Apuntes de clase
Semana 5 - 09 de Marzo de 2023
MC6104 - Diseño de Experimentos

Juan José Cordero Gómez
Escuela de computación
Instituto Tecnológico de Costa Rica

Anova Monofactorial

Se retoma el ejercicio iniciado en Semana 4 sobre la prueba de Anova Monofactorial, este método estadístico permite encontrar las diferencias entre dos o más grupos.

Se asemeja a la prueba T o "T-Test" en cuanto a ejecución.

Los supuestos para análisis por Anova Monofactorial son los siguientes:

1) Que los datos obedezcan a una distribución normal, si los datos no son enteramente normales, se acepta algunas desviaciones, pero deben ser moderadas.
2) Que las variables de los datos sean independientes, lo que significa que los sus valores no son afectados por el resto de los valores.
3) Los grupos deben tener homocedasticidad, es decir, la misma varianza.

Si los puntos anteriores se cumplen, el método de Anova Monofactorial puede ser aplicado sin problemas.

Para una correcta interpretación de los resultados es necesario clarificar los siguiente conceptos:

- Hipótesis nula: se dice que es verdadera si los promedios de las variable medidas de cada grupo son iguales.
- Hipótesis alternativa: se que es verdadera si los promedios de las variables medidas de cada grupo son diferentes.

Se debe tener en cuenta que el análisis no va a arrojar directamente el resultado de donde están las diferencias, en caso de haberlas, para ello se debe realizar un ejercicio de interpretación post-hoc específicamente entre los grupos que más nos interesen durante la revisión.

Ejemplo de Anova Monofactorial

Se toma como premisa el mismo experimento realizado para prueba T, donde se pretende analizar los tiempos de entrenamiento de varios algoritmos en una red neuronal. Ver la Tabla I.

TABLE I
Premisa de experimento Anova.

	Factores
	Método de entrenamiento
Niveles	Algoritmo A
	Algoritmo B
	Algoritmo C

Se emplean los datos disponibles en el archivo "Datos-monofactorial.txt", que están disponibles en el TEC Digital.

El formato de los datos es como se muestra en la figura 1.

A continuación se desarrolla el ejercicio completo en R-Studio.

Se inicia instalando los paquetes necesarios de R:

```
if(!require(psych)){install.packages("psych")}
if(!require(FSA)){install.packages("FSA")}
if(!require(Rmisc)){install.packages("Rmisc")}
if(!require(ggplot2)){install.packages("ggplot2")}
if(!require(car)){install.packages("car")}
if(!require(multcompView)){install.packages("
    multcompView")}
if(!require(multcomp)){install.packages("multcomp")}
if(!require(lsmeans)){install.packages("lsmeans")}
if(!require(rcompanion)){install.packages("
    rcompanion")}
```

A este punto se exploran dos formas de cargar los datos.

Cargar los desde un archivo:

```
Data = read.table("PATH al Archivo",header = TRUE)
```

O asignar la totalidad de los datos a una variable en R:

```
In = ("Algoritmo        Ejecucion  Tiempo
'Algoritmo A'      '1'    12060
'Algoritmo A'      '2'    14089
'Algoritmo A'      '3'    13502
'Algoritmo A'      '4'     9574
'Algoritmo A'      '5'    14056
'Algoritmo A'      '6'    11569
'Algoritmo A'      '7'    13047
'Algoritmo A'      '8'    13275
'Algoritmo A'      '9'    14257
```

Algoritmo	Ejecución	Tiempo
'Algoritmo A'	'1'	12060
'Algoritmo A'	'2'	14089
'Algoritmo A'	'3'	13502
'Algoritmo A'	'4'	9574
'Algoritmo A'	'5'	14056
'Algoritmo A'	'6'	11569
'Algoritmo A'	'7'	13047
'Algoritmo A'	'8'	13275
'Algoritmo A'	'9'	14257
'Algoritmo A'	'10'	15075
'Algoritmo A'	'11'	12506
'Algoritmo A'	'12'	11557
'Algoritmo A'	'13'	9548
'Algoritmo A'	'14'	11514
'Algoritmo A'	'15'	16015
'Algoritmo A'	'16'	13004
'Algoritmo A'	'17'	10510

