Tema 3
ANOVA y tablas de contingencia
(Comparación de poblaciones)

1. Introducción

La Ecología se puede definir como la ciencia que estudia las relaciones entre los componentes bióticos y abióticos de la naturaleza. Gran parte de los estudios ecológicos tienen por finalidad investigar cómo responden las especies frente a las variables ambientales. Para ello, la Estadística dispone de métodos y técnicas útiles para detectar y describir estas relaciones. El objetivo de este tema, y el siguiente, es proporcionar un conocimiento adecuado de los procedimientos estadísticos que pueden ser empleados en la investigación ecológica.

Aunque en la mayoría de las técnicas que se estudiarán a continuación los procedimientos de cálculo se ejemplifican con datos reales, en la práctica debe recurrirse al uso del ordenador. El objetivo fundamental del tema es, por tanto, proporcionar un conocimiento adecuado acerca del significado de las diferentes técnicas, su correcta aplicación, y la interpretación de los resultados que proporcionan los programas informáticos (el manejo de algunos de éstos se aprenderá en las clases prácticas).

En esta introducción es conveniente hacer hincapié en el hecho de que la mayoría de las pruebas o tests estadísticos que se presentan en este tema requieren las siguientes condiciones para los datos:

- que sean independientes
- que sigan una distribución normal
- que sus varianzas sean homogéneas (homocedásticas) e independientes del valor de las variables ambientales.

En primer lugar es necesario que los datos sean independientes entre sí. Como se recordará del Tema 2, la no independencia es una consecuencia de la pseudorreplicación, y en los estudios de campo ocurre frecuentemente cuando los datos proceden de transectos o muestreos repetidos a lo largo del tiempo. En estos casos las observaciones consecutivas (en el espacio o en el tiempo) tienden a presentar valores similares entre sí. La solución a este problema hay que buscarla en un diseño del muestreo (o del experimento) que evite esta circunstancia.

El segundo supuesto, de normalidad, rara vez se cumple en los datos que maneja el ecólogo, pero pueden encontrarse soluciones mediante la transformación de los mismos. En concreto, el empleo de logaritmos (decimales o neperianos) produce resultados satisfactorios en la mayoría de los casos.

Por último, en los datos biológicos es frecuente observar cómo la varianzas no son homogéneas, sino que aumentan generalmente con el valor de la media; esta falta de independencia entre varianzas y medias debe evitarse para una correcta aplicación de numerosos tests estadísticos (los denominados **test paramétricos**). Como se verá en el siguiente ejemplo, la transformación de los datos suele ser también eficaz para evitar el incumplimiento de este último requisito.

Los datos que aparecen en la Tabla 1 forman parte de un muestreo de aves realizado en diversas sierras de la Región de Murcia. En concreto se presenta el número de individuos de *Sylvia undata* (Curruca rabilarga) contados en 5 transectos realizados en 4 sierras diferentes.

Autor: Carlos González Revelles

Tabla 1. Número de individuos de *Sylvia undata* presentes en muestras de cuatro sierras de la Región de Murcia. Las varianzas de los datos originales son mayores cuanto mayores son las medias. La transformación logarítmica, ln(x), corrige esta circunstancia.

DATOS ORIGINALES						$\bar{x}$	s^2
SIERRA DE LA TERCIA	9	4	6	1	6	5,2	8,7
SIERRA DE LA TORRECILLA	2	9	9	4	5	5,8	9,7
SIERRA DE LA MUELA	2	7	6	8	20	8,6	45,8
SIERRA DE CARRASCOY	35	18	21	8	8	18,0	124,5
DATOS TRANSFORMADOS						$\bar{x}$	s^2
SIERRA DE LA TERCIA	2,20	1,39	1,79	0,00	1,79	1,43	0,72
SIERRA DE LA TORRECILLA	0,69	2,20	2,20	1,39	1,61	1,62	0,40
SIERRA DE LA MUELA	0,69	1,95	1,79	2,08	3,00	1,90	0,68
SIERRA DE CARRASCOY	3,56	2,89	3,04	2,08	2,08	2,73	0,41

En definitiva, una transformación previa de los datos es, en la mayoría de los casos, imprescindible para la correcta aplicación de numerosos tests. En los diferentes ejemplos que se analizan en este tema se utilizará, en general, la transformación ln(x+1) si hay ceros en los datos, o ln(x) si no los hay. Cuando los datos no cumplen los requisitos mencionados, incluso después de transformados, pueden utilizarse pruebas alternativas (test **no paramétricos**). Los tests no paramétricos no son exigentes con las características de los datos y su utilización es cada vez más frecuente en la investigación científica. La mayor complejidad de su cálculo ya no representa un inconveniente con el uso generalizado de ordenadores y paquetes estadísticos.

