Análisis estadístico - COVID 19

2025-02-06

El siguiente taller aplicado lo realizaron: Miguel Ángel Peña Cedeño, Gabriel Puello, Valentina Torres Jaimes y Juan Jose Maldonado García.

1. SELECCIÓN Y DESCRIPCIÓN DE LA BASE DE DATOS

La base de datos seleccionada fue de "Casos_positivos_de_COVID-19_en_Colombia._(fecha de descarga).csv" tomada del siguiente link: https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia-/gt2j-8ykr/about_data la hemos seleccionado ya que nos resulto interesante de analizar porque tiene una gran cantidad de observaciones.

Se trata de un dataset realizado por el Instituto Nacional de Salud acerca de los casos positivos de COVID-19 en Colombia, estos datos han sido recopilados desde 2020 y la ultima actualización fue realizada el 14 de agosto de 2024. Son sumamente relevantes en materia de salud para el país y su analisis puede arrojar datos valiosos para preparar mejor a la sociedad colombiana en caso de una futura pandemia ya que este se trata de un fenomeno que eventualmente volverá a ocurrir.

2. IDENTIFICACIÓN Y CLASIFICACIÓN DE LAS VARIABLES

Iniciamos nuestro analisis cargando las 3 librerias que usaremos a lo largo del documento, las dos primeras vistas en clase y la tercera nos ha resultado de utilidad para poder generar graficos mas amigables a la vista al momento de tener que representar valores grandes en variables. Adicionalmente vamos a crear una variable "db" que va a permitir trabajar mas facil con la base de datos y va a leer de un archivo en formato .csv, por ultimo, realizamos str(db) para empezar a ver de que forma está estructurado el dataset.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(scales)

db <- read.csv("Casos_positivos_de_COVID-19_en_Colombia._20250210.csv")
str(db)
```

```
## 'data.frame':    6390971 obs. of  23 variables:
##  $ fecha.reporte.web           : chr  "2020-12-24 00:00:00" "2020-12-24 00:00:00" "2020-12-24 00:00:0
##  $ ID.de.caso                  : int  1556979 1556980 1556981 1556982 1556983 1556984 1556985 155698
##  $ Fecha.de.notificación       : chr  "2020-12-22 00:00:00" "2020-12-19 00:00:00" "2020-12-19 00:00:0
##  $ Código.DIVIPOLA.departamento: int  76 76 76 76 76 76 76 76 76 76 ...
```

