


Primera parte: Diferencia de medias de poblaciones normales para muestras independientes

Caso 1: muestras normales independientes, varianzas conocidas

*Importante!! En esta unidad **no trabajamos** con técnicas descriptivas o exploratorias, trabajamos con pruebas de hipótesis. Por ello, el cumplimiento de los supuestos es fundamental para la validez de la prueba!!!*

Se sospecha que el **pH medido en la superficie del suelo de la Región 1 es diferente al de la Región 2**. Un geólogo determina electromecánicamente el pH, en la superficie del suelo de 20 puntos elegidos al azar, de cada una de las dos regiones de interés.

Se supone que el pH en la superficie del suelo de las regiones 1 y 2 se distribuye normalmente, con varianzas 0,85 y 1,22 respectivamente. Los resultados muestrales fueron:

	Región 1	Región 2	
Media	6,58	5,74	
designamos	X: "pH de un punto en la superficie del suelo de la región 1"	Y: "pH de de un punto en la superficie del suelo de la región 2"	

Entonces **nuestro modelo** puede definirse como dos muestras normales independientes:

X_1, X_2, \dots, X_{n1} con $X \sim N(\mu_1, \sigma_1^2)$ con σ_1^2 conocida

Y_1, Y_2, \dots, Y_{n2} con $Y \sim N(\mu_2, \sigma_2^2)$ con σ_2^2 conocida

Estamos interesados en realizar inferencias acerca del parámetro **diferencia de medias** de las dos poblaciones: $\mu_1 - \mu_2$.

Un estimador puntual insesgado para este parámetro es $\bar{X} - \bar{Y}$

Como ambas poblaciones son normales con varianzas conocidas, sabemos que

$$\bar{X} \sim N\left(\mu_1; \frac{\sigma_1^2}{n_1}\right) \quad \text{e} \quad \bar{Y} \sim N\left(\mu_2; \frac{\sigma_2^2}{n_2}\right) \quad \text{independientes}$$

Entonces:

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Y por lo tanto, estandarizando el estimador puntual, tendremos una distribución conocida y tabulada:

$$z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Las hipótesis para testear en nuestro ejemplo serán entonces:

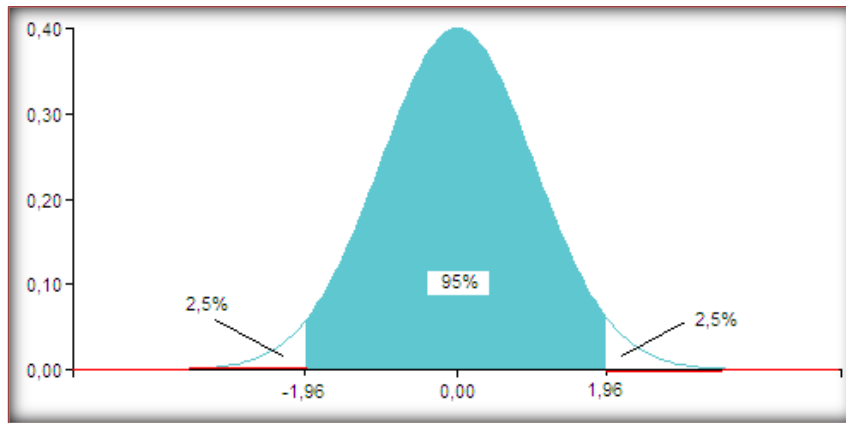
$$\mathbf{H_0:} \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad \mathbf{H_1:} \mu_1 - \mu_2 \neq 0$$

Como se trata de una hipótesis alternativa bilateral; es decir que valores muy grandes o muy pequeños del estadístico de contraste nos conducirán a rechazar la hipótesis de nulidad.

Si establecemos un nivel de significación $\alpha = 0,05$, es decir una probabilidad máxima de rechazar H_0 siendo ésta cierta del 0,05, la región de rechazo será:

$$z_{obs} \geq 1,96 \text{ o } z_{obs} \leq -1,96$$

Es decir que la decisión será rechazar H_0 si $z_{obs} \geq 1,96$ o bien si $z_{obs} \leq -1,96$



El valor de la variable pivotal o estadístico de contraste bajo la hipótesis nula para este ejemplo es:

$$z_{obs} = \frac{6,58 - 5,74 - (0)}{\sqrt{\frac{0,85}{20} + \frac{1,22}{20}}} = 2,61$$

Como $2,61 > 1,96$ entonces la **decisión** es rechazar H_0 , es decir que existe evidencia empírica en contra de la hipótesis de que las medias poblacionales de los pH de los suelos de las dos regiones son iguales, con un nivel de significación del 5%.

Puede ser de interés cuantificar la fuerza del rechazo de la hipótesis nula, o bien la probabilidad de encontrar un valor tan extremo o más que el hallado en esta muestra siendo cierta la hipótesis nula, a esta probabilidad se la denomina **p valor** y en nuestro ejemplo es:

$$\mathbf{p.valor = 2 * P(z > 2,61) = 2 * 0,00453 = 0,00906}$$

Este valor es pequeño, lo que nos indica que tenemos bastante seguridad en la decisión, ya que es muy poco probable que siendo cierta H_0 , nos encontremos con un par de muestras con estos valores medios.

Si quisiéramos construir un **intervalo de confianza** de nivel $1-\alpha$ para la diferencia de las medias de pH de estas dos poblaciones deberíamos utilizar la información:

$$P \left(z_{\alpha/2} < \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{1-\alpha/2} \right) = 1 - \alpha$$

Despejando la diferencia de los parámetros a estimar que resulta ser un nuevo parámetro, tenemos la expresión del **intervalo de confianza** de nivel $1-\alpha$ para la **diferencia de medias de poblaciones normales** con muestras independientes y varianzas conocidas:

$$\left[\bar{X} - \bar{Y} + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}; \bar{X} - \bar{Y} + z_{1-\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

Recordemos que un intervalo de confianza de nivel 95% para un parámetro se interpreta como que, de cada 100 intervalos contruidos a partir de muestras de igual tamaño, alrededor de 95 cubrirán el valor verdadero del parámetro, (diferencia de medias poblacionales de pH de las dos regiones).

En nuestro ejemplo este intervalo de nivel 95% para la diferencia de medias de pH resulta:


$$[6,58-5,74 + /- 1,96 \sqrt{\frac{0,85}{20} + \frac{1,22}{20}}] = [0,84-0.63; 0,84+0.63] = [0.21; 1,47]$$

Se debe interpretar que con una confianza del 95% que el intervalo [0.71; 1.97] cubra el valor verdadero de la diferencia entre los valores medios de pH de las regiones 1 y 2.

Como ambos extremos son positivos, el test basado en el intervalo de confianza también rechaza la hipótesis de igualdad y se puede pensar que el valor medio poblacional de pH de la región 1 es superior al valor medio poblacional del pH de la región 2.

Caso 2: muestras normales -varianzas desconocidas pero supuestamente iguales

El tiempo que le toma a la habichuela en duplicar su peso es una medida de su calidad para enlatar. Un experimento con 15 repeticiones independientes de cada una de dos variedades produjo los resultados siguientes:

Variedad A	Variedad B	
$\bar{X} = 17,2$ horas	$\bar{Y} = 18,3$ horas	
$s_x = 0,7$ horas	$s_y = 0,8$ horas	

Interesa decidir si la calidad de la variedad B es superior a la calidad de la variedad A, utilizando para probar estas hipótesis un nivel de significación del 0,01.

Si podemos asegurar que ambas muestras provienen de distribuciones normales con varianza común entonces **nuestro modelo** puede definirse como dos muestras normales independientes:

$$X_1, X_2, \dots, X_{n1} \text{ con } E(X_i) = \mu_1, \text{ y } V(X_i) = \sigma^2$$

con σ^2 desconocida

$$Y_1, Y_2, \dots, Y_{n2} \text{ con } E(Y_i) = \mu_2, \text{ y } V(Y_i) = \sigma^2$$

Las hipótesis de interés para este caso son:

$$H_0: \mu_1 - \mu_2 \geq 0 \quad \text{vs.} \quad H_1: \mu_1 - \mu_2 < 0$$

o equivalentemente

$$(H_0: \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1: \mu_1 - \mu_2 < 0)$$

Se trata de una prueba unilateral izquierda, es decir que rechazaremos valores bajos del estadístico de contraste.

