Estadística Básica con R y R–Commander

(Versión Febrero 2008)

Autores:
A. J. Arriaza Gómez
F. Fernández Palacín
M. A. López Sánchez
M. Muñoz Márquez
S. Pérez Plaza
A. Sánchez Navas

UCA Universidad de Cádiz
Servicio de Publicaciones

Edita: Servicio de Publicaciones de la Universidad de Cádiz
C/ Dr. Marañón, 3
11002 Cádiz

http://www.uca.es/publicaciones

ISBN:

Depósito legal:

Estadística Básica con R y R-commander
(Versión Febrero 2008)
Autores: A. J. Arriaza Gómez, F. Fernández Palacín, M. A. López Sánchez, M. Muñoz Márquez, S. Pérez Plaza, A. Sánchez Navas

http://knuth.uca.es/ebrcmdr

Capítulo 7

Introducción al Análisis de la Varianza

1. Conceptos básicos

Aunque en origen el *Análisis de la Varianza* (ANOVA) fue introducido por Fisher para evaluar los efectos de los distintos niveles de un factor sobre una variable respuesta continua, desde un punto de vista puramente abstracto el ANOVA va a permitir generalizar el contraste de igualdad de medias de dos a k poblaciones. Y esa es la perspectiva en la que se va a centrar este último capítulo. No se propondrá pues ningún modelo teórico, sino que el objetivo se limitará a usar la técnica para contrastar la hipótesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$. Eso sí, al igual que se ha hecho para una y dos poblaciones, se evaluarán las hipótesis previas relativas a la calidad de la muestra, a la estructura de probabilidad, normal o no, de la población y a si las distintas poblaciones tienen varianzas iguales o distintas, propiedad esta última conocida como *homocedasticidad*.

El ANOVA en su versión paramétrica del *test de la F*, como todos los procedimientos estadísticos, tiene un cierto grado de *robustez* frente a un relativo incumplimiento de alguna(s) de sus hipótesis. En concreto, el test de la F soporta mejor las deficiencias respecto a la normalidad que las relacionadas con la homocedasticidad. En todo caso, los test son menos sensibles a las desviaciones de las hipótesis exigidas cuando el

número de observaciones de las muestras es aproximadamente el mismo.

Como libro de ruta se propone que, cuando se verifiquen todas las hipótesis exigidas la alternativa preferida sea el test de la F. Cuando se dé la normalidad pero no la homocedasticidad, se recomienda el uso del *test de Welch* o el *test de Kruskal Wallis*. Si falla, aunque no de forma drástica la normalidad, con valores de p entre $0,01$ y $0,05$, la robustez del test de la F le hace seguir siendo una buena opción. Por último, si fallara fuertemente la normalidad, se recomienda el uso del test de Kruskal Wallis.

Si la conclusión del test aplicado fuera el rechazo de la hipótesis nula, no ocurriría como en el caso de dos poblaciones en el que claramente una de ellas tendría media superior a la otra, sino que habría que evaluar las relaciones entre las k poblaciones, bien dos a dos o a través de combinaciones entre ellas, mediante los denominados *test de comparaciones múltiples*. El resultado final de estas comparaciones desembocará en un mapa de relaciones que, debido a la naturaleza intrínseca de los test, no verificará en general el principio de transitividad.

Existe una gran cantidad de test que realizan las comparaciones múltiples, tratando cada uno de ellos de adaptarse mejor a determinadas circunstancias. Cabe destacar, por ser de uso más extendido, los contrastes de Duncan, Newman-Keuls, Bonferroni, Scheffé y HSD de Tukey. Dependiendo de que las comparaciones sean entre parejas de medias o más generales, combinaciones de las mismas, será más aconsejable el test de Tukey o el de Scheffé. En el caso de comparaciones de parejas de medias, puesto que el de Tukey proporciona intervalos de confianza de menor longitud, se preferirá al de Scheffé.

