Análisis de la varianza de un factor

El **test t de 2 muestras** se aplica cuando se quieren comparar las medias de dos poblaciones con **distribuciones normales** con **varianzas iguales** y se observan **muestras independientes** para cada población. Ahora consideraremos una generalización para el caso en que se quieren comparar tres o más medias.

Ejemplo 7: En la tabla siguiente se muestran los resultados obtenidos en una investigación acerca de la estabilidad de un reactivo fluorescente en diferentes condiciones de almacenamiento.

Se conservaron tres muestras en cada una de 4 condiciones. Supongamos (porque a veces puede ocurrir) que para una de las condiciones, la medición no pudo realizarse o se detectó una falla y fue eliminada. Los datos observados son:

Condiciones	Mediciones observadas (señales de fluorescencia)	Media Muestral
Recientemente preparada	102 100 101	101
Una hora en la oscuridad	101 101 104	102
Una hora con luz tenue	97 95 99	97
Una hora con luz brillante	92 94	93

Mirando los promedios muestrales se ven diferencias y nos preguntamos si las condiciones de almacenamiento no influyeron sobre la fluorescencia de las muestras (ésta será nuestra H_0), ¿cuál es la probabilidad de que por simple azar se observen diferencias entre las medias muestrales de esta magnitud?

Para generalizar podemos pensar que observamos k muestras (en el ejemplo k=4). Suponemos el siguiente modelo:

Modelo de k muestras normales independientes con varianzas iguales.

Muestra 1: $X_{11}, X_{12}, ..., X_{1\,n1}$ v. a. i.i.d $N(\mu_1, \sigma^2)$

.....

Muestra i: $X_{i1}, X_{i2}, ..., X_{i\,ni}$ v. a. i.i.d $N(\mu_i, \sigma^2)$

.....

Muestra k: $X_{i1}, X_{k2}, ..., X_{k\,nk}$ v. a. i.i.d $N(\mu_k, \sigma^2)$

y asumimos que las v. a. de una muestra son independientes de las v. a. de otra muestra.

Llamaremos $\bar{X}_i$ y s_i^2 a la media y la varianza muestrales de la muestra i = 1,2, ...,k.

Parece natural proponer un estimador de σ^2 basado en un promedio ponderado de las varianzas de cada muestra s_i^2 , tal como se hacemos con el s_P^2 cuando comparamos dos muestras. Se puede demostrar que el mejor estimador insesgado de σ^2 bajo el modelo anterior es:

$$s_p^2 = \frac{(n_1 - 1) * s_1^2 + \ldots + (n_k - 1) * s_k^2}{n_1 + \ldots + n_k - k} = \frac{\sum_{i=i}^{k} (n_i - 1) * s_i^2}{n - k} = \frac{SS_W}{n - k} \qquad (1)$$

En la última expresión hemos llamado

$$n = \sum_{i=i}^{k} n_i$$

al número total de observaciones.

Vamos a estudiar la hipótesis nula:

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

La hipótesis alternativa es H_1: no es cierta H_0

Llamemos

$$\overline{X} = \frac{\sum_{i=i}^{k} n_i \overline{X}_i}{n} = \frac{\sum_{i=i}^{k} \sum_{j=1}^{n_i} X_{ij}}{n}$$

a la media general de todas las observaciones

El estadístico para el test óptimo para este problema, tiene al estimador de la varianza (dado por (1)) en el denominador y una medida de las diferencias (similar a la variancia) entre las medias de las distintas muestras en el numerador. Esta medida es:

$$\frac{\sum_{i=i}^{k} n_i (\overline{X}_i - \overline{X})^2}{k-1} = \frac{SS_B}{k-1} \qquad (2)$$

El estadístico del test se obtiene dividiendo (2) sobre (1):

$$F = \frac{\left(\sum_{i=i}^{k} n_i (\overline{X}_i - \overline{X})^2 \right) / (k-1)}{s_p^2} = \frac{SS_B / k-1}{SS_W / n-k} \qquad (3)$$

Si H_0 fuera cierta, entonces el denominador y el numerador serían parecidos, por lo tanto el cociente sería cercano a 1.

Si las medias poblacionales no son todas iguales, el numerador tiende a ser mayor que el denominador y por lo tanto, el cociente será mayor a 1.

Test *F*:

1er. paso: Calculo el estadístico *F* dado por (3)

Nota: Si H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ es cierta, este estadístico tiene distribución F con **k-1** grados de libertad en el **numerador** y **n-k** g.l. en el **denominador**.

¿De donde surgen los grados de libertad? Se puede demostrar, que si se satisfacen los supuestos del análisis de la varianza resulta:

i) $\frac{SS_W}{\sigma^2} \sim \chi^2_{n-K}$ ii) $\frac{SS_B}{\sigma^2} \sim \chi^2_{K-1}$ Bajo H_0 y además SS_B y SS_W son independientes

2do. paso: Si $F > F_{k-1,n-k;\alpha}$ rechazamos H_0.

Las cuentas de este test pueden hacerse con el Statistix. Para ello hay que ir a "Statistics", "One, Two, Multi-Sample Tests", "One-Way AOV" y se obtiene:

ONE-WAY AOV FOR FLUORESCE BY CONDICION

SOURCE	DF	SS	MS	F	P
BETWEEN	3	122.182	40.7273	**15.84**	**0.0017**
WITHIN	7	18.0000	2.57143		
TOTAL	10	140.182			

	CHI-SQ	DF	P
BARTLETT'S TEST OF EQUAL VARIANCES	0.75	3	0.8610

¿Para que sirve este test?

CONDICION	MEAN	SAMPLE SIZE	GROUP STD DEV
1	101.00	3	1.0000
2	102.00	3	1.7321
3	97.000	3	2.0000
4	93.000	2	1.4142
TOTAL	98.727	11	1.6036

Rechazamos la hipótesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ al nivel 0.01 y concluimos que la media de la fluorescencia depende de las condiciones de almacenamiento.

