Manipulación y Representación de Datos Estadísticos con R

Carlos G.

Manipulación y representación de datos estadísticos con R

Contenido

Manipulación y representación de datos estadísticos con R

3. INTRODUCCIÓN

3.1. Objetivo del trabajo:

El objetivo de este documento es explorar técnicas de manipulación y visualización de datos estadísticos utilizando R. Se busca que el lector adquiera habilidades prácticas para manejar estructuras de datos y generar visualizaciones que permitan interpretar conjuntos de datos complejos de manera efectiva.

3.2. Justificación:

En la ingeniería informática, procesar y visualizar datos es crucial para tomar decisiones basadas en información confiable. R es una herramienta reconocida por su capacidad en análisis estadístico y visualización de datos, brindando soluciones robustas desde la exploración básica hasta modelos predictivos avanzados. Dominar R capacita a los profesionales para enfrentar desafíos en ciencia de datos, análisis estadístico y desarrollo de software enfocado en la gestión y representación de datos.

3.3. Metodología:

Este trabajo se enfoca en el análisis de datos utilizando R, comenzando con conjuntos de datos predefinidos como state.area y discoveries, seleccionados por su accesibilidad y reproducibilidad. Los datos son organizados mediante estructuras nativas de R, tales como vectores y tablas, y procesados mediante funciones integradas, como table() para calcular frecuencias y cumsum() para obtener valores acumulados.Para la visualización, se emplean herramientas básicas como hist() para crear histogramas, barplot() para diagramas de barras y plot() para polígonos de frecuencias. Estas técnicas permiten identificar patrones, tendencias y distribuciones en los datos, facilitando su interpretación. En cuanto al análisis estadístico, se calculan medidas de tendencia central, como la moda, combinando enfoques gráficos y funciones avanzadas del paquete modeest, como mfv(). Esto posibilita comparar los resultados obtenidos por diferentes métodos para evaluar su consistencia y precisión. Finalmente, los resultados se interpretan analizando las visualizaciones y estadísticas generadas, extrayendo conclusiones sobre las características de los datos. Este proceso reflexiona sobre la coherencia y aplicabilidad de los métodos utilizados, destacando las capacidades de R para el análisis y la visualización de datos.

4. DESARROLLO DE LAS ACTIVIDADES

4.1 ACTIVIDAD 1: REPRESENTACIÓN DE LOS DATOS.

***Preámbulo:Histogramas.Ejecuciones extraídos de los ejemplos*:**

Ejecución de sentencias precip, vectores con datos almacenados

```
> ?precip
> precip
```

Mobile	Juneau	Phoenix	Little Rock
67.0	54.7	7.0	48.5
Los Angeles	Sacramento	San Francisco	Denver
14.0	17.2	20.7	13.0
Hartford	Wilmington	Washington	Jacksonville
43.4	40.2	38.9	54.5
Miami	Atlanta	Honolulu	Boise
59.8	48.3	22.9	11.5
Chicago	Peoria	Indianapolis	Des Moines
34.4	35.1	38.7	30.8
Wichita	Louisville	New Orleans	Portland
30.6	43.1	56.8	40.8
Baltimore	Boston	Detroit	Sault Ste. Marie
41.8	42.5	31.0	31.7
Duluth	Minneapolis/St Paul	Jackson	Kansas City
30.2	25.9	49.2	37.0
St Louis	Great Falls	Omaha	Reno
35.9	15.0	30.2	7.2
Concord	Atlantic City	Albuquerque	Albany
36.2	45.5	7.8	33.4
Buffalo	New York	Charlotte	Raleigh
36.1	40.2	42.7	42.5
Bismark	Cincinnati	Cleveland	Columbus
16.2	39.0	35.0	37.0
Oklahoma City	Portland	Philadelphia	Pittsburg
31.4	37.6	39.9	36.2
Providence	Columbia	Sioux Falls	Memphis
42.8	46.4	24.7	49.1
Nashville	Dallas	El Paso	Houston
46.0	35.9	7.8	48.2
Salt Lake City	Burlington	Norfolk	Richmond
15.2	32.5	44.7	42.6
Seattle Tacoma	Spokane	Charleston	Milwaukee
38.8	17.4	40.8	29.1
Cheyenne	San Juan		
14.6	59.2		

Ilustración 1: Lluvia media en 70 ciudades de Norteamérica

Histograma de frecuencias relativas. Ejecución sentencia hist(precip, freq=FALSE) y hist(precip,breaks=10), presentado en siguientes ilustraciones en orden respetivamente.

