Introducción al Análisis Exploratorio de Datos (EDA) en R

Módulo 2

2024-03-16

Continuación de EDA univariante: supuestos

Distribución normal univariada

Una variable aleatoria continua X está normalmente distribuida si su **función de densidad** sigue la forma:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{\frac{-1}{2\sigma^2}(x-\mu)^2\right\}$$

donde

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$\int_{a}^{b} f(x)dx = P(a \leq x \leq b)$$

Parámetros de la distribución: media(μ) y varianza (σ^2) de la distribución.

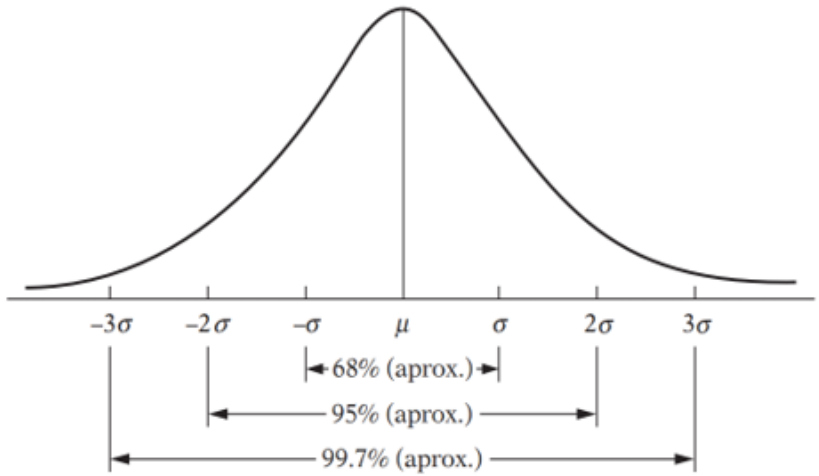
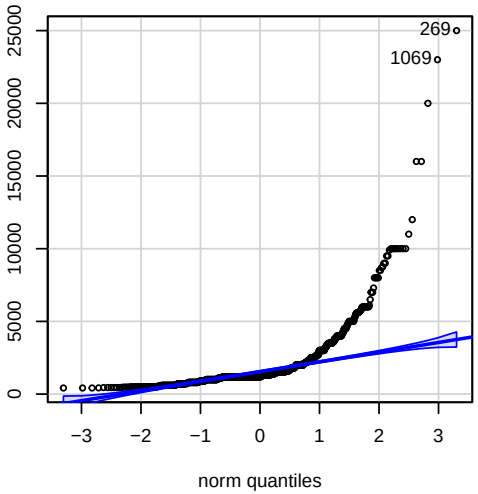


Figure 1: Áreas bajo la curva normal

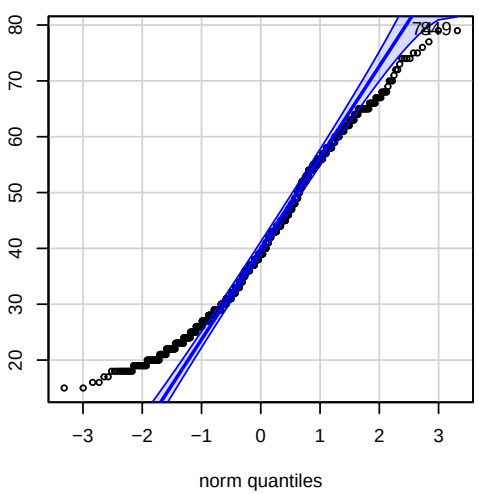
Normalidad univariada

El gráfico **cuantil-cuantil** (gráfico Q-Q) compara dos distribuciones de probabilidad usando sus cuantiles. Usando la librería `car`, comparamos la distribución de probabilidad de una muestra aleatoria con la distribución normal.