Comentarios sobre la "tabla del análisis de la varianza".
Se puede demostrar que vale la siguiente igualdad:

$$\sum_{i=i}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2 \quad = \quad \sum_{i=i}^{k} n_i(\overline{X}_i-\overline{X})^2 \quad + \quad \sum_{i=i}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2$$

En la expresión anterior aparecen tres "sumas de cuadrados":

"suma de cuadrados entre grupos" (SS_B: Between)
"suma de cuadrados dentro de grupos" (SS_W: Within)
"suma de cuadrados total" (SS_T: Total)

Statistix calcula estas tres sumas de cuadrados para el ejemplo y las muestra en la tabla que aparece al principio de la salida anterior (llamada tabla del Análisis de la Varianza). DF es la abreviatura de "degrees of freedom", SS de "sum of squares" y MS de "mean square". En castellano seria gl, SC y CM.

Analysis of Variance					
Source	SS	Df	MS	F	Prob > F
Between	SSB	k-1	MSB = SSB/k-1	MSB/MSW	
Within	SSW	n-k	MSW = SSW/n-k		
Total	SST	n-1	MST = SST/n-1		

Comparación de pares de medias

Supongamos que hemos aplicado el test F y hemos rechazado la H_0.

¿Qué quiere decir la alternativa? Que no todas la medias son iguales pero, ¿cuáles son diferentes?

Cuando no se puede rechazar H_0 generalmente el análisis termina ahí, pero cuando se rechaza generalmente el experimentador no se conforma con esa respuesta, sino que desea comparar las medias, frecuentemente (no siempre) de a pares, como para identificar cuáles son las que difieren.

Intervalo de confianza para la diferencia de dos medias.

Queremos comparar las medias de los grupos i y j. Empecemos por construir un IC para $\mu_i - \mu_j$

El estimador puntual es $\overline{X}_i - \overline{X}_j$

¿Cuál es su varianza? ¿Como se estima?

Puede demostrarse que

$$[\overline{X}_i - \overline{X}_j - t_{n-k,\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)};\ \overline{X}_i - \overline{X}_j + t_{n-k,\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}] \quad (4)$$

es un Intervalo de Confianza con nivel 1-α .

Si en vez de intervalo queremos estudiar la H_0: $\mu_i = \mu_j$ también es fácil deducir un test.

¿Se pueden calcular muchos IC o aplicar muchos tests?

¿Cuál es la crítica que se suele hacer a los IC “usando la distribución t” (de la forma (4)) y a los tests deducidos de estos intervalos?

Si hiciéramos unos pocos intervalos **elegidos a priori** (antes de observar los datos) la probabilidad de equivocarnos será >5%, pero no sería tan alta...

Si por ejemplo tenemos 6 tratamientos y hacemos todas las comparaciones de a pares, el número de intervalos de confianza será 15, ¿cuál será la probabilidad de que alguno no contenga al verdadero valor del parámetro? Aunque no la sepamos calcular exactamente, es evidente que esta probabilidad es mucho > que 0.05.

Por eso cuando uno planea de antemano hacer uno o muy pocos intervalos o tests puede usar (4), pero en caso contrario conviene utilizar un método de intervalos de confianza simultáneos.

Intervalos de confianza simultáneos (concepto general, no sólo para el análisis de varianza de un factor)

¿Cuál es la **definición** de IC para un parámetro θ?
Recordemos que si $X=(X_1,X_2,...,X_n)$ es la muestra observada, un intervalo $[a(X),b(X)]$ es un IC para θ con nivel $1-\alpha$ si

$$\mathbf{P(\ a(X) \le \theta \le b(X)\ ) = 1-\alpha}$$

Ahora deseamos calcular IC para cada uno de los parámetros θ_j (digamos j=1...,m). Se dice que el intervalo $[a_j(X),b_j(Y)]$ es un IC para θ_j calculado por un método simultáneo si

$$P\left(\bigcap_{j=1}^{m} [a_j(X) \le \theta_j \le b_j(X)] \right) \quad \ge \quad 1-\alpha \qquad (5)$$

o sea que la probabilidad de que todos los IC sean correctos (contengan al verdadero valor del parámetro) es $\ge 1-\alpha$. La probabilidad de que alguno sea incorrecto es $\le \alpha$.

Método de Bonferroni.

Un método muy general (para cualquier modelo) para obtener intervalos de confianza simultáneos es calcular cada uno de ellos con nivel 1-α/m, donde m es el número de IC que se desea calcular.

Este método tiene la ventaja de ser muy simple y muy general, pero sólo se usa en la práctica si m es pequeño, porque para valores moderados de m da IC de gran longitud.

Para el caso particular del análisis de la varianza de un factor, basta usar (4), pero reemplazando $t_{n-k,\alpha/2}$ por $t_{n-k,\alpha/2m}$ donde m es el número de IC que se desea calcular:

$$[\overline{X}i - \overline{X}_j - t_{n-k,\alpha/2m}\sqrt{s_p^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)};\ \overline{X}_i - \overline{X}_j + t_{n-k,\alpha/2m}\sqrt{s_p^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}]$$

BONFERRONI COMPARISON OF MEANS OF FLUORESCE BY CONDICION

CONDICION	MEAN	HOMOGENEOUS GROUPS
2	102.00	I
1	101.00	I
3	97.000	I I
4	93.000	.. I

THERE ARE 2 GROUPS IN WHICH THE MEANS ARE
NOT SIGNIFICANTLY DIFFERENT FROM ONE ANOTHER.

CRITICAL T VALUE 4.944 REJECTION LEVEL 0.010
STANDARD ERRORS AND CRITICAL VALUES OF DIFFERENCES
VARY BETWEEN COMPARISONS BECAUSE OF UNEQUAL SAMPLE SIZES.

Método de Tukey.

Los intervalos de Tukey son similares a los dados, pero reemplazando $t_{n-k,\alpha/2}$ por el valor $\mathbf{q_{k,n-k,\alpha}} / \sqrt{2}$, resultando

$$\overline{X}_i - \overline{X}_j \pm \frac{1}{\sqrt{2}} q_{k,n-k,\alpha} \sqrt{S_p^2 (\frac{1}{n_i} + \frac{1}{n_j})}$$

donde los valores "q" están tabulados y corresponden a la distribución estudiada por Tukey, llamada distribución del "**rango studentizado**" de k variables normales independientes. El $\sqrt{2}$ que aparece se debe simplemente a como se construyó la tabla.

