ANOVA análisis de varianza para comparar múltiples medias

Joaquín Amat Rodrigo *j.amatrodrigo@gmail.com*

Enero, 2016

Índice

Idea intuitiva del ANOVA 3
ANOVA de una vía para datos independientes 5
Introducción 5
Condiciones para ANOVA de una vía con datos independientes 8
Comparación múltiple de medias. Contrastes POST-HOC 9
Intervalos *LSD* de Fisher (Least Significance Method) 9
Bonferroni adjustment 10
Holm–Bonferroni Adjustment 11
Tukey-Kramer Honest Significant Difference (HSD) 11
Dunnett's correction (Dunnett's test) 12
Inconvenientes de controlar el false positive rate mediante Bonferroni 12
False Discovery Rate (FDR), Benjamini & Hochberg (BH) 13
Tamaño del efecto $\boldsymbol{\eta 2}$ 15
Potencia (*power*) ANOVA de una vía 16
Resultados de un ANOVA 16
Ejemplo 16
ANOVA de dos vías para datos independientes 27
Introducción 27
Condiciones ANOVA de dos vías para datos independientes 28
Tamaño del efecto 28
Ejemplo 1 28
Ejemplo 2 35
ANOVA con variables dependientes (ANOVA de medidas repetidas) 42

Introducción....42
Condiciones para ANOVA de variables dependientes....42
Ejemplo....43
Bibliografía....49

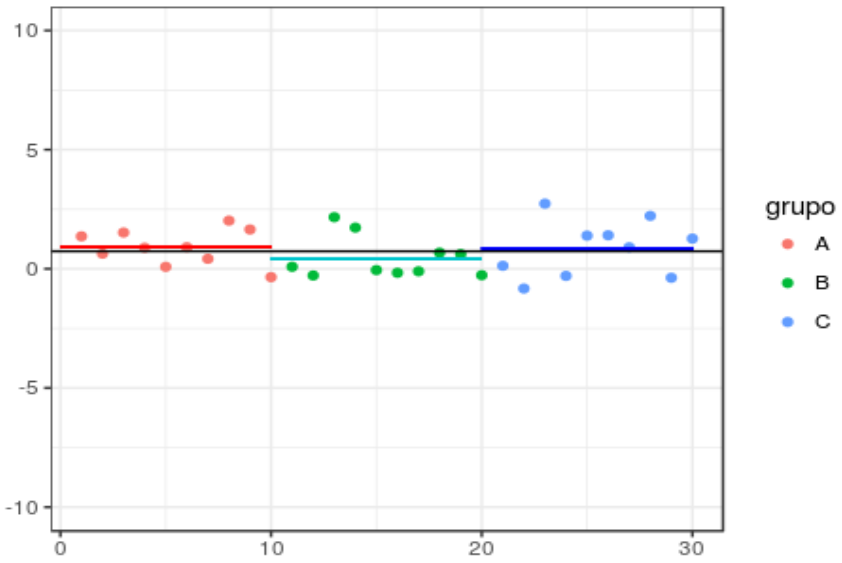
Idea intuitiva del ANOVA

La técnica de análisis de varianza (ANOVA) también conocida como análisis factorial y desarrollada por Fisher en 1930, constituye la herramienta básica para el estudio del efecto de uno o más factores (cada uno con dos o más niveles) sobre la media de una variable continua. Es por lo tanto el test estadístico a emplear cuando se desea comparar las medias de dos o más grupos. Esta técnica puede generalizarse también para estudiar los posibles efectos de los factores sobre la varianza de una variable.

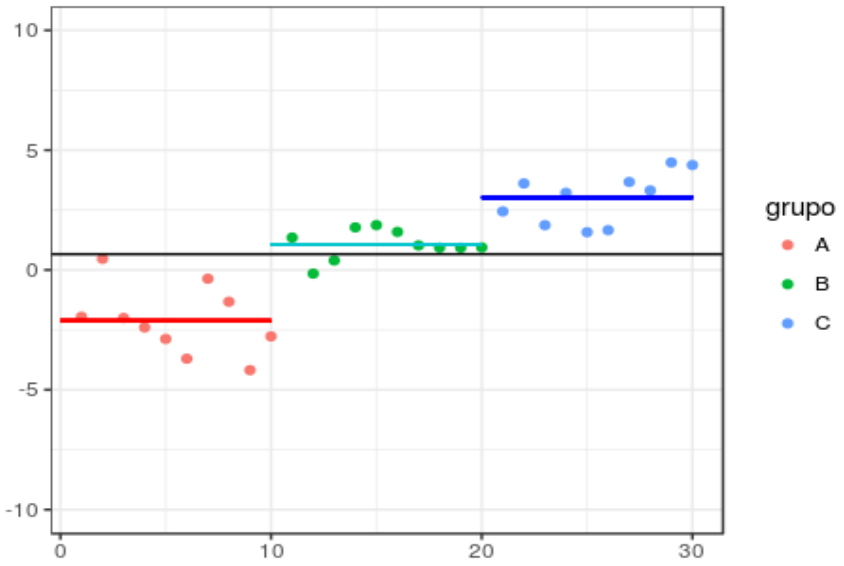
La hipótesis nula de la que parten los diferentes tipos de ANOVA es que la media de la variable estudiada es la misma en los diferentes grupos, en contraposición a la hipótesis alternativa de que al menos dos medias difieren de forma significativa. ANOVA permite comparar múltiples medias, pero lo hace mediante el estudio de las varianzas.

El funcionamiento básico de un ANOVA consiste en calcular la media de cada uno de los grupos para a continuación comparar la varianza de estas medias (varianza explicada por la variable grupo, *intervarianza*) frente a la varianza promedio dentro de los grupos (la no explicada por la variable grupo, *intravarianza*). Bajo la hipótesis nula de que las observaciones de los distintos grupos proceden todas la misma población (tienen la misma media y varianza), la varianza ponderada entre grupos será la misma que la varianza promedio dentro de los grupos. Conforme las medias de los grupos estén más alejadas las unas de las otras, la varianza entre medias se incrementará y dejará de ser igual a la varianza promedio dentro de los grupos.

Tres muestras de una población con media 1 y desviación estándar 1



Tres muestras de tres poblaciones distintas con medias -2, 1, 3 y desviación estándar 1



La línea negra es la media para todas las observaciones.

El estadístico estudiado en el ANOVA, conocido como F_{ratio}, es el ratio entre la varianza de las medias de los grupos y el promedio de la varianza dentro de los grupos. Este estadístico sigue una distribución conocida como "F de Fisher-Snedecor". Si se cumple la hipótesis nula, el estadístico F adquiere el valor de 1 ya que la *intervarianza* será igual a la *intravarianza*. Cuanto más difieran las medias de los grupos mayor será la varianza entre medias en comparación al promedio de la varianza dentro de los grupos, obteniéndose valores de F superiores a 1 y por lo tanto menor la probabilidad de que la distribución adquiera valores tan extremos (menor el *p-value*).

En concreto, si S_1^2 es la la varianza de una muestra de tamaño N_1 extraida de una población normal de varianza σ_1^2 y S_2^2 es la la varianza de una muestra de tamaño N_2 extraida de una población normal de varianza σ_2^2, y ambas muestras son independientes, el cociente:

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

se distribuye como una variable F de Snedecor con (N_1 y N_2) grados de libertad. En el caso del ANOVA, dado que dos de las condiciones son la normalidad de los grupos y la

homocedasticidad de varianza ($\sigma_1^2 = \sigma_1^2$), el valor F se puede obtener dividiendo las dos varianzas calculadas a partir de las muestras (intervariaza y intravarianza).

Existen diferentes tipos de ANOVA dependiendo de la si se trata de datos independientes (ANOVA entre sujetos), si son pareados (ANOVA de mediciones repetidas), si comparan la variable cuantitativa dependiente contra los niveles de una única variable explicatoria o factor (ANOVA de una vía) o frente a dos factores (ANOVA de dos vías). Este último puede ser a su vez aditivo o de interacción (los factores son independientes o no lo son). Cada uno de estos tipos de ANOVA tiene una serie de requerimientos.

ANOVA de una vía para datos independientes

Introducción

El ANOVA de una vía, ANOVA con un factor o modelo factorial de un solo factor es el tipo de análisis que se emplea cuando los datos no están pareados y se quiere estudiar si existen diferencias significativas entre las medias de una variable aleatoria continua en los diferentes niveles de otra variable cualitativa o factor. Es una extensión de los *t-test independientes* para más de dos grupos.

Las hipótesis contrastadas en un ANOVA de un factor son:

- Ho: No hay diferencias entre las medias de los diferentes grupos : $\mu_1 = \mu_2 \ldots = \mu_k = \mu$
- Ha: Al menos un par de medias son significativamente distintas la una de la otra.

Otra forma de plantear las hipótesis de un ANOVA es la siguiente. Si se considera μ como el valor esperado para una observación cualquiera de la población (la media de todas las observaciones sin tener en cuenta los diferentes niveles), y α_i el efecto introducido por el nivel i. La media de un determinado nivel (μ_i) se puede definir como:

$$\mu_i = \mu + \alpha_i$$

- Ho: Ningún nivel introduce un efecto sobre la media total: $\alpha_1 = \alpha_2 = \ldots \alpha_k = 0$
- Ha: Al menos un nivel introduce un efecto que desplaza su media: Algún $\alpha_i \neq 0$

Como se ha mencionado anteriormente, la diferencia entre medias se detecta a través del estudio de la varianza entre grupos y dentro de grupos. Para lograrlo, el ANOVA requiere de una descomposición de la varianza basada en la siguiente idea:

$$Variabilidad\ total =$$

$$variabilidad\ debida\ a\ los\ diferentes\ niveles\ del\ factor + variabilidad\ residual$$

lo que es equivalente a:

$$Variabilidad\ total =$$

$$variabilidad\ explicada\ por\ el\ factor + variabilidad\ no\ explicada\ por\ el\ factor$$

lo que es equivalente a:

$$Varianza\ total = (varianza\ entre\ niveles) + (varianza\ dentro\ de\ los\ niveles)$$

Para poder calcular las diferentes varianzas en primer lugar se tienen que obtener las Sumas de Cuadrados (SS o Sc):

Suma de Cuadrados Total o Total Sum of Squares (TSS): mide la variabilidad total de los datos, se define como la suma de los cuadrados de las diferencias de cada observación respecto a la media general de todas las observaciones. Los grados de libertad de la suma de cuadrados totales es igual al número total de observaciones menos uno (N-1).

Suma de cuadrados del factor o Sum of Squares due to Treatment (SST): mide la variabilidad en los datos asociada al efecto del factor sobre la media (la diferencia de las medias entre los diferentes niveles o grupos). Se obtiene como la suma de los cuadrados de las desviaciones de la media de cada proveedor respecto de la media general, ponderando cada diferencia al cuadrado por el número de observaciones de cada grupo. Los grados de libertad correspondientes son igual al número niveles del factor menos uno (k-1).

Suma de cuadrados residual/error o Sum of Squares of Errors (SSE): mide la variabilidad dentro de cada nivel, es decir, la variabilidad que no es debida a variable cualitativa o factor. Se calcula como la suma de los cuadrados de las desviaciones de cada observación respecto a la media del nivel al que pertenece. Los grados de libertad asignados a la suma de cuadrados residual equivale la diferencia entre los grados de libertad totales y los

grados de libertad del factor, o lo que es lo mismo (N-k). En estadística se emplea el termino error o residual ya que se considera que esta es la variabilidad que muestran los datos debido a los errores de medida. Desde el punto de vista biológico tiene más sentido llamarlo *Suma de cuadrados dentro de grupos* ya que se sabe que la variabilidad observada no solo se debe a errores de medida, si no a los muchos factores que no se controlan y que afectan a los procesos biológicos.

$$TSS = SSE + SST$$

Una vez descompuesta la suma de cuadrados se puede obtener la descomposición de la varianza dividiendo la Suma de Cuadrados entre los respectivos grados de libertad. De forma estricta, al cociente entre la Suma de Cuadrados y sus correspondientes grados de libertad se le denomina *Cuadrados Medios o Mean Sum of Squares* y pueden ser empleado como estimador de la varianza:

ANOVA se define como análisis de varianza, pero en un sentido estricto, se trata de un análisis de la Suma de Cuadrados Medios.

$\hat{S}_T^2 = \frac{TSS}{N-1}$ = Cuadrados Medios Totales = Cuasivarianza Total (varianza muestral total)

$\hat{S}_t^2 = \frac{SST}{k-1}$ = Cuadrados Medios del Factor = Intervarianza (varianza entre las medias de los distintos niveles)

$\hat{S}_E^2 = \frac{SSE}{N-k}$ = Cuadrados Medios del Error = Intravarianza (varianza dentro de los niveles, conocida como varianza residual o de error)

Una vez descompuesta la estimación de la varianza, se obtiene el estadístico F_{ratio} dividiendo la intervarianza entre la intravarianza:

$$F_{ratio} = \frac{Cuadrados\ Medios\ del\ Factor}{Cuadrados\ Medios\ del\ Error} = \frac{\hat{S}_t^2}{\hat{S}_E^2} = \frac{intervarianza}{intravarianza} \sim F_{k-1,N-k}$$

Dado que por definición el estadístico F_{ratio} sigue una distribución *F Fisher-Snedecor* con $k-1$ y $N-t$ grados de libertad, se puede conocer la probabilidad de obtener valores iguales o más extremos que los observados.

Condiciones para ANOVA de una vía con datos independientes

Independencia: Las observaciones deben ser aleatorias. El tamaño total de la muestra de cada grupo debe de ser < 10% de la población a la que representa. Los grupos (niveles del factor) deben de ser independientes entre ellos.

Distribución normal de cada uno de los niveles o grupos: La variable cuantitativa debe de distribuirse de forma normal en cada uno de los grupos, siendo menos estricta esta condición cuanto mayor sea el tamaño de cada grupo. La mejor forma de verificar la normalidad es estudiar los residuos de cada observación respecto a la media del grupo al que pertenecen.

