Taller 1

Valentina Gonzalez Ruiz - Andres Felipe Beltran Rodriguez

10/23/2021

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

1. 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)](https://www.tandfonline.com/doi/abs/10.1080/01621459.1926.10502161)

* 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:

Podemos entonces introducir la ecuación a R. y calcular la cantidad de intervalos a utilizar:

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

## [1] 7.491994

Dado que no podemos seleccionar 7.491994 intervalos, podemos rendondear al número entero más cercano con 0 cifras decimales:

round(k,0)

## [1] 7

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

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

## [1] 15565.86

La longitud del intervalo es .

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 7 intervalos, Teniendo encuenta que ya tenemos el primero, min(gastos), tenemos que calcular los 6 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 que calculamos en el paso anterior):

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 están 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:

for(i in 1:nrow(TDF)){  
   
 TDF$Fi[i] <- sum(TDF$fi[1:i])   
   
}

En donde sum(TDF$fi[1:i]) suma todas todas las frecuencias absolutas presentes en la columna fi desde la posición 1, hasta la posición de la nueva columna Fi que se está llenando. Si el contador está en la tercera posición, i = 3 entonces la tercera posición de la columna Fi, TDF$Fi[3], será la suma de todos los elementos desde el primero, TDF$fi[1] hasta el tercero, TDF$fi[3], por esto el uso del operador dos puntos :.

* También para la frecuencia relativa acumulada:

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")

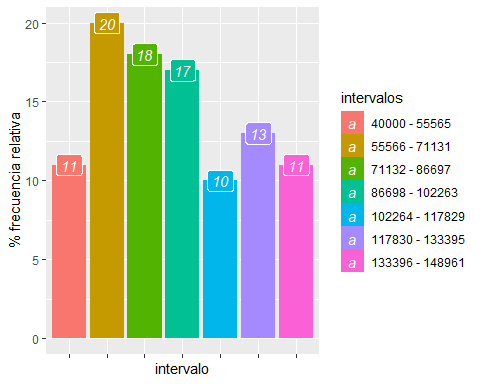
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| min. | max. | fi | fr | Fi | Fr |
| 40000 | 55565 | 10 | 0.11 | 10 | 0.11 |
| 55566 | 71131 | 18 | 0.20 | 28 | 0.31 |
| 71132 | 86697 | 16 | 0.18 | 44 | 0.49 |
| 86698 | 102263 | 15 | 0.17 | 59 | 0.66 |
| 102264 | 117829 | 9 | 0.10 | 68 | 0.76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

A partir de la tabla de frecuencias podemos afirmar que:

* 76% de los empleados entrevistados para la muestra analizada tienen unos gastos semanales de entre 40.000 pesos y 117.829 pesos.
* hay solo 9 empleados que presentan gastos semanales entre 102.264 y 117.829 pesos.
* 80 empleados presentan gastos semanales de entre 40.000 y 133.395 pesos.

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())



1. Interpretar los resultados mostrados por el gráfico. Qué puede decir de la distribución de los datos?

* La distribución muestra que las mayores frecuencias estan en el intervalo de 55566 a 71131, y que hay un sesgo hacia la cola derecha.
* podemos confirmar las afirmaciones sobre sesgo y dispersión al calcular los momentos de curtosis y sesgo:

library(e1071)  
kurtosis(TDF$fr)

## [1] -1.936019

skewness(TDF$fr)

## [1] 0.2388275

En efecto, presentando un momento de curtosis negativo esta distribucion es platicurtica o pesada en los hombros, presentando alta dispersión.

A su vez, presenta un sesgo positivo o con cola hacia la derecha.

Esto quiere decir que a pesar de que el rango de gastos semanales de los empleados es amplio, existe una tendencia entre los empleados hacia gastos bajos por semana.

### Medidas descriptivas de localización y dispersión

1. 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 una semana, los gastos promedio de los empleados de la empresa entrevistados para la muestra utilizada son 91307.11

La mediana se calcula con la funcion median()

median(gastos)

## [1] 88021

Para su interpretación, también podemos hallar el mínimo y el máximo usando la función quantile()

quantile(gastos)

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

Dentro de los empleados de la empresa, que fueron entrevistados en la muestra analizada, la mitad, o el 50% tiene gastos semanales entre 40000 y 88021, mientras que la otra mitad, o el otro 50% tienen gastos semanales entre 88021 y 148961.

