Laboratorio 1 - Estadística Computacional

Benjamin Jorquera Jorquera 201473521 − 9 Malla nueva

7 de abril de 2017

1. Virus Zika

(*a*) Elimine todos aquellas filas con datos faltantes en las columna *report date* o *value*

R: Para poder comenzar a analizar los datos en R, primero se debe conocer la ruta de la carpeta contenedora para la lectura de archivos, donde se copiarán las plantillas de Excel (csv) y se abrirán ejecutando la función definida en el script. Debido a la gran cantidad de datos en este problema, se usara la función *summary()*, generando así un dataframe resumido"para llevar a cabo un mejor seguimientos de estos.

Ahora bien, para hacer una limpieza de las filas donde no existe información en las columnas *report date* o *value*, se debe generar un *subset* de los datos, usando las funciones definidas en el script. Luego se hace uso de *complete.cases*.

(*b*) Transforme el tipo de datos de la columna *value* a *numeric*.

R: Primero se debe transformar el tipo de datos de la columna *value* de *integer* a *character*, esto para no causar anomalías en los datos. Luego se podrá transformar a *numeric* y completar los casos faltantes. La función está definida en el script.

(*c*) Elimine las columnas sin información.

R: Se puede observar que los datos de las columnas *time period* y *time period type* no presentan información (NA's), por ende se eliminan completamente usando la función definida en el script.

(*d*) Genere un subset para cada uno de los siguientes países que contenga solo los datos del mismo:

- Estados Unidos
- Argentina
- Brazil
- Colombia
- Mexico

R: Usando las ocurrencias de *string* correspondientes a cada país, se podrán generar los 5 *subsets* correspondientes a los datos llamando a la funcíon definida en el script, simplificando así la muestra.

(*e*) Elimine de sus subsets los datos de aquellas locaciones en las que no hubo casos reportados.

R: Función definida en el script.

(*f*) Redisee sus subsets de forma tal que contengan solo la información de los casos confirmados totales para cada país, es decir, solo los valores pertenecientes a *cumulative confirmed local cases* en la columna de *data field*.

R: Los nuevos subsets son generados a partir de las ocurrencias *confirmed* y *reported* en la columna *data field*, pero evitando el string *microcephaly*. La primeras condiciones se debe a que en los subsets de cada

país la columna *data field* no ocupa la misma sintaxis en sus datos para çonfirmarüna persona infectada de virus Zyka; La segunda condición deriva de que estos no corresponden a un caso de infección del virus, sino de la aparición de microcefalia en el nacimiento de los bebes, producto del efecto colateral del virus en la madre.

Las funciones con sus respectivas expresiones regulares están definidas en el script.

(g) Renombre los países de sus subsets con su nombre correcto, por ejemplo, si dice *United States Virgin Islands*, debe decir simplemente *United States*.

R: Cada valor de la columna *location* es evaluada como *character* y reemplazada por el nombre de su respectivo país, indicado en cada subset, utilizando la función descrita en el script.

(h) Para cada país nombrado anteriormente, obtenga las siguientes medidas de tendencia y dispersión, concluya de acuerdo a lo que le dicen estos estadísticos:

- Media de casos reportados.
- Varianza de la cantidad de casos reportados para cada país.
- Mediana de los casos reportados.
- Cuartiles 1 y 3 de cada país para los casos reportados.

R: Utilizando las funciones descritas en el script, se obtienen de los dataframes de cada subset los siguientes datos:

1. USA:

   Media = 23.40824
   Varianza = 11219.82
   Mediana = 5
   Primer cuartil (25 %) = 2
   Tercer cuartil (75 %) = 12

2. Argentina:

   Media = 5.833333
   Varianza = 167.9985
   Mediana = 2
   Primer cuartil (25 %) = 2
   Tercer cuartil (75 %) = 12

3. Brazil:

   Media = 9446.245
   Varianza = 674471038
   Mediana = 1561.5
   Primer cuartil (25 %) = 401.25
   Tercer cuartil (75 %) = 2652.75

4. Colombia:

   Media = 39.16895
   Varianza = 50580
   Mediana = 3
   Primer cuartil (25 %) = 1
   Tercer cuartil (75 %) = 11

5. Mexico:

   Media = 5.321101
   Varianza = 88.99779
   Mediana = 2
   Primer cuartil (25 %) = 1
   Tercer cuartil (75 %) = 5

(i) Realice un boxplot para cada país de acuerdo al número de casos reportados en las distintas localidades, ubique todos los boxplot en la misma gráfica, recuerde ajustar los ejes y las etiquetas para hacer su grafico más claro y fácil de analizar, se descontará puntaje por gráficos poco prolijos. Concluya sobre lo que ve en los distintos boxplot.

