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 (modelo B). Ahora consideraremos un problema similar, pero cuando 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 un error grosero y fue eliminada. Los datos observados son:

Condiciones	Mediciones observadas (señales de fluorescencia)	Media
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 las medias se ve que son diferentes. Pero nos preguntamos, si las condiciones de almacenamiento no influyeran sobre la fluorescencia de las muestras (ésta será nuestra H_0), cuál es la probabilidad de que por azar se observen diferencias entre las medias muestrales de estas magnitudes?

Para generalizar supongamos que observamos k muestras (en el ejemplo k=4). Suponemos el siguiente modelo, que es una generalización del modelo B de la clase anterior:

Modelo de k muestras normales independientes con varianzas iguales.

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

.....

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

.....

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

y las vs. as. de una muestra son independientes de las vs. as. de otra muestra.

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

Parece natural que el estimador de σ^2 se obtenga calculando un promedio ponderado de las varianzas de cada muestra s_i^2 . Se puede demostrar que el mejor estimador insesgado de σ^2 bajo el modelo anterior es:

$$s_p^2 = \frac{(n_1 - 1)*s_1^2 + ... + (n_k - 1)*s_k^2}{n_1 + ... + n_k - k} = \frac{\sum_{i=i}^{k}(n_i - 1)*s_i^2}{n - k} \qquad (23)$$

En la última expresión hemos llamado

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

al número total de observaciones.

Vamos a estudiar la hipótesis nula:

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

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 (23)) 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:

$$\frac{\sum_{i=i}^{k} n_i(\overline{X}_i - \overline{X})^2}{k-1} \qquad (24)$$

El estadístico del test se obtiene dividiendo (24) sobre (23):

$$F = \frac{\left(\sum_{i=i}^{k} n_i(\overline{X}_i - \overline{X})^2\right)/(k-1)}{s_p^2} \qquad (25)$$

Test *F*:
1er paso: Calculo el estadístico *F* dado por (25)
Nota: Si H_0: $\mu_1 = \mu_2 = ... = \mu_k$ es cierta, este estadístico tiene distribución F con k-1 grados de libertad en el numerador y n-k g.l. en el denominador.
2do. paso: Si $F > F_{k-1,n-k;\alpha}$ rechazo 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

por lo que se rechaza la hipótesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ y se concluye que la media de la fluorescencia depende de las condiciones de almacenamiento.

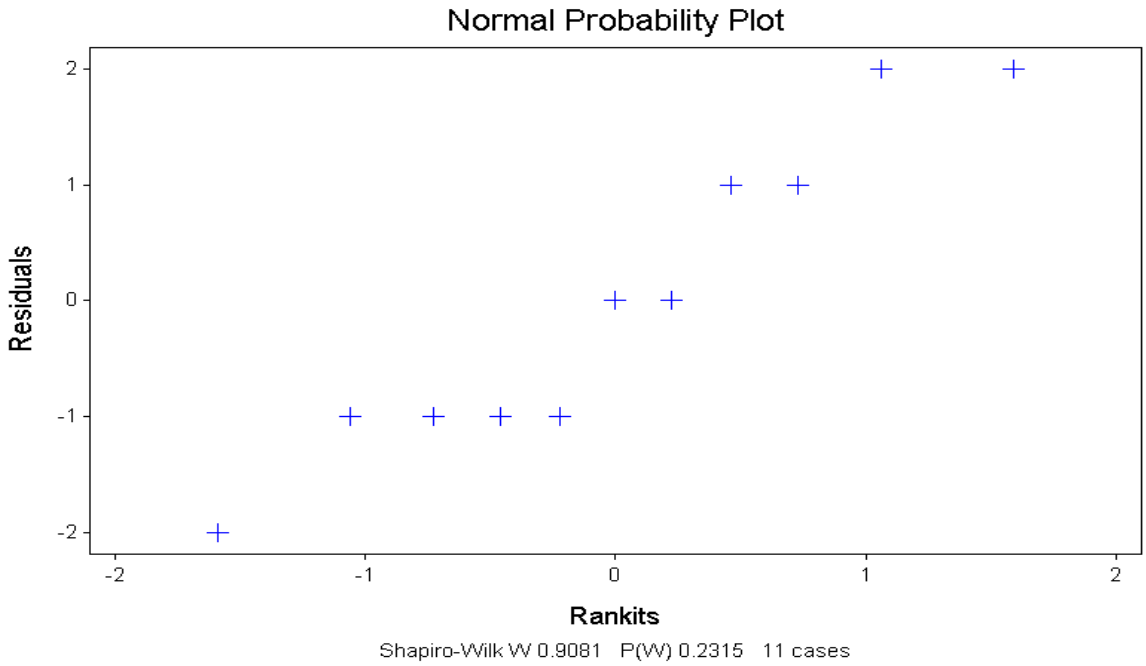
Comentarios sobre la "tabla del análisis de la varianza".
Se puede demostrar que vale la siguiente igualdad:

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

En la expresión anterior aparecen tres "sumas de cuadrados", que se llaman "suma de cuadrados entre grupos", "sc dentro de grupos" y "sc total". Diga usted cuál es cuál.
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.

