7

INTRODUCCIÓN AL ANÁLISIS DE VARIANZA

7.1 INTRODUCCIÓN

En las Ciencias Naturales son comunes los experimentos cuyo objetivo es comparar las medias de más de dos poblaciones. Por ejemplo, se puede desear medir el efecto de la temperatura sobre el tiempo que tarda en completarse una determinada reacción o transformación química. Con éste propósito un investigador puede seleccionar cuatro temperaturas distintas y medir el tiempo que tarda en ocurrir la transformación estudiada. Para aumentar la confiabilidad de los resultados la reacción puede repetirse varias veces, digamos cinco, para cada una de las temperaturas seleccionadas. En este ejemplo la variable independiente es la temperatura (T_j) que ha sido clasificada en cuatro niveles diferentes o tratamientos (T_1, T_2, T_3, T_4). Cada una de las mediciones efectuadas para una misma temperatura es una repetición (n_i) y el tiempo (t_{ij}) que tarda en completarse el proceso es la variable dependiente, donde $i = 1,2,3,4,5$ y $j = 1,2,3,4$. De modo que un valor cualquiera, por ejemplo t_{32}, es el tiempo que tarda en ocurrir la tercera repetición del proceso bajo el efecto del segundo nivel de temperatura. Los resultados obtenidos se suelen organizar como se muestra en la Tabla 7.1.

Tabla 7.1. Tiempo de reacción de una transformación química para cuatro temperaturas

	Temperaturas			
Repeticiones	T_1	T_2	T_3	T_4
1	t_{11}	t_{12}	t_{13}	t_{14}
2	t_{21}	t_{22}	t_{23}	t_{24}
3	t_{31}	t_{32}	t_{33}	t_{34}
4	t_{41}	t_{42}	t_{43}	t_{44}
5	t_{51}	t_{52}	t_{53}	t_{54}
Promedio	$\bar{X}_{.1}$	$\bar{X}_{.2}$	$\bar{X}_{.3}$	$\bar{X}_{.4}$

Una forma de verificar si las temperaturas afectan la velocidad de reacción, es comparando los valores promedios del tiempo de reacción de los grupos de datos. Si al menos uno de estos promedios es diferente se concluye que la temperatura afecta la velocidad de la reacción. Alternativamente se concluiría que los tiempos promedios son iguales y que la temperatura no afecta la velocidad de la reacción. El método más utilizado para tomar decisiones de éste tipo es el Análisis de Varianza (Andeva), el cual fue desarrollado por Ronald Fisher en 1921.

Aunque éste método originariamente se creó para analizar los resultados de experimentos agrícolas su uso se ha extendido a casi todas las disciplinas científicas.

7.2 FUNDAMENTOS DEL ANÁLISIS DE VARIANZA (ANDEVA)

Del nombre del método se colige inmediatamente que de alguna manera el examen de la variabilidad de un conjunto de muestras sirve para hacer inferencias acerca de la relación entre las varianzas de dos o más poblaciones de datos, pero como veremos más adelante, el método se utiliza más con el propósito de detectar si al menos, una de varias medias muestrales proviene de una población de valores con una media diferente. En la Tabla 7.2 se muestra el arreglo formal que deben tener los datos para efectuar un análisis de varianza para un único factor.

Tabla 7.2. Arreglo de los datos en un análisis de varianza de un único factor.

	Muestras					
Observaciones	1	2	3	.	k	
1	x_{11}	x_{12}	x_{13}	.	x_{1k}	
2	x_{21}	x_{22}	x_{23}	.	x_{2k}	
3	x_{31}	x_{32}	x_{33}	.	x_{3k}	
.	.	.	.	.	.	
n	x_{3n_1}	x_{3n_2}	x_{3n_3}	.	x_{3n_k}	
Medias	$\bar{x}_{.1}$	$\bar{x}_{.2}$	$\bar{x}_{.2}$	.	$\bar{x}_{.k}$	$\bar{\bar{x}}_{..}$
Varianzas	$s^2_{.1}$	$s^2_{.2}$	$s^2_{.3}$		$s^2_{.k}$	

Donde:

x_{ij} *: valor de la variable estudiada en la observación i-ésima de la j-ésima muestra.*

$i = 1,2,3,\ldots,n.$

$j = 1,2,3,\ldots,k.$

$\bar{x}_{.j}$ *= media de la j-ésima muestra.*

$\bar{\bar{x}}_{..}$ *= media de todas las medias de muestras.*

$s^2_{.j}$ *= varianza de la j-ésima muestra.*

Parece un poco extraño que se pueda llegar a conclusiones sobre las medias a través de un examen de las relaciones y magnitudes de las varianzas. Esta situación se tratará de aclarar a continuación, analizando a través de dos ejemplos, la lógica del Andeva.

Ejemplo 7.1

Un investigador planificó un experimento para determinar si diferentes contenidos de proteínas en la dieta afectan el crecimiento corporal de los ratones de laboratorio. Con tal fin, decidió seleccionar aleatoriamente ratones de una misma edad, sexo, raza y con un peso muy similar, para evitar que la variabilidad de esta medida interfiriera en la detección de los efectos de las dietas que se querían probar. Con este sentido, el investigador seleccionó 16 ratones de un tamaño similar y los asignó aleatoriamente en cuatro grupos de cuatro individuos. Después de pesar cada individuo, decidió comprobar estadísticamente si los pesos promedios de cada grupo estimaban la misma media poblacional para asegurarse que no había diferencias significativas en el tamaño promedio de los grupos. Los valores de peso inicial de los ratones se presentan en la Tabla 7.3.

Tabla 7.3. Peso inicial de ratones de laboratorio

	Peso (grs.)			
Individuo N°	Grupo 1	Grupo 2	Grupo 3	Grupo 4
1	56,3	58,2	56,1	56,9
2	57,0	57,2	54,2	55,9
3	54,0	58,4	56,4	54,0
4	56,7	55,8	55,9	55,0
Media	56,0	57,4	55,65	55,45
Varianza	1,8600	1,4133	0,9767	1,5367

Antes de analizar el método usado para resolver el problema de las posibles diferencias entre los pesos promedio, es importante analizar algunos aspectos relacionados con la naturaleza del método que debe usarse.

7.2.1 ¿Una o varias pruebas de hipótesis?

Suponiendo que la variable peso se distribuye normalmente con una media μ_X y una varianza σ_X^2, y que la muestra de 16 pesos es representativa, se puede considerar que cada grupo de pesos es una muestra aleatoria de la población de pesos y que las medias y las varianzas de cada grupo, estiman respectivamente la misma media μ_X y la misma varianza σ_X^2.

De cumplirse los supuestos anteriores, no deben existir diferencias estadísticamente significativas entre las medias de los grupos. Para comprobar esta presunción, el primer método que pudiera ensayarse sería una prueba de hipótesis para dos medias (Ej. Prueba de t). Como la prueba de t solo puede hacerse de dos en dos, para cuatro muestras (A, B, C, D), se requieren un total de seis comparaciones (AB, AC, AD, BC, BD, CD). Esto tiene un problema. Sí las muestras provienen de una misma población, en cada comparación hay una probabilidad de 0,95 de no existir diferencias significativas entre las medias muestrales, entonces la probabilidad de no diferencia en todos los casos, por la regla de multiplicación de probabilidades para eventos independientes sería $(0,95)^6 = 0,74$. Esto significa que la probabilidad de aceptar en cada comparación que dos medias son diferentes cuando en realidad no lo son, es igual $1 - 0,74 = 0,26$. Es decir que la posibilidad de equivocarse en cada prueba pasa de 5% a un 26%. Esta probabilidad aumenta considerablemente al aumentar el número de pruebas de t. Si se comparan cinco muestras la probabilidad de equivocarse al tomar una decisión es un poco mayor al 40%.

Dada esta complicación, la solución está en tener un método que determine simultáneamente la existencia de diferencias significativas entre las medias muestrales. Este método fue propuesto por R. Fisher y se fundamenta en el cálculo de dos varianzas: la varianza dentro de las muestras y la varianza entre las muestras, por lo cual se le conoce con el nombre de Análisis de Varianza (Andeva). A continuación se explicará como es posible comparar varias medias usando estas varianzas.

7.2.2 Varianza dentro de los grupos

La varianza dentro de las muestras se conoce comúnmente como varianza dentro de grupos, término que continuaremos usando. Para el caso del ejemplo 7.1, al provenir los datos de cada grupo de una misma población de valores, cada varianza de grupo estima la misma varianza poblacional cuyo valor se desconoce (Figura 7.1)

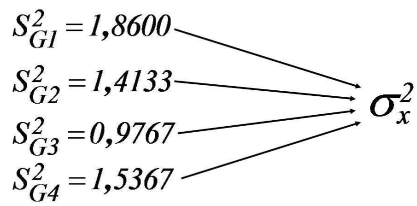


Figura 7.1

Una mejor estimación de la varianza poblacional se puede obtener promediando las varianzas de todos grupos y obtener una varianza promedio ponderada (S^2_p).

$$S^2_p = \frac{(n_1 - 1)\ S^2_{G1} + (n_2 - 1)\ S^2_{G2} + (n_3 - 1)\ S^2_{G3} + (n_4 - 1)\ S^2_{G4}}{n_1 + n_2 + n_3 + n_4 - 4} =$$

$$S^2_p = \frac{(4-1)\ 1,86 + (4-1)\ 1,4133 + (4-1)\ 0,9767 + (4-1)\ 1,5367}{12} = 1,4467$$

Esta varianza ponderada representa la variación en peso que existe entre los individuos dentro de cada grupo, por lo cual se denominará Varianza Dentro de Grupos (S^2_{DG}).

$$S^2_{DG} = 1,4467$$

7.2.3 Varianza entre grupos

Una segunda estimación de la varianza poblacional puede hacerse a partir de las medias de cada grupo. Puesto que cada grupo de pesos representa una muestra extraída de una misma población, entonces cada media de grupos es una media muestral. Es sabido que las medias muestrales obtenidas de una población distribuida normalmente también se distribuyen normalmente con una media $\mu_{\bar{x}} = \mu_x$ y varianza $\sigma^2_{\bar{x}} = \sigma^2_x / n$ (Figura 7.2)

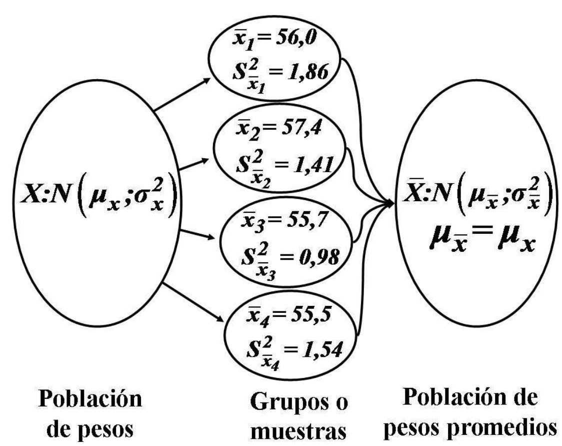


Figura 7.2

De la relación $\sigma^2_{\bar{x}} = \sigma^2_x / n$ se deduce que es posible calcular el valor de la varianza poblacional σ^2_x si se conoce $\sigma^2_{\bar{x}}$ y n. Aunque en el ejemplo que se viene trabajando se desconoce $\sigma^2_{\bar{x}}$, se puede estimar a partir del valor de la varianza muestral ($S^2_{\bar{x}}$).

$$S^2_{\bar{x}} = \frac{\sum_{j=1}^{k}\left(\bar{x}_{.j} - \bar{\bar{x}}_{..}\right)^2}{k-1} = \frac{\sum_{j=1}^{k}\left(\bar{x}_{.j}\right)^2 - \frac{\left(\sum_{j=1}^{K}\bar{x}_{.j}\right)^2}{k}}{k-1}$$

Donde:

$\bar{x}_{.j} =$ *el promedio de cada grupo j.*

$\bar{\bar{x}}_{..} =$ *el promedio total o media de las medias de los j grupos.*

k = número de grupo = 4. j = 1, 2, 3, 4

Al aplicar la ecuación anterior se tiene:

$$S^2_{\bar{x}} = \frac{12602{,}39 - \frac{(224{,}5)^2}{4}}{3} = 0{,}7742$$

Puesto que $S^2_{\bar{x}} = \frac{S^2_x}{n}$ se tiene que $S^2_x = nS^2_{\bar{x}}$; por lo tanto: $S^2_x = 4(0{,}7742) = 3{,}0967$. Esta varianza que representa la variación de peso que existe entre los grupos se denomina varianza entre grupos (S^2_{EG}) y es una segunda estimación de la varianza poblacional σ^2_x.

$$S^2_{EG} = 3{,}0967$$

7.2.4 Prueba de hipótesis para dos varianzas

Las dos varianzas que se acaban de calcular, la varianza dentro de grupos y la varianza entre grupos estiman la misma varianza poblacional. No hay razón alguna para pensar que no sea así puesto que, las diferencias observadas en el peso de los ratones dentro y entre grupo son simplemente aleatorias. En otras palabras, lo que hace distintas a dos medidas de peso dentro de un mismo grupo es lo mismo que hace distintas a dos medidas de peso de dos grupos diferentes. Es oportuno recordar que la varianza entre grupos refleja las diferencias existentes entre las medias de los grupos, pero a su vez estas medias se calcularon a partir de los valores dentro de cada grupo. De modo que al ser aleatorias las diferencias de los valores entre los grupos también lo son las diferencias entre las medias de esos mismos grupos. Por lo tanto si estadísticamente probamos que la varianza dentro de grupos y la varianza entre grupos son iguales, estaríamos probando que las medias de los grupos también son iguales. La igualdad de los dos tipos de varianza calculados en el ejemplo es fácil de comprobar con una prueba de hipótesis, como se verá a continuación:

Hipótesis:

$$H_o : \sigma^2_{EG} = \sigma^2_{DG}$$
$$H_1 : \sigma^2_{EG} \neq \sigma^2_{DG}$$

Nivel de significación: $1-\alpha=0,95$

Estadístico de prueba:

$$F_o=\frac{S^2_{EG}}{S^2_{DG}}=\frac{3,0967}{1,4467}=2,1406$$

Zona de aceptación de Ho:

$$ZA:\left\{F\ /\ f_{\left(\alpha/2\ ;\ k-1/N-k\right)}\leq F\leq f_{\left(1-\alpha/2\ ;\ k-1/N-k\right)}\right\}$$

$$ZA:\left\{F\ /\ f_{\left(0,025\ ;\ 3/12\right)}\leq F\leq f_{\left(0,975\ ;\ 3/12\right)}\right\}$$

$$ZA:\left\{F/0,0698\ \leq\ F\ \leq\ 4,47\right\}$$

Decisión: Como el estadístico de prueba $F_0=2,1406$ se encuentra dentro de la zona de aceptación de H_0, se concluye que los datos no aportan evidencia para rechazar H_0, por lo tanto se puede considerar que no existen diferencias significativas entre las varianzas entre y dentro de grupos. Este resultado se puede extrapolar y afirmar que al ser estas varianzas iguales las medias de los grupos son iguales, que era el resultado que se esperaba encontrar.

Con el procedimiento anterior se comprobó que tanto la varianza entre grupos como la varianza dentro de grupos estiman la misma varianza poblacional. A continuación se verá como este mismo procedimiento sirve también para someter a prueba la hipótesis de igualdad de dos o más medias poblacionales

Ejemplo 7.2
Continuando con el problema de los ratones, supongamos ahora que una vez que se comprobó que no existen diferencias significativas entre el peso promedio de los grupos de ratones se desea determinar sí diferentes contenidos de proteínas en la dieta afecta el crecimiento corporal de los individuos. Como variable independiente se puede usar la concentración de proteína aplicada en cuatro niveles, por ejemplo en dietas con 20%, 25%; 30% y 35% de proteína cruda respectivamente y la variable dependiente sigue siendo el peso. El experimento se inicia con la alimentación de los ratones de cada grupo con uno de los cuatro tipos de dieta y después de cierto tiempo se determina el peso de cada ratón. Con un sentido pedagógico supóngase que el efecto de la concentración proteica al 20% no tuvo ningún efecto, por lo que los pesos iniciales de los individuos del primer grupo no cambiaron y que además las otras concentraciones incrementaron el peso de los individuos de cada grupo en dos, tres y cuatro gramos respectivamente. El autor de éste texto les ruega a los lectores que temporalmente no

se den por enterados del modo en que fueron afectados los pesos de los ratones presentados en la Tabla 7.3. Los nuevos pesos finales se muestran en la Tabla 7.4.

Tabla 7.4. Peso final de ratones alimentados con cuatro dietas con distinto contenido de proteína.

	Peso final (gr)			
Individuo N°	20%	25%	30%	35%
1	56,3	60,20	59,10	60,90
2	57,0	59,20	57,20	59,90
3	54,0	60,40	59,40	58,00
4	56,7	57,80	58,90	59,00
Media	56,00	59,40	58,65	59,45
Varianza	1,8600	1,4133	0,9767	1,5367

Ahora es necesario comprobar si las diferentes concentraciones de proteína en la dieta modificaron el crecimiento de los ratones. Si las dietas tuvieron efecto, las medias de los grupos deben diferir significativamente. Como explicamos anteriormente, no es posible hacer comparaciones entre pares de medias porque aumenta sustancialmente la probabilidad de cometer el error tipo I, sin embargo mediante el análisis de varianza se puede determinar si efectivamente alguna de las medias es diferente. Por lo tanto se procede a calcular nuevamente las varianzas dentro y entre grupos para los datos de la Tabla 7.4.

7.2.5 Cálculo de la varianza dentro de los grupos

Esta varianza se calcula como el promedio ponderado de las varianzas de cada grupo. Como se puede notar en la Tabla 7.4 las varianzas de cada grupo no se modificaron después de aplicadas las dietas con relación a las varianzas de grupo de los datos de la Tabla 7.3 por lo tanto la varianza dentro de grupos (S^2_{DG}) no debe haber cambiado:

$$S^2_p = \frac{(n_1 - 1)\ S^2_{G1} + (n_2 - 1)\ S^2_{G2} + (n_3 - 1)\ S^2_{G3} + (n_4 - 1)\ S^2_{G4}}{n_1 + n_2 + n_3 + n_4 - 4} =$$

$$S^2_p = \frac{(4-1)\ 1{,}86 + (4-1)\ 1{,}4133 + (4-1)\ 0{,}9767 + (4-1)\ 1{,}5367}{12} = 1{,}4467$$

$$S^2_p = S^2_{DG} = 1{,}4467$$

Sr. Lector, como sabemos que los grupos 2, 3 y 4 fueron afectados por las diferentes dietas (secreto que compartimos) deberíamos preguntarnos ¿Cómo es posible que la varianza dentro de estos grupos no se alteró?. La respuesta la encontramos en aquella propiedad de la varianza que establece que la adición o sustracción de una constante a cada valor de un conjunto de datos no altera su valor, que fue lo que precisamente se hizo.

