Capítulo 13 El análisis de varianza

13.2 Procedimiento del análisis de varianza

El procedimiento de *análisis de varianza* o ANOVA trata de analizar la variación de un conjunto de respuestas y asignar partes de esta variación a cada variable en un conjunto de variables independientes. El objetivo es identificar variables independientes importantes y determinar la forma en que afectan la respuesta.

La variabilidad de un conjunto de n mediciones es cuantificada por la suma de cuadrados de las desviaciones $\sum_{i=1}^{n}(y_i-\overline{y})^2$. El procedimiento ANOVA divide en partes esta suma de cuadrados de las desviaciones, llamada *suma total de cuadrados*, cada una de las cuales se atribuye a una de las variables independientes del experimento, más un residuo que está asociado con erro aleatorio.

El mecanismo del ANOVA puede ilustrarse mejor con un ejemplo.

Ejemplo

Suponga que deseamos usar información en muestras independientes de tamaños $n_1 = n_2$ para comparar las medias de dos poblaciones distribuidas normalmente con medias μ_1 y μ_2 y varianzas iguales $\sigma_1^2 = \sigma_2^2 = \sigma^2$. La variación total de las mediciones de respuesta de las dos muestras es cuantificada por

$$\text{SS}_{\text{total}} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y})^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_1}(Y_{ij}-\overline{Y})^2$$

donde Y_{ij} denota la j-ésima observación de la i-ésima muestra y $\overline{Y}$ es la media de todas las $n = 2n_1$ observaciones. Esta cantidad puede dividirse en dos partes de la siguiente manera:

$$\begin{aligned}\text{SS}_{\text{total}} &= \sum_{i=1}^{2}\sum_{j=1}^{n_1}(Y_{ij}-\overline{Y})^2 \\ &= \underbrace{n_1\sum_{i=1}^{2}(\overline{Y}_i-\overline{Y})^2}_{\text{SST}} + \underbrace{\sum_{i=1}^{2}\sum_{j=1}^{n_1}(Y_{ij}-\overline{Y}_i)^2}_{\text{SSE}}\end{aligned}$$

donde $\overline{Y}_i$ es el promedio de las observaciones de la i-ésima muestra, para $i = 1, 2$. Examinaremos la cantidad SSE, recordemos que se supuso que las varianzas poblacionales subyacente son iguales y que $n_1 = n_2$.

$$\begin{aligned}\text{SSE} &= \sum_{i=1}^{2}\sum_{j=1}^{n_1}(Y_{ij}-\overline{Y}_i)^2 = \sum_{i=1}^{2}(n_1-1)S_i^2 \\ &= (n_1-1)S_1^2 + (n_1-1)S_2^2\end{aligned}$$

donde $S_i^2 = \dfrac{1}{n_1-1}\sum_{j=1}^{n_1}(Y_{ij}-\overline{Y}_i)^2$. Recuerde que, en el caso $n_1 = n_2$, el estimador 'agrupado' para la varianza común σ^2 está dado por

$$\begin{aligned}S_p^2 &= \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} \\ &= \frac{(n_1-1)S_1^2+(n_1-1)S_2^2}{n_1+n_1-2} = \frac{\text{SSE}}{2n_1-2}\end{aligned}$$

Debido a que sólo hay dos tratamientos (o poblaciones) y $n_1 = n_2$, la otra parte,

$$\text{SST} = n_1\sum_{i=1}^{2}(\overline{Y}_i-\overline{Y})^2 = \frac{n_1}{2}(\overline{Y}_1-\overline{Y}_2)^2$$

la *suma de cuadrados de los tratamientos*, será grande si la diferencia de las medias muestrales es grande. En consecuencia, cuanto más grande sea SST, mayor será la evidencia de que existe una diferencia entre μ_1 y μ_2.