Antes de iniciar el estudio de este tema es conveniente recordar las nociones elementales relacionadas con el contraste de hipótesis y los errores estadísticos (tipo I y tipo II). Los errores tipo I ocurren cuando se rechaza una hipótesis nula verdadera, mientras que los errores tipo II se cometen cuando no se rechaza una hipótesis nula que en realidad es falsa (Figura 1). Usualmente, los científicos trabajan minimizando el error tipo I, y de forma generalizada se establece el nivel de significación 0,05 como criterio de decisión para el rechazo de la hipótesis nula. Los valores de probabilidad (P) que proporcionan los tests estadísticos representan precisamente la probabilidad de obtener nuestros datos siendo cierta la hipótesis; de esta forma se considera que un valor de P menor que 0,05 es suficientemente bajo como para rechazarla.

Las diferentes pruebas estadísticas que se presentan a continuación pueden clasificarse según la naturaleza cualitativa o cuantitativa de los datos biológicos y ambientales. Así, para el desarrollo de este tema y el siguiente se seguirá el esquema de la Figura 2.

	Decisión:	
Realidad:	**No rechazar H_0**	**Rechazar H_0**
H_0 cierta	Decisión correcta (probabilidad = $1-\alpha$)	Error Tipo I (probabilidad = α)
H_0 falsa	Error Tipo II (probabilidad = β)	Decisión correcta (probabilidad = $1-\beta$)

Figura 1. Errores estadísticos tipo I y tipo II.

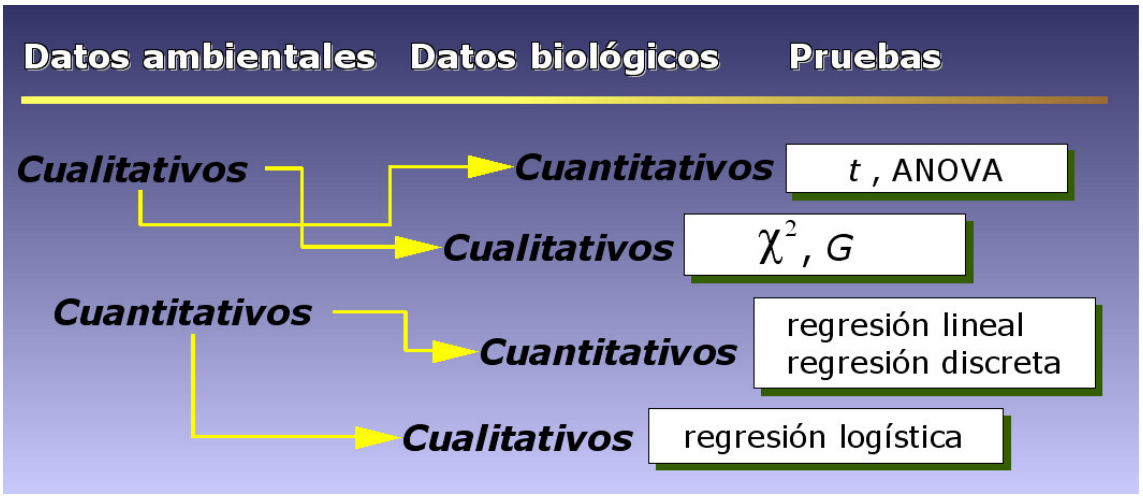

Figura 2. Pruebas estadísticas aplicables en la investigación ecológica, clasificadas según las características de los datos biológicos y ambientales.

2. Test de la *t* y análisis de la varianza (ANOVA)

En numerosas ocasiones el ecólogo se plantea la comparación entre poblaciones que presentan características ambientales diferentes. Supóngase, por ejemplo, que se quiere comparar las densidades de una planta en diferentes tipos de sustratos. En este caso, la variable ambiental (tipo de sustrato) es de naturaleza cualitativa, y presenta un determinado número de modalidades (por ejemplo: calizas, arcillas y arenas). El objetivo de la investigación se centra en determinar si existen diferencias significativas de abundancia entre los distintos tipos de sustratos, es decir, si la planta en cuestión se ve afectada por la naturaleza del sustrato. En Estadística, una variable ambiental de tipo cualitativo o nominal se denomina **factor**, aunque en Ecología este término se puede utilizar indistintamente para variables ambientales cualitativas o cuantitativas.