Suposiciones del modelo. Diagnóstico.
El test *F* se ha deducido suponiendo el modelo de k muestras normales independientes con varianzas iguales. Cuando el tamaño de la muestra de cada grupo es grande, el test *F* es válido (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 sea completamente cierto, pero sí aproximadamente. Al igual que con el test t, hay que observar los datos para detectar si el modelo es aproximadamente cierto o es falso.

Normalidad: Si, luego de obtener la salida anterior, vamos a "Results", "Plots", "Normal Probability Plot", obtenemos el siguiente gráfico:



¿Cómo se calculan los "residuos" que se representan en el gráfico anterior?
Tanto este gráfico como el test de Shapiro-Wilk (que se muestra abajo del mismo), sirven para decidir si puede rechazarse la normalidad.

Test para estudiar si las varianzas son iguales: Para estudiar la suposición de igualdad de varianzas, además del gráfico también se puede hacer algún test. Algunos paquetes estadísticos hacen automáticamente algún test de la hipótesis de igualdad de varianzas cuando uno le pide el análisis de la varianza de un factor.

El problema es considerar el modelo

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

y la hipótesis H: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_I^2$.

Hay varios tests. El más antiguo es el **test de Bartlett** (es el que hace Statistix). Se basa en un estadístico que tiene distribución aproximadamente χ^2_{I-1} bajo H. Tiene un inconveniente serio: es muy sensible a la falta de normalidad.

Otro test que es poco sensible a la falta de normalidad es el **test 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: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_I^2$ es cierta y todos los n_i "no son muy pequeños", el estadístico tiene distribución aproximadamente F con I-1 y n-I g.l. Esto permite aplicar un test aproximado de la hipótesis de igualdad de varianzas.
Otros paquetes estadísticos (SPSS por ejemplo) usan el test de Levene y no el test de Bartlett. Para aplicarlo con el Statistix, habría que calcular primero los valores d_{ij}.

Test no paramétrico para comparer 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.

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.

Intervalo de confianza para la diferencia de dos medias.

Queremos comparar las medias de los grupos i y i*. Empecemos por construir un IC para $\mu_i - \mu_{i*}$

El estimador puntual es $\overline{X}i - \overline{X}_{i*}$.

¿Cuál es su varianza? ¿Como se estima?

Puede demostrarse que

$$[\overline{X}i - \overline{X}_{i*} - t_{n-k,\alpha/2}\sqrt{s_p{}^2\left(\frac{1}{n_i}+\frac{1}{n_{i*}}\right)};\ \overline{X}_i - \overline{X}_{i*} + t_{n-k,\alpha/2}\sqrt{s_p{}^2\left(\frac{1}{n_i}+\frac{1}{n_{i*}}\right)}] \qquad (26)$$

es un IC con nivel 1-α (¿exacto o asintótico?).

Si en vez de intervalo queremos estudiar la H_0: $\mu_i = \mu_{i*}$ también es fácil deducir un test (hacerlo).

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 (26)) y a los tests deducidos de estos intervalos. Si hacemos unos pocos intervalos **elegidos a priori** (antes de observar los datos) la probabilidad de equivocarnos será >5%, pero no será tan alta... Pero si por ejemplo tenemos 6 tratamientos y hacemos todas las comparaciones de a pares, el nro. de IC será 15, ¿cuál será la probabilidad de que alguno no contenga al verdadero valor del parámetro? Aunque no la sepamos calcular, 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 (26), 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-α 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

$$P\left(\bigcap_{j=1}^{m}[a_j(X) \le \theta_j \le b_j(X)]\right) \quad \ge \quad 1-\alpha \qquad (27)$$

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 $1-\alpha/m$, donde m es el número de IC que se desea calcular. (Ej.: demostrar que de este modo se consigue (27), usando la desigualdad de Bonferroni).
Este método tiene la ventaja de ser muy simple y muy general, pero sólo se usa en la práctica si m es muy pequeño, porque para valores moderados de m da IC de mucha longitud.

Para el caso particular del análisis de la varianza de un factor, basta usar (26), pero reemplazando $t_{n-k,\alpha/2}$ por $t_{n-k,\alpha/2m}$ donde m es el número de IC que se desea calcular.

Método de Tukey.
Los intervalos de Tukey (o Tukey-Kramer) son similares a los dados en (26) pero reemplazando $t_{n-k,\alpha/2}$ por el valor

$$\mathbf{q_{k,n-k,\alpha} / \sqrt{2}}$$

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.

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 $\ge$ en (27) 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 (27) 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.

Comparación de los métodos considerados
Si se desea calcular un IC o aplicar un test para una sola diferencia de medias elegida **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 (dá intervalos de menor longitud que Bonferroni).

Para elegir entre Bonferroni y Tukey, no es "trampa" elegir el método que dá 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" (entre $t_{n-k,\alpha/2m}$ y $q_{k,n-k,\alpha}/\sqrt{2}$).