ANALISIS DE SUPERVIVENCIA PARA DATOS AGRUPADOS. PARTE I: IMPUTACIÓN DE VALORES EN CASO DE CENSURA A INTERVALOS•

Servy, Elsa
Hachuel, Leticia
Boggio, Gabriela
Cuesta, Cristina
Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística

1. INTRODUCCIÓN

El análisis de supervivencia estudia un grupo de individuos para los cuales está definido un evento, frecuentemente llamado "falla" o "muerte", que ocurre después de un tiempo. La variable de interés en este tipo de estudio es el tiempo transcurrido desde un origen (que debe estar definido de manera clara) hasta la ocurrencia de la falla.

Una fuente especial de dificultad en el análisis de supervivencia es la posibilidad de que algunos individuos no puedan ser observados adecuadamente.

En este trabajo se considera el caso en que el origen ha sido definido con claridad pero el momento de la falla se conoce aproximadamente, más explícitamente, sólo se sabe que ocurrió dentro de un intervalo de tiempo. Este tipo de ambigüedad se conoce como "censura a intervalos" y los datos que la adolecen se dicen " censurados por intervalo".

El objetivo del presente estudio es obtener una manera de imputar el tiempo exacto de supervivencia en cada uno de los individuos, para luego realizar el análisis de los datos por métodos que suponen que las mediciones son exactas.

El problema que sirvió de motivación a este estudio proviene de una investigación entomológica cuyo material experimental se describe a continuación.

2. MATERIAL Y MÉTODO

El material experimental está compuesto por 138 larvas de la oruga *Colias lesbia fabrizius*, que fueron estudiadas desde que se produce la eclosión del huevo - evento que define el origen- hasta la formación de la crisálida o pupa - que representa la "falla"-. Es decir, el tiempo que se estudia es el de la duración del período larval, medido en días.

Las larvas recién nacidas se colocaban en hojas de alfalfa en el campo y se las cubría con pequeños dispositivos que además de protegerlas permitían su identificación. Los entomólogos visitaban el experimento a intervalos de tiempo no equiespaciados, de uno a siete días de duración.

Por el régimen de visitas empleado, se consideraba que una larva había terminado su período larval en el tiempo t si la formación de la crisálida se había producido entre el momento de la visita anterior y t. Las larvas se estudiaron a través de los años 1982 y 1984. Debido a que el estudio no se realizó en forma simultánea para las 138 larvas, los intervalos de censura no fueron siempre los mismos. Si la longitud de los intervalos hubiese sido común a todas las larvas, se podrían haber utilizado procedimientos estándares para el análisis del fenómeno bajo estudio.

El método usado en este trabajo fue sugerido por Cox y consiste en estimar el día exacto de formación de la crisálida mediante el algoritmo EM. El estudio hace uso de variables climáticas que pueden ayudar a explicar la duración del estado larval. A cada larva le corresponde un valor de la variable climática que es el promedio de los valores que ella toma durante los días de vida de la larva. Desde que existen imprecisiones en la longitud de la vida, las mismas se trasladan a las variables explicativas. De allí, fue necesario realizar imputaciones para las variables explicativas también.

• Financiado por el Programa de Fomento a la Investigación Científica y Tecnológica . SECYT.

3. EL ALGORITMO EM

A los fines de este estudio se considera que el dato no está censurado si la transformación en crisálida se produjo entre dos visitas realizadas en días consecutivos.

Si los datos no tuviesen censura, las frecuencias correspondientes a cada uno de los días de duración serían conocidos. En nuestro problema, sería conocido el número de crisálidas que se formaron exactamente a los 11 días, 12 días, etc. Estas frecuencias –no observadas en virtud del imperfecto modo de medición - se denominan con:

(3.1) $$k_{11}, k_{12}, ..., k_{28}$$

La distribución de probabilidad de esas frecuencias, bajo la hipótesis de independencia del momento de abandono del estado larval de las larvas es multinomial,

(3.2) $$f(k_{11},....,k_{28} / N\ \pi_1,...,\pi_{28}) = \frac{138!}{k_1!...k_{28}!}\pi_{11}^{k_{11}}...\pi_{28}^{k_{28}}$$

y,

(3.3) $$\log\ f(k_{11},....,k_{28} / N\ \pi_1,...,\pi_{28}) = \log\frac{138!}{k_1!...k_{28}!} + \sum_{j=11}^{28} k_j\phi_j \quad {}_j$$

donde $\phi_j = \log\pi_j$

3.1. PASO E DEL ALGORITMO EM.

Los datos realmente observados son S_1, S_2, ..., S_{138} , siendo S_i el intervalo, descrito en días, dentro del cual, según el informe del entomólogo, se formó la i-ésima crisálida. Por ejemplo, si la penúltima visita se realizó a los 15 días y la última a los 18 (allí se encontró formada la crisálida), S_i ={16, 17}. Cada S_i es un subconjunto del conjunto {11, ..., 28}.

k y {S_i , i=1, ...138} están vinculados por la siguiente igualdad:

(3.1.1) $$k_j = \sum_{i=1}^{138} Y_{ij}^* \quad ; \quad Y_{ij}^* = \begin{cases} 1 \text{ con probabilidad } \pi_j \Big/ |S_i| & \text{si } j \in S \\ 0 \text{ en caso contrario} \end{cases}$$

(3.1.2) $$|S_i| = \sum_{j \varepsilon S_i} \pi_j \ ; \quad j = 11,...,28$$

De donde,

(3.1.3) $$E(k_j) = \sum E(Y_{ij}^* / S_1,....,S_{138};\pi_{11},...,\pi_{28}) = \sum_{i=1}^{138} g_{ij}\frac{\pi_j}{|S_i|} \ ; \quad \text{donde } g_{ij} = \begin{cases} 1 \text{ si } j \in S \\ 0 \text{ en caso contrario} \end{cases}$$

La última expresión define el paso E del algoritmo.

3.2. PASO M DEL ALGORITMO EM.

El paso de maximización implica maximizar la función de verosimilitud (3.