2. Diagnosis del modelo

Como se ha puesto de manifiesto, los primeros pasos a dar son los de comprobar si las muestras son aleatorias y las poblaciones normales a través de los test descritos en el capítulo anterior. A continuación, si la muestra no está contaminada y no hay desviaciones importantes de normalidad, se comprobará la hipótesis de homocedasticidad y a la vista

de ambas pruebas se elegirá el contraste adecuado. Puesto que ya se han visto los test de aleatoriedad y de normalidad, se dedicará este epígrafe a validar la hipótesis de homocedasticidad. Para ello, se empleará el *test de homogeneidad de varianzas de Barlett*.

Ejemplo 7.1

El archivo `cebada.dat` *contiene información sobre la producción de cuatro variedades de cebada. Utilizando el test de Barlett se estudiará la homocedasticidad de los datos. En* **Rcmdr**, *una vez cargados los datos, se selecciona:* `Estadísticos`→`Varianzas`→`Test de Barlett`, *tomando en la ventana de diálogo, en* `Grupos`, *el factor tipo de cebada,* `tipo`, *y en la* `variable explicada` *la producción de la misma,* `prod`.

```
> bartlett.test(prod~tipo, data=Datos)
Bartlett test of homogeneity of variances
data: prod by tipo
Bartlett's K-squared = 5.9371, df = 3, p-value = 0.1147
```

Dado que `p-valor`$= 0,1147$ *no se rechaza la hipótesis de igualdad de varianzas para los cuatro tipos del factor.*

En muchas ocasiones las muestras que se emplean son de tamaño muy pequeño, menores de 10 elementos, y dado que los test son en general muy conservativos, van a tender a no rechazar la hipótesis nula debido a la escasez de información. Por ello, en este tipo de situaciones, además de la aplicación del contraste para validar la hipótesis, es bueno analizar la naturaleza de los datos. En particular, cuando se trata de validar la normalidad de los datos, si éstos no se han obtenido por un procedimiento de medición sino por observación o conteo, los datos no van a ser intrínsecamente normales aunque pasen el test de normalidad. Para mitigar el problema se recomienda realizar una transformación de los datos. Entre las transformaciones más importantes destacan la raíz cuadrada y la arco seno. La transformación raíz cuadrada se emplea cuando los datos se obtienen a partir de un conteo de elementos, pues en ese caso la distribución de los mismos suele ser de tipo Poisson. Por otra parte, cuando se tienen los datos en forma de tanto por uno, p, es decir que proceden de una binomial, se aconseja la transformación $\text{arcsen}\sqrt{p}$.

3. Test de la F

En este epígrafe se estudiará el contraste de igualdad de medias suponiendo que los datos son normales y homocedásticos. El test que se utilizará será el de la F, que no es sino la generalización del test de la t de student a k poblaciones.

Ejemplo 7.2

Para evaluar el índice de alfabetización de cuatro municipios de una determinada comarca, se ha pasado un test a varios habitantes de cada una de ellas con los siguientes resultados.

Pueblo 1	*Pueblo 2*	*Pueblo 3*	*Pueblo 4*
78	52	82	57
85	48	91	61
90	60	85	45
77	35	74	46
69	51	70	
	47		

Los datos se han recogido en el fichero `alfabeto.dat`*. Suponiendo que los datos son normales y que las varianzas son iguales se aplicará el test de la F. En* ***Rcmdr****, una vez cargados los datos, se selecciona* `Estadísticos`→`Medias`→`ANOVA de un factor...`*, lo que da acceso a la ventana de diálogo del procedimiento donde se indicarán las variables a tratar, obteniendo en* ***Rcmdr*** *la siguiente salida:*

```
> .Anova <- lm(Ind~Pueblo, data=Datos)
> anova(.Anova)
 Analysis of Variance Table
Response: Ind
             Df  Sum Sq  Mean Sq  F value     Pr(> F)
   Pueblo     3  4499.0   1499.7   22.433  5.632e-06 ***
 Residuals   16  1069.6     66.8
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> tapply(Datos$Ind, Datos$Pueblo, mean, na.rm=TRUE) # means
     P1         P2         P3         P4
 79.80000   48.83333   80.40000   52.25000
```