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 los empleados entrevistados en la muestra analizada son 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 los gastos semanales de los empleados utilizados en la muestra, respecto al valor promedio es de 29527.13 pesos

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

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

## [1] 108961

\*El mayor gasto semanal de los empleados entrevistados en la muestra 108961 más costoso que el gasto semanal minimo de la muestra analizada.

Para calcular el coeficiente de variación creamos la funcion Coefvar():

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

## [1] 32.33826

* Dado que el coeficiente de variación de gastos semanales en la muestra de empleados analizada es mayor al 20% los gastos semanales en la muestra son heterogeneos.

Ahora, para cada turno de trabajo, podemos separar la muestra en turnos.

Primero revisamos cuantos turnos hay:

turnos <- unique(tarea1$turno)  
turnos

## [1] "diurno" "tarde" "nocturna"

Luego, podemos separar la información de gastos semanales por turno:

library(dplyr)   
gastosDiurno <- filter(tarea1, turno == 'diurno') %>% select(gastos) %>% unlist

Una vez tenemos todos los datos de gastos semanales correspondientes al turno diurno de la muestra analizada, podemos calcular el coeficiente de variación utilizando la función CoefVar() pasando como argumento el vector gastosDiurno:

CoefVar(gastosDiurno)

## [1] 37.21381

Podemos realizar una operación análoga para los otros dos turnos.

gastosTarde <- filter(tarea1, turno == 'tarde') %>% select(gastos) %>% unlist  
gastosNocturna <- filter(tarea1, turno == 'nocturna') %>% select(gastos) %>% unlist  
CoefVarTurno <- round(data.frame(diurno = CoefVar(gastosDiurno),  
 tarde =CoefVar(gastosTarde), nocturna =CoefVar(gastosNocturna)),2)

y visualizamos los resultados:

knitr::kable(CoefVarTurno, 'simple')

|  |  |  |
| --- | --- | --- |
| diurno | tarde | nocturna |
| 37.21 | 25.51 | 34.68 |
|  |  |  |

Gracias a la tabla anterior, podemos afirmar que en los tres turnos de trabajo, se observa heterogeneidad en los gastos semanales de los empleados entrevistados para la muestra analizada.

**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 e 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 78065.4 88021.0 97707.0 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

1. 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 una mayor dispersión que la de una distribución normal estándar.

1. 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)+(1.5\*sd(gastos)))  
length(which(gastos>Rango1.5sd[1] & gastos <Rango1.5sd[2]))/length(gastos)\*100

## [1] 87.77778

Rango2sd<-c(mean(gastos)-(2\*sd(gastos)),mean(gastos)+(2\*sd(gastos)))  
length(which(gastos>Rango2sd[1] & gastos <Rango2sd[2]))/length(gastos)\*100

## [1] 100

Rango3sd<-c(mean(gastos)-(3\*sd(gastos)),mean(gastos)+(3\*sd(gastos)))  
length(which(gastos>Rango3sd[1] & gastos <Rango3sd[2]))/length(gastos)\*100

## [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.

1. 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.

1. 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:

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

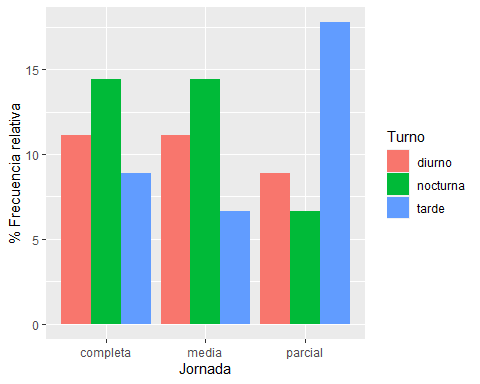
library(gmodels)  
CrossTable(tarea1$turno,tarea1$jornada,prop.chisq=F)

##   
##   
## 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 Total |   
## -------------|-----------|-----------|-----------|-----------|  
## diurno | 10 | 10 | 8 | 28 |   
## | 0.357 | 0.357 | 0.286 | 0.311 |   
## | 0.323 | 0.345 | 0.267 | |   
## | 0.111 | 0.111 | 0.089 | |   
## -------------|-----------|-----------|-----------|-----------|  
## nocturna | 13 | 13 | 6 | 32 |   
## | 0.406 | 0.406 | 0.188 | 0.356 |   
## | 0.419 | 0.448 | 0.200 | |   
## | 0.144 | 0.144 | 0.067 | |   
## -------------|-----------|-----------|-----------|-----------|  
## tarde | 8 | 6 | 16 | 30 |   
## | 0.267 | 0.200 | 0.533 | 0.333 |   
## | 0.258 | 0.207 | 0.533 | |   
## | 0.089 | 0.067 | 0.178 | |   
## -------------|-----------|-----------|-----------|-----------|  
## Column Total | 31 | 29 | 30 | 90 |   
## | 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 jornada parcial. Al parecer no se asocian mucho las variables