Se basa en la distribución de

$$Q = \max \left| \frac{\overline{X}_i - \overline{X}_j - (\mu_i - \mu_j)}{\sqrt{S_p^2 (\frac{1}{n_i} + \frac{1}{n_j})}} \right|$$

Para el caso originalmente pensado por Tukey en el que los tamaños de muestras son iguales ($n_1=n_2=...=n_I$), este método hace que se cumpla el = en vez del ≥ en (5) cuando se realizan todas los comparaciones de a pares.

El método de Tukey es óptimo (da IC de la menor longitud posible) cuando se desea calcular IC para todos los pares posibles y los n_j's son iguales.

Para el caso en que los tamaños de muestras no son iguales, se demostró que sigue valiendo (5) pero con ">". En este caso el método se conoce también como "método de Tukey-Kramer".

Tests simultáneos: son los derivados de IC simultáneos. Tienen la propiedad de que la probabilidad de cometer algún error tipo I es menor o igual que α.

TUKEY (HSD) COMPARISON OF MEANS OF FLUORESCE BY CONDICION

CONDICION	MEAN	HOMOGENEOUS GROUPS
2	102.00	I
1	101.00	I
3	97.000	I I
4	93.000	.. I

THERE ARE 2 GROUPS IN WHICH THE MEANS ARE
NOT SIGNIFICANTLY DIFFERENT FROM ONE ANOTHER.

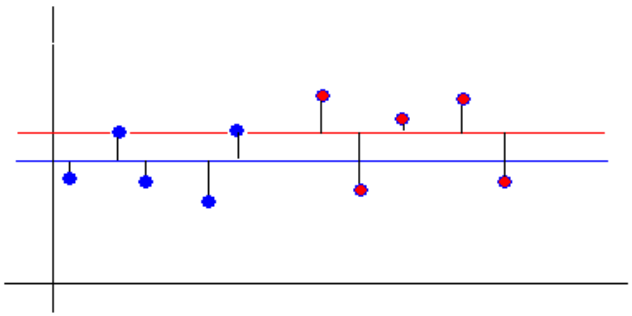
CRITICAL Q VALUE 6.552 REJECTION LEVEL 0.010
STANDARD ERRORS AND CRITICAL VALUES OF DIFFERENCES
VARY BETWEEN COMPARISONS BECAUSE OF UNEQUAL SAMPLE SIZES.

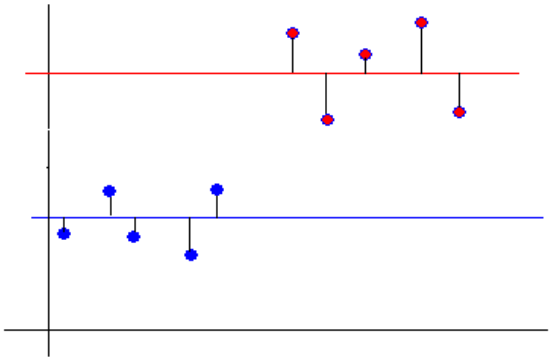
Comparación de los métodos considerados

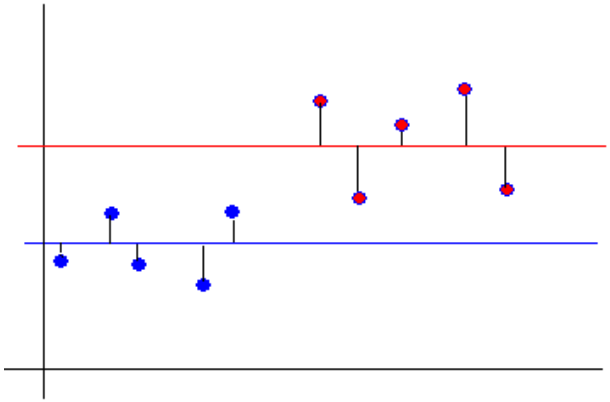
Si se desea calcular un IC o aplicar un test para una sola diferencia de medias elegidas **a priori**, evidentemente el método de elección es el basado en la distribución t.

Si son unos pocos, elegidos a priori conviene usar Bonferroni. Si se hacen muchas comparaciones de a pares (o algunas elegidas a posteriori, que es “igual que hacer muchas”) conviene usar Tukey pues da intervalos de menor longitud que Bonferroni.

Para elegir entre Bonferroni y Tukey, no es "trampa" elegir el método que da IC de menor longitud. No se necesita hacer las cuentas del IC para elegir el método: basta comparar quien es menor entre los valores de la tabla de "t" y de la tabla de "q", es decir entre $t_{n-k,\alpha/2m}$ y $q_{k,n-k,\alpha}/\sqrt{2}$.







Suposiciones del modelo. Diagnóstico.

El test *F* ha sido deducido bajo el supuesto de que las k muestras son normales independientes con igual varianza. Cuando el tamaño de la muestra de cada grupo es grande, el test *F* es válido en forma aproximada (el valor p calculado es aproximado) aunque la variable no tenga distribución normal, gracias al Teorema Central del Límite.

En la practica no es esperable que el modelo se cumpla exactamente, pero sí en forma aproximada. Al igual que con el test t, hay que analizar los datos para detectar si el modelo es aproximadamente cierto o si en cambio es falso.

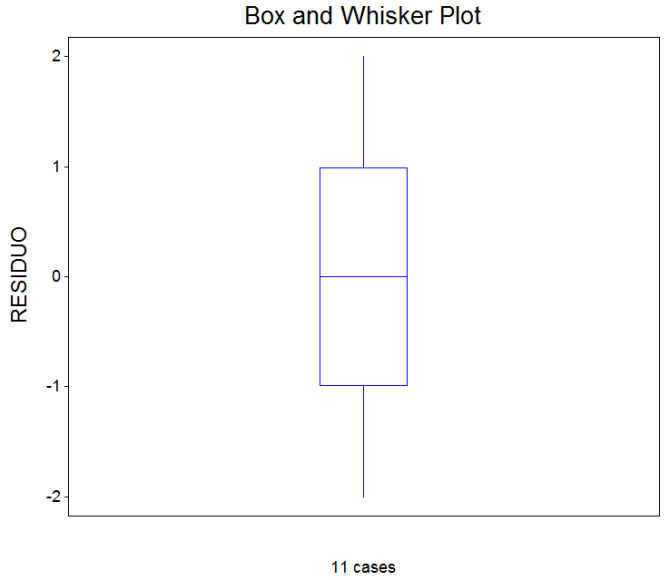
Normalidad

A menos que hubiera una gran cantidad de datos para cada nivel del factor, lo aconsejable es estudiar los residuos obtenidos a partir de la predicción que obtenemos prediciendo la media de cada nivel por el promedio muestral en cada casilla. Bajo los supuestos del modelo, es esperable que estos residuos sean aproximadamente normales y podríamos realizar un boxplot o un histograma para tener una idea de cómo se distribuyen.

