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 7.1.1.

Tabla 7.1.1: 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	$\overline{X}_{.1}$	$\overline{X}_{.2}$	$\overline{X}_{.3}$	$\overline{X}_{.4}$

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

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. Puede parecer un contrasentido el 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 que quería determinar si diferentes contenidos de proteínas en la dieta afectaban el crecimiento corporal en los ratones de laboratorio, inicio su experimento seleccionando aleatoriamente ratones de una misma edad, sexo y raza. Sin embargo la condición más importante a controlar era el tamaño de los ratones, que debía ser mas o menos parecido, para evitar que su variabilidad 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. La variable que utilizó para estimar el tamaño corporal fue el peso, debido a que esta característica es muy fácil de medir, además de que estima el crecimiento corporal con mayor exactitud y precisión que lo hace la talla del cuerpo. Después de pesar cada individuo, para asegurarse que no había diferencias significativas en el tamaño, decidió comprobar estadísticamente si los pesos promedios de cada grupo estimaban la misma media poblacional. Los valores de peso inicial de los ratones se presentan en la Tabla 7.2.1.

Tabla 7.2.1: 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 de peso, es importante puntualizar algunos aspectos relacionados con la naturaleza de los datos.

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 equivocarse al tomar una decisión es un poco mayor al 40%.
Dada esta complicación, lo que se necesita es un método que determine simultáneamente si entre las medias muestrales existen diferencias significativas. Este método fue el propuesto por R. Fisher y se fundamenta a partir del cálculo de dos varianzas: la varianza dentro de los grupos y la varianza entre grupos.

Varianza dentro de los grupos
Para el caso del ejemplo, 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 7.1.1)

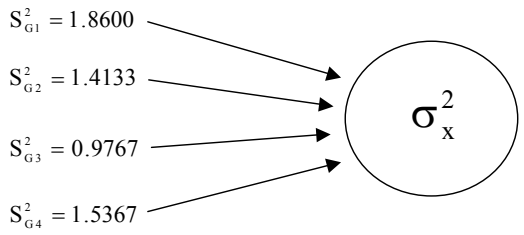


Figura 7.1.1

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}).

$$S^2_{DG} = 1.4467$$

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 $\sigma^2_{\bar{x}} = \dfrac{\sigma^2_x}{n}$ (Figura 7.1.2)

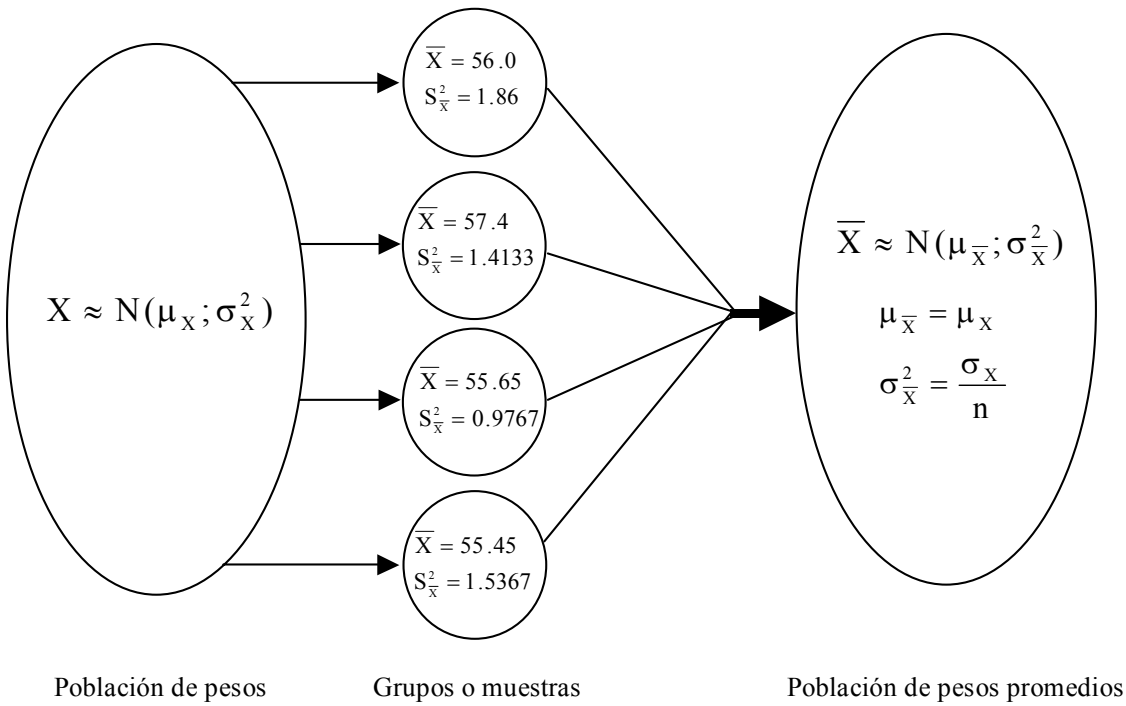


Figura 7.1.2

De la relación $\sigma^2_{\bar{x}} = \dfrac{\sigma^2_x}{n}$ 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 ($S^2_{\bar{X}}$).

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

donde:

$\bar{X}_j$ = el promedio de cada grupo;

$\bar{\bar{X}}$ = el promedio total o media de todas las medias de grupo

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

Al aplicar la ecuación anterior se tiene:

$$S_{\bar{x}}^2 = \frac{12602.39 - \frac{(224.5)^2}{4}}{3} = 0.7742$$

Puesto que $S_{\bar{X}}^2 = \frac{S_x^2}{n}$ se tiene que $S_{\bar{x}}^2 = nS_{\bar{X}}^2$; por lo tanto

$$S_{\bar{X}}^2 = 4(0.7742) = 3.0967$$

Esta varianza que representa la variación de peso que existe entre los grupos se denomina varianza entre grupos (S_{EG}^2) y es una segunda estimación de la varianza poblacional σ_X^2.

$$S_{EG}^2 = 3.0967$$

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:

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

$$H_1 : \sigma_{EG}^2 \neq \sigma_{DG}^2$$

Nivel de significación: $1 - \alpha = 0.95$

Estadístico de prueba:

$$F_o = \frac{S_{EG}^2}{S_{DG}^2} = \frac{3.0967}{1.4467} = 2.1406$$

Zona de aceptación de Ho:

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

$$ZA:\left\{ F \,/\, f_{\left(0.025\,;\,3/12\right)} \le F \le f_{\left(0.975\,;\,3/12\right)} \right\}$$

$$ZA:\left\{ F \,/\, 0.0698 \;\le\; F \;\le\; 4.47 \right\}$$

Decisión: Como el estadístico de prueba F_o = 2.1406 se encuentra dentro de la zona de aceptación de Ho, se concluye que los datos no aportan evidencia para rechazar Ho, 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. Los pesos finales se muestran en la tabla 7.2.2.

Tabla 2.2.2: 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 2.2.2.