- A pesar de que el ANOVA es bastante robusto aun cuando existe cierta falta de normalidad, si la simetría es muy pronunciada y el tamaño de cada grupo no es muy grande, se puede recurrir en su lugar al test no paramétrico *prueba H de Kruskal-Wallis*. En algunos libros recomiendan mantenerse con ANOVA a no ser que la falta de normalidad sea muy extrema.
- Los datos atípicamente extremos pueden invalidar por completo las conclusiones de un ANOVA. Si se observan residuos extremos hay que estudiar con detalle a que observaciones pertenecen, siendo aconsejable recalcular el ANOVA sin ellas y comparar los resultados obtenidos.

Varianza constante entre grupos (homocedasticidad): La varianza dentro de los grupos debe de ser aproximadamente igual en todos ellos. Esto es así ya que en la hipótesis nula se considera que todas las observaciones proceden de la misma población, por lo que tienen la misma media y también la misma varianza.

- Esta condición es más importante cuanto menor es el tamaño de los grupos.
- El ANOVA es bastante robusto a la falta de homodedasticidad si el diseño es equilibrado (mismo número de observaciones por grupo).
- En diseños no equilibrados, la falta de homocedasticidad tiene mayor impacto. Si los grupos de menor tamaño son los que presentan mayor desviación estándar, aumentará el número de falsos positivos. Si por el contrario los grupos de mayor tamaño tienen mayor desviación estándar aumentarán los falsos negativos.
- Si no se puede aceptar la homocedasticidad, se recurre a lo que se conoce como ANOVA heterodástica que emplea la corrección de Welch (Welch test), en R su función es `oneway.test()`.

En el libro *Handbook of Biological Statistics* se considera altamente recomendable emplear diseños equilibrados. Siendo así, consideran fiable el ANOVA siempre y cuando el número de observaciones por grupo no sea menor de 10 y la desviación estándar no varíe más de 3 veces entre grupos. Para modelos no equilibrados recomiendan examinar con detalle la homocedasticidad, si las varianzas de los grupos no son muy semejantes es mejor emplear *Welch's ANOVA*.

Comparación múltiple de medias. Contrastes POST-HOC

Si un Análisis de Varianza resulta significativo, implica que al menos dos de las medias comparadas son significativamente distintas entre sí, pero no se indica cuáles. Para identificarlas hay que comparar dos a dos las medias de todos los grupos introducidos en el análisis mediante un *t-test* u otro test que compare 2 grupos, ha esto se le conoce como análisis *post-hoc*. Debido a la inflación del error de tipo I, cuantas más comparaciones se hagan más aumenta la probabilidad de encontrar diferencias significativas (para α = 0.05, de cada 100 comparaciones se esperan 5 significativas solo por azar). Los niveles de significancia pueden ser ajustados en función del número de comparaciones (corrección de significancia). Si no se hace ningún tipo de corrección se aumenta la posibilidad de falsos positivos (error tipo I) pero si se es muy estricto con las correcciones se pueden considerar como no significativas diferencias que realmente podrían serlo (error tipo II). La necesidad de corrección o no, y de qué tipo, se ha de estudiar con detenimiento en cada caso. Los principales métodos de comparación *post-hoc* (algunas con corrección y otros no) son:

Intervalos *LSD* de Fisher (Least Significance Method)

Siendo $\overline{x}_i$ la media muestral de un grupo, la desviación típica estimada de dicha media (asumiendo la homocedasticidad de los grupos) es igual a la raíz cuadrada de los Cuadrados Medios del Error (que como se ha visto es la estimación de la intravarianza o varianza del error) dividida por el número de observaciones de dicho grupo. Asumiendo también la normalidad de los grupos, se puede obtener el intervalo LSD como:

$$\overline{x}_i \pm \frac{\sqrt{2}}{2} t^{\alpha}_{gl(error)} \sqrt{\frac{SSE}{n}}$$

Los intervalos *LSD* son básicamente un conjunto de *t-test* individuales con la única diferencia de que en lugar de calcular una *pooled SD* empleando solo los dos grupos comparados, calcula la *pooled SD* a partir de todos los grupos.

Cuanto más se alejen los intervalos de dos grupos más diferentes son sus medias, siendo significativa dicha diferencia si los intervalos no se solapan. Es importante comprender que los intervalos *LSD* se emplean para comparar las medias pero no se pueden interpretar como el intervalo de confianza para cada una de las medias. El método *LSD* no conlleva ningún tipo de corrección de significancia, es por esto que su uso parece estar desaconsejado para determinar significancia aunque sí para identificar que grupos tienen las medias más distantes.

En R se pueden obtener los intervalos *LSD* y su representación gráfica mediante la función `LSD.test {agricolae}`.

Bonferroni adjustment

Este es posiblemente el ajuste de significancia más extendido a pesar de que no está recomendado para la mayoría de las situaciones que se dan en el ámbito de la biomedicina. La corrección de *Bonferroni* consiste en dividir el nivel de significancia α entre el número de comparaciones dos a dos realizadas.

$$\alpha_{corregido} = \frac{\alpha}{numero\ de\ grupos}$$

Con esta corrección se asegura que la probabilidad de obtener al menos un falso positivo entre todas las comparaciones (*family-wise error rate*) es $\leq \alpha$. Permite por lo tanto contrastar una hipótesis nula general (la de que toda las hipótesis nulas testadas son verdaderas) de forma simultanea, cosa que raramente es de interés en las investigaciones. Se considera un método excesivamente conservativo sobre todo a medida que se incrementa el número de comparaciones. Se desaconseja su utilización excepto en situaciones muy concretas. *False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies, Mark E. Glickman.*

En R se puede calcular mediante la función `pairwise.t.test()` indicando en los argumentos *p.adj = "bonferroni"*.

Holm–Bonferroni Adjustment

Con este método, el valor de significancia α se corrige secuencialmente haciéndolo menos conservativo que el de *Bonferroni*. Aun así, parece que tampoco es indicado si se realizan más de 6 comparaciones.

El proceso consiste en realizar un *t-test* para todas las comparaciones y ordenarlas de menor a mayor *p-value*. El nivel de significancia para la primera comparación (la que tiene menor *p-value*) se corrige dividiendo α entre el número total de comparaciones, si no resulta significativo se detiene el proceso, si sí que lo es, se corrige el nivel de significancia de la siguiente comparación (segundo menor *p-value*) dividiendo entre el número de comparaciones menos uno. El proceso se repite hasta detenerse cuando la comparación ya no sea significativa.

Tukey-Kramer Honest Significant Difference (HSD)

Es el ajuste recomendado cuando el número de grupos a comparar es mayor de 6 y el diseño es equilibrado (mismo número de observaciones por grupo). En el caso de modelos no equilibrados el método *HSD* es conservativo, requiere diferencias grandes para que resulte significativo. Solo aplicable si se trata de datos no pareados.

El *Tukey's test* es muy similar a un *t-test*, excepto que corrige el *experiment wise error rate*. Esto lo consigue empleando un estadístico que sigue una distribución llamada *studentized range distribution* en lugar de una *t-distribution*. El estadístico empleado se define como:

$$q_{calculado} = \frac{\overline{x}_{max} - \overline{x}_{min}}{S\sqrt{2/n}}$$

Donde: $\overline{x}_{max}$ es la mayor de las medias de los dos grupos comparados, $\overline{x}_{min}$ la menor, S la *pooled SD* de estos dos grupos y n el número total de observaciones en los dos grupos.

Para cada par de grupos, se obtiene el valor $q_{calculado}$ y se compara con el esperado acorde a una *studentized range distribution* con los correspondientes grados de libertad. Si la probabilidad es menor al nivel de significancia α establecido, se considera significativa la diferencia de medias. Al igual que con los intervalos *LSD*, es posible calcular intervalos *HSD* para estudiar su solapamiento.

En R, las funciones `TukeyHSD()` y `plot(TukeyHSD)` permiten calcular los *p-value* corregidos por Tukey y representar los intervalos.

Dunnett's correction (Dunnett's test)

Es el equivalente al test Tukey-Kramer (HSD) recomendado cuando en lugar de comparar todos los grupos entre sí ($\frac{(k-1)k}{2}$ comparaciones) solo se quieren comparar frente a un grupo control ($k-1$ comparaciones). Se emplea con frecuencia en experimentos médicos.

Inconvenientes de controlar el false positive rate mediante Bonferroni

Tal como se ha visto en los ejemplos, al realizar múltiples comparaciones es importante controlar la inflación del error de tipo I. Sin embargo, correcciones como las de *Bonferroni* o similares conllevan una serie de problemas. La primera es que el método se desarrolló para contrastar la hipótesis nula universal de que los dos grupos son iguales para todas las variables testadas, no para aplicarlo de forma individual a cada test. A modo de ejemplo, supóngase que un investigador quiere determinar si un nuevo método de enseñanza es efectivo empleando para ello estudiantes de 20 colegios. En cada colegio se selecciona de forma aleatoria un grupo control, un grupo que se somete al nuevo método y se realiza un test estadístico entre ambos considerando un nivel de significancia $\alpha = 0.05$. La corrección de *Bonferroni* implica comparar el *p-value* obtenido en los 20 test frente a $0.05/20 = 0.0025$. Si alguno de los *p-values* es significativo, la conclusión de *Bonferroni* es que la hipótesis nula de que el nuevo sistema de enseñanza no es efectivo en todos los grupos (colegios) queda rechazada, por lo que se puede afirmar que el método de enseñanza es efectivo para alguno de los 20 grupos, pero no cuáles ni cuántos. Este tipo de información no es de interés en la gran mayoría de estudios, ya que lo que se desea conocer es qué grupos difieren.

El segundo problema de la corrección de *Bonferroni* es que una misma comparación será interpretada de forma distinta dependiendo del número de test que se hagan. Supóngase que un investigador realiza 20 contrastes de hipótesis y que todos ellos resultan en un *p-value* de 0.001. Aplicando la corrección de *Bonferroni* si el límite de significancia para un test individual es de $\alpha = 0.05$, el nivel de significancia corregido resulta ser $0.05/20 = 0.0025$, por lo que el investigador concluye que todos los test son significativos. Un segundo investigador reproduce los mismos análisis en otro laboratorio y llega a los mismos resultados, pero para confirmarlos todavía más, realiza 80 test estadísticos adicionales con lo que su nivel de

significancia corregido pasa a ser de 0.05/100 = 0.0005. Ahora, ninguno de los test se puede considerar significativo, por lo que debido a aumentar el número de contrastes las conclusiones son totalmente contrarias.

Viendo los problemas que implica ¿Para qué sirve entonces la corrección de Bonferroni?

Que su aplicación en las disciplinas biomédicas no sea adecuada no quita que pueda serlo en otras áreas. Imagínese una factoría que genera bombillas en lotes de 1000 unidades y que testar cada una de ellas antes de repartirlas no es práctico. Una alternativa consiste en comprobar únicamente una muestra de cada lote, rechazando cualquier lote que tenga más de x bombillas defectuosas en la muestra. Por supuesto, la decisión puede ser errónea para un determinado lote, pero según la teoría de Neyman-Pearson, se puede encontrar el valor x para el que se minimiza el ratio de error. Ahora bien, la probabilidad de encontrar x bombillas defectuosas en la muestra depende del tamaño que tenga la muestra, o en otras palabras, del número de test que se hagan por lote. Si se incrementa el tamaño también lo hace la probabilidad de rechazar el lote, es aquí donde la corrección de *Bonferroni* recalcula el valor de x que mantiene minimizado el ratio de errores.

False Discovery Rate (FDR), Benjamini & Hochberg (BH)

Los métodos descritos anteriormente se centran en corregir la inflación del el error de tipo I (*false positive rate*), es decir, la probabilidad de rechazar la hipótesis nula siendo esta cierta. Esta aproximación es útil cuando se emplea un número limitado de comparaciones. Para escenarios de *large-scale multiple testing* como los estudios genómicos en los que se realizan miles de test de forma simultánea, el resultado de estos métodos es demasiado conservativo e impide que se detecten diferencias reales. Una alternativa es controlar el *false discovery rate*.

El *false discovery rate (FDR)* se define como: (*todas las definiciones son equivalentes*)

- La proporción esperada de test en los que la hipótesis nula es cierta, de entre todos los test que se han considerado significativos.
- *FDR* es la probabilidad de que una hipótesis nula sea cierta habiendo sido rechazada por el test estadístico.
- De entre todos los test considerados significativos, el *FDR* es la proporción esperada de esos test para los que la hipótesis nula es verdadera.
- Es la proporción de test significativos que realmente no lo son.
- La proporción esperada de falsos positivos de entre todos los test considerados como significativos.

El objetivo de controlar el *false discovery rate* es establecer un límite de significancia para un conjunto de test tal que, de entre todos los test considerados como significativos, la proporción de hipótesis nulas verdaderas (falsos positivos) no supere un determinado valor. Otra ventaja añadida es su fácil interpretación, por ejemplo, si un estudio publica resultados estadísticamente significativos para un *FDR* del 10%, el lector tiene la seguridad de que como máximo un 10% de los resultados considerados como significativos son realmente falsos positivos.

Cuando un investigador emplea un nivel de significancia α, por ejemplo de 0.05, suele esperar cierta seguridad de que solo una pequeña fracción de los test significativos se correspondan con hipótesis nulas verdaderas (falsos positivos). Sin embargo, esto no tiene por qué ser así. La razón por la que un *false positive rate* bajo no tiene por qué traducirse en una probabilidad baja de hipótesis nulas verdaderas entre los test significativos (*false discovery rate*) se debe a que esta última depende de la frecuencia con la que la hipótesis nula contrastada es realmente verdadera. Un caso extremo sería el planteado en el ejemplo 1, en el que todas las hipótesis nulas son realmente ciertas por lo que el 100% de los test que en promedio resultan significativos son falsos positivos. Por lo tanto, la proporción de falsos positivos (*false discovery rate*) depende de la cantidad de hipótesis nulas que sean ciertas de entre todas los contrastes.