R: Ya que se dispone de un conjunto de datos sobre los afectados del virus Zika, se generan boxplots para representar los casos reportados de cada muestra, estos se observan en (1). Debido a la gran cantidad de habitantes en Brazil, se muestran 2 perspectivas de los boxplots con distinto rango; Se puede concluir evidentemente que Brazil tiene mayor número de afectados por el virus Zika, seguido por Colombia, se puede deducir que gran parte de la población no cumple con los requisitos mínimos de higiene; por otro lado Argentina parece tener la menor cantidad de afectados por el virus.

(j) Para finalizar, realice al menos una conclusión general utilizando los estadísticos y gráficos que obtuvo sobre este dataset, puede realizar más gráficos o investigación en el dataset para apoyar sus conclusiones.

R: En base a la información recaudada, el diagrama de caja o *boxplot* fue de gran ayuda para manejar la gran cantidad de datos proporcionada. Aun así se tuvo que eliminar gran cantidad de *outliers* presentes en el diagrama, también se observa que estos no presentan una distribución simétrica, ya que la mediana no se encuentra en el centro del rectángulo. Finalmente proporciono herramientas para facilitar el análisis descriptivo del *dataset*.

2. Análisis de los sobrevivientes del Titanic.

(a) Genere un subset que contenga solo a hombres y otro solo a mujeres.

R: Se separa la muestra en 2 dataframes. Funciones definidas en el script.

(b) Genere boxplot para las edades de los supervivientes de ambos sexos, uno al lado del otro, concluya a partir de lo que ve.

R: Primero se limpian los datos obteniendo solo los supervivientes al incidente, luego se analiza el diagrama de *boxplots* en (2). Ambos tienden aproximadamente a los mismos resultados de acuerdo a la edad, sin embargo, se puede deducir que las mujeres de mayor edad tuvieron mejores oportunidades para sobrevivir, así mismo con los hombres mas jóvenes; También se puede concluir debido al tamaño del *boxplot* que las mujeres tuvieron mayor probabilidades de sobrevivir, ya que probablemente fueron más inteligentes al momento de tomar decisiones o puede que se haya priorizado su supervivencia por encima de la de los hombres.

(c) Construya un histograma para los precios pagados por todos aquellos supervivientes, haga que su histograma sea de frecuencias relativas, de forma de poder visualizar la probabilidad de sobrevivir de acuerdo a la cantidad de dinero pagado.

R: Se adjunta el histograma en (3) del anexo y su respectiva función en el script.

(d) Elimine aquellos registros que no contienen valor para la variable Cabin.

R: Función definida en el script.

(e) Obtenga todas las cabinas para las cuales se tiene registro, para simplificar este resultado, no haga distinción entre los números de cabina, solo la letra de esta, por ejemplo, la cabina C123 debe ser considerada simplemente como una cabina de tipo "C".

R: Discriminando el número adjunto a cada cabina, se obtiene que en el barco solo habían 3 tipos: A, B y C. Las funciones están definidas en el script.

(f) Genere un gráfico de barras que muestre todos los supervivientes para cada tipo de cabina.

R: Se adjunta el grafico de barras en (4) del anexo y su respectiva función en el script.

(g) Para finalizar, realice al menos una conclusión general utilizando los estadísticos y gráficos que realizó sobre este dataset, puede realizar más gráficos o investigación en el dataset para apoyar sus conclusiones.

R: Se puede concluir en base a los gráficos analizados anteriormente, que los pasajeros a bordo del Titanic que viajaban en las cabinas de bajos más precios tenían menos posibilidades de sobrevivir, de hecho, casi nulas.

