EVIDENCIAS EN PEDIATRÍA
Toma de decisiones clínicas basadas en las mejores pruebas científicas
www.evidenciasenpediatria.es

Fundamentos de medicina basada en la evidencia

Comparación de más de dos medias. Análisis de la varianza

Molina Arias M[1], Ochoa Sangrador C[2], Ortega Páez E[3]
[1]Servicio de Gastroenterología. Hospital Universitario La Paz. Madrid. España.
[2]Servicio de Pediatría. Hospital Virgen de la Concha. Zamora. España.
[3]UGC de Maracena. Distrito Granada-Metropolitano. Granada. España.

Correspondencia: Manuel Molina Arias, mma1961@gmail.com

Palabras clave en español: estadística; inferencia estadística; comparación de medias; análisis de la varianza; ANOVA.
Palabras clave en inglés: statistics; statistical inference; comparing means; analysis of variance; ANOVA.

Fecha de recepción: 21 de diciembre de 2020 • **Fecha de aceptación:** 8 de enero de 2021
Fecha de publicación del artículo: 13 de enero de 2021

Evid Pediatr. 2021;17:11.

CÓMO CITAR ESTE ARTÍCULO

Molina Arias M, Ochoa Sangrador C, Ortega Páez E. Comparación de dos medias. Análisis de la varianza. Evid Pediatr. 2021;17:11.

Este artículo está disponible en: http://www.evidenciasenpediatria.es/EnlaceArticulo?ref=2021;17:11.

Comparación de más de dos medias. Análisis de la varianza

Molina Arias M[1], Ochoa Sangrador C[2], Ortega Páez E[3]
[1]*Servicio de Gastroenterología. Hospital Universitario La Paz. Madrid. España.*
[2]*Servicio de Pediatría. Hospital Virgen de la Concha. Zamora. España.*
[3]*UGC de Maracena. Distrito Granada-Metropolitano. Granada. España.*

Correspondencia: Manuel Molina Arias, mma1961@gmail.com

Vimos en un artículo anterior de esta serie las técnicas disponibles tanto para comparar una media con un valor poblacional como para comparar dos medias entre sí, teniendo en cuenta, en este segundo caso, si se trataba de medias de muestras independientes o relacionadas. Para ello, la prueba que utilizamos con más frecuencia es la de la t de Student, basada en el valor de significación del estadístico t calculado a partir de la diferencia de las dos medias y el error estándar de esta diferencia.

En el presente artículo abordaremos la comparación de más de dos medias. La primera posibilidad que se nos puede ocurrir en estos casos es la de comparar las medias dos a dos utilizando una de las pruebas descritas para la comparación de medias, como la de la t de Student. Sin embargo, como veremos más adelante, esto no sería correcto, ya que las sucesivas comparaciones aumentarían la probabilidad de cometer un error de tipo I, esto es, de detectar un falso positivo y obtener diferencias significativas solo por azar.

Por ello, para comparar más de dos medias, utilizaremos una técnica denominada análisis de la varianza (ANOVA). A pesar de su nombre, el ANOVA no compara varianzas, sino medias, aunque se sirve de las varianzas de las variables para realizarlo.

Existen varios tipos de ANOVA, aunque en este artículo nos centraremos en su forma más sencilla, el ANOVA de un factor para medias independientes.

CÁLCULO DEL ANOVA

La hipótesis nula del ANOVA plantea que no existen diferencias entre las medias de los grupos estudiados. Por su parte, la hipótesis alternativa plantea que sí existen diferencias, al menos, entre dos de las medias estudiadas, pero no nos dice entre cuáles. Esto tendremos que determinarlo en un paso adicional, denominado habitualmente análisis *post hoc*, y que veremos más adelante.

Veamos en qué se basa y cómo se lleva a cabo el procedimiento del ANOVA.

Los primeros estadísticos que podemos calcular son la media y la varianza globales de toda la muestra. A continuación, podemos calcular las medias agrupadas por el factor que queremos estudiar. Son estas medias las que estamos interesados en comparar.

A continuación, podemos suponer que el valor de la variable cuantitativa de un individuo concreto puede igualarse a la suma de tres componentes: la media de su grupo, la variabilidad debida a las características de ese grupo y la variabilidad debida al azar (aleatoria):

$$x_{ij} = \bar{x}_j + s_j + e,$$

donde x_{ij} representa el valor del individuo *i* del grupo *j*, $\bar{x}_j$ la media del grupo *j* y s_j la varianza debida al grupo y *e* expresa el error aleatorio.