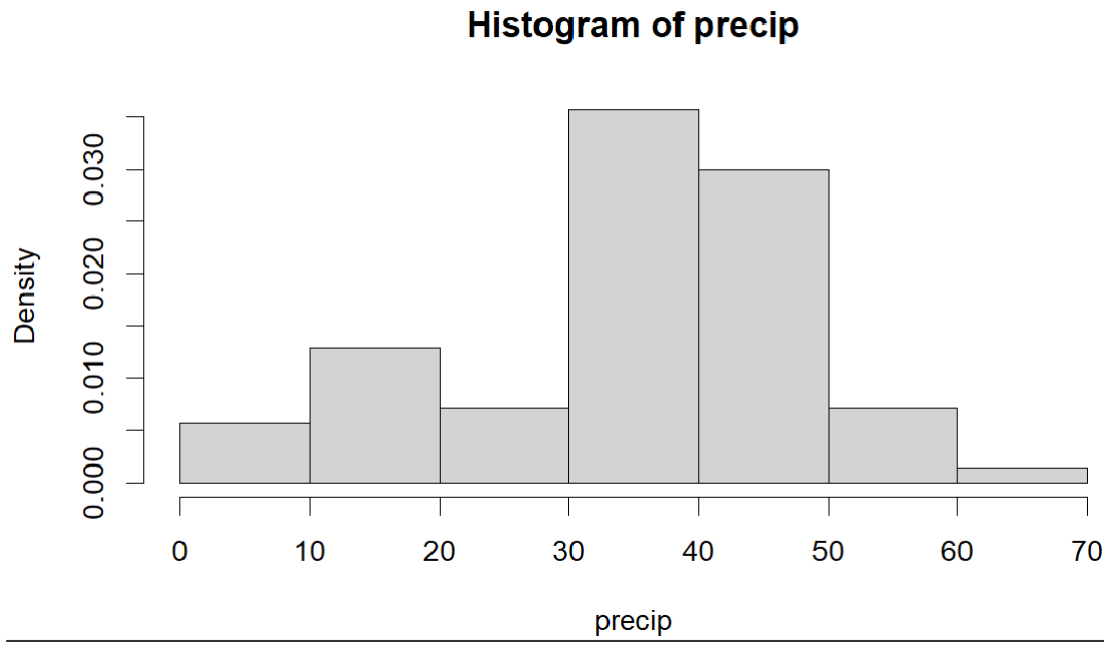


Ilustración 2

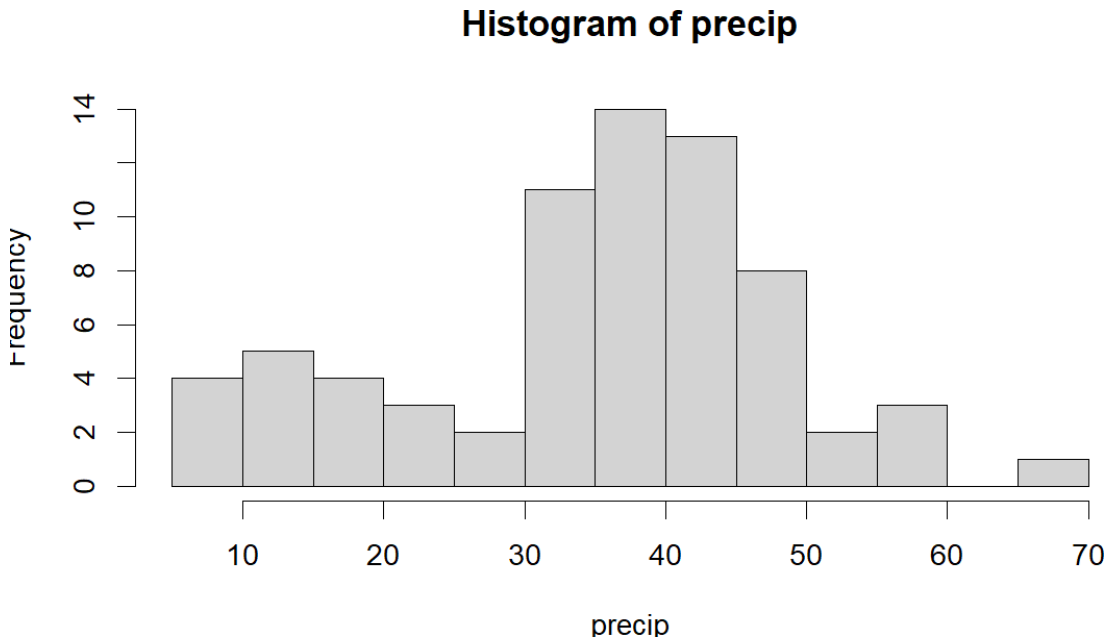


Ilustración 3

Histograma de Diagramas de barras

ventas <- c(345.3, 452.1, 395.6);names(ventas) <- c("Pedro", "Juan", "Maria");

Forma 1: barplot(ventas, main="Ventas España", ylab="Miles de euros");

Forma 2 : pie(ventas, main="Ventas España").

Forma 3: dotchart(ventas,main="Ventas España", xlab="Miles de euros")

