ANÁLISIS DE LA VARIANZA
PARTE PRIMERA

Febrero de 2012

Índice general

1. INTRODUCCIÓN

Los modelos de Análisis de la Varianza **(ANOVA)** fueron inicialmente desarrollados por R.A. Fisher como método de análisis de los resultados obtenidos en los **Diseños Experimentales.** En dichos modelos, la variable dependiente de tipo continuo **(variable de respuesta)** se expresa en función de las variables independientes **(factores)** que son variables de tipo categórico. Esto último es lo que los diferencia de los modelos de Regresión dado que en éstos las variables independientes son de tipo continuo.

Definido el modelo, el método descompone la variabilidad total existente en la variable de respuesta según las distintas fuentes de variación consideradas en el modelo: Por una parte, las variabilidades correspondientes a cada factor, las variabilidades originadas por la acción combinada de los diferentes factores **(interacciones)** y por otra, una variabilidad residual originada por el conjunto de los restantes factores, conocidos o desconocidos, que no han sido introducidos en el modelo.

Para distinguir cuales son los factores o interacciones con influencia significativa, se efectúan contrastes mediante el estadístico **"F" de Fisher-Snedecor**. Dicho estadístico también se utiliza para discernir si el modelo, en su conjunto, explica una parte significativa de la variabilidad presente en la respuesta. Tanto la variabilidad total como cada una de las variabilidades en la que ésta se descompone, se disponen en una tabla denominada **Tabla de Análisis de la Varianza**. Dicha tabla recoge, además, el número de grados de libertad (g.d.l.) que corresponden a cada fuente de variación, las medias cuadráticas y los resultados de los contrastes "F". Aunque trataremos a continuación con más extensión sobre el asunto, anticipamos que los

g.d.l. asignados a cada fuente de variación corresponden al número de datos **independientes** que se emplean en su cálculo.

Cuando el diseño es complejo por presencia conjunta en el modelo de factores de diversos tipos (fijos, aleatorios, anidados, etc.) es habitual y útil añadir a la tabla una columna con los valores de la Esperanza media cuadrática. De esta forma, como veremos, tendremos un criterio claro para el cálculo de los estadísticos "F".

Aunque el método "ANOVA" es también aplicable a los casos en que se desea extraer información útil a partir de una base de datos ya existente y en cuya generación el investigador no ha participado **(Análisis Observacional),** nuestra atención se centrará en los casos, mucho más importantes, en los que se utiliza el Análisis de la Varianza para analizar e interpretar los resultados obtenidos al combinar **intencionadamente** los diferentes factores, es decir, para analizar los resultados obtenidos en la realización de un **Diseño Experimental.**

2. FUNDAMENTOS DEL DISEÑO COMPLETAMENTE ALEATORIZADO

El modelo ANOVA más sencillo es el que está ligado al Diseño Completamente Aleatorizado. Dicho modelo considera solamente un factor como causa de variabilidad. Este factor está constituído por varios niveles fijos. La ecuación del modelo es la siguiente:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

y_{ij} : Son los valores obtenidos para la variable de respuesta. El subindice "i" indica los diferentes niveles del factor: 1, 2 , ... k. El subindice "j" se refiere al número de orden en la repetición: 1,2 ... r de un determinado nivel "i".

μ : Valor medio de la variable de respuesta.

τ_i : Es el efecto del nivel "i". Para un nivel "i" con un valor medio μ_i, se define el efecto τ_i como la diferencia $\mu_i - \mu$.

ε_{ij} : Término residual que se supone conforme con una distribución Normal de media cero y desviación típica σ que representaremos por

$N(0, \sigma)$. El modelo supone que esta distribución es idéntica para todos los niveles.

Los verdaderos valores de los parámetros: μ, μ_i, τ_i y σ son desconocidos y para simplificar supondremos que el número de repeticiones "r" de cada nivel es el mismo. Tras realizar "n" experimentos elementales: $n = k \cdot r$ obtendremos las estímaciones de los parámetros: $\hat{\mu}, \hat{\mu}_i, \hat{\tau}_i$ y $\hat{\sigma}$, la variabilidad total y las variabilidades en que ésta se descompone.

Resaltamos el hecho de que los "n" experimentos deben ser efectuados en un orden **completamente aleatorizado**. De otra forma, el término residual podría no asimilarse a una distribución Normal N (o, σ) de valores independientes y el análisis no sería correcto. Los programas informáticos, entre ellos el programa estadístico STATGRAPHICS que es el que utilizamos, facilitan automáticamente el orden aleatorio en la ejecución de los "n" experimentos.