7.26 Cálculo de la varianza entre grupos

Como se vio anteriormente, el valor de la varianza entre grupos depende del valor de las medias muestrales:

$$S_{EG}^2 = S_{\bar{x}}^2 = nS_{\bar{x}}^2 = \frac{n\sum_{j=1}^{k}\left(\bar{x}_{.j} - \bar{\bar{x}}_{..}\right)^2}{k-1}$$

Si por alguna causa la separación entre las medias de los grupos aumenta, se incrementará la diferencia del término $(\bar{x}_{.j} - \bar{\bar{x}}_{..})^2$ y consecuentemente también lo hará la varianza entre grupos (S_{EG}^2) Esta dispersión de los valores de las medias de grupo puede llegar a ser lo suficientemente grande para hacer la S_{EG}^2 mucho mayor que la S_{DG}^2, llegada ésta situación se puede concluir que las dos varianzas no estiman la misma varianza poblacional.

Como vimos la varianza dentro de grupos (S_{DG}^2) sigue estimando la varianza poblacional de los pesos antes de aplicar los tratamientos, pero la varianza entre grupos estima una varianza poblacional distinta, mucho mayor que la estimada por la varianza dentro de grupos. Para verificar si los tratamientos afectaron suficientemente las medias de los grupos para hacer la varianza entre grupos significativamente superior a la varianza dentro de grupos procedamos a calcular su valor y efectuar una prueba de hipótesis para la igualdad de las varianzas.

$$S_{\bar{X}}^2 = \frac{\sum_{j=1}^{k}\left(\bar{x}_{.j} - \bar{\bar{x}}_{..}\right)^2}{k-1} = \frac{\sum_{j=1}^{k}\left(\bar{x}_{.j}\right)^2 - \frac{\left(\sum_{j=1}^{K}\bar{x}_{.j}\right)^2}{k}}{k-1} = \frac{13638{,}49 - \frac{(233{,}5)^2}{4}}{3} = 2{,}6408$$

$$S_{EG}^2 = S_{\bar{x}}^2 = nS_{\bar{x}}^2 = 4(2{,}6408) = 10{,}5633$$

7.2.7 Prueba de hipótesis para las varianzas

Hipótesis:

$$H_o : \sigma_{EG}^2 = \sigma_{DG}^2$$
$$H_1 : \sigma_{EG}^2 > \sigma_{DG}^2$$

En éste caso la hipótesis alternativa siempre propone que la varianza entre grupo es mayor que la varianza dentro de grupo, porque cualquier cambio en los valores de la variable dependiente tiende a hacer mayor la varianza entre grupos.

Estadístico de prueba:

$$F_o = \frac{S_{EG}^2}{S_{DG}^2} = \frac{10{,}5633}{1{,}4467} = 7{,}3018$$

Zona de aceptación de H_0:
La zona de aceptación en éste caso es de una cola a la derecha, porque la hipótesis alternativa establece una relación de mayor valor para σ^2_{EG}.

$$ZA: \left\{ F \,/\, F \le f_{\left(1-\alpha;\, k-1 \big/ N-k\right)} \right\}$$

$$ZA: \left\{ F \,/\, F \le f_{\left(0{,}95\ ;\ 3 \big/ 12\right)} \right\}$$

$$ZA: \left\{ F \big/ F \le 3{,}5 \right\}$$

Decisión:
Como el valor observado $F_0 = 7{,}3018$ se encuentra fuera de la zona de aceptación de H_0, se concluye que los datos aportan evidencia para rechazar H_0, por lo tanto se puede considerar que existen diferencias significativas entre las varianzas entre y dentro de grupos. Este resultado puede extrapolarse y afirmar que al ser estas varianzas diferentes las medias de los grupos son diferentes. Por lo tanto se puede decir que se tiene un 95% de confianza de que al menos una de las dietas incrementó el peso corporal de los ratones.

De todo lo explicado anteriormente se puede concluir que cuando se comparan varios grupos de datos, dentro de los mismos están presentes dos fuentes de variación. Una denominada Variación Dentro de Grupo, la cual refleja las diferencias entre las observaciones dentro de cada grupo. Estas diferencias son simplemente aleatorias y pueden ser de naturaleza genética, ambiental y de medición. Las mismas pueden minimizarse pero no eliminarse totalmente. En la medida que los datos provengan de elementos más homogéneos, por ejemplo de individuos que sean lo más parecido posible en cuanto a atributos como el sexo, la edad, progenitores, talla, etc., disminuyen las diferencias entre ellos. Sin embargo siempre habrá un remanente de diferencias que no es posible eliminar, por ésta razón la varianza dentro de grupo se le denomina en forma genérica Varianza Residual o Remanente o simplemente Error. Esta varianza no cambia aún después de alterar los valores por el efecto de aplicar algún tipo de tratamiento y sigue estimando la varianza poblacional común a todos los grupos (σ^2).

La segunda fuente de variación se denominó Variación Entre Grupo y la misma evidencia la diferencia de los valores entre grupos. Cuando las medidas se obtienen de una misma población y no se les aplica tratamiento alguno, las diferencias entre ellas son simplemente aleatorias, es decir que la variabilidad entre grupos tiene la misma naturaleza que la variabilidad dentro de grupos. En este caso la prueba de hipótesis de la razón de varianzas no muestra diferencias significativas entre las varianzas.

$$\frac{S^2_{EG}}{S^2_{DG}} \Rightarrow \frac{\sigma^2_{EG}}{\sigma^2_{DG}} = \frac{\sigma^2}{\sigma^2} = 1$$

Por el contrario, cuando los grupos de valores provienen de poblaciones diferentes o si proviniendo de una misma población son alterados por la aplicación de algún tratamiento (τ), la varianza entre grupos deja de estimar la varianza poblacional inicial (σ^2)

$$\frac{S^2_{EG}}{S^2_{DG}} \Rightarrow \frac{\sigma^2_{EG}}{\sigma^2_{DG}} = \frac{\sigma^2 + \tau}{\sigma^2} > 1$$

La varianza entre grupos por lo general se denomina varianza debido a tratamientos.

7.3 PARTICIÓN DE LA SUMA TOTAL DE CUADRADOS

Hasta ahora sabemos que para un dado conjunto de datos es posible identificar dos diferentes fuentes de variación: una es la variación dentro de grupos que deja ver el promedio de las diferencias aleatorias que existen entre los valores dentro de los grupos; la otra es la variación entre grupos que evidencia además de las diferencias aleatorias de los valores entre grupos, las eventuales diferencias debido a los efectos de los tratamientos. Pero además de las dos varianzas anteriores es posible calcular la varianza total si se consideran todos los valores como un único gran conjunto de datos. Calculemos dichas varianzas para los datos que se presentan en la Tabla 7.5.

Tabla 7.5. Valores seleccionados aleatoriamente de una misma población.

	Grupo 1	Grupo 2	Grupo 3
	4,33	4,10	4,00
	3,85	4,33	4,16
	4,32	4,24	4,29
	3,96	4,48	3,89
	3,74	4,42	4,20
Media	4,040	4,314	4,108
Varianza	0,07375	0,02258	0,02587

Puesto que los datos fueron escogidos aleatoriamente de la misma población, las diferencias entre los valores dentro o entre grupos son aleatorias, por lo tanto las tres varianzas (entre, dentro y total) que vamos a calcular a continuación estiman la misma varianza poblacional (σ^2).

1) Varianza dentro Grupo

$$S^2_{DG} = \frac{(n_1-1)S^2_{G1} + (n_2-1)S^2_{G2} + (n_3-1)S^2_{G3}}{n_1+n_2+n_3-3} = \frac{(4)0,07375+(4)0,02258+(4)0,02587}{12} = 0,04073$$

2) Varianza entre grupos

$$S^2_{EG} = nS^2_{\bar{x}} = n\frac{\sum_{j=1}^{3}\left(\bar{x}_{.j} - \bar{\bar{x}}_{..}\right)^2}{k-1} = 5\frac{(0,040712)}{2} = \frac{0,20356}{2} = 0,10178$$

3) Varianza total

$$S^2_T = \frac{\sum_{j=1}^{N}\left(x_{.j} - \bar{\bar{x}}_{..}\right)^2}{N-1} = \frac{0,69236}{14} = 0,04945$$

Es razonable pensar que si la varianza total se calcula usando todos los datos, la magnitud de la misma debería incluir las otras dos varianzas. Sin embargo, como las varianzas no son aditivas no es posible establecer una relación aritmética entre las tres varianzas.

$$S^2_{EG} + S^2_{DG} = 0.10178 + 0.04073 = 0.14251 \quad >> \quad S^2_T = 0.04945$$

Esta incongruencia aparente, se puede aclarar si se presta atención a las sumas de cuadrados, que es el término genérico que se le da al numerador de cualquier varianza. La suma de cuadrados (SC) es una manera de medir la dispersión de un conjunto de n datos y se expresa como la sumatoria del cuadrado de la diferencia entre cada dato y su valor promedio.

$$SC = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Para el caso del Andeva, las sumas de cuadrados se identificarán de la forma siguiente:

Suma de cuadrados Total = SCT
Suma de cuadrados entre grupos = SCEG
Suma de cuadrados dentro de grupo = SCDG

De modo que las respectivas varianzas se pueden representar con las fórmulas siguientes:

$$S^2_T = \frac{SCT}{N-1} \qquad S^2_{EG} = \frac{SCEG}{k-1} \qquad S^2_{DG} = \frac{SCDG}{N-k}$$

Es evidente que las sumas de cuadrados es el término que dentro de la ecuación de la varianza mide la dispersión de los datos. La propiedad más importante de la suma de cuadrados es que son aditivas, es decir que la adición de las sumas de cuadrados entre y dentro de grupo es igual a la suma de cuadrados total. Aunque esto puede ser demostrado algebraicamente, puede bastar como comprobación verificar la relación entre las sumas de cuadrados calculadas en el ejemplo que se viene trabajando. En este caso los resultados de las varianzas fueron los siguientes:

$$S_{EG}^{2} = \frac{SCEG}{k-1} = \frac{0,20356}{2} = 0,10178$$

$$S_{DG}^{2} = \frac{SCDG}{N-k} = \frac{0,4888}{12} = 0,04072$$

$$S_{T}^{2} = \frac{SCT}{N-1} = \frac{0,69236}{14} = 0,04945$$

Es fácilmente comprobable que la Suma de Cuadrados Total esta conformada por las otras dos sumas de cuadrados.

$$\begin{gathered} SCEG \quad + \quad SCDG \quad = \quad SCT \\ 0,20356 \quad + \quad 0,48880 = \quad 0,69236 \end{gathered}$$

La propiedad anterior aclara la incongruencia planteada anteriormente acerca de la variabilidad total que, cuando es medida como una varianza, no incluye la variabilidad dentro y entre los grupos. Las relaciones que se acaban de analizar ofrecen la ventaja de facilitar los cálculos de las medidas involucradas en el Análisis de Varianza.

7.4 NOTACIÓN BÁSICA Y CÁLCULOS NECESARIOS

Supongamos que tenemos k poblaciones de una variable aleatoria que se distribuye normalmente con la misma varianza σ^2 y con medias $\mu_{x_1}; \mu_{x_2}; \mu_{x_3}; ....; \mu_{x_k}$. De cada población se extrae en forma aleatoria e independiente una muestra de tamaño n y los datos se ordenan como se muestra en la Tabla 7.6.

Tabla 7.6. Ordenación de valores muestrales en el cálculo de un Andeva

	K Muestras					
Observaciones	1	2	3	.	k	
1	x_{11}	x_{12}	x_{13}	.	x_{1k}	
2	x_{21}	x_{22}	x_{23}	.	x_{2k}	
3	x_{31}	x_{32}	x_{33}	.	x_{3k}	
.	.	.	.	.	.	
n_j	$x_{n_1 1}$	$x_{n_2 2}$	$x_{n_3 3}$	.	$x_{n_k k}$	Gran Total
Total	$\sum_{i=1}^{n_1} x_{i1}$	$\sum_{i=1}^{n_1} x_{i2}$	$\sum_{i=1}^{n_1} x_{i3}$	.	$\sum_{i=1}^{n_1} x_{ik}$	$\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}$
Medias	$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	$\bar{X}_{\cdot 3}$	.	$\bar{X}_{\cdot k}$	$\bar{\bar{x}}_{\cdot\cdot}$

Siendo:

$$x_{ij} = \text{una observación o valor cualquiera} \qquad \left\{ \begin{matrix} i = 1, 2, 3, \ldots, n_j \\ j = 1, 2, 3, \ldots, k \end{matrix} \right\}$$

$$\sum_{i=1}^{n_j} x_{ij} = \text{total de la j-ésima columna}$$

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij} = \text{total de todas las observaciones}$$

$$\bar{x}_{.j} = \frac{\sum_{i=1}^{n_j} x_{i.}}{n_j} = \text{media de la j-ésima muestra}$$

$$\bar{\bar{x}}_{..} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{N} = \text{media de todas las observaciones}$$

7.4.1 Sumas de cuadrados (SC)

Las sumas de cuadrados, términos necesarios para calcular las varianzas entre y dentro de grupos se obtienen relacionando los promedios y valores totales presentados en la Tabla 7.6.

$$\text{Suma de cuadrados total} = SCT = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{\bar{x}}_{..} \right)^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} \right)^2 - \frac{\left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij} \right)^2}{N}$$

$$\text{Suma de cuadrados entre grupos} = SCEG = n_j \left[\sum_{j=1}^{k} \left(\bar{x}_{.j} - \bar{\bar{x}}_{..} \right)^2 \right] = \sum_{j=1}^{k} \frac{\left(\sum_{i=1}^{n_j} x_{ij} \right)^2}{n_j} - \frac{\left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij} \right)^2}{N}$$

$$\text{Suma de cuadrados dentro de grupos} = SCDG = SCT - SCEG = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_{.j} \right)^2$$

7.4.2 Cuadrados medios (CM)

Dividiendo las respectivas sumas de cuadrados entre los grados de libertad se obtienen las varianzas entre y dentro de grupos. En el Andeva las varianzas se denominan cuadrados medios, entonces se habla del cuadrado medio entre grupos (CMEG) o del cuadrado medio dentro de grupos (CMDG).

$$Varianza\ entre\ grupos = CMEG = \frac{SCEG}{k\text{-}1}$$

$$Varianza\ dentro\ de\ grupos\ = CMDG = \frac{SCDG}{k(n\text{-}1)}$$

7.4.3 Grados de libertad
Los grados de libertad (denominador en la ecuación de la varianza) también son aditivos, de modo que:

Grados de libertad total = Grados libertad entre grupos + Grados libertad dentro grupos

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

7.5 ANÁLISIS DE VARIANZA DE UN FACTOR

El modelo de Andeva que se ha venido analizando se conoce como diseño de un solo factor, de una sola vía o clasificación única, porque interviene una sola variable independiente que puede tener varios niveles de acción. Estos niveles de acción, se le denomina comúnmente como tratamientos. Cuando se consideran dos variables independientes a la vez, el diseño se denomina de dos factores, de dos vías o de clasificación doble. Para un número mayor de variables independientes los arreglos se denominan multifactoriales.

A continuación se presenta el procedimiento completo para efectuar un análisis de varianza de un factor.

1. Hipótesis:

$$H_o:\ \mu_{x_1} = \mu_{x_2} = \mu_{x_3} = ... = \mu_{x_k}$$
$$H_1:\ al\ menos\ una\ de\ las\ \mu_{x_j}\ \ es\ diferente$$

2. Sumas de cuadrados
   Se obtienen los datos básicos necesarios para calcular las sumas de cuadrados y los cuadrados medios, para lo cual se deben calcular las tres cantidades siguientes:

$$a)\ \sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij} \qquad b)\ \sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}^2 \qquad c)\ \sum_{j=1}^{k}\frac{\left(\sum_{i=1}^{n_j} x_{ij}\right)^2}{n_j}$$

   Sumas de cuadrados total (SCT)

$$SCT = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}\right)^2 - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N}$$

Sumas de cuadrados entre grupos (SCEG)

$$SCEG = \sum_{j=1}^{k} \frac{\left(\sum_{i=1}^{n_j} x_{ij} \right)^2}{n_j} - \frac{\left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij} \right)^2}{N}$$

Sumas de cuadrados dentro de grupos (SCEG) o residual

$$SCDG = SCT - SCEG$$

3. Tabla de Andeva
   Se construye la tabla resumen del análisis de varianza (Tabla 7.7)

Tabla 7.7. Tabla de Andeva de una vía

Fuente de variación	Suma de cuadrados	Grados de Libertad	Cuadrados Medios	F_o
Entre grupos (Entre tratamientos)	$\sum_{j=1}^{k} \frac{\left(\sum_{i=1}^{n_j} x_{ij} \right)^2}{n_j} - \frac{\left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij} \right)^2}{N}$	$k-1$	$\frac{SCEG}{k-1}$	$\frac{CMEG}{CMDG}$
Dentro de grupos (Residual o error)	$SCT - SCEG$	$N-k$	$\frac{SCDG}{N-k}$	
Total	$\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} \right)^2 - \frac{\left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij} \right)^2}{N}$	$N-1$		

4. Se establece la zona de aceptación para la hipótesis de igualdad de las varianzas

$$ZA: \left\{ F/F < f_{[1-\alpha\ ;\ k-1/\ N-k]} \right\}$$

Ejemplo 7.3
En un estudio sobre el SIDA se quiere saber si hay diferencias en los niveles de la droga AZT en la sangre a los 60, 90, 120 y 150 minutos de haberse aplicado la misma. El tratamiento se le administró a cuatro grupos de pacientes de la misma edad, raza, sexo y que no absorben bien las grasas. Los resultados se muestran en la Tabla 7.8.

Tabla 7.8. Concentración de AZT en la sangre en distintos tiempos desde su aplicación.

Paciente N°	60 min.	90 min.	120 min.	150 min.
1	2,69	1,91	1,72	0,22
2	3,37	1,89	2,11	1,40
3	2,42	1,61	1,41	1,09
4	3,30	1,81	1,16	0,69
5	2,61	1,90	1,24	1,01
6	2,17	1,88	1,34	0,24
7	3,65	2,32		1,02
8	2,37	2,07		1,34
9		2,29		0,97
Media	2,8225	1,9644	1,4967	0,8867

a. Hipótesis:

$$H_o: \ \mu_{x_1} = \mu_{x_2} = \mu_{x_3} = ... = \mu_{x_k}$$
$$H_1: \ \text{al menos una de las } \mu_{x_j} \text{ es diferente}$$

La hipótesis nula desde el punto de vista biológico presume que la concentración de AZT en la sangre no cambia después de los 60 minutos de haberse aplicado.

b. Sumas de cuadrados
Se efectúan los cómputos necesarios para calcular las sumas de cuadrados, los cuadrados medios y el valor de F (Tabla 7.9.)

