SESION 3

***Sesión 3: Estadística descriptiva con rstudió***

**Contenido**

Tablas de distribución de frecuencias para variables cualitativas y cuantitativas, gráficos estadísticos en R.

**Actividad 3:**

Aplicaciones en R utilizando los datos de esta sesión.

**Tablas de frecuencia**

Las tablas de frecuencia son muy utilizadas en estadística y R permite crear tablas de una forma sencilla. En este capítulo se explican las principales funciones para la elaboración de tablas.

**Tabla de contingencia con table**

La función table sirve para construir tablas de frecuencia de una vía, a continuación la estrctura de la función.

table(..., exclude, useNA)

Los parámetros de la función son:

* ... espacio para ubicar los nombres de los objetos (variables o vectores) para los cuales se quiere construir la tabla.
* exclude: vector con los niveles a remover de la tabla. Si exclude=NULL implica que se desean ver los NA, lo que equivale a useNA = 'always'.
* useNA: instrucción de lo que se desea con los NA. Hay tres posibles valores para este parámetro: 'no' si no se desean usar, 'ifany' y 'always' si se desean incluir.

### Ejemplo: tabla de frecuencia de una vía

Considere el vector fuma mostrado a continuación y construya una tabla de frecuencias absolutas para los niveles de la variable frecuencia de fumar.

fuma <- c('Frecuente', 'Nunca', 'A veces', 'A veces', 'A veces',

'Nunca', 'Frecuente', NA, 'Frecuente', NA, 'hola',

'Nunca', 'Hola', 'Frecuente', 'Nunca')

A continuación se muestra el código para crear la tabla de frecuencias para la variable fuma.

table(fuma)

## fuma

## A veces Frecuente hola Hola Nunca

## 3 4 1 1 4

De la tabla anterior vemos que NO aparece el conteo de los NA, para obtenerlo usamos lo siguiente.

table(fuma, useNA='always')

## fuma

## A veces Frecuente hola Hola Nunca <NA>

## 3 4 1 1 4 2

Vemos que hay dos niveles errados en la tabla anterior, Hola y hola. Para construir la tabla sin esos niveles errados usamos lo siguiente.

table(fuma, exclude=c('Hola', 'hola'))

## fuma

## A veces Frecuente Nunca <NA>

## 3 4 4 2

Por último construyamos la tabla sin los niveles errados y los NA, a esta última tabla la llamaremos tabla1 para luego poder usarla. Las instrucciones para hacer esto son las siguientes.

tabla1 <- table(fuma, exclude=c('Hola', 'hola', NA))

tabla1

## fuma

## A veces Frecuente Nunca

## 3 4 4

Al crear una tabla con la instrucción table(var1, var2), la variable 1 quedará por filas mientras que la variable 2 estará en las columnas.

### Ejemplo: tabla de frecuencia de dos vías

Considere otro vector sexo mostrado a continuación y construya una tabla de frecuencias absolutas para ver cómo se relaciona el sexo con fumar del ejemplo anterior.

sexo <- c('Hombre', 'Hombre', 'Hombre', NA, 'Mujer',

'Casa', 'Mujer', 'Mujer', 'Mujer', 'Hombre', 'Mujer',

'Hombre', NA, 'Mujer', 'Mujer')

Para construir la tabla solicitada usamos el siguiente código.

table(sexo, fuma)

## fuma

## sexo A veces Frecuente hola Hola Nunca

## Casa 0 0 0 0 1

## Hombre 1 1 0 0 2

## Mujer 1 3 1 0 1

De la tabla anterior vemos que aparecen niveles errados en fuma y en sexo, para retirarlos usamos el siguiente código incluyendo en el parámetro exclude un vector con los niveles que **NO** deseamos en la tabla.

tabla2 <- table(sexo, fuma, exclude=c('Hola', 'hola', 'Casa', NA))

tabla2

## fuma

## sexo A veces Frecuente Nunca

## Hombre 1 1 2

## Mujer 1 3 1

## 9.2 Función prop.table

La función prop.table se utiliza para crear tablas de frecuencia relativa a partir de tablas de frecuencia absoluta, la estructura de la función se muestra a continuación.

prop.table(x, margin=NULL)

* x: tabla de frecuencia.
* margin: valor de 1 si se desean proporciones por filas, 2 si se desean por columnas, NULL si se desean frecuencias globales.

### Ejemplo: tabla de frecuencia relativa de una vía

Obtener la tabla de frencuencia relativa para la tabla1.

Para obtener la tabla solicitada se usa el siguiente código.

prop.table(x=tabla1)

## fuma

## A veces Frecuente Nunca

## 0.2727273 0.3636364 0.3636364

### Ejemplo: tabla de frecuencia relativa de dos vías

Obtener la tabla de frencuencia relativa para la tabla2.

