Evaluación 4- Diseño de un factor

Diana Laura Javier Garcia

11/11/2020

# Diseño con un factor

Ejemplo 10.4, 10.5 y 10.6

Considere la situación del kilometraje de la gasolina. Sea , y las medias y sean , y las desviaciones estándar de las poblaciones de todos los posibles kilómetrajes de gasolina usando gasolina de los tipos A, B y C.Para estimar estas medias y desviaciones estándar, North American Oil ha empleado una diseño experimental aleatorio y ha obtenido las muestras de kilometraje en la Tabla 10.1. Las medias de estas muestras , y — son las estimaciones puntuales de , y . Las desviaciones estándar de estas muestras , y — son las estimaciones puntuales de , y . Con estas estimaciones puntuales, se probará si existen diferencias estadísticamente significativas entre las medias de tratamiento , y . Si existen tales diferencias, se estimará la magnitud de estas diferencias. Esto permitirá a North American Oil juzgar si estas diferencias tienen importancia práctica.

Como el enunciado lo indica, este experimento constiste en evaluar el kilometraje de tres tipos de gasolina con el objetivo de saber si las difrencias entre estas son significativas.

Para tener una primer idea del comportamiento de los datos, y a manera de corroborar la media y la desviación estandar, se realizó el análisis descriptivo. Para ello los datos se reescribieron en un documento de excel con extensiòn *csv*, el cual se nombre *mileage*, de esta manera se leyeron los datos:

mileage<- read.csv("mileage.csv")  
names(mileage)

## [1] "n" "A" "B" "C"

Organizando los datos en forma vertical, en las columnas *gasolina* y *kilometraje*:

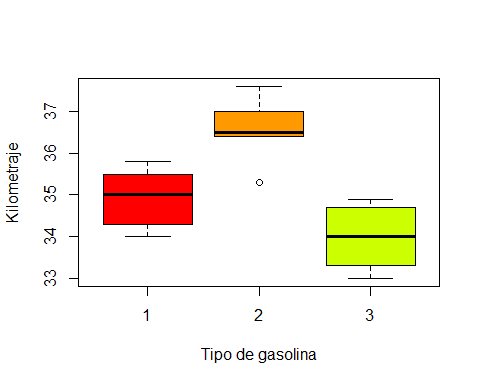
mileage.vert<-data.frame(gasolina=gl(3,5),kilometraje=c(mileage$A,mileage$B,mileage$C))  
  
mileage.vert

## gasolina kilometraje  
## 1 1 34.0  
## 2 1 35.0  
## 3 1 34.3  
## 4 1 35.5  
## 5 1 35.8  
## 6 2 35.3  
## 7 2 36.5  
## 8 2 36.4  
## 9 2 37.0  
## 10 2 37.6  
## 11 3 33.3  
## 12 3 34.0  
## 13 3 34.7  
## 14 3 33.0  
## 15 3 34.9

Posteriormente se generó el diagrama de caja de cada una de las muestras, donde 1 se refiere a la gasolina A, 2 a la B, y 3 a la C. De los diagramas de caja se observa que el gasolina 2 presenta la mayor media, seguida de la gasolina tipo 1 y por último la gasolina tipo 3. Es importante notar que la gasolina 1 solo presenta un bigote, lo cual estaría asociado a una dispersiòn de los datos entre la media y el máximo. Sin embargo, es importante notar que la dispersión de la gasolina 1 y 3 son similares, lo cual es un buen indicativo de la presición de los datos.

Por otra parte, retomando el objetivo de este trabajo, se pueden observar diferencias claras entre las muestras, sin embargo, aunque parece que son significativas, hasta este punto no se puede concluir. El estudio de las diferencias se analizó en la siguiente sección.

boxplot(kilometraje ~ gasolina, xlab = "Tipo de gasolina", ylab = "Kilometraje", data = mileage.vert, col = rainbow(10))



A fin de corroborar númericamente las observaciones anteriores, se obtuvo el resumen estadístico de las muestras como se presenta a continuación:

summary(mileage)

## n A B C   
## Min. :1 Min. :34.00 Min. :35.30 Min. :33.00   
## 1st Qu.:2 1st Qu.:34.30 1st Qu.:36.40 1st Qu.:33.30   
## Median :3 Median :35.00 Median :36.50 Median :34.00   
## Mean :3 Mean :34.92 Mean :36.56 Mean :33.98   
## 3rd Qu.:4 3rd Qu.:35.50 3rd Qu.:37.00 3rd Qu.:34.70   
## Max. :5 Max. :35.80 Max. :37.60 Max. :34.90