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 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 3, por lo tanto la varianza dentro de grupos (S^2_{DG}) no debe haber cambiado:

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

$$S_p^2 = \frac{(4-1)\ 1.86 + (4-1)\ 1.4133 + (4-1)\ 0.9767 + (4-1)\ 1.5367}{12} = 1.4467$$

$$S_p^2 = S_{DG}^2 = 1.4467$$

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. Esto fue lo que precisamente se hizo.

Cálculo de la varianza entre grupos

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

$$S_{EG}^2 = S_x^2 = nS_{\overline{X}}^2 = \frac{n\sum_{j=1}^{k}\left(\overline{X}_j - \overline{\overline{X}}\right)^2}{k-1}$$

Si por alguna causa la separación entre las medias de los grupos aumenta, se incrementará la diferencia del término $\left(\overline{X}_{.j} - \overline{\overline{X}}\right)^2$ 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.

$$S^2_{\overline{X}} = \frac{\sum_{j=1}^{k}\left(\overline{X}_{.j} - \overline{\overline{X}}\right)^2}{k-1} = \frac{\sum_{j=1}^{k}\left(\overline{X}_{.j}\right)^2 - \frac{\left(\sum_{j=1}^{K}\overline{X}_{.j}\right)^2}{k}}{k-1} = \frac{13638.49 - \frac{(233.5)^2}{4}}{3} = 2.6408$$

$$S^2_{EG} = S^2_{\overline{X}} = nS^2_{\overline{X}} = 4(2.6408) = 10.5633$$

Prueba de hipótesis

Hipótesis:

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

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: $F_o = \frac{S^2_{EG}}{S^2_{DG}} = \frac{10.5633}{1.4467} = 7.3018$

Zona de aceptación de Ho:
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}.

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

$$ZA : \left\{ F / F \leq f_{\left(0.95\ ;\ 3/12\right)} \right\}$$

$$ZA : \left\{ F / F \leq 3.5 \right\}$$

Decisión:
Como el valor observado $F_o = 7.3018$ se encuentra fuera de la zona de aceptación de Ho, se concluye que los datos aportan evidencia para rechazar Ho, 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 7.3.1.

Tabla 7.3.1: 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 $\left(\sigma^2\right)$.

1) Varianza dentro Grupo

$$S_{DG}^2 = \frac{\left(n_1 - 1\right)\ S_{G1}^2 + \left(n_2 - 1\right)\ S_{G2}^2 + \left(n_3 - 1\right)\ S_{G3}^2}{n_1 + n_2 + n_3 - 3} = \frac{(4)\ 0.07375 + (4)\ 0.02258 + (4)\ 0.02587}{12} =$$

$$= \frac{0.4888}{12} = 0.04073$$

2) Varianza entre grupos

$$S_{EG}^2 = nS_{\overline{X}}^2 = n\frac{\sum_{j=1}^{3}\left(\overline{X}_j - \overline{\overline{X}}\right)^2}{k-1} = 5\frac{(0.040712)}{2} = \frac{0.20356}{2} = \ 0.10178$$

3) Varianza total

$$S_T^2 = \frac{\sum_{j=1}^{N}\left(X_j - \overline{\overline{X}}\right)^2}{N-1} = \frac{0.69236}{14} = 0.04945$$

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.

$$S_{EG}^2 + S_{DG}^2 = 0.10178 + 0.04073 = 0.14251 \quad >> \quad S_T^2 = 0.04945$$

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.

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

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:

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

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, para no complicar mucho la explicación, 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:

$$\mathrm{S}_{\mathrm{EG}}^2 = \frac{SCEG}{k-1} = \frac{0.20356}{2} = 0.10178$$

$$S_{DG}^2 = \frac{SCDG}{N-k} = \frac{0.4888}{12} = 0.04072$$

$$S_T^2 = \frac{SCT}{N-1} = \frac{0.69236}{14} = 0.04945$$

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

$$SCEG + SCDG = SCT$$

$$0.20356 + 0.48880 = 0.69236$$

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 7.4.1

Tabla 7.4.1: Ordenación de valores muestrales en el cálculo de un Andeva

Observaciones	K Muestras					
1	1	2	3	.	k	
2	x_{11}	x_{12}	x_{13}	.	x_{1k}	
3	x_{21}	x_{22}	x_{23}	.	x_{2k}	
4	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	$\sum_{i=1}^{n_1} x_{i1}$	$\sum_{i=1}^{n_1} x_{i2}$	$\sum_{i=1}^{n_1} x_{i3}$	.	$\sum_{i=1}^{n_1} x_{ik}$	$\sum_{j=1}^{k}\sum_{i=1}^{n_k} x_{ij}$
Medias	$\overline{X}_{.1}$	$\overline{X}_{.2}$	$\overline{X}_{..3}$	.	$\overline{X}_{.k}$	$\overline{\overline{X}}$

siendo:

x_{ij} = una observación o valor cualquiera $\left\{\begin{array}{l} i = 1, 2, 3, ..., n_j \\ j = 1, 2, 3, ..., k \end{array}\right\}$

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

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

$\overline{X}_{.j} = \dfrac{\sum_{i=1}^{n_j} x_{i.}}{n_j}$ = media de la j-ésima muestra

$\overline{\overline{X}} = \dfrac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{N}$ = *media* de todas las observaciones

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 anterior.

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

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

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

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).

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

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

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: CASO GENERAL

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

1. Hipótesis: $H_o: \mu_{x1} = \mu_{x2} = \mu_{x3} = ... = \mu_{xk}$

   H_1 : al menos una de las μ_{xj} es diferente
2. Se establece el nivel de significación (1-α) para la aceptación de H_o. En caso de no especificarse, se considera 1-α = 0.95.
3. Se define el estadístico de prueba a usar: $F_o = \frac{S^2_{EG}}{S^2_{DG}} = \frac{CMEG}{CMDG}$

4. 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:

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

Sumas de cuadrados:

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

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

$$SCDG = SCT - SCEG$$

Cuadrados medios:

$$CMEG = \frac{SCRG}{k\text{-}1} \qquad CMDG = \frac{SCDG}{k(n\text{-}1)}$$

5. Se construye la tabla resumen del análisis de varianza (Tabla 7.5.1)

Tabla 7.5.1: 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		

6. 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 7.5.2.

Tabla 7.5.2: 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: al menos una de las μ_{x_j} 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. Se establece el nivel de significación para la aceptación de H_o : $1-\alpha = 0.95$

c. Se define y calcula el estadístico de prueba: $F_o = \dfrac{S^2_{EG}}{S^2_{DG}} = \dfrac{CMEG}{CMDG}$

d. Se efectúan los cómputos necesarios para calcular las sumas de cuadrados, los cuadrados medios y el valor de F (Tabla 7.5.3)

Tabla 7.5.3: Cálculos necesarios para el Andeva del ejemplo 7.3.

