Capítulo 1

Diseños completamente aleatorizados

1.1. Introducción. Ejemplos

En la investigación científica es frecuente encontrarse con la necesidad de comparar entre sí diversas alternativas. Algunas de estas situaciones pueden ser las siguientes:

- Una compañía algodonera que emplea diversos fertilizantes desea comprobar si éstos tienen efectos diferentes sobre el rendimiento de la semilla de algodón.
- Una profesora de estadística que imparte en grupos experimentales de alumnos, en los que explica la misma materia pero siguiendo distintos métodos de enseñanza, desea comprobar si el método de enseñanza utilizado influye en las calificaciones de los alumnos.
- Una industria química, que obtiene un determinado producto, está interesada en comprobar si los cambios de temperatura influyen en la cantidad de producto obtenido.

Todas estas situaciones tienen en común que su interés está centrado en un solo factor con varios niveles o tratamientos que pueden producir efectos distintos y, por ello, pueden ser abordadas mediante la técnica estadística del *Análisis de la Varianza de un factor o una vía.*

El análisis de la varianza fue desarrollado por Fisher en 1925 con el objetivo de comparar entre sí varios grupos o tratamientos mediante la descomposición de la variabilidad total de un experimento en componentes independientes que puedan atribuirse a distintas causas. Esencialmente este análisis determina si la discrepancia *entre las medias de*

los tratamientos es mayor de lo que podría esperarse razonablemente de la discrepancia existente *dentro de los tratamientos.*

En los ejemplos anteriores, aparte del factor mencionado, también pueden influir otros muchos factores que se suponen de poca importancia. Por ejemplo:

- En el rendimiento de la planta de algodón, además del tipo de fertilizante, también pueden influir, pequeñas variaciones en la cantidad de riego, en la pureza de los insecticidas suministrados, etc.
- En las calificaciones de los alumnos, además del método de enseñanza, también pueden influir, el nivel cultural del alumno, el grado de atención y de interés del alumno, etc.
- En la cantidad de producto obtenido, además de la temperatura, también pueden influir, la pureza de la materia prima, la habilidad de los operarios, etc.

El resultado de todas estas causas o factores no controlados influyen en la variable respuesta; en el caso concreto de la compañía algodonera, en las diferencias de los rendimientos, en la *variabilidad* de los rendimientos. El análisis de esta *variabilidad* es la base fundamental de esta técnica estadística.

Refiriéndonos de nuevo al ejemplo de la compañía algodonera, supongamos que deseamos comprobar el efecto del fertilizante en el rendimiento del cultivo de algodón, cuya variabilidad también es debida a la presencia de otros factores. Teóricamente es posible dividir esta variabilidad en dos partes, la originada por el factor de interés, el tipo de fertilizante, y la producida por los restantes factores que entran en juego, conocidos o no, controlables o no, que recibe el nombre de *perturbación o error experimental.*

En Estadística Básica se aborda el problema de contrastar la igualdad de medias de dos poblaciones. Por ejemplo, comparar entre sí dos tipos de fertilizantes, dos métodos de enseñanza, dos tratamientos médicos, dos tipos de insecticidas, dos temperaturas empleadas en el proceso de obtención de un determinado producto, etc. Estos tipos de tests son de uso muy frecuente y se denominan contrastes para dos muestras independientes. El Análisis de la Varianza generaliza estos procedimientos a más de dos poblaciones.

Para abordar esta situación, seguiremos la siguiente metodología:

(i) Establecer un modelo de comportamiento, que podemos formalizar matemáticamente.

(ii) Estimar los parámetros del modelo.

(iii) Contrastar la hipótesis de igualdad de medias de los tratamientos.

(iv) Comprobar la idoneidad del modelo.

Como hemos dicho anteriormente, presentaremos en este capítulo el modelo con un solo factor y en capítulos posteriores generalizaremos esta idea a más de un factor.

A lo largo de este capítulo vamos a considerar algunos de los ejemplos citados en la introducción como situaciones ilustrativas de referencia.

1.2. Planteamiento del modelo

Para desarrollar esta sección consideramos como ejemplo ilustrativo la situación de la compañía algodonera. A lo largo de las sucesivas secciones, seguiremos haciendo referencia a este ejemplo. Dicha situación daría lugar, con unos datos concretos, al siguiente enunciado:

Ejemplo 1.1

Una compañía algodonera, interesada en maximizar el rendimiento de la semilla de algodón, desea comprobar si dicho rendimiento depende del tipo de fertilizante utilizado para tratar la planta. A su disposición tiene 5 tipos de fertilizantes. Para comparar su eficacia fumiga, con cada uno de los fertilizantes, un cierto número de parcelas de terreno de la misma calidad y de igual superficie. Al recoger la cosecha se mide el rendimiento de la semilla, obteniéndose las siguientes observaciones que se muestran en la Tabla 1-1

Tabla 1-1. *Rendimiento del algodón*

Fertilizantes	*Rendimiento*					
1	*51*	*49*	*50*	*49*	*51*	*50*
2	*56*	*60*	*56*	*56*	*57*	
3	*48*	*50*	*53*	*44*	*45*	
4	*47*	*48*	*49*	*44*		
5	*43*	*43*	*46*	*47*	*45*	*46*

En este experimento, se han considerado 5 tipos de fertilizantes que se han aplicado, respectivamente, a 6, 5, 5, 4 y 6 parcelas. La variable de interés o variable respuesta es el rendimiento de la semilla en peso por unidad de superficie.

Todo este planteamiento se puede formalizar de manera general para cualquier experimento unifactorial. Supongamos un factor con I niveles y que para el nivel i-ésimo se obtienen n_i observaciones de la variable respuesta. Entonces podemos postular el siguiente modelo

$$y_{ij} = \mu + \tau_i + u_{ij} \quad , \tag{1.1}$$

donde

- y_{ij} es la variable aleatoria que representa la observación j-ésima del i-ésimo tratamiento (nivel i-ésimo del factor).
- μ es un efecto constante, común a todos los niveles, denominado media global.
- τ_i es la parte de y_{ij} debida a la acción del nivel i-ésimo, que será común a todos los elementos sometidos a ese nivel del factor, ("aportación cuantitativa del nivel i-ésimo del factor al valor total de la variable y_{ij}"), llamado *efecto del tratamiento* i-ésimo.
- u_{ij} son variables aleatorias que engloban un conjunto de factores, cada uno de los cuales influye en la respuesta sólo en pequeña magnitud pero que de forma conjunta debe tenerse en cuenta en la especificación y tratamiento del modelo; es decir, las perturbaciones o error experimental pueden interpretarse como las variaciones causadas por todos los factores no analizados y que dentro del mismo tratamiento variarán de unos elementos a otros. Estas perturbaciones deben verificar las siguientes condiciones:
  - Que tengan media cero
    $$\mathrm{E}[u_{ij}] = 0 \qquad \forall i, j \quad .$$
  - Que tengan varianza constante (hipótesis de homocedasticidad)
    $$\mathrm{Var}[u_{ij}] = \sigma^2 \qquad \forall i, j \quad .$$
  - Que sean estadísticamente independientes entre sí
    $$\mathrm{E}[u_{ij}u_{rk}] = 0 \qquad i \neq r \;\; \text{ó} \;\; j \neq k \quad .$$
  - Que su distribución sea normal.

Nuestro objetivo es estimar los efectos de los tratamientos y contrastar la hipótesis de que todos los niveles del factor producen el mismo efecto, frente a la alternativa de que al menos dos difieren significativamente entre sí. Para ello, se supone que los errores experimentales son variables aleatorias independientes con distribución normal, con media cero y varianza constante σ^2.

En este modelo, que estudia el efecto que produce un solo factor en la variable respuesta, la asignación de las unidades experimentales a los distintos niveles del factor se debe realizar de forma completamente al azar. Este modelo, junto con este procedimiento de asignación, recibe el nombre de *Diseño Completamente Aleatorizado* y está basado en el modelo estadístico de *Análisis de Varianza de un Factor o una Vía.* Para aplicar este

diseño adecuadamente las unidades experimentales deben ser lo más homogéneas posible.

En el modelo estadístico dado por la ecuación (1.1), se distinguen dos situaciones según la selección de los tratamientos: *modelo de efectos fijos* y *modelo de efectos aleatorios*.

(i) En el modelo de *efectos fijos* el experimentador decide qué niveles concretos se van a considerar y las conclusiones obtenidas son aplicables sólo a dichos niveles, no pudiéndose hacer extensivas a otros niveles no incluidos en el estudio.

En la situación de referencia, la compañía algodonera decide utilizar unos determinados fertilizantes. Se trata de un modelo de efectos fijos y la compañía algodonera aplicará los resultados de la investigación *exclusivamente* a los fertilizantes considerados en el estudio.

El caso de las calificaciones de los alumnos también se trata de un modelo unifactorial de efectos fijos, ya que la profesora sólo está interesada en averiguar si unos determinados métodos de enseñanza influyen en las calificaciones de los alumnos y aplicará los resultados de la investigación *exclusivamente* a los métodos de enseñanza empleados.

(ii) En el modelo de *efectos aleatorios*, los niveles del factor se seleccionan al azar; es decir, los niveles estudiados son una *muestra aleatoria* de una población de niveles. En este modelo se generalizan las conclusiones (basadas en la muestra de niveles), a todos los posibles niveles del factor, hayan sido explícitamente considerados en el análisis o no.