1. Un histograma de frecuencias relativas (darlas en porcentajes).
2. 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","Frecuencia Relativa")  
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")

 > \* 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.

1. 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.

1. 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 Total |   
## -------------|-----------|-----------|-----------|-----------|  
## no | 9 | 12 | 7 | 28 |   
## | 0.321 | 0.429 | 0.250 | 0.311 |   
## | 0.290 | 0.414 | 0.233 | |   
## | 0.100 | 0.133 | 0.078 | |   
## -------------|-----------|-----------|-----------|-----------|  
## si | 22 | 17 | 23 | 62 |   
## | 0.355 | 0.274 | 0.371 | 0.689 |   
## | 0.710 | 0.586 | 0.767 | |   
## | 0.244 | 0.189 | 0.256 | |   
## -------------|-----------|-----------|-----------|-----------|  
## Column Total | 31 | 29 | 30 | 90 |   
## | 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.

1. 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

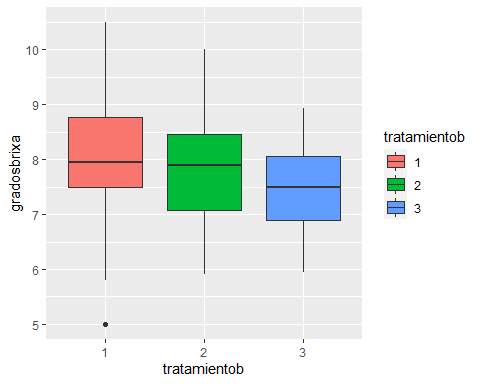
* En este caso como el valor de alfa es 0.05 y el p-valor es 0.311 es mayor la hipotesis nula se rechaza es decir que no hay asociacion entre la jornada y el habito de fumar de los empleados.

## 3. Con el archivo adjunto de fresa

1. Realizar el boxplot de la variable gradosbrixa para frutos de fresa por tratamientob (0.25 g/m2 de Nitrogeno (N), 0.20 g/m2 de anhídrido fosfórico (P2O5), 0.15 g/m2 de óxido de potasa (K2O) e interprete el gráfico correspondiente.

library(readxl)  
fresa <- read\_excel("./tarea1.xlsx", sheet =2)

nas <- which(is.na(fresa$gradosbrixa))  
  
colores<-c("#00AFBB", "#E7B800", "#FC4E07")  
  
df1 <- data.frame(gradosbrixa = fresa$gradosbrixa[-nas],  
 tratamientob = as.factor(fresa$tratamientob[-nas]))  
library(ggplot2)  
ggplot(df1,  
 aes( x = tratamientob,  
 y = gradosbrixa,  
 fill = tratamientob)) +  
   
 geom\_boxplot()

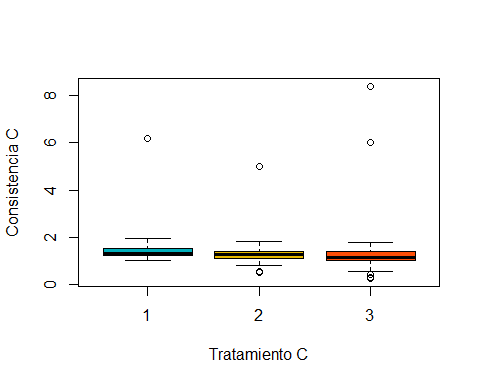


* Podemos ver que para el tratamiento B 1 los grados bix A presentan un outlier, además la mediana no esta en el centro de la caja indicando un ligero sesgo hacia la cola derecha; tiene el rango mas grande de los tres tratamientos.

para el tratamiento B 2 el rango es mas pequeño pero estos datos tienen mas dispersion, como se puede ver por el ancho de la caja, tambien presenta un ligero sesgo hacia la izquierda. Finalmente, el tratamiento B 3 es el que tiene menor rango y se ve simetrico. Estas afirmaciones se pueden soportar calculando los momentos de curtosis y de sesgo.