Debido a que hemos supuesto que Y_{ij} está distribuida normalmente con $E(Y_{ij}) = \mu_i$, para $i = 1, 2$ y $V(Y_{ij}) = \sigma^2$ y como $\text{SSE}/(2n_1-2)$ es un estimador agrupado de σ^2 se deduce que,

$$E\left(\frac{\text{SSE}}{2n_1-2}\right) = \sigma^2$$

y que SSE/σ^2 tiene una distribución χ^2 con $2n_1-2$ grados de libertad. En la sección 13.6 se obtendrá un resultado que implica que

$$E(\text{SST}) = \sigma^2 + \frac{n_1}{2}(\mu_1-\mu_2)^2$$

Entonces SST estima a la varianza si las medias poblacionales son iguales y una cantidad mayor si $\mu_1 \neq \mu_2$. Dada la hipótesis de que $\mu_1 = \mu_2$ se deduce que

$$Z = \frac{\overline{Y}_1-\overline{Y}_2}{\sqrt{2\sigma^2/n_1}}$$

tiene una distribución normal estándar y por lo tanto,

$$Z^2 = \left(\frac{n_1}{2}\right)\left[\frac{(\overline{Y}_1-\overline{Y}_2)^2}{\sigma^2}\right] = \frac{\text{SST}}{\sigma^2}$$

tiene una distribución χ^2 con 1 grado de libertad. Es claro ver como la SST es una función que sólo depende de las medias muestrales y que la SSE es una función que solo depende de las varianzas muestrales.

Ejemplo

Del Teorema 7.3 se tiene que la media muestral y la varianza muestral son independientes y por lo tanto se puede deducir que la SST y la SSE son variables aleatorias independientes. De acuerdo con la definición 7.3 se tiene que,

$$\frac{\dfrac{(\text{SST})/\sigma^2}{1}}{\dfrac{(\text{SSE}/\sigma^2)}{(2n_1-2)}} = \frac{\text{SST}/1}{\text{SSE}/(2n_1-2)}$$

tiene una distribución F con $v_1 = 1$ grado de libertad en el numerador y $v_2 = (2n_1-2)$ grados de libertad en el denominador. Al dividir las sumas de cuadrados entre sus respectivos grados de libertad se les denomina *cuadrados medios* y estan dados por,

$$\text{MSE} = \frac{\text{SSE}}{2n_1-2} \qquad \text{MST} = \frac{\text{SST}}{1}$$

De acuerdo con la hipótesis nula de igualdad de medias poblaciones, los cuadrados medios representan estimaciones de σ^2. Cuando la hipótesis nula es falsa y $\mu_1 \neq \mu_2$ entonces el cuadrado medio de los tratamientos (MST) es una estimación mayor que σ^2

Entonces para la prueba de $H_0 : \mu_1 = \mu_2$ contra $H_a : \mu_1 \neq \mu_2$ usamos

$$F = \frac{\text{MST}}{\text{MSE}}$$

como el estadístico de prueba y la región de rechazo con nivel de significancia α es $F > F_\alpha$. Entonces la prueba ANOVA resulta en una prueba F de una cola con grados de libertad asociados con MST y MSE.

Para el caso de comparar dos medias resultaría más sencillo utilizar la prueba t ya vista en otras secciones. Cuando se desea comprar varias medias esto ya no es factible y resulta más sencillo utilizar la prueba F como se verá en siguiente sección.

13.3 Comparación de más de dos medias: análisis de variana para un diseño de un factor

En esta sección se vera una generalización del método ANOVA visto en la sección anterior.
Suponga que muestras aleatorias independientes se han sacado de k poblaciones normales con medias $\mu_1, \mu_2, \ldots, \mu_k$, respectivamente y varianza común σ^2. Para que sean completamente generales, permitiremos que los tamaños muestrales sean desiguales y que n_i, para $i = 1, 2, \ldots, k$ sea el número de observaciones de la muestra tomadas de la i-ésima población. El número total de observaciones del experimento es $n = n_1 + n_2 + \cdots + n_k$.

Denotemos con Y_{ij} la respuesta para la j-ésima unidad experimental de la i-ésima muestra y representemos con $Y_{i\bullet}$ y $\overline{Y}_{i\bullet}$ el total y la media, respectivamente, de las n_i respuestas en la i-ésima muestra. El punto en la segunda posición del subíndice de $Y_{i\bullet}$ indica que esta cantidad se calcula sumando todos los posibles valores del subíndice que es sustituido por el punto, j en este caso. Además se denotaran con letras minúsculas las observaciones de las variables, por ejemplo y_{ij} representa las observaciones de Y_{ij}. Con esto se tiene lo siguiente,

$$Y_{i\bullet} = \sum_{i=1}^{n_i} Y_{ij} \qquad \overline{Y}_{i\bullet} = \left(\frac{1}{n_i}\right) \sum_{j=1}^{n_i} Y_{ij} = \left(\frac{1}{n_i}\right) Y_{i\bullet}$$

