3. Análisis elemental de datos: codificación, transformación y representación

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

Curso 2008–09

Índice

Antes de empezar

Antes de empezar deben cargarse las funciones necesarias para desarrollar la práctica utilizando:

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

1. Introducción

El objetivo de la siguiente práctica es revisar, aplicar y discutir en distintos tipos de casos (los más habituales en general), las principales opciones para simplificar, expresar y representar los datos y sus relaciones.

También se introducirá el uso de los métodos de simulación. Se trata de un elemento importante en ecología ya que por una parte facilita la evaluación y la compresión del papel que juegan los distintos procedimientos de análisis de datos, y por otra parte, permite explorar modelos de funcionamiento de los sistemas y procesos estudiados.

2. Simulación

Resulta de gran interés en ecología poder recurrir a métodos de simulación para el estudio de los procesos estocásticos, es decir, en aquello en donde existe un componente aleatorio. Para ello, la posibilidad de simular

variables procedentes de un modelo teórico dado, de forma automática, facilita la tarea. Así, en ocasiones, es interesante disponer de un conjunto de valores procedentes de una distribución aleatoria binomial, uniforme, normal, ...

Disponer de valores procedentes de un modelo teórico, ya sea procedente de una distribución estadística teórica o de un modelo más biológico nos muestran resultados posibles que proporcionaría la distribución teórica: podríamos decir que el modelo es una suerte de "genotipo" y los datos (experimentales o simulados) el "fenotipo".

Por ejemplo, puede simularse la proporción sexual de una camada tirando una moneda por cada individuo que la compone y asignado a cada sexo uno de los posibles resultados: (macho=cara, hembra=cruz). Repitiendo infinidad de veces el experimento tendríamos la frecuencia teórica de encontar 1 macho, 2 machos, ... En la figura 1 se refleja la distribución teórica para camadas de 4 ejemplares y la simulación para 100, 1000, 10000 y 100000 repeticiones.

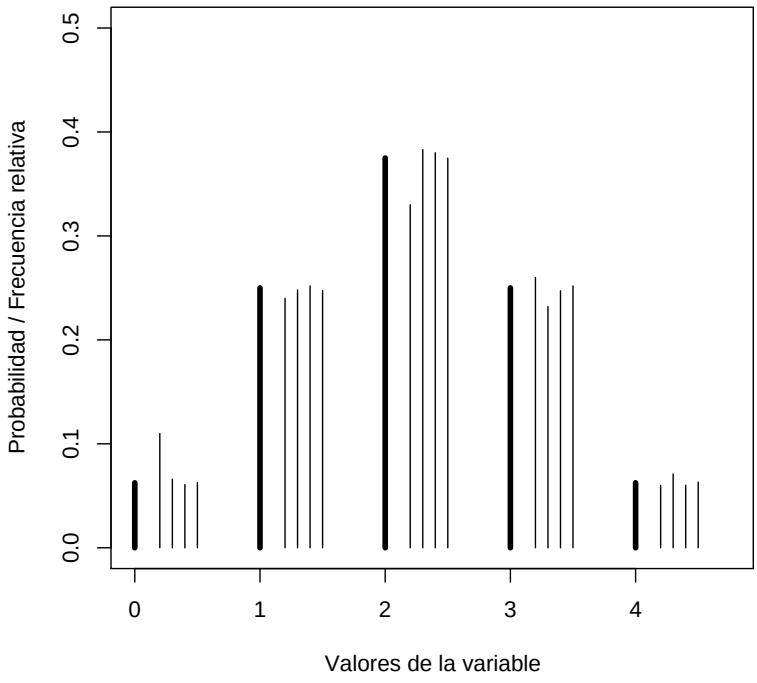

Figura 1: Datos de probabilidad teórica para una binomial, p=0.5 y n=4 (linea gruesa), simulaciones con 100, 1000, 10000 y 100000 repeticiones de izquierda a derecha (line delgada).

