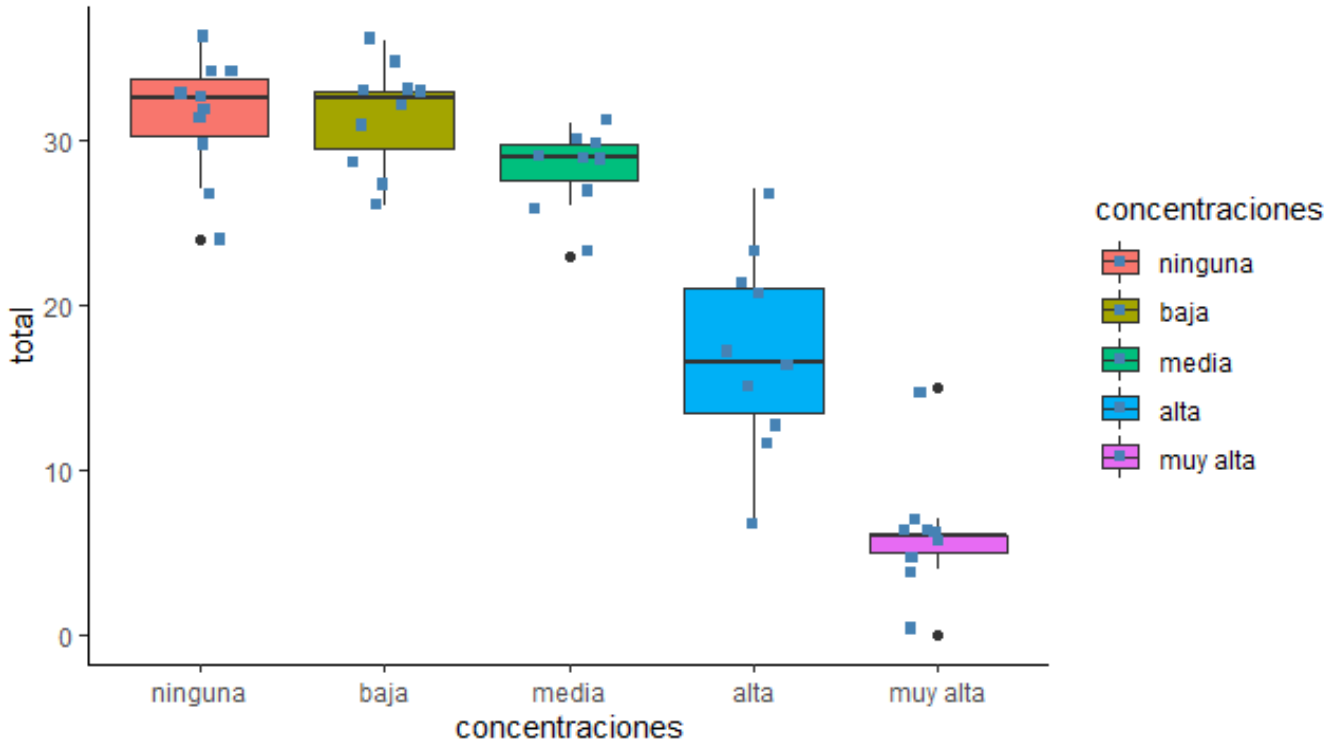
Ejemplo de Análsis de ANOVA de un solo factor

De la base de datos:

https://vincentarelbundock.github.io/Rdatasets/articles/data.html

- Se seleccionan los datos denominados “Toxicity of Nitrofen in Aquatic Systems” y se estudia la variable “Total” que cuenta el número total de crías vivas de una especie de zooplancton bajo diferentes concentraciones (variable “conc”) de un herbicida (nitrofen). Se realiza una descripción de los datos con un diagrama de cajas y bigotes.



Dónde la escala de concentraciones es:

Concentración	Ninguna	baja	media	alta	muy alta
Valor	0	80	160	235	310

Se graficaron 50 valores de los cuales 10 se asocian a cada concentración analizada. Es decir que hay 10 réplicas para cada concentración.

De la gráfica se observa que a bajas concentraciones del herbicida hay un mayor número total de crías vivas de zooplancton, aunque a concentración “alta” y “muy alta” ya hay una caída en el número total de crías vivas de zooplancton. Esto se puede observar también en los valores de las medias.

Además, se observan algunos valores extremos a concentración cero, media y muy alta.

Concentraciones	Media de Total (crías vivas)
ninguna	31.4
baja	31.5
media	28.3
alta	17.2
muy alta	6

- A continuación se realiza el test de ANOVA para explicar si alguno de los tratamientos es significativamente diferente a la media del control.

  Para tratar el problema de forma general se define

  y_{ij} a la j – ésima observación tomada bajo la concentración i.