Si la hipótesis nula es cierta y no hay diferencias entre los distintos grupos, el componente s_j será pequeño, similar al componente aleatorio (e) o incluso menor, así que el cociente entre ambos valdrá 1 o menos de 1. En el caso de que sí existan diferencias entre grupos, el valor del cociente será mayor que 1, ya que la varianza entre grupos será más alta mientras que el error aleatorio se mantendrá de forma similar. Una vez entendido esto, el siguiente paso será identificar los componentes de la varianza.

Como ya sabemos, la varianza es el promedio de los cuadrados de la distancia de cada valor respecto a su media, por lo que su numerador se conoce como **suma total de cuadrados**. Así, podemos entender una varianza como el cociente entre una suma de cuadrados y sus grados de libertad. Por tanto, la varianza total puede definirse como la suma de cuadrados total dividida por el número de grados de libertad, que son $n - 1$ (n es el tamaño de la muestra total).

Esta **suma total de cuadrados** puede descomponerse en los dos componentes ya referidos al hablar del valor de cada individuo. El primer componente es la suma de cuadrados entre grupos diferentes. También se conoce como variabilidad entre los tratamientos y como variabilidad del modelo o de la regresión. El segundo componente es la suma de cuadrados

intragrupos, también denominado suma de cuadrados de los errores o residuales.

Así, la suma de cuadrados (SC) total puede expresarse de la forma siguiente:

SC total = SC entre grupos + SC intragrupos (residual),

siendo *n* el tamaño muestral total y *k* el número de grupos a analizar, el número total de grados de libertad (*n* – 1) se divide entre los dos componentes, siendo de *k* – 1 los grados de libertad de la variabilidad entre grupos y *n* – *k* la de la variabilidad intragrupos.

Existen fórmulas para calcular las diferentes sumas de cuadrados, pero no merece la pena entrar en detalle, ya que lo recomendable es utilizar uno de los programas estadísticos que lo calculan de forma automática.

Con todos estos componentes podremos, finalmente, elaborar la tabla del ANOVA (tabla 1).

Como vemos, dividiendo la suma de cuadrados entre grupos e intragrupos por sus respectivos grados de libertad, obtendremos las varianzas entre grupos e intragrupos, respectivamente. El cociente de las dos varianzas sigue una distribución de la F de Snedecor con *k* – 1, *n* – *k* grados de libertad, lo que nos permitirá calcular la probabilidad del valor encontrado. Bajo el supuesto de la hipótesis nula, el valor de *F* debe estar próximo a la unidad. Cuanto más se aleje de 1, más probable será que la diferencia entre las dos varianzas no sea debida al azar y podamos, así, rechazar la hipótesis nula.

ANÁLISIS DE MEDIAS *POST HOC*

Tras realizar el ANOVA, rechazar la hipótesis nula implicará que hay diferencia estadísticamente significativa entre, al menos, dos de las medias comparadas, pero no sabremos entre cuáles, por lo que tendremos que comparar las medias dos a dos para saber qué pareja o parejas difieren de forma significativa.

Como vimos en un artículo previo, la comparación de medias es sencilla. El problema surge al tener que realizar comparaciones múltiples. Aunque la probabilidad de cometer un error de tipo I al comparar dos medias es de 0,05, si realizamos comparaciones múltiples está probabilidad de encontrar un falso positivo aumenta. Para hacernos una idea, si hacemos 5 comparaciones, la probabilidad de encontrar un falso positivo solo por azar será de 0,22, pero si son 20 comparaciones, esta sube hasta 0,64.

Por este motivo, es necesario aplicar una corrección para comparaciones múltiples que mantenga la probabilidad global de error de tipo I por debajo del límite establecido habitualmente de 0,05.

La más sencilla es la corrección de Bonferroni, que calcula un valor de *p* "penalizado" que se establece como nuevo umbral de significación estadística, en lugar del habitual $p < 0,05$. De forma sencilla, podemos decir que el nuevo valor de p se obtendrá dividiendo 0,05 entre el número de contrastes. Por ejemplo, si hacemos 6 contrastes (comparamos 6 parejas de medias), el valor de *p* para considerar la diferencia como estadísticamente significativa será de 0,05 / 6 = 0,008 (y no 0,05).

Hay muchos más métodos para realizar una corrección para comparaciones múltiples. Habitualmente, el programa estadístico nos especificará cuál ha utilizado. Algunos de los más usados son los de Tukey, de Scheffé, de Dunnett o de Sidak.