```
##  $ Nombre.departamento         : chr  "VALLE" "VALLE" "VALLE" "VALLE" ...
##  $ Código.DIVIPOLA.municipio   : int  76001 76001 76001 76001 76001 76001 76001 76001 76001 76001 ..
##  $ Nombre.municipio            : chr  "CALI" "CALI" "CALI" "CALI" ...
##  $ Edad                        : int  67 66 68 74 65 66 74 66 64 65 ...
##  $ Unidad.de.medida.de.edad    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Sexo                        : chr  "F" "F" "F" "F" ...
##  $ Tipo.de.contagio            : chr  "Comunitaria" "Comunitaria" "Comunitaria" "Comunitaria" ...
##  $ Ubicación.del.caso          : chr  "Casa" "Casa" "Casa" "Fallecido" ...
##  $ Estado                      : chr  "Leve" "Leve" "Leve" "Fallecido" ...
##  $ Código.ISO.del.país         : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Nombre.del.país             : chr  "" "" "" "" ...
##  $ Recuperado                  : chr  "Recuperado" "Recuperado" "Recuperado" "Fallecido" ...
##  $ Fecha.de.inicio.de.síntomas : chr  "2020-12-21 00:00:00" "2020-12-07 00:00:00" "2020-12-18 00:00:0
##  $ Fecha.de.muerte             : chr  "" "" "" "2020-12-30 00:00:00" ...
##  $ Fecha.de.diagnóstico        : chr  "2020-12-23 00:00:00" "2020-12-23 00:00:00" "2020-12-22 00:00:0
##  $ Fecha.de.recuperación       : chr  "2021-01-04 00:00:00" "2020-12-25 00:00:00" "2021-01-01 00:00:0
##  $ Tipo.de.recuperación        : chr  "Tiempo" "Tiempo" "Tiempo" "" ...
##  $ Pertenencia.étnica          : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ Nombre.del.grupo.étnico     : chr  "" "" "" "" ...
```

Encontramos un total de 23 variables y 6.390.971 observaciones en el dataset, de estas se realizo una selección intencional de las variables que consideramos mas relevantes para el analisis, ya que variables como "ID.de.caso" no resultaban especialmente utiles a la hora del analisis:

a. **fecha.reporte.web**: Cuantitativa – Intervalo

b. **ID.de.caso**: Cualitativa – Nominal

c. **Fecha.de.notificación**: Cuantitativa – Intervalo

d. **Código.DIVIPOLA.departamento**: Cualitativa – Nominal

e. **Nombre.departamento**: Cualitativa – Nominal

f. **Código.DIVIPOLA.municipio**: Cualitativa – Nominal

g. **Nombre.municipio**: Cualitativa – Nominal

h. **Edad**: Cuantitativa – Razón

i. **Unidad.de.medida.de.edad**: Cualitativa – Nominal

j. **Sexo**: Cualitativa – Nominal

k. **Tipo.de.contagio**: Cualitativa – Nominal

l. **Ubicación.del.caso**: Cualitativa – Nominal

m. **Estado**: Cualitativa – Ordinal

n. **Código.ISO.del.país**: Cualitativa – Nominal

o. **Nombre.del.país**: Cualitativa – Nominal

p. **Recuperado**: Cualitativa – Nominal

q. **Fecha.de.inicio.de.síntomas**: Cuantitativa – Intervalo

r. **Fecha.de.muerte**: Cuantitativa – Intervalo

s. **Fecha.de.diagnóstico**: Cuantitativa – Intervalo

t. **Fecha.de.recuperación**: Cuantitativa – Intervalo

u. **Tipo.de.recuperación**: Cualitativa – Nominal

v. **Pertenencia.étnica**: Cualitativa – Nominal

w. **Nombre.del.grupo.étnico**: Cualitativa – Nominal

3. ORGANIZACIÓN Y PRESENTACIÓN DE LOS DATOS

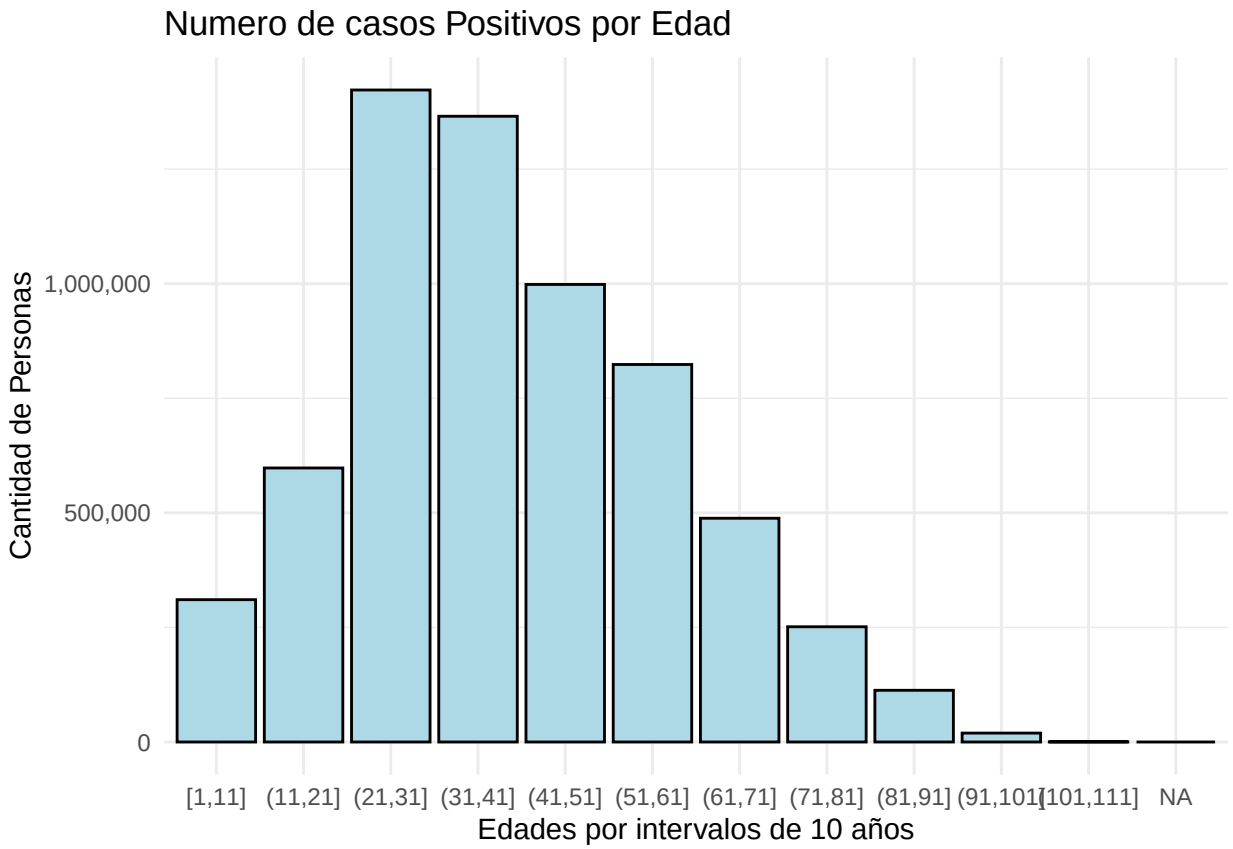
A continuacion hacemos la presentacion de los datos sirviendonos de una tabla de frecuencia que nos muestre la edad en intervalos de 10 años por el numero de casos de infección, esto para no tener que hacer 114 filas distintas en la tabla sino solo 12.