También podemos simular procesos "más biológicos"; por ejemplo: la curva de supervivencia de individuos de una cohorte. Dado un conjunto inicial de individuos de edad 0, cohorte, se simula el proceso de muerte de cada uno de la siguiente forma: la mortalidad es constante en cada unidad de tiempo y su valor es de 0.4; en cada unidad de tiempo todos y cada uno de los individuos "juega" con un dado de 10 caras para determinar su supervivencia. Para sobrevivir necesita obtener valores de 1 a 6, en el caso contrario el individuo muere esa edad.

Partiendo de 100 individuos, en cada unidad de tiempo el número de supervivientes se determina tras el resultado del juego, en promedio, tendremos el 60 % de los individuos vivos, pero en cada caso la cantidad puede variar mucho, tal como sucede en la realidad, pues a la proporción teórica sólo se aproximaría cuando el tamaño de la cohorte fuera muy grande (figura 2).

En resumen, la simulación nos permite analizar la variabilidad que produce asumir un determinado modelo. En la realidad nos enfrentaremos a la situación contraria: desde la realidad queremos obtener el modelo.

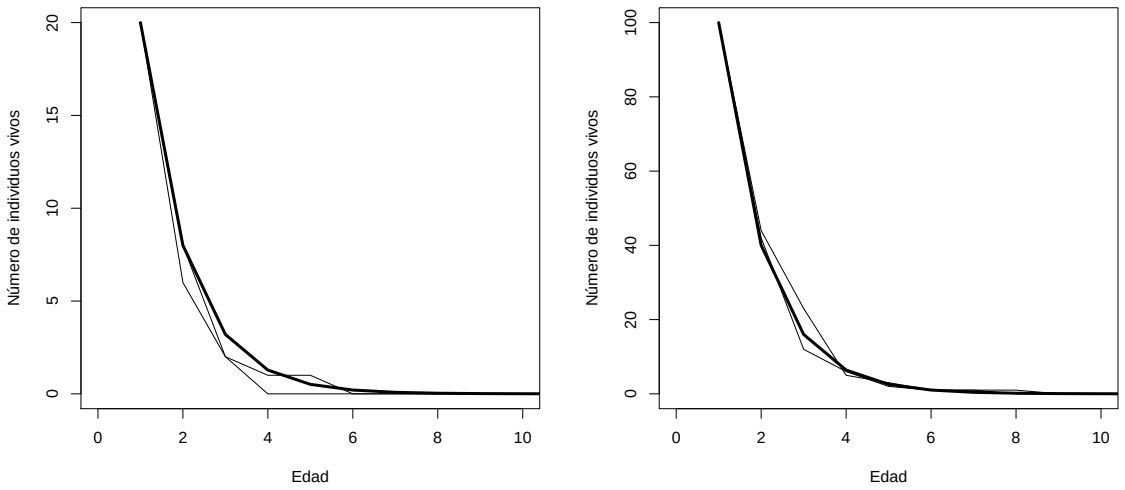

Figura 2: Simulación de la supervivencia de una cohorte con la edad, con 20 individudos (izquierda) y 100 (derecha), la mortalidad en todas las edades es de 0.6.

Ejercicios. Bloque 1:

1. Puede utilizarse la función `sample()` para tomar muestras aleatorias de una variable o también para generar valores aleatorios de un espacio muestral dado, por ejemplo:

   ```
   sample(c(0,1), 100, replace=T)
   sample(c(0,1), 100, replace=T, prob=c(0.75,0.25))
   ```

   Verificar que los resultados se corresponden razonablemente con lo esperado, tabulando los resultados mediante al función `table()`.