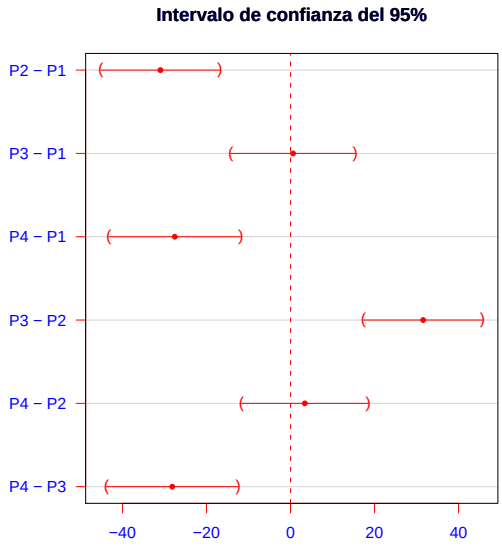

Figura 7.1: Intervalos de confianza de Tukey

```
> tapply(Datos$Ind, Datos$Pueblo, sd, na.rm=TRUE) # std.
deviations
      P1       P2       P3       P4
  8.043631 8.183316 8.443933 7.973916
> tapply(Datos$Ind, Datos$Pueblo, function(x) sum(!is.na(x))) #
counts
 P1  P2  P3  P4
  5   6   5   4
> remove(.Anova)
```

Puesto que el p-valor $< 0,001$ *se rechaza la hipótesis de igualdad de medias en el índice de alfabetización de los cuatro municipios.*

3.1. Comparaciones múltiples

Bajo las mismas hipótesis del test de la F, si se rechaza la hipótesis nula de igualdad de medias se debe proceder a la realización de contrastes de medias dos a dos.

Ejemplo 7.3

Con los datos del ejemplo anterior y puesto que se ha rechazado la hipótesis de igualdad global se realizarán las comparaciones de medias dos a dos. Se accede mediante la misma secuencia de menú, `Estadísticos→Medias→ANOVA de un factor...`*, a la ventana de introducción de datos y opciones, marcando ahora* `Comparaciones dos a dos de las medias`.

Además de la salida anterior **Rcmdr** *crea dos bloques de instrucciones, una que genera la salida numérica de intervalos para las diferencias de medias y otra que construye el gráfico de dichos intervalos.*

Análisis numérico:

El siguiente grupo de instrucciones crea la salida numérica.

```
> .Pairs <- glht(.Anova, linfct = mcp(Pueblo = ''Tukey''))
> confint(.Pairs)
Simultaneous Confidence Intervals for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: lm(formula = Ind~Pueblo, data = Datos)
Estimated Quantile = 2.8607
Linear Hypotheses:
                 Estimate     lwr        upr
  P2 - P1 == 0   -30.9667   -45.1295   -16.8038
  P3 - P1 == 0    0.6000    -14.1926    15.3926
  P4 - P1 == 0   -27.5500   -43.2399   -11.8601
  P3 - P2 == 0    31.5667    17.4038    45.7295
  P4 - P2 == 0    3.4167    -11.6810    18.5143
  P4 - P3 == 0   -28.1500   -43.8399   -12.4601
95% family-wise confidence level
```

El análisis de la salida lleva a que P_1 es igual a P_3 y mayor que P_2 y P_4, que P_2 es igual a P_4 y menor que P_3 y que P_3 es mayor que P_4.

Análisis gráfico:

Por otra parte, el siguiente grupo de instrucciones crea el gráfico de intervalos de confianza para la diferencia de medias (figura 7.1).

```
> old.oma <- par(oma=c(0,5,0,0))
> plot(confint(.Pairs), col= ''red'', main=''Intervalo
de confianza del 95%'',col.main=''blue'', xlab='' '',
col.axis=''blue'')
> par(old.oma)
> remove(.Pairs)
```

4. Alternativa no paramétrica. Test de Kruskal Wallis

Como se ha indicado, si fallan las hipótesis de normalidad y/o homocedasticidad se debe recurrir a una alternativa no paramétrica para realizar el test de igualdad de medias. La solución más extendida la proporciona el test de Kruskal Wallis. Dicho test es una prueba basada en rangos con signos y es una generalización del test de Wilcoxon al caso de k muestras.

