ANOVA de dos vías

Ignacio Díaz Oreiro

CI0131 Diseño de Experimentos

Agradecimiento a la profesora Dra. Kryscia Ramírez Benavides, por facilitar material usado en esta presentación.

Diseño factorial de dos factores

- Muchos experimentos involucran estudiar los efectos de dos o más factores, para lo que se utilizan los diseños factoriales.
- Los tipos más simples de diseños factoriales incluyen únicamente dos factores o conjuntos de tratamientos.
- Hay a niveles del factor A y b niveles del factor B, los cuales se disponen en un diseño factorial;
- Es decir, cada réplica del experimento contiene todas las $a \cdot b$ combinaciones de los tratamientos. En general, hay n réplicas.

Diseño factorial de dos factores

- Por ejemplo, un ingeniero de software está optimizando el tiempo de ejecución de un algoritmo de ordenamiento que se ejecutará en sistemas con tamaños de datos de entrada variables.
- Selecciona como factor el diseño la estructura de datos auxiliar para el algoritmo, y tiene tres opciones posibles: Hash Table, Árbol Binario de Búsqueda o Array Dinámico.
- El ingeniero también quiere verificar si el nivel de concurrencia afectará su tiempo de ejecución, que será otro factor de diseño

Diseño factorial de dos factores

- El ingeniero decide probar las tres estructuras de datos con niveles de concurrencia: 1 hilo, 4 hilos, 16 hilos, ya que estos rangos son representativos de los escenarios reales donde se usará el algoritmo.
- Se ejecutan cuatro pruebas con cada combinación de estructura de datos y concurrencia, y las 36 ejecuciones se realizan de manera aleatoria para evitar sesgos.

Diseño factorial de dos factores

- En la tabla se presentan los datos del experimento y de los tiempos de ejecución de los algoritmos, expresados en milisegundos:
- Los tiempos de ejecución obtenidos están en el rango de 40 a 188 milisegundos.

Tiempo de ejecución (en milisegundos)

Nivel de concurrencia (hilos)

Estructura	1		4		16	
Tabla hash	130 74	155 180	34 80	40 75	20 82	70 58
Árbol binario	150 159	188 126	136 106	122 115	25 58	70 45
Array dinámico	138 168	110 160	174 150	120 139	96 82	104 60

Diseño factorial de dos factores

- El ingeniero quiere responder :
- ¿Qué efectos tienen el tipo de estructura de datos y la cantidad de hilos sobre el tiempo de ejecución del algoritmo?
- ¿Existe alguna estructura de datos que ofrezca un rendimiento consistente (tiempo bajo) independientemente del nivel de concurrencia?

Tiempo de ejecución (en milisegundos)

	Nivel de concurrencia (hilos)					
Estructura	1		4		16	
Tabla hash	130	155	34	40	20	70
	74	180	80	75	82	58
Árbol binario	150	188	136	122	25	70
	159	126	106	115	58	45
Array dinámico	138	110	174	120	96	104
	168	160	150	139	82	60

Diseño factorial de dos factores

- Quizá se pueda identificar una estructura que no se degrade significativamente sin importar la cantidad de hilos.
- Así, se podría garantizar que el algoritmo será robusto ante variaciones en el nivel de concurrencia.

Tiempo de ejecución (en milisegundos)

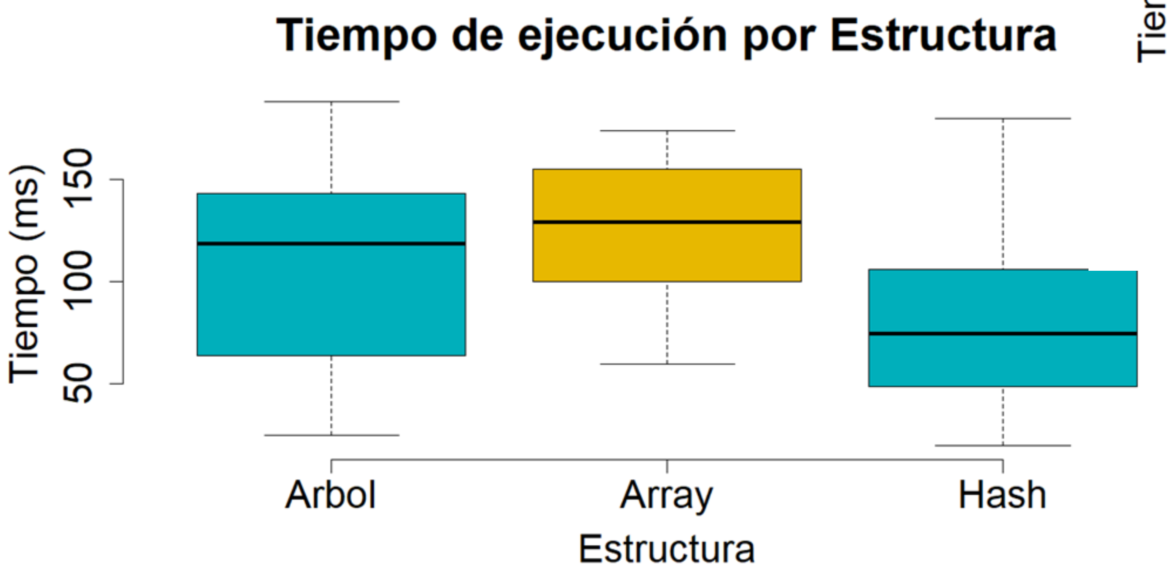
Estructura	Nivel de concurrencia (hilos): 1		4		16	
Tabla hash	130 74	155 180	34 80	40 75	20 82	70 58
Árbol binario	150 159	188 126	136 106	122 115	25 58	70 45
Array dinámico	138 168	110 160	174 150	120 139	96 82	104 60

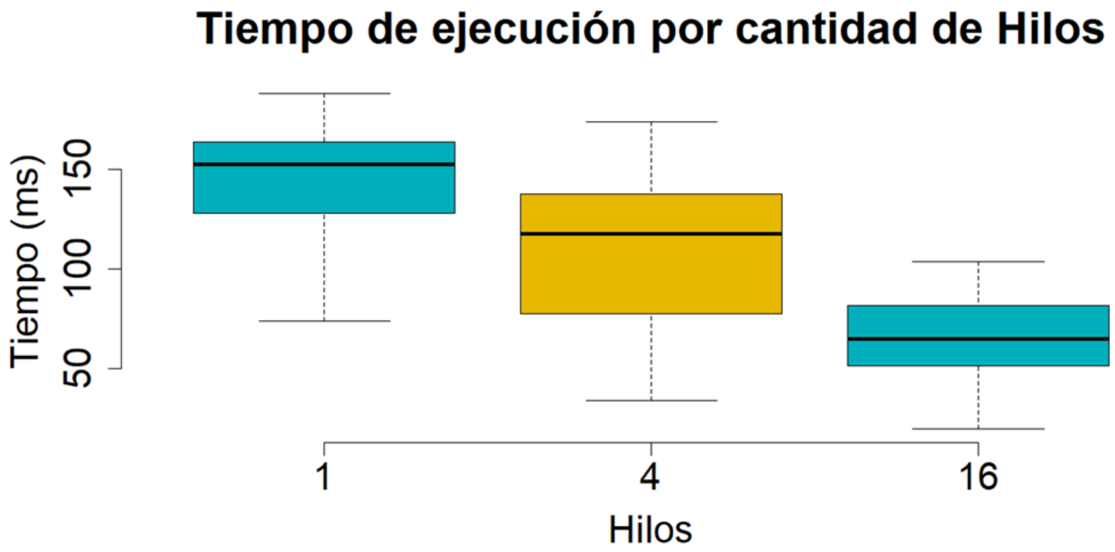
- Se trata de la aplicación del diseño experimental estadístico (factorial de dos factores) en la optimización de algoritmos, un problema clave en computación.

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

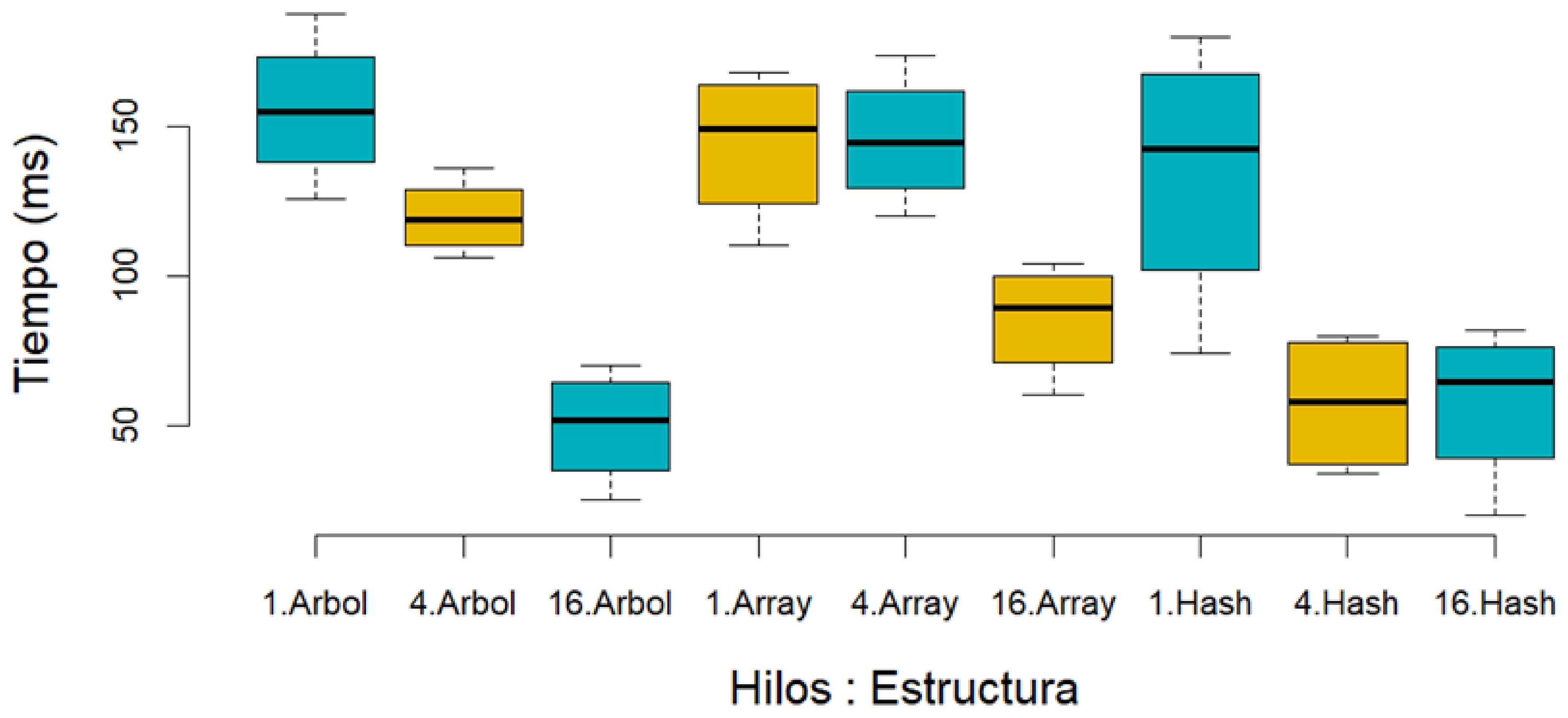
- Se realiza un análisis exploratorio de datos:





