3. Muestreo, estadísticos y simulación

Modelización de Sistemas Ambientales (07M1)
Dpto. de Ecología e Hidrología y Dpto. Geografía
Facultad de Biología
Universidad de Murcia

Curso 2006–07

Índice

1. Introducción

En esta sesión se estudiarán algunos casos de simulación sencillos. Además, se realizará un análisis estadístico de los resultados para comprobar el comportamiento de los resultados obtenidos.

2. Algo de genética (1.5 horas)

Los trabajos clásicos de Mendel se basan en un modelo conceptual sencillo: los padres poseen alelos por pares, cada uno, aporta sólo uno al descendiente, aleatoriamente y de forma equiprobable.

Para simular la descendencia de parentales heterocigotos, asumiendo un gen con dos alelos A, dominante, y a, recesivo. puede utilizarse el siguiente procedimiento:

```
genotipo<-c("A","a")
g<-1000
sample(genotipo,g,replace=T)->m
sample(genotipo,g,replace=T)->p
(table (m,p)->td)
```

En la primera expresión definimos la variable `genotipo` con los dos alelos. Posteriormente generaremos aleatoriamente 1000 gametos para el padre, `p`, y otros tantos para la madre, `m`. Finalmente utilizaremos la función `table()` para realizar el "cruzamiento" de los gametos, y determinar cuantos individuos presentan cada uno de los genotipos posibles.

EJERCICIOS

1. ¿Qué resultados cabe esperar desde un punto de vista teórico? ¿Qué significan las celdillas de la tabla obtenida? ¿Qué diferencias hay con los valores observados? ¿Cómo influye el tamaño muestral (número de descendientes simulados)?

2. La teoría predice que de ser cierta la hipótesis nula de un test de χ^2 de bondad de ajuste las diferencias entre los valores esperados y los observados deben ser menor que un determinado valor de la función χ^2, concretamente el correspondiente a 1 grado de libertad y que deja a la izquierda el 0.95 de los valores teóricos, es decir, $\chi^2_{1,0.05}$. El test puede aplicarse en R mediante la función `chisq.test()`:

   ```
   chisq.test(td,0.25)
   ```

siendo 0.25 el valor esperado en cada celdilla.

¿Puede afirmarse que los valores simulados están estadísticamente de acuerdo con los esperados?

¿Qué ocurre si repetimos varias veces la simulación? ¿Obtenemos resultados similares? ¿Sería razonable encontrar un resultado no significativo? ¿Por qué?

3. Representar la función `dchisq()` para un grado de libertad mediante:

```
i<-seq(0,5,0.01)
plot(i,dchisq(i,1),type="l",asp=1)
abline(h=0,v=0,col="blue")
```

primero definimos el intervalo y la resolución mediante la variable `i`, después representamos la curva y posteriormente representamos los ejes en el origen.

¿Cuanto vale el área encerrada entre la curva y el eje de abscisas? ¿Por qué?

¿Donde deben representarse los valores experimentales de χ^2 obtenidos en los experimentos anteriores?

4. Para automatizar el experimento podemos definir una función, a la que llamaremos `experimento`:

```
experimento<-function(g=1000){
  genotipo<-c("A","a")
  sample(genotipo,g,replace=T)->m
  sample(genotipo,g,replace=T)->p
  table (m,p)->td
  chisq.test(td,0.25)$p.value
  }
```

Puede repetirse el experimento un gran número de veces mediante:

```
re<-NULL;for (i in 1:1000) re[i]<-experimento()
```

es decir, crear una variable `re` para almacenar los resultados experimentales, y repetir el experimento 1000 veces guardando el valor de probabilidad que devuelve el test en cada simulación.

¿Se producen valores menores de 0.05? ¿Con qué frecuencia? Sugerencia utilizar las siguientes expresiones para evaluar los resultados experimentales:

```
hist(re,seq(0,1,0.01))
sum(re<0.05)
```

¿Cuantos experimentos presentan un valor de p (`p-value`) menor de 0.05? ¿Qué ocurre con más o menos experimento? ¿Se mantiene constante la proporción?

