Anova un factor y Kruskal-Wallis

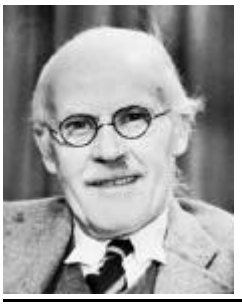
Introducción

El análisis de la varianza (Anova) se debe al estadístico-genético Sir Ronald Aylmer Fisher (1890-1962), autor del libro "Statistics Methods for Research Workers" publicado en 1925 y pionero de la aplicación de métodos estadísticos en el diseño de experimentos, introduciendo el concepto de aleatorización.

El Anova se puede utilizar en las situaciones en las que nos interesa analizar una respuesta cuantitativa, llamada habitualmente variable dependiente, medida bajo ciertas condiciones experimentales identificadas por una o más variables categóricas (por ejemplo tratamiento, sexo), llamadas variables independientes. Cuando hay una sola variable que proporciona condiciones experimentales distintas, el análisis recibe el nombre de Anova de un factor.

Entre las pruebas de comparación múltiples a posteriori, que se utilizan a continuación de las técnicas del Anova, se encuentra la prueba HSD de Tukey. John Tukey es, asímismo, conocido por introducir la transformación rápida de Fourier, aunque trabajó en muchas áreas incluyendo sobre todo la filosofía de

R. A. Fisher



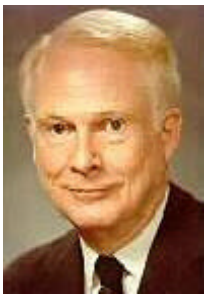
John Tukey

la estadística.
Cuando el análisis de la varianza no es aplicable debido a incumplimientos de las suposiciones del modelo, es necesario aplicar la prueba de Kruskal-Wallis para el contraste de k medianas. Esta prueba es una ampliación de la prueba de Mann-Whitney-Wilcoxon para dos medianas.

La prueba de Kruskal-Wallis fue propuesta por William Henry Kruskal (1919-) y W. Allen Wallis (1912-1998) en el artículo "Use of ranks in one-criterion variance analysis" publicado en el “Journal of

William H. Kruskal

W. Allen Wallis



American Statistics Association” en 1952.

Fórmulas básicas

En el análisis de la varianza, la variación en la respuesta se divide en la variación entre los diferentes niveles del factor (los diferentes tratamientos) y la variación entre individuos dentro de cada nivel. Suponiendo que las medias de los grupos son iguales, la variación entre grupos es comparable a la variación entre individuos. Si la primera es mucho mayor que la segunda, puede indicar que las medias en realidad no son iguales.

El objetivo principal del Anova es contrastar si existen diferencias entre las diferentes medias de los niveles de las variables (factores). Cuando sólo hay dos medias, el Anova es equivalente a la prueba t-Student para el contraste de dos medias.

La variación observada en la respuesta se asume que es debida al efecto de las variables categóricas, aunque también se asume que existe cierto error aleatorio independiente que explica la variación residual. Se asume también que dicho error aleatorio sigue una distribución normal con media 0 y varianza constante. Estas asunciones son análogas a las exigidas para la prueba t-Student para contrastar la igualdad de dos medias, donde se asumía normalidad de la respuesta en cada grupo e igualdad de varianzas (contrastada mediante la F-Snedecor).

Para estudiar la validez del modelo es necesario confirmar estas hipótesis mediante el estudio de los residuos (valores predichos - valores observados): normalidad, tendencias, etc. y la realización de un contraste de homocedasticidad (homogeneidad de varianzas entre los grupos).

Para el estudio de la normalidad de los errores, se puede recurrir al estudio de la normalidad de cada grupo (al igual que en la prueba t-Student) pero no es recomendable, debido a que puede requerir un gran número de pruebas. La solución utilizada habitualmente es el estudio del gráfico de dispersión entre los residuos y los valores predichos. Este gráfico permite estudiar la simetría, si existen patrones de comportamiento, la independencia entre observaciones y tendencias en general. Si se observa algún comportamiento de los mencionados, el modelo no es válido y se debe cambiar de modelo, de técnica estadística o transformar las variables.

Uno de los posibles contrastes para la homocedasticidad es la prueba de Barlett propuesta por Barlett en 1937. Esta prueba presupone que los datos provienen de variables con distribución normal. Otra alternativa menos sensible a la falta de normalidad y por este motivo recomendada por diversos autores es la prueba de Levene propuesta por Levene en 1960.