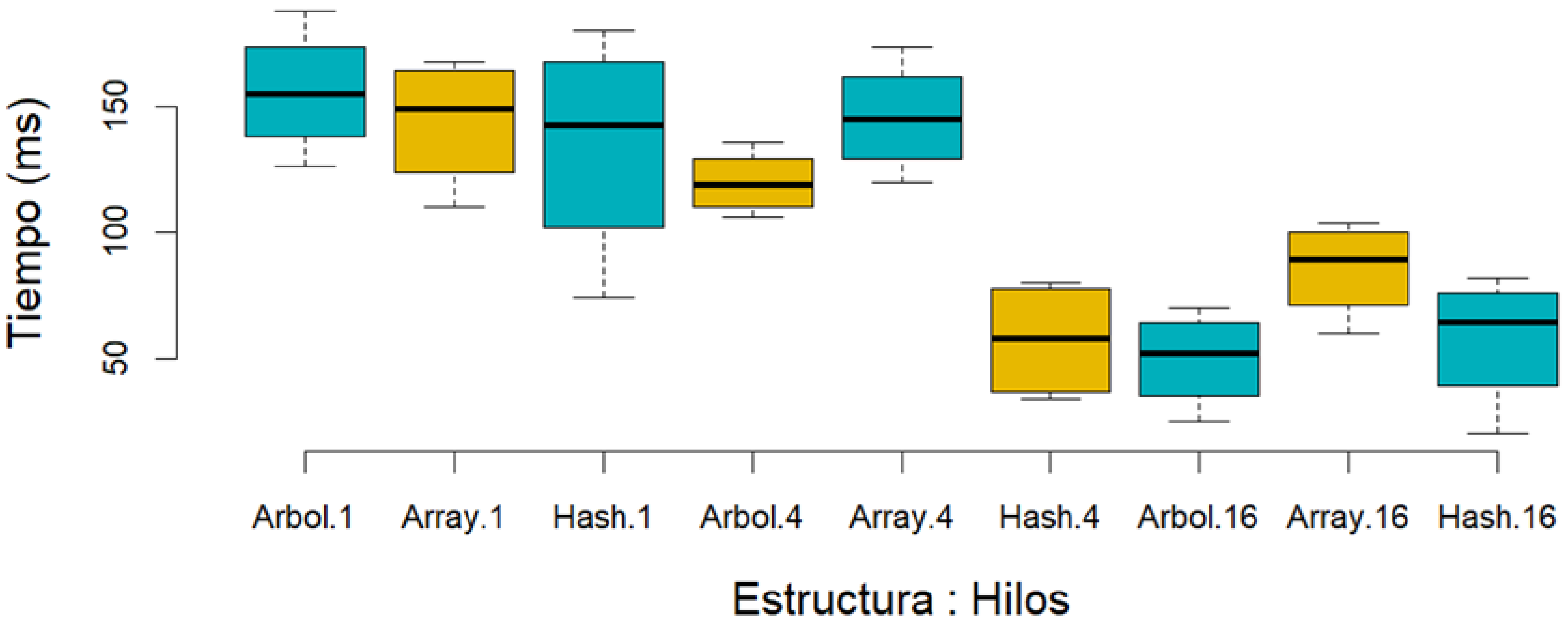
Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías



Diseño factorial de dos factores

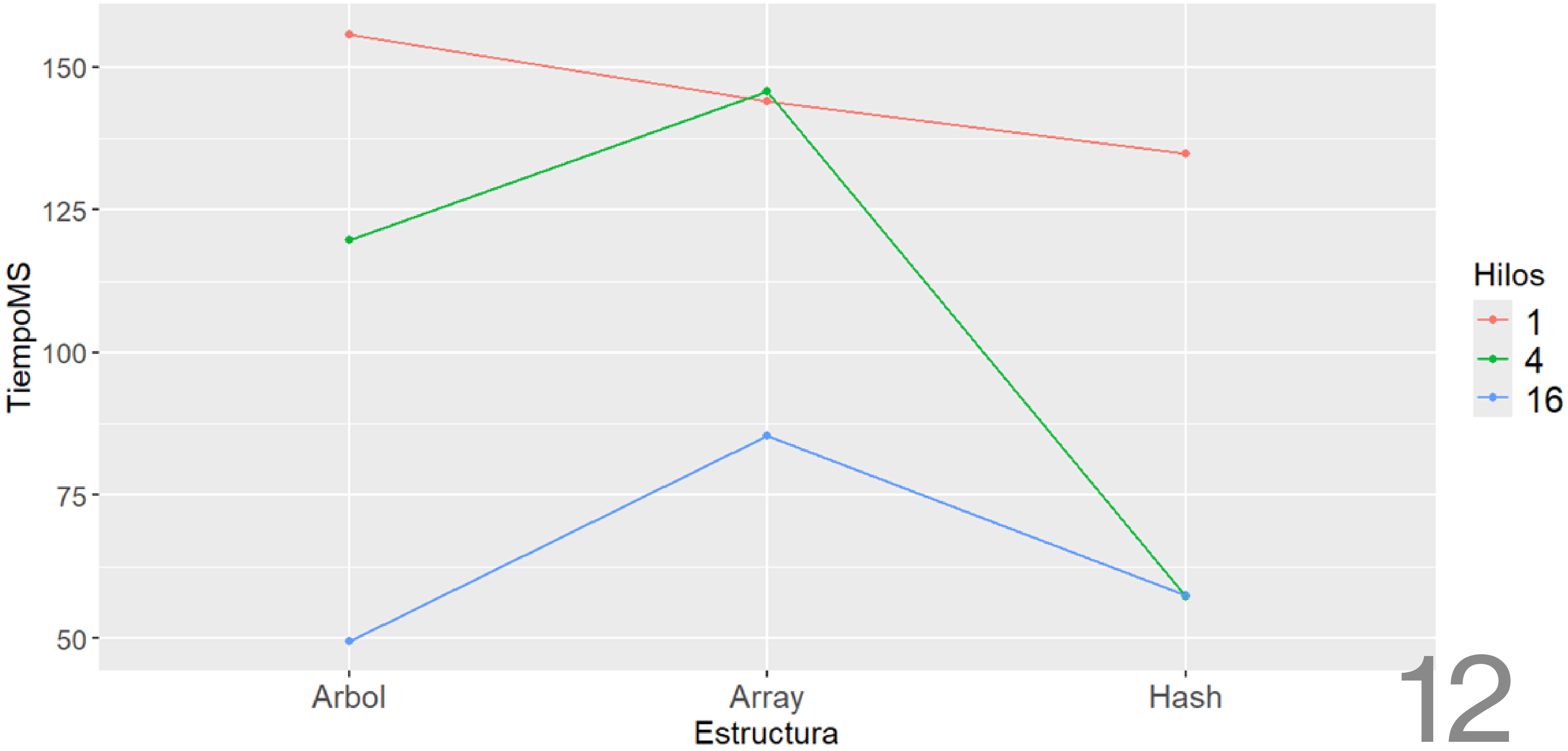
Ejemplo de análisis de varianza de dos vías



Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

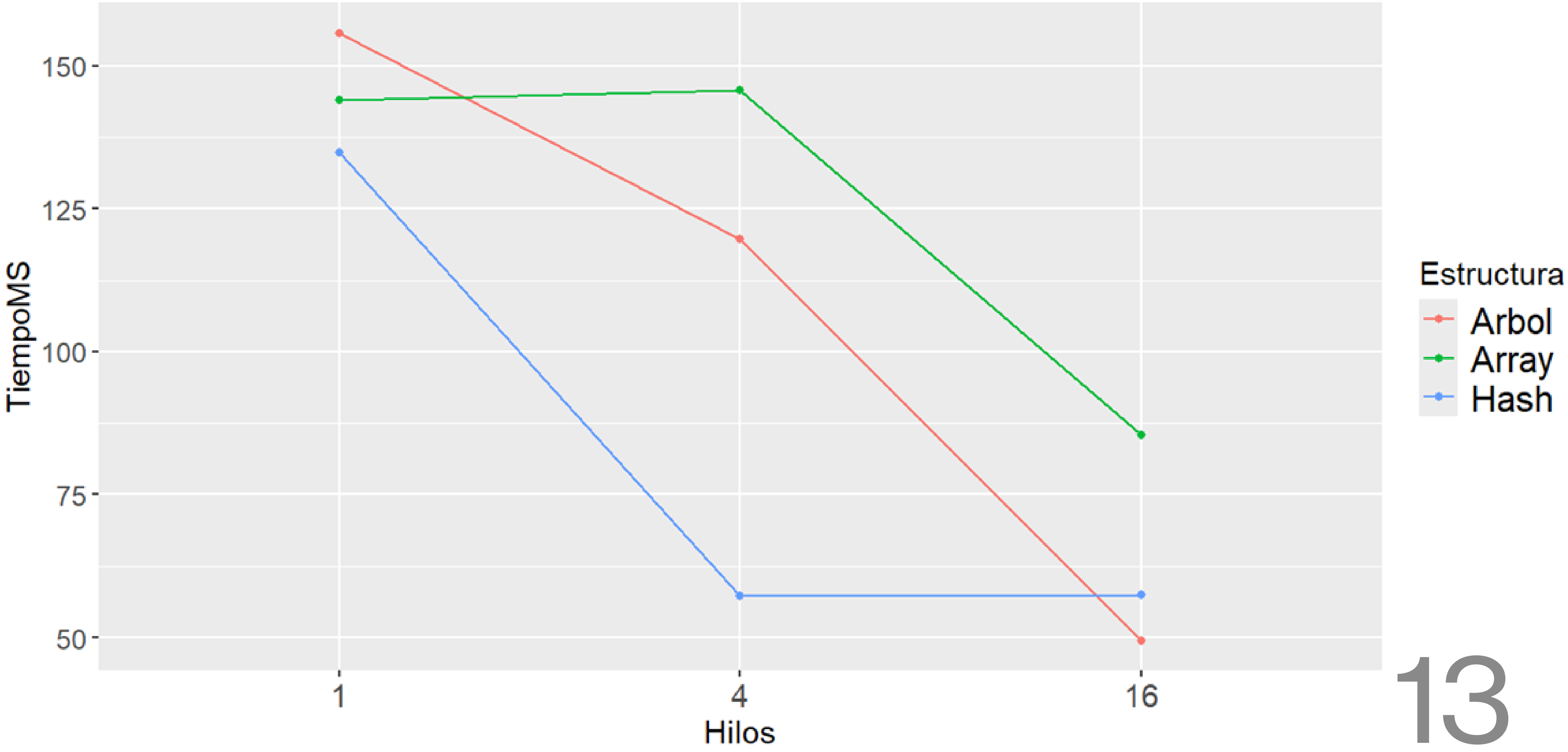
Se pueden graficar las interacciones:



Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

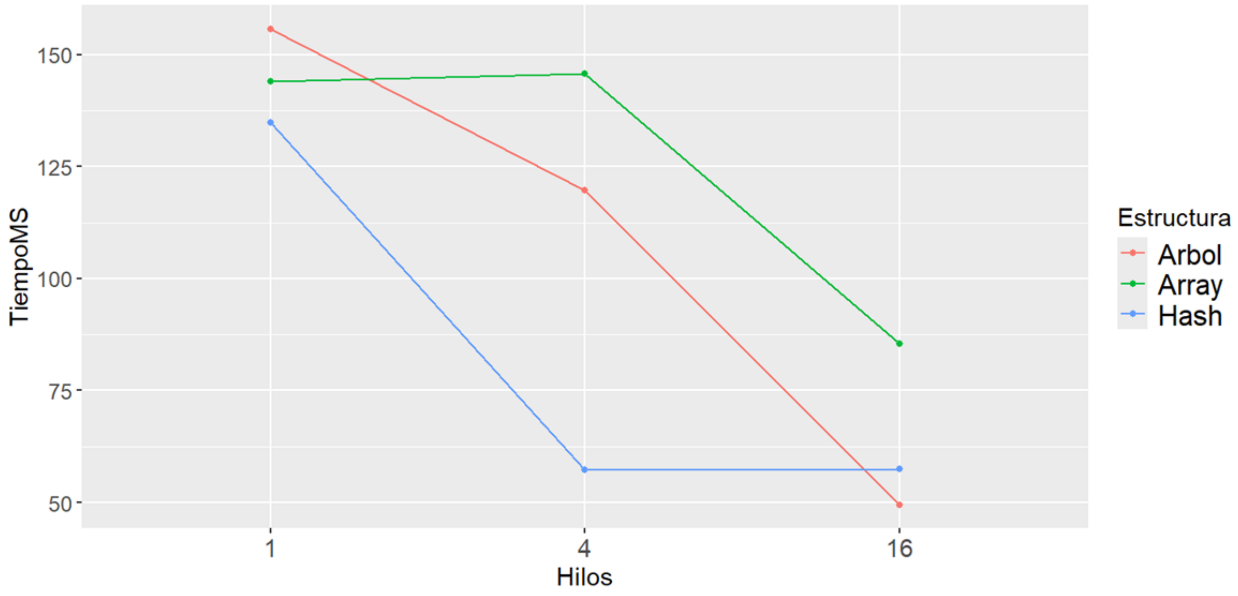
Se pueden graficar las interacciones:



Diseño factorial de dos factores

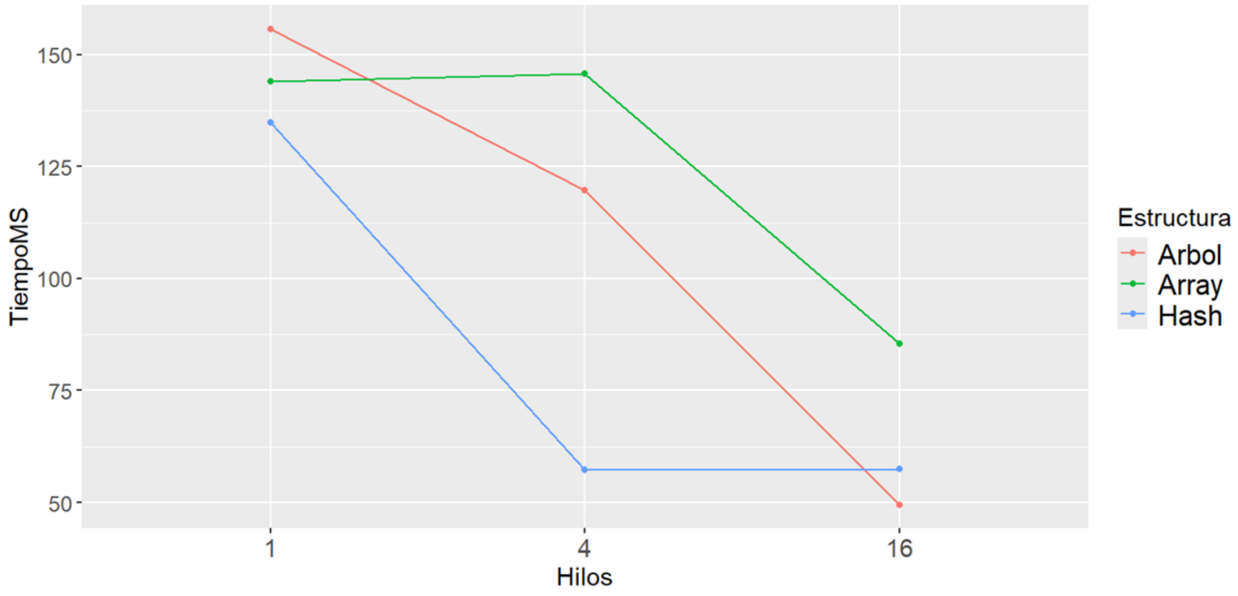
Ejemplo de análisis de varianza de dos vías

- La interacción significativa está indicada por la falta de paralelismo de las líneas.
- En general, se obtienen tiempos mayores sin concurrencia, independientemente del tipo de estructura.



Ejemplo de análisis de varianza de dos vías

- Al cambiar de 1 a 4 hilos, el tiempo para la estructura Árbol puede aumentar, mientras que disminuye para el Array y Hash.
- De 4 a 16 hilos, el tiempo disminuye para Árbol y Hash, y prácticamente no cambia Array.



- La estructura Árbol parece dar los tiempos más altos mejores a medida que aumenta la concurrencia.

Diseño factorial de dos factores

- Para pasar al caso general, sea y_{ijk} la respuesta observada cuando el factor A tiene el nivel *i*-ésimo (i = 1, 2, ..., a) y el factor B tiene el nivel *j*-ésimo (j = 1, 2, ..., b) en la réplica *k*-ésima (k = 1, 2, ..., n).

		Factor B			
		1	2	$\dots$	b
Factor A	1	$y_{111}, y_{112}, \dots, y_{11n}$	$y_{121}, y_{122}, \dots, y_{12n}$		$y_{1b1}, y_{1b2}, \dots, y_{1bn}$
	2	$y_{211}, y_{212}, \dots, y_{21n}$	$y_{221}, y_{222}, \dots, y_{22n}$		$y_{2b1}, y_{2b2}, \dots, y_{2bn}$
	$\vdots$				
	a	$y_{a11}, y_{a12}, \dots, y_{a1n}$	$y_{a21}, y_{a22}, \dots, y_{a2n}$		$y_{ab1}, y_{ab2}, \dots, y_{abn}$

Diseño factorial de dos factores

- El orden en que se hacen las *abn* observaciones se selecciona al azar, por lo que este diseño es un diseño completamente aleatorizado.

Factor A \ Factor B	1	2	. . .	b
1	$y_{111}, y_{112}, \ldots, y_{11n}$	$y_{121}, y_{122}, \ldots, y_{12n}$		$y_{1b1}, y_{1b2}, \ldots, y_{1bn}$
2	$y_{211}, y_{212}, \ldots, y_{21n}$	$y_{221}, y_{222}, \ldots, y_{22n}$		$y_{2b1}, y_{2b2}, \ldots, y_{2bn}$
⋮				
a	$y_{a11}, y_{a12}, \ldots, y_{a1n}$	$y_{a21}, y_{a22}, \ldots, y_{a2n}$		$y_{ab1}, y_{ab2}, \ldots, y_{abn}$

Diseño factorial de dos factores

Modelo de datos

- Las observaciones de un experimento factorial pueden describirse con un modelo.
- El modelo de los efectos para dos factores es:

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n \end{cases}$$

- donde y_{ijk} representa la observación *ijk*-ésima, μ es el efecto promedio global, τ_i es el efecto del nivel i-ésimo del factor A, β_j es el efecto del nivel j-ésimo del factor B, $(\tau\beta)_{ij}$ es el efecto de la interacción entre τ_i y β_j y ϵ_{ijk} es un componente del error aleatorio.

