

PRÁCTICA DE ESTADÍSTICA

FUNDAMENTOS DE BIOLOGÍA APLICADA I

PUNTUACIÓN: 2,6 DE 3

El Departamento de Ecología de la Universidad de Granada ha proporcionado un conjunto de datos en los cuales se recoge información sobre ciertas variables medidas en hojas de encina en distintas zonas de la provincia de Granada.

Del conjunto de datos total se ha seleccionado una muestra de 90 hojas de encina, tomadas en 5 zonas diferentes. Estas zonas son:

- Cortijuela
- Molinillo
- Fardes
- Ladihonda
- Fazares

Se sabe que las zonas 1, 2 y 3 son zonas húmedas (denominado 1) frente a las zonas 4 y 5 que son muy secas (denominado 2). (Las dos primeras columnas del archivo de datos adjunto hacen referencia a estas dos variables: LOCALIZACIÓN y CLIMA).

En cada localización se han tomado 18 hojas de encina y en ellas se ha medido la longitud y la anchura total. Además, con esta información se ha calculado un índice de asimetría. (Esta información viene en la tabla de datos adjunto, columnas 3 a 5: LONGITUD, ANCHURA TOTAL, ASIMETRÍA).

El **primer objetivo** es comprobar si la asimetría de las hojas de encina se ven afectadas por la localización de los árboles.

Para ello, se considera un diseño de un experimento en el que la localización de los árboles es el factor, (las distintas localizaciones son los niveles del factor), y la asimetría es la variable respuesta. Es un modelo equilibrado de efectos fijos y se plantea el siguiente contraste:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

$$H_1: \mu_i \neq \mu_j \text{ al menos para un par } (i,j) \text{ i } j$$

Los datos obtenidos son:

ANOVA de un factor

ANOVA

Asimetría

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	,027	4	,007	2,906	,026
Intra-grupos	,198	85	,002		

Total	,225	89			
-------	------	----	--	--	--

El valor del estadístico F para el contraste es 2,906

El p-valor asociado al contraste es 0,026, por lo que a un nivel de significación del 5%, se rechaza la hipótesis nula, por lo que al menos alguna de las localizaciones afecta a la asimetría de las hojas.

Para precisar qué zonas difieren estadísticamente entre sí, se usan los contrastes de Tukey, (nivel de significación 1%):

H0: $\mu_1 = \mu_2$

H1: $\mu_1 \neq \mu_2$

Comparaciones múltiples

Comparaciones múltiples

Variable dependiente: Asimetria

HSD de Tukey

(I) Zona	(J) Zona	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 90%					Límite inferior	Límite superior
	1	2	-,000556	,016085	1,000	-,04078	,03967				
		3	-,002611	,016085	1,000	-,04284	,03762				
		4	,033333	,016085	,242	-,00690	,07356				
		5	,035222	,016085	,194	-,00501	,07545				
2	1	,000556	,016085	1,000	-,03967	,04078					
		3	-,002056	,016085	1,000	-,04228	,03817				
		4	,033889	,016085	,227	-,00634	,07412				
		5	,035778	,016085	,181	-,00445	,07601				
3	1	,002611	,016085	1,000	-,03762	,04284					
		2	,002056	,016085	1,000	-,03817	,04228				
		4	,035944	,016085	,177	-,00428	,07617				
		5	,037833	,016085	,139	-,00240	,07806				
4	1	-,033333	,016085	,242	-,07356	,00690					
		2	-,033889	,016085	,227	-,07412	,00634				
		3	-,035944	,016085	,177	-,07617	,00428				
		5	,001889	,016085	1,000	-,03834	,04212				
5	1	-,035222	,016085	,194	-,07545	,00501					
		2	-,035778	,016085	,181	-,07601	,00445				
		3	-,037833	,016085	,139	-,07806	,00240				
		4	-,001889	,016085	1,000	-,04212	,03834				

A un nivel de significación del 1%, se puede decir que las zonas no difieren estadísticamente entre sí.

El **segundo objetivo** es comprobar si la longitud de las hojas sirve para predecir la asimetría.

Vemos el diagrama de dispersión de la variable independiente (longitud), con respecto a la dependiente (asimetría).