Estamos nuevamente interesados en realizar inferencias acerca del parámetro **diferencia de medias** poblacionales de las dos poblaciones: $\mu_1 - \mu_2$.

$$\text{Sabemos que: } \bar{X} \sim N\left(\mu_1; \frac{\sigma^2}{n_1}\right) \quad \text{y} \quad \bar{Y} \sim N\left(\mu_2; \frac{\sigma^2}{n_2}\right)$$

$$\text{Luego: } \bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2; \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right) = N\left(\mu_1 - \mu_2; \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Entonces estandarizando el estimador puntual propuesto, obtenemos la expresión de la variable pivotal:

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Como la varianza es común a las dos poblaciones, tiene sentido construir un estimador insesgado de la varianza común, basado en ambas muestras, que usualmente se conoce como varianza amalgamada(o pooleada):

$$s_a^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\text{Además: } \frac{(n_1 - 1)s_1^2}{\sigma^2} \sim \chi_{n_1 - 1}^2 \quad \text{y} \quad \frac{(n_2 - 1)s_2^2}{\sigma^2} \sim \chi_{n_2 - 1}^2 \quad \text{independientes}$$

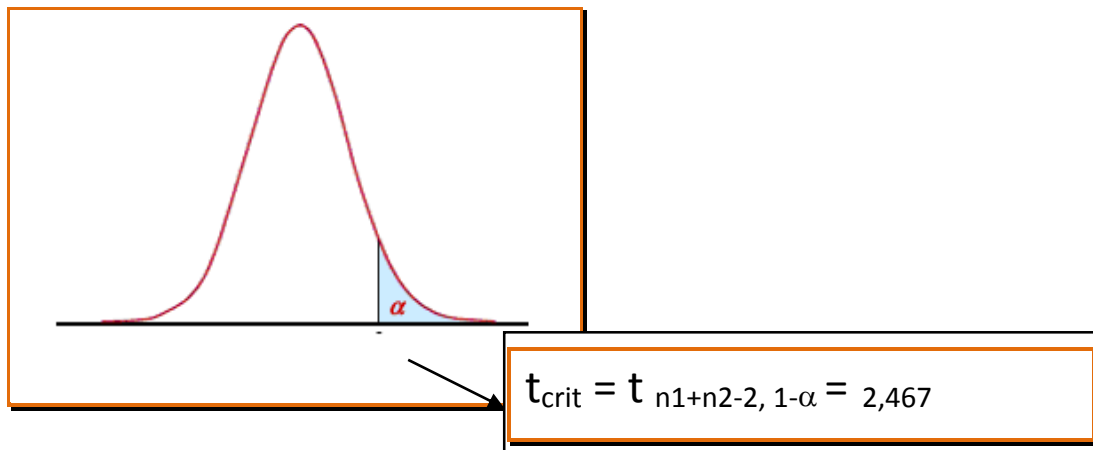
Y la suma de variables aleatorias Chi cuadrado independientes es otra variable aleatoria Chi cuadrado cuyos grados de libertad son la suma de los grados de libertad de las variables sumadas.

Entonces:
$$U = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{\sigma^2} \sim \chi_{n_1 + n_2 - 2}^2$$

Además Z y U son independientes, entonces nuestra variable pivotal puede ser:

$$t = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{s_a^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1 + n_2 - 2}$$

Y la región de rechazo del test de nivel 0,01 es $t_{\text{obs}} > t_{28; 0,99} = 2,467$



El valor observado del estadístico de contraste en nuestro ejemplo es:

$$t_{\text{obs}} = \frac{18.3 - 17.2}{\sqrt{\frac{14 * 0.8^2 + 14 * 0.7^2}{28} \left(\frac{1}{15} + \frac{1}{15} \right)}} = 4.008$$

Como $4,008 > 2,467$ **rechazamos la hipótesis de nulidad** con un nivel de significación del 1%, lo cual significa que hay evidencia en contra de la hipótesis nula que sostiene que el valor medio poblacional del tiempo que tarda la variedad A de habichuelas en duplicar su tamaño es igual o mayor que el tiempo medio poblacional que tarda la variedad B.

Para construir un intervalo de confianza de nivel $1-\alpha$ para la diferencia de medias de poblaciones normales provenientes de muestras independientes, con varianzas desconocidas, pero que pueden suponerse iguales, debemos utilizar la variable pivotal de la expresión:

$$P\left(t_{n_1+n_2-2, \alpha/2} < \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < t_{n_1+n_2-2, 1-\alpha/2}\right) = 1-\alpha$$

Despejando de esta expresión el parámetro de interés obtenemos el intervalo de confianza de nivel $1-\alpha$ para la diferencia de medias de poblaciones independientes con varianzas desconocidas pero iguales, resulta:

$$\left[\bar{X} - \bar{Y} + t_{n_1+n_2-2, \alpha/2} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}; \bar{X} - \bar{Y} + t_{n_1+n_2-2, 1-\alpha/2} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \right]$$

Para nuestro ejemplo el intervalo de confianza para la diferencia de medias de nivel 99% es:

$$\left[17.2 - 18.3 + / - 2.763 * \sqrt{\frac{14 * 0.7^2 + 14 * 0.8^2}{28} \left(\frac{1}{15} + \frac{1}{15}\right)} \right] = [-1,85835; -0,34164]$$

Este intervalo tiene ambos extremos negativos lo que indica que la estimación de la diferencia de medias poblacionales es negativa, vale decir que la media poblacional de la región 1 es superior a la de la región 2.

Observación: si las varianzas de las dos poblaciones no pueden suponerse iguales, no puede utilizarse la prueba del caso 2, el problema lo resolvieron Fisher & Behrems que pusieron su nombre a la prueba y es también una prueba basada en la distribución t, pero con grados de libertad calculados mediante la siguiente formula:

$$w = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2 - 1}}$$

Algunos paquetes estadísticos hacen previamente la prueba de homogeneidad de varianzas y luego deciden cuál de los dos casos debe aplicarse e informan el resultado sin decir qué prueba han aplicado para obtenerlo.

Otros paquetes hacen la prueba de varianzas y dan las dos pruebas posibles para que el usuario decida cuál toma.

Esta decisión es importante porque en alguno de los test podríamos estar rechazando y en el otro no!!.


Caso 3: muestras de poblaciones cualesquiera

Si las muestras son suficientemente grandes, es posible aplicar una prueba asintótica, basada en el Teorema del Límite Central.

Este es el caso más usual para data mining donde se dispone de mucha información y en general la información no satisface el supuesto de normalidad.

Si se desea aplicar una prueba que suponga normalidad y los datos disponibles no la satisfacen una alternativa viable es aplicar transformaciones de Box & Cox o de transformaciones de Jhonson para normalizar los datos y que dichos test sean válidos.

Los datos siguientes corresponden a una muestra aleatoria de varones y otra de mujeres, estudiantes universitarios cuyas edades oscilan entre 20 y 30 años, que hacen algún tipo de actividad física a los que se les preguntó sobre el número promedio de horas semanales que dedican a este tipo de actividades (deportes, gimnasia, caminata, etc.).

		Varones	Mujeres
	Número de observaciones	124	110
	Media muestral	6.6	5.4
	Desvío estándar muestral	4.3	3.6

Queremos testear la hipótesis de que la cantidad de horas semanales dedicadas a la actividad deportiva es la misma para ambos grupos definidos por la variable sexo con un nivel de significación del 1%.