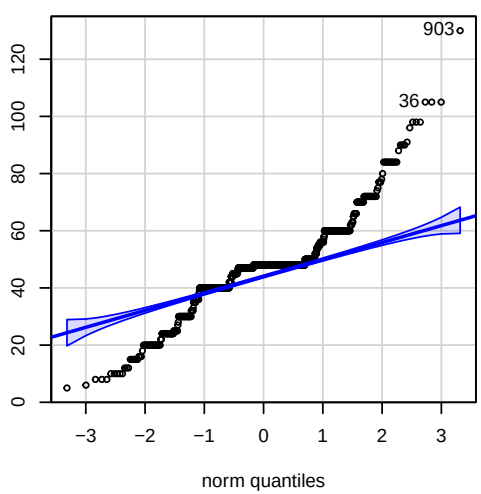
Ingreso laboral (miles $)



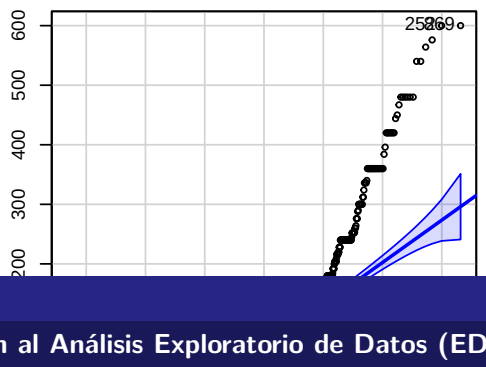
Edad (años)



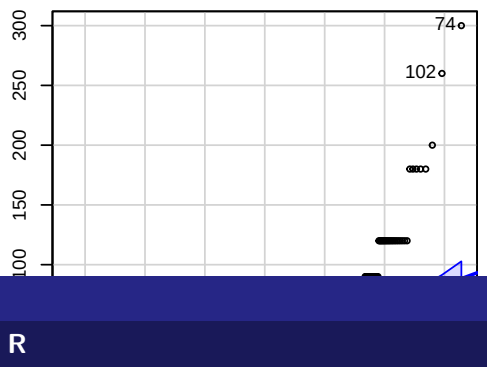
Horas trabajadas (semana)



Tiempo en el trabajo actual



Tiempo de desplazamiento



Normalidad univariada (cont.)

La intuición del método gráfico es verificada mediante pruebas formales. Usando la librería `nortest`:

variable	Anderson-Darling	Lilliefors (Kolmogorov-Smirnov)	Pearson chi-square
ingreso	0	0	0
edad	0	0	0
horas_semana	0	0	0
t_actual	0	0	0
t_viaje	0	0	0

Se examinan las siguientes transformaciones

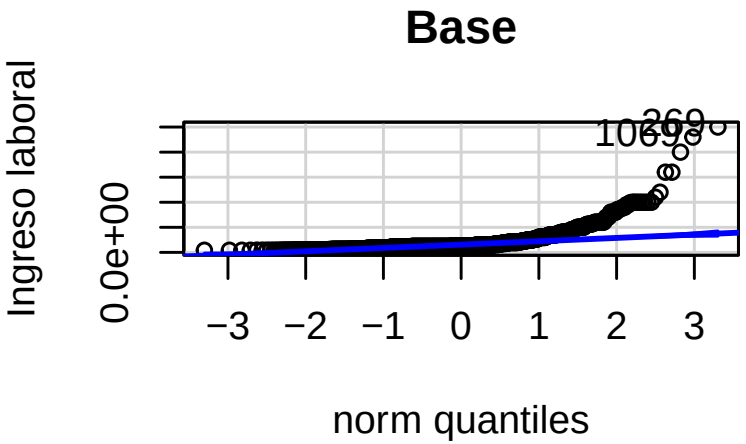
log(x)	A-D	K-S)	P
ingreso	0	0	0
edad	0	0	0
horas_semana	0	0	0
t_actual	0	0	0
t_viaje	0	0	0

