Capítulo 3

Experimentos con un solo factor (análisis de varianza)

Sumario

- Diseño completamente al azar y ANOVA
- Comparaciones o pruebas de rango múltiples
- Verificación de los supuestos del modelo
- Elección del tamaño de la muestra
- Uso de software computacional

Objetivos de aprendizaje

- Explicar los elementos de los diseños completamente al azar y el análisis de varianza; asimismo, conocer la importancia del tamaño de la muestra.
- Describir las diversas pruebas de rangos múltiples y la comparación por contrastes.
- Realizar la verificación de los supuestos del modelo.

Mapa conceptual

Conceptos clave

- Análisis de varianza
- Contraste
- Contrastes ortogonales
- Cuadrados medios
- Diagramas de cajas
- Diferencia mínima significativa (LSD)
- Diseño balanceado
- Gráfica de probabilidad en papel normal
- Método de Sheffé
- Métodos de comparaciones múltiples
- Modelo de efectos fijos
- Notación de puntos
- Residuos
- Tabla de análisis de varianza
- Tratamiento control
- Varianza constante

En el capítulo anterior vimos los métodos para comparar dos tratamientos o condiciones (poblaciones o procesos). En este capítulo, aunque se sigue considerando un solo factor, se presentan los diseños experimentales que se utilizan cuando el objetivo es comparar más de dos tratamientos. Puede ser de interés comparar tres o más máquinas, varios proveedores, cuatro procesos, tres materiales, cinco dosis de un fármaco, etcétera.

Es obvio que, al hacer tales comparaciones, existe un interés y un objetivo claro. Por ejemplo, una comparación de cuatro dietas de alimentación en la que se utilizan ratas de laboratorio, se hace con el fin de estudiar si alguna nueva dieta que se propone es mejor o igual que las ya existentes; en este caso, la variable de interés es el peso promedio alcanzado por cada grupo de animales después de ser alimentado con la dieta que le tocó.

Por lo general, el interés del experimentador está centrado en comparar los tratamientos en cuanto a sus medias poblacionales, sin olvidar que también es importante compararlos con respecto a sus varianzas. Así, desde el punto de vista estadístico, la hipótesis fundamental a probar cuando se comparan varios tratamientos es:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu$$
$$H_A : \mu_i \neq \mu_j \text{ para algún } i \neq j \qquad \textbf{(3.1)}$$

con la cual se quiere decidir si los tratamientos son iguales estadísticamente en cuanto a sus medias, frente a la alternativa de que al menos dos de ellos son diferentes. La estrategia natural para resolver este problema es obtener una muestra representativa de mediciones en cada uno de los tratamientos, y construir un estadístico de prueba para decidir el resultado de dicha comparación.

Se podría pensar que una forma de probar la hipótesis nula de la expresión (3.1) es mediante pruebas T de Student aplicadas a todos los posibles pares de medias; sin embargo, esta manera de proceder incrementaría de manera considerable el error tipo I (rechazar H_0 siendo verdadera). Por ejemplo, supongamos que se desea probar la igualdad de cuatro medias a través de pruebas T de Student. En este caso se tienen seis posibles pares de medias, y si la probabilidad de aceptar la hipótesis nula para cada prueba individual es de $1 - \alpha = 0.95$, entonces la probabilidad de aceptar las seis hipótesis nulas es de $0.95^6 = 0.73$, lo cual representa un aumento considerable del error tipo I. Aunque se utilice un nivel de confianza tal que $(1 - \alpha)^6 = 0.95$, el procedimiento resulta inapropiado porque se pueden producir sesgos por parte del experimentador. Por otra parte, existe un método capaz de probar la hipótesis de igualdad de las k medias con un solo estadístico de prueba, éste es el denominado *análisis de varianza*, el cual se estudiará más adelante.

Diseño completamente al azar y ANOVA

Muchas comparaciones, como las antes mencionadas, se hacen con base en el diseño completamente al azar (DCA), que es el más simple de todos los diseños que se utilizan para comparar dos o más tratamientos, dado que sólo consideran dos fuentes de variabilidad: los *tratamientos* y el *error aleatorio*. En el siguiente capítulo veremos diseños que consideran la influencia de otras fuentes de variabilidad (bloques).

Este diseño se llama *completamente al azar* porque todas las corridas experimentales se realizan en orden aleatorio completo. De esta manera, si durante el estudio se hacen en total N pruebas, éstas se corren al azar, de manera que los posibles efectos ambientales y temporales se vayan repartiendo equitativamente entre los tratamientos.

Ejemplo 3.1

Comparación de cuatro métodos de ensamble. Un equipo de mejora investiga el efecto de cuatro *métodos de ensamble* A, B, C y D, sobre el *tiempo de ensamble* en minutos. En primera instancia, la estrategia experimental es aplicar cuatro veces los cuatro métodos de ensamble en orden completamente aleatorio (las 16 pruebas en orden aleatorio). Los tiempos de ensamble obtenidos se muestran en la tabla 3.1. Si se usa el *diseño completamente al azar* (DCA), se supone que, además del método de ensamble, no existe ningún otro factor que influya de manera significativa sobre la variable de respuesta (tiempo de ensamble).

Más adelante veremos cómo investigar si las diferencias muestrales de la tabla 3.1 garantizan diferencias entre los métodos.

Ejemplo 3.2

Comparación de cuatro tipos de cuero. Un fabricante de calzado desea mejorar la calidad de las suelas, las cuales se pueden hacer con uno de los cuatro tipos de cuero A, B, C y D disponibles en el mercado. Para ello, prueba los cueros con una máquina que hace pasar los zapatos por una superficie abrasiva; la suela de éstos se desgasta al pasarla por dicha superficie. Como criterio de desgaste se usa la pérdida de peso después de un número fijo de ciclos. Se prueban en orden aleatorio 24 zapatos, seis de cada tipo de cuero. Al hacer las pruebas en orden completamente al azar se evitan sesgos y las mediciones en un tipo de cuero resultan independientes de las demás. Los datos (en miligramos) sobre el desgaste de cada tipo de cuero se muestran en la tabla 3.2.

Tabla 3.1 Diseño completamente al azar, ejemplo 3.1.

Método de ensamble			
A	***B***	***C***	***D***
6	7	11	10
8	9	16	12
7	10	11	11
8	8	13	9

Tabla 3.2 Comparación de cuatro tipos de cuero (cuatro tratamientos).

Tipo de cuero	Observaciones						Promedio
A	264	260	258	241	262	255	256.7
B	208	220	216	200	213	206	209.8
C	220	263	219	225	230	228	230.8
D	217	226	215	227	220	222	220.7

Tabla 3.3 Diseño completamente al azar.

Tratamientos				
T_1	T_2	T_3	…	T_k
Y_{11}	Y_{21}	Y_{31}	…	Y_{k1}
Y_{12}	Y_{22}	Y_{32}	…	Y_{k2}
Y_{13}	Y_{23}	Y_{33}	…	Y_{k3}
⋮	⋮	⋮	⋱	⋮
Y_{1n_1}	Y_{2n_2}	Y_{3n_3}	…	Y_{kn_k}

La primera interrogante a despejar es si existen diferencias entre el desgaste promedio de los diferentes tipos de cuero. A continuación veremos la teoría general del diseño y análisis de este tipo de experimentos (DCA), y más adelante se analizarán los datos de los ejemplos planteados.

Supongamos que se tienen k poblaciones o tratamientos, independientes y con medias desconocidas $\mu_1, \mu_2, \ldots, \mu_k$, así como varianzas también desconocidas pero que se suponen iguales $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$. Las poblaciones pueden ser k métodos de producción, k tratamientos, k grupos, etc., y sus medias se refieren o son medidas en términos de la variable de respuesta.

Se decide realizar un experimento completamente al azar para comparar las poblaciones, en principio mediante la hipótesis de igualdad de medias (relación 3.1). Los datos generados por un diseño completamente al azar para comparar dichas poblaciones se pueden escribir como en la tabla 3.3. El elemento Y_{ij} en esta tabla es la j-ésima observación que se hizo en el tratamiento i; n_i es el tamaño de la muestra o las repeticiones observadas en el tratamiento i. Es recomendable utilizar el mismo número de repeticiones ($n_i = n$) en cada tratamiento, a menos que hubiera alguna razón para no hacerlo.[1] Cuando $n_i = n$ para toda i se dice que el *diseño es balanceado*.

Diseño balanceado
Es cuando se utiliza el mismo número de repeticiones en cada tratamiento.

El número de tratamientos k es determinado por el investigador y depende del problema particular de que se trata. El número de observaciones por tratamiento (n) debe escogerse con base en la variabilidad que se espera observar en los datos, así como en la diferencia mínima que el experimentador considera que es importante detectar. Con este tipo de consideraciones, por lo general se recomiendan entre 5 y 30 mediciones en cada tratamiento. Por ejemplo, se usa $n = 10$ cuando las mediciones dentro de cada tratamiento tienen un comportamiento consistente (con poca dispersión). En el otro extremo, se recomienda $n = 30$ cuando las mediciones muestran bastante dispersión. Cuando es costoso o tardado realizar las pruebas para cada tratamiento se puede seleccionar un número menor de repeticiones, con lo cual sólo se podrán detectar diferencias grandes entre los tratamientos.

En caso de que los tratamientos tengan efecto, las observaciones Y_{ij} de la tabla 3.3 se podrán describir con el modelo estadístico lineal dado por:

[1] Si uno de los tratamientos resulta demasiado caro en comparación con los demás, se pueden plantear menos pruebas con éste. Por otra parte, cuando uno de los tratamientos es un control (tratamiento de referencia) muchas veces es el más fácil y económico de probar, y como es de interés comparar a todos los tratamientos restantes con el control, se recomienda realizar más corridas en éste para que sus parámetros queden mejor estimados.

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \textbf{(3.2)}$$

donde μ es el parámetro de escala común a todos los tratamientos, llamado *media global*, τ_i; es un parámetro que mide el efecto del tratamiento i y ε_{ij} es el error atribuible a la medición Y_{ij}. Este modelo implica que en el diseño completamente al azar actuarían a lo más dos fuentes de variabilidad: los *tratamientos* y el *error aleatorio*. La media global μ de la variable de respuesta no se considera una fuente de variabilidad por ser una constante común a todos los tratamientos, que hace las veces de punto de referencia con respecto al cual se comparan las respuestas medias de los tratamientos (véase figura 3.2). Si la respuesta media de un tratamiento particular μ_i es "muy diferente" de la respuesta media global μ, es un síntoma de que existe un efecto de dicho tratamiento, ya que como se verá más adelante, $\tau_i = \mu_i - \mu$. La diferencia que deben tener las medias entre sí para concluir que hay un efecto (que los tratamientos son diferentes), nos lo dice el análisis de varianza (ANOVA).

En la práctica puede suceder que los tratamientos que se desea comparar sean demasiados como para experimentar con todos. Cuando esto sucede es conveniente comparar sólo una muestra de la población de tratamientos, de modo que τ_i pasa a ser una variable aleatoria con su propia varianza σ_τ^2 que deberá estimarse a partir de los datos (véase sección "Modelos de efectos aleatorios" del capítulo 5). En este capítulo sólo se presenta el caso en que todos los tratamientos que se tienen se prueban, es decir, se supone una población pequeña de tratamientos, lo cual hace posible compararlos a todos. En este caso, el modelo dado por la ecuación (3.2) se llama *modelo de efectos fijos*.

Modelo de efectos fijos
Es cuando se estudian todos los posibles tratamientos.

ANOVA para el diseño completamente al azar (DCA)

El *análisis de varianza* (ANOVA) es la técnica central en el análisis de datos experimentales. La idea general de esta técnica es separar la variación total en las partes con las que contribuye cada fuente de variación en el experimento. En el caso del DCA se separan la variabilidad debida a los tratamientos y la debida al error. Cuando la primera predomina "claramente" sobre la segunda, es cuando se concluye que los tratamientos tienen efecto (figura 3.1*b*), o dicho de otra manera, las medias son diferentes. Cuando los tratamientos no dominan contribuyen igual o menos que el error, por lo que se concluye que las medias son iguales (figura 3.1*a*). Antes de comenzar

Análisis de varianza
Consiste en separar la variación total observada en cada una de las fuentes que contribuye a la misma.