Para cada observación, el residuos r_{ij} se calcula como:

$$r_{ij} = X_{ij} - \overline{X}_i$$

El siguiente gráfico muestra el boxplot correspondiente a los residuos del ejemplo de fluorescencia:



Los residuos parecen tener una distribución simétrica y no se observan datos atípicos, por lo que no parece haber importantes apartamientos de la normalidad.

QQ-plot

Otro posible gráfico para analizar la distribución de una muestra aleatoria es el QQ-plot y a la vez nos sirve para evaluar la cercanía a la distribución normal.

Para realizarlo se consideran los estadísticos de orden:

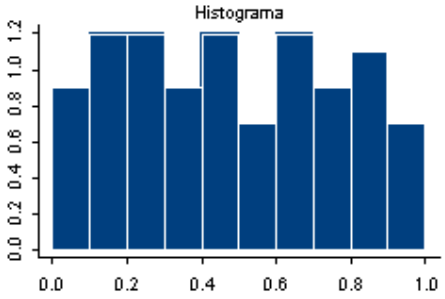
$$x_{(1)} \leq x_{(2)} \leq \ldots\ldots \leq x_{(n)}$$

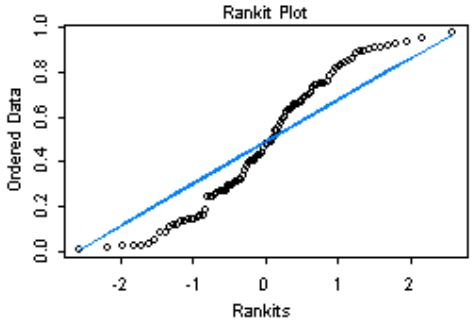
que se grafican versus los percentil $\frac{i-1/3}{n+1/3}$ teóricos de la normal, es decir $\phi^{-1}(\frac{i-1/3}{n+1/3})$.

Si los datos provienen de una distribución normal esperamos que el gráfico sea parecido a una recta. El alejamiento de la normalidad se ve reflejado por la forma del gráfico.

