Práctica 0. FMAD 2021-2022

ICAI. Master en Big Data. Fundamentos Matemáticos del Análisis de Datos (FMAD).

Monsalve Rodilla, Ignacio

Curso 2021-22. Última actualización: 2021-09-24

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.4 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.1 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

# EJERCICIO I

## APARTADO I

Se ofrece una tabla de densidad de probabilidad de la variable aleatoria X1.

Creamos un vector llamado valor y otro llamado probabilidad

valor = c(0,1,2,3)  
  
prob = c(64/125, 48/125, 12/125, 1/125)

La media se calcula simplemente multiplicando los valores por sus respectivas probabilidades:

media = valor %\*% prob  
(media = media[1,1])

## [1] 0.6

La varianza teórica se puede calcular de la siguiente forma:

sum((valor-media)^2\*prob)

## [1] 0.48

## APARTADO 2

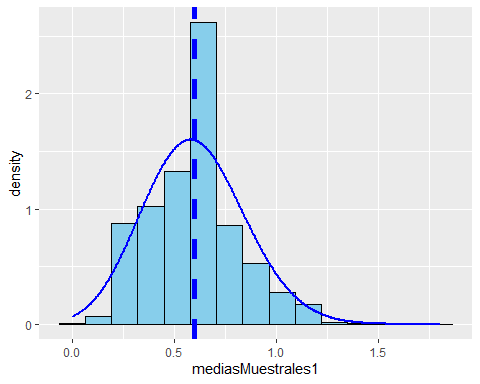
valor = c(0,1,2,3)  
  
prob = c(64/125, 48/125, 12/125, 1/125)

Se combinan ‘sample’ y ‘replicate’ para simular 100.000 muestras de tamaño 10 y estudiar la distribución de las medias muestrales.

k = 100000  
n1 = 10   
mediasMuestrales1 = replicate (k, {  
 muestra = sample(valor, n1, replace = TRUE, prob = prob)  
 mean(muestra)  
})

ggplot() +   
 geom\_histogram(aes(x = mediasMuestrales1, y=..density..), bins = 15, fill="skyblue", color="black") +   
 geom\_vline(xintercept = mean(mediasMuestrales1),   
 col="blue", linetype="dashed", size=2)+  
 geom\_density(mapping = aes(mediasMuestrales1), adjust=6,color="blue",size=1, adjust=1.5)

## Warning: Duplicated aesthetics after name standardisation: adjust



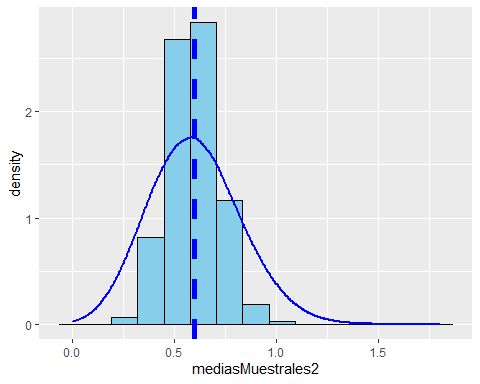
Se repite con tamaño 30 en vez de 10, por lo que es de esperar que la acentuación del histograma hacia la normal sea mayor que antes.

k = 100000  
n1 = 30  
mediasMuestrales2 = replicate (k, {  
 muestra = sample(valor, n1, replace = TRUE, prob = prob)  
 mean(muestra)  
})

De nuevo, se puede representar con ggplot:

ggplot() +   
 geom\_histogram(aes(x = mediasMuestrales2, y=..density..), bins = 15, fill="skyblue", color="black") +   
 geom\_vline(xintercept = mean(mediasMuestrales1),   
 col="blue", linetype="dashed", size=2)+  
 geom\_density(mapping = aes(mediasMuestrales1), adjust=3,color="blue",size=1, adjust=1.5)