Fig. 1. Datos de ejemplo para prueba de Anova monofactorial.

```
'Algoritmo A'      '10'   15075
'Algoritmo A'      '11'   12506
'Algoritmo A'      '12'   11557
'Algoritmo A'      '13'    9548
'Algoritmo A'      '14'   11514
'Algoritmo A'      '15'   16015
'Algoritmo A'      '16'   13004
'Algoritmo A'      '17'   10510
'Algoritmo A'      '18'   13040
'Algoritmo A'      '19'   17098
'Algoritmo A'      '20'   13080
'Algoritmo B'      '1'    11080
'Algoritmo B'      '2'    12089
'Algoritmo B'      '3'    12538
'Algoritmo B'      '4'    10571
'Algoritmo B'      '5'    12010
'Algoritmo B'      '6'    12598
'Algoritmo B'      '7'    13543
'Algoritmo B'      '8'    13547
'Algoritmo B'      '9'    13217
'Algoritmo B'      '10'   15297
'Algoritmo B'      '11'   12210
'Algoritmo B'      '12'   11299
'Algoritmo B'      '13'   10067
'Algoritmo B'      '14'   11279
'Algoritmo B'      '15'   14006
'Algoritmo B'      '16'   12099
'Algoritmo B'      '17'   11581
'Algoritmo B'      '18'   14012
'Algoritmo B'      '19'   15069
'Algoritmo B'      '20'   12000
'Algoritmo C'      '1'     9081
'Algoritmo C'      '2'    11012
'Algoritmo C'      '3'    11529
'Algoritmo C'      '4'     9569
'Algoritmo C'      '5'    11092
'Algoritmo C'      '6'    11524
'Algoritmo C'      '7'    12522
'Algoritmo C'      '8'    12588
'Algoritmo C'      '9'    12241
'Algoritmo C'      '10'   13257
'Algoritmo C'      '11'   11294
'Algoritmo C'      '12'   10226
'Algoritmo C'      '13'    9591
'Algoritmo C'      '14'    9224
'Algoritmo C'      '15'   12033
'Algoritmo C'      '16'   11063
'Algoritmo C'      '17'    9537
'Algoritmo C'      '18'   13014
'Algoritmo C'      '19'   14033
'Algoritmo C'      '20'   11093
      ")
Data = read.table(textConnection(In),header = TRUE)
```

Cualquiera que sea la opción elegida se comportara de la misma forma.

Ahora se ordenan los datos según la forma en que se ingresaron y explícitamente se evita que R lo ordene alfabéticamente.

```
Data$Algoritmo = factor(Data$Algoritmo)
levels = unique(Data$Algoritmo)
```

Se verifican los datos en la tabla.

```
library(psych)
headTail(Data)

---Salida del comando
      Algoritmo Ejecucion Tiempo
1   Algoritmo A         1  12060
2   Algoritmo A         2  14089
3   Algoritmo A         3  13502
4   Algoritmo A         4   9574
...         <NA>       ...    ...
57  Algoritmo C        17   9537
58  Algoritmo C        18  13014
59  Algoritmo C        19  14033
60  Algoritmo C        20  11093

str(Data)

---Salida del comando
'data.frame': 60 obs. of  3 variables:
 $ Algoritmo: Factor w/ 3 levels "Algoritmo A",..: 1
     1 1 1 1 1 1 1 1 1 ...
 $ Ejecucion: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Tiempo   : int  12060 14089 13502 9574 14056
    11569 13047 13275 14257 15075 ...

summary(Data)

---Salida del comando
       Algoritmo    Ejecucion          Tiempo
 Algoritmo A:20   Min.   : 1.00   Min.   : 9081
 Algoritmo B:20   1st Qu.: 5.75   1st Qu.:11093
 Algoritmo C:20   Median :10.50   Median :12094
                  Mean   :10.50   Mean   :12234
                  3rd Qu.:15.25   3rd Qu.:13262
                  Max.   :20.00   Max.   :17098

#rm(In) #Necesario solo si se cargan los datos en el
     mismo script.
```

Con el comando summarize se hace un resumen de los datos para facilitar la lectura.

```
Summarize(Tiempo ~ Algoritmo,
```

```
         data = Data,
         digits = 3)

---Salida del comando
    Algoritmo  n     mean       sd   min       Q1
    median       Q3   max
1 Algoritmo A 20 12918.80 1941.192  9548 11566.00
    13043.5 14064.25 17098
2 Algoritmo B 20 12505.60 1414.667 10067 11510.50
    12154.5 13544.00 15297
3 Algoritmo C 20 11276.15 1424.242  9081 10067.25
    11193.5 12311.25 14033
```

En este punto se realiza el diagrama de cajas y bigotes con el fin de realizar el análisis visual de los grupos.

```
M = tapply(Data$Tiempo,
           INDEX = Data$Algoritmo,
           FUN = mean)

boxplot(Tiempo ~ Algoritmo, data = Data)

points(M,
       col="red",
       pch="+",
       cex=2)
```

El código anterior genera el siguiente plot. Figura 2.