Refiriéndonos a la situación de la industria química, interesada en la influencia de la temperatura en la cantidad de producto obtenido, se podría considerar como un modelo de efectos aleatorios si las temperaturas comparadas son una muestra entre las posibles temperaturas que se podrían utilizar.

Es importante distinguir claramente las diferencias entre ambos modelos. En el primero, tanto la compañía algodonera como la profesora podían estudiar un número de fertilizantes y métodos de enseñanza, respectivamente, más amplio pero sólo están interesados en unos determinados. El interés se centra en la comparación de las medias de los niveles considerados, por lo cual los resultados sólo pueden aplicarse a dichos niveles. Por el contrario, en la industria química el interés se centra en la variabilidad que produce el cambio de temperaturas en la cantidad de producto obtenido.

1.3. Modelo de efectos fijos

En este modelo, los efectos τ_i son constantes desconocidas que estamos interesados en estimar y en contrastar determinadas hipótesis relativas a dichos efectos. Para ello, para cada nivel i del factor, tomamos una muestra aleatoria simple de tamaño n_i. En principio podemos reescribir el modelo 1.1 en la forma

$$y_{ij} = \mu_i + u_{ij} \quad , \tag{1.2}$$

donde y_{ij} será la observación correspondiente al elemento j-ésimo ($j = 1, 2, \cdots, n_i$) sujeto al nivel i-ésimo del factor ($i = 1, 2, \cdots, I$) y μ_i es la media correspondiente al nivel i-ésimo.

Las condiciones anteriores de este modelo se resumen en:

1º)	$y_{ij} = \mu_i + u_{ij}$
2º)	$u_{ij} \rightsquigarrow N(0, \sigma)$
3º)	u_{ij} son independientes entre sí .

Al ser μ_i constante para el tratamiento i, toda la fuente de aleatoriedad del modelo descansa en las variables de perturbación y la variable y_{ij} toma el carácter aleatorio de ellas; por ello, las hipótesis establecidas para las variables de perturbación pueden ser formuladas en términos de la variable respuesta. En otras palabras las variables y_{ij} son variables aleatorias independientes con distribución normal, con media μ_i y varianza σ^2; es decir:

- La esperanza de la variable respuesta en el nivel i-ésimo es μ_i

$$\mathrm{E}[y_{ij}] = \mu_i \qquad \forall j \quad .$$

  Esta hipótesis exige que las n_i observaciones correspondientes al tratamiento i-ésimo tengan la misma media μ_i.

- La varianza de las variables y_{ij} es constante

$$\mathrm{Var}[y_{ij}] = \sigma^2 \qquad \forall i, j \quad .$$

- Las observaciones y_{ij} son independientes entre sí

$$\text{Cov}\,[y_{ij}, y_{rk}] = 0 \qquad i \neq r \;\; \text{ó} \;\; j \neq k \quad .$$

- La variable respuesta tiene distribución normal.

Si expresamos μ_i como suma de dos términos: μ, común a todas las observaciones, y τ_i, específica de cada nivel, es decir

$$\mu_i = \mu + \tau_i \quad , \tag{1.3}$$

sustituyendo (1.3) en la ecuación (1.2), obtenemos la primera expresión del modelo dada en (1.1), es decir

$$y_{ij} = \mu + \tau_i + u_{ij} \quad ,$$

donde, el efecto producido por el nivel i-ésimo se define como la diferencia entre la media μ_i, del nivel i, y la media general μ; es decir, los efectos de los tratamientos τ_i son las desviaciones de la media de cada nivel con respecto a la media general, por esta razón se debe verificar la relación

$$\sum_{i=1}^{I} n_i \tau_i = 0 \quad . \tag{1.4}$$

De esta manera el valor esperado de la respuesta en el i-ésimo tratamiento es,

$$\text{E}\,[y_{ij}] \equiv \mu_i = \mu + \tau_i \quad ,$$

la suma de la media general y el efecto del i-ésimo tratamiento.

En este modelo se trata de contrastar si todos los niveles del factor producen el mismo efecto

$$H_0: \quad \tau_i = 0 \qquad \forall i$$

frente a la alternativa

$$H_1: \quad \tau_i \neq 0 \quad \text{por lo menos para algún } i \;,$$

o, equivalentemente, si todos los tratamientos tienen la misma media

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_I = \mu$$

frente a la alternativa

$$H_1: \quad \mu_i \neq \mu_j \quad \text{por lo menos para algún par } (i, j) \quad .$$

Si H_0 es cierta, todos los tratamientos tienen la misma media, μ, la pertenencia a un grupo u otro es irrelevante, y podemos considerar todas las observaciones como provenientes de una única población.

Hemos introducido dos formas de expresión del modelo:

$$y_{ij} = \mu_i + u_{ij}$$
$$y_{ij} = \mu + \tau_i + u_{ij} \quad ,$$

ambas formas son equivalentes. Para nuestro estudio emplearemos la segunda expresión.

Consideremos un factor con I niveles y que para cada nivel i se toman n_i observaciones. Los datos se organizan en forma tabular como se muestra en la Tabla 1-2

Tabla 1-2. Experimento unifactorial

Tratamiento (nivel)	Observaciones						Nº Observ. n_i	Totales $y_{i.}$	Promedios $\bar{y}_{i.}$
1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1n_1}	n_1	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2n_2}	n_2	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{in_i}	n_i	$y_{i.}$	$\bar{y}_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	y_{I1}	y_{I2}	$\cdots$	y_{Ij}	$\cdots$	y_{In_I}	n_I	$y_{I.}$	$\bar{y}_{I.}$
							N	$y_{..}$	$\bar{y}_{..}$

donde utilizamos la siguiente notación:

- N es el número total de observaciones, es decir, $N = \sum_{i=1}^{I} n_i$
- y_{ij} es la observación j-ésima del tratamiento i-ésimo; donde el subíndice i varía desde 1 hasta I y el subíndice j varía desde 1 hasta n_i
- $y_{i.}$ es el total de las observaciones bajo el i-ésimo tratamiento, es decir

$$y_{i.} = \sum_{j=1}^{n_i} y_{ij} \qquad i = 1, \cdots, I$$

- $\bar{y}_{i.}$ es el promedio de las observaciones bajo el i-ésimo tratamiento, es decir

$$\bar{y}_{i.} = \frac{y_{i.}}{n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \quad i = 1, \cdots, I$$

- $y_{..}$ es la suma de todas las observaciones, denominado el total general, es decir

$$y_{..} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij} = \sum_{i=1}^{I} y_{i.}$$

- $\bar{y}_{..}$ es la media general de las observaciones, es decir

$$\bar{y}_{..} = \frac{y_{..}}{N} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{N}\sum_{i=1}^{I} n_i\bar{y}_{i.} \quad .$$

En estas definiciones, la notación de "punto" en el subíndice indica la suma que se ha realizado sobre el subíndice reemplazado por el punto.

Si los tamaños n_i de las muestras son distintos, el modelo recibe el nombre de *modelo no-equilibrado o no-balanceado.* Si todas las muestras tienen el mismo tamaño, $n_i = n$, el modelo se llama *modelo equilibrado o balanceado.*

En primer lugar, analizaremos el caso más general, el modelo no-equilibrado.

1.3.1. Estimación de los parámetros del modelo

Estimación por máxima verosimilitud

Como ya se ha explicado anteriormente, la hipótesis de normalidad sobre los términos de error conlleva el hecho de que las variables y_{ij} sean normales e independientes, por lo que es inmediato construir la función de verosimilitud asociada a la muestra $\mathbf{y} = (y_{11}, \cdots, y_{1n_1}, y_{21}, \cdots, y_{2n_2}, \cdots, y_{I1}, \cdots, y_{In_I})$:

$$\mathbb{L}(\mu, \tau_i, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}[y_{ij} - \mu - \tau_i]^2\right) \tag{1.5}$$

Los estimadores máximo-verosímiles para los parámetros μ, τ_i y σ^2 son los valores para los cuales la función de verosimilitud alcanza su máximo. Para determinarlos habrá que obtener los puntos críticos de la función (1.5). Por conveniencia, en vez de maximizar

la función de verosimilitud, se maximiza el logaritmo, ya que el logaritmo conserva los puntos críticos por ser una función creciente. En este caso,

$$\ln(\mathbb{L}(\mu, \tau_i, \sigma^2)) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}[y_{ij} - \mu - \tau_i]^2 \tag{1.6}$$

Las derivadas parciales respecto de los parámetros del modelo son las siguientes:

$$\begin{aligned}
\frac{\partial \ln \mathbb{L}}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}[y_{ij} - \mu - \tau_i] \\
\frac{\partial \ln \mathbb{L}}{\partial \tau_i} &= \frac{1}{\sigma^2}\sum_{j=1}^{n_i}[y_{ij} - \mu - \tau_i] \qquad i = 1, \cdots, I \\
\frac{\partial \ln \mathbb{L}}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}[y_{ij} - \mu - \tau_i]^2 \quad .
\end{aligned} \tag{1.7}$$

Igualando estas derivadas parciales a cero, se obtiene un sistema de ecuaciones que proporciona los estimadores máximo verosímiles.

Veamos las ecuaciones y las soluciones que vamos obteniendo.