1. Realizar el boxplot para la variable consistenciac en frutos de fresa por tratamientoc (0.10 g/m2 óxido de magnesio (MgO), 20 g/m2 de Nitrógeno, 10 g/m2 de anhídrido fosfórico (P2O5)), e interprete el gráfico correspondiente.

boxplot(fresa$consistenciac ~ fresa$tratamientoc, xlab = "Tratamiento C", ylab = "Consistencia C", col= colores)



* En los 3 tratamientos vemos que se encuentran outliers, estos datos de consistencia C son tan grandes que no permiten ver muy bien la distribución de los demas datos, sin embargo, se alcanza a ver que el tratamiento C 1 tiene un ligero sesgo hacia la cola derecha y es el de menor rango, luego le sigue el tratamiento C 2 tiene un ligero sesgo hacia la cola izquierda, que tiene un rango un poco mas grade y por ultimo el tratamiento C 3 que tiene el rango mas grande de los 3 y

Para confirmar las afirmaciones sobre el sesgo de las distribuciones podemos intentar eliminar los datos atipicos:

library(dplyr)  
fresa2 <- fresa  
  
tratc1 <- filter(fresa2, tratamientoc == 1) %>% select(consistenciac) %>% unlist   
IQR1 <- IQR(tratc1)  
CritInf1 <- quantile(tratc1)[2] - 1.5\*IQR1  
CritSup1 <- quantile(tratc1)[4] + 1.5\*IQR1  
  
out1 <- which(tratc1 < CritInf1 | tratc1 > CritSup1)  
consistenciac1 <- unlist(fresa2$consistenciac[-c(out1)])

En el chunk anterior retiramos el valor en la posicion 70, que coincide con el maximo para el tratamiento C nivel 1 y a su vez un dato atipico. Podemos realizar una operacion analoga para los otros dos niveles del tratamiento C:

* Tratamiento C 2

tratc2 <- filter(fresa2, tratamientoc == 2) %>% select(consistenciac) %>% unlist   
IQR2 <- IQR(tratc2)  
CritInf2 <- quantile(tratc2)[2] - 1.5\*IQR2  
CritSup2 <- quantile(tratc2)[4] + 1.5\*IQR2  
  
out2 <- which(tratc2 < CritInf2 | tratc2 > CritSup2)  
out2

## consistenciac38 consistenciac69 consistenciac71   
## 38 69 71

consistenciac2 <- unlist(fresa2$consistenciac[-c(out2)])

Para el nivel 2 del tratamiento C identificamos 3 valores atipicos.

* Tratamiento C 3

tratc3 <- filter(fresa2, tratamientoc == 3) %>% select(consistenciac) %>% unlist   
IQR3 <- IQR(tratc3)  
CritInf3 <- quantile(tratc3)[3] - 1.5\*IQR3  
CritSup3 <- quantile(tratc3)[4] + 1.5\*IQR3  
  
out3 <- which(tratc3 < CritInf3 | tratc3 > CritSup3)  
out3

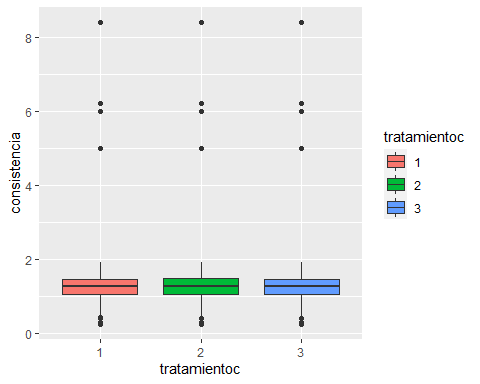
## consistenciac47 consistenciac48 consistenciac51 consistenciac52 consistenciac53   
## 47 48 51 52 53   
## consistenciac54 consistenciac56 consistenciac73 consistenciac74 consistenciac75   
## 54 56 73 74 75

consistenciac3 <- unlist(fresa2$consistenciac[-c(out3)])

Para el nivel 3 del tratamiento C identificamos 10 valores atipicos.