CONDICIONES PARA REALIZAR UN ANOVA

Para poder realizar un ANOVA de una vía o factor debemos verificar que se cumplen tres condiciones:

1. Independencia de las observaciones: las observaciones deben ser aleatorias y los grupos deben ser independientes.

2. La variable cuantitativa debe distribuirse de forma normal en cada uno de los grupos de la variable cualitativa.

3. Homocedasticidad: la varianza dentro de los grupos debe ser similar en todos ellos.

Aunque el ANOVA es bastante robusto en la ausencia de normalidad, si la asimetría es muy grande (grupos con tamaños muy diferentes), será conveniente realizar su equivalente no paramétrico, que es la prueba de Kruskal-Wallis. ANOVA es aún más robusto ante la ausencia de homocedasticidad, aunque si los grupos son de tamaño desigual convendrá realizarlo aplicando la corrección de Welch (lo que en algunos programas se denomina ANOVA heterocedástico).

Tabla 1. Tabla resumen del ANOVA.

Variabilidad	Suma de cuadrados (SC)	gl	Varianzas	F
Entre grupos	SC entre grupos (SC_E)	$k–1$	$S_E = \frac{SC_E}{k-1}$	$\frac{S_E}{S_I}$
Intragrupos (residual)	SC intragrupos (SC_I)	$n–k$	$S_I = \frac{SC_I}{n-k}$	
Total	SC total (SC_T)	$n–1$		

