Nombre: *Grupo:*

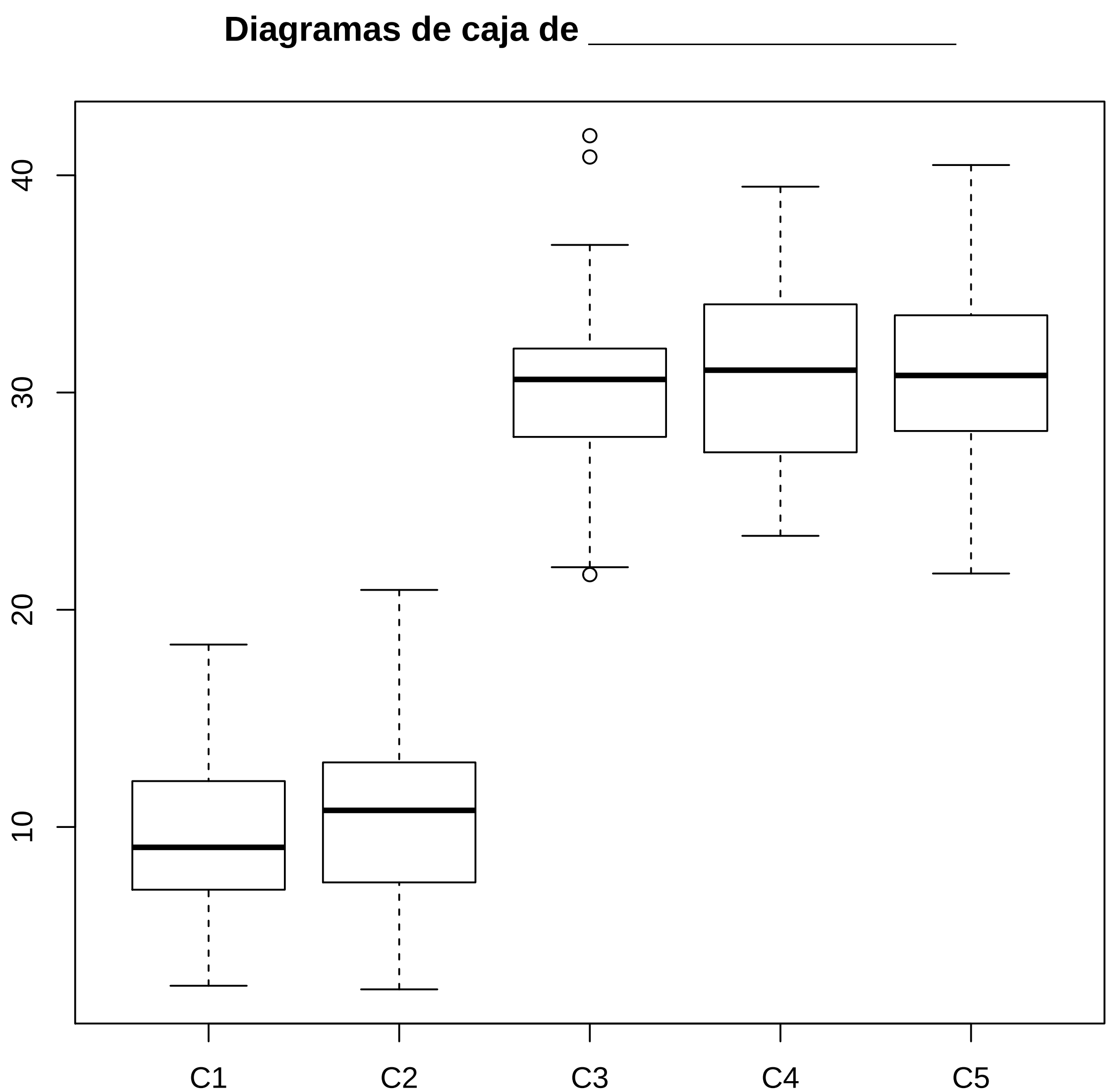
Matemáticas III. GMAT. Control 2 11 junio 2017-2018. Ejercicios.