La siguiente tabla representa los resultados de los "n" experimentos elementales agrupados en los "k" niveles del factor, los totales y los valores medios en cada nivel así como el total general T y la media general $(\overline{y})$.

Factor (k niveles)	Resultados (r repeticiones)	Totales	Medias
1	$y_{11}y_{12}...y_{1r}$	T_1	$\overline{y_1}$
2	$y_{21}y_{22}...y_{2r}$	T_2	$\overline{y_2}$
3	$y_{31}y_{32}...y_{3r}$	T_3	$\overline{y_3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$y_{k1}y_{k2}...y_{kr}$	T_k	$\overline{y_k}$
		T	$\overline{y}$

La estimación para el valor medio será: $\hat{\mu} = \overline{y}$ que se calcula según las expresiones:

$$\overline{y} = \frac{T}{n} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{r} y_{ij}}{n} = \frac{\sum_{i=1}^{k} \overline{y_i}}{k}$$

La variabilidad total presente en los resultados vendrá dada por la suma de los cuadrados (SC) de las desviaciones de los resultados respecto de la media general que representaremos por $SC_{TOTAL} = \sum_{i=1}^{k} \sum_{j=1}^{r} (y_{ij} - \overline{y})^2$

La variabilidad asignable al único factor con k niveles (k tratamientos aplicados cada uno r veces) estará representada por $SC_{TRATAMIENTOS}$ y vendrá dada por la suma de los cuadrados de las desviaciones de las medias de cada nivel a la media general:

$$SC_{TRATAMIENTOS} = \sum_{i=1}^{k} \sum_{j=1}^{r} (\overline{y_i} - \overline{y})^2 = r \sum_{i=1}^{k} (\overline{y_i} - \overline{y})^2$$

La variabilidad residual está representada por $SC_{RESIDUAL}$ y vendrá dada por la suma de los cuadrados de las desviaciones de los resultados **dentro** de cada nivel respecto de su correspondiente media:

$$SC_{RESIDUAL} = \sum_{i=1}^{k} \sum_{j=1}^{r} (y_{ij} - \overline{y_i})^2$$

Demostraremos a continuación que la variabilidad total $SC_{TOTAL} = \sum_{i=1}^{k} \sum_{j=1}^{r} (y_{ij} - \overline{y})^2$ equivale a la suma de la variabilidad **entre los tratamientos** $SC_{TRATAMIENTOS} = \sum_{i=1}^{k} \sum_{j=1}^{r} (\overline{y_i} - \overline{y})^2$ y la variabilidad **dentro de los tratamientos** o **variabilidad residual** $SC_{RESIDUAL} = \sum_{i=1}^{k} \sum_{j=1}^{r} (y_{ij} - \overline{y_i})^2$. Esta igualdad:

$$\mathbf{SC_{TOTAL}} = \mathbf{SC_{TRATAMIENTOS}} + \mathbf{SC_{RESIDUAL}} \tag{1}$$

es la base del Análisis de la varianza.

Para efectuar la demostración partiremos de la expresión SC_{TOTAL} a la que sumaremos y restaremos $\overline{y_i}$ que es el valor medio del tratamiento "i":

$$SC_{TOTAL} = \sum_{i=1}^{k} \sum_{j=1}^{r} (y_{ij} - \overline{y})^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} (y_{ij} - \overline{y} + \overline{y_i} - \overline{y_i})^2$$

Reagrupando convenientemente la última expresión quedará:

$$SC_{TOTAL} = \sum_{i=1}^{k} \sum_{j=1}^{r} [(\overline{y_i} - \overline{y}) + (y_{ij} - \overline{y_i})]^2$$

$$SC_{TOTAL} = \sum_{i=1}^{k}\sum_{j=1}^{r}(\overline{y_i} - \overline{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{r}(y_{ij} - \overline{y_i})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{r}(\overline{y_i} - \overline{y})(y_{ij} - \overline{y_i}) \quad (2)$$

Vamos a comprobar que el término $\sum_{i=1}^{k}\sum_{j=1}^{r}(\overline{y_i} - \overline{y})(y_{ij} - \overline{y_i})$ se anula ya que es equivalente a $\sum_{i=1}^{k}(\overline{y_i} - \overline{y})\sum_{j=1}^{r}(y_{ij} - \overline{y_i})$ donde el primer factor es nulo dado que:

$$\sum_{i=1}^{k}(\overline{y_i} - \overline{y}) = (\overline{y_1} - \overline{y}) + (\overline{y_2} - \overline{y}) + \cdots + (\overline{y_k} - \overline{y}) =$$
$$(\overline{y_1} + \overline{y_2} + \cdots + \overline{y_k}) - k \cdot \overline{y} = k\overline{y} - k\overline{y} = 0$$

