5. Análisis de datos en ecología (I): introducción a las pruebas estadísticas

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")
```

Cargar también las funciones del muestreador:

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

1. Introducción

En ecología, como en otras ciencias experimentales, las hipótesis se contrastan con los datos obtenidos a través de un muestreo o un experimento. Para analizar estos datos se puede recurrir a un gran número de pruebas o *tests* estadísticos. La naturaleza de las hipótesis y de los datos condicionan el procedimiento del análisis. En todos los casos es muy importante tener presente la necesidad de trasladar las hipótesis de trabajo a preguntas concretas, a las que puedan dar respuesta los datos y las pruebas estadísticas.

Existe un conjunto reducido, pero básico, de pruebas estadísticas que se utilizan frecuentemente en ecología. Entre ellas cabe destacar las siguientes:

- pruebas de normalidad, utilizadas para comprobar la normalidad de los datos.
- pruebas de comparación de varianzas, para comparar varianzas muestrales.
- pruebas de comparación de medias, para determinar si las medias de las muestras son distintas.
- pruebas de bondad de ajuste, destinadas a contrastar distribuciones teóricas y distribuciones muestrales.
- pruebas de independencia, utilizadas para determinar el comportamiento de dos variables cualitativas.
- técnicas de regresión, que permiten determinar las relaciones entre variables de respuesta (dependientes), cuantitativas o cualitativas, y variables ambientales (independientes) cuantitativas.

Con otro criterio, también podemos clasificar las pruebas estadísticas en dos grandes grupos:

Paramétricas Asumen hipótesis sobre los parámetros poblacionales y la distribución de la variable (principalmente normalidad y homogeneidad de varianza).

No paramétricas No precisan asunciones estrictas sobre la distribución de la variable. Su mayor inconveniente, como pago a las pocas exigencias previas, es la falta de sensibilidad a pequeñas diferencias.

2. Errores estadísticos e intervalos de confianza

Consideremos en primer lugar un concepto básico relacionado con la pruebas estadísticas: *el error estadístico*. Así, para cada estadístico se puede definir un error estadístico o error estándar (en inglés: *standard error, SE*).

Para la media muestral $\bar{x}$, en el caso de datos procedentes de una distribución normal, con varianza conocida, el error se define como:

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n}},$$

donde σ es la desviación típica poblacional y n el tamaño muestral. Los errores permiten construir intervalos de confianza para el estadístico. En este caso podemos construir un intervalo de confianza centrado en la media, que contenga a la media poblacional con la probabilidad o confianza $(1 - \alpha)$ deseada; hablamos, también, de un nivel de significación (α) para el intervalo. Este es:

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}};$$

siendo z el valor correspondiente a la distribución normal (en R, puede calcularse con `qnorm()`).

Para verificar la validez de este intervalo podemos generar un número elevado de muestras, mediante simulación, procedentes de una población $\mathcal{N}(0, 1)$, de un tamaño dado, con un nivel de significación prefijado. Calcularemos los límites de confianza y comprobaremos el número de éstos que dejan fuera a la media poblacional ($\mu = 0$).

Ejercicios. Bloque 1:

1. En primer lugar generamos una muestra de tamaño 30, con datos procedentes de una población normal, de $\mu = 0$ y $\sigma = 1$; para ello construimos un vector con treinta valores normales:

   `x<-rnorm(30)`

   Para determinar el intervalo de confianza calculamos en primer lugar la media de los datos muestrales: `mx<-mean(x)`. A continuación los límites inferior `lim` y superior `lsm`:

   ```
   lim<-mx-1.96/sqrt(30)
   lsm<-mx+1.96/sqrt(30)
   ```

   El intervalo así calculado contendrá el valor 0, es decir μ, 95 de cada 100 veces que se repita el procedimiento.

2. Para realizar automáticamente repeticiones del procedimiento anterior utilícese la función `limnorm(n, número de repeticiones, alfa)`. Por ejemplo:

   `limnorm(30,100,0.05)`

   La función devuelve el número de casos en los que el intervalo deja fuera el valor cero, y representa gráficamente los intervalos de confianza calculados.

   Repite varias veces la orden: ¿sobre qué valor oscila el número de límites que contienen al cero? ¿Cabe esperar una gran variación de este valor? ¿Qué efecto tiene aumentar n?

En el caso más común de desconocer la varianza poblacional, el cálculo de los límites de confianza se realiza mediante la expresión:

$$\bar{x} - t_{\nu,1-\alpha/2}\frac{s}{\sqrt{n}}, \qquad \bar{x} + t_{\nu,1-\alpha/2}\frac{s}{\sqrt{n}};$$

siendo t el valor correspondiente de la distribución t–Student, con ν grados de libertad (donde $\nu = n - 1$); este valor puede calcularse con la función `qt()`. Por ejemplo: `qt(0.975,29)`.

Ejercicios. Bloque 2:

1. Volvemos en esta práctica a utilizar el muestreador de poblaciones artificiales para calcular los errores en la estimación de las medias. Haremos un muestreo para cada lámina:

   ```
   md.cuadrado(30, loc=lam1)->md1
   ```

   Ahora estimaremos las medias y sus intervalos de confianza. Para ello utilizaremos la función `apply()` y las ecuaciones de la página anterior.

   ```
   apply(md1, 2, mean); apply(md1, 2, sd)