   ```
   table(sample(c(0,1), 100, replace=T))
   ```

---

3. Análisis de variables aisladas

3.1. Tabulación de datos

Considerando una variable podemos estudiar su comportamiento en relación con la distribución de la frecuencia de aparición de las observaciones del espacio muestral: **distribución de frecuencias**. Utilizamos una tabla de distribución de frecuencias para reflejarlo:

Valores Variable	frecuencia absoluta	frec. relativa	frec. abs. acumulada	frec. rel. acum.
x_1	f_1	h_1	F_1	H_1
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_i	f_i	h_i	F_i	H_i
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_k	f_k	h_k	n	1

$$n = \sum_{i=1}^{k} f_i$$

siendo n el número total de observaciones; k el número de clases consideradas; x_i los distintos valores que toma la variable; f_i el número de veces que aparece ese valor (frecuencia absoluta); h_i es la frecuencia relativa:

f_i/n; F_i es la frecuencia absoluta acumulada, suma de las frecuencias absolutas de esta clase más las de las clases anteriores, F_k coincide con n. H_i es la frecuencia relativa acumulada, suma de las frecuencias relativas de esta clase más las anteriores; H_k coincide con 1. Resulta obvio que las frecuencias acumuladas carecen de sentido en el caso de variables nominales.

Este tipo de tabulación se puede realizar para todas las variables, cualitativas y las cuantitativas, si bien en el caso de continuas es preciso determinar intervalos en los que considerar la frecuencia. Si la variable es discreta pero presenta un gran número de valores puede procederse también de esta manera.

Con R a partir de una variable `x` pueden calcularse las frecuencias absolutas mediante la función `table()`: `table(x)`. Las frecuencias acumuladas se obtiene mediante la función `cumsum()` que se aplica a los valores tabulados: `cumsum(table(x))`. Las frecuencias relativas se obtienen de las anteriores dividiendo por el número de elementos del vector: `table(x)/length(x)`

Ejercicios. Bloque 2:

1. Utilizando la simulación de 100 observaciones para una variable que toma valores de 1 a 10 equiprobables:

   ```
   sample(1:10,100, replace=T)->a
   ```

   ```
   table(a)
   ```

   Calcular la tabla de frecuencias absolutas, relativas absolutas acumuladas y relativas acumuladas. Se recomienda el uso de las funciones `sum` y `cumsum`.

   ```
   sum(table(a))
   ```

   ```
   cumsum(table(a))
   ```

   ¿Coinciden los valores observados con los que cabría esperar de las reglas impuestas en la simulación?

   ¿Qué ocurre si aumentamos el tamaño de la muestra?

2. Utilizando la función `runif()` que proporciona valores de una distribución aleatoria uniforme en el rango de 0 a 1, calcular 100 valores mediante:

   ```
   runif(100)
   ```

   Calcular como en el caso anterior la tabla de frecuencias absolutas, relativas absolutas acumuladas y relativas acumuladas. Se recomienda el uso de las funciones `sum` y `cumsum`.

3. Comprobar el resultado al trabajar con simulación de números normales.

   ```
   z<-rnorm(200)
   ```

   ```
   tz<-table(z)
   ```

   ```
   plot(names(tz),cumsum(tz))
   ```

3.2. Representación de frecuencias

Las tablas de frecuencias pueden representarse por:

diagramas de barras: se representan los valores de la variable (cualitativas o discretas) en abscisas y la frecuencia en ordenadas, utilizando una barra para cada clase con la altura correspondiente a la frecuencia (absoluta o relativa, acumulada o no). Han de cuidarse especialmente algunos aspectos para que el diagrama no resulte engañoso: a) el eje de abscisas debe empezar en cero, o indicar claramente lo contrario; b) las barras deber ser delgadas (mejor una línea gruesa); c) evitar colorear o resaltar algunas barras.

diagramas de sectores: las clases de las variables cualitativas se representan por sectores circulares cuyo ángulo es proporcional al la frecuencia. La representación en perspectiva altera la sensación de proporcionalidad "correcta".

pictogramas: una figura (por ejemplo un árbol) representa la frecuencia; para evitar el efecto longitud/área utiliza una figura por cada cantidad (o proporción) de observaciones.

histogramas: en el caso de variables continuas se representa la variable sobre abscisas; sobre cada intervalo de clase se levanta un rectángulo de área proporcional a la frecuencia.

Ejercicios. Bloque 3:

1. Representar gráficamente las frecuencias (absolutas y absolutas acumuladas) calculadas anteriormente utilizando la funciones `plot(,type="h")`, `barplot()` y `pie()`. Por ejemplo:

   ```
   barplot(cumsum(table(cut(runif(100),10))))
   ```

   ¿Alguna representación mejora las posibilidades de interpretación del comportamiento de la variable?