Diseño factorial de dos factores

Modelo de datos

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n \end{cases}$$

- Se supone que ambos factores A y B son fijos, y los efectos de los tratamientos se definen como las desviaciones de la media global, por lo que:

$$\Sigma_{i=1}^{a}\tau_i = 0 \quad \Sigma_{j=1}^{b}\beta_j = 0$$

- Se manera similar, los efectos de las interacciones son fijos y se definen de tal modo que: $\Sigma_{i=1}^{a}(\tau\beta)_{ij} = \Sigma_{j=1}^{b}(\tau\beta)_{ij} = 0$

- Puesto que hay n réplicas del experimento, hay *abn* observaciones en total.

Diseño factorial de dos factores

Modelo de datos

- Otro modelo posible de un experimento factorial es el modelo de las medias:

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n \end{cases}$$

- donde la media de la celda ij-ésima es:

$$\mu_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij}$$

Diseño factorial de dos factores

Modelo de datos

- En el diseño factorial de dos factores, los factores (o tratamientos) de los renglones y las columnas, A y B, son de igual interés.
- El interés se encuentra en probar hipótesis acerca de la igualdad de los efectos de los tratamientos de los renglones:

$$H_0\colon \tau_1 = \tau_2 = \cdots = \tau_a = 0$$
$$H_1\colon \text{at least one } \tau_i \neq 0$$

- y de la igualdad de los efectos de los tratamientos de las columnas:

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_b = 0$$
$$H_1\colon \text{at least one } \beta_j \neq 0$$

Diseño factorial de dos factores

Modelo de datos

- También interesa determinar si los tratamientos de los renglones y las columnas interactúan. Por lo tanto, también querría probarse:

$$H_0: (\tau\beta)_{ij} = 0 \qquad \text{for all } i, j$$

$$H_1: \text{at least one } (\tau\beta)_{ij} \neq 0$$

Diseño factorial de dos factores

Análisis estadístico del modelo con efectos fijos

- El procedimiento de prueba puede resumirse en una tabla del análisis de varianza:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
A treatments	SS_A	$a - 1$	$MS_A = \dfrac{SS_A}{a - 1}$	$F_0 = \dfrac{MS_A}{MS_E}$
B treatments	SS_B	$b - 1$	$MS_B = \dfrac{SS_B}{b - 1}$	$F_0 = \dfrac{MS_B}{MS_E}$
Interaction	SS_{AB}	$(a - 1)(b - 1)$	$MS_{AB} = \dfrac{SS_{AB}}{(a - 1)(b - 1)}$	$F_0 = \dfrac{MS_{AB}}{MS_E}$
Error	SS_E	$ab(n - 1)$	$MS_E = \dfrac{SS_E}{ab(n - 1)}$	
Total	SS_T	$abn - 1$		

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

	Observ	Estructura	Hilos	TiempoMS
1	1	Arbol	4	122
2	2	Hash	4	40
3	3	Array	4	139
4	4	Array	4	174
5	5	Hash	4	34
6	6	Array	1	110
7	7	Arbol	1	126
8	8	Hash	16	70
9	9	Hash	1	130

- Vamos a construir el modelo en R para el análisis de varianza de dos vías, considerando la interacción entre los dos factores.
- El archivo con los datos ya fue cargado y se llama ord.
- La sintaxis del modelo ANOVA es la siguiente:

```
aov_orden <- aov(TiempoMS ~ Estructura * Hilos, data = ord)
```

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

- En el caso del tiempo de ejecución, el análisis de varianza reporta:

```
> aov_orden <- aov(TiempoMS ~ Estructura * Hilos, data = ord)
> summary(aov_orden)
                 Df Sum Sq Mean Sq F value   Pr(>F)    
Estructura        2  10684    5342   7.911  0.00198 ** 
Hilos             2  39119   19559  28.968 1.91e-07 ***
Estructura:Hilos  4   9614    2403   3.560  0.01861 *  
Residuals        27  18231     675                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ejemplo de análisis de varianza de dos vías

```
                   Df Sum Sq Mean Sq F value   Pr(>F)    
Estructura          2  10684    5342   7.911  0.00198 ** 
Hilos               2  39119   19559  28.968 1.91e-07 ***
Estructura:Hilos    4   9614    2403   3.560  0.01861 *  
Residuals          27  18231     675                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Puesto que $F_{0.05, 2, 27} = 3.35$, el efecto de la Estructura sí es significativo ($F_0 > F_{0.05, 2, 27}$).
- También el efecto de la cantidad de Hilos es significativo
- Y puesto que $F_{0.05, 4, 27} = 2.73$, se concluye que hay una interacción significativa entre la Estructura y la cantidad de Hilos.

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

```
                   Df  Sum Sq  Mean Sq  F value    Pr(>F)
Estructura          2   10684     5342    7.911   0.00198 **
Hilos               2   39119    19559   28.968  1.91e-07 ***
Estructura:Hilos    4    9614     2403    3.560   0.01861 *
Residuals          27   18231      675
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Los valores-p también muestran que los efectos de la estructura y la cantidad de hilos son significativos, así como la interacción entre estos.

Diseño factorial de dos factores

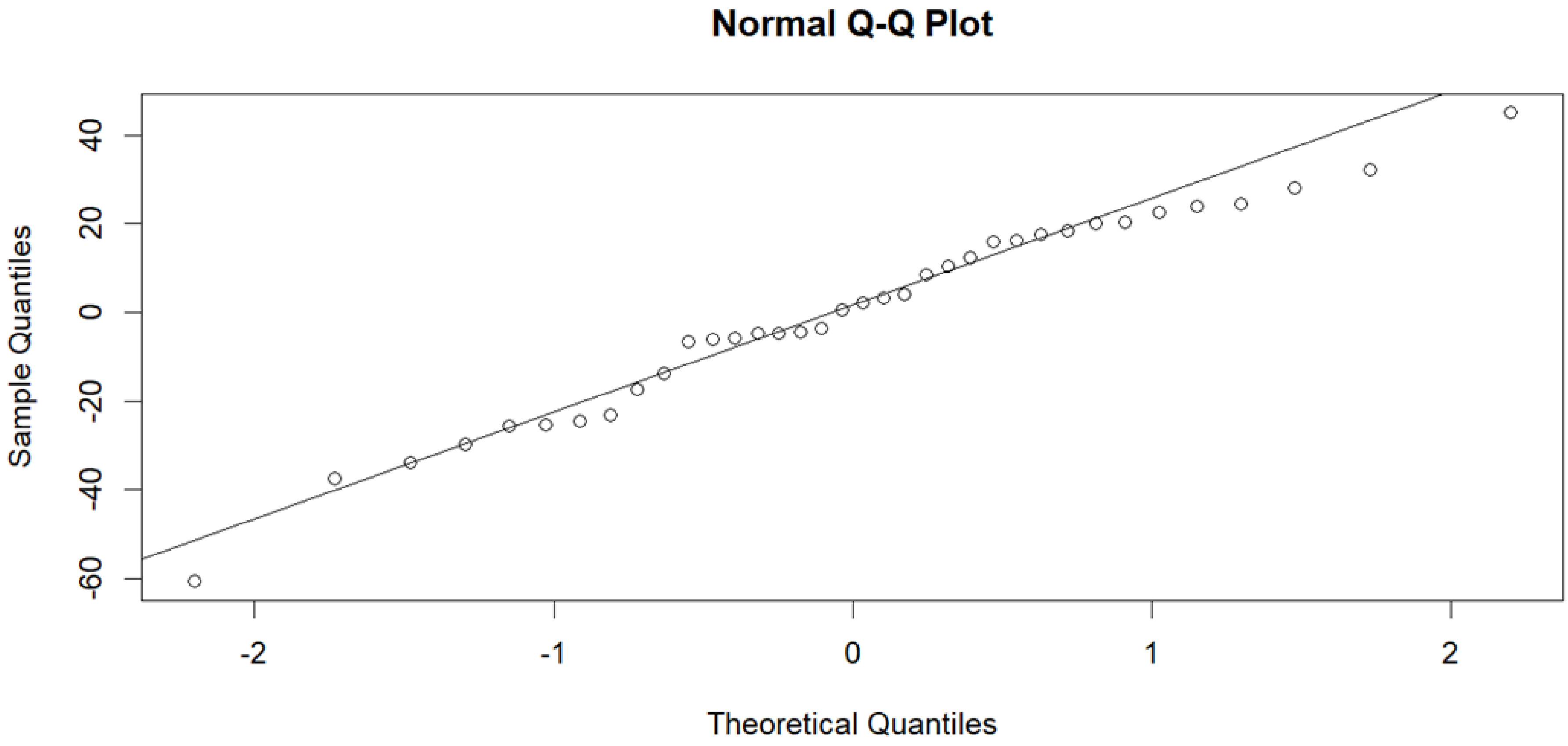
Ejemplo de análisis de varianza de dos vías

- Debemos verificar los supuestos del ANOVA:
  - Normalidad
  - Homocedasticidad
  - Independencia
- Esto se debe realizar sobre los residuales.

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

Para normalidad podemos inicialmente ver el gráfico qq-plot:



Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

- Ahora ejecutaremos el estadístico Shapiro-Wilks sobre los residuales:

```
> anova_orden <- aov(TiempoMS ~ Estructura * Hilos, data = ord)
> shapiro.test(anova_orden$residuals)