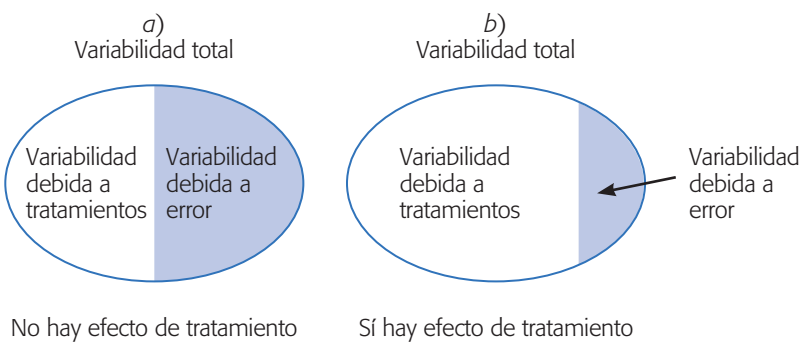

Figura 3.1 Partiendo la variación total en sus componentes en un DCA.

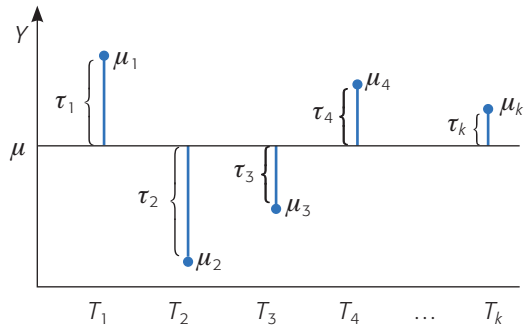

Figura 3.2 Representación de los efectos de los tratamientos en el DCA.

con el análisis del DCA se introduce alguna notación que simplifica la escritura de las expresiones involucradas en dicho análisis.

Notación de puntos
Sirve para representar sumas y medias que se obtienen a partir de los datos experimentales.

Notación de puntos

Sirve para representar de manera abreviada cantidades numéricas que se pueden calcular a partir de los datos experimentales, donde Y_{ij} representa la j-ésima observación en el tratamiento i, con $i = 1, 2, \ldots, k$ y $j = 1, 2, \ldots, n_i$. Las cantidades de interés son las siguientes:

$Y_{i\bullet}$ = Suma de las observaciones del tratamiento i.
$\bar{Y}_{i\bullet}$ = Media de las observaciones del i-ésimo tratamiento.
$Y_{\bullet\bullet}$ = Suma total de las $N = n_1 + n_2 + \ldots + n_k$ mediciones.
$\bar{Y}_{\bullet\bullet}$ = Media global o promedio de todas las observaciones.

Note que el punto indica la suma sobre el correspondiente subíndice. Así, algunas relaciones válidas son,

$$Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij}; \bar{Y}_{i\bullet} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}; Y_{\bullet\bullet} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$$

$$\bar{Y}_{\bullet\bullet} = \frac{Y_{\bullet\bullet}}{N}; i = 1,2,\ldots,k$$

donde $N = \Sigma_{i=1}^{k} n_i$ es el total de observaciones.

ANOVA

El objetivo del análisis de varianza en el DCA es probar la hipótesis de igualdad de los tratamientos con respecto a la media de la correspondiente variable de respuesta:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu$$
$$H_A : \mu_i \neq \mu_j \text{ para algún } i \neq j \quad \textbf{(3.3)}$$

la cual se puede escribir en forma equivalente como:

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_k = 0$$
$$H_A : \tau_i \neq 0 \text{ para algún } i \quad \textbf{(3.4)}$$

donde τ_i es el efecto del tratamiento i sobre la variable de respuesta. Si se acepta H_0 se confirma que los efectos sobre la respuesta de los k tratamientos son estadísticamente nulos (iguales a cero), y en caso de rechazar se estaría concluyendo que al menos un efecto es diferente de cero.

La equivalencia de las hipótesis (3.3) y (3.4) se deduce directamente del modelo asociado al diseño (ecuación 3.2),[2] pero se observa más fácilmente en la figura 3.2, que es una manera de representar el diseño completamente al azar. En dicha figura se ve que $\tau_i = \mu_i - \mu$, el efecto del tratamiento i, es la distancia entre la respuesta media del tratamiento, μ_i, y la respuesta media global, μ, y cuando un efecto es igual a cero equivale a decir que la media del tratamiento correspondiente es igual a la media global. Así, se observa que para que todas las respuestas medias de tratamientos sean iguales a la respuesta media global μ, representada por la línea horizontal, se requiere que todos los efectos τ_i sean iguales a cero.

Para probar la hipótesis dada por las relaciones (3.3) o (3.4) mediante la técnica de ANOVA, lo primero es descomponer la variabilidad total de los datos en sus dos componentes: la variabilidad debida a tratamientos y la que corresponde al error aleatorio, como se hace a continuación.

Una medida de la variabilidad total presente en las observaciones de la tabla 3.3 es la *suma total de cuadrados* dada por,

$$SC_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{N}$$

donde $Y_{\bullet\bullet}$ es la suma de los $N = \Sigma_{i=1}^{n_i} n_i$ datos en el experimento. Al sumar y restar adentro del paréntesis la media del tratamiento i, $(\bar{Y}_{i\bullet})$:

$$SC_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[(Y_{ij} - \bar{Y}_{i\bullet}) + (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) \right]^2$$

y desarrollando el cuadrado, la SC_T se puede partir en dos componentes como:

$$SC_T = \sum_{i=1}^{k} n_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2$$

donde el primer componente es la *suma de cuadrados de tratamientos* (SC_{TRAT}) y el segundo es la *suma de cuadrados del error* (SC_E). Al observar con detalle estas sumas de cuadrados se aprecia que la SC_{TRAT} mide la variación o diferencias *entre*

[2] Basta observar que $E(Y_{ij}) = \mu + \tau_i = \mu$, de modo que $\tau_i = \mu_i - \mu$.

tratamientos, ya que si éstos son muy diferentes entre sí, entonces la diferencia $\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$ tenderá a ser grande en valor absoluto, y con ello también será grande la SC_{TRAT}. Mientras que la SC_E mide la variación *dentro de tratamientos*, ya que si hay mucha variación entre las observaciones de cada tratamiento entonces $Y_{ij} - \bar{Y}_{i\cdot}$ tenderá a ser grande en valor absoluto. En forma abreviada, esta descomposición de la suma total de cuadrados se puede escribir como:

$$SC_T = SC_{TRAT} + SC_E \tag{3.5}$$

Como hay un total de $N = \Sigma_{i=1}^{n_i} n_i$ observaciones, la SC_T tiene $N - 1$ grados de libertad. Hay k tratamientos o niveles del factor de interés, así que SC_{TRAT} tiene $k - 1$ grados de libertad, mientras que la SC_E tiene $N - k$. Los grados de libertad que corresponden a los términos de la igualdad (3.5) cumplen una relación similar dada por:

$$N - 1 = (k - 1) + (N - k)$$

Cuadrados medios
Es la suma de cuadrados divididos entre sus respectivos grados de libertad.

Las sumas de cuadrados divididas entre sus respectivos grados de libertad se llaman *cuadrados medios*. Los dos que más interesan son el *cuadrado medio de tratamientos* y el *cuadrado medio del error*, que se denotan por

$$CM_{TRAT} = \frac{SC_{TRAT}}{k-1} \text{ y } CM_E = \frac{SC_E}{N-k}$$

Los valores esperados de los cuadrados medios están dados por

$$E(CM_E) = \sigma^2 \text{ y } E(CM_{TRAT}) = \sigma^2 + \frac{\sum_{i=1}^{k} n_i \tau_i^2}{N-k} \tag{3.6}$$

En estas expresiones se aprecia que cuando la hipótesis nula es verdadera, ambos cuadrados medios estiman la varianza σ^2, ya que el segundo término de la expresión para el $E(CM_{TRAT})$ sería igual a cero. Con base en este hecho se construye el estadístico de prueba como sigue: se sabe que SC_E y SC_{TRAT} son independientes, por lo que SC_E/σ^2 y SC_{TRAT}/σ^2 son dos variables aleatorias independientes con distribución ji-cuadrada con $N - k$ y $k - 1$ grados de libertad, respectivamente. Entonces, bajo el supuesto de que la hipótesis H_0 (relaciones 3.3 y 3.4) es verdadera, el estadístico

$$F_0 = \frac{CM_{TRAT}}{CM_E} \tag{3.7}$$

sigue una distribución F con $(k - 1)$ grados de libertad en el numerador y $(N - k)$ grados de libertad en el denominador. De las ecuaciones (3.6) y (3.7) se deduce que

si F_0 es grande, se contradice la hipótesis de que no hay efectos de tratamientos; en cambio, si F_0 es pequeño se confirma la validez de H_0. Así, para un nivel de significancia α prefijado, se rechaza H_0 si $F_0 > F_{\alpha, k-1, N-k}$, donde $F_{\alpha, k-1, N-k}$ es el percentil $(1-\alpha) \times 100$ de la distribución F. También se rechaza H_0 si el valor-$p < \alpha$, donde el valor-p es el área bajo la distribución $F_{k-1, N-k}$ a la derecha del estadístico F_0, es decir, el valor-$p = P(F > F_0)$.

Toda la información necesaria para calcular el estadístico F_0 hasta llegar al valor-p se escribe en la llamada *tabla de análisis de varianza* (ANOVA) que se muestra en la tabla 3.4. En esta tabla, las abreviaturas significan lo siguiente: FV = fuente de variabilidad (efecto), SC = suma de cuadrados, GL = grados de libertad, CM = cuadrado medio, F_0 = estadístico de prueba, valor-p = significancia observada.

Tabla de análisis de varianza
En ésta se resume el análisis de varianza de un experimento, que sirve para probar las hipótesis de interés.

Debemos señalar que el caso particular de comparar dos tratamientos suponiendo varianzas desconocidas pero iguales (prueba T de Student presentada en el capítulo 2), también se puede analizar con el ANOVA y se obtiene el mismo valor del valor-p que con la prueba T. Es fácil comprobar que el estadístico t_0 de la prueba T elevado al cuadrado es igual al estadístico F_0 (3.7) de la prueba F del ANOVA. Por último, es importante resaltar que el ANOVA supone que la variable de respuesta se distribuye *normal*, con *varianza constante* (los tratamientos tienen varianza similar) y que las mediciones son *independientes* entre sí. Estos supuestos deben verificarse para estar más seguros de las conclusiones obtenidas.

Análisis del ejemplo 3.2 (comparación de cuatro tipos de cuero). La interrogante que se planteó en el problema de la comparación entre los cuatro tipos de cuero fue: ¿existen diferencias entre el desgaste promedio de los diferentes tipos de cuero? La respuesta a esta pregunta es el resultado de contrastar las hipótesis:

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu$$
$$H_A : \mu_i \neq \mu_j \text{ para algún } i \neq j \qquad \textbf{(3.8)}$$

En la tabla 3.5 se muestra el análisis de varianza para este ejemplo. Como el valor-$p = 0.0000$ es menor que la significancia prefijada $\alpha = 0.05$, se rechaza H_0 y se

Tabla 3.4 Tabla de ANOVA para el DCA.

FV	SC	GL	CM	F_0	Valor-p
Tratamientos	$SC_{TRAT} = \sum_{i=1}^{k} \frac{Y_{i\bullet}^2}{n_i} - \frac{Y_{\bullet\bullet}^2}{N}$	$k-1$	$CM_{TRAT} = \frac{SC_{TRAT}}{k-1}$	$\frac{CM_{TRAT}}{CM_E}$	$P(F > F_0)$
Error	$SC_E = SC_T - SC_{TRAT}$	$N-k$	$CM_E = \frac{SC_E}{N-k}$		
Total	$SC_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{N}$	$N-1$			

acepta que al menos un par de tipos de cuero tienen un desgaste promedio diferente (la verificación de supuestos se deja al lector como ejercicio).

