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")
```

Luego podemos separar solo la variable gastos en un vector:

```
gastos <- tarea1$gastos
```

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

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

- 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:

$$k = 1 + 3.332 \cdot log(90)$$

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

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

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

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

$$7 \approx 7.491994 = 1 + 3.332 \cdot log(90) \quad (1)$$

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

$$longitud\ del\ intervalo = \frac{max(gastos) - min(gastos)}{7} \quad (2)$$

```
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:

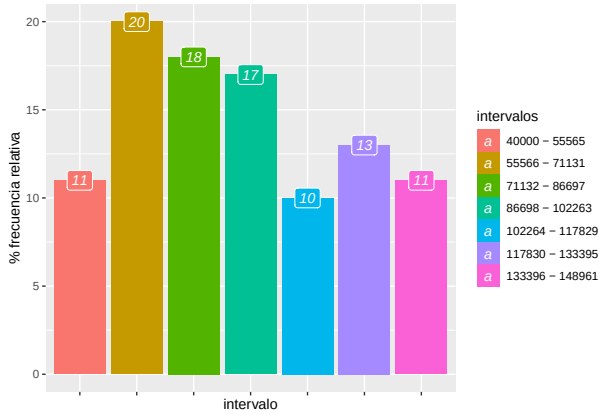
```
knitr::kable(TDF2,"simple")
```

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 <- sort(factor(paste(TDF$min,'-',TDF$max),
                          levels = paste(TDF$min,'-',TDF$max)))
ggplot(TDF,
aes(x = reorder(intervalos,intervalos, function(x) length(x)),
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())
```



Medidas descriptivas de localización y dispersión d. Calcule las medidas descriptivas de localización y variabilidad incluyendo el primer cuartil y el decil siete. Interpretar los resultados obtenidos. La media se calcula con la funcion `mean()`

```
mean(gastos)
```

```
## [1] 91307.11
```

*En promedio los gastos de la empresa son de 91307.11

La mediana se calcula con la funcion `median()`

```
median(gastos)
```

```
## [1] 88021
```

*El 50% de los gastos de la empresa son menores a 88021 y el otro 50% de los gastos son mayores

Para calcular la moda creamos una funcion `moda()`

```
moda <- function(x) {
  v <- unique(x)
  t <- tabulate(match(x, v))
  v[t == max(t)]
}
moda(gastos)
```

```
## [1] 50000 60000 65000 75000
```

*Los gastos mas frecuentes en la empresa tienen valores de 50000, 60000, 65000 y 75000

La desviacion estandar y la varianza muestral se calcula con las funciones `sd()` y `var()`

```
sd(gastos)
```

```
## [1] 29527.13
```

```
var(gastos)
```

```
## [1] 871851382
```

*La desviacion de las observaciones respecto al valor promedio es de 29527.13

Para calcular el rango se le resta al maximo dato el minimo:

```
Rango<-max(gastos)-min(gastos)
Rango
```

```
## [1] 108961
```

*El mayor gasto de la empresa es 108961 mas costoso que el gasto minimo de la empresa

Para calcular el coeficiente creamos la funcion `Coefvar()`

```
CoefVar<-function(x){
  sd(x)/mean(x)*100
}
CoefVar(gastos)
```

```
## [1] 32.33826
```

*Como el coeficiente de variacion es mayor al 20% los datos son heterogeneos

Cuartiles y deciles

Para calcular cuartiles se usa la función `quantile()`

```
quantile(gastos)
```

```
##       0%      25%      50%      75%     100%
##  40000.0  66050.0  88021.0 115041.2 148961.0
```

*Primer cuartil :el 25% de los gastos de la empresa tienen un valor de 40000 hasta 66050 y el 75% restante tienen un valor de 66050 hasta 148961

Para calcular el rango intercuartilico se usa la función `IQR()`

```
IQR(gastos)
```

```
## [1] 48991.25
```

*El 50% de los gastos de la empresa tienen un valor de 66050 hasta 115041.2

Para calcular deciles se usa la función `quantile()`