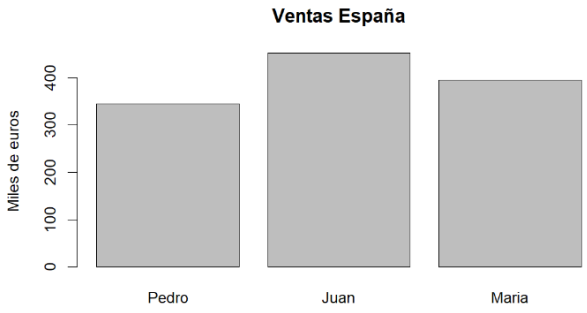


Ilustración 3: Forma 1

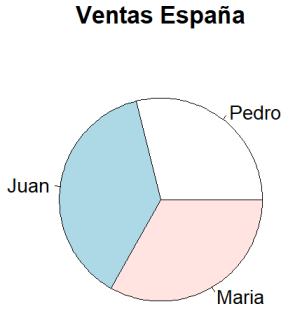


Ilustración 4: Forma 2

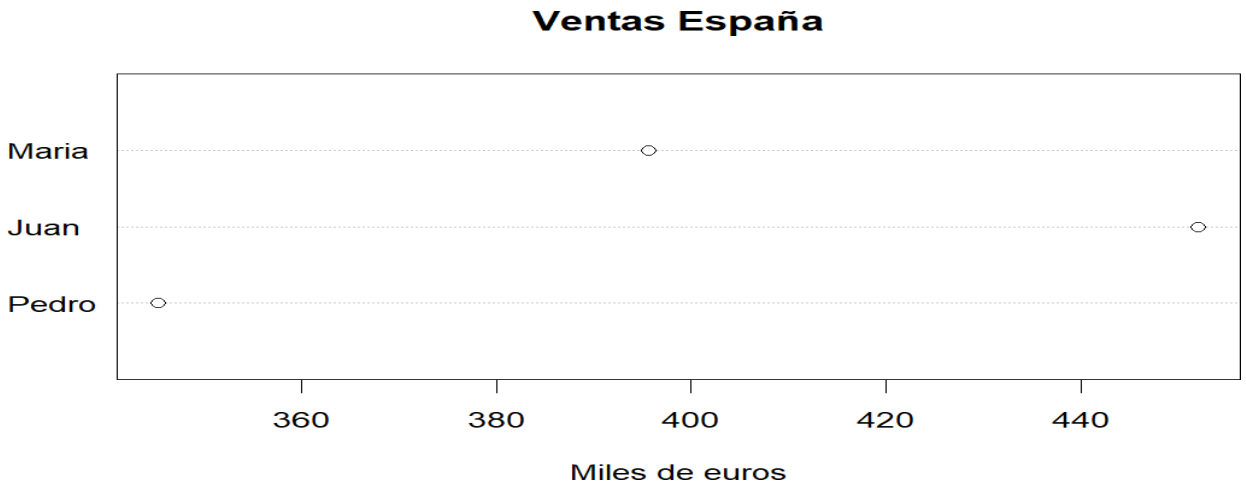


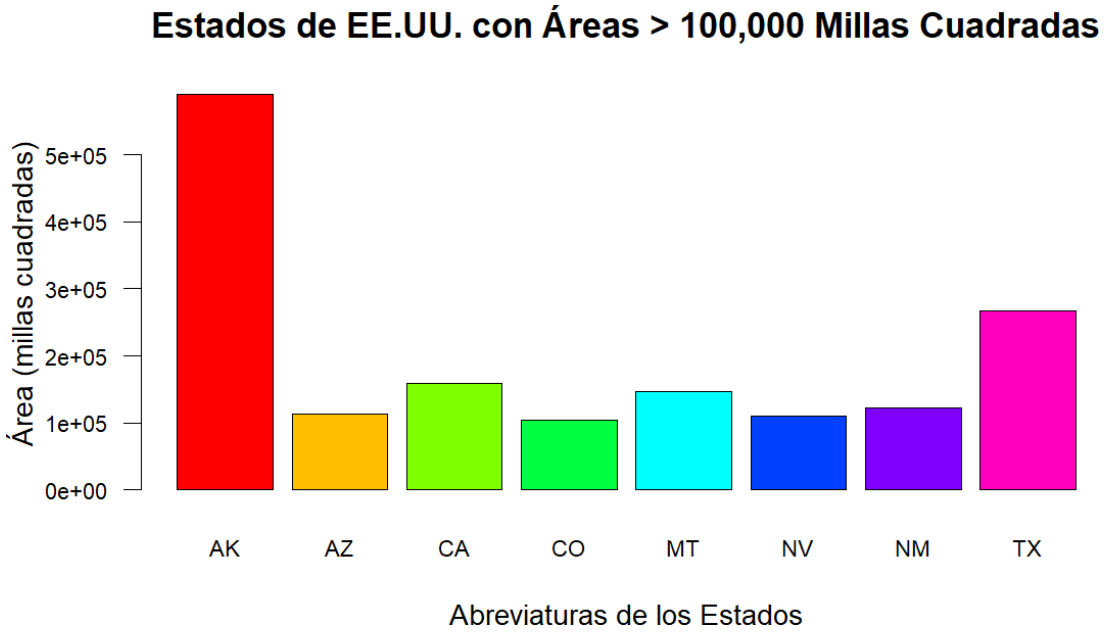
Ilustración 5: Forma 3

4.1 ACTIVIDAD 1: DIAGRAMA DE BARRAS DE ESTADOS CON MÁS DE 100.000 MILLAS CUADRADAS.

SENTENCIAS DE DESARROLLO DEL DIAGRAMA.

```
> # Filtrar los estados con un área mayor a 100,000 millas cuadradas
> estados_filtrados <- state.abb[state.area > 100000]
> areas_filtradas <- state.area[state.area > 100000]
>
> # Crear un diagrama de barras con personalización
> barplot(
+     areas_filtradas,
+     names.arg = estados_filtrados,
+     xlab = "Abreviaturas de los Estados",
+     ylab = "Área (millas cuadradas)",
+     main = "Estados de EE.UU. con Áreas > 100,000 Millas Cuadradas",
+     col = rainbow(length(areas_filtradas)),  # Colores en un espectro arcoíris
+     border = "black",  # Borde negro para las barras
+     horiz = FALSE,  # Barras verticales
+     cex.names = 0.8,  # Tamaño de las etiquetas de los estados
+     cex.axis = 0.8,  # Tamaño de las etiquetas del eje
+     las = 1          # Etiquetas del eje vertical en orientación horizontal
+ )
```

REPRESENTACIÓN DE DIAGRAMA DE BARRAS.