Los análisis de tipo exploratorio en los que el investigador trata de identificar resultados significativos sin apenas conocimiento previo se caracterizan por una proporción alta de hipótesis nulas falsas. Los análisis que se hacen para confirmar hipótesis, en los que el diseño se ha orientado en base a un conocimiento previo, suelen tener una proporción de hipótesis nulas verdaderas alta. Idealmente, si se conociera de antemano la proporción de hipótesis nulas verdaderas de entre todos los contrastes se podría ajustar con precisión el límite significancia adecuado a cada escenario, sin embargo, esto no ocurre en la realidad.

La primera aproximación para controlar el *FDR* fue descrita por Benjamini y Hochberg en 1995. Acorde a su publicación, si se desea controlar que en un estudio con n comparaciones el *FDR* no supere un porcentaje d hay que:

- Ordenar los n test de menor a mayor *p-value* ($p_1,\ p_2,\ \dots\ p_n$)
- Se define k como la última posición para la que se cumple que $p_i \leq d\frac{i}{n}$
- Se consideran significativos todos los *p-value* hasta la posición k ($p_1,\ p_2,\ \dots\ p_k$)

La principal ventaja de controlar el *false discovery rate* se hace más patente cuantas más comparaciones se realicen, por esta razón se suele emplear en situaciones con cientos o miles de comparaciones. Sin embargo, el método puede aplicarse también a estudios de menor envergadura.

El método propuesto por Benjamini & Hochberg asume a la hora de estimar el número de hipótesis nulas erróneamente consideradas falsas que todas las hipótesis nulas son ciertas. Como consecuencia, a estimación del *FDR* está inflada y por lo tanto es conservativa. A continuación se describen métodos más sofisticados que estiman la frecuencia de hipótesis nulas verdaderas a partir de la distribución de los *p-values*.

Para información más detallada sobre comparaciones múltiples consultar el capítulo: Comparaciones múltiples: corrección de p-value y FDR .

Tamaño del efecto η^2

El tamaño del efecto de un ANOVA, 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.

En el ANOVA la medida del tamaño del efecto más empleada es η^2 que se define como:

$$\eta^2 = \frac{SumaCuadrados_{entre\ grupos}}{SumaCuadrados_{total}}$$

Los niveles de clasificación más empleados para el tamaño del efecto son:

- 0.01 = pequeño
- 0.06 = mediano
- 0.14 = grande

Los valores necesarios para calcular η^2 se obtienen del summary del ANOVA. En R puede obtenerse mediante la función `etaSquared()` de paquete *lsr*.

Potencia (*power*) ANOVA de una vía

Los test de potencia permiten determinar la probabilidad de encontrar diferencias significativas entre las medias para un determinado α indicando el tamaño de los grupos, o bien calcular el tamaño que deben de tener los grupos para ser capaces de detectar con una determinada probabilidad una diferencia en las medias si esta existe. En aquellos casos que se quiere conocer el tamaño que han de tener las muestras, es necesario conocer (bien por experimentos previos o por muestras piloto) una estimación de la varianza de la población.

La función `power.anova.test()` del paquete *stats* realiza el cálculo de potencia para modelos de ANOVA equilibrados.

Resultados de un ANOVA

A la hora de comunicar los resultados de un ANOVA hay que indicar el valor obtenido para el estadístico F, los grados de libertad, el *p-value* y el tamaño del efecto (η^2).

Puede ocurrir que un análisis ANOVA indique que hay diferencias significativas entre las medias y luego que los t-test no encuentren ninguna comparación que lo sea. Esto no es contradictorio, simplemente es debido a que se trata de dos técnicas distintas.

Ejemplo

Supóngase que se un estudio quiere comprobar si existe una diferencia significativa entre el % de bateos exitosos de los jugadores de béisbol dependiendo de la posición en la que juegan. En caso de que exista diferencia se quiere saber qué posiciones difieren del resto. La siguiente tabla contiene una muestra de jugadores seleccionados aleatoriamente.

```
posicion <- c("OF", "IF", "IF", "OF", "IF", "IF", "OF", "OF", "IF", "IF", "OF",
    "OF", "IF", "OF", "IF", "IF", "IF", "OF", "IF", "OF", "IF", "OF", "IF",
    "OF", "IF", "DH", "IF", "IF", "IF", "OF", "IF", "IF", "IF", "IF", "OF",
    "IF", "OF", "IF", "IF", "IF", "IF", "OF", "OF", "IF", "OF", "OF", "IF",
    "IF", "OF", "OF", "IF", "OF", "OF", "OF", "IF", "DH", "OF", "OF", "OF",
    "IF", "IF", "IF", "IF", "OF", "IF", "IF", "OF", "IF", "IF", "IF", "OF",
    "IF", "IF", "OF", "IF", "IF", "IF", "IF", "IF", "IF", "OF", "DH", "OF",
    "OF", "IF", "IF", "IF", "OF", "IF", "OF", "IF", "IF", "IF", "IF", "OF",
```

```
    "OF", "OF", "DH", "OF", "IF", "IF", "OF", "OF", "C", "IF", "OF", "OF", "IF",
    "OF", "IF", "IF", "IF", "OF", "C", "OF", "IF", "C", "OF", "IF", "DH", "C",
    "OF", "OF", "IF", "C", "IF", "IF", "IF", "IF", "IF", "IF", "OF", "C", "IF",
    "OF", "OF", "IF", "OF", "IF", "OF", "DH", "C", "IF", "OF", "IF", "IF", "OF",
    "IF", "OF", "IF", "C", "IF", "IF", "OF", "IF", "IF", "IF", "OF", "OF", "OF",
    "IF", "IF", "C", "IF", "C", "C", "OF", "OF", "OF", "IF", "OF", "IF", "C",
    "DH", "DH", "C", "OF", "IF", "OF", "IF", "IF", "IF", "C", "IF", "OF", "DH",
    "IF", "IF", "IF", "OF", "OF", "C", "OF", "OF", "IF", "IF", "OF", "OF", "OF",
    "OF", "OF", "OF", "IF", "IF", "DH", "OF", "IF", "IF", "OF", "IF", "IF",
    "IF", "IF", "OF", "IF", "C", "IF", "IF", "C", "IF", "OF", "IF", "DH", "C",
    "OF", "C", "IF", "IF", "OF", "C", "IF", "IF", "IF", "C", "C", "C", "OF",
    "OF", "IF", "IF", "IF", "IF", "OF", "OF", "C", "IF", "IF", "OF", "C", "OF",
    "OF", "OF", "OF", "OF", "OF", "OF", "OF", "OF", "OF", "OF", "C", "IF", "DH",
    "IF", "C", "DH", "C", "IF", "C", "OF", "C", "C", "IF", "OF", "IF", "IF",
    "IF", "IF", "IF", "IF", "IF", "IF", "OF", "OF", "OF", "IF", "OF", "OF",
    "IF", "IF", "IF", "OF", "C", "IF", "IF", "IF", "IF", "OF", "OF", "IF", "OF",
    "IF", "OF", "OF", "OF", "IF", "OF", "OF", "IF", "OF", "IF", "C", "IF", "IF",
    "C", "DH", "OF", "IF", "C", "C", "IF", "C", "IF", "OF", "C", "C", "OF")
bateo <- c(0.359, 0.34, 0.33, 0.341, 0.366, 0.333, 0.37, 0.331, 0.381, 0.332,
    0.365, 0.345, 0.313, 0.325, 0.327, 0.337, 0.336, 0.291, 0.34, 0.31, 0.365,
    0.356, 0.35, 0.39, 0.388, 0.345, 0.27, 0.306, 0.393, 0.331, 0.365, 0.369,
    0.342, 0.329, 0.376, 0.414, 0.327, 0.354, 0.321, 0.37, 0.313, 0.341, 0.325,
    0.312, 0.346, 0.34, 0.401, 0.372, 0.352, 0.354, 0.341, 0.365, 0.333, 0.378,
    0.385, 0.287, 0.303, 0.334, 0.359, 0.352, 0.321, 0.323, 0.302, 0.349, 0.32,
    0.356, 0.34, 0.393, 0.288, 0.339, 0.388, 0.283, 0.311, 0.401, 0.353, 0.42,
    0.393, 0.347, 0.424, 0.378, 0.346, 0.355, 0.322, 0.341, 0.306, 0.329, 0.271,
    0.32, 0.308, 0.322, 0.388, 0.351, 0.341, 0.31, 0.393, 0.411, 0.323, 0.37,
    0.364, 0.321, 0.351, 0.329, 0.327, 0.402, 0.32, 0.353, 0.319, 0.319, 0.343,
    0.288, 0.32, 0.338, 0.322, 0.303, 0.356, 0.303, 0.351, 0.325, 0.325, 0.361,
    0.375, 0.341, 0.383, 0.328, 0.3, 0.277, 0.359, 0.358, 0.381, 0.324, 0.293,
    0.324, 0.329, 0.294, 0.32, 0.361, 0.347, 0.317, 0.316, 0.342, 0.368, 0.319,
    0.317, 0.302, 0.321, 0.336, 0.347, 0.279, 0.309, 0.358, 0.318, 0.342, 0.299,
    0.332, 0.349, 0.387, 0.335, 0.358, 0.312, 0.307, 0.28, 0.344, 0.314, 0.24,
    0.331, 0.357, 0.346, 0.351, 0.293, 0.308, 0.374, 0.362, 0.294, 0.314, 0.374,
    0.315, 0.324, 0.382, 0.353, 0.305, 0.338, 0.366, 0.357, 0.326, 0.332, 0.323,
    0.306, 0.31, 0.31, 0.333, 0.34, 0.4, 0.389, 0.308, 0.411, 0.278, 0.326,
    0.335, 0.316, 0.371, 0.314, 0.384, 0.379, 0.32, 0.395, 0.347, 0.307, 0.326,
    0.316, 0.341, 0.308, 0.327, 0.337, 0.36, 0.32, 0.372, 0.306, 0.305, 0.347,
    0.281, 0.281, 0.296, 0.306, 0.343, 0.378, 0.393, 0.337, 0.327, 0.336, 0.32,
    0.381, 0.306, 0.358, 0.311, 0.284, 0.364, 0.315, 0.342, 0.367, 0.307, 0.351,
    0.372, 0.304, 0.296, 0.332, 0.312, 0.437, 0.295, 0.316, 0.298, 0.302, 0.342,
    0.364, 0.304, 0.295, 0.305, 0.359, 0.335, 0.338, 0.341, 0.3, 0.378, 0.412,
    0.273, 0.308, 0.309, 0.263, 0.291, 0.359, 0.352, 0.262, 0.274, 0.334, 0.343,
    0.267, 0.321, 0.3, 0.327, 0.313, 0.316, 0.337, 0.268, 0.342, 0.292, 0.39,
    0.332, 0.315, 0.298, 0.298, 0.331, 0.361, 0.272, 0.287, 0.34, 0.317, 0.327,
    0.354, 0.317, 0.311, 0.174, 0.302, 0.302, 0.291, 0.29, 0.268, 0.352, 0.341,
    0.265, 0.307, 0.36, 0.305, 0.254, 0.279, 0.321, 0.305, 0.35, 0.308, 0.326,
    0.219, 0.23, 0.322, 0.405, 0.321, 0.291, 0.312, 0.357, 0.324)

datos <- data.frame(posicion = posicion, bateo = bateo)
```

1.Estudio de los datos: Número de grupos, observaciones por grupo y distribución de las observaciones.

Se identifica el número de grupos y cantidad de observaciones por grupo para determinar si es un modelo equilibrado. También se calculan la media y desviación típica de caga grupo.

```
table(datos$posicion)
```

```
##
##   C  DH  IF  OF
##  39  14 154 120
```

```
aggregate(bateo ~ posicion, data = datos, FUN = mean)
```

```
##   posicion     bateo
## 1        C 0.3226154
## 2       DH 0.3477857
## 3       IF 0.3315260
## 4       OF 0.3342500
```

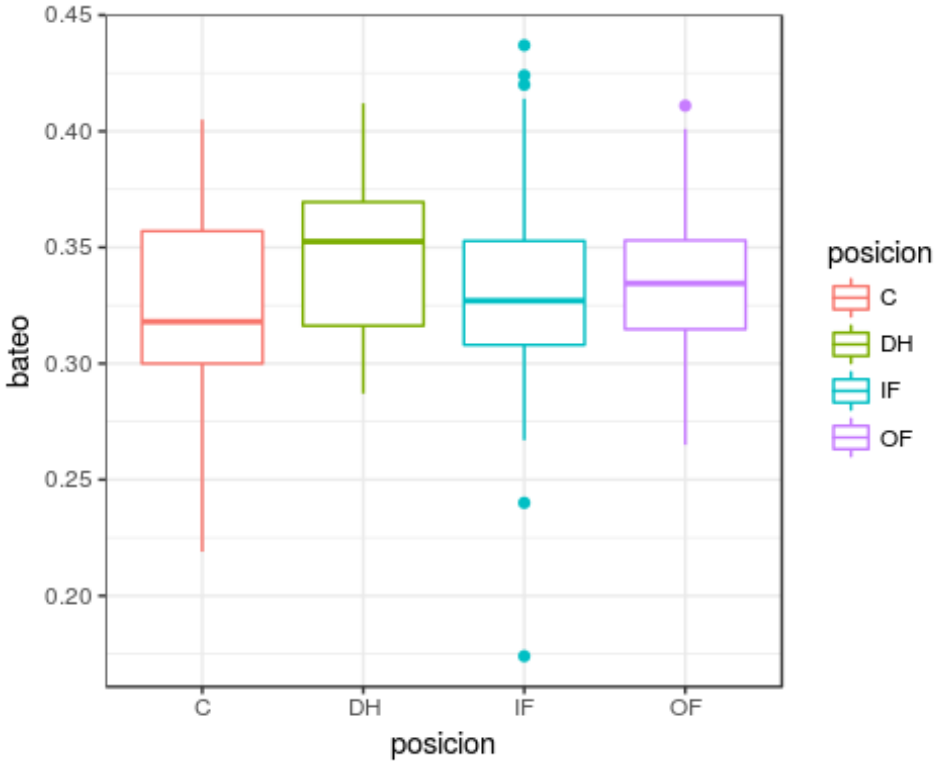
```
aggregate(bateo ~ posicion, data = datos, FUN = sd)
```

```
##   posicion      bateo
## 1        C 0.04513175
## 2       DH 0.03603669
## 3       IF 0.03709504
## 4       OF 0.02944394
```

Dado que el número de observaciones por grupo no es constante, se trata de un modelo no equilibrado. Es importante tenerlo en cuenta cuando se comprueben las condiciones de normalidad y homocedasticidad.

