MÉTODOS ESTADÍSTICOS

Proyecto de análisis estadístico de Variables en R estudio

Segundo informe

Jorge Eduardo Rodriguez

Carol Valentina Villanueva

Universidad de la Salle

04/04/2022

2) Introducción

Cotidianamente se pueden observar diversidad de escenarios donde se puede atribuir el desenlace de un evento al azar, incluso por encima de la frecuencia de aquellos que pueden acontecer con una exactitud previsible. Esencialmente la aplicación de las diferentes ramas de la probabilidad reside en la capacidad para estimar o predecir eventos, incluso si estos se dan por acontecimientos del azar.

La probabilidad se centra en el porcentaje de posibilidad de que un hecho suceda, siendo 0 un evento imposible probabilísticamente hablando y 1 el valor que tendría un evento cuyo desenlace se conoce con exactitud. Las distribuciones de probabilidad para una variable en específico se pueden obtener a partir de datos experimentales con la frecuencia de ocurrencia del valor de la variable en específico, por esta razón, entre más datos o información se tengan para calcular un evento, más acertado será el resultado calculado.

Dadas las características de esta rama de la ciencia, esta tiene gran variedad de aplicación en todas las ramas del conocimiento., como, por ejemplo, el ámbito de la biología donde es de suma importancia la determinación de las características de los árboles para identificar la viabilidad de un ecosistema y de las posibles especies que se encontrarían allí. Por ello, la adquisición de probabilidades asociadas a la edad, diámetro de tronco, altura e incluso especie de los árboles deben ser predecidos lo más acertadamente con probabilidad, posibilidades que se estudiarán en el siguiente informe.

3) Análisis de la tabla de contingencia.

Empezamos importando los datos para hacer el tratamiento estadístico pertinente

```
censo <- read.csv2("CENSO ARBOREO COMUNA 15 CALI COLOMBIA.csv")
```

Se realiza la tabla de contingencia de los datos:

```
tabla_veged<-table(censo$TIPO.DE.VEGETACIÓN, censo$EDAD)
tabla_veged
```

```
##
##                     Juvenil Maduro
##   Arbol                  21     98
##   Arbusto                 7     39
##   Palma                   5     21
##   Planta arbustiva        0      9
```

Reorganizando los datos en una tabla de contingencia con las frecuencias relativas y las sumas por filas y columnas encontramos la siguiente información

```
fr <- addmargins(tabla_veged/sum(tabla_veged))
fr
```

```
##
##                     Juvenil Maduro   Sum
##   Arbol               0.105  0.490 0.595
##   Arbusto             0.035  0.195 0.230
##   Palma               0.025  0.105 0.130
##   Planta arbustiva    0.000  0.045 0.045
##   Sum                 0.165  0.835 1.000
```

Donde cada valor dentro de esta tabla se interpreta como la probabilidad de eventos en especificos:

$P(Arbol \cap Juvenil)$	$P(Arbol \cap Maduro)$	$P(Arbol)$
$P(Abusto \cap Juvenil)$	$P(Arbusto \cap Maduro)$	$P(Arbusto)$
$P(Palma \cap Juvenil)$	$P(Palma \cap Maduro)$	$P(Palma)$
$P(Planta\ arb. \cap Juvenil)$	$P(Planta\ arb. \cap Maduro)$	$P(Planta\ int.)$
$P(Juvenil)$	$P(Maduro)$	1

Calculamos las probabilidades

```
PArbol = fr[1,3] ;
PArbolJuvenil = fr[1,1] ;
PArbolMaduro = fr[1,2] ;

PArbusto = fr[2,3] ;
PArbustoJuvenil = fr[2,1] ;
PArbustoMaduro = fr[2,2] ;

PPalma = fr[3,3] ;
PPalmaJuvenil = fr[3,1] ;
PPalmaMaduro = fr[3,2] ;

PPlantaArb = fr[4,3] ;
PPlantaArbJuvenil = fr[4,1] ;
PPlantaArbMaduro = fr[4,2] ;

PJuvenil = fr[5,1] ;
PMaduro = fr[5,2] ;
```

a) La probabilidad de que al elegir un individuo al azar de la muestra este pertenezca a la categoría de maduro, dado que no pertenece a los arboles, se calcula de la siguiente forma:

$$P(Maduro|No\ es\ arbol) = \frac{P(Maduro \cap No\ es\ arbol)}{P(No\ es\ arbol)}$$