Tabla 7.9. Cálculos necesarios para el Andeva del ejemplo 7.3.

Paciente N°	60 min.	90 min.	120 min.	150 min.	Gran total
$\bar{X}_j$	2,8225	1,9644	1,4967	0,8867	
$\sum_{j=1}^{k} x_{ij}$	22,5800	17,6800	8,9800	7,9800	57,2200
$\sum_{j=1}^{k} x_{ij}^2$	65,7998	35,1442	14,0774	8,5272	123,5486
$\sum_{j=1}^{k} \frac{\left(\sum_{i=1}^{n_j} x_{ij}\right)^2}{n_j}$	63,7321	34,7314	13,4401	7,0756	118,9791

Suma de cuadrados total (SCT)

$$SCT=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}\right)^2-\frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j}x_{ij}\right)^2}{N}=123{,}5486-\frac{\left(57{,}22\right)^2}{32}=123{,}5486-102{,}3165=21{,}2321$$

Suma de cuadrados entre grupos (SCEG)

$$SCEG=\sum_{j=1}^{k}\frac{\left(\sum_{i=1}^{n_j}x_{ij}\right)^2}{n_j}-\frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j}x_{ij}\right)^2}{N}=118{,}9791-\frac{\left(57{,}22\right)^2}{32}=118{,}9791-102{,}3165=16{,}6626$$

Suma de cuadrados dentro grupos (SCDG)

$$SCDG\ =\ SCT\ -\ SCEG=21{,}2321-16{,}6626=4{,}5695$$

c. Tabla de Andeva
Se construye la tabla resumen del análisis de varianza (Tabla 7.10.)

Tabla 7.10. Tabla de Andeva para el ejemplo 7.3.

Fuente de variación	Suma de cuadrados	Grados de Libertad	Cuadrados Medios	F_o
Entre grupos (Entre tiempos)	16,6626	3	5,5542	34,03
Dentro de grupos (Residual o Error)	4,5695	28	0,1632	
Total	21,2321	31		

d. Se establece la zona de aceptación para la hipótesis de igualdad de las varianzas

$$ZA:\left\{F\ /\ F<f_{[1-\alpha;k-1/N-k]}\right\}=\left\{F\ /\ F<f_{[0{,}95;3/28]}\right\}=\left\{F\ /\ F<2{,}95\right\}$$

e. Decisión: como el valor del estadístico de prueba $F_0=34{,}03$ es mucho mayor que el límite crítico $(\ f=2{,}95\ )$, se rechaza H_0, por lo tanto se acepta la hipótesis alternativa, que propone que al menos una de las muestras proviene de una población de valores con una media diferente.

f. Conclusión: la concentración de AZT en la sangre de pacientes afectados de SIDA y con mala absorción de grasas presenta un valor distinto en al menos uno de los lapsos transcurridos desde la aplicación del tratamiento.

Ejemplo 7.4.
En un estudio sobre la extracción de iones metálicos por ciertos compuestos, se determinó el % de eficiencia de extracción del hierro por cuatro agentes quelantes. El experimento se repitió tres veces para cada compuesto. Se quiere saber si entre los quelantes existen diferencias en su capacidad de extracción ¿Cuál es la conclusión si se quiere sólo se aceptan un error menor al 1% al tomar una decisión? Los resultados se presentan en la Tabla 7.11.

Tabla 7.11. Cantidad de Hierro extraído (%) por cuatro agentes quelantes.

Experimento N°	Quelante 1	Quelante 2	Quelante 3	Quelante 4
1	84	80	83	79
2	79	77	80	79
3	83	78	80	78
Media	82	78,33	81	78,67

a) Hipótesis:

$$H_o: \ \mu_{x_1} = \mu_{x_2} = \mu_{x_3} = \ldots\ldots = \mu_{x_k}$$
$$H_1: \ \textit{al menos una de las } \mu_{x_j} \textit{ es diferente}$$

La hipótesis nula desde el punto de vista químico presume que la eficiencia de extracción de los cuatro quelantes es la misma

b) Sumas de cuadrados
Se efectúan los cómputos necesarios para calcular las sumas de cuadrados, los cuadrados medios y el valor de F (Tabla 7.12.)

Tabla 7.12. Cálculos necesarios para el Andeva del ejemplo 7.4.

	Quelante 1	Quelante 2	Quelante 3	Quelante 4	Gran total
$\bar{X}_j$	82,00	78,33	81,00	78,67	
$\sum_{i=1}^{n_j} x_{ij}$	246	235	243	236	960
$\sum_{j=1}^{n_j} x_{ij}^2$	20186	18413	19689	18566	76854
$\dfrac{\left(\sum_{i=1}^{n_j} x_{ij}\right)^2}{n_j}$	20172	18408,33	19683	18565,33	76828,67

Sumas de cuadrados total (SCT)

$$SCT = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}\right)^2 - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N} = 76854 - \frac{(960)^2}{12} = 76854 - 76800 = 54$$

Sumas de cuadrados entre grupos (SCEG)

$$SCEG = \sum_{j=1}^{k}\frac{\left(\sum_{i=1}^{n_j} x_{ij}\right)^2}{n_j} - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N} = 76828{,}6 - \frac{(960)^2}{12} = 76826{,}6 - 76800 = 28{,}6$$

Sumas de cuadrados dentro de grupos (SCEG) o residual

$$SCDG = SCT - SCEG = 54{,}0 - 28{,}6 = 25{,}40$$

c) Tabla de Andeva
Se construye la tabla resumen del análisis de varianza (Tabla 7.13.).

Tabla: 7.13. Tabla de Andeva para el ejemplo 7.4.

Fuente de variación	Suma de cuadrados	Grados de Libertad	Cuadrados Medios	F_o
Entre grupos (Entre tratamientos)	28,60	3	9,53	3,0
Dentro de grupos (Residual o Error)	25,40	8	3,175	
Total	54,00	11		

d) Se establece la zona de aceptación para la hipótesis de igualdad de las varianzas

$$ZA: \left\{F \,/\, F < f_{[1-\alpha\ ; k-1/N-k\ ]}\right\} = \left\{F \,/\, F < f_{[0{,}99\ ;\ 3/8]}\right\} = \left\{F \,/\, F < 7{,}59\right\}$$

e) Decisión: Como el valor del estadístico de prueba $F_0 = 3{,}0$ es menor al límite crítico $(\ f = 7{,}59\ )$, se acepta H_0.

f) Conclusión: se acepta con un 99% de confianza que los valores promedios del % de extracción de los cuatro quelantes no difieren significativamente.

7.6. COMPARACIÓN MÚLTIPLE DE MEDIAS

Como hemos visto el Andeva se usó para contrastar la hipótesis nula de no diferencia entre las medias de k poblaciones:

$$H_0 = \mu_1 = \mu_2 = \mu_3 = \cdots\cdots = \mu_k$$

El Andeva finaliza cuando se acepta H_0, pero si se rechaza H_0 y se concluye que al menos una de las medias μ_j es diferente. Conocido este resultado, es muy frecuente que se quiera saber cual o cuales son esas medias, respuesta que no ofrece el Andeva. Por ello se han desarrollado una serie de métodos que en forma genérica se identifican como Comparaciones Múltiples de Medias. Estos intentan identificar las medias o grupos de medias que son diferentes. Algunos autores agrupan las comparaciones múltiples en dos categorías: Pruebas *a priori* y Pruebas *a posteriori*.

Con las pruebas *a priori* las comparaciones deben ser planificadas antes de efectuar el experimento y decidir de antemano que medias se van a comparar. Son muy pocas las comparaciones que deben efectuarse, para algunos autores no más de tres. Las pruebas *a posteriori* se realizan una vez obtenidos los resultados y sólo si H_0 ha sido rechazada. En este tipo de prueba las comparaciones se hacen entre todas las parejas posibles.

La razón de la distinción anterior se verá a continuación: supóngase que se tiene una población de valores con una distribución normal con media μ_x y varianza $\sigma_{\bar{x}}^2$; sí de esta población se extraen pares de muestras todas con el mismo tamaño n y calculamos sus medias $\bar{x}_j$, hemos visto que es posible originar una nueva variable $\Delta\bar{x} = \bar{x}_1 - \bar{x}_2$ denominada diferencia de medias muestrales, que por la propiedad reproductiva de la distribución normal sabemos que también se distribuye normalmente con una media esperada $\mu_{(\bar{x}_1 - \bar{x}_2)} = 0$ y una varianza esperada $2\sigma_{\bar{x}}^2$. Dado que la distribución es normal se espera que la mayoría de pares de medias tendrán diferencias pequeñas y se ubicarán alrededor del valor 0, y que algunas pocos pares tendrán diferencias lo suficientemente grandes para ubicarse en los extremos o colas de la distribución. Si desconociéramos de donde provienen las medias muestrales y deseáramos probar si dos medias muestrales se extrajeron de dos poblaciones diferentes, podríamos contrastar la hipótesis nula $H_o : \mu_{x_1} = \mu_{x_2}$ con un nivel de significación $1 - \alpha = 0,95$. Bajo esta situación todos aquellos pares de media con diferencias muy grandes permitirán rechazar H_0, pero esto ocurrirá sólo en el 5% de los casos. En otras palabras, si escogiéramos aleatoriamente las medias muestrales, la probabilidad de rechazar H_0 siendo cierta sería igual a 0,05. Pero sí la selección de las medias muestrales no es al azar y se escogieran a conciencia las medias muestrales con mayor separación en sus valores, siempre se tendrían pares de medias con grandes diferencias y la hipótesis nula H_0 se rechazaría una y otra vez, a pesar de ser cierta. Esto mismo es lo que pasa si después de obtener los resultados de un experimento se escogen en forma deliberada las medias muestrales que se van a comparar. Por tal razón en las pruebas *a priori* se deben seleccionar previo al experimento las muestras cuyas medias se desean comparar. Para el caso de las comparaciones *a posteriori* donde se puede escoger las muestras a comparar, al no ser ésta selección aleatoria no se puede seguir usando la

distribución de probabilidades sobre la cual se basa el Andeva, es necesario usar una distribución de probabilidad diferente la cual cambia de uno a otro método.
Son numerosos los métodos de comparación múltiple. Entre las pruebas *a priori* se encuentran la Mínima Diferencia Significativa (MDS); los Contrastes Ortogonales y la Prueba de Dunnett. Entre las pruebas *a posteriori* se pueden mencionar las pruebas de Tukey o Diferencia Verdadera Significativa (DVS); de Student, Newman y Keuls (NKS); de amplitudes múltiples de Duncan; de Scheffé, de Bonferroni y la de Gabriel. A continuación estudiaremos un método de cada tipo: la Prueba de la Mínima Diferencia Significativa (MDS) y la Prueba de Tukey o de la Diferencia Verdaderamente Significativa (DVS).

7.6.1. Prueba de la Mínima Diferencia Significativa (MDS).
Esta fue la primera prueba de comparación múltiple y fue introducida por Sir Ronald Fisher. La misma es una prueba *a priori*. Es recomendable que con la MDS no se efectúen más de tres comparaciones, pues la posibilidad de equivocarse al rechazar H_o siendo cierta es superior al 18%.
El fundamento de la MDS es muy sencillo. Supongamos que un Andeva aplicado a varias muestras fue significativo y queremos comparar si dos de las medias muestrales provienen de poblaciones diferentes. Se puede recurrir a una prueba de hipótesis para dos medias poblacionales (Prueba de t).

Hipótesis:
$$H_o : \mu_{x_1} = \mu_{x_2}$$
$$H_1 : \mu_{x_1} > \mu_{x_2}$$

Sabemos que al comparar dos medias cuyas muestras provienen de dos poblaciones con la misma varianza el estadístico a usar es:

$$T = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_{x_1} - \mu_{x_2})}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$$

Como $n_1 = n_2$; $S_p^2 = CMDG$ y $\mu_{x_1} = \mu_{x_2}$, la expresión anterior queda igual a:

$$T = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{2CMDG}{n}}}$$

La zona de aceptación de H_o es:

$$ZA : \left\{ T \,/\, T < t_{(1-\alpha\ ;\ N\text{-}k)} \right\}$$

Esta ZA se puede presentar en su forma equivalente:

$$ZA : \left\{ T \,/\, \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{2CMDG}{n}}} < t_{(1-\alpha\ ;\ N\text{-}k)} \right\}$$

La regla de decisión se puede expresar de la forma siguiente:

$$\text{Sí } (\bar{x}_1 - \bar{x}_2) > t_{(1-\alpha\ ;\ N\text{-}k)}\sqrt{\frac{2CMDG}{n}} \text{ se rechaza } H_0$$

Esta diferencia es el menor valor que puede existir entre dos medias para aceptar H_0, si la diferencia es mayor se rechaza H_0. Esta es la razón por lo que tal diferencia se denomina Mínima Diferencia Significativa (MDS). Al extrapolarse esta comparación a todas las medias, la suposición de igualdad de las medias poblacionales es rechazada cada vez que se cumple:

$$MDS > t_{(1-\alpha;\ N\text{-}k)}\sqrt{\frac{2CMDG}{n}}$$

Ejemplo 7.5

Un ecólogo que estudia los requerimientos nutricionales de una especie de monos, quiere determinar si la calidad de la dieta de tres poblaciones de monos que viven en forma salvaje en tres localidades diferentes (L1, L2, L3), es igual a la de una población de monos que está bajo protección especial en un parque nacional (PN). Con tal propósito colectó cinco muestras de sangre de hembras adultas para cada una de las cuatro poblaciones y determinó el contenido de ácido fólico ($\mu g / l$) en la sangre. Compruebe con un 99% de confianza si el contenido de ácido fólico en la sangre de las poblaciones silvestres es diferente al de la población protegida. Los resultados del análisis de sangre se muestran en la Tabla 7.14.

Tabla 7.14. Contenido de Ácido fólico $(\mu g / l)$ en la sangre de monos provenientes de cuatro poblaciones diferentes.

Individuo N°	PN	L1	L2	L3
1	257,20	174,40	221,20	175,20
2	294,90	185,00	231,00	165,90
3	283,70	166,40	228,60	174,80
4	310,00	172,10	215,80	171,60
5	305,20	184,50	205,10	191,10
Media	290,20	176,48	220,34	175,72

A) Andeva

a) Hipótesis: $H_o :\ \mu_{x_1} = \mu_{x_2} = \mu_{x_3} = ... = \mu_{x_k}$

$H_1 :\ \text{al menos una de las } \mu_{x_j} \text{ es diferente}$

La hipótesis nula desde el punto de vista biológico presume que la concentración de Ácido Fólico en la sangre es la misma en las cuatro poblaciones de monos.

b) Sumas de cuadrados

Se efectúan los cómputos necesarios para calcular las sumas de cuadrados, los cuadrados medios y el valor de F (Tabla 7.15.)

Tabla 7.15. Cálculos necesarios para el Andeva del ejemplo 7.5.

	PN	LI	L2	L3	Gran total
$\sum_{j=1}^{k} x_{ij}$	1451,00	882,40	1101,70	878,60	4313,70
$\sum_{j=1}^{k} x_{ij}^2$	422850,58	155987,98	243184,05	154738,66	976761,27
$\sum_{j=1}^{k} \frac{\left(\sum_{i=1}^{n_j} x_{ij} \right)^2}{n_j}$	421080,20	155725,95	242748,58	154387,59	973942,32

Suma de cuadrados total (SCT)

$$SCT = \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}^2 - \frac{\left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij} \right)^2}{N} = 976761,27 - \frac{(4313,70)^2}{20} = 46360,89$$

Suma de cuadrados entre grupos (SCT)

$$SCEG = \sum_{j=1}^{k} \frac{\left(\sum_{i=1}^{n_j} x_{ij} \right)^2}{n_j} - \frac{\left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij} \right)^2}{N} = 973942,32 - \frac{(4313,70)^2}{20} = 43541,94$$

Suma de cuadrados dentro de grupos (SCDG) o residual

$$SCDG = SCT - SCEG = 46360,89 - 43541,94 = 2818,95$$

c) Se construye la tabla resumen del análisis de varianza (Tabla 7.16)

Tabla 7.16. Tabla de Andeva para el ejemplo 7.5.

Fuente de variación	Suma de cuadrados	Grados de Libertad	Cuadrados Medios	F_o
Entre grupos (Entre localidades)	43541,94	3	14513,98	82,38
Dentro de grupos (Residual o Error)	2818,95	16	176,18	
Total	46360,89	19		

d) Se establece la zona de aceptación para la hipótesis de igualdad de las varianzas

$$ZA: \left\{ F \big/ F < f_{[1-\alpha\ ; k-1/N-k\ ]} \right\} = \left\{ F \big/ F < f_{[0{,}99\ ; 3/16\ ]} \right\} = \left\{ F \big/ F < 5{,}29 \right\}$$

e) Decisión: como el valor del estadístico de prueba $F_0 = 82{,}38$ es mucho mayor que el límite crítico $(f = 5{,}29)$, se rechaza H_0, por lo tanto se acepta la hipótesis alternativa, que propone que al menos una de las muestras proviene de una población de valores con una media diferente.

f) Conclusión: la concentración de Ácido Fólico en la sangre de monos provenientes de diferentes localidades presenta un valor distinto en al menos una de las poblaciones.

Comparación Múltiple de Medias para el ejemplo 7.5.

Son tres las comparaciones que se quieren efectuar: la media de la población protegida contra la media de cada una de las tres poblaciones silvestres. Esto se decidió antes de efectuar la experiencia, por lo tanto se puede usar una prueba *a priori* como la de la Mínima Diferencia Significativa.

a) Se calcula el valor de MDS.

$$MDS = t_{[1-\alpha\ ;\ N\text{-}k]}\sqrt{\frac{2CMDG}{n}} = t_{(0{,}99;\ 16\ )}\sqrt{\frac{2\left(176{,}184\right)}{5}} = 2{,}583\ \left(8{,}395\right) = 21{,}68$$

Se prepara una tabla con las diferencias a probar, las medias deben estar ordenadas en un sentido creciente o decreciente (Tabla 7.17.)

Tabla 7.17. Diferencias entre las medias muestrales del ejemplo 7.5.