La representación gráfica más útil antes de realizar un ANOVA es el modelo Box-Plot.

```
ggplot(data = datos, aes(x = posicion, y = bateo, color = posicion)) +
    geom_boxplot() +
    theme_bw()
```



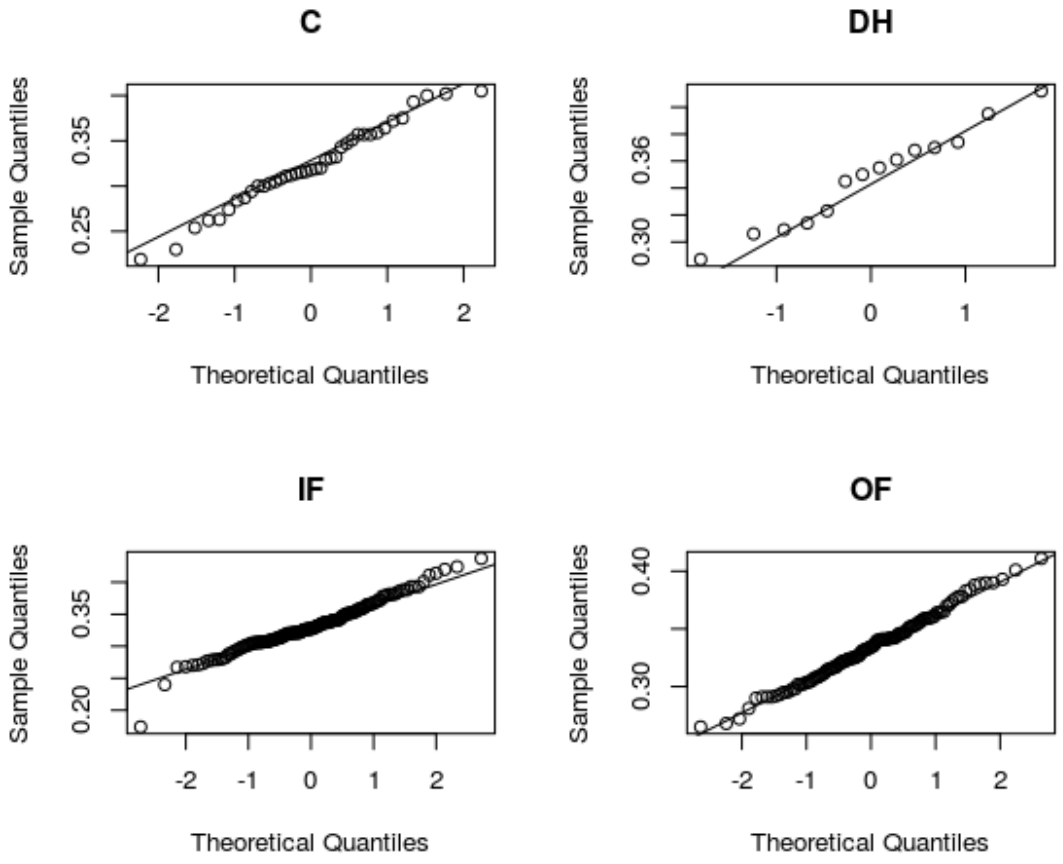
Este tipo de representación permite identificar de forma preliminar si existen asimetrías, datos atípicos o diferencia de varianzas. En este caso, los 4 grupos parecen seguir una distribución simétrica. En el nivel IF se detectan algunos valores extremos que habrá que estudiar con detalle por si fuese necesario eliminarlos. El tamaño de las cajas es similar para todos los niveles por lo que no hay indicios de falta de homocedasticidad.

2.Verificar condiciones para un ANOVA

Independencia:

El tamaño total de la muestra es < 10% de la población de todos los bateadores de la liga. Los grupos (variable categórica) son independientes entre ellos ya que se ha hecho un muestreo aleatorio de jugadores de toda la liga (no solo de un mismo equipo). Distribución normal de las observaciones: La variable cuantitativa debe de distribuirse de forma normal en cada uno de los grupos. El estudio de normalidad puede hacerse de forma gráfica (qqplot) o con test de hipótesis.

```
par(mfrow = c(2,2))
qqnorm(datos[datos$posicion == "C","bateo"], main = "C")
qqline(datos[datos$posicion == "C","bateo"])
qqnorm(datos[datos$posicion == "DH","bateo"], main = "DH")
qqline(datos[datos$posicion == "DH","bateo"])
qqnorm(datos[datos$posicion == "IF","bateo"], main = "IF")
qqline(datos[datos$posicion == "IF","bateo"])
qqnorm(datos[datos$posicion == "OF","bateo"], main = "OF")
qqline(datos[datos$posicion == "OF","bateo"])
```




```
par(mfrow = c(1,1))
```

Dado que los grupos tienen más de 50 eventos se emplea el test de *Kolmogorov-Smirnov* con la corrección de Lilliefors. La función en R se llama `lillie.test()` y se encuentra en el paquete *nortest*. Si fuesen menos de 50 eventos por grupo se emplearía el test Shapiro-Wilk.

```
require(nortest)
by(data = datos,INDICES = datos$posicion,FUN = function(x){ lillie.test(x$bateo)})
```

```
## datos$posicion: C
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x$bateo
## D = 0.087208, p-value = 0.6403
##
## ------------------------------------------------------------
## datos$posicion: DH
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x$bateo
## D = 0.11205, p-value = 0.9046
##
## ------------------------------------------------------------
## datos$posicion: IF
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x$bateo
## D = 0.070653, p-value = 0.05787
##
## ------------------------------------------------------------
## datos$posicion: OF
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x$bateo
## D = 0.044082, p-value = 0.8213
```

Los test de hipótesis no muestran evidencias de falta de normalidad.

Varianza constante entre grupos (homocedasticidad):

Dado que hay un grupo (IF) que se encuentra en el límite para aceptar que se distribuye de forma normal, el test de Fisher y el de Bartlett no son recomendables. En su lugar es mejor emplea un test basado en la mediana *test de Levene* o *test de Fligner-Killeen*.

```r
fligner.test(bateo ~ posicion,datos)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  bateo by posicion
## Fligner-Killeen:med chi-squared = 6.9724, df = 3, p-value =
## 0.07278
```

```r
require(car)
leveneTest(bateo ~ posicion,datos,center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value Pr(>F)
## group   3  2.6057 0.0518 .
##       323
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No hay evidencias significativas de falta de homocedasticidad en ninguno de los dos test.

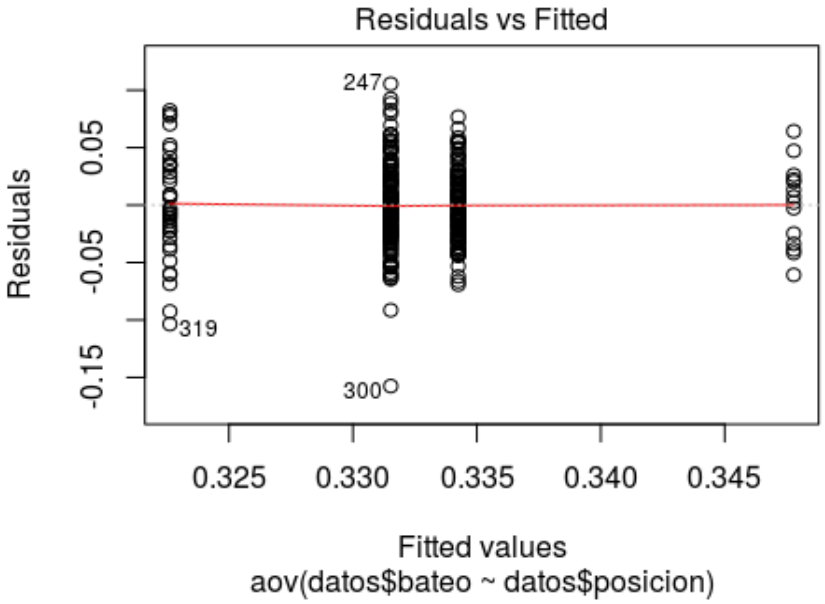
El estudio de las condiciones puede realizarse previo cálculo del ANOVA, puesto que si no se cumplen no tiene mucho sentido seguir adelante. Sin embargo la forma más adecuada de comprobar que se satisfacen las condiciones necesarias es estudiando los residuos del modelo una vez generado el ANOVA. R permite graficar los residuos de forma directa con la función *plot(objeto anova)*.

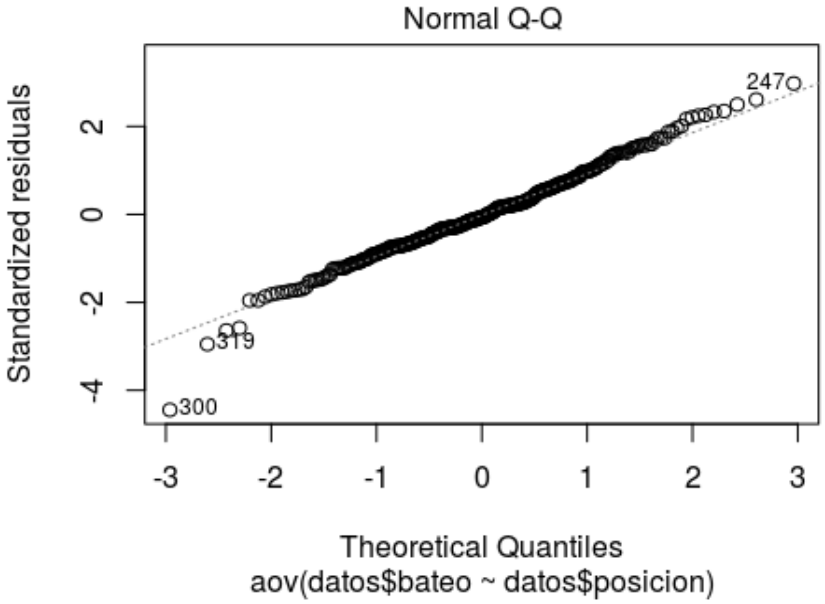
3.Análisis de varianza ANOVA

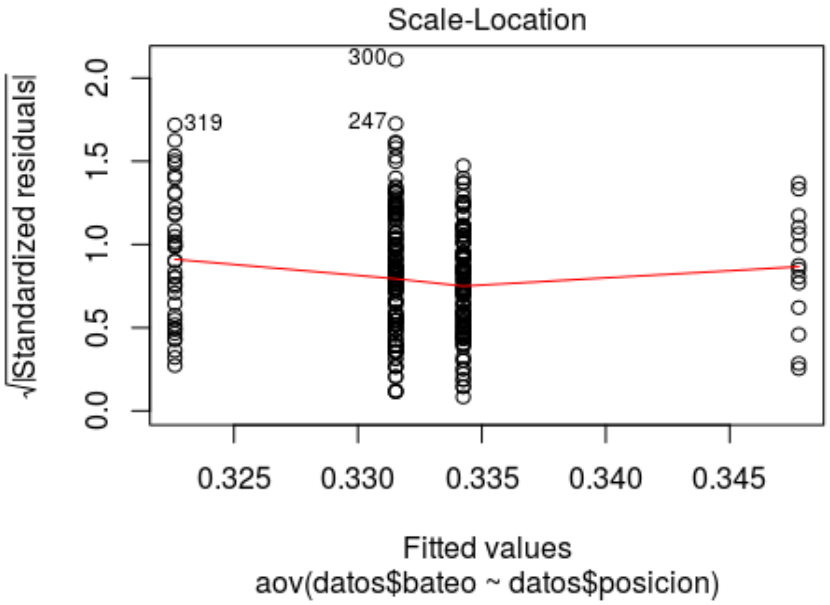
```r
anova <- aov(datos$bateo ~ datos$posicion)
summary(anova)
```

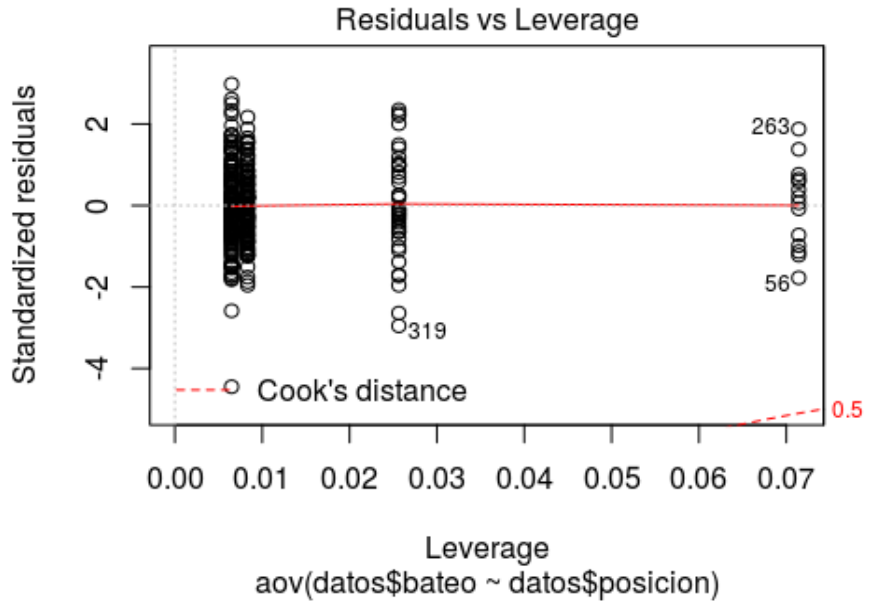
```
##                Df Sum Sq  Mean Sq F value Pr(>F)
## datos$posicion   3 0.0076 0.002519   1.994  0.115
## Residuals      323 0.4080 0.001263
```