Al anularse el tercer término de la expresión (2) queda:

$$SC_{TOTAL} = \sum_{i=1}^{k}\sum_{j=1}^{r}(\overline{y_i} - \overline{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{r}(y_{ij} - \overline{y_i})^2$$

Puesto que $SC_{TRATAMIENTOS} = \sum_{i=1}^{k}\sum_{j=1}^{r}(\overline{y_i} - \overline{y})^2$ y $SC_{RESIDUAL} = \sum_{i=1}^{k}\sum_{j=1}^{r}(y_{ij} - \overline{y_i})^2$ queda demostrada la igualdad (1):

$$\mathbf{SC_{TOTAL} = SC_{TRATAMIENTOS} + SC_{RESIDUAL}}$$

En cuanto a los grados de libertad (g.d.l.) también se cumple la igualdad:

$$\underset{\text{(n - 1)}}{\nu_{TOTAL}} = \underset{\text{(k - 1)}}{\nu_{TRATAMIENTOS}} + \underset{\text{(n - k)}}{\nu_{RESIDUAL}}$$

ν_{TOTAL} : Los g.d.l. en la $SC_{TOTAL} = n - 1$

$\nu_{TRATAMIENTOS}$: Los g.d.l. en la $SC_{TRATAMIENTOS} = k - 1$

$\nu_{RESIDUAL}$: Los g.d.l. en la $SC_{RESIDUAL} = n - k$

Los grados de libertad asignados a cada fuente de variación proceden de lo siguiente:

SC_{TOTAL}: (n - 1) g.d.l. ya que de las "n" desviaciones $(y_{ij} - \overline{y})$ existentes en $SC_{TOTAL} = \sum_{i=1}^{k} \sum_{j=1}^{r} (y_{ij} - \overline{y})^2$ tan solo (n-1) son independientes puesto que la suma de las "n" desviaciones $\sum(y_{ij} - \overline{y}) = 0$ y dadas (n-1) de ellas, la desviación restante queda impuesta.

$SC_{TRATAMIENTOS}$: (k -1) g.d.l. puesto que de los "k" efectos $\hat{\tau}_i = (\overline{y_i} - \overline{y})$ existentes en $SC_{TRATAMIENTOS} = \sum_{i=1}^{k} \sum_{j=1}^{r} (\overline{y_i} - \overline{y})^2 = r \sum_{i=1}^{k} (\overline{y_i} - \overline{y})^2 = r \sum_{i=1}^{k} \hat{\tau}_i{}^2$ tan sólo (k-1) son independientes. Al existir la restricción de que la suma de los "k" efectos ha de ser nula: $\sum(\overline{y_i} - \overline{y}) = 0$, dados k-1 de ellos, el efecto restante queda impuesto.

$SC_{RESIDUAL}$: (n-k) g.d.l. De los "n" residuales existentes en $SC_{RESIDUAL} = \sum_{i=1}^{k} \sum_{j=1}^{r} (y_{ij} - \overline{y_i})^2$ tan solo (n-k) son independientes. Dentro de cada uno de los "k" niveles, existen "r" residuales y solamente (r-1) de ellos son independientes puesto que se da la restricción $\sum(y_{ij} - \overline{y_i}) = 0$ y dados (r-1) residuales de un nivel, el residual restante queda impuesto. Habrá, por tanto, k(r-1) residuales independientes equivalentes a (n-k):

$$\text{k(r-1)} = \text{kr-k} = \text{(n-k) g.d.l.}$$

Por otra parte, las medias cuadráticas (MC) de cada fuente de variación se calculan a partir de las sumas de cuadrados (SC) correspondientes dividiéndolas por sus respectivos g.d.l.(ν): $MC = \dfrac{SC}{\nu}$. Todos estos valores se recogen en la llamada **Tabla de Análisis de la Varianza:**

Fuente de Variación	Sumas de cuadrados	(g.d.l.)	Medias Cuadráticas
TRATAMIENTOS	$SC_{TRAT} = r \sum_{i=1}^{k} (\overline{y_i} - \overline{y})^2$	(k-1)	$MC_{TRAT.} = \dfrac{SC_{TRAT}}{k-1}$
RESIDUAL	$SC_{RESIDUAL} = \sum_{i=1}^{k} \sum_{j=1}^{r} (y_{ij} - \overline{y_i})^2$	(n-k)	$MC_{RESIDUAL} = \dfrac{SC_{RESIDUAL}}{n-k}$
TOTAL	$SC_{TOTAL} = \sum_{i=1}^{k} \sum_{j=1}^{r} (y_{ij} - \overline{y})^2$	(n-1)	