Con la desviación estandar:

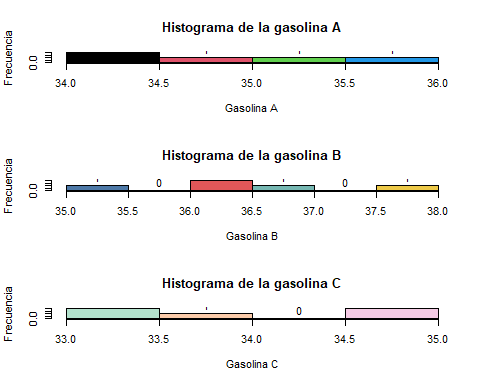
sapply(mileage,sd)

## n A B C   
## 1.5811388 0.7661593 0.8502941 0.8348653

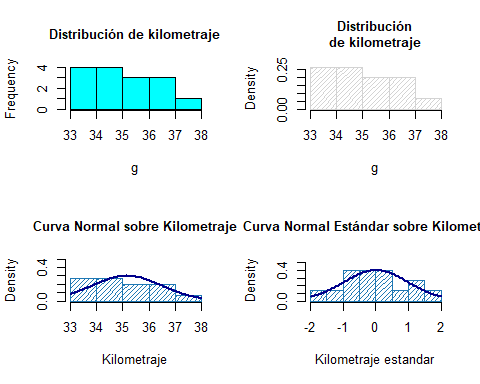
La primera observación que se tiene es que estos resultados coinciden con los presentados en el enunciado. Y tal como se observó en el diagrama de caja, la desviaciones entre las muestras son muy parecidas, lo que está asociado a una distribuión similar de los datos entre los cuartiles.

Por otra parte, si se analizan los histogramas de cada una de las muestras se observa que, la muestra 1, con 4 clases, presentan la misma frecuencia, a excepciòn de la primera clase que tiene una frecuencia de 2. A diferencia de esta, las muestras 2 y 3 presentan frecuencias muy diferentes en cada una de sus muestras. Ahora bien, se analiza en conjunto el kilometraje, como se muestra en la siguiente figura, se observa que tanto la frecuencia absoluta como la relativa se ajustan a una distribución normal. La prueba de normalidad se realiza en la siguiente sección.

par(mfrow=c(3,1))  
hist(mileage$A, main = "Histograma de la gasolina A", ylab = "Frecuencia", xlab = "Gasolina A",col=palette("Tableau 10"), labels = TRUE)  
hist(mileage$B, main = "Histograma de la gasolina B", ylab = "Frecuencia", xlab = "Gasolina B",col=palette("Pastel 2"), labels = TRUE)  
hist(mileage$C, main = "Histograma de la gasolina C", ylab = "Frecuencia", xlab = "Gasolina C",col=palette("Paired"), labels = TRUE)

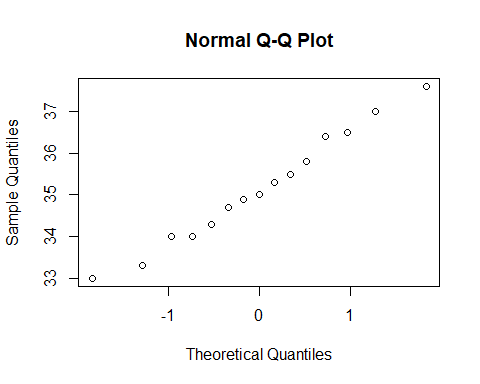


par (mfrow=c(2,2))  
g = mileage.vert$kilometraje  
m<-mean(g)  
std<-sqrt(var(g))  
  
hist(g, breaks = "Sturges", col = "cyan", main= "Distribución de kilometraje",  
cex.main=1.0)  
  
hist(g, density=30, breaks = "Sturges", prob=TRUE, main= "Distribución  
de kilometraje", cex.main=1.0)  
  
hist(g, density=30, prob=TRUE, col = 12,  
xlab="Kilometraje", ylim=c(0, 0.5),  
main="Curva Normal sobre Kilometraje", cex.main=1.0)  
  
curve(dnorm(x, mean=m, sd=std),  
col="darkblue", lwd=2, add=TRUE, yaxt="n")  
zg<-(g-m)/std  
  
hist(zg, density=30, prob=TRUE, col = 12,  
xlab="Kilometraje estandar", ylim=c(0, 0.55),  
main="Curva Normal Estándar sobre Kilometraje", cex.main=1.0)  
curve(dnorm(x, mean=0, sd=1),  
col="darkblue", lwd=2, add=TRUE, yaxt="n")

