XII. ANALISIS DE VARIANZA (ANOVA)

El análisis de varianza es una prueba paramétrica que nos permite docimar en forma eficiente la hipótesis nula:

$$Ho: \quad \mu_1 = \mu_2 = \mu 3 \ldots\ldots = \mu_k; \qquad k \geq 3$$

en base a k muestras extraídas al azar y en forma independiente de poblaciones normales con varianza común σ^2.

El nombre de análisis de varianza o (ANOVA) se deriva del hecho de que el análisis se basa en la comparación de las varianzas estimadas de las varias causas que las pueden originar. El método se originó en el estudio agronómico, por lo cual se usan términos tales como "bloques" y "tratamiento" (refiriéndose a poblaciones o muestras que se diferencian en términos de semillas, variedades, cantidades utilizadas de fertilizantes, métodos de siembra, ...). Pero hoy en día, este método de análisis se aplica a casi todo tipo de diseño experimental ya sea en ciencias naturales o sociales. La razón para el rápido incremento del uso del análisis de varianza como una herramienta básica se debe principalmente a su gran versatilidad: se puede fácilmente adaptar, dentro de un amplio campo, para evaluar los datos obtenidos dentro de un número grande de experimentos, los cuales tiene varias variables aleatorias contínuas. También, nos ayuda a ver si diferentes datos muestrales clasificados en términos de una sola variable, tienen o no algún significado. Además, nos facilita las compara

ciones de datos muestrales los cuales han sido clasificados de acuerdo a dos o más variables. En cada caso, las respuestas finales comprenden la significancia o la falta de ella, para cada clasificación de acuerdo con una dada variable, así también como los efectos conjuntos de variables (combinación de tratamientos) (9,10,42).

12.1. Clasificación de Acuerdo a una Variable completamente Aleatorizada.

Suponga que tenemos c poblaciones, A_1, A_2, ..., A_c cada una normalmente distribuída con una media igual a μ_i y con una varianza común σ^2. Frecuentemente, estas poblaciones se denominan tratamientos representando, por ejemplo, c variedades diferentes de soya, c distancias de siembra, c regiones diferentes, ..., etc. Estas c poblaciones se asume que tienen una varianza común, ya que se cree que diferentes tratamientos usados para el mismo propósito difieren en las medidas de tendencia central pero no en su dispersión. Más aún, imaginemos que todas estas poblaciones juntas constituyen una gran población con una media μ, llamada "la media de la gran población", la cual se define como:

$$\mu = (1/c) \sum_{i=1}^{c} \mu_i \qquad \{12.1\}$$

Bajo los dos supuestos anteriores queremos probar que todas las medias de los tratamientos son iguales. Es decir que:

$$Ho: \quad \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_c$$

Si Ho es cierta, podemos esperar que cada una de las medias poblacionales sea igual a la gran media, $\mu = \mu_i$. En caso contrario, cada μ_i difiere de μ en la cantidad ω_i, la cual representaría el efecto del tratamiento.

$$\omega_i = \mu_i - \mu \quad ; \quad i = 1, 2, \ldots, c \qquad \{ 12.2 \}$$

El modelo analizado es de la forma :

$$X_{ij} = \mu_i + e_{ij} = \omega_i + \mu + e_{ij} \qquad \{ 12.3 \}$$

La fórmula $\{ 12.3 \}$ implica que cada observación se aleja de la gran media (μ) debido a: 1) Un factor que representa el efecto del tratamiento (ω_i) y 2) Un factor completamente aleatorio ó error (e_{ij}).

La parte central del ANOVA, no importando el tipo de modelo usado está en la división de la suma de cuadrados totales entre diferentes partes que tienen un significado completo. En la primera etapa se organizan los datos tal como se indica en la Tabla 12.1.

En la tabla 12.1, cada observación X_{ij} tiene doble subíndice. El primer sub-índice indica a que muestra o tratamiento pertenece la observación y el segundo numera la observación dentro de una muestra determinada. Por ejemplo, X_{42} es la segunda observación del cuarto tratamiento.

Con la información de la tabla 12.1 podemos calcular dos estadísticos básicos en el ANOVA: 1) La media muestral de cada columna: $\bar{X}_{icol.}$ $(i = 1, 2, \ldots, c)$, donde:

$$\bar{X}_{icol.} = (1/n_i) \sum_{j=1}^{n_i} X_{ij} = (1/n_i)(X_{il} + X_{i2} + \ldots + X_{ij} + \ldots + X_{in_i}) \quad \{ 12.4 \}$$

TABLA 12.1. ANOVA : Datos Muestrales del Modelo Lineal

OBSERVACION	MUESTRA			
	1	2		c
1	X_{11}	X_{21}	.....	X_{c1}
2	X_{12}	X_{22}	.....	X_{c2}
.	.	.		.
.	.	.		..
n_i	X_{1n_1}	X_{2n_2}		X_{cn_c}
TOTALES	$\sum_{j=1}^{n_1} X_{1j}$	$\sum_{j=1}^{n_2} X_{2j}$		$\sum_{j=1}^{n_c} X_{cj}$
n_i	n_1	n_2		n_c

2) La gran media muestral $(\bar{\bar{X}})$

$$\bar{\bar{X}} = (1/N) \sum_{i=1}^{c} \sum_{j=1}^{n_i} X_{ij} = (1/N) \left(\sum_{j=1}^{n_i} X_{1j} + \ldots + \sum_{j=1}^{n_c} X_{cj} \right) \quad \{ 12.5 \}$$

Donde: $N = \sum_{i=1}^{c} n_i = n_1 + n_2 + \ldots + n_c$.

La suma de cuadrados totales $\sum_{i=1}^{c} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$ (de aquí en adelante referida como SCT), se puede descomponer en dos partes fundamentales, una debida a la variación dentro de la muestra (ó sea debido a los errores, la cual denotamos por SCE) y la otra debida a la variación existente de una muestra a otra (ó debido a los tratamientos, la cual denotamos por SCC). Es decir:

$$\sum_{i=1}^{c} \sum_{j=1}^{n_i} (X_{ij}-\bar{X})^2 = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (X_{ij}-\bar{X}_{icol})^2 + \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\bar{X}_{icol}-\bar{X})^2 \qquad \{12.6\}$$

$$\text{SCT} = \text{SCE} + \text{SCC}$$

Lo lógico de esta separación de la suma de cuadrados totales en dos partes diferentes se puede notar intuitivamente. Las observaciones dentro de cada muestra van a mostrar siempre una variabilidad. Las diferencias entre las observaciones individuales puede provenir de dos causas. Diferencias entre individuos que pertenecen a diferentes tratamientos muestrales pueden resultar ya sea por los diferentes tratamientos o por las variaciones aleatorias o por ambos. La suma de cuadrados entre muestras, (SCC), refleja la contribución de las diferencias debidas a los tratamientos así como la contribución de la variabilidad aleatoria de muestra a muestra. Las observaciones individuales, en el mismo tratamiento muestral, sin embargo, pueden diferir una de otras solamente por las variaciones aleatorias, ya que cada individuo dentro del grupo recibe el mismo tratamiento. La suma de los cuadrados

dentro de las muestras (SCE), mide aquellas diferencias intermuestrales que se deben solamente al azar. Por lo consiguiente, en cualquier grupo de c muestras es posible aislar estas dos clases de variabilidad: la suma de cuadrados entre grupos que refleja la variación debida tanto a tratamientos como al azar, más la suma de cuadrados dentro de los grupos que refleja la variación debida solamente al azar (9,18).

Las fórmulas simplificadas para encontrar las diferentes sumas de cuadrados son:

$$SCT = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (X_{ij} - \bar{\bar{X}})^2 = \sum_{i=1}^{c} \sum_{j=1}^{n_i} X_{ij}^2 - N\bar{\bar{X}}^2 \qquad \{ 12.7 \}$$

$$SCE = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\ col})^2 = \sum_{i=1}^{c} \sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{c} n_i \bar{X}_{i\ col}^2 \qquad \{ 12.8 \}$$

$$SCC = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\bar{X}_{i\ col} - \bar{\bar{X}})^2 = \sum_{i=1}^{c} n_i \bar{X}_{i\ col}^2 - N\bar{\bar{X}}^2 \qquad \{ 12.9 \}$$

Dado que SCT = SCC + SCE, tenemos que SCE = SCT - SCC {12.10}

Por consiguiente, usualmente se calculan SCT y SCC y se encuentra por diferencia SCE.