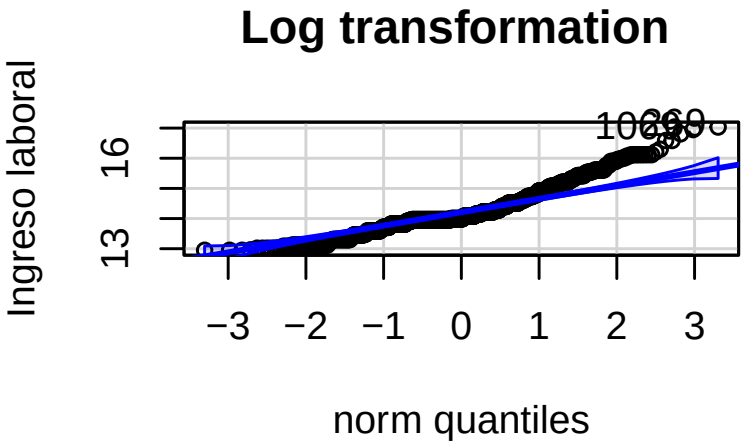
sqrt(x)	A-D	K-S)	P
ingreso	0	0	0
edad	0	0	0
horas_semana	0	0	0
t_actual	0	0	0
t_viaje	0	0	0

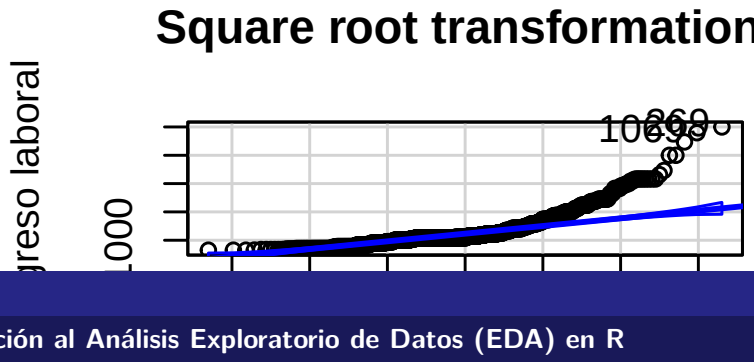
cuberoot(x)	A-D	K-S)	P
ingreso	0	0	0
edad	0	0	0
horas_semana	0	0	0
t_actual	0	0	0
t_viaje	0	0	0

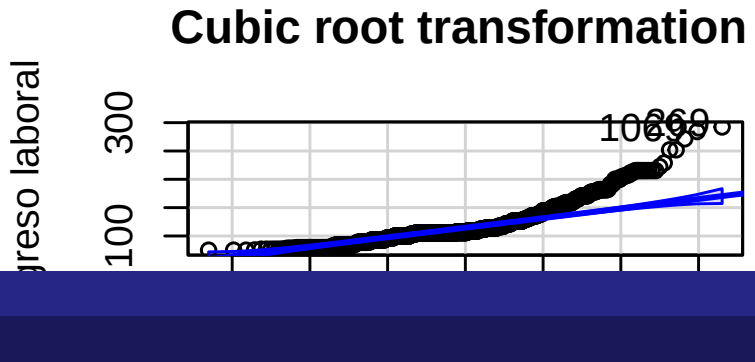
Normalidad univariada (cont.)

Considérese el análisis sobre las transformacions del ingreso laboral (miles $). Como se verificó en el **Módulo 2**, la transformación logarítmica reduce la influencia de los valores atípicos. En lo sucesivo, nuestro análisis emplea el **logaritmo del ingreso laboral**.









Resumen descriptivo (mediana e IQR)

Se verificó que las variables continuas no siguen una distribución normal y, en consecuencia, el resumen descriptivo reporta la mediana y el rango intercuartílico.