```
db <- read.csv("Casos_positivos_de_COVID-19_en_Colombia._20250210.csv")

bins <- seq(min(db$Edad, na.rm = TRUE), max(db$Edad, na.rm = TRUE), by = 10)

tabla_frecuencia <- db %>%
  mutate(Edad = cut(Edad, breaks = bins, include.lowest = TRUE)) %>%
  group_by(Edad) %>%
  summarise(n_personas = n())
tabla_frecuencia
```

```
## # A tibble: 12 x 2
##    Edad      n_personas
##    <fct>          <int>
##  1 [1,11]        310534
##  2 (11,21]       597888
##  3 (21,31]      1422643
##  4 (31,41]      1365345
##  5 (41,51]       998354
##  6 (51,61]       823714
##  7 (61,71]       488231
##  8 (71,81]       251461
##  9 (81,91]       112843
## 10 (91,101]       19463
## 11 (101,111]        487
## 12 <NA>               8
```

A continuacion vamos a visualizar la misma tabla de frecuencia haciendo uso de un grafico de barras, aquí es donde nos resulto util la inclusion de la libreria (scales) en los valores del eje y.

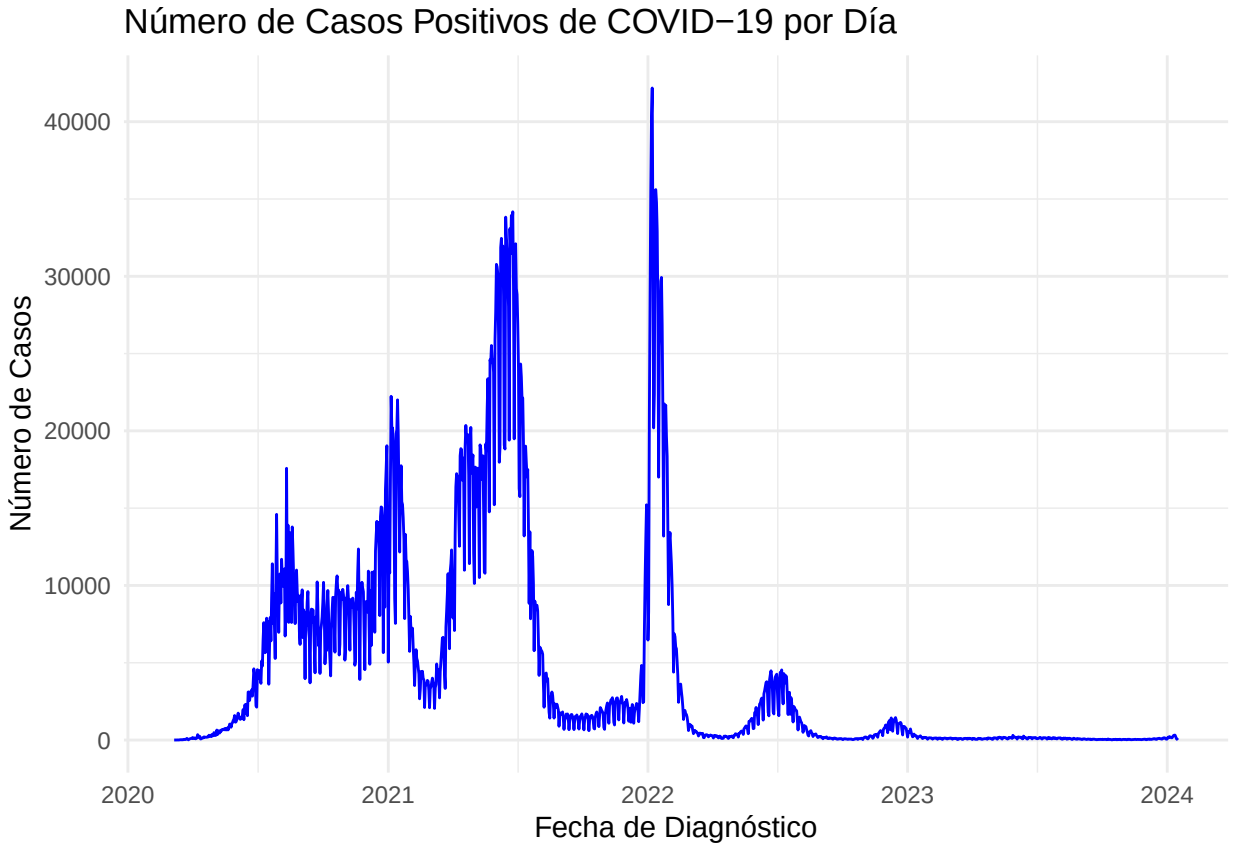
```
ggplot(tabla_frecuencia, aes(x = Edad, y = n_personas)) +
  geom_bar(stat = "identity", fill = "lightblue", color = "black", na.rm = TRUE) +
  labs(title = "Numero de casos Positivos por Edad", x = "Edades por intervalos de 10 años", y = "Canti
  scale_y_continuous(labels = comma) +
  theme_minimal()
```



Podemos observar como los intervalos con mayor numero de personas infectadas son los que van desde los 21 hasta los 61 años.

Para continuar con la presentacion de los datos quisimos agrupar el numero de casos de infección reportados por las fechas, para esto agrupamos el numero de casos por la variable de "Fecha.de.diagnóstico"