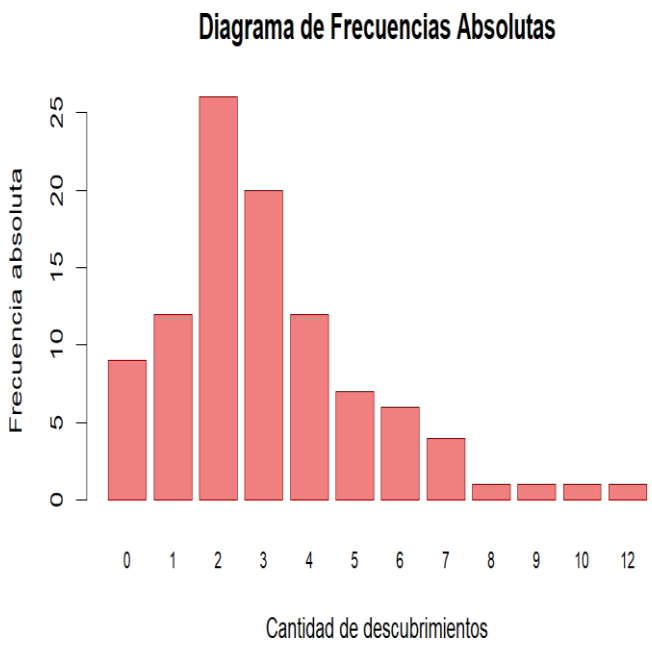
4.2 **ACTIVIDAD 2:** DIAGRAMAS Y POLÍGONOS DE LAS FRECUENCIAS ABSOLUTAS, RELATIVAS Y ACUMULADAS EN DATASET DISCOVERIES.

SENTENCIAS DE DESARROLLO DEL CÁLCULO DE LAS FRECUENCIAS

```
# Cargar el dataset discoveries
data("discoveries")
# Obtener las frecuencias absolutas
frecuencias_absolutas <- table(discoveries)
# Calcular las frecuencias relativas
frecuencias_relativas <- frecuencias_absolutas / sum(frecuencias_absolutas)
# Calcular las frecuencias absolutas acumuladas
frecuencias_absolutas_acum <- cumsum(frecuencias_absolutas)
# Calcular las frecuencias relativas acumuladas
frecuencias_relativas_acum <- cumsum(frecuencias_relativas)
```

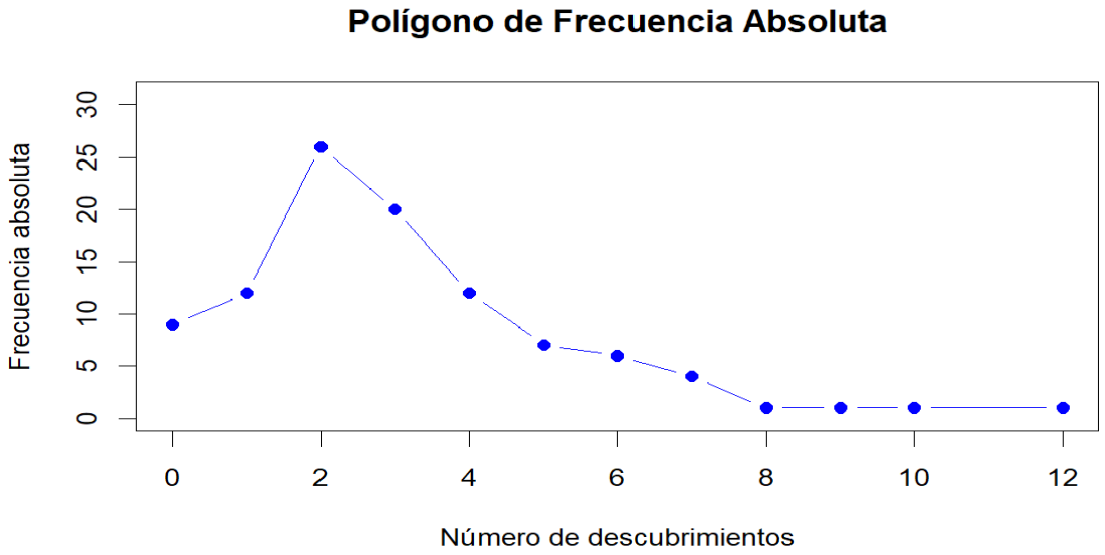
REPRESENTACIÓN DIAGRAMA FRECUENCIA ABSOLUTA.

```
barplot(
    frecuencias_absolutas,
    xlab = "Cantidad de descubrimientos",
    ylab = "Frecuencia absoluta",
    main = "Diagrama de Frecuencias Absolutas",
    col = "lightcoral",
    border = "darkred",
    cex.names = 0.8
)
```