Área	Ingreso	Edad	Horas (semana)	Tiempo actual	Tiempo de viaje
Total	1200000 (1100000 - 2e+06)	39 (29 - 51)	48 (40 - 48)	36 (12 - 108)	25 (15 - 40)
Barranquilla	1200000 (1160000 - 1875000)	36 (29 - 47.5)	48 (47.25 - 48)	33 (11.25 - 81)	30 (20 - 45)
Bogota	1360000 (1160000 - 2337500)	42 (28 - 50)	48 (42 - 56)	36 (11 - 84)	40 (20 - 60)
Bucaramanga	1250000 (1160000 - 2e+06)	38 (29 - 50.5)	47 (40 - 48)	47 (12 - 120)	28 (15 - 30)
Cali	1160000 (1112500 - 1675000)	40.5 (31 - 54)	47 (40 - 48)	36 (12 - 114)	30 (20 - 45)
Cartagena	1200000 (9e+05 - 2e+06)	39 (29 - 53)	48 (40 - 48)	60 (24 - 120)	20 (15 - 30)
Cucuta	1160000 (8e+05 - 1487500)	38 (31 - 48)	48 (40 - 60)	24 (8 - 86)	20 (15 - 30)
Ibague	1450000 (1160000 - 3400000)	40 (27 - 46)	48 (40 - 48)	28 (11 - 120)	20 (10 - 30)
Manizales	1200000 (1160000 - 2e+06)	42 (32 - 53)	48 (42 - 48)	36 (12 - 108)	20 (15 - 30)
Medellin	1303000 (1160000 - 2e+06)	36.5 (28 - 48)	48 (47 - 48)	24 (10 - 108)	35 (20 - 48.75)
Monteria	1160000 (780000 - 2050000)	37 (28 - 47)	47 (40 - 48)	36 (12 - 73.5)	15 (11 - 20)
Pasto	1200000 (9e+05 - 1820000)	44 (32.5 - 54)	48 (40 - 50)	36 (12 - 144)	20 (10 - 30)
Pereira	1275000 (1160000 - 2e+06)	38 (30 - 47)	48 (40 - 48)	36 (12 - 72)	20 (12 - 40)
Villavicencio	1225500 (1160000 - 2e+06)	43 (28 - 55)	48 (47 - 54)	24 (4 - 120)	25 (15 - 30)

Introducción al EDA multivariante

EDA multivariante

El **EDA multivariante** está fundamentado en la exploración, de manera simultánea, de dos o más características (variables) medidas en un conjunto de casos (Kachigan, 1991). El análisis univariante se centra en la variación, mientras que el análisis multivariante se centra en la **covariación** y **correlaciones** que refleja un conjunto de variables.

El análisis multivariante no sólo incluye estrategias de análisis exploratorio; sino, además, las siguientes técnicas:

- Componentes principales
- Análisis factorial
- Regresión múltiple
- Análisis discriminante múltiple
- Análisis multivariante de varianza
- Análisis cluster
- Análisis de correspondencias
- etc.

Independencia de variables catégoricas

La siguiente gráfica representa la asociación entre el nivel educativo máximo alcanzado por el trabajador (**educación**) y el área metropolitana en que reside.