Para analizar las relaciones entre poblaciones y factores ambientales existen diferentes técnicas estadísticas que se comentan a continuación. En el caso más simple, el factor ambiental presenta únicamente dos modalidades mutuamente excluyentes. El

método tradicional para comparar dos medias es el **test de la *t***. Este estadístico sigue la distribución de la *t* de Student, y se calcula según la expresión:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{(n_1 - 1)\cdot s_1^2 + (n_2 - 1)\cdot s_2^2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 \cdot n_2}}} \quad (1)$$

A pesar de su aparente complejidad, cuando los tamaños de las muestras son grandes (>30), la expresión (1) puede simplificarse notablemente:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_2} + \frac{s_2^2}{n_1}}} \quad (2)$$

Y en cualquier caso, siempre que los tamaños de muestra sean iguales ($n = n_1 = n_2$), la ecuación (1) se reduce a:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}} \quad (3)$$

La hipótesis nula (H_0) para este test es que las medias son iguales, y los grados de libertad que hay que considerar son $\nu = (n_1 + n_2 - 2)$ para las dos primeras expresiones, y $\nu = (2n - 2)$ para la última.

Por otra parte, los límites de confianza para la diferencia entre medias se pueden calcular según

$$I_1 = (\overline{x}_1 - \overline{x}_2) - t_{\alpha[\nu]} s_{\overline{x}_1 - \overline{x}_2} \quad (4a)$$

$$I_2 = (\overline{x}_1 - \overline{x}_2) + t_{\alpha[\nu]} s_{\overline{x}_1 - \overline{x}_2} \quad (4b)$$

siendo $t_{\alpha[\nu]}$ el valor de la distribución *t* de Student para el nivel de significación α y ν grados de libertad, y $s_{\overline{x}_1 - \overline{x}_2}$ el error típico de la diferencia entre dos medias, es decir, el denominador de la expresión (1), (2) ó (3).

Antes de su aplicación, hay que tener en cuenta que el test de la *t* requiere, como ya se ha comentado anteriormente, normalidad en los datos e igualdad de varianzas para ambas variables (solana y umbría). Para comprobar la normalidad puede utilizarse el denominado **test de Shapiro-Wilk**, de cálculo complejo, mientras que para comprobar que las varianzas poblacionales no son significativamente distintas puede recurrirse el **test de la *F***. Este último test es aplicable en el caso simple de dos únicas varianzas, y se calcula como el cociente entre ambas:

$$F = \frac{s_1^2}{s_2^2} \quad (5)$$

Los grados de libertad son n_1-1 en el numerador y n_2-1 en el denominador. La H_0 en este test es que las varianzas poblacionales son iguales (y, por tanto, el cociente igual a 1). El siguiente ejemplo servirá para ilustrar la aplicación del test de la *t* y el test de la *F*.

Como parte de una investigación sobre las causas de la expansión del alga *Caulerpa prolifera* en el Mar Menor, se realizó un experimento consistente en someter fragmentos de talo a diferentes tratamientos en el laboratorio, con objeto de analizar su crecimiento. Los datos que se presentan en la Tabla 2 se obtuvieron midiendo el crecimiento de los fragmentos en dos grupos de acuarios. El primer grupo (acuarios control) se mantuvo exclusivamente con agua de mar, mientras que el otro grupo fue "fertilizado" (*tratado*) con una determinada cantidad de nitrógeno. La cuestión que se plantea resolver es si la adición de este recurso influencia el crecimiento de *Caulerpa*.

Tabla 2. Crecimiento (cm) de fragmentos de talo de *Caulerpa prolifera* en acuarios control y acuarios con tratamiento de nitrógeno.

Núm.	Control	Nitrógeno
1	3,32	8,47
2	2,52	4,70
3	4,05	3,08
4	3,47	4,46
5	1,78	5,19
6	2,38	8,00
7	1,42	5,31
8	1,89	2,81
9	4,21	5,21
10	1,63	6,00
11	2,74	9,28
12	5,92	20,12
13	4,76	20,99
14	8,42	6,57
15	3,32	5,35
16	5,00	6,20
17	2,08	6,85
18	7,59	3,67
19	4,37	4,71
20	1,91	4,43
21	2,52	2,56
22	2,32	4,71
23	5,79	6,86
24	4,76	2,82
25	3,23	4,14
26	1,92	1,21
27	4,36	15,83
28	6,26	4,91

Fuente: http://es.wikipedia.org/wiki/Caulerpa

Para el análisis de los datos seguiremos el procedimiento comentado anteriormente. En primer lugar aplicaremos el test de normalidad de Shapiro-Wilk, que proporciona los siguientes resultados:

- *control:* $W = 0,9171$; $P = 0,0294$
- *nitrógeno:* $W = 0,7295$; $P = 0,0000$

Dado que en ambos casos el valor de P es inferior a 0,05 se rechaza la hipótesis nula y se concluye que las variables en cuestión no siguen una distribución normal. No obstante, mediante una transformación logarítmica de los datos sí se obtiene un ajuste significativo:

- ln(*control*): $W = 0,9695$; $P = 0,5677$
- ln(*nitrógeno*): $W = 0,9437$; $P = 0,1371$

El siguiente paso, por tanto, es comprobar la homogeneidad de varianzas de los datos transformados:

- $F = \frac{s_1^2}{s_2^2} = \frac{0,2391}{0,3660} = 0,6534$; $P = 0,2750$

Aceptado el requisito de homocedasticidad, puede aplicarse finalmente la prueba de la *t*:

- $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2 + s_2^2}{n}}} = \dfrac{1{,}1966 - 1{,}6969}{\sqrt{\dfrac{0{,}2391 + 0{,}3660}{28}}} = \dfrac{-0{,}5003}{0{,}1470} = -3{,}4027\,; \quad P = 0{,}0013$

Con estos resultados se rechaza la hipótesis nula de igualdad de medias y se concluye que el aporte de nitrógeno al medio favorece el crecimiento de los talos de *Caulerpa*.

A continuación se presenta otro ejemplo para la comparación de medias procedentes de dos poblaciones:

En un muestreo realizado por alumnos de la asignatura de Ecología en el Parque Regional de Calblanque, se pretendía analizar las diferencias existentes entre las abundancias de diversas plantas en laderas orientadas al norte (umbrías) y laderas con orientación sur (solanas). Para cada tipo de ladera se tomó un total de 20 unidades de muestreo de 2 x 2 m. Una de las especies muestreadas fue el tomillo (*Thymus hyemalis*), de la que se obtuvieron los datos (número de individuos por unidad de muestreo) que aparecen en la Tabla 3.

Tabla 3. Número de tomillos (*Thymus hyemalis*) presentes en muestras de laderas con exposición sur (solanas) y exposición norte (umbrías), en el Parque Regional de Calblanque.

SOLANA	15	6	15	0	10	24	0	0	5	7	4	17	4	5	1	0	14	3	6	10
UMBRÍA	0	0	5	6	0	6	11	0	3	11	7	0	3	8	5	5	1	1	0	4

El objetivo que se plantea es conocer si el factor orientación determina la existencia de diferencias significativas entre las poblaciones de *Thymus hyemalis* en solanas y umbrías. Desde el punto de vista estadístico, el problema es determinar si ambos conjuntos de datos pertenecen a la misma población, es decir, si presentan la misma media e igual varianza.

En primer lugar debe comprobarse la normalidad de los variables, por lo que se aplica el test de Shapiro-Wilk a ambas variables. Como resultado se obtiene:

- *solana:* $W = 0{,}9003$; $P = 0{,}0418$
- *umbría:* $W = 0{,}8852$; $P = 0{,}0220$

De esta forma se comprueba que los datos no se ajustan a una distribución normal. La transformación de los datos mediante la expresión ln(*x*+1) tampoco produce resultados satisfactorios:

- ln(*solana*+1): $W = 0{,}8964$; $P = 0{,}0353$
- ln(*umbría*+1): $W = 0{,}8547$; $P = 0{,}0064$

Dadas las circunstancias, en este caso no debería utilizarse ningún test paramétrico, y como alternativa puede aplicarse un test no paramétrico: el **test de suma de rangos de Wilcoxon** (equivalente al **test de Mann-Withney**). Esta prueba permite comparar las medias de dos variables que no cumplen los requistos de normalidad u homocedasticidad. Así, los resultados del test (no hace falta utilizar los datos transformados), son los siguientes:

- $W = 143{,}5$; $P = 0{,}126$

A la vista del valor de probabilidad se concluye que no existen diferencias significativas entre las medias de solana y umbría: las dos poblaciones no son significativamente distintas y, por tanto, la orientación no influye en la abundancia de tomillos.

Los ejemplos analizados anteriormente representan el caso más sencillo de comparación de dos medias, pero en numerosas ocasiones es necesario hacer comparaciones entre más de dos poblaciones. En estos casos no es posible aplicar el test de la *t*, por lo que hay que recurrir al denominado **análisis de la varianza** o **ANOVA** (*ANalysis Of VAriance*). De entre los numerosos métodos de ANOVA, presentaremos aquí únicamente el más sencillo, conocido como la **clasificación única del análisis de la varianza** (*one way ANOVA*), en el que los grupos se clasifican según un único factor o tratamiento (Figura 3). El análisis de la varianza es una generalización del test de la *t*, pero también puede emplearse para analizar las diferencias entre dos medias; de hecho, en estos casos se obtienen los mismos resultados que con el test de la *t*.

