Mínimo y máximo
- Valores atípicos
- Normalidad: Shapiro Wilk
- Simetría: Test de Mira y Wilcoxon
- Igualdad de medias: Student o Bootstrap

Respecto a la variable *Tiempo de articulación (logaritmo)*, resulta que se dispone de 3561 casos válidos, ya que se produce un 6.71 % de casos perdidos en esta magnitud. El valor medio se alcanza en 6.3 milisegundos, con una desviación típica de 0.37, mientras que la mediana disminuye hasta 6.27 milisegundos. El valor mínimo se alcanza en 5.08 y el máximo asciende a 8.1. El 50 % de las observaciones centrales se encuentran entre 6.07 y 6.51, estando estos datos agrupados en un intervalo de amplitud 0.44 milisegundos, si bien el 25 % de los datos inferiores se encuentran en una amplitud de 0.99 y el 25 % de las valores más grandes se ubican en una distancia de 1.6 milisegundos. Se detectan 19 y 65 observaciones extremas en la tramo inferior y superior de la distribución, respectivamente. La presencia de estas observaciones atípicas no influye significativamente en los estadísticos calculados. En la tabla aparece como NA los valores perdidos, y %(NA+) y %(NA-) representan la distribución porcentual incluyendo o no los casos perdidos, respectivamente.

	Frec.	%(NA+)	%(NA-)
[5.08,5.38)	17	0.4	0.5
[5.38,5.68)	108	2.8	3.0
[5.68,5.99)	511	13.4	14.3
[5.99,6.29)	1220	32.1	34.3
[6.29,6.59)	1041	27.4	29.2
[6.59,6.89)	471	12.4	13.2
[6.89,7.2)	135	3.6	3.8
[7.2,7.5)	37	1.0	1.0
[7.5,7.8)	16	0.4	0.4
[7.8,8.1]	5	0.1	0.1
NA's	239	6.3	0.0
Total	3800	100.0	100.0



Respecto a la distribución de los datos, se rechaza la hipótesis de normalidad (test de Shapiro-Wilk, p-valor=0).

Valoración

El valor medio obtenido es consistente.

Estadística Univariante

Variables nominales

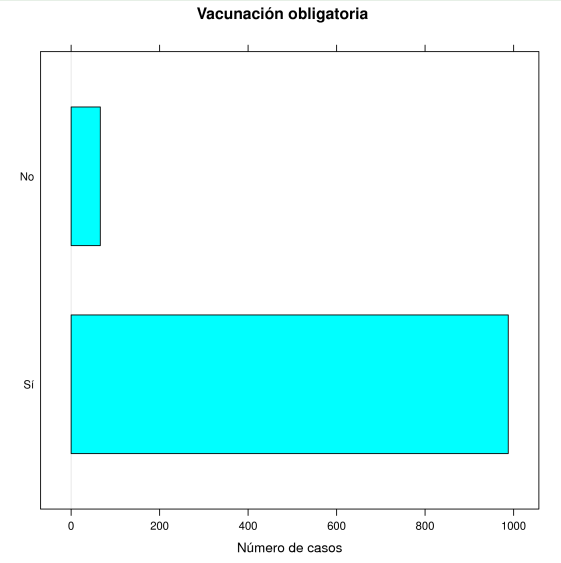
```
univariantenominallatexparagraph <- function
      ( x
      , labelx = "etiqueta x"
      , namex = "x"
      , isordered = FALSE
      , nombrefiltro =""
      , unitsx = "unidades"
      , ndigits = 2
      , printtable = TRUE
      , printgraph = TRUE
      , latexwidthblock = 0.485
      , dirgraph = "Dibujos"
      , latexwidthgraph = 0.485
      , significancelevelalpha = 0.05
      , ... )
```

Análisis univariante variable nominal

- Frecuencias
- Porcentajes

Vacunación obligatoria

	Frec.	%	Acum. %
No	66	6.3	6.3
Sí	988	93.7	100.0
Total	1054	100.0	100.0



Estadística bivariante

Tipos de análisis

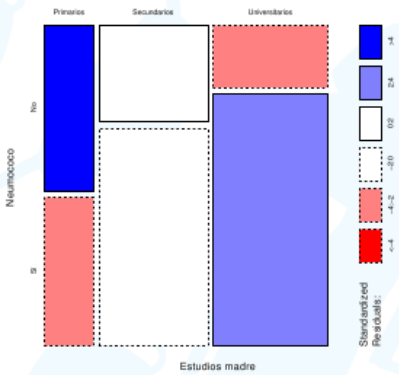
- Nominal- Nominal
- Nominal- Continua
- Continua- Continua

Estadística bivariante: nominal nominal

Test implementados

- Número sufiente de casos:
  - Test de Pearson
- Número insufiente de casos:
  - Test de Fisher
  - Test de Barnard (2x2)

Se realizado el análisis para estudiar la relación entre *Neumococo* y *Estudios madre*, obteniéndose que se rechaza la hipótesis de independencia (test de Pearson, p-valor=0). Se detectan relaciones en las siguientes modalidades: *No* con *Universitarios* (residuo=-3.93), *Sí* con *Primarios* (residuo=-3.69), *Sí* con *Universitarios* (residuo=2.57) y *No* con *Primarios* (residuo=5.64).