Podemos una vez retirados los valores atipicos volver a graficar los box plot

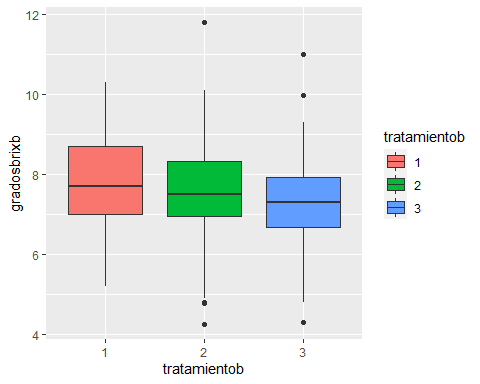
consistenciasC <- c(consistenciac1,consistenciac2,consistenciac3)  
tratamientoC <- c(rep(1,233),rep(2,231),rep(3,224))  
  
consistenciaSinOut <- data.frame(consistencia = consistenciasC, tratamientoc = as.factor(tratamientoC))  
  
library(ggplot2)  
ggplot(consistenciaSinOut,  
 aes( x = tratamientoc ,  
 y = consistencia,  
 fill = tratamientoc)) +  
   
 geom\_boxplot()



Podemos observar que al retirar datos atipicos cambian las distribuciones, y es dificil diferenciarlas entre si. Esto indica que tal vez retirar los datos atipicos en este caso no es la mejor opción.

1. Realizar el boxplot para la variable gradosbrixb en frutos de fresa por tratamientob (0.25 g/m2 de Nitrogeno (N), 0.20 g/m2 de anhídrido fosfórico (P2O5), 0.15 g/m2 de óxido de potasa (K2O) de anhídrido fosfórico (P2O5)), e interprete el gráfico correspondiente.

nasb <- which(is.na(fresa$gradosbrixb))  
  
df2 <- data.frame(gradosbrixb = fresa$gradosbrixb[-nasb],  
 tratamientob = as.factor(fresa$tratamientob[-nasb]))  
library(ggplot2)  
ggplot(df2,  
 aes( x = tratamientob,  
 y = gradosbrixb,  
 fill = tratamientob)) +  
   
 geom\_boxplot()



En este histograma podemos observar que el tratamiento 1 presenta mayores grados brixb que 2 y 2 a su vez que 3, teniendo en cuenta las medias. También el tratamiento 1 presenta mayor dispersión, el tratamiento 2 un sesgo con cola hacia la derecha y el 3 un sesgo con cola hacia la izquierda.

1. Construya una tabla de frecuencias completa para la variable consistenciaa en frutos de fresa y realice el gráfico de barras que le permita ver la distribución de frecuencias relativas en porcentaje e interprete.

Primero separamos la variable consistenciaa en un vector:

nas3 <- which(is.na(fresa$consistenciaa))  
consistenciaa <- fresa$consistenciaa[-nas3]  
  
length(consistenciaa)

## [1] 84

Para determinar la cantidad de intervalos de nuevo utilizamos la ecuación de sturges:

k2 <- 1 + (3.322 \* log10(84))  
k2

## [1] 7.392456

round(k2,0)

## [1] 7

Tendremos para esta variable 7 intervalos. Para saber la longitud del intervalo:

LongInt2 <- (max(consistenciaa)-min(consistenciaa))/7  
round(LongInt2,4)

## [1] 0.2304

mins2 <- seq(min(consistenciaa),   
 min(consistenciaa)+((LongInt2))\*6,  
 by = LongInt2)

maxs2 <- mins2 + LongInt2-0.0001  
maxs2[7] <- maxs2[7] + 0.0001  
TDFc <- data.frame(  
 min = round(mins2,4),  
 max = round(maxs2,4))  
  
TDFc

## min max  
## 1 0.2625 0.4928  
## 2 0.4929 0.7231  
## 3 0.7232 0.9535  
## 4 0.9536 1.1838  
## 5 1.1839 1.4142  
## 6 1.4143 1.6445  
## 7 1.6446 1.8750

realizamos los conteos:

for(i in 1: nrow(TDFc)){  
   
 TDFc$fi[i] <- length(  
 which(  
TDFc$min[i] <= consistenciaa & consistenciaa <= TDFc$max[i]  
 )  
 )  
   
}  
TDFc

## min max fi  
## 1 0.2625 0.4928 1  
## 2 0.4929 0.7231 1  
## 3 0.7232 0.9535 2  
## 4 0.9536 1.1838 11  
## 5 1.1839 1.4142 44  
## 6 1.4143 1.6445 18  
## 7 1.6446 1.8750 7

