6. Análisis de datos en ecología (II): ANOVA y tablas de contingencia.

Ecología Metodológica y Cuantitativa (5C1)
Departamento de Ecología e Hidrología

Curso 2008–09

Índice

Antes de empezar

Iniciar R. Cargar el archivo de funciones de la asignatura, `funciones.R`:

```
source("http://www.um.es/docencia/emc/datos/funciones.R")
```

1. Análisis de la varianza

Ejercicios. Bloque 1:

Otra prueba para comparar las diferencias de medias entre dos poblaciones es el **análisis de la varianza** (**ANOVA**), que tiene la ventaja, entre otras, de ser aplicable en el caso general de más de dos poblaciones o tratamientos. De igual forma, el **test de Bartlett** permite contrastar la homogeneidad de más de dos varianzas. A continuación, siguiendo con los datos analizados en la sesión anterior correspondientes al fichero `caulerpa`, utilizaremos ambas pruebas con los datos transformados.

```
read.table("http://www.um.es/docencia/emc/datos/caulerpa.dat")->caulerpa
attach(caulerpa)
```

Aplicaremos primero el **test de Bartlett**:

```
bartlett.test(log(crec)~trat)
```

y luego el **ANOVA**:

```
aov(log(crec)~trat)->caulerpa.aov; summary(caulerpa.aov)
```

Cuando se aplica un ANOVA es conveniente asignar un nombre al análisis (por ejemplo `caulerpa.aov`); de esta forma, los resultados podrán emplearse posteriormente. La función `summary` proporciona la información detallada de los resultados del análisis.

Como puede apreciarse, las conclusiones son idénticas a las alcanzadas con las pruebas realizadas en el ejercicio 3 de la sesión anterior.

```
t.test(log(crec)~trat,var.equal=T)
```

1. Comprueba que los valores de probabilidad del test de la *t* y el de ANOVA son idénticos.
2. Comprueba que el cuadrado del valor de *t* es igual a la F de la prueba de ANOVA.

Ejercicios. Bloque 2:

Plantearemos en este ejercicio un caso de comparación de medias entre tres poblaciones. Los datos (archivo `pinos.dat`) proceden de un muestreo sobre las características del bosque como hábitat de aves rapaces. En concreto se presentan las estimas de dbh medio de los pinos presentes en parcelas localizadas en en tres tipos de ambientes: zonas llanas (1), laderas (2) y cumbres (3). Recuerda utilizar la función `factor()`.

Realizaremos las pruebas habituales: test de normalidad, test de Bartlett y ANOVA, transformando los datos si fuese necesario.

En el caso de comparar más de dos medias, si la prueba de ANOVA resulta significativa, surge un problema adicional: hay que determinar qué medias son distintas entre sí. Podemos utilizar el test de Tukey:

```
TukeyHSD(pinos.aov)
plot(TukeyHSD(pinos.aov))
```

Esta función proporciona las diferencias entre los posibles pares de medias y el intervalo de confianza (al 95 %) de dicha diferencia. Al igual que en el caso de la t, si el intervalo contiene el valor 0 concluiremos que las medias en cuestión no son significativamente distintas.

También para estos casos de comparación de más de dos medias existe una prueba no paramétrica alternativa al ANOVA. Se trata del **test de Kruskal-Wallis**, que es una generalización del **test de Mann-Whitney**. Podemos aplicar esta prueba a los datos originales:

```
kruskal.test(dbh~factor(zona))
```

1. ¿Cual es la interpretación ecológica de los resultados?
2. Idea un método para representar los resultados del test de Tukey.

2. Análisis de tablas de contingencia

Las tablas de contingencia son el resultado de la obtención de datos correspondientes a estudios observacionales o experimentales en los que se investiga las relaciones entre dos variables de carácter cualitativo.

Analizaremos como ejemplo la siguiente tabla en la que se presenta el número de presencias y ausencias (variable biológica cualitativa) de *Sedum sediforme* en unidades de muestreo (cuadrados de 2x2 m) localizadas en función de las características del sustrato (variable ambiental cualitativa):

	presencias	ausencias	
calizas	5	10	15
pizarras	13	7	20
	18	17	35

Las celdas de la tabla son las frecuencias observadas (o_{ij}), que suman un total de $n = 35$ unidades de muestreo. En estas tablas, la hipótesis nula considerada es la independencia (no relación) de ambas variables; bajo este supuesto, las frecuencias esperadas (e_{ij}) se calculan como el producto de las marginales (totales de filas y columnas) dividido por el total (n). Para contrastar esta hipótesis de independencia utilizaremos dos pruebas estadísticas: el test de χ^2 y el test de la G.

