ÍNDICE

II
INTRODUCCIÓN A LA TEORÍA Y APLICACIONES
DE LA ESTADÍSTICA

IV
SIMULACIÓN DE PROCESOS Y VALIDACION DE MÉTODOS

C. ANÁLISIS DE LA VARIANZA

En esta última sección se consideran cuestiones complementarias importantes en el tratamiento de datos, como son las relacionadas con la *homogeneidad* y con ellas se generaliza lo ya visto en el análisis de muestras anteriormente. Ya se mencionó al discutir la distribución de Fisher (Cap.6) que ésta constituía la base de una técnica muy poderosa conocida como el *análisis de la varianza* (ANOVA). Esta es una técnica perfectamente adecuada para realizar comparaciones entre varias series de datos que pudieran haberse obtenido bajo diferentes condiciones, equipamientos, calidades o purezas de los productos empleados, etc. La potencia de ANOVA radica en que permite estudiar simultáneamente estos diversos *factores* de variación que influyen en los datos obtenidos, separando sus efectos particulares, tanto si son puramente *aditivos* (efectos independientes) como si *interaccionan* entre sí. Estos factores de variación se clasifican en *controlables* (o de efecto fijo) y *no controlables* (aleatorios). Ejemplos típicos de factores de variación, en principio, controlables son los equipos o máquinas de medición o producción, los operarios que realizan las mediciones, y las determinadas condiciones experimentales seguidas para la obtención de la muestra (los posibles agentes reductores, oxidantes, quelantes, etc.). Por otra parte, ejemplos típicos de factores de variación no controlables son las inhomogeneidades en el sistema del que se toman las muestras y el día en el que se realizan las mediciones (por los cambios aleatorios en la presión, temperatura, etc.).

Es interesante aquí fijarse en que además de los factores anteriores siempre va a estar presente el *error aleatorio del proceso de medición* (factor aleatorio o residual) como un factor o fuente de inevitables variaciones en los datos tomados. ANOVA también permite evaluar el efecto de este último factor. Incluso ANOVA puede utilizarse ventajosamente en la decisión del mejor ajuste de regresión entre varios de ellos. La complejidad que puede alcanzarse con ANOVA es muy grande y de ahí que goce hoy de una popularidad creciente en las ciencias experimentales (¡fue introducida por Fisher entre 1937 y 1950!). Aquí se van a considerar sólo algunas aplicaciones comunes y sencillas de ANOVA, remitiendo al lector, como de costumbre, a la bibliografía especializada para más detalles.

La aplicación significativa de ANOVA a un conjunto de datos $\{x_{ij}\}$ procedentes de varias muestras (i = muestra, j = dato) requiere de ciertas condiciones generales:

— El carácter Gaussiano de la distribución de probabilidad de la que se hayan extraído las muestras, para poder aplicar con fiabilidad el test F de Fisher.

— La homogeneidad de las varianzas de todas las medidas analizadas, pues en el análisis se mezclan todas estas medidas.

— Las variaciones debidas a factores no controlados deben ser verdaderamente aleatorias.

— Los errores aleatorios tienen que distribuirse de forma aproximadamente Gaussiana.

Como una parte importante es la homogeneidad de las varianzas se va a comenzar presentando un criterio que la garantiza.

10.6. Homogeneidad de un conjunto de varianzas muestrales

Dado un conjunto de k varianzas muestrales $\{s_1^2, s_2^2, \ldots, s_k^2\}$ cada una fijada con un número diferente de mediciones $\{N_1, N_2, \ldots, N_k\}$ extraídas de una población normal, su homogeneidad se puede analizar con el estadístico B_s de Bartlett

$$B_s = \frac{1}{c}\left[N \ln S^2 - \sum_{i=1}^{k} (N_i - 1) \ln s_i^2\right] \tag{10.6.1}$$

en donde suele exigirse que $N_i \geq 5$ y cada símbolo es

$$N = \sum_{i=1}^{k} (N_i - 1) \tag{10.6.2}$$

$$S^2 = \frac{1}{N} \sum_{i=1}^{k} (N_i - 1) s_i^2 \tag{10.6.3}$$

$$c = 1 + \frac{-\dfrac{1}{N} + \displaystyle\sum_{i=1}^{k} \frac{1}{N_i - 1}}{3(k-1)} \tag{10.6.4}$$

El estadístico B_s sigue aproximadamente una distribución χ^2 con $\nu = k - 1$ grados de libertad. Una vez seleccionado el nivel de significación unilateral α (= 0,05, 0,01, etc.) se ensaya la hipótesis nula como es habitual. Así, para

H_0 : el conjunto $\{s_1^2, s_2^2, ..., s_k^2\}$ es homogéneo y todas las varianzas proceden de una misma población (μ, σ^2)

$$\left\langle s_1^2 \right\rangle = \left\langle s_2^2 \right\rangle = ... = \left\langle s_k^2 \right\rangle = \sigma^2 \qquad (10.6.5)$$

se tiene la regla de decisión siguiente

— Si $B_s < \chi^2_{1-\alpha}(k - 1)$, se acepta H_0 al nivel α.

— Si $B_s > \chi^2_{1-\alpha}(k - 1)$ se rechaza H_0 al nivel α.

Este criterio puede aplicarse al caso de diferentes poblaciones con la misma σ^2.