3. Anexo.

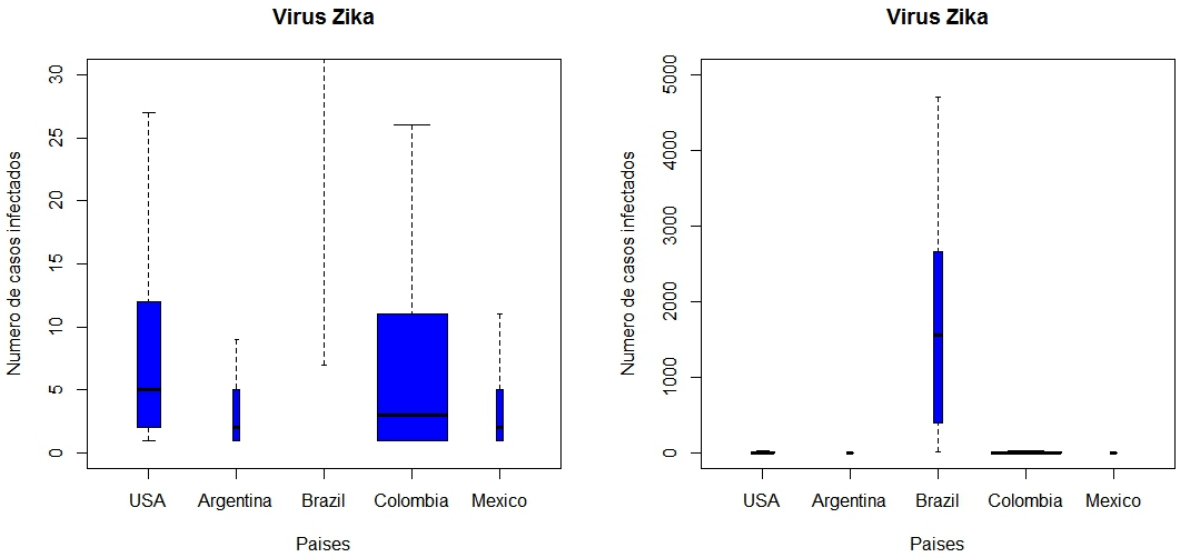

Figura 1: Boxplots subsets de paises. Casos infectados con virus Zika

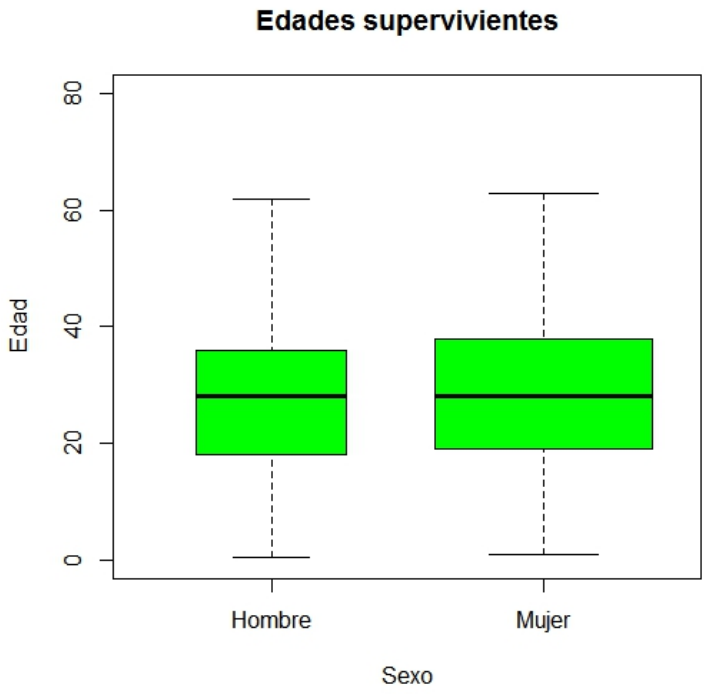

Figura 2: Boxplot sobrevivientes: Hombres y Mujeres.

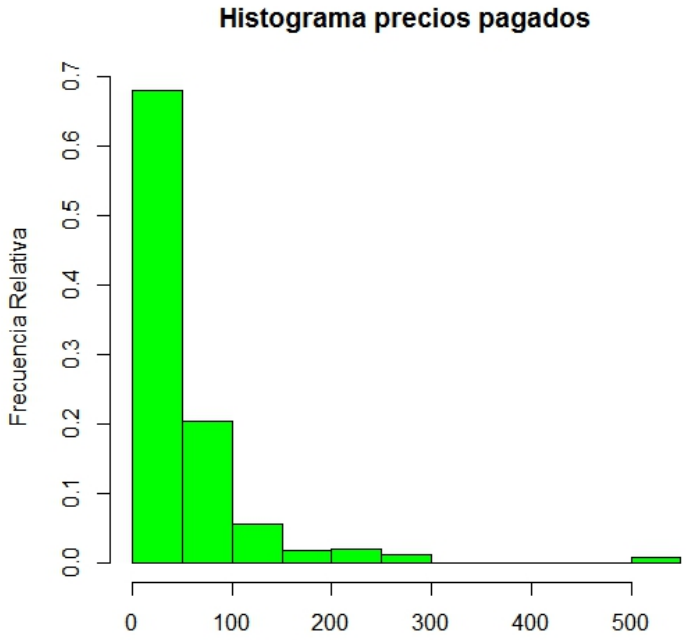

Figura 3: Histograma casos de sobrevivientes por precio pagado de cabina.