## Warning: Duplicated aesthetics after name standardisation: adjust

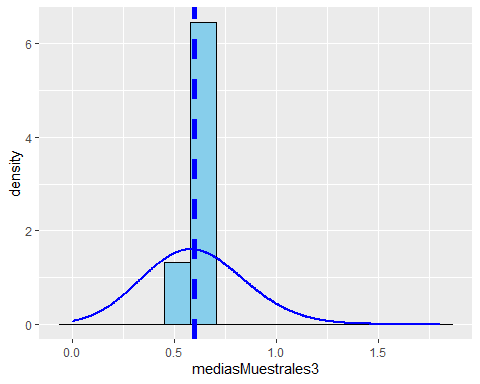


No se pide, pero por reforzar la idea, se incluye también un tamaño de muestra igual a 1000, y se puede observar como se ve claramente la normal

k = 10000  
n1 = 1000  
mediasMuestrales3 = replicate (k, {  
 muestra = sample(valor, n1, replace = TRUE, prob = prob)  
 mean(muestra)  
})

ggplot() +   
 geom\_histogram(aes(x = mediasMuestrales3, y=..density..), bins = 15, fill="skyblue", color="black") +   
 geom\_vline(xintercept = mean(mediasMuestrales1),   
 col="blue", linetype="dashed", size=2)+  
 geom\_density(mapping = aes(mediasMuestrales1), adjust=6,color="blue",size=1, adjust=1.5)

## Warning: Duplicated aesthetics after name standardisation: adjust



## APARTADO III

La variable aleatoria discreta X2 tiene otra tabla de probabilidad, por lo que se guardan dos nuevos vectores, nombrados como valor2 y prob2, y también se vuelven a incluir los del apartado I ya que se necesitarán y está bien recordarlos.

valor = c(0,1,2,3)  
  
prob = c(64/125, 48/125, 12/125, 1/125)  
  
  
  
  
valor2 = c(0,1,2)  
  
prob2 = c(1/2, 1/4, 1/4)

Suponiendo que X1 y X2 son independientes:

**¿Qué valores puede tomar la suma X1+X2?**

Los valores que puede tomar son los comprendidos entre 0 y 5

**¿Cual es tu tabla de probabilidad?**

Para la tabla de probabilidad se utilizan bucles for y se multiplican las distintas probabilidades de X1 y X2 para cada valor que pueden tomar.

probf <- vector(length=6)  
for(i in seq(0,5,1)){  
 for(j in seq(0,3,1)){  
 for(k in seq(0,2,1)){  
 if((j+k)==i){  
 probf[i+1] <- probf[i+1] + (prob[j+1]\*prob2[k+1])  
 }  
 }  
 }  
}  
probf

## [1] 0.256 0.320 0.272 0.124 0.026 0.002

sum(probf)

## [1] 1

## APARTADO IV

La media teórica de la suma se calcula :

media\_ej\_tot = valor%\*%prob + valor2%\*%prob2  
media\_ej\_tot[1,1]

## [1] 1.35

También podría hacerse de esta otra forma, es lo mismo.

a = seq(0,5,1)  
a

## [1] 0 1 2 3 4 5

probf

## [1] 0.256 0.320 0.272 0.124 0.026 0.002

media\_ej = a%\*%probf  
media\_ej[1,1]

## [1] 1.35

A continuación, se simulan cien mil valores:

k = 100000   
mediasMuestrales4 = replicate(k, {  
X1 = sample(0:3, size = 1, replace = TRUE, prob = c(64, 48, 12, 1))  
X2 = sample(0:2, size = 1, replace = TRUE, prob = c(1/2, 1/4, 1/4))  
mean(X1)+mean(X2)  
})

(medias\_ambas\_variables = mean(mediasMuestrales4))

## [1] 1.34359

## ENUNCIADO II

En primer lugar, se descarga el fichero del enlace y se guarda en una variable llamada ‘datos2’

library(tidyverse)

datos2 = read\_csv('testResults.csv')

## Rows: 200 Columns: 9

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): name, gender\_age  
## dbl (7): id, test\_number, week1, week2, week3, week4, week5

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View(datos2)

Se puede ver como la información de gender y age se encuentra en la misma tabla, separada por un \_, por lo que no se cumplen los principios ‘tidy’. Lo que se quiere conseguir es de esa columna, obtener dos columnas en las que se pueda diferenciar por una parte la variable ‘gender’ y por otra ‘age’. Esto se realiza con separate.

datosLimpios <- datos2 %>%   
 separate(gender\_age, into = c("gender", "age"), sep = "\_", convert = TRUE) %>%   
 pivot\_longer(c("week1", "week2", "week3", "week4", "week5"), names\_to = "Type of week",   
 values\_to = "value")

También se utiliza “pivot\_longer” ya que las distintas semanas aparecen en diferentes columnas, y nos gustaría tener todo ello en una única columna que se nombra como ‘Type of week’

View(datosLimpios)

# EJERCICIO III

## EJERCICIO 2 (sección 7.5.1.1 de R4DS)

**¿Qué variables en el dataset Diamonds son las más importantes para predecir el precio del diamante?**

En primer lugar, se deberían escoger aquellas variables que pueden ser posibles candidatas.