Paciente N°	60 min.	90 min.	120 min.	150 min.	Gran total
Media	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

Cálculo de las Sumas de cuadrados.

$$SCT = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}\right)^2 - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N} = 123.5486 - \frac{(57.22)^2}{32} = 123.5486 - 102.3165 = 21.2321$$

$$SCEG = \sum_{j=1}^{k}\frac{\left(\sum_{i=1}^{n_j} x_{ij}\right)^2}{n_j} - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N} = 118.9791 - \frac{(57.22)^2}{32} = 118.9791 - 102.3165 = 16.6626$$

$$SCDG = SCT - SCEG = 21.2321 - 16.6626 = 4.5695$$

Cálculo de los Cuadrados medios.

$$CMEG = \frac{SCEG}{k\text{-}1} = \frac{16.6626}{3} = 5.5542 \qquad CMDG = \frac{SCDG}{N\text{-}k} = \frac{4.5695}{28} = 0.1632$$

e. Se construye la tabla resumen del análisis de varianza (Tabla 7.5.4.)

Tabla 7.5.4: Tabla de Andeva para el ejemplo 7.3.

Fuente de variación	**Suma de cuadrados**	**Grados de Libertad**	**Cuadrados Medios**	**F_o**
Entre grupos **(Entre tratamientos)**	16.6626	3	5.5542	34.03
Dentro de grupos **(Residual o Error)**	4.5695	28	0.1632	
Total	21.2321	31		

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

$$ZA: \left\{ F \,/\, F < f_{\left[1-\alpha\ ;\ k-1 / N-k\right]} \right\} = \left\{ F \,/\, F < f_{\left[0.95\ ;\ 3/28\right]} \right\} = \left\{ F \,/\, F < 2.95 \right\}$$

g. Decisión: como el valor del estadístico de prueba Fo = 34.03 es mucho mayor que el límite crítico (f = 2.95), se rechaza Ho, 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.

h. 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 7.5.5.

Tabla 7.5.5: 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: $H_o: \mu_{x_1} = \mu_{x_2} = \mu_{x_3} = ... = \mu_{x_k}$

H_1: al menos una de las μ_{x_j} 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) Se establece el nivel de significación para la aceptación de H_o : $1-\alpha = 0.99$

c) Se define el estadístico de prueba: $F_o = \dfrac{S^2_{EG}}{S^2_{DG}} = \dfrac{CMEG}{CMDG}$

d) Se efectúan los cómputos necesarios para calcular las sumas de cuadrados, los cuadrados medios y el valor de F (tabla 7.5.6)

Tabla 7.5.6: Cálculos necesarios para el Andeva del ejemplo 7.4.

	Quelante 1	**Quelante 2**	**Quelante 3**	**Quelante 4**	**Gran total**
Media	82	78,33	81	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

Cálculo de las Sumas de cuadrados.

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

$$SCEG = \sum_{j=1}^{k}\frac{\left(\sum_{i=1}^{n_j} x_{ij}\right)^2}{n_j} - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N} = 76828.6 - \frac{(960)^2}{12} = 76826.6 - 76800 = 28.6$$

$$SCDG = SCT - SCEG = 54.0 - 28.6 = 25.40$$

Cálculo de los Cuadrados medios

$$CMEG = \frac{SCEG}{k\text{-}1} = \frac{28.60}{3} = 9.53 \qquad CMDG = \frac{SCDG}{N\text{-}k} = \frac{25.40}{8} = 3.175$$

e) Se construye la tabla resumen del análisis de varianza (Tabla 7.5.6).

Tabla: 7.5.6: 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		

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

$$ZA: \left\{ F \,/\, F < f_{\left[1-\alpha\ ;\ k-1/N-k\right]} \right\} = \left\{ F \,/\, F < f_{\left[0.99\ ;\ 3/8\right]} \right\} = \left\{ F \,/\, F < 7.59 \right\}$$

g) Decisión: como el valor del estadístico de prueba F_o = 3.0 es menor al límite crítico (f = 7.59), se acepta Ho.

h) 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:

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

El Andeva finaliza cuando se acepta H_o, pero si se rechaza H_o y se concluye que al menos una de las medias μ_j es diferente, es muy frecuente que se quiera conocer 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*. La diferencia básica entre los dos tipos de comparación estriba en con las pruebas *a priori* son muy pocas las comparaciones que deben efectuarse, para algunos autores no más de tres, por lo tanto deben ser planificadas antes de efectuar el experimento y decidir de antemano que medias se van a comparar. Las pruebas *a posteriori* se realizan una vez obtenidos los resultados y sólo si H_o 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 σ_x^2; sí de esta población se extraen pares de muestras todas con el mismo tamaño n y calculamos sus medias $\overline{X}_j$, es posible originar una nueva variable $\Delta\overline{X} = \overline{X}_1 - \overline{X}_2$ que se denomina diferencia de medias muestrales, que por la propiedad reproductiva de la distribución normal sabemos que también se distribuye normalmente con una media esperada $\mu_{(\overline{X}_1 - \overline{X}_2)} = 0$ y una varianza esperada $2\sigma_x^2$. 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 $H_o : \mu_{x1} = \mu_{x2}$ con un nivel de significación $1\text{-}\alpha = 0.95$. Bajo esta situación todos aquellos pares de media con diferencias muy grandes permitirán rechazar H_o, pero esto ocurrirá sólo en el 5% de los casos. En otras palabras, si escogiéramos aleatoriamente las medias muestrales, la probabilidad de rechazar H_o 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_o 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 numerosas 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:

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_{x_x} - \mu_{x_2}\right)}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$$

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

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

La zona de aceptación de H_o es:

$$ZA : \left\{ T / T < t_{[1-\alpha\ ;\ \text{N-k}]} \right\}$$

Esta ZA se puede presentar en su forma equivalente:

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

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

$$\text{Sí } \left(\overline{X}_1 - \overline{X}_2\right) > t_{[1-\alpha\ ;\ \text{N-k}]}\sqrt{\frac{2CMDG}{n}} \text{ se rechaza } H_o$$

Esta diferencia es el menor valor que puede existir entre dos medias para aceptar H_o, si la diferencia es mayor se rechaza H_o. 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:

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

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 7.6.1.

Tabla 7.6.1: Contenido de Ácido fólico (µg/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) Se establece el nivel de significación para la aceptación de H_o : $1-\alpha = 0.99$

c) Se define el estadístico de prueba: $RV = \frac{S_{EG}^2}{S_{DG}^2} = \frac{CMEG}{CMDG}$

d) Se efectúan los cómputos necesarios para calcular las sumas de cuadrados, los cuadrados medios y el valor de F (Tabla 7.6.2.)

Tabla 7.6.2.: 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

Cálculo de las Sumas de cuadrados.

$$\text{SCT} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}^2 - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N} = 976761.27 - \frac{(4313.70)^2}{20} = 46360.89$$

$$\text{SCEG} = \sum_{j=1}^{k} \frac{\left(\sum_{i=1}^{n_j} x_{ij}\right)^2}{n_j} - \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right)^2}{N} = 973942.32 - \frac{(4313.70)^2}{20} = 43541.94$$

$$\text{SCDG} = \text{SCT - SCEG} = 46360.89 - 43541.94 = 2818.95$$

Cálculo de los Cuadrados medios

$$\text{CMEG} = \frac{\text{SCEG}}{\text{k-1}} = \frac{43541.94}{3} = 14513.98 \qquad \text{CMDG} = \frac{\text{SCDG}}{\text{N-k}} = \frac{2818.95}{16} = 176.18$$

e) Se construye la tabla resumen del análisis de varianza (Tabla 7.6.3)

Tabla 7.6.3: Tabla de Andeva para el ejemplo 7.5.