En este escenario la intersección entre estar maduro y no ser árbol es la probabilidad de que el resto de las especies sean maduros. Por tanto,

```
Pa    <-    (PArbustoMaduro+PPalmaMaduro+PPlantaArbMaduro)/(1-PArbol)    ;
Pa

## [1] 0.8518519
```

b) La probabilidad de que al elegir un individuo al azar este no pertenezca a la categoría de palma, dado que pertenece maduro es:

$$P(No\ es\ palma|Maduro) = \frac{P(No\ es\ palma \cap Maduro)}{P(Maduro)}$$

La intersección entre no ser Palma y estar maduro es la probabilidad de tener el resto de las especies maduras, por tanto,

```
Pb    <-    (PArbustoMaduro+PArbolMaduro+PPlantaArbMaduro)/(PMaduro)    ;
Pb

## [1] 0.8742515
```

c) La probabilidad de que al elegir un individuo al azar de la muestra este sea arbusto o palma, dado que no es maduro, es

$$P(Arbusto \cup Palma|No\ es\ maduro) = \frac{P\big((Arbusto \cup Palma) \cap No\ es\ maduro\big)}{P(No\ es\ maduro)}$$

En este caso la intersección entre ser arbusto o palma se puede obtener bajo la suposición de que ser arbusto o palma son eventos independientes, por tanto,

Luego, la probabildad

```
Pb <- (PArbustoJuvenil+PPalmaJuvenil-PArbustoJuvenil*PPalmaJuvenil)/(1-PM
aduro) ;
Pb

## [1] 0.3583333
```

4) Análisis de la regresión lineal.

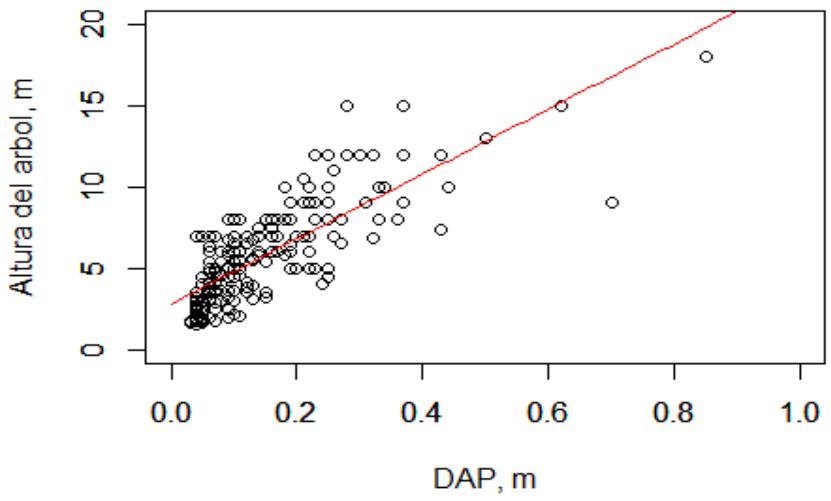
a) Hacemos el modelo

```
mod<-lm(censo$ALTURA.ARBOL~censo$DAP)
```

Y para observar el arreglo tenemos

```
plot(censo$DAP,censo$ALTURA.ARBOL,xlim=c(0,1),ylim=c(0,20),xlab="",ylab="
")

par(new=TRUE)
```

```
fun = function (x) mod$coefficients[1]+mod$coefficients[2]*x
plot(fun,from=0,to=1,xlim=c(0,1),ylim=c(0,20),ylab="Altura del arbol, m",
xlab="DAP, m",col="red")
```



b) Con el resumen del ajuste lineal de los datos podemos identificar un valor de r^2 de 0.6258.

```
summary(mod)

##
## Call:
## lm(formula = censo$ALTURA.ARBOL ~ censo$DAP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7862 -1.4254 -0.1269  1.3760  6.5858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8329     0.2091   13.55   <2e-16 ***
## censo$DAP    19.9333     1.0910   18.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.879 on 198 degrees of freedom
## Multiple R-squared:  0.6277, Adjusted R-squared:  0.6258
## F-statistic: 333.8 on 1 and 198 DF,  p-value: < 2.2e-16
```