Las fórmulas anteriores se aplican para la suma de cuadrados que tengan o nó igual tamaño de muestra. Si la hipótesis nula que los tratamientos son nulos es cierta, entonces, cada una de estas tres sumas de cuadrados, divididas por sus respectivos grados de libertad, es un estimativo insesgado de la varianza poblacional común σ^2. El núme

ro total de grados de libertad en este modelo es (N-1), ya que SCT se calcula de un total de N observaciones usando la media muestral como un estimativo de μ (se pierde un grado de libertad). La estimación de σ^2 en base de SCC (variación entre las medias de las columnas), tiene (c-1) grados de libertad, ya que hay c medias para comparar (cada columna da una media y tenemos c columnas). El número de grados de libertad para el estimativo de σ^2 basándonos en SCE es (N-c), ya que SCE se obtiene de N observaciones con las c medias muestrales como un estimativo de μ. Tal como en el caso de la suma de cuadrados, los grados de libertad también son aditivos.

$$(N-1) = (c-1) + (N - c) \qquad \{12.11\}$$

Los estimadores de la varianza muestral se suelen denominar los cuadrados medios. En el modelo discutido se puede ver que estamos interesados en comparar dos cuadrados medios: los cuadrados medios entre columnas, CMC = SCC/(c-1) y los cuadrados medios de los errores (SCE), CME = SCE/(N-c). Nótese que ambos cuadrados medios están distribuídos con una distribución chi-cuadrada con (c-1) y (N-c) grados de libertad, respectivamente.

La información necesaria para realizar el ANOVA, se lleva a una tabla denominada la Tabla de ANOVA (Ver tabla 12.2)

Para el modelo a analizar, el par de hipótesis a probar son:

TABLA 12.2. Análisis de Varianza: Modelo Lineal.

CAUSA DE VARIACION	Suma de Cuadrados SC	Grados de libertad δ	Cuadrados medios CM
Entre columnas	SCC	c-1	CMC = SCC/(c-1)
Dentro de columnas (error)	SCE	N-c	CME = SCE/(N-c)
TOTAL	SCT	N-1	

Ho: Todas las medias de los tratamientos son iguales, o sea que $\omega_i = 0$

H_1: No todas las medias de tratamientos son iguales, o sea que no todas los ω_i son iguales a cero.

El estadístico utilizado es el cociente F formado por la razón de CMC y CME. Esto porque, bajo el supuesto de que todos los e_{ij} están normalmente distribuídos en forma independiente con media cero y varianza σ^2, CME debe ser un estimador insesgado de σ^2 ya sea que Ho sea cierta o nó, ya que SCE mide solamente las variaciones aleatorias. Pero CMC es un estimador insesgado de σ^2 si y solo si Ho es cierta. Cuando los efectos de los tratamientos no son nulos, CMC, realmente estima la cantidad $\sigma^2 + \underline{a}$ donde $\underline{a}$ es un sesgo positivo y mide la magnitud de las diferencias entre tratamientos. Bajo estas circunstancias esperemos que $F = (CMC/CME) = 1$, o muy cercano a uno si Ho es cierta, y se espera que esta razón sea significativamente mayor que uno si Ho es falsa. Por consiguiente el estadístico a utilizarse es:

$$F = \frac{SCC/(c-1)}{SCE/(N-c)} = \frac{CMC}{CME} \sim F_{(c-1),\ (N-c)} \qquad (12.12)$$

Puesto que F sigue una distribución de F con (c-1) grados de libertad para el numerador y (N-c) grados de libertad para el denominador, se rechaza Ho al $\alpha\%$ si y sólo si $F > F^*_{\alpha;(c-1),(N-c)}$ (9,18,42).

Ejemplo 12.1. (Muestras de Igual Tamaño). Las edades promedias (en años cumplidos) de agricultores jefes de familia en tres regiones diferentes se presentan en la Tabla 12.3. Se puede concluír que desde el punto de vista de la edad promedia del jefe de familia, las tres regiones no difieren significativamente?

TABLA 12.3. Edad Promedia del Jefe de familia en tres regiones diferentes. 1977

OBSERVACION	REGION 1	2	3	
1	40	20	50	
2	37	27	47	
3	20	31	37	
4	28	32	28	
5	32	40	30	
6	29	27	32	
7	31	21	20	
8	60	29	54	
9	48	31	24	
10	30	26	39	
TOTAL	355	284	361	1000
n_i	10	10	10	30

Solución: 1) Ho: $\mu_1 = \mu_2 = \mu_3$

H_1: Al menos una de las medias es diferente a las otras

2) $\alpha = 0,01$; $N = 30$

3)y 4) Se emplea la prueba F para ANOVA. El estadístico utilizado es F definido por la fórmula { 12.12 } el cual sigue una distribución de F con (c-1) y (N-c) grados de libertad.

5. $\delta_1 = (c-1) = (3-1) = 2$ = Grados de libertad del numerador

$\delta_2 = (N-c) = (30-3) = 27$= Grados de libertad del denominador

Se rechaza Ho al 1% si $F > F^*_{1\%(2, 27)} = 5,49$.

6. Los totales de cada año de los tratamientos se tabulan en la tabla 12.3. Se tiene que: $k = 3$; $n_1 = n_2 = n_3 = 10$

$N = n_1 + n_2 + n_3 = 30$; $\sum_{i=1}^{3} \sum_{j=1}^{10} X_{ij} = 1000$

Por consiguiente:

$\bar{X}_1 = 355/10 = 35,5$; $\bar{X}_2 = 284/10 = 28,4$; $\bar{X}_3 = 361/10 = 36,1$

$$\bar{\bar{X}} = (1/N) \sum_{i=1}^{3} \sum_{j=1}^{10} X_{ij} = (1/30)\ (1000) = 33,33$$

$$\sum_{i=j}^{3} \sum_{j=1}^{10} X_{ij}^2 = 40^2 + 37^2 + ... + 24^2 + 39^2 = 36.344,00$$ (Se deben

de elevar cada una de las 30 observaciones de la Tabla 12.3 al cuadrado)

$$N\bar{\bar{X}}^2 = (30)\ (33,33)^2 = 33.326,67$$

$$\sum_{i=1}^{c} n_i \bar{X}_{i\ col.}^2 = (10)\ (35,5)^2 + (10)\ (28,4)^2 + 10(36,1)^2 = 33.700,2$$

Aplicando las fórmulas {12.7}, {12.8} y {12.9} tenemos que:

SCT = 36.344.00 - 33.326,67 = 3.017,33

SCE = 36.344,00 - 33.700,20 = 2643,80

SCC = 33700,20 - 33.326,67 = 373,53

Nótese que:

SCE = SCT - SCC = 3017,33 - 373,53 = 2643,80

Los grados de libertad asociados con cada una de estas tres sumas de cuadrados son:

$\delta_{SCT} = (N-1) = 30 - 1 = 29$; $\delta_{SCC} = (c-1) = 3-1 = 2$; $\delta_{SCE} = (N-c) = 30-3=27$

Los cuadrados medios (SC / δ_i) son:

CMT = 3017,17 / 29 = 104,05; CME = 2643,80 / 27 = 97,918

CMC = 373,53 / 2 = 186,765

La tabla de ANOVA es:

TABLA 12.4 Tabla de ANOVA

CAUSA DE VARIACION	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS	F
Entre tratamientos	373,53	2	186,765	1,907
ERROR	2643,80	27	97,918	
TOTAL	3017,33	29	-	

$F = (186,765) / (97,918) = 1,907$

7. Dado que $F = 1,907 < F^{*}_{1\%;2,27} = 5,49$, no se rechaza H_o al 1%.

Ejemplo 12.2 (Muestras con Tamaños Diferentes) Escobar (11) presenta el siguiente ejemplo: La población de campesinos en la región A se dividió en tres grupos (según el ingreso neto por hectárea), denominados Cabeza, Medio y Cola. Si se tiene la información de la tabla 12.5., podemos concluír que los tres grupos difieren significativamente en su ingreso medio por hectárea?

TABLA 12.5 Ingreso Promedio de los Agricultores en la Región A, 1977

	GRUPO 1 (cola)		Grupo 2 (medio)		Grupo 3 (cabeza)	
	Y_{1j}	Y^2_{1j}	Y_{2j}	Y^2_{2j}	Y_{3j}	Y_{3j}^2
	635	403.225	781	609.961	835	697.225
	670	448.900	788	620.944	847	717.409
	683	466.489	794	630.436	852	725.904
	716	512.656	801	641.601	858	736.164
	731	534.361	803	644.809	879	772.641
	742	550.564	805	648.025	892	795.664
	753	567.009	807	651.249	913	833.569
	760	577.600	809	654.481	936	876.096
	768	589.824	812	659.344	1035	1.071.225
	770	592.900	818	669.124		
			826	682.276		
TOTALES	7228	5.243.528	8844	7.112.250	8047	7.225.897
n_i	10		11		9	
$\bar{X}_{icol}$	722,8		804,0		894,11	

Solución: Ho: $\mu_1 = \mu_2 = \mu_3$

1. H_1: Al menos uno de los promedios difiere de los otros.

2. $\alpha = 0,05$; $n_1 = 10$, $n_2 = 11$, $n_3 = 9$; $N = 30$.

3 y 4. Iguales a los del problema 12.1.

5. $\delta_1 = (c-1) = (3-1) = 2$ = grados de libertad del numerador.

$\delta_2 = (N-c) = (30-3) = 27$ = grados de libertad del denominador.

Se rechaza H_o al 5% si $F > F^*_{5\%,\ \delta_1,\ \delta_2} = 3,35$.