Luego las hipótesis de interés en este caso son:

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1: \mu_1 - \mu_2 \neq 0$$

Donde :

μ_1 = media del tiempo semanal dedicado a la gimnasia por los hombres

μ_2 = media del tiempo semanal dedicado a la gimnasia por las mujeres

En nuestro modelo tenemos dos muestras aleatorias independientes de tamaños grandes ($n_i > 30$) cuya distribución no podemos garantizar que sea normal. Simbólicamente.

$$\begin{array}{ll} X_1, X_2, \dots, X_{n_1} & \text{con} \quad E(X_i) = \mu_1 \quad \text{y} \quad V(X_i) = \sigma_1^2 \\ Y_1, Y_2, \dots, Y_{n_2} & \text{con} \quad E(Y_j) = \mu_2 \quad \text{y} \quad V(Y_j) = \sigma_2^2 \end{array} \quad \text{Independientes}$$

Sabemos por el teorema central del límite que:

$$\frac{\bar{X} - \mu_1}{\sigma_1 / \sqrt{n_1}} \approx N(0,1) \quad \text{y que} \quad \frac{\bar{Y} - \mu_2}{\sigma_2 / \sqrt{n_2}} \approx N(0,1)$$

Y como desconocemos los valores de las varianzas poblacionales podemos utilizar la propiedad:

$$\frac{\sigma_1}{s_1} \rightarrow 1 \quad \text{y que} \quad \frac{\sigma_2}{s_2} \rightarrow 1 \quad \text{conforme } n \rightarrow \infty$$

Luego:

$$\frac{\bar{X} - \mu_1}{s_1 / \sqrt{n_1}} \approx N(0,1) \quad \text{y que} \quad \frac{\bar{Y} - \mu_2}{s_2 / \sqrt{n_2}} \approx N(0,1)$$

Y entonces:

$$Z^{obs} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx N(0,1)$$

Al tratarse de una hipótesis bilateral, rechazaremos valores muy altos o muy bajos de estadístico de contraste.

La región de rechazo del test para un nivel de significación del 1% será: $z^{obs} > z_{0.995}$ o $z^{obs} < z_{0.005}$ es decir que rechazaremos valores del estadístico de contraste superiores a 2.58 o inferiores a -2.58.

El valor del estadístico del test observado para nuestros datos es:

$$Z^{obs} = \frac{(\bar{X} - \bar{Y}) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(6.6 - 5.4) - 0}{\sqrt{\frac{(4.3)^2}{124} + \frac{(3.6)^2}{110}}} = \frac{1.2}{0.516} = 2.32$$

Por lo tanto la decisión es no rechazar H_0 , vale decir que no existe evidencia en contra de que los tiempos destinados semanalmente a la actividad

deportiva sean iguales en ambos sexos. **Concluimos que la media del tiempo semanal que dedican a actividades físicas los varones no es significativamente diferente que el que dedican las mujeres.**

¿Qué ocurre si deseamos ahora comparar las medias de varios grupos??

Si hacemos la comparación de a pares, es decir si tenemos n grupos, deberíamos realizar $k(k-1)/2$ contraste y si tomamos un nivel de significación α que es bien habitual, la probabilidad de no cometer error de tipo I en todos ellos, es decir el nivel de significación global es: $1-(1-\alpha)^n$.

Si el nivel de significación fuera por ejemplo del 0.05 y la cantidad de grupos $k=5$ el nivel global es $1-(1-0.05)^5 = 0.2262$

Es decir que la probabilidad de cometer algún error **crece notablemente!!!**


Una mejor respuesta para este problema desarrollada por Fisher es comparar las medias de tres o más poblaciones con distribuciones normales de igual varianza y observamos muestras independientes para cada población, el análisis que corresponde es el que desarrollaremos a continuación y se denomina análisis de la varianza.(ADEVA) o analysis of variance (ANOVA)

Análisis de la varianza de un factor

El té es la bebida más usual en el mundo entero después del agua, actualmente se ha difundido mucho el consumo del té verde. La folacina es la **única vitamina B presente en cualquier cantidad importante de té**, y recientes avances en métodos de ensayo han determinado de manera más precisa el contenido de folacina.

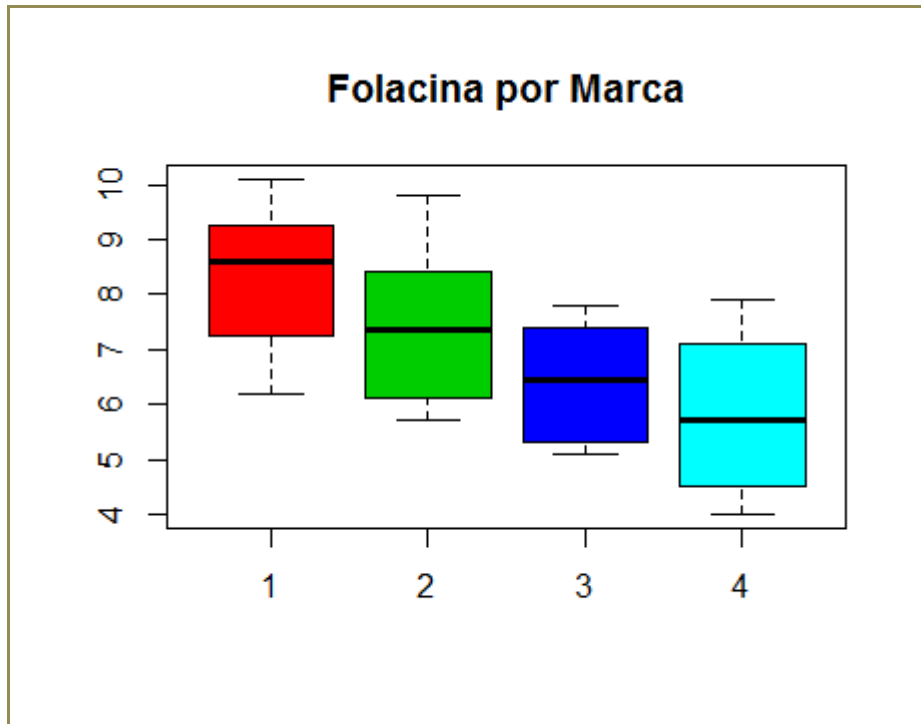
Ejemplo 1

Consideremos los siguientes datos acerca del contenido de folacina para especímenes recogidos al azar de las cuatro marcas principales de té verde muy conocidas en el mercado argentino.



	Marca 1	Marca 2	Marca 3	Marca 4
	7.9	5.7	6.8	6.4
	6.2	7.5	7.8	7.1
	6.6	9.8	5.1	7.9
	8.6	6.1	7.4	4.5
	8.9	8.4	5.3	5.0
	10.1	7.2	6.1	4
	9.6			
Media	8.2714	7.4500	6.4167	5.8167
Desvío Standard	1.4625	1.5083	1.1053	1.5510

- a) Analicemos gráficamente si se observan diferencias importantes entre los contenidos medios de folacina en las distintas marcas.



En este gráfico se aprecian diferencias en los valores medianos de las distribuciones del contenido porcentual de folacina en estas tres marcas.

Lo que deberemos investigar es si estas diferencias que apreciamos visualmente en el boxplot comparativo son estadísticamente significativas o no.

Para generalizar supongamos el siguiente **modelo**: observamos **k** muestras:

k muestras normales independientes con varianzas iguales.

Muestra 1: $X_{11}, X_{12}, \dots, X_{1n_1}$ vs.as. i.i.d. $N(\mu_1, \sigma^2)$

.....

Muestra i : $X_{i1}, X_{i2}, \dots, X_{i n_i}$ vs.as. i.i.d. $N(\mu_i, \sigma^2)$

.....

Muestra k : $X_{k1}, X_{k2}, \dots, X_{k n_k}$ vs.as. i.i.d. $N(\mu_k, \sigma^2)$

Las variables aleatorias observadas son independientes entre sí dentro de las muestras y entre las muestras.

Este es un supuesto bastante fuerte que si no se satisface habrá que realizar una transformación de los datos o aplicar técnicas no paramétricas que no suponen homocedasticidad ni normalidad.