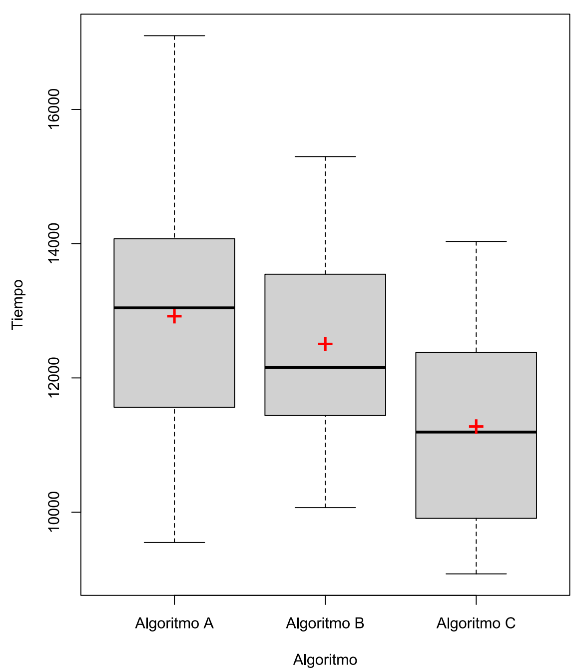

Fig. 2. Diagrama de cajas y bigotes.

El diagrama permite visualizar los rangos de los grupos a nivel de los bigotes, la cruz interna representa el promedio. La línea negra representa los cuartiles (es la division de los datos al 25%) y estos permiten ir viendo una ligera diferencia sobre cual grupo podría ser estadísticamente distinto.

A este punto es requerido continuar con el análisis ya que no es posible sacar conclusiones exactas con respecto a los grupos, por lo que se propone realizar el gráfico de promedio e intervalos de confianza.

Primero se preparan los datos:

```
Sum = groupwiseMean(Tiempo ~ Algoritmo,
                    data = Data,
                    conf = 0.95,
                    digits = 3,
                    traditional = FALSE,
                    percentile = TRUE)

Sum

---Salida del comando
    Algoritmo  n  Mean Conf.level Percentile.lower
    Percentile.upper
1 Algoritmo A 20 12900       0.95            12100
            13800
2 Algoritmo B 20 12500       0.95            11900
            13100
3 Algoritmo C 20 11300       0.95            10700
            11900
```

Y se procede a gráficar:

```
library(ggplot2)

#Plot ajustado para evitar el use de parametros
    deprecados.
ggplot(Sum,
       aes(x = Algoritmo,
           y = Mean)) +
  geom_errorbar(aes(ymin = Percentile.lower,
                    ymax = Percentile.upper),
                width = 0.05,
                linewidth = 0.5) +
  geom_point(shape = 15,
             size = 4) +
  theme_bw() +
  theme(axis.title = element_text(face = "bold")) +
  ylab("Tiempo promedio, s")
```

Se obtiene la figura 3.

Realizando el análisis “C” parece ser suficientemente diferente a nivel estadístico, esto porque los intervalos de confianza no se traslapan.

Eventualmente si los intervalos de confianza se traslapan podrían ser iguales, no obstante, será necesario hacer la prueba.

Es importante mencionar que al emplear intervalos de confianza al 95%, significa que tenemos un 5% de probabilidades de equivocarnos.

ProTip: Cuando el gráfico de bigotes no indica que representan los bigotes, se puede evaluar visualmente, si son simétricos generalmente van a representar el rango, de lo contrario serían los intervalos de confianza.

INTERVALOS DE CONFIANZA

Sirven para indicar que tan precisa es la estadística que se calculo y está directamente relacionado con la diferencia entre el promedio de la población y la muestra, también llamado error de muestreo. Básicamente son utilizados para medir la incertidumbre en una variable muestreada.