En general, el Anova es un procedimiento muy robusto que ofrece buenas aproximaciones en el caso que las premisas del modelo no se cumplan rigurosamente.

Muchas veces interesa saber qué medias difieren entre sí después de realizar el Anova. Para realizar contrastes a posterior es necesario ajustar el error alfa, y para este objetivo existen diferentes métodos, siendo la corrección de Tukey propuesta por el matemático John Tukey (1915-2000) la más habitual de todas ellas.

Los contrastes de comparaciones múltiples (o comparaciones a posteriori) proporcionan información detallada sobre las diferencias entre las medias dos a dos. Para este objetivo una primera intuición nos llevaría a realizar los correspondientes pruebas t-Student (o pruebas de Mann-Whitney-Wilcoxon para medianas) entre todas las posibles parejas de grupos. El problema reside en la repetición de múltiples contrastes. Si se tienen 5 medias, se necesitaría realizar 10 comparaciones 2 a 2 y cada una de ellas tendría un error alfa o de tipo I (probabilidad de rechazar la hipótesis nula cuando en realidad es

cierta) del 5%. Se puede comprobar que al realizar 10 contrastes al 5%, la probabilidad de rechazar al menos una de las hipótesis nulas es aproximadamente del 40%. De manera que con un 40% de probabilidades se llegaría a alguna conclusión falsa.

Existen diversos métodos para ajustar este tipo de error y conseguir que efectivamente el error conjunto no sea superior al 5%. La primera aproximación es debida a Fisher, quien propuso que sólo se han de comparar las diferencias entre medias 2 a 2 si el precedente Anova ha resultado significativo. Estas comparaciones a posteriori se realizan sin corrección alguna. Este método es conocido cómo LSD ("Least Significant Difference"). El método de Bonferroni es extremadamente conservador pero no depende de la muestra, sólo del número de comparaciones. Consiste en substituir el error alfa por α/nc siendo nc el número de comparaciones. En el método de Sidak se sustituye α por $1 - (1 - \alpha)^{1/nc}$, siendo uno de los más utilizados cuando sólo nos interesa contrastar si algunas de las diferencias son significativas. Existen otros métodos para controlar el error de cada comparación debidos a Scheffé (1953) y el método HSD ("Honestly Signifficant Difference") de Tukey (1953). Cuando todas las diferencias que se quieren estudiar son contra un mismo grupo control, es habitual realizar el ajuste de Dunnett (1955). También existen métodos de comparación de grupos de medias que permiten detectar grupos homogéneos de medias cómo el ajuste de Duncan y el de SNK ("Student-Newman-Keuls"), que son adecuados cuando los grupos son balanceados y el interés reside en obtener una comparación global.

En general, el método más conveniente es: después de realizar un Anova realizar el ajuste de Tukey y si se quiere contrastar todos los grupos con un control realizar el de Dunnett.

Cuando una comparación a posteriori no es significativa, la conclusión es: no ha sido posible rechazar la hipótesis nula, no que sea cierta. Por este motivo, es posible encontrar un modelo Anova significativo y que al mismo tiempo no haya diferencias entre medias dos a dos. Muchas veces esta situación es debida a tamaños de muestra reducidos.

En el caso de que no se cumplan las suposiciones del análisis de la varianza, es necesario aplicar la prueba de Kruskal-Wallis para el contraste de k medianas, que generaliza a la prueba de Mann-Whitney-Wilcoxon para dos medianas.

Cuando se compara medianas a través de la prueba de Kruskal-Wallis, las comparaciones 2 a 2 no suelen estar implementadas en los paquetes estadísticos, aunque se puede utilizar el método de Dunn por su sencillez de aplicación.

Anova de un factor

La prueba Anova nos permite comparar las medias de r grupos, siendo r mayor o igual a 2. El modelo Anova presupone que las varianzas de los grupos son iguales y que los residuos o errores son aleatorios, independientes e idénticamente distribuidos siguiendo una ley normal con media 0 y desviación constante. La hipótesis nula de la prueba Anova de un factor es:

H_0: Las medias de los k grupos son todas iguales

H_1: Al menos una de las medias es diferente

Esta prueba se basa en la comparación de las sumas de cuadrados medias debidas a la variabilidad entre grupos y la debida a la variabilidad intra grupos (dentro de los grupos). Ambas sumas son estimaciones independientes de la variabilidad global, de manera que, si el cociente entre la primera y la segunda es grande, se tendrá mayor probabilidad de rechazar la hipótesis nula. Este cociente sigue una distribución F con r - 1 y n - r grados de libertad.