10.7. Homogeneidad de un conjunto de medias (ANOVA-1)

En un experimento con un único *factor* de variación, controlable o no, cuando se dispone de un conjunto de medias $\{\bar{x}_1, \bar{x}_2, ..., \bar{x}_k\}$ cada una de ellas fijada con una serie de N mediciones extraídas de una población normal, x_{ij} (i = 1, 2, ..., k; j = 1, 2, ..., N), su homogeneidad puede evaluarse con un estadístico F de Fisher definido como

$$F = \frac{s_n^2}{s_d^2} = \frac{\dfrac{N}{k-1}\displaystyle\sum_{i=1}^{k}\left(\bar{x}_i - \bar{x}\right)^2}{\dfrac{1}{k(N-1)}\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{N}\left(x_{ij} - \bar{x}_i\right)^2}; \quad s_n^2 > s_d^2 \qquad (10.7.1)$$

en donde $\bar{x}$ es la media global de todas las $\bar{x}_i$

$$\bar{x} = \frac{1}{k}\sum_{i=1}^{k}\bar{x}_i \qquad (10.7.2)$$

Tabla 1. ANOVA-1

Muestra	Mediciones				Medias
1	x_{11}	x_{12}	......	x_{1N}	$\bar{x}_1$
2	x_{21}	x_{22}	......	x_{2N}	$\bar{x}_2$
......	......	......	......	......	......
k	x_{k1}	x_{k2}	......	x_{kN}	$\bar{x}_k$
Media global					$\bar{x}$

Este estadístico sigue la distribución de Fisher con $\nu_n = k - 1$ y $\nu_d = k\,(N - 1)$ grados de libertad. Fijado el nivel de significación unilateral α la verificación de la hipótesis nula se realiza de la forma habitual. Esquemáticamente

H_0 : el conjunto $\{\bar{x}_1, \bar{x}_2, ..., \bar{x}_k\}$ es homogéneo y todas las medias proceden de una misma población (μ, σ^2)

$$\langle \bar{x}_1 \rangle = \langle \bar{x}_2 \rangle = ... = \langle \bar{x}_k \rangle = \mu \tag{10.7.3a}$$

$$\langle s_1^2 \rangle = \langle s_2^2 \rangle = ... = \langle s_k^2 \rangle = \sigma^2 \tag{10.7.3b}$$

se tiene la regla de decisión siguiente

— Si $F < F^C_{\nu_1,\nu_2}(\alpha)$, se acepta H_0 al nivel α.

— Si $F > F^C_{\nu_1,\nu_2}(\alpha)$, se rechaza H_0 al nivel α.

El procedimiento anterior contiene de manera compacta lo que se conoce como la aplicación ANOVA-1 y es muy instructivo analizar la naturaleza de F que debe ser un cociente entre dos varianzas como se indica en (10.7.1).

Como la población es común se puede estimar σ^2 de dos maneras significativas: *entre muestras* y *dentro de la muestra global*. Estos dos cálculos van a dar origen al numerador y al denominador de (10.7.l) respectivamente y se detallan a continuación.

i) *Estimación entre muestras*

Ya se vio en el Cap. 6 que la varianza de una media muestral se relaciona con la varianza poblacional como

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{N} \tag{10.7.4}$$

Por otra parte, en el supuesto de que H_0 fuera cierta es claro que la varianza de las medias muestrales $s_{\bar{X}}^2$ es también un estimador de $\sigma_{\bar{X}}^2$ con $k-1$ grados de libertad

$$s_{\bar{X}}^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(\bar{x}_i - \bar{x}\right)^2 \approx \frac{\sigma^2}{N} \qquad (10.7.5)$$

De este modo la estimación de σ^2 por esta vía es el numerador de (10.8.1)

$$s_n^2 = \frac{N}{k-1}\sum_{i=1}^{k}\left(\bar{x}_i - \bar{x}\right)^2 \; ; \quad \nu = k-1 \quad \text{grados de libertad} \qquad (10.7.6)$$

ii) *Estimación dentro de la muestra*

En cada muestra una estimación de la varianza poblacional σ^2 está dada por

$$s_i^2 = \frac{1}{N-1}\sum_{j=1}^{N}\left(x_{ij} - \bar{x}_i\right)^2 ; \quad i = 1,2,\ldots,k \qquad (10.7.7)$$

y promediando los k valores anteriores se tiene otro estimador para σ^2 que es justamente el denominador de (10.8.1)

$$s_d^2 = \frac{1}{k(N-1)}\sum_{i=1}^{k}\;\sum_{j=1}^{N}\left(x_{ij} - \bar{x}_i\right)^2 ; \quad \nu = k(N-1) \quad \text{grados de libertad} \qquad (10.7.8)$$

en donde hay que notar que hay k muestras cada una con $N-1$ grados de libertad.

Observaciones adicionales

Las dos explicaciones anteriores detallan la construcción de F (10.7.1) y la regla de decisión asociada. Conviene además notar en este contexto las observaciones que se indican a continuación.

— Al estar fijada con las diferencias entre valores medios la estimación s_n^2 no depende explícitamente de las variaciones dentro de cada muestra.

— Igualmente, la estimación s_d^2 al depender de las diferencias $x_{ij} - \bar{x}_i$ no depende de las medias muestrales $\bar{x}_i$ (ni de $\bar{x}$).