2. Representar `runif(100)` mediante la función `hist()`:

   ```
   hist(runif(100))
   ```

   Repetir el histograma pero con la opción `plot=F`:

   ```
   hist(runif(100),plot=F)
   ```

   ¿Cuantas clases se están considerando? ¿qué significan: `breaks`, `$mids` y `$counts`?

   construir nuevos histogramas considerando un distinto número de clases, p. ej.: 2, 8, 13, . . . :

   ```
   hist(runif(100),2)
   ```

   ¿Cuantas clases conviene utilizar para construir un histograma? ¿Qué papel juega el tamaño muestral?

3. Utilizando el fichero de datos `iris.dat` tabular y representar gráficamente la distribución de frecuencias de las 5 variables.

   ```
   read.table("http://www.um.es/docencia/emc/datos/iris.dat",header=T)->iris
   attach(iris)
   hist(lonsep)
   ```

3.3. Estadísticos

Además de las tablas de frecuencia podemos describir las variables mediante estadísticos —funciones de los datos muestrales— y parámetros —funciones de los datos poblacionales—; los primeros se anotan mediante letras latinas: la media muestral para x es $\bar{x}$; la varianza para la misma variable es: s_x^2. Para los parámetros poblacionales recurrimos a letras griegas: μ_x y σ_x^2. Considerando el objetivo para el que utilizamos el estadístico, hablamos de dos grandes tipos: *estadísticos de centralización* y *estadísticos de dispersión*, que describen, respectivamente, la tendencia central de los datos (como la media, o valor central o centro de gravedad de los datos) y la tendencia a alejarse de ese valor central (como la varianza o inercia de los datos).

El uso de estadísticos muestrales tiene como objetivo aproximar los valores poblacionales y por lo tanto predecir el comportamiento de la población.

Habitualmente se suelen describir los datos muestrales mediante: *máximo y mínimo*, que se corresponden con los valores extremos observados; *primer cuartil, mediana,* y *tercer cuartil*, ordenados los valores de menor a mayor, el primer cuartil es el que supera al 25 % de los datos, la mediana al 50 % y el segundo al 75 %; *media*, valor central o centro de gravedad de los datos; *varianza*, medida de alejamiento de la media o inercia de los datos.

Ejercicios. Bloque 4:

1. Utilizar la función `summary()` para describir las distintas variables utilizadas anteriormente, `iris`.

   ¿Para que sirven las funciones: `range()`, `sd()`, `var()` y `summary()`?

2. Utilizar la función `quantile()` para calcular los cuartiles de la variable `ancsep`.

   Representar los valores de la media y de los cuartiles sobre el histograma de la variable:

   ```
   hist(ancsep)
   abline(v=mean(ancsep),col=2)
   abline(v=quantile(ancsep),col=3)
   ```

3. Para calcular los estadísticos, y en general aplicar una función, de todas las columnas (o filas) de una matriz se utiliza la función `apply()`, `apply(datos,indice,estadístico)`

   ```
   var(iris)
   apply(iris,2,var)
   ```

4. Crear la matriz la matriz `territorios` a partir del fichero de datos:

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

   ¿Qué ocurre cuando calculamos la media de pollos producidos en cada nido? ¿que efecto tiene incluir en la función `mean()` la opción na.rm=T?

3.4. Distribuciones teóricas

La variables experimentales pueden tener, en su distribución de frecuencias, un comportamiento similar al de ciertas distribuciones teóricas: distribución de Poisson, binomial, binomial negativa, uniforme, normal, ... El significado que se puede atribuir en el caso de que una variable experimental siga un determinado modelo teórico depende de la naturaleza de la propiedad observada.