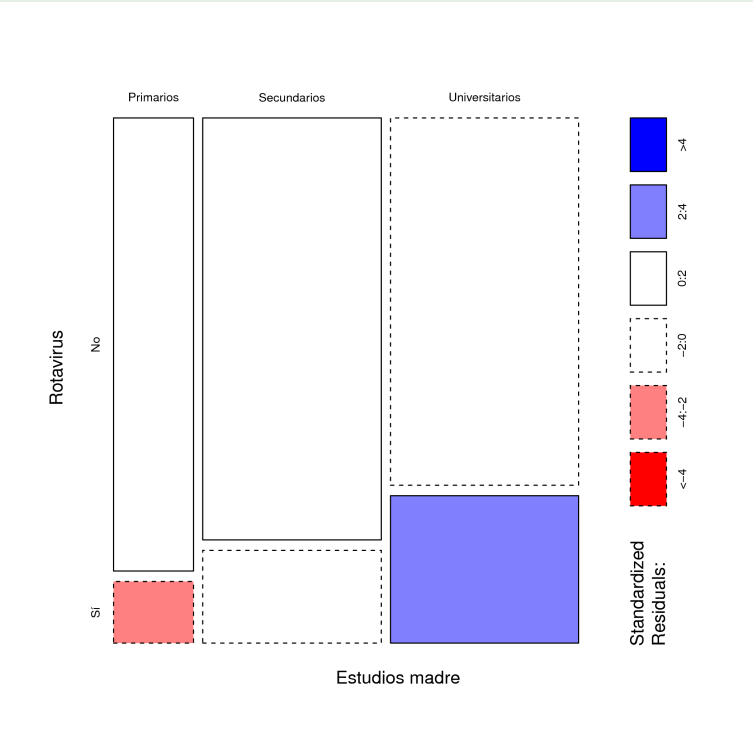


	Primarios				Secundarios				Universitarios			
	n	%Fila	%Col.	Resid.	n	%Fila	%Col.	Resid.	n	%Fila	%Col.	Resid.
No	97	52.72	31.49	5.64	126	30.66	40.91	0.26	85	19.63	27.60	-3.93
Sí	87	47.28	12.08	-3.69	285	69.34	39.58	-0.17	348	80.37	48.33	2.57

Estadística bivariante: nominal nominal

Relación tipo de estudios y vacunación

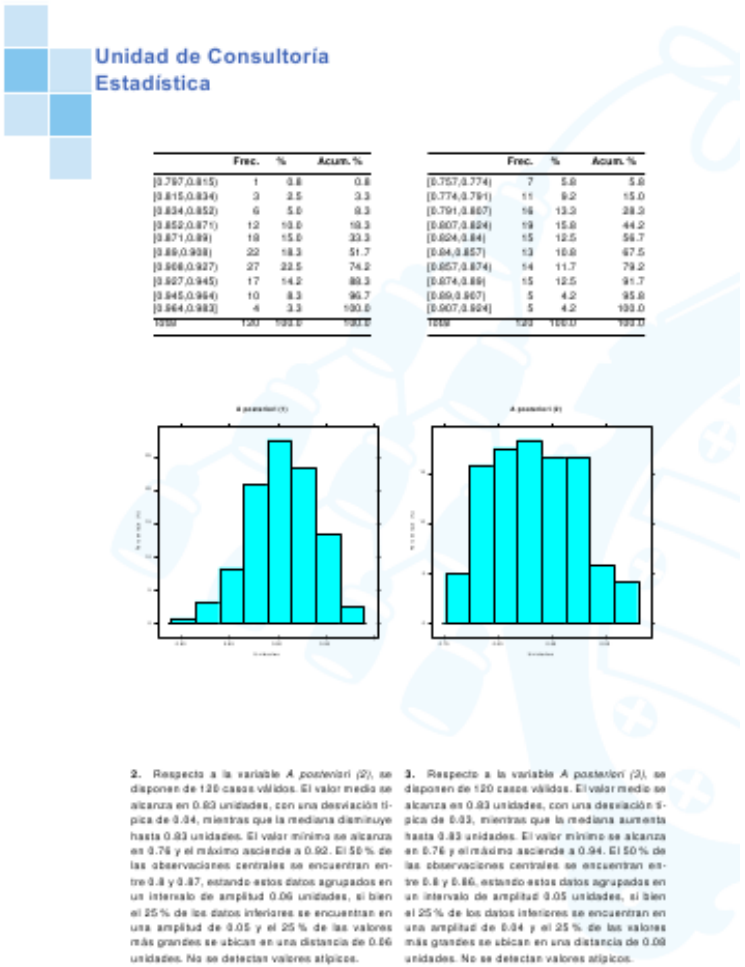
	Primarios				Secundarios				Universitarios			
	n	**%Fila**	**%Col.**	**Resid.**	**n**	**%Fila**	**%Col.**	**Resid.**	**n**	**%Fila**	**%Col.**	**Resid.**
No	162	88.04	20.05	1.44	337	82.00	41.71	0.78	309	71.36	38.24	-1.7
Sí	22	11.96	10.00	**-2.77**	74	18.00	33.64	-1.49	124	28.64	56.36	**3.26**