Fuente de variación	**Suma de cuadrados**	**Grados de Libertad**	**Cuadrados Medios**	**F_o**
Entre grupos **(Entre tratamientos)**	43541.94	3	14513.98	82.38
Dentro de grupos **(Residual o Error)**	2818.95	16	176.18	
Total	46360.89	19		

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

$$ZA: \left\{ F \,/\, F < f_{\left[1-\alpha \;;\; k-1/N-k\right]} \right\} = \left\{ F \,/\, F < f_{\left[0.99 \;;\; 3/16\right]} \right\} = \left\{ F \,/\, F < 5.29 \right\}$$

g) Decisión: como el valor del estadístico de prueba Fo = 82.38 es mucho mayor que el límite crítico (f = 5.29), se rechaza Ho, 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.

h) 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.

$$MDS = t_{[1-\alpha \;;\; \text{N-k}]}\sqrt{\frac{2CMDG}{n}} = t_{[0.99 \;;\; 16]}\sqrt{\frac{2(176.184)}{5}} = 2.583\ (8.395) = 21.68$$

Se prepara una tabla con las diferencias a probar, las medias deben estar ordenadas en un sentido creciente o decreciente (Tabla 7.6.3.)

Tabla 7.6.4: 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_o 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.
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 debe sacar conclusiones sobre las diferencias de medias que no fueron previamente seleccionadas, pues esto aumentaría el número de comparaciones y la probabilidad de equivocarse al contrastar H_o aumenta considerablemente.

7.6.2. Prueba de Tukey o de la Diferencia Verdaderamente Significativa (DVS). Este método se puede usar 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:

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

El valor q_α = estadístico de Tukey, se encuentra en una tabla cuyos argumentos de entrada son k y N-k. Siendo k el número total de muestras de involucradas en el Andeva; N-k los grados de libertad con los cuales se calculó el CMDG; n es el tamaño de las muestras y α = probabilidad de rechazar Ho 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.

$$DVS = q_{\alpha[k;N-k]}\sqrt{\frac{CMDG}{n}} = q_{0.01[4;16]}\sqrt{\frac{CMDG}{n}} = 5.19\sqrt{\frac{176.18}{5}} = 30.81$$

Se prepara una tabla con las diferencias a probar, las medias deben estar ordenadas en un sentido creciente o decreciente (Tabla 7.6.4.).

Tabla 7.6.5: 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_o 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 Ho.
L2 – L3 = 46.62 es mayor a 30.81. Se rechaza Ho.
L2 – L1 = 43.86 es mayor a 30.81. Se rechaza Ho.
PN – L3 = 116.48 es mayor a 30.81. Se rechaza Ho.
PN – L1 = 113.72 es mayor a 30.81. Se rechaza Ho.
PN – L2 = 69.86 es mayor a 30.81. Se rechaza Ho.

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,

$$\mu_1 = \mu_2 = \mu_3 = \mu$$

En la Figura 7.7.1 se muestra gráficamente la situación planteada.

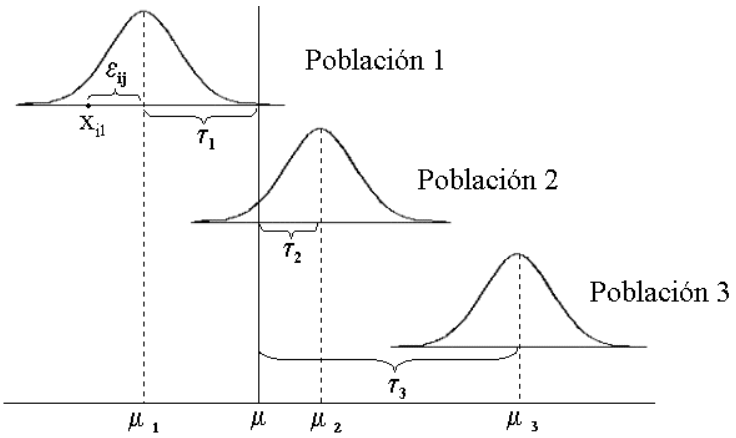


Figura 7.7.1**:** Esquema de los efectos de tratamientos (τ_j) y los errores aleatorios (ε_{ij}) presentes en un Andeva de efectos fijos.

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,

$$\boldsymbol{x_{ij} = \mu + \tau_j + \varepsilon_{ij}}$$

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.

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 ($\boldsymbol{\varepsilon_{ij}}$) son aleatorios y se distribuyan normal e independientemente con una media $\boldsymbol{\mu = 0}$ y una misma varianza $\boldsymbol{\sigma^2}$. 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".*

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 7.7.1 y 7.7.2 se muestra como el efecto multiplicativo afecta las varianzas de los grupos. En la Tabla 7.7.1 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 7.7.1: 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 7.7.2: 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 7.7.2 cada valor de los grupos 2 y 3 se multiplicó por un factor constante. La varianza de estos dos grupos aumentaron 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.

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.

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 7.7.2.

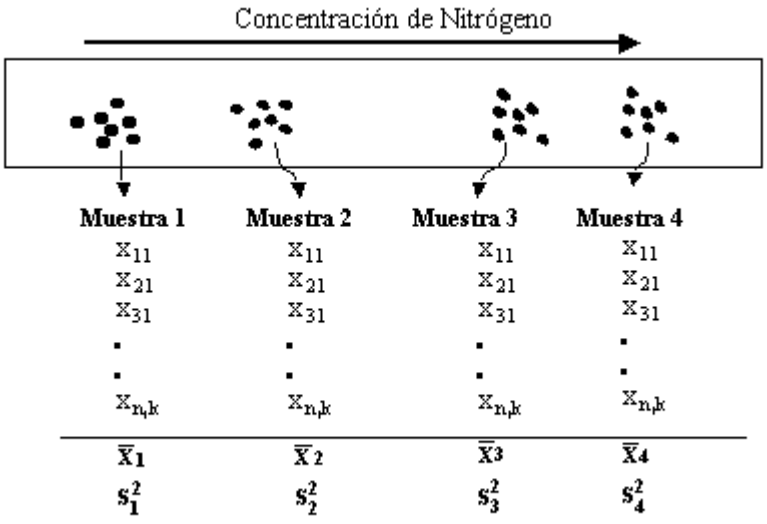


Figura 7.7.3: Gradiente en la distribución del contenido de nitrógeno del suelo.

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.

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 7.7.3.

Tabla 7.7.3: 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 7.7.4.