Al igual que el ANOVA con dos medias se tiene,

$$\text{SS}_{\text{total}} = \text{SST} + \text{SSE}$$

donde

$$\text{SS}_{\text{total}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2 - \text{CM}$$

El símbolo CM denota *corrección para la media* y esta definido como sigue,

$$\text{CM} = \frac{(\text{total de todas las observaciones})^2}{n} = \frac{1}{n}\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}\right)^2 = n\overline{Y}^2$$

$$\text{SST} = \sum_{i=1}^{k} n_i(\overline{Y}_{i\bullet} - \overline{Y})^2 = \frac{1}{n_i} \sum_{i=1}^{k} Y_{i\bullet}^2 - \text{CM}$$
$$\text{SSE} = \text{SS}_{\text{total}} - \text{SST}$$

En el recuadro anterior se muestra una forma más sencilla de encontrar la suma de cuadrados del error. A pesar de esto se observa que la SSE es la suma agrupada de cuadrados para todas las k muestras y es igual a,

$$\text{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet})^2 = \sum_{i=1}^{k} (n_i - 1) S_i^2$$

donde $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet})^2$. Cada uno de los S_i^2 valores proporciona un estimador insesgado para $\sigma_i^2 = \sigma^2$ con $n_i - 1$ grados de libertad. Entonces un estimador insesgados de σ^2 basado en $(n_1 + n_2 + \cdots + n_k - k) = n - k$ grados de libertad está dado por,

$$S^2 = \text{MSE} = \frac{\text{SSE}}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1)} = \frac{\text{SSE}}{n - k}$$

La cuadrado medio de tratamientos posee $(k-1)$ grados de libertad y es,

$$\text{MST} = \frac{\text{SST}}{k - 1}$$

Para probar la hipótesis nula, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ contra la alternativa de que al menos una de las igualdades no se cumple, usamos el estadístico F descrito en la sección pasada con base en $v_1 = k - 1$ y $v_2 = n - k$ grados de libertad en el numerador y denominador respectivamente. La hipótesis nula será rechzada si,

$$F = \frac{\text{MST}}{\text{MSE}} > F_\alpha$$

donde F_α es el valor crítico de F par auna prueba de nivel α.

Las desviaciones moderadas respecto a las suposiciones que se hicieron al inicio no afectarán de manera grave las propiedades de la prueba. Esto es particularmente cierto en la suposición de normalidad. La suposición de varianzas poblacionales iguales es menos crítica si los tamaños de las muestras de las poblaciones respectivas son todos iguales ($n_1 = n_2 = \cdots = n_k$). Se dice que un diseño de un factor con igual número de observaciones por tratamiento está *balanceado*.

13.4 Tabla de análisis de varianza para un diseño de un factor

Fuente	gl	SS	MS	F
Tratamientos	$k-1$	SST	$\text{MST} = \frac{\text{SST}}{k-1}$	$\frac{\text{MST}}{\text{MSE}}$
Error	$n-k$	SSE	$\text{MSE} = \frac{\text{SSE}}{n-k}$	
Total	$n-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2$		

13.5 Modelo estadístico para el diseño de un factor

Igual que antes, denotamos con Y_{ij} las variables aleatorias que generan los valores observados y_{ij}, para $i = 1, 2, \ldots, k$ y $j = 1, 2, \ldots, n_i$. Los valores Y_{ij} corresponden a muestras aleatorias independientes de poblaciones normales con $E(Y_{ij}) = \mu_i$ y $V(Y_{ij}) = \sigma^2$, para $i = 1, 2, \ldots, k$ y $j = 1, 2, \ldots, n_i$.