(i) De la primera ecuación, obtenemos

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij} - N\mu - \sum_{i=1}^{I}\sum_{j=1}^{n_i}\tau_i = 0 \quad . \tag{1.8}$$

Teniendo en cuenta que $\sum_i n_i\tau_i = 0$, de la ecuación (1.8) se obtiene

$$\widehat{\mu} = \frac{\displaystyle\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}}{N} = \bar{y}_{..} \tag{1.9}$$

(ii) Para τ_i, obtenemos

$$\sum_{j=1}^{n_i} y_{ij} - n_i\widehat{\mu} - \sum_{j=1}^{n_i} \tau_i = 0 \qquad i = 1, 2, \cdots, I \tag{1.10}$$

De las ecuaciones (1.10) se obtienen los siguientes estimadores máximo-verosímiles para los parámetros τ_i

$$\widehat{\tau}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} - \widehat{\mu} = \bar{y}_{i.} - \bar{y}_{..} \quad . \tag{1.11}$$

Estas soluciones

$$\begin{aligned} \widehat{\mu} &= \quad \bar{y}_{..} \\ \widehat{\tau}_i &= \quad \bar{y}_{i.} - \bar{y}_{..} \qquad i = 1, 2, \cdots, I \quad , \end{aligned} \tag{1.12}$$

se pueden intuir fácilmente, ya que indican que la media general se puede estimar utilizando el promedio de todas las observaciones y cualquiera de los efectos de los tratamientos usando la diferencia entre el promedio correspondiente al tratamiento y el promedio total.

Finalmente, sustituyendo $\widehat{\mu}$ y $\widehat{\tau}_i$ en la última ecuación de (1.7), obtenemos el estimador de máxima verosimilitud para la varianza poblacional

$$\widehat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i} [y_{ij} - \widehat{\mu} - \widehat{\tau}_i]^2 = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i} [y_{ij} - \bar{y}_{i.}]^2 = \frac{1}{N}\sum_{i=1}^{I} n_i s_i^2 \quad , \tag{1.13}$$

donde s_i^2 es la varianza muestral del nivel i-ésimo.

Residuos

Los residuos se definen como las diferencias entre los valores observados y_{ij} y los valores previstos por el modelo $\widehat{y}_{ij}$ y los denotamos por e_{ij},

$$e_{ij} = y_{ij} - \widehat{y}_{ij} = y_{ij} - \widehat{\mu} - \widehat{\tau}_i = y_{ij} - \bar{y}_{i.} \quad .$$

Por lo tanto, el estimador máximo-verosímil, $\widehat{\sigma}^2$, se puede escribir como

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i} e_{ij}^2}{N} \quad .$$

Los residuos son los estimadores de los errores aleatorios $u_{ij} = y_{ij} - \mu - \tau_i$, los cuales son variables aleatorias no observables. Se verifica que la suma de los residuos es cero, en efecto

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i} e_{ij} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.}) = \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij} - \sum_{i=1}^{I} n_i\bar{y}_{i.} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij} - \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij} = 0 \quad .$$

Propiedades de los estimadores máximo verosímiles

A continuación vamos a ver algunas propiedades que verifican los estimadores del modelo. Concretamente, vamos a determinar su esperanza, su varianza y su distribución en el muestreo.

1) Propiedades de $\widehat{\mu}$

a) $\widehat{\mu}$ es un estimador centrado de μ, puesto que

$$\mathrm{E}\left[\widehat{\mu}\right] = \quad \mathrm{E}\left[\bar{y}_{..}\right] = \frac{1}{N}\sum_{i,j}\mathrm{E}\left[y_{ij}\right] = \frac{1}{N}\sum_{i,j}(\mu + \tau_i) = \frac{1}{N}\left(N\mu + \sum_i n_i\tau_i\right) = \mu$$

b) La varianza de $\widehat{\mu}$ es σ^2/N, puesto que al ser independientes las observaciones se verifica:

$$\mathrm{Var}\left[\widehat{\mu}\right] = \quad \mathrm{Var}\left[\sum_{i,j}\frac{y_{ij}}{N}\right] = \sum_{i,j}\mathrm{Var}\left[\frac{y_{ij}}{N}\right] = \sum_{i,j}\frac{\mathrm{Var}\left[y_{ij}\right]}{N^2} = \sum_{i,j}\frac{\sigma^2}{N^2} = \frac{\sigma^2}{N} \qquad (1.14)$$

c) $\widehat{\mu}$ se distribuye según una Normal, puesto que dicho estimador es combinación lineal de las variables y_{ij} y éstas son variables aleatorias independientes con distribución Normal.

2) Propiedades de $\widehat{\tau}_i$

a) $\widehat{\tau}_i$ es un estimador centrado de τ_i, puesto que

$$\mathrm{E}\left[\widehat{\tau}_i\right] = \quad \mathrm{E}\left[\bar{y}_{i.}\right] - \mathrm{E}\left[\bar{y}_{..}\right] = \mathrm{E}\left[\frac{1}{n_i}\sum_j y_{ij}\right] - \mu = \frac{1}{n_i}\,\mathrm{E}\left[\sum_j(\mu + \tau_i + u_{ij})\right] - \mu =$$

$$\frac{1}{n_i}\left[n_i\mu + n_i\tau_i + \sum_j \mathrm{E}[u_{ij}]\right] - \mu = \tau_i \qquad (1.15)$$

b) La varianza de $\widehat{\tau}_i$ es $(N-n_i)\dfrac{\sigma^2}{Nn_i}$, puesto que

$$\begin{aligned}
\mathrm{Var}\,[\widehat{\tau}_i] = \quad & \mathrm{Var}\,[\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}] = \mathrm{Var}\left[\frac{1}{n_i}\sum_j y_{ij} - \frac{1}{N}\sum_{i,j} y_{ij}\right] = \\
& \frac{1}{n_i^2}\sum_j \mathrm{Var}\,[y_{ij}] + \frac{1}{N^2}\sum_{i,j}\mathrm{Var}\,[y_{ij}] - \frac{2}{Nn_i}\,\mathrm{Cov}\left[\sum_j y_{ij}, \sum_{i,j} y_{ij}\right] = \\
& \frac{1}{n_i^2}\sum_j \sigma^2 + \frac{1}{N^2}\sum_{i,j}\sigma^2 - \frac{2}{Nn_i} n_i\sigma^2 = \\
& \frac{1}{n_i}\sigma^2 + \frac{1}{N}\sigma^2 - \frac{2\sigma^2}{N} = \frac{\sigma^2}{n_i} - \frac{\sigma^2}{N} = (N-n_i)\frac{\sigma^2}{Nn_i}
\end{aligned} \tag{1.16}$$

En el caso del modelo equilibrado se deduce fácilmente que la varianza de $\widehat{\tau}_i$ es $(I-1)\sigma^2/N$.

c) $\widehat{\tau}_i$ se distribuye según una Normal, puesto que dicho estimador está expresado como función lineal de variables aleatorias con distribución Normal.

3) Propiedades de $\widehat{\sigma}^2$

$\widehat{\sigma}^2$ no es un estimador insesgado de σ^2. Para demostrarlo veamos en primer lugar que $\dfrac{N\widehat{\sigma^2}}{\sigma^2}$ se distribuye según una χ^2 con $N-I$ grados de libertad.

$$\begin{aligned}
\frac{N\widehat{\sigma^2}}{\sigma^2} = \quad & \frac{1}{\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i} e_{ij}^2 = \frac{1}{\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2 = \\
& \sum_{j=1}^{n_1}\left(\frac{y_{1j}-\bar{y}_{1\cdot}}{\sigma}\right)^2 + \sum_{j=1}^{n_2}\left(\frac{y_{2j}-\bar{y}_{2\cdot}}{\sigma}\right)^2 + \cdots + \sum_{j=1}^{n_I}\left(\frac{y_{Ij}-\bar{y}_{I\cdot}}{\sigma}\right)^2 \quad .
\end{aligned} \tag{1.17}$$

Estos sumandos son estadísticamente independientes entre sí al serlo las observaciones

muestrales. Además se verifica que

$$\sum_{j=1}^{n_i}\left(\frac{y_{ij}-\bar{y}_{i.}}{\sigma}\right)^2=\frac{n_i s_i^2}{\sigma^2} \quad ,$$

se distribuye según una χ^2 con n_i-1 grados de libertad.

Por tanto,

$$\frac{N\widehat{\sigma}^2}{\sigma^2}\rightsquigarrow\chi^2_{\Sigma_i(n_i-1)}=\chi^2_{N-I} \quad . \tag{1.18}$$

Puesto que la esperanza matemática de una distribución χ^2 coincide con sus grados de libertad, se concluye que

$$\mathrm{E}\left[\frac{N\widehat{\sigma}^2}{\sigma^2}\right]=N-I\Rightarrow\mathrm{E}\left[\widehat{\sigma}^2\right]=\frac{N-I}{N}\sigma^2$$

luego, como queríamos demostrar, $\widehat{\sigma^2}$ no es un estimador insesgado de σ^2. Ahora bien, a partir de este resultado se construye fácilmente un estimador centrado simplemente considerando

$$\widetilde{\sigma}^2=\frac{N}{N-I}\widehat{\sigma}^2$$

y por tanto,

$$\mathrm{E}\left[\widetilde{\sigma}^2\right]=\sigma^2 \quad .$$

Dicho estimador recibe el nombre de *varianza residual*, pues da información acerca de cuanta variabilidad deja de explicar el modelo y se acumula en los términos de error o residuos. La varianza residual se denota por $\widehat{S}_R^2$ y se expresa, por tanto, de la siguiente forma

$$\widetilde{\sigma^2}=\widehat{S}_R^2=\frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}[y_{ij}-\bar{y}_{i.}]^2}{N-I}=\frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}e_{ij}^2}{N-I} \quad . \tag{1.19}$$

También se puede determinar el valor esperado de $\widehat{\sigma}^2$ de la siguiente forma:

En primer lugar, calculamos $\mathrm{E}\left[e_{ij}^2\right]$

$$\begin{aligned}
\mathrm{E}\left[e_{ij}^2\right] = \quad & \mathrm{Var}\left[e_{ij}\right] = \mathrm{Var}\left[y_{ij} - \bar{y}_{i.}\right] = \\
& \mathrm{Var}[y_{ij}] + \mathrm{Var}\left[\bar{y}_{i.}\right] - 2\,\mathrm{Cov}[y_{ij}, \bar{y}_{i.}] = \\
& \sigma^2 + \frac{\sigma^2}{n_i} - 2\,\mathrm{Cov}\left[y_{ij}, \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}\right] = \sigma^2 + \frac{\sigma^2}{n_i} - 2\frac{\sigma^2}{n_i} = \\
& \sigma^2 - \frac{\sigma^2}{n_i}
\end{aligned}$$

Por lo tanto,

$$\begin{aligned}
\mathrm{E}\left[\sum_{i,j} e_{ij}^2\right] = \quad & \sum_{i,j} \mathrm{E}\left[e_{ij}^2\right] = \sum_{i,j}\left[\sigma^2 - \frac{\sigma^2}{n_i}\right] = \\
& N\sigma^2 - \sum_{i,j}\frac{\sigma^2}{n_i} = N\sigma^2 - \sigma^2\sum_i \frac{1}{n_i} n_i = \\
& (N-I)\sigma^2 \quad .
\end{aligned}$$

Entonces

$$\mathrm{E}\left[\widehat{\sigma}^2\right] = \mathrm{E}\left[\sum_{i,j}\frac{e_{ij}^2}{N}\right] = \frac{N-I}{N}\sigma^2 \quad .$$

En resumen,

$$\boxed{\begin{array}{c}
\widehat{\mu} \rightsquigarrow N\left(\mu, \dfrac{\sigma^2}{N}\right) \\[2ex]
\widehat{\tau}_i \rightsquigarrow N\left(\tau_i, (N-n_i)\dfrac{\sigma^2}{N n_i}\right) \\[2ex]
N\dfrac{\widehat{\sigma}^2}{\sigma^2} \rightsquigarrow \chi^2_{N-I}
\end{array}}$$

Estimación por mínimos cuadrados

Hemos planteado el modelo de efectos fijos como

$$y_{ij} = \mu + \tau_i + u_{ij} \quad ,$$

en el supuesto de que las perturbaciones, u_{ij}, son variables aleatorias independientes e idénticamente distribuidas según una Normal de media 0 y varianza σ^2. Las hipótesis de dicho modelo se pueden relajar en el siguiente sentido:

Las perturbaciones son variables aleatorias que verifican

1º)
$$\mathrm{E}\left[u_{ij}\right] = 0 \qquad \forall i, j \quad .$$

2º)
$$\mathrm{Var}\left[u_{ij}\right] = \sigma^2 \qquad \forall i, j \quad .$$

3º)
$$\mathrm{Cov}\left[u_{ij}, u_{rk}\right] = \mathrm{E}\left[u_{ij}u_{rk}\right] = 0 \qquad i \neq r \;\; \text{ó} \;\; j \neq k \quad .$$

Hay que hacer notar que entre las hipótesis del modelo no se hace ninguna referencia a la distribución específica de las perturbaciones. En estas condiciones, la estimación de los parámetros se aborda mediante el *método de mínimos cuadrados.*

Con el fin de obtener los estimadores de μ y τ_i mediante el método de mínimos cuadrados, consideremos la suma de cuadrados de los errores, ecuación (1.20), y determinemos los valores de μ y τ_i, que notaremos por $\widehat{\mu}$ y $\widehat{\tau}_i$, que minimizan dicha expresión

$$\Lambda = \sum_{i=1}^{I}\sum_{j=1}^{n_i} u_{ij}^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \mu - \tau_i)^2 \quad . \tag{1.20}$$

Para ello, se deriva Λ respecto de μ y τ_i y se particulariza en $\widehat{\mu}$ y $\widehat{\tau}_i$ obteniéndose un sistema de $I + 1$ ecuaciones con $I + 1$ incógnitas

$$\left.\begin{array}{rcl} \left.\dfrac{\partial \Lambda}{\partial \mu}\right|_{\widehat{\mu},\widehat{\tau}_i} & = & 0 \\ \left.\dfrac{\partial \Lambda}{\partial \tau_i}\right|_{\widehat{\mu},\widehat{\tau}_i} & = & 0 \\ i = 1, 2, \cdots, I & & \end{array}\right\}$$

Dichas ecuaciones dan lugar al sistema

$$\left.\begin{array}{ll} -2\displaystyle\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\widehat{\mu}-\widehat{\tau}_i) & =0 \\ -2\displaystyle\sum_{j=1}^{n_i}(y_{ij}-\widehat{\mu}-\widehat{\tau}_i) & =0 \\ i=1,2,\cdots,I & \end{array}\right\} \tag{1.21}$$

que se denomina *sistema de ecuaciones normales de mínimos cuadrados*. Este sistema de ecuaciones coincide con el sistema que verifican los estimadores de máxima verosimilitud de los parámetros μ y τ_i, cuando se impone la hipótesis de normalidad, cuyas soluciones vienen dadas por las expresiones (1.9) y (1.11), respectivamente. Obsérvese que por este método no se obtiene ninguna ecuación para estimar σ^2. Por analogía con el "caso normal", se utiliza como estimador de σ^2 la expresión

$$\widetilde{\sigma}^2=\widehat{S}_R^2=\frac{\displaystyle\sum_{i=1}^{I}\sum_{j=1}^{n_i}e_{ij}^2}{N-I} \quad . \tag{1.22}$$

Aunque la hipótesis de normalidad no es necesaria para estimar los parámetros mediante el método de mínimos cuadrados, dicha hipótesis resultará imprescindible para establecer las distribuciones de los estadísticos involucrados en el proceso de contraste de hipótesis.

Observación 1.1

Nótese que la inclusión de la hipótesis de normalidad de las perturbaciones conduce a la independencia entre dichas variables, puesto que en caso de normalidad, incorrelación implica independencia.

1.3.2. Descomposición de la variabilidad

Para comparar los efectos de los distintos niveles de un factor se emplea la técnica estadística denominada *análisis de la varianza*, abreviadamente *ANOVA,* que está basada en la descomposición de la variabilidad total de los datos en distintas componentes.

Se considera la siguiente identidad:

$$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.}) \quad , \tag{1.23}$$

que expresa cada variable y_{ij} como la suma de tres términos:

- La media total $\bar{y}_{..}$, es decir el estimador de μ
- El efecto producido por el tratamiento i-ésimo, (desviación de la media del i-ésimo nivel del factor respecto de la media total), $\bar{y}_{i.} - \bar{y}_{..}$, es decir el estimador de τ_i
- La desviación entre los valores observados y los valores previstos por el modelo, $y_{ij} - \bar{y}_{i.}$, es decir el estimador de u_{ij}.

Por tanto, la expresión (1.23) también se puede poner en la forma

$$y_{ij} = \widehat{\mu} + \widehat{\tau}_i + e_{ij} \tag{1.24}$$

Consideramos esta descomposición para todas las observaciones, que expresada en forma vectorial resulta

$$\mathbf{Y} = \underline{\widehat{\mu}} + \underline{\widehat{\tau}} + \mathbf{e} \quad , \tag{1.25}$$

siendo

$$\begin{array}{ll}
\mathbf{Y} = & (y_{11}, \ldots,\ y_{1n_1},\ y_{21}, \ldots,\ y_{2n_2}, \ldots,\ y_{I1}, \ldots,\ y_{In_I})' \\
\underline{\widehat{\mu}} = & (\bar{y}_{..}, \ldots..,\ \bar{y}_{..},\ \bar{y}_{..}, \ldots..,\ \bar{y}_{..}, \ldots..,\ \bar{y}_{..}, \ldots..,\ \bar{y}_{..})' \\
\underline{\widehat{\tau}} = & (\bar{y}_{1.} - \bar{y}_{..}, \ldots..,\ \bar{y}_{1.} - \bar{y}_{..},\ \bar{y}_{2.} - \bar{y}_{..}, \ldots..,\ \bar{y}_{2.} - \bar{y}_{..}, \ldots..,\ \bar{y}_{I.} - \bar{y}_{..}, \ldots..,\ \bar{y}_{I.} - \bar{y}_{..})' \\
e = & (y_{11} - \bar{y}_{1.}, \ldots, y_{1n_1} - \bar{y}_{1.}, y_{21} - \bar{y}_{2.}, \ldots, y_{2n_2} - \bar{y}_{2.}, \ldots, y_{I1} - \bar{y}_{I.}, \ldots, y_{In_I} - \bar{y}_{I.})'
\end{array}$$

donde

- $\mathbf{Y}$: Contiene N términos independientes y_{ij}. Tiene, por tanto, N grados de libertad.
- $\underline{\widehat{\mu}}$: Contiene N coordenadas iguales a $\bar{y}_{..}$. Tiene, por tanto, un grado de libertad.
- $\underline{\widehat{\tau}}$: Contiene I valores distintos $\bar{y}_{i.} - \bar{y}_{..}$, cada uno repetido n_i veces y sujetos a una ecuación de restricción, $\sum_i n_i(\bar{y}_{i.} - \bar{y}_{..}) = 0$. Tiene, por tanto, $I-1$ grados de libertad.
- $\mathbf{e}$: Contiene los N residuos sujetos a I ecuaciones de restricción, $\sum_j (y_{ij} - \bar{y}_{i.}) = 0$ para $i = 1, \ldots, I$. Tiene, por tanto, $N - I$ grados de libertad.