Cálculo de la suma de cuadrados

Las sumas de cuadrados son un paso previo para el cálculo del Anova. Si se denotan por r al número de grupos, por n_j el número de individuos en cada grupo j = 1,...,r, $\bar{x}_{.j}$ la media de cada grupo y $\bar{x}_{..}$ la media global. La suma de cuadrados entre grupos SCE, la suma de cuadrados dentro de grupos SDE y la suma de cuadrados total SCT se calculan del siguiente modo:

$$SCE = \sum_{j=1}^{r} n_j \left(\bar{x}_{.j} - \bar{x}_{..}\right)^2$$

$$SCD = \sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_{.j}\right)^2 = \sum_{j=1}^{r}\sum_{i=1}^{n_j} x_{ij}^2 - \sum_{j=1}^{r} n_j \bar{x}_{.j}^2$$

$$SCT = \sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_{..}\right)^2$$

Utilizando la siguiente igualdad que permite expresar las desviaciones entre los datos observados x_{ij} y la media total ("grand mean") $\bar{x}_{..}$ como suma de las desviaciones de la media del grupo $\bar{x}_{.j}$ y la media total más las desviaciones entre los datos observados y la media del grupo, de forma que:

$$x_{ij} - \bar{x}_{..} = \left(\bar{x}_{.j} - \bar{x}_{..}\right) + \left(x_{ij} - \bar{x}_{.j}\right)$$

se puede demostrar que SCT = SCE + SCD y por tanto la variabilidad de los datos (dada por SCT) se expresa como la suma de la variabilidad debida a los grupos (a las medias) o variabilidad explicada (dada por SCE) más la variabilidad dentro de los grupos (variabilidad residual) o variabilidad no explicada (dada por SCD).

Cálculo de los grados de libertad

Los grados de libertad entre grupos GLE, dentro de los grupos GLD y total GLT se calculan de la manera siguiente:

$$GLE = r - 1$$
$$GLD = n - r$$
$$GLT = n - 1$$

Cálculo de los cuadrados medios

El cuadrado medio entre grupos CME y el cuadrado medio dentro de grupos se calculan de la manera siguiente:

$$CME = \frac{SCE}{GLE}$$

$$CMD = \frac{SCD}{GLD}$$

Estadístico de contraste F

El estadístico de contraste para realizar la prueba Anova se construye de la forma siguiente:

$$F = \frac{CME}{CMD}$$

que se distribuye según una F-Snedecor con GLE grados de libertad del numerador y GLD grados de libertad del denominador.

Cálculo del coeficiente de determinación

Una medida relativa de la variabilidad explicada por los grupos es el cociente:

$$R^2 = \frac{SCE}{SCT}$$

que se denomina coeficiente de determinación , este coeficiente estará entre cero y uno. Queda claro que cuanto más próximo esté de uno, más variabilidad explica el modelo, y por tanto menos variabilidad no explicada o residual.

Tabla del Anova

La información anterior se suele disponer en forma de tabla:

	Suma de Cuadrados	G.L.	Cuadrado Medio	F-valor	p-valor
Entre Grupos	SCE	GLE	CME	F	p
Dentro Grupos	SCD	GLD	CMD		
Total	SCT	GLT			

Kruskal-Wallis

La prueba de Kruskal-Wallis es el método más adecuado para comparar poblaciones cuyas distribuciones no son normales. Incluso cuando las poblaciones son normales, este contraste funciona muy bien. También es adecuado cuando las desviaciones típicas de los diferentes grupos no son iguales entre sí, sin embargo, el Anova de un factor es muy robusto y sólo se ve afectado cuando las desviaciones típicas difieren en gran magnitud.

La hipótesis nula de la prueba de Kruskal-Wallis es:

H_0: Las k medianas son todas iguales

H_1: Al menos una de las medianas es diferente

Cálculo de los rangos para cada observación

Para cada observación se le asigna el rango según el orden que ocupa la observación en el conjunto total de los datos, asignando el rango medio en caso de empates.