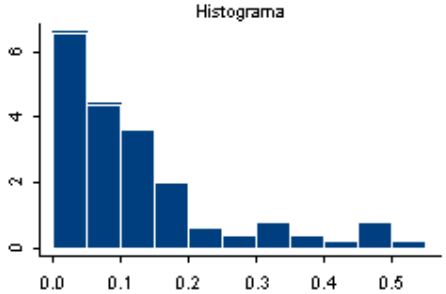
Algunos ejemplos QQ-plots

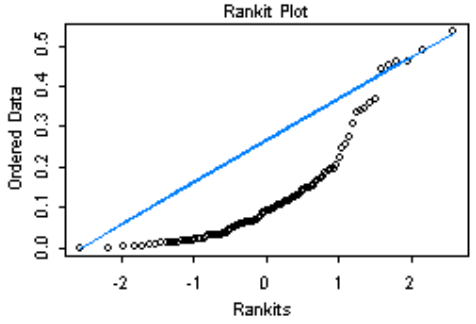
Simétrica con Colas Livianas



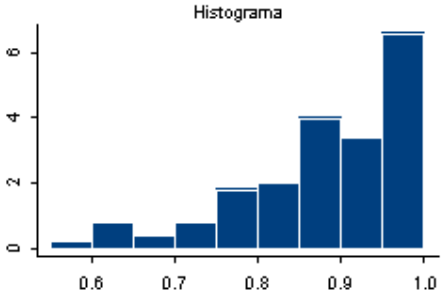


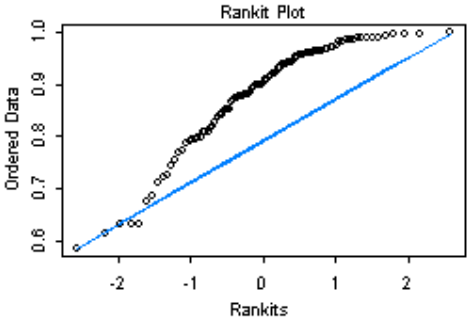
Asimetrica a Derecha



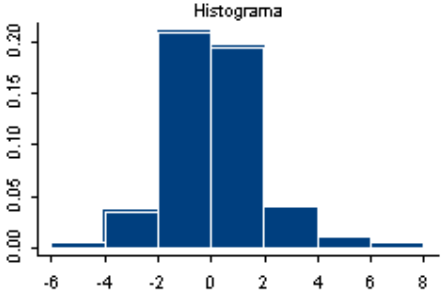


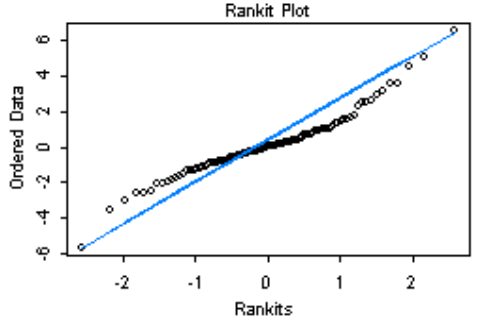
Asimetrica a Izquierda

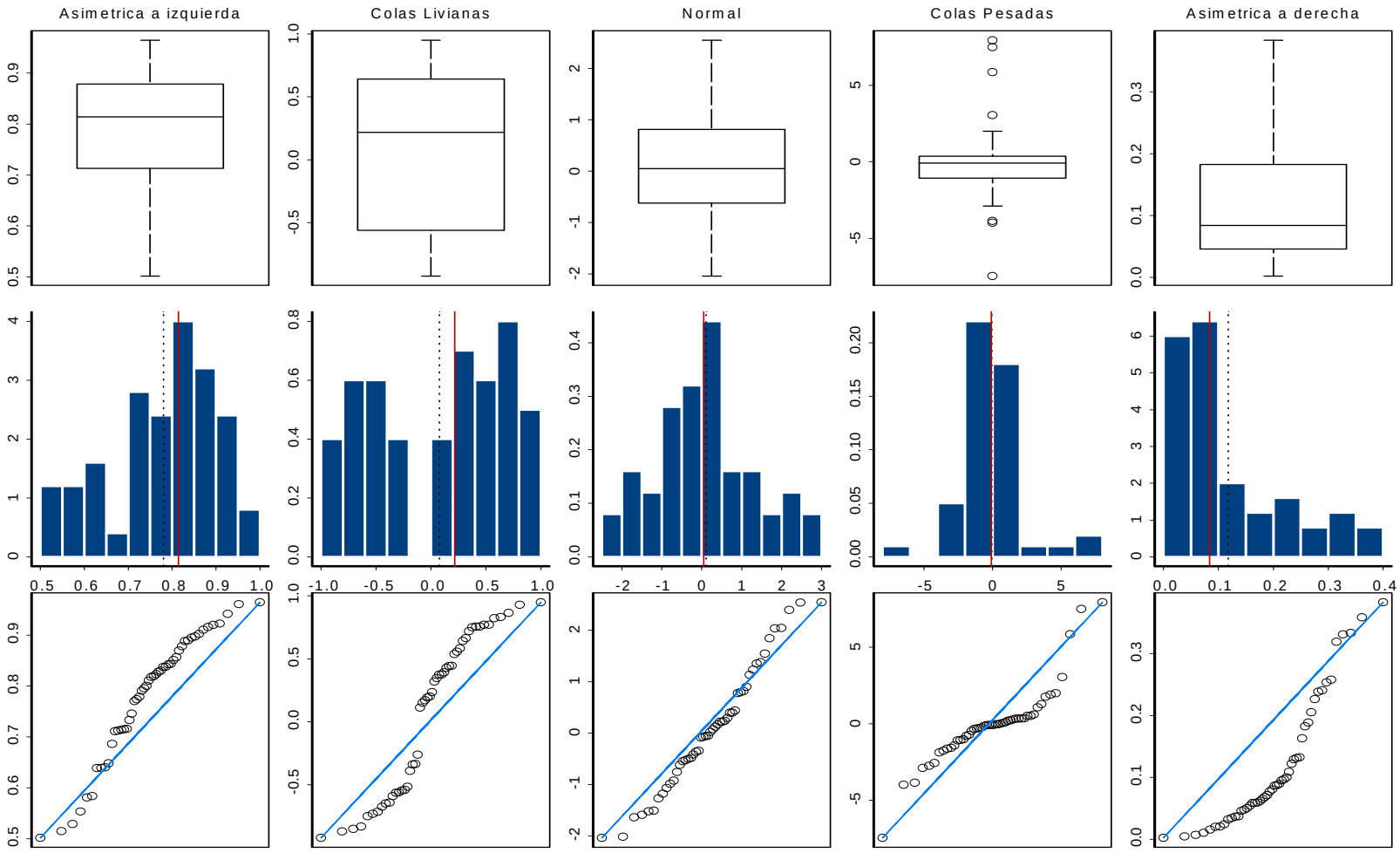




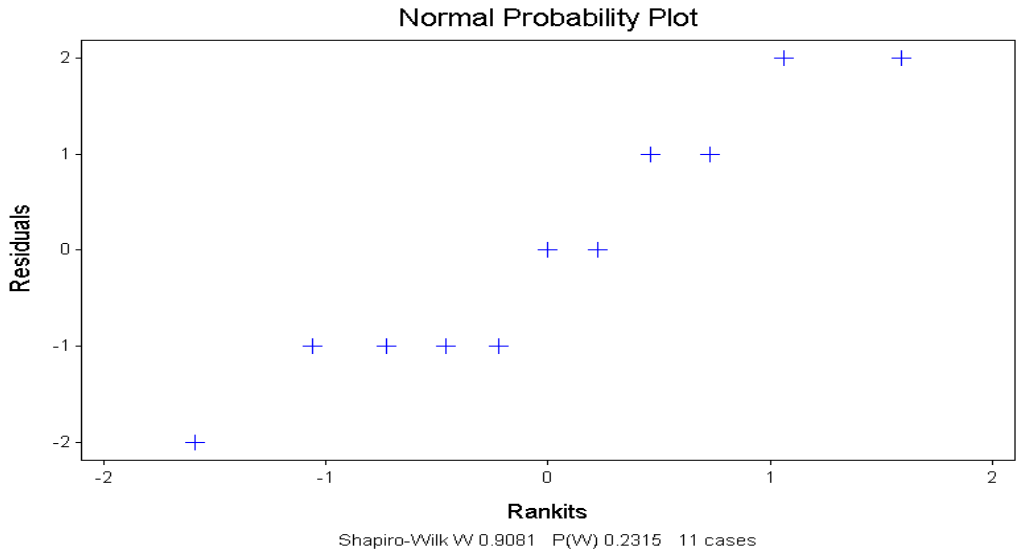
Simétrica con Colas Pesadas






Asimetrica a izquierda
Colas Livianas
Normal
Colas Pesadas
Asimetrica a derecha
Rojo=Mediana, Negro=Media

Si una vez que calculamos la Tabla de ANOVA con SX, vamos a "Results", "Plots", "Normal Probability Plot", obtenemos el siguiente gráfico:



Debajo del gráfico SX nos da una información adicional que corresponde al estadístico de test de Shapiro-Wilk y su correspondiente p-valor. Con este test podemos chequear la hipótesis de normalidad y podemos rechazar el supuesto de normalidad si el p-valor que nos brinda es muy pequeño. Esencialmente lo que hace este test es medir cuán cerca de una recta esta la curva que describen los puntos graficados en el QQ-plot.