Tabla 7.7.4: 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.

$$\overline{X}_A = 5.4 \qquad \overline{X}_B = 11.0 \qquad \overline{X}_C = 217.6 \qquad \overline{X}_D = 231.4$$

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 7.7.5) separados de los tratamientos C y D (Tabla 7.7.6)

Tabla 7.7.5: 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 7.7.6: 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.

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 TRANSFORMACIONES

Por lo general, la experiencia indica que para la mayoría de los datos biológicos el no cumplimiento de los supuestos anteriores no es de importancia y el Andeva puede efectuarse sin mayores problemas. Sin embargo, hay ocasiones en las cuales no es posible obviar el incumplimiento de dichos supuestos. En estos casos se tienen dos alternativas.

Una vía es recurrir a la estadística no paramétricas y usar métodos equivalentes al Andeva, como las pruebas de Kruskall-Wallis (una vía) o de Friedman (dos vías). Este tipo de estadística no requiere de suposiciones previas acerca de la distribución de los datos. Sin embargo, cuando se cumplen los supuestos, aunque sea en forma aproximada, el Andeva es mucho más potente que las pruebas no paramétricas para verificar diferencias significativas entre las medias poblacionales. La otra solución es la de transformar los datos de tal forma que los nuevos valores cumplan con los supuestos. Veamos como funciona la transformación con un grupo de datos artificiales. En la Tabla 7.8.1 se muestran tres grupos de datos no transformados. Los valores de los grupos B y C son el resultado de multiplicar los valores de A por un factor de 2 y 3 respectivamente.

Tabla 7.8.1: Datos ficticios no transformados

	Valores originales (x)		
	A	B	C
	4	8	12
	6	12	18
	7	14	21
	5	10	15
	3	6	9
	8	16	24
Promedio	5,5	11	16,5
Varianza	3,5	14	31,5

Bajo esta situación se violan dos de los supuestos del modelo del Andeva, por un lado existe un efecto multiplicativo, el cual incumple la condición de aditividad; por otro lado las varianzas de los grupos son muy diferentes. La aplicación de una transformación logarítmica puede resolver estos dos problemas (Tabla 7.8.2).

Tabla 7.8.2: Datos ficticios transformados

	Valores transformados (Log x)		
	A	B	C
	0,6021	0,9031	1,0792
	0,7782	1,0792	1,2553
	0,8451	1,1461	1,3222
	0,6990	1,0000	1,1761
	0,4771	0,7782	0,9542
	0,9031	1,2041	1,3802
Media	0,7174	1,0184	1,1945
Varianza	0,0252	0,0252	0,0252

El efecto multiplicativo se convierte en un efecto aditivo puesto que el log xy es equivalente a log x + log y. Por otro lado las varianzas se hacen más parecidas. En el caso del ejemplo son exactamente iguales por tratarse de datos artificiales.

El paso siguiente es efectuar el análisis de varianza con los datos transformados. En este punto es importante reflexionar sobre el concepto de transformación, el cual es difícil de aceptar porque da la impresión que se quiere ver lo que se desea y no lo que es. Pero esta es una idea equivocada que posiblemente surge de nuestra costumbre de ver las relaciones cuantitativas en una escala lineal o aritmética. La transformación no es sino un cambio en la escala de observación, que da una perspectiva distinta y permite detectar relaciones que no se observaban en la escala original. Vamos a ilustrar esta idea con un ejemplo no matemático. Supóngase un viajero que desea trasladarse por primera vez desde un sitio A hacia otros dos sitios B y C. En la Figura 7.8.1 se muestra el trayecto de la carretera entre los tres puntos.

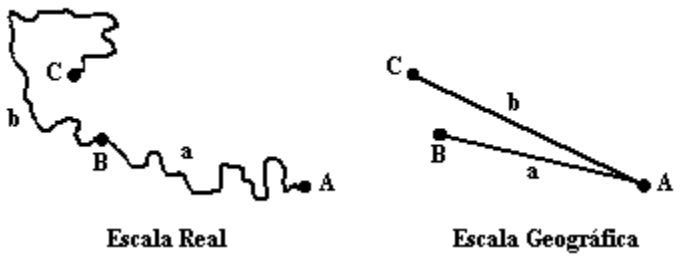


Figura 7.8.1

Para el viajero el punto C esta tan lejos del punto B como éste del punto A. Esto es verdad bajo la perspectiva de la situación que él debe resolver. El traslado entre los tres puntos se debe hacer a través de una carretera. Los accidentes del terreno determinan que las distancias a recorrer sean mayores que las distancia verdaderas. Si el viajero, se cambia a una escala de observación mayor, por ejemplo desde un avión o sobre un mapa, donde se observa simultáneamente la ubicación de los tres puntos, se dará cuenta que el punto C está mucho más cerca de B y A de lo que parecía al viajar por la carretera, tal y como se muestra en la Figura 7.8.1. Las dos situaciones son verdaderas y la conclusión que obtuvo de ambas dependieron de la perspectiva o el nivel de la escala en la cual se colocó.
Los datos biológicos pueden producir situaciones caracterizadas por falta de aditividad, de heterogeneidad de las varianzas y no normalidad de los datos. Los tres tipos de problemas pueden ser resueltos mediante algunas transformaciones estándar, las cuales veremos a continuación.

Transformación logarítmica.
Esta transformación se produce por la conversión de los datos originales en logaritmos, usualmente se utilizan logaritmos decimales. Cuando existen valores menores a 1, se puede usar el log (x+1) para evitar trabajar con cantidades negativas. La transformación logarítmica ayuda a resolver situaciones de falta de aditividad e independencia de los datos, como hemos visto anteriormente. También es muy útil cuando existe dependencia de la varianza con respecto al valor de la media. Es decir, que a mayores valores de las medias le corresponden mayores varianzas. Considérese, por ejemplo, el caso del número de presas consumidas por un depredador. Esta relación varía desde cero (ninguna presa consumida), hasta valores extremadamente grandes cuando un solo depredador consume muchas presas, teóricamente este número puede ser infinitamente grande puesto que no hay límite para el número de presas consumidas. Si el registro de la relación presa/depredador se efectúa a lo

largo del tiempo, es posible que en aquellos fechas con escasez de presas el numero promedio de esa relación sea bajo y también su varianza. Por el contrario, en las épocas con abundancias de presas, el promedio de la relación presa/depredador será grande y consecuentemente su varianza también será grande. Esta situación se puede clarificar graficando la desviación estándar vs el promedio. En la Figura 7.82 se muestra el promedio y la desviación estándar del número de presas consumidas por individuo de la trucha Arco iris en ocho fechas diferentes, para valores no transformados (arriba) y valores transformados a logaritmos (abajo).