Cálculo de la suma de rangos R_m

Para cada grupo m = 1,…,r, siendo r el número de grupos, se define R_m como la suma de rangos de cada grupo m

Cálculo del valor medio de los rangos $E[R_m]$ y de los rangos medios $\bar{R}_m$

El valor medio de los rangos $E[R_m]$ se calcula como:

$$E[R_m] = \frac{n_m(n+1)}{2}$$

y el rango medio $\overline{R}_m$ como:

$$\overline{R}_m = \frac{R_m}{n_m}$$

Estadístico de contraste H'

El estadístico de contraste de Kruskal-Wallis H' se calcula como:

$$H' = \frac{\frac{12}{n(n+1)} \sum_{m=1}^{r} \frac{1}{n_m} [R_m - E[R_m]]^2}{1 - \frac{\sum_{j=1}^{k} (d_j^3 - d_j)}{n^3 - n}}$$

siendo d_j el número de empates en j = 1,...,k siendo k el número de valores distintos de la variable respuesta, que sigue una distribución Chi-Cuadrado con r - 1 grados de libertad.

Ejemplos

Anova de un factor

Se tienen los siguientes datos experimentales, correspondientes a 40 individuos de los que se ha recogido información de dos variables: la variable explicativa Status es nominal y la variable respuesta Fc2 es cuantitativa. Los datos se presentan de forma que en las filas hay varios individuos para facilitar la lectura:

Fc2	Status	Fc2	Status	Fc2	Status	Fc2	Status
155	1	144	1	126	2	120	3
154	1	136	1	160	2	126	3
148	1	134	1	136	2	116	3
132	1	142	1	158	2	142	3
126	1	138	1	142	2	144	3
132	1	140	1	134	2	112	3
156	1	136	1	148	2	116	3
138	1	165	2	146	2	120	3
158	1	148	2	126	3	122	3
144	1	128	2	128	3	132	3

Calcular la prueba Anova de comparación de medias para los datos anteriores.

Cálculo de la suma de cuadrados

Las sumas de cuadrados son un paso previo para el cálculo del Anova. Si se denotan por r al número de grupos, por n_j el número de individuos en cada grupo j = 1,...,r, $\overline{x}_{.j}$ la media de cada grupo y $\overline{x}_{..}$ la media global. La suma de cuadrados entre grupos SCE, la suma de cuadrados dentro de grupos SDE y la suma de cuadrados total SCT se calculan del siguiente modo:

$$SCE = \sum_{j=1}^{r} n_j \left(\overline{x}_{.j} - \overline{x}_{..}\right)^2 = 17 \cdot (141.9412 - 137.7)^2 + 11 \cdot (144.6364 - 137.7)^2 +$$

$$12 \cdot (125.3333 - 137.7)^2 = 2670.2467$$

$$SCD = \sum_{j=1}^{r} \sum_{i=1}^{n_j} \left(x_{ij} - \overline{x}_{.j}\right)^2 = \sum_{j=1}^{r} \sum_{i=1}^{n_j} x_{ij}^2 - \sum_{j=1}^{r} n_j \overline{x}_{.j}^2 =$$

$$765330 - \left(17 \cdot 141.9412^2 + 11 \cdot 144.6364^2 + 12 \cdot 125.3333^2\right) = 765330 - 761121.85 = 4208.1533$$

$$SCT = \sum_{j=1}^{r} \sum_{i=1}^{n_j} \left(x_{ij} - \overline{x}_{..}\right)^2 = SCE + SCD = 6878.4$$

Cálculo de los grados de libertad

Los grados de libertad entre grupos GLE, dentro de los grupos GLD y total GLT se calculan de la manera siguiente:

GLE = r - 1 = 2

GLD = n - r = 40 - 3 = 37

GLT = n – 1 = 40 - 1 = 39

Cálculo de los cuadrados medios

El cuadrado medio entre grupos CME y el cuadrado medio dentro de grupos se calculan de la manera siguiente:

$$\text{CME} = \frac{\text{SCE}}{\text{GLE}} = 1335.1234$$

$$\text{CMD} = \frac{\text{SCD}}{\text{GLD}} = 113.7339$$

Estadístico de contraste F

El estadístico de contraste para realizar la prueba Anova se construye de la forma siguiente:

$$\text{F} = \frac{\text{CME}}{\text{CMD}} = 11.7390$$

que se distribuye según una F-Snedecor con GLE = 2 grados de libertad del numerador y GLD = 37 grados de libertad del denominador, que tiene asociado un p-valor de 0.0001

Cálculo del coeficiente de determinación

Una medida relativa de la variabilidad explicada por los grupos es el cociente:

$$\text{R}^2 = \frac{\text{SCE}}{\text{SCT}} = 39\%$$

con lo que se tendría, al igual que en modelos de regresión, que el modelo Anova, o más específicamente, la variable que forma los grupos, explica un 39% de la variabilidad de la variable respuesta.