```
db$Fecha.de.diagnóstico <- as.Date(db$Fecha.de.diagnóstico, format = "%Y-%m-%d")

casos_por_dia <- db %>%
  group_by(Fecha.de.diagnóstico) %>%
  summarise(n_casos = n())

ggplot(casos_por_dia, aes(x = Fecha.de.diagnóstico, y = n_casos)) +
  geom_line(color = "blue", na.rm = TRUE) +
  labs(title = "Número de Casos Positivos de COVID-19 por Día",
       x = "Fecha de Diagnóstico",
       y = "Número de Casos") +
  theme_minimal()
```



Gracias a la graficación de este dataframe nos es visible los momentos donde se presentaron los picos de infección, podriamos establecer 3 principales que serían: a inicios de 2021, a mitades de 2021 y finalmente a inicios de 2022.

4. MEDIDAS DE TENDENCIA CENTRAL

A continuacion vamos a calcular las medidas de tendencia central para la variable de edad ya que es la única variable cuantitativa de razón en el dataset, para poder llegar a la moda nos vamos a ayudar de la tabla de frecuencias de la edad realizada anteriormente, en ella encontramos que:

La clase modal (con mayor frecuencia) es el intervalo: [21,31] con frecuencia = 1,422,643

Dado que esa es la clase modal sacamos los datos:

L = 21 = límite inferior del intervalo modal fm = 1,422,643 = frecuencia de la clase modal f_anterior = 597,888 = frecuencia de la clase anterior (11,21] f_siguiente = 1,365,345 = frecuencia de la clase siguiente [31,41] c = 10 = tamaño del intervalo