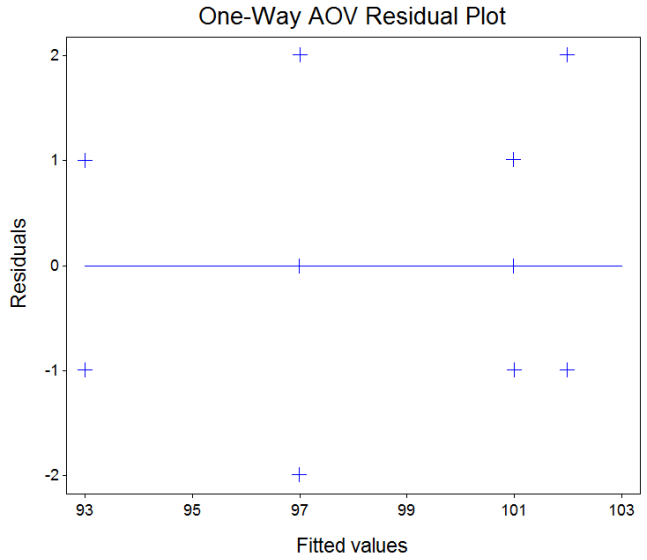
En nuestro ejemplo el estadístico del test de Shapiro-Wilk es 0.9081 y el p-valor correspondiente es de 0.2315, con lo cual no rechazamos el supuesto de normalidad.

Tests para estudiar si las varianzas son iguales

Para estudiar la suposición de igualdad de varianzas podemos graficar y también se pueden realizar algunos tests.

Respecto del gráfico podemos considerar un scatter-plot o diagrama de dispersión de los promedios muestrales versus los residuos.
En el ejemplo de Fluorescencia resultaría:

No parece haber apartamientos del supuesto de homocedasticidad en este caso.



Respecto de tests existen algunas alternativas.

En principio consideraremos el modelo

$$X_{ij} \sim N(\mu_i, \sigma_i^2) \quad (i=1,...,k;\ j=1,...,n_i) \text{ independientes}$$

y la hipótesis a testear será

$$H_0: \ \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 \ .$$

Hay varios tests. El más antiguo es el **test de Bartlett**. Se basa en un estadístico que tiene distribución aproximadamente χ^2_{k-1} bajo H_0.

Test de Barlett. Compara a S_p^2 y las S_j^2 correspondientes a cada grupo, pero en escala logarítmica. El test se basa en el estadístico $\ln 10 \frac{q}{c}$

con

$$q = (n-K)\log_{10} S_p^2 - \sum_{j=1}^{K}(n_j - 1)\log_{10} S_j^2$$

y

$$c = 1 + \frac{1}{3(K-1)}\left(\sum_{j=1}^{K}\frac{1}{n_j - 1} - \frac{1}{n-K}\right),$$

q es grande cuando las varianzas muestrales S_j^2 difieren mucho y vale cero cuando las S_j^2 son iguales.

La <u>región de rechazo</u> es: $\ln 10 \frac{q}{c} > \chi^2_{k-1,\ \alpha}$, donde $\chi^2_{k-1,\ \alpha}$ es el valor que deja a la derecha un área α bajo la curva de densidad χ^2 con k-1 grados de libertad.

Test Q de Cochran. Para el caso en que los n_j son todos iguales Cochran propuso el siguiente estadístico:

$$Q = \frac{\max(S_1^2, S_2^2, ..., S_K^2)}{S_1^2 + S_2^2 + ...S_K^2}$$

Se rechaza Ho cuando Q es muy grande, es decir cuando Q es más grande que el percentil 1-α de la distribución de Q cuando las varianzas son iguales (H_0).

Test de Hartley. El estadístico es:

$$F_{\max} = \frac{\max(S_1^2, S_2^2, ..., S_K^2)}{\min(S_1^2, S_2^2, ..., S_K^2)}$$

Región de rechazo: $F_{max} > f_{\max K, \bar{n}-1, \alpha}$

F_{max} tiene una distribución llamada $F_{\max K, \bar{n}-1}$ con dos parámetros K y $\bar{n}$ -1

($\bar{n}$ = parte entera (promedio de los n_i)

Este test tiene nivel exacto para igual número de observaciones por muestra y si los tamaños muestrales difieren tiene nivel aproximado.

	CHI-SQ	DF	P	
BARTLETT'S TEST OF	------	------	------	
EQUAL VARIANCES	0.75	3	**0.8610**	**No rechazamos la homocedaticidad**
COCHRAN'S Q		0.4000		
LARGEST VAR / SMALLEST VAR		4.0000		

Sin embargo, estos tres tests tienen un problema en común y es su alta sensibilidad a la falta de normalidad. Por esta razón, es necesario disponer de alguna alternativa más resistente a la falta de normalidad.

Un test que es poco sensible a la falta de normalidad es el **test de Modificado de Levene**. Para aplicarlo, primero se calculan

$$d_{ij} = | X_{ij} - \tilde{X}_i |$$

donde $\tilde{X}_i$ denota la mediana del tratamiento i .

Luego se calcula el estadístico *F* del análisis de un factor a los d_{ij}.

Si la hipótesis H: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ es cierta y los n_i "no son muy pequeños", el estadístico tiene distribución aproximadamente F con k-1 y n-k grados de libertad. Esto permite aplicar un test aproximado de la hipótesis de igualdad de varianzas. Rechazamos la igualdad de varianzas si el estadístico toma un valor muy grande.

Para aplicarlo con el SX, habría que calcular primero los valores d_{ij} y luego el test de F.

SOURCE	DF	SS	MS	F	P	
BETWEEN	3	0.90909	0.30303	0.53	**0.6757**	**No rechazamos la homocedaticidad**
WITHIN	7	4.00000	0.57143			
TOTAL	10	4.90909				

Test no paramétrico para comparar 3 o más muestras: test de Kruskal-Wallis.

Este test es una generalización del test de Wilcoxon- Mann Whitney al caso de más de 2 muestras. Igual que el test de Mann Whitney no requiere que los datos sean normales, y el estadístico de este test no se calcula con los datos originales, sino con los rangos de los datos.

Los supuestos en que se basa el test son:

- Los datos son por lo menos ordinales, es decir los datos pueden ordenarse en forma creciente de acuerdo con algún criterio.

- Además de la independencia entre las observaciones de una misma muestra suponemos independencia entre las observaciones de las distintas muestras.