La descomposición (1.25) está formada por componentes ortogonales dos a dos, ya que se verifica

$$\begin{aligned}
\underline{\widehat{\mu}}' \times \underline{\widehat{\tau}} = &\quad \bar{y}_{..} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \widehat{\tau}_i = \quad 0 \\
\underline{\widehat{\mu}}' \times \mathbf{e} = &\quad \bar{y}_{..} \sum_{i=1}^{I} \sum_{j=1}^{n_i} e_{ij} = \quad 0 \\
\underline{\widehat{\tau}}' \times \mathbf{e} = &\quad \sum_{i=1}^{I} \widehat{\tau}_i \sum_{j=1}^{n_i} e_{ij} = \quad 0
\end{aligned}$$

La ecuación (1.23) también se puede expresar de la siguiente forma

$$y_{ij} - \bar{y}_{..} = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.}) \quad , \tag{1.26}$$

si elevamos al cuadrado los dos miembros de la expresión anterior y sumamos para todas las observaciones, tenemos

$$\begin{aligned}
\sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = &\quad \sum_{i=1}^{I} \sum_{j=1}^{n_i} [(\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})]^2 = \sum_{i=1}^{I} n_i \, (\bar{y}_{i.} - \bar{y}_{..})^2 + \\
&\sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + 2 \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..}) \, (y_{ij} - \bar{y}_{i.}) \quad .
\end{aligned} \tag{1.27}$$

Los dobles productos se anulan, ya que los términos son ortogonales, por lo que dicha ecuación queda en la forma

$$\sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \quad \sum_{i=1}^{I} n_i \, (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 \tag{1.28}$$

que representa la *ecuación básica del análisis de la varianza*, que simbólicamente podemos escribir

$$SCT = SCTr + SCR \quad ,$$

donde hemos desglosado la *variabilidad total* de los datos

$$SCT = \sum_{ij} (y_{ij} - \bar{y}_{..})^2 \quad ,$$

denominada *suma total de cuadrados*, en dos partes:

1) $SCTr = \sum_{i=1}^{I} n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$, la suma de cuadrados de las desviaciones de las medias de los tratamientos respecto de la media general, denominada *suma de cuadrados entre tratamientos* o *variabilidad explicada*

2) $SCR = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.}\right)^2$, la suma de cuadrados de las desviaciones de las observaciones de cada nivel respecto de su media, denominada *suma de cuadrados dentro de los tratamientos*, *variabilidad no-explicada o residual.*

A partir de las sumas de cuadrados anteriores se pueden construir los denominados *cuadrados medios,* definidos como los cocientes entre dichas sumas de cuadrados y sus correspondientes grados de libertad.

* Cuadrado medio total[1]

$$\widehat{S}_T^2 = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2}{N-1} \tag{1.29}$$

* Cuadrado medio entre tratamientos

$$\widehat{S}_{Tr}^2 = \frac{\sum_{i=1}^{I} n_i (\bar{y}_{i.} - \bar{y}_{..})^2}{I-1} \tag{1.30}$$

* Cuadrado medio residual

$$\widehat{S}_R^2 = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}{N-I} \quad , \tag{1.31}$$

Una notación muy utilizada también en la práctica para los cuadrados medios total, entre tratamientos y residual es, respectivamente, CMT, $CMTr$ y CMR o CME.

A continuación vamos a calcular las esperanzas matemáticas de estos cuadrados medios. En primer lugar, recordemos la expresión del modelo (1.1)

$$y_{ij} = \mu + \tau_i + u_{ij} \quad .$$

[1] El número de grados de libertad asociados a SCT es $N-1$ ya que $\sum_{i,j}(y_{ij} - \overline{y}_{..}) = 0$.

Consideremos las expresiones de $y_{i.}$, $\bar{y}_{i.}$, $y_{..}$ e $\bar{y}_{..}$, en función de los parámetros del modelo, con objeto de poder hallar las esperanzas de las varianzas muestrales. También tengamos en cuenta que $\sum_i n_i\tau_i = 0$. Así tenemos:

$$\begin{aligned} y_{i.} &= n_i\mu + n_i\tau_i + u_{i.} \;\; ; \qquad \bar{y}_{i.} = \mu + \tau_i + \overline{u}_{i.} \\ y_{..} &= N\mu + \textstyle\sum_i n_i\tau_i + u_{..} \;\; ; \;\; \bar{y}_{..} = \mu + \overline{u}_{..} \end{aligned} \tag{1.32}$$

1º) El cuadrado medio entre grupos lo podemos expresar como:

$$\widehat{S}^2_{Tr} = \frac{\displaystyle\sum_{i=1}^{I} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{I-1} = \frac{\displaystyle\sum_{i=1}^{I} n_i\left[\tau_i + (\overline{u}_{i.} - \overline{u}_{..})\right]^2}{I-1} =$$

$$\frac{\displaystyle\sum_{i=1}^{I} n_i\tau_i^2}{I-1} + \frac{\displaystyle\sum_{i=1}^{I} n_i(\overline{u}_{i.} - \overline{u}_{..})^2}{I-1} + \frac{\displaystyle 2\sum_{i=1}^{I} n_i\tau_i(\overline{u}_{i.} - \overline{u}_{..})}{I-1}$$

y su esperanza matemática será la suma de las esperanzas matemáticas de cada sumando; es decir,

$$\mathrm{E}\left[\widehat{S}^2_{Tr}\right] = \mathrm{E}\left[\frac{\displaystyle\sum_{i=1}^{I} n_i\tau_i^2}{I-1}\right] + \mathrm{E}\left[\frac{\displaystyle\sum_{i=1}^{I} n_i(\overline{u}_{i.} - \overline{u}_{..})^2}{I-1}\right] + \mathrm{E}\left[\frac{\displaystyle 2\sum_{i=1}^{I} n_i\tau_i(\overline{u}_{i.} - \overline{u}_{..})}{I-1}\right] \quad . \tag{1.33}$$

Ahora bien, puesto que:

a) El modelo es de efectos fijos $\mathrm{E}[\tau_i] = \tau_i$, entonces

$$\mathrm{E}\left[\frac{\displaystyle\sum_{i=1}^{I} n_i\tau_i^2}{I-1}\right] = \frac{1}{I-1}\sum_{i=1}^{I} n_i\,\mathrm{E}\left[\tau_i^2\right] = \frac{1}{I-1}\sum_{i=1}^{I} n_i\tau_i^2 \quad . \tag{1.34}$$

b) Como $\mathrm{E}\left[\widehat{\tau}_i - \mathrm{E}\left[\widehat{\tau}_i\right]\right]^2$ es la $\mathrm{Var}(\widehat{\tau}_i)$, cuya expresión, determinada en la subsección 1.3.1, es $(N - n_i)\sigma^2/(Nn_i)$, luego

$$\mathrm{E}\left[\frac{\sum_{i=1}^{I} n_i(\overline{u}_{i.}-\overline{u}_{..})^2}{I-1}\right] = \sum_{i=1}^{I}\frac{n_i}{I-1}\mathrm{E}\left[\overline{u}_{i.}-\overline{u}_{..}\right]^2 =$$

$$\sum_{i=1}^{I}\frac{n_i}{I-1}\mathrm{E}\left[(\bar{y}_{i.}-\bar{y}_{..})-\tau_i\right]^2 =$$

$$\sum_{i=1}^{I}\frac{n_i}{I-1}\mathrm{E}\left[\widehat{\tau}_i-\mathrm{E}\left[\widehat{\tau}_i\right]\right]^2 = \sum_{i=1}^{I}\frac{n_i}{I-1}\mathrm{Var}(\widehat{\tau}_i) =$$

$$\sum_{i=1}^{I}\frac{n_i}{I-1}(N-n_i)\frac{\sigma^2}{Nn_i} = \frac{\sigma^2}{N(I-1)}(IN-N) = \sigma^2 \tag{1.35}$$

c) Como $\mathrm{E}(\overline{u}_{i.}-\overline{u}_{..})=0$, entonces

$$\mathrm{E}\left[\frac{2\sum_{i=1}^{I} n_i\tau_i(\overline{u}_{i.}-\overline{u}_{..})}{I-1}\right] = \frac{2}{I-1}\sum_{i=1}^{I} n_i\tau_i\,\mathrm{E}\left[\overline{u}_{i.}-\overline{u}_{..}\right]=0 \quad . \tag{1.36}$$

Por lo tanto, sustituyendo las expresiones (1.34), (1.35) y (1.36) en (1.33) tenemos que el valor esperado del cuadrado medio entre grupos es:

$$\mathrm{E}\left[\widehat{S}^2_{Tr}\right] = \frac{\sum_{i=1}^{I} n_i\tau_i^2}{I-1}+\sigma^2 \quad . \tag{1.37}$$