6. De acuerdo a la tabla 12.5 tenemos que: $\bar{X}_1 = 722,8$; $\bar{X}_2 = 804,0$; $\bar{X}_3 = 894,11$; $N = 30$; $\bar{\bar{X}} = 24119,0 / 30 = 803,97$

$$\sum_{i=1}^{3} \sum_{j=1}^{n_i} X_{ij} = 24.119,0 \quad ; \quad \sum_{i=1}^{3} \sum_{j=1}^{n_i} X_{ij}^2 = 19.581.675,0.$$

$$N \bar{\bar{X}}^2 = 30\ (803,97)^2 = 19.391.032,00$$

$$\sum_{i=1}^{3} n_i \bar{X}_{icol}^2 = 10(722,8)^2 + 11(804,0)^2 + 9(894,11)^2 = 19.529.868,6$$

Entonces, SCT = 19.581.675,0 - 19.391.032,0 = 190.643,0

SCC = 19.529.868,6 - 19.391.032,0 = 138.836,6

SCE = 190.643,0 - 138.836,6 = 51.806,4

TABLA 12.6. Tabla de ANOVA

CAUSA	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS	F
Variac. entre grupos	138.836,6	2	69.418,3	36,179
Variac. dentro del grupo	51.806,4	27	1.918,76	
TOTAL	190.643,0	29	-	

7. Puestos que 36,179 > 3,35, se rechaza H_o al 5% y concluímos que las medias difieren significativamente entre sí.

12.2. La Mínima Diferencia Significativa

El análisis de varianza nos ayuda decir si o no hay una diferencia significativa entre las medias de los tratamientos, pero no nos indica donde está la diferencia. Cuando el estadístico F indica una diferencia significativa entre las medias, lo natural es tratar de conocer donde está la causa de la diferencia. Si se tienen muestras de igual tamaño, ésto se puede conocer mediante el procedimiento denominado de la mínima diferencia significativa (MDS).

Cuando $n_1 = n_2 = \ldots = n_c$, y F nos lleva a rechazar la hipótesis nula, la MDS se define como la diferencia más pequeña que pueda existir entre dos medias muestrales estadísticamente diferentes.

$$MDS = \sqrt{(2/n)\ (\ CME\ )\ (\ F^{*}_{1,(N-c);\alpha}\ )} \qquad \{\ 12.13\ \}$$

Donde, n = Tamaño muestral común

CME = Cuadrados medios debidos al error

$F^{*}_{1,(N-c);\ \alpha}$ = Valor crítico de la distribución de F con $\delta_1 = 1$ y $\delta_2 = (N-c)$ y a un nivel del α%.

Cualesquiera dos medias de los tratamientos muestrales en el ANOVA, se dice que difieren significativamente, a un nivel de significancia (α) determinado, si la diferencia absoluta entre ellas es mayor que el va-

lor de MSD calculado utilizando la fórmula { 12.13 } .

Ejemplo 12.3. Asuma que la edad de cinco jefes de familia en tres veredas diferentes son los tabulados en la tabla 12.7. Existe diferencia significativa en la edad promedia de una vereda a la otra ? Si la respuesta es afirmativa, cuales de las medias son significativamente diferentes entre sí?

TABLA 12.7. Edad Promedia de Cinco Agricultores en Tres Veredas, 1977

Observaciones	V_1	V_2	V_3	Total
1	25	31	24	
2	30	39	30	
3	36	38	28	
4	38	42	25	
5	31	35	28	
Total	160	185	135	480
$\bar{X}_{icol.}$	32	37	27	$32 = \bar{\bar{X}}$

Solución:

Las hipótesis y el procedimiento son similares a los realizados en el problema anterior. Tras realizar los cálculos encontramos que, $n = 5$; CME = 16,17; $F_{2,12} = 7,50$. Dado que $F^{*}_{1\%;2,12} = 6,93$; rechazamos la hipótesis que las medias no difieren significativamente al unc por ciento.

$$F^{*}_{1\%,(1,12)} = 9,33 \quad \text{(ver tabla 6.5.c)}$$

$$MDS = \sqrt{(2/5)\ (16,17)\ (9,33)} = 7,768$$

$$| X_1 - X_2 | = | 32 - 37 | = 5 < 7,768$$

$$| X_1 - X_3 | = | 32 - 27 | = 5 < 7,768$$

$$| X_2 - X_3 | = | 37 - 27 | = 10 > 7,768$$

Solamente las poblaciones dos y tres difieren significativamente en la edad promedio del jefe de familia.

La prueba de MSD se denomina una prueba de parejas y sólo debe emplearse cuando el estadístico F de la fórmula { 12.12 } ha llevado a rechazar la hipótesis nula. Cuando se usa la prueba de MDS sin que se pueda rechazar H_o, pueden resultar algunas diferencias entre las mayores de las medias, pero estas son explicadas por cambios aleatorios.

Si las muestras no son de igual tamaño, no hay un procedimiento general para determinar cual pareja de medias causa la diferencia significativa en la prueba. Por consiguiente, se debe de evaluar todos los posibles pares de medias por medio de la prueba de t explicada en el capítulo X (9).

12.3. Clasificación de una Sola Variable Mediante Bloques Aleatorios

El modelo de ANOVA discutido arriba, es el diseño más sencillo y es el adecuado cuando los elementos analizados son relativamente ho

mogeneos con respecto a todas las variables menos en la que se está probando que hay una diferencia significativa. Cuando esto no sucede, se debe emplear otra clase de modelo. Uno de los diseños alternos al completamente aleatorio, es el de clasificación de una variable en bloques aleatorios.

Experimentos aleatorizados en bloques es un término agronómico utilizado para denotar que varias variables o tratamientos se aplican a diferentes bloques de tierra para replicaciones de un experimento. El objetivo principal es el de establecer las diferencias significativas entre los efectos de los tratamientos, tales como productos de diferentes tipos de semillas , o la calidad de diferentes marcas de fertilizantes. Pero las diferencias en las producciones de semillas o fertilizantes pueden atribuirse no solamente a las diferentes marcas empleadas, sino también a las diferancias en la calidad de los bloques. Para aislar el efecto de éstos se emplea la aleatoridad, la cual se logra asignando tratamientos en forma aleatoria entre los diferentes bloques. Aunque el diseño se originó en las ciencias naturales, tiene amplia aplicabilidad en otras áreas del conocimiento humano. Por ejemplo, el diseño se puede emplear para determinar si c marcas diferentes de fertilizantes dan igual productividad, con datos provenientes de encuestas socioeconómicas, aislando el efecto que tenga la capacidad del administrador de la finca (bloques), mediante asignación de las marcas de fertilizantes en forma aleatoria entre los diferentes administradores, los cuales han sido seleccionados de antemano en forma aleatoria. La idea básica del

diseño es comparar todos los efectos del tratamiento dentro de un bloque de material experimental, eliminando los efectos del ambiente.

Para elaborar el modelo, consideremos c tratamientos A_i (i = 1, 2,...,c) y r bloques B_j (j = 1, 2,..,r). Entonces, cada observación X_{ij} se puede considerar como una muestra de tamaño uno, sacada de una población con media igual a μ_{ij}. Por consiguiente, existen rc poblaciones, una por cada combinación de bloques con tratamientos. Se asume que las rc poblaciones están normalmente distribuidas con una varianza común σ^2. La gran media poblacional se define como:

$$\mu = (1 / rc) \sum_{i=1}^{c} \sum_{j=1}^{r} \mu_{ij} \qquad \{12.14\}$$

Los datos se tabulan tal como aparecen en la tabla 12.8.

TABLA 12.8 Clasificación de una Variable Mediante Bloques Aleatorios

	A_1	A_2	A_3	A_c	$t_{\cdot j}$	$\bar{X}_{\cdot j}$
B_1	X_{11}	X_{21}	X_{31}	X_{c1}	$t_{\cdot 1}$	$\bar{X}_{\cdot 1}$
B_2	X_{12}	X_{22}	X_{32}	X_{c2}	$t_{\cdot 2}$	$\bar{X}_{\cdot 2}$
.	.	.	. .	.	.	.
.	.	.	. .	.	.	.
B_r	X_{1r}	X_{2r}	X_{3r}	X_{cr}	$t_{\cdot r}$	$\bar{X}_{\cdot r}$
$t_{i\cdot}$	$t_{1\cdot}$	$t_{2\cdot}$	$t_{3\cdot}$	$t_{c\cdot}$	T = Gran total	
$\bar{X}_{i\cdot}$	$\bar{X}_{1\cdot}$	$\bar{X}_{2\cdot}$	$\bar{X}_{3\cdot}$	$\bar{X}_{c\cdot}$	$\bar{\bar{X}} = \frac{1}{rc} T$	

En la tabla 12.8., cada observación X_{ij} tiene doble subíndice. El subíndice i indica el tratamiento y el j indica el bloque. Por ejemplo, X_{34} es la observación que está en el cuarto bloque y pertenece al tercer tratamiento. Para elaborar el estadístico, buscamos: a) los totales de cada columna ($t_{i.}$) y los totales de cada fila ($t_{.j}$); b) las medias de los tratamientos ($\bar{X}_{i.}$) y la de los bloques ($\bar{X}_{.j}$) .