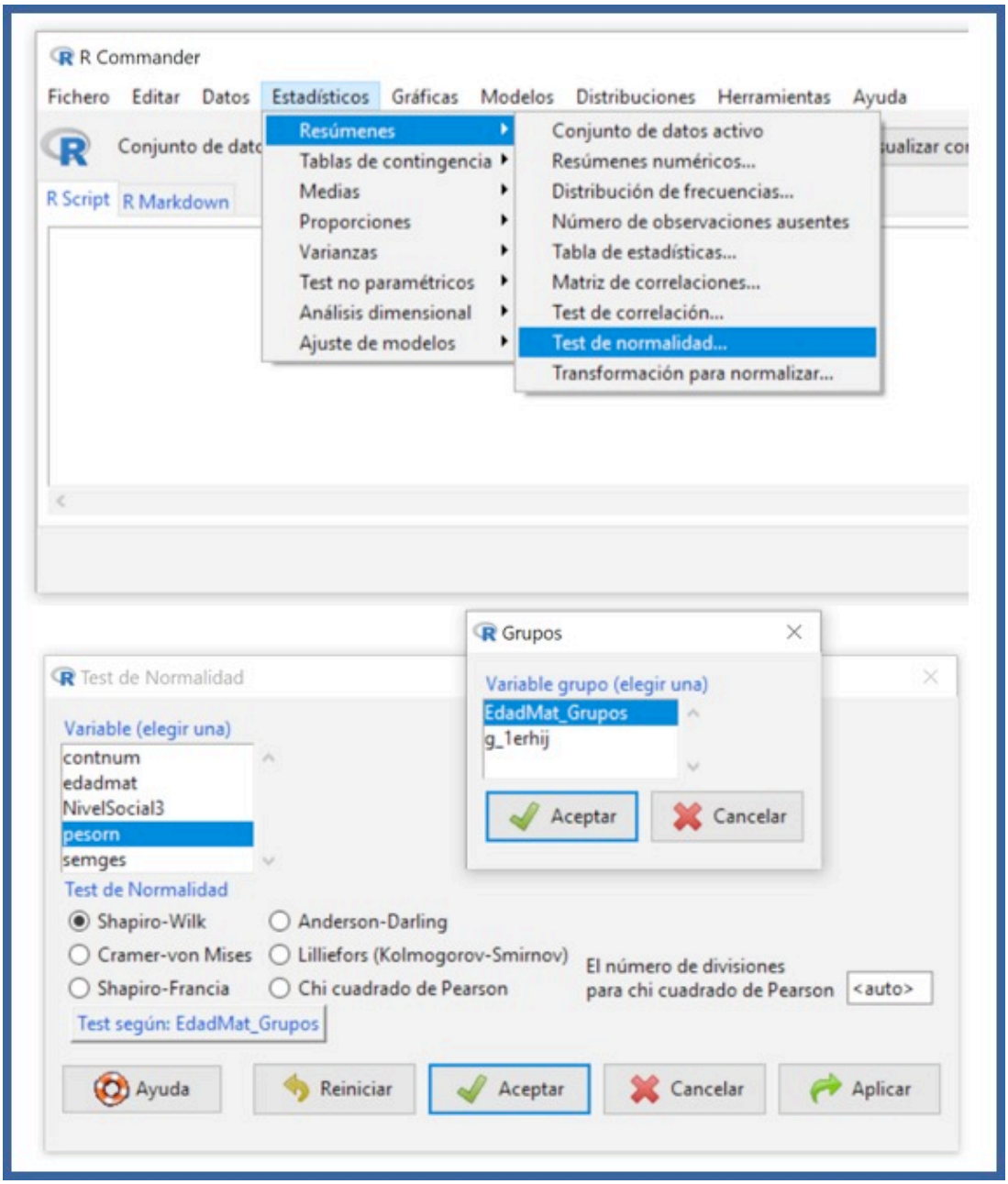
EJEMPLO DE ANOVA

Tenemos una base de datos con 197 registros de pesos de recién nacidos a término con la que trataremos de ver si la edad de la madre influye en el peso al nacer de estos niños. Para ello, compararemos las medias del peso al nacimiento (variable cuantitativa) entre tres grupos diferentes en los que hemos categorizado la edad materna (variable cualitativa con 3 factores o categorías: "menor de 19 años", "de 19 a 34 años" y "mayor o igual a 35 años"). Puede descargarse la base de datos en http://archivos.evidenciasenpediatria.es/files/43-216-RUTA/PesoRN_EdadMadre.RData. Realizaremos un ANOVA utilizando el programa estadístico R con la interfaz RCommander.

Una vez cargada la base de datos, comprobaremos que la variable cuantitativa sigue una distribución normal en las tres categorías de la variable cualitativa. Realizamos, por ejemplo, una prueba de Shapiro-Wilk seleccionando la opción de RCommander "Estadísticos/Resúmenes/Test de normalidad" (figura 1). En la ventana emergente marcamos la variable "pesorn", la opción "Shapiro-Wilk" y pulsamos el botón "Test por grupos" para seleccionar la variable "EdadMat_Grupos".

El resultado nos dice que los valores de p del contraste son 0,25, 0,64 y 0,31 para los grupos de menor de 19 años, de 19 a 34 y mayor o igual a 35 años, respectivamente. Al ser los valores de p >0,05 no podemos rechazar la hipótesis nula que, en el caso de la prueba de Shapiro-Wilk, supone que los datos se ajustan a una distribución normal.

Figura 1. Prueba de Shapiro-Wilk para comprobar si se cumple el supuesto de normalidad.



Para mayor seguridad, completamos esta prueba con un método gráfico, como el del gráfico de quantiles. Para ello, seleccionamos las opciones "Gráficas/Gráfica de comparación de quantiles". En la ventana emergente marcamos la variable "pesorn" y pulsamos el botón "Gráfica por grupos" para seleccionar la variable "EdadMat_Grupos" (figura 2). Podemos ver los gráficos en la figura 3: los datos se ajustan razonablemente a la diagonal del gráfico, lo que quiere decir que los cuantiles son similares a los teóricos si se ajustasen a una distribución normal. Corroboramos así que el peso al nacimiento se distribuye de forma normal en los tres grupos de la variable "EdadMat_Grupos".

Comprobaremos el supuesto de homocedasticidad realizando una prueba de Barlett. Seleccionamos "Estadísticos/Varianzas/Test de Barlett" y seleccionaremos las variable "EdadMat_Grupos" y "pesorn". Obtenemos un valor de p = 0,13, con lo que no podemos rechazar la hipótesis nula de igualdad de varianzas.

Nuestras observaciones son independientes y, como hemos comprobado, se cumplen los supuestos de normalidad y homocedasticidad. Podemos proceder con el ANOVA.

Seleccionamos las opciones "Estadísticos/Medias/ANOVA de un factor". En la ventana emergente seleccionamos las dos variables, "EdadMat_Grupos" y "pesorn". Debajo de las variables tenemos dos casillas que podemos marcar. La primera indicará al programa que realice el estudio *post hoc* de comparación de medias. La segunda, que aplique al ANOVA la corrección de Welch si no se cumple el supuesto de homocedasticidad. Marcamos solo la primera opción y pulsamos "Aceptar" (figura 4).

Figura 2. Obtención de los gráficos de comparación de quantiles de la variable longitud del recién nacido por grupos de edad materna.