TDFc$fr <- TDFc$fi/length(consistenciaa)

for(i in 1:nrow(TDFc)){  
   
 TDFc$Fi[i] <- sum(TDFc$fi[1:i])   
   
}

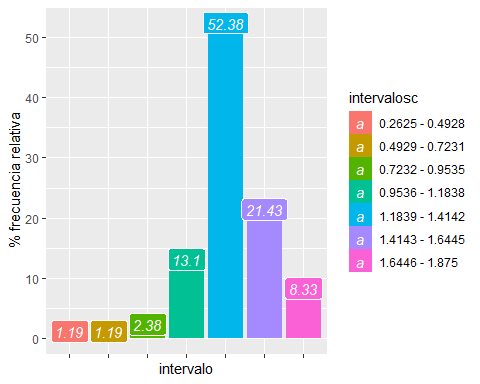
for(i in 1:nrow(TDFc)){  
   
 TDFc$Fr[i] <- sum(TDFc$fr[1:i])   
   
}

TDFc2 <- TDFc  
colnames(TDFc2) <- c('min.',  
 'max.',  
 'f~i~',  
 'f~r~',  
 'F~i~',  
 'F~r~'  
   
 )

knitr::kable(TDFc2, 'simple')

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| min. | max. | fi | fr | Fi | Fr |
| 0.2625 | 0.4928 | 1 | 0.0119048 | 1 | 0.0119048 |
| 0.4929 | 0.7231 | 1 | 0.0119048 | 2 | 0.0238095 |
| 0.7232 | 0.9535 | 2 | 0.0238095 | 4 | 0.0476190 |
| 0.9536 | 1.1838 | 11 | 0.1309524 | 15 | 0.1785714 |
| 1.1839 | 1.4142 | 44 | 0.5238095 | 59 | 0.7023810 |
| 1.4143 | 1.6445 | 18 | 0.2142857 | 77 | 0.9166667 |
| 1.6446 | 1.8750 | 7 | 0.0833333 | 84 | 1.0000000 |

library(ggplot2)  
  
intervalosc <- sort(factor(paste(TDFc$min,'-',TDFc$max),  
 levels = paste(TDFc$min,'-',TDFc$max),) )  
ggplot(TDFc,  
aes(x = reorder(intervalosc,intervalosc, function(x)-length(x)),  
y = fr\*100,  
fill=intervalosc,  
label = round(fr\*100,2)  
)  
) +  
geom\_bar(stat="identity") +  
xlab("intervalo") +  
ylab('% frecuencia relativa')+  
geom\_label(aes(fill = intervalosc),  
colour = "white",  
fontface = "italic") +  
 theme(axis.text.x = element\_blank())

 El mayor porcentaje de frecuencias relativas se encuentra en el intervalo entre 1.1839 y 1.4142. Es decir que las muestras de fresa en su mayoría presentan una consistencia dentro de este rango. A su vez la distribucion tiene cola hacia la izquierda.

1. Construya la matriz de correlaciones de Pearson para las variables gradosbrixa consistenciaa, gradosbrixb consistenciab, gradosbrixc y consistenciac y concluya.

primero revisamos cual variable tienen mas valores faltantes, asi escogeremos las muestras restantes para calcular la correlacion entre variables

which(is.na(fresa$gradosbrixa))

## [1] 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102  
## [19] 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120  
## [37] 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138  
## [55] 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156  
## [73] 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174  
## [91] 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192  
## [109] 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210  
## [127] 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228  
## [145] 229 230 231 232 233 234

which(is.na(fresa$consistenciaa))

## [1] 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102  
## [19] 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120  
## [37] 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138  
## [55] 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156  
## [73] 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174  
## [91] 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192  
## [109] 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210  
## [127] 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228  
## [145] 229 230 231 232 233 234

which(is.na(fresa$gradosbrixb))

## [1] 230 231 232 233 234

which(is.na(fresa$consistenciab))

## [1] 230 231 232 233 234

which(is.na(fresa$gradosbrixc))

## integer(0)

which(is.na(fresa$consistenciac))

## integer(0)

nasCor <- which(is.na(fresa$gradosbrixa))