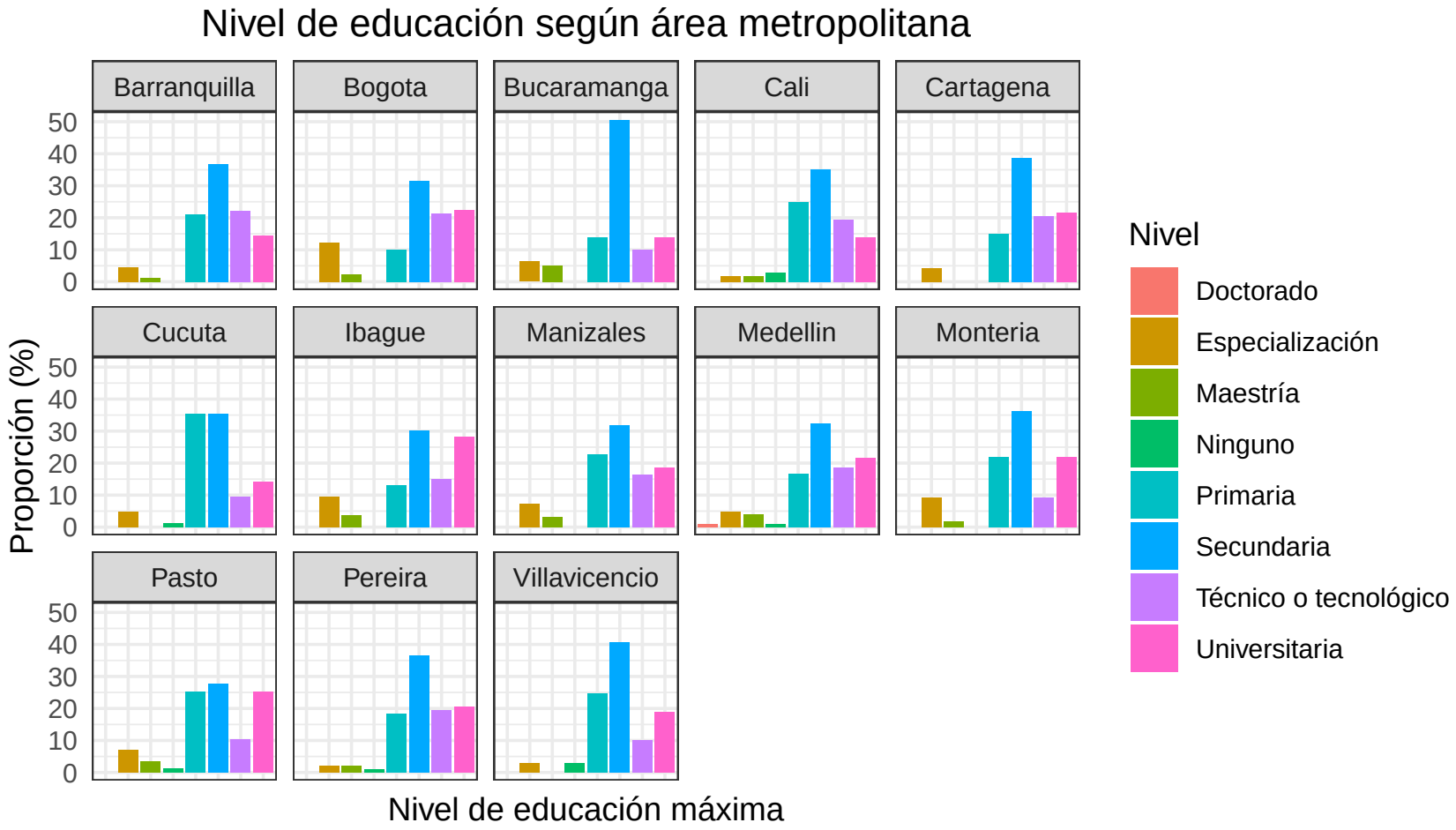


Tabla de contingencia

La **tabla de contingencia** permite resumir la información reportada en la figura anterior:

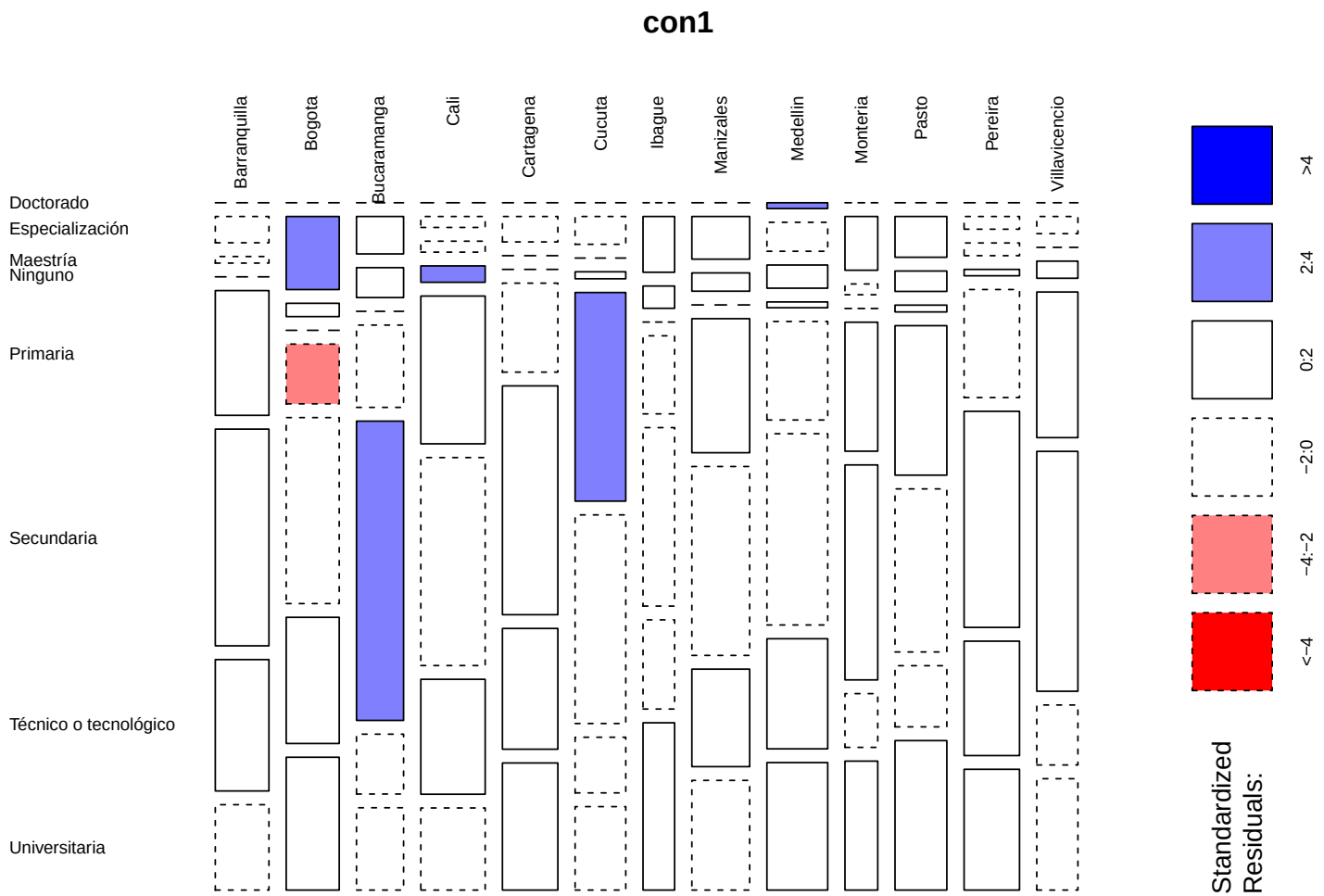
	Doctorado	Especialización	Maestría	Ninguno	Primaria	Secundaria	Técnico o tecnológico	Universit
Barranquilla	0	4	1	0	19	33	20	
Bogota	0	11	2	0	9	28	19	
Bucaramanga	0	5	4	0	11	40	8	
Cali	0	2	2	3	27	38	21	
Cartagena	0	4	0	0	14	36	19	
Cucuta	0	4	0	1	30	30	8	
Ibague	0	5	2	0	7	16	8	
Manizales	0	7	3	0	22	31	16	
Medellin	1	5	4	1	17	33	19	
Monteria	0	5	1	0	12	20	5	
Pasto	0	6	3	1	22	24	9	
Pereira	0	2	2	1	17	34	18	
Villavicencio	0	2	0	2	17	28	7	
Sum	1	62	24	9	224	391	177	

La conclusión es verificada mediante la implementación de una prueba χ^2 de independencia