Si al menos un tipo de cuero se desgasta de forma diferente de otro, entonces ¿cuáles tipos de cuero son diferentes entre sí? Para responder esta pregunta se realizan todas las comparaciones posibles, dos a dos entre las medias de tratamientos, para lo cual existen varios métodos de prueba conocidos genéricamente como *métodos de comparaciones múltiples*, algunos de los cuales se presentan más adelante, junto con otros análisis gráficos que permiten entender mejor los resultados.

Método de comparaciones múltiples
Técnicas para comparar todos los posibles pares de medias de tratamientos.

Además de la tabla 3.5 del ANOVA se observa que la variación total en 24 datos de este experimento fue de 9 101. De esta cantidad, 7 072 se debe a las diferencias entre los tipos de cuero y 2 029 corresponde a la diferencia entre los cueros del mismo tipo. Al ponderar esto por los correspondientes grados de libertad, se obtienen los cuadrados medios que reflejan la magnitud real de cada fuente de variación. Así, vemos que las diferencias debido al tipo de cuero es de 2 357 y que el error es de 101; por lo tanto, la primera es 23.2 veces más grande que la segunda, lo cual indica que las diferencias observadas entre los tipos de cuero son significativas y que no se deben a pequeñas variaciones muestrales (error).

Ejemplo 3.3

Comparación de cuatro métodos de ensamble. Consideremos los datos del DCA dados en el ejemplo 3.1, donde el interés era comparar cuatro métodos de ensamble en cuanto al tiempo promedio en minutos que requiere cada uno de ellos. Se hicieron cuatro observaciones del tiempo de ensamble en cada método. Los resultados se muestran en la tabla 3.1.

Una manera de comparar los métodos de ensamble (tratamientos) es probar la hipótesis:

$$H_0 : \tau_A = \tau_B = \tau_C = \tau_D = 0$$
$$H_A : \tau_i \neq 0 \text{ para algún } i = A, B, C, D \qquad \textbf{(3.9)}$$

En caso de no rechazar H_0 se concluye que los tiempos promedio de los cuatro métodos de ensamble son estadísticamente iguales; pero si se rechaza, se concluye que al menos dos de ellos son diferentes. En la tabla 3.6 se muestra el análisis de varianza correspondiente, en donde se aprecia que el valor del valor-p = 0.0018 es menor que $\alpha = 0.05$, por lo que se rechaza H_0 en este nivel de significancia en particular. No obstante, también se rechazaría para cualquier otro nivel de significancia

Tabla 3.5 ANOVA para los tipos de cuero.

FV	SC	GL	CM	F_0	Valor-p
Tipo de cuero	7 072.33	3	2 357.44	23.24	0.0000
Error	2 029.0	20	101.45		
Total	9 101.33	23			

Tabla 3.6 ANOVA para los métodos de ensamble.

FV	SC	GL	CM	F_0	Valor-*p*
Tratamientos	69.5	3	23.17	9.42	0.0018
Error	29.5	12	2.46		
Total	99.0	15			

prefijado, α, que cumpla con $\alpha > 0.0018$, ya que en esos casos el estadístico de prueba $F_0 = 9.42$ caería en la región de rechazo.

Cálculos manuales

Hay personas que, cuando hacen los cálculos de forma manual, complementan el entendimiento de un análisis con el apoyo de una calculadora de bolsillo, al menos para los casos más simples. Para el caso del ANOVA, estos cálculos se facilitan si primero se obtiene la información básica desplegada en la tabla 3.7. Con esta información se pueden calcular las sumas de cuadrados, como se hace a continuación:

1. Suma total de cuadrados o variabilidad total de los datos:

$$SC_T = \sum_{i=j}^{4}\sum_{j=1}^{4} Y_{ij}^2 - \frac{Y_{\cdot\cdot}^2}{N} = 1\,620 - \frac{156^2}{16} = 99.0$$

2. Suma de cuadrados de tratamientos o variabilidad debida a la diferencia entre métodos de ensamble:

$$SC_{TRAT} = \sum_{i=1}^{4} \frac{Y_{i\cdot}^2}{4} - \frac{Y_{\cdot\cdot}^2}{N} = \frac{(29^2 + 34^2 + 51^2 + 42^2)}{4} - \frac{156^2}{16} = 69.5$$

3. Suma de cuadrados del error o variabilidad dentro de métodos de ensamble:

$$SC_E = SC_T - SC_{TRAT} = 99 - 69.5 = 29.5$$

Tabla 3.7 Detalles de los cálculos para el ANOVA en el DCA para el tiempo de ensamble, ejemplo 3.3.

Métodos de ensamble					Operaciones básicas
	A	*B*	*C*	*D*	
Observaciones ⇒	6	7	11	10	$\Sigma_{i=1}^{4}\Sigma_{j=1}^{4} Y_{ij}^2 = 6^2 + 7^2 + \ldots + 9^2 = 1\,620$
	8	9	16	12	= suma de los cuadrados de todas las observaciones o datos
	7	10	11	11	
	8	8	13	9	
Total por tratamiento $(Y_{i\cdot})$ ⇒	29	34	51	42	$Y_{\cdot\cdot} = \Sigma_{i=1}^{4}\Sigma_{j=1}^{4} Y_{ij} = 6 + 7 + \ldots + 9 = 156$ suma de los datos
Número de datos en cada tratamiento (n_i) ⇒	4	4	4	4	$N = \Sigma_{i=1}^{4} n_i = 16$ total de mediciones
Media muestral por tratamiento $(\bar{Y}_{i\cdot})$ ⇒	7.25	8.50	12.75	10.50	$\bar{Y}_{\cdot\cdot} = \frac{Y_{\cdot\cdot}}{N} = \frac{156}{16} = 9.75$ media global
Desviaciones respecto a la media global $(\hat{\tau}_i)$ ⇒	–2.50	–1.25	3.0	0.75	$\hat{\tau}_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$ efecto estimado del método *i*

4. Cuadrados medios de tratamientos y del error (efecto ponderado de cada fuente de variación):

$$CM_{TRAT} = \frac{SC_{TRAT}}{k-1} = \frac{69.5}{3} = 23.17 \text{ y } CM_E = \frac{SC_E}{N-k} = \frac{29.5}{12} = 2.46$$

5. Estadístico de prueba:

$$F_0 = \frac{CM_{TRAT}}{CM_E} = \frac{23.17}{2.46} = 9.42$$

Con toda esta información se procede a llenar la tabla 3.6 de ANOVA. El valor de la significancia observada o valor-*p* es el área bajo la curva de la distribución $F_{3, 12}$ a la derecha de $F_0 = 9.42$, lo cual es difícil de calcular de forma manual. Sin embargo, cuando esto no sea posible, recordemos que otra forma de rechazar o no una hipótesis es comparar el estadístico de prueba contra un número crítico de tablas. En el caso de las tablas de la distribución *F* en el apéndice, se lee que el valor crítico para $\alpha = 0.05$ es $F_{0.05, 3, 12} = 3.49$. Como $F_0 = 9.42 > F_{0.05, 3, 12} = 3.49$, entonces se rechaza H_0, con lo cual se concluye que sí hay diferencia o efecto de los métodos de ensamble en cuanto a su tiempo promedio.

Diagramas de caja
Gráficos basados en los cuartiles de un conjunto de datos.

Diagramas de cajas simultáneos

Los *diagramas de cajas*[3] *simultáneos* representan una manera descriptiva de comparar tratamientos. En la figura 3.3 se presentan los diagramas de cajas simultáneos para los cuatro métodos de ensamble del ejemplo 3.3. Se observa que el método *C* parece diferente a los métodos *A* y *B* en cuanto a sus medias; la media del método *D* también se ve diferente a la media del método *A*. Por otra parte, se observa un poco más de variabilidad en el método *C* que en todos los demás. Lo que sigue es verificar que lo que se observa en el diagrama de caja implica diferencias significativas entre los distintos tratamientos; por lo tanto, es necesario hacer pruebas estadísticas porque los datos que se analizan en los diagramas de cajas son muestras.

En general, cuando los diagramas no se traslapan es probable que los tratamientos correspondientes sean diferentes entre sí, y la probabilidad es mayor en la medida que los diagramas están basados en más datos. Cuando se traslapan un poco puede ser que haya o no diferencias significativas, y en cualquier caso es conveniente utilizar una prueba estadística para determinar cuáles diferencias son significativas. Estas pruebas se verán en la siguiente sección.

Gráficos de medias

Cuando se rechaza H_0 mediante el ANOVA, y se concluye que no hay igualdad entre las medias poblacionales de los tratamientos, pero no se tiene información específica

[3] El diagrama de caja es una herramienta para describir el comportamiento de unos datos, y es de suma utilidad para comparar procesos, tratamientos y, en general, para hacer análisis por estratos (lotes, proveedores, turnos). El diagrama de caja se basa en los cuartiles y parte el rango de variación de los datos en cuatro grupos, cada uno de los cuales contiene 25% de las mediciones. De esta forma se puede visualizar dónde empieza 25% de los datos mayores, dónde 25% de los datos menores y de dónde a dónde se ubica 50% de los datos que están al centro.

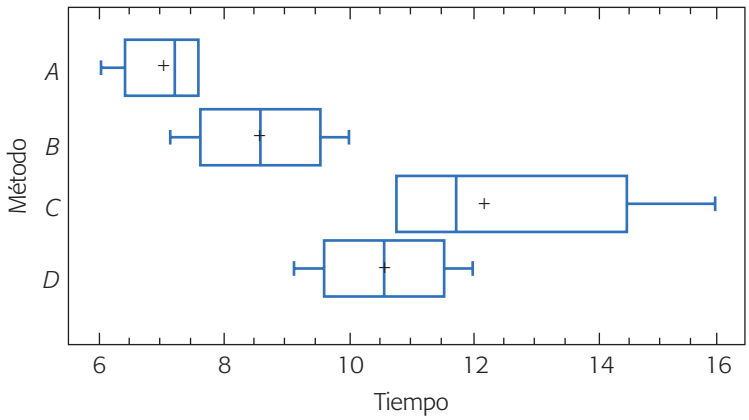

Figura 3.3 Diagramas de cajas para los métodos de ensamble.

sobre cuáles tratamientos son diferentes entre sí, el gráfico de medias (*means plot*) permite hacer una comparación visual y estadística de las medias de los tratamientos (métodos de ensamble). En la figura 3.4 se presenta el gráfico de medias con intervalos de confianza de acuerdo con la prueba LSD, la cual se estudiará más adelante.

Como se explicó en el capítulo anterior, si dos intervalos de confianza se traslapan, los tratamientos correspondientes son estadísticamente iguales en cuanto a sus medias; pero si no se traslapan, entonces son diferentes. Así, podemos ver que el método LSD detecta con una confianza de 95% que $A \neq C$, $A \neq D$ y $B = C$. De esta forma, la conclusión práctica del experimento es que el mejor método de ensamble parece ser el *A*, ya que estadísticamente sus tiempos son menores que los de los métodos *C* y *D*. Le sigue el método *B*, ya que éste es mejor que el *C*. Pero no es posible concluir que el método *A* sea mejor que el método *B*, ya que sus intervalos se traslapan. Si se quisiera decidir en forma estadística sobre la diferencia entre los métodos *A* y *B*, una forma de hacerlo es tomar más datos para incrementar la potencia de la prueba, o bien, recurrir a otros criterios para tomar la decisión.

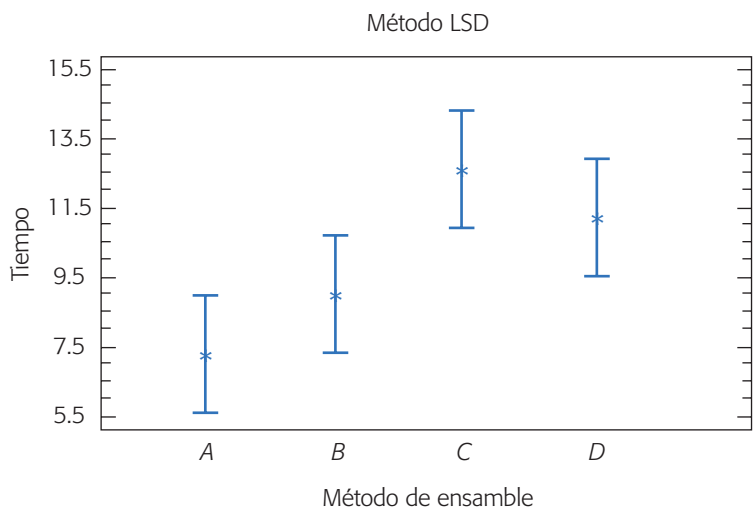

Figura 3.4 Gráfico de medias con el método LSD (ejemplo 3.3).