   apply(md1, 2, mean) + qt(0.975, 29)*apply(md1, 2, sd)/sqrt(30)
   apply(md1, 2, mean) - qt(0.975, 29)*apply(md1, 2, sd)/sqrt(30)
   ```

   Comprobar si el intervalo estimado contiene la media real de la población: `densidad.real(lam1)/100`

2. Estimar ahora el tamaño de muestra necesario para obtener un error relativo del 10 % (D = 0.1) en cada caso. Para ello hay que aplicar la siguiente ecuación:

   $$n = \frac{s^2\, t^2}{\bar{x}^2\, D^2}$$

   donde t es el valor de la distribución t de Student para α =0.05 y $n - 1$ grados de libertad. En el caso de asumir una distribución aleatoria de los individuos (como es el caso de la lámina 1), la expresión anterior se simplifica:

   $$n = \frac{t^2}{\bar{x}\, D^2}$$

   Utilizaremos:

   ```
   qt(0.975, 29)^2 / (apply(md1, 2, mean) * 0.1^2)
   ```

3. Repetir el procedimiento con el muestreo de la lámina 2.

   ```
   md.cuadrado(30, loc=lam2)->md2
   ```

3. Pruebas para la comparación de medias

Ejercicios. Bloque 3:

Analizaremos los datos de un experimento sobre la influencia del nitrógeno en el crecimiento (en cm) de talos del alga *Caulerpa prolifera*, disponibles en el archivo `caulerpa.dat`. Se trata de averiguar si el crecimiento de los talos es mayor en acuarios con aporte suplementario de nitrógeno, que en los talos de acuarios control (sin tratamiento). Para ello debemos comprobar si en ambos acuarios el crecimiento medio de los talos es el mismo (**hipótesis nula**, H_0) o no (hipótesis alternativa), por lo que utilizaremos diferentes tipos de pruebas estadísticas que permiten comparar las medias de dos poblaciones.

Leer el archivo y asignar a la matriz el nombre `caulerpa`:

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

Ver los datos leídos:

```
caulerpa
```

Para que R reconozca los nombres de variables es necesario utilizar:

```
attach(caulerpa)
```

Dado que `trat` es una variable de tipo factor, es conveniente que R la maneje como tal:

```
trat<-factor(trat)
```

Ya estamos en disposición de analizar los datos del experimento. Comenzaremos utilizando el **test de la *t***, pero antes de su aplicación debemos comprobar que los datos cumplen el requisito de normalidad:

```
normalidad(crec)
```

Esta función calcula el estadístico de **Shapiro-Wilk** y el gráfico **Q-Q normal**. Observaremos que la variable no sigue la distribución normal, por lo que optaremos por realizar una transformación logarítmica:

```
normalidad(log(crec))
```

Si lo preferimos podemos crear una nueva variable para trabajar con más comodidad:

```
lcrec<-log(crec)
```

Podemos ahora calcular algunos estadísticos descriptivos de la nueva variable, en función del tipo de tratamiento (*control* y *nitrógeno*), y representar los histogramas de frecuencias. Por ejemplo:

```
mean(lcrec[trat==0]);var(lcrec[trat==0]);hist(lcrec[trat==0])
```

```
mean(lcrec[trat==1]);var(lcrec[trat==1]);hist(lcrec[trat==1])
```

Efectivamente, la media de crecimiento en los acuarios con nitrógeno es mayor que en los acuarios control, pero como los datos provienen de una muestra, es necesario comprobar la significación estadística de esta diferencia. Pero antes también debe comprobarse el cumplimiento de otro requisito fundamental para la aplicación de los test paramétricos: la homogeneidad de las varianzas. Podemos utilizar el cociente entre ambas varianzas (**test de la** F):

```
var.test(lcrec~trat)
```

Con el resultado obtenido aceptaremos la hipótesis nula de igualdad de varianzas, por lo que podremos aplicar finalmente el **test de la** t, que en su expresión más sencilla ($n_1 = n_2$) se calcula según:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2+s_2^2}{n}}}$$

Comenzaremos calculando el numerador (la diferencia de medias):

```
dif<-mean(lcrec[trat==0])-mean(lcrec[trat==1])
```

Luego continuamos con el cálculo del **error estándar de la diferencia de medias** (el denominador):

```
error<-sqrt((var(lcrec[trat==0])+var(lcrec[trat==1]))/28)
```

Finalmente el valor de t se obtiene como:

```
dif/error
```

La función `t.test` realiza los mismos cálculos, pero además proporciona el valor de probabilidad correspondiente y el intervalo de confianza de la diferencia de medias:

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

La expresión `var.equal=T` indica que el test asume igualdad de varianzas (existe otra versión de esta prueba para la comparación de medias con varianzas no homogéneas: `var.equal=F`).

- ¿Aceptamos o rechazamos la hipótesis nula? ¿Cuál es la interpretación ecológica de los resultados del test?
- Discutir el efecto de las varianzas y el tamaño muestral (n), en el cálculo del estadístico t, y su influencia sobre la posible aceptación o rechazo de la hipótesis nula.
- Discutir el significado del intervalo de confianza de la diferencia de medias. Si sólo conociéramos este intervalo, en ausencia del valor de P, ¿cómo decidiríamos la significación estadística de la diferencia de medias?

Ejercicios. Bloque 4:

Los datos del archivo `thymus.dat`, corresponden a un muestreo para comparar la densidad de tomillos (*Thymus hyemalis*) en laderas de solana ($solana = 1$) y umbría ($solana = 0$) en el Parque Regional de Calblanque. Los datos representan el número de tomillos por unidad de muestreo (2x2 m).

Leer el archivo y seguir el mismo procedimiento que en el ejercicio anterior. Como hay ceros en los datos será necesario utilizar la transformación $\log(x + 1)$.

Aunque en este caso la transformación logarítmica no es suficiente para conseguir que los datos se ajusten a una distribución normal, realizar las pruebas anteriores de comparación de varianzas y medias. No obstante. como alternativa más apropiada, utilizaremos también un test no paramétrico con los datos originales: el **test de suma de rangos de Wilcoxon** (equivalente al **test de Mann-Whitney**)

```
wilcox.test(tomillos~solana)
```

[La opción `paired=T` ejecutaría otra variante de este test, para datos "pareados". También es posible utilizar esta opción en el test de la t.]

- ¿Cual es la interpretación ecológica de los resultados?
- Discutir la posible razón de la ineficacia, en este caso, de la transformación logarítmica para la normalización de los datos.
- Comparando los valores de P obtenidos con el test de la t y el de Wilcoxon, ¿qué podríamos comentar sobre las características de los test no paramétricos, por lo que se refiere a la aceptación o rechazo de la hipótesis nula?

Con el siguiente ejemplo utilizaremos otro tipo de prueba estadística: un test de aleatorización. Requiere un cálculo algo más laborioso, pero representa una buena alternativa, cada vez más utilizada por los investigadores.

El archivo `litorina.dat` contiene datos sobre el número de litorinas (*Melaraphe neritoides*) presentes en unidades de muestreo (20x20 cm) localizadas en las zonas mediolitoral (batida por el oleaje, $batida = 1$) y supralitoral (no batida $batida = 0$) de un área rocosa del litoral murciano. ¿Existen diferencias significativas en la abundancia de la especie en una y otra zona?

Procedimiento. Primero leemos el archivo de datos.

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

Ahora calculamos la diferencia observada en las medias

```
dif.obs<-mean(datos[batida==1])-mean(datos[batida==0])
```

La aleatorización consiste en asignar el tratamiento aleatoriamente a cada dato (en este caso el tipo de zona). Este procedimiento de permutación puede hacerse con la función `sample()`:

```
sample(batida)->batida.a
```

Ahora calculamos de nuevo la diferencia; en este caso:

```
mean(datos[batida.a==1])-mean(datos[batida.a==0])
```

El proceso ha de repetirse muchas veces, por ejemplo 1000. Para ello hay que crear una variable (en la que iremos almacenando los resultados) y utilizar un bucle:

```
dif<-rep(0,1000)
for (i in 1:1000) {sample(batida)->batida.a;
                   mean(datos[batida.a==1])-mean(datos[batida.a==0])->dif[i]}
```

Para calcular la significación estadística tendremos que calcular cuántos valores obtenidos mediante la aleatorización son mayores que el valor de diferencia observado. El valor de P se obtiene dividiendo dicho número por el número de permutaciones

```
sum(abs(dif)>abs(dif.obs))
```

```
hist(abs(dif)
abline(abs(dif.obs))
```

4. Ejercicios adicionales:

1. Siguiendo el procedimiento del ejercicio 1, simular muestras bajo el supuesto de desconocimiento de la varianza poblacional; utilizar en este caso la función `limnormvd(n, número de repeticiones, alfa)`.

   Comparando los resultados obtenidos en el caso anterior ¿qué coste tiene desconocer la varianza poblacional (σ)?

2. Demostrar la equivalencia entre las expresiones: $\bar{x} \pm t\frac{s}{\sqrt{n}}$ y $n = \frac{s^2\ t^2}{\bar{x}^2\ D^2}$

3. El archivo `romero.dat` contiene los datos de un muestreo de romeros (*Rosmarinus officinalis*) en el Parque Regional de Calblanque. Se pretendía analizar la influencia del tipo de suelos sobre la abundancia de la especie. Para ello se situaron 20 unidades de muestreo de 2x2 m sobre sustrato calizo ($sust = 0$) y otras 20 sobre sustrato de pizarras ($sust = 1$). ¿Existe relación entre el tipo de sustrato y la abundancia de la especie?

   Utilizar los tres métodos (paramético, no paramétrico y test de aleatorización), comparando los resultados y discutiendo su adecuación a las características de los datos.