Kruskal-Wallis

Se tienen los siguientes datos experimentales, correspondientes a 22 individuos de los que se ha recogido información de dos variables: una variable explicativa Exp nominal y otra variable respuesta Rta cuantitativa. Los datos se presentan de forma que en las filas hay varios individuos para facilitar la lectura:

Rta	Exp	Rta	Exp
15	1	28	2
15	1	28	2
25	1	28	2
25	1	35	2
25	1	43	2
33	1	13	3
43	1	15	3
15	2	25	3
16	2	25	3
16	2	35	3
25	2		
28	2		

Calcular la prueba de Kruskal-Wallis de comparación de medianas para los datos anteriores.

Cálculo de los rangos para cada observación

Para cada observación se le asigna el rango según el orden que ocupa la observación en el conjunto total de los datos, asignando el rango medio en caso de empates:

Rta	Exp	Rango (Rta)	Rta	Exp	Rango (Rta)
15	1	3.5	28	2	15.5
15	1	3.5	28	2	15.5
25	1	10.5	28	2	15.5
25	1	10.5	35	2	19.5
25	1	10.5	43	2	21.5
33	1	18	13	3	1
43	1	21.5	15	3	3.5
15	2	3.5	25	3	10.5
16	2	6.5	25	3	10.5
16	2	6.5	35	3	19.5
25	2	10.5			
28	2	15.5			

Cálculo de la suma de rangos R_m

Para cada grupo m = 1,...,r, siendo r el número de grupos, se define R_m como la suma de rangos de cada grupo m, que para los datos del ejemplo resultan ser:

$$R_1 = \sum_{\text{grupo1}} \text{rangos} = 3.5 + 3.5 + 10.5 + 10.5 + 10.5 + 18 + 21.5 = 78.00$$

$$R_2 = \sum_{\text{grupo2}} \text{rangos} = 3.5 + 6.5 + 6.5 + 10.5 + 15.5 + 15.5 + 15.5 + 15.5 + 19.5 + 21.5 = 130.00$$

$$R_3 = \sum_{\text{grupo3}} \text{rangos} = 1 + 3.5 + 10.5 + 10.5 + 19.5 = 45.00$$

Cálculo del valor medio de los rangos $E[R_m]$ y de los rangos medios $\overline{R}_m$

El valor medio de los rangos $E[R_m]$ se calcula como:

$$E[R_m] = \frac{n_m(n+1)}{2}$$

y el rango medio $\overline{R}_m$ como:

$$\overline{R}_m = \frac{R_m}{n_m}$$

Para los datos del ejemplo resultan ser:

$$E[R_1] = \frac{n_1(n+1)}{2} = \frac{7 \cdot (22+1)}{2} = 80.50$$

$$E[R_2] = \frac{n_2(n+1)}{2} = \frac{10 \cdot (22+1)}{2} = 115.50$$

$$E[R_3] = \frac{n_3(n+1)}{2} = \frac{5 \cdot (22+1)}{2} = 57.50$$

$$\overline{R}_1 = \frac{R_1}{n_1} = 11.14$$

$$\overline{R}_2 = \frac{R_2}{n_2} = 13.00$$

$$\overline{R}_3 = \frac{R_3}{n_3} = 9.00$$

Estadístico de contraste H'

El estadístico de contraste H' se calcula como:

$$H' = \frac{\frac{12}{n(n+1)} \sum_{m=1}^{r} \frac{1}{n_m} [R_m - E[R_m]]^2}{1 - \frac{\sum_{j=1}^{k} (d_j^3 - d_j)}{n^3 - n}}$$

siendo d_j el número de empates en j = 1,…,k siendo k el número de valores distintos de la variable respuesta, que para los datos del ejemplo resulta ser:

$$\sum_{j=1}^{k} (d_j^3 - d_j) = (4^3 - 4) + (2^3 - 2) + (6^3 - 6) + (4^3 - 4) + (2^3 - 2) + (2^3 - 2) = 348$$

con lo que:

$$H' = \frac{\frac{12}{22 \cdot 23} \left(\frac{1}{7} [78 - 80.5]^2 + \frac{1}{10} [130 - 115]^2 + \frac{1}{5} [45 - 57.5]^2 \right)}{1 - \frac{348}{22^3 - 22}} = 1.3398$$

que sigue una distribución Chi-Cuadrado con r –1 = 2 grados de libertad, que tiene asociada un p-valor de 0.5118