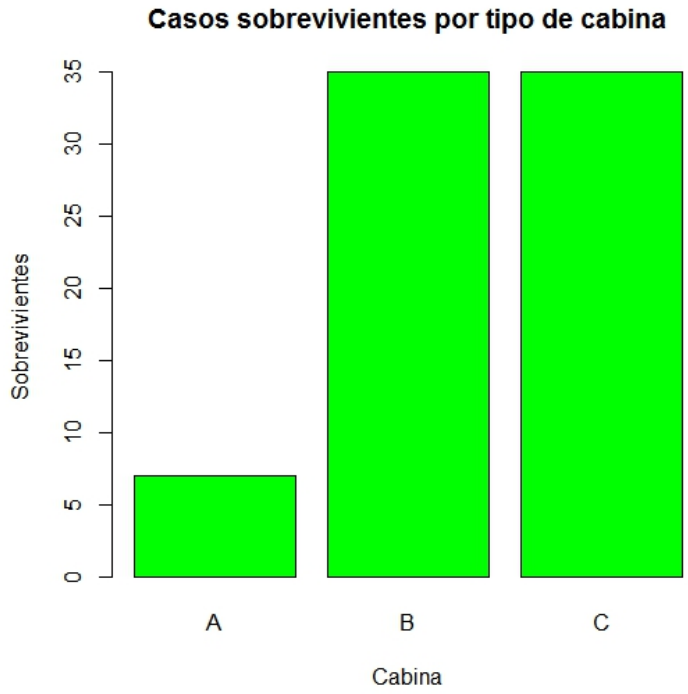

Figura 4: Diagrama de barras sobrevivientes v/s tipo de cabina.

CÓDIGOS

- Pregunta 1

```
getwd()
Datos = read.csv(file=çdc_zika.csv", sep = ",")
summary(Datos)

(a)
Datos = subset(Datos, value != ¨¨ & report_date != ¨¨ )
Datos = Datos[complete.cases(Datos$value),]
Datos = Datos[complete.cases(Datos$report_date),]

(b)
Datos$value = as.numeric(as.character(Datos$value))
Datos = Datos[complete.cases(Datos$value),]

(c)
Datos = Datos[,!sapply(Datos,function(x) any(is.na(x)))]

(e)
Datos.sub.usa = Datos[grep(Ünited_States", Datos$location),]
Datos.sub.arg = Datos[grep(.ªrgentina", Datos$location),]
Datos.sub.bra = Datos[grep("Brazil", Datos$location),]
Datos.sub.col = Datos[grep(Çolombia", Datos$location),]
Datos.sub.mex = Datos[grep("Mexico", Datos$location),]
```

(f)
Datos.sub.usa = subset(Datos.sub.usa, value != 0)
Datos.sub.arg = subset(Datos.sub.arg, value != 0)
Datos.sub.bra = subset(Datos.sub.bra, value != 0)
Datos.sub.col = subset(Datos.sub.col, value != 0)
Datos.sub.mex = subset(Datos.sub.mex, value != 0)

(g)
Datos.sub.usa = Datos.sub.usa[grepl(çonfirmed", Datos.sub.usa$data_field) & !grepl("microcephaly", Datos.sub.usa$data_field) — grepl(reported", Datos.sub.usa$data_field),]

Datos.sub.arg = Datos.sub.arg[grepl(çonfirmed", Datos.sub.arg$data_field) & !grepl("microcephaly", Datos.sub.arg$data_field) — grepl(reported", Datos.sub.arg$data_field),]

Datos.sub.bra = Datos.sub.bra[grepl(çonfirmed", Datos.sub.bra$data_field) & !grepl("microcephaly", Datos.sub.bra$data_field) — grepl(reported", Datos.sub.bra$data_field),]