El modelo general utilizado es:

$$X_{ij} = \mu + \omega_i + \Omega_j + e_{ij} \qquad \{ 12.15 \}$$

La fórmula {12.15} nos indica que cada observación difiere de la gran media (μ) por un factor debido a los tratamientos (ω_i), otro debido a las diferencias entre los bloques (Ω_j) y por un factor completamente aleatorio (e_{ij}) o de error.

Las medias de las filas y de las columnas se encuentran mediante las siguientes fórmulas:

$$\bar{X}_{i.} = (1 / r) \; t_{i.} \qquad \{ 12.16 \}$$

$$\bar{X}_{.j} = (1 / c) \; t_{.j} \qquad \{ 12.17 \}$$

Otros estadísticos importantes son:

$$T = \text{Gran total} = \sum_{i=1}^{c} t_{i.} = \sum_{j=1}^{r} t_{.j} \qquad \{ 12.18 \}$$

$\bar{\bar{X}}$ = Estimativo de la gran media poblacional $\mu = (1 / rc) \; T$

$$C = \text{Factor de corrección} = (1 / rc) \; T^2 \qquad \{ 12.19 \}$$

La suma de cuadrados totales (SCT) se divide en: a) La suma de cuadrados debida a los tratamientos o columnas (SCC); b) La suma de

cuadrados debida a los bloques o filas (SCF); y c) La suma de cuadrados debida a los términos de error (SCE).

$$SCT = SCC + SCF + SCE \qquad \{ 12.20 \}$$

Para encontrar estas sumas de cuadrados tenemos las fórmulas siguientes:

$$SCT = \sum_{i=1}^{c} \sum_{j=1}^{r} X_{ij}^2 - C \qquad \{ 12.21 \}$$

$$SCC = (1 / r) \sum_{i=1}^{c} t_{i\cdot}^2 - C \qquad \{ 12.22 \}$$

$$SCF = (1 / c) \sum_{j=1}^{r} t_{\cdot j}^2 - C \qquad \{ 12.23 \}$$

$$SCE = SCT - (SCC + SCF) \qquad \{ 12.24 \}$$

La tabla de ANOVA es:

TABLA 12.9 Análisis de Varianza, tabla de ANOVA

CAUSA DE VARIACION	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS
Columnas (tratamientos)	SCC	c - 1	CMC = SCC/(c-1)
Filas (bloques)	SCF	r - 1	CMF = SCF/(r-1)
Error	SCE	(r-1)(c-1)	CME = SCE/(r-1)(c-1)
TOTAL	SCT	rc - 1	-

Existen dos conjuntos de hipótesis que podemos formular con los datos:

a) H_o : $\mu_1 = \mu_2 = \ldots = \mu_c$ (ó $\omega_i = 0$)

H_1 : Al menos una de las medias es diferente a las otras.

b) H_o : $\Omega_1 = \Omega_2 = \ldots\ldots = \Omega_r = 0$

H_1 : No todos los Ω_j son iguales a cero.

El primer par de hipótesis implica que no hay diferencias significativas entre los tratamientos y la segunda que no hay diferencia entre los bloques (y que por tanto, éstos ejercen una influencia neutral). Sin embargo, en el presente diseño sólo estamos interesados en el primer tipo de hipótesis, las cuales se pueden docimar mediante el estadístico:

$$F = CMC \;/\; CME \qquad \{12.25\}$$

el cual sigue una distribución de F con δ_1 = (c-1) grados de libertad para el numerador y δ_2 = (r-1) (c-1) grados de libertad para el denominador. Se rechaza la hipótesis nula al α por ciento si $F > F^*_{\alpha,(\delta_1,\delta_2)}$ (9,18,42).

Ejemplo 12.4. Se tienen cinco métodos diferentes de comunicación y se desea saber cual de ellos es el más efectivo para la Vereda Aguas. Al azar se seleccionan cinco agricultores y a cada uno de ellos se les somete a los cinco métodos (el método escogido en una ocasión determinada es seleccionado al azar) y al final se realiza un examen de comprensión, el cual se califica sobre 20 puntos . Dados los resultados de la tabla 12.10, podemos concluír que los cinco métodos son igualmente eficientes (en cuanto a su facilidad de comprensión) en la zona estudiada?

TABLA 12.10. Puntajes de comprensión de cinco métodos de Comunicación en la vereda Aguas, 1977.

AGRICULTOR	MÉTODOS 1	2	3	4	5	$t_{.j}$	$\bar{X}_{.j}$
1	15	20	12	17	8	72	14.4
2	12	13	20	12	18	75	15.0
3	5	18	8	10	12	53	10.6
4	18	12	10	20	13	73	14.6
5		15	18	12	7	61	12.2
$t_{i.}$	59	78	68	71	58	334 = T	
$\bar{X}_{i.}$	11,8	15,6	13,6	14,2	11,6	13,36 = $\bar{\bar{X}}$	

Solución: Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

1. H_1: No todos los promedios son iguales

2. $\alpha = 0,05$; $n_i = 5$ (i = 1, 2, 3, 4, 5,); $N = \sum_{i=1}^{5} n_i = 25$

3, 4 Análisis de varianza, diseño de clasificación de una sola variable mediante bloques aleatorios. El interés se centra en ver si hay diferencia significativa entre los tratamientos pero se quiere quitar la diferencia de los bloques (agricultor). Se utiliza el estadístico F (fórmula {12.25} el cual sigue una distribución de F con

$\delta_1 = (c-1) = (5-1) = 4$ y $\delta_2 = (r-1)\ (c-1) = (5-1)(5-1) = 16$

grados de libertad.

5. Se rechaza Ho al 5% si $F > F^*_{\delta_1, \delta_2} = 3,01$

6. Tenemos que: r = 5; c = 5; $\Sigma\ t_{i.} = \Sigma\ t_{.j} = 334 = T$

$\bar{\bar{X}} = 13,36$; $N = 25 = rc$.

$$\sum_{i=1}^{c=5}\sum_{j=1}^{r=5} X_{ij}^2 = (15^2 + 20^2 + 12^2 + \ldots + 7^2) = 4.920,0$$

C = Factor de corrección = $(1/rc)T^2 = (1/25)(334)^2 = 4.462,24$

$$\sum_{i=1}^{5} t_{i\cdot}^2 = (59^2 + 78^2 + 68^2 + 71^2 + 58^2) = 22.594,0$$

$$\sum_{i=1}^{5} t_{\cdot j}^2 = (72^2 + 75^2 + 53^2 + 73^2 + 61) = 22.668,0$$

Aplicando las fórmulas {12.21} a {12.24} obtenemos que:

SCT = 4.920,0 - 4.462,24 = 457,76

SCC = (2/5) (22.594,0) - 4.462,24 = 56,56

SCF = (1/5) (22.668,0) - 4,462,24 = 71,36

SCE = 457,76 - (56,56+ 71,36) = 329,84

La Tabla de ANOVA es:

TABLA 12.11. Tabla de ANOVA

Causa de variación	Suma de cuadrados	Grados de libertad	Cuadrados Medios
Columnas (tratamientos)	56,56	(c-1) = 4	56,56/4 = 14,14
Filas (bloques)	71,36	(r-1)= 4	71,36/4 = 17,84
Error	329,84	(c-1)(r-1)=16	329,84/16= 20,615
TOTAL	457,76	rc-1 = 24	

F = CMC/CME = 14,14/20,615 = 0,686

7. Puesto que F = 0,686 < $F^*_{3\%;\ 4,16}$ = 3,01, no se rechaza Ho al 5%

y concluímos que no hay diferencia significativa entre los cinco pro

medios.

Tal como en el caso del diseño simple, cuando los tamaños de muestras son iguales y al rechazar la hipótesis nula, se puede saber cuales de las diferencias son significativas, utilizando el criterio de la mínima diferencia significativa (MDS). Donde,

$$MDS = \sqrt{(2/r)(CME)(F^{*}_{\alpha\%;1,(r-1)(c-1)})} \qquad \{12.26\}$$

El procedimiento de decisión es similar al explicado anteriormente.

12.4. Modelo Completamente Aleatorizado a Dos Criterios, sin Replicación

Este modelo se refiere a un diseño aleatorio, en el cual los datos muestrales se han clasificado de acuerdo a dos variables aleatorias independientes, y dónde sólo hay una observación en cada celda. El marco teórico de este modelo es exactamente el mismo del modelo discutido en la sección anterior (Modelo de Bloques Aleatorios) , excepto que se diferencia en la interpretación de los datos.