	PN = 290,20
L3 = 175,72	116,48**
L1 = 176,48	113,72**
L2 = 220,34	69,86**

(**) = Las diferencias son muy significativas (p < 0,01)

b) Se establece la regla decisión:

Sí $MDS > 21{,}68$ se rechaza la hipótesis H_0 de igualdad de las medias poblacionales.

Al aplicar la regla de decisión se observa que todas las diferencias de medias de la tabla son mayores que el valor de la MDS, por lo tanto se acepta que las medias que se están comparando se diferencian significativamente.

c) Conclusión: Se puede concluir que los datos proporcionan suficiente evidencia para aceptar con un 99% de confianza que el contenido promedio de Ácido Fólico en la

sangre de la población de monos bajo protección es mayor que el de las poblaciones silvestres.
Es importante advertir que con esta prueba no se deben sacar conclusiones sobre las diferencias de medias que no fueron previamente seleccionadas, pues esto aumentaría el número de comparaciones y también aumentaría considerablemente la probabilidad de cometer el error tipo I.

7.6.2. Prueba de Tukey o de la Diferencia Verdaderamente Significativa (DVS).

Este método se usa para efectuar todas las posibles comparaciones entre pares de medias muestrales con el propósito de contrastar la hipótesis de igualdad de sus medias poblacionales. La prueba consiste en calcular un único valor crítico o DVS contra el cual se comparan todas las diferencias entre los pares de medias. La DVS se obtiene con la fórmula siguiente:

$$DVS = q_{\alpha[k;N\text{-}k]}\sqrt{\frac{CMDG}{n}}$$

El valor $q_{\alpha} =$ estadístico de Tukey, se encuentra en una tabla cuyos argumentos de entrada son k y $N\text{-}k$. Siendo k el número total de muestras de involucradas en el Andeva; $N\text{-}k$ los grados de libertad con los cuales se calculó el CMDG; n es el tamaño de las muestras y α = probabilidad de rechazar H_0 siendo cierta (Error tipo I). Las diferencias de medias que sean superiores al valor de DVS son significativamente diferentes. Cuando las muestras no tienen el mismo tamaño, se puede sustituir el valor de n por el promedio siguiente:

$$n \approx \frac{2n_M n_m}{(n_M + n_m)};$$

Donde : $n_M =$ muestra de mayor tamaño y $n_m =$ muestra de menor tamaño

Ejemplo 7.6.

Este ejemplo es una variante de ejemplo 7.5 de las poblaciones de monos. Supóngase que el ecólogo ahora le interesa en efectuar todas las comparaciones posibles. En este caso puede recurrirse a una prueba *a posteriori* como la de Tukey. Se usarán los mismos datos

a) Se calcula el valor de DVS.

$$DVS = q_{\alpha[k;N\text{-}k]}\sqrt{\frac{CMDG}{n}} = q_{0,01[4;16]}\sqrt{\frac{CMDG}{n}} = 5,19\sqrt{\frac{176,18}{5}} = 30,81$$

Se prepara una tabla con las diferencias a probar, las medias deben estar ordenadas en un sentido creciente o decreciente (Tabla 7.18).

Tabla 7.18. Diferencias entre pares de medias muestrales del ejemplo 7.6

	L3 = 173,72	L1 = 176,48	L2 = 220,34	PN = 290,20
L3 = 173,72	-	2,76^{ns}	46,62**	116,48**
L1 = 176,48	-	-	43,86**	113,72**
L2 = 220,34	-	-	-	69,86**

(**ns**) = no existen diferencias significativas (P > 0,01) (******) = Las diferencias son muy significativas (P < 0,01)

b) Se establece la regla de decisión.

Sí $DVS > 30{,}18$ se rechaza la hipótesis H_0 de igualdad de las medias poblacionales.

c) Se toma la decisión para cada comparación:

L1 - L3 = 2,76 es menor a 30,81. Se acepta H_0.
L2 – L3 = 46,62 es mayor a 30,81. Se rechaza H_0.
L2 – L1 = 43,86 es mayor a 30,81. Se rechaza H_0.
PN – L3 = 116,48 es mayor a 30,81. Se rechaza H_0.
PN – L1 = 113,72 es mayor a 30,81. Se rechaza H_0.
PN – L2 = 69,86 es mayor a 30,81. Se rechaza H_0.

En lugar de especificar una por una las comparaciones, se puede indicar en la tabla de comparaciones de medias cuales son las diferentes usando la llamada (*) o cuales medias son iguales mediante la llamada (ns). Otra forma mucho más práctica es la siguiente: las medias se ordenan con una secuencia creciente o decreciente y aquellas medias muestrales que provienen de poblaciones con la misma media poblacional se subrayan con la misma línea.

L3 = 173,72 L1 = 176,48 L2 = 220,34 PN = 290,20

En la representación anterior se advierte inmediatamente que L1 y L2 tienen medias que no se diferencian significativamente, mientras que las medias de L2 y PN se diferencian significativamente entre sí y con las otras medias.

d) Conclusión
Se concluye que los datos proporcionan suficiente evidencia para aceptar con un 99% de confianza que el contenido promedio de Ácido Fólico en la sangre de los monos que habitan las localidades 1 y 3 son iguales y es diferente entre los monos que viven en las localidades L2 y PN y entre estos y los monos de las otras poblaciones.

7.7 MODELO Y SUPUESTOS BÁSICOS DEL ANÁLISIS DE VARIANZA

7.7.1 Modelo lineal

Para facilitar el conocimiento del modelo, usaremos un ejemplo sencillo. Supóngase que se tienen tres poblaciones de valores de una variable cualquiera que se distribuye normalmente con medias μ_1, μ_2 y μ_3 respectivamente. La media general será igual a:

$$\mu = \frac{\sum_{j=1}^{3} \mu_j}{3}$$

Las desviaciones entre cada media y la media general es igual a $\tau_j = \mu_j - \mu$ siendo $j = 1, 2, \cdots, k$. Por lo tanto, μ_1, μ_2 y μ_3 difieren de μ en τ_1, τ_2 y τ_3. Esto es,

$$\mu_1 = \mu + \tau_1 \qquad \mu_2 = \mu + \tau_2 \qquad \mu_3 = \mu + \tau_3$$

Los τ_j representan los efectos o desviación debido a los tratamientos. Como cada τ_j es una desviación respecto a una media, se debe cumplir:

$$\sum_{j=1}^{3} \tau_j = \sum_{j=1}^{3} (\mu_1 - \mu) = 0$$

Se puede notar que cuando $\tau_j = 0$, todas las medias son iguales: $\mu_1 = \mu_2 = \mu_3$

En la Figura 7.3 se muestra gráficamente la situación planteada. Dentro de cada población los valores x_{ij} se distribuyen alrededor del promedio μ_j y la desviación de cada x_{ij} respecto a μ_j es producto del azar. Tales desviaciones se denominan errores aleatorios y se expresan como, $\varepsilon_{ij} = x_{ij} - \mu_j$. De ésta ecuación se deduce que el valor de cada x_{ij} es igual a $x_{ij} = \mu_j + \varepsilon_{ij}$. Como $\mu_j = \mu + \tau_j$ se tiene que $x_{ij} = \mu + \tau_j + \varepsilon_{ij}$. Esta ecuación constituye el modelo básico para el caso de un Andeva de una vía, donde un factor ejerce efectos fijos sobre las diferentes muestras.

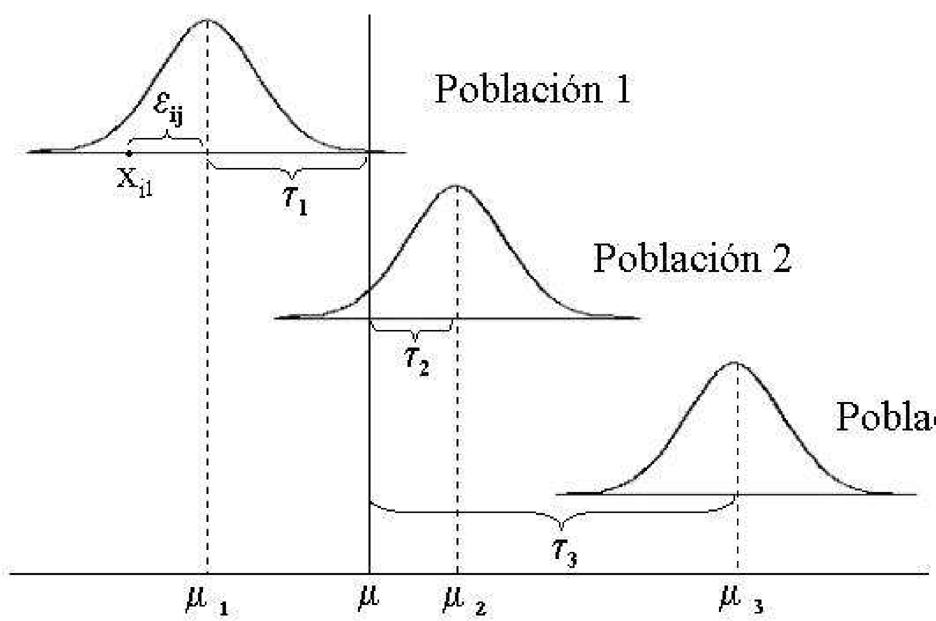


Figura 7.3**.** Esquema de los efectos de tratamientos (τ_{ij}) y los errores aleatorios (ε_{ij} ε_{ij}) presentes en un Andeva de efectos fijos.

7.7.2 Supuestos básicos

La validez del modelo anterior depende del cumplimiento de los supuestos siguientes: a) los tratamientos y los efectos ambientales son aditivos; y b) las desviaciones o errores (ε_{ij}) son aleatorios y se distribuyan normal e independientemente con una media $\mu = 0$ y una misma varianza σ^2. El incumplimiento de uno o más de estos supuestos puede conducir a la toma de decisiones equivocadas con una mayor frecuencia que la fijada por el nivel de probabilidad escogido. Al respecto Steel y Torrie (1988) señalan *"Los experimentadores pueden pensar que está usando un nivel del 5 por ciento cuando el nivel puede ser en realidad de 7 u 8 por ciento"*.

7.7.2.1 Aditividad

La aditividad ocurre cuando los tratamientos aplicados a un grupo de muestras afectan los valores de cada muestra en forma aritmética o lineal. Si el resultado de aplicar cada tratamiento modifica los valores de cada muestra en forma geométrica o no lineal el efecto se dice que es multiplicativo.

La falta de aditividad conduce a una heterogeneidad de las varianzas, de manera que la varianza dentro de grupos no estima una varianza comun a todas las poblaciones. En las Tablas 7.19 y 7.20 se muestra como el efecto multiplicativo afecta las varianzas de los grupos. En la Tabla 7.19 para lograr un efecto aditivo a cada valor de los grupos 2 y 3 se le añadió un valor constante que simula el efecto de cada tratamiento. Como se puede observar las varianzas de cada grupo no cambia.

Tabla 7.19. Efecto aditivo de los tratamientos sobre los valores de tres muestras.

	Tratamientos		
	x_{ij+0}	x_{ij+2}	x_{ij+3}
	4	6	7
	5	7	8
	7	9	10
	3	5	6
Media	4,75	6,75	7,75
Varianza	2,92	2,92	2,92
Varianza dentro de grupo = 2,92			

Tabla 7.20. Efecto multiplicativo de los tratamientos sobre los valores de tres muestras.

	Tratamientos		
	x_{ij}	$2x_{ij}$	$3x_{ij}$
	4	8	12
	5	10	15
	7	14	21
	3	6	9
Media	4,75	9,50	14,25
Varianza	2,92	11,67	26,25
Varianza dentro de grupo = 13,61			

En la Tabla 7.20 cada valor de los grupos 2 y 3 se multiplicó por un factor constante. La varianza de estos dos grupos aumentó cerca de 4 y 9 veces respectivamente. Por su parte la varianza dentro de grupos (13,61) es muy superior al valor esperado, que debe ser muy parecido al estimado por la varianza del primer grupo (2,92), donde no se aplicó ningún tratamiento. La falta de aditividad afecta especialmente las comparaciones individuales entre pares de medias. La transformación de los datos en logaritmos resuelve, por lo general, los efectos multiplicativos.

7.7.2.2 Aleatoriedad

Un requisito fundamental para efectuar un Andeva es que la selección de las muestras debe ser aleatoria, De lo contrario, el Andeva pueden ser un método ineficiente en la detección de diferencias entre medias por falta de independencia de los datos, de homogeneidad en las varianzas y de normalidad en la distribución. De modo que nunca están de más todos los cuidados que aseguren la aleatoriedad del muestreo tanto en experimentos de laboratorio, como cuando se hacen ensayos u observaciones de campo. Tal como lo señala Cochran y Cox (1957) la aleatorización no es sino una precaución contra errores que pueden o no ocurrir y que pueden ser o no graves si ocurren.

7.7.2.3 Independencia

La violación del supuesto de independencia afecta la validez de la prueba de hipótesis sobre igualdad de varianzas. Veamos un ejemplo**,** supongamos que en un estudio sobre el contenido de nitrógeno en el suelo, se quiere determinar su contenido en cuatro zonas de un lote de terreno, tal como se muestra en la Figura 7.4.

Como en el lote existe un gradiente en la distribución del contenido de nitrógeno, se esperaría que las mediciones hechas en sitios muy próximos proporcionen valores muy similares entre ellos y menos parecidos con los sitios más alejados, por lo tanto las n observaciones de cada muestra deben tener un mayor parecido entre ellas que con las de las otras muestras. Una consecuencia de este sesgo en el muestreo sería la pérdida de independencia de un dato respecto a otro, puesto que la probabilidad de seleccionar algunos individuos depende de la elección previa de otros. En nuestro ejemplo la falta de independencia determinaría una disminución de las diferencias entre las mediciones dentro de una muestra y consecuentemente del valor de la varianza dentro de grupo. Igualmente las diferencias entre las medias de las muestras serían mayores a lo esperado y se sobrestimaría la varianza entre grupos. Estos dos hechos afectarían los resultados de un Andeva al incrementar el valor de la razón de varianzas y la probabilidad de rechazar la hipótesis de igualdad de las varianzas. La falta de independencia en el caso anterior se debe básicamente al procedimiento usado para obtener los datos, de modo que el problema se podría resolver modificando dicho procedimiento mediante un muestreo aleatorio. Otra situación que ejemplifica como el diseño de un experimento puede determinar una pérdida de independencia en el registro de los datos es la que se describe a continuación. Supóngase que se van a suministrar diferentes dosis de una droga a animales de laboratorio de diferentes tamaños. Si el suministro de las dosis a los animales es aleatorio sin considerar su tamaño, también es posible que ocurra una pérdida de independencia cuando se hacen experimentos más relacionados con el tiempo que con el espacio. Por ejemplo aquellos ensayos que requieren varios días para completarse deben efectuarse aleatorizando en el tiempo la aplicación del tratamiento o la toma de mediciones.

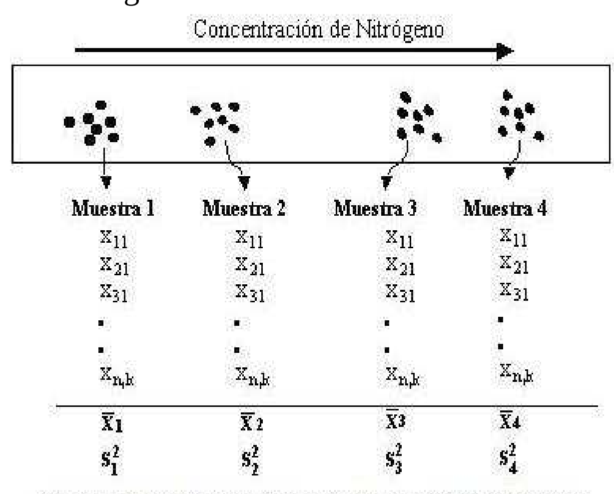


Figura 7.4. Distribución del contenido de nitrógeno del suelo.

7.7.2.4 Homogeneidad de las Varianzas

Como se vio anteriormente la varianza dentro de grupos se calcula promediando las varianzas de cada grupo o muestra. Esto se hace bajo el supuesto que dichas varianzas estén estimando la misma varianza poblacional independientemente de que las muestras provengan de la misma o diferentes poblaciones. De modo que la igualdad de las varianzas en un grupo de muestras es una condición indispensable en el Andeva. Esta condición también se conoce bajo el nombre de Homogeneidad u Homoscedasticidad de las varianzas. La falta de homogeneidad puede enmascarar eventuales diferencias estadísticas entre las medias de varias muestras. Una situación como la anterior la ilustraremos con un ejemplo.

Ejemplo 7.7.

En un experimento hipotético se aplicaron cuatro tratamientos, cada uno repetido cinco veces como se muestra en la Tabla 7.21.

Tabla 7.21. Tabla de resultados del ejemplo 7.7

	Tratamientos			
Repetición	A	B	C	D
1	4	8	215	231
2	6	9	205	227
3	2	11	221	245
4	8	17	212	229
5	7	10	235	225
Total	27,0	55,0	1088,0	1157,0
Media	5,4	11,0	217,6	231,4
Varianza	5,8	12,5	127,8	62,8

Los resultados del Andeva, se presentan en la Tabla 7.22.

Tabla 7.22. Tabla de Andeva para el ejemplo 7.7

Fuente de variación	SC	GL	CM	Fo
Entre Tratamientos	234483	3	78161,0	14963,62***
Residual	865,6	16	52,225	
Total	235319	19		

Los resultados anteriores determinaron la aceptación de la hipótesis alternativa de que al menos una de las medias proviene de una población de valores diferentes con un nivel de confianza del 95%. Para conocer cuales medias difirieron se aplicó una prueba de Tukey. Después de calcular el valor del estadístico de Tukey (DVS = 13,08) se encontró que las medias de los tratamientos A y B son iguales entre sí y que las medias de los tratamientos C y D son diferentes entre sí.