 ## Varianza y normalidad

Para estudiar normalidad de los datos se realizó se obtuvo la grafica cuantil-cualtil que permite observar cuan cerca está la distribución de un conjutno de datos a alguna distribución. Además se realizó la prueba de Kolmorogov y de Shapiro Wilk. Ambas pruebas plantean la hipótesis nula que una muestra proviene de una distribución normal, y por ende plantea una hipótesis alternativa que sostiene que la distribución no es normal. Para este trabajo se encontrò que dado que el p-valor de ambas es mayor a 0.05, se puede decir que los datos se distribuyen normalmente, tal como se estimó anteriormente.

qqnorm(mileage.vert$kilometraje)



### Prueba de Shapiro Wilk

sw<-shapiro.test(mileage.vert$kilometraje)  
print(sw)

##   
## Shapiro-Wilk normality test  
##   
## data: mileage.vert$kilometraje  
## W = 0.98013, p-value = 0.9704

### Prueba de Kolmogorov

ks<-ks.test(mileage.vert$kilometraje, "pnorm", mean=mean(mileage.vert$kilometraje),sd=sd(mileage.vert$kilometraje))

## Warning in ks.test(mileage.vert$kilometraje, "pnorm", mean =  
## mean(mileage.vert$kilometraje), : ties should not be present for the Kolmogorov-  
## Smirnov test

print(ks)

##   
## One-sample Kolmogorov-Smirnov test  
##   
## data: mileage.vert$kilometraje  
## D = 0.090864, p-value = 0.9997  
## alternative hypothesis: two-sided

### Prueba de Bartlett

Para verificar la igualdad de varianzas se aplicó la Prueba de Bartlett, la cual permite contrastar la igualdad de varianza en 2 o más poblaciones sin necesidad de que el tamaño de los grupos sea el mismo. Como se puede observar, el p-valor es mayor a 0.05, por lo tanto se cumple la hipotesis de que las varianzas no presentan diferencias significativas.

bartlett.test(kilometraje ~ gasolina, data=mileage.vert)

##   
## Bartlett test of homogeneity of variances  
##   
## data: kilometraje by gasolina  
## Bartlett's K-squared = 0.043678, df = 2, p-value = 0.9784

## ANOVA

En esta sección se presentan los resultados del ANOVA. En este punto, conviene recordar las hipótesis del análisis. En primer lugar, se parte de que la hipótesis nula es que la media de la variable estudiada es la misma en los diferentes grupos, en contra posición a la hipótesis alternativa de que al menos dos medias son diferentes. La instrucción para realizar el ANOVA es: *aov*. En este trabajo se guarda la instrucción como *modelo*. Y posteriormente , los resultados se ven con el comando *summary*. En términos matemáticos:

modelo <- aov(kilometraje ~ gasolina, data = mileage.vert)  
modelo

## Call:  
## aov(formula = kilometraje ~ gasolina, data = mileage.vert)  
##   
## Terms:  
## gasolina Residuals  
## Sum of Squares 17.04933 8.02800  
## Deg. of Freedom 2 12  
##   
## Residual standard error: 0.8179242  
## Estimated effects may be unbalanced

summary(modelo)

## Df Sum Sq Mean Sq F value Pr(>F)   
## gasolina 2 17.049 8.525 12.74 0.00108 \*\*  
## Residuals 12 8.028 0.669   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Analizando los resultados de la tabla ANOVa tenemos las siguientes observaciones:

1. Los grados de libertad del factor son 2
2. Los grados de libertad residuales con 12
3. La suma de cuadrado de los grupos son 17.049
4. La suma de cuadrados del error son 8.028.
5. El valor estadístico F es 12.74, que es valor critico.
6. El valor de p es 0.00108

De la observación 6, dado que p es menor que 0.05 se concluye que existe un efecto significativo de la variable gasolina. Con el fin de corroborar el resultado obtenido de la prueba ANOVA, se presentan los cálculos correspondientes a la *suma de cuadrados y medias al cuadrado*:

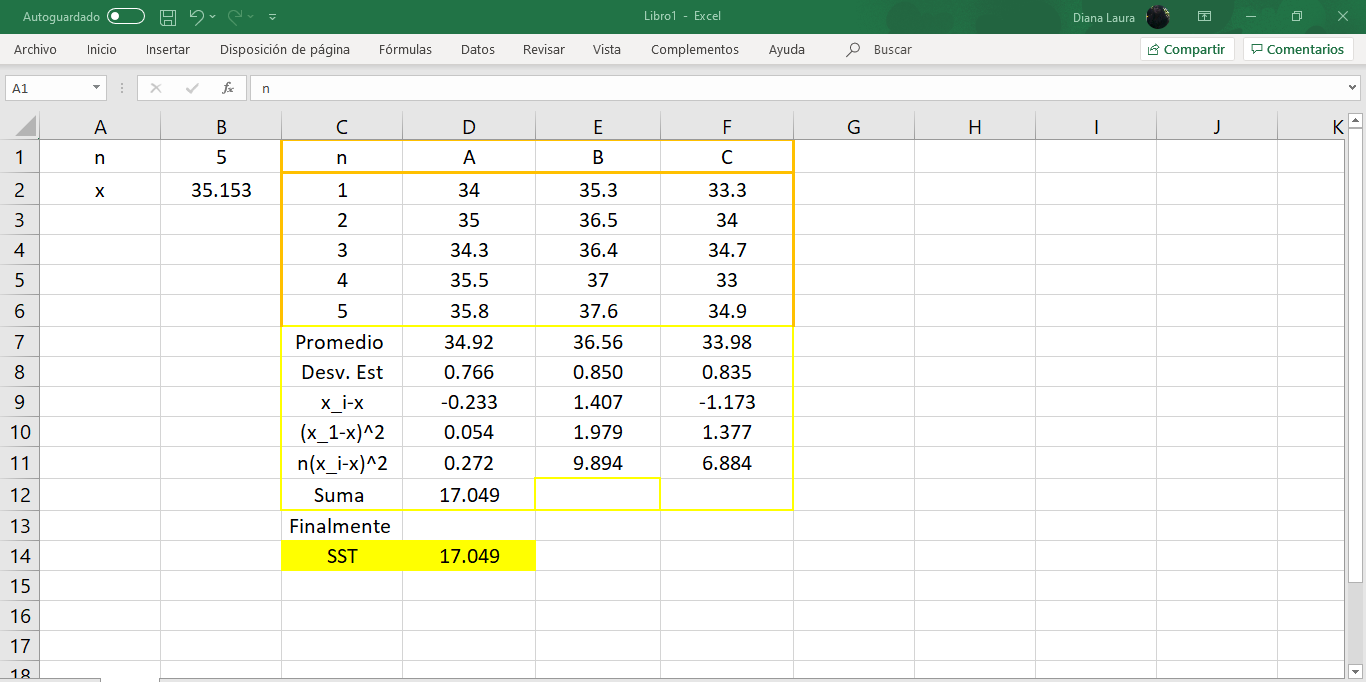
Donde

El valor de SST se obtuvo calculando la diferencia entre la media de cada muestra y la media , luego se elvaron al cuadrado estas diferencias y se multiplicaron por el número de observaciones y se sumo el total. El resultado obtenido es una medida de la variabilidad de las medias de las muestras. Los resultados se realizaron en Excel y se presentan a continuación:

Entonces:

Los resultados se presentan en Excel

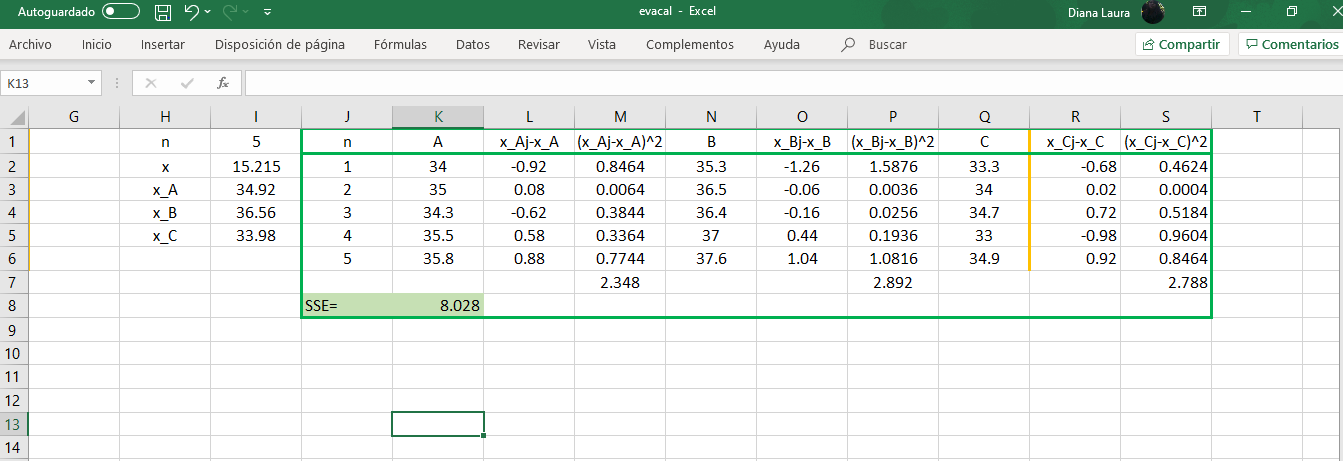
knitr::include\_graphics("cap1.png" )



Mientras que el error de la suma de cuadrados es:

$=\sum\_{j=1}^{n\_{A}}\left(x\_{Aj}-\bar{xA}\right)^{2} + \sum\_{j=1}^{n\_{B}}(x\_{Bj}-\bar{x\_B}\right)^{2}+\sum{j=1}^{n\_{c}}\left(x\_{C j}-\bar{x\_C}\right)^{2}$

knitr::include\_graphics("cap2.png" )



Finalmente, se define la suma de los cuadrados SSTO:

Se definen además, *treatment mean square, MST*

Y partir de estos, se define F como:

$F=

En este problema p=3, ya que es el número de muestras:

Además:

En la siguiente sección se obtuvieron estas diferencias por el método de Tukey, para ello se empleó la función: *TukeyHSD* y posteriormente se graficaron los intervalos de confianza.

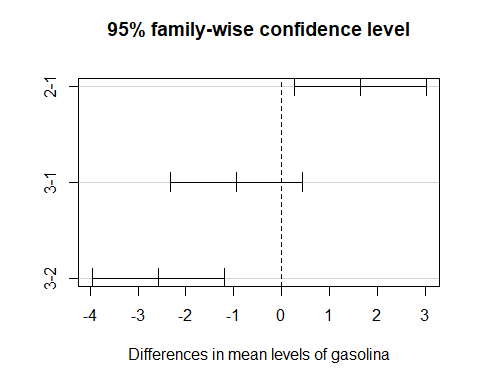
intervals = TukeyHSD(modelo)  
intervals

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = kilometraje ~ gasolina, data = mileage.vert)  
##   
## $gasolina  
## diff lwr upr p adj  
## 2-1 1.64 0.2599123 3.0200877 0.0204273  
## 3-1 -0.94 -2.3200877 0.4400877 0.2056606  
## 3-2 -2.58 -3.9600877 -1.1999123 0.0008518

print(intervals)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = kilometraje ~ gasolina, data = mileage.vert)  
##   
## $gasolina  
## diff lwr upr p adj  
## 2-1 1.64 0.2599123 3.0200877 0.0204273  
## 3-1 -0.94 -2.3200877 0.4400877 0.2056606  
## 3-2 -2.58 -3.9600877 -1.1999123 0.0008518

plot(intervals)



El método de Turkey indica que las diferencias entre medias en las que el intervalo de confianza engloba los límites inferior y superior que no contienen el valor cero, son estadísticamente significativas, además de que el valor P es menor a 0.05. Para este problema, vemos que los grupos con diferencias significativas son los siguientes:

Gasolinas 3 y 2 Gasolinas 2 y 1

Mientras que el grupo que toca el cero, es decir, que tiene una diferencia entre las medias no estadisticamente significativa y con valor de p mayor a 0.05 es:

Gasolina 3 y 1.

Realizando los cálculos:

Considerando una diferencia entre para un intervalo de confianza entre:

Donde depende de los grados de libertad y MSE es el error definido anteriormente.