En la gráfica no se observa que se cumplan todas las hipótesis básicas de regresión (linealidad, homocedasticidad, normalidad e independencia). Lo interesante sería proceder a la transformación de los datos mediante logaritmos, cuadrados, etc. Incluso tras la transformación de los mismos se puede comprobar que sigue existiendo un diagrama de dispersión sin estructura alguna.

A pesar de no observarse una tendencia lineal y homocedastica en la distribución de los puntos, se estiman los parámetros de cada uno de los modelos siguientes (en que y siempre será la asimetría, y x será longitud):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i=1, \dots, n$$

Variables introducidas/eliminadas(b)

Modelo	Variables introducidas	Variables eliminadas	Método
1	Longitud(a)	.	Introducir

a Todas las variables solicitadas introducidas

b Variable dependiente: Asimetría

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,061(a)	,004	-,008	,050469

a Variables predictoras: (Constante), Longitud

Coeficientes(a)

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Error típ.	Beta		
				B						
1	(Constante)	,077		,023		3,307	,001			
	Longitud	-,001		,001	-,061	-,576	,566			

a Variable dependiente: Asimetria

La recta de regresión sería:

$$Y = 0,077 - 0,001 X$$

El valor del coeficiente de determinación $R^2 = 0,004$ indica que sólo el 0,4% de la variabilidad de la asimetría es explicada por su relación lineal con la longitud de las hojas.

El valor del coeficiente de correlación $r = 0,061$ no denota una buena asociación lineal positiva entre ambas variables.

Si planteamos el siguiente contraste para testar la significatividad de la variable independiente:

$H_0 : = 0$

$H_1 : \neq 0$

Los datos obtenidos son:

ANOVA(b)

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.	
1	Regresión	,001	1	,001	,332	,566(a)	
	Residual	,224	88	,003			
	Total	,225	89				

a Variables predictoras: (Constante), Longitud

b Variable dependiente: Asimetría

El p-valor = 0,566 es mayor que 0,05, por lo que no se rechaza la hipótesis nula de no linealidad del modelo, es decir, las variables no tienen relación lineal.

El **tercer objetivo** es pronosticar a qué clima es más probable que pertenezca una hoja de encina a la cual se le midió la longitud y resultó ser 25,62 y su asimetría 0,072. Cuantificar la probabilidad de que la hoja sea de uno y otro clima. ¿Es una predicción fiable?

Realizamos un análisis discriminante y obtenemos los siguientes datos:

Resumen de las funciones canónicas discriminantes

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	,154(a)	100,0	100,0	,365

a Se han empleado las 1 primeras funciones discriminantes canónicas en el análisis.

Si contrastamos H_0 : los centros son iguales

H_1 : los centros son distintos

Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,867	12,463	2	,002

A un nivel de significación de 0,05, el p-valor asociado al contraste = 0,002 es menor, por lo que se rechaza la hipótesis nula. Los centros no son iguales.

Coefficientes estandarizados de las funciones discriminantes canónicas

	Función	
		1
Longitud	,342	
Asimetría	,975	

La función discriminante canónica es:

$$D = 0,342 \cdot \text{longitud} + 0,975 \cdot \text{asimetría}$$

Matriz de estructura

	Función	
		1
Asimetría	,940	
Longitud	,244	

Correlaciones intra-grupo combinadas entre las variables discriminantes y las funciones discriminantes canónicas tipificadas Variables ordenadas por el tamaño de la correlación con la función.

El valor 0,940 indica una relación positiva fuerte entre la asimetría y pertenecer a un grupo u otro.

Coefficientes de las funciones canónicas discriminantes

	Función	
		1
Longitud	,061	
Asimetría	20,550	
(Constante)	-2,764	

Coefficientes no tipificados

Las variables longitud y asimetría presentan una correlación positiva con la función. Hay que tener en cuenta la constante, por lo que la función será por tanto:

$$D = 0,342 \cdot \text{longitud} + 0,975 \cdot \text{asimetría} - 2,764$$

El que se clasifiquen en un grupo u otro dependerá de si se obtiene un valor menor o mayor que el punto de corte cuando se calcule la puntuación discriminante.