```
mean(db$Edad)
```

```
## [1] 39.98171
```

```
median(db$Edad)
```

```
## [1] 38
```

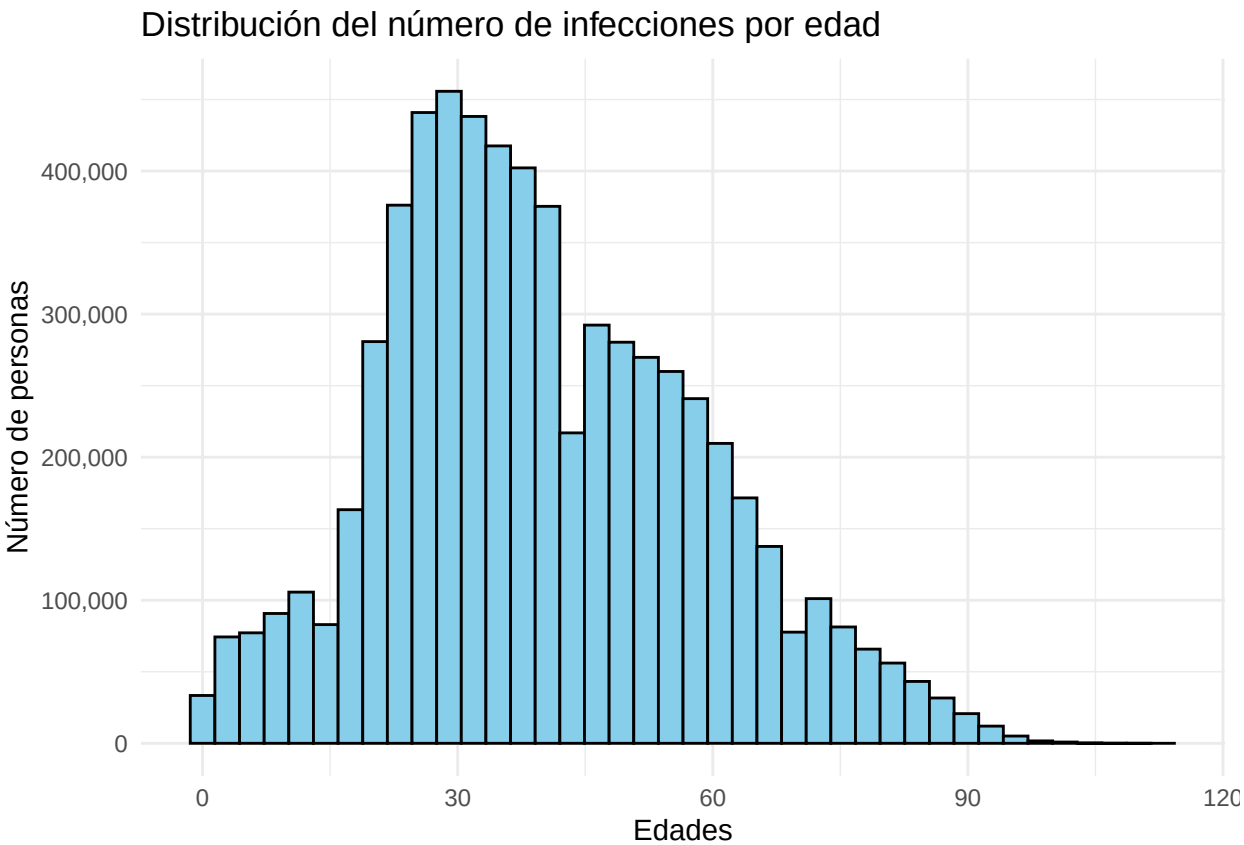
```
L <- 21
fm <- 1422643
f_anterior <- 597888
f_siguiente <- 1365345
c <- 10
```

```r
d1 <- fm - f_anterior
d2 <- fm - f_siguiente

moda <- L + (d1 / (d1 + d2)) * c

print(moda)
```

```
## [1] 30.3504
```

Encontramos que el promedio de edad de los casos positivos es de aproximadamente 40 años, el dato que mas se repite en el dataset es 30 años y la mediana es de 38 años, lo que sugiere una distribución sesgada ya que, aunque los valores no estan muy distanciados en el caso de la media y la mediana, deben ser iguales para sugerir que la distribucion es simétrica. Vamos a graficar la edad sin intervalos y con mas columnas para observar de que forma se distribuyen los datos:

```r
ggplot(db, aes(x = Edad)) +
  geom_histogram(bins = 40, color = "black", fill = "skyblue", na.rm = TRUE) +
  labs(title = "Distribución del número de infecciones por edad",
    x = "Edades", y = "Número de personas") +
    scale_y_continuous(labels = comma) +
  theme_minimal()
```



La distribución tiene forma asimétrica hacia la derecha (cola larga en edades mayores), lo que significa que la mayoría de los contagios se concentran en edades más jóvenes o medias. Se observa un pico entre los 25 y 40 años, lo que indica que esa es la franja etaria con mayor cantidad de contagios registrados. Después de los 60 años, la frecuencia comienza a disminuir gradualmente, aunque siguen existiendo contagios hasta los 90 años o más.

5. MEDIDAS DE VARIABILIDAD

Ahora vamos a proceder con calcular las medidas de variabilidad:

```
rango <- max(db$Edad) - min(db$Edad)
rango
```

```
## [1] 113
```

```
sd(db$Edad)
```

```
## [1] 18.47426
```

```
var(db$Edad)
```

```
## [1] 341.2984
```

Encontramos que el rango de edades es de 113 años, lo cual indica una población muy diversa en cuanto a edad, desde recién nacidos hasta personas mayores de 100 años. La desviación estándar de aproximadamente 18.5 años, esto indica que las edades de los casos positivos presentan una alta dispersión respecto a la media, que es de 40 años. Esto sugiere que el COVID-19 afectó a personas de un amplio rango etario, no concentrado en un solo grupo de edad.

Por ultimo, la varianza nos indica cuánto varían las edades respecto al promedio. Aunque sus valores no nos dicen mucho de forma intuitiva como la desviación estándar, una varianza alta como esta refleja una dispersión considerable en los datos, hay personas muy jóvenes y personas muy mayores entre los casos positivos.

6. MEDIDAS DE POSICIÓN

Para las medidas de posición vamos a partir de la variable de edad:

```
quantile(db$Edad, probs = c(0.25, 0.5, 0.75), na.rm = TRUE)
```

```
## 25% 50% 75%
##  27  38  53
```

```
quantile(db$Edad, seq(0.1, 0.9, 0.1), na.rm = TRUE)
```

```
## 10% 20% 30% 40% 50% 60% 70% 80% 90%
##  19  24  29  33  38  43  49  56  65
```

```
quantile(db$Edad, probs = c(0.10, 0.25, 0.50, 0.75, 0.90), na.rm = TRUE)
```

```
## 10% 25% 50% 75% 90%
##  19  27  38  53  65
```

Las medidas de posición nos confirman lo anteriormente mencionado, pues muestran que los casos positivos de COVID-19 se distribuyen principalmente entre los 27 y 53 años. La edad mediana es de 38 años, lo que indica que la mitad de los casos se concentran por debajo de esa edad. Además, el 90% de los casos tiene 65 años o menos, lo cual sugiere que los contagios afectan principalmente a personas adultas jóvenes y de mediana edad.

7. APLICACIÓN DE DISTRIBUCIONES DE PROBABILIDAD

Para aplicar una distribución de probabilidad nos intereso ver cual era la probabilidad en una distribución geometrica de encontrar por primera vez una mujer y a un hombre en la primera y en la sexta observación consecutiva:

```
library(dplyr)
p_mujer <- mean(db$Sexo == "F", na.rm = TRUE)
```

```r
p_hombre <- mean(db$Sexo == "M", na.rm = TRUE)
dgeom(0, prob = p_mujer)
```

```
## [1] 0.5346742
```

```r
dgeom(5, prob = p_mujer)
```

```
## [1] 0.01166474
```

```r
dgeom(0, prob = p_hombre)
```

```
## [1] 0.465324
```

```r
dgeom(5, prob = p_hombre)
```

```
## [1] 0.02033335
```

Mirando los datos concluimos que hay un 53.5% de probabilidad de que la primera persona observada en el dataset sea una mujer. Por otro lado, la probabilidad de que la sexta persona observada sea la primera mujer, y que las cinco anteriores observaciones no lo sean, es apenas del 1.17%. Por otro lado, Hay un 46.5% de probabilidad de que la primera persona que revisemos sea un hombre. y la probabilidad de tener que revisar 6 personas para encontrar al primer hombre es del 2.03%.

8. APLICACIÓN DEL MODELO NORMAL

Para el modelo normal quisimos hacerlo con la variable de edad ya que esta es la que se comporta de forma mas aproximadamente normal, sin embargo hay que mencionar que los calculos realizados abajo parten de la suposición de que la edad en efecto sigue una distribución normal dentro del dataset.