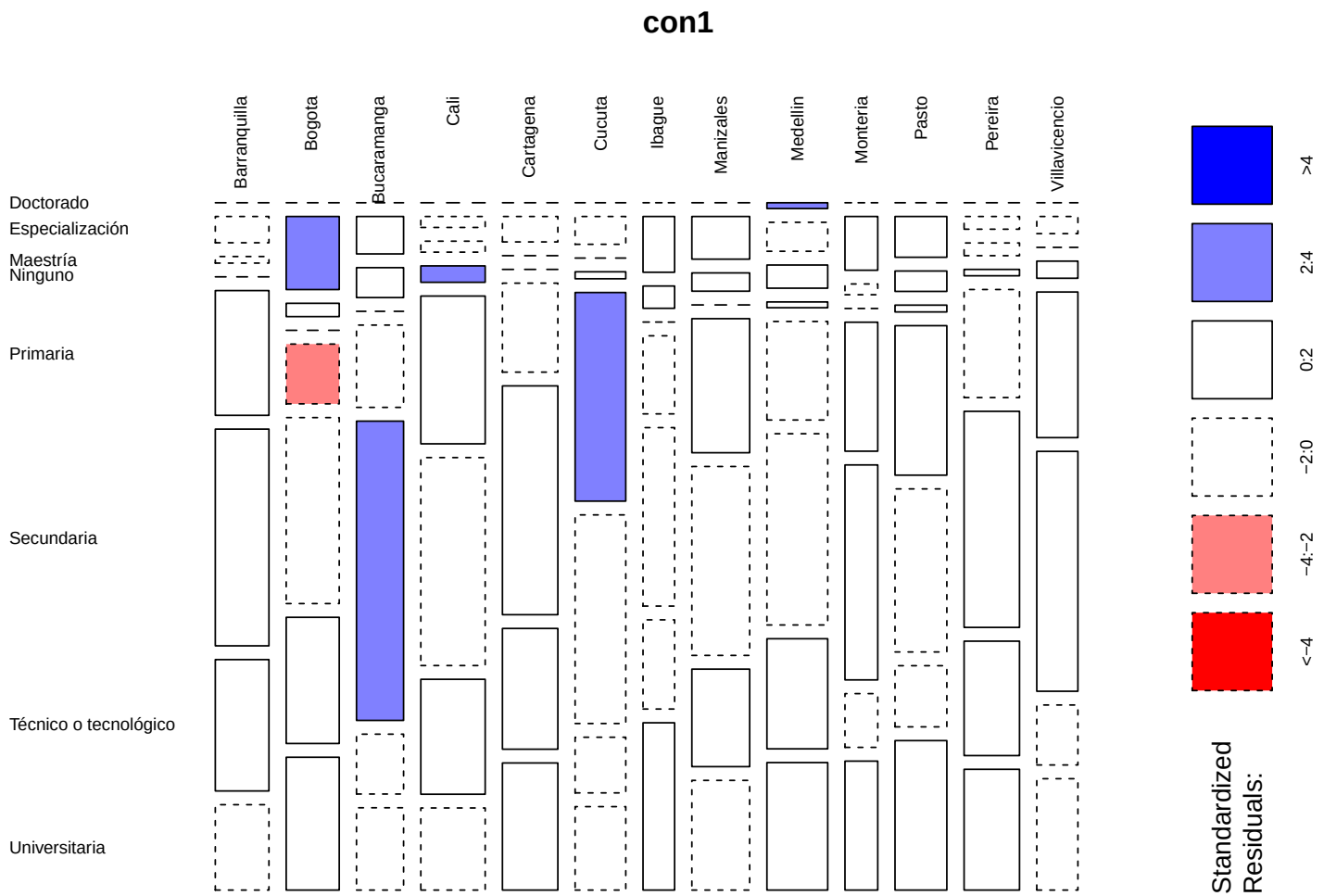
AQUÍ VA LA DESCRIPCIÓN FORMAL DE LA PRUEBA CHI-2 DE INDEPENDENCIA

```
##
##  Pearson's Chi-squared test
##
## data:  dataset$area and dataset$edu
## X-squared = 109.12, df = 84, p-value = 0.03415
```

Tabla de contingencia (cont.)



Resumen: independencia de variables categóricas



Resumen: independencia de variables categóricas

La siguiente es una forma útil de incluir la prueba χ^2 de independencia en la visualización de los datos:

```
##
## Attaching package: 'ggpubr'

## The following object is masked from 'package:plyr':
##
##     mutate

## Warning in chisq.test(dataset$sexo, dataset$posic): Chi-squared approximation
## may be incorrect
```