```
plot(anova)
```


Residuals vs Fitted
Residuals
0.05
-0.05
-0.15
247
319
300
0.325
0.330
0.335
0.340
0.345
Fitted values
aov(datos$bateo ~ datos$posicion)


Normal Q-Q
Standardized residuals
2
0
-2
-4
247
319
300
-3
-2
-1
0
1
2
3
Theoretical Quantiles
aov(datos$bateo ~ datos$posicion)





Dado que el *p-value* es superior a 0.05 no hay evidencias suficientes para considerar que al menos dos medias son distintas. La representación gráfica de los residuos no muestra falta de homocedasticidad (gráfico 1) y en el *qqplot* los residuos se distribuyen muy cercanos a la línea de la normal (gráfico 2).

4.Calcular el tamaño del efecto de un ANOVA

En el caso de ANOVA la mediada más comúnmente empleada es *eta cuadrado*.

$$\eta^2 = \frac{SC_{efecto}}{SC_{total}}$$

Los valores necesarios para calcular η^2 se obtienen del *summary* del ANOVA

```
eta_cuadrado <- 0.0076/(0.0076 + 0.4080)
eta_cuadrado
```

```
## [1] 0.01828681
```

5.Comparaciones múltiples

En este caso el ANOVA no ha resultado significativo por lo que no tiene sentido realizar comparaciones dos a dos. Sin embargo con fines didácticos se muestra como se harían. De entre los diferentes métodos de comparaciones múltiples y correcciones, se van a emplear las dos más recomendadas: Corrección de Holm y TukeyHSD.

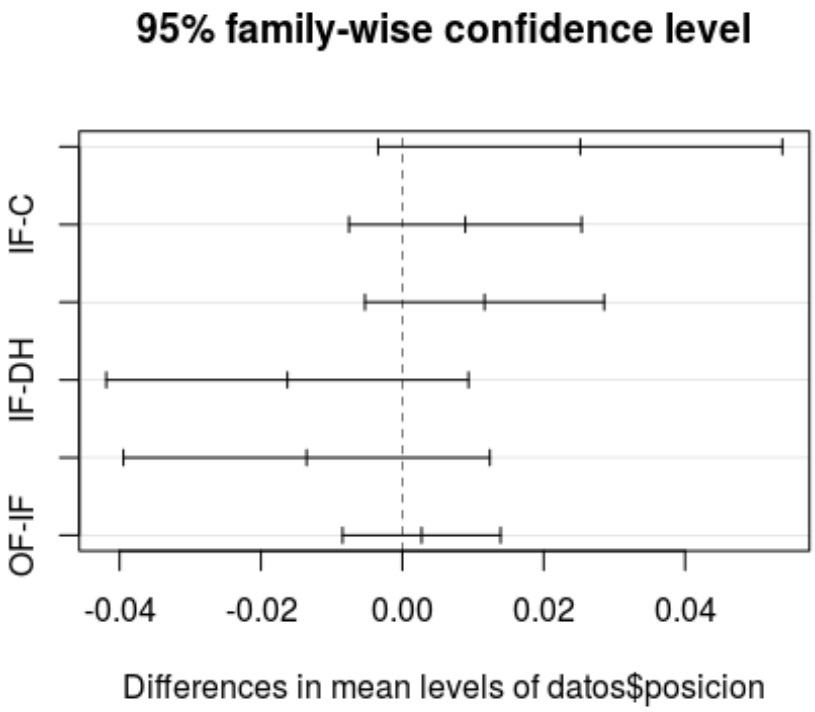
```
pairwise.t.test(x = datos$bateo, g = datos$posicion, p.adjust.method = "holm",
    pool.sd = TRUE, paired = FALSE, alternative = "two.sided")
```

```
##  Pairwise comparisons using t tests with pooled SD
##
## data:  datos$bateo and datos$posicion
##
##    C    DH   IF
## DH 0.14 -    -
## IF 0.49 0.41 -
## OF 0.38 0.49 0.53
##
## P value adjustment method: holm
```

```
TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = datos$bateo ~ datos$posicion)
##
## $`datos$posicion`
##                diff          lwr         upr     p adj
## DH-C   0.025170330 -0.003424987 0.053765647 0.1064933
## IF-C   0.008910589 -0.007542191 0.025363369 0.5011093
## OF-C   0.011634615 -0.005282602 0.028551833 0.2867635
## IF-DH -0.016259740 -0.041880002 0.009360521 0.3582394
## OF-DH -0.013535714 -0.039456673 0.012385244 0.5327045
## OF-IF  0.002724026 -0.008451748 0.013899800 0.9225240
```

```
plot(TukeyHSD(anova))
```



Como era de esperar no se encuentra diferencia significativa entre ningún par de medias.

6.Conclusión

En el estudio realizado se ha observado un tamaño de efecto pequeño y la técnicas de inferencia ANOVA no han encontrado significancia estadística para rechazar que las medias son iguales entre todos los grupos.

ANOVA de dos vías para datos independientes

Introducción

El análisis de varianza de dos vías, también conocido como plan factorial con dos factores, sirve para estudiar la relación entre una variable dependiente cuantitativa y dos variables independientes cualitativas (factores) cada uno con varios niveles. El ANOVA de dos vías permite estudiar cómo influyen por si solos cada uno de los factores sobre la variable dependiente (modelo aditivo) así como la influencia de las combinaciones que se pueden dar entre ellas (modelo con interacción).

Supóngase que se quiere estudiar el efecto de un fármaco sobre la presión sanguínea (variable cuantitativa dependiente) dependiendo del sexo del paciente (niveles: hombre, mujer) y de la edad (niveles: niño, adulto, anciano).

El efecto simple de los factores consiste en estudiar cómo varía el efecto del fármaco dependiendo del sexo sin diferenciar por edades, así como estudiar cómo varia el efecto del fármaco dependiendo de la edad sin tener en cuenta el sexo.

El efecto de la interacción doble consiste en estudiar si la influencia de uno de los factores varía dependiendo de los niveles del otro factor. Es decir, si la influencia del factor sexo sobre la actividad del fármaco es distinta según la edad del paciente o lo que es lo mismo, si la actividad del fármaco para una determinada edad es distinta según si se es hombre o mujer.

Condiciones ANOVA de dos vías para datos independientes

Las condiciones necesarias para que un ANOVA de dos vías sea válido, así como el proceso a seguir para realizarlo son semejantes al ANOVA de una vía. Las únicas diferencias son:

Hipótesis: El ANOVA de dos vías con repeticiones combina 3 hipótesis nulas, que las medias de las observaciones agrupadas por un factor son iguales, que las medias de las observaciones agrupadas por el otro factor son iguales; y que no hay interacción entre los dos factores.

Requiere calcular la la Suma de Cuadrados y Cuadrados Medios para ambos factores. Es frecuente encontrar que a un factor se le llama "tratamiento" y al otro "bloque o *block*"

Es importante tener en cuenta que el orden en el que se multiplican los factores no afecta al resultado del ANOVA **únicamente** si el tamaño de los grupos es igual (modelo equilibrado) de lo contrario sí importa. Por esta razón es muy recomendable que el diseño sea equilibrado.

El estudio de la interacción de los dos factores solo es posible si se dispone varias observaciones para cada una de las combinaciones de los niveles.

En R se puede realizar este tipo de ANOVA con las funciones:

- Modelo aditivo: `aov(variable_respuesta ~ factor1 + factor2, data)`
- Modelo con interacción: `aov(variable_respuesta ~ factor1 x factor2, data)`

Tamaño del efecto

En el caso del ANOVA con dos factores se puede calcular el tamaño del efecto η^2 para cada uno de los dos factores así como para la interacción.

Ejemplo 1

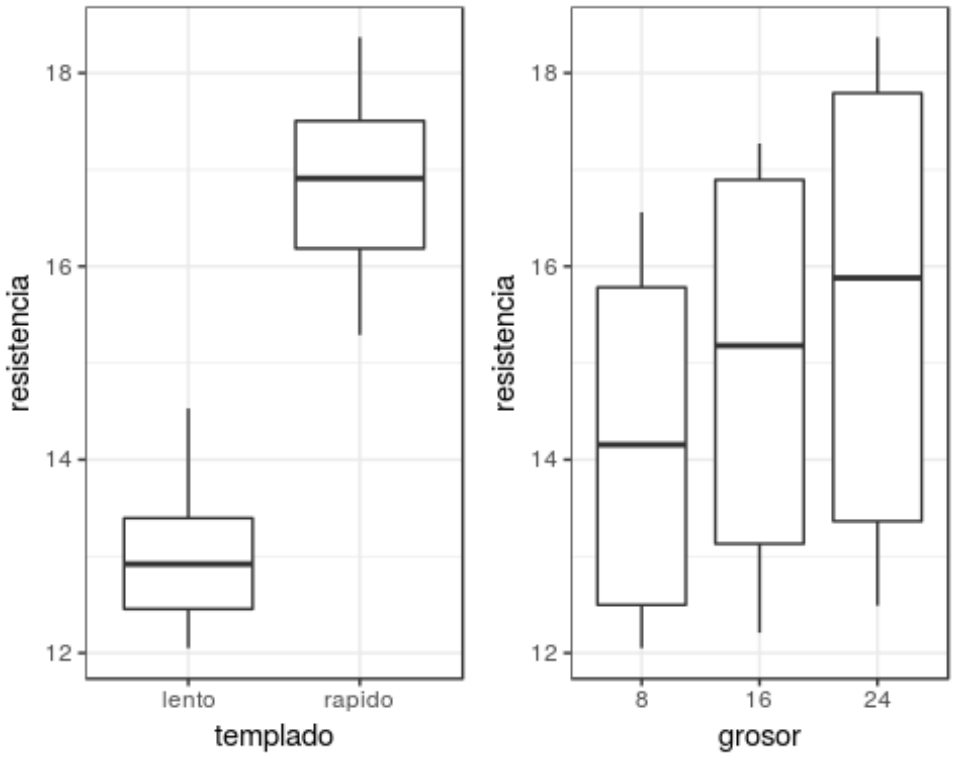
Una empresa de materiales de construcción quiere estudiar la influencia que tienen el grosor y el tipo de templado sobre la resistencia máxima de unas láminas de acero. Para ello miden el estrés hasta la rotura (variable cuantitativa dependiente) para dos tipos de templado (lento y rápido) y tres grosores de lámina (8mm, 16mm y 24 mm).

```
resistencia <- c(15.29, 15.89, 16.02, 16.56, 15.46, 16.91, 16.99, 17.27, 16.85,
    16.35, 17.23, 17.81, 17.74, 18.02, 18.37, 12.07, 12.42, 12.73, 13.02, 12.05,
    12.92, 13.01, 12.21, 13.49, 14.01, 13.3, 12.82, 12.49, 13.55, 14.53)
templado <- c(rep(c("rapido", "lento"), c(15, 15)))
grosor <- rep(c(8, 16, 24), each = 5, times = 2)
datos <- data.frame(templado = templado, grosor = as.factor(grosor), resistencia =
resistencia)
head(datos)
```

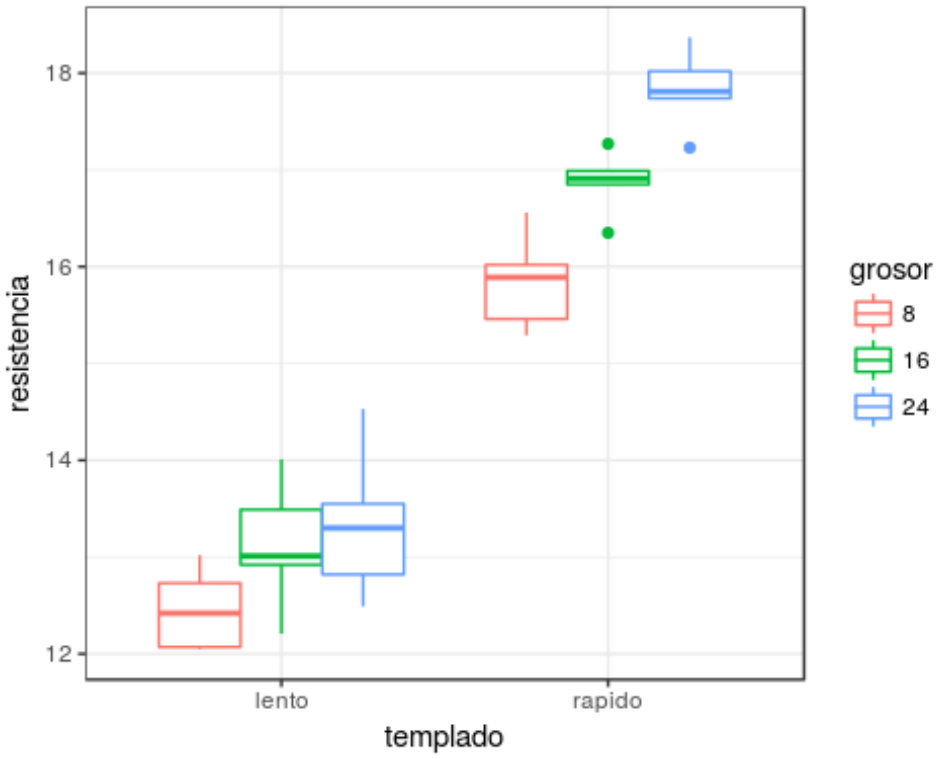
```
##   templado grosor resistencia
## 1   rapido      8       15.29
## 2   rapido      8       15.89
## 3   rapido      8       16.02
## 4   rapido      8       16.56
## 5   rapido      8       15.46
## 6   rapido     16       16.91
```

En primer lugar se generan los diagramas "Box-plot" para identificar posibles diferencias significativas, asimetrías, valores atípicos y homogeneidad de varianza entre los distintos niveles. Se puede acompañar a los gráficos con las medias y varianza de cada grupo.