Ejemplo 7.4

Suponga que se desea comparar el rendimiento de 5 tipos de neumáticos, A, B, C, D y E, para lo que decide probarlos en distintos coches de similares características. Sus vidas medias en rodaje, medidas en miles de kilómetros, vienen dadas en la siguiente tabla:

Llantas	*Vidas medias*				
A	68	72	77	42	53
B	72	53	63	53	48
C	60	82	64	75	72
D	48	61	57	64	50
E	64	65	70	68	53

Para contrastar que no hay diferencias entre los cinco tipos de neumáticos se elige el test de Kruskal Wallis. Los datos han sido almacenados en el fichero `neumaticos.dat` *dentro del repositorio de datos. En* **Rcmdr** *se activa la secuencia de menú* `Estadísticos`→ `Test no paramétricos`→`Test de Kruskal Wallis`, *abriéndose la correspondiente ventana de diálogo donde se seleccionan variable y factor, en este caso* `Km` *y* `Neum`. **Rcmdr** *proporciona en primer lugar las medianas de cada grupo y seguidamente el estadístico de Kruskal Wallis junto con su p-valor.*

```
> tapply(DatosKm, DatosNeum, median, na.rm=TRUE)
  A   B   C   D   E
 68  53  72  57  65
```

```
> kruskal.test(Km~Neum, data=Datos)

Kruskal-Wallis rank sum test
data: Km by Neum
Kruskal-Wallis chi-squared = 6.4949, df = 4, p-value = 0.1651
```

A la vista de los resultados, p-valor $= 0,1651$*, se concluye que no hay diferencias significativas entre los rendimientos de los cinco tipos de neumáticos.*

5. Ejercicios

7.1 Estudie, a partir de la tabla de datos porcentuales que se da, si las medias de los tres niveles de un determinado factor son iguales.

Nivel I	Nivel II	Nivel III
8,1	8,6	12
9,2	8,9	13,2
9,5	7,4	13,1

7.2 Una empresa tiene en un establecimiento cuatro vendedores y pretende asignar primas en función de las ventas. A la vista de la tabla de ventas en los últimos cinco meses (miles de euros), indique si los cuatro vendedores son igualmente eficaces. De no ser así elabore el ranking en razón de las ventas.

Vend. 1	Vend. 2	Vend. 3	Vend. 4
6,46	5,79	8,37	4,94
4,83	5,13	7,57	4,11
5,89	6,17	8,69	5,45
5,30	4,72	8,06	5,21
6,33	5,60	7,23	5,00

7.3 A partir de la cuenta de resultados que presentaban 13 entidades financieras englobadas en los ámbitos europeo, nacional y regional se ha calculado el porcentaje destinado a la generación bruta de fondos, con los siguientes resultados:

Ámbito	Generación bruta de fondos
Europeo	0,4 3,8 2,5 2,9
Tipo II	4,7 2,0 1,8 2,8
Tipo III	0,9 3,7 3,1 6,2 2,7

¿Puede considerarse que la proporción de fondos es igual indepen-

dientemente del ámbito de actuación?

7.4 Una cierta planta ha sido cultivada con cinco fertilizantes distintos. Se desea estudiar si el tipo de fertilizante influye en la longitud de la planta, para lo cual se han medido las longitudes de cinco series de 10 plantas, obteniéndose para cada serie los resultados que aparecen en el fichero plantas.dat. ¿Hay evidencia estadística suficiente para afirmar que las medias son diferentes? De ser así, ¿existen tipos de fertilizante que no se diferencien entre sí?

7.5 Un fabricante está interesado en la resistencia a la tensión de una fibra sintética. Se sospecha que la resistencia está relacionada con el porcentaje de algodón en la fibra. Suponer que la distribución para cada porcentaje son aproximadamente normales y se da la homogeneidad de las varianzas. Para ello, se emplean cinco niveles de porcentaje de algodón. De 5 réplicas aleatorias se obtienen los siguientes datos:

Porcentaje de algodón	1	2	3	4	5
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11

¿Puede considerarse que la resistencia de las prendas es la misma independiente del porcentaje de algodón presente en sus fibras?