1) El `data frame` `datos_vuelos` contiene información del retraso en minutos de vuelos de varias compañías aéreas diferentes.

```
head(datos_vuelos)

##     retraso compania
## 1  8.308064       C1
## 2  3.800487       C1
## 3  9.742283       C1
## 4 11.083525       C1
## 5 16.941135       C1
## 6  8.941155       C1

str(datos_vuelos)

## 'data.frame': 250 obs. of  2 variables:
##  $ retraso : num  8.31 3.8 9.74 11.08 16.94 ...
##  $ compania: Factor w/ 5 levels "C1","C2","C3",..: 1 1 1 1 1 1 1 1 1 1 ...

anova_sol=aov(retraso~compania,data=datos_vuelos)
summary(anova_sol)

##              Df Sum Sq Mean Sq F value Pr(>F)
## compania      4  25174    6293   375.5 <2e-16 ***
## Residuals   245   4106      17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

pairwise.t.test(datos_vuelos$retraso,datos_vuelos$compania,"none")

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  datos_vuelos$retraso and datos_vuelos$compania
##
##    C1     C2     C3   C4
## C2 0.32   -      -    -
## C3 <2e-16 <2e-16 -    -
## C4 <2e-16 <2e-16 0.52 -
## C5 <2e-16 <2e-16 0.59 0.91
##
## P value adjustment method: none

library(agricolae)
duncan.test(anova_sol,"compania",group=TRUE)$groups

##      retraso groups
## C4 30.766867      a
## C5 30.671788      a
## C3 30.235084      a
## C2 10.490940      b
## C1  9.678596      b
```

```
duncan.test(anova_sol,"compania",group=FALSE)$comparison

##           difference pvalue signif.        LCL         UCL
## C1 - C2  -0.81234391 0.3221          -2.425113   0.8004253
## C1 - C3 -20.55648850 0.0000     *** -22.254224 -18.8587532
## C1 - C4 -21.08827137 0.0000     *** -22.884526 -19.2920168
## C1 - C5 -20.99319169 0.0000     *** -22.747649 -19.2387340
## C2 - C3 -19.74414459 0.0000     *** -21.356914 -18.1313754
## C2 - C4 -20.27592746 0.0000     *** -22.030385 -18.5214698
## C2 - C5 -20.18084778 0.0000     *** -21.878583 -18.4831125
## C3 - C4  -0.53178287 0.5449          -2.229518   1.1659524
## C3 - C5  -0.43670319 0.5943          -2.049472   1.1760660
## C4 - C5   0.09507968 0.9077          -1.517690   1.7078489

library(car)

## Loading required package: carData

leveneTest(datos_vuelos$retraso,datos_vuelos$compania)

## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  0.3552 0.8403
##       245

bartlett.test(datos_vuelos$retraso,datos_vuelos$compania)

##
##  Bartlett test of homogeneity of variances
##
## data:  datos_vuelos$retraso and datos_vuelos$compania
## Bartlett's K-squared = 0.38658, df = 4, p-value = 0.9836

library(nortest)
sapply(levels(datos_vuelos$compania),FUN=function(x){
  lillie.test(datos_vuelos[datos_vuelos$compania==x,"retraso"])}
  )

##           C1
## statistic 0.09160773
## p.value   0.3683079
## method    "Lilliefors (Kolmogorov-Smirnov) normality test"
## data.name "datos_vuelos[datos_vuelos$compania == x, "retraso"]"
##           C2
## statistic 0.07794665
## p.value   0.6287935
## method    "Lilliefors (Kolmogorov-Smirnov) normality test"
## data.name "datos_vuelos[datos_vuelos$compania == x, "retraso"]"
##           C3
## statistic 0.09137701
## p.value   0.3722495
## method    "Lilliefors (Kolmogorov-Smirnov) normality test"
## data.name "datos_vuelos[datos_vuelos$compania == x, "retraso"]"
##           C4
## statistic 0.08060674
## p.value   0.5754615
```