Como primer análisis exploratorio, se mirará por encima el dataset.

View(diamonds)

En este caso existe un número de columnas razonable, y será más o menos sencillo limitar la búsqueda, sin embargo en el mundo real podemos encontrar datasets con muchas más columnas y será también importante el conocimiento que se tenga sobre el tema para poder descartar algunas variables que se sepa que no influyen tanto y ahorrarse mucho tiempo y coste computacional.

También puede ocurrir que no se sepa qué es exactamente cada columna, y una de las ventajas de trabajar con este tipo de datasets es que está muy bien organizado y estandarizado, por lo que si se utiliza el comando ?diamonds, nos dará una breve descripción del asunto.

?diamonds

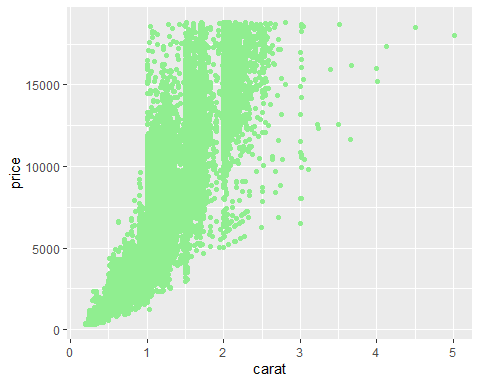
## starting httpd help server ... done

Se puede ver una pequeña descripción en la que se incluye, por ejemplo: carat: peso del diamante cut: calidad del diamante desde ‘Normal’ hasta ‘Ideal’ color: una escala en letras desde D(lo mejor) hasta J (lo peor) clarity: mide cómo de transparente es x, y, z: son coordinadas, que de alguna forma se puede interpretar que se encuentran “agrupadas” dentro de la variable ‘carat’

En primer lugar, se debe realizar una distinción entre variables continuas y otro tipo de variables como pueden ser factores.

Para la variable carat, por ejemplo, que es el peso del diamante, se puede realizar un gráfico de puntos:

ggplot(diamonds, aes(x = carat, y = price)) +  
 geom\_point(colour = 'lightgreen')



Para terminar de cuantificar el análisis, podría realizarse una regresión lineal.

regresion <- lm(carat ~ price, data = diamonds)  
summary(regresion)

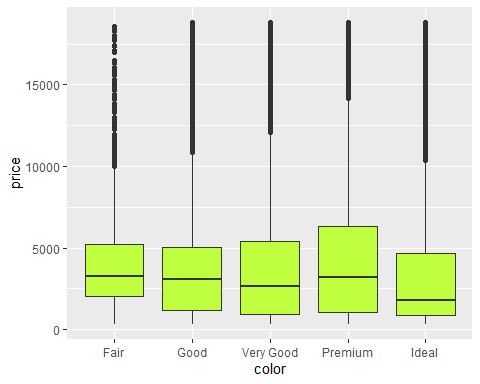
##   
## Call:  
## lm(formula = carat ~ price, data = diamonds)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.35765 -0.11329 -0.02442 0.10344 2.66973   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.673e-01 1.112e-03 330.2 <2e-16 \*\*\*  
## price 1.095e-04 1.986e-07 551.4 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.184 on 53938 degrees of freedom  
## Multiple R-squared: 0.8493, Adjusted R-squared: 0.8493   
## F-statistic: 3.041e+05 on 1 and 53938 DF, p-value: < 2.2e-16

Se puede ver como el R^2 = 85%, que es bastante elevado.

Esto también se puede realizar integrando un boxplot como se explica en R for Data Scientist.