```
quantile(gastos,probs = seq(0,1,0.1))
```

```
##       0%      10%      20%      30%      40%      50%      60%      70%
##  40000.0  55163.3  62790.2  69116.9  78065.0  88021.0  98977.0 110067.8
##      80%      90%     100%
## 121277.2 135264.4 148961.0
```

*Septimo decil: El 70% de los gastos de la empresa tienen un valor de 40000 hasta 110067.8 y el 30% restante tienen un valor de 110067.8 hasta 148961

e. Qué puede decir respecto a la simetría y curtosis en la distribución de los datos, justificar la respuesta e interprete estas medidas. ### Símetria y curtosis

```
library(e1071)
skewness(gastos)
```

```
## [1] 0.2963071
```

```
kurtosis(gastos)
```

```
## [1] -1.072836
```

Como el resultado del coeficiente de asimetria es mayor a 0, la distribucion es asimetria a la derecha, es decir que hay un sesgo hacia la derecha. La curtosis dio un resultado negativo indicando una distribucion platocurtica, mas aplastada que la distribucion normal y con mas desviacion estandar.

f. Qué proporción de mediciones está dentro de 1.5 desviaciones estándar de la media, a dos desviaciones estándar de la media y a tres desviaciones estándar de la media, concluya? Para calcular que proporcion de los datos estan en 1,5 2 y 3 desviaciones estandar de la media, obtenemos estos rangos de datos:

```
Rango1.5sd<-c(mean(gastos)-(1.5*sd(gastos)),mean(gastos)+(
length(which(gastos>Rango1.5sd[1]  & gastos <Rango1.5sd[2]
```

```
## [1] 87.77778
```

```
Rango2sd<-c(mean(gastos)-(2*sd(gastos)),mean(gastos)+(2*sd
length(which(gastos>Rango2sd[1]  & gastos <Rango2sd[2]))/1
```

```
## [1] 100
```

```
Rango3sd<-c(mean(gastos)-(3*sd(gastos)),mean(gastos)+(3*sd
length(which(gastos>Rango3sd[1]  & gastos <Rango3sd[2]))/1
```

```
## [1] 100
```

*El 87,7% de las mediciones esta dentro de 1.5 desviaciones estándar de la media y el 100% de las mediciones esta dentro de 2 y 3 desviaciones estándar de la media.

g. Cuál de las medidas de localización y de variabilidad recomendaría y porque? > *Para estos datos la mejor medida de tendencia central

serian las modas, ya que como se ve en el histograma la mayoria de gastos se encuentran en el intervalo entre 55566 y 71131 que coincide con las modas, la media y la mediana se ven afectadas por el sesgo a la derecha y sus valores son mayores a 8000. La mejor medida de varianza podria ser el rango intercuartilico ya que no se ve tan afectado por la media como la desviacion estandar.

h. Calcule los coeficientes de variación por turno para la variable gastos semanales y diga en cuál de los turnos hay más homogeneidad, interprete. Coeficientes de variacion por turno

```
empresa<-tarea1[c(2,4)]
empresa$turno<-as.factor(empresa$turno)
diurno<-subset(empresa,turno == "diurno")
tarde<-subset(empresa,turno == "tarde")
nocturna<-subset(empresa,turno == "nocturna")

CoefVar(diurno$gastos)
```

```
## [1] 37.21381
```

```
CoefVar(tarde$gastos)
```

```
## [1] 25.51264
```

```
CoefVar(nocturna$gastos)
```

```
## [1] 34.67709
```

*El menor coeficiente de variacion lo tiene el turno de la tarde (25.5%) por lo que es el turno con mayor homogeneidad, le sigue el turno nocturno (34.7%) que es mas heterogeneo y finalmente el turno diurno (37.2) que es el mas heterogeneo de los 3.

2. De las variables jornada y turno construya:

a. Una tabla de frecuencias para el cruce de las dos variables y concluya.