  Si se lo piensa como una variable aleatoria, se puede expresar

  $Y_{ij} = \mu + \tau_i + E_{ij}$ para i = 1, 2, …, a y j = 1, 2, …, n

  aquí μ es la media de toda la población (media total), τi es el parámetro asociado al tratamiento i, y Eij es el error que debe ser N (0, σ^2) e independiente (iid)

  Para este caso i = 5 y j = 10

  Si se define la media poblacional de cada tratamiento

  $\mu_i = \mu + \tau_i$

  Lo que significa que cada concentración es una variable aleatoria con distribución N (0, σ^2).

  El problema que se está analizando se denomina de efectos fijos (también pueden ser aleatorios). Es decir, se fijó previamente los niveles de concentración (en este caso a = 5). Como τ_i se puede pensar como desviaciones de la media total μ entonces se lo utiliza para probar

  $\sum_{i=1}^{a} \tau_i = 0$

  Lo que significa que no hay diferencia de total de crías vivas bajo ninguna concentración de herbicida, es decir, no modificará la media total μ.

  Se define el test de ANOVA como:

  $H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5 = 0$ hipótesis nula

  $H_1: \tau_i \neq 0$ para al menos una i (concentración)

  Definiendo la variabilidad total del ANOVA particionada en dos componentes

$SS_T = n\ \sum_{i=1}^{a}(\bar{y}_{i\blacksquare} - \bar{y}_{\blacksquare\blacksquare})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\blacksquare}\right)^2$

Donde

$SS_{conc} = n\ \sum_{i=1}^{a}(\bar{y}_{i\blacksquare} - \bar{y}_{\blacksquare\blacksquare})^2$ cuantifica la diferencia de concentraciones

$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\blacksquare}\right)^2$ cuantifica el error aleatorio solamente

La idea principal es **comparar** estas variaciones, si la variación explicada por las diferencias en μ_i es grande respecto de la no explicada entonces se rechaza H_0.

Los grados de libertad de SS_{conc} y SS_E son respectivamente, $a - 1$ y $a(n - 1)$. Por lo tanto, para obtener magnitudes cuadráticas medias se divide por los grados de libertad respectivos:

$MSConc = \frac{SS_{conc}}{a-1}$ y $MSE = \frac{SS_E}{a(n-1)}$

Tomando de base lo anterior se define el **Estadístico** del test como

$F_0 = \frac{MSConc}{MSE}$

De donde se puede decir que si H_0 es verdadera F_0 tiene distribución F con $a - 1$ y $a(n - 1)$ grados de libertad para el numerador y denominador, respectivamente.

Finalmente, el test será: si $f_0 > f_{\alpha,(a-1),a(n-1)}$ (donde $f_{\alpha,(a-1),a(n-1)}$ es el valor crítico) la Hipótesis nula se rechazará.

Para la realización del análisis mediante el programa R se toma el p-valor para evaluar el resultado del test.

Como salida del test se obtiene

	Df	Sum Sq	Mean Sq	F value	P-valor
Concentraciones	4	4935	1233.7	79.24	2 e-16
Residuals	45	701	15.6		

Se concluye que al ser el p-valor tan pequeño se rechaza la Hipótesis nula, es decir, alguna de las concentraciones produce valores con media distinta a la de control.

- Se computan los intervalos de confianza para las diferencias de medias, a fin de determinar y cuáles son significativamente diferentes.

Se pueden crear intervalos de confianza (IC) tomando las medias de los tratamientos para así poder compararlos. Esto se hace utilizando $(\mu_i - \mu_j)$.

Un estimador puntual de esta diferencia es $\hat{\mu}_i - \hat{\mu}_j = \bar{Y}_{i\blacksquare} - \bar{Y}_{j\blacksquare}$. MSE es estimador de σ^2, pero $V\{\bar{Y}_{i\blacksquare} - \bar{Y}_{j\blacksquare}\} = 2\sigma^2/n$, por lo tanto el estadístico será

$$T = \frac{\bar{Y}_{i\blacksquare} - \bar{Y}_{j\blacksquare} - (\mu_i - \mu_j)}{\sqrt{2MSE/n}}$$

Tiene distribución t-student con $a(n-1)$ grados de libertad. Es importante remarcar que si se comparan dos concentraciones y el IC contiene el cero entonces no se puede descartar la hipótesis de igualdad de esas medias. Es decir, no habrá diferencia entre las dos concentraciones.

De la salida del software se obtienen los siguientes valores

	2.5 %	97.5 %
Intercept	28.886895	33.9131046
Concentraciones baja	-3.454067	3.6540666
Concentraciones media	-6.654067	0.4540666
Concentraciones alta	-17.754067	-10.6459334
Concentraciones muy alta	-28.954067	-21.8459334

Dado que el cero está en el IC para las comparaciones entre “ninguna” concentración con “baja” y “media” concentración, entonces no se puede descartar la hipótesis de igualdad de esas medias. Es decir que no habrá diferencias entre esos pares comparados de concentraciones.