```
library(ggplot2)
library(gridExtra)
p1 <- ggplot(data = datos, mapping = aes(x = templado, y = resistencia)) +
geom_boxplot() +
    theme_bw()
p2 <- ggplot(data = datos, mapping = aes(x = grosor, y = resistencia)) +
geom_boxplot() +
    theme_bw()
p3 <- ggplot(data = datos, mapping = aes(x = templado, y = resistencia, colour =
grosor)) +
    geom_boxplot() + theme_bw()
grid.arrange(p1, p2, ncol = 2)
```


resistencia
18
16
14
12
lento
rapido
templado
resistencia
18
16
14
12
8
16
24
grosor

p3


resistencia
18
16
14
12
lento
rapido
templado
grosor
8
16
24

```
with(data = datos, expr = tapply(resistencia, templado, mean))
```

```
##    lento   rapido
## 12.97467 16.85067
```

```
with(data = datos, expr = tapply(resistencia, templado, sd))
```

```
##     lento    rapido
## 0.7113455 0.9276427
```

```
with(data = datos, expr = tapply(resistencia, grosor, mean))
```

```
##      8     16     24
## 14.151 15.001 15.586
```

```
with(data = datos, expr = tapply(resistencia, grosor, sd))
```

```
##        8       16       24
## 1.836993 2.036797 2.442354
```

```
with(data = datos, expr = tapply(resistencia, list(templado, grosor), mean))
```

```
##             8     16     24
## lento  12.458 13.128 13.338
## rapido 15.844 16.874 17.834
```

```
with(data = datos, expr = tapply(resistencia, list(templado, grosor), sd))
```

```
##                8        16        24
## lento  0.4207969 0.6724730 0.7833709
## rapido 0.5000300 0.3341856 0.4171690
```

A partir de la representación gráfica y el cálculo de las medias se puede intuir que existe una diferencia en la resistencia alcanzada dependiendo del tipo de templado. La resistencia parece incrementarse a medida que aumenta el grosor de la lámina, si bien no está clara que la diferencia en las medias sea significativa. La distribución de las observaciones de cada nivel parece simétrica sin presencia de valores atípicos. A priori parece que se satisfacen las condiciones necesarias para un ANOVA, aunque habrá que confirmarlas estudiando los residuos.

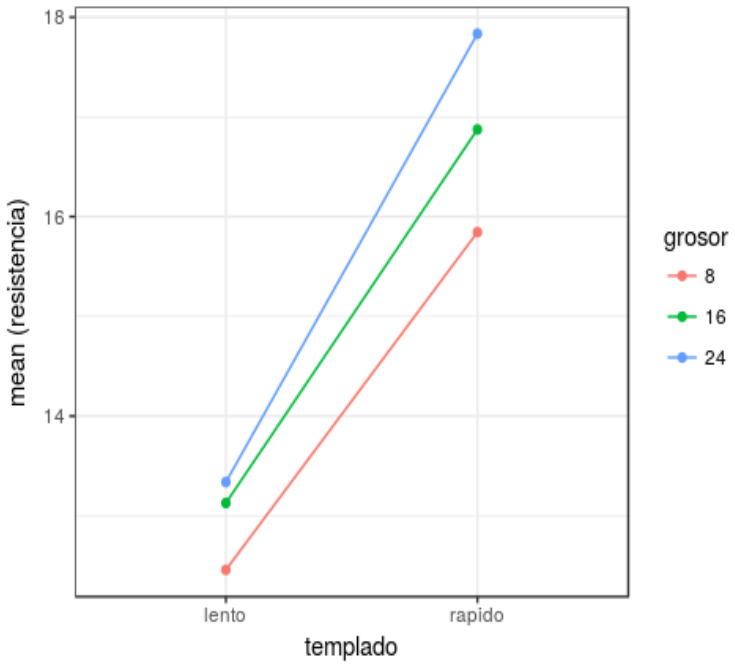
También es posible identificar posibles interacciones de los dos factores de forma gráfica mediante lo que se conocen como "gráficos de interacción". Si las líneas que describen los datos para cada uno de los niveles son paralelas significa que el comportamiento es similar independientemente del nivel del factor, es decir, no hay interacción.

Obtención de gráficos de interacción con los gráficos base de R

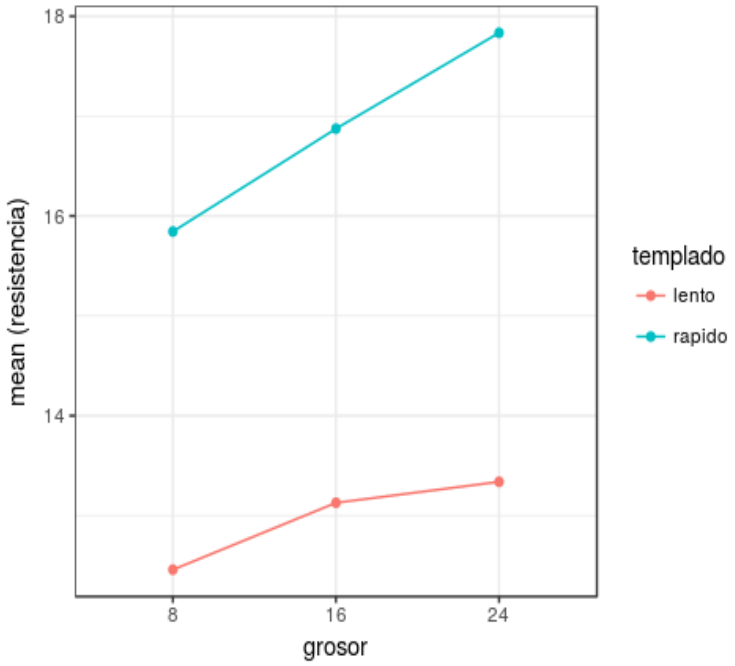
```
interaction.plot(templado, grosor, resistencia, data = datos, col = 1:3, type =
"b")
interaction.plot(grosor, templado, resistencia, data = datos,col = 2:3, type = "b")
```

Obtención de gráficos de interacción con ggplot2

```
ggplot(data = datos, aes(x = templado, y = resistencia, colour = grosor, group =
grosor)) +
 stat_summary(fun.y = mean, geom = "point") +
 stat_summary(fun.y = mean, geom = "line") +
 labs(y = "mean (resistencia)") +
 theme_bw()
```



```
ggplot(data = datos, aes(x = grosor, y = resistencia, colour = templado, group =
templado)) +
   stat_summary(fun.y = mean, geom = "point") +
   stat_summary(fun.y = mean, geom = "line") +
   labs(y = "mean (resistencia)") +
   theme_bw()
```



Los gráficos de interacción parecen indicar (a falta de obtener los *p-values* mediante el ANOVA) que el incremento de resistencia entre los dos tipos de templado es proporcional para los tres grosores. Al representar la resistencia en función del grosor para los dos tipos de templado, parece observarse cierta desviación en el grosor 24mm. Esta ligera desviación podría deberse a simple variabilidad o porque existe interacción entre las variables grosor y templado, por lo que tiene que ser confirmada mediante el ANOVA.

```
anova <- aov(resistencia ~ templado * grosor, data = datos)
summary(anova)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## templado        1 112.68  112.68 380.082 3.19e-16 ***
## grosor          2  10.41    5.21  17.563 2.00e-05 ***
## templado:grosor  2   1.60    0.80   2.705   0.0873 .
## Residuals       24   7.11    0.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tamaño de efecto

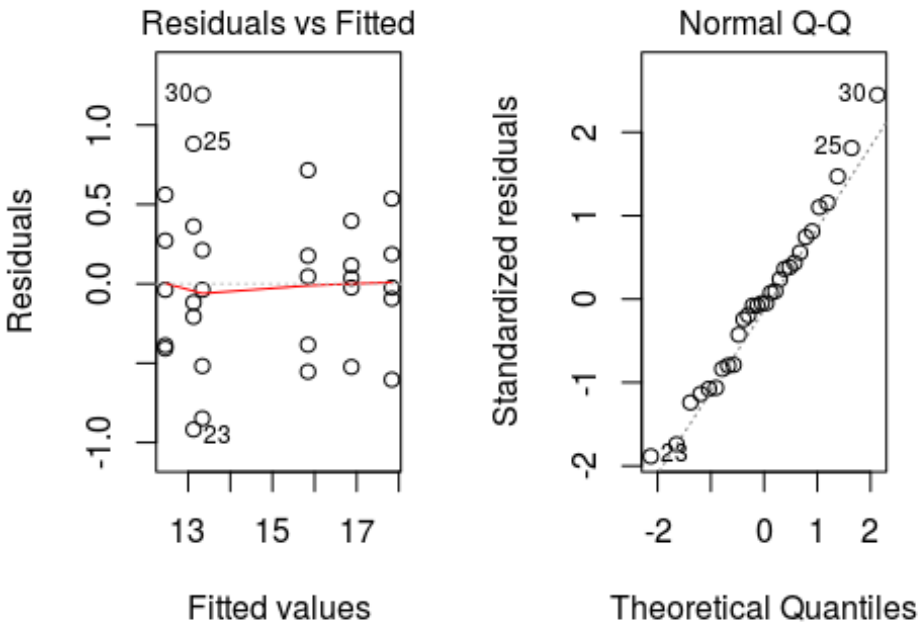
```
library(lsr)
etaSquared(anova)
```

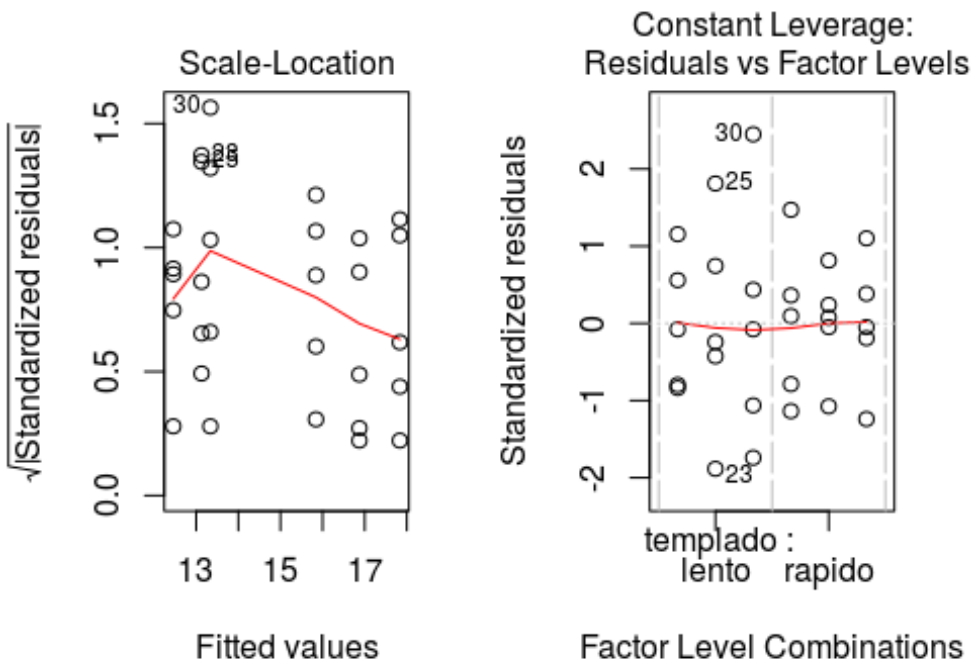
```
##                     eta.sq eta.sq.part
## templado        0.85485219   0.9406061
## grosor          0.07900327   0.5940887
## templado:grosor 0.01216553   0.1839235
```

El análisis de varianza confirma que existe una influencia significativa sobre la resistencia de las láminas por parte de ambos factores (templado y grosor) con tamaños de efecto η^2 grande y mediano respectivamente, pero que no existe interacción significativa entre ellos.

Para poder dar por validos los resultados del ANOVA es necesario verificar que se satisfacen las condiciones de un ANOVA.

```
par(mfrow = c(1,2))
plot(anova)
```






```
par(mfrow = c(1,1))
```

Los residuos muestran la misma varianza para los distintos niveles (homocedasticidad) y se distribuyen de forma normal.

Ejemplo 2

Supóngase un estudio clínico que analiza la eficacia de un medicamento teniendo en cuenta dos factores, el sexo (masculino y femenino) y la juventud (joven, adulto). Se quiere analizar si el efecto es diferente entre alguno de los niveles de cada variable por si sola o en combinación.

Este estudio implica comprobar si el efecto medio del fármaco es significativamente distinto entre alguno de los siguientes grupos: hombres, mujeres, jóvenes, adultos, hombres jóvenes, hombres adultos, mujeres jóvenes y mujeres adultas.