```
library(gmodels)
CrossTable(tarea1$turno,tarea1$jornada,prop.chisq=F)
```

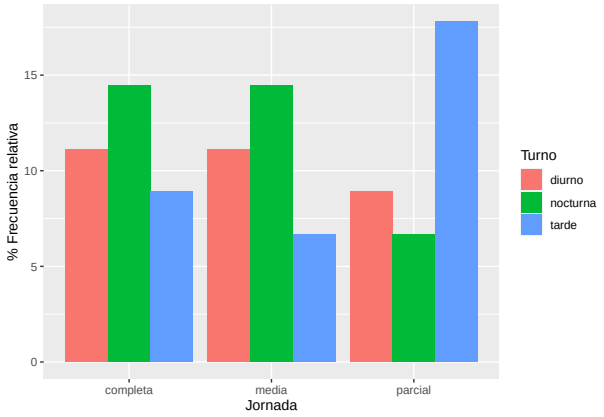
```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  90
##
##
##              | tarea1$jornada
## tarea1$turno |  completa |     media |   parcial | Row
## -------------|-----------|-----------|-----------|-----
##       diurno |        10 |        10 |         8 |
##              |     0.357 |     0.357 |     0.286 |
##              |     0.323 |     0.345 |     0.267 |
##              |     0.111 |     0.111 |     0.089 |
## -------------|-----------|-----------|-----------|-----
##     nocturna |        13 |        13 |         6 |
##              |     0.406 |     0.406 |     0.188 |
##              |     0.419 |     0.448 |     0.200 |
##              |     0.144 |     0.144 |     0.067 |
## -------------|-----------|-----------|-----------|-----
##        tarde |         8 |         6 |        16 |
##              |     0.267 |     0.200 |     0.533 |
##              |     0.258 |     0.207 |     0.533 |
##              |     0.089 |     0.067 |     0.178 |
## -------------|-----------|-----------|-----------|-----
## Column Total |        31 |        29 |        30 |
##              |     0.344 |     0.322 |     0.333 |
## -------------|-----------|-----------|-----------|-----
##
##
```

*El 31% de las personas esta en el turno diurno, de este grupo un 35.7% esta en jornada completa, otro 35.7 % esta en la jornada media y el 28.6% restante esta en jornada parcial. El 35.6% de las personas esta en el turno nocturno, de este grupo un 40.6% esta en jornada completa, otro 40.6 % esta en la jornada media y el 18.8% restante esta en jornada parcial. El 33.3% de las personas esta en el turno tarde, de este grupo un 26.7% esta en jornada completa, un 20 % esta en la jornada media y el 53.3% restante esta en jornada parcial. En el turno diurno y nocturno hay menor cantidad de personas en jornada parcial, en cambio en el turno tarde la mayoria de personas esta en la jor-

nada parcial. Al parecer no se asocian mucho las variables

b. Un histograma de frecuencias relativas (darlas en porcentajes).
c. Interpretar los resultados mostrados por el gráfico. Qué puede decir de la distribución de los datos?

```
tarea1$turno<-as.factor(tarea1$turno)
tarea1$jornada<-as.factor(tarea1$jornada)
tablaf<-table(tarea1$turno,tarea1$jornada)
tFrecRel<-as.data.frame(tablaf/90*100)
colnames(tFrecRel)<-c("Turno","Jornada","Frec
ggplot(tFrecRel, aes(fill=tFrecRel$Turno, y=tFrecRel$`Frecuencia Relativa` , x= tFrecRel$Jornada)) +
    geom_bar(position="dodge", stat="identity")+
xlab("Jornada") +
ylab('% Frecuencia relativa')+ scale_fill_discrete(name = "Turno")
```



> *La distribucion de los datos cambia dependiendo de la jornada, en la jornada completa y media siguen la misma tendencia, en donde la mayoria de personas estan en el turno nocturno y la minoria en el turno tarde, mientras que en la jornada parcial la mayoria de personas esta en el turno tarde y la minoria en el turno noche. Los graficos muestran que las variables no estan asociadas.

d. Diga si hay asociación entre las dos variables con un nivel de significancia de 0.05 e interprete el resultado. Para determinar si hay asociacion se hace un test Ji- cuadrado de Pearson con la funcion `chisq.test()`