Se pueden medir con desviaciones estándar de la siguiente forma:

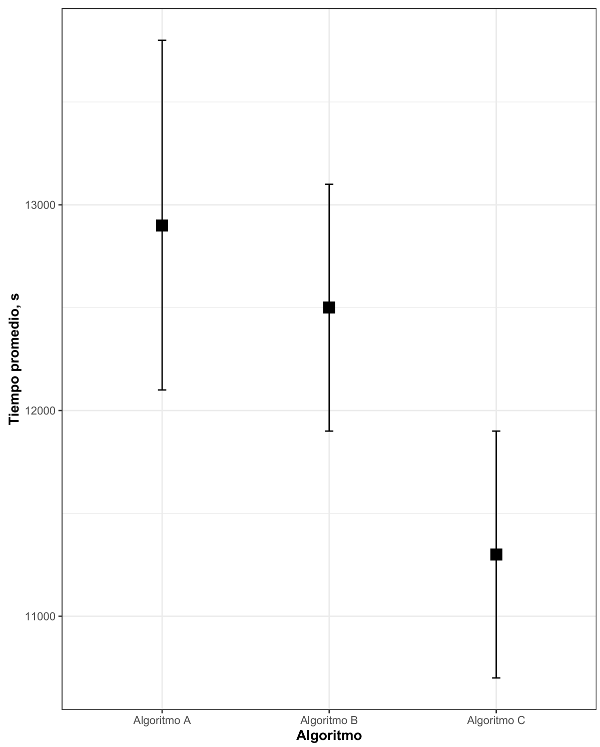

Fig. 3. Diagrama de promedios e intervalos de confianza.

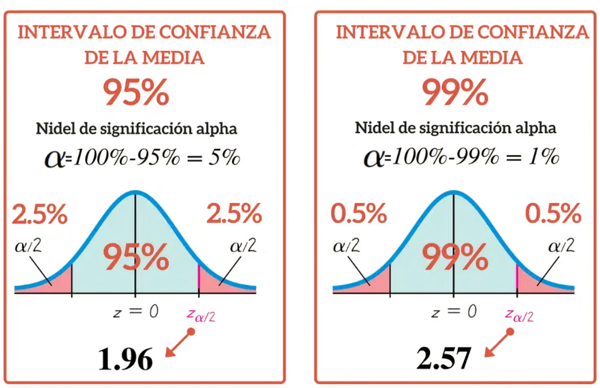

Fig. 4. Intervalos de confianza.

- A dos desviaciones estándar se tiene 95% de confianza.
- A tres desviaciones estándar se tiene 99% de confianza.

Si durante una investigación se hacen conclusiones sobre el promedio de una población es recomendado que se haga indicando que intervalos de confianza se tienen, así mismo estos nos indican que tan certera podría ser esta conclusión o inferencia.

Se debe distinguir la diferencia de intervalos de confianza contra el rango que nos dice los valores mínimos y máximos de un grupo.

La ventaja de hacer el análisis del intervalo de confianza es que entre dos grupos o más, si estos no se traslapan es posible decir con certeza que los grupos son estadísticamente diferentes y no sería necesario hacer la prueba numérica.

También ayudan a definir el nivel de confianza, que es básicamente que tan probable es que la estimación realizada basada en muestra, sea verdadera para el resto de la población.

MODELO LINEAL

Se define como “Una variable dependiente es predicha por un conjunto de variables independientes a través de una relación lineal”, son comúnmente utilizados en la estadística ya que emplean supuestos directamente sobre los datos empleados.

$$Y_i = \beta_0 + \beta_1\phi_1(X_{i1}) + \cdots + \beta_p\phi_p(X_{ip}) + \varepsilon_i \qquad i = 1, \ldots, n$$

Fig. 5. Formula del modelo lineal.

Algunos modelos lineales generales son:

- Regresión lineal.
- Regresión lineal múltiple.
- Análisis de varianza.

Para realizar el modelo lineal en R se emplea el siguiente código:

```
model = lm(Tiempo ~ Algoritmo,
           data = Data)

summary(model)

---Salida del comando
Call:
lm(formula = Tiempo ~ Algoritmo, data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-3370.8 -1211.6    25.1  1065.4  4179.2

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           12918.8      360.5  35.835  < 2e-16 ***
AlgoritmoAlgoritmo B   -413.2      509.8  -0.810  0.42105
AlgoritmoAlgoritmo C  -1642.7      509.8  -3.222  0.00211 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1612 on 57 degrees of freedom
Multiple R-squared:  0.1647, Adjusted R-squared: 0.1353
F-statistic: 5.618 on 2 and 57 DF,  p-value: 0.005932
```

R-CUADRADO (R-SQUARED)

Se emplea para determinar que tan eficientemente el modelo explica los datos, es del tipo “goodness-of-fit” que utiliza un resumen de las diferencias entre los datos observados y los esperados.