Si las transformaciones disponibles no son efectivas para que los supuestos se satisfagan, veremos luego que una alternativa interesante es el test de **Kruskal Wallis**.

Denotemos con:

\bar{X}_i y s_i^2 a la media y la varianza de la muestra i (para $i = 1, 2, \dots, k$)

Parece natural que el estimador de σ^2 se obtenga calculando un promedio ponderado de las varianzas de cada muestra s_i^2 (es una generalización de la idea de la varianza amalgamada o pooleada). Se puede demostrar que el mejor estimador insesgado de σ^2 bajo el modelo anterior es:

$$S_p^2 = SSW / (n - k) = \frac{(n_1 - 1) * s_1^2 + \dots + (n_k - 1) * s_k^2}{n_1 + \dots + n_k - k} = \frac{\sum_{i=1}^k (n_i - 1) * s_i^2}{n - k}$$

SSW (sum squares within) suma de cuadrados dentro de los grupos

En la última expresión hemos denotado con $n = \sum_{i=1}^k n_i$ al número total de observaciones.

Vamos a estudiar la hipótesis nula:

$$H_0 = \mu_1 = \mu_2 = \dots = \mu_k$$

Llamemos $\bar{X}_{..} = \frac{\sum_{i=1}^k n_i \bar{X}_i}{n} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}}{n}$ a la media general de todas las observaciones.

El estadístico para el test óptimo para este problema, tiene al estimador de la varianza en el denominador y una medida de las diferencias (similar a la varianza) entre las medias de las distintas muestras en el numerador. Esta medida es:

$$SSB = \frac{\sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2}{k-1}$$

SSB(sum squares between) suma de cuadrados entre los grupos

El estadístico del test se obtiene dividiendo (SSB) sobre (SSW):

$$F^{obs} = \frac{\left(\sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2 \right) / (k-1)}{s_p^2}$$

Para decidir si las medias son iguales en las distintas subpoblaciones o no, debemos aplicar un test F.

El estadístico F es el cociente de dos variables Chi cuadrado, es decir el cociente de dos varianzas. Como las variables Chi cuadrado son positivas, la variable F asume solamente valores positivos también. Para tomar la decisión:

1er paso establecemos la hipótesis de nulidad y la alternativa

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

H1: existe al menos un par i, j tal que $\mu_i \neq \mu_j$

2do paso: calculamos el estadístico F^{obs} definido; este estadístico tiene distribución $F_{k-1, n-k}$ que son los grados de libertad de los dos estimadores de la varianza, que son dos variables Chi cuadrado del numerador y del denominador.

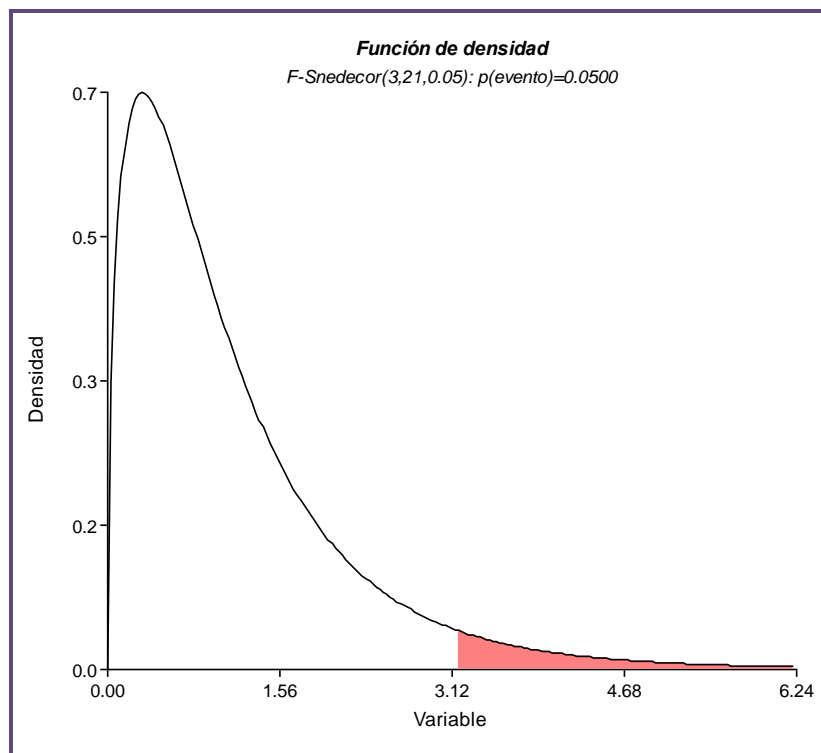
3^{er} paso: Se decide con la siguiente regla: $F > F_{k-1, n-k, \alpha}$ rechazo H_0 .

Por qué rechazamos valores grandes del estadístico? O bien por qué se trata de una prueba unilateral derecha?

Porque estamos comparando dos estimadores de la misma varianza, en el numerador utilizamos las diferencias entre las medias de los grupos y la media general mientras que en el denominador amalgamamos las varianzas estimadas para cada subgrupo.

Si el numerador es mucho más grande que el denominador indica que las medias son muy distintas entre sí.

En el siguiente grafico señalamos la región de rechazo de la prueba:



Suponiendo en primera instancia que se verifican los supuestos del modelo del Análisis de la Varianza. Construimos la tabla y aplicamos el test F para decidir si existen diferencias entre las medias del contenido de folacina de las distintas marcas a nivel 0.05.

En R definimos el contraste:

```
te.aov<-aov(te$folac~te$marca)
summary(te.aov)
```

	Df	Sum of Sq	Mean Sq	F	p value	Pr(>F)
te\$marca	3	22.93	7.645	3.791	0.0256	**
Residuals	21	42.35	2.016			

En esta tabla el test F rechaza la igualdad de varianzas a nivel 0.05

Ahora, antes de decidir, debemos estudiar si los supuestos del contraste se satisfacen para ver si la conclusión es válida.

Para ello se realiza el **diagnóstico del modelo**.

Diagnóstico del modelo: Para que este test F sea válido el modelo de k muestras normales independientes con varianzas iguales tiene que ser aproximadamente cierto. Al igual que con el test t , hay que observar los datos para detectar si hay alguna razón para pensar que este modelo es falso.

Analicemos en primera instancia el supuesto de homocedasticidad (igualdad de varianzas de los grupos).

En este ejemplo los diagramas de caja aparecen a diferentes alturas pero el tamaño de las cajas se ve muy similar, así que desde este grafico no hay motivos para sospechar que no se cumple el supuesto de homocedasticidad.

Para realizar el análisis cuantitativo, existen diferentes pruebas alternativas para esta hipótesis.

1- Test de Bartlett

$$H_0 \sigma^2_1 = \sigma^2_2 = \dots = \sigma^2_k$$

H_1 : existe al menos un par i, j tal que $\sigma^2_i \neq \sigma^2_j$

```
bartlett.test(te$folac,te$marca)
```

```
Bartlett test of homogeneity of variances
```

```
data: te$folac and te$marca
```

```
Bartlett's K-squared = 0.6168, df = 3, p-value = 0.8926
```

El test de Bartlett, no rechaza la hipótesis de nulidad, es decir que no hay evidencia de que la varianza de alguno de los subgrupos difiera de las otras.

El problema de este test es su **sensibilidad a la falta de normalidad**.

Esto implica que puede que rechace la hipótesis nula por no cumplirse el supuesto de normalidad en lugar de rechazarla por no cumplirse el supuesto de homocedasticidad.

Una alternativa más robusta, es decir que no es sensible a la falta de normalidad o a la presencia de algún valor atípico la brinda el test de Levene.

2- Test de Levene

```
library(Rcmdr)
```

```
library(reshape2)
```

```
library(car)
```

```
marca<-as.factor(te$marca)
```

```
leveneTest(te$folac~marca)
```

```
Levene's Test for Homogeneity of Variance (center = median)
```

```
  Df F value Pr(>F)
```

```
group 3  0.2949 0.8286
```

```
21
```

El test de Levene realiza un nuevo análisis de la varianza para los valores absolutos de los residuos de las observaciones respecto de la mediana de su grupo.