Comparaciones o pruebas de rango múltiples

Después de que se rechazó la hipótesis nula en un análisis de varianza, es necesario ir a detalle y ver cuáles tratamientos son diferentes. A continuación veremos tres estrategias distintas para ir a ese detalle.

Comparación de parejas de medias de tratamientos

Cuando no se rechaza la hipótesis nula $H_0 : \mu_1 = \mu_2 = \ldots \mu_k = \mu$, el objetivo del experimento está cubierto y la conclusión es que los tratamientos no son diferentes. Si por el contrario se rechaza H_0, y por consiguiente se acepta la hipótesis alternativa $H_A : \mu_i \neq \mu_j$ para algún $i \neq j$, es necesario investigar cuáles tratamientos resultaron diferentes, o cuáles provocan la diferencia. Como se acaba de ilustrar en la gráfica de medias, estas interrogantes se responden probando la igualdad de todos los posibles pares de medias, para lo que se han propuesto varios métodos, conocidos como *métodos de comparaciones múltiples* o *pruebas de rango múltiple*. La diferencia primordial entre los métodos radica en la potencia que tienen para detectar las diferencias entre las medias. Se dice que una prueba es más potente si es capaz de detectar diferencias más pequeñas.

Diferencia mínima significativa (LSD)
Es la diferencia mínima que debe haber entre dos medias muestrales para considerar que dos tratamientos son diferentes.

Método LSD (diferencia mínima significativa)

Una vez que se rechazó H_0 en el ANOVA, el problema es probar la igualdad de todos los posibles pares de medias con la hipótesis:

$$\begin{aligned} H_0 &: \mu_i = \mu_j \\ H_A &: \mu_i \neq \mu_j \end{aligned} \tag{3.10}$$

para toda $i \neq j$. Para k tratamientos se tienen en total $k(k-1)/2$ pares de medias. Por ejemplo, si $k = 4$ existen $4 \times 3/2 = 6$ posibles pares de medias. El estadístico de prueba para cada una de las hipótesis dadas en (3.11) es la correspondiente diferencia en valor absoluto entre sus medias muestrales $\left|\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}\right|$. Se rechaza la hipótesis $H_0 : \mu_i = \mu_j$ si ocurre que

$$\left|\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}\right| > t_{\alpha/2,\,N-k}\sqrt{CM_E\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = LSD \tag{3.11}$$

donde el valor de $t_{\alpha/2,\,N-k}$ se lee en las tablas de la distribución T de Student con $N-k$ grados de libertad que corresponden al error, el CM_E es el cuadrado medio del error y se obtiene de la tabla de ANOVA, n_i y n_j son el número de observaciones para los tratamientos i y j, respectivamente. La cantidad LSD se llama *diferencia mínima*

significativa (*least significant difference*), ya que es la diferencia mínima que debe existir entre dos medias muestrales para considerar que los tratamientos correspondientes son significativamente diferentes. Así, cada diferencia de medias muestrales en valor absoluto que sea mayor que el número LSD se declara significativa. Note que si el diseño es balanceado, es decir, si $n_1 = n_2 = \ldots = n_k = n$, la diferencia mínima significativa se reduce a:

$$LSD = t_{\alpha/2,\ N-k}\sqrt{2CM_E/n} \tag{3.12}$$

En caso de rechazar H_0 se acepta la hipótesis alternativa $H_A : \mu_i \neq \mu_j$, la cual nos dice que las medias de los tratamientos i y j son diferentes. El método LSD tiene una potencia importante, por lo que en ocasiones declara significativas aun pequeñas diferencias.

Ejemplo 3.4

Ilustremos esta prueba continuando con el ejemplo 3.3, en el cual, con el ANOVA se rechazó la hipótesis $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$ y se acepta que al menos un par de medias de tratamientos (métodos de ensamble) son diferentes entre sí. Para investigar cuáles pares de medias son estadísticamente diferentes se prueban los seis posibles pares de hipótesis:

$$\begin{aligned}
&H_0 : \mu_A = \mu_B \quad vs. \quad H_A : \mu_A \neq \mu_B \\
&H_0 : \mu_A = \mu_C \quad vs. \quad H_A : \mu_A \neq \mu_C \\
&H_0 : \mu_A = \mu_D \quad vs. \quad H_A : \mu_A \neq \mu_D \\
&H_0 : \mu_B = \mu_C \quad vs. \quad H_A : \mu_B \neq \mu_C \\
&H_0 : \mu_B = \mu_D \quad vs. \quad H_A : \mu_B \neq \mu_D \\
&H_0 : \mu_C = \mu_D \quad vs. \quad H_A : \mu_C \neq \mu_D
\end{aligned} \tag{3.13}$$

utilizando el método de LSD. En el ANOVA de la tabla 3.6 se observa que los grados de libertad del error son $N - k = 12$, y que el cuadrado medio del error es $CM_E = 2.46$. Si usamos una significancia predefinida de $\alpha = 0.05$, de la tabla de la distribución T de Student con 12 grados de libertad, se obtiene que $t_{0.025,\ 12} = 2.18$. Como en cada tratamiento se hicieron $n = 4$ pruebas, entonces:

$$LSD = t_{\alpha/2,\ N-k}\sqrt{2CM_E/n} = 2.18\sqrt{\frac{2 \times 2.46}{4}} = 2.42$$

La decisión sobre cada una de las seis hipótesis listadas arriba se obtiene al comparar las correspondientes diferencias de medias muestrales en valor absoluto con el número LSD = 2.42. Se declaran significativas aquellas diferencias que son mayores a este número. Los resultados se muestran en la tabla 3.8, de donde se concluye que $\mu_A = \mu_B, \mu_B = \mu_D, \mu_C = \mu_D$, mientras que $\mu_A \neq \mu_C, \mu_B \neq \mu_C$ y $\mu_A \neq \mu_D$. Note que son los mismos resultados que previamente se obtuvieron en la gráfica de medias (figura 3.4), cuyos intervalos están basados en este método LSD. De manera

Tabla 3.8 Aplicación de la prueba *LSD* a métodos de ensamble.

Diferencia poblacional	Diferencia muestral en valor absoluto	Decisión
$\mu_A - \mu_B$	$1.25 < 2.42$	No significativa
$\mu_A - \mu_C$	$* 5.50 > 2.42$	Significativa
$\mu_A - \mu_D$	$* 3.25 > 2.42$	Significativa
$\mu_B - \mu_C$	$* 4.25 > 2.42$	Significativa
$\mu_B - \mu_D$	$2.00 < 2.42$	No significativa
$\mu_C - \mu_D$	$2.25 < 2.42$	No significativa

específica, los intervalos en la gráfica de medias (*means plot*) con el método LSD se obtienen con:

$$\bar{Y}_{i\bullet} \pm t_{\alpha/2,\, N-k}\sqrt{\frac{CM_E}{n_i}}$$

De esta forma, si dos intervalos se traslapan, entonces no habrá diferencias entre las medias de los tratamientos correspondientes. Note que $\sqrt{CM_E/n}$ se está considerando como el error estándar o desviación estándar de la correspondiente media muestral.

Método de Tukey

Un método más conservador para comparar pares de medias de tratamientos es el *método de Tukey*, el cual consiste en comparar las diferencias entre medias muestrales con el valor crítico dado por:

$$T_\alpha = q_\alpha\,(k, N-k)\ \sqrt{CM_E/n_i}$$

donde CM_E es el cuadrado medio del error, n es el número de observaciones por tratamiento, k es el número de tratamientos, $N - k$ es igual a los grados de libertad para el error, α es el nivel de significancia prefijado y el estadístico $q_\alpha(k, N - k)$ son puntos porcentuales de la distribución del rango estudentizado, que se obtienen de la correspondiente tabla en el apéndice. Se declaran significativamente diferentes los pares de medias cuya diferencia muestral en valor absoluto sea mayor que T_α. A diferencia de los métodos LSD y Duncan, el método de Tukey trabaja con un error α muy cercano al declarado por el experimentador.

Ejemplo 3.5

Para aplicar el método de Tukey al ejemplo de los métodos de ensamble, a partir del ANOVA de la tabla 3.6, se toma la información pertinente y de las tablas del rango estudentizado dadas en el apéndice, para $\alpha = 0.05$, se obtiene $q_{0.05}(4, 12) = 4.20$, de manera que el valor crítico es:

$$T_{0.05} = q_{0.05}(4, 12)\sqrt{CM_E/n} = 4.20 \times \sqrt{2.46/4} = 3.27$$

que al compararlo con las diferencias de medias muestrales, los resultados sobre las seis hipótesis son:

Diferencia poblacional	Diferencia muestral	Decisión
$\mu_A - \mu_B$	$1.25 < 3.27$	No significativa
$\mu_A - \mu_C$	$* 5.50 > 3.27$	Significativa
$\mu_A - \mu_D$	$3.25 > 3.27$	No significativa
$\mu_B - \mu_C$	$* 4.25 > 3.27$	Significativa
$\mu_B - \mu_D$	$2.00 < 3.27$	No significativa
$\mu_C - \mu_D$	$2.25 < 3.27$	No significativa

De esta tabla se concluye que $\mu_A = \mu_B = \mu_D$, $\mu_C = \mu_D$, $\mu_A \neq \mu_C$ y $\mu_B \neq \mu_C$. Observe que esta prueba no encuentra diferencia entre los métodos de ensamble *A* y *D*, la cual sí se detectó con el método LSD. Esto es congruente con el hecho de que la prueba de Tukey es menos potente que la prueba LSD, por lo que las pequeñas diferencias no son detectadas como significativas. Asimismo, el riesgo de detectar una diferencia que no existe es menor con el método de Tukey. En la práctica, después de que se ha rechazado H_0 con el ANOVA, conviene aplicar ambos métodos (LSD y Tukey) u otros, cuando haya dudas sobre cuál es el tratamiento ganador. Cuando la diferencia entre dos tratamientos es clara, ambos métodos coinciden.

Método de Duncan

En este método para la comparación de medias, si las k muestras son de igual tamaño, los k promedios se acomodan en orden ascendente y el error estándar de los promedios se estima con $S_{\bar{Y}_{i\cdot}} = \sqrt{CM_E/n}$. Si alguno o todos los tratamientos tienen tamaños diferentes, se reemplaza n por la media armónica de las $\{n_i\}$, que está dada por,

$$n_{AR} = \frac{k}{\sum_{i=1}^{k} \frac{1}{n_i}}$$

Nótese que cuando $n_1 = n_2 = \ldots = n_k = n$, ocurre que $n_{AR} = n$. De la tabla de rangos significantes de Duncan dada en el apéndice, se obtienen los valores críticos $r_\alpha(p, l)$, $p = 2, 3, \ldots, k$, donde α es el nivel de significancia prefijado y l son los grados de libertad para el error. Con estos $k - 1$ valores se obtienen los rangos de significancia mínima dados por

$$R_p = r_\alpha(p, l) S_{\bar{Y}_{i\cdot}};\ p = 2, 3, \ldots, k$$

Las diferencias observadas entre las medias muestrales se comparan con los rangos R_p de la siguiente manera: primero se compara la diferencia entre la media más grande y la más pequeña con el rango R_k. Luego, la diferencia entre la media más grande y la segunda más pequeña se compara con el rango R_{k-1}. Estas comparaciones continúan hasta que la media mayor se haya comparado con todas las demás. Enseguida, se compara la diferencia entre la segunda media más grande y la media

menor con el rango R_{k-1}. Después, la diferencia entre la segunda media más grande y la segunda más pequeña se compara con el valor de R_{k-2}, y así sucesivamente hasta que se comparan los $k(k-1)/2$ pares de medias posibles con el rango que les corresponda. En las comparaciones donde la diferencia observada es mayor que el rango respectivo, se concluye que esas medias son significativamente diferentes. Si dos medias caen entre otras dos que no son muy diferentes, entonces esas dos medias poblacionales también se consideran estadísticamente iguales.