El p-value, indica si hay una relación significativa descrita por el modelo, el r-cuadrado, es el grado sobre el cual los datos son explicados por el modelo.

Por ejemplo, decir que se tiene un 0.6 de R-Cuadrado, significa que el 60% de la variabilidad de la variable dependiente es explicada por el modelo.

Si se tiene p-values y r-squared bajos, esto significa que hay mucha variabilidad, no obstante, los datos explican que si hay relaciones significativas.

Continuación - Ejemplo de Anova Monofactorial

Se realiza la prueba de Anova para confirmar el p-value y validar si hay diferencia estadistica en los grupos.

```
library(car)

Anova(model, #Tipo 2 es el valor por defecto
      type = "II") # Suma de cuadrados

---Salida del comando
Anova Table (Type II tests)

Response: Tiempo
             Sum Sq Df F value   Pr(>F)
Algoritmo  29203870  2  5.6176 0.005932 **
Residuals 148161499 57
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.
    ' 0.1 ' ' 1
```

Se puede ver que el resultado indica que si hay diferencia estadística en al menos alguno de los grupos, esto debido al valor del p-value, que haciendo uso de los “*” nos dice también con precisión se puede afirmar esto, para el caso particular la prueba pasa con un $\alpha = 0.01$.

Lo anterior se puede visualizar en un histograma y en gráfico de dispersion que se realizan en R de la siguiente forma:

```
x = residuals(model)

library(rcompanion)

plotNormalHistogram(x)

plot(fitted(model),
     residuals(model))

plot(model)
```

El histograma permite observar la normalidad en los residuos.

Viendo la figura 7 se debe considerar que si el cono que forman los tres grupos lleva a ser muy pronunciado, esto significa que no hay homocedasticidad, es decir, lo grupos no tienen la misma varianza.

Análisis post-hoc

Como el Anova solamente nos dice que si hay un eventual diferencia estadística pero no dá más detalles, es necesario explorar dos conceptos:

- **Least Square Means:** Son promedios para grupos que son ajustados para promedios de otros factores. Son comúnmente utilizados para estudios desbalanceados.
- **Least Square Means for multiple comparisons:** Se usan en comparaciones de valores entre grupos en modelos lineales. Resume los efectos de factores y se usan para comprobar contrastes lineales entre predicciones.

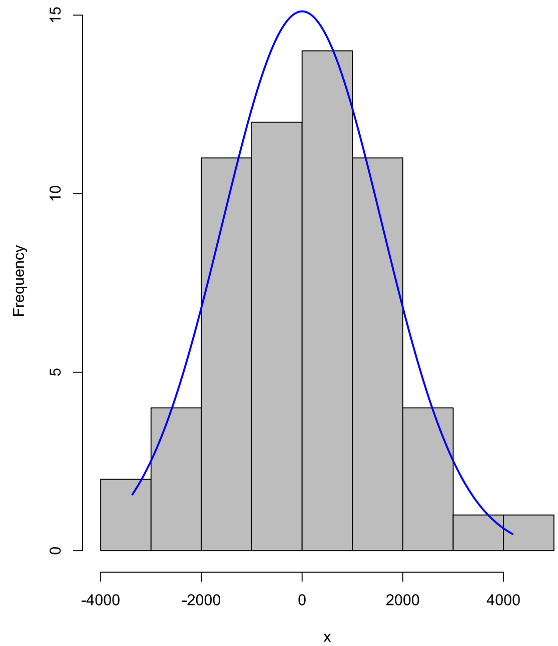

Fig. 6. Histograma de residuos.