Modelo estadístico para un diseño de un factor
Para $i = 1, 2, \ldots, k$ y $j = 1, 2, \ldots, n_i$,

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$
$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

donde

Y_{ij} = la j-ésima observación de la población (tratamiento) i,
μ = la media general
τ_i = el efecto no aleatorio del tratamiento i,
donde $\sum_{i=1}^{k} \tau_i = 0$,
ε_{ij} = término de error aleatorio tales que ε_{ij} son variables aleatorias independientes, distribuidas normalmente, con $(\varepsilon_{ij}) = 0$ y $V(\varepsilon_{ij}) = \sigma^2$

13.6 Prueba de aditividad de las sumas de cuadrados y $E(\text{MST})$ para un diseño de un factor

Demostración de que $\text{SS}_{\text{total}} = \text{SST} + \text{SSE}$

Esta demostración usa resultado elementales en sumatorias y la técnica de sumar y restar $\overline{Y}_{i\bullet}$ dentro de la expresión para la suma de cuadrados totales. Entonces se tiene lo siguiente,

$$\begin{aligned}
\text{SS}_{\text{total}} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i\bullet}+\overline{Y}_{i\bullet}-\overline{Y})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(Y_{ij}-\overline{Y}_{i\bullet})+(\overline{Y}_{i\bullet}-\overline{Y})]^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big[(Y_{ij}-\overline{Y}_{i\bullet})^2+2(Y_{ij}-\overline{Y}_{i\bullet})(\overline{Y}_{i\bullet}-\overline{Y}) \\
&\quad +(\overline{Y}_{i\bullet}-\overline{Y})^2\Big]
\end{aligned}$$

Sumando primero para j, obtenemos

$$\begin{aligned}
\text{SS}_{\text{total}} = \sum_{i=1}^{k}\Bigg[&\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i\bullet})^2 \\
&+2(\overline{Y}_{i\bullet}-\overline{Y})\sum_{i=1}^{n_i}(Y_{ij}-\overline{Y}_{i\bullet})+n_i(\overline{Y}_{i\bullet}-\overline{Y})^2\Bigg]
\end{aligned}$$

donde

$$\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i\bullet}) = Y_{i\bullet}-n_i\overline{Y}_{i\bullet} = Y_{i\bullet}-Y_{i\bullet} = 0$$

En consecuencia, el término medio de la expresión para el SS_{total} es igual a cero. Entonces, sumando para i obtenemos,

$$\begin{aligned}
\text{SS}_{\text{total}} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i\bullet})^2+\sum_{i=1}^{k}n_i(\overline{Y}_{i\bullet}-\overline{Y})^2 \\
&= \text{SSE}+\text{SST}
\end{aligned}$$

La demostración aditividad de las sumas de cuadrados ANOVA para otros diseños experimentales se puede obtener de un modo semejante, aunque el procedimiento es a veces tedioso.

Demostración del valor esperado del MST

Se demostrara esto para un diseño de un factor. Usando el modelo estadístico para el diseño de un factor presentado en la sección anterior se deduce que,

$$\overline{Y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i}Y_{ij} = \frac{1}{n_i}\sum_{j=1}^{n_i}(\mu+\tau_i+\varepsilon_{ij}) = \mu+\tau_i+\overline{\varepsilon}_i$$

donde $\overline{\varepsilon} = \frac{1}{n_i}\sum_{j=1}^{n_i}\varepsilon_{ij}$. Como las ε_{ij} son variables independientes con $E(\varepsilon_{ij})=0$ y $V(\varepsilon_{ij})=\sigma^2$, el Teorema 5.12 implica que $E(\overline{\varepsilon}_i)$ y $V(\overline{\varepsilon}_i)=\sigma^2/n_i$

De un modo análogo, $\overline{Y}$ está dada por,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}Y_{ij} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mu+\tau_i+\varepsilon_i) = \mu+\overline{\tau}+\overline{\varepsilon}$$

donde

$$\overline{\tau} = \frac{1}{n}\sum_{i=1}^{k}n_i\tau_i \qquad\qquad \overline{\varepsilon} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\varepsilon_{ij}$$

Como los valores τ_i son constantes, $\overline{\tau}$ es simplemente una constante; de nuevo, usando el Teorema 5.12, obtenemos $E(\overline{\varepsilon})=0$ y $V(\overline{\varepsilon})=\sigma^2/n$. Por lo tanto, con respecto a los términos del modelo para el diseño de un factor,