Ejemplo 3.6

De nuevo, supongamos que interesa probar las seis hipótesis dadas en (3.13) para los cuatro métodos de ensamble. En la tabla de ANOVA (tabla 3.6) se lee que $CM_E = 2.46$, lo cual se basa en 12 grados de libertad. Así, el error estándar de cada promedio es $S_{\bar{Y}_{i\bullet}} = \sqrt{CM_E/n} = \sqrt{2.46/4} = 0.78$, dado que se hicieron $n = 4$ observaciones en cada tratamiento. De la tabla de rangos significantes de Duncan dada en el apéndice, para $\alpha = 0.05$ y 12 grados de libertad, se leen los rangos $r_{0.05}(2, 12) = 3.08$, $r_{0.05}(3, 12) = 3.23$ y $r_{0.05}(4, 12) = 3.33$. Con esta información, los rangos mínimos significantes son:

$$R_2 = r_{0.05}(2, 12)S_{\bar{Y}_{i\bullet}} = (3.08)(0.78) = 2.40$$
$$R_3 = r_{0.05}(3, 12)S_{\bar{Y}_{i\bullet}} = (3.23)(0.78) = 2.52$$
$$R_4 = r_{0.05}(4, 12)S_{\bar{Y}_{i\bullet}} = (3.33)(0.78) = 2.60$$

Estos rangos se comparan con las diferencias de medias de acuerdo al método descrito arriba.

Las cuatro medias muestrales acomodadas en orden ascendente son: $\bar{Y}_A = 7.25$, $\bar{Y}_B = 8.50$, $\bar{Y}_D = 10.50$ y $\bar{Y}_C = 12.75$. De aquí se obtienen las diferencias en el orden dado por el método de Duncan y se van comparando con el rango correspondiente. En la siguiente tabla se resumen los resultados obtenidos.

Diferencia poblacional	Diferencia muestral comparada con su rango R_p	Decisión
$\mu_C - \mu_A$	$12.75 - 7.25 = 5.5^* > 2.60 = R_4$	Significativa
$\mu_C - \mu_B$	$12.75 - 8.50 = 3.27^* > 2.52 = R_3$	Significativa
$\mu_C - \mu_D$	$12.75 - 10.50 = 2.25 < 2.40 = R_2$	No significativa
$\mu_D - \mu_A$	$10.50 - 7.25 = 3.25^* > 2.60 = R_3$	Significativa
$\mu_D - \mu_B$	$10.50 - 8.50 = 2.0 < 2.40 = R_2$	No significativa
$\mu_B - \mu_A$	$8.50 - 7.25 = 1.25 < 2.40 = R_2$	No significativa

De esta tabla se concluye que $\mu_A = \mu_B$, $\mu_B = \mu_D$ y $\mu_C = \mu_D$, mientras que $\mu_A \neq \mu_C$, $\mu_B \neq \mu_C$ y $\mu_A \neq \mu_D$, que son las mismas conclusiones que se obtuvieron con el método LSD. En general, las pruebas de Duncan y LSD tienen un desempeño similar.

Tratamiento control
Se refiere a un tratamiento estándar de referencia o a la ausencia de tratamiento.

Comparación de tratamientos con un control (método de Dunnet)

Una vez que se rechaza H_0 con el ANOVA, en ocasiones uno de los k tratamientos a comparar es el llamado *tratamiento control* y el interés fundamental es comparar los

$k-1$ tratamientos restantes con dicho control. En muchos casos el tratamiento control se refiere a un tratamiento estándar de referencia o también a la ausencia de tratamiento (véase ejercicio 3.12). Por ejemplo, al comparar varios medicamentos para el resfriado es conveniente que uno de los tratamientos sea que los pacientes no utilicen ningún medicamento; esto sirve como referencia para decidir la posible utilidad de los medicamentos.

Por facilidad, denotemos como tratamiento control al k-ésimo tratamiento. Hacer comparaciones con respecto al control implica probar las $k-1$ hipótesis dadas por:

$$H_0 : \mu_i = \mu_k$$
$$H_A : \mu_i \neq \mu_k$$

con $i = 1, 2, \ldots, k-1$, donde k es el tratamiento control. La hipótesis nula se rechaza si,

$$\left|\bar{Y}_{i\bullet} - \bar{Y}_{k\bullet}\right| > D_\alpha(k-1, l)\sqrt{CM_E\left(\frac{1}{n_i}+\frac{1}{n_k}\right)}$$

donde $D_\alpha(k-1, l)$ se encuentra en las tablas del apéndice; l son los grados de libertad del cuadrado medio del error. Se recomienda que el tamaño de muestra del tratamiento control sea grande, a fin de estimar su media con mayor precisión.

Comparación por contrastes

No siempre interesa probar sólo las $k(k-1)/2$ hipótesis dos a dos dadas por $H_0 : \mu_i = \mu_j$ *vs.* $H_0 : \mu_i \neq \mu_j$ para $i \neq j$, y no siempre estas hipótesis dos a dos interesan todas por igual. En ocasiones, el objetivo del estudio lleva a contrastar hipótesis que involucran a más de dos medias. En esta sección se presentan este tipo de alternativas en la comparación de medias, pero antes se definen los conceptos de *contraste* y *contrastes ortogonales*.

Contraste

Una expresión de la forma $C = \Sigma_{i=1}^{k} c_i \mu_i$ es una combinación lineal de las medias poblacionales de interés, donde los coeficientes c_i son números reales. La combinación lineal C se llama *contraste* si cumple que la suma de los coeficientes es igual a cero ($\Sigma_{i=1}^{k} c_i = 0$). Muchas hipótesis estadísticas de interés son contrastes, como por ejemplo las hipótesis de comparación de medias. En efecto, ya hemos visto que la hipótesis nula $H_0 : \mu_i = \mu_j$ para $i \neq j$ se puede escribir de manera equivalente como $H_0 : \mu_i - \mu_j = 0$, donde se observa que el contraste correspondiente es la combinación lineal $c_i\mu_i + c_j\mu_j$ con $c_i = 1$ y $c_j = -1$, e interesa verificar si es estadísticamente igual a cero.

Contraste
Combinación lineal de medias poblacionales donde la suma de los coeficientes es igual a cero.

En general, supongamos que interesa probar si el contraste definido por $C = \Sigma_{i=1}^{k} c_i \mu_i$ es igual a cero. Si las poblaciones objeto de estudio son normales ($N(\mu_i, \sigma_i^2); i = 1, 2, \ldots, k$) el contraste C sigue una distribución normal con media $\mu_C = \Sigma_{i=1}^{k} c_i \mu_i$ y varianza $V_C = \Sigma_{i=1}^{k} \frac{c_i^2}{n_i}\sigma_i^2$. Cuando las varianzas de los tratamientos

son iguales y el diseño experimental es balanceado ($n_i = n$ para cada i), la varianza del contraste se reduce a $V_C = \frac{\sigma^2}{n}\Sigma_{i=1}^{k} c_i^2$. Al usar el CM_E para estimar a σ^2 y $\bar{Y}_{i\bullet}$ para estimar a la media μ_i, se puede ver que un intervalo al $100(1-\alpha)\%$ de confianza para el contraste C está dado por:

$$\sum_{i=1}^{k} c_i \bar{Y}_{i\bullet} \pm t_{\alpha/2,\, N-k} \sqrt{\frac{CM_E}{n} \sum_{i=1}^{k} c_i^2}$$

donde $t_{\alpha/2,\, N-k}$ es un punto porcentual de la distribución T de Student con $N-k$ grados de libertad. En caso de que el intervalo contenga al cero se concluye que el contraste C es estadísticamente igual a cero.

Contrastes ortogonales
Cuando la suma del producto de los coeficientes de dos contrastes es igual a cero.

Contrastes ortogonales

En el caso de un diseño balanceado, dos contrastes $C_1 = \Sigma_{i=1}^{k} c_{1i}\mu_i$ y $C_2 = \Sigma_{i=1}^{k} c_{2i}\mu_i$ son *ortogonales* si la suma del producto de los coeficientes es igual a cero, esto es, si $\Sigma_{i=1} c_{1i}c_{2i} = 0$; para el diseño desbalanceado son ortogonales si $\Sigma_{i=1} n_i c_{1i}c_{2i} = 0$. Dadas las k medias de interés correspondientes a k tratamientos objeto de estudio, se pueden construir una infinidad de conjuntos de $k-1$ contrastes ortogonales entre sí. En particular, con el uso de contrastes ortogonales es posible construir un grupo de hipótesis de interés independientes entre sí. Por ejemplo, en el problema de los $k = 4$ métodos de ensamble se pueden construir grupos de contrastes ortogonales de tamaño tres. Una posibilidad de elección se muestra en la siguiente tabla:

c_1	c_2	c_3	c_4	Contrastes ortogonales
2	–1	–1	0	$2\mu_A - \mu_B - \mu_C$
0	1	–1	0	$\mu_B - \mu_C$
1	1	1	–3	$\mu_A + \mu_B + \mu_C - 3\mu_D$

Es fácil ver que los tres contrastes definidos en esta tabla son ortogonales entre sí. Por ejemplo, el primero y el segundo son ortogonales porque $(2 \times 0) + (-1 \times 1) + (-1 \times -1) + (0 \times 0) = 0$, y lo mismo pasa con los otros dos posibles productos. Observe también que con cada contraste se puede definir una hipótesis estadística, como se hace en el siguiente método de Sheffé.

Método de Sheffé
Sirve para probar todos los contrastes de medias que pudieran ser de interés, en particular aquellos que involucran a más de dos medias.

Método de Sheffé

Este método está diseñado para probar todos los contrastes de medias que pudieran interesar al experimentador, sin el inconveniente de inflar por ello el error tipo I (detección de diferencias que no existen). Supongamos que interesa contrastar las hipótesis

$$H_0 : 2\mu_A = \mu_B + \mu_C$$
$$H_A : 2\mu_A \neq \mu_B + \mu_C \qquad \textbf{(3.14)}$$

donde la hipótesis nula se puede escribir alternativamente como $H_0 : 2\mu_A - \mu_B - \mu_C = 0$, lo cual implica que la hipótesis está definida por el contraste $C_0 = 2\mu_A - \mu_B - \mu_C$. De manera que el contraste estimado está dado por

$$\hat{C}_0 = 2\bar{Y}_A - \bar{Y}_B - \bar{Y}_C$$

y su varianza estimada es

$$V(\hat{C}_0) = CM_E \sum \frac{c_i^2}{n_i}$$

donde n_i es el número de mediciones en el tratamiento $i = A, B, C$. Intervalos simultáneos al $100(1-\alpha)\%$ de confianza para todos los contrastes tienen la forma

$$\hat{C} \pm \sqrt{(k-1)V(\hat{C})F_{\alpha,\, k-1,\, N-k}}$$

donde $\hat{C}$ representa la estimación de cualquier posible contraste y $F_{\alpha,\, k-1,\, N-k}$ es el cuantil $100(1-\alpha)$ de una distribución F con $k-1$ grados de libertad en el numerador, y $N-k$ grados de libertad en el denominador. Si el intervalo resultante para un contraste particular, digamos C_0, no contiene al cero, se concluye que el contraste es significativamente diferente de cero, lo cual lleva a rechazar H_0. De manera equivalente, el método de Sheffé rechaza la hipótesis nula si el contraste asociado es

$$\left|\hat{C}_0\right| > \sqrt{(k-1)V(\hat{C})F_{\alpha,\, k-1,\, N-k}}$$

Supongamos que en el ejemplo de los métodos de ensamble se quieren contrastar las hipótesis dadas en la ecuación (3.14). Con las medias muestrales (tabla 3.7) se calcula el estadístico $\hat{C}_0 = 2(7.25) - 8.50 - 12.75 = -6.75$. La varianza del contraste es $V(\hat{C}_0) = 2.46(6)/4 = 3.69$. Como $\sqrt{(k-1)V(\hat{C})F_{\alpha,\, k-1,\, N-k}} = \sqrt{3 \times 3.69 \times 3.49} = 6.21$ y $\left|\hat{C}_0\right| = 6.75$, se rechaza la hipótesis $H_0 : 2\mu_A = \mu_B + \mu_C$ y se acepta la $H_A : 2\mu_A \neq \mu_B + \mu_C$.