2º) Ya hemos visto en la subsección 1.3.1 que la varianza residual es un estimador insesgado de la varianza poblacional, es decir

$$\mathrm{E}\left[\widehat{S}^2_{R}\right] = \sigma^2 \quad .$$

3º) Por último, calculemos el valor esperado del cuadrado medio total. Para ello nos basaremos en la ecuación básica del ANOVA que podemos poner en función de los cuadrados medios de la siguiente forma:

$$(N-1)\widehat{S}_T^2 = (I-1)\widehat{S}_{Tr}^2 + (N-I)\widehat{S}_R^2 \quad ,$$

tomando esperanzas matemáticas en ambos miembros y aplicando la linealidad del valor esperado, tenemos

$$(N-1)\,\mathrm{E}\left[\widehat{S}_T^2\right] = (I-1)\,\mathrm{E}\left[\widehat{S}_{Tr}^2\right] + (N-I)\,\mathrm{E}\left[\widehat{S}_R^2\right] \quad ,$$

de donde, sustituyendo los valores obtenidos anteriormente para $\mathrm{E}\left[\widehat{S}_{Tr}^2\right]$ y $\mathrm{E}\left[\widehat{S}_R^2\right]$, obtenemos

$$\mathrm{E}\left[\widehat{S}_T^2\right] = \frac{\sum_{i=1}^{I} n_i \tau_i^2}{N-1} + \sigma^2 \quad . \tag{1.38}$$

1.3.3. Análisis estadístico

El contraste estadístico de interés en este modelo, como mencionamos al principio de esta sección, es el que tiene como hipótesis nula la igualdad de medias de los tratamientos:

$$H_0 \equiv \mu_1 = \mu_2 = \cdots = \mu_I = \mu$$

o equivalentemente

$$H_0 \equiv \tau_1 = \tau_2 = \cdots = \tau_I = 0$$

Como hemos comprobado anteriormente se verifica que:

a) $\widehat{S}_R^2 = SCR/(N-I)$ es un estimador insesgado de la varianza σ^2 independientemente de que se verifique la hipótesis nula.

b) Y si no hay diferencia entre las medias de los I tratamientos; es decir, si es cierta la hipótesis nula, el primer sumando de $\mathrm{E}\left[\widehat{S}_{Tr}^2\right]$ es nulo, y entonces $\widehat{S}_{Tr}^2$ es un estimador insesgado de σ^2.

Sin embargo, hay que notar que si existe diferencia en las medias de los tratamientos, el valor esperado del cuadrado medio entre tratamientos es mayor que σ^2. De todo ésto podemos deducir que el contraste puede efectuarse comparando $\widehat{S}_{Tr}^2$ y $\widehat{S}_R^2$.

Para ello, vamos a estudiar la distribución de SCT, $SCTr$ y SCR en la hipótesis de que los tratamientos no influyen, es decir bajo la hipótesis de que las variables aleatorias $y_{ij} \rightsquigarrow N(\mu, \sigma^2)$.

Tipificando las variables aleatorias y_{ij} en la descomposición (1.23) , se tiene

$$\frac{y_{ij}-\mu}{\sigma} = \frac{\bar{y}_{..}-\mu}{\sigma} + \frac{\bar{y}_{i.}-\bar{y}_{..}}{\sigma} + \frac{y_{ij}-\bar{y}_{i.}}{\sigma} \quad . \tag{1.39}$$

Considerando esta descomposición para todas las observaciones y expresándola en forma vectorial, tenemos

$$\mathbf{Z} = \mathbf{Z}_1 + \mathbf{Z}_2 + \mathbf{Z}_3 \tag{1.40}$$

siendo

$$\mathbf{Z} = \frac{1}{\sigma}(y_{11}-\mu, \ldots, y_{1n_1}-\mu, y_{21}-\mu, \ldots, y_{2n_2}-\mu, \ldots, y_{I1}-\mu, \ldots, y_{In_I}-\mu)'$$

$$\mathbf{Z}_1 = \frac{1}{\sigma}(\bar{y}_{..}-\mu, \ldots.., \bar{y}_{..}-\mu, \bar{y}_{..}-\mu, \ldots.., \bar{y}_{..}-\mu, \ldots.., \bar{y}_{..}-\mu, \ldots.., \bar{y}_{..}-\mu)'$$

$$\mathbf{Z}_2 = \frac{1}{\sigma}(\bar{y}_{1.}-\bar{y}_{..}, \ldots.., \bar{y}_{1.}-\bar{y}_{..}, \bar{y}_{2.}-\bar{y}_{..}, \ldots, \bar{y}_{2.}-\bar{y}_{..}, \ldots, \bar{y}_{I.}-\bar{y}_{..}, \ldots, \bar{y}_{I.}-\bar{y}_{..})'$$

$$\mathbf{Z}_3 = \frac{1}{\sigma}(y_{11}-\bar{y}_{1.}, \ldots, y_{1n_1}-\bar{y}_{1.}, y_{21}-\bar{y}_{2.}, \ldots, y_{2n_2}-\bar{y}_{2.}, \ldots., y_{I1}-\bar{y}_{I.}, \ldots, y_{In_I}-\bar{y}_{I.})'$$

donde

- $\mathbf{Z}$: Contiene N términos independientes $\frac{1}{\sigma}(y_{ij}-\mu)$. Tiene, por tanto, N grados de libertad.
- $\mathbf{Z}_1$: Contiene N coordenadas iguales a $\frac{1}{\sigma}(\bar{y}_{..}-\mu)$. Tiene, por tanto, un grado de libertad.
- $\mathbf{Z}_2$: Contiene I valores distintos $\frac{1}{\sigma}(\bar{y}_{i.}-\bar{y}_{..})$, cada uno repetido n_i veces y sujetos a una ecuación de restricción, $\sum_i n_i(\bar{y}_{i.}-\bar{y}_{..}) = 0$. Tiene, por tanto, $I-1$ grados de libertad.
- $\mathbf{Z}_3$: Contiene N coordenadas $\frac{1}{\sigma}(y_{ij}-\bar{y}_{i.})$, sujetas a I ecuaciones de restricción, $\sum_j(y_{ij}-\bar{y}_{i.}) = 0$ para $i = 1, \ldots, I$. Tiene, por tanto, $N-I$ grados de libertad.

Bajo la hipótesis nula hemos realizado una descomposición del vector $\mathbf{Z}$, de variables $N(0,1)$ independientes, en componentes ortogonales. Por lo tanto, podemos aplicar el Teorema de Cochran de descomposición en formas cuadráticas cuyo enunciado es el siguiente:

Teorema 1.1

Consideremos un vector $\mathbf{X}$ *de dimensión* n *cuyas coordenadas son variables aleatorias independientes con distribución Normal de media 0 y desviación típica 1,* $N(0,1)$. *Supongamos que:*

a)

$$\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_r \qquad (r \leq n)$$

donde $\mathbf{X}_j$ *tiene* n_j *grados de libertad,* $(j = 1, 2, \cdots, r)$.

b) Los vectores $\mathbf{X}_j$ *son ortogonales entre sí y por tanto* $n = \sum_{j=1}^{r} n_j$.

En estas condiciones se verifica que los cuadrados de los módulos de cada uno de los vectores se distribuyen como variables aleatorias χ^2 *independientes con* n_j *grados de libertad.*

Puesto que la descomposición (1.40) cumple las condiciones del Teorema de Cochran, se verifica que

i)

$$\frac{SCTr}{\sigma^2} = \frac{\sum_i n_i(\overline{y}_{i.} - \bar{y}_{..})^2}{\sigma^2} \rightsquigarrow \chi^2_{I-1}$$

ii)

$$\frac{SCR}{\sigma^2} = \frac{\sum_{i,j}(y_{ij} - \bar{y}_{i.})^2}{\sigma^2} \rightsquigarrow \chi^2_{N-I}$$

y además estas dos distribuciones son independientes entre sí.

Hay que notar que:

- SCR/σ^2 se distribuye como una χ^2 con $N - I$ grados de libertad, se verifique o no la hipótesis nula, como ya vimos en la subsección 1.3.1
- $SCTr/\sigma^2$ se distribuye como una χ^2 con $I - 1$ grados de libertad, solamente cuando se verifique la hipótesis nula.

Por consiguiente, bajo la hipótesis nula, el cociente

$$F = \frac{\dfrac{SCTr/\sigma^2}{I-1}}{\dfrac{SCR/\sigma^2}{N-I}} = \frac{\widehat{S}^2_{Tr}}{\widehat{S}^2_R} \tag{1.41}$$

sigue una distribución F de Snedecor con $I-1$ y $N-I$ grados de libertad y será el estadístico de contraste para probar dicha hipótesis nula. Por otra parte, si H_0 es cierta, tanto el numerador como el denominador del estadístico de contraste (1.41) son estimadores insesgados de σ^2, mientras que si H_0 no es cierta, la esperanza matemática de $\widehat{S}^2_{Tr}$ será mayor que σ^2. Por tanto, rechazaremos H_0 cuando el valor de dicho estadístico sea mayor que el correspondiente valor teórico de la distribución F con $I-1$ y $N-I$ grados de libertad al nivel de significación α.

El procedimiento práctico para efectuar el contraste es el siguiente:

1º) Se fija un nivel de significación α

2º) Se calcula el valor experimental de F, F_{exp}, dado por $\widehat{S}^2_{Tr}/\widehat{S}^2_R$

3º) Se compara el valor F_{exp} con la F teórica al nivel de significación α, tomándose la siguiente decisión

$$\text{aceptar } H_0 \text{ si} \quad F_{exp} \leq F_{\alpha;I-1,N-I}$$

$$\text{rechazar } H_0 \text{ si} \quad F_{exp} > F_{\alpha;I-1,N-I} \quad .$$

La hipótesis H_0 que se contrasta es que simultáneamente los $\tau_i = 0$, de modo que rechazar dicha hipótesis quiere decir que al menos uno de los efectos es distinto de cero.