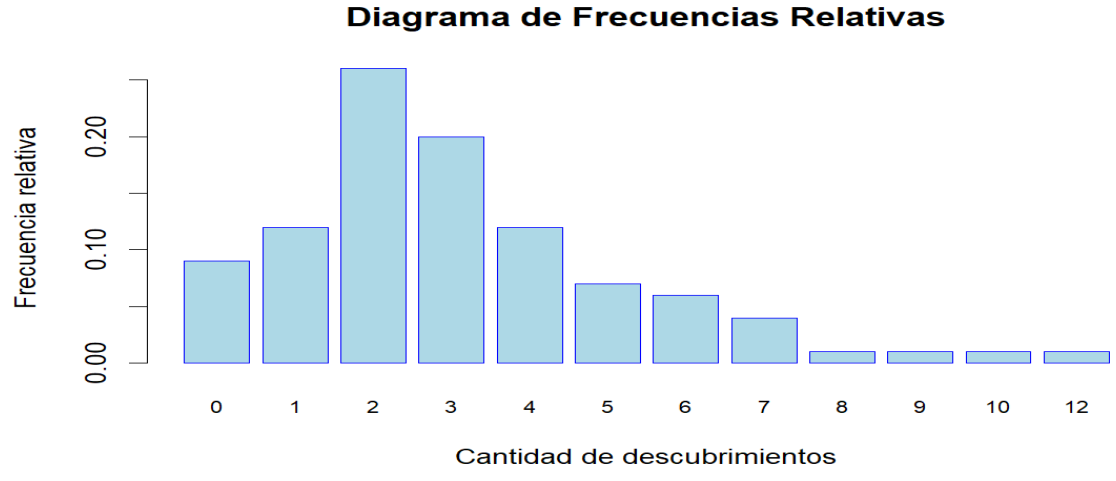
REPRESENTACIÓN DE POLÍGONO FRECUENCIA ABSOLUTA.

```
plot(
    as.numeric(names(frecuencias_absolutas)), frecuencias_absolutas,
    type = "o",
    col = "navy",
    pch = 16,
    xlab = "Número de descubrimientos",
    ylab = "Frecuencia absoluta",
    main = "Polígono de Frecuencia Absoluta"
)
```



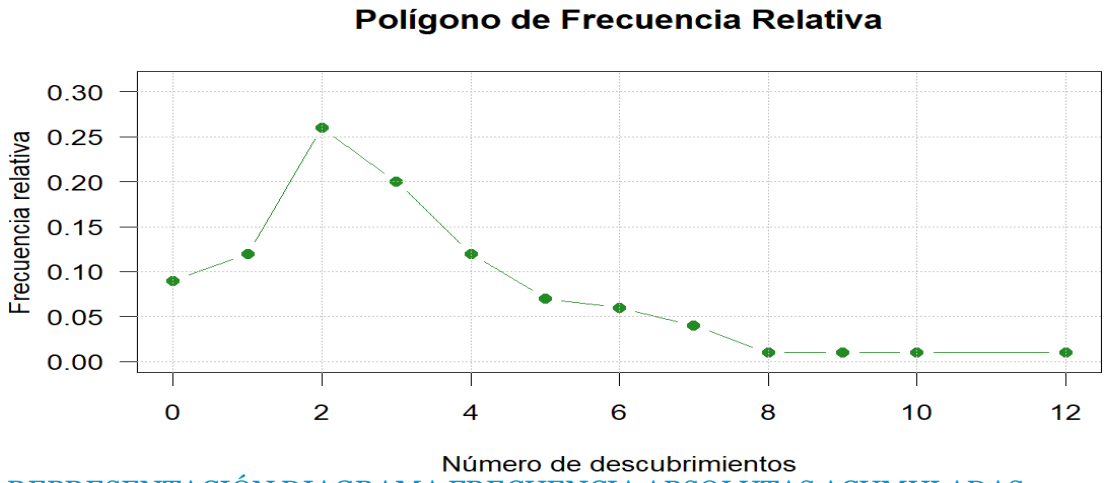
REPRESENTACIÓN DE DIAGRAMA FRECUENCIA RELATIVA

```
barplot(
      frecuencias_relativas,
      xlab = "Cantidad de descubrimientos",
      ylab = "Frecuencia relativa",
      main = "Diagrama de Frecuencias Relativas",
      col = "lightblue",
      border = "blue",
      cex.names = 0.8
)
```



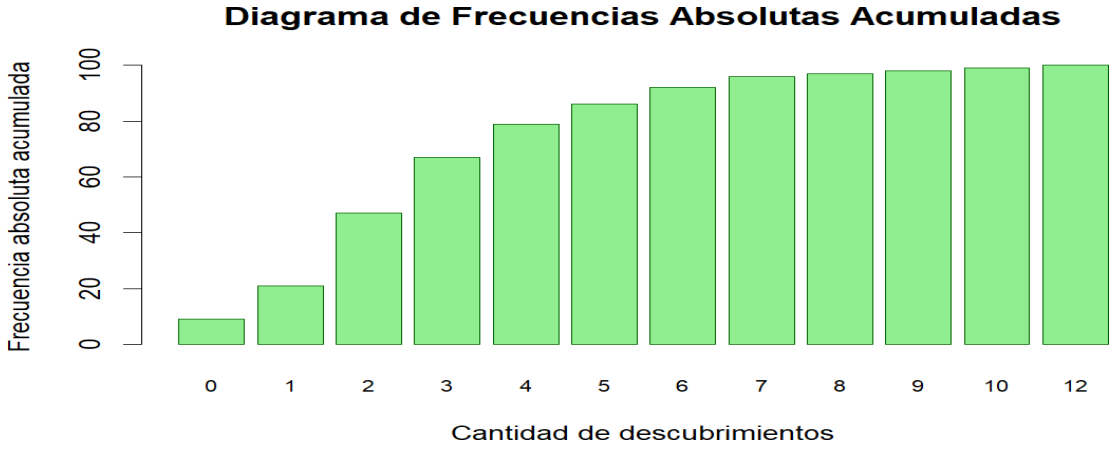
REPRESENTACIÓN POLÍGONO DE FRECUENCIA RELATIVA.

```
plot(
    as.numeric(names(frecuencias_relativas)), frecuencias_relativas,
    type = "o",
    col = "forestgreen",
    pch = 16,
    xlab = "Número de descubrimientos",
    ylab = "Frecuencia relativa",
    main = "Polígono de Frecuencia Relativa"
)
```