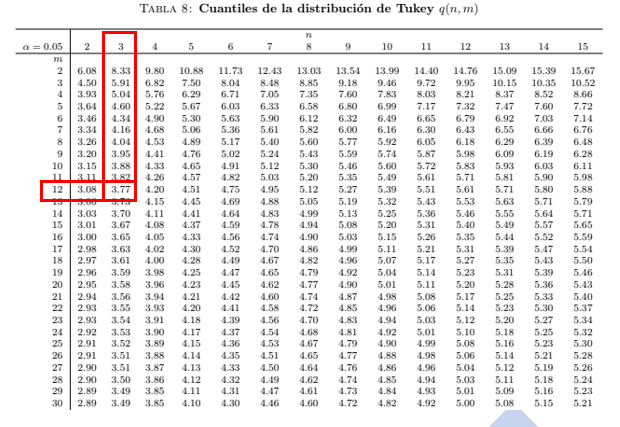
Además, el porcentaje de confianza de TUkey para :

Donde es el valor de tabla correspondiente a la intersección entre y .

Y el estimador puntual de la media, para un intervalo de confianza :

Para este trabajo, p=3, ya que es el número de muestras, cada una de tamaño 5, con un total de 15 observaciones (n=15), y MSE=0.669 como se encontró anteriormente. Además, consultando la Tabla de Tukey cuando n=3 y m=n-p=15-3=12, se obtiene

knitr::include\_graphics("tukey.png" )

 Así, para el intervalo entre la muestra 1 y 2 es:

El intervalo entre las muestras 1 y 2 (A y C)

Y finalmente, el intervalo entre la muestra B y C:

Como se puede observar, el intervalo entre las muestras A y C contiene el valor de 0, esto significa que existe una posibilidad de que ambas medias de la población sean iguales. A diferencia de esto, las muestras A con B y A con C, no contienen el valor cero, por lo que se puede afirmar que existen diferencias significativas. Se observa, además que estos resultados coinciden con los obtenidos con la prueba de Tukey realizada al inicio de esta sección.

Para finalizar se obtuvo el análisis LSD

library(agricolae)

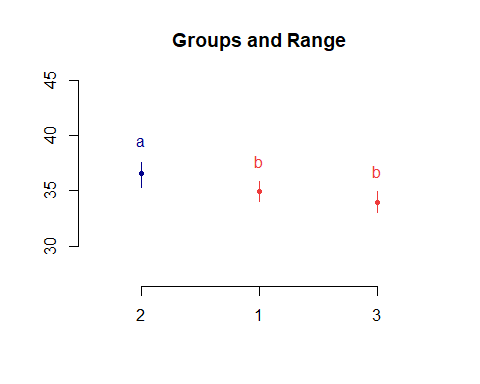
## Warning: package 'agricolae' was built under R version 4.0.3

attach(mileage.vert)

LSD <-LSD.test(modelo,"gasolina")  
LSD

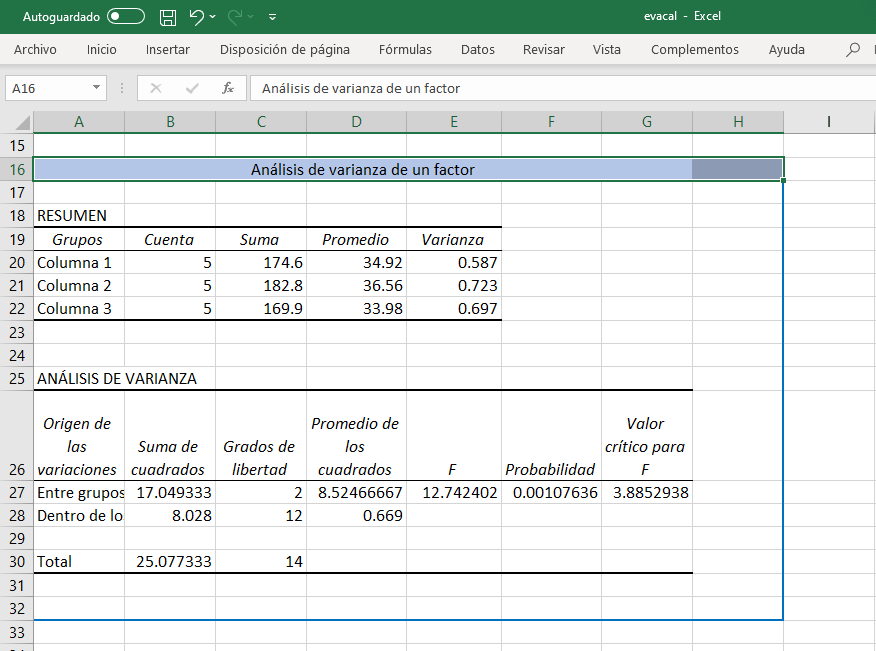
## $statistics  
## MSerror Df Mean CV t.value LSD  
## 0.669 12 35.15333 2.326733 2.178813 1.127101  
##   
## $parameters  
## test p.ajusted name.t ntr alpha  
## Fisher-LSD none gasolina 3 0.05  
##   
## $means  
## kilometraje std r LCL UCL Min Max Q25 Q50 Q75  
## 1 34.92 0.7661593 5 34.12302 35.71698 34.0 35.8 34.3 35.0 35.5  
## 2 36.56 0.8502941 5 35.76302 37.35698 35.3 37.6 36.4 36.5 37.0  
## 3 33.98 0.8348653 5 33.18302 34.77698 33.0 34.9 33.3 34.0 34.7  
##   
## $comparison  
## NULL  
##   
## $groups  
## kilometraje groups  
## 2 36.56 a  
## 1 34.92 b  
## 3 33.98 b  
##   
## attr(,"class")  
## [1] "group"