$\overline{X}_A = 5,4$ $\quad \overline{X}_B = 11,0$ $\quad \overline{X}_C = 217,6$ $\quad \overline{X}_D = 231,4$

Si volvemos a examinar la tabla de datos se puede notar que existe una gran diferencia entre las varianzas, con una relación aproximada de 20 a 1 para los dos grupos de varianzas (C-D/A-B). Existen algunas pruebas estadísticas, como la de Bartlett y la F_{max}, para comprobar la igualdad de varias varianzas, las cuales no trataremos, pero que son de fácil aplicación y aparecen en muchos de textos de estadística básica. Bajo la circunstancia de no tener varianzas homogéneas, lo recomendable es comparar los tratamientos A y B (Tabla 7.23) separados de los tratamientos C y D (Tabla 7.24)

Tabla 7.23. Tabla de Andeva para la comparación entre los tratamientos A y B

Fuente de variación	SC	GL	CM	Fo
Entre Tratamientos	78,4	1	78,4	8,57*
Residual	73,2	8	9,15	
Total	151,6	9		

Tabla 7.24. Tabla de Andeva para la comparación entre los tratamientos C y D

Fuente de variación	SC	GL	CM	Fo
Entre Tratamientos	476,1	1	476,1	$5{,}0^{ns}$
Residual	762,4	8	95,3	
Total	1238,5	9		

Los resultados de estas dos tablas muestran como la comparación por separado, produjo resultados totalmente opuestos a los obtenidos con la comparación conjunta de los cuatro tratamientos. Ahora los promedios de A y B son diferentes y los promedios de C y D son iguales. En otras ocasiones, la desigualdad de las varianzas se produce por la tendencia de muchas variables a tener medias y varianzas correlacionadas positivamente. En este caso se puede usar una transformación logarítmica de los datos para homogeneizar las varianzas. Es oportuno llamar la atención que cuando las diferencias entre las varianzas no son muy pronunciadas se puede efectuar sin mayores inconvenientes el Andeva de efectos fijos, puesto que la prueba de razón de varianzas es bastante robusta a esta situación.

7.7.2.5 Normalidad

La falta de normalidad en la distribución de los errores afecta las pruebas de significación de la razón de varianzas, siempre y cuando la distribución de los datos sea fuertemente asimétrica y/o multimodal. De lo contrario, cuando el sesgo de las distribuciones es moderado los resultados y conclusiones del Andeva no son afectados de manera importante. El método más usado para corregir la no normalidad es la transformación de los datos. De no solucionarse el problema se debe recurrir a las pruebas no paramétricas.

7.8 ANÁLISIS DE VARIANZA DE DOS FACTORES

Hasta ahora el análisis de varianza (Andeva) se ha utilizado para determinar o descubrir si dos o más conjuntos de mediciones que han estado sometidas al efecto de diferentes niveles de acción de un único factor presentan diferencias estadísticas significativas. En este tipo de Andeva los datos de las muestras se arreglan y presentan en una matriz como la siguiente:

	Nivel de acción del factor A				
Observación	1	2	3	.	a
1	x_{11}	x_{12}	x_{13}	.	x_{1a}
2	x_{21}	x_{22}	x_{23}	.	x_{2a}
3	x_{31}	x_{32}	x_{33}	.	x_{3a}
.	.	.	.	.	.
n	x_{n1}	x_{n2}	x_{n3}	.	x_{na}

Donde:

$x_{ij} =$ *valor de la variable medida para la repetición i y el nivel j del factor principal*

$i = 1,2,3, \ldots, n$ *repeticiones*

$j = 1,2,3, \ldots, a$ *niveles de acción del factor principal*

El diseño anterior supone que otros factores diferentes al factor controlado que pueden afectar los datos estudiados se mantienen constantes. Si se desea evaluar el efecto del factor A bajo otro nivel de acción de algún otro factor, es necesario repetir completamente el experimento para un nuevo nivel de acción del segundo factor o factor B. Si por ejemplo se escogen tres niveles de acción del factor B, se tendrían tres matrices de datos (Figura 7.5 A,B,C).

	b_1		
n	a_1	a_2	a_3
1	x_{111}	x_{121}	x_{131}
2	x_{211}	x_{221}	x_{231}
3	x_{311}	x_{321}	x_{331}

A

	b_2		
n	a_1	a_2	a_3
1	x_{112}	x_{122}	x_{132}
2	x_{212}	x_{222}	x_{232}
3	x_{312}	x_{322}	x_{332}

B

	b_3		
n	a_1	a_2	a_3
1	x_{113}	x_{123}	x_{133}
2	x_{213}	x_{223}	x_{233}
3	x_{313}	x_{323}	x_{333}

C

Figura 7.5

En este arreglo el término x_{ijk} representa el valor de la variable medida para la repetición i, bajo el nivel j del factor A y el nivel k del factor B$(k = 1,2,3)$. En total se requerirían 27 ensayos para medir el efecto del factor A para tres condiciones diferentes del factor B.
Si contrariamente lo que se quiere medir es el efecto del factor B bajo tres diferentes niveles del factor A, se tendría que diseñar un experimento similar al anterior donde el factor B actuaría en tres niveles diferentes, manteniendo constante el factor A (Figura 7.6).

	a_1		
n	b_1	b_2	b_3
1	x_{111}	x_{112}	x_{113}
2	x_{211}	x_{212}	x_{213}
3	x_{311}	x_{312}	x_{313}

A

	a_2		
n	b_1	b_2	b_3
1	x_{121}	x_{122}	x_{123}
2	x_{221}	x_{222}	x_{223}
3	x_{321}	x_{322}	x_{323}

B

	a_3		
n	b_1	b_2	b_3
1	x_{131}	x_{132}	x_{133}
2	x_{231}	x_{232}	x_{233}
3	x_{331}	x_{332}	x_{333}

C

Figura 7.6

En total, se requerirían efectuar 54 ensayos, para evaluar el efecto de tres niveles de acción del factor A y tres niveles de acción del factor B, sobre tres observaciones. Afortunadamente, el Andeva permite evaluar simultáneamente el efecto de dos o más factores, disminuyendo grandemente el número de mediciones o ensayos. En el caso del ejemplo anterior, solo se necesitarán nueve ensayos para efectuar la evaluación conjunta de los dos factores sobre la variable estudiada. Bajo este diseño experimental, sólo se necesita medir el valor de la variable dependiente cuando está sometida al efecto conjunto de los dos factores, siguiendo como modelo el arreglo siguiente:

	Factor A		
Factor B	a_1	a_2	a_3
b_1	x_{11}	x_{12}	x_{13}
b_2	x_{21}	x_{22}	x_{23}
b_3	x_{31}	x_{32}	x_{33}

Donde:

x_{jk} = *tiempo de desarrollo bajo el nivel* a_i *y el nivel* b_k

$j = 1,2,3.$

$k = 1,2,3.$

A continuación se usará un ejemplo para explicar el procedimiento usado cuando se aplica un Andeva de dos factores.

Ejemplo 7.8

Un investigador que estudia las características vitales de una especie de insecto que transmite una enfermedad viral, desea evaluar el efecto que tiene la humedad relativa y la temperatura sobre el tiempo de desarrollo de las larvas del insecto vector. Con tal fin el investigador cría larvas del insecto combinando tres humedades relativas y cinco temperaturas. Los resultados se muestran en la Tabla 7.25.

Tabla 7.25. Tiempo de desarrollo de las larvas de un insecto alimentados bajo tres humedades relativas y cinco temperaturas diferentes.

	Tiempo de desarrollo (días)		
Temperaturas (°C)	70% HR	75% HR	80% HR
22	9,68	12,65	10,68
24	8,22	10,69	9,55
26	7,95	9,95	9,04
28	7,41	8,82	8,55
30	7,39	7,68	7,45

El proceso de resolución de un Andeva de dos factores, básicamente consiste en sustraer en forma independiente la varianza de cada factor de la varianza total, para lo cual es necesario considerar que los valores de la variable medidos bajo el efecto de los diferentes niveles de acción de un factor se toman como las repeticiones efectuadas bajo uno de los niveles de acción del otro factor. Por ejemplo, los valores de la primera columna de datos (9,68; 8,22; 7,95; 7,41; 7,39) que son los valores del tiempo de desarrollo a cinco diferentes temperaturas, también son las cinco repeticiones para la humedad relativa de 70%.

	Tiempo de desarrollo (días)		
Temperaturas (°C)	70% HR	75% HR	80% HR
22	9,68	.	.
24	8,22	.	.
26	7,95	.	.
28	7,41	.	.
30	7,39	.	.

Igualmente, los valores de tiempo de desarrollo para las diferentes humedades relativas de la primera fila de los datos (9,68; 12,65; 10,68) son las repeticiones para la temperatura de 22°C.

	Tiempo de desarrollo (días)		
Temperaturas (°C)	70% HR	75% HR	80% HR
22	9,68	12,65	10,68
24	.	.	.
26	.	.	.
28	.	.	.
30	.	.	.

A continuación se desarrollará paso a paso el procedimiento usado para efectuar el Andeva de dos factores.

a) <u>Hipótesis de investigación.</u>

a.1) Primer factor: Humedad Relativa

H_0: *El tiempo de desarrollo no es afectado por la humedad relativa.*
H_1 : *La humedad relativa afecta el tiempo de desarrollo.*

a.2) Segundo factor: Temperatura

H_0: *El tiempo de desarrollo es indedependiente de la temperatura.*
H_1 : *El tiempo de desarrollo depende de la temperatura.*

b) <u>Hipótesis estadísticas.</u>

b.1) Primer factor: Humedad Relativa

H_o : $\mu_1 = \mu_2 = \mu_3$ *(No hay diferencias entre los tiempos promedios de desarrollo)*
H_1 : *al menos una de las HR determina un tiempo promedio de desarrollo* μ_j *diferente.*

Donde:

$j = 1,2,3$ *niveles o tratamientos del factor 1 (Humedad Relativa).*

b.2) Segundo factor: Temperatura

H_o : $\mu_1 = \mu_2 = \mu_3 = ... = \mu_i$ *(No hay diferencias en los tiempos promedios de desarrollo)*
H_1: *al menos una de las T°C determina un tiempo promedio de desarrollo* μ_i *diferente.*

c) Sumas de cuadrados

c.1) Cálculos necesarios para el Andeva

T°C	*70%HR*	*75%HR*	*80%HR*	$\sum_{j=1}^{k_i} X_{ij}$	$\sum_{j=1}^{k_i} X_{ij}^2$	$\frac{\left(\sum_{j=1}^{k_i} X_{ij}\right)^2}{k_i}$	$\bar{X}_{filas}$
22	9,98	10,68	12,65	33,31	373,685	369,852	11,10
24	8,22	9,55	10,69	28,46	273,047	269,991	9,49
26	7,95	9,04	9,95	26,94	243,927	241,921	8,98
28	7,41	8,55	8,72	24,68	204,049	203,034	8,23
30	7,39	8,45	8,68	24,52	201,357	200,410	8,17
$\sum_{i=1}^{n_j} X_{ij}$	40,95	46,27	50,69	137,91	1296,06	1285,21	
$\sum_{j=1}^{n_j} X_{ij}^2$	339,892	431,492	524,682	1296,065			
$\frac{\left(\sum_{i=1}^{n} X_{ij}\right)^2}{n_j}$	335,381	428,183	513,895	1277,458			
$\bar{X}_{columnas}$	8,19	9,25	10,14				

c.2) Sumas de Cuadrados total (SCT)

$$SCT = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}\right)^2 - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N} = 1296{,}065 - \frac{(137{,}91)^2}{15} = 28{,}120$$

c.3) Suma de cuadrados entre columnas (SCEC)

$$SCEC = \sum_{j=1}^{k}\frac{\left(\sum_{i=1}^{n_j} x_{ij}\right)^2}{n_j} - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N} = 1277{,}458 - \frac{(137{,}91)^2}{15} = 9{,}514$$

c.4) Suma de cuadrados entre filas (SCEF)

$$SCEF = \sum_{i=1}^{n}\frac{\left(\sum_{j=1}^{k} x_{ij}\right)^2}{k} - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N} = 1285{,}21 - \frac{(137{,}91)^2}{15} = 17{,}27$$

c.5) Suma de cuadrados del error o residual (SCE)

$$SCE = SCT\text{-}SCEC\text{ - }SCEF = 28{,}12 - 9{,}514 - 17{,}27 = 1{,}34$$

c.6) Tabla de Andeva

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	F_0
Humedad relativa	9,514	2	4,757	28,33
Temperatura	17,263	4	4,316	25,71
Dentro	1,343	8	0,168	
Total	28,120	14		

d) <u>Zona de aceptación para la hipótesis de igualdad de las varianzas</u>.

d.1) <u>Primer factor: Humedad Relativa.</u>

$$ZA: \left\{ F \,/\, F < f_{\left[1-\alpha\ ;\ {}^{k-1}\!/\!_{(k-1)(n-1)}\right]} \right\} = \left\{ F / F < f_{\left[0.95\ ;\ {}^{2}\!/\!_{8}\right]} \right\} = \left\{ F / F < 4.46 \right\}$$

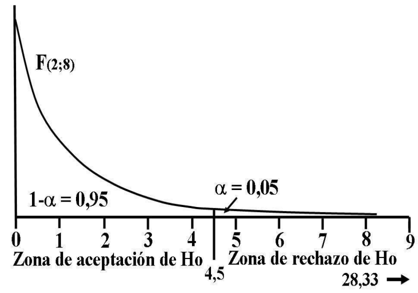
d.2) Segundo factor: Temperatura.

$$ZA: \left\{ F \,/\, F < f_{\left[1-\alpha\ ;\ n-1/(k-1)(n-1)\right]} \right\} = \left\{ F \,/\, F < f_{\left[0.95\ ;\ 4/8\right]} \right\} = \{ F \,/\, F < 3.84 \}$$

e) Decisión.

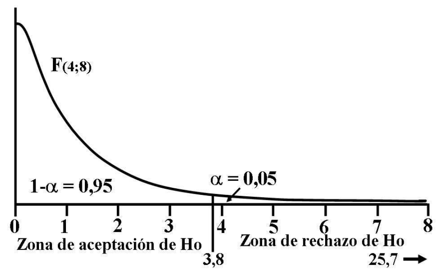
e.1) Primer factor: Humedad relativa

Dado que el valor del estadístico de prueba $F_o = 28,33$ es mucho mayor que el límite crítico $F_{(0,95\ ;\ 2/8)} = 4,5$ se rechaza Ho y se acepta con un 95% de confianza que al menos una de las HR determina que el promedio del tiempo de desarrollo correspondiente es diferente.



e.2) Segundo factor: Temperaturas

Dado que el valor del estadístico de prueba $F_o = 25,71$ es mayor que el límite crítico $F_{(0,95\ ;\ 4/8)} = 3,8$ se rechaza Ho y se acepta con un 95% de confianza que al menos una de las temperaturas determina que el promedio del tiempo de desarrollo correspondiente es diferente.



f) Conclusión.

f.1) Primer factor: una o más de las humedades relativas determina un tiempo de desarrollo diferente.

f.2) Segundo factor: una o más de las temperaturas determina un tiempo de desarrollo diferente.

El Andeva se completa mediante una prueba de comparación múltiple de medias a fin de determinar las humedades relativas y temperaturas que producen diferentes tiempos de desarrollo.

7.9. ¿UNO O DOS FACTORES?

Una mejor comprensión de los fundamentos del análisis de varianza de dos factores, se obtiene al comparar sus resultados con los obtenidos con los dos análisis de varianza que sería necesario efectuar para examinar el efecto de cada factor individualmente.

A continuación se muestran las tablas de resultados para los tres análisis de varianza:

TABLA DE ANDEVA UN FACTOR: HUMEDAD RELATIVA

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	F_o
Humedad	9,514	2	4,757	$3{,}07^{ns}$
Error o Residual	18,607	12	1,551	
Total	28,120	14		

TABLA DE ANDEVA UN FACTOR: TEMPERATURA

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	F_o
Temperatura	17,263	4	4,316	$3{,}98^{*}$
Error o Residual	10,857	10	1,086	
Total	28,120	14		

TABLA ANDEVA DOS FACTORES: HUMEDAD RELATIVA Y TEMPERATURA

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	F_0
Humedad relativa	9,514	2	4,757	28,33
Temperatura	17,263	4	4,316	25,71
Dentro	1,343	8	0,168	
Total	28,120	14		

Al analizar con mayor cuidado la suma de cuadrados de las tres tablas se observa lo siguiente:

a) La suma de cuadrados total es igual en las tres tablas.
b) La suma de cuadrados debido tanto a la humedad relativa como a la temperatura medida en los Andeva individuales, fue obtenida en forma simultánea e independiente en el Andeva de dos factores.
c) La suma de cuadrados debido a la humedad relativa y a la temperatura no es igual a la suma de cuadrados total, por lo tanto hay cierta cantidad de variación que no es explicada por estos dos factores, la cual necesariamente debe ser atribuida a otros factores no identificados por lo que se denomina variación residual. De modo que la suma de cuadrados residual en el Andeva de dos vías se obtiene por sustraerle a la suma de cuadrados total la suma de cuadrados debido a los dos factores estudiados.

$$SCTotal - SCHumedad - SCTemperatura = SC\,Residual$$

$$28{,}120 - 9{,}514 - 17{,}263 = 1{,}343$$

Conociendo la magnitud de variación que es producida por los factores no identificados es fácil comprobar para el caso de los Andeva de un factor, que la variación residual incluye la variación no identificada y la variación debido al segundo factor bajo control.

En el caso del Andeva para la Humedad relativa se tiene:

$$SC\,Residual - SCTemperatura = 18{,}607 - 17{,}263 = 1{,}34$$

En el caso del Andeva para la temperatura relativa se tiene:

$$SC\,Residual - SCHumedad = 10{,}857 - 9{,}514 = 1{,}34$$

Ahora debe quedar más claro que el análisis de varianza de dos factores se fundamenta en la descomposición de la varianza total en tres varianzas parciales: la varianza debido al primer factor; la varianza debida al segundo factor y la varianza residual debido a los factores no identificados o no controlados.

7.10. ANÁLISIS DE VARIANZA DE DOS FACTORES CON REPETICIONES

El análisis de varianza se puede utilizar para verificar si los efectos combinados de los factores actuantes son independientes. Si los factores son independientes, sus efectos combinados sólo se suman, pero si ellos interactúan, el efecto combinado no es aditivo y la respuesta de la variable medida aumenta o disminuye más de lo que cabría esperar con la simple suma de los efectos individuales de cada factor. Veamos los ejemplos siguientes, tomados de la química, para ilustrar el concepto de interacción. Si dos soluciones se mezclan y la temperatura que resulta es un valor intermedio al presentado por las temperaturas de las soluciones antes de la mezcla, se dice que el efecto de ambas sustancias es aditivo. Sin embargo si se une ácido sulfúrico con una solución acuosa, la mezcla resultante alcanza una temperatura con un valor muy por encima de la temperatura de las dos soluciones antes de mezclarse. En este caso se dice que el efecto de ambas soluciones sobre la temperatura es sinérgico, vale decir que el efecto es superior a la suma de los efectos individuales. Contrariamente, en el caso de otras mezclas (Ej. urea + cloruro de amonio) ocurre una disminución de la temperatura, mucho mayor a lo esperado. En este caso se dice que los efectos son antagónicos.
En este punto es importante resaltar que la importancia de valorar el efecto de la interacción radica en que las conclusiones finales que se pueden obtener de un análisis de varianza están condicionadas a que los factores sean independientes o interactúen. Veamos un ejemplo de esta situación. Se midió el tiempo en horas que tarda en aliviarse el dolor articular en las rodillas en pacientes que fueron sometidos a la acción combinada de aspirina y una crema analgésica. Los resultados se muestran en la tabla 7.26:

Tabla 7.26. Efecto combinado de dos factores sin interacción.