Tanto la MC_{TRAT} como la $MC_{RESIDUAL}$ son variables aleatorias. Si aplicamos a ambas el concepto estadístico de **Esperanza Matemática (E)** obtendremos los valores esperados de dichas variables que serán la Esperanza media cuadrática de los tratamientos (EMC_{TRAT}) y la Esperanza media cuadrática residual ($EMC_{RESIDUAL}$):

$$EMC_{TRAT} = E\left[\frac{r\sum_{i=1}^{k}(\overline{y_i}-\overline{y})^2}{k-1}\right]$$

$$EMC_{RESIDUAL} = E\left[\frac{\sum_{i=1}^{k}\sum_{j=1}^{r}(y_{ij}-\overline{y_i})^2}{n-k}\right]$$

Sustituyendo y_{ij} por la ecuación del modelo $y_{ij} = \mu+\tau_i+\varepsilon_{ij}$ y teniendo en cuenta que $E(\varepsilon_{ij}^2) = \sigma^2$ y que $E(\varepsilon_{ij}) = 0$ resultan tras un laborioso cálculo que omitimos, las siguientes expresiones:

$$EMC_{TRAT} = \sigma^2 + \frac{r\sum_{i=1}^{k}{\tau_i}^2}{k-1}$$

$$EMC_{RESIDUAL} = \sigma^2$$

Cuando los modelos son más complejos que el presente, el cálculo de las esperanzas se hace aún más largo. Por ello se recurre a algoritmos como el que presentaremos en otro artículo.

Como veremos en el ejemplo que sigue, los programas estadísticos informáticos calculan las Esperanzas medias cuadráticas de cada una de las fuentes de variabilidad. La determinación de dichas esperanzas es fundamental para efectuar correctamente los contrastes de hipótesis mediante el estadístico F de Fisher-Snedecor.

Si enunciamos para nuestro modelo: $y_{ij} = \mu+\tau_i+\varepsilon_{ij}$ la hipótesis nula (Ho) consistente en suponer que todos los efectos son nulos: $\tau_1 = \tau_2 = ... = \tau_k = 0$, tanto la $EMC_{TRAT} = \sigma^2 + \frac{r\sum_{i=1}^{k}\tau_i^2}{k-1}$ como la $EMC_{RESIDUAL} = \sigma^2$ quedan reducidas a σ^2 que es la varianza del modelo.

Bajo dicha hipótesis nula, las medias cuadráticas siguen distribuciones ji-cuadrado (χ^2) y su cociente se ajustará a una distribución F de Fisher-Snedecor. Es decir, supuesto que se cumple Ho, la $MC_{TRAT.} = \frac{SC_{TRAT}}{k-1}$ sigue una distribución χ^2 con (k-1) g.d.l. mientras que $MC_{RESIDUAL} = \frac{SC_{RESIDUAL}}{n-k}$ sigue otra distribución χ^2 con (n-k) g.d.l. En consecuencia,

el cociente $\frac{MC_{TRAT.}}{MC_{RESIDUAL}}$ seguirá una distribución F de Fisher-Snedecor con (k-1) y (n-k) g.d.l.

En el caso de que el ratio F sea superior al valor crítico correspondiente al nivel de significación α, habrá que rechazar la hipótesis nula $H_0 : \tau_i = 0$. Es decir, se rechaza H_0 cuando el ratio F $> F_\alpha$, k-1, n-k. Entonces, alguno de los efectos $\tau_i \neq 0$ y, logicamente, $EMC_{TRAT} = \sigma^2 + \frac{r\sum_{i=1}^{k} {\tau_i}^2}{k-1}$ sería significativamente mayor que σ^2.

Los valores de las EMC se incluyen también en la tabla de Análisis de la Varianza que para el presente modelo: $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ quedará finalmente:

Fuente de Variación	Suma de cuadrados	g.d.l	Medias cuadráticas	Esperanzas medias cuadráticas	F
TRATAMIENTOS	$SC_{TRAT} = r\sum_{i=1}^{k}(\overline{y_i} - \overline{y})^2$	(k-1)	$MC_{TRAT} = \frac{SC_{TRAT}}{k-1}$	$\sigma^2 + \frac{r\sum_{i=1}^{k}\tau_i^2}{k-1}$	$\frac{MC_{TRAT}}{MC_{RESIDUAL}}$
RESIDUAL	$SC_{RESIDUAL} = \sum\sum(y_{ij} - \overline{y_i})^2$	(n-k)	$MC_{RESIDUAL} = \frac{SC_{RESIDUAL}}{n-k}$	σ^2	
TOTAL	$SC_{TOTAL} = \sum_{i=1}^{k}\sum_{j=1}^{r}(y_{ij} - \overline{y})^2$	(n-1)			