Estadística bivariante: nominal continua

Test estadísticos

- Univariante para cada uno de los niveles
- Normalidad
- Homocedasticidad:
  - Test F
  - Test de Ansary Bradley
  - Test de Barlett
  - Test de Flinger-Killen

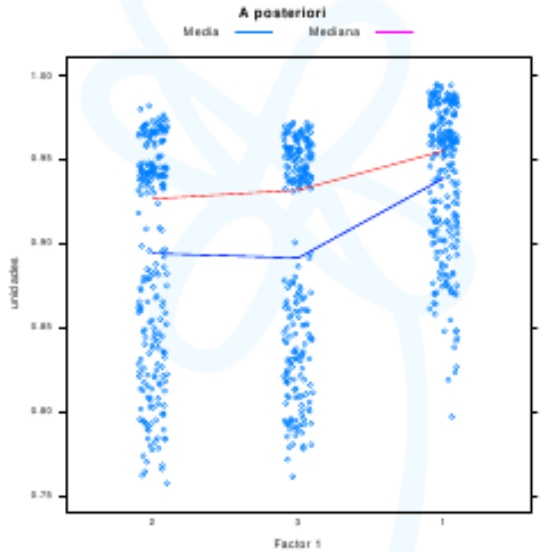


Estadística bivariante: nominal continua

Test Igualdad de medias

- Test paramétricos
  - Student
  - Welch
  - Anova
- Test no paramétricos
  - Wilcoxon
  - Kruskall-Wallis
  - Permutaciones

El test de permutaciones indica que en ciertos pares de niveles se producen diferencias significativas entre sí. Si los ordenados por orden de significatividad, resulta la siguiente prelación: *2* y *1* (p-valor=0) y *3* y *1* (p-valor=0), respectivamente. Por el contrario, no se producen diferencias entre los siguientes niveles: *3* y *2* (p-valor=0.75), respectivamente.



Valoración

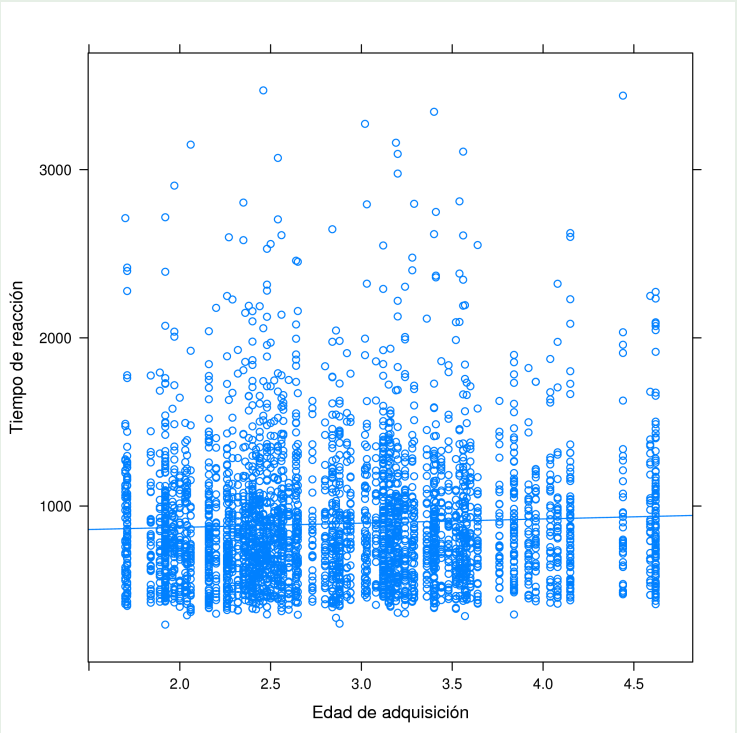
Se detectan diferencias significativas entre los grupos considerados.

Estadística bivariante: continua continua

Análisis

Regresión lineal

Gráfico



¡Gracias por vuestra atención!