Los fundamentos del análisis de la varianza pueden expresarse mediante una ecuación o modelo de la forma

$$y_{ij} = \mu + T_j + e_{ij} \quad (6)$$

o bien

$$y_{ij} - \mu = T_j + e_{ij} \quad (7)$$

donde y_{ij} representa el valor observado en la unidad experimental o de muestreo *i* correspondiente al tratamiento o factor *j*, μ es la media global, T_j es el efecto debido al tratamiento o factor *j* y e_{ij} es el error asociado a cada observación.

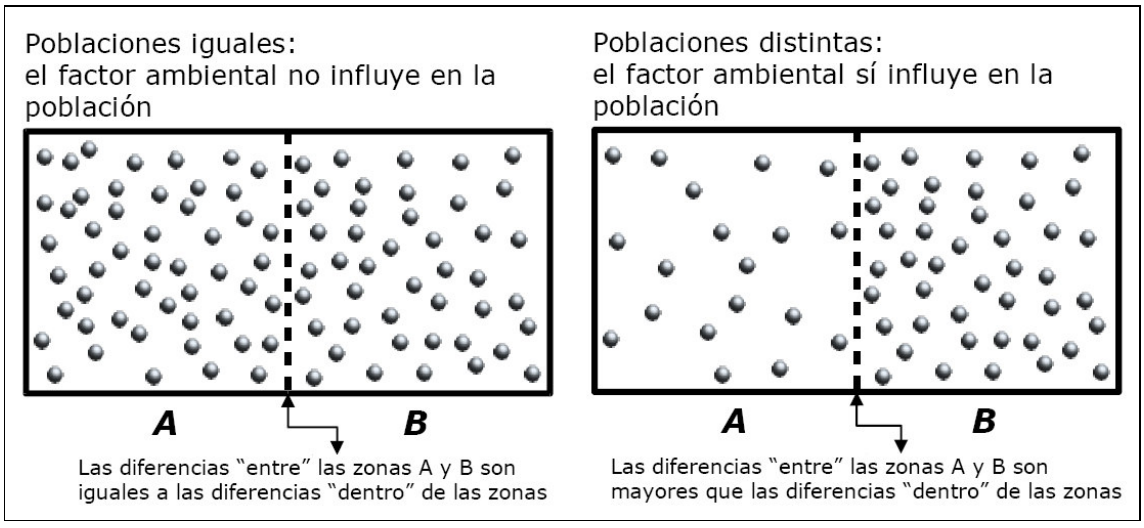

Figura 3. Esquema representativo de los fundamentos comparativos del análisis de la varianza.

Para ilustrar el cálculo de esta prueba se emplearán los datos de la Tabla 4, donde se muestran los valores de crecimiento radicular (en mm) de plántulas de *Halocnemum strobilaceum*, en condiciones experimentales de laboratorio, sometidas a diferentes tratamientos de salinidad.

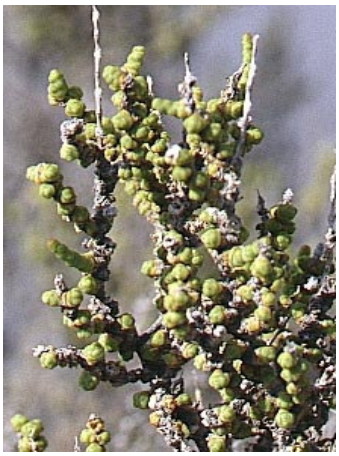

Tabla 4. Crecimiento de raíces de plántulas de *Halocnemum strobilaceum* (en mm) sometidas a diferentes tratamientos de salinidad.

Núm.	Control	Salinidad baja	Salinidad media	Salinidad alta
1	20	31	29	16
2	27	38	26	18
3	23	28	25	19
4	27	36	25	23
5	29	25	27	12
6	25	25	25	19
7	24	32	21	16
8	29	28	24	19
9	23	28	22	21
10	22	30	28	19

Puede comprobarse previamente, uilizando la prueba de Shapiro-Wilk, que los datos se ajustan a una distribución normal. Sin embargo, en este caso, para contrastar la homocedasticidad se empleará el **test de Bartlett**, que permite comparaciones de varianzas entre dos o más grupos:

- $\chi^2 = 5{,}6191$; $\nu = 3$; $P = 0{,}1317$

El valor de probabilidad, mayor de 0,05, sugiere la aceptación de la hipótesis nula de igualdad de varianzas, por lo que finalmente se puede aplicar el ANOVA sin necesidad de transformar los datos. Los resultados presentan en la Tabla 5:

Tabla 5. Tabla del ANOVA de los datos de crecimiento de raíces de *Halocnemum strobilaceum*. [g.l.: grados de libertad; s.c.: suma de cuadrados; m.c.: media de cuadrados.]