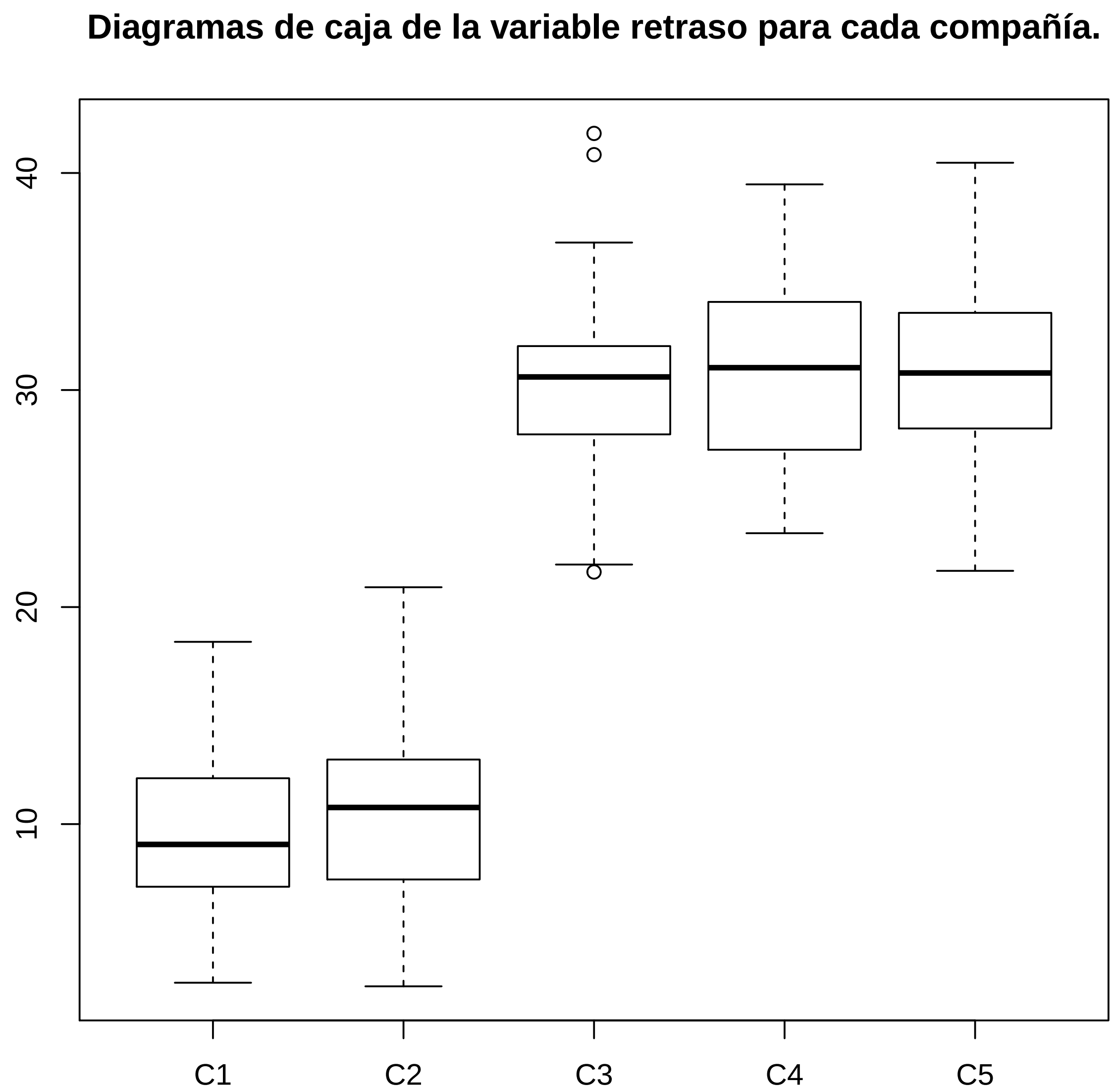
```
## method    "Lilliefors (Kolmogorov-Smirnov) normality test"
## data.name "datos_vuelos[datos_vuelos$compania == x, "retraso"]"
##           C5
## statistic 0.05257725
## p.value   0.9799692
## method    "Lilliefors (Kolmogorov-Smirnov) normality test"
## data.name "datos_vuelos[datos_vuelos$compania == x, "retraso"]"

boxplot(datos_vuelos$retraso~datos_vuelos$compania,
        main="Diagramas de caja de ___________________")
```



Contestad a las siguientes cuestiones justificando que parte del código utilizáis

1) Interpretar y poner un título adecuado al diagrama de cajas ¿Qué nos dice el diagrama sobre la igualdad de medias del retraso? (**0.5 puntos**)

```r
boxplot(datos_vuelos$retraso~datos_vuelos$compania,
        main="Diagramas de caja de la variable retraso para cada compañía.")
```



En este gráfico se muestran los diagramas de caja para cada una de las compañías.

Se observa que la dispersión medida gráficamente por la altura de la caja (diferencia ente el tercer y el primer cuartil) es semejante entre las ciudades, quizá la C_4 tiene mayor dispersión.

Respecto a la igualdad de medias por demos decir que las dos primeras compañías tienen retrasos menores y mediana menores que las compañías C_3, C_4, C_5, C_6. Estas últimas compañía presentan diagramas de caja semejantes y medianas semejantes. Todo esto nos hace sospechar que las distribuciones de los retrasos de las dos primeras compañías son similares y menores que las de las otras tres compañías que a su vez parecen tener distribuciones similares.