En el modelo de clasificación de una variable mediante bloques aleatorios, estabamos interesados únicamente en los efectos de los tratamientos (la variable independiente A_i). Los bloques se consideraban solamente un material experimental. En el modelo de clasificación de acuerdo a dos variables, tanto los tratamientos

como los bloques son variables independientes y se evaluan en forma simultánea. Consecuentemente, mientras A_i son los tratamientos en el modelo de una variable; en el modelo de dos variables, tratamientos son las combinaciones de bloques y tratamientos A_i B_j. Más aún, en el modelo de clasificación de acuerdo a dos variables sin un modelo de réplica, se seleccionan muestras de tamaño uno de cada tratamiento combinado y los datos muestrales se presentan así:

Tratamiento Combinado	A_1B_1	A_1B_2	A_2B_1....	A_cB_r
Observación	x_{11}	x_{12}	x_{21}	x_{cr}

Expresada en una tabla de dos clasificaciones se tiene:

	A_1	A_2	A_c
B_1	x_{11}	x_{21}	x_{c1}
B_2	x_{12}	x_{22}	x_{c2}
.	.	.	.
B_r	x_{1r}	x_{2r}	x_{cr}

La segunda forma de presentación es exactamente igual al caso anterior. Sin embargo, tiene este marco una interpretación diferente. En el modelo de clasificación de una sola variable, A_i constituyen la variable independiente y B_j son los bloques a los cuales se les

asigna A_i en forma aleatoria. En el método de clasificación de dos variables, tanto A_i como B_j son variables aleatorias de tamaño uno sacadas de una poblacion que corresponde a Ai Bj.

El análisis de varianza y el procedimiento de prueba para este modelo, es el mismo que para el modelo anterior, excepto que aquí tienen un significado ambos grupos de pruebas : $\omega_i = 0$ y $\Omega_j = 0$, los cuales se deben probar simultáneamente.

Hipótesis

1) Prueba para A_i (tratamientos: columnas) Ho: $\omega_i = 0$

H_i: no todos los ω_i son cero

Estadístico: $F_{(c-1),(r-1)(c-1)} = \frac{CMC}{CME}$

2) Prueba para B_j (tratamientos:filas)

Ho: $\Omega_j = 0$

H_1: No todas las Ω_j son iguales a cero

Estadístico: $F_{(r-1);\ (r-1)(c-1)} = \frac{CMF}{CME}$

Ejemplo 12.5. Con referencia a los datos del problema 12.4, podemos concluír que hay diferencia significativa entre los diferentes métodos de comunicación y entre los diferentes agricultores?

Solución: a)

Ho: $\omega_i = 0$

H_1: No todos los ω_i son nulos

1. b) Ho: $\Omega_j = 0$

H_1: no todas las Ω_j son nulas

2. $\alpha = 0,05$; $N = 25 = rc$

Contemos directamente nuestro interés en el estadístico y criterio de decisión:

a) $F_{4,16} = 14,14/20,615 = 0,686$

b) $F_{4,16} = 17,84/20,615 = 0,865$

Dado que ambos estadísticos F tienen iguales grados de libertad y son menores que el $F^*_{5\%;\ 4,16} = 3,01$, no se rechaza Ho al 5% y aceptamos que no hay diferencia significativa ni entre los tratamientos ni entre los bloques.

Si se rechaza la hipótesis nula, Ho: $\Omega_j = 0$ se puede hallar cuales promedios (de los bloques) son significativos mediante la MDS.

$$MDS = \sqrt{(2/c)(CME)\ F^*_{\alpha\%;(r-1)(c-1)}} \qquad \{ 12.27 \}$$

Aparte de los diseños discutidos existen otros tipos de diseños más complicados que pueden ser utilizados en el análisis de los datos socioeconómicos. Sin embargo, los tres modelos discutidos en el presente capítulo son los de aplicación más inmediata a los datos de la Ficha Técnica.

XIII. EL MODELO DE REGRESION LINEAL BIVARIADA

Si a los elementos de una población le observamos dos medidas en forma simultánea, tenemos una población bivariante y si observamos más de dos medidas tenemos una población multivariante. Por ejemplo cuando observamos el ingreso y los gastos en fertilizantes de cada uno de los agricultores en una región determinada, originamos una población bivariada. El problema principal al analizar los datos bivariantes o multivariantes es describir y medir la asociación o covariación que existe entre las variables analizadas y determinar cómo las variables varían juntas. Por ejemplo, podemos observar que a mayor ingreso, los agricultores tienden por lo general a gastar más en fertilizantes y viceversa a menor ingreso compran menos fertilizantes. Lógicamente, ésto puede ser cierto en el promedio, puesto que hay excepciones, donde a altos ingresos se gaste menos en fertilizantes. Sin embargo, si se puede establecer la relación promedia que existe entre las dos variables de una forma matemática, estaremos en posición de estimar en modo bastante preciso, en promedio, el gasto en fertilizantes de un agricultor en base de su nivel de ingreso. Tal procedimiento se denomina estimación por asociación.

Existen dos aspectos distintos pero relacionados en el estudio de asociación entre variables. El primero se denomina <u>análisis de regresión</u>, el cual trata de establecer la "naturaleza de la relación"

entre variables . Es decir, en el análisis de regresión estudiamos la relación funcional entre las variables de modo que podamos predecir el valor de una variable (denominada la variable dependiente o explicada) en base a los valores conocidos de otra u otras variables (llamadas las variables independientes o explicatorias). El segundo aspecto del análisis de asociación se denomina el <u>análisis de correlación</u> y se ocupa en determinar el grado de relación entre las variables analizadas (9,12,18,23).

13.1. El Modelo de Regresión Lineal Bivariante.

En el análisis de regresión, como en otros tipos de estudios estadísticos, generalmente procedemos a observar los datos muestrales, utilizando los resultados obtenidos como estimación de la correspondiente relación poblacional.

Para llegar a determinar una ecuación que nos relacione las variables, un primer paso que nos sirve de ayuda es la colección de datos que muestren los valores correspondientes de las variables analizadas. Por ejemplo, si se desea estudiar la relación que existe entre el ingreso (variable X) y los gastos en fertilizantes (variable Y), se obtendría una muestra de n observaciones, donde se tendrían los ingresos $X_1, X_2, X_3, \ldots, X_n$ y los gastos $Y_1, Y_2, \ldots, Y_n$. El paso siguiente es representar los puntos $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ en un sistema de coordenadas rectangulares para obtener el

Diagrama de dispersión, el cual facilita representar la curva que se ajuste a los datos. Por ejemplo, en el diagrama de dispersión de la figura 13.1 se puede observar que los datos se ajustan a una línea recta de pendiente positiva.

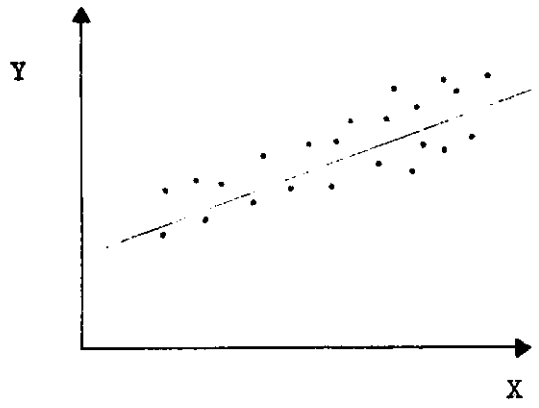


Figura 13.1.

El modelo más sencillo que se puede estimar es el modelo de regresión lineal bivariante, el cual tiene la forma funcional:

$$Y_i = \alpha + \beta X_i \qquad \{ 13.1 \}$$

Lo que deseamos es conocer los valores de α y β, de tal modo que para un dado valor de X (la variable independiente) podamos predecir el valor de Y (la variable dependiente). Puesto que no se tiene toda la serie completa, lo que tomamos es una muestra aleatoria en el tiempo para estimar los parámetros α y β mediante los estadísticos $\hat{\alpha}$ y $\hat{\beta}$.

En la práctica no podemos obtener una relación exacta (tal como la sugerida por la fórmula { 13.1 }), puesto que es muy difícil que obtengamos que una variable sea explicada por una sola variable explicatoria. Por consiguiente, cometemos un error al no considerar otras variables explicatorias. Además y especialmente cuando Y se

refiere a una variable socioeconómica, es muy difícil predecir con plena certeza su valor, ya que Y contiene un elemento aleatorio, que la hace cambiar de observación a observación de la variable X. Finalmente,no tenemos instrumentos de medidas que nos permitan medir con exactitud el valor de Y. Todo esto implica que debemos tener un término aleatorio, el cual esperamos se distribuya alrededor de cero (los unos tienden a compensarse con los otros). Por consiguiente, el modelo poblacional es:

$$Y_i = \alpha + \beta X_i + e_i \qquad \{13.1a\}$$

Para estimar los valores de los parámetros de α y β mediante el modelo de regresión, consideramos los siguientes supuestos (21,23):

1) X es una variable fija medida en forma correcta, no correlacionada con los términos de error y con una varianza finita diferente a cero. 1/

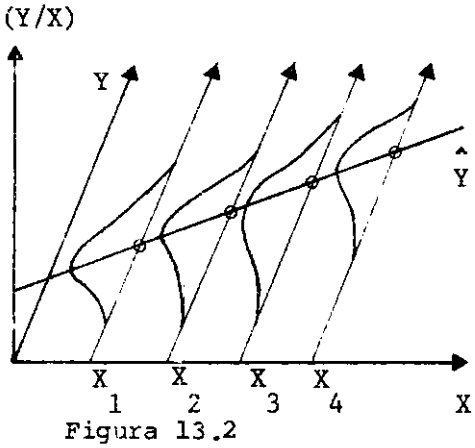
2) 

Figura 13.2

Para cada valor de X existe una serie de valores aleatorios de Y, los cuales se distribuyen aleatoriamente con media igual a la línea de regresión y varianza finita.