Con base en que este valor oscila entre 0 y 1, donde 0 establece que no hay relación alguna entre el modelo y los datos y, por el contrario, 1 será el valor donde el modelo predice todos los datos empleados para el ajuste. El resultado obtenido de R^2, sugiere que sí existe una relación lineal entre los datos. Sin embargo, incluso si este valor se

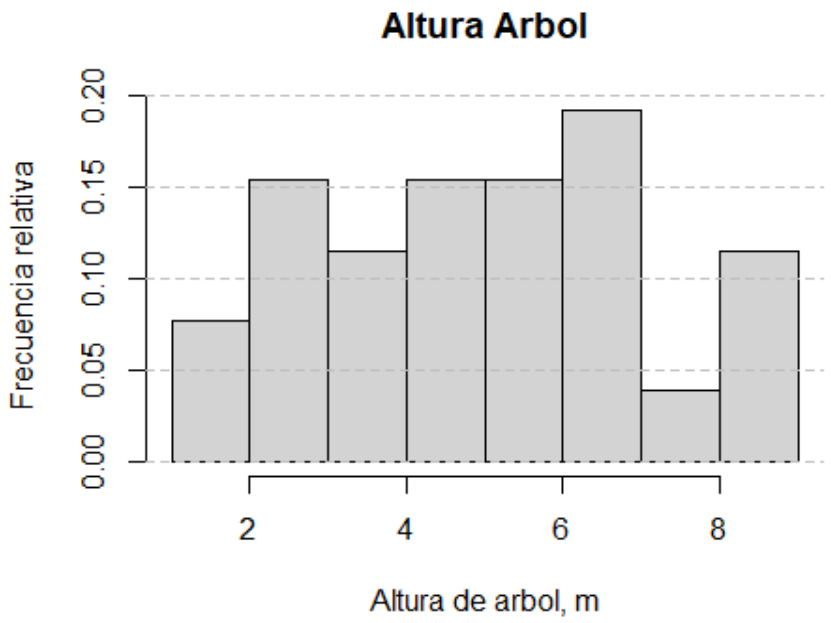
encuentra por encima de 0.5, resulta evidente que hay un gran margen de datos que no son representados por el modelo, en especial por la alta varianza que estos datos tienen.

El ajuste realizado y el modelo lineal empleado establece una relación acertada entre DAP y la altura de los árboles, no obstante, es importante resaltar que, dada la naturaleza de las variables analizadas, los datos presentan una variación muy elevada. Por tanto, es indispensable tener en cuenta la desviación para las variables del modelo, este dato permite establecer con precisión un intervalo en el que se encontrarían los valores de las variables de estudio.

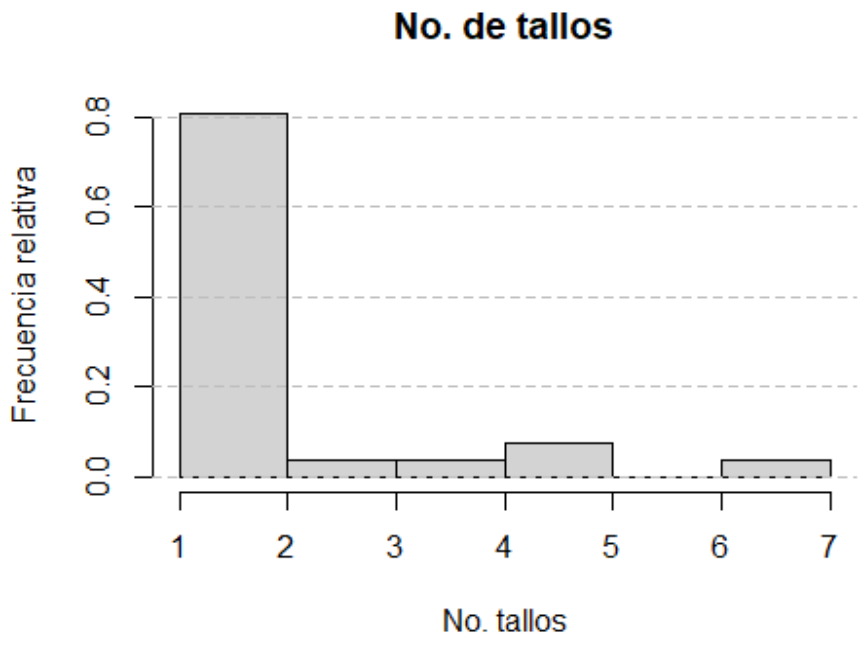
5. Construcción de la distribución.