Aunque los programas informáticos facilitan todos los componentes de la tabla de Análisis de la Varianza, es conveniente conocer algunas expresiones de las sumas de cuadrados más apropiadas para el cálculo manual:

$$SC_{TOTAL} = \sum_{i=1}^{k}\sum_{j=1}^{r} y_{ij}^2 - \frac{T^2}{n}$$

$$SC_{TRAT} = \frac{1}{r}\sum_{i=1}^{k} T_i^2 - \frac{T^2}{n}$$

donde T es el total general, T_i es el Total en cada nivel, "k" es número de niveles, "r" el número de repeticiones en cada nivel y $n = k \cdot r$.

La suma de cuadrados residual se calcula por diferencia:

$$SC_{RESIDUAL} = SC_{TOTAL} - SC_{TRAT}.$$

Intervalos de confianza de los parámetros:

En general, para definir el intervalo de confianza bilateral de un parámetro θ con un nivel de confianza $(1-\alpha)$ siendo α el nivel de significación, se utiliza la siguiente expresion:

$$\hat{\theta} \pm \text{ (Error estándar del parametro)} \cdot t_{\frac{\alpha}{2}}, \nu$$

donde $\hat{\theta}$ es la estima del parámetro, $t_{\frac{\alpha}{2}}, \nu$ es el valor de la distribución t de Student para ν g.d.l. y nivel de significación α.

El valor de ν corresponde a los g.d.l. del residual del modelo ajustado que en nuestro caso es $\nu = n - k$. El valor del error estándar dependerá de cual sea el parámetro y se calcula, siempre, a partir de la estima de la desviación típica del modelo: $\hat{\sigma}$. Por ejemplo, tendremos:

$$\text{Error estándar de } \mu = \frac{\hat{\sigma}}{\sqrt{n}}; \text{ Error estándar de } \mu_i = \frac{\hat{\sigma}}{\sqrt{r}}$$

$$\text{Error estándar del contraste } (\mu_i - \mu_j) = \sqrt{\frac{2}{r}} \cdot \hat{\sigma}$$

3. EJEMPLO DE DISEÑO COMPLETAMENTE ALEATORIZADO: POROSIDAD DEL COQUE

En el presente ejemplo se aplican los fundamentos expuestos en el punto anterior a un caso industrial en el que se quiere comparar la porosidad del coque obtenido en la destilación de tres mezclas diferentes de carbones. La porosidad del coque es una importante característica con vistas a su utilización en el Horno Alto o en la Fundición.

Los resultados se obtuvieron tras planificar y ejecutar un Diseño Completament Aleatorizado con un solo factor (mezcla de carbón) a k=3 niveles y con r=5 repeticiones en cada nivel. La variable dependiente es el porcentaje de porosidad del coque que ha sido determinada por picnometría. En total se han efectuado n=kr=15 experimentos elementales de coquización de las

mezclas y determinación de la porosidad del coque. Los resultados se analizaron mediante Análisis de la Varianza para un solo factor. Vemos que el método de análisis **(Análisis de la Varianza simple)** está ligado al diseño al que se aplica **(Diseño Completamente aleatorizado).**

Los cálculos se han efectuado con el programa estadístico STATGRAPHICS. La discusión de resultados, conclusiones y cálculos se recogen en las páginas 11 a 18 y la base de datos en las páginas 15 y 18.

DISCUSIÓN DE RESULTADOS

El modelo de Análisis de la Varianza (ANOVA) al que se ajustan los resultados obtenidos es:

Porosidad $= \mu + \tau_i + \varepsilon$ equivalente a Porosidad $= \mu_i + \varepsilon$. Los parámetros del modelo son los siguientes: μ es la media general, τ_i son los efectos, μ_i el valor medio en cada nivel y ε es el término residual que sigue una distribución Normal $(0, \sigma)$.

- Los 15 experimentos elementales han sido efectuados en orden aleatorio según lo indicado en la pag 15.

  El único factor considerado (factor A) tiene tres niveles: 1,2,3 correspondientes a cada mezcla de carbón y clase de coque resultante.