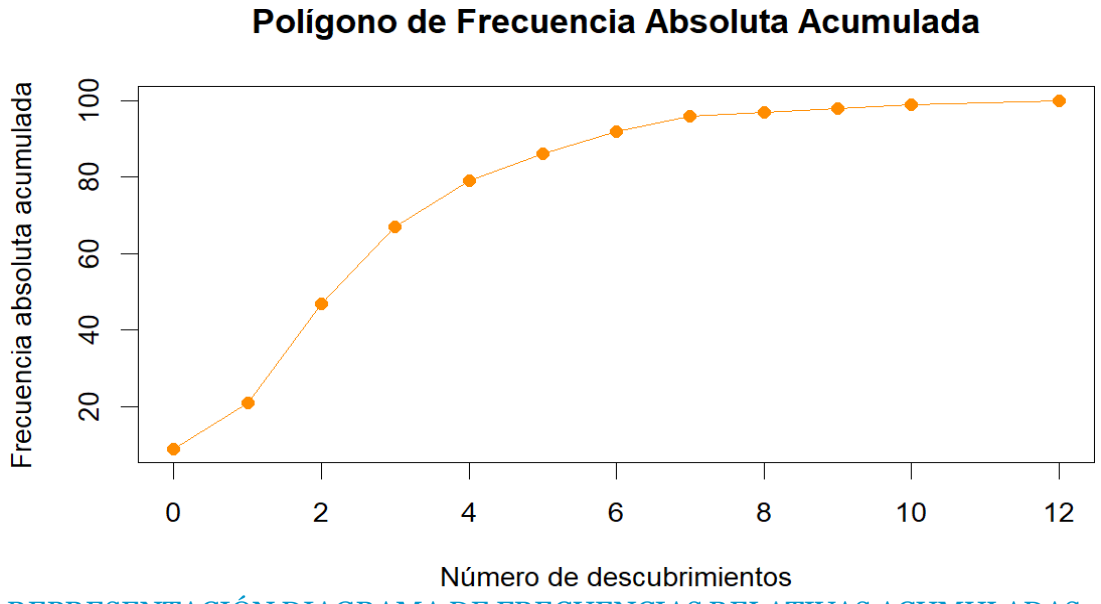
REPRESENTACIÓN DIAGRAMA FRECUENCIA ABSOLUTAS ACUMULADAS.

```
barplot(
    frecuencias_absolutas_acum,
    xlab = "Cantidad de descubrimientos",
    ylab = "Frecuencia absoluta acumulada",
    main = "Diagrama de Frecuencias Absolutas Acumuladas",
    col = "lightgreen",
    border = "darkgreen",
    cex.names = 0.8
)
```



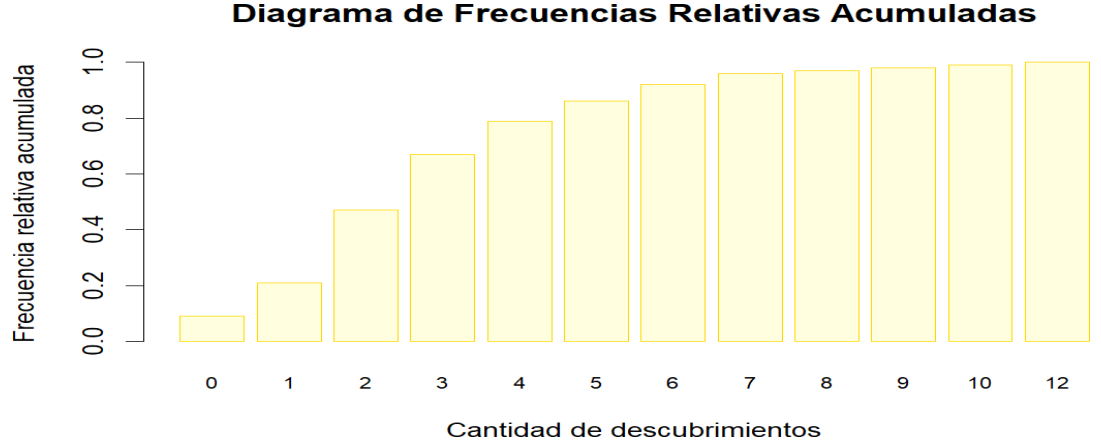
REPRESENTACIÓN POLÍGONO FRECUENCIA ABSOLUTAS ACUMULADAS.

```
plot(
    as.numeric(names(frecuencias_absolutas_acum)), frecuencias_absolutas_acum,
    type = "o",
    col = "darkorange",
    pch = 16,
    xlab = "Número de descubrimientos",
    ylab = "Frecuencia absoluta acumulada",
    main = "Polígono de Frecuencia Absoluta Acumulada"
)
```