Fuente de variación	g.l.	s.c.	m.c.	*F*	*P*
Entre grupos	3	821,67	273,89	21,52	3,70e-08
Dentro de grupos	36	458,10	12,72		
Total	39	1279,77	32,81		

Para conocer cómo se calculan y qué significan los diferentes valores de esta tabla, recurriremos a una explicación que permita comprender el método de forma intuitiva. En primer lugar, se puede expresar cada dato de la Tabla 4 como la suma de la media del grupo (tratamiento) al que corresponde más un término de error:

$$\textit{valor observado} = \textit{media estimada de cada grupo} + \textit{error} \qquad (8)$$

De esta forma, por ejemplo, el primer valor de la Tabla 4 puede escribirse como:

20,00 = 24,90 – 4,90

y el último:

19,00 = 18,20 + 0,80

Para cada uno de los términos de la ecuación (8) ha de calcularse la **suma de cuadrados** (**s.c.**), restando previamente la media del total de datos (24,83 en este caso) a los valores observados y a las medias estimadas de cada grupo; de esta forma, siguiendo con el ejemplo, las expresiones anteriores quedarían de la siguiente forma:

$(20{,}00 - 24{,}83) = (24{,}90 - 24{,}83) - 4{,}90$

$(19{,}00 - 24{,}83) = (18{,}20 - 24{,}83) + 0{,}80$

[Nótese que estas dos últimas expresiones se corresponden con la ecuación (7).] La suma de cuadrados total es la suma de cuadrados correspondiente a primer término:

$(20{,}00 - 24{,}83)^2 + (27{,}00 - 24{,}83)^2 + ... + (19{,}00 - 24{,}83)^2 = 1279{,}77$

La suma de cuadrados entre grupos es la que corresponde al segundo término:

$10\times(24{,}90 - 24{,}83)^2 + ... + 10\times(18{,}20 - 24{,}83)^2 = 821{,}67$

Finalmente, la suma de cuadrados dentro de grupos es la correspondiente al término de errores (residuos):

$4{,}90^2 + 2{,}18^2 + ... + 0{,}80^2 = 458{,}10$

Los grados de libertad son n-1 para la s.c. total (n es el número total de observaciones), q-1 para la s.c. entre grupos (siendo q el número de grupos), y n-q para la s.c. dentro de grupos. Las **medias de cuadrados** (**m.c.**) se calculan dividiendo la correspondiente suma de cuadrados por sus grados de libertad. Nótese que las m.c. son varianzas, por lo que la m.c. dentro de grupos se denomina también **varianza residual**, y la m.c. total es la varianza total de los datos.

El valor del estadístico F se obtiene dividiendo la varianza entre grupos por la varianza residual, y debe compararse con los valores críticos de la distribución F. Si las medias de cada población son iguales, la variación debida a las diferencias entre grupos no será mucho mayor que la variación dentro de los grupos, y el cociente F tendrá un valor próximo a 1. Este test, a diferencia del test de la F para comparación de varianzas, es de una cola, ya que la hipótesis alternativa es: *m.c. entre grupos > m.c. dentro de grupos*.

Para el ejemplo, F = 21,52 con 3 g.l. en el numerador y 36 en el denominador. La probabilidad que se obtiene es prácticamente 0, por lo que debemos rechazar la hipótesis nula y concluir que las medias de crecimiento radicular de las plántulas sometidas a diferentes tratamientos no son iguales. El problema añadido cuando existen más de dos medias a comparar es determinar cuáles de esas medias son distintas entre sí. Se necesita, por tanto, algún tipo de prueba adicional que permita responder a esta pregunta.

Los test de comparación de medias que se aplican tras un ANOVA significativo se fundamentan en el cálculo de límites de confianza para dichas medias. Uno de los más utilizados es el **test de Tukey**. Su aplicación en este caso proporciona los siguientes intervalos:

Tabla 6. Resultados del test de Tukey aplicado tras el análisis de la varianza de los datos de Tabla 4. Los intervalos que contienen el 0, corresponden a medias de grupos (tratamientos) que no son significativamente distintas entre sí.

Grupos a comparar	Diferencia	Límite inferior	Límite superior
1 – 0 *	6,1	1,8035	10,3965
2 – 0	0,3	-3,9965	4,5965
3 – 0 *	-6,7	-10,9965	-2,4035
2 – 1 *	-5,8	-10,0965	-1,5035
3 – 1 *	-12,8	-17,0965	-8,5035
3 – 2 *	-7,0	-11,2965	-2,7035

* Medias significativamente distintas entre sí (P<0,05).

Como alternativa no paramétrica al ANOVA, en los casos en los que se incumplan los requisitos de normalidad y homogeneidad de varianzas, puede utilizarse el denominado **test de Kruskal-Wallis**.