De cada una de las k poblaciones tenemos una muestra aleatoria de tamaño n_i, es decir:

Muestra de la población i	Tamaño de muestra
$Y_{11}, Y_{12},\ldots\ldots, Y_{1n1}$	n_1
$Y_{21}, Y_{22},\ldots\ldots, Y_{2n2}$	n_2
.....	.....
.....	.....
.....	.....
$Y_{k1}, Y_{k2},\ldots\ldots, Y_{knk}$	n_k
Total de observaciones	$n = n_1 + n_2 + \ldots. n_k$

La hipótesis nula a testear es

H_o: todas las poblaciones tienen la misma distribución

Bajo H_o, todas las observaciones provienen de poblaciones idénticas. Si hacemos un pool con todas las n observaciones Y_{ij} y las ordenamos de menor a mayor, obtendremos los rangos

$$R_{ij}$$

Si H_o es cierta , las observaciones **Y_{ij} provienen de una misma distribución** y por lo tanto, todas las asignaciones de los rangos a las k muestras tienen la misma chance de ocurrir.

Si H_o es falsa, algunas muestras tenderán a tomar los rangos más pequeños, mientras que otras tenderán a tomar los rangos más grandes.

El **estadístico G_{KW} test de Kruskal-Wallis** mide la discrepancia entre los promedios observados de los rangos para cada tratamiento $\overline{R}_i$ y el valor que esperaríamos si Ho fuera cierta. En este test rechazamos la hipótesis nula de igualdad de medias si G_{KW} es grande.

El SX nos da el valor del estadístico G_{KW}.

El estadístico puede calcularse como

$$G_{KW} = \frac{\frac{SSB}{SST}}{n-1}$$

donde *SSB* y *SST* son, respectivamente, la **suma de cuadrados between** y la **suma de cuadrados total** para la tabla de análisis de la varianza correspondiente a los rangos de las observaciones.

SX nos da el p-valor usando la aproximación por una distribución: χ^2_{k-1}

Esta aproximación es válida cuando:

k=3 $n_i \geq 6$ para las k muestras

o bien

k>3 $n_i \geq 5$ para las k muestras

Para el caso en que k=3 y los $n_i \leq 5$ se debe usar la tabla con los percentiles de la distribución exacta.

Veamos salidas de SX para un nuevo ejemplo.

Ejemplo:

Interesa comparar el volumen espiratorio forzado en 1 segundo (FEV), en pacientes con enfermedad arterial coronaria que provienen de 3 centros médicos. La pregunta planteada es: *¿Hay alguna diferencia entre las medias de FEV de estos tres centros?*

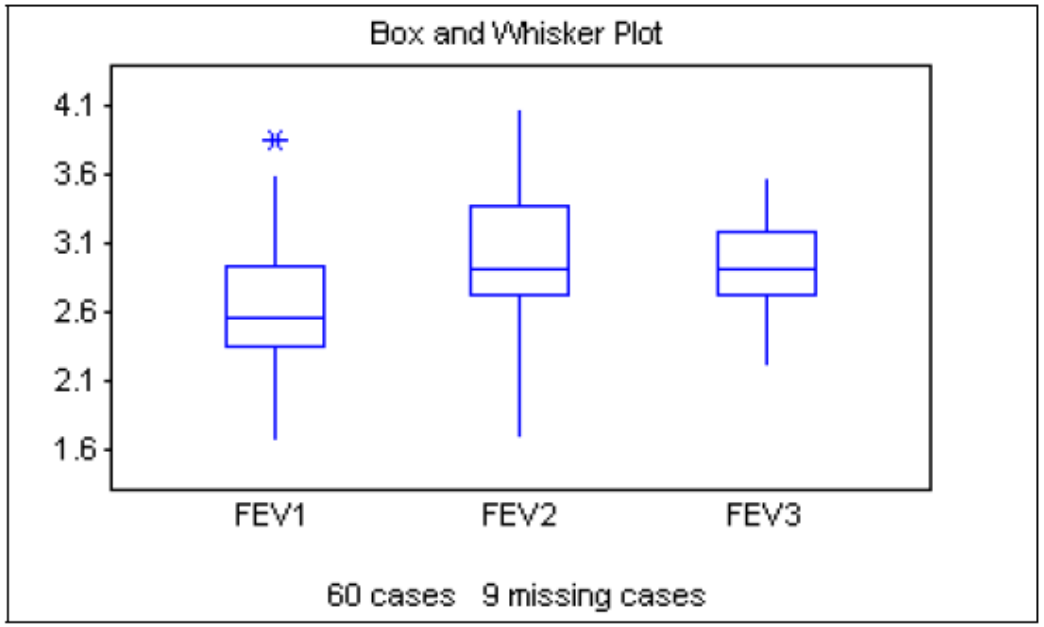
La tabla siguiente muestra el volumen espiratorio forzado en un segundo de pacientes de los tres centros y algunas medidas de resumen

FEV1	FEV2	FEV3
3.13	3.17	3.53
3.57	3.22	2.79
1.86	2.88	3.22
2.27	1.71	2.25
3.01	2.92	2.98
1.69	3.77	2.47
2.40	3.29	2.77
2.51	3.39	2.95
3.86	3.86	3.56
3.36	2.94	2.88
2.56	2.61	2.63
2.55	2.71	3.38
1.98	3.41	3.07
2.57	2.89	2.81
2.08	2.59	3.17
2.47	3.39	2.23
2.47	2.19	M
2.74	4.06	M

FEV1	FEV2	FEV3
2.88	1.98	M
2.60	2.81	M
2.45	2.85	M
2.23	M	M
3.20	M	M

	FEV1	FEV2	FEV3
N	23	21	16
MEAN	2.6278	2.9829	2.9181
SD	0.5386	0.5892	0.4031
MINIMUM	1.6900	1.7100	2.2300
1ST QUARTI	2.2700	2.6600	2.6650
MEDIAN	2.5500	2.9200	2.9150
3RD QUARTI	3.0100	3.3900	3.2075
MAXIMUM	3.8600	4.0600	3.5600
MAD	0.3200	0.3300	0.2700