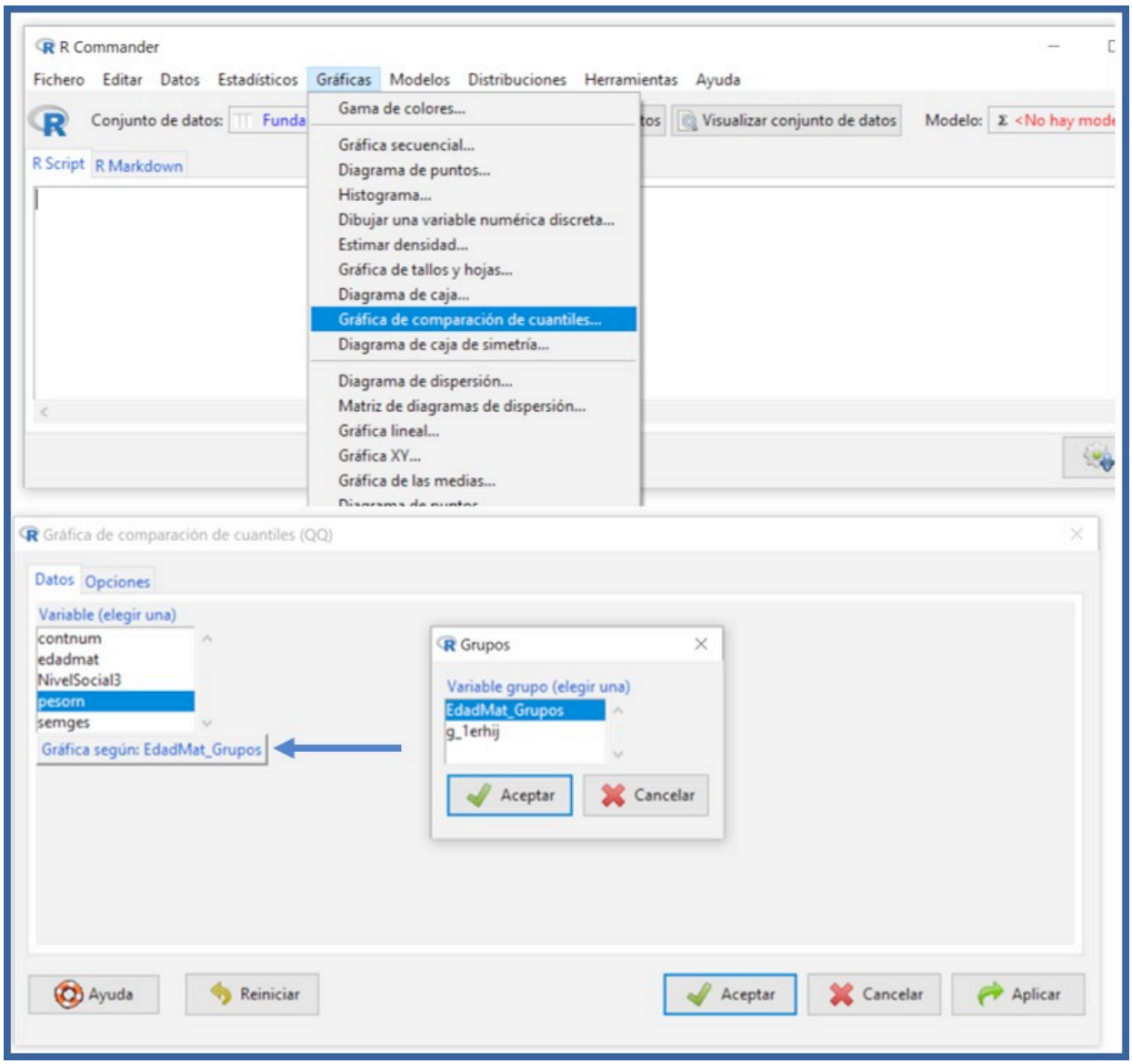
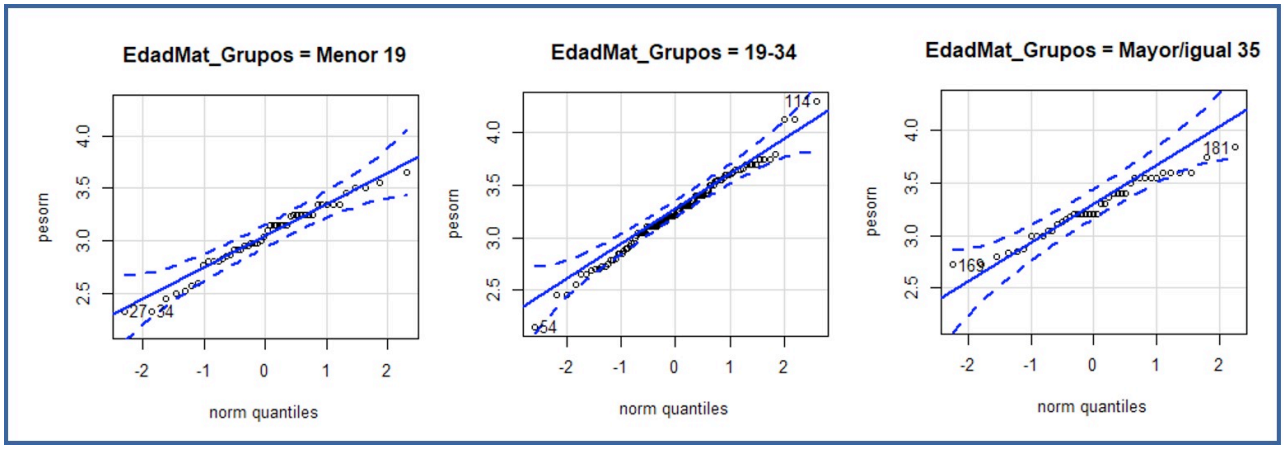


Figura 3. Gráficos de comparación de quantiles de la variable longitud del recién nacido por grupos de edad materna.