Para una mayor sencillez en el cálculo se utilizan las expresiones abreviadas de SCT, $SCTr$ y SCR

$$\begin{aligned} SCT &= \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N} \\ SCTr &= \sum_{i=1}^{I} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N} \\ SCR &= \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{I} \frac{y_{i.}^2}{n_i} \quad , \end{aligned} \tag{1.42}$$

que se obtienen de forma inmediata de la definición de cada uno de los términos.

El contraste básico del análisis de la varianza utiliza la descomposición, (1.28), *ecuación básica del análisis de la varianza,* cuyos términos se pueden disponer en forma tabular de la siguiente manera

Tabla 1-3. Tabla ANOVA para el modelo de efectos fijos unifactorial

Fuentes de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F_{exp}
Entre grupos	$\sum_{i=1}^{I} n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 = SCTr$	$I-1$	$\widehat{S}^2_{Tr}$	$\widehat{S}^2_{Tr}/\widehat{S}^2_R$
Dentro de grupos	$\sum_{i=1}^{I}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.}\right)^2 = SCR$	$N-I$	$\widehat{S}^2_R$	
TOTAL	$\sum_{i=1}^{I}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{..}\right)^2 = SCT$	$N-1$	$\widehat{S}^2_T$	

Alternativamente, utilizando las expresiones abreviadas de SCT, $SCTr$ y SCR, dadas en (1.42), la Tabla ANOVA se expresa de la siguiente forma

Tabla 1-4. Forma práctica de la tabla ANOVA para el modelo de efectos fijos unifactorial

Fuentes de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F_{exp}
Entre grupos	$\sum_{i=1}^{I} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N} = SCTr$	$I-1$	$\widehat{S}^2_{Tr}$	$\widehat{S}^2_{Tr}/\widehat{S}^2_R$
Dentro de grupos	$\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{I} \frac{y_{i.}^2}{n_i} = SCR$	$N-I$	$\widehat{S}^2_R$	
TOTAL	$\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N} = SCT$	$N-1$	$\widehat{S}^2_T$	

Coeficiente de determinación

La adecuación de los datos al modelo se podría comprobar mediante la varianza residual, pero esta cantidad tiene el inconveniente de depender de la escala de medida de los datos. Por ello, una medida más apropiada es el coeficiente de determinación, denotado por R^2 y definido como el cociente entre la variabilidad explicada y la variabilidad total

$$R^2 = \frac{SCTr}{SCT} \quad .$$

Esta cantidad es adimensional y se interpreta como la proporción de la variabilidad total presente en los datos que es explicada por el modelo de análisis de la varianza.

Para ilustrar el análisis de la varianza unifactorial de efectos fijos (caso no-equilibrado), vamos a considerar el Ejemplo 1-1, en el que se desea comprobar si se aprecian diferencias significativas en el rendimiento de la semilla de algodón con los distintos fertilizantes.

Para ello, construimos la Tabla 1-5, organizando los datos de la siguiente manera

Tabla 1-5. Datos del rendimiento del algodón

Fertiliz.	Observaciones						n_i	$y_{i.}$	$\bar{y}_{i.}$	$\sum y_{ij}^2$	$y_{i.}^2/n_i$
1	51	49	50	49	51	50	6	300	50	15004	15000
2	56	60	56	56	57		5	285	57	16257	16245
3	48	50	53	44	45		5	240	48	11574	11520
4	47	48	49	44			4	188	47	8850	8836
5	43	43	46	47	45	46	6	270	45	12164	12150
							26	$y_{..} = 1283$		63849	63751

que facilita los cálculos del análisis de la varianza.

Las sumas de cuadrados necesarias para el análisis de la varianza se calculan como sigue:

$$SCT = \sum_{i=1}^{5}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N} = 63849 - \frac{(1283)^2}{26} = 537{,}88$$

$$SCTr = \sum_{i=1}^{5} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N} = 63751 - \frac{(1283)^2}{26} = 439{,}88$$

$$SCR = SCT - SCTr = 98$$

El análisis de la varianza resultante se presenta en la siguiente tabla.

Tabla 1-6. Análisis de la varianza para los datos del Ejemplo 1-1

Fuentes de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F_{exp}
Entre grupos	439.88	4	109.97	23.55
Dentro de grupos	98.00	21	4.67	
TOTAL	537.88	25		

Obsérvese en esta tabla como el cuadrado medio entre tratamientos (109.97) es mucho mayor que el cuadrado medio dentro de los tratamientos (4.67), entonces debe ser muy improbable que los efectos de los tratamientos sean iguales. Efectivamente si realizamos el contraste al 5 % y comparamos el cociente $F_{exp} = 109{,}97/4{,}67 = 23{,}55$ con la F teórica ($F_{0{,}05;4,21} = 2{,}84$), se concluye que se rechaza H_0; en otras palabras, concluimos que, a un nivel de significación del 5 %, el rendimiento de la semilla de algodón difiere significativamente dependiendo del tipo de fertilizante utilizado. Igualmente ocurriría al nivel de significación del 1 %, ($F_{0{,}01;4,21} = 4{,}36$) o incluso a un nivel de significación mucho más pequeño. Más adelante, veremos otra forma de decidir en un contraste de hipótesis por medio del nivel mínimo de significación.

Comprobamos mediante el coeficiente de determinación, cuyo valor es

$$R^2 = \frac{SCTr}{SCT} = \frac{439{,}88}{537{,}88} = 0{,}8178 \quad ,$$

que el factor "tipo de fertilizante" explica el 81.78 % de la variabilidad en el rendimiento de la semilla de algodón.

Para la ejecución práctica de estos cálculos se suele requerir el empleo del ordenador y el uso de un software apropiado. En la sección **??**, mostraremos la utilización del paquete estadístico STATGRAPHICS.

En el caso de que se rechace la hipótesis nula, resulta de interés estudiar que tratamientos son distintos entre sí. Este tema será tratado con detalle en el Capítulo 2. Por otro lado, se puede ampliar el análisis estadístico realizado incluyendo intervalos de confianza para los parámetros del modelo, en especial μ_i y σ^2.

Intervalos de confianza para μ_i y σ^2

a) En primer lugar construyamos un intervalo de confianza para estimar la media del i-ésimo tratamiento,

$$\mu_i = \mu + \tau_i \quad .$$

Como hemos mencionado el estimador puntual de μ_i es $\widehat{\mu}_i = \widehat{\mu} + \widehat{\tau}_i = \bar{y}_{i.}$. Al ser las y_{ij} variables aleatorias independientes con distribución Normal de media μ_i y varianza σ^2, entonces las $\bar{y}_{i.}$ son también variables aleatorias independientes con distribución Normal de media μ_i y varianza σ^2/n_i. Por lo tanto, si σ^2 es conocida, podría usarse la distribución normal para construir el intervalo de confianza para μ_i. Como generalmente σ^2 es desconocida, se debe utilizar la varianza residual, $\widehat{S}_R^2$, como estimador de σ^2 y el intervalo de confianza, en este caso, se construye a partir

de la distribución t de Student. Así, un intervalo de confianza al nivel de confianza $(1-\alpha)$ para la media μ_i del i-ésimo tratamiento, es:

$$\left[\bar{y}_{i.} \pm t_{\alpha/2;N-I}\sqrt{\widehat{S}_R^2/n_i}\right] \quad . \tag{1.43}$$

En efecto, utilizando el resultado $N\widehat{\sigma}^2/\sigma^2 \rightsquigarrow \chi^2_{N-I}$ que obtuvimos en la subsección 1.3.1 y como $\widehat{\sigma}^2$ y $\widehat{S}_R^2$ están relacionadas de la siguiente forma:

$$\widehat{S}_R^2 = \frac{N}{N-I}\widehat{\sigma}^2 \quad ,$$

se deduce que:

$$\frac{(N-I)\widehat{S}_R^2}{\sigma^2} \rightsquigarrow \chi^2_{N-I} \quad , \tag{1.44}$$

y por tanto

$$\frac{\dfrac{\bar{y}_{i.}-\mu_i}{\sigma/\sqrt{n_i}}}{\sqrt{\dfrac{(N-I)\widehat{S}_R^2}{\sigma^2(N-I)}}} = \frac{\bar{y}_{i.}-\mu_i}{\sqrt{\widehat{S}_R^2/n_i}} \rightsquigarrow t_{N-I}$$

b) A continuación, construyamos un intervalo de confianza para la varianza poblacional σ^2, para ello utilizamos el resultado (1.44), obteniendo el siguiente intervalo para σ^2

$$\left(\frac{(N-I)\widehat{S}_R^2}{\chi^2_{\alpha/2;N-I}} \quad , \quad \frac{(N-I)\widehat{S}_R^2}{\chi^2_{1-\alpha/2;N-I}}\right) \tag{1.45}$$

donde $\chi^2_{1-\alpha/2}$ y $\chi^2_{\alpha/2}$ son, respectivamente, los puntos críticos inferior y superior de una variable χ^2 con $N-I$ grados de libertad y con una probabilidad de $\alpha/2$ en cada cola de la distribución.

Con los datos del Ejemplo 1-1 vamos a obtener intervalos de confianza para la media de uno de los niveles y para la varianza poblacional.

a) Usando la ecuación (1.43), un intervalo de confianza al 95 % para la media, por ejemplo, del tratamiento 5 es

$$\left[45 \pm t_{0,025;21}\sqrt{4{,}67/6}\right] = (45 \pm 1{,}835) \quad .$$

Por tanto, el intervalo deseado para μ_5 es (43,164 , 46,835).

b) Usando la ecuación (1.45), un intervalo de confianza al 95 % para la varianza poblacional es

$$\left(\frac{21(4{,}67)}{\chi^2_{0{,}025;21}} \ , \ \frac{21(4{,}67)}{\chi^2_{0{,}975;21}}\right) = (2{,}764 \ , \ 9{,}539) \quad .$$

1.3.4. Modelo equilibrado

Un caso muy importante del modelo unifactorial es el *modelo equilibrado o balanceado*, en el que para cada nivel del factor se toma el mismo número de observaciones. Este modelo presenta las siguientes ventajas sobre el modelo no-equilibrado:

1) Se simplifica el proceso de cálculo y además permite hacer la transición sencilla al modelo en bloques completos al azar, que estudiaremos en el Capítulo 4.