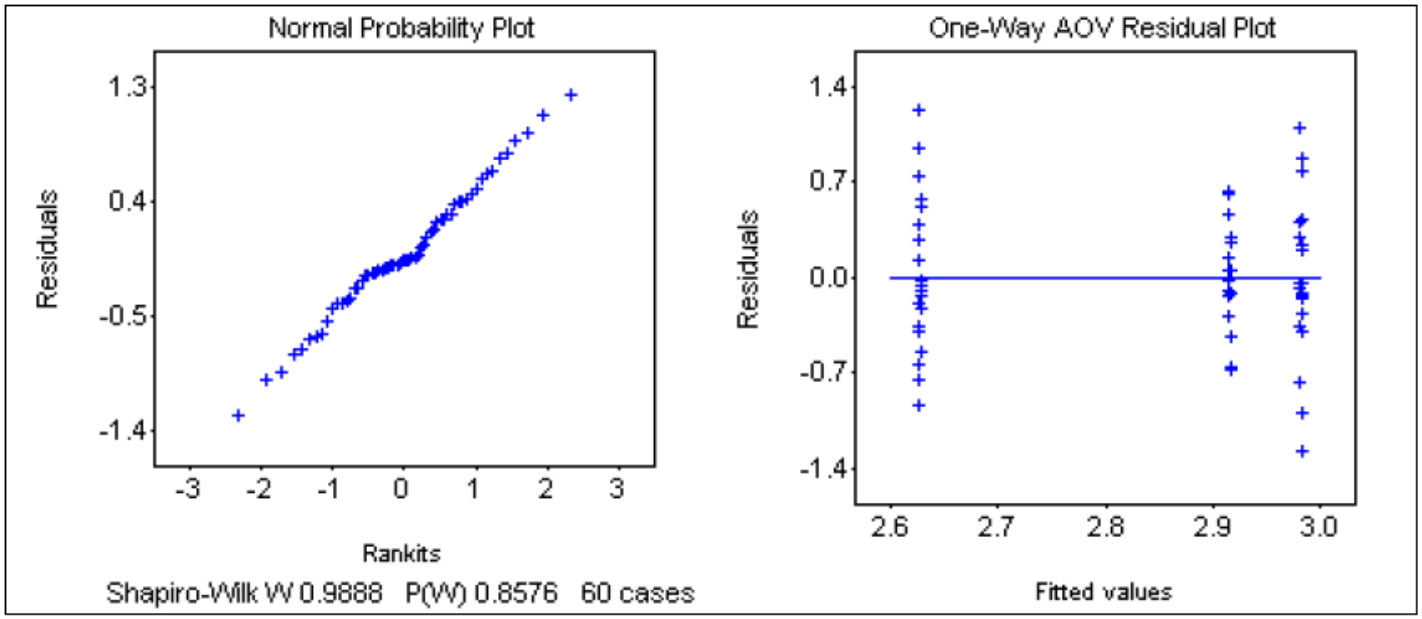
Veamos los boxplots paralelos.



¿Qué resultados obtendríamos si aplicásemos el test de F a estos datos?

```
ONE-WAY AOV FOR: FEV1 FEV2 FEV3

SOURCE     DF        SS          MS         F        P
-------   ----   ---------   ---------   ------   ------
BETWEEN     2      1.55502     0.77751     2.81   0.0685 ⇐  mayor a 0.05

WITHIN     57      15.7621     0.27653
TOTAL      59      17.3171

                     CHI-SQ     DF        P
BARTLETT'S TEST OF   ------   ------   ------
   EQUAL VARIANCES     2.32        2   0.3138

COCHRAN'S Q                      0.4341
LARGEST VAR / SMALLEST VAR       2.1366

COMPONENT OF VARIANCE FOR BETWEEN GROUPS    0.02532
EFFECTIVE CELL SIZE                           19.8
```

Analizamos los residuos.



KRUSKAL-WALLIS ONE-WAY NONPARAMETRIC AOV FOR FEV1 BY HOSPITAL

VARIABLE	MEAN RANK	SAMPLE SIZE
FEV1	23.2	23
FEV2	35.9	21
FEV3	33.9	16
TOTAL	30.5	60

KRUSKAL-WALLIS STATISTIC 6.5695

P-VALUE, USING CHI-SQUARED APPROXIMATION **0.0374** ⇐ menor a 0.05

PARAMETRIC AOV APPLIED to RANKS

SOURCE	DF	SS	MS	F	P
BETWEEN	2	2002.86	1001.43	3.57	**0.0346**
WITHIN	57	15984.6	280.432		
TOTAL	59	17987.5			

TOTAL NUMBER OF VALUES THAT WERE TIED 20

MAX. DIFF. ALLOWED BETWEEN TIES 0.00001

Como vemos, a partir del test de Kruskal-Wallis rechazaríamos la hipótesis H_0 de igualdad de las distribuciones con un nivel de significación de 0.05.

Como vemos la salida de SX incluye la tabla de ANOVA para los rangos de las observaciones. Esto se basa en que los dos estadísticos están relacionados. Si llamamos F_R al estadístico del test de F aplicado a los rangos tenemos que:

$$F_R = \frac{(N-k)G_{KW}}{(k-1)(N-1-G_{KW})}$$

Como en el caso del test de F, si rechazamos la hipótesis de igualdad seguramente vamos a desear identificar cuáles son las poblaciones que tiene distribución diferente. Si el número de comparaciones es moderado podemos utilizar el método de Bonferroni.

Si deseamos comparar *m* distribuciones podríamos calcular los intervalos de confianza para los promedios de los rangos que están dados por

$$\overline{R}_i - \overline{R}_j \pm z_{\alpha/2m}\left[\frac{n(n+1)}{12}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)\right]^{1/2}$$

Volviendo al ejemplo de FEV obtendríamos

FEV en 1 segundo

COMPARISONS OF MEAN RANKS

VARIABLE	MEAN RANK	HOMOGENEOUS GROUPS
FEV2	35.881	I
FEV3	33.875	I I
FEV1	23.239	.. I

THERE ARE 2 GROUPS IN WHICH THE MEANS ARE NOT SIGNIFICANTLY DIFFERENT FROM ONE ANOTHER.

Concluiríamos que los valores de FEV en el centro 2 difieren de los del centro 1 al nivel 5%.