La variable nominal elegida serán el tipo de vegetación y, por tanto, se examinará la tercera categoría asociada a la palma. La variable de esta categoría podría ser DAP, altura árbol o incluso el No. de tallos.

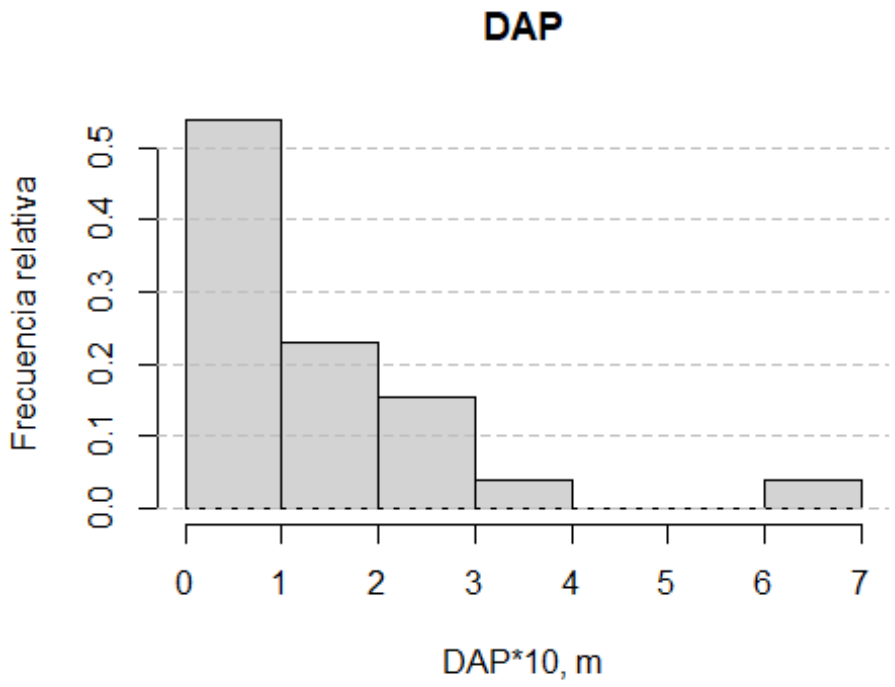
```
DatosPalma =  filter(censo, censo$TIPO.DE.VEGETACIÓN == "Palma")

hist(DatosPalma$ALTURA.ARBOL, prob = TRUE,xlab = "Altura de arbol, m", yl
ab = "Frecuencia relativa", main="Altura Arbol")
grid(nx = NA, ny = NULL, lty = 2, col = "gray", lwd = 1)
```




```
hist(DatosPalma$N..DE.TALLOS, prob = TRUE,xlab = "No. tallos", ylab = "Fr
ecuencia relativa",main="No. de tallos")
grid(nx = NA, ny = NULL, lty = 2, col = "gray", lwd = 1)
```