$$\begin{aligned}
\text{MST} &= \left(\frac{1}{k-1}\right)\sum_{i=1}^{k}n_i(\overline{Y}_{i\bullet}-\overline{Y})^2 \\
&= \left(\frac{1}{k-1}\right)\sum_{i=1}^{k}n_i(\tau_i+\overline{\varepsilon}_i-\overline{\tau}-\overline{\varepsilon})^2 \\
&= \left(\frac{1}{k-1}\right)\sum_{i=1}^{k}n_i((\tau_i-\overline{\tau})^2 \\
&\quad +\left(\frac{1}{k-1}\right)\sum_{i=1}^{k}2n_i(\tau_i-\overline{\tau})(\overline{\varepsilon}_i-\overline{\varepsilon}) \\
&\quad +\left(\frac{1}{k-1}\right)\sum_{i=1}^{k}n_i(\overline{\varepsilon}_i-\overline{\varepsilon})^2
\end{aligned}$$

como $\overline{\tau}$ y τ_i, para $i=1,2,\ldots,k$, son constantes y $E(\varepsilon_{ij}) = E(\overline{\varepsilon}_i) = E(\overline{\varepsilon}) = 0$, se deduce que.

$$\begin{aligned}
E(\text{MST}) &= \left(\frac{1}{k-1}\right)\sum_{i=1}^{k}n_i(\tau_i-\overline{\tau})^2 \\
&\quad +\left(\frac{1}{k-1}\right)E\left[\sum_{i=1}^{k}n_i(\overline{\varepsilon}_i-\overline{\varepsilon})^2\right]
\end{aligned}$$

Demostración del valor esperado del MST

Observe que,

$$\begin{aligned}
\sum_{i=1}^{k}n_i(\overline{\varepsilon}_i-\overline{\varepsilon})^2 &= \sum_{i=1}^{k}(n_i\overline{\varepsilon}_i^2-2n_i\overline{\varepsilon}_i\overline{\varepsilon}+n_i\overline{\varepsilon}^2) \\
&= \sum_{i=1}^{k}n_i\overline{\varepsilon}_i^2-2n\overline{\varepsilon}^2+n\overline{\varepsilon}^2 = \sum_{i=1}^{k}n_i\overline{\varepsilon}_i^2-n\overline{\varepsilon}^2
\end{aligned}$$

Como $E(\overline{\varepsilon}_i)=0$ y $V(\overline{\varepsilon}_i)=\sigma^2/n_i$ se deduce que $E(\overline{\varepsilon}_i^2) = \sigma^2/n_i$, para $i=1,2,\ldots,k$. Del mismo modo, $E(\overline{\varepsilon}^2) = \sigma^2/n$ y por lo tanto,

$$\begin{aligned}
E\left[\sum_{i=1}^{k}n_i(\overline{\varepsilon}_i-\overline{\varepsilon})^2\right] &= \sum_{i=1}^{k}n_iE(\overline{\varepsilon}_i^2)-nE(\overline{\varepsilon}^2) \\
&= k\sigma^2-\sigma^2 = (k-1)\sigma^2
\end{aligned}$$

Resumiendo, obtenemos

$$E(\text{MST}) = \sigma^2+\left(\frac{1}{k-1}\right)\sum_{i=1}^{k}n_i(\tau_i-\overline{\tau})^2$$

donde $\overline{\tau} = \frac{1}{n}\sum_{i=1}^{k}n_i\tau_i$. Dada $H_0: \tau_1=\tau_2=\cdots=\tau_k = 0$, se deduce que $\overline{\tau}=0$ y por lo tanto, $E(\text{MST})=\sigma^2$. Entonces cuando la H_0 es verdadera, MST/MSE es la proporción entre dos estimadores insesgados para σ^2. Cuando la H_0 es falsa se entonces MST tendrá un valor más grande.

13.7 Estimación en un diseño de un factor

Se pueden desarrollar intervalos de confianza para una media de tratamientos y para la diferencia entre un par de medias de tratamiento, basados en datos obtenidos en un diseño de un factor. Los métodos son análogos al capítulo 8, la única diferencia es que se usa el MSE como estimador de la(s) varianza(s) poblacional(es) σ^2.

$$\overline{Y}_{i\bullet}\pm t_{\alpha/2}\frac{S}{\sqrt{n_i}} \quad \text{y} \quad (\overline{Y}_{i\bullet}-\overline{Y}_{i'\bullet})\pm t_{\alpha/2}S\sqrt{\frac{1}{n_i}+\frac{1}{n_{i'}}}$$

donde

$$S = \sqrt{S^2} = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n_1+n_2+\cdots+n_k-k}}$$

y t_α está basada en $(n-k)$ grados de libertad.