Verificación de los supuestos del modelo

La validez de los resultados obtenidos en cualquier análisis de varianza queda supeditado a que los supuestos del modelo se cumplan. Estos supuestos son: *normalidad*, *varianza constante* (*igual varianza de los tratamientos*) e *independencia*. Esto es, la respuesta (Y) se debe distribuir de manera normal, con la misma varianza en cada tratamiento y las mediciones deben ser independientes. Estos supuestos sobre Y se traducen en supuestos sobre el término error (ε) en el modelo [véase expresión (3.2)]. Es una práctica común utilizar la muestra de *residuos* para comprobar los supuestos del modelo, ya que si los supuestos se cumplen, los residuos o residuales se pueden ver como una muestra aleatoria de una distribución normal con media cero y varianza constante. Los residuos, e_{ij}, se definen como la diferencia entre la respuesta observada (Y_{ij}) y la respuesta predicha por el modelo ($\hat{Y}_{ij}$), lo cual permite hacer un diagnóstico más directo de la calidad del modelo, ya que su magnitud señala qué tan bien describe a los datos el modelo. Veamos.

Residuos
Son generados por la diferencia entre la respuesta observada y la respuesta predicha por el modelo en cada prueba experimental.

Recordemos de (3.2), que el modelo que se espera describa los datos en el DCA está dado por:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \tag{3.15}$$

donde Y_{ij} ($i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, n$) es el j-ésimo dato en el tratamiento i; μ es la media global, τ_i es el efecto del tratamiento i y ε_{ij} representa al error asociado con la observación Y_{ij}. Cuando se realiza el ANOVA, y sólo cuando éste resulta significativo, entonces se procede a estimar el modelo ajustado o modelo de trabajo dado por:

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\tau}_i \tag{3.16}$$

donde $\hat{Y}_{ij}$ es la respuesta predicha, $\hat{\mu}$ es la media global estimada y $\hat{\tau}_i$ es el efecto estimado del tratamiento i; los gorros indican que son *estimadores*, es decir, valores calculados a partir de los datos del experimento. El término del error desaparece del modelo estimado, por el hecho de que su valor esperado es igual a cero ($E(\varepsilon_{ij}) = 0$). Como la media global se estima con $\bar{Y}_{\bullet\bullet}$ y el efecto del tratamiento con $\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}$, el modelo ajustado del DCA se puede escribir como:

$$\hat{Y}_{ij} = \bar{Y}_{\bullet\bullet} + (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) = \bar{Y}_{i\bullet} \tag{3.17}$$

Esto es, la respuesta predicha para cada observación es la media muestral del tratamiento correspondiente. De esta manera, el residual o *residuo asociado a la observación* Y_{ij} está dado por

$$e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{i\bullet}$$

Los supuestos del modelo lineal (3.15), en términos de los residuos, son:

1. *Los* e_{ij} *siguen una distribución normal con media cero.*
2. *Los* e_{ij} *son independientes entre sí.*
3. *Los residuos de cada tratamiento tienen la misma varianza* σ^2.

Para comprobar cada supuesto existen pruebas analíticas y gráficas que veremos a continuación. Por sencillez, muchas veces se prefieren las pruebas gráficas. Éstas tienen el inconveniente de que no son "exactas", pero aun así, en la mayoría de las situaciones prácticas proporcionan la evidencia suficiente en contra o a favor de los supuestos. El uso de las pruebas gráficas requiere una fuerte evidencia visual para concluir que el supuesto en cuestión no se cumple, ya que se requiere que la evidencia en contra de un supuesto esté soportada por más de dos puntos. Cuando son uno o dos los puntos que se salen del comportamiento esperado de las gráficas se puede tratar de un problema de puntos aberrantes, no de violación del supuesto en cuestión. En ese caso debe investigarse la obtención de dichas mediciones atípicas, ya que ese tipo de puntos pueden afectar sensiblemente los resultados del análisis.

Se puede utilizar una prueba analítica para subsanar las ambigüedades que surjan en la interpretación visual (subjetiva) de las gráficas.

Es mejor prevenir en lo posible que los supuestos no se violen, para ello se aplican los tres principios básicos del diseño de experimentos: repetición, aleatorización y bloqueo. Es fácil encontrar situaciones en las que por no aplicar alguno de estos principios no se cumplen los supuestos del modelo. Por ejemplo, por no aleatorizar el orden en que se corren las pruebas pueden surgir problemas con el supuesto de independencia.

Normalidad

Gráfica de probabilidad
Sirve para verificar visualmente si los datos siguen una distribución de probabilidad específica.

Un procedimiento gráfico para verificar el cumplimiento del supuesto de normalidad de los residuos consiste en graficar los residuos en *papel* o en la *gráfica de probabilidad normal* que se incluye casi en todos los paquetes estadísticos. Esta gráfica del tipo *X-Y* tiene las escalas de tal manera que si los residuos siguen una distribución normal, al graficarlos tienden a quedar alineados en una línea recta; por lo tanto, si claramente no se alinean se concluye que el supuesto de normalidad no es correcto. Cabe enfatizar el hecho de que el ajuste de los puntos a una recta no tiene que ser perfecto, dado que el análisis de varianza resiste pequeñas y moderadas desviaciones al supuesto de normalidad. En las figuras 3.6*a* y 3.6*b* se representan, en la gráfica de probabilidad normal, dos aspectos de los residuos, en los cuales el supuesto de normalidad no se cumple.

Gráfica de probabilidad en papel normal

Consideremos los N residuos e_i que resultan del análisis de una varianza, o cualquier conjunto de N datos de los cuales se quiere verificar su procedencia de una distribución normal. Los pasos en la construcción de la gráfica de probabilidad normal para los residuos son los siguientes:

1. Ordenar los N valores del menor al mayor y asignarles los rangos de 1 a N. Sean r_i, $i = 1, 2,\ldots, N$, los datos en orden creciente.
2. Calcular una posición de graficación para cada dato en función de su rango y del total de observaciones como $(i - 0.5)/N$, $i = 1, 2,\ldots, N$.
3. El papel de probabilidad normal es un formato para realizar una gráfica del tipo *X-Y*, donde una de las escalas es lineal y la otra es logarítmica. Sobre el papel de probabilidad normal se dibujan las parejas $(r_i, (i - 0.5)/N)$.
4. Dibujar una línea recta sobre los puntos para tratar de dilucidar si se ajustan a ella o no. La interpretación de la gráfica es subjetiva, pero muchas veces es suficiente para llegar a una conclusión razonable sobre la distribución que siguen los datos.

Para ilustrar lo anterior, supongamos que los residuos son los siguientes 10 datos: 48.8, 51.5, 50.6, 46.5, 41.7, 39.9, 50.4, 43.9, 48.6, 48.6. Los cálculos necesarios para obtener las parejas a graficar se muestran en la tabla 3.9.

En el papel de probabilidad normal se grafican las parejas dadas por la primera y tercera columnas $(r_i, (i - 0.5)/N)$, y la gráfica resultante se muestra en la figura 3.5*a*. En ésta no hay evidencia suficiente en contra de la normalidad de los datos.

Tabla 3.9 Cálculos para realizar una gráfica de probabilidad normal.

Dato r_i	Rango i	$(i - 0.5)/N$	$Z_i = \Phi^{-1}((i - 0.5)/N)$
39.9	1	0.05	–1.64
41.7	2	0.15	–1.03
43.9	3	0.25	–0.67
46.5	4	0.35	–0.38
48.6	5	0.5	0.00
48.6	6	0.5	0.00
48.8	7	0.65	0.38
50.4	8	0.75	0.67
50.6	9	0.85	1.03
51.5	10	0.95	1.64

Gráfica de probabilidad normal en papel ordinario

A falta de papel de probabilidad normal, la gráfica de probabilidad también se puede hacer en papel ordinario con escalas equiespaciadas en ambos ejes. Para ello, primero se obtiene el valor normal estandarizado Z_i que cumple la relación:

$$\frac{(i-0.5)}{N} = P(Z < Z_i) = \Phi(Z_i) \qquad \textbf{(3.18)}$$

donde $\Phi(Z_i)$ es la función de distribución normal estándar acumulada evaluada en Z_i. Es decir, $Z_i = \Phi^{-1}(\frac{i-0.5}{N})$. Las parejas a dibujar en el papel ordinario son (r_i, Z_i) (ver tabla 3.9). En la figura 3.5*b* se muestra la gráfica de probabilidad en papel ordinario para los mismos datos graficados en papel normal. Observe que es básicamente la misma gráfica. Los cálculos necesarios para los Z_i se pueden hacer fácilmente en Excel con la función: DISTR.NORM.ESTAND.INV y en *Statgraphics* con la función INVNORMAL.

Además de la evaluación visual basada en la gráfica de probabilidad normal, existen varios métodos analíticos para contrastar la hipótesis H_0 : *Hay normalidad* contra H_A : *No hay normalidad*. Entre dichas pruebas se encuentran la ji-cuadrada para bondad de ajuste, la prueba de Shapiro-Wilks y la prueba de Anderson-Darling,

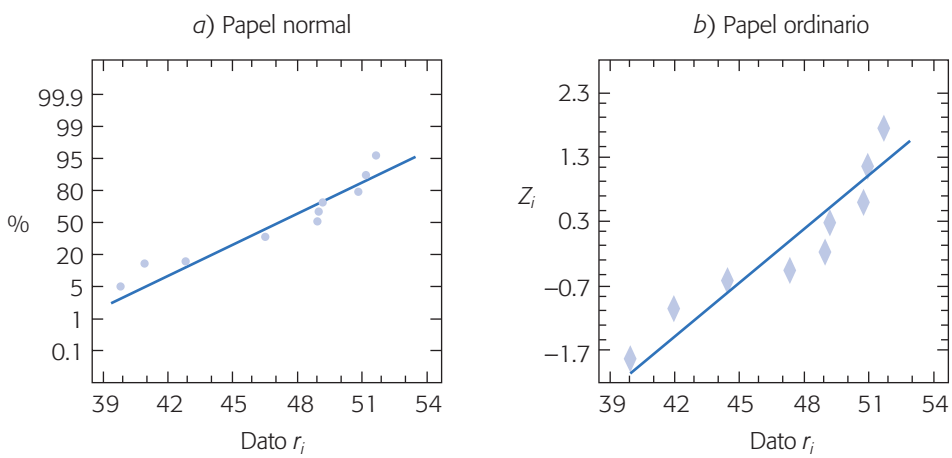

Figura 3.5 Gráfica de probabilidad en papel normal y en papel ordinario.

de las cuales, la de Shapiro-Wilks es una de las más recomendadas y que presentamos a continuación.

Prueba de Shapiro-Wilks para normalidad

Consideremos una muestra aleatoria de datos $x_1, x_2, \ldots, x_n$ que proceden de cierta distribución desconocida denotada por $F(x)$. Se quiere verificar si dichos datos fueron generados por un proceso normal, mediante las hipótesis estadísticas:

H_0 : *Los datos proceden de una distribución normal* ($F(x)$ *es normal*).
H_A : *Los datos no proceden de una distribución normal* ($F(x)$ *no es normal*).

Los pasos para la prueba de Shapiro-Wilks son: 1) Se ordenan los datos de menor a mayor. Denotemos los datos ordenados por $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. 2) De la tabla dada en el apéndice para este procedimiento se obtienen los coeficientes $a_1, a_2, \ldots, a_k$, donde k es aproximadamente $n/2$. 3) Se calcula el estadístico W definido como:

$$W = \frac{1}{(n-1)S^2}\left[\sum_{i=1}^{k} a_i (X_{(n-i+1)} - X_{(i)})\right]^2 \qquad \textbf{(3.19)}$$

donde S^2 es la varianza muestral. 4) Por último, si el valor del estadístico es mayor que su valor crítico al nivel α seleccionado en la tabla del apéndice, se rechaza la normalidad de los datos.

Para ilustrar la prueba de Shapiro-Wilks consideremos otra vez los mismos datos de las gráficas de probabilidad normal. De acuerdo con los datos ordenados, parte del procedimiento posterior al paso 2 para calcular el estadístico W se resume en la tabla que se presenta más adelante.