Lo primero que nos muestra R es la tabla del ANOVA (tabla 2). En las filas podemos ver los datos que hacen referencia a la variabilidad entre grupos (EdadMat_Grupos) e intragrupo (residuals). En columnas se muestran los grados de libertad (Df), las sumas de cuadrados (Sum Sq), las varianzas de cada componente (Mean Sq = Sum Sq/Df), el valor de F (F value, el

Figura 4. Obtención del ANOVA.

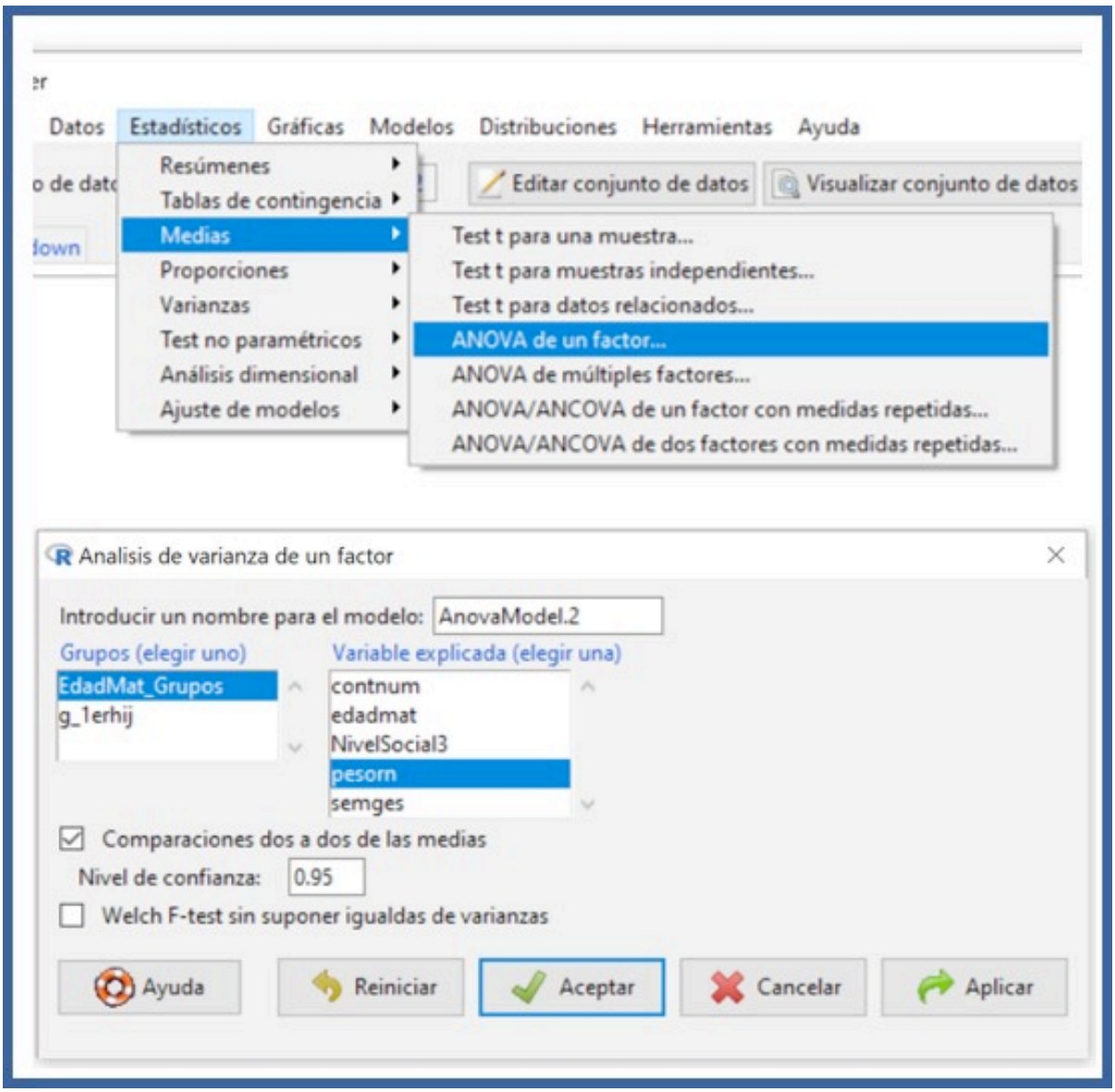


Tabla 2. Tabla resumen del ANOVA de la variable longitud del recién nacido por grupos de edad materna.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
EdadMat_Grupos	2	1,562	0,7812	6,655	0,0016
Residuals	194	28,23	1,0455		