```
chisq.test(tarea1$turno,tarea1$jornada)
```

```
##
##  Pearson's Chi-squared test
## data:  tarea1$turno and tarea1$jornada
## X-squared = 8.925, df = 4, p-value = 0.063
```

*En este caso como el valor de alfa es 0.05 y el p-valor es 0.063 es mayor la hipotesis nula se rechaza es decir que no hay asociacion entre la jornada y los turnos.

e. Construya la tabla de frecuencias correspondiente al hábito de fumar y jornada laboral, concluya.

```
tarea1$fuma <-as.factor(tarea1$fuma)
CrossTable(tarea1$fuma,tarea1$jornada,prop.chisq=F)
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  90
##
##
##              | tarea1$jornada
##  tarea1$fuma |  completa |     media |   parcial | Row
## -------------|-----------|-----------|-----------|-----
##           no |         9 |        12 |         7 |
##              |     0.321 |     0.429 |     0.250 |
##              |     0.290 |     0.414 |     0.233 |
##              |     0.100 |     0.133 |     0.078 |
## -------------|-----------|-----------|-----------|-----
##           si |        22 |        17 |        23 |
##              |     0.355 |     0.274 |     0.371 |
##              |     0.710 |     0.586 |     0.767 |
##              |     0.244 |     0.189 |     0.256 |
## -------------|-----------|-----------|-----------|-----
## Column Total |        31 |        29 |        30 |
##              |     0.344 |     0.322 |     0.333 |
## -------------|-----------|-----------|-----------|-----
##
##
```

*El 28% de las personas no fuma, de este grupo un 32% esta en jornada completa, otro 42.9 % esta en la jornada media y el 25% restante esta en jornada parcial. El 68.9% de las personas fuma, de este grupo

un 35.5% esta en jornada completa, otro 27.4 % esta en la jornada media y el 37.1% restante esta en jornada parcial. La mayoria de las personas que no fuman estan en la jornada media, mientras que las que fuman estan en su mayoria en la jornada parcial o completa. Las variables no estan muy asociadas entre si.

f. Hay asociación entre el hábito de fumar de los empleados y la jornada laboral con un nivel de significancia de 0.05. Interprete el resultado?

```
chisq.test(tarea1$fuma,tarea1$jornada)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tarea1$fuma and tarea1$jornada
## X-squared = 2.3359, df = 2, p-value = 0.311
```

*En este caso como el valor de alfa es 0.05 y el p-valor es 0.311 es mayor la hipotesis nula se rechaza es decir que no hay asociacion entre la jornada y el habito de fumar de los empleados.

4. Con el archivo daños por sitio, correspondiente a una muestra de 480 cardones en la alta Guajira. Construya una tabla de frecuencias para el cruce de las variables tipo de daño (0= no hay daño, 1= corte con machete, 2= daño por insectos, 3= comido de cabras, 4= daño por viento, 5= daño por aves) y sitio (1= intervenido, 0= no intervenido).

a. Realizar el análisis respectivo, construya las tablas de frecuencias correspondientes