	Shapiro-Wilk normality test

data:  anova_orden$residuals
W = 0.97606, p-value = 0.6117
```

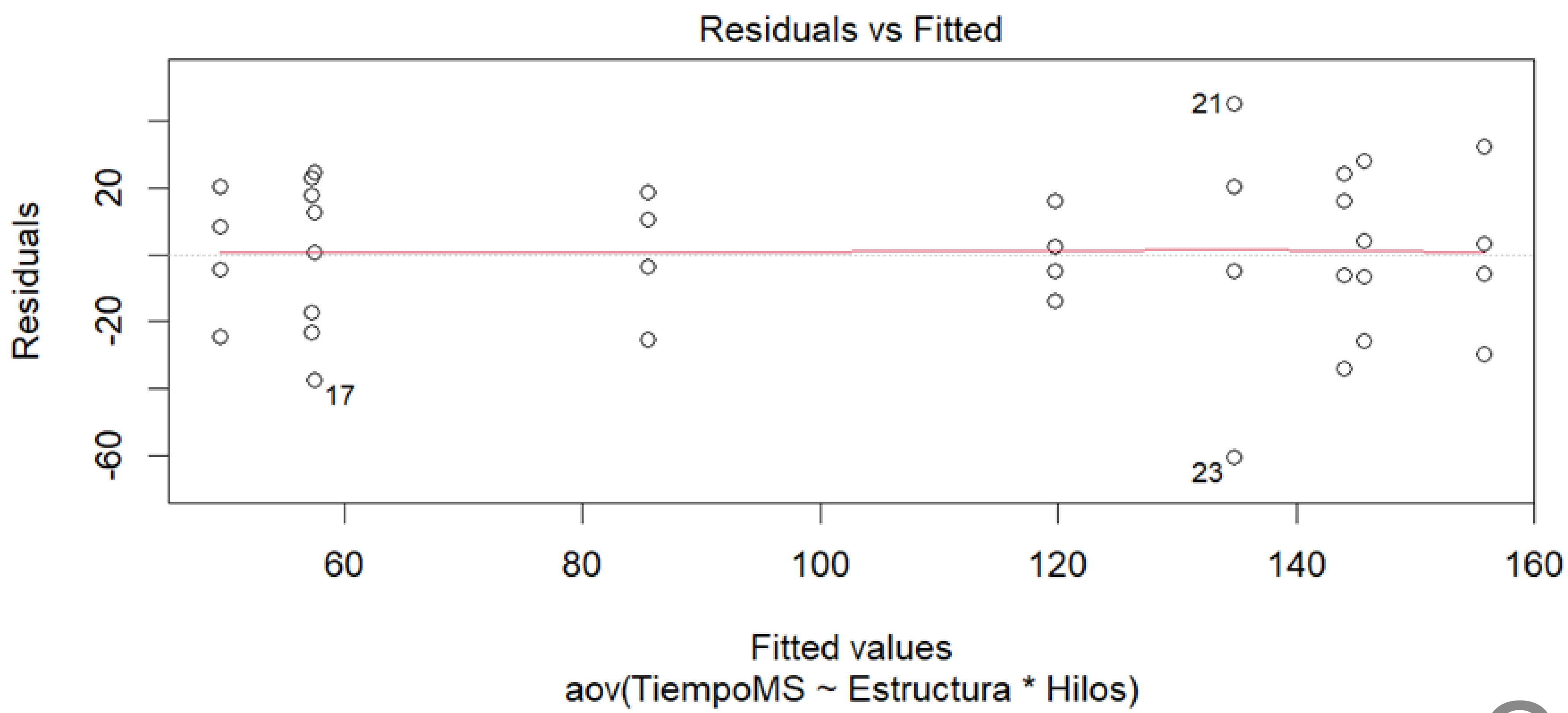
- Como el valor p es mayor que 0.05, no hay razones para rechazar la hipótesis nula de normalidad de los residuales.

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

Para la igualdad de varianzas podemos ver uno de los gráficos de residuales vrs valores ajustados, que se obtiene con:

plot(aov_orden)



Ejemplo de análisis de varianza de dos vías

- Ahora ejecutaremos un estadístico sobre los residuales, la prueba Levene:

```
> residuals_aov_orden <- resid(aov_orden)
> leveneTest(residuals_aov_orden ~ Estructura * Hilos, data = ord)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  8  0.7996 0.6081
      27
```

- Podemos suponer que las varianzas son homogéneas dado que el valor p = 0.6081, mayor que 0.05.

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

- Otra forma de ejecutar la prueba de Levene:

```
> leveneTest(TiempoMS ~ Estructura * Hilos, data = ord)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  8  0.7996 0.6081
      27
```

- El resultado es el mismo: podemos suponer que las varianzas son homogéneas.

Ejemplo de análisis de varianza de dos vías

- Para usar la prueba de Bartlett, más robusta cuando hay normalidad, debemos crear una columna con la combinación de los factores, usando la función *interaction*:

```
ord$Group <- interaction(ord$Estructura, ord$Hilos, sep = " - ")
```

	Observ	Estructura	Hilos	TiempoMS	Group
1	1	Arbol	4	122	Arbol - 4
2	2	Hash	4	40	Hash - 4
3	3	Array	4	139	Array - 4
4	4	Array	4	174	Array - 4
5	5	Hash	4	34	Hash - 4
6	6	Array	1	110	Array - 1

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

- Ejecutamos la prueba de Bartlett, creando primero un vector con los residuales:

```
> residuals_aov_orden <- resid(aov_orden)
> bartlett.test(residuals_ord ~ Group, data = ord)

        Bartlett test of homogeneity of variances

data:  residuals_ord by Group
Bartlett's K-squared = 5.2354, df = 8, p-value = 0.7321
```

- El resultado es el mismo: podemos suponer que las varianzas son homogéneas.

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

- También podemos ejecutar la prueba Fligner-Killeen:

```
> fligner.test(residuals_ord ~ Group, data = ord)

        Fligner-Killeen test of homogeneity of variances

data:  residuals_ord by Group
Fligner-Killeen:med chi-squared = 5.5878, df = 8, p-value = 0.6933
```

- Podemos suponer que las varianzas son homogéneas.

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

- Bartlett si se cuenta con datos normales y se quiere mayor potencia.
- Levene si no es segura la normalidad o hay ligero incumplimiento a la normalidad.
- Fligner-Killeen si los datos son claramente no normales o incluso ordinales.

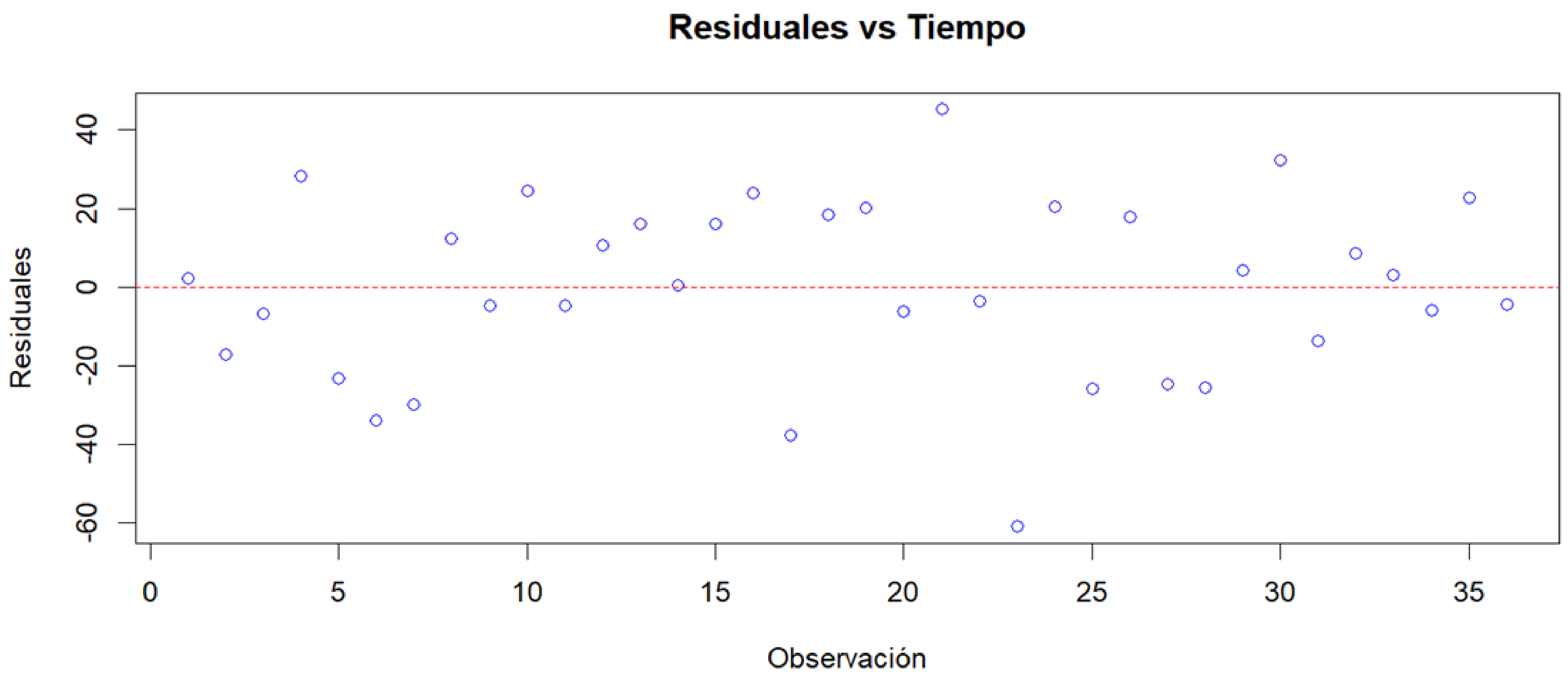
Ejemplo de análisis de varianza de dos vías

Para la independencia podemos ver el gráfico de los residuales en el tiempo, es decir, en el orden que fueron tomados los resultados de acuerdo a la aleatorización del muestreo:

```r
plot(ord$Observ, anova_orden$residuals,
     main = "Residuales vs Tiempo",
     xlab = "Observación",
     ylab = "Residuales",
     pch = 1,       # Tipo de punto
     col = "blue")  # Color de los puntos
abline(h = 0, col = "red", lty = 2)
```

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías



No se aprecian patrones, por lo que podemos suponer independencia de las observaciones.

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

- Este gráfico de residuales vrs tiempo solo tiene sentido si la muestra fue tomada de forma aleatoria y el orden en que se tomaron las muestras se ve reflejado en el conjunto de datos (data frame).
- Dado el gráfico y el hecho de que la muestra fue completamente aleatorizada, podemos suponer independencia de residuales.

Ejemplo de análisis de varianza de dos vías

- Dado que los supuestos se cumplen, podemos utilizar los resultados del ANOVA:

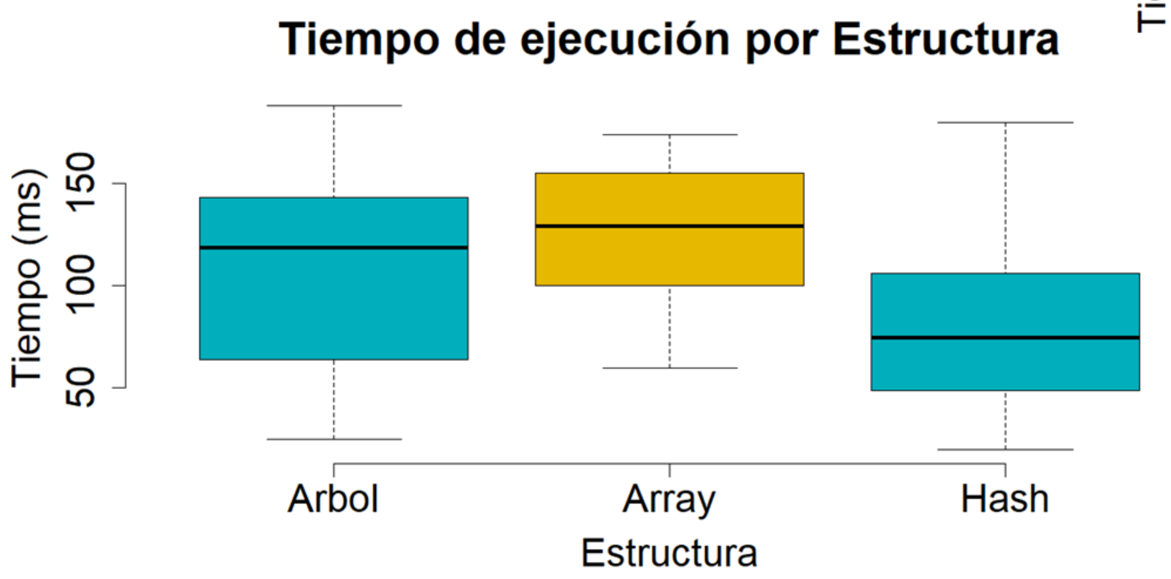
```
                 Df Sum Sq Mean Sq F value   Pr(>F)    
Estructura        2  10684    5342   7.911  0.00198 ** 
Hilos             2  39119   19559  28.968 1.91e-07 ***
Estructura:Hilos  4   9614    2403   3.560  0.01861 *  
Residuals        27  18231     675                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

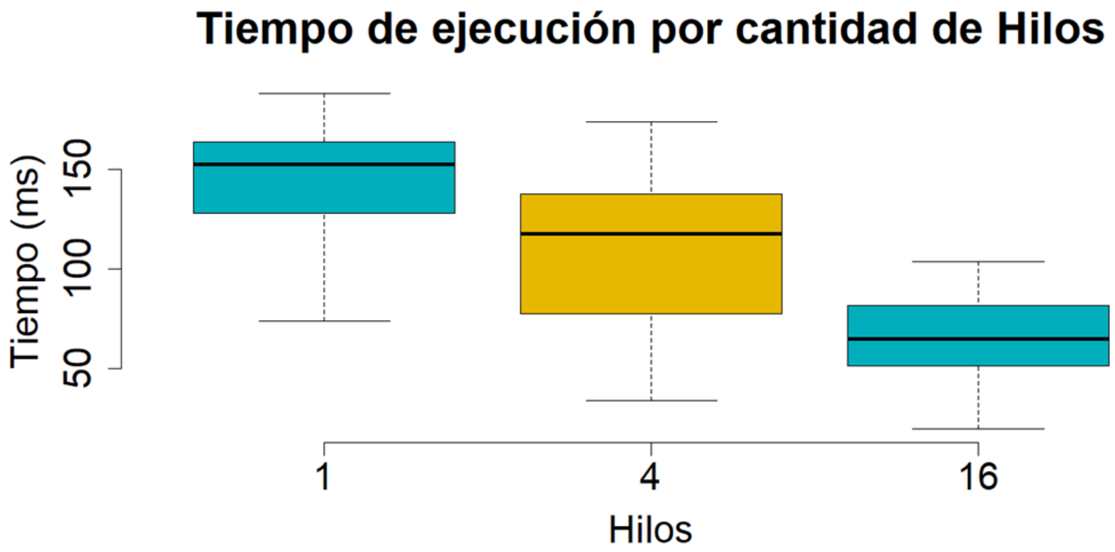
- Sí hay diferencias significativas tanto en la Estructura como en la cantidad de Hilos.

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

- Sí hay diferencias significativas en los factores.





Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

- El **tamaño del efecto** de un ANOVA (η^2 ó eta al cuadrado) es el valor que permite medir cuanta varianza en la variable dependiente cuantitativa es resultado de la influencia de la variable cualitativa independiente, o lo que es lo mismo, cuanto afecta la variable independiente (factor) a la variable dependiente.
- Cuando **solo** tenemos **un factor**, se calcula dividiendo la suma de cuadrados del factor (SSFactor) entre la suma de cuadrados total (SSTotal):

$$\eta^2 = \frac{SSFactor}{SSTotal}$$

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

- Cuando tenemos **varios factores**, se calcula dividiendo la suma de cuadrados del factor (SSFactor) entre la suma de cuadrados del factor más la suma de cuadrados del error (SSFactor + SSError):

$$\eta_p^2 = \frac{SSFactor}{SSFactor + SSError}$$

- Se le llama **eta al cuadrado parcial**, y mide la proporción de la varianza explicada por un factor considerando **solo** su impacto, **sin incluir** la variabilidad debida a otros factores o interacciones.

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

- En R podemos usar la función etaSquared de la biblioteca lsr:

```
> library(lsr)
> etaSquared(aov_orden, anova = TRUE)
                    eta.sq eta.sq.part        SS df        MS         F            p
Estructura       0.1375935   0.3694939 10683.722  2  5341.861  7.911372 1.976083e-03
Hilos            0.5038023   0.6821113 39118.722  2 19559.361 28.967692 1.908596e-07
Estructura:Hilos 0.1238139   0.3452663  9613.778  4  2403.444  3.559535 1.861117e-02
Residuals        0.2347902          NA 18230.750 27   675.213        NA           NA
```

- Se considera al eta.sq.part alrededor de 0.01 como un efecto pequeño, alrededor de 0.06 como mediano, y mayor a 0.14 como un efecto grande.
- En este caso los tres valores de eta al cuadrado parcial son grandes.

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

```
                 Df Sum Sq Mean Sq F value   Pr(>F)    
Estructura        2  10684    5342   7.911  0.00198 ** 
Hilos             2  39119   19559  28.968 1.91e-07 ***
Estructura:Hilos  4   9614    2403   3.560  0.01861 *  
Residuals        27  18231     675                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- También hay interacción significativa entre los dos factores.
- Es necesario ver entre qué grupos se da la diferencia.
- Los métodos de comparaciones múltiples, como la prueba Tukey, son útiles a este respecto.

Ejemplo de análisis de varianza de dos vías

```
> TukeyHSD(aov_orden)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = TiempoMS ~ Estructura * Hilos, data = ord)

$Estructura
                 diff        lwr        upr     p adj
Array-Arbol  16.75000  -9.552344  43.052344 0.2717815
Hash-Arbol  -25.16667 -51.469011   1.135677 0.0627571
Hash-Array  -41.91667 -68.219011 -15.614323 0.0014162

$Hilos
          diff        lwr       upr     p adj
4-1  -37.25000  -63.55234 -10.94766 0.0043788
16-1 -80.66667 -106.96901 -54.36432 0.0000001
16-4 -43.41667  -69.71901 -17.11432 0.0009787
```

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