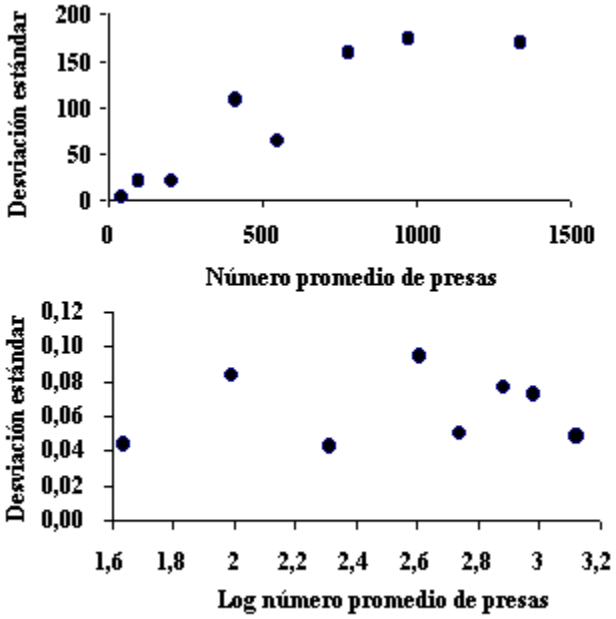


Figura 7.8.2: Dispersión del número de presas consumidas en una escala lineal (arriba) y una escala logarítmica (abajo)

En la parte superior del gráfico se observa que existe una relación aproximadamente lineal entre la desviación estándar y el promedio de presas consumidas por individuo. En la parte inferior se puede ver que la transformación a los logaritmos naturales eliminó la dependencia de la varianza con la media, es decir que las varianzas se hicieron homogéneas.

Muchas veces no es la homogenización de las varianzas lo que determina una transformación logarítmica, sino la necesidad práctica de disminuir las diferencias de magnitud que pueden existir en un conjunto de datos. Por ejemplo, los cultivos de bacterias presentan un crecimiento exponencial, que es más apropiado representarlo en una escala logarítmica que en una escala aritmética, dada la naturaleza no lineal de este proceso.

Transformación raíz cuadrada

Los resultados de muchos experimentos se expresan como el número de veces que ocurre un resultado en un tiempo determinado o en un espacio dado. Por ejemplo: número de partículas desintegradas en una unidad de tiempo; número de electrones emitidos en una unidad de tiempo; número de glóbulos por campo; número de casos de una enfermedad en

un año; número de anímales por unidad de área; número de bacterias por unidad de volumen y número de plantas por unidad de longitud. Usualmente, la distribución de este tipo de resultados se ajusta al modelo de Poción, por lo que sus varianzas y medias son muy similares. Esta falta de independencia de la varianza compromete los resultados del Andeva. Afortunadamente, la transformación de los datos en sus raíces cuadradas puede resolver este problema. En la Tabla 7.8.3 se expone un ejemplo, en el cual se aplicó la transformación raíz cuadrada.

Tabla 7.8.3: número de ninfas por hoja (datos originales y transformados) en dos fechas diferentes durante el desarrollo de cierto cultivo.

a. Datos originales

N° ninfas (x)	Fecha 1	Fecha 2
0	13	0
1	27	1
2	28	5
3	18	9
4	9	13
5	4	16
6	1	18
7		14
8		10
9		7
10		4
11		2
12		1
13		1
Σfx	199	606
Σfx^2	581	4206
Σf	100	101
Media	1.99	6.00
Varianza	1.87	5.70
Razón de Varianza	3.05^{***}	
$F_{(0.05;\ 99/100)}$	1.87	

b. Datos transformados

N° ninfas $(\sqrt{x})$	Fecha 1	Fecha 2
0,000	13	0
1,000	27	1
1,414	28	5
1,732	18	9
2,000	9	13
2,236	4	16
2,449	1	18
2,646		14
2,828		10
3,000		7
3,162		4
3,317		2
3,464		1
3,606		1
Σfx	127,17	242,20
Σfx^2	199	606
Σf	100	101
Media	1,27	2,40
Varianza	0,38	0,25
Razón de Varianza	0.67^{ns}	
$F_{(0.05;\ 99/100)}$	1.87	

Los resultados de la Tabla 7.8.3ª muestran que la segunda fecha tiene una varianza significativamente mayor que la primera. En la Tabla 7.8.3b se observa que una vez aplicada la transformación $\sqrt{\mathbf{x}}$ no hay diferencias significativas entre las varianzas. La transformación homogenizó las varianzas de las dos muestras.

La $\sqrt{\mathbf{x}}$ también puede usarse para transformar datos porcentuales, cuando el intervalo de variación se encuentra entre un 0 y un 20 por ciento. Sí el intervalo va de 80 a 100 por ciento, los porcentajes deberán restarse de 100 antes de la transformación.

Transformación Angular (Arco-seno).

Esta transformación se utiliza para valores expresados en porcentajes o proporciones. Este tipo de datos por lo general se distribuye siguiendo el modelo binomial. Como sabemos las distribuciones binomiales se caracterizan porque la varianza es función de la media.

Media = μ = np Varianza = σ^2 = npq = μq

La Figura 7.8.3 muestra como las varianzas de distintas distribuciones binomial tienden a ser mayores para valores intermedios de las medias y son menores para valores pequeños o grandes de la media.
De la condición anterior se vislumbra que distribuciones de datos con medias diferentes pueden ser asimétricas y con varianzas diferentes. La transformación arco seno puede solucionar esta situación, puesto que al aplicarse alarga los extremos de la distribución y angosta la parte central.

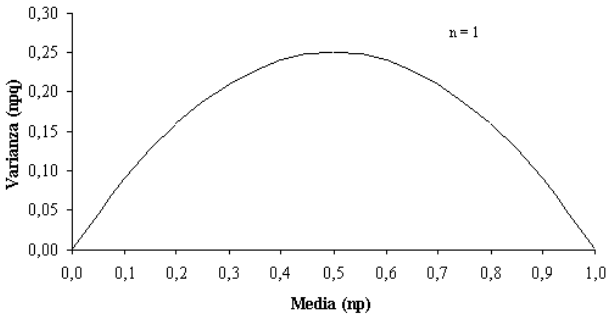


Figura 7.8.3. Distribución de la varianza para varias distribuciones binomial con diferentes medias.

Para transformar los datos se obtiene el arco seno (inverso del seno) de la raíz de la proporción ($\text{arcoseno}\sqrt{p}$), siendo p es el valor proporcional de los datos originales (los porcentajes deben dividirse entre 100). Las unidades de los valores transformados son grados o radianes. En la Tabla 7.8.4 se muestra la distribución del número de truchas como un porcentaje del total de presas encontradas en los estómagos de 246 individuos, antes y después de aplicarles la transformación angular.

Tabla 7.8.4. Distribución de la proporción de presas en el estómago de 246 truchas, para los datos originales (% presas) y transformados en el arcoseno $\sqrt{(\%\text{ de presas})/100}$.

% presas	Grados	N° truchas
0	0,000	5
10	18,435	25
20	26,565	59
30	33,211	60
40	39,232	45
50	45,000	20
60	50,768	12
70	56,789	8
80	63,435	6
90	71,565	4
100	90,000	2

La aplicación de la transformación a los % de las presas aproxima la distribución a una normal, tal como lo muestra la Figura 7.8.4.