  De acuerdo con los resultados de la página 16, los valores estimados para los parámetros son los siguientes:

$$\begin{array}{lll} \hat{\mu} = 48,4193 & \hat{\tau}_1 = 0,7067 & \\ \hat{\mu_1} = 49,126 & \hat{\tau}_2 = 0,3187 & \hat{\sigma} = \sqrt{MC_{RESIDUAL}} \\ \hat{\mu_2} = 48,738 & \hat{\tau}_3 = -1,0253 & \hat{\sigma} = \sqrt{0,88816} = 0,94242 \\ \hat{\mu_3} = 47,394 & & \end{array}$$

  Obsérvese que los efectos estimados $(\hat{\tau}_i)$ son las diferencias de los valores medios estimados para cada nivel $(\hat{\mu}_i)$ y la media general estimada $(\hat{\mu})$. Se aprecia, también, que los efectos están sometidos a la restricción $\sum \hat{\tau}_i = 0$. Por ello, los efectos (tratamientos) tienen 2 g.d.l. ya que dados dos de ellos, el tercero queda impuesto por la restricción.

  La estima $\hat{\sigma}$ de la desviación típica del modelo nos permite calcular intervalos de confianza para los diferentes parámetros: μ, μ_i así como para los contrastes $(\mu_i - \mu_j)$.

- En la tabla de Análisis de la Varianza (pag 16) se recogen las Sumas de cuadrados, grados de libertad, medias cuadráticas y el ratio F. Estos valores coinciden con los que pueden obtenerse por cálculo:

$$SC_{TRAT} = r \sum_{i=1}^{3} \hat{\tau_i^2} = 5 \cdot 1,65223 = 8,26117$$

$$SC_{TOTAL} = \sum_{i=1}^{3} \sum_{j=1}^{5} (y_{ij} - \overline{y})^2 = \sum_{i=1}^{3} \sum_{j=1}^{5} (y_{ij} - 48,4193)^2 = 18,9191$$

$$SC_{RESIDUAL} = SC_{TOTAL} - SC_{TRAT} = 10,6579$$

- Los g.d.l. de la SC_{TOTAL} son 14, ya que de las 15 desviaciones a la media total (ver última columna de la pag 18) tan solo 14 de ellas son independientes puesto que su suma es nula.

- De los 15 valores residuales obtenidos del modelo (pag 18) tan solo 12 son independientes ya que están sometidos a tres restricciones (dentro de cada nivel, la suma de residuales es nula). De cada 5 residuales por nivel, solamente 4 son independientes estando el quinto impuesto. En total habrá 4x3=12 residuales independientes, es decir, 12 g.d.l. Como comprobación, puede verse que $\sum residuales^2 = 10,6579$

- Las medias cuadráticas y el ratio F de la tabla de Análisis de la Varianza (pag 16) coinciden con:

$$MC_{TRAT} = \frac{SC_{TRAT}}{g.d.l} = \frac{8,26117}{2} = 4,13059$$

$$MC_{RESIDUAL} = \frac{SC_{RESIDUAL}}{g.d.l} = \frac{10,6579}{12} = 0,88816$$

$$F = \frac{MC_{TRAT}}{MC_{RESIDUAL}} = \frac{4,13059}{0,88816} = 4,65$$

Al ratio F=4,65 le corresponde, bajo la hipótesis nula (H_0) de que todos los efectos son nulos, una probabilidad del 3,20 % que es inferior al nivel de significación 5 %. Por ello, se infiere que existen diferencias significativas entre los niveles y que algunos de los efectos no son nulos.

En la pag 17 se recogen los valores de las Esperanzas medias cuadráticas de los tratamientos y de los residuales así como los intervalos de confianza para la media general μ y los valores medios de los tres niveles. Por otra parte tenemos:

$$\text{Desviación típica del modelo: } \hat{\sigma} = 0,94242$$
$$\text{Error estándar de la media total: } \frac{\hat{\sigma}}{\sqrt{15}} = 0,24333$$
$$\text{Error estándar de las medias de los niveles} = \frac{\hat{\sigma}}{\sqrt{5}} = 0,42146$$

Intervalos de confianza

a) **Para la media general:** $\hat{\mu} \pm \frac{\hat{\sigma}}{\sqrt{n}} \cdot t_{\frac{\alpha}{2}}, \nu$ donde $\hat{\mu} = 48,4193$; $\frac{\hat{\sigma}}{\sqrt{n}} = 0,24333$; para $\alpha = 0,05$ y $\nu = 12$ resulta $t_{\frac{\alpha}{2}}, \nu = 2,179$. Con ello el intervalo será:

$$48,4193 \pm 0,24333 \cdot 2,179 = 48,4193 \pm 0,530216$$
$$47,8892 < \mu < 48,9495$$

b) **Para las medias de los niveles μ_i :** $\hat{\mu_i} \pm \frac{\hat{\sigma}}{\sqrt{n}} \cdot t_{\frac{\alpha}{2}}, \nu$