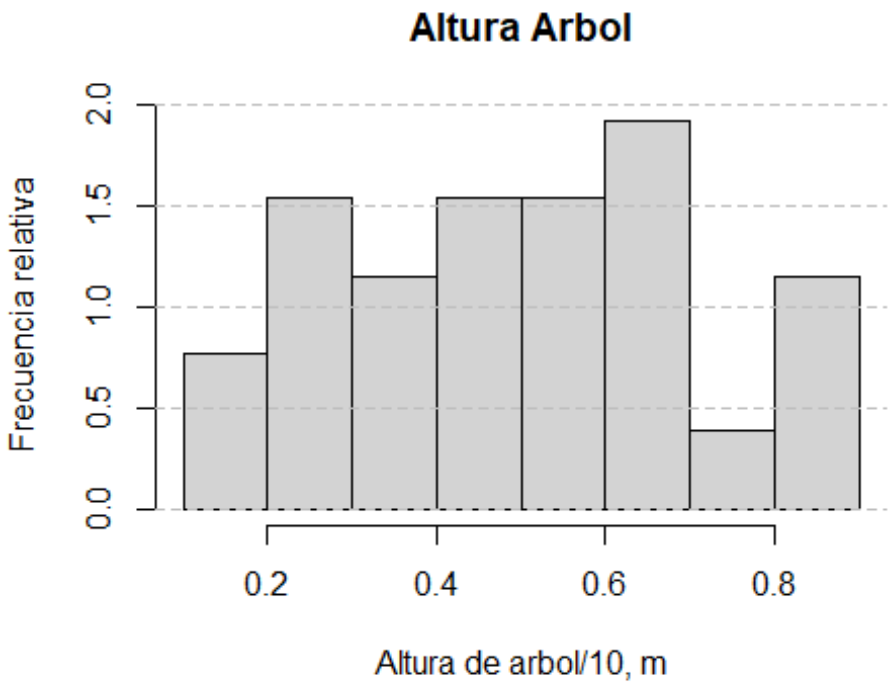



```
hist(DatosPalma$DAP*10, freq = FALSE,xlab = "DAP*10, m", ylab = "Frecuenc
ia relativa", main="DAP")
grid(nx = NA, ny = NULL, lty = 2, col = "gray", lwd = 1)
```

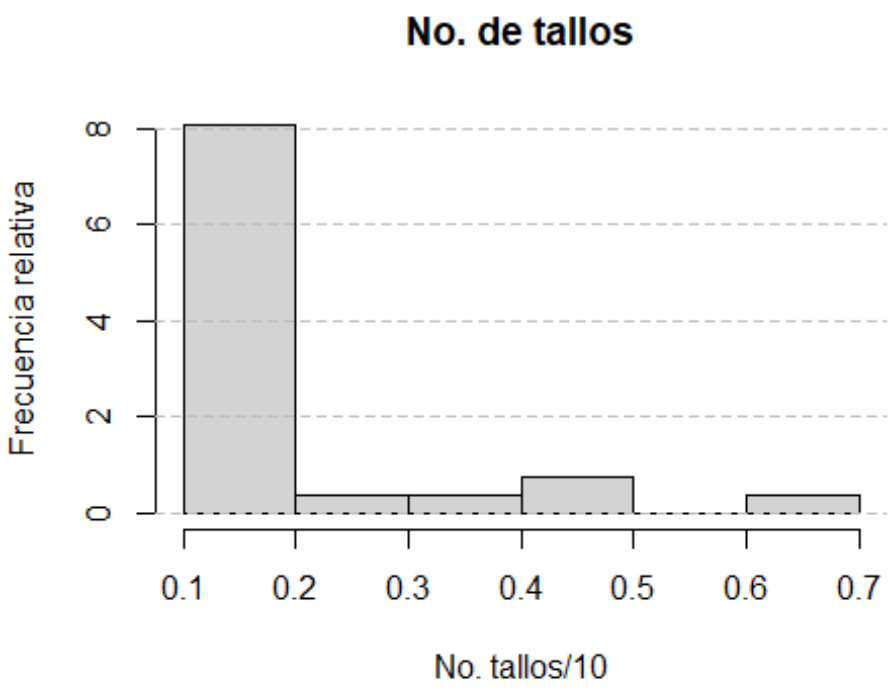


Estos histogramas están dados para el censo original, donde se tiene un total de 26 individuos. Si deseamos que estos resultados sean para 10 tendremos que multiplicar las frecuencias relativas por el valor de los individuos a tomar, encontrando las siguientes distribuciones para cada variable