En nuestro ejemplo no rechaza la hipótesis nula que es la misma que la del test de Bartlett.

Es decir que podemos suponer que se cumple la hipótesis de homocedasticidad.

Falta analizar el cumplimiento del supuesto de normalidad de la distribución de los residuos(o de la variable, son equivalentes).

Para ello disponemos de test de normalidad y de un gráfico que compara los cuantiles empíricos con los esperados bajo el supuesto de normalidad. Este grafico se denomina QQplot.(gráfico de cuantil-cuantil).

R tiene implementada una batería de test de normalidad en la librería nortest. Dos muy conocidos y potentes son el test de Shapiro Wilk y el test de Anderson Darling.

```
library(nortest)
shapiro.test(residuals(te.aov))
Shapiro-Wilk normality test
data: residuals(te.aov)
W = 0.9518, p-value = 0.2747
```

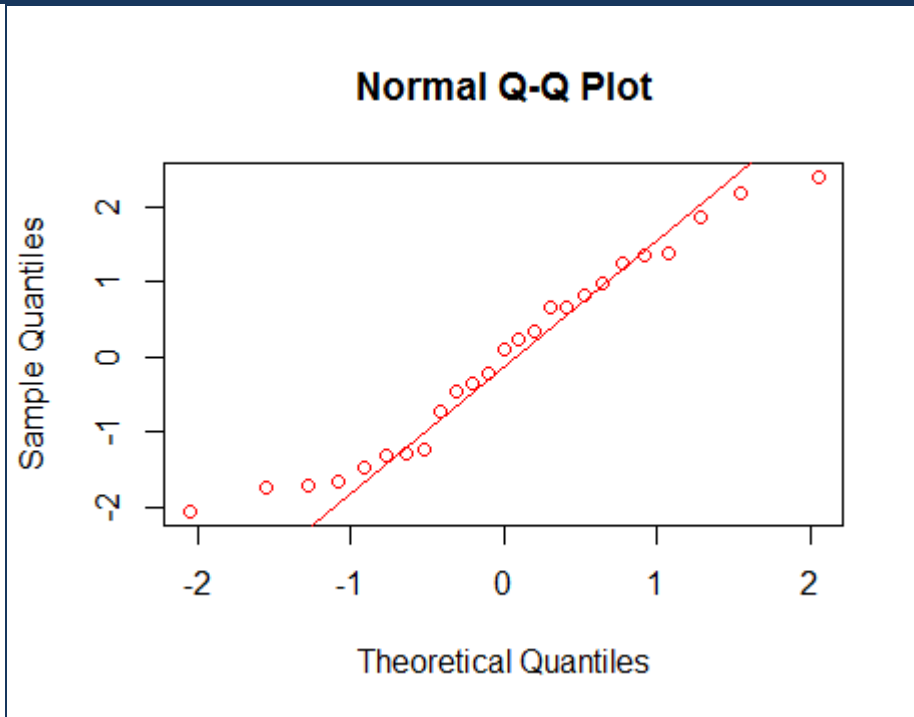
```
Anderson-Darling normality test
data: residuals(te.aov)
A = 0.3751, p-value = 0.3873
ad.test(residuals(te.aov))
```

```
library(moments)
agostino.test(residuals(te.aov))
D'Agostino skewness test
data: residuals(te.aov)
skew = 0.1108, z = 0.1781, p-value = 0.8587
alternative hypothesis: data have a skewness
```

Gráficos de cuantil- cuantil

```
qqnorm(residuals(te.aov),col=2)
```

```
qqline(residuals(te.aov),col=2)
```



Si los datos dibujados fueran normales, deberían posicionarse sobre la recta dibujada. Esto no ocurrirá nunca en la realidad, dado que se trata de una muestra aleatoria. Lo que debemos observar es si los puntos se alejan mucho de la recta o no.

En base a las salidas de R no existe evidencia empírica en contra de la normalidad de la distribución de la variable(o de los residuos).

Luego podemos dar por valido el rechazo de la hipótesis de nulidad del test F de análisis de la varianza.

¿Qué hacemos si no son normales o no resultan heterocedásticos?



Una primera opción es transformar los datos para que resulten normales y/o homocedásticos. En este contexto disponemos de los siguientes lineamientos para elegir la transformación más adecuada.

- Si la media y la dispersión están relacionadas de modo tal que a mayor media, mayor dispersión y /o la distribución presenta asimetría positiva. → **Transformación logarítmica**.
- Para variables con distribución Poisson → **Transformación $\sqrt{x} + 1$** .
- Para variables con distribución binomial o porcentajes → **Transformación arco seno**.

También existen **transformaciones de potencia o de Box y Cox**.

Analizando la relación entre la variabilidad y la media de los grupos, teniendo en cuenta la siguiente información (transformación de Box-Cox):

Relación	Transformación	y^p	p
$\sigma \propto \mu$	ninguna	y^1	1
$\sigma \propto \sqrt{\mu}$	raíz	$y^{0,5}$	0,5
$\sigma \propto \mu$	logaritmo	$\log(y)$	0
$\sigma \propto \mu^2$	recíproca	y^{-1}	-1

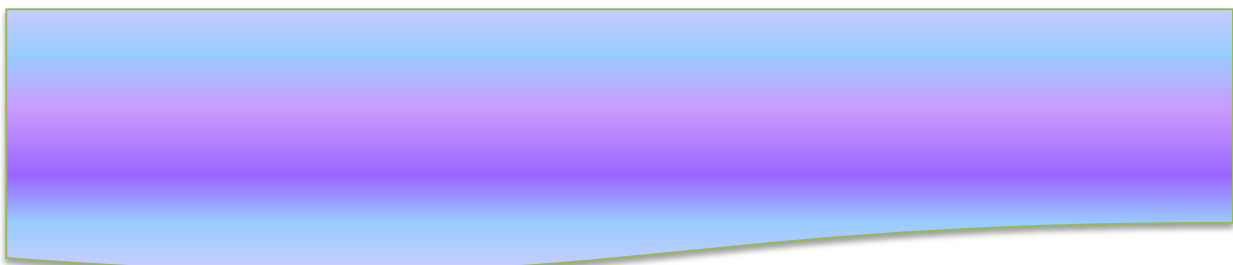
Encontramos
evidencia

en contra de la
igualdad

de contenido
medio de
folacina

en las
diferentes
marcas

de té...



Cabe entonces preguntarnos ahora... cuál/es son las medias diferentes

Cuando no se puede rechazar H_0 generalmente el análisis termina ahí, pero cuando se rechaza resulta lógico que el experimentador no se conforme con esa respuesta, sino que desee comparar las medias, frecuentemente de a pares en general y de algunas otras formas en casos específicos.

Intervalo de confianza para la diferencia de dos medias

Queremos comparar las medias de los grupos i y i^* .

A estas comparaciones se las conoce como **comparaciones a posteriori**.

Empecemos por construir un IC para μ_i y μ_{i^*} .

El estimador puntual es $\bar{Y}_i - \bar{Y}_{i^*}$. ¿Cuál es su varianza? ¿Cómo se estima?

Puede demostrarse que:

$$\left[\bar{Y}_i - \bar{Y}_{i^*} - t_{n-k, \alpha/2} \sqrt{s_p^2 \left(\frac{1}{n_i} + \frac{1}{n_{i^*}} \right)}; \bar{Y}_i - \bar{Y}_{i^*} + t_{n-k, \alpha/2} \sqrt{s_p^2 \left(\frac{1}{n_i} + \frac{1}{n_{i^*}} \right)} \right]$$

es un IC con nivel $1 - \alpha$.