```
cardones <- read_excel("./tarea1.xlsx",
    sheet = "cardones")
cardones$`Tipo de Daño`<-as.factor(cardones$`Tipo de Daño`)
cardones$sitio<-as.factor(cardones$sitio)
attach(cardones)
CrossTable(`Tipo de Daño`,sitio,prop.chisq=F)
```

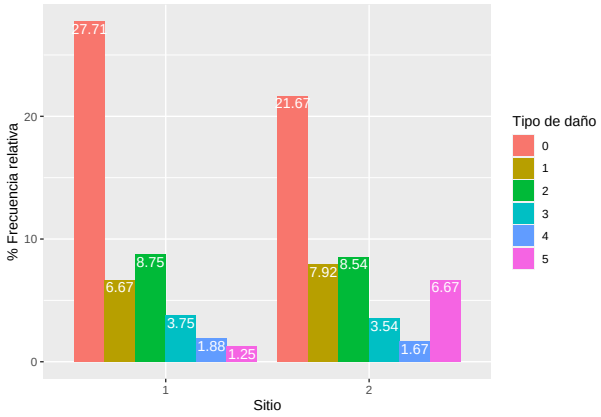
```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  480
##
##
##                | sitio
## Tipo de Daño |         1 |         2 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |       133 |       104 |       237 |
##              |     0.561 |     0.439 |     0.494 |
##              |     0.554 |     0.433 |           |
##              |     0.277 |     0.217 |           |
## -------------|-----------|-----------|-----------|
##            1 |        32 |        38 |        70 |
##              |     0.457 |     0.543 |     0.146 |
##              |     0.133 |     0.158 |           |
##              |     0.067 |     0.079 |           |
## -------------|-----------|-----------|-----------|
##            2 |        42 |        41 |        83 |
##              |     0.506 |     0.494 |     0.173 |
##              |     0.175 |     0.171 |           |
##              |     0.087 |     0.085 |           |
## -------------|-----------|-----------|-----------|
##            3 |        18 |        17 |        35 |
##              |     0.514 |     0.486 |     0.073 |
##              |     0.075 |     0.071 |           |
##              |     0.037 |     0.035 |           |
## -------------|-----------|-----------|-----------|
##            4 |         9 |         8 |        17 |
##              |     0.529 |     0.471 |     0.035 |
##              |     0.037 |     0.033 |           |
##              |     0.019 |     0.017 |           |
## -------------|-----------|-----------|-----------|
##            5 |         6 |        32 |        38 |
##              |     0.158 |     0.842 |     0.079 |
##              |     0.025 |     0.133 |           |
##              |     0.012 |     0.067 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       240 |       240 |       480 |
##              |     0.500 |     0.500 |           |
## -------------|-----------|-----------|-----------|
##
##
```

*Lo que se pueder ver en la tabla de contingencia es que al comparar el tipo de daño en los sitios internevidos y sin intervenir todos tienen valores muy similares excepto el daño 5, que es daño por aves, el 7% de los cardones sufre daño por aves y de este grupo

el 82% esta en sitios no intervenidos y solo un 15.8% esta en sitios intervenidos. Las variables estan asociadas entre si.

b. Hacer el histograma de las frecuencias relativas en porcentaje y concluya.

```
tablef<-table(`Tipo de Daño`,sitio)
tableFrecRel<-as.data.frame(tablef/480*100)
ggplot(tableFrecRel, aes(fill=tableFrecRel$Tipo.de.Daño, y=tableFrecRel$Freq , x= tableFrecRel$sitio)) +
xlab("Sitio") +ylab('% Frecuencia relativa')+ scale_fill_discrete(name = "Tipo de daño")+  geom_text(aes
```



> *Como lo muestra el histograma al comparar los sitios intervenido y los no internevidos las tendencias son similares, excepto para el daño 5 o daño por aves que tiene una frecuencia relativa mucho mayor en los sitios no intervenidos que en los intervenidos. Las variables sitios y tipo de daños estan relacionadas.

c. Hay asociación entre el tipo de daño y el sitio. Utilice un nivel de significancia del 0.05 y concluya?

```
chisq.test(`Tipo de Daño`,sitio)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Tipo de Daño and sitio
## X-squared = 21.952, df = 5, p-value = 0.0005348
```

*En este caso como el valor de alfa es 0.05 y el p-valor es 0.00053 es menor, la hipotesis nula no se rechaza es decir que si hay asociacion entre la intervencion del sitio y el tipo de daño.

Referencias

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