2) La restricción $\sum_i n_i\tau_i = 0$ del modelo no-equilibrado se simplifica a $\sum_i \tau_i = 0$, que resulta mucho más natural.

3) Los contrastes resultantes son más robustos, es decir, más insensibles al incumplimiento de las hipótesis de normalidad y homocedasticidad.

4) La potencia del contrate de comparación de medias es máxima.

5) Las comparaciones múltiples, que veremos en el Capítulo 2, se abordan de manera exacta con cualquiera de los métodos posibles.

En este modelo, la Tabla ANOVA 1-4 del modelo no-equilibrado se simplifica, obteniéndose la Tabla 1-7, donde n es el tamaño común de cada muestra y, por tanto, In es el número total de elementos N.

Tabla 1-7. Forma práctica de la tabla ANOVA para el modelo equilibrado

Fuentes de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F_{exp}
Entre grupos	$\sum_{i=1}^{I} \frac{y_{i.}^2}{n} - \frac{y_{..}^2}{N} = SCTr$	$I-1$	$\widehat{S}^2_{Tr}$	$\widehat{S}^2_{Tr}/\widehat{S}^2_R$
Dentro de grupos	$\sum_{i=1}^{I}\sum_{j=1}^{n} y_{ij}^2 - \sum_{i=1}^{I} \frac{y_{i.}^2}{n} = SCR$	$N-I$	$\widehat{S}^2_R$	
TOTAL	$\sum_{i=1}^{I}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{N} = SCT$	$N-1$	$\widehat{S}^2_T$	

Para ilustrar el análisis de la varianza unifactorial de efectos fijos en el caso equilibrado, vamos a considerar la segunda situación con unos datos concretos:

Ejemplo 1.2

Una profesora de estadística imparte clase en 4 grupos de alumnos, en los que explica la misma materia pero siguiendo distintos métodos de enseñanza. Desea averiguar si el método de enseñanza utilizado influye en las calificaciones de los alumnos. Las calificaciones medias obtenidas por los alumnos correspondientes a los 4 grupos fueron

***Tabla 1-8**. Datos para el Ejemplo 1-2*

Grupos	*Calificaciones*											
1	*8.2*	*7.3*	*7.2*	*6.1*	*3.2*	*8.5*	*2.5*	*5.5*	*5.3*	*4.4*	*3.8*	*10*
2	*6.4*	*3.8*	*3.5*	*9.1*	*8.2*	*7.5*	*3.6*	*2.5*	*6.5*	*5.3*	*5.2*	*5.1*
3	*9.2*	*10*	*8.1*	*5.3*	*2.5*	*2.6*	*6.1*	*9.5*	*10*	*4.2*	*2.1*	*0.0*
4	*8.4*	*7.1*	*6.3*	*4.1*	*3.4*	*5.2*	*6.1*	*4.3*	*3.3*	*3.5*	*9.2*	*8.2*

Construimos la Tabla 1-9, organizando los datos de la siguiente manera

Tabla 1-9. Datos del Ejemplo 1-2

Grupos	Observaciones	n	$y_{i.}$	$\bar{y}_{i.}$	$\sum_j y_{ij}^2$	$y_{i.}^2$
1	$8.2 \cdots 10$	12	72.0	6.00	490.46	5184.00
2	$6.4 \cdots 5.1$	12	66.7	5.55	416.55	4448.88
3	$9.2 \cdots 0.0$	12	69.6	5.80	540.86	4844.16
4	$8.4 \cdots 8.2$	12	69.1	5.75	446.79	4771.20
		48	$y_{..} = 277.4$		1894.66	19248.24

Las sumas de cuadrados necesarias para el análisis de la varianza se calculan como sigue:

$$SCT = \sum_{i=1}^{4}\sum_{j=1}^{12} y_{ij}^2 - \frac{y_{..}^2}{N} = 1894{,}66 - \frac{(277{,}4)^2}{48} = 291{,}51$$

$$SCTr = \sum_{i=1}^{4} \frac{y_{i.}^2}{n} - \frac{y_{..}^2}{N} = \frac{19248{,}24}{12} - \frac{(277{,}4)^2}{48} = 1{,}18$$

$$SCR = SCT - SCTr = 290{,}33 \quad .$$

El análisis de la varianza se presenta en la Tabla 1-10.

Tabla 1-10. Análisis de la varianza para los datos del Ejemplo 1-2

Fuentes de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F_{exp}
Entre grupos	1.18	3	0.39	0.060
Dentro de grupos	290.33	44	6.59	
TOTAL	291.51	47		

Al nivel de significación del 5 % se debe aceptar la hipótesis H_0 y concluir que las medias de los tratamientos no difieren significativamente ya que $F_{exp} = 0{,}39/6{,}59 = 0{,}060$ es menor que la F teórica ($F_{0{,}05;3,44} = 2{,}81$); en otras palabras, decidimos que, a un nivel de significación del 5 %, las calificaciones obtenidas por los alumnos en los 4 grupos no difieren significativamente[2].

Comportamiento de los datos frente a un cambio de origen y de escala

En esta sección vamos a comprobar que el análisis de la varianza se obtiene de forma equivalente cuando se transforman los datos mediante un cambio de origen y un cambio de escala. Para ello, utilizaremos los datos del Ejemplo 1-1.

Cambio de origen

Supongamos que se efectúa un cambio de origen en las observaciones, como valor conveniente del origen se debe tomar un valor próximo a la media de los datos; en este ejemplo podemos tomar 49. Los resultados se presentan en la Tabla 1-11.

[2] Al ser F_{exp} menor que la unidad no es necesario considerar la $F_{teorica}$ ya que siempre se acepta la hipótesis nula para cualquier α.

Tabla 1-11. Cambio de origen en los datos del Ejemplo 1-1

Fertiliz.	Valores transformados						n_i	$y_{i.}$	$\bar{y}_{i.}$	$\sum_j y_{ij}^2$	$y_{i.}^2/n_i$
1	2	0	1	0	2	1	6	6	1	10	6
2	7	11	7	7	8		5	40	8	332	320
3	−1	1	4	−5	−4		5	−5	−1	59	5
4	−2	−1	0	−5			4	−8	−2	30	16
5	−6	−6	−3	−2	−4	−3	6	−24	4	110	96
							26	$y_{..} = 9$		541	443

Las sumas de cuadrados son:

$$SCT = \sum_{i=1}^{5}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N} = 541 - \frac{(9)^2}{26} = 537{,}88$$

$$SCTr = \sum_{i=1}^{5} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N} = 443 - \frac{(9)^2}{26} = 439{,}88$$

$$SCR = SCT - SCTr = 98 \quad .$$

Observamos que al hacer un cambio de origen en los datos las sumas de cuadrados permanecen invariantes.

Cambio de escala

Veamos ahora el comportamiento de los datos frente a un cambio de escala. Para ello, supongamos que dividimos cada observación, por simplicidad, por 10. Así, se obtiene la siguiente tabla

Tabla 1-12. Cambio de escala en los datos del Ejemplo 1-1

Fertil.	Valores transformados						n_i	$y_{i.}$	$\bar{y}_{i.}$	$\sum y_{ij}^2$	$y_{i.}^2/n_i$
1	5.1	4.9	5.0	4.9	5.1	5.0	6	30.0	5.0	150.04	150.00
2	5.6	6.0	5.6	5.6	5.7		5	28.5	5.7	162.57	162.45
3	4.8	5.0	5.3	4.4	4.5		5	24.0	4.8	115.74	115.20
4	4.7	4.8	4.9	4.4			4	18.8	4.7	88.50	88.36
5	4.3	4.3	4.6	4.7	4.5	4.6	6	27.0	4.5	121.64	121.50
							26	$y_{..} = 128{,}3$		638.49	637.51

La correspondiente Tabla ANOVA es

Tabla 1-13.

Fuentes de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F_{exp}
Entre grupos	4.3988	4	1.0997	23.55
Dentro de grupos	0.9800	21	0.0467	
TOTAL	5.3788	25		

Comprobamos que la relación de las sumas de cuadrados en los datos del ejemplo original, ($SCT = 537{,}88$, $SCTr = 439{,}88$ y $SCR = 98$), con los valores transformados es la unidad de escala al cuadrado. Además el valor del estadístico de contraste es el mismo.

Todo ésto, que hemos visto para casos particulares se puede demostrar que es cierto para un cambio de origen y escala arbitrario. La finalidad práctica de estas transformaciones es simplificar las operaciones necesarias para obtener la tabla ANOVA.

Bibliografía utilizada

* **García Leal, J. & Lara Porras, A.M.** (1998). *"Diseño Estadístico de Experimentos. Análisis de la Varianza."* Grupo Editorial Universitario.

* **Lara Porras, A.M.** (2000). *"Diseño Estadístico de Experimentos, Análisis de la Varianza y Temas Relacionados: Tratamiento Informático mediante SPSS"* Proyecto Sur de Ediciones.