Por otro lado, se puede decir que existe evidencia para decir que para las concentraciones “alta” y “muy alta” las medias son diferentes ya que no contienen al cero en el IC.

Si quisiéramos saber en cual tratamiento en particular hay diferencia se puede comprar de a pares mediante el procedimiento denominado Least Significant Difference de Fisher.

Se compara $LSD = t_{\alpha/2,a(n-1)}\sqrt{2MSE/n}$ con las diferencias $\left|\bar{y}_{i\blacksquare} - \bar{y}_{j\blacksquare}\right|$. Si esto último es mayor, habrá diferencia significativa.

La salida de R es directamente los p-valores de todas las diferencias de a pares, utilizando el método de Bonferroni para corregir significancia dividiéndola entre el número de grupos (concentraciones).

	ninguna	baja	media	alta
baja	1.00	-	-	-
media	0.86	0.76	-	-
alta	2.9 e-09	2.4 e-09	1.2 e-06	-
muy alta	< 2 e-16	< 2 e-16	< 2.1 e-15	9.5 e-07

De la tabla de p-valores se observa que hay diferencias significativas entre los pares de concentraciones “alta” con todas las demás y de “muy alta” con todas las demás concentraciones, debido a que el p-valor para dichos casos es muy pequeño.

- Se verifica normalidad y el comportamiento de los residuos con el objetivo de justificar si es válido utilizar el método de Análisis de Varianza.

ANOVA asume:

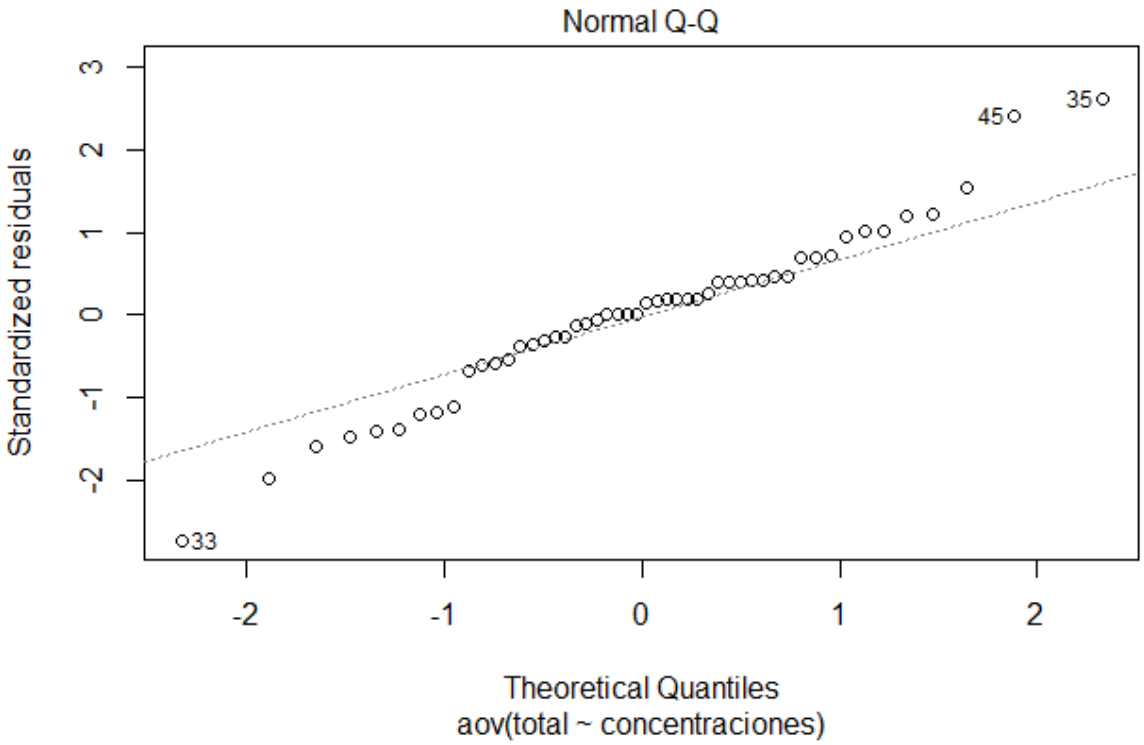
(1) que las observaciones tienen distribución Normal y son independientes.

(2) que la varianza es homogénea, constante.