Hasta ahora todos los ejemplos analizados consideraban únicamente un factor de clasificación (una única variable ambiental de naturaleza cualitativa). Sin embargo, los diseños de muestreos o experimentos pueden incorporar más de una variable ambiental. En estos casos, el análisis estadístico requiere la utilización de técnicas de ANOVA más complejas.

En el siguiente ejemplo (Tabla 7) se consideran los datos de un experimento (diseño de bloques completos al azar) en el que se investigó la respuesta de crecimiento de la planta *Eriophorum angustifolium* a cuatro tratamientos de fertilización (N, N+P, N+P+K y control) en cinco localidades de tundra (B, M, R, S, Q) en Alaska. En este caso, el modelo del ANOVA es:

$$y_{ijk} = \mu + T_j + B_k + e_{ijk} \qquad (9)$$

donde B_k representa el efecto asociado al bloque *k*. Los resultados del ANOVA se presentan en la Tabla 8.

Tabla 7. Crecimiento de *Eriophorum angustifolium* en cinco localidades de tundra en Alaska y bajo diferentes tratamientos de fertilización.

	Localidad				
Tratamiento	B	M	R	S	Q
Control	10	6	11	2	5
N	58	45	55	50	37
N + P	63	43	68	41	39
N + P + K	68	47	63	43	40

Tabla 8. Resultados del ANOVA de dos vías aplicado a los datos de la Tabla 7. [g.l.: grados de libertad; s.c.: suma de cuadrados; m.c.: media de cuadrados.]

Fuente de variación	g.l.	s.c.	m.c.	*F*	*P*
Tratamiento	3	7241,8	2413,9	79,31	3,52e-08
Localidades (bloques)	4	1335,2	333,8	10,97	5,61e-04
Error	12	365,2	30,4		
Total	19	8942,2	470,64		

En este caso, existen diferencias significativas entre tratamientos y entre bloques (localidades). Aunque las diferencias entre estas últimas no son de interés para la investigación, su efecto sí debe considerarse en el análisis para obtener unos resultados adecuados al diseño en bloques del experimento.

3. Test de χ^2 y test de la *G*

En numerosas ocasiones la información que se obtiene en el campo o en el laboratorio no es cuantitativa. Muchas respuestas biológicas y ecológicas son de naturaleza cualitativa, no requieren cuantificación, o es excesivamente costoso obtenerla. Datos sobre el éxito reproductor de una especie, sobre la composición de su dieta, o simplemente sobre su presencia o ausencia en determinadas áreas, requieren un tratamiento estadístico distinto al planteado en el apartado anterior. Cuando las variables ambientales son

también de tipo cualitativo, los datos se presentan en forma de **tablas de contingencia**, con *R* filas y *C* columnas. Este tipo de tablas resumen, en definitiva, información sobre frecuencias (proporciones) de aparición u ocurrencia de los diferentes tipos de respuesta biológica considerados (por ejemplo la frecuencia de aparición de una especie en un tipo de sustrato, o la proporción de reptiles en la dieta de un depredador en diferentes tipos de hábitat).

Los datos de la Tabla 9 correspondientes a un muestreo de vegetación en el Parque Regional de Calblanque en el que se pretendía analizar la relación entre la frecuencia de aparición de diferentes especies en diferentes tipos de sustrato (calizas y pizarras) presentes en la zona. En concreto los datos que se presentan corresponden a la presencia de individuos de *Sedum sediforme* en 35 unidades de muestreo (cuadrados de 2x2 m).

Tabla 9. Presencia de *Sedum sediforme* en unidades de muestreo sobre sustrato de calizas y pizarras en el Parque Regional de Calblanque.

	Presencias	Ausencias	
Calizas	5	10	15
Pizarras	13	7	20
	18	17	35

Un test apropiado para analizar estos datos es el **test de χ^2** (chi-cuadrado o ji-cuadrado), que se calcula mediante la expresión

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} \qquad (10)$$

donde o_i y e_i son respectivamente los valores observados y esperados de cada celda de la tabla. El valor esperado de una celda se calcula multiplicando el total de la fila y columna correspondiente y dividiendo por el total de la tabla; en nuestro ejemplo, los valores esperados son los que aparecen en la Tabla 10.

Tabla 10. Valores de frecuencias esperadas correspondientes a los datos de la Tabla 7.

	Presencias	Ausencias	
Calizas	(15·18) / 35 = 7,71	(15·17) / 35 = 7,29	15
Pizarras	(20·18) / 35 = 10,29	(20·17) / 35 = 9,71	20
	18	17	35

El valor de χ^2 sería, por tanto:

- $\chi^2 = (5 - 7{,}71)^2 / 7{,}71 + \ldots + (7 - 9{,}71)^2 / 9{,}71 = 3{,}44$

que debe compararse con los valores críticos de la distribución Chi-cuadrado con 1 grado de libertad. Como $\chi^2_{0,05[1]} = 3{,}84$ es mayor que el valor obtenido, aceptamos la hipótesis nula de independencia, es decir, que la presencia de *Sedum sediforme* es independiente del tipo de sustrato.