Nivel $\mu_1 : 49,126 \pm 0,42146 \cdot 2,179 = 49,126 \pm 0,91836$
Nivel $\mu_2 : 48,738 \pm 0,42146 \cdot 2,179 = 48,738 \pm 0,91836$
Nivel $\mu_3 : 47,394 \pm 0,42146 \cdot 2,179 = 47,394 \pm 0,91836$

48,2077 $<\mu_1 <$50,0443
47,8197 $<\mu_2 <$49,6563
46,4757 $<\mu_3 <$48,3123

Todos estos intervalos coincidan con los señalados en la tabla al final de la pag 17

c) **Para los contrastes $\mu_i - \mu_j$:**

$$\text{Error estándar de los contrastes} = \sqrt{\frac{2 \cdot MC_{RESIDUAL}}{r}}$$
$$\text{Error estándar} = \sqrt{\tfrac{2 \cdot 0,88816}{5}} = 0,59604; t_{\frac{\alpha}{2}}, \nu = 2,179$$

$$\text{Intervalos } (\mu_i - \mu_j) = (\hat{\mu}_i - \hat{\mu}_j) \pm \text{Error estándar} \cdot t_{\frac{\alpha}{2}, \nu}$$

$$\text{Intervalo } (\mu_1 - \mu_2) : (\hat{\mu_1} - \hat{\mu_2}) \pm 0,59604 \cdot 2,179 = (\hat{\mu_1} - \hat{\mu_2}) \pm 1,2987$$

$$\text{Intervalo } (\mu_1 - \mu_3) : (\hat{\mu_1} - \hat{\mu_3}) \pm 0,59604 \cdot 2,179 = (\hat{\mu_1} - \hat{\mu_3}) \pm 1,2987$$

$$\text{Intervalo } (\mu_2 - \mu_3) : (\hat{\mu_2} - \hat{\mu_3}) \pm 0,59604 \cdot 2,179 = (\hat{\mu_2} - \hat{\mu_3}) \pm 1,2987$$

Para $(\mu_1 - \mu_2) = 0{,}388 \pm 1{,}2987$

Para $(\mu_1 - \mu_3) = 1{,}732 \pm 1{,}2987$

Para $(\mu_2 - \mu_3) = 1{,}344 \pm 1{,}2987$

El primero de los tres intervalos comprende el valor cero por lo que no denota una diferencia significativa entre los niveles 1 y 2. Sin embargo, los dos restantes señalan, al no contener el valor cero, diferencias significativas tanto entre los niveles 1 y 3 como entre los 2 y 3.

Todos estos resultados coinciden con los señalados al final de la página 16.

Finalmente, si reflejamos los resultados según la disposición de la tabla de la pag 3:

Niveles del factor	Observaciones	Totales	Medias
1	51,24 49,61 47,77 48,42 48,59	245,63	49,126
2	48,34 47,64 48,86 49,86 48,99	243,69	48,738
3	47,28 47,79 46,83 47,31 47,76	236,97	47,394
		726,26	48,4193

y aplicamos las fórmulas de la pag 8, podemos obtener mediante **cálculo manual** lo siguiente:

$$SC_{TOTAL} = \sum\sum y_{ij}^2 - \frac{t^2}{n} = 35185,3967 - \frac{726,29^2}{15} = 18,9191$$

$$SC_{TRAT} = \frac{1}{r}\sum T_i^2 - \frac{T^2}{n} = \frac{175873,6939}{5} - \frac{726,29^2}{15} = 8,26117$$

$$SC_{RESIDUAL} = SC_{TOTAL} - SC_{TRAT} = 10,6579$$

Obtenidas las sumas de cuadrados, se podría continuar el análisis manualmente según lo ya expuesto.

CONCLUSIONES

Como conclusiones a la discusión efectuada en el ejemplo destacamos lo siguiente:

- Los valores asignados por el modelo ANOVA a cada nivel para la porosidad son, simplemente, los respectivos valores medios muestrales: $\hat{\mu_1} = 49,126; \hat{\mu_2} = 48,738$ y $\hat{\mu_3} = 47,394$.
- A partir de la estimación de la desviación típica del modelo: $\hat{\sigma} = 0,94242$ se han establecido intervalos de confianza del 95 % para los valores medios de los tres niveles, efectuado contrastes y comprobado que es significativa la diferencia de porosidad entre los coques 1 y 3 así como la existente entre los coques 2 y 3. También, se ha contrastado que no hay diferencia significativa entre la porosidad de los coques 1 y 2.
- El modelo ha exigido la aleatorización completa en el orden de efectuar los experimentos elementales (ver orden en pag 15).