2) Escribid hipótesis del contraste de ANOVA y discutid si se cumplen las condiciones necesarias para realizarlo. (**0.5 puntos**)

Las condiciones del ANOVA de efectos fijos son una muestra aleatoria simple para cada nivel del factor, poblaciones normales para la muestra de cada nivel del factor y de la misma varianza (homocedásticidad).

La condición del muestra alatoria simple viene dada por el diseño experimalntal del enunciado. La normalidad de las muestras para cada nivel del factor se comprueba con el test de Lilliefors (con la función `lillie` de la librería `nortest`), todos los p-valores son altos, el más pequeño es del 0rden de 0.07 para el nivel C_2. Así que no hay evidencias fuertes para rechazar la normalidad de las distribuciones en cada ciudad.

La igualdad de varianzas entre ciudades se resuleve con el test de Levene con la función `levene.tes` de la librería `car` el que se obtine un p-valor alto del orden de 0.984 conformado por el test de homogeneidad de varianzas de Bartlett (función `bartlett.test`) con un p-valor del orden de 0.38

Así que o hay evidencias fuertes en contra de la hocedasticidad y normalidad de cada muestra.

3) Escribid la tabla (estándar, la de los apuntes) del ANOVA con toda la información de qué es y cómo se calcula cada valor. Concluid en base a ello el resultado del ANOVA (**0.5 puntos**)

4) Sea cual sea el resultado del ANOVA, realizad el ajuste por Bonferroni para $\alpha = 0.1$ y discutid los resultados obtenidos a partir la salida del código. (**0.5 puntos**)

5) Discutid el resultado de la salida del código del test de Duncan. (**0.5 puntos**)

Solución:

2) Para estudiar si hay evidencia de que el retraso de un vuelo en la salida aumenta el retraso de su llegada se toma una muestra aleatoria simple de 100 vuelos y se anota para cada vuelo si tuvo retraso en la salida y en la llegada (en minutos). La tabla siguiente resume los resultados:

Salida /Llegada	No Retraso	Retraso
No Retraso	75	15
Retraso	6	4

1) Plantear un contraste de igualdad de proporciones entre la proporción de vuelos retrasados en la salida y en la llegada. ¿Qué diseño experimental estamos utilizando? (**0.5 puntos.**)

2) Resolver el contraste al nivel de significación $\alpha = 0.1$ (**0.5 puntos.**)

3) Calcular el p-valor del contraste anterior. (**0.5 puntos.**)

4) Calcular e interpretar un intervalo de confianza para la diferencia de proporciones al nivel 99 %. (**0.5 puntos.**)

```
set.seed(2018)
salida =rbinom(100,size=1,prob=0.1)
llegada=rbinom(100,size=1,prob=0.2)
aux=table(salida,llegada)
b=aux[2,1]
b

## [1] 6

d=aux[1,2]
d

## [1] 15

n=sum(aux)
n

## [1] 100

t=(b/n-d/n)/sqrt((b+d)/n^2)
t

## [1] -1.963961

2*(1-pt(abs(t),100-1))

## [1] 0.05233902

t^2

## [1] 3.857143
```

```
mcnemar.test(aux,correct=FALSE)

##
##  McNemar's Chi-squared test
##
```

```
## data:  aux
## McNemar's chi-squared = 3.8571, df = 1, p-value = 0.04953

mcnemar.test(aux,correct=FALSE)$statistic

## McNemar's chi-squared
##              3.857143
```

3) Se piensa que el tiempo en segundos transcurrido entre dos reservas de vuelos de avión en un mismo día podría seguir una distribución exponencial con una reserva cada cinco segundos. Se toma una muestra de 10 tiempos en segundos.

Vuelo	1	2	3	4	5	6	7	8	9	10
Retraso	0.50	1.40	1.60	2.20	2.40	3.70	3.90	4.50	5.20	7.10

1) ¿Cuál es y qué parámetros tiene la función de distribución teórica propuesta? Escribid correctamente la función de distribucion. (**0.5 puntos**)

