Taller 1

Andres Felipe Beltran Rodriguez

10/23/2021

1. Con la variable gastos semanales de los empleados en una empresa construya:

a. una distribución de frecuencias, de frecuencias relativas, acumuladas, y relativas acumuladas.

Para resolver este punto debemos ingresar los datos a R:

```
library(readxl)
tarea1 <- read_excel("./tarea1.xlsx")

gastos <- tarea1$gastos
```

Para definir la cantidad de intervalos en los cuales separar los datos, podemos utilizar la ecuacion de sturges (Sturges 1926)

$$k = 1 + 3.332 \cdot log(n) \tag{1}$$

- n = tamaño de muestra, número de observaciones
- k = número de intervalos

podemos revisar el tamaño de muestra con la función `length()` de R:

```
length(gastos)
```

```
## [1] 90
```

Para esta muestra de 90 individuos:

$$k = 1 + 3.332 \cdot log(90) \tag{2}$$

```
k <- 1 + (3.322 * log10(90))
k
```

```
## [1] 7.491994
```

```
round(k,0)
```

```
## [1] 7
```

$$7 \approx 7.491994 = 1 + 3.332 \cdot log(90) \tag{3}$$

Para este conjunto de datos podemos utilizar 7 intervalos. Para saber el tamaño del intervalo, podemos dividir el rango en el número de intervalos:

$$longitud\ del\ intervalo = \frac{max(gastos) - min(gastos)}{7} \tag{4}$$

```
LongInt <- (max(gastos)-min(gastos))/7
LongInt
```

```
## [1] 15565.86
```

La longitud del intervalo es $13620.12 \approx 13620$.

```
LongInt <- round(LongInt,0)
LongInt
```

```
## [1] 15566
```

Podemos entonces construir la tabla de valores minimos y máximos para este intervalo:

Primero calculamos los límites inferiores de los 16 intervalos, Teniendo encuenta que ya tenemos el primero, `min(gastos)`, tenemos que calcular los 15 restantes.

```
mins <- seq(min(gastos),
            min(gastos)+((LongInt))*6,
            by = LongInt)
```

Para calcular los límites superiores, basta con sumar la longitud del intervalo - 1, ya que uno de los valores del intervalo ya está (el límite inferior):

```
maxs <- mins + LongInt-1

TDF <- data.frame(
  min = mins,
  max = maxs
)
TDF
```

```
##      min    max
## 1  40000  55565
## 2  55566  71131
## 3  71132  86697
## 4  86698 102263
## 5 102264 117829
## 6 117830 133395
## 7 133396 148961
```

Ahora podemos iterar a lo largo de las filas de la tabla de frecuencias `TDF` buscando cuantos elementos de `gastos` estan dentro de cada intervalo definido por cada fila.

```
for(i in 1: nrow(TDF)){

  TDF$fi[i] <- length(
    which(
TDF$min[i] <= gastos & gastos <= TDF$max[i]
    )
  )

}
TDF
```

```
##      min    max fi
## 1  40000  55565 10
## 2  55566  71131 18
## 3  71132  86697 16
## 4  86698 102263 15
## 5 102264 117829  9
## 6 117830 133395 12
## 7 133396 148961 10
```

Una vez tenemos la frecuencia absoluta, podemos calcular la frecuencia relativa dividiendo por el numero de observaciones:

```
TDF$fr <- round(TDF$fi/length(gastos),2)
TDF
```

```
##      min    max fi   fr
## 1  40000  55565 10 0.11
## 2  55566  71131 18 0.20
## 3  71132  86697 16 0.18
## 4  86698 102263 15 0.17
## 5 102264 117829  9 0.10
## 6 117830 133395 12 0.13
## 7 133396 148961 10 0.11
```

Una vez tenemos las frecuencias absolutas y relativas, podemos calcular las acumuladas de la siguiente manera:

- Primero para la frecuencia absoluta acumulada(F_i):

```
for(i in 1:nrow(TDF)){

  TDF$Fi[i] <- sum(TDF$fi[1:i])

}
```

- También para la frecuencia relativa acumulada(F_r):

```
for(i in 1:nrow(TDF)){

  TDF$Fr[i] <- sum(TDF$fr[1:i])

}
```

Una vez hemos calculado todas las frecuencias, podemos imprimir la tabla final:

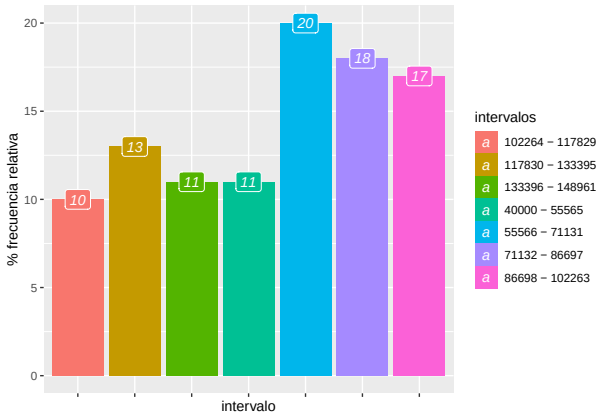
```
knitr::kable(TDF2,"simple")
```

min.	max.	f_i	f_r	F_i	F_r
40000	55565	10	0.11	10	0.11
55566	71131	18	0.20	28	0.31
71132	86697	16	0.18	44	0.49
86698	102263	15	0.17	59	0.66
102264	117829	9	0.10	68	0.76
117830	133395	12	0.13	80	0.89
133396	148961	10	0.11	90	1.00

Para observar los resultados de la tabla de frecuencias, podemos hacer un histograma del porcentaje de frecuencia relativa en funcion de los intervalos:

```
library(ggplot2)

intervalos <- factor(paste(TDF$min,'-',TDF$max))
ggplot(TDF,
aes(x = intervalos,
y = fr*100,
fill=intervalos,
label = round(fr*100,2)
)
) +
geom_bar(stat="identity") +
xlab("intervalo") +
ylab('% frecuencia relativa')+
geom_label(aes(fill = intervalos),
colour = "white",
fontface = "italic") +
 theme(axis.text.x = element_blank())
```



Referencias

- Sturges, H. A. (1926). The choice of a class interval. Journal of the american statistical association, 21(153), 65-66.