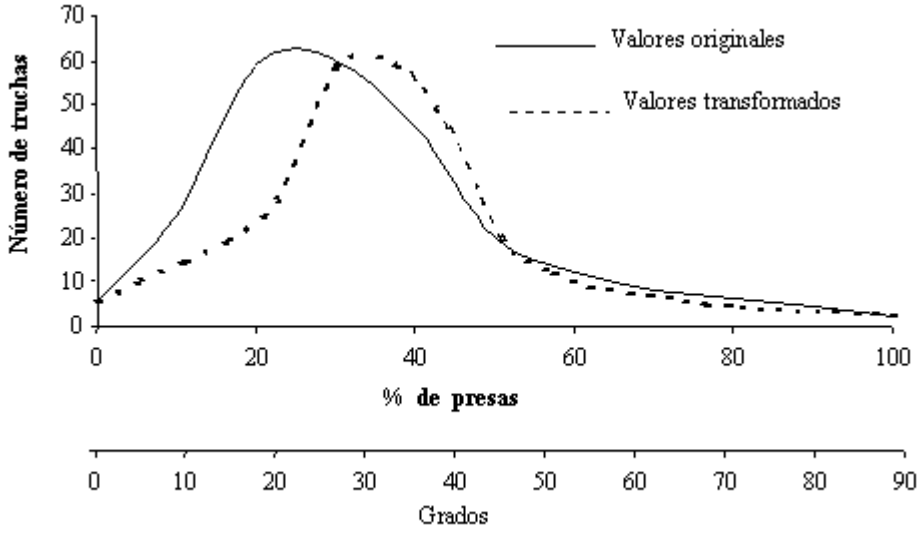


Figura 7.8.4: Distribución del número de presas (en % de presas ó en grados) en el contenido estomacal de 246 truchas.

8. EJERCICIOS

1) 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. La probabilidad de cometer el error tipo I no debe ser mayor al 1%.

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

¿La densidad de siembra tienen algún efecto sobre el rendimiento promedio del maíz.?

2) Utilice los datos de la tabla siguiente para determinar si el valor promedio del ancho cefálico (mm) de alguna de las cuatro especies de insectos es diferente.

	Ancho cefálico (mm)			
Individuo N°	***Squallidus sp.***	***Afligidus sp.***	***Chavitensis sp.***	***Linaronuss sp.***
1	7.67	7.58	8.17	7.08
2	7.04	7.09	7.54	6.19
3	7.32	7.12	7.82	6.62
4	7.46	7.11	7.96	
5	7.33		7.76	
6			7.92	

3) Un investigador quiere probar como afectan las dietas ricas en grasas el peso del higado. Para tal efecto seleccionó cuatro grupos de patos de la especie *"Patus donald"* 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
3.42	3.17	3.34	3.64
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

3.1) ¿Cuáles son las hipótesis biológicas a probar?
3.2) ¿Cuáles son las hipótesis estadísticas a probar?
3.3) ¿Desde el punto de vista biológico que mide la variación dentro de grupos?
3.4) ¿Desde el punto de vista biológico que mide la variación entre grupos?
3.5) ¿Compruebe si las dietas tienen algún efecto sobre el peso promedio?

4) Se está investigando el efecto de la concentración inicial de un fertilizante sobre el tamaño de las plantas de un determinado cultivo. Para tal fín se fertilizaron cuatro parcelas de terreno con cuatro distintas concentraciones del producto. Despues de seis semanas, se midió la altura en cuatro plantas elegidas aleatoriamente dentro de cada parcela encontrándose los valores siguientes:

	Concentración inicial fertilizante (mg/l)			
Planta N°	C1	C2	C3	C4
1	58.2	56.3	50.1	52.9
2	57.2	54.5	54.2	49.9
3	58.4	57.0	55.4	50.0
4	55.8	55.3	54.9	51.7

¿Tiene la concentración inicial del fertilizante algún efecto sobre el tamaño promedio de las plantas?

5) Cuatro grupos de ratas 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 A	DIETA B	DIETA C	DIETA D
3.42	3.17	3.34	3.64
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) ¿Desde el punto de vista biológico que mide la variación dentro de grupos?
c) ¿Desde el punto de vista biológico que mide la variación entre grupos?
d) ¿Compruebe si las dietas tienen algún efecto sobre el peso del higado?

6) 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. 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. 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 (28.4)	A (16.7)	E (26.3)	C (26.3)	B (23.6)
B (26.8)	C (25.5)	D (25.3)	B (22.6)	A (19.7)
A (21.1)	E (24.5)	B (21.4	D (26.3)	C (26.6)
C (30.4)	D (28.2)	C (27.1)	E (27.0)	D (32.6)
E (27.6)	B (23.8)	A (14.9)	A (15.5)	E (30.1)

Verifique si las concentraciones de fertilizante tienen efecto sobre la producción de maíz.

7) A fín de 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 fué estimulado con una concentración diferente de la droga. Los resultados obtenidos se analizaron mediante un análisis de varianza. 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:

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

8) En un ensayo para determinar el efecto de la concentracion de nitrógeno en el suelo sobre la producción de un cultivo, se sembraron 10 parcelas con la planta estudiada. Nueve parcelas se fertilizaron con diferentes concentraciones de nitrógeno y una no se trató y sirvió como control. Despues de 12 semanas, se tomaron aleatoriamente, de cada parcela, 10 plantas y se les determinó el peso promedio. El investigador tiene interés en conocer lo siguiente: a) Si la concentración de nitrógeno tiene algún efecto sobre el crecimiento de las plantas y b) Si existen diferencias en cuanto al peso promedio de las plantas de las parcelas tratadas con relación a la parcela control.
   Suponiendo que las dos variables en cuestión se distribuyen normalmente con varianzas homógeneas desconocidas, responda para cada caso los siguiente:

   8.1) Tipo de análisis que debe realizar
   8.2) Las hipótesis estadísticas a probar.
   8.3) Las hipótesis biológicas a probar.
   8.4) Test estadístico

9) El examen de la movilidad electroforética de las proteínas del suero de diferentes poblaciones de venados, dió 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 estan basados en muestras de 12 individuos. Suponga que la movilidad electroforética es una variable normalmente distribuída. Haga el análisis de varianza y responda lo siguiente:

9.1) El examen de la movilidad electroforética de las proteínas del suero de diferentes poblaciones de venados, dió los resultados siguientes:
9.2) Desde el punto de vista biológico ¿Cual es el objetivo de éste análisis de varianza?
9.3) ¿A que se debe la variación dentro de grupos?
9.4) ¿A que se debe la variación entre grupos?
9.5) Determine si las varianza entre poblaciones se diferencia significativamente de la varianza dentro de las poblaciones.

10) 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 y el oxígeno disuelto en el agua en 10 ocasiones entre los meses de julio y noviembre de 1996. Los resultados se presentan en la tabla 1 y 2. ¿Varía significativamente la temperatura promedio del agua con la altura sobre el nivel de mar (m.s.n.m)?.