1/ Con un cambio de interpretación en las distribuciones probabilísticas, podemos asumir que X es una variable aleatoria y llegar a los mismos resultados que cuando se asume que es fija.

$$E(Y/X) = \mu_{Y/X} = \alpha + \beta X \qquad \{ 13.2 \}$$

$$V(Y) = \sigma^2, \qquad 0 < \sigma^2 < \infty \qquad \{ 13.3 \}$$

3) Los términos aleatorios se distribuyen con media cero y varianza finita, la cual es igual a la varianza de la variable aleatoria independiente Y.

$$E(e_i) = 0; \quad V(e_i) = \sigma^2 = V(Y) \qquad \{ 13.4 \}$$

4) Los términos de error no están correlacionados entre sí. Teniéndo en cuenta estos supuestos, queremos estimar el modelo poblacional de la fórmula {13.1 a} mediante el modelo muestral

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i \qquad \{ 13.5 \}$$

Para un dado valor de X_i, nosotros predecimos $\hat{Y}_i$ pero lo que en realidad se observa es Y_i. Por ende cometemos un error igual a $e_i = (Y_i - \hat{Y}_i)$. Algunos de estos errores son positivos (i. g e_1 en la figura 13.3) y otros son negativos (i. e. e_2 en la figura 13.3). Deseamos encontrar los valores de $\hat{\alpha}$ y $\hat{\beta}$ minimizando la suma de los errores elevados al cuadrado. El procedimiento matemático que nos permite hacer ésto se denomina el método de los mínimos cuadrados. Por consiguiente

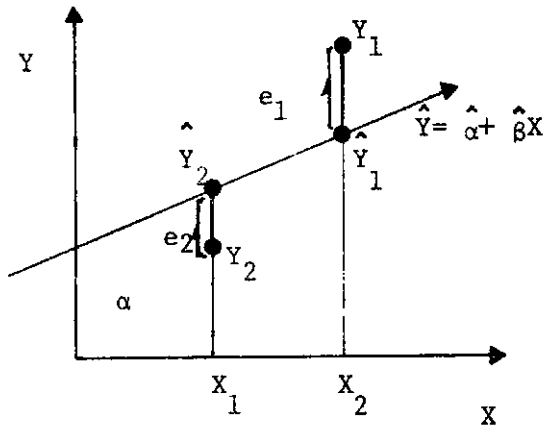


Figura 13.3

nuestra función objetivo es:

$$\text{Minimizar} \quad \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X)^2 \qquad \{13.6\}$$

Aplicación de las reglas de cálculo nos llevan al sistema de ecuaciones

$$\Sigma Y = n \hat{\alpha} + \hat{\beta} \Sigma X$$

$$\Sigma YX = \hat{\alpha} \Sigma X + \hat{\beta} \Sigma X^2 \qquad \{13.7\}$$ 2/

El sistema anterior se denomina las ecuaciones normales del modelo de regresión lineal bivariante. La solución algebráica del sistema es la siguiente (21,29):

$$\hat{\beta} = \frac{\Sigma (X-\bar{X})(Y-\bar{Y})}{\Sigma (X-\bar{X})^2} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}} \qquad \{13.8\}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} \qquad \{13.9\}$$

Los valores de $\hat{\alpha}$ y $\hat{\beta}$ se estiman mediante una muestra aleatoria determinada. Si se sacan todas las posibles muestras de tamaño n de la población estudiada, cada una de ellas va a dar una estimación diferente de α y β . Por consiguiente, podemos formar distribuciones probabilísticas de $\hat{\alpha}$ y $\hat{\beta}$, las cuales tienen las siguientes características:

a) Distribución muestral de $\hat{\beta}$

2/ Todas las sumatorias van desde i = 1 hasta n (número de observaciones muestrales de las dos variables).

$$E(\hat{\beta}) = \beta \qquad \{13.10.\}$$

$$V(\hat{\beta}) = \frac{\sigma^2}{\Sigma(X_i-\bar{X})^2} = \frac{\sigma^2}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}} \qquad \{13.11\}$$

b) Distribución muestral de $\hat{\alpha}$

$$E(\hat{\alpha}) = \alpha \qquad \{13.12\}$$

$$V(\hat{\alpha}) = \sigma^2 \left\{ \frac{\Sigma X_i^2}{n(\Sigma X^2 - (\Sigma X)^2/n)} \right\} \qquad \{13.13\}$$

Puesto que σ^2 no se conoce, se estima utilizando los datos muestrales mediante la fórmula:

$$\hat{\sigma}^2 = \frac{\Sigma e_i^2}{n-2} = \frac{\Sigma Y^2 - \hat{\alpha}\,\Sigma Y - \hat{\beta}\,\Sigma XY}{n-2} \qquad \{13.14\}$$

Ejemplo 13.1. En la vereda Aguas existían en 1977, doce cultivadores de papa. La información del capital líquido disponible y los gastos de insumos mejorados se presenta en la Tabla 13.1. Encuentre la relación funcional que existe entre el gasto en insumos (Y) y el capital líquido disponible (X) de los agricultores cultivadores de papa de la región. Si un agricultor tiene un capital disponible de 17 (en miles de pesos), que nivel de insumos mejorados se puede esperar que compre?

En la Tabla 13.1 se presentan los valores de X y de Y en las columnas 2 y 3, respectivamente. En la columna 4 se multiplica cada valor de X por el valor de Y, mientras que en las 5 y 6 se tabulan

TABLA 13.1. Capital Disponible y Gastos en Insumos Mejorados en la Vereda Aguas, 1977

Agric. (1)	Capital X ($000) (2)	Gastos Y ($00) (3)	XY (4)	X^2 (5)	Y^2 (6)	$\hat{Y}$ (7)	$(Y-\hat{Y})$ (8)	$(Y-\hat{Y})^2$ (9)
1	50	25	1250	2500	625	28,194	-3,194	10,202
2	45	20	900	2025	400	25,984	-5,984	35,808
3	15	12	180	225	144	12,724	-0,724	0,524
4	18	12	216	324	144	14,050	-2,050	4,203
5	58	30	1740	3364	900	31,730	-1,730	2,993
6	15	10	150	225	100	12,724	-2,724	7,420
7	10	12	120	100	144	10,514	+1,486	2,208
8	5	10	50	25	100	8,304	+1,696	2,876
9	57	35	1995	3249	1225	31,288	+3,712	13,779
10	28	19	532	784	361	18,470	+0,530	0,281
11	30	20	600	900	400	19,354	+0,646	0,417
12	42	33	1386	1764	1089	24,658	+8,342	69,589
Totales	373	238	9119	15485	5632	--	+0,006	150,300

los valores de X y de Y al cuadrado. Por consiguiente:

ΣX = total columna 2 = 373; ΣY = total columna 3 = 238;

ΣXY = total columna 4 = 9.119,0; ΣX^2 = total columna 5 = 15.485,0

ΣY^2 = total columna 6 = 5.632.0 ; n = 12

$\bar{X}$ = 373 / 12 = 31,083 ; $\bar{Y}$ = 238 / 12 = 19,833

Aplicando las fórmulas (13.8) y (13.9) Obtenemos:

$$\hat{\beta} = \frac{(9.119,0 - (373) (238) / 12)}{15.485,0 - (373)^2 / 12} = 0,442$$

$$\hat{\alpha} = 19,833 - (0,442) (31,083) = 6,094$$

El modelo estimado es:

$$\hat{Y}_i = 6,094 + 0,442 \; X_i$$

El valor de $\hat{\alpha} = 6,094$ es el nivel autónomo de Y, o sea aquel nivel de Y que no depende de X, e implica que cuando X=0 se gasta \$609,4 en insumos mejorados (puesto que la información de Y viene en cientos de pesos). El valor de $\hat{\beta} = 0,442$ es el valor de la pendiente de la linea estimada e indica que por cada unidad extra en la cual se incremente X, el valor de Y se aumenta en 0,442 unidades. En otras palabras, por cada \$1000 extra de capital liquido disponible, el agricultor promedio de la zona aumenta sus gastos en insumos mejorados en \$442.