Tabla 3. Análisis *post hoc* de las diferencias de medias de la variable longitud del recién nacido. Valores de significación de las diferencias de medias.

	Estimate	Std. Error	t value	Pr(>|t|)
Mayor/igual 35 – 19-34 == 0	0,02295	0,06285	0,365	0,92854
Menor 19 – 19-34 == 0	-0,20009	0,05944	-3,367	0,00268
Menor 19 – Mayor/igual 35 == 0	-0,22304	0,07286	-3,061	0,00692

Tabla 4. Análisis *post hoc* de las diferencias de medias de la variable longitud del recién nacido. Intervalos de confianza de las diferencias de medias.

	Estimate	lwr	upr
Mayor/igual 35 – 19-34 == 0	0,02295	-0,12513	0,17102
Menor 19 – 19-34 == 0	-0,20009	-0,34012	-0,06006
Menor 19 – Mayor/igual 35 == 0	-0,22304	-0,39470	-0,05138

cociente de las dos varianzas) y el valor de p del contraste. En este caso, p = 0,0016, por lo que podemos rechazar la hipótesis nula de igualdad de medias y concluir que al menos una de las medias es estadísticamente diferente a las demás.

Nos queda revisar el análisis *post* hoc comparando las medias dos a dos que nos ofrece el programa a continuación.

En primer lugar, nos informa del método utilizado para realizar el ajuste de la significación para comparaciones múltiples que, en este caso, es el de Tukey. A continuación, nos muestra la tabla con las diferencias de medias, comparadas dos a dos (tabla 3), realizando una prueba de la t de Student bajo la hipótesis nula de igualdad de medias (diferencia igual a cero). Como podemos ver, al comparar los dos grupos mayores de 19 años el valor de p = 0,92, por lo que asumimos que no hay diferencias en los pesos de los recién nacidos de madres de estos dos grupos.

Sin embargo, al comparar los peso de las madres de menos de 19 años con los otros dos grupos se obtienen valores de p estadísticamente significativos: 0,002 y 0,006 al comparar con los grupos de madres de 19 a 34 y de 35 o más, respectivamente. Podemos rechazar la hipótesis nula de igualdad de medias y concluir que el peso al nacimiento de los hijos de madres de menos de 19 años es inferior al de los hijos de madres con edad superior.

El análisis *post hoc* se completa con la tabla de los valores estimados de las tres diferencias de medias y los límites de sus intervalos de confianza (tabla 4), que R nos proporciona también de modo gráfico (figura 5). Podemos ver, tanto en el análisis numérico como en el gráfico, que únicamente los intervalos de las diferencias de pesos al nacimiento entre los hijos de madres menores de 19 años y los otros dos grupos son estadísticamente significativos, con valores de p <0,05 e intervalos que no cruzan el valor nulo que, para una diferencia de medias, es cero.