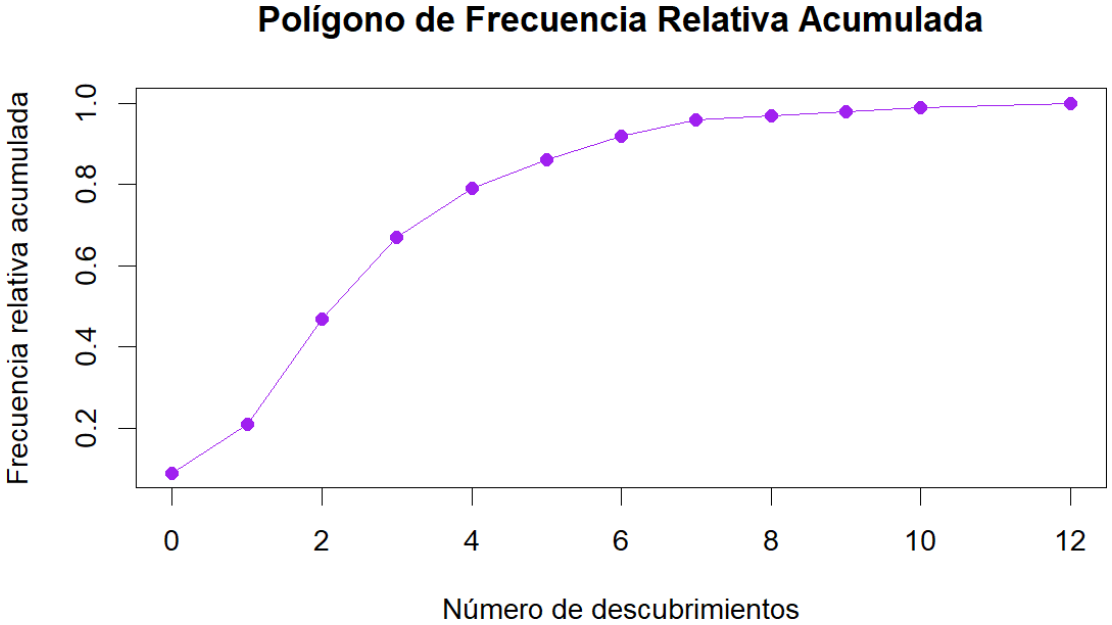
REPRESENTACIÓN DIAGRAMA DE FRECUENCIAS RELATIVAS ACUMULADAS

```
barplot(
    frecuencias_relativas_acum,
    xlab = "Cantidad de descubrimientos",
    ylab = "Frecuencia relativa acumulada",
    main = "Diagrama de Frecuencias Relativas Acumuladas",
    col = "lightyellow",
    border = "gold",
    cex.names = 0.8
)
```



REPRESENTACIÓN POLÍGONO DE FRECUENCIAS RELATIVAS ACUMULADAS

```
plot(
    as.numeric(names(frecuencias_absolutas_acum)), frecuencias_absolutas_acum,
    type = "o",
    col = "darkorange",
    pch = 16,
    xlab = "Número de descubrimientos",
    ylab = "Frecuencia absoluta acumulada",
    main = "Polígono de Frecuencia Absoluta Acumulada"
)
```



4.3 ACTIVIDAD 3: LA MODA COMPARACION DE DATOS

PREGUNTA 1.

```
# Ordenar los datos de discoveries
datos_ordenados <- sort(discoveries)
print(datos_ordenados)
```

```
 [1]  0  0  0  0  0  0  0  0  0  1  1  1  1  1  1  1  1  1  1  1  2  2  2  2  2  2  2
[29]  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  3  3  3
[57]  3  3  3  3  3  3  3  3  3  3  3  3  4  4  4  4  4  4  4  4  4  4  4  5  5  5  5  5
[85]  5  5  6  6  6  6  6  6  7  7  7  7  8  9 10 12
```

Como se puede analizar, tras ordenar los números para una mayor facilidad visual, el número moda es el 2 ya que es el que más se repite.

PREGUNTA 2.

```
# Instalar la biblioteca modeest
if (!require(modeest)) install.packages("modeest")
# Cargar la biblioteca
library(modeest)
# Calcular la moda usando mfv (Most Frequent Value)
moda_mfv <- mfv(discoveries)
print(moda_mfv)
```

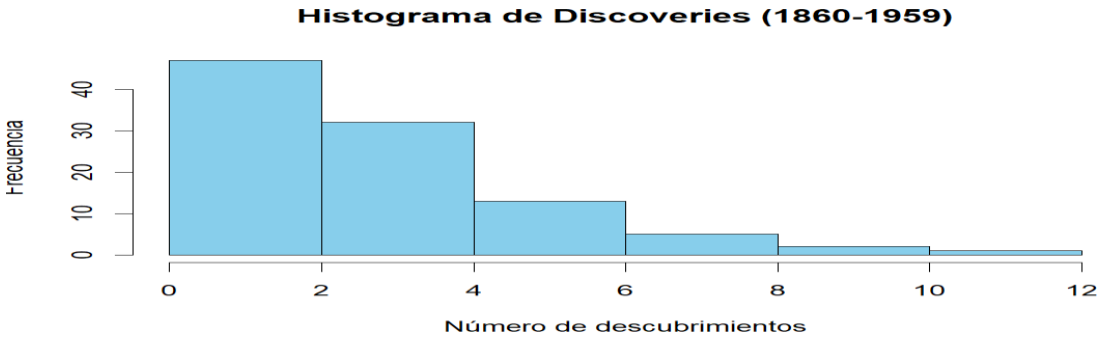
Se muestra el resultado: `[1] 2`

PREGUNTA 3.