De este intervalo podemos deducir un test para estudiar las hipótesis $H_0: \mu_i = \mu_{i^*}$ vs $H1: \mu_i \neq \mu_{i^*}$

El problema de este intervalo o de este test es que tiene nivel $1 - \alpha$ para una comparación de un par... pero deja de tener este nivel cuando queremos comparar muchos pares, como analizamos en la introducción del análisis de la varianza.

Por eso cuando uno planea de antemano hacer uno o muy pocos intervalos o test puede usar intervalos de a pares elevando adecuadamente el nivel de confianza de los mismos, pero en caso contrario conviene utilizar un método de intervalos de confianza simultáneos.

En este sentido se dispone de muchas alternativas de comparación de nivel global; como las de **Dunnet**, **Newman – Keuls**, **Tuckey o LSD** (least significative difference).

Todos los procedimientos involucran el cálculo de un valor que es comparado con las diferencias entre pares de promedios.

Si este valor es menor que la diferencia significa que esos dos grupo no son significativamente diferentes.

Tradicionalmente, las comparaciones múltiples se realizan al mismo nivel de significancia que el ANOVA. Por ejemplo, para un ANOVA significativo a un nivel de 5% ($\alpha = 0,05$), se realizan comparaciones múltiples al 5%. Sin embargo, esto puede variarse según la necesidad del estudio.

Algunos ejemplos los exhibimos en la siguiente tabla:

Prueba	Fórmula	Tabla
LSD	$T\sqrt{2 * SCE/n}$	Student
Dunnet	$D\sqrt{2 * SCE/n}$	D Dunnet
Tuckey HSD	$Q\sqrt{SCE/n}$	Q Tuckey
Newman Keuls	$Q_{\max}\sqrt{SCE/n}$	Q_{\max} Newman Keuls
Duncan	$Q_o\sqrt{SCE/n}$	Q_o Duncan
Sheffe	$s\sqrt{SCE/n}$	$S=\sqrt{(n-1)F}$

Básicamente todas las pruebas son mejoras de la prueba original de t de Student.

Dunnet se emplea cuando el interés es comparar todos los competidores contra uno original, por ejemplo todas la mejoras de tratamiento al tradicional.

LSD debe emplearse solo si se desean unas pocas comparaciones establecidas a priori(antes de realizar el análisis de la varianza).

Scheffe y Tuckey están diseñadas para comparar todos los pares de medias.

Ejemplo 2:

Badimon y cols (J.Clin.Invest 85: 1234, 1990) llevaron a cabo un estudio a fin de determinar el efecto de la fracción lipoproteica HDL-VHDL sobre lesiones ateroscleróticas en conejos. Para ello, 24 conejos Nueva Zelanda fueron asignados aleatoriamente y en forma balanceada a una de las siguientes dietas aterogénicas:

Dieta 1: 60 días de dieta rica en colesterol 0,5%

Dieta 2: 90 días de dieta rica en colesterol 0,5%

Dieta 3: 90 días de dieta rica en colesterol 0,5% y luego 30 días con 50 mg de fracción lipoproteica HDL-VHDL por semana.

Los animales fueron sacrificados. En todos los casos se comprobaron lesiones aterógenicas en aorta. Se midió el contenido de colesterol en aorta (en mg/g) con los siguientes resultados:

Dieta1	Dieta2	Dieta3
13,4	10,4	7,5
11,0	14,2	7,2
15,3	20,5	6,7
16,7	19,6	7,6
13,4	18,5	11,2
20,1	24,0	9,6
13,6	23,4	6,8
18,3	13,6	8,5


El modelo para este ejemplo es:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad 1 \leq i \leq 3 \quad 1 \leq j \leq 8$$

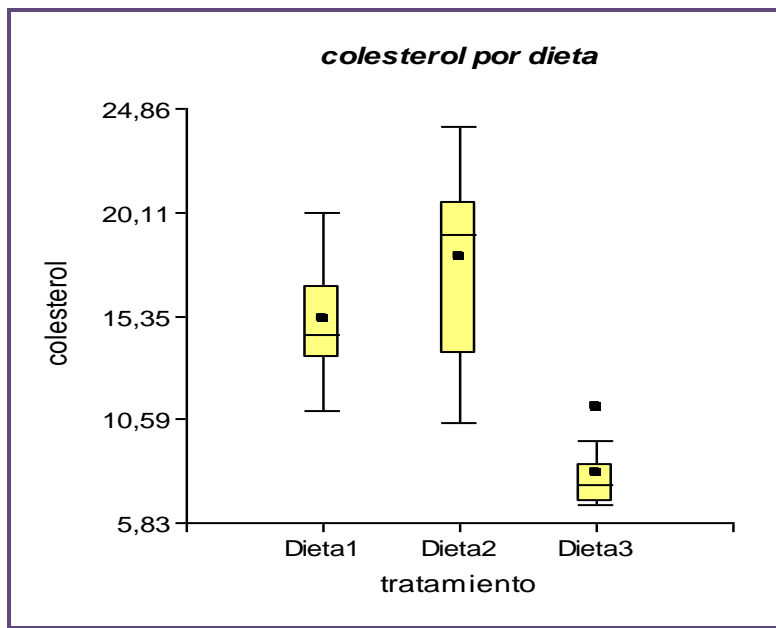
$$\text{Con } \varepsilon_{ij} \sim N(0, \sigma^2)$$

Calculamos la media y el desvío del contenido de colesterol en aorta correspondiente a cada una de las dietas.

Variable dependiente:colesterol

tratamiento	Media	Desviación típica	N	
Dieta1	15,225	2,9894	8	
Dieta2	18,025	4,8664	8	
Dieta3	8,138	1,5620	8	
Total	13,796	5,3608	24	

Graficamos un boxplot para apreciar gráficamente si hay diferencia entre los contenidos medios de colesterol en aorta de las dietas y también ver si existen outliers en las distribuciones o asimetrías y si tiene sentido pensar que tienen varianzas iguales.



Se aprecia en el grafico que las varianzas no son similares. También se aprecia la presencia de un outlier para la dieta 3. Para ver si estas sospechas tiene significación estadística ensayamos la prueba de Levene.

Contraste de Levene sobre la igualdad de las varianzas error^a(SPSS)

Variable dependiente:colesterol

F	gl1	gl2	Sig.
5,128	2	21	,015

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

No se satisface el supuesto de homocedasticidad!!!

Testeamos la normalidad, para ver si ese supuesto se satisface para este conjunto de datos:

Prueba de Kolmogorov-Smirnov para una muestra(SPSS)

		Residuo para colesterol
N		24
Parámetros	Media	,0000
normales ^{a,,b}	Desviación típica	3,26649
Diferencias más	Absoluta	,122
extremas	Positiva	,067
	Negativa	-,122
Z de Kolmogorov-Smirnov		,595
Sig. asintót. (bilateral)		,870

a. La distribución de contraste es la Normal.

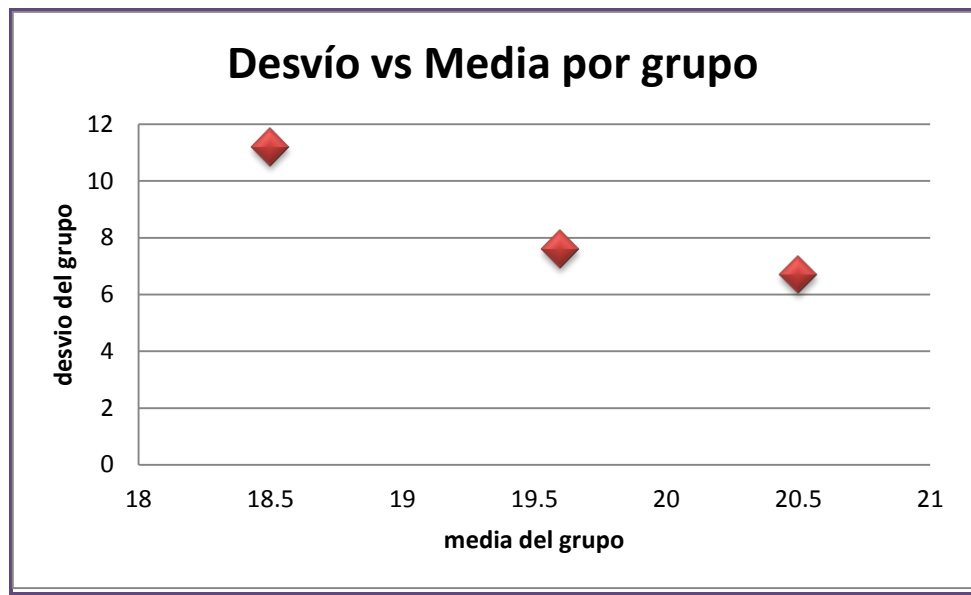
b. Se han calculado a partir de los datos.