El valor experimental de χ^2 se calcula mediante:

$$\chi^2_{exp} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Este estadístico sigue una distribución χ^2 con $\nu = (R-1) \times (C-1)$ grados de libertad (R = número de filas y C = número de columnas).

Por su parte, el estadístico G se calcula según:

$$G = 2\left[\left(\sum o_{ij} \ln o_{ij}\right) - \left(\sum o_t \ln o_t\right) + n \ln n\right]$$

donde o_t son los totales de filas y columnas.

Ejercicios. Bloque 3:

Leeremos el archivo de datos correspondiente a la tabla de *Sedum:*

```
read.table("http://www.um.es/docencia/emc/datos/sedum.dat")->sedum
```

En primer lugar cronstruiremos la tabla de frecuencias:

```
sedum<-table(sedum)
```

Posteriormente calcularemos el valor de χ^2, para lo cual utilizaremos la función:

```
chisq.test(sedum)
```

Cabe reseñar que la función aplica automáticamente la denominada corrección de Yates (necesaria cuando $n < 200$). La función también proporciona las frecuencias esperadas mediante: `chisq.test(sedum)$exp`. A esta tabla de frecuencias esperadas podemos asignarle un nombre; por ejemplo:

```
sedum.e<-chisq.test(sedum)$exp
```

Aplicar ahora el test de la G:

```
g.test(sedum)
```

Esta función no aplica directamente la corrección de Yates, por lo que la realizaremos nosotros:

```
sedum.y<-sedum
sedum.y[1,1]<-sedum[1,1]+0.5
sedum.y[1,2]<-sedum[1,2]-0.5
sedum.y[2,1]<-sedum[2,1]-0.5
sedum.y[2,2]<-sedum[2,2]+0.5
```

Comprobaremos el resultado de la corrección y aplicaremos de nuevo el test:

```
sedum.y
g.test(sedum.y)
```

Comprueba que ahora los resultados de la χ^2 y de la G son más parecidos.

- ¿Qué interpretación ecológica podemos extraer de estos análisis?
- Interpreta el resultado de las siguientes órdenes:

  ```
  (sedum.y-sedum.e)^2/sedum.e
  sum((sedum.y-sedum.e)^2/sedum.e)
  ```

---

3. Ejercicios adicionales

1. En el archivo `raiz.dat` se presentan los datos de un experimento en el que se analizó la influencia de la salinidad en el crecimiento (mm) de las raíces de *Halocnemum strobilaceum*. Los tratamientos considerados fueron: control ($trat = 0$), salinidad baja ($trat = 1$), media ($trat = 2$ y alta ($trat = 3$). ¿Qué conclusiones ecológicas pueden extraerse de los resultados del experimento? Cuidado al utilizar la función `attach()`: los nombres de las columnas no deben coincidir con variables preexistentes.
2. En el archivo `testudo.dat` aparecen los datos de un muestreo de Tortuga mora (*Testudo graeca*) en tres tipos de hábitat: matorral ($hab = 1$), espartal ($hab = 2$) y pinar abierto ($hab = 3$). Con estos datos ¿se encuentran diferencias significativas en la abundancia de tortugas en los distintos tipos de hábitat considerados?

3. Ejemplo de ANOVA con dos factores. En el archivo `bloques.dat` se presentan 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. Analiza el efecto del tratamiento, considerando el diseño en bloques (localidades). Utiliza el signo + para incluir las dos variables en el ANOVA: `summary(aov(crec ~ loc + trat))`.

4. Utilizando la matriz del archivo `territorios.dat`, analizar las diferencias de productividad (variable `pollos`) existentes entre años (variable `factor(year)`) y entre territorios con diferente orientación (variable `orient`).

5. En el archivo `carrasca.dat` se presentan la tabla de contingencia correspondiente a un muestreo de presencia/ausencia de *Quercus rotundifolia* en cuadrículas UTM de 250x250 m en la Sierra de la Torrecilla. Cada cuadrícula fue asignada a una determinada categoría de uso (agrícola, forestal o mixto).

   ¿Es independiente la presencia de carrascas del tipo de uso del territorio?

6. El archivo `calzada.dat` proporciona el número de presas de diferentes categorías (aves, reptiles y mamíferos) aportadas al nido por machos y hembras de Águilas Calzadas (*Hieraætus pennatus*) durante la época de cría de los pollos.

   ¿Influye el sexo en el tipo de presas capturadas por estas águilas?

7. En el archivo `cenizo.dat` se muestra el número de presas capturadas por Águiluchos Cenizos (*Circus pygargus*) durante diferentes períodos de la época de reproducción en el espacio protegido de Ajauque-Rambla Salada (abril: celo; mayo: incubación; junio: cría de pollos; julio: alimentación de los jóvenes).

   ¿Varía la composición de la dieta a lo largo de la época de reproducción?