Estas dos condiciones deben ser verificadas. Para esto debemos analizar **los residuos**, que son las diferencias entre la observación y_{ij} y el valor estimado ($\hat{y}_{ij} = \bar{y}_{i\bullet}$). Si analizamos, por ejemplo, la concentración i entonces el residuo $e_{ij} = y_{ij} - \bar{y}_{i\bullet}$.

El residuo tendrá la información que no explica el modelo.

(1) Para verificar la condición **Normalidad**, se puede realizar un qqplot.



Del gráfico anterior se observa que el residuo tiene una distribución aproximadamente normal.

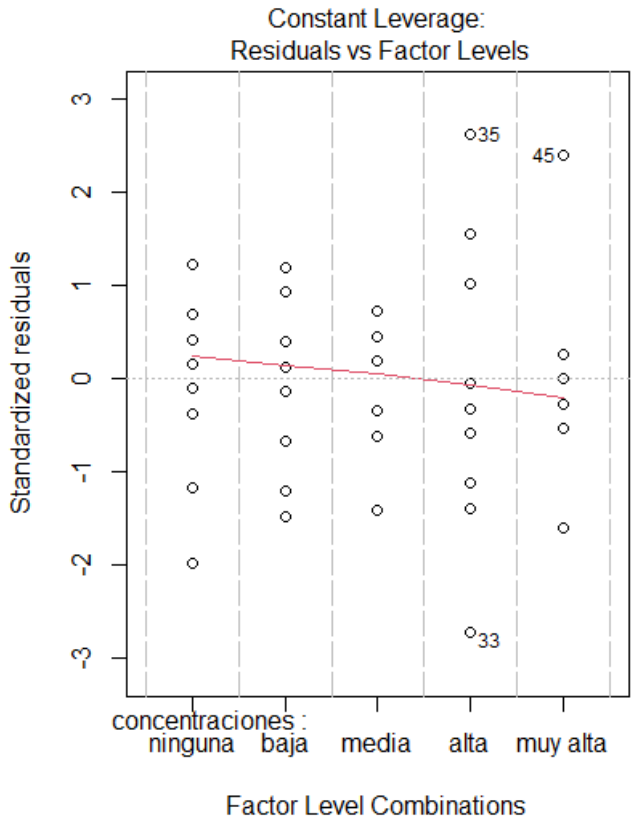
<u>También se puede utilizar el test de Shapiro-Wilk para verificar Normalidad</u>

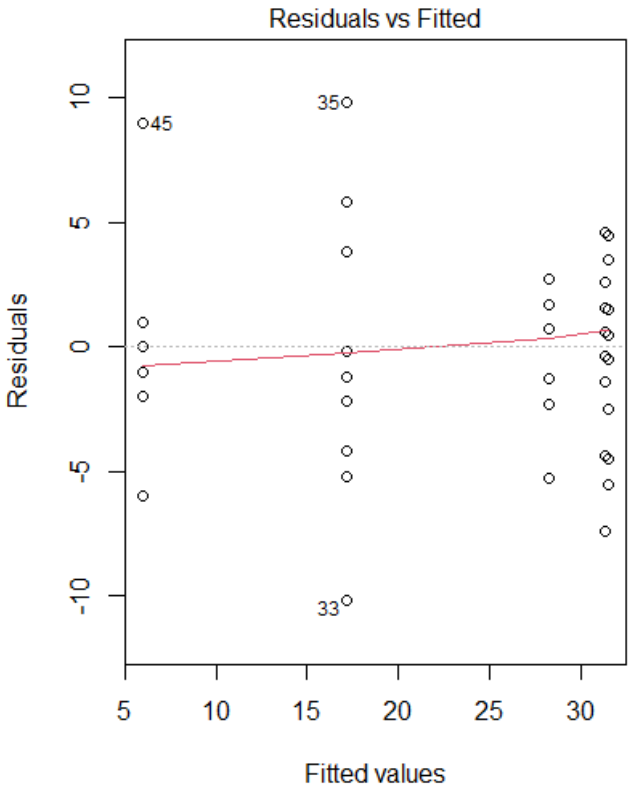
Aplicando el test en R se obtiene un p-value = 0.2987 (mayor a un α = 0.05) lo cual indica que no hay evidencia para decir que no es normal.

(2) Para verificar la condición de ***varianza homogénea (constante)***, se recomienda el análisis de los residuos de la siguiente manera

- d_i (residuo estandarizado) vs concentraciones
- $e_{ij}\ Vs\ \hat{y}_i$

Las siguientes figuras nos ayudan a detectar comportamientos que se separen de los supuestos formulados.





De las gráficas anteriores de puede notar que el residuo estandarizado no sale del intervalo (-2,2) por lo cual se puede tomar la varianza constante.

En conclusión dado que se cumple la condición de Normalidad y la varianza es homogénea (constante) es Válido utilizar el Método de Análisis de Varianza.