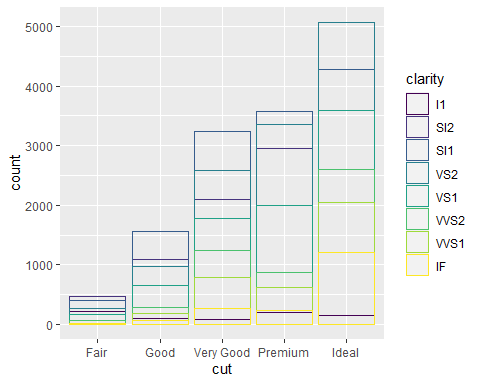
Para mostrar esta opción, se analizará la variable ‘color’ con los boxplot, de forma que el gráfico resulta:

diamonds %>%  
 mutate(color = fct\_rev(color)) %>%  
 ggplot(aes(x = color, y = price)) +  
 geom\_boxplot(mapping = aes(x = cut, y = price), fill = 'olivedrab1')



En el caso de la variable ‘clarity’ ocurre algo similar, en este caso va a representarse con un gráfico distinto:

ggplot(data = diamonds, mapping = aes(x = cut, colour = clarity)) +   
 geom\_bar(fill = NA, position = "identity")



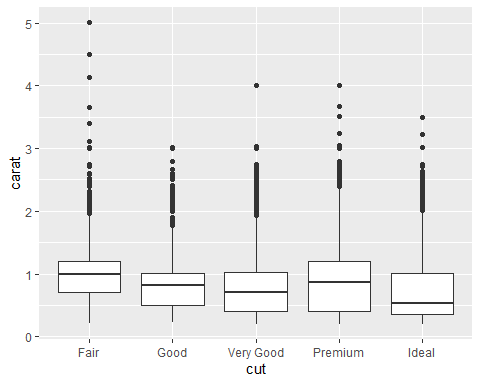
En este caso, el mejor predictor sería la variable carat.

Por ir un paso más allá, se puede realizar también una regresión lineal para cuantificar de alguna forma cómo de bueno sería el modelo.

**¿Cómo está dicha variable relacionada con ‘cut’?**

En esta segunda pregunta, la relación entre ‘cut’ y ‘carat’ se puede representar de la siguiente forma:

ggplot(diamonds, aes(x = cut, y = carat)) +  
 geom\_boxplot()



**¿Por qué la combinación de las dos relaciones hace que los diamantes de peor calidad sean más caros?**

En el gráfico expuesto anteriormente puede verse que existe mucha variabilidad dentro de cada categoría. Se puede ver como en la primera categoría, los valores son más elevados que en resto, lo cual explica el hecho de que el precio sea más caro.

# EJERCICIO 4 (Sección 12.6.1)

Para cado país, año y sexo, computar el número total de casos de TB. Realizar también un gráfico explicativo.

En primer lugar, se van a ver los datos.

library(tidyr)

data(who)  
head(who, 15)

## # A tibble: 15 x 60  
## country iso2 iso3 year new\_sp\_m014 new\_sp\_m1524 new\_sp\_m2534 new\_sp\_m3544  
## <chr> <chr> <chr> <int> <int> <int> <int> <int>  
## 1 Afghanistan AF AFG 1980 NA NA NA NA  
## 2 Afghanistan AF AFG 1981 NA NA NA NA  
## 3 Afghanistan AF AFG 1982 NA NA NA NA  
## 4 Afghanistan AF AFG 1983 NA NA NA NA  
## 5 Afghanistan AF AFG 1984 NA NA NA NA  
## 6 Afghanistan AF AFG 1985 NA NA NA NA  
## 7 Afghanistan AF AFG 1986 NA NA NA NA  
## 8 Afghanistan AF AFG 1987 NA NA NA NA  
## 9 Afghanistan AF AFG 1988 NA NA NA NA  
## 10 Afghanistan AF AFG 1989 NA NA NA NA  
## 11 Afghanistan AF AFG 1990 NA NA NA NA  
## 12 Afghanistan AF AFG 1991 NA NA NA NA  
## 13 Afghanistan AF AFG 1992 NA NA NA NA  
## 14 Afghanistan AF AFG 1993 NA NA NA NA  
## 15 Afghanistan AF AFG 1994 NA NA NA NA  
## # ... with 52 more variables: new\_sp\_m4554 <int>, new\_sp\_m5564 <int>,  
## # new\_sp\_m65 <int>, new\_sp\_f014 <int>, new\_sp\_f1524 <int>,  
## # new\_sp\_f2534 <int>, new\_sp\_f3544 <int>, new\_sp\_f4554 <int>,  
## # new\_sp\_f5564 <int>, new\_sp\_f65 <int>, new\_sn\_m014 <int>,  
## # new\_sn\_m1524 <int>, new\_sn\_m2534 <int>, new\_sn\_m3544 <int>,  
## # new\_sn\_m4554 <int>, new\_sn\_m5564 <int>, new\_sn\_m65 <int>,  
## # new\_sn\_f014 <int>, new\_sn\_f1524 <int>, new\_sn\_f2534 <int>, ...