Funciones en los centroides de los grupos

Clima	Función	
		1
1	,317	
2	-,475	

Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos

Estos son los valores discriminantes en el centro de cada grupo, lo que sirve para determinar el punto de corte de ambos grupos.

$$0,317 - 0,475 = - 0,079$$

2

Estadísticos de clasificación

Resumen del proceso de clasificación

Procesados		91
Excluidos	Código de grupo perdido o fuera de rango	0
	Perdida al menos una variable discriminante	0
Usados en los resultados		91

Total	1,000	90	90,000
2	Previas	500	ponderados
1	,500	No	54
		análisis	
		Casos utilizados en el	

Probabilidades previas para los grupos

Caso mal clasificado

**.							
		ado desagrup	1	,975	1	,572	,001
91							
90	2		2	,667	1	,658	,186
89	2		2	,860	1	,612	,031
88	2		1**	,559	1	,685	,341
87	2		1**	,838	1	,617	,042
86	2		2	,259	1	,770	1,275
85	2		2	,858	1	,543	,032
84	2		2	,828	1	,619	,047
83	2		2	,726	1	,644	,123
82	2		2	,438	1	,717	,603
81	2		2	,365	1	,737	,820
80	2		1**	,723	1	,508	,125
79	2		1**	,977	1	,572	,001
78	2		2	,746	1	,639	,105
77	2		1**	,995	1	,579	,000
76	2		2	,463	1	,710	,540
75	2		2	,714	1	,647	,135
74	2		2	,849	1	,541	,036
73	2		1**	,631	1	,667	,231
72	2		1**	,870	1	,546	,027
71	2		1**	,751	1	,516	,101
70	2		2	,166	1	,804	1,921
69	2		2	,506	1	,699	,442
68	2		1**	,467	1	,709	,529
67	2		2	,651	1	,662	,204
66	2		2	,887	1	,550	,020
65	2		2	,767	1	,520	,088
64	2		2	,600	1	,675	,275
63	2		2	,519	1	,695	,416
62	2		2	,973	1	,571	,001
61	2		1**	,742	1	,513	,108
60	2		2	,914	1	,557	,012
59	2		2	,995	1	,577	,000
58	2		2	,420	1	,722	,650
57	2		2	,781	1	,523	,078
56	2		1**	,818	1	,533	,053
55	2		2	,522	1	,694	,409
54	1		1	,938	1	,593	,006
53	1		1	,753	1	,516	,099
52	1		2*	,322	1	,750	,979
51	1		1	,080	1	,845	3,056
50	1		1	,918	1	,558	,011
49	1		1	,727	1	,509	,122
48	1		2*	,987	1	,581	,000
47	1		1	,883	1	,606	,022
46	1		2*	,748	1	,515	,104
45	1		1	,035	1	,879	4,439
44	1		1	,972	1	,585	,001
43	1		1	,078	1	,847	3,108
42	1		2*	,689	1	,653	,160
41	1		1	,328	1	,748	,955
40	1		1	,695	1	,501	,154
39	1		1	,920	1	,595	,000

En la tabla anterior se muestra para cada hoja, a qué grupo pertenece realmente, seguido del grupo pronosticado para el individuo. Los casos clasificados de forma incorrecta se señalan con **.

Clasificados correctamente el 65,6% de los casos agrupados originales.

a.		Casos desagrupados	100,0	,0	100,0
		2	30,6	69,4	100,0
%		1	63,0	37,0	100,0
		Casos desagrupados	1	0	1
		2	11	25	36
Clima		1	34	20	Total
Original Recuento		1	34	20	54
		2	11	25	
		Grupo de pertenencia			

Resultados de la clasificación

Se han clasificado correctamente el 65,6 % de los casos.

De las 54 hojas de clima húmedo, 34 han sido clasificadas correctamente, y 20 de ellas las clasifica en el grupo 2 (clima seco).

De las 36 hojas de clima seco, 25 han sido clasificadas correctamente, y 11 han sido clasificadas en clima húmedo.

La clasificación de la última hoja es en el grupo 1 (clima húmedo) con una probabilidad del 57,17 %. La puntuación discriminante es 0,28525.

La probabilidad de esta hoja de pertenecer al grupo 2 es de 42,83 %.

A continuación se adjunta el archivo de datos, en el que se han agregado las siguientes columnas:

dis_1: indica, para cada caso, el grupo pronosticado.

dis1_1: indica la puntuación discriminante de cada caso según la función discriminante.

dis1_2: indica la probabilidad de pertenencia al grupo 1.

dis2_2: indica la probabilidad de pertenencia al grupo 2.