Reemplazando los valores de X en la ecuación estimada, obtenemos los valores de $\hat{Y}$, los cuales se tabulan en la columna 7 de la tabla 13.1. Por ejemplo, $\hat{Y}_1 = 6,094 + 0,442 \; (50) = 28,194$.
En la columna 8 se tabulan los errores cometidos en la estimación realizada. Donde $e_i = (Y_i - \hat{Y}_i)$ (se restan las columnas 3 y 7). Por ejemplo, $e_1 = 25 - 28,194 = - 3,194$. Finalmente, en la última columna se presentan los valores de los errores elevados al cuadrado (valores de la columna 8 elevados al cuadrado).
Aplicando la fórmula { 13.14 } tenemos que:

$$\hat{\sigma}^2 = \Sigma e_i^2 \; / \; (n-2) = 150,3 \; / \; (12 - 2) = 15,03$$

O también,

$$\hat{\sigma}^2 = \frac{\Sigma Y^2 - \hat{\alpha} \Sigma Y - \hat{\beta} \Sigma YX}{n - 2} = \frac{5632 - 6,094(238) - 0,442(9119)}{12 - 2} = 15,10$$

La segunda fórmula para estimar σ^2 es más fácil de computar (puesto

que no tenemos que calcular $\hat{Y}_i$ ni e_i^2) y además más exacta (al estimar e_i cometemos un error de redondeo el cual se acrecenta al elevar los términos de error al cuadrado) . Por consiguiente, $\hat{\sigma}^2 = 15,1$. Aplicando las fórmulas { 13.11 } y { 13.13 } estimamos las varianzas de $\hat{\alpha}$ y $\hat{\beta}$, respectivamente.

$$V (\hat{\beta}) = \frac{15,10}{15485,0 - (373)^2 / 12} = \frac{15,10}{3890,917} = 0,00388$$

$$V (\hat{\alpha}) = 15,10 \left\{ \frac{15485,0}{12 (3890,917)} \right\} = 5,00789$$

Los errores estándar de $\hat{\beta}$ y $\hat{\alpha}$ son:

$$S_{\hat{\beta}} = \sqrt{V (\hat{\beta})} = \sqrt{0,00388} = 0,062$$

$$S_{\hat{\alpha}} = \sqrt{V (\hat{\alpha})} = \sqrt{5,00789} = 2,238$$

Las fórmulas { 13.10 } y { 13.12 } implican que $\hat{\beta}$ y $\hat{\alpha}$, son estimadores insesgados de β y α , respectivamente. Además, por el llamado teorema de GAUSS-MARKOV , se puede concluír que son estimadores eficientes, consistentes y suficientes. En otras palabras, si se cumplen todos los supuestos enunciados anteriormente, el método de los mínimos cuadrados (MMC) nos da estimadores que tienen todas la propiedades deseables de un buen estimador (21,32,40,51).

13.2 Intervalos de Confianza para α , β y σ^2

En el modelo de regresión lineal simple (ó bivariante) se pueden calcular intervalos de confianza (I.C.) del (1 - α) por

ciento para α, β y σ^2, de igual modo como se definieron en el capítulo VIII para otros estadísticos muestrales. Estos intervalos son: 3/

$$P(\hat{\alpha} - t^*_{\alpha/2;(n-2)} S_{\hat{\alpha}} < \alpha < \hat{\alpha} + t^*_{\alpha/2;(n-2)} S_{\hat{\alpha}}) = 1-\alpha \quad \{13.15\}$$

$$P(\hat{\beta} - t^*_{\alpha/2;(n-2)} S_{\hat{\beta}} < \beta < \hat{\beta} + t^*_{\alpha/2;(n-2)} S_{\hat{\beta}}) = 1-\alpha \quad \{13.16\}$$

$$P\left\{\frac{(n-2)\,\hat{\sigma}^2}{\chi^2_L} < \sigma^2 < \frac{(n-2)\,\hat{\sigma}^2}{\chi^2_c}\right\} = 1-\alpha \quad \{13.17\}$$

Ejemplo 13.2. Encuentre un I.C. del 95% para α, β y σ^2, utilizando los datos del problema 13.1.

Solución: De acuerdo a los datos del problema 13.1., tenemos que:

$\hat{\alpha} = 6,094$; $\hat{\beta} = 0,442$; $\hat{\sigma}^2 = 15,1$; $S_{\hat{\alpha}} = 2,238$;

$S_{\hat{\beta}} = 0,062$; $n = 12$; $(n-2) = 10$

$t^*_{\alpha/2;(n-2)} = t^*_{0,025;10} = 2,23$ (ver tabla 6.4)

$\chi^2_L = \chi^2_{10;0.975} = 20,5$; $\chi^2_c = \chi^2_{10;0,025} = 3,25$ (tabla 6.3)

Por consiguiente, aplicando las fórmulas { 13.15 } a { 13.17 }, tenemos que:

a) $P(6,094 - 2,23(2,238) < \alpha < 6,094 + 2,23(2,238)) = 0,95$

$P(1,103 < \alpha < 11,085) = 0,95$

Podemos esperar con un 95% de confianza que el verdadero intercepto de

---

3/ Para realizar pruebas de hipótesis e intervalos de confianza, se asume que las distribuciones muestrales de $\hat{\alpha}$, $\hat{\beta}$, $\hat{Y}$ y e_i son normales.

la linea en la población esté entre $ 110,3 y $ 1.108,5 (puesto que la información está en cientos de pesos).

b) $P(0,442 - 2,23(0,062) < \beta < 0,442 + 2,23(0,062)) = 0,95$

$$P (0,304 < \beta < 0,580) = 0,95$$

Con un 95% de confianza podemos afirmar que la verdadera pendiente de la línea en la población está entre 0,304 y 0,580. Es decir, que por cada mil pesos que se incremente el capital líquido, el agricultor promedio gastará entre $ 304 y $ 580 adicionales en insumos mejorados.

c) $$P \left\{ \frac{(12 - 2)(15,1)}{20,5} < \sigma^2 < \frac{(12 - 2)(15,1)}{3,25} \right\} = 0,95$$

$$P (7,419 < \sigma^2 < 46,462) = 0,95$$

$$P (2,714 < \sigma < 6,816) = 0,95$$

Podemos afirmar con un 95% de confianza que la verdadera varianza poblacional alrededor de la línea de regresión está entre 7,419 y 46,462; mientras que la desviación típica poblacional (la raíz cuadrada de los límites de la varianza) está entre 2,714 y 6,816.

13.3. Prueba de Hipótesis para α, β y σ^2

Utilizando los estadísticos:

$$t_c = (\hat{\alpha} - \alpha_o) / S_{\hat{\alpha}} \sim t_{n-2} \quad \{ 13.18 \}$$

$$t_c = (\hat{\beta} - \beta_o) / S_{\hat{\beta}} \sim t_{n-2} \quad \{ 13.19 \}$$

$$\chi^2_c = (n-2) \hat{\sigma}^2 / \sigma^2_o \sim \chi^2_{n-2} \quad \{ 13.20 \}$$

Podemos probar hipótesis acerca de los parámetros α , β y σ^2 tal como se ilustra en el ejemplo 13.3.

Ejemplo 13.3. Con referencia a los datos del problema 13.1., docime las siguientes hipótesis:

a) H_o: $\beta = 0$ $\quad$ H_1: $\beta > 0$

b) H_o: $\beta = 0{,}50$ $\quad$ H_1: $\beta \neq 0{,}50$

c) H_o: $\alpha = 0$ $\quad$ H_1: $\alpha > 0$

d) H_o: $\alpha = 12{,}0$ $\quad$ H_1: $\alpha \neq 12{,}0$

e) H_o: $\sigma_2^2 = 20{,}0$ $\quad$ H_1: $\sigma > 20{,}0$

Solución: Tenemos que, $t^*_{n-2;\,\alpha} = t^*_{10;,05} = 1{,}81$ (prueba de una cola a la derecha); y $t_{n-2;\,\alpha/2} = t^*_{10;0{,}025} = 2{,}23$ (prueba de dos colas).

a) $t_c = (0{,}442 - 0) / 0{,}062 = 7{,}129 > 1{,}81$. Se rechaza H_o al 5% y aceptamos que $\beta > 0$. Esto implica que la variable X si influencia significativamente a Y.

b) $t_c = (0{,}442 - 0{,}5) / 0{,}062 = -0{,}935 > -2{,}23$. No se rechaza H_o al 5%. Nótese que este valor está incluído en el intervalo de confianza del 95% para β y sería otro criterio para no rechazar H_o).

c) $t_c = (6{,}094 - 0) / 2{,}238 = 2{,}723 > 1{,}81$. Se rechaza H_o al 5% y concluimos que la línea no pasa por el origen.

d) $t_c = (6{,}094 - 12) / 2{,}238 = -2{,}638 < -2{,}23$. Se rechaza H_o al 5% . Nótese que este valor no está incluído en el I.C. del 95% para α estimado anteriormente.

e) $\chi^2_c = \frac{10(15,1)}{20} = 7,55 < \chi^{2*}_{10;,95} = 18,3$. Por tanto, no se rechaza H_o al 5% de significancia.

13.4. Predicción de un Valor Promedio de la Variable Dependiente

Conociendo los valores de $\hat{\alpha}$ y $\hat{\beta}$, se puede predecir el valor promedio de Y para un valor determinado de X. Si se toma un valor dentro del rango de los valores observados de X se dice que se realiza una interpolación . Si se toma un valor de X que no está dentro del rango observado de esta variable, realizamos una extrapolación.