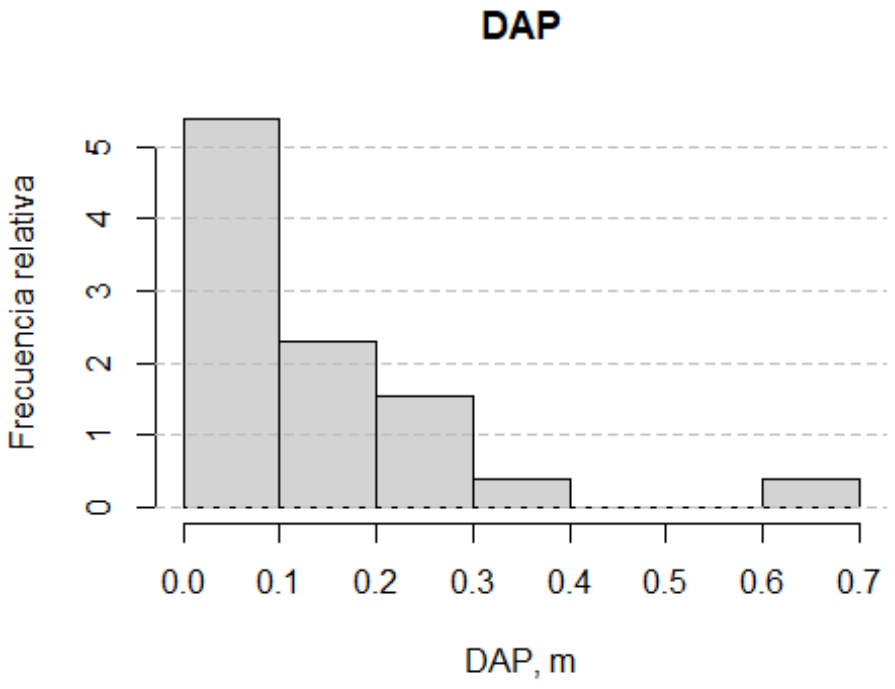
```
DatosPalma =  filter(censo, censo$TIPO.DE.VEGETACIÓN == "Palma")

hist(DatosPalma$ALTURA.ARBOL/10, prob = TRUE,xlab = "Altura de arbol/10,
m", ylab = "Frecuencia relativa", main="Altura Arbol")
grid(nx = NA, ny = NULL, lty = 2, col = "gray", lwd = 1)
```




```
hist(DatosPalma$N..DE.TALLOS/10, prob = TRUE,xlab = "No. tallos/10", ylab
= "Frecuencia relativa",main="No. de tallos")
grid(nx = NA, ny = NULL, lty = 2, col = "gray", lwd = 1)
```




```
hist(DatosPalma$DAP, freq = FALSE,xlab = "DAP, m", ylab = "Frecuencia rel
ativa", main="DAP")
grid(nx = NA, ny = NULL, lty = 2, col = "gray", lwd = 1)
```



Conclusiones:

- Con los resultados obtenidos en el censo, es posible evidenciar que escogiendo un individuo al azar se tiene una mayor probabilidad para que este se encuentre entre árbol o arbusto. De igual forma, las características menos favorecidas en términos probabilísticos, al menos con los resultados del censo, son las condiciones de planta arbustiva y palma. Por otro lado, en la muestra estudiada hay una mayor proporción de especies en una edad madura, llegando a representar más del 80%.
- El ajuste realizado con el modelo lineal empleado para predecir la altura de los árboles en función del DAP, obtuvo un de $R^2 \sim 0.6$. Este dato sugiere que la varianza del modelo lineal representa alrededor del 60% de la varianza de los datos; desentendiendo del escenario y el objetivo de cálculo, esta regresión puede ser útil para predecir algunos datos o al menos una tendencia, sin embargo, es importante resaltar que dada la naturaleza de las variables analizadas, los datos presentan una variación muy elevada y, por tanto, es indispensable tener en cuenta la desviación de las variables del modelo para aumentar la probabilidad de encontrar un valor acertado en un intervalo de posibles soluciones de las variables de estudio.
- Con la distribución discreta de la muestra analizada se puede observar que si se eligen de manera aleatoria 10 palmas se tiene poco más del 50% de probabilidad de que su DAB se encuentre entre 0-0.1 y casi el 80% de que tenga 1-2 tallos. No obstante, la altura del árbol tiene una distribución cuasi uniforme, almenos para los valores comprendidos entre 2-7m, que tienen un poco más de frecuencia relativa en comparación con el resto de los datos.

Bibliografia

Alcandía Santiago de Cali. (Octubre de 2013). *Censo arbóreo de Santiago de Cali*. Obtenido de http://datos.cali.gov.co/dataset/censo-arboreo-de-santiago-de-cali