Una vez sabemos que escoger, podemos proceder a crear la tabla:

df3 <- data.frame(gradosbrixa = fresa$gradosbrixa[-nasCor],  
consistenciaa = fresa$consistenciaa[-nasCor],  
 gradosbrixb = fresa$gradosbrixb[-nasCor],  
consistenciab = fresa$consistenciab[-nasCor],  
 gradosbrixc = fresa$gradosbrixc[-nasCor],  
consistenciac = fresa$consistenciac[-nasCor])  
  
cor(df3)

## gradosbrixa consistenciaa gradosbrixb consistenciab gradosbrixc  
## gradosbrixa 1.00000000 0.11881497 0.181963924 -0.06787461 0.12012597  
## consistenciaa 0.11881497 1.00000000 0.036315066 -0.05049572 -0.09230665  
## gradosbrixb 0.18196392 0.03631507 1.000000000 -0.09071009 0.32372985  
## consistenciab -0.06787461 -0.05049572 -0.090710089 1.00000000 -0.04702223  
## gradosbrixc 0.12012597 -0.09230665 0.323729849 -0.04702223 1.00000000  
## consistenciac 0.03401680 -0.11179596 -0.001546841 0.34529670 0.03066735  
## consistenciac  
## gradosbrixa 0.034016800  
## consistenciaa -0.111795957  
## gradosbrixb -0.001546841  
## consistenciab 0.345296697  
## gradosbrixc 0.030667354  
## consistenciac 1.000000000

Para visualizar mejor la matriz de correlacion podemos utilizar la funcion chart.correlation()

library(PerformanceAnalytics)

## Loading required package: xts

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## Attaching package: 'xts'

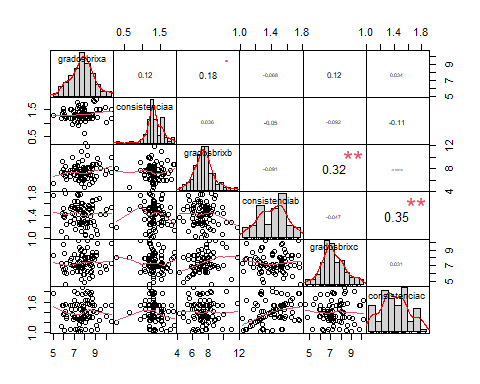
## The following objects are masked from 'package:dplyr':  
##   
## first, last

##   
## Attaching package: 'PerformanceAnalytics'

## The following objects are masked from 'package:e1071':  
##   
## kurtosis, skewness

## The following object is masked from 'package:graphics':  
##   
## legend

chart.Correlation(df3, histogram=TRUE, pch=19)



Las variables que mayor correlación presentan son gradosbrixb con grados brixc, y consistenciab con consistenciac. Esto puede indicar que hay relación entre los experimentos b y c, pero gracias a los graficos de dispersión también vemos que hay mucha dispersión a lo largo de la linea diagonal entre x e y. Es importante anotar que correlación no implica causalidad, y para determinar las relaciones causa y efecto habría que hacer más análisis posteriores, como un diseño experimental.

Gracias a los histogramas podemos ver la unimodalidad para los grados brix en b y c, y multimodalidad en las consistencias b y c.

## 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).

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

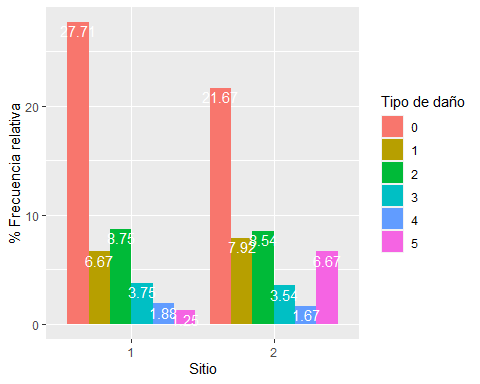
cardones <- read\_excel("./tarea1.xlsx",   
 sheet = "cardones")  
cardones$`Tipo de Daño`<-as.factor(cardones$`Tipo de Daño`)  
cardones$sitio<-as.factor(cardones$sitio)  
attach(cardones)  
CrossTable(`Tipo de Daño`,sitio,prop.chisq=F)

##   
##   
## 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.

1. 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$sitio)) + geom\_bar(position="dodge", stat="identity")+  
xlab("Sitio") +ylab('% Frecuencia relativa')+ scale\_fill\_discrete(name = "Tipo de daño")+ geom\_text(aes(label = round(tableFrecRel$Freq,2) ), vjust = 1.2, colour = "white", position = position\_dodge(.9))



* 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.

1. 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.