Shapiro-Wilks (modificado) InfoStat

Variable	n	Media	D.E.	W*	p (una cola)
RDUO colesterol	24	0,00	3,27	0,97	0,8697

No hay motivo para rechazar el supuesto de normalidad.

Debemos transformar los datos para aplicar el análisis de la varianza. Para seleccionar la transformación adecuada, graficamos el desvío de cada grupo en función de la media del mismo.



La relación entre desvío y media parece lineal, debemos pensar entonces en la transformación logaritmo.

Transformamos los datos y vemos si con los datos transformados se satisfacen o no los supuestos.

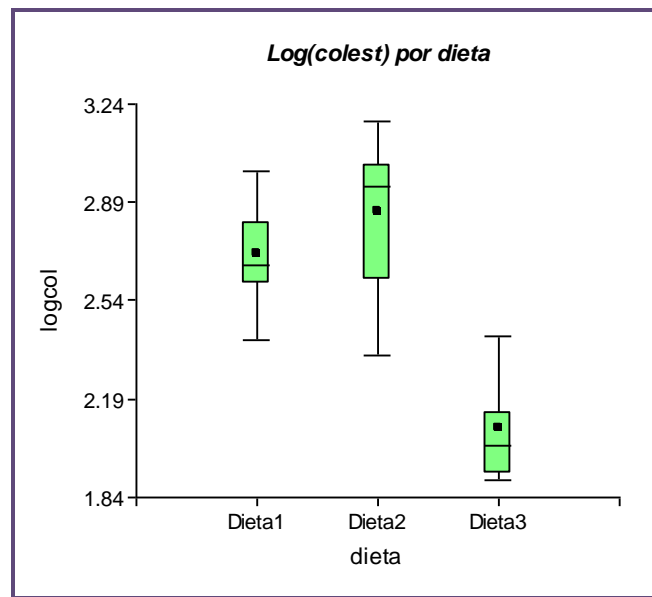
Análisis de la varianza (de la variable transformada)

Variable	N	R ²	R ² Aj	CV
logcol	24	0.71	0.68	8.97

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	2.7	2	1.35	25.83	<0.0001
dieta	2.7	2	1.35	25.83	<0.0001
Error	1.1	21	0.05		
Total	3.8	23			

La **conclusión** es que si se satisfacen los supuestos del modelo, las medias de los logaritmos del colesterol no son iguales para las tres dietas.

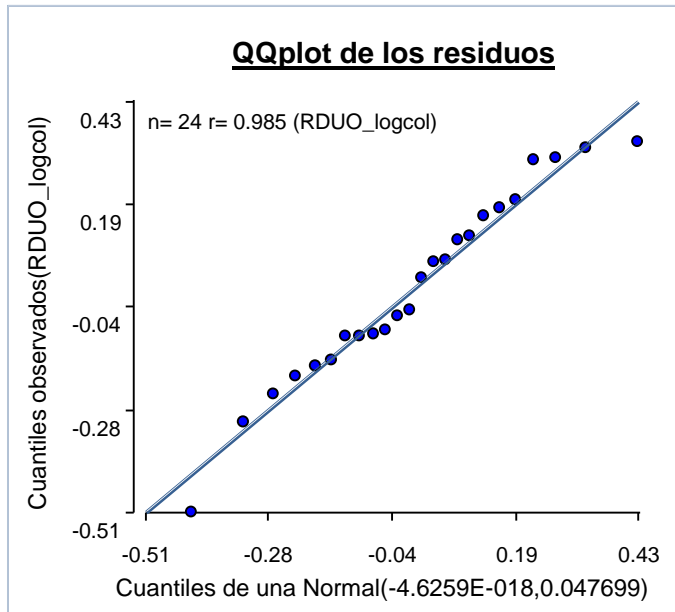


Diagnóstico del modelo con los datos transformados

Shapiro-Wilks (modificado)

Variable	n	Media	D.E.	W*	p(Unilateral D)
RDUO logcol	24	0.00	0.22	0.95	0.5468

No se rechaza la normalidad de los residuos del modelo con datos transformados.



Prueba de Levene (Análisis de la varianza sobre el valor absoluto de los residuos respecto de la mediana del grupo)

Variable	N	R ²	R ² Aj	CV
abs(res-med)	24	0.08	0.008	7.79

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	0.04	2	0.02	0.87	0.4334
dieta	0.04	2	0.02	0.87	0.4334
Error	0.47	21	0.02		
Total	0.51	23			

No se rechaza la hipótesis de homocedasticidad de los residuos.

Luego, la conclusión extraída es válida!!.

Nos preguntamos por último cuáles tratamientos difieren entre sí. Utilizamos los intervalos de confianza simultáneos para las diferencias de medias de

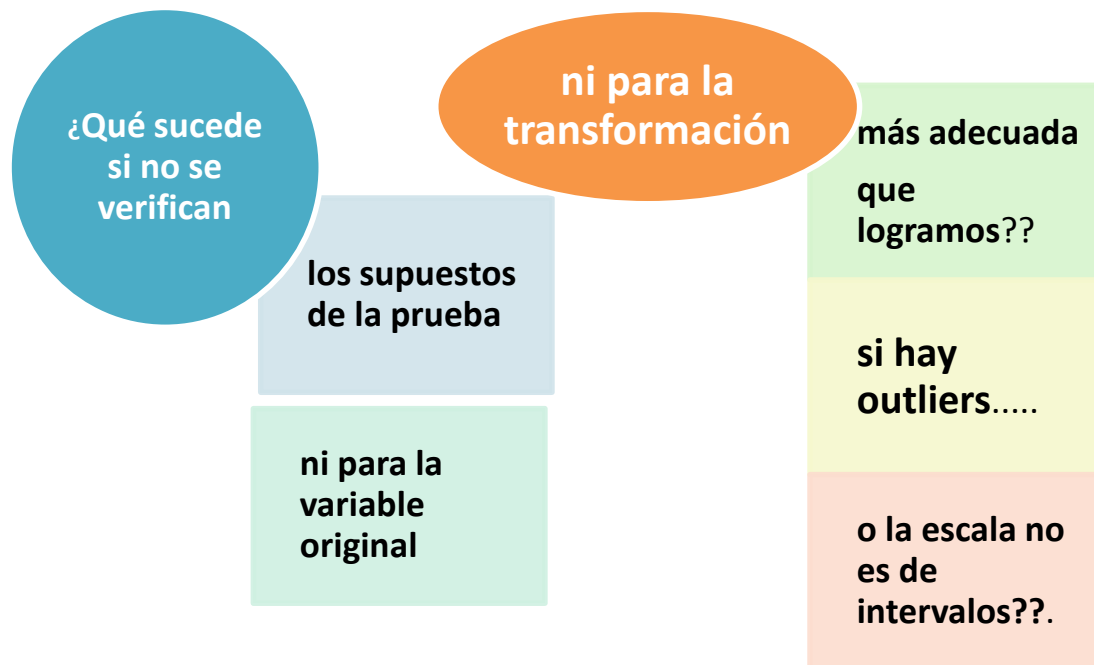
Tuckey y se aprecia que la dieta 3 produce niveles de colesterol inferiores a los de las otras dos dietas.