Datos.sub.col = Datos.sub.col[grepl(çonfirmed", Datos.sub.col$data_field) & !grepl("microcephaly", Datos.sub.col$data_field) — grepl(reported", Datos.sub.col$data_field),]

Datos.sub.mex = Datos.sub.mex[grepl(çonfirmed", Datos.sub.mex$data_field) & !grepl("microcephaly", Datos.sub.mex$data_field) — grepl(reported", Datos.sub.mex$data_field),]

(h)
Datos.sub.usa$location = as.character(Datos.sub.usa$location)
Datos.sub.usa$location = Ünited_States"
Datos.sub.usa$location = as.factor(Datos.sub.usa$location)

Datos.sub.arg$location = as.character(Datos.sub.arg$location)
Datos.sub.arg$location = .ªrgentina"
Datos.sub.arg$location = as.factor(Datos.sub.arg$location)

Datos.sub.bra$location = as.character(Datos.sub.bra$location)
Datos.sub.bra$location = "Brazil"
Datos.sub.bra$location = as.factor(Datos.sub.bra$location)

Datos.sub.col$location = as.character(Datos.sub.col$location)
Datos.sub.col$location = Çolombia"
Datos.sub.col$location = as.factor(Datos.sub.col$location)

Datos.sub.mex$location = as.character(Datos.sub.mex$location)
Datos.sub.mex$location = "Mexico"
Datos.sub.mex$location = as.factor(Datos.sub.mex$location)

(i)
mean(Datos.sub.usa$value)
var(Datos.sub.usa$value)
median(Datos.sub.usa$value)
quantile(Datos.sub.usa$value)

mean(Datos.sub.arg$value)
var(Datos.sub.arg$value)

median(Datos.sub.arg$value)
quantile(Datos.sub.usa$value)

mean(Datos.sub.bra$value)
var(Datos.sub.bra$value)
median(Datos.sub.bra$value)
quantile(Datos.sub.bra$value)

mean(Datos.sub.col$value)
var(Datos.sub.col$value)
median(Datos.sub.col$value)
quantile(Datos.sub.col$value)

mean(Datos.sub.mex$value)
var(Datos.sub.mex$value)
median(Datos.sub.mex$value)
quantile(Datos.sub.mex$value)

(j)
boxplot(Datos.sub.usa$value, Datos.sub.arg$value, Datos.sub.bra$value, Datos.sub.col$value, Datos.sub.mex$value, main="Virus Zika", xlab = "Paises", ylab = "Numero de casos infectados", outline = FALSE ,ylim=c(0,30), col = c("blue"), varwidth = TRUE, names = c(ÜSA",.ªrgentina","Brazil",Çolombia","Mexico"))

boxplot(Datos.sub.usa$value, Datos.sub.arg$value, Datos.sub.bra$value, Datos.sub.col$value, Datos.sub.mex$value, main="Virus Zika", xlab = "Paises", ylab = "Numero de casos infectados", outline = FALSE ,ylim=c(0,5000), col = c("blue"), varwidth = TRUE, names = c(ÜSA",.ªrgentina","Brazil",Çolombia","Mexico"))

- Pregunta 2

(a)
Datos = read.csv(file="titanic_data.csv", sep = ",")
Datos = subset(Datos, Survived != 0)
Datos.sub.h = subset(Datos, Sex == "male")
Datos.sub.m = subset(Datos, Sex == "female")

(b)
boxplot(Datos.sub.h$Age, Datos.sub.m$Age, main = .ªdades supervivientes", xlab = "Sexo", ylab = .ªdad", outline = FALSE ,ylim=c(0,80), col = c("green"), varwidth = TRUE, names = c("Hombre", "Mujer"))

(d)
g = hist(Datos$Fare, plot=F)
g$counts = g$counts / sum(g$counts)
plot(g, freq=TRUE, ylab="Frecuencia Relativa", main = "Histograma precios pagados", xlab = , col=c("green"))

(e)
Datos = subset(Datos, Cabin != ¨¨)

(f)
Datos.sub.cab1 = Datos[grep(.ª", Datos$Cabin),]
Datos.sub.cab2 = Datos[grep("B", Datos$Cabin),]
Datos.sub.cab3 = Datos[grep(Ç", Datos$Cabin),]

(g)
L = matrix(c(7,35,35),ncol=3,byrow=TRUE)
colnames(L) = c(.ª", "B", Ç")
barplot(L, main=Çasos sobrevivientes por tipo de cabina", xlab = Çabina", ylab = "Sobrevivientes", col = c("green"))