Valores de temperatura (°C) del agua del Río Chama a diferentes alturas sobre el nivel del mar (m.s.n.m)

Altitud (m.s.n.m)	**3500**	**3325**	**3125**	**3000**	**2700**	**2300**	**1900**	**1750**
JUL 02	6.6	6.8	7.5	9.0	10.1	12.5	16.2	17.0
JUL 18	7.7	8.8	9.3	9.6	10.2	12.2	14.6	17.8
JUL 31	8.6	10.1	10.3	10.6	10.6	15.5	18.5	20.5
AGO 14	10.1	12.1	12.3	12.3	12.8	15.3	18.7	20.8
AGO 29	10.2	11.6	11.4	11.7	12.5	14.8	16.2	18.3
SEP 11	8.5	9.5	9.6	10.6	11.3	13.0	15.4	17.3
SEP 25	9.4	10.2	10.9	11.4	12.1	13.7	16.0	17.8
OCT 10	8.2	9.2	9.9	10.0	10.7	12.8	14.4	16.7
OCT 23	8.3	9.5	9.9	10.3	10.9	12.7	15.6	17.8
NOV 07	9.9	11.0	10.8	10.7	11.4	14.5	16.5	18.5

11) A fín 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 concentracion diferente de N_2. Despues 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 signicativas entre los pesos promedios de las plantas de cada parcela.

11.1) ¿Por qué a través de un análisis de varianza se puudo determinar que existen diferencias entre los pesos promedios de las distintas parcelas?

11.2) ¿Explique si la producción promedio de cada parcela es afectada por el tratamiento con nitrógeno?
11.3) ¿Explique si la varianza del peso promedio entre parcelas es afectada por el tratamiento con nitrógeno?
11.4) ¿Explique si la varianza del peso de las plantas dentro de cada parcela es afectada por el tratamiento con nitrógeno?

12) A fín de medir el efecto de una droga sobre la presión sanguínea en una raza de ratones, se eligieron aleatoriamente 15 ratones de una misma camada y se formaron tres grupos de 5 ratones. Cada grupo fué estimulado con una concentración diferente de la droga. Los resultados obtenidos se analizaron mediante un análisis de varianza. 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:

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

13) 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		

Saque conclusiones con respecto a los tratamientos.

14) Se está investigando el efecto de la concentración inicial de un fertilizante sobre el tamaño de las plantas de un determinado cultivo. Para tal fín se fertilizaron cuatro parcelas de terreno con distintas concentraciones del producto (800, 600, 400 y 200 mg/l). Despues de seis semanas, se midió la altura en cuatro plantas elegidas aleatoriamente dentro de cada parcela encontrándose los valores siguientes:

	Altura de las plantas (cm)			
Planta N°	800 mg/l	600 mg/l	400 mg/l	200 mg/l
1	58.2	56.3	50.1	52.9
2	57.2	54.5	54.2	49.9
3	58.4	57.0	55.4	50.0
4	55.8	55.3	54.9	51.7

¿Tiene la concentración inicial del fertilizante algún efecto sobre el tamaño promedio de las plantas?

15) 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.

	Rendimiento (Kg/m^2)			
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:

15.1) ¿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%.

15.2) ¿Cuales son los rendimientos estadísticamente diferentes?

16) 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 cerdos cada uno. A cada lote se le da una alimentación distinta. 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

Analice los datos y diga si las dietas producen diferentes ganacias de peso.

17) El propietario de una droguería sospecha que los lotes de materia prima suministrados por distintos proveedores difieren significativamente en el contenido de calcio. Despues de recibir cuatro lotes de la materia prima adquiridos a los proveedores sospechosos decide comprobrar su hipótesis. El farmaceútico de la droguería seleccionó aleatoriamente cuatro muestras del lote comprado a cada proveedor y determinó el contenido de calcio para cada muestra. Despues 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

17.1) Determine si es fundada la sospecha del propietario de la droguería
17.2) ¿Tendrá algún efecto sobre los resultados el hecho de haber realizado los análisis en diferentes dias?

18) Con el propósito de determinar como afectan cuatro catalizadores la concentración de un componente en una mezcla reaccionante. Se prepararon 20 mezclas reaccionantes y se formaron cuatro grupos de cinco mezclas. A las mezclas reaccionantes de cada grupo se le añadió uno de los catalizadores y se determinó la concentración de la sustancia investigada. Los resultados obtenidos se analizaron mediante un análisis de varianza. 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:

18.1) Formalice en una tabla el diseño del experimento.
18.2) ¿Cuál modelo de análisis de varianza debe ser utilizado para interpretar los resultados? Razone su respuesta.
18.3) ¿Que factores determinan la varianza entre y dentro de grupos?
18.4) ¿Como se interpreta el hecho de que exista diferencia significativa entre las dos varianzas calculadas?

19) 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 de C (µg/ml) se determinó en cuatro ocasiones para diferentes concentraciones iniciales de A (µg/ml) obteniendose 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:

19.1) ¿Tiene la concentración inicial de A algún efecto sobre la concentración final promedio de la sustancia C?

19.2) ¿Cuales son las concentraciones cuyos efectos son diferentes?

20) Se midió la concentración de una sustancia contaminante en muestras recolectadas en tres zonas del río Albarregas. Los resultados se analizaron mediante un ANDEVA a partir de 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

Complete el análisis de varianza

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

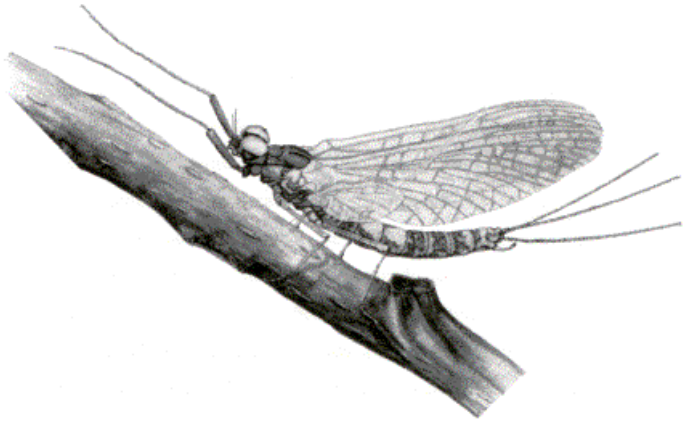
Ancho cefálico (mm)			
Especie 1	Especie 2	Especie 3	Especie 4
7.67	7.58	8.17	7.08
7.04	7.09	7.54	6.19
7.32	7.12	7.82	6.62
7.46	7.11	7.96	
7.33		7.76	
		7.92	

¿Determine cuales especies tendrán un ancho cefálico diferente?

22) Cuatro grupos de ratas 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:

Peso (%)			
Dieta 1	Dieta 2	Dieta 3	Dieta 4
3.42	3.17	3.34	3.64
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

22.1) ¿Cuáles son las hipótesis biológicas a probar?
22.2) ¿Desde el punto de vista biológico que mide la variación dentro de grupos?
22.3) ¿Desde el punto de vista biológico que mide la variación entre grupos?
22.4) ¿Compruebe si las dietas tienen algún efecto sobre el peso del higado?



Adulto de Ephemeroptera (Insecta)