Test: Tukey Alfa=0.05 DMS=0.28806

Error: 0.0522 gl: 21

dieta	Medias	n	E.E.	
Dieta3	2.08	8	0.08	A
Dieta1	2.71	8	0.08	B
Dieta2	2.86	8	0.08	B

Medias con una letra común no son significativamente diferentes ($p > 0.05$)



La alternativa en este caso son las pruebas no paramétricas:

Entre ellas podemos utilizar la prueba de la mediana o la prueba de Kruskal Wallis también conocida como análisis de la varianza no paramétrico.

Entre ellas la más potente es la prueba de Kruskal Wallis, pero solo podremos aplicarla si la distribución de la variable es similar en los distintos grupos.

Test de Kruskal Wallis (no paramétrico para muestras independ.)

Esta prueba que es una extensión de la prueba de la suma de rangos de Wilcoxon para comparar la homogeneidad de dos poblaciones a partir de dos muestras aleatorias simples e independientes de ambas, contrasta la hipótesis nula de que las k muestras independientes proceden de la misma población y, en particular, todas ellas tienen la misma esperanza

Se basa en los rangos de las observaciones es el procedimiento alternativo a la prueba F del análisis de la varianza que no dependa de la hipótesis de normalidad.

El modelo que supone este test es:

Pobl 1: Y_{11}, \dots, Y_{1J} v.a.i.id con escala al menos ordinal

.....

Pobl k: Y_{k1}, \dots, Y_{kJ} v.a.i.id con escala al menos ordinal

donde las variables de las k poblaciones también son independientes entre sí.

Las distribuciones de todas las subpoblaciones deben ser semejantes, de lo contrario el rechazo de la hipótesis de nulidad implicaría que las distribuciones son distintas y no que sus medianas difieren

La hipótesis nula que planteamos es:

$$H_0: \theta_1 = \theta_2 = \dots = \theta_k$$

$$H_1: \text{existe al menos un par } i, j \text{ tal que } \theta_i \neq \theta_j$$

El estadístico de contraste para esta prueba es:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1)$$

Donde:

N es el total de observaciones

R_{ij} es el rango en la distribución conjunta de la observación j del tratamiento i

R_{i.} es el rango total de la muestra i

Y bajo H₀, el estadístico H tiene distribución aproximada Chi cuadrado con k-1 grados de libertad.

Por eso la regla de decisión es:

- ⊗ Rechazo H₀ cuando $H_{obs} \geq \chi^2_{k-1; 1-\alpha}$
- ⊗ No rechazo H₀ cuando $H_{obs} < \chi^2_{k-1; 1-\alpha}$

Procedimiento

1. Se ordenan todas las observaciones en sentido creciente y se reemplazan por su rango R_{it}, i = 1,...,I, t = 1,...,n_i, en la muestra conjunta ordenada.
2. En caso de empates se asigna a cada una de las observaciones empatadas el rango promedio de todas ellas.
3. Se calcula la suma de los rangos de cada grupo de observaciones. La suma de los rangos en la muestra combinada del i-ésimo tratamiento se designa con R_{i.}
4. Se calcula el estadístico de contraste H.
5. Se decide y se concluye.

Aclaración: decir que “las poblaciones tienen la misma posición” es equivalente a decir que tienen “el mismo valor esperado o media aritmética de los rangos”, o que “las poblaciones tienen igual mediana”.

Veamos un ejemplo:

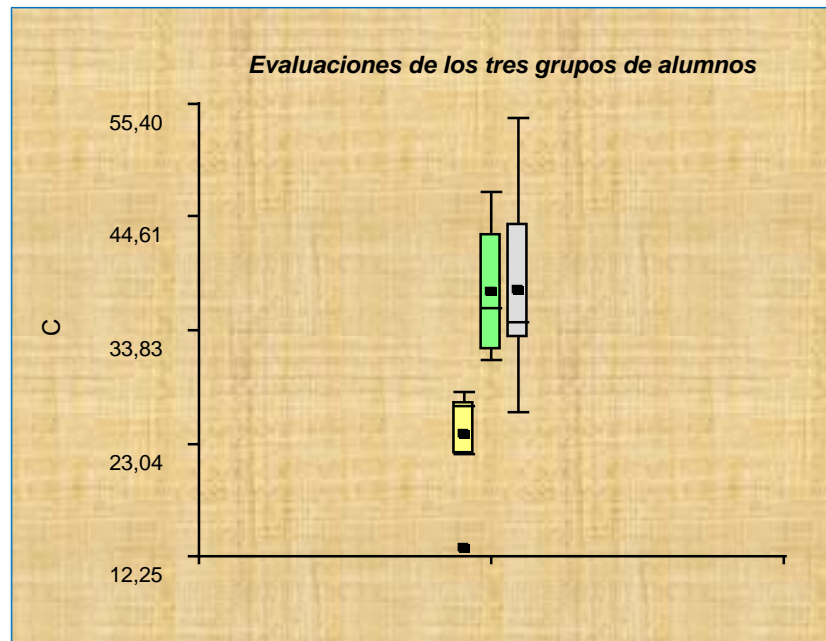
Ejemplo 3

Se realizó una intervención educativa innovadora para mejorar el rendimiento de los estudiantes. El grupo A es el grupo de control y los restantes son los grupos con distintas innovaciones. Se evaluó a los alumnos mediante una evaluación objetiva sobre un total de 60 puntos. Las puntuaciones logradas por los alumnos se presentan en la siguiente tabla:

Condición	Recuentos						
A	13	27	26	22	28	27	
B	43	35	47	32	31	37	
C	33	33	33	26	44	33	54

Graficamos los datos correspondientes a las tres distribuciones de los valores observados:





Testeamos la normalidad de la variable.

Variable	n	Media	D.E.	W*	p (una cola)
A	6	23,83	5,71	0,77	0,0342
B	6	37,50	6,32	0,90	0,4532
C	6	37,67	9,85	0,93	0,6281

Se rechaza la hipótesis de normalidad para la distribución de los recuentos correspondientes al grupo A.

Aplicamos por lo tanto un análisis no paramétrico mediante el test de Kruskal Wallis.

Planteamos las hipótesis:

H_0 : las k poblaciones tienen la misma posición para la variable en estudio

H_1 : al menos una población tiene diferente posición para la variable en estudio

Ordenamos los datos y los rankeamos:

Puntuación	grupo	rango	
13	A	1	
22	A	2	
26	A	3,5	
27	A	5,5	
27	A	5,5	
28	A	7	$R_A=24,5$
31	B	8	
32	B	9	
35	B	12,5	
37	B	14	
43	B	15	
47	B	17	$R_B=75,5$
26	C	3,5	
33	C	10	
34	C	11	
35	C	12,5	
44	C	16	
54	C	18	$R_C=71$

Establecemos la regla de decisión: Rechazamos H_0 si $\chi^2_{\text{obs}} > \chi^2_{(3-1),0.95}$ es decir que rechazaremos H_0 si $\chi^2_{\text{obs}} > 5,99$

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1) = \frac{12}{18(18+1)} \left[\frac{24,5^2}{6} + \frac{71^2}{6} + \frac{75,5^2}{6} \right] - 3(18+1) = 9,325$$

La decisión es rechazar H_0 pues $9,325 > 5,99$ que es el valor crítico establecido. Luego no puede suponerse que la mediana de los recuentos de linfocitos sea similar en los distintos grupos de pacientes considerados.

Si aplicamos la prueba en InfoStat, obtenemos también las comparaciones a posteriori del test de KW cuando se rechaza la hipótesis de nulidad:

Prueba de Kruskal Wallis

Variable	grupo	N	Medias	D.E.	Medianas	H	p
Puntuación A		6	23.83	5.71	26.50	9.32	0.0093
Puntuación B		6	37.50	6.32	36.00		
Puntuación C		6	37.67	9.85	34.50		

Trat. Ranks

A	4.08	A
C	11.83	B
B	12.58	B

Medias con una letra común no son significativamente diferentes ($p > 0.05$)