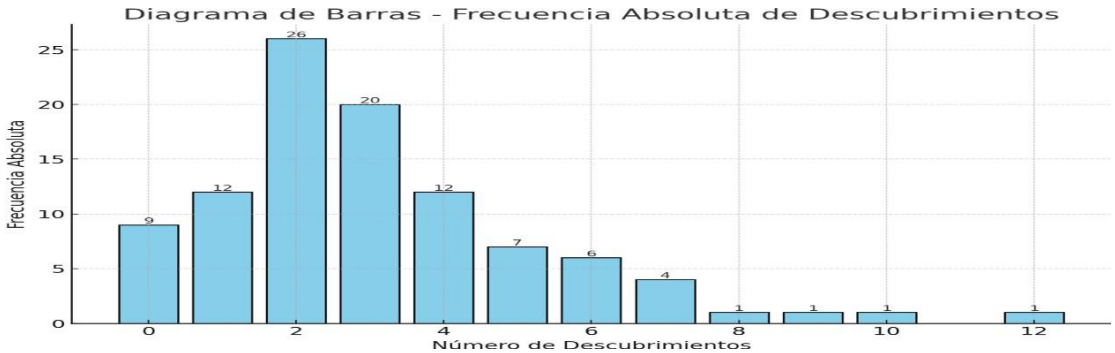
Para corroborarlo, se observa que la moda calculada con el método mfv() es coincidente con el análisis visual previo y, por lo tanto, se concluye que son coherentes en cuanto al resultado obtenido.

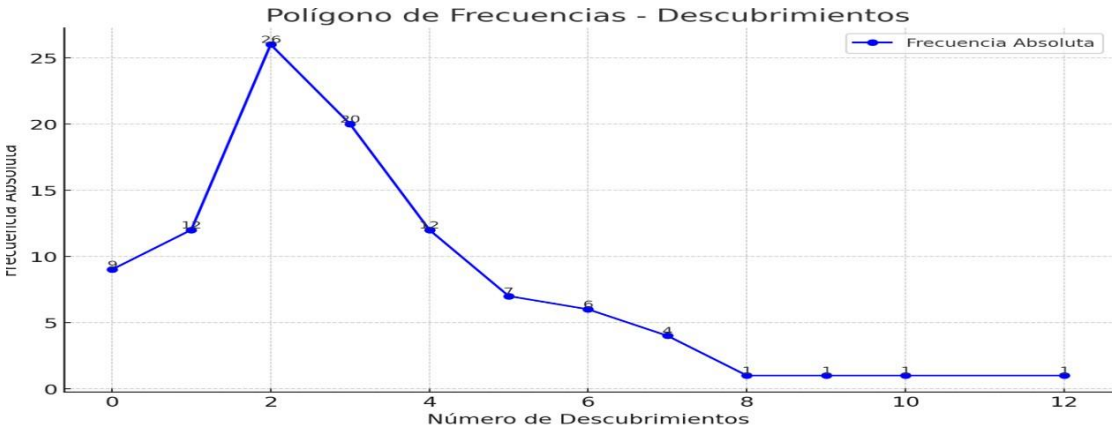
PREGUNTA 4.

```
hist(
     discoveries,
     main = "Histograma de Discoveries (1860-1959)",
     xlab = "Número de descubrimientos",
     ylab = "Frecuencia",
     col = "skyblue",
     border = "black"
)
```



PREGUNTA 5: Análisis diagrama y polítono frecuencia absoluta.





Análisis

Se puede determinar que el número de descubrimientos con mayor frecuencia es el valor 2, con 26 puntos respecto de la frecuencia absoluta, esto indica que la moda equivale a 2 descubrimientos anuales.

El resultado de calcular la **mediana** (posición central = 50 y total de observaciones=100) **es 2**, valores que abarca del 22º al 47º año en la distribución acumulada y la **media** como resultado de (Núm. Total de descubrimientos * Frecuencia absoluta) / Total de observaciones **es 2.76**. Por lo que , la distribución se considera **ligeramente asimétrica positiva**, (cola hacia la derecha), porque la mayoría de los valores están en el rango de 0 a 5 descubrimientos, mientras que hay valores atípicos (8, 9, 10 y 12 descubrimientos) que son poco frecuentes.

PREGUNTA 6.

El valor más frecuente mostrado por el histograma (2 descubrimientos) es coherente con los resultados obtenidos. Tanto el cálculo de la moda (26 años con 2 descubrimientos) como la mediana y los gráficos (diagrama de barras y polígono de frecuencias) refuerzan que 2 descubrimientos es el valor predominante en el conjunto de datos. Por todo lo anterior, se puede afirmar que el valor que muestra el histograma más frecuente es coherente.

PREGUNTA 7.

El diagrama más representativo es el diagrama de barras ya que se puede identificar con la altura de la barra representando el valor de frecuencia absoluta que indica con claridad visual **la moda**. Con respecto al polígono de frecuencia absoluta es menos representativo en término de valores discretos que el diagrama de barras.

REFERENCIAS

Blog R. *Funciones de visualización en R*. Disponible en: https://blog-r.es/visualizacion_de_datos/funciones-de-visualizacion-en-r/

Blog R. *Funciones de manipulación de datos en R*. Disponible en: https://blog-r.es/analisis_de_datos/funciones-de-manipulacion-de-datos-en-r/

Blog R. *Análisis de datos en R*. Disponible en: https://blog-r.es/analisis_de_datos/analisis-datos-r/

Machine Learning para Todos. *¿Qué es R y para qué utilizarlo?* Disponible en: https://machinelearningparatodos.com/que-es-r-y-para-que-utilizarlo/

Conectando Ideas. *Análisis estadístico con R*. Disponible en: https://conectandoideas.net/analisis-estadistico-con-r/

Análisis de Datos con R. *Introducción al análisis de datos con R*. Disponible en: https://analisis-de-datos-con-r.github.io/CRISP-DM/001-Introduccion.html