La varianza es $S^2 = 15.72$. Con la fórmula de la ecuación (3.19) se obtiene que

$$W = \frac{1}{(10-1)15.72}[11.26]^2 = 0.896$$

i	a_i	$(X_{(n-i+1)} - X_{(i)})$	$a_i(X_{(n-i+1)} - X_{(i)})$
1	0.5739	51.5 – 39.9 = 11.6	6.66
2	0.3291	50.6 – 41.7 = 8.9	2.93
3	0.2141	50.4 – 43.9 = 6.5	1.39
4	0.1224	48.8 – 46.5 = 2.3	0.28
5	0.0399	48.6 – 48.6 = 0	0.00

Con el tamaño de muestra $n = 10$, en la tabla de valores críticos dada en el apéndice se lee que el cuantil 95 es $W_{1-0.05} = 0.987$. Como W es menor que $W_{1-\alpha}$ se acepta que los datos proceden de una distribución normal, que concuerda con lo que se observó en las gráficas de probabilidad de la figura 3.5.

Varianza constante

Varianza constante
Supuesto del ANOVA que se cumple cuando los tratamientos tienen la misma varianza.

Una forma de verificar el supuesto de *varianza constante* (o que los tratamientos tienen la misma varianza) es graficando los predichos contra los residuos ($\hat{Y}_{ij}$ *vs.* e_i), por lo general $\hat{Y}_{ij}$ va en el eje horizontal y los residuos en el eje vertical. Si los puntos

en esta gráfica se distribuyen de manera aleatoria en una banda horizontal (sin ningún patrón claro y contundente), entonces es señal de que se cumple el supuesto de que los tratamientos tienen igual varianza. Por el contrario, si se distribuyen con algún patrón claro y contundente, como por ejemplo una forma de "corneta o embudo", entonces es señal de que no se está cumpliendo el supuesto de varianza constante (figura 3.6*c*). Un claro embudo en los residuales indicará que el error de pronóstico del modelo tiene una relación directa (positiva o negativa) con la magnitud del pronóstico (predicho).

Otra gráfica que ayuda a verificar el supuesto es la gráfica de niveles de factor contra residuos. En el eje *X* de esta gráfica se ponen los tratamientos o los niveles de un factor, y en el eje vertical se agregan los residuos correspondientes a cada tratamiento o nivel de factor. Si se cumple el supuesto de varianza constante, se espera que la amplitud de la dispersión de los puntos en cada nivel de factor tenderá a ser similar; y no se cumplirá el supuesto si hay diferencias fuertes en esta amplitud,

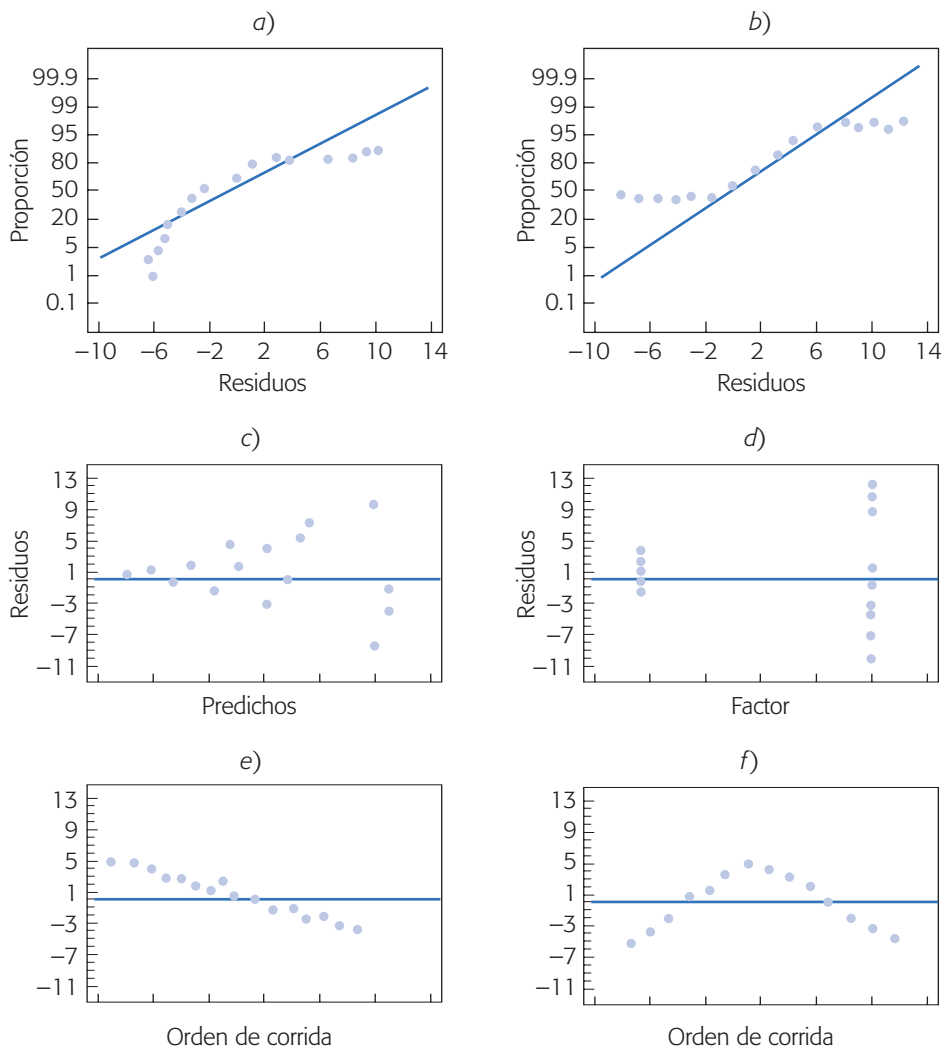

Figura 3.6 Ejemplos de gráficas de residuos donde no se cumplen los supuestos para el ANOVA.

como se muestra en la figura 3.6*d*. En la interpretación de esta gráfica debe considerarse que, en estadística, las pequeñas diferencias por lo general no son significativas, y también debe tomarse en cuenta la cantidad de observaciones hechas en cada nivel del factor, puesto que este hecho puede impactar la dispersión aparente en cada tratamiento.

Otra interpretación de la gráfica de factor contra residuos es que cuando los tratamientos o niveles muestran una dispersión diferente de sus residuales correspondientes (como en la figura 3.6*d*), es que el factor o los tratamientos tienen un efecto significativo sobre la variabilidad de la respuesta.

Con base en esta información se podría proponer un nivel de operación para dicho factor que minimice la dispersión y optimice la media.

Así, cuando hay una evidencia contundente en las gráficas anteriores, donde no se cumple el supuesto de varianza constante, entonces se debe ver en qué sentido resultan afectadas las conclusiones que se obtienen con el ANOVA y las pruebas de rangos múltiples. Por ejemplo, si se aprecia que el mejor tratamiento también es el que tiene menor dispersión, entonces se debe mantener tal tratamiento como la elección correcta, y ver si es de interés investigar por qué la diferencia en variabilidad con algunos de los otros tratamientos. Pero, si al que se le considera el mejor tratamiento es el que tiene la varianza más grande, entonces es difícil mantenerlo como la elección correcta. En este caso se debe replantear la decisión y el análisis. Una forma de volver a hacer el análisis y reconsiderar la situación es transformar los datos u observaciones Y_{ij}, de manera que se disminuyan las diferencias en dispersión y se pueda ver más claramente lo que ha pasado en el experimento. Existe una gran cantidad de transformaciones propuestas que logran lo anterior, entre las más frecuentes se encuentran la logarítmica y la raíz cuadrada. La transformación se hace de la siguiente manera: se saca logaritmo a los datos u observaciones por ejemplo, y con los datos transformados se vuelve a hacer el análisis completo. En la sección "Transformaciones para estabilizar varianzas" del capítulo 5 aborda el tema con detalle.

En general, siempre se debe investigar por qué no se ha cumplido el supuesto de varianza constante, ya que eso ayuda a entender mejor el proceso o sistema con el que se experimenta. Por ejemplo, una razón frecuente que hace que tal supuesto no se cumpla es que algunas variables tienen una dispersión directamente proporcional a su magnitud, de tal forma que si sus valores son pequeños, éstos tienden a ser más homogéneos en comparación con la variabilidad que entre sí tienen los valores grandes. Ahora veamos una prueba analítica para la igualdad de varianzas.

Prueba de Bartlett para homogeneidad de varianzas

Supongamos que se tienen k poblaciones o tratamientos independientes, cada uno con distribución normal ($N(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, k$), donde las varianzas son desconocidas. Se quiere probar la hipótesis de igualdad de varianzas dada por:

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2 \\ H_A &: \sigma_i^2 \neq \sigma_j^2 \text{ para algún } i \neq j \end{aligned} \qquad \textbf{(3.20)}$$

Mediante un diseño completamente al azar se obtienen k muestras aleatorias de tamaños n_i ($i = 1, 2, \ldots, k$) de dichas poblaciones, de modo que el total de mediciones

es $N = n_1 + n_2 + \ldots + n_k$. El estadístico de prueba para la hipótesis (3.20) está dado por

$$\chi_0^2 = 2.3026\frac{q}{c}$$

donde

$$q = (N-k)\log_{10} S_p^2 - \sum_{i=1}^{k}(n_i - 1)\log_{10} S_i^2$$

y

$$c = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}(n_i - 1)^{-1} - (N-k)^{-1}\right)$$

con

$$S_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)S_i^2}{N-k}$$

donde S_i^2 es la varianza muestral del tratamiento i. Bajo la hipótesis nula de igualdad de varianza, el estadístico χ_0^2 sigue una distribución ji-cuadrada con $k - 1$ grados de libertad, por lo que se rechaza H_0 cuando χ_0^2 es más grande que $\chi^2_{(\alpha,\, k-1)}$. Observe que el estadístico q, en el numerador del estadístico χ_0^2, es grande en la medida de que las varianzas muestrales S_i^2 son diferentes y es igual a cero cuando éstas son iguales.

La prueba de Bartlett que acabamos de describir es sensible a la falta de normalidad de las poblaciones de interés, por lo que debe comprobarse el cumplimiento de este supuesto.

Independencia

La suposición de independencia en los residuos puede verificarse si se grafica el orden en que se colectó un dato contra el residuo correspondiente. De esta manera, si al graficar en el eje horizontal el tiempo (orden de corrida) y en el eje vertical los residuos, se detecta una tendencia o patrón no aleatorio claramente definido, esto es evidencia de que existe una correlación entre los errores y, por lo tanto, el supuesto de independencia no se cumple (véanse figuras 3.6*e* y 3.6*f*). Si el comportamiento de los puntos es aleatorio dentro de una banda horizontal, el supuesto se está cumpliendo. La violación de este supuesto generalmente indica deficiencias en la planeación y ejecución del experimento; asimismo, puede ser un indicador de que no se aplicó en forma correcta el principio de aleatorización, o de que conforme se fueron realizando las pruebas experimentales aparecieron factores que afectaron la respuesta observada. Por ello, en caso de tener problemas con este supuesto, las conclusiones que se obtienen del análisis son endebles y por ello es mejor revisar lo hecho y tratar de investigar por qué no se cumplió con ese supuesto de independencia, a fin de reconsiderar la situación.

Una prueba analítica para verificar la independencia entre residuos consecutivos es la prueba de Durbin-Watson, que se presenta en el capítulo 11. El problema con dicha prueba es que no es capaz de detectar otros patrones de correlación entre residuos (no consecutivos) que también son violatorios del supuesto de independencia.

Tabla 3.10 Residuos para ejemplo 3.2.