2) Contrastar la hipótesis del enunciado con el test KS, al nivel de significación $\alpha = 0.1$. (**1 puntos**)

```
set.seed(2018)
datos=sort(round(rexp(10,1/5),1))
ks.test(datos,"pexp",1/5)

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  datos
## D = 0.25345, p-value = 0.4669
## alternative hypothesis: two-sided

teoricas=pexp(datos,1/5)
teoricas

##  [1] 0.09516258 0.24421626 0.27385096 0.35596358 0.38121661 0.52288608
##  [7] 0.54159399 0.59343034 0.64654532 0.75828598

obs=(1:10)/10
obs

##  [1] 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

abs(teoricas- ((1:10)-1)/10)

##  [1] 0.09516258 0.14421626 0.07385096 0.05596358 0.01878339 0.02288608
##  [7] 0.05840601 0.10656966 0.15345468 0.14171402

abs(teoricas- (1:10)/10)

##  [1] 0.004837418 0.044216259 0.026149037 0.044036421 0.118783392
##  [6] 0.077113916 0.158406011 0.206569660 0.253454682 0.241714017

D=pmax(abs(teoricas- ((1:10)-1)/10),abs(teoricas- (1:10)/10))
D
```

```
##  [1] 0.09516258 0.14421626 0.07385096 0.05596358 0.11878339 0.07711392
##  [7] 0.15840601 0.20656966 0.25345468 0.24171402

max(D)

## [1] 0.2534547
```

4) La siguiente tabla contiene los valores de `retraso_llegada, retraso_salida` y `distancia` del trayecto del vuelo para cuatro vuelos. Las distancias vienen dadas en centenas de kilómetros y los retrasos en decenas de minutos.

```
df=data.frame(retraso_llegada,retraso_salida,distancia)
df

##   retraso_llegada retraso_salida distancia
## 1              27              3        10
## 2              -9             -1        15
## 3              18              2        20
## 4              46              5         5

X=cbind(rep(1,4),df$retraso_salida,df$distancia)
X

##      [,1] [,2] [,3]
## [1,]    1    3   10
## [2,]    1   -1   15
## [3,]    1    2   20
## [4,]    1    5    5

Y=matrix(df$retraso_llegada,ncol=1)
Y

##      [,1]
## [1,]   27
## [2,]   -9
## [3,]   18
## [4,]   46

t(X)%*%X

##      [,1] [,2] [,3]
## [1,]    4    9   50
## [2,]    9   39   80
## [3,]   50   80  750

det(t(X)%*%X)

## [1] 5150

solve(t(X)%*%X)

##            [,1]        [,2]        [,3]
## [1,]  4.4368932 -0.53398058 -0.23883495
## [2,] -0.5339806  0.09708738  0.02524272
## [3,] -0.2388350  0.02524272  0.01456311
```

```
solve(t(X)%*%X)%*%t(X)%*%Y

##              [,1]
## [1,]  0.57281553
## [2,]  9.07766990
## [3,] -0.03980583

X%*%solve(t(X)%*%X)%*%t(X)%*%Y

##            [,1]
## [1,] 27.407767
## [2,] -9.101942
## [3,] 17.932039
## [4,] 45.762136

sum((X%*%solve(t(X)%*%X)%*%t(X)%*%Y)^2)

## [1] 3249.762
```

Usad el código anterior cuando pertoque para contestar a las siguientes preguntas.

1) Escribid y explicad la ecuación del modelo de regresión lineal múltiple que predice el `retraso_llegada` a partir de las otras dos variables. (**0.5 puntos.**)

2) Calcular R^2 y R^2 ajustado de la anterior regresión. (**0.5 puntos.**)

```
sumYhat_square=sum((X%*%solve(t(X)%*%X)%*%t(X)%*%Y)^2)# ya se daba
meanY=mean(Y)# a mano
SST=4*(sum(Y^2)/4-mean(Y)^2)#a mano
SST

## [1] 1569

Error=Y-X%*%solve(t(X)%*%X)%*%t(X)%*%Y
SSR=4*(sumYhat_square/4-meanY^2)
SSR

## [1] 1568.762

R2=SSR/SST
R2

## [1] 0.9998484

summary(sol)$r.squared

## [1] 0.9998484

R2adj=1-(1-R2)*(4-1)/(4-2-1)
R2adj

## [1] 0.9995452

summary(sol)$adj.r.squared

## [1] 0.9995452
```

3) Calcula el AIC de este modelo. (**0.5 puntos.**)

```
SSE=SST-SSR
SSE

## [1] 0.2378641

AIC_value=4*log(SSE/4)+2*2
AIC_value

## [1] -7.289401
```

4) Calcular el intervalo de confianza al 95 % para el coeficiente de la variable distancia ¿Qué se puede deducir de su presencia en el modelo? (**0.5 puntos.**)

```
confint(sol)

##                      2.5 %     97.5 %
## (Intercept)    -12.4804678 13.6260989
## retraso_salida   7.1467615 11.0085783
## distancia       -0.7876434  0.7080318
```

```
sol$coefficients[3]+c(-1,1)*qt(0.975,4-2-1)*
  sqrt(SSE/(4-2-1)*solve(t(X)%*%X)[3,3])

## [1] -0.7876434  0.7080318
```