Si se desea la tabla de frecuencias relativas global se usa el siguiente código. El resultado se almacena en el objeto tabla3 para ser usado luego.

tabla3 <- prop.table(x=tabla2)

tabla3

## fuma

## sexo A veces Frecuente Nunca

## Hombre 0.1111111 0.1111111 0.2222222

## Mujer 0.1111111 0.3333333 0.1111111

Si se desea la tabla de frecuencias relativas marginal por **columnas** se usa el siguiente código.

tabla4 <- prop.table(x=tabla2, margin=2)

tabla4

## fuma

## sexo A veces Frecuente Nunca

## Hombre 0.5000000 0.2500000 0.6666667

## Mujer 0.5000000 0.7500000 0.3333333

## 9.3 Función addmargins

Esta función se puede utilizar para agregar los totales por filas o por columnas a una tabla de frecuencia absoluta o relativa. La estructura de la función es la siguiente.

addmargins(A, margin)

* A: tabla de frecuencia.
* margin: valor de 1 si se desean proporciones por columnas, 2 si se desean por filas, NULL si se desean frecuencias globales.

### Ejemplo

Obtener las tablas tabla3 y tabla4 con los totales margines global y por columnas respectivamente.

Para hacer lo solicitado usamos las siguientes instrucciones.

addmargins(tabla3)

## fuma

## sexo A veces Frecuente Nunca Sum

## Hombre 0.1111111 0.1111111 0.2222222 0.4444444

## Mujer 0.1111111 0.3333333 0.1111111 0.5555556

## Sum 0.2222222 0.4444444 0.3333333 1.0000000

addmargins(tabla4, margin=1)

## fuma

## sexo A veces Frecuente Nunca

## Hombre 0.5000000 0.2500000 0.6666667

## Mujer 0.5000000 0.7500000 0.3333333

## Sum 1.0000000 1.0000000 1.0000000

Note que los valores de 1 y 2 en el parámetro margin de las funciones prop.table y addmargins significan lo contrario.

## 9.4 Función hist

Construir tablas de frecuencias para variables cuantitativas es necesario en muchos procedimientos estadísticos, la función hist sirve para obtener este tipo de tablas. La estructura de la función es la siguiente.

hist(x, breaks='Sturges', include.lowest=TRUE, right=TRUE,

plot=FALSE)

Los parámetros de la función son:

* x: vector numérico.
* breaks: vector con los límites de los intervalos. Si no se especifica se usar la regla de Sturges para definir el número de intervalos y el ancho.
* include.lowest: valor lógico, si TRUE una observación xi�� que coincida con un límite de intervalo será ubicada en el intervalo izquierdo, si FALSE será incluída en el intervalo a la derecha.
* right: valor lógico, si TRUE los intervalos serán cerrados a derecha de la forma (liminf,limsup](������,������], si es FALSE serán abiertos a derecha.
* plot: valor lógico, si FALSE sólo se obtiene la tabla de frecuencias mientras que con TRUE se obtiene la representación gráfica llamada histograma.

### Ejemplo

Genere 200 observaciones aleatorias de una distribución normal con media μ=170�=170 y desviación σ=5�=5, luego construya una tabla de frecuencias para la muestra obtenida usando (a) la regla de Sturges y (b) tres intervalos con límites 150, 170, 180 y 190.

Primero se construye el vector x con las observaciones de la distribución normal por medio de la función rnorm y se especifica la media y desviación solicitada. Luego se aplica la función hist con el parámetro breaks='Sturges', a continuación el código utilizado.

x <- rnorm(n=200, mean=170, sd=5)

res1 <- hist(x=x, breaks='Sturges', plot=FALSE)

res1

## $breaks

## [1] 155 160 165 170 175 180 185

##

## $counts

## [1] 4 30 61 71 26 8

##

## $density

## [1] 0.004 0.030 0.061 0.071 0.026 0.008

##

## $mids

## [1] 157.5 162.5 167.5 172.5 177.5 182.5

##

## $xname

## [1] "x"

##

## $equidist

## [1] TRUE

##

## attr(,"class")

## [1] "histogram"

El objeto res1 es una lista donde se encuentra la información de la tabla de frecuencias para x. Esa lista tiene en el elemento breaks los límites inferior y superior de los intervalos y en el elemento counts están las frecuencias de cada uno de los intervalos.

Para obtener las frecuencias de tres intervalos con límites 150, 170, 180 y 190 se especifica en el parámetros breaks los límites. El código para obtener la segunda tabla de frecuencias se muestra a continuación.

res2 <- hist(x=x, plot=FALSE,

breaks=c(150, 170, 180, 190))

res2

## $breaks

## [1] 150 170 180 190

##

## $counts

## [1] 95 97 8

##

## $density

## [1] 0.02375 0.04850 0.00400

##

## $mids

## [1] 160 175 185

##

## $xname

## [1] "x"

##

## $equidist

## [1] FALSE

##

## attr(,"class")

## [1] "histogram"

### Ejemplo

Construya el vector x con los siguientes elementos: 1.0, 1.2, 1.3, 2.0, 2.5, 2.7, 3.0 y 3.4. Obtenga varias tablas de frecuencia con la función hist variando los parámetros include.lowest y right. Use como límite de los intervalos los valores 1, 2, 3 y 4.