plot(LSD)



Para finalizar este trabajo, se comprobaron los resultados mediante un análisis de ANOVA en excel. Los resultados se muestran a continuación:

knitr::include\_graphics("cap3.png" )

 Como se puede observar, el análisis ANOVA proporciona la suma de los cuadrados, los grados de libertad, el promedio de los cuadrados, el valor y la probabilidad, los cuales si se comparan con el análisis de R, se encuentran resultados similares, salvo por algunas diferencias de decimas, las cuales pueden deberse a redondeo o al mismo programa. De esta manera, se comprueba que dado que el valor de la probabilidad es menor que 0.05 (correspondiente a un nivel de confianza del 95%) existe diferencia entre usar cualquiera de las gasolinas del experimento. El análisis puede consultarse en el documento de excel *evacal.exc* en la hoja *ANOVA E1*.

### Presentar claramente sus conclusiones para cada tipo de experimento, de acuerdo con los resultados obtenidos, explicando las diferencias entre los tres tipos de experimentos.

Con base a los análisis presentados se concluye de manera general que tanto, el análisis realizado con R, con fórmulas y con Excel, se encuentran los mismos resultados, lo que permitió corroborar los cálculos del libro consultado. Es importante mencionar, que a partir del análisis descriptivo por medio de diagramas de caja e histogramas se pueden inferir las primeras hipotesis, las cuales en este trabajo en partícular fueron comprobadas en su totalidad con cada uno de los cálculos. Algunas de las conclusiones que se pueden desprender son las siguientes:

A partir del diagrama de caja y del resumen de datos, se encontró que la gasolina que tiene mayor kilometraja fue la tipo 2 (o B), y la diferencia con las otras dos es dos unidades. Se encontró además que no existe una dispersión de los datos de manera importante, por lo que los datos son recabados con precisión. Esto es importante porque proporciona una medida de confiabilidad. Con respecto a la distribución, se observa una tendencia de normalidad, sin embargo a simple vista no se puede concluir.

Las pruebas de normalidad (Shapiro y Kolmorogov) e igualdad de varianza (Bartlett) pudieron verificar que los datos siguen una distribución normal y que se puede afirmar la hipotesis de que las varianzas no presentan diferencias significativas.

Con respecto al análisis de ANOVA, este rechazó la hipotesis nula de que la media de la variable kilometraje es la misma en las diferentes muestras, esto significa que existe una diferencia entre usar una de las tres gasolinas. Para detectar los grupos que generan estas diferencias su realizó la prueba de Tukey. El análisis de esta prueba, apartir de la evaluación de p-valor, mostró que el par de gasolinas 3 y 2, asi como las gasolinas 2 y 1 son la que no tienen una diferencia significativa, a diferencia de estas, el par que si presenta diferencia son las 3 y 1. Como parte de este trabajo, además se calcularón los intervalos con las fórmulas matematicas, las cuales permitieron corroborar el análisis de R.

Finalmente, se obtuvo el mismo anális ANOVA en R, el cual fue consistente con todos los cálculos anteriores. A manera de conclusión, para este experimento en particular, se recomienda usar la gasolina tipo 2 ya que es la que mayor kilometraje rinde, y dado que no existe una diferencia significativa entre las gasolinas 1 y 3, estas pueden usarse indistintamente, pero no se recomienda su empleo.