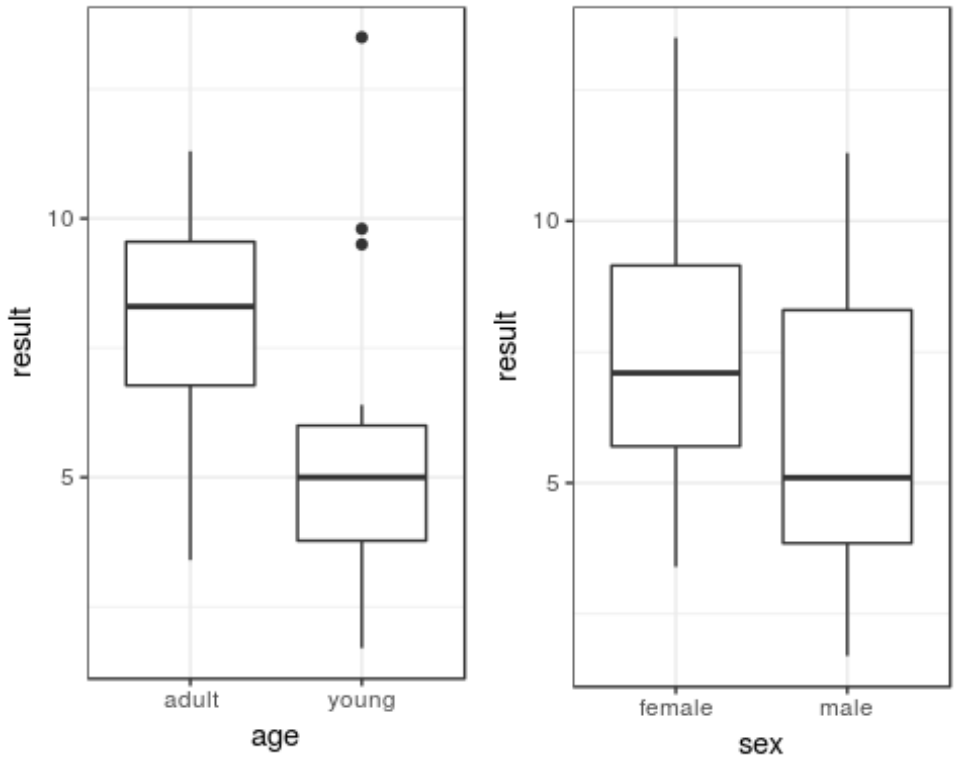
```
subject <- as.factor(c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,
    17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30))
sex <- c("female", "male", "male", "female", "male", "male", "male", "female",
    "female", "male", "male", "male", "male", "female", "female", "female",
    "male", "female", "female", "male", "male", "female", "male", "male", "male",
    "male", "male", "male", "female", "male")
age <- c("adult", "adult", "adult", "adult", "adult", "adult", "young", "young",
    "adult", "young", "young", "adult", "young", "young", "young", "adult",
    "young", "adult", "young", "young", "young", "young", "adult", "young",
    "young", "young", "young", "young", "young", "adult")
result <- c(7.1, 11, 5.8, 8.8, 8.6, 8, 3, 5.2, 3.4, 4, 5.3, 11.3, 4.6, 6.4,
    13.5, 4.7, 5.1, 7.3, 9.5, 5.4, 3.7, 6.2, 10, 1.7, 2.9, 3.2, 4.7, 4.9, 9.8,
    9.4)

datos <- data.frame(subject, sex, age, result)
head(datos, 4)
```

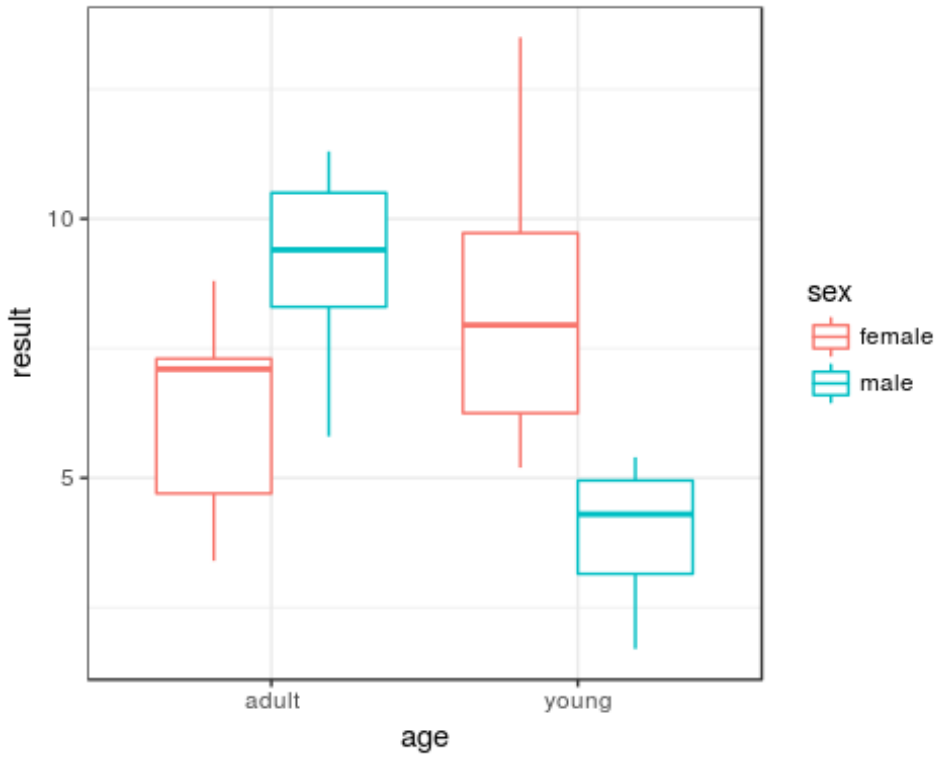
```
##   subject    sex   age result
## 1       1 female adult    7.1
## 2       2   male adult   11.0
## 3       3   male adult    5.8
## 4       4 female adult    8.8
```

En primer lugar se generan los diagramas "Box-plot" para identificar posibles diferencias significativas, asimetrías, valores atípicos y homogeneidad de varianza entre los distintos niveles. Se acompaña a los gráficos de la media y varianza de cada grupo.

```
p1 <- ggplot(data = datos, mapping = aes(x = age, y = result)) + geom_boxplot() +
    theme_bw()
p2 <- ggplot(data = datos, mapping = aes(x = sex, y = result)) + geom_boxplot() +
    theme_bw()
p3 <- ggplot(data = datos, mapping = aes(x = age, y = result, colour = sex)) +
    geom_boxplot() + theme_bw()
grid.arrange(p1, p2, ncol = 2)
```


result
10
5
adult
young
age
result
10
5
female
male
sex

p3


result
10
5
adult
young
age
sex
female
male

```
with(data = datos, expr = tapply(result, sex, mean))
```

```
##   female     male
## 7.445455 5.926316
```

```
with(data = datos, expr = tapply(result, sex, sd))
```

```
##   female     male
## 2.828202 2.906858
```

```
with(data = datos, expr = tapply(result, age, mean))
```

```
##    adult    young
## 7.950000 5.505556
```

```
with(data = datos, expr = tapply(result, age, sd))
```

```
##    adult    young
## 2.431049 2.871047
```

```
with(data = datos, expr = tapply(result, list(sex, age), mean))
```

```
##           adult    young
## female 6.260000 8.433333
## male   9.157143 4.041667
```

```
with(data = datos, expr = tapply(result, list(sex, age), sd))
```

```
##           adult    young
## female 2.170944 3.106552
## male   1.900752 1.157158
```

A partir de la representación gráfica y el cálculo de las medias se puede intuir que existe una diferencia en el efecto del fármaco dependiendo de la edad y también del sexo. El efecto parece ser mayor en mujeres que en hombres y en adultos que en jóvenes, si bien la significancia se tendrá que confirmar con el ANOVA. La distribución de las observaciones de cada nivel parece simétrica con la presencia de un único valor atípico. A priori parece que se satisfacen las condiciones necesarias para un ANOVA, aunque habrá que confirmarlas estudiando los residuos.

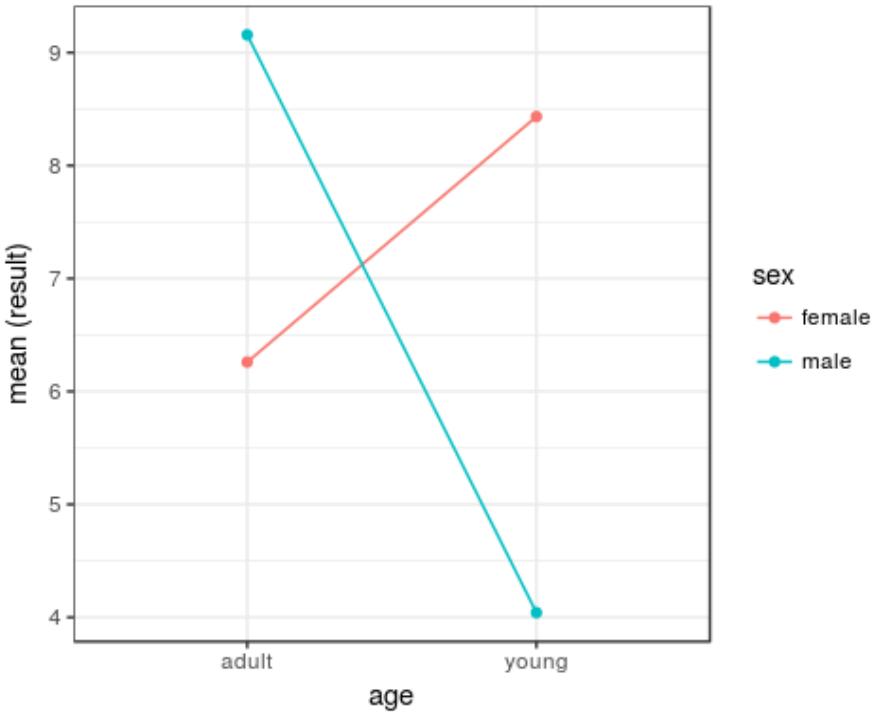
Es posible identificar posibles interacciones de los dos factores de forma gráfica mediante lo que se conocen como "gráficos de interacción". Si las líneas que describen los datos para cada uno de los niveles son paralelas significa que el comportamiento es similar independientemente del nivel del factor, es decir, no hay interacción.

Obtención de gráficos de interacción con los gráficos base de R

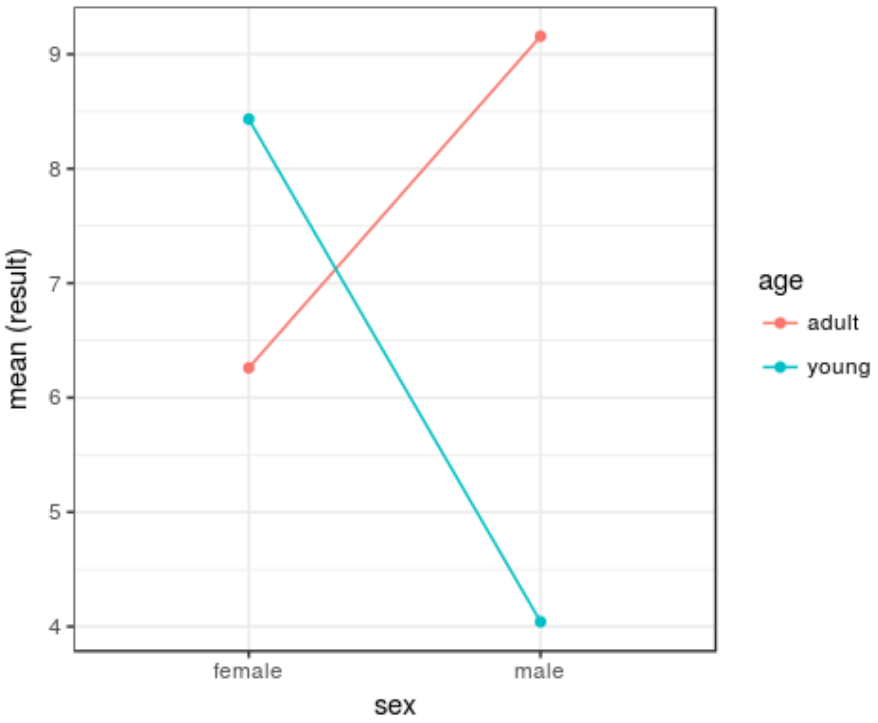
```
interaction.plot(trace.factor = datos$sex, x.factor = datos$age, response =
datos$result, fun = "mean", legend = TRUE, col = 2:3, type = "b")
interaction.plot(trace.factor = datos$age, x.factor = datos$sex, response =
datos$result,
    fun = "mean", legend = TRUE, col = 2:3, type = "b")
```

Obtención de gráficos de interacción con ggplot2

```
ggplot(data = datos, aes(x = age, y = result, colour = sex, group = sex)) +
    stat_summary(fun.y = mean, geom = "point") + stat_summary(fun.y = mean,
    geom = "line") + labs(y = "mean (result)") + theme_bw()
```




```
ggplot(data = datos, aes(x = sex, y = result, colour = age, group = age)) +
    stat_summary(fun.y = mean, geom = "point") +
 stat_summary(fun.y = mean, geom = "line") +
 labs(y = "mean (result)") +
 theme_bw()
```



Se observa una clara interacción entre ambos factores. La respuesta al fármaco es distinta entre adultos y jóvenes, y de tendencia inversa dependiendo del sexo. En mujeres, la respuesta es mayor cuando son jóvenes que cuando son adultas y en hombres mayor cuando son adultos y menor cuando son jóvenes. El ANOVA permite saber si las diferencias observadas son significativas.

```
anova_2vias <- aov(formula = result ~ sex*age, data = datos)
summary(anova_2vias)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## sex          1  16.08   16.08   4.038  0.0550 .
## age          1  38.96   38.96   9.786  0.0043 **
## sex:age      1  89.61   89.61  22.509 6.6e-05 ***
## Residuals   26 103.51    3.98
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

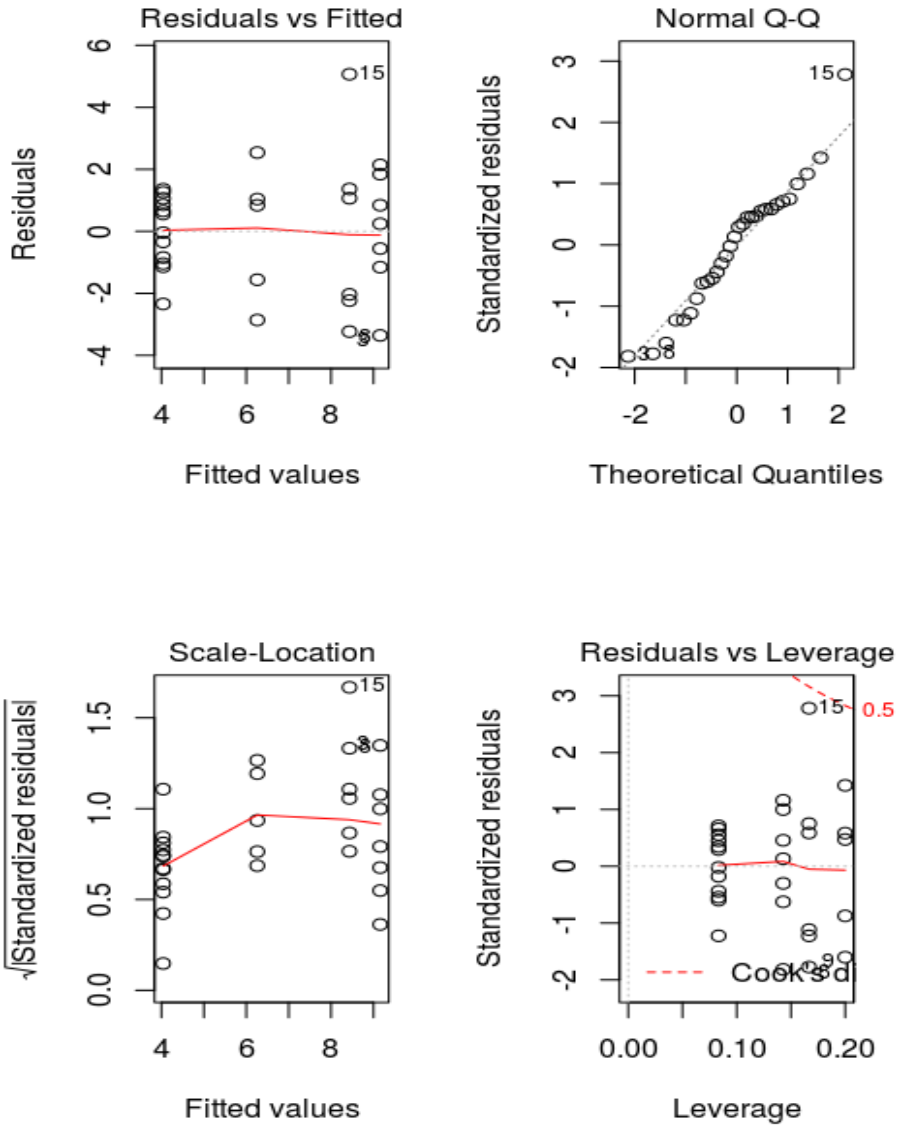
```
etaSquared(anova_2vias)
```