Así, por ejemplo, la composición de una camada (número de machos y hembras nacidos) sigue una distribución binomial. Conocido el número de individuos nacidos n y la probabilidad p de uno de los sexos (p. ej.: hembra), es posible determinar la frecuencia con la que este sexo aparece en las camadas (0 hembras, 1 hembra, ..., n hembra). Así, tenemos una distribución binomial positiva: $\mathcal{B}(n,p)$ donde n es el máximo número posible de éxitos que puede obtenerse (en este ejemplo el número total de individuos en la camada); y p, es la probabilidad del éxito (que el individuo sea hembra). Es posible obtener la probabilidad para cada uno de los valores de la variable que van de 0 a n utilizando la expresión:

$$P(X = x) \; = \; \binom{n}{x} p^x \, (1-p)^{n-x} \qquad x = 1, 2, 3, \ldots, n$$

En R resulta fácil realizar este cálculo mediante la función:

```
dbinom(x, n, p)
```

siendo: `x` el número de éxitos deseados; `n` el número máximo de éxitos; y `p` la probabilidad del éxito. Otro situación en la que se presenta una distribución binomial se da cuando la variable es *el número de individuos de una especie por unidad de muestreo*, siempre que la distribución espacial de los organismo sea regular.

En el caso de variables continuas la distribución más habitual es la distribución normal, cuya función de densidad es:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

siendo μ y σ la media y varianza poblacionales. El valor $f(f)$ se obtiene fácilmente en R mediante la expresión:

```
dnorm(x, μ, σ)
```

En el caso de variables continuas debe calcularse la probabilidad para un intervalo, y calculado la integral definida entre los límites del intervalo para la función de distribución de manera que:

$$P(a \leq x \leq b) \; = \int_a^b \frac{1}{\sqrt{2\pi\sigma}} \, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx$$

Este cálculo se realiza de forma sencilla en R mediante la función

`pnorm(`x`, `μ`, `σ`)`

que proporciona la probabilidad, p, en el intervalo $]-\infty$,x]. La función inversa: calcula el valor de x para una probabilidad dada, en el intervalo $]-\infty$,x], es:

`qnorm(`p`, `μ`, `σ`)`

Además puede utilizarse la función `rnorm(`k`, `μ`, `σ`)` para simular k valores de una normal de media μ y desviación típica σ.

Ejercicios. Bloque 5:

1. Utilizando la función `dbinom()`, calcular la distribución de frecuencias para la variable *número de hembras en la camada* de 7 individuos, sabiendo que la probabilidad de nacimiento de una hembras es 0.56.

   ```
   dbinom(0:7,7,0.56)
   ```

   ¿cual es la probabilidad de que todos los nacidos sean hembras? ¿si aumenta p a 0.60 aumenta o disminuye la probabilidad de que todos sean hembras? ¿cuanto varía esta?

2. utilizando la función `dnorm()` representar la función normal, $\mathcal{N}(0,1)$, entre -3 y +3 mediante:

   ```
   plot (dnorm, -3,3)
   ```

   Con la ayuda de la función `pnorm()` calcular la probabilidad de que una variable $\mathcal{N}(0,1)$ tome valores: menores que 0, menores 1, y entre -1 y 1.

   ```
   pnorm (0)
   ```

3. Utilizando la función `qnorm()` determinar que valor de una variable $\mathcal{N}(0,1)$ deja entre $-\infty$ y este una probabilidad de 0.975.

   ```
   qnorm(0.975)
   ```

4. Calcular los estadísticos descriptivos habituales de una variable normal obtenida por simulación.

   ```
   rnorm(100)->z
   summary(z)
   ```

   ¿se obtienen los valores que cabría esperar?

5. Superponer al histograma de la *longitud del sépalo* la distribución normal que podría asociarse utilizando la función `histnorm()`. Para utilizar esta función son necesarias las funciones de prácticas incluidas el fichero `funciones.R`, después de descargarlo de la página web de la asignatura:

   ```
   histnorm(lonsep,30)
   ```

   ¿sugiere el gráfico un comportamiento normal de la variable?

---

4. Análisis de dos variables

En ecología el comportamiento de un sistema puede expresarse por una sola variable (cualquier tipo de respuesta biológica), pero generalmente el interés de las investigaciones se orienta a determinar si existe dependencia de un factor ambiental, u otro factor biótico. Así pues, en la mayoría de ocasiones, trataremos de establecer relaciones entre, al menos, dos variables.