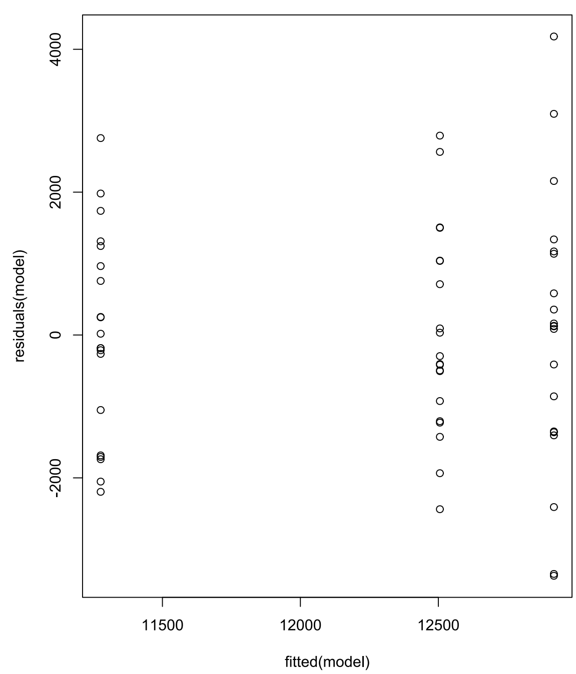

Fig. 7. Gráfico de dispersión.

Pruebas de separación de promedios

Para realizar la prueba de separación de promedios se puede emplear el Leas Square Mean o lsmeans en R (comparación de pares), de la siguiente forma:

```
library(multcompView)
library(lsmeans)

marginal = lsmeans(model,
                   ~ Algoritmo)

# Ejecuta todas las comparaciones de pares.
# Para la variable Algoritmo.

pairs(marginal,
      adjust="tukey")

---Salida del comando
 contrast                  estimate  SE df t.ratio p.value
 Algoritmo A - Algoritmo B      413 510 57   0.810 0.6981
 Algoritmo A - Algoritmo C     1643 510 57   3.222 0.0059
 Algoritmo B - Algoritmo C     1229 510 57   2.411 0.0494

P value adjustment: tukey method for comparing a family of 3 estimates
```

Para este análisis de debe considerar el p-value, cuando este es muy alto no hay diferencia entre los grupos, pero si el valor es menor a 0.05, si hay diferencia.

Protip: Se hace Anova $->$ Post-hoc $->$ Análisis de pares.

LSMEANS

Permite ejecutar las comparaciones por pares para la variable independiente que llamamos a“Algoritmo” en este ejemplo.

- **Error estándar:** Estima que tan cerca está el promedio calculado, de ser el verdadero promedio de la población. Es la desviación estándar de la raíz del número de observaciones.
- **Ajuste de Tukey:** La prueba de múltiples hipótesis suele acumular errores de tipo 1, por lo que se realiza el ajuste de Tukey, lo que permite realizar múltiples análisis Anova mientras mantiene el Error tipo 1 por debajo de los valores que sean significativos. Se considera mejor que la prueba T entre grupos (práctica que es científicamente deshonesta).

Usando la función CLD

El cld o Compact Letter Display, permite hacer una clasificación de los grupos, en otros grupos indicados por letras según su semejanza estadística, por lo que facilita el análisis.

R permite hacerlo de la siguiente forma:

```
library(multcomp)

CLD = cld(marginal,
          alpha = 0.05,
          Letters = letters,
          adjust = "tukey")

CLD

---Salida del comando
 Algoritmo   lsmean  SE df lower.CL upper.CL .group
 Algoritmo C  11276 361 57    10389    12163  a
 Algoritmo B  12506 361 57    11619    13392   b
 Algoritmo A  12919 361 57    12032    13806   b

Confidence level used: 0.95
Conf-level adjustment: sidak method for 3 estimates
P value adjustment: tukey method for comparing a family of 3 estimates
significance level used: alpha = 0.05
NOTE: If two or more means share the same grouping symbol,
      then we cannot show them to be different.
      But we also did not show them to be the same.
```

En este caso vemos que el Algoritmo A y B, comparten la letra “b”, lo que indica que **NO** son diferentes estadísticamente, no obstante si lo son con respecto a C.

Fin del análisis

A este punto es posible graficar los promedios con los intervalos de confianza y las letras de separación de grupos.