13.8 Modelo estadístico para el diseño de bloques aleatorizado

El diseño de bloques aleatorizado se emplea para compara k tratamientos usando b bloques. Los bloques se seleccionan de modo que, con optimismo, las unidades experimentales dentro de cada bloque sean homogéneas en esencia. Los tratamientos se asignan de manera aleatoria a las unidades experimentales de cada bloque, en forma tal que cada tratamiento aparece exactamente una vez en cada uno de los b bloques. Entonces, el número total de observaciones obtenidas en un diseños de bloques aleatorizado es $n = bk$.

Modelo estadístico para un diseño de bloques aleatorizado
Para $i = 1, 2, \ldots, k$ y $j = 1, 2, \ldots, b$,

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

donde
Y_{ij} = observación del tratamiento i en el bloque j,
μ = la media general
τ_i = efecto no aleatorio del tratamiento i, donde $\sum_{i=1}^{k} \tau_i = 0$
β_j = el efecto no aleatorio del bloque j, donde $\sum_{j=1}^{b} \beta_j = 0$
ε_{ij} = término de erro aleatorios tales que ε_{ij} son variables aleatorias independientes, distribuidas normalmente, con $E(\varepsilon_{ij}) = 0$ y $V(\varepsilon_{ij}) = \sigma^2$.

Este modelo difiere del de diseño completamente aleatorizado sólo en que contiene parámetro asociados con los diferentes bloques. Ya que se supone que los efectos de los bloques son fijos pero desconocidos, este modelo por lo general se conoce como modelo de *bloques de efectos fijos*.

Consideremos la observación Y_{ij} hecha en el tratamiento i en el bloque j. Observe que las suposiciones del modelo implican que $E(Y_{ij}) = \mu + \tau_i + \beta_j$ y $V(Y_{ij}) = \sigma^2$ para $i = 1, 2, \ldots, k$ y $j = 1, 2, \ldots, b$. Es fácil demostrar que las medias de dos observaciones hechas en bloques distintos solo difieren por la diferencia de los efectos de bloques, esto es $j \neq j'$

$$E(Y_{ij}) - E(Y_{ij'}) = \mu + \tau_i + \beta_j - (\mu + \tau_i + \beta_{j'}) = \beta_j - \beta_{j'}$$

Del mismo modo, dos observaciones que se tomen del mismo bloque tienen medias que difieren sólo por la diferencia de los efectos de tratamiento. Esto es $i \neq i'$,

$$E(Y_{ij}) - E(Y_{i'j}) = \mu + \tau_i + \beta_j - (\mu + \tau_{i'} + \beta_j) = \tau_i - \tau_{i'}$$

Observaciones que se tomen en diferentes tratamientos y diferentes bloques tienen medias que difieren en los efectos de bloques y tratamientos.

$$E(Y_{ij}) - E(Y_{i'j'}) = (\tau_i - \tau_{i'}) + (\beta_j - \beta_{j'})$$

13.9 El análisis de varianza para el diseño de bloques aleatorizado

El ANOVA para un diseño de bloques aleatorizado es similar al diseño completamente aleatorizado. La diferencia principal es el hecho de que la suma total de cuadrados, SS_{total} se divide en tres partes: la suma de cuadrados por bloques, tratamientos y error.

Denote el total y promedio de todas las observaciones del bloque j como $\overline{Y}_{\bullet j}$ y $\overline{Y}_{\bullet j}$, respectivamente. Del mismo modo, $Y_{i\bullet}$ y $\overline{Y}_{i\bullet}$ representan el total y el promedio de todas las observaciones que reciben el tratamiento i. Entonces para un diseño de bloques aleatorizado que contenga b bloques y k tratamientos tenemos las siguientes sumas de cuadrados:

$$SS_{total} = \sum_{i=1}^{k}\sum_{j=1}^{b} (Y_{ij} - \overline{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{b} Y_{ij}^2 - CM$$
$$= SSB + SST + SSE \quad \text{donde}$$
$$SSB = k\sum_{j=1}^{b} (\overline{Y}_{\bullet j} - \overline{Y})^2 = \frac{1}{k}\sum_{j=1}^{b} Y_{\bullet j}^2 - CM$$
$$SST = b\sum_{i=1}^{k} (\overline{Y}_{i\bullet} - \overline{Y})^2 = \frac{1}{b}\sum_{i=1}^{k} Y_{i\bullet}^2 - CM$$
$$SSE = SS_{total} - SSB - SST$$