Un opción para describir y visualizar, de forma sencilla una relación entre dos variables (cualitativas o cuantitativas) es una simple tabla de contingencia, donde las filas y las columnas se corresponden con las modalidades de cada una de las variables; las celdillas se corresponden al número de observaciones que presentan las dos modalidades simultáneamente.

Otra aproximación es la representación gráfica de las muestras en el plano xy y a partir del estudio de la forma y distribución de la nube de puntos podemos plantear posibles relaciones y los análisis pertinentes para evaluarlas.

También es frecuente el cálculo de estadísticos condicionando una variable a la otra (habitualmente cualitativa).

En resumen se trata de encontrar indicios de la existencia de la respuesta de una variable a la otra, es decir, un modelo de respuesta.

Ejercicios. Bloque 6:

1. Construir tablas de contingencia para la matriz `iris` usando la función `table()`. Representar las tablas utilizando la función `barplot()`. Si utilizamos el parámetro `beside=T` ¿cuál es la diferencia? ¿cuál es más útil?

   ```
   barplot(table(cut(lonsep,4),especie))
   barplot(table(cut(lonsep,4),especie),beside=T)
   ```

2. Puede representarse el comportamiento de los datos condicionados a una variable cualitativa utilizando:

   ```
   f<-factor(especie,labels=c("setosa","versicolor","virginica"))
   dotchart(ancpet, groups=f)
   ```

3. Representar los estadísticos de algunas variables mediante la función `boxplot()`. ¿Como se interpretan este tipo de gráficos?

   ```
   boxplot(lonpet~especie)
   quantile(lonpet[especie==1])
   abline(h=5.0,col=3)
   ```

4. Representar dos variables mediante la función `plot()`, colorear los puntos por una tercera variable usando el parámetro `col=variable`. Realiza el mismo procedimiento con las distintas columnas de una matriz utilizando:

   ```
   plot(iris[,1:2], col=especie)
   ```

   ¿qué ventajas tendría una representación de este tipo? ¿puedes indicar algunas aplicaciones?

   Añadir a los gráficos anteriores:

   ```
   points(mean(lonsep),mean(ancsep),col=2,pch=16,cex=1.5)
   abline(v=mean(lonsep),h=mean(ancsep),col=2)
   ```

   ¿Qué representan los puntos y las líneas de color rojo?

5. En algunos casos es necesario utilizar un subconjunto de la matriz de datos. Puede utilizarse operadores lógicos para proporcionar un vector de selección, por ejemplo, si queremos la media de la *anchura del sépalo* de los ejemplares de *Iris setosa* (código 1):

   ```
   summary(iris[especie==1,2])
   ```

   Puede resultar más claro y cómodo, si se utiliza reiteradamente la misma selección utilizar una variable de ayuda:

   ```
   sel<-especie==1
   summary(ancsep[sel])
   ```

6. Calcular estadísticos, o realizar gráficos, para subconjuntos de los valores de una variable condicionados por los de otra se puede realizar con comodidad mediante:

   ```
   by(lonsep,especie,median)
   aggregate (ancsep,list(especie),mean)
   ```

   La ventaja de la función agregate es que puede trabajar con toda la matriz:

   ```
   aggregate (iris,list(sp=especie),mean)
   ```

   En algunos casos puede ser necesario definir la función a aplicar a todos los elementos, como puede ser el caso del cálculo con variables donde se presentan valores *missing*:

   ```
   by(pollos,year,function(x) mean(x,na.rm=T))
   ```

7. La representación gráfica de estos valores condicionados puede resultar muy interesante:

   ```
   plot(as.matrix(by(pollos,year,function(x) mean(x,na.rm=T))),type="o")
   ```

   ¿Cómo se interpreta la gráfica obtenida? ¿Cuáles son los valores representados en el eje de abscisas? ¿Y en el de ordenadas? ¿Qué ventajas tiene la siguiente representación?