El valor promedio que predecimos de Y depende de los valores muestrales (puestos que éstos determinan los valores de $\hat{\alpha}$ y $\hat{\beta}$ y $\hat{Y}$ es una función lineal de éstos), por consiguiente, podemos formar una distribución muestral del estadístico $\hat{Y}$, cuyas características principales son:

$$E(\hat{Y}) = \alpha + \beta X \qquad \{13.21\}$$

$$V(\hat{Y}) = \hat{\sigma}^2 \left\{ \frac{1}{n} + \frac{(X_o - \bar{X})^2}{\Sigma X^2 - (\Sigma X)^2/n} \right\} \qquad \{13.22\}$$

Si queremos analizar, ya no un valor promedio, sino un valor individual de Y (Y_o), el estadístico obtenido se distribuye con media igual a la distribución de un valor promedio, pero con una mayor variabilidad (21,29,51).

$$V(Y_o) = \hat{\sigma}^2 \left\{ 1 + \frac{1}{n} + \frac{(X_o - \bar{X})^2}{\Sigma X^2 - (\Sigma X)^2/n} \right\} \qquad \{13.23\}$$

Ejemplo 13.4. Si un agricultor tiene un capital disponible de $17 (en miles de pesos) que nivel de insumos mejorados se puede esperar que él compre?

Solución: $Y = 6,094 + 0,442\ (\ 17\ ) = 13,61$

Por consiguiente, podemos esperar que compre $ 13,61 en insumos mejorados. La anterior es una estimación puntual de Y . Para encontrar una estimación intervalo del 95% tenemos que:

$$V(\hat{Y}) = 15,10 \left\{ \frac{1}{12} + \frac{(17 - 31,083)^2}{3890,917} \right\} = 2,027$$

$$S_{\hat{Y}} = \sqrt{V(\hat{Y})} = \sqrt{2,027} = 1,424$$

$$P(13,61 - 2,23(1,424) < Y < 13,61 + 2,23(1,424)) = 0,95$$

$$(\ 10,435 < Y < 16,785\ ) = 0,95$$

Podemos concluír con un 95% que el verdadero gasto en insumos mejorados cuando X = 17, está entre $1043,5 y $ 1678,5.

Ejemplo 13.5 En 1978, un agricultor de la vereda Aguas tenía un capital disponible de $ 20 (en miles de pesos) y gastaba $15 (en cientos de pesos) en insumos mejorados, se habría podido predecir este valor con los datos de 1977 (tabla 13.1.) o ha cambiado la relación de un año a otro ?

Solución: Observamos en 1977 un valor de $Y_o = 15$, cuando $X = 20$. Aquí tenemos un valor particular de Y, puesto que es uno observado y no

un valor predicho por la ecuación. El valor que se predice para X=20 es igual a: $\hat{Y} = 6,094 + 0,442(20) = 14,93$.

Lo que deseamos probar es si o no hay diferencia significativa entre $Y_o = 15$, y $\hat{Y} = 14,93$. Esto lo realizamos mediante el estadístico,

$$t_c = (Y_o - \hat{Y}) / S_{Y_o} \sim t_{n-2} \qquad \{13.24\}$$

En nuestro ejemplo,

$$V(Y_o) = 15,1 \{ 1 + (1/12) + (20 - 31,083)^2 / 3890,917 \} = 16,835$$

$$S_{Y_o} = \sqrt{ V (Y_o) } = \sqrt{16,835} = 4,103$$

$t_c = (15 - 14,93) / 4,103 = 0,017 < t^*_{,025,10} = 2,23$. No se rechaza H_o al 5% y concluímos que nuestro modelo si predice el valor observado en 1978.

13.5. Análisis de Varianza en el Modelo Lineal Bivariado.

La suma de cuadrados totales (SCT) alrededor de Y (la variable dependiente) se puede descomponer en dos partes fundamentales. Una explicada por la regresión (SCR) o tratamiento y la otra debida al error (SCE) o a términos netamente aleatorios.

$$SCT = SCR + SCE \qquad \{ 13.25 \}$$

Donde,

$$SCT = \Sigma (Y - \bar{Y})^2 = \Sigma Y^2 - (\Sigma Y)^2 / n \qquad \{13.26\}$$

$$SCR = \Sigma (\hat{Y} - \bar{Y})^2 = \hat{\beta} \{ \Sigma XY - (\Sigma X)(\Sigma Y)/n \} \qquad \{13.27\}$$

$$SCE = SCT - SCR \qquad \{13.28\}$$

Con la suma de cuadrados totales existen (n-1) grados de libertad asociados con ella. Con la SCR hay uno y con la SCE existen (n-2). Por consiguiente, la tabla de ANOVA es igual a:

TABLA 13.2 Modelo de Regresión Lineal Bivariante,tabla de ANOVA

CAUSA	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS	F
REGRESION	SCR	1	CMR=SCR/1	CMR/CME
ERROR	SCE	n-2	CME=SCE/(n-2)	
TOTALES	SCT	n-1	-	-

Con el estadístico $F = CMR / CME$ { 13.29 } , probamos la hipótesis nula que la variable independiente X no influye a la variable dependiente Y. Esto implica que en el caso de la regresión lineal bivariante, la prueba de F, desempeña la misma función que la prueba de t del estadístico t_c (fórmula 13.13) para docimar que $\beta = 0$. Pero en el caso de regresión múltiple (donde tenemos más de una variable independiente), la prueba de F sirve para observar el efecto de todas las variables independientes, mientras que la de t nos sirve para probar la importancia de cada una de las variables independientes (en forma individual) en la variable dependiente (variable Y)(9,23,42,51).

Ejemplo 13.6. Realice el ANOVA con los datos del problema 13.1.

Solución: De acuerdo a los problemas anteriores tenemos que:

$$SCT = \Sigma Y^2 - (\Sigma Y)^2/n = 5632 - (238)^2/12 = 911,67$$

$$\{ \Sigma XY - (\Sigma X) (\Sigma Y) / n \} = 9119 - \{ (373)(238)/12 \} = 1721,17$$

$$SCR = (0,442) (1721,17) = 760,76$$

$$SCE = 911,67 - 760,76 = 150,91$$

TABLA 13.3. Tabla de ANOVA

CAUSA	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS	F
REGRESION	760,76	1	760,76	50,415
ERROR	150,91	10	15,09	
TOTALES	911,67	11	-	-

$$F = 760,76 / 15,09 = 50,415$$

El estadístico F de la fórmula { 13.29 } , sigue una distribución de F con $\delta_1 = 1$ y $\delta_2 = (n-2)$. Se rechaza H_o al α % (H_o: la(s) variable(s) independiente(s) no influye(n) a Y) si $F > F^*_{\alpha ; \delta_1, \delta_2}$

En nuestro ejemplo, $F^*_{5\%;1,10} = 4,96 < 50,415 = F$, por consiguiente, se rechaza H_o al 5% de significancia.

13.6. Comparación de Dos Pendientes

Asuma que en una región determinada (ó en un año específico) se ha estimado la regresión, $\hat{Y}_i = \hat{\alpha}_1 + \hat{\beta}_1 X_i$

y que en otra región (o en otro año), se estimó para las mismas variables la regresión: $\hat{Y}_i = \hat{\alpha}_2 + \hat{\beta}_2 X_i$

Se desea saber si hay o no una diferencia significativa del efecto de X en Y de una región a otra (ó sea se desea saber si hay o no una diferencia significativa entre β_1 Y β_2). Si tanto β_1 como β_2 son estadísticamente diferente de cero, las hipótesis:

$$H_o: \quad \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \quad \beta_1 \neq \beta_2$$ 4/

Se pueden docimar mediante el estadístico t_c, donde:

$$t_c = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{V(\hat{\beta}_1)\ V(\hat{\beta}_2)}} \qquad \{ 13.30 \}$$

El estadístico t_c definido por la fórmula { 13.30 } sigue una distribución de t con $\delta = (n_1 + n_2 - 4)$ grados de libertad. Se rechaza H_o al α % si $| t_c | > t^*_{\alpha/2;\ \delta}$ (21,23).

Ejemplo 3.7. En 1977 existían 12 cultivadores de papa en la vereda Aguas y 15 en la vereda el Cisne, que eran usuarios del ICA. Para cada una de las explotaciones agropecuarias usuarias del ICA de estas dos veredas se observó el capital líquido disponible X (medido en miles de pesos) y el nivel de gastos en insumos mejorados Y (medido en cientos de pesos). Los datos suministrados por la Ficha Técnica

---

4/ También se pueden realizar docimasias unilaterales

nos permitieron calcular que:

Vereda 1 (Aguas)

$\hat{\beta}_1 = 0,442$; $V(\hat{\beta}_1) = 0,00388$; $n_1 = 12$

Vereda 2 (El Cisne)

$\hat{\beta}_2 = 0,585$; $V(\hat{\beta}_2) = 0,0065$; $n_2 = 15$

Podemos concluír que los agricultores de papa en la vereda 2 gastan más en insumos mejorados por unidad de X que los de la vereda 1 ?

Solución: Tenemos que:

$$H_o: \beta_1 = \beta_2 \quad ; \quad H_1: \beta_2 > \beta_1$$

$\delta = (12 + 15 - 4) = 23$ = grados de libertad.

$t^*_{0,05;23} = 1,71$ (prueba de una cola)

$$t_c = (0,585 - 0,442) / \sqrt{0,00388 + 0,0065} = 0,992$$

Puesto que $t_c = 0,992 < t^*_{0,05;23} = 1,71$, no se rechaza H_o al 5% y concluímos que las dos pendientes son estadísticamente iguales.