5. ¿Qué efecto tendría un aumento o disminución en el experimento del número de descendiente? ¿Cuál sería el procedimiento para comproblarlo?

3. Tamaño y forma de organismos (1.5 horas)

Simularemos el comportamiento de dos variables relacionadas en los organismos, las que describen la talla y el peso, considerando algunos aspectos de la relación y como estudiar los resultados obtenidos. Para ello utilizaremos un modelo básico: la regresión lineal. Se trata de considerar que las variables x e y se relacionan según:

$$y = a \times x + b + \epsilon$$

siendo y la variable dependiente, x la variable independiente, a la pendiente de la recta, b la ordenada en el origen y ϵ el error asociado a la medida o a la propia naturaleza de la relación. En nuestro caso puede escribirse el modelo como:

$$peso = a \times talla + b + \epsilon$$

EJERCICIOS

Para simular valores de una población con talla media de 100 y desviación típica de 15 y asumiendo un comportamiento normal de la variable, utilizamos:

```
rnorm(30,100,15) -> talla
```

1. Puede verificarse la normalidad de la simulación con las siguientes expresiones, en primer lugar, para los datos originales, depués para los valores estandarizados.

   ```
   qqnorm(talla)
   qqnorm(scale(talla))
   ```

   ¿Qué resultado cabe esperar para datos normales en una representación `qqplot()`? ¿Son los resulados "normales"? ¿Cuales son las diferencias entre las dos gráficas? ¿Cual es el papel de la funcion `scale()`?

2. Para generar los valores de peso consideraremos el modelo anterior. Los valores se obtienen de la siguiente expresión:

   ```
   peso<-0.45*talla+15
   plot(talla,peso)
   x11()
   peso<-0.45*talla+15+rnorm(30)
   plot(talla,peso)
   ```

   ¿Qué diferencia existe entre los dos gráficos obtenidos? ¿A qué se término se puede asociar el ruido? ¿Cómo podemos calificar el añadir valores aleatorios de una distribución normal? ¿Qué efecto pueden tener distintos valores de μ y σ?

   ¿Qué valor de peso esperamos para una talla dada?

3. Evaluaremos las relaciones entre las dos variables con un análisis de regresión. Utilizaremos la función `lm()`.

   ```
   lm(peso~talla)
   ```

   Puede representarse gráficamente la recta de regresión utilizando:

   ```
   abline(lm(peso~talla))
   ```

   ¿Cuáles son los valores para la pendiente y el término independiente estimados por el modelo? ¿Coinciden con los teóricos empleados en la simulación? ¿Por qué?

4. Puede obtenerse una descripción mas detallada del análisis utilizando la función `summary()`:

   ```
   summary(lm(peso~talla))
   ```

   La proporción de información contenida en los datos y que queda reflejada en el modelo esta indicada en el término: `Adjusted R-squared`, con valores próximos a uno los que indican un modelo muy ajustado a los datos.

   ¿Qué efecto tedría un mayor papel ruido en el muestreo? Para discutir esta cuestión pueden considerarse el sigiente procedimiento, con distitos valores de `sruido`:

   ```
   peso<-0.45*talla+15
   sruido<-5
   peso<-0.45*talla+15+rnorm(30,0,sruido)
   plot(talla,peso)
   lm(peso~talla)->lm.tp
   summary(lm.tp)
   abline(lm.tp)
   ```

   `sruido` indica la desviación típica del ruido. A mayor valor de esta mayor efecto del ruido y por lo tanto debería obtenerse una menor calidad del modelo. ¿Puede verificarse esta afirmación? ¿Cómo?

4. Para entregar

Cuando utilizamos la expresión:

```
lm(lm.tp)$coef[2]
```

obtenemos el coeficiente de la variable dependiente en el análisis de regresión realizado y que se almacena en la variable `lm.tp`.

Diseñar un experimento para obtener numerosas simulaciones del proceso descrito en la sección 3, y así, estudiar el comportamiento del coeficiente de la variable independiente (Pistas: ¿Qué valor esperamos? ¿Sigue una distribución normal?).