```
##             eta.sq eta.sq.part
## sex     0.04842176   0.1040130
## age     0.15699885   0.2734635
## sex:age 0.36110080   0.4640119
```

El análisis de varianza no encuentra diferencias significativas en el efecto del fármaco entre hombres y mujeres (factor sex) pero sí encuentra diferencias significativas entre jóvenes y adultos y entre al menos dos grupos de las combinaciones de sexo y edad, es decir, hay significancia para la interacción. El tamaño del efecto η^2 es grande tanto para edad como para la interacción de edad y sexo. *Importante: El orden en el que se multiplican los factores no afecta únicamente si el tamaño de los grupos es igual, de lo contrario sí afecta.*

Para poder dar por válidos los resultados del ANOVA es necesario verificar que se satisfacen las condiciones de un ANOVA.

```
par(mfrow = c(1,2))
plot(anova_2vias)
```



```
par(mfrow = c(1,1))
```

Los residuos muestran la misma varianza para los distintos niveles (homocedasticidad) y se distribuyen de forma normal. La observación número 15 tiene un residuo atípicamente grande. Sería conveniente repetir el ANOVA sin esta observación para comprobar el impacto.

ANOVA con variables dependientes (ANOVA de medidas repetidas)

Introducción

Cuando las variables a comparar son mediciones distintas pero sobre los mismos sujetos, no se cumple la condición de independencia, por lo que se requiere un ANOVA específico que realice comparaciones considerando que los datos son pareados (de forma similar como se hace en los t-test pareados pero para comparar más de dos grupos).

Condiciones para ANOVA de variables dependientes

Esfericidad:

Es la única condición para poder aplicar este tipo de análisis. La esfericidad implica que la varianza de las diferencias entre todos los pares de variables a comparar sea igual. Si el ANOVA se realiza para comparar la media de una variable cuantitativa entre tres niveles (A, B, C), se aceptará la esfericidad si la varianza de las diferencias entre A-B = A-C = B-C.

- La esfericidad se puede analizar mediante el test de *Mauchly* cuya hipótesis nula es que sí existe esfericidad.
- Con frecuencia se viola la condición de esfericidad, pero se pueden aplicar dos tipos de correcciones que pueden hacer posible el seguir adelante con el ANOVA. Son la corrección de *Greenhouse-Geisser* y la de *Huynh-Feldt*.
- Otra alternativa en caso de no cumplirse la esfericidad es el test no paramétrico *test de Friedman*.

La función `aov()` permite realizar ANOVA de medidas repetidas pero no comprueba la esfericidad ni permite incorporar las correcciones. La función `Anova()` del paquete *car* es más completa, devuelve junto con los resultados del anova, el test de esfericidad de *Mauchly* y los valores que se obtiene con las correcciones de *Greenhouse-Geisser* y la de *Huynh-Feldt*. La función `Anova()` requiere que los datos estén almacenados en una matriz en formato de "tabla ancha".

Ejemplo

Supóngase un estudio en el que se quiere comprobar si el precio la compra varía entre 4 cadenas de supermercado distintas. Para ellos se seleccionan una serie de elementos de la compra cotidiana y se registra su valor en cada uno de los supermercados ¿Existen evidencias de que el precio medio de la compra es diferente dependiendo del supermercado?

Se trata de distintas mediciones sobre un mismo elemento por lo tanto son datos pareados.

```
elemento <- c("lettuce", "potatoes", "milk", "eggs", "bread", "cereal",
"ground.beef",
    "tomato.soup", "laundry.detergent", "aspirin")
tienda_A <- c(1.755, 2.655, 2.235, 0.975, 2.37, 4.695, 3.135, 0.93, 8.235, 6.69)
tienda_B <- c(1.78, 1.98, 1.69, 0.99, 1.7, 3.15, 1.88, 0.65, 5.99, 4.84)
tienda_C <- c(1.29, 1.99, 1.79, 0.69, 1.89, 2.99, 2.09, 0.65, 5.99, 4.99)
tienda_D <- c(1.29, 1.99, 1.59, 1.09, 1.89, 3.09, 2.49, 0.69, 6.99, 5.15)

datos <- data.frame(elemento, tienda_A, tienda_B, tienda_C, tienda_D)
datos
```

```
##             elemento tienda_A tienda_B tienda_C tienda_D
## 1            lettuce    1.755     1.78     1.29     1.29
## 2           potatoes    2.655     1.98     1.99     1.99
## 3               milk    2.235     1.69     1.79     1.59
## 4               eggs    0.975     0.99     0.69     1.09
## 5              bread    2.370     1.70     1.89     1.89
## 6             cereal    4.695     3.15     2.99     3.09
## 7        ground.beef    3.135     1.88     2.09     2.49
## 8        tomato.soup    0.930     0.65     0.65     0.69
## 9  laundry.detergent    8.235     5.99     5.99     6.99
## 10           aspirin    6.690     4.84     4.99     5.15
```

Para poder visualizar los datos con *ggplot2* y realizar la exploración inicial se pasa de formato de "tabla ancha" a "tabla larga" con una columna por variable.

```
# Instalar paquete tidyr
require(tidyr)
datos_tabla_larga <- gather(data = datos, key = "tienda", value = "precio", 2:5)
head(datos_tabla_larga, 5)
```

```
##   elemento   tienda precio
## 1  lettuce tienda_A  1.755
## 2 potatoes tienda_A  2.655
## 3     milk tienda_A  2.235
## 4     eggs tienda_A  0.975
## 5    bread tienda_A  2.370
```

Es recomendable calcular el precio total de cada uno de los grupos, así como una representación gráfica para hacerse una idea de cuales podrían diferir significativamente.

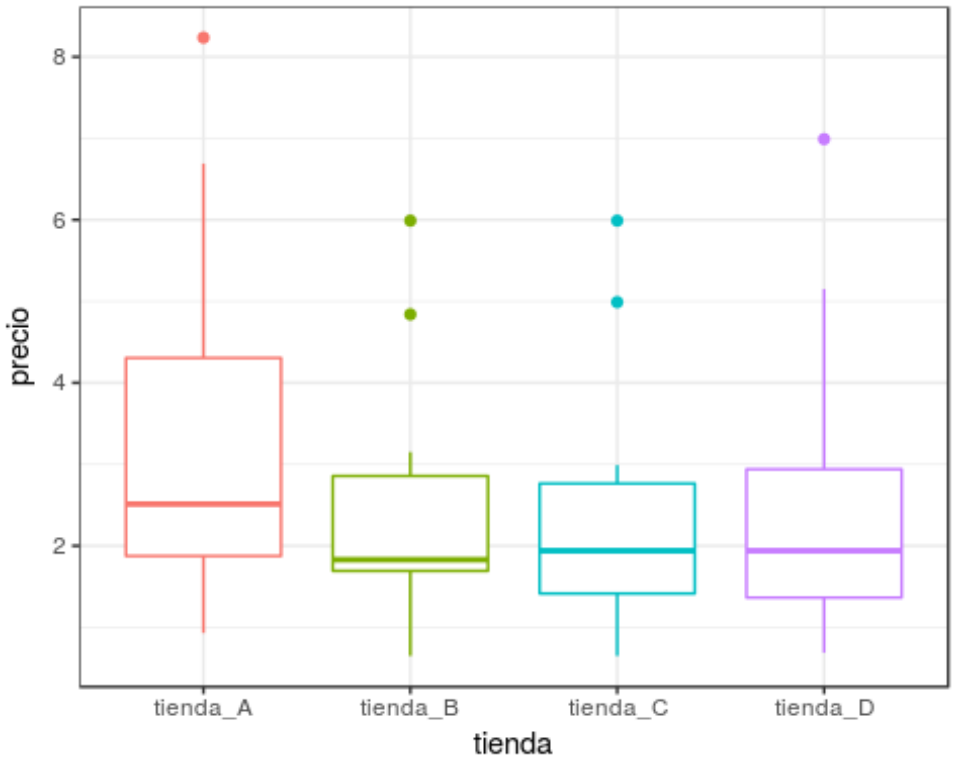
```
with(data = datos_tabla_larga, expr = tapply(precio, tienda, mean))
```

```
## tienda_A tienda_B tienda_C tienda_D
##   3.3675   2.4650   2.4360   2.6260
```

```
with(data = datos_tabla_larga, expr = tapply(precio, tienda, sd))
```

```
## tienda_A tienda_B tienda_C tienda_D
## 2.440371 1.707430 1.765296 1.987758
```

```
ggplot(data = datos_tabla_larga, mapping = aes(x = tienda, y = precio, colour =
tienda)) +
    geom_boxplot() + theme_bw() + theme(legend.position = "none")
```



ANOVA de datos pareados aov()v

```
anova_pareado <- aov(formula = precio ~ tienda + Error(elemento/tienda), data =
datos_tabla_larga)
summary(anova_pareado)
```

```
##
## Error: elemento
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals  9  139.5    15.5
##
## Error: elemento:tienda
##           Df Sum Sq Mean Sq F value  Pr(>F)
## tienda     3  5.737  1.9124   13.03 1.9e-05 ***
## Residuals 27  3.964  0.1468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
eta_cuadrado <- 5.610/(5.610 + 3.506)
eta_cuadrado
```

```
## [1] 0.6154015
```

El análisis de varianza resulta significativo con un tamaño de efecto grande, pero al no haberse comprobado la condición de esfericidad no se pueden considerar válidos.

ANOVA de datos pareados Anova()

Realizar un ANOVA de datos pareados con la función *Anova()* requiere tener los datos almacenados en un formato específico. Los valores de la variable continua se almacenan en una matriz en la que cada columna representa un nivel distinto (grupo) de la variable cualitativa.

```
datos <- as.matrix(datos[-1])
# se excluye la primera columna porque contiene los niveles del factor.
```

El siguiente paso es crear un modelo lineal multivariable. Los coeficientes estimados coinciden con la media de cada grupo.

```
modelo_lm <- lm(datos ~ 1)
```

Se define el diseño del estudio, es decir, definir los diferentes grupos.

```
tienda <- factor(c("tienda_A", "tienda_B", "tienda_C", "tienda_D"))
```

Por último se emplea la función `Anova()` para realizar el análisis de varianza, el test de esfericidad y las correcciones.

```r
library(car)
anova_pareado <- Anova(modelo_lm, idata = data.frame(tienda), idesign = ~tienda,
    type = "III")
summary(anova_pareado, multivariate = F)
```

```
## Univariate Type III Repeated-Measures ANOVA Assuming Sphericity
##
##                  SS num Df Error SS den Df      F    Pr(>F)
## (Intercept) 296.725      1  139.479      9 19.146  0.001782 **
## tienda        5.737      3    3.964     27 13.025 1.898e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Mauchly Tests for Sphericity
##
##        Test statistic   p-value
## tienda        0.12901 0.0078973
##
##
## Greenhouse-Geisser and Huynh-Feldt Corrections
##  for Departure from Sphericity
##
##         GG eps Pr(>F[GG])
## tienda 0.46824   0.001747 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           HF eps Pr(>F[HF])
## tienda 0.5287146 0.00103603
```

No es posible crear los intervalos Tukey HSD para datos pareados. Para las comparaciones múltiples después de que el ANOVA haya resultado significativo hay que recurrir a otras correcciones.

```r
pairwise.t.test(x = datos_tabla_larga$precio, g = datos_tabla_larga$tienda,
    p.adjust.method = "holm", paired = TRUE, alternative = "two.sided")
```

```
##
##  Pairwise comparisons using paired t tests
##
## data:  datos_tabla_larga$precio and datos_tabla_larga$tienda
```

```
##
##           tienda_A tienda_B tienda_C
## tienda_B 0.022    -        -
## tienda_C 0.014    0.695    -
## tienda_D 0.014    0.491    0.331
##
## P value adjustment method: holm
```

Conclusión

El análisis ANOVA (incluyendo las correcciones dada la falta de esfericidad) encuentra diferencias significativas en el precio de los alimentos entre al menos 2 tiendas, siendo el tamaño del efecto grande. La posterior comparación dos a dos por *t-student* con corrección de significancia *holm* identifica como significativas las diferencias entre las tiendas A-B, A-C y A-D pero no entre B-C, B-D y C-D.

Bibliografía

Open Intro Statistics

Statistics Using R with Biological Examples

TheRBook Michael J Crawley

Handbook of Biological Statistics

Métodos estadísticos en ingenieria Rafael Romero Villafranca, Luisa Rosa Zúnica Ramajo

R Tutorials by William B. King, Ph.D http://ww2.coastal.edu/kingw/statistics/R-tutorials/