Este es el código en R empleado:

```
#Preparan los datos para graficar

CLD$Algoritmo = factor(CLD$Algoritmo,
                       levels = c("Algoritmo A",
                                  "Algoritmo B",
                                  "Algoritmo C"))

CLD$.group=gsub(" ", "",CLD$.group)

#Se realiza el grafico con los intervalos de confianza y las letras para separar los grupos.
library(ggplot2)

ggplot(CLD,
       aes(x = Algoritmo,
           y = lsmean,
           label = .group)) +
  geom_point(shape = 15,
             size = 4) +
  geom_errorbar(aes(ymin = lower.CL,
                    ymax = upper.CL),
                width = 0.2,
                size = 0.7) +
  theme_bw() +
  theme(axis.title = element_text(face = "bold"),
        axis.text = element_text(face = "bold"),
        plot.caption = element_text(hjust = 0)) +
  ylab("Promedio del minimo cuadrado \ n
       Tiempo de ejecucion") +
  geom_text(nudge_x = c(0, 0, 0),
            nudge_y = c(1100, 1100, 1100),
            color = "black")
```

El código anterior resulta en el siguiente gráfico 8

Se cita el análisis final (leyenda) visto en clase con respecto a las conclusiones del análisis:

Tiempo de entrenamiento de algoritmos de redes neuronales. Las cajas representan el promedio del mínimo cuadrado para las tres clases, seguido por anova

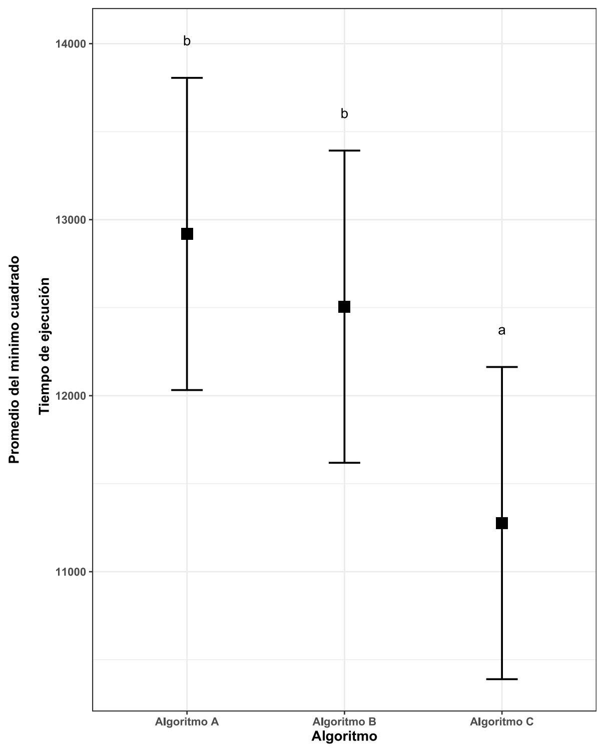

Fig. 8. Gráfico para análisis completo.

monofactorial.
Las barras de error indican los intervalos de confianza de los promedios del mínimo cuadradro al 95%. Promedios que comparten letra no son significativamente diferentes (alfa = 0.05, ajustados por Tukey).

Concepto de experimentos

Método científico

Richard Feynman:

Fue un físico teórico que vivió del año 1918 al 1988, era un excelente comunicador (se recomienda ver algunas de sus charlas en Youtube). Además trabajó en mecánica cuántica y fue reconocido por sus trabajos de la física del helio súper enfriado, física de partículas.

También participo en el proyecto Manhattan, ganó el premio Nobel en física en el año 1965.

Hace observaciones muy importantes con respecto a lo que es la Ciencia.

Ciencia

La ciencia implica que:

- Método.
- Conocimiento.
- Aplicaciones.

La ciencia es naturalmente incierta y esto porque puede ser falseable, si hay algo que no es congruente es suficiente razón para realizar mayores investigaciones.

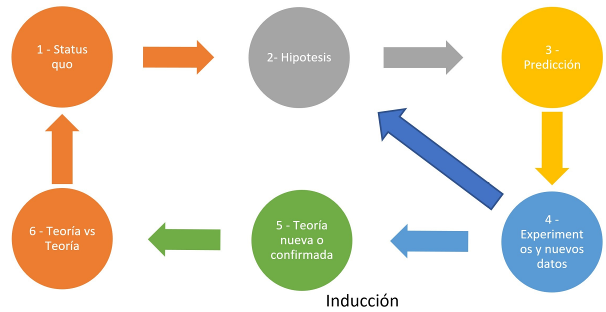

Fig. 9. Método científico.

References

[1] “Intervalos de confianza.” [Online]. Available: https://conceptosclaros.com/intervalo-confianza/