	Tiempo de alivio (Horas)		
	Con aspirina (a_1)	Sin aspirina (a_1)	Promedio
Con crema (b_1)	1	6	3,5
Sin crema (b_2)	4	7	5,5
Promedio	2,5	6,5	

Los valores promedios indican que al aplicar la aspirina se reduce el tiempo en conseguir alivio al dolor articular $(2{,}5h < 6{,}5h)$ y que el tiempo de alivio sigue siendo menor tanto al aplicar crema $(1h < 6h)$, como al no aplicar crema $(4h < 7h)$.
Supongamos que los tiempos de alivio en ausencia de aspirina son diferentes, y que resultan contrarios a lo esperado, para lo cual se intercambiaron intencionalmente los valores de la segunda fila de datos, como se muestra en la Tabla 7.27.

Tabla 7.27. Efecto combinado de dos factores con interacción.

	Tiempo de alivio (Horas)		
	Con aspirina (a_1)	Sin aspirina(a_1)	Promedio
Con crema (b_1)	1	6	3,5
Sin crema (b_2)	7	4	5,5
Promedio	4,0	5,0	

El análisis de los resultados muestra que, la relación entre los valores promedios del tiempo de alivio en presencia o ausencia del analgésico se mantiene igual a los presentados en la Tabla 7.25 del primer grupo de datos, es decir que la aspirina disminuye el tiempo en conseguir alivio $(4h < 5h)$. Sin embargo el sentido de esta relación se pierde en los niveles particulares de acción. En efecto, aún cuando el tiempo de alivio sigue siendo menor en presencia de la aspirina y la crema $(1h < 6h)$, el sentido de esta relación cambia en ausencia de la crema, al aumentar el tiempo de alivio que ahora es mayor $(7h > 4h)$.
Igualmente sucede cuando se analiza el segundo factor (crema analgésica). Para los dos grupos de datos se observa que el tiempo promedio en lograr alivio es menor en presencia de la crema que en su ausencia $(3{,}5h < 5{,}5h)$. Sin embargo el sentido de esta relación se pierde en ausencia de la aspirina, al ser mayor el tiempo de alivio con la crema que en su ausencia $(6h > 4h)$.
La conclusión que se puede obtener de los dos grupos de resultados que se acaban de analizar, es la siguiente. Si la respuesta de una variable a los efectos de un factor en cada nivel de acción del otro factor, es consistente en cuanto a las relaciones de magnitud, como es el caso del primer grupo de datos, sólo basta con analizar los valores promedios para inferir conclusiones acerca de las eventuales diferencias entre los distintos niveles de acción del factor considerado, independientemente del otro factor actuante. Por ejemplo, en el caso de la aspirina, los valores promedios indican que el tiempo para conseguir alivio es menor cuando se usa la aspirina, y esto es cierto tanto en ausencia como en presencia de la crema analgésica. En términos estadísticos se dice que los dos factores son independientes y por tanto sus efectos son aditivos.
Si la respuesta de una variables a los efectos de un factor en cada nivel de acción del otro factor no es consistente, como en el caso del segundo conjunto de datos, no basta con analizar los valores promedios para inferir conclusiones acerca de los efectos de los diferentes niveles de acción del factor estudiado, sino que es necesario examinar los resultados para cada nivel de acción del otro factor. Por ejemplo, en el caso de la aspirina en el segundo conjunto de

datos, los valores promedios indican que el tiempo para conseguir alivio es menor cuando se usa la aspirina. Sin embargo, esta es una verdad a medias, porque sólo es cierta cuando se usa crema, pero no es cierta cuando no se usa la crema. Esto significa que el efecto de la aspirina sobre tiempo de alivio depende del nivel de acción del otro factor, es decir de la presencia o ausencia de la crema. En términos estadísticos se dice que los dos factores interaccionan y por tanto sus efectos no son aditivos. Este concepto puede ilustrase gráficamente. En la Figura 7.7A se muestra como responde la variable estudiada (Tabla 7.26), a los cambios del factor B, en los dos niveles de acción del factor A.

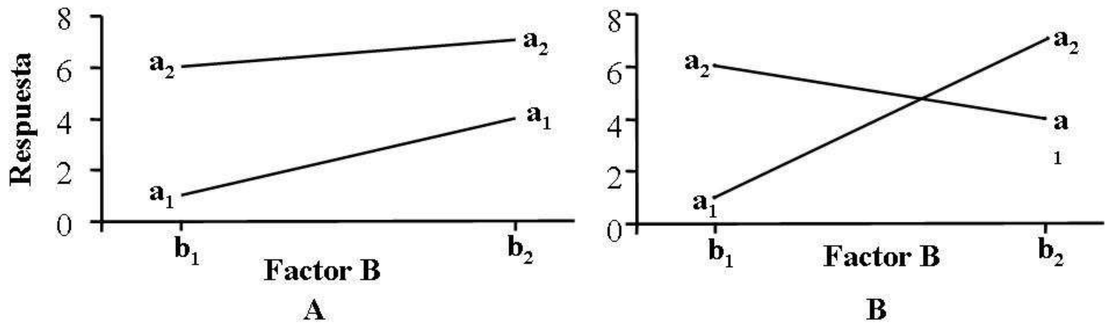


Figura 7.7. Respuesta de una variable en un experimento sin interacción (A) y con interacción (B)

Se observa que las dos rectas se mantienen casi paralelas al cambiar el nivel de acción del factor B, lo que indica ausencia de interacción entre los dos factores actuantes. Si se construye un gráfico similar (Figura 7.7B) para los datos del segundo grupos de datos (Tabla 7.27), se ve como las rectas de respuesta dejan de ser paralelas y se entrecruzan, indicando que hay interacción entre los dos factores.

Cuando hay pocos niveles de acción y son pocos los factores estos gráficos son útiles para descubrir la existencia de interacción, pero su interpretación se complica cuando aumenta tanto el número los factores como de niveles de acción.
Una mejor forma de detectar el efecto de interacción entre dos factores es mediante un análisis de varianza con observaciones repetidas. En este tipo de Andeva en lugar de medir el efecto combinado de cada nivel de acción de cada factor sobre la variable dependiente una sola vez, es necesario hacerlo en más de una ocasión.

7.11 FUNDAMENTO DEL ANDEVA DE DOS FACTORES CON OBSERVACIONES REPETIDAS

A continuación se usará un ejemplo para explicar en forma sencilla el fundamento del Andeva de dos factores con observaciones repetidas. Supóngase que se quiere determinar el efecto que tienen dos factores sobre una determinada variable. El primer factor se clasifica en tres niveles de acción $(a = 3)$, el segundo factor se clasifica en dos niveles de acción $(b = 2)$, y para cada combinación de niveles de los dos factores se efectúan tres mediciones $(n = 3)$. Los resultados se muestran en la tabla siguiente:

Factor A	a_1		a_2		a_3	
Factor B	b_1	b_2	b_1	b_2	b_1	b_2
Observaciones	4	9	2	8	5	10
	7	8	3	7	6	8
	5	8	2	5	4	7

El procedimiento del Andeva de dos factores con observaciones repetidas tiene como objetivo cuantificar la varianza residual, las varianzas debida a cada factor y la varianza debido a la interacción. La razón de usar observaciones repetidas es que con ellas es posible determinar la variación residual producida sólo por el error experimental y no por otras fuentes extrañas de variación como puede ser la interacción entre los factores actuantes. Para que resulte instructivo, la cuantificación de las diferentes varianzas se efectuará en varias etapas, la cuales se desarrollaran a continuación.

7.11.1 Varianza residual

Para calcular la variación residual debida al error experimental es necesario considerar cada subgrupo de datos como muestras independientes y efectuar los mismos cálculos requeridos en el análisis de varianza de un solo factor. Con este primer análisis, además de la varianza residual se obtiene la variación entre los subgrupos. Si la varianza entre los subgrupos no es significativa es porque ninguno de los dos factores tiene efectos significativos sobre la variable estudiada y no es necesario continuar con el análisis.

	Factor A			Factor B			
Observaciones	a_1	a_2	a_3	b_1	b_2	b_3	
1	4	2	5	9	8	10	
2	7	3	6	8	7	8	
3	5	2	4	8	5	7	Total
$\sum_{i=1}^{n} x_i$	16	7	15	25	20	25	108
$\sum_{i=1}^{n} x_i^2$	90	17	77	209	138	213	744
$\sum_{j=1}^{k} \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n_j}$	85,33	16,33	75,00	208,33	133,33	208,30	726,67

Suma de cuadrados total

$$SCT = \sum_{j=1}^{6}\sum_{i=1}^{3}\left(x_{ij}\right)^2 - \frac{\left(\sum_{j=1}^{6}\sum_{i=1}^{3} x_{ij}\right)^2}{N} = 744 - \frac{(108)^2}{18} = 744 - 648 = 96,0$$

Suma de cuadrados entre subgrupos

$$SCESG = \sum_{j=1}^{6}\frac{\left(\sum_{i=1}^{3} x_{ij}\right)^2}{n_j} - \frac{\left(\sum_{j=1}^{6}\sum_{i=1}^{3} x_{ij}\right)^2}{N} = 726,67 - \frac{(108)^2}{18} = 726,67 - 648 = 78,67$$

Suma de cuadrados dentro de subgrupos

$$SCDSG = SCT - SCESG = 96,00 - 78,67 = 17,33$$

Tabla de ANDEVA

Fuente de variación	Suma de cuadrados	Grados de Libertad	Cuadrados Medios	F_o
Entre subgrupos (Entre factores)	78,67	$ab - 1 = 6 - 1 = 5$	15,73	82,38**
Dentro de subgrupos (Residual o Error)	17,33	$ab(n - 1) = 12$	1,44	
Total	96,0	$abn - 1 = 18 - 1 = 17$		

Este primer Andeva finaliza al comprobar si la varianza entre los subgrupos se diferencia significativamente de la variación residual mediante una prueba de hipótesis para las dos varianzas.
Al ser el valor $F_o = 82,38$ mayor que el valor crítico $f_{(0,95;\ 5/12)} = 3,11$ se acepta que la variación entre los subgrupos es significativamente mayor que la variación residual. Esto indica que existe un componente de variación añadida por otros factores distintos al error experimental, que es necesario identificar.

7.11.2 Variación añadida

Parte de la variación de los subgrupos es producida por el efecto independiente o combinado de los factores actuantes. Entonces, es necesario estimar en forma independiente la variación debido a cada factor y a la interacción.

7.11.2.1 Variación añadida debido al factor A

La varianza (cuadrado medio) debido al factor A se obtiene como el cociente entre la suma de cuadrados debido al factor A y sus grados de libertad. Esto requiere considerar que para cada nivel de acción del factor A, existen seis observaciones individuales y no dos grupos de tres mediciones o repeticiones. A continuación se muestra la tabla de datos y las cantidades requeridas para calcular la varianza debida al factor A.

Tabla de datos para el factor A

	Factor A			
Observaciones (n_i)	a_1	a_2	a_3	
1	4,0	2,0	5,0	
2	7,0	3,0	6,0	
3	5,0	2,0	4,0	
4	9,0	8,0	10,0	
5	8,0	7,0	8,0	
6	8,0	5,0	7,0	Total
$\sum_{i=1}^{n} x_i$	41,00	27,00	40,00	108,00
$\sum_{j=1}^{a} \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n_j}$	280,17	121,50	266,67	668,33

Suma de cuadrados del factor A

$$SCFA = \sum_{j=1}^{3} \frac{\left(\sum_{i=1}^{6} x_{ij}\right)^2}{n_j} - \frac{\left(\sum_{j=1}^{3}\sum_{i=1}^{6} x_{ij}\right)^2}{N} = 668,33 - \frac{(108)^2}{18} = 668,33 - 648 = 20,33$$

Cuadrado medio del factor A

$$CMFA = \frac{SCFA}{GL} = \frac{20,33}{2} = 10,12$$

7.11.2.2 Variación añadida debido al factor B

La varianza (cuadrado medio) debido al factor B se obtiene como el cociente entre la suma de cuadrados debido al factor B y sus grados de libertad. Esto requiere asumir que cada nivel de acción del factor B está conformado por nueve observaciones individuales y no tres grupos de tres mediciones o repeticiones. A continuación se muestra la tabla de datos y las cantidades requeridas para calcular la varianza debida al factor B.

Tabla de datos para el factor B

	Factor B		
Observaciones (n_i)	b_1	b_2	
1	4,0	9,0	
2	2,0	8,0	
3	5,0	10,0	
4	7,0	8,0	
5	3,0	7,0	
6	6,0	8,0	
7	5,0	8,0	
8	2,0	5,0	
9	4,0	7,0	Total
$\sum_{i=1}^{n} x_i$	38,0	70,0	108
$\sum_{j=1}^{b} \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n_j}$	560,0	544,4	704,89

Suma de cuadrados del factor B

$$SCFB = \sum_{j=1}^{2} \frac{\left(\sum_{i=1}^{9} x_{ij}\right)^2}{n_j} - \frac{\left(\sum_{j=1}^{2}\sum_{i=1}^{9} x_{ij}\right)^2}{N} = 704,89 - \frac{(108)^2}{18} = 704,89 - 648 = 56,89$$

Cuadrados medio del factor B

$$CMFB = \frac{SCFB}{Gl} = \frac{56,89}{1} = 56,89$$

7.11.2.3 Variación añadida debida a la interacción.

Las sumas de cuadrados de los factores A y B son parte de la suma de cuadrados entre los subgrupos la cual fue calculada en el primer análisis de varianza $(SCESG = 78,67)$. Sin embargo, este valor es un poco mayor que las dos sumas de cuadrados juntas:

$$SCFA + SCFB = 20,33 + 56,89 = 77,22$$

La diferencia entre estas dos cantidades: *SCESG- (SCFA+ SCFB)=78,67- 77,22 =1,45* es una tercera suma de cuadrados que recibe el nombre de suma de cuadrados de la interacción entre los factores A y B.

7.11.3 Partición de la suma de cuadrados total

En resumen se puede ver que la suma de cuadrados total, en el caso de un Andeva de dos factores con observaciones repetidas esta formada por cuatro sumas de cuadrados (Figura 7.9)

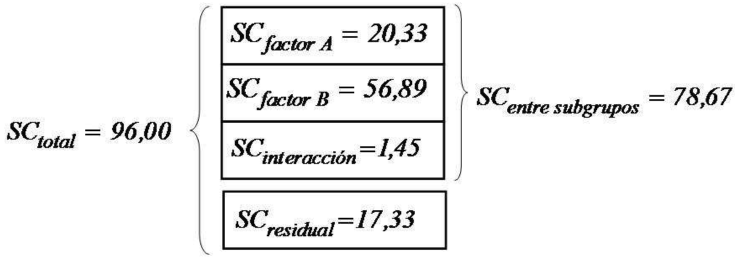


Figura 7.9. Partición de la suma total de cuadrados en un ANDEVA de dos factores con observaciones repetidas.

7.11.4 Tabla de resultados

A continuación se muestra como se tabulan los resultados.

Tabla de ANDEVA

Fuente de variación	Suma de cuadrados	Grados de Libertad	Cuadrados Medios	F_o
Entre subgrupos	78,67	$ab-1=6-1=5$		
a) Factor A	20,33	$a-1=3-1=2$	10,17	7,1*
b) Factor B	56,89	$b-1=2-1=1$	56,89	39,5*
c) Interacción AxB	1,45	$(a-1)(b-1)=(2)(1)=2$	0,73	$0,5^{ns}$
Dentro de subgrupos (Residual o Error)	17,33	$ab(n-1)=12$	1,44	
Total	96,00	$abn-1=18-1=17$		

7.11.5 Significación de las varianzas o cuadrados medios

El Andeva de dos factores finaliza al comprobar si los tres componentes de la varianza añadida (factor A, factor B e interacción) se diferencian significativamente de la varianza residual, mediante una prueba de hipótesis dos varianzas.

7.11.5.1 Significación de la varianza debida al Factor A

Dado que el valor del estadístico de prueba $F_O=7,1$ es mayor que el límite crítico $F_{(0,95\ ;\ 2/12)}=4,75$ se rechaza H_0 y se acepta con un 95% de confianza que al menos uno de los niveles de acción del factor A tiene un promedio diferente o afecta significativamente los valores de la muestra. En este caso se debe efectuar una comparación múltiple de medias, tal como la prueba de Tukey, a fin de determinar cual o cuales medias son diferentes.

7.11.5.2 Significación de la varianza debida al Factor B

Dado que el valor del estadístico de prueba $F_O=39,5$ es menor que el límite crítico $F_{(0,95\ ;\ 1/12)}=4,75$ se rechaza H_0 y se acepta con un 95% de confianza que al menos uno de los niveles de acción del factor B tiene un promedio diferente o afecta significativamente los valores de la muestra. En este caso como sólo son dos niveles de acción, no es necesario efectuar una comparación múltiple de medias, puesto que resulta obvio que las dos únicas medias son diferentes.

7.11.5.3 Significación de la varianza debida a la Interacción

Dado que el valor del estadístico de prueba $F_O = 0{,}50$ es menor que el límite crítico $F_{(0{,}95\ ;\ 2/12)} = 3{,}88$ se rechaza H_0 y se acepta con un 95% de confianza que no existe interacción entre los dos factores actuantes.

7.12 PROCEDIMIENTO GENERAL PARA EL ANDEVA DE DOS FACTORES CON OBSERVACIONES REPETIDAS

A continuación se resolverá un Andeva de dos factores con repeticiones, mediante un procedimiento similar al que se explicó anteriormente, con algunas ligeras modificaciones que le dan mayor formalidad al proceso.

Ejemplo

Un investigador desea evaluar el efecto combinado del pH y el tiempo de exposición sobre la capacidad que tiene un método, que usa tolueno como solvente orgánico, para extraer la enzima β-D-galactosidasa contenida en células de levadura. Con tal fin diseño un experimento en el cual midió la actividad enzimática en tres ocasiones, para tres diferentes valores de pH y tres tiempos de extracción. La actividad enzimática fue medida como la cantidad necesaria para producir un μmol x 10^{-1} de o-nitrofenol (ONP) a partir de un sustrato específico. Los resultados se muestran en la Tabla 7.29.