COMPARACIÓN DE MÁS DE DOS MEDIAS EN OTROS SUPUESTOS

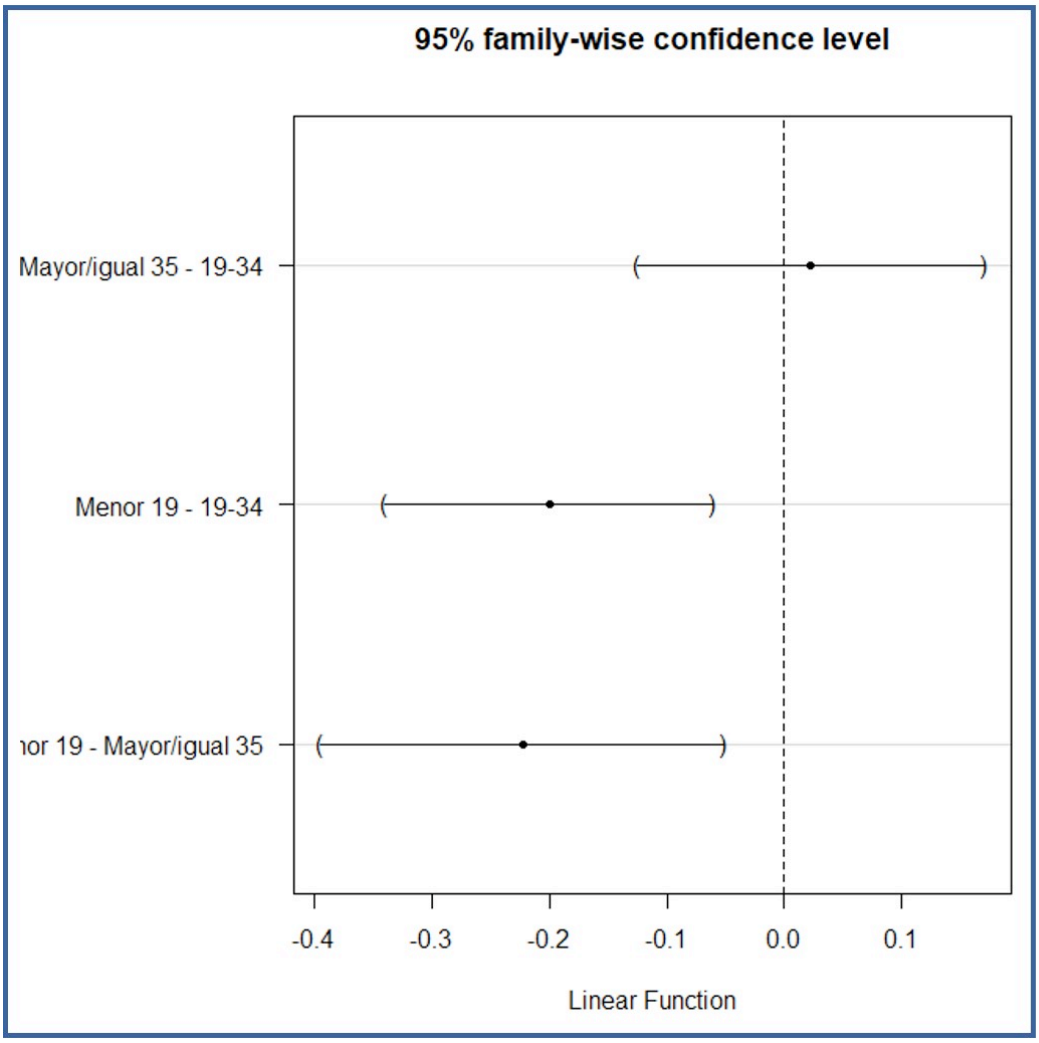
Ya hemos comentado que la alternativa no paramétrica para comparar múltiples medias independientes es la prueba de Kruskal-Wallis.

Cuando se trate de medias de datos apareados, tendremos que realizar la prueba de ANOVA para muestras apareadas. En este caso no es necesario que se cumpla el supuesto de homocedasticidad, pero sí el de esfericidad, que implica que las varianzas entre todos los pares de variables a comparar sean iguales. Los programas estadísticos comprueban este supuesto mediante la prueba de Mauchly.

Las alternativas, cuando no se cumple el supuesto de esfericidad, son dos. La primera, realizar el ANOVA aplicando una corrección como la de Greenhouse-Geisser o la de Huynh-Feldt. La segunda, realizar el equivalente no paramétrico que, en este caso, es la prueba de Friedman.

Para terminar, decir que es posible realizar un ANOVA comparando las medias de una variable cuantitativa según las categorías o factores de más de una variable cualitativa. Este es el método del ANOVA de doble vía o con dos factores.

Figura 5. Representación gráfica de los intervalos de confianza de las diferencias de medias.



BIBLIOGRAFÍA

- Arriaza Gómez AJ, Fernández Palacín F, López Sánchez MA, Muñoz Márquez M, Pérez Plaza S, Sánchez Navas S. Estadística Básica con R y R-Commander. Cádiz: Universidad de Cádiz; 2008.
- Martínez González MA, Martín Calvo N, Toledo JB. Comparaciones de k medias (tres o más grupos). En: Martínez González MA, Sánchez-Villegas A, Toledo Atucha E, Faulin Fajardo J (eds.). Bioestadística amigable. 3.ª edición. Barcelona: Elsevier; 2014. pp. 213-39.
- Molina Arias M, Ochoa Sangrador C, Ortega Páez E. Comparación de dos medias. Pruebas de la t de Student. Evid Pediatr. 2020;16:51.
- Ochoa Sangrador C, Molina Arias M, Ortega Páez E. Inferencia estadística: contraste de hipótesis. Evid Pediatr. 2020;16:11.