En el caso de tablas de 2 x 2, cuando $n < 200$, se recomienda realizar la corrección de Yates, que consiste en sumar o restar 0,5 a los valores observados, según el siguiente criterio:

- cuando $a \cdot d > b \cdot c$: $a - 0{,}5$; $d - 0{,}5$; $b + 0{,}5$; $c + 0{,}5$
- cuando $a \cdot d < b \cdot c$: $a + 0{,}5$; $d + 0{,}5$; $b - 0{,}5$; $c - 0{,}5$

donde los valores a, b, c y d se corresponden en la tabla con el siguiente esquema:

a	*b*	*a+b*
c	*d*	*c+d*
a+c	*b+d*	*n*

Para el ejemplo de *Sedum sediforme*, la aplicación de la corrección de Yates modificaría la Tabla 9 de la siguiente forma:

Tabla 11. Frecuencias observadas correspondientes a la Tabla 7 una vez aplicada la corrección de Yates.

	Presencias	Ausencias	
Calizas	5,5	9,5	15
Pizarras	12,5	7,5	20
	18	17	35

El valor de χ^2 corregido sería de 2,29, que no modifica la conclusión anterior.

En el siguiente ejemplo se analizará otra tabla de contingencia, en este caso de 4x3. Se trata de un estudio sobre la influencia de los diferentes periodos de la estación reproductora en el tipo de presas capturadas por Aguiluchos Cenizos (*Circus pygargus*) en el espacio natural de Ajauque y Rambla Salada. Los datos se presentan en la Tabla 12.

Tabla 12. Número de presas de tres categorías (aves, reptiles e invertebrados), capturadas por Aguiluchos Cenizos (*Circus pygargus*) en diferentes periodos de la estación reproductora. Abril: celo; mayo: incubación; junio: crianza de pollos; julio: periodo de dependencia de los jóvenes.

	aves	reptiles	invertebrados
abril	22	4	11
mayo	31	7	6
junio	15	4	15
julio	13	1	8

Autor: Carlos González Revelles

En este caso, la variable dependiente es el tipo de presa capturada, y el factor ambiental es el período de tiempo considerado. El análisis de χ^2 proporciona el siguiente resultado:

- $\chi^2 = 10{,}54$; $\nu = 6$; $P = 0{,}1037$

Como consecuencia se acepta la hipótesis nula, y se concluye que la composición de la dieta es independiente del período de la estación reproductora. Como regla general, los grados de libertad para una tabla de contingencia son $\nu = (R - 1) \cdot (C - 1)$.

Como alternativa a la prueba de χ^2 existe la posibilidad de utilizar el denominado **test de la *G***, que proporciona resultados similares. La expresión para el cálculo del estadístico *G*, aplicado a tablas de contingencia, es la siguiente:

$$G = 2 \cdot \left[\left(\sum f_i \cdot \ln f_i\right) - \left(\sum f_t \cdot \ln f_t\right) + n \cdot \ln n\right] \qquad (11)$$

donde f_i son los valores observados de cada celda de la tabla y f_t son los totales de filas y columnas. Para el ejemplo de *Sedum*, la aplicación del test de la *G*, con la correción de Yates (Tabla 9), proporciona:

- $G = 2\,[(5{,}5 \ln 5{,}5 + 12{,}5 \ln 12{,}5 + 9{,}5 \ln 9{,}5 + 7{,}5 \ln 7{,}5) -$
  $(18 \ln 18 + 17 \ln 17 + 15 \ln 15 + 20 \ln 20) + 35 \ln 35] = 2{,}31$

El valor de *G* hay que compararlo con $\chi^2_{0,05[1]} = 3{,}81$, por lo que se acepta H_0, comprobándose además que los resultados de ambos tests son muy parecidos. Para el ejemplo de las presas del Aguilucho Cenizo (Tabla 10), el resultado que se obtiene es:

- $G = 11{,}36$; $\nu = 6$; $P = 0{,}0779$

Hay que señalar que los tests de *G* y la χ^2 no deben utilizarse en aquellos casos en los que el tamaño de la muestra es muy pequeño ($n < 25$). Así mismo, cuando el valor esperado para una celda es menor de 1 (o menor de 5, según otros autores) es conveniente agrupar dos o más clases de la variable ambiental, de forma que los nuevos valores esperados superen esa cifra.