Tabla 7.29. Actividad de la β-D-galactosidasa para diferentes valores de pH y tiempo de extracción

	μmol/ 10^{-1} de ONP		
Tiempo	**pH 6,0**	**pH 8,0**	**pH 10,0**
5 horas	0,4587	2,5641	1,0137
	0,4937	2,7598	1,0911
	0,3568	1,9945	0,7885
10 horas	0,8578	4,7949	1,8957
	0,9232	5,1608	2,0403
	0,6672	3,7297	1,4745
20 horas	1,0091	5,6411	2,2302
	1,0861	6,0715	2,4004
	0,7850	4,3879	1,7348

El investigador desea determinar si la actividad de la enzima es afectada tanto por el pH del medio como por el tiempo de exposición de los extractos de células de levadura al proceso de extracción.

Este es un problema típico para usar un Andeva de dos factores con repeticiones. Este ejemplo permitirá mostrar el procedimiento utilizado en la resolución de este tipo de análisis.

Procedimiento

a) Hipótesis de investigación:

H_0: *La acidez y el tiempo de extracción no afectan la actividad de la enzima.*

H_1: *La acidez y el tiempo de extracción afectan la actividad de la enzima.*

b) Hipótesis estadísticas

En éste caso, en lugar de plantear las hipótesis sobre las medias poblacionales, es más conveniente plantear las hipótesis sobre los efectos de los factores, En éste caso, el efecto del primer factor (pH) se identifica con la letra griega α, el efecto del segundo factor (tiempo) con la letra griega β, y el efecto de la interacción entre los dos factores (pH x tiempo) como $\alpha\beta$.

b.1) Efecto del pH.

$$H_o: \ \alpha_1 = \alpha_2 = \alpha_3 = 0$$
$$H_1: \text{ al menos uno de los } \alpha_j \neq 0.$$

b.2) Efecto del tiempo.

$$H_o: \ \beta_1 = \beta_2 = \beta_3 = 0$$
$$H_1: \text{ al menos una de las } \beta_i \neq 0.$$

b.3) Efecto de la Interacción.

$$H_0: (\alpha\beta) = 0 \qquad \text{para todo } i, j$$
$$H_1: (\alpha\beta) \neq 0 \qquad \text{para todo } i, j$$

c) Cómputos necesarios para calcular las sumas de cuadrados total, entre subgrupos y residual.

c.1) Organice los datos en una tabla de una única entrada. Utilice las columnas para ubicar los valores de la variable estudiada (ONP) obtenidos para las diferentes combinaciones de los a niveles del factor 1 (pH) y los b niveles del factor B (tiempo). Obtenga las cantidades necesarias para calcular las sumas de cuadrados total (SCT), entre subgrupos (SCESG) y dentro de subgrupos o residual (SCDSG).

n_i	a_1b_1	a_2b_1	a_3b_1	a_1b_2	a_2b_2	a_3b_2	a_1b_3	a_2b_3	a_3b_3	
1	0,4587	2,5641	1,0137	0,8578	4,7949	1,8957	1,0091	5,6411	2,2302	
2	0,4937	2,7598	1,0911	0,9232	5,1608	2,0403	1,0861	6,0715	2,4004	
3	0,3568	1,9945	0,7885	0,6672	3,7297	1,4745	0,785	4,3879	1,7348	Total
$\sum_{i=1}^{n} x_i$	1,3092	7,3184	2,8933	2,4482	13,6854	5,4105	2,8802	16,1005	6,3654	58,41
$\sum_{i=1}^{n} x_i^2$	0,5815	18,1691	2,8398	2,0333	63,5356	9,9307	2,8141	87,9388	13,7452	201,59
$\sum_{i=1}^{n} \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n_i}$	0,5713	17,8530	2,7904	1,9979	62,4301	9,7578	2,7652	86,4087	13,5061	198,08

c.2) Suma de cuadrados total (SCT)

$$SCT = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{abn} = 201{,}59 - \frac{(58{,}41)^2}{27} = 201{,}59 - 126{,}36 = 75{,}23$$

c.3) Suma de cuadrados entre subgrupos (SCESG)

$$SCESG = \sum_{i=1}^{n} \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n_j} - \frac{\left(\sum_{i=1}^{n} x_i\right)}{abn} = 198{,}08 - \frac{(58{,}41)^2}{27} = 198{,}08 - 126{,}36 = 71{,}72$$

c.4) Suma de cuadrados residual o dentro de subgrupos (SCDSG)

$$SCDSG = SCT - SCESG = 75{,}23 - 71{,}72 = 3{,}51$$

d) <u>Cómputos necesarios para calcular las sumas de cuadrados debido a los dos factores</u>.

d.1) Los datos se organizan en una tabla de doble entrada con los totales de cada grupo, en la forma siguiente:

Tiempo	pH 6	pH 8	pH 10	$\sum_{j=1}^{3} x_j$	$\sum_{i=1}^{3}\frac{\left(\sum_{j=1}^{3} x_i\right)^2}{na}$
5 horas	1,3092	7,3184	2,8933	11,521	14,748
10 horas	2,4482	13,685	5,4105	21,544	51,572
20 horas	2,8802	16,101	6,3654	25,346	71,380
$\sum_{i=1}^{3} x_i$	6,638	37,104	14,669	58,411	137,700
$\sum_{j=1}^{3}\frac{\left(\sum_{i=1}^{3} x_i\right)^2}{nb}$	4,8953	152,97	23,909	181,77	

d.2) Suma de cuadrados debido al pH ó factor A (SCFA)

$$SCFA = \sum_{j=1}^{3}\frac{\left(\sum_{i=1}^{3} x_i\right)^2}{nb} - \frac{\left(\sum_{j=1}^{3}\sum_{i=1}^{3} X_i\right)^2}{nab} = 181{,}77 - \frac{(58{,}41)^2}{27} = 181{,}77 - 126{,}36 = 55{,}41$$

d.3) Suma de cuadrados debido al tiempo ó factor B (SCFB)

$$SCFB = \sum_{i=1}^{3}\frac{\left(\sum_{j=1}^{3} x_j\right)^2}{nb} - \frac{\left(\sum_{j=1}^{3}\sum_{i=1}^{3} X_i\right)^2}{nab} = 137{,}70 - \frac{(58{,}41)^2}{27} = 137{,}70 - 126{,}36 = 11{,}34$$

d.4) Suma de cuadrados debido a la interacción (SCI)

$$SCI = SCESG - SCFA - SCFB = 71{,}72 - 55{,}41 - 11{,}34 = 4{,}97$$

f.3) Tabla de Andeva:

Tabla de Andeva

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	F_o
Entre subgrupos	71,72	8	8,97	46,00
a) pH	55,41	2	27,71	142,10
b) Tiempo	11,34	2	5,67	29,08
c) Interacción	4,97	4	1,24	6,36
Residual	3,51	18	0,195	
Total	75,23	26		

e) <u>Zona de aceptación para la hipótesis de igualdad de las varianzas</u>

e.1) pH

$$ZA: \left\{ F \big/ F < f_{[1-\alpha;\ (a-1)/ab(n-1)]} \right\} = \left\{ F \big/ F < f_{[[0,95\ ;\ 2/18\ ]]} \right\} = \left\{ F \big/ F < 3{,}55 \right\}$$

e.2) Tiempo

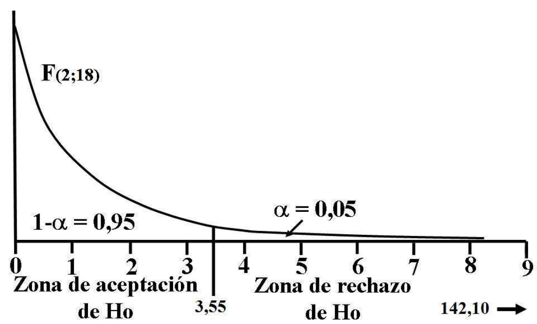
$$ZA: \left\{ F \big/ F < f_{[[1-\alpha;\ (b-1)/ab(n-1)]]} \right\} = \left\{ F \big/ F < f_{[0,95;\ 2/18\ ]} \right\} = \left\{ F \big/ F < 3{,}55 \right\}$$

e.3) Interacción

$$ZA: \left\{ F \big/ F < f_{[1-\alpha;\ (a-1)(b-1)/ab(n-1)]} \right\} = \left\{ F \big/ F < f_{[[0,95;\ 4/18\ ]]} \right\} = \left\{ F \big/ F < 2{,}93 \right\}$$

f) <u>Decisión:</u>

f.1) pH: dado que el valor del estadístico de prueba $Fo = 142{,}10$ es mucho mayor que el límite crítico $F_{(0,95\ ;\ 2/18)} = 3{,}55$ se rechaza Ho y se acepta con un 95% de confianza que al menos en uno de los niveles de pH el promedio de la actividad enzimática es diferente.



f.2) Tiempo: dado que el valor del estadístico de prueba $Fo = 29{,}08$ es mayor que el límite crítico $F_{(0,95;\ 2/18)} = 3{,}55$ se rechaza Ho y se acepta con un 95% de confianza que al menos en uno de los tiempos de extracción el promedio de la actividad enzimática es diferente.

f.3) Interacción: dado que el valor del estadístico de prueba $Fo = 6{,}36$ es mayor que el límite crítico $F_{(0,95\ ;\ 4/18)} = 2{,}93$ se rechaza Ho y se acepta con un 95% de confianza que existe interacción entre los factores: ph y tiempo de extracción.

g) <u>Conclusión:</u>

g.1) El factor pH afecta la conductividad del agua,

g.2) El factor tiempo de extracción afecta la conductividad del agua.

g.3) Existe interacción entre los dos factores

Al ser la interacción significativa es necesario examinar individualmente la respuesta promedio de la variable estudiada, en este caso la actividad enzimática, para cada combinación de los niveles de acción de ambos factores. En esta situación resulta muy útil representar gráficamente las respuestas promedios (Figura 7.10)

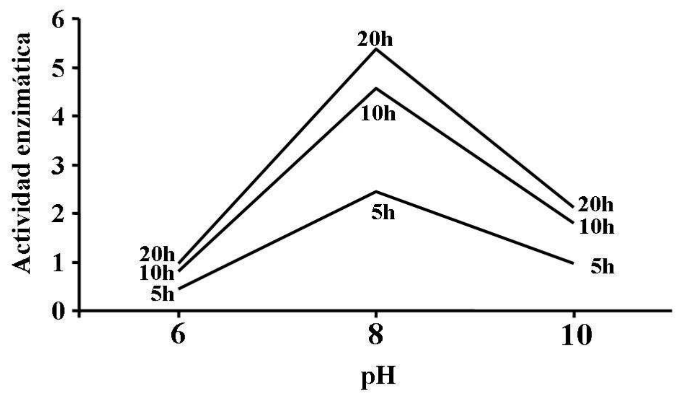


Figura 7.10

La Figura 7.10 muestra que aunque las rectas son paralelas, la actividad enzimática promedio aumenta considerablemente cuando el medio tiene pH 8, especialmente para los dos mayores tiempos de extracción. Esto indica que en este nivel de pH la cantidad de enzima extraída esta muy por encima del valor que se esperaría si los efectos de los dos factores fuesen aditivos. Por el contrario, existe un efecto sinérgico entre los tiempos de extracción y el valor de pH 8.

Puesto que el efecto de los dos factores fue significativo, resulta de interés efectuar una prueba de comparación múltiple de medias, como la prueba de Tukey, para cada factor, con el propósito de descubrir cuales de los niveles de acción de los dos factores determinan una actividad enzimática promedio, significativamente diferente.

7.13 EJERCICIOS

1. ¿Por que en el análisis de varianza se puede someter a prueba una hipótesis de igualdad de medias de dos o más grupos de datos si el análisis de estos datos se basa en una comparación de varianzas?

2. ¿Por que en el análisis de varianza se acepta la hipótesis nula de igualdad de varianzas cuando el valor de la razón varianzas entre grupos/varianza dentro de grupos es menor al valor crítico de F para un dado valor de α y un determinado número de grados de libertad para el numerador y el denominador?

3. A fin de determinar el efecto de la concentración de nitrógeno en el suelo sobre la producción de un cultivo, se sembraron 10 parcelas con la planta estudiada. Cada parcela se fertilizó con una concentración diferente de N_2. Después de 12 semanas, se tomaron aleatoriamente, de cada parcela, 10 plantas y se pesaron. Los datos se examinaron mediante un análisis de varianza y se encontraron diferencias significativas entre los pesos promedios de las plantas de cada parcela. Responda:

a. ¿Por qué a través de un análisis de varianza se pudo determinar que existen diferencias entre los pesos promedios de las distintas parcelas?
b. ¿Explique si la producción promedio de cada parcela es afectada por el tratamiento con nitrógeno?
c. ¿Explique si la varianza del peso promedio entre parcelas es afectada por el tratamiento con nitrógeno?
d. ¿Explique si la varianza del peso de las plantas dentro de cada parcela es afectada por el tratamiento con nitrógeno?

4. Con el propósito de determinar como afectan cuatro catalizadores la concentración de un componente en una reacción química. Se prepararon 20 mezclas de reacción y se formaron cuatro grupos de cinco. A las mezclas de cada grupo se le añadió uno y sólo uno de los catalizadores y se determinó la concentración resultante de la sustancia investigada. Los datos obtenidos se analizaron mediante un Andeva. Se encontró que la varianza entre grupos difiere significativamente de la varianza dentro de grupos. En función de la experiencia anterior se le pide lo siguiente:

a. Formalice simbólicamente en una tabla el diseño del experimento.
b. ¿Que factores determinan la varianza entre y dentro de grupos?
c. ¿Como se interpreta el hecho de que exista diferencia significativa entre las dos varianzas calculadas?

5. Para medir el efecto de una droga sobre la presión sanguínea en una raza de ratones de laboratorio, se eligieron aleatoriamente 15 ratones de una misma camada y se formaron tres grupos de 5 ratones. Cada grupo fue estimulado con una concentración diferente de la droga. Los resultados obtenidos se analizaron mediante un Andeva. Se encontró que la varianza entre grupos difiere significativamente de la varianza dentro de grupos. En función de la experiencia anterior se le pide los siguiente:

a. Formalice simbólicamente en una tabla el diseño del experimento.
b. Desde el punto de vista biológico a que se debe la variación dentro de los grupos y entre los grupos?
c. Desde el punto de vista biológico como se puede interpretar que exista diferencia significativa entre las dos varianzas calculadas.

6. Se quiere conocer si la concentración inicial de un fertilizante afecta el tamaño de las plantas de un determinado cultivo. Para tal fin se fertilizaron cuatro parcelas de terreno con cuatro distintas concentraciones del producto (C1, C2, C3, C4). Después de seis semanas, se midió la altura en cuatro plantas elegidas al azar dentro de cada parcela encontrándose los valores siguientes:

C4 (49,9)*	C1(55,8)	C3 (55,4)	C1(58,2)
C3 (54,2)	C2 (57,0)	C2 (55,3)	C2 (54,5)
C1(58,4)	C4 (52,9)	C1(57,2)	C3 (54,9)
C4 (51,7)	C3(50,1)	C4 (50,0)	C2(56,3)

Entre paréntesis se indica la concentración del fertilizante en mg/l.

a. ¿Hay efecto de la concentración del fertilizante sobre el tamaño de las plantas?

7. Un investigador quiere probar como afectan las dietas ricas en grasas el peso del hígado de patos de cierta especie. Para tal efecto seleccionó cuatro grupos de individuos que se sometieron a cuatro dietas que difieren en el contenido de lípidos. Después de cierto tiempo se determinó el peso del hígado como un tanto por ciento (%) del peso del cuerpo, obteniéndose los resultados siguientes:

Dieta 1	Dieta 2	Dieta 3	Dieta 4
3,96	3,63	3,72	3,93
3,87	3,38	3,81	3,77
4,19	3,47	3,66	4,18
3,58	3,39	3,55	4,21
3,76	3,41	3,51	3,88
3,84	3,55		3,96
	3,44		3,91

a. ¿Cuáles son las hipótesis biológicas a probar?
b. ¿Cuáles son las hipótesis estadísticas a probar?
c. ¿Desde el punto de vista biológico que mide la variación dentro de grupos?
d. ¿Desde el punto de vista biológico que mide la variación entre grupos?
e. ¿Compruebe si las dietas tienen algún efecto sobre el peso promedio?

8. El examen de la movilidad electroforética de las proteínas del suero de diferentes poblaciones de venados, produjo los resultados siguientes:

	Movilidad electroforética (x 10^{-5} cm^2/voltio segundos)	
Población	Media	Desviación
A	2,8	0,07
B	2,5	0,05
C	2,9	0,05
D	2,5	0,05
E	2,8	0,07

Los datos están basados en muestras de 12 individuos. Suponga que la movilidad electroforética es una variable normalmente distribuida. Haga el análisis de varianza y responda lo siguiente:

a. Desde el punto de vista biológico como se puede interpretar que exista diferencia significativa entre las dos varianzas calculadas.
b. Desde el punto de vista biológico ¿Cual es el objetivo de éste análisis de varianza?
c. ¿A que se debe la variación dentro de grupos?
d. ¿A que se debe la variación entre grupos?
e. Determine si las varianza entre poblaciones se diferencia significativamente de la varianza dentro de las poblaciones.

9. En una investigación para determinar el efecto combinado de la hormona estradiol y la droga acetato de nortisterona sobre los niveles un marcador de inflamación como la proteína C reactiva (PCR), en mujeres menopáusicas sanas a las cuales se les había extraído quirúrgicamente los ovarios. Se seleccionaron seis pacientes, las cuales fueron tratadas con 2 mg de estradiol y 1 mg de acetato de noretisterona diario por 12 meses. Los niveles de PCR se midieron mediante un inmunoensayo enzimático en muestras de suero sanguíneo a los 0, 3, 6, 9 y 12 meses de tratamiento. Los resultados se muestran en la tabla siguiente:

Niveles de la proteína C reactiva (mg/dl)				
0 meses	3 meses	6 meses	9 meses	12 meses
0,12	0,58	0,49	0,35	0,39
0,14	0,68	0,65	0,54	0,65
0,27	0,78	0,66	0,60	0,68
0,93	0,87	0,70	0,74	0,78
0,98	0,87	0,87	0,82	0,78
1,20	1,18	0,92	0,95	0,85

a. ¿Tiene el tratamiento algún efecto sobre los niveles de PCR?