Lo primero que debemos hacer es crear el vector x solicitado así:

x <- c(1.1, 1.2, 1.3, 2.0, 2.0, 2.5, 2.7, 3.0, 3.4)

En la Figura [9.1](https://fhernanb.github.io/Manual-de-R/tablas.html#fig:dots) se muestran los 9 puntos y con color azul se representan los límites de los intervalos.

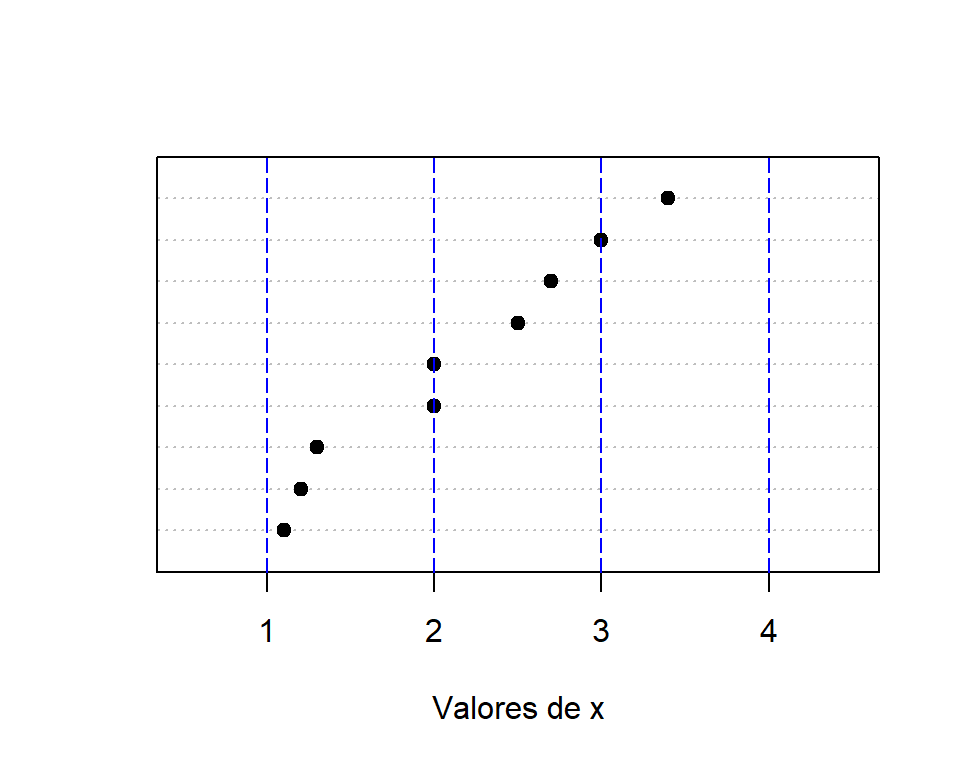


Figure 9.1: Ubicación de los puntos del ejemplo con límites en color azul.

A continuación se presenta el código para obtener la tabla de frecuencia usando rigth=TRUE, los resultados se almacenan en el objeto res3 y se solicitan sólo los dos primeros elementos que corresponden a los límites y frecuencias.

res3 <- hist(x, breaks=c(1, 2, 3, 4), right=TRUE, plot=FALSE)

res3[1:2]

## $breaks

## [1] 1 2 3 4

##

## $counts

## [1] 5 3 1

Ahora vamos a repetir la tabla pero usando rigth=FALSE para ver la diferencia, en res4 están los resultados.

res4 <- hist(x, breaks=c(1, 2, 3, 4), right=FALSE, plot=FALSE)

res4[1:2]

## $breaks

## [1] 1 2 3 4

##

## $counts

## [1] 3 4 2

Al comparar los últimos dos resultados vemos que la primera frecuencia es 5 cuando right=TRUE porque los intervalos se consideran cerrados a la derecha.

Ahora vamos a construir una tabla de frecuencia usando FALSE para los parámetros include.lowest y right.

res5 <- hist(x, breaks=c(1, 2, 3, 4),

include.lowest=FALSE, right=FALSE,

plot=FALSE)

res5[1:2]

## $breaks

## [1] 1 2 3 4

##

## $counts

## [1] 3 4 2

De este último resultado se ve claramente el efecto de los parámetros include.lowest y right en la construcción de tablas de frecuencia.

## EJERCICIOS

Use funciones o procedimientos (varias líneas) de R para responder cada una de las siguientes preguntas.

En el Cuadro [8.2](https://fhernanb.github.io/Manual-de-R/read.html#tab:toy) se presenta una base de datos sencilla. Lea la base de datos usando la funcion read.table y construya lo que se solicita a continuación.

1. Construya una tabla de frecuencia absoluta para la variable pasatiempo.
2. Construya una tabla de frecuencia relativa para la variable fuma.
3. Construya una tabla de frecuencia relativa para las variables pasatiempo y fuma.
4. ¿Qué porcentaje de de los que no fuman tienen como pasatiempo la lectura.
5. ¿Qué porcentaje de los que corren no fuman?