library(tidyverse)

A continuación, se deben realizar algunos pasos para limpiar y ordenar la tabla

who <- who %>%  
 pivot\_longer(  
 cols = new\_sp\_m014:newrel\_f65,  
 names\_to = "key",  
 values\_to = "cases",  
 values\_drop\_na = TRUE  
) %>%  
mutate(  
 key = stringr::str\_replace(key, "newrel", "new\_rel")  
) %>%  
separate(key, c("new", "var", "sexage")) %>%  
select(-new, -iso2, -iso3) %>%  
separate(sexage, c("sex", "age"), sep = 1)

Para calcular el número total de casos de tuberculosis:

who %>%  
 group\_by(country, year, sex) %>%  
 summarise(cases = sum(cases))

## `summarise()` has grouped output by 'country', 'year'. You can override using the `.groups` argument.

## # A tibble: 6,921 x 4  
## # Groups: country, year [3,484]  
## country year sex cases  
## <chr> <int> <chr> <int>  
## 1 Afghanistan 1997 f 102  
## 2 Afghanistan 1997 m 26  
## 3 Afghanistan 1998 f 1207  
## 4 Afghanistan 1998 m 571  
## 5 Afghanistan 1999 f 517  
## 6 Afghanistan 1999 m 228  
## 7 Afghanistan 2000 f 1751  
## 8 Afghanistan 2000 m 915  
## 9 Afghanistan 2001 f 3062  
## 10 Afghanistan 2001 m 1577  
## # ... with 6,911 more rows

La representación por gráfico:

# Esta podría ser una opción, pero realmente no queda nada claro, por lo que se realizan otros gráficos.

who %>%  
 group\_by(country, year, sex) %>%  
 summarise(cases = sum(cases)) %>%  
 unite(country\_sex, country, sex, remove = FALSE) %>%  
 ggplot(aes(x = year, y = cases, group = country\_sex, colour = sex)) +  
 geom\_line()

## `summarise()` has grouped output by 'country', 'year'. You can override using the `.groups` argument.

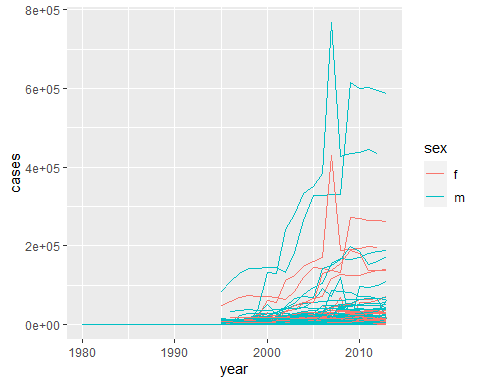


tabla <- who %>%  
 group\_by(country) %>%  
 summarise(cases = sum(cases)) %>%  
 arrange(cases) %>%  
 filter(cases > mean(cases))

View(who)

Un posible gráfico podría ser este:

who %>%  
 group\_by(country, year, sex) %>%  
 filter(year>1995 & country %in% tabla$country) %>%  
 summarise(cases = sum(cases)) %>%  
 ggplot(aes(x = year, y = cases, group = sex, colour = country)) +  
 geom\_line()

## `summarise()` has grouped output by 'country', 'year'. You can override using the `.groups` argument.



Otra opción sería utilizar los gráficos desglosados de diferente forma, como por ejemplo por paises, y utilizar una escala diferente para cada país para poder visualizarlo de una forma correcta.