10. Un agrónomo intentando determinar el efecto de la concentración de un fertilizante sobre la producción de maíz planificó y efectuó el experimento siguiente: i) escogió cinco concentraciones diferentes del fertilizante, las cuales denominó A, B, C, D y E; ii) seleccionó 5 lotes de terreno (I, II, III, IV y V), cada uno con la misma superficie y ubicados uno al lado del otro. Cada lote lo dividió en cinco parcelas del mismo tamaño y en cada parcela sembró el mismo número de plantas de maíz de cierta variedad. A cada una de las parcelas de cada uno de los lotes le asignó aleatoriamente una concentración de fertilizante y después de cierto tiempo midió la producción de maíz en Kg/Ha. Verifique si las concentraciones de fertilizante tienen efecto sobre la producción de maíz. En la tabla siguiente se presenta el esquema del diseño experimental. Los valores entre paréntesis indican la producción de maíz obtenida para el tratamiento respectivo.

LOTES				
I	II	III	IV	V
D (6843)	A (5670)	E (6633)	C (6636)	B (6367)
B (6685)	C (6557)	D (6534)	B (6263)	A (5976)
A (6116)	E (6455)	B (6141	D (6630)	C (5662)
C (7040)	D (6825)	C (6719)	E (6705)	D (7260)
E (6762)	B (6382)	A (5490)	A (5556)	E (7010)

11. Complete el siguiente cuadro de análisis de la varianza.

Fuente de Variación	Suma de Cuadrados	Grados de libertad	Cuadrados Medios	F
Tratamientos		6	4,73	
Error	37,8			
Total		20		

a. Saque conclusiones con respecto a los tratamientos.

12. En un estudio sobre las condiciones fisicoquímicas del agua de la parte alta del río Chama se seleccionaron varios puntos de muestreo a diferentes alturas sobre el nivel de mar. En cada estación se midió la temperatura del agua en 10 ocasiones entre los meses de julio y noviembre de 1996. Los resultados se presentan en la tabla siguiente. ¿Varía significativamente la temperatura promedio del agua con la altura sobre el nivel de mar (m.s.n.m)?. En la tabla siguiente se muestran los valores de temperatura (°C) del agua del Río Chama a diferentes alturas sobre el nivel del mar (m.s.n.m)

	Temperatura del agua (°C)							
Altitud (m.s.n.m)	3500	3325	3125	3000	2700	2300	1900	1750
Mayo	6,6	6,8	7,5	9,0	10,1	12,5	16,2	17,0
Junio	7,7	8,8	9,3	9,6	10,2	12,2	14,6	17,8
Julio	8,6	10,1	10,3	10,6	10,6	15,5	18,5	20,5
Agosto	10,1	12,1	12,3	12,3	12,8	15,3	18,7	20,8
Septiembre	10,2	11,6	11,4	11,7	12,5	14,8	16,2	18,3
Octubre	8,5	9,5	9,6	10,6	11,3	13,0	15,4	17,3
Noviembre	9,4	10,2	10,9	11,4	12,1	13,7	16,0	17,8
Septiermbre	8,2	9,2	9,9	10,0	10,7	12,8	14,4	16,7
Octubre	8,3	9,5	9,9	10,3	10,9	12,7	15,6	17,8
Diciembre	9,9	11,0	10,8	10,7	11,4	14,5	16,5	18,5

13. El experimento siguiente fue diseñado para determinar el valor relativo de cuatro alimentaciones diferentes respecto a la ganancia en peso de un grupo de cochinos. Veinte cerdos con el mismo peso aproximadamente se reparten al azar en cuatro lotes de cinco individuos. A cada lote se le da una alimentación distinta. Analice los datos y diga si las dietas producen diferentes ganancias de peso. En la tabla siguiente se da el peso en Kg de cada uno de los cochinos después de cierto tiempo.

Peso (Kg)			
Dieta A	Dieta B	Dieta C	Dieta D
133	163	210	195
144	148	233	184
135	152	220	199
149	146	226	187
143	157	229	193

14. El experimento siguiente fue diseñado para determinar el efecto de la densidad de siembra (n° de plantas /m²) sobre el rendimiento del maíz (kg/m²). Se sembraron veinte parcelas con maíz y se formaron cuatro grupos de cinco parcelas cada uno. Cada grupo tiene una densidad de siembra diferente. En la tabla siguiente se da el rendimiento en Kg/m² de cada una de las parcelas después de cierto tiempo.

	RendimientoKg/m²			
Parcela N°	20 plantas/m²	30 plantas/m²	40 plantas/m²	50 plantas/m²
1	21,0	19,5	16,3	13,3
2	23,3	18,4	14,8	14,4
3	22,0	19,9	15,2	13,5
4	22,6	18,7	14,6	14,9
5	22,9	19,3	15,7	14,3

Se tiene interés en responder las interrogantes siguientes:

a. ¿La densidad de siembra tienen algún efecto sobre el rendimiento promedio del maíz? La probabilidad de rechazar la hipótesis nula no debe ser mayor al 1%.

b. ¿Cuales son los rendimientos estadísticamente diferentes?

15. Se midió el ancho cefálico (mm) a individuos de cuatro especies de insectos del mismo género. Encontrándose los resultados siguientes.

Ancho cefálico (mm)			
Especie 1	Especie 2	Especie 3	Especie 4
7,67	7,58	8,17	7,08
7,04	7,09	7,54	6,19
7,32	7,12	7,82	6,62
7,46	7,11	7,96	
7,33		7,76	
		7,92	

a. Determine si existe diferencia del ancho cefálico entre las especies.

b. En caso de existir diferencias, investigue cuál o cuales especies difieren.

16. Se midió la concentración de una sustancia contaminante en muestras recolectadas en tres zonas de un río. Los resultados obtenidos se muestran en la tabla siguiente:

Zona	Tamaño de muestra	Media	Desviación
A	5	22,36	0,8961
B	5	21,16	0,6066
C	5	19,08	0,6648

a. Use los datos de la tabla anterior para efectuar un análisis de varianza.

17. En el transcurso de una investigación se determinó electrométricamente el pH del suelo a diferentes profundidades en cinco días consecutivos. En cada profundidad se hicieron cinco mediciones de pH. Los resultados se presentan en la tabla siguiente:

	Valores de pH			
Dia	05 cm	10 cm	15 cm	20 cm
1	4.46	4.49	4.51	4.42
2	4.48	4.46	4.64	4.40
3	4.46	4.42	4.46	4.62
4	4.71	4.49	4.42	4.46
5	4.40	4.50	4.49	4.39

a. ¿Tiene la profundidad algún efecto sobre la concentración de H^+?
b. ¿Cuáles son las fuentes de variación que determinan, para este caso, la varianza dentro de los grupos?
c. ¿Existen otros factores distintos a los mencionados por Ud. en la parte b, para que la varianza entre muestras sea mayor a la varianza dentro de las muestras?
d. ¿Determine si los días tiene efecto sobre la concentración de H^+?

18. Se está investigando el efecto de la concentración inicial de una sustancia A sobre la concentración final del producto (C) en la reacción siguiente: A + B $\rightarrow$ C + D. La concentración final del producto C se determinó en cuatro ocasiones para diferentes concentraciones iniciales de la sustancia reaccionante A obteniéndose los resultados siguientes:

	Concentración final de C (μg/ml)			
Observación N°	Sustancia A Conc. inicial 1	Sustancia A Conc. inicial 2	Sustancia A Conc. inicial 3	Sustancia A Conc. inicial 4
1	28,2	26,0	26,1	21,9
2	26,2	24,1	25,2	29,9
3	27,4	27,8	25,7	20,0
4	25,8	25,9	24,9	21,7

Se tiene interés en responder las interrogantes siguientes:
a. ¿Tiene la concentración inicial de A algún efecto sobre la concentración final promedio de la sustancia C?
b. ¿Cuales son las concentraciones cuyos efectos son diferentes?

19. El propietario de una droguería sospecha que los lotes de materia prima suministrados por distintos proveedores difieren significativamente en el contenido de calcio. Después de recibir cuatro lotes de la materia prima adquiridos a los proveedores dudosos decide comprobar su hipótesis. Con tal fin, seleccionó aleatoriamente cuatro muestras del lote comprado a cada proveedor y determinó el contenido de calcio para cada muestra. Después de realizadas todas las determinaciones obtuvo los resultados siguientes:

	Concentración de calcio (mg/g)			
Muestras	Proveedor 1	Proveedor 2	Proveedor 3	Proveedor 4
1	23,46	23,59	23,51	23,28
2	23,48	23,46	23,64	23,40
3	23,56	23,42	23,46	23,37
4	23,39	23,49	23,52	23,46
5	23,40	23,50	23,49	23,39

a. Determine si es fundada la sospecha del propietario de la droguería

b. ¿Tendrá algún efecto sobre los resultados el hecho de haber realizado los análisis en diferentes días?

20. Se efectúo un experimento para determinar la estabilidad de la vitamina C en el concentrado de jugo de naranja reconstituidos que fueron almacenados hasta por una semana en un refrigerador. Se probaron tres tipos comerciales de concentrados utilizando tres lapsos diferentes (número de días que permanecen en el refrigerador). Los resultados, expresado mg/l de ácido ascórbico, fueron los siguientes:

C[**0**(53,6)]	A[**0**(54,2)]	A[**3**(42,8)]	A[**7**(48,8)]	C[**0**(52,0)]	A[**3**(49,4)]
A[**7**(40,4)]	B[**0**(48,0)]	B[**0**(48,4)]	C[**3**(49,6)]	C[**3**(47,0)]	B[**3**(44,0)]
C[**7**(45,2)]	B[**3**(42,4)]	C[**3**(48,0)]	C[**0**(52,5)]	A[**0**(52,6)]	C[**7**(43,4)]
A[**3**(53,2)]	A[**3**(49,2)]	B[**7**(42,0)]	C[**7**(47,6)]	B[**7**(43,2)]	B[**0**(49,6)]
C[**0**(51,8)]	B[**7**(49,2)]	B[**3**(48,8)]	A[**0**(46,5)]	A[**0**(49,8)]	B[**3**(44,0)]
B[**7**(44,0)]	C[**7**(48,5)]	A[**7**(42,7)]	B[**0**(56,0)]	C[**3**(48,2)]	A[**7**(47,6)]

En los datos de la tabla anterior la letra identifica el tipo de concentrado, el primer número (resaltado en negritas) indica el total de días de refrigeración y el segundo número (entre paréntesis) expresa la cantidad de ácido ascórbico medido en el concentrado. Organice los datos de modo que, utilizando un nivel de significación $\alpha = 0{,}05$, pueda someter a prueba las hipótesis siguientes:

a. No existe diferencia en los contenidos de ácido ascórbico entre los tipos de concentrado

b. No existe diferencia en el contenido de ácido ascórbico debido a los tiempos de refrigeración.

c. No existe interacción entre los tipos de concentrado y el tiempo de refrigeración.

d. Dada la situación que encuentre que son significativos los efectos de alguno de los dos factores, determine mediante una prueba de comparación múltiple de medias, cuales son los niveles de acción del factor que se diferencian significativamente.

21. En un laboratorio de análisis se desean comparar tres técnicas para la determinación de Glucosa. Con este propósito, se obtiene suero de tres pacientes escogidos al azar entre los paciente que asisten diariamente a consulta. El suero de cada paciente se fracciona en 9 alícuotas iguales, las cuales se dividieron aleatoriamente en tres grupos de tres. A cada una de las tres alícuotas de suero de un mismo grupo se le midió la concentración de

glucosa usando sólo una de las tras técnicas. Este procedimiento se repitió para cada paciente. Los veintisiete valores medidos se presentan en la tabla siguiente.

Paciente	Técnica	Glucosa	Paciente	Técnica	Glucosa	Paciente	Técnica	Glucosa
C	2	0,81	A	2	0,95	B	3	1,02
A	3	1,07	A	2	0,99	C	1	0,79
B	1	0,92	B	1	0,86	B	3	1,03
C	2	0,79	C	3	0,98	A	3	1,13
C	2	0,86	A	1	0,98	C	1	0,84
B	2	0,90	C	3	1,01	B	2	0,91
B	1	0,92	C	3	1,01	A	1	0,98
A	3	1,07	B	2	1,01	B	3	1,07
A	2	1,01	C	1	0,80	A	1	1,01

Se desea responder las interrogantes siguientes:

a. ¿Hay diferencias entre las 3 técnicas analizadas?

b. ¿Es significativa la variación entre pacientes?

c. ¿Existe interacción entre las técnicas y los pacientes?

d. ¿Cuál o cuales técnicas tienen valores promedios significativamente diferentes?

e. ¿Por que es irrelevante responder la pregunta: cuál o cuales de los pacientes presentan promedios significativamente diferentes?

22. Un investigador quiere determinar si la eficiencia para extraer benceno a partir de compuestos orgánicos mediante un nuevo método es afectada por diferentes niveles de temperatura y/o presión a los cuales se puede someter el proceso. Para probar esto, decidió realizar 12 extracciones de benceno, a cuatro temperaturas diferentes (T_1, T_2, T_3 y T_4) y a tres valores de presión (P_1, P_2, y P_3). Cada ensayo lo realizó a un determinado valor de temperatura y a un determinado valor de presión. Los % de rendimiento después de realizados los 12 ensayos fueron los siguientes:

$T_1P_1 = 88$	$T_1P_3 = 96$	$T_2P_2 = 92$	$T_2P_3 = 97$
$T_3P_2 = 95$	$T_4P_3 = 96$	$T_4P_1 = 85$	$T_1P_2 = 94$
$T_4P_2 = 91$	$T_3P_1 = 87$	$T_2P_1 = 89$	$T_3P_3 = 98$

a. Organice los datos y determine si la temperatura y/o la presión afectan el rendimiento en el proceso de extracción del benceno.

b. Determine, de ser necesario, cuales valores de temperatura y/o de presión están afectando en forma diferente el rendimiento del proceso.

23. Suponiendo que en el análisis de varianza del ejercicio anterior, se hubiese considerado únicamente el factor presión, construya la nueva tabla de Andeva a partir de las sumas de cuadrados y grados de libertad obtenidos en dicho ejercicio.

24. En un estudio sobre el tiempo de cura del dolor de cabeza por la acción de dos factores: la aspirina y el café. Se eligieron al azar tres pacientes para cada una de las cuatro

combinaciones posibles de los factores estudiados y se les midió el tiempo en horas que se produce el alivio del dolor de cabeza. Los datos se muestran en la tabla siguiente:

Tratamiento	**Tiempo (horas)**		
Aspirina + café	1,6 (individuo 1)	1,6 (individuo 2)	1,8 (individuo 3)
Aspirina – café	3,6 (individuo 4)	3,2 (individuo 5)	3,8 (individuo 6)
Café – aspirina	4,3 (individuo 7)	4,1 (individuo 8)	3,6 (individuo 9)
Sin café y sin aspirina	4,3 (individuo 10)	5,1 (individuo 11)	4,9 (individuo 12)

a. Hay interacción entre el café y la spirina

25. Se quiere probar la eficacia de un somnífero estudiando posibles diferencias de la misma por el sexo de los sujetos. Se eligen al azar dos grupos de insomnes varones y otros dos de mujeres y tanto para los hombres como para las mujeres se suministra a un grupo el somnífero y a otro un placebo y se mide, en minutos, el tiempo que tardan en dormirse. Los resultados son:

Hombre		Mujer	
Placebo	Somnífero	Placebo	Somnífero
30	35	50	42
50	32	35	30
45	30	46	15
47	25	25	18
38	30	32	23

Responda lo siguiente:

a. ¿El somnífero es realmente efectivo?.

b. ¿Afecta el somnífero de forma diferente a las personas según su sexo?

c. ¿Hay interacción entre los dos factores?

26. En dos cultivos de algas, desarrollados en dos volúmenes diferentes de dos medios gaseosos: uno rico en nitrógeno y el otro rico en hidrógeno, se determinó el contenido de nitrógeno por mg de tejido utilizando el método de Micro-Kjeldhal. Los resultados fueron los siguientes:

Contenido de nitrógeno (mg de N_2/mg de tejido)

40 ml		**60 ml**	
$CO_2 + N_2$	**$CO_2 + H_2$**	**$CO_2 + N_2$**	**$CO_2 + H_2$**
0,613	0,317	0,514	0,216
0,668	0,069	0,586	0,104
0,581	0,069	0,310	0,214
0,624	0,259	0,888	0,104
0,460	0,627	1,210	0,104
0,501	0,380	1,199	0,104

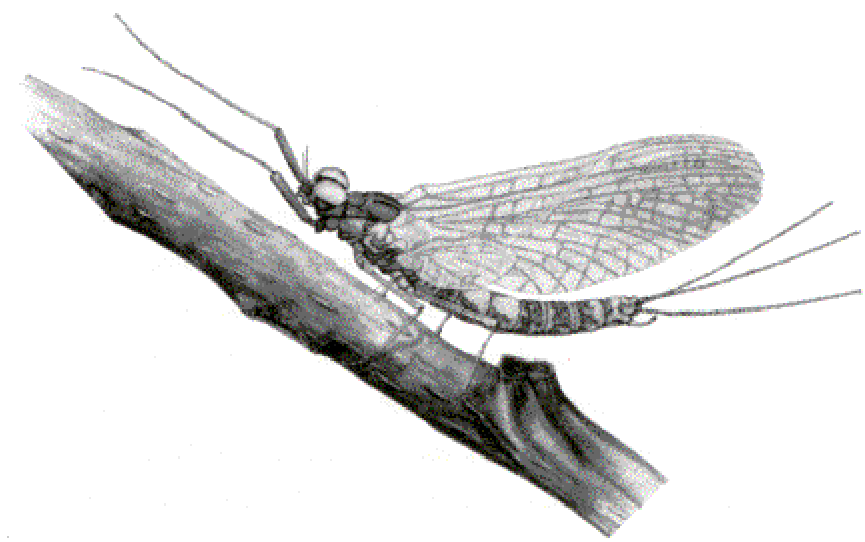
a. Establezca las hipótesis estadísticas probar.

b. Establezca las hipótesis científicas a probar.
c. Determine si el volumen y el tipo de medio interactúan.
d. Identifique las fuentes de variación más importantes que originan la variación residual.
e. Identifique las fuentes que originan la variación entre los factores estudiados.
f. Determine para cada factor, cuales son los valores promedios diferentes.

27. Un farmacéutico a cargo de tres farmacias, escoge al azar dos días de un mes, y contabiliza el total de ventas por día para cada sucursal. Los datos los presenta en una tabla como la siguiente.

	Ingresos Bs x 10^6/día		
Sucursal	Mes 1	Mes 2	Mes 3
1	2,5 3,4	2,8 3,4	2,7 3,2
2	2,9 5,0	2,8 5,7	2,4 5,2
3	1,8 3,8	1,9 4,0	1,8 3,9

Responda lo siguiente:
a. ¿Hay diferencia significativa entre las ventas de las sucursales?
b. ¿Hay diferencia significativa entre los días?



Adulto de Ephemeroptera (Insecta)