Cuero	Observado Y_{ij}	Predicho $\bar{Y}_{i\cdot}$	Residuo $e_{ij} = Y_{ij} - \bar{Y}_{i\cdot}$	Cuero	Observado Y_{ij}	Predicho $\bar{Y}_{i\cdot}$	Residuo $e_{ij} = Y_{ij} - \bar{Y}_{i\cdot}$
A	264	256.7	7.33	*A*	262	256.7	5.33
C	220	230.8	–10.83	*D*	220	220.7	–0.67
B	208	209.8	–2.5	*A*	255	256.7	–1.67
B	220	209.8	9.5	*B*	200	209.8	–10.5
A	260	256.7	3.33	*D*	222	220.7	1.33
A	258	256.7	1.33	*B*	213	209.8	2.5
D	217	220.7	–3.67	*A*	241	256.7	–15.67
C	263	230.8	32.17	*C*	228	230.8	–2.83
D	229	220.7	5.83	*B*	206	209.8	–4.5
C	219	230.8	–11.83	*C*	230	230.8	–0.83
B	216	209.8	5.5	*D*	215	220.7	–5.67
C	225	230.8	–5.83	*D*	224	220.7	3.33

Ejemplo 3.6

(Continuación del análisis para comparar cuatro tipos de cuero). En el ejemplo 3.2 se compararon cuatro tipos de cuero en cuanto a su desgaste, y mediante el ANOVA se concluyó que los cueros tienen un desgaste promedio diferente (ver tabla 3.5). Falta ver que se cumplan los supuestos del ANOVA. Para ello, primero se calculan los residuos de las 24 mediciones, restando a cada valor observado su correspondiente predicho, que en este caso como $\hat{Y}_{ij} = \bar{Y}_{i\cdot}$ se debe restar la media del tratamiento correspondiente. Los 24 residuos se listan en la tabla 3.10.

Con la muestra de 24 residuos se procede a dibujar las gráficas de residuos en papel de probabilidad normal, residuos contra predichos y residuos contra orden de corrida. Las gráficas resultantes se muestran en las figuras 3.7*a*, *b* y *c*. Se observa el cumplimiento de los supuestos de normalidad, varianza constante e independencia, respectivamente. Sin embargo, en las tres gráficas es notorio un punto que se aleja bastante del resto, el cual es un punto aberrante cuyo origen debe investigarse. En la tabla 3.10 se encuentra que este residuo grande de valor 32.17 y que corresponde a la prueba 8 con una medición de 263 en el tipo de cuero *C*. Debe verificarse que no haya ningún error con este dato. Cuando un punto aberrante no se percibe, puede afectar sensiblemente las conclusiones del análisis del experimento.

Elección del tamaño de la muestra

Una decisión importante en cualquier diseño de experimentos es decidir el número de réplicas que se hará por cada tratamiento (tamaño de muestra). Por lo general, si se esperan diferencias pequeñas entre tratamientos será necesario un mayor tamaño de muestra. Aunque existen varios métodos para estimar el tamaño muestral, muchas veces tienen poca aplicabilidad porque requieren cierto conocimiento previo sobre la varianza del error experimental.

Si recurrimos a la experiencia vemos que el número de réplicas en la mayoría de las situaciones experimentales en las que se involucra un factor varía entre cinco

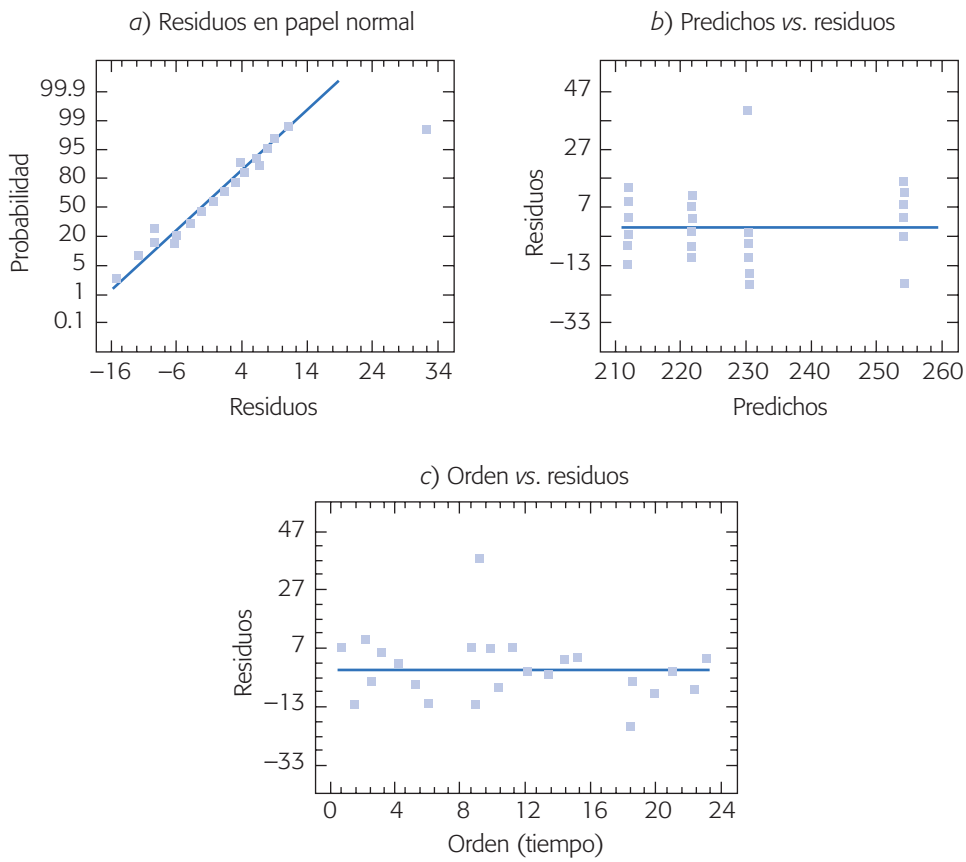

Figura 3.7 Gráficas de residuos para los tipos de cuero.

y diez; incluso, en algunos casos puede llegar hasta 30. La tendencia podría inclinarse por un extremo de este rango e incluso salirse de éste, de acuerdo con las siguientes consideraciones:

- A menor diferencia que se espera en los tratamientos, mayor será la cantidad de réplicas si se quieren detectar diferencias significativas, y viceversa, es decir, si se esperan grandes diferencias quizá con pocas réplicas sea suficiente.
- Si se espera mucha variación dentro de cada tratamiento, debido a la variación de fuentes no controladas como métodos de medición, medio ambiente, materia prima, etc., entonces se necesitarán más réplicas.
- Si son varios tratamientos (cuatro o más), entonces éste es un punto favorable para reducir el número de réplicas.

Además de lo anterior, es preciso considerar los costos y el tiempo global del experimento. De aquí que si se toman en cuenta las consideraciones antes expuestas se podrá establecer el tamaño de muestra que permita responder en una primera fase las preguntas más importantes que se plantearon con el experimento.

Elección del tamaño de muestra por intervalo de confianza

Supongamos que el experimentador ya tiene el número de tratamientos que desea probar, k, y que tomando en cuenta las consideraciones antes citadas tiene una propuesta inicial del número de réplicas por tratamiento que va a utilizar, n_0. También tiene una idea aproximada del valor de σ (la desviación estándar del error aleatorio), así como una idea de la magnitud de las diferencias, d_T, entre tratamientos que le interesa detectar. Por ejemplo, supongamos que en el caso de los tiempos promedio de los $k = 4$ métodos de ensamble (ejemplo 3.1), tiene idea de realizar $n_0 = 5$ pruebas; en cuanto a las diferencias, le interesa detectar 2 minutos, $d_T = 2$ entre un método y otro, y espera que cada método tenga una variabilidad intrínseca de $\sigma = 1.5$; esto debido a factores no controlados (habilidad del operador, cansancio, variabilidad de las partes a ensamblar, error de medición del tiempo de ensamble, etcétera).

Ahora recordemos que en las comparaciones o pruebas de rangos múltiples, la diferencia mínima significativa entre tratamientos está dada por la expresión (3.12):

$$LSD = t_{(\alpha/2,\, N-k)}\sqrt{2CM_E/n}$$

despejando n de aquí, obtenemos:

$$n = \frac{2\left(t_{(\alpha/2,\, N-k)}\right)^2 CM_E}{(LSD)^2}$$

Si la significancia es $\alpha = 0.05$, entonces en esta fórmula se hacen las siguientes sustituciones: $N = k \times n_0$, $CM_E = \sigma^2$, $LSD = d_T$; de esta forma, el tamaño de muestra que tentativamente se debe usar está dado por,

$$n = \frac{2\left(t_{(0.025,\, k\times n_0 - k)}\right)^2 \sigma^2}{(d_T)^2}$$

El valor de n arrojado por esta fórmula dará una idea del número de réplicas por tratamiento, de acuerdo con las consideraciones iniciales que se reflejan a través de (k, n_0, σ, d_T), y sobre todo por el número total de corridas experimentales, $N = k \times n$, que es lo que muchas veces interesa más al experimentador debido a los costos y tiempos. Si N está fuera del presupuesto se podrán revisar algunas consideraciones y quizá pensar en un número menor de tratamientos.

Al aplicar esta expresión al caso de los cuatro métodos de ensamble obtenemos:

$$n = \frac{2(t_{(0.025,\,15)})^2\,(1.5)^2}{(2)^2} = \frac{2(2.131)^2\,(1.5)^2}{(2)^2} = 5.1$$

Por lo tanto, $n = 5$ se debería utilizar como tamaño de muestra (número de pruebas por tratamiento).

Uso de software computacional

Casi cualquier *software estadístico* incluye procedimientos para realizar análisis de varianza, comparar tratamientos y hacer análisis relacionados. En términos generales,

en una columna se registra el código para cada tratamiento corrido (se ponen tantos renglones como pruebas hechas), y en otra columna se registran los valores correspondientes obtenidos para *Y*. Con esto, en *Statgraphics* el análisis de los diseños comparativos se realiza básicamente en la opción *Compare* del menú principal.

La secuencia para un diseño completamente al azar es: *Compare* → *Analysis of variance* → *One-way anova*. En las opciones del procedimiento aparecen todas las pruebas y análisis que se han descrito en este capítulo.

Otra posibilidad en *Statgraphics* es accesar con la siguiente secuencia de opciones: *Special* → *Experimental Design* → *Create Design*, después de esto se debe elegir el tipo de diseño, que en este caso es *Single Factor Categorical*. Enseguida se define el número de niveles (tratamientos) y el nombre de los mismos. También se debe definir el nombre de la(s) variable(s) de respuesta(s). En la siguiente pantalla se pedirá el número de réplicas adicionales a la básica (si se pide una, en total se tendrán dos al considerar la réplica básica) y también aparece la opción de aleatorizar el orden para correr las pruebas, que siempre debe utilizarse en un diseño completamente aleatorizado. Todo esto permitirá generar una columna en la que se incluyen todas las pruebas a ser corridas, y una columna en blanco para cada variable de respuesta, la cual debe ser llenada en la medida que se vayan obteniendo los resultados del experimento. De la versión 15 de *Statgraphics* en adelante, la secuencia para crear diseños es *DOE* → *Design Creation*.

Para hacer el análisis, una vez generado el archivo de datos con los tratamientos y las respuestas, se siguen las opciones: *Special* → *Experimental Design* → *Analyze Design*, después se da el nombre de la variable de respuesta a analizar, y entonces se tendrá acceso a un conjunto de opciones de análisis tanto gráficas como analíticas, entre ellas las que hemos comentado en este capítulo.

En Minitab se registran los datos en dos columnas, como ya se dijo, y al ANOVA se accesa con la secuencia *Stat* → *Anova* → *One way*, y se da el nombre de las columnas que contienen los datos. También se eligen las comparaciones de medias deseadas y las gráficas.

Uso de Excel

El ANOVA de un diseño con un criterio de clasificación se realiza con la secuencia: *Herramientas* → *Análisis de datos* → *Análisis de varianza con un factor*. Si no estuviera activada la opción de *Análisis de datos*, se utiliza la opción de *Complementos* dentro del mismo menú de *Herramientas*. Entonces, se declara el rango de los datos, que pueden estar acomodados por columnas o por renglones. La salida contiene las estadísticas básicas de cada una de las muestras y el ANOVA correspondiente.

Preguntas y ejercicios

1. Explique en qué consiste y cuándo se debe aplicar el diseño completamente al azar con un solo criterio de clasificación.
2. Supongamos que se desea probar la igualdad entre sí de cinco medias. Una alternativa para hacer esto sería comparar de dos en dos las medias, utilizando la prueba *T* de Student y al final tomar una decisión. Explique por qué esto aumenta el error tipo I.