En las fórmulas anteriores, $\overline{Y}$ es el promedio de todas las $n = bk$ obsrvaciones, o $\overline{Y} = \frac{1}{bk}\sum_{j=1}^{b}\sum_{i=1}^{k} Y_{ij}$ y

$$CM = \frac{(\text{total de todas las observaciones})^2}{n} = \frac{1}{bk}\left(\sum_{j=1}^{b}\sum_{i=1}^{k} Y_{ij}\right)^2$$

Al igual que en el diseño completamente aleatorizado se puede construir una tabla ANOVA de la siguiente manera,

Fuente	G.L.	SS	MS	F
Bloques	$b-1$	SSB	$MSB = \frac{SSB}{b-1}$	$\frac{MSB}{MSE}$
Tratamientos	$k-1$	SST	$MST = \frac{SST}{k-1}$	$\frac{MST}{MSE}$
Error	$(b-1)(k-1)$	SSE	MSE	
Total	$n-1$	SS_{total}		

Al igual que en la sección 13.3 se utiliza el estadístico de prueba F, solo que ahora se pueden hacer dos pruebas. La primera es la prueba de igualdad de medias o tratamientos y el estadístico de prueba es $F\frac{MST}{MSE}$. Este estadístico tiene región de rechazo $F > F_\alpha$ con base en $v_1 = (k-1)$ y $v_2 = (b-1)(k-1)$ grados de libertad en numerado y denominador, respectivamente.

La segunda prueba que se puede hacer es con respecto al efecto que tienen los bloques en la respuesta media. Entonces la hipótesis nula sería para $\beta_j = 0$ con $j = 1, 2, \ldots, b$. Al igual que las anteriores se utiliza el estadístico de prueba F ahora con el cuadrado medio de los bloques, esto es: $F = \frac{MSB}{MSE}$. Tiene región de rechazo $F > F_\alpha$ con base en $v_1 = b-1$ y $v_2 = (b-1)(k-1)$ grados de libertad en el numerador y denominador respectivamente.

13.10 Estimación en el diseño de bloques aleatorizado

El intervalo de confianza para la diferencia entre un par de medias de tratamiento en un diseño de bloques aleatorizado es análogo al del diseño visto en la sección 13.7. Un intervalo de confianza de $100(1-\alpha)\%$ para $\tau_i - \tau_{i'}$ es

$$(\overline{Y}_{i\bullet} - \overline{Y}_{i'\bullet}) \pm t_{\alpha/2} S\sqrt{\frac{2}{b}}$$

donde $n_i = n_{i'} = b$, el número de observaciones contenido en una media de tratamiento y $S = \sqrt{MSE}$. En este caso el valor de $t_{\alpha/2}$ está basado en $v = (b-1)(k-1)$ grados de libertad y que S se obtiene de la tabla ANOVA asociada con el diseño de bloques aleatorizado.

13.11 Selección del tamaño muestral

La determinación de los tamaños muestrales sigue un procedimiento similar para ambos diseños. Se hace un resumen de un método general. Primero el experimentador debe decidir sobre el parámetro (o parámetros) de interés principal. Por lo general esto comprende la comparación de un par de medias de un tratamiento. En el segundo paso, el experimentador debe especificar un límite del error de estimación que pueda ser tolerado. Una vez que esto se haya determinado, lo siguiente es seleccionar n_i (el tamaño de la muestra de la población o tratamiento i) o bien, de manera correspondiente, b (el número de bloques para un diseño de bloques aleatorizado) que reducirá el semi ancho del intervalo de confianza para el parámetro de modo que, en un nivel de confianza prescrito, sea menor o igual al límite especificado del error de estimación. Debe destacarse que la solución del tamaño muestra *siempre* será una aproximación porque σ es desconocido y una estimación para este es desconocida hasta que se obtenga la muestra. La mejor estimación disponible para σ se usará para producir una solución aproximada.