   ```
   y<-as.matrix(by(pollos,year,function(x) mean(x,na.rm=T)))
   plot(2000:2006,y,type="o",main="Número de pollos",
   xlab="Año",ylab="Pollos")
   ```

---

5. Transformaciones

A menudo resulta imprescindible transformar las variables para adecuar la escala, o conseguir propiedades convenientes en los datos: cambios de escala, transformaciones logarítmicas, estandarización, cálculo de rangos, discretización o cualifiación de variables continuas, etc.

Ejercicios. Bloque 7:

1. Para obtener una variable presencia/ausencia puede recurrirse a una operación lógica, sabiendo que los valores `FALSE` y `TRUE` se corresponden con cero y uno respectivamente:

   ```
   >setosa<-(especie==1)+0
   ```

   ¿Se ha obtenido el código 1 para los ejemplares que corresponde a la especies *Iris setosa* y 0 para los demás?

2. Cualificar una variable.....Utilizando `cut(runif(100),seq(0,1,0.1))` pueden construirse intervalos en la variable y posteriormente tabular.

   ¿Qué ventajas tiene definir intervalos en el rango de la variable?

3. En ciertas ocasiones conviene considerar una variable cuantitativa como un factor cualitativo, por ejemplo, los niveles de concentración de una sustancia experimental: 5 gr, 10 gr, ... 100 gr. Esta tarea se realiza mediante la función `factor()`.

   ```
   summary(especie)
   summary(factor(especie))
   ```

4. El cambio de escala de una variable consiste en multiplicar los valores por una constante, un caso particular es expresar los datos en unidades de desviación típica, si los datos previamente se centran sobre su media hablamos de estandarización de la variable.

   ```
   ( ancpet-mean(ancpet) ) / sd(ancpet) ->aps
   ```

   o más sencillamente utilizando la función `scale()`:

   ```
   scale(ancpet) ->aps
   ```

   ¿Cuales son la media y la varianza de la nueva variable `aps`?

5. Para tamaños de muestra muy pequeños se recurre, en muchos casos, a pruebas estadísticas no paramétricas. Este tipo de muestras requieren las transformación de los valores en rangos:

   ```
   v <- c(3,0,2,1,1,3,3,6)
   rank(v)
   ```

   ¿Cómo se interpretan los valores de rango? Sugerencia, ordenar los valores de `v` en otra variable (`vo`), y repetir el ejercicio con esta variable.

---

6. Ejercicios adicionales

1. Realizar la simulación siguiente:

   ```
   a<-0;b<-1;error<-0.1;n<-20
   x<-runif(n)
   y<-a+b*x+error*(rnorm(n))
   plot(x,y)
   abline(a,b)
   ```

   Describir el gráfico resultante. Modificar los valores de error (por ejemplo: 0.01 y 2) y analizar el efecto que produce.

2. Se sospecha que la matriz de datos `territorios` refleja para los distintos años un numero medio de pollos distinto y un número menor ocupación de territorios.

   ¿Qué estadísticos deberían calcularse para poder reflejar la respuesta a estas cuestiones? ¿Qué resultados deberían obtenerse para reafirmar las sospechas?

3. Considerando el siguiente protocolo:

   ```
   read.table("http://www.um.es/docencia/emc/datos/evolpobmun.dat",header=T)->murcia
   attach(murcia)
   seleccion<-!is.na(apply(murcia,1,sum))
   x<-(cpa1950[seleccion]-cpa1900[seleccion])/cpa1900[seleccion]
   y<-(cpa1991[seleccion]-cpa1950[seleccion])/cpa1950[seleccion]
   plot(x,y,asp=1,main="Título",xlab="x",ylab="y")
   text(x,y,rownames(murcia)[seleccion],pos=3)
   abline(h=0,v=0)
   abline(0,1,col=2)
   ```

   ¿Qué representan las variables `x` y `y`? ¿Cómo se pueden interpretar las cuatro regiones definidas por los ejes de coordenadas? ¿Qué sentido tiene la recta representada con color rojo donde $y = x$? ¿Cuáles deben ser las etiquetas de los ejes y el título del gráfico?