```r
1 - pnorm(40, mean(db$Edad, na.rm = TRUE), sd(db$Edad, na.rm = TRUE))
```

```
## [1] 0.4996051
```

```r
1 - pnorm(65, mean(db$Edad, na.rm = TRUE), sd(db$Edad, na.rm = TRUE))
```

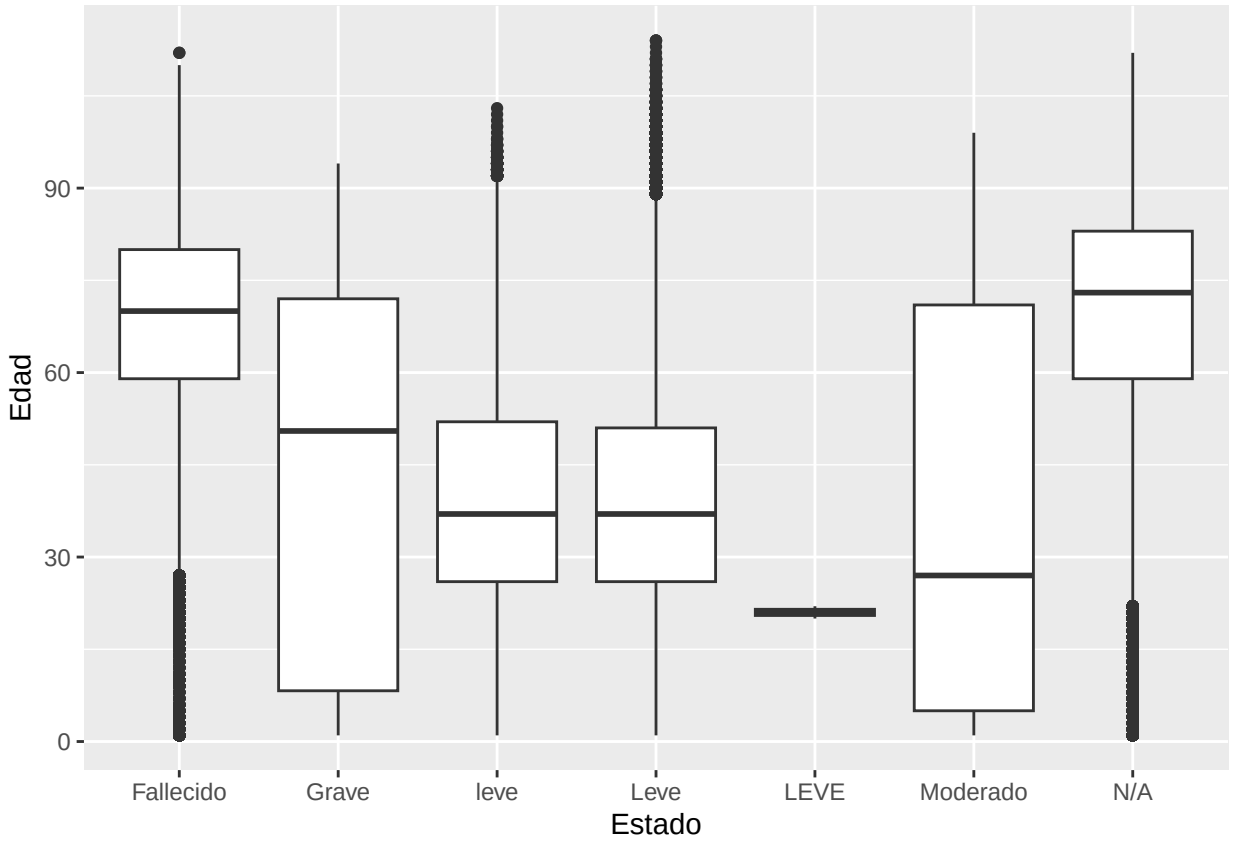
```
## [1] 0.08783248
```

Segun los datos, la probabilidad de que una persona infectada tenga más de 40 años es aproximadamente del 50%. Sin embargo, esto se reduce drásticamente cuando queremos observar cual es la probabilidad de que tenga mas de 65 años, pues pasa a ser del 8.7%

9. BONO (BOXPLOT)

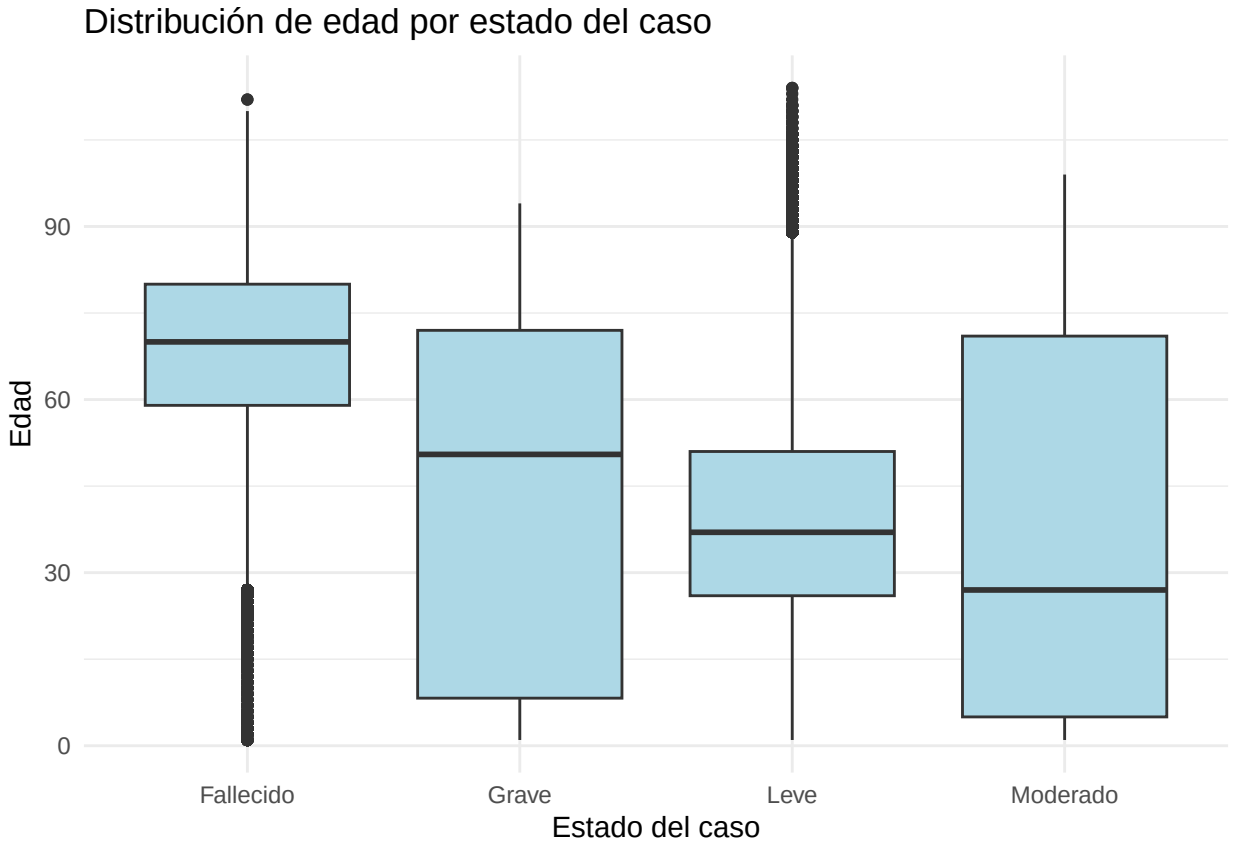
Para el bono hemos querido mirar un boxplot por estado para comparar las edades en el dataset, sin embargo nos hemos encontrado con el problema de que hay tanto valores nulos, como valores que se repiten de forma erronea unicamente porque se escriben con un caracter diferente, como es el caso de las mayusculas entre “leve”, “Leve” y “LEVE”, esto lo podemos observar aquí:

```r
ggplot(db, aes(x = Estado, y = Edad)) + geom_boxplot()
```



Debido a esto, hemos eliminado los valores nulos y agrupado las categorias para que no se repitan y podamos tener una visualización de los datos mejor:

```r
db_limpio <- db %>%
  filter(!is.na(Estado)) %>%
  mutate(Estado = tolower(Estado)) %>%
  mutate(Estado = recode(Estado,
                         "LEVE" = "Leve",
                         "leve" = "Leve",
                         "grave" = "Grave",
                         "moderado" = "Moderado",
                         "fallecido" = "Fallecido")) %>%
  filter(Estado %in% c("Leve", "Grave", "Moderado", "Fallecido"))

ggplot(db_limpio, aes(x = Estado, y = Edad)) +
  geom_boxplot(fill = "lightblue") +
  labs(
    title = "Distribución de edad por estado del caso",
    x = "Estado del caso",
    y = "Edad"
  ) +
  theme_minimal()
```



El gráfico muestra una relación clara entre edad y gravedad del COVID-19. A medida que la edad aumenta, encontramos mas casos de gravedad o fallecimiento. Por el contrario, los casos leves y moderados predominan entre personas más jóvenes. Esto sugiere que la edad es un factor de riesgo importante en la gravedad que pueda tener una infección de COVID-19 en una persona.

10. INTERPRETACIÓN DE RESULTADOS Y CONCLUSIONES

Este análisis nos permitió extraer conclusiones relevantes en torno al comportamiento de la enfermedad según edad, sexo, y estado clínico del paciente. En primer lugar, se identificó que la variable Edad concentra la mayor frecuencia de contagios en el grupo de personas entre los 21 y 40 años, siendo la moda de la distribución aproximadamente 30.35 años. Esto concuerda con los grupos social y laboralmente más activos. Sin embargo, los casos más graves y los fallecimientos se asocian con edades mayores, especialmente por encima de los 65 años, tal como se evidenció en los boxplots y en los valores de mediana por grupo clínico.

Desde el punto de vista de variabilidad, la desviación estándar (~18 años) y el rango de edades (113 años) muestran una alta dispersión, lo cual confirma que el virus afectó a una población muy heterogénea en términos de edad.

Se aplicaron modelos probabilísticos sobre distintas variables. Se aplicó la distribución geométrica para modelar el número de personas a revisar hasta encontrar la primera mujer (o el primer hombre) en la base. Esto permitió explorar probabilidades discretas basadas en variables categóricas, extendiendo el análisis más allá de variables cuantitativas.

Finalmente, se estimaron probabilidades bajo el modelo normal para edades, observando que alrededor del 50% de los casos están por encima de los 40 años, y que solo el 8.7% supera los 65 años. Aunque se asumió normalidad para este análisis, se aclaró que la distribución real de edad es ligeramente asimétrica.

Para finalizar, pensamos que un analisis de este tipo es relevante en materia de salud para mejorar la capacidad clinica del país en casos de pandemias futuras, vimos como habían picos de infección de hasta 40.000 casos diarios y esta es una cifra que puede saturar completamente la capacidad hospitalaria.