13.12 Intervalos de confianza simultáneos para más de un parámetro

En esta sección se presentara un procedimiento para formar conjuntos de intervalos de confianza para que el coeficiente de confianza *simultáneo* sea no menor que $(1-\alpha)$ para cualquier valor especificado de α.
Suponga que deseamos hallar intervalos de confianza $I_1, I_2, \ldots I_m$ para parámetros $\theta_1, \theta_2, \ldots, \theta_m$ para que

$$P(\theta_j \in I_j \quad \text{para toda } j = 1, 2, \ldots, m) \geq 1 - \alpha$$

Esta meta se puede alcanzar si se usa una desigualdad de probabilidad simple, conocida como *desigualdad de Bonferroni*. Para cualesquiera eventos $A_1, A_2, \ldots, A_m$ tenemos,

$$\overline{A_1 \cap A_2 \cap \cdots \cap A_m} = \overline{A}_1 \cup \overline{A}_2 \cup \cdots \cup \overline{A}_m$$

Por lo tanto, se tiene que

$$P(A_1 \cap A_2 \cap \cdots \cap A_m) = 1 - P(\overline{A}_1 \cup \overline{A}_2 \cup \cdots \cup \overline{A}_m)$$

También, de la ley aditiva de probabilidad sabemos que

$$P(\overline{A}_1 \cup \overline{A}_2 \cup \cdots \cup \overline{A}_m) \leq \sum_{j=1}^{m} P(\overline{A}_j)$$

En consecuencia, obtenemos la *desigualdad de Bonferroni*

$$P(A_1 \cap A_2 \cap \cdots \cap A_m) \geq 1 - \sum_{j=1}^{m} P(\overline{A}_j)$$

Suponga que $P(\theta_j \in I_j) = 1 - \alpha_j$ y sea A_j el evento $\{\theta_j \in I_j\}$. Entonces

$$P(\theta_1 \in I_1, \ldots, \theta_m \in I_m) \geq 1 - \sum_{j=1}^{m} P(\theta_j \notin I_j) = 1 - \sum_{j=1}^{m} \alpha_j$$

Si todas las α_j, para $j = 1, 2, \ldots, m$, se seleccionan iguales a α, podemos ver que el coeficiente de confianza simultáneo de los intervalos I_j, para $j = 1, 2, \ldots, m$ podría sera de sólo $(1 - m\alpha)$ que es menor que $(1-\alpha)$ si $m > 1$. Un coeficiente simultáneo de confianza de al menos $(1 - \alpha)$ puede asegurarse si se seleccionan los intervalos de confianza I_j, para que $\sum_{j=1}^{m} \alpha_j = \alpha$. Una forma de lograr este objetivo es que cada intervalo se construya par que tenga coeficiente de confianza $1 - (\alpha/m)$.

13.13 Análisis de varianza usando modelos lineales

En el capítulo 11 se vieron métodos para analizar los modelos lineales, estos se pueden adaptar para su uso en el ANOVA. Esto se ilustra al formular un modelo lineal para datos obtenidos de un diseño completamente aleatorizado que comprende $k = 2$ tratamientos.

Sea Y_{ij} la variable aleatoria en la j-ésima observación del tratamiento i, para $i = 1, 2$. Definamos una *variable x ficticia o indicadora*, de la siguiente manera:

$$x = \begin{cases} 1, & \text{si la observación es de la población 1} \\ 0, & \text{de otro modo} \end{cases}$$

Si usamos x como variable independiente en un modelo lineal, podemos modelar Y_{ij} como

$$Y_{ij} = \beta_0 + \beta_1 x + \varepsilon_{ij}$$

donde ε_{ij} es un error aleatorio distribuido normalmente con $E(\varepsilon_{ij}) = 0$ y $V(\varepsilon_{ij}) = \sigma^2$. En este modelo,

$$\mu_1 = E(Y_{1j}) = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$$

y

$$\mu_2 = E(Y_{2j}) = \beta_0 + \beta_1(0) = \beta_0$$

Entonces se deduce que $\beta_1 = \mu_1 - \mu_2$ y una prueba de la hipótesis $mu_1 - \mu_2 = 0$ es equivalente a la prueba de que β_1. Se puede demostrar que $\hat{\beta}_0 = \overline{Y}_{2\bullet}$ y $\hat{\beta}_1 = \overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}$ son los estimadores de mínimos cuadrados obtenidos al ajustar el modelo lineal anterior.