```
$`Estructura:Hilos`
                     diff        lwr        upr     p adj
Array:1-Arbol:1    -11.75  -73.57318  50.073184 0.9991463
Hash:1-Arbol:1     -21.00  -82.82318  40.823184 0.9616404
Arbol:4-Arbol:1    -36.00  -97.82318  25.823184 0.5819453
Array:4-Arbol:1    -10.00  -71.82318  51.823184 0.9997369
Hash:4-Arbol:1     -98.50 -160.32318 -36.676816 0.0003449
Arbol:16-Arbol:1  -106.25 -168.07318 -44.426816 0.0001152
Array:16-Arbol:1   -70.25 -132.07318  -8.426816 0.0172076
Hash:16-Arbol:1    -98.25 -160.07318 -36.426816 0.0003574
Hash:1-Array:1      -9.25  -71.07318  52.573184 0.9998527
Arbol:4-Array:1    -24.25  -86.07318  37.573184 0.9165175
Array:4-Array:1      1.75  -60.07318  63.573184 1.0000000
Hash:4-Array:1     -86.75 -148.57318 -24.926816 0.0018119
```

… un total de 36 filas

Ejemplo de análisis de varianza de dos vías

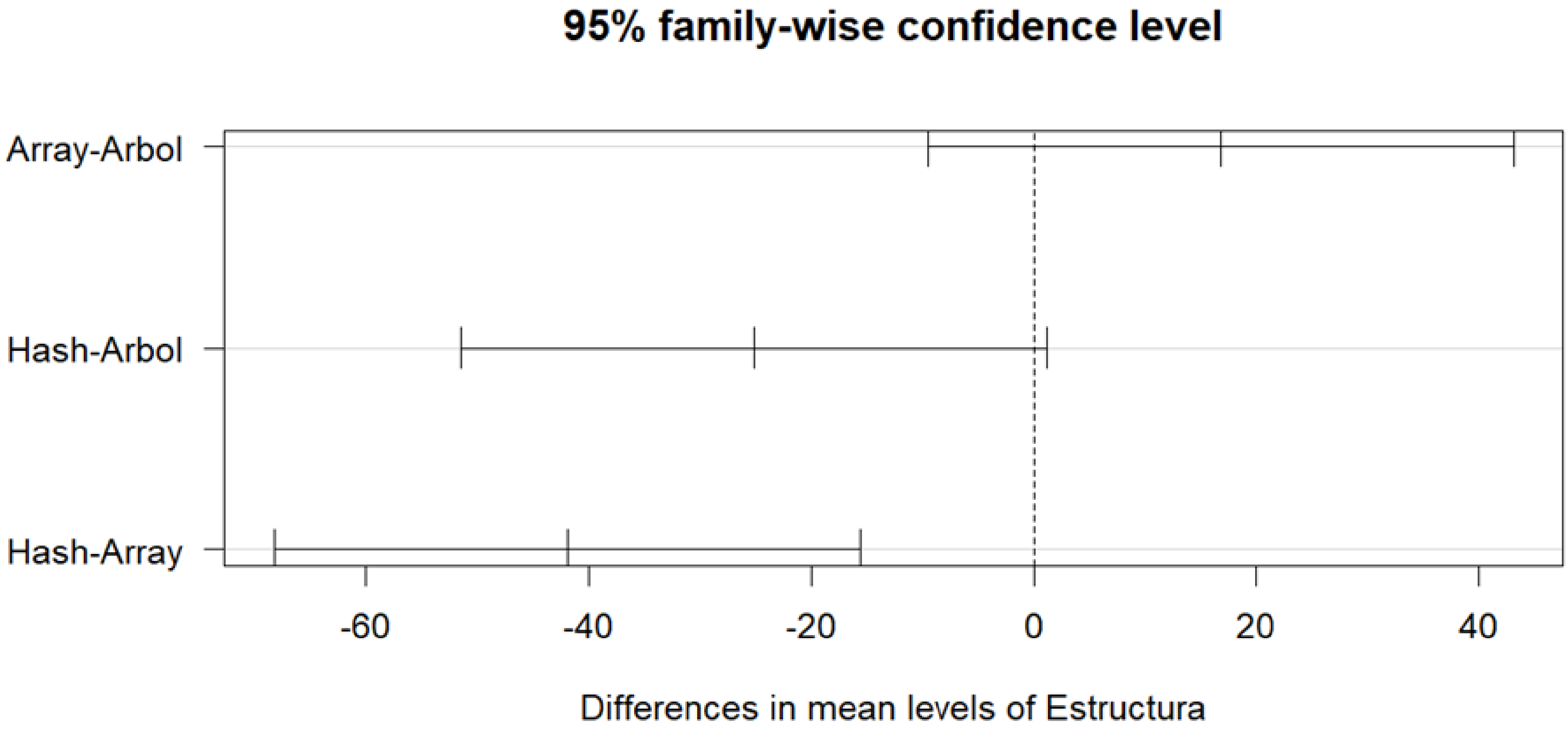
- También se puede analizar estas comparaciones visualmente:

```
tukey_result <- TukeyHSD(aov_orden)
plot(tukey_result, las = 1)
```

- Se generan tres gráficos:

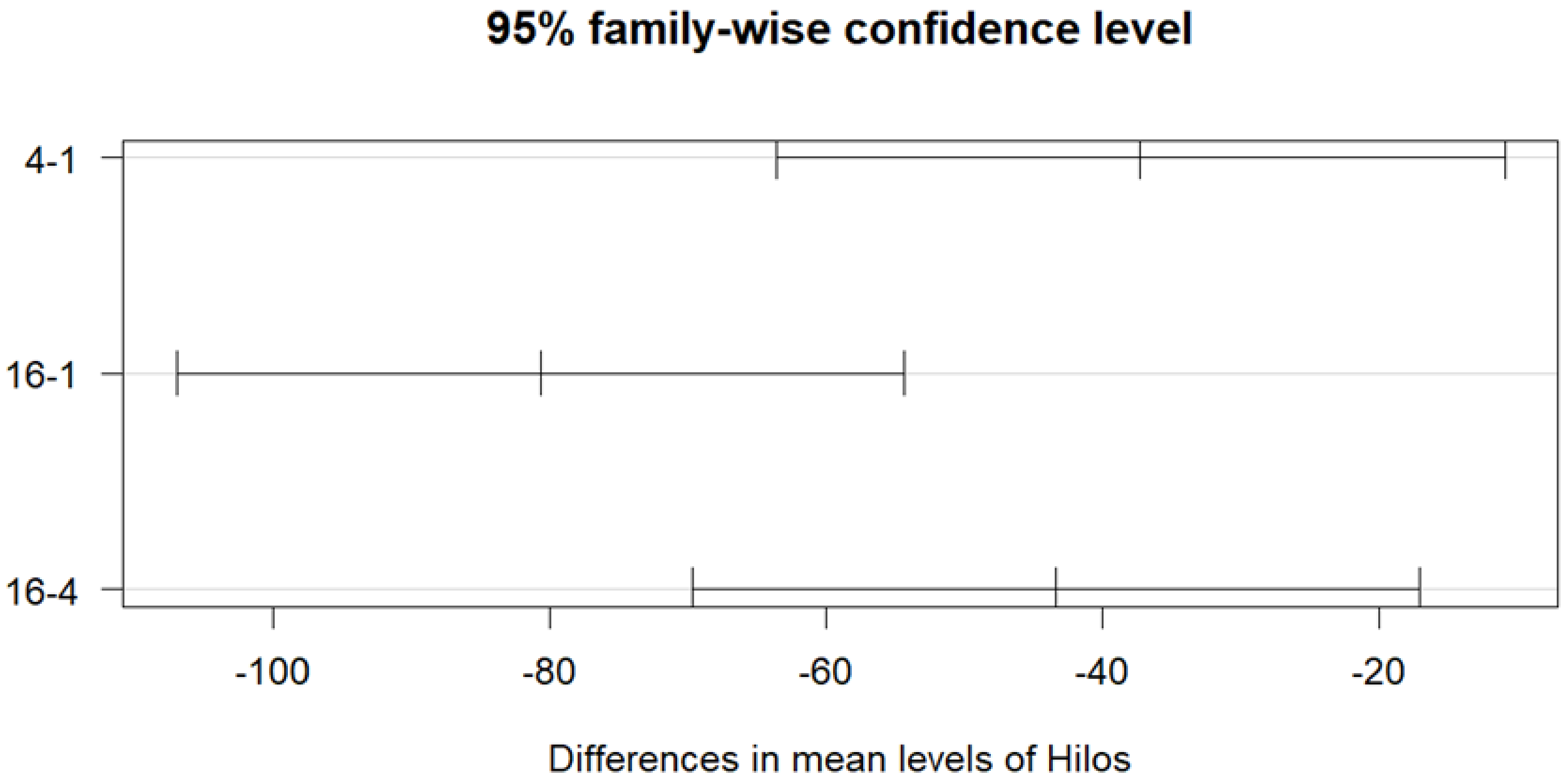
Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías



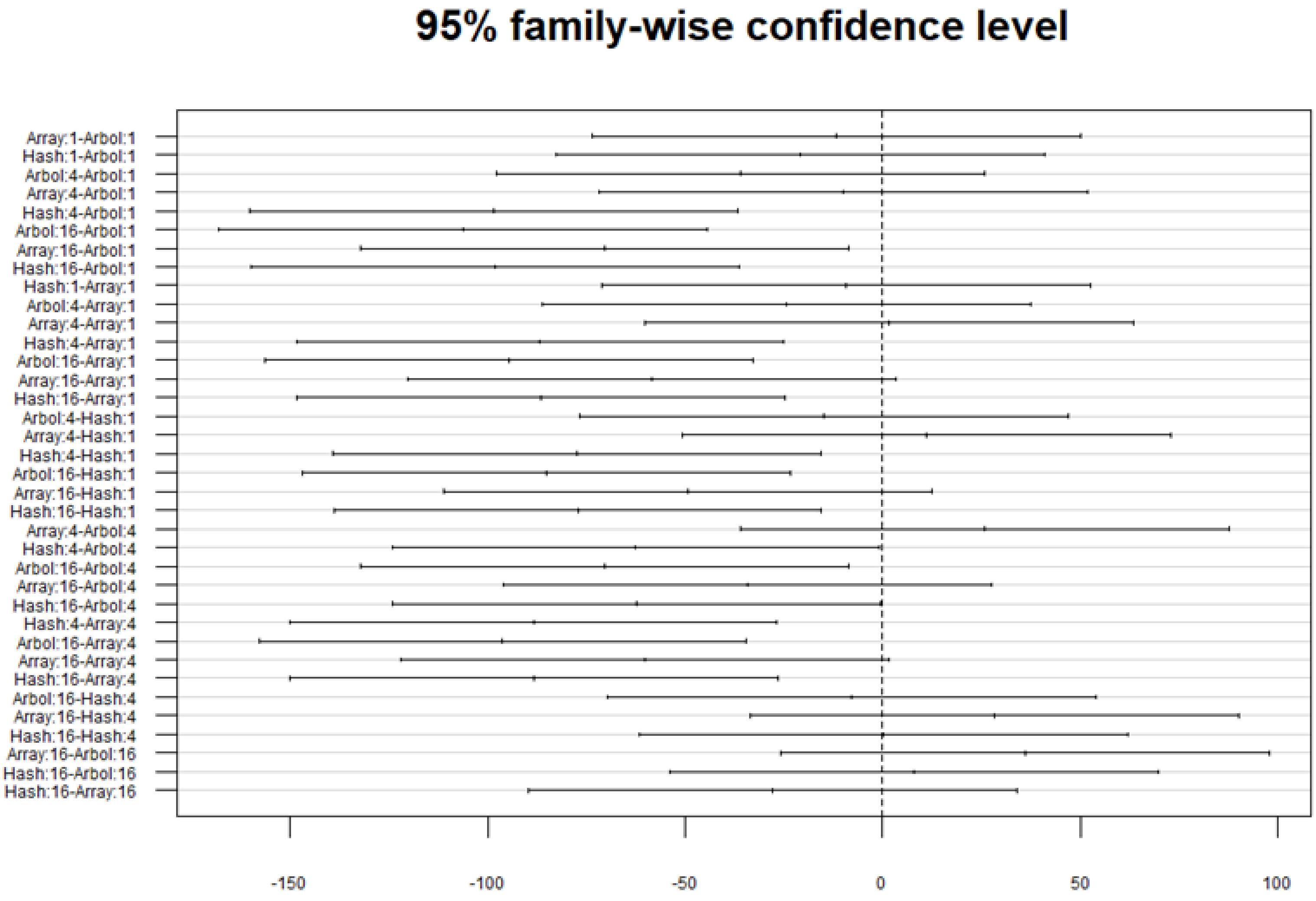
Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías



Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías



Ejemplo de análisis de varianza de dos vías

- También se puede revisar solo las comparaciones con valor p < 0.05:

```
tukey_result <- TukeyHSD(aov_orden)

# Convertir el resultado a un data frame
tukey_df <- as.data.frame(tukey_result$`Estructura:Hilos`)

# Filtrar para obtener solo las comparaciones con valores p < 0.05
filtered_tukey_df <- tukey_df[tukey_df$`p adj` < 0.05, ]

# Mostrar el resultado filtrando solo los p < 0.05
print(filtered_tukey_df)
```

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

	diff	lwr	upr	p adj
Hash:4-Arbol:1	-98.50	-160.3232	-36.676816	0.0003449173
Arbol:16-Arbol:1	-106.25	-168.0732	-44.426816	0.0001151507
Array:16-Arbol:1	-70.25	-132.0732	-8.426816	0.0172076171
Hash:16-Arbol:1	-98.25	-160.0732	-36.426816	0.0003573540
Hash:4-Array:1	-86.75	-148.5732	-24.926816	0.0018118976
Arbol:16-Array:1	-94.50	-156.3232	-32.676816	0.0006077696
Hash:16-Array:1	-86.50	-148.3232	-24.676816	0.0018764896
Hash:4-Hash:1	-77.50	-139.3232	-15.676816	0.0065212148
Arbol:16-Hash:1	-85.25	-147.0732	-23.426816	0.0022350689
Hash:16-Hash:1	-77.25	-139.0732	-15.426816	0.0067471137
Hash:4-Arbol:4	-62.50	-124.3232	-0.676816	0.0460387808
Arbol:16-Arbol:4	-70.25	-132.0732	-8.426816	0.0172076171
Hash:16-Arbol:4	-62.25	-124.0732	-0.426816	0.0474674659
Hash:4-Array:4	-88.50	-150.3232	-26.676816	0.0014172574
Arbol:16-Array:4	-96.25	-158.0732	-34.426816	0.0004743996
Hash:16-Array:4	-88.25	-150.0732	-26.426816	0.0014679433

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

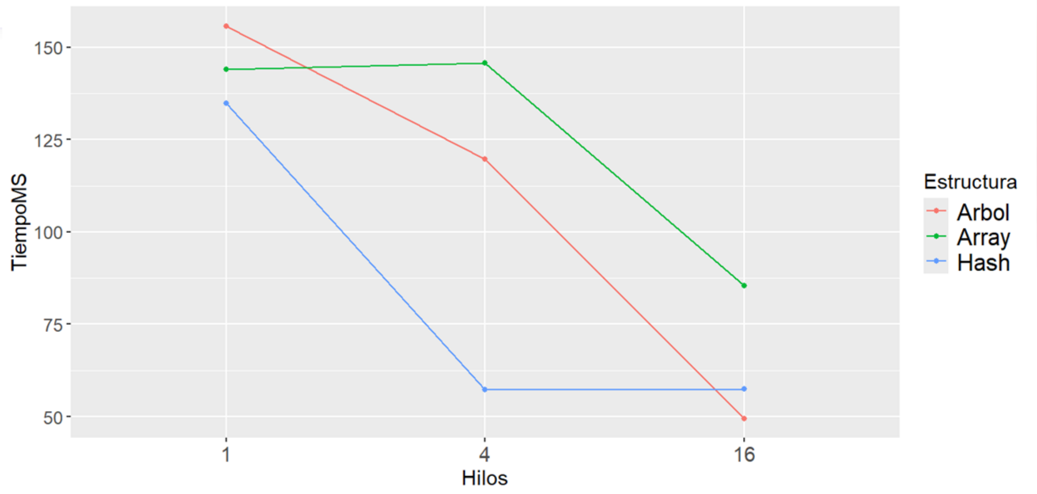
- Cuando la interacción es significativa, las comparaciones entre las medias de uno de los factores (por ejemplo A) pueden ser nubladas por la interacción AB.
- Una forma de abordar esta cuestión consiste en fijar el factor B en un nivel específico y revisar la prueba de Tukey del factor A con ese nivel B.
- En el ejemplo, supongamos que el interés es detectar las diferencias entre las medias de los tres tipos de Estructura.
- Puesto que la interacción es significativa, esta comparación se hace con un solo nivel de la cantidad de hilos, por ejemplo el segundo nivel (4 hilos).

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

- De la prueba Tukey obtenemos:

	diff	lwr	upr	p adj
Array:4-Arbol:4	26.00	-35.82318	87.823184	0.8822881
Hash:4-Arbol:4	-62.50	-124.32318	-0.676816	0.0460388
Hash:4-Array:4	-88.50	-150.32318	-26.676816	0.0014173



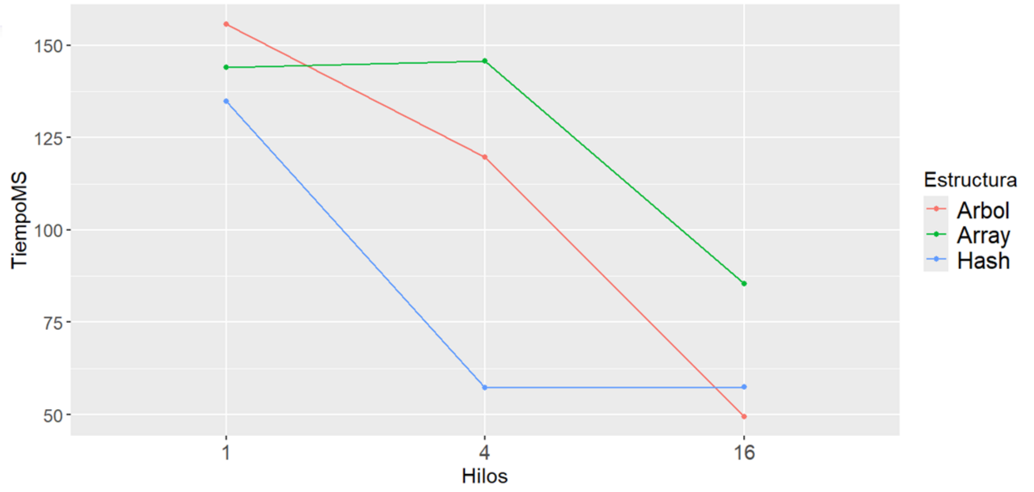
- Este análisis indica que, con el nivel de 4 hilos, el tiempo de respuesta es el mismo para las estructuras Array y Árbol, y que el tiempo de Hash es significativamente menor (comparado con Array y Árbol).

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

- Si vemos los resultados de la prueba Tukey para 16 hilos:

	diff	lwr	upr	p adj
Array:16-Arbol:16	36.00	-25.82318	97.823184	0.5819453
Hash:16-Arbol:16	8.00	-53.82318	69.823184	0.9999508
Hash:16-Array:16	-28.00	-89.82318	33.823184	0.8347331



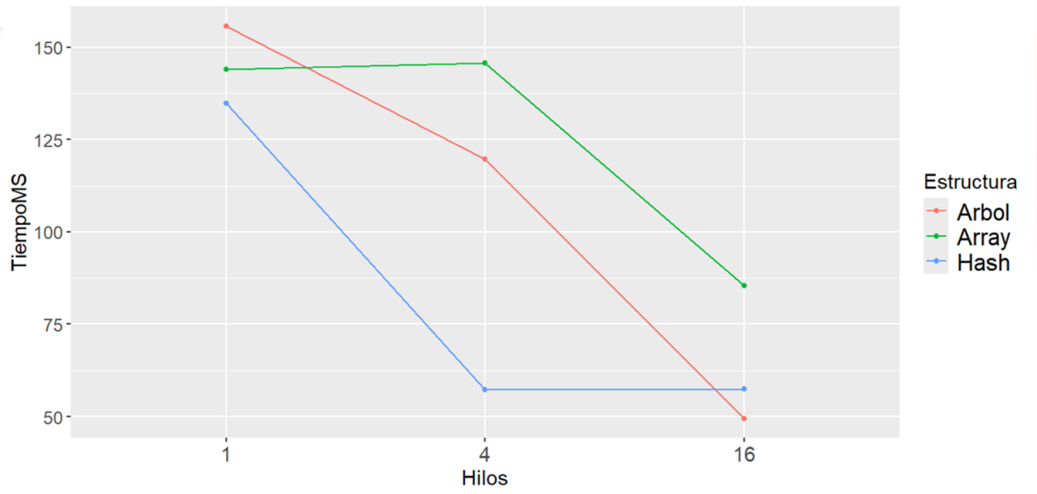
- Se puede observar que los tiempos promedio no presentan diferencias significativas para ninguna de las tres estructuras.

Diseño factorial de dos factores

Ejemplo de análisis de varianza de dos vías

- De forma equivalente se puede fijar el nivel de la estructura y ver las variaciones en la cantidad de hilos:

	diff	lwr	upr	p adj
Arbol:4-Arbol:1	-36.00	-97.82318	25.823184	0.5819453
Arbol:16-Arbol:1	-106.25	-168.07318	-44.426816	0.0001152
Arbol:16-Arbol:4	-70.25	-132.07318	-8.426816	0.0172076



- Se puede observar que con la estructura Árbol, el tiempo promedio es el mismo para 1 y 4 hilos, y que el tiempo promedio para 16 hilos es significativamente menor (comparado con 1 y 4 hilos).

Referencias

Montgomery, D.C. (2013).

Design of Experiments.

John Wiley & sons.