run	Factor_A	POROSIDAD (%)
1	3	47.28
2	3	47.79
3	1	51.24
4	2	48.34
5	2	47.64
6	1	49.61
7	3	46.83
8	1	47.77
9	3	47.31
10	1	48.42
11	2	48.86
12	2	49.86
13	3	47.76
14	2	48.99
15	1	48.59

ANOVA Table for POROSIDAD by Factor_A

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	8.26117	2	4.13059	4.65	0.0320
Within groups	10.6579	12	0.88816		
Total (Corr.)	18.9191	14			

Table of Means for POROSIDAD by Factor_A

Factor_A	Count	Mean
1	5	49.126
2	5	48.738
3	5	47.394
Total	15	48.4193

Multiple Range Tests for POROSIDAD by Factor_A

Method: 95.0 percent LSD

Factor_A	Count	Mean	Homogeneous Groups
3	5	47.394	X
2	5	48.738	X
1	5	49.126	X

Contrast	Difference	+/- Limits
1 - 2	0.388	1.29866
1 - 3	*1.732	1.29866
2 - 3	*1.344	1.29866

* denotes a statistically significant difference.

```
General Linear Models
---------------------
Number of dependent variables: 1
Number of categorical factors: 1
Number of quantitative factors: 0


Analysis of Variance for POROSIDAD
-------------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-------------------------------------------------------------------------------
Model                    8.26117      2      4.13059       4.65       0.0320
Residual                 10.6579     12      0.88816
-------------------------------------------------------------------------------
Total (Corr.)            18.9191     14

Type III Sums of Squares
-----------------------------------------------------------------------------------------
Source                       Sum of Squares     Df   Mean Square    F-Ratio    P-Value
-----------------------------------------------------------------------------------------
Factor_A                           8.26117      2       4.13059       4.65     0.0320
Residual                           10.6579     12       0.88816
-----------------------------------------------------------------------------------------
Total (corrected)                  18.9191     14

Expected Mean Squares
-----------------------------------------------------------------------------------------
Source                       EMS
-----------------------------------------------------------------------------------------
Factor_A                     (2)+Q1
Residual                     (2)
-----------------------------------------------------------------------------------------

F-Test Denominators
-----------------------------------------------------------------------------------------
Source                       Df        Mean Square     Denominator
-----------------------------------------------------------------------------------------
Factor_A                     12.00     0.88816         (2)
-----------------------------------------------------------------------------------------

Variance Components
-----------------------------------------------------------------------------------------
Source                       Estimate
-----------------------------------------------------------------------------------------
Residual                     0.88816
-----------------------------------------------------------------------------------------

R-Squared = 43.6658 percent
R-Squared (adjusted for d.f.) = 34.2768 percent
Standard Error of Est. = 0.942422

Table of Least Squares Means for POROSIDAD
with 95.0 Percent Confidence Intervals
-------------------------------------------------------------------------------
                                          Stnd.       Lower       Upper
Level                  Count    Mean      Error       Limit       Limit
-------------------------------------------------------------------------------
GRAND MEAN             15       48.4193   0.243332    47.8892     48.9495
Factor_A
1                      5        49.126    0.421464    48.2077     50.0443
2                      5        48.738    0.421464    47.8197     49.6563
3                      5        47.394    0.421464    46.4757     48.3123
-------------------------------------------------------------------------------
```

	BLOCK	Factor_A	POROSIDAD	ESTIMACIONES	RESIDUALES	MEDIAS	DESVIACIONES
1	1	3	47.28	47.394	-0.114	49.126	-1.1393
2	1	3	47.79	47.394	0.396	48.738	-0.6293
3	1	1	51.24	49.126	2.114	47.394	2.8207
4	1	2	48.34	48.738	-0.398		-0.0793
5	1	2	47.64	48.738	-1.098		-0.7793
6	1	1	49.61	49.126	0.484		1.1907
7	1	3	46.83	47.394	-0.564		-1.5893
8	1	1	47.77	49.126	-1.356		-0.6493
9	1	3	47.31	47.394	-0.084		-1.1093
10	1	1	48.42	49.126	-0.706		0.0007
11	1	2	48.86	48.738	0.122		0.4407
12	1	2	49.86	48.738	1.122		1.4407
13	1	3	47.76	47.394	0.366		-0.6593
14	1	2	48.99	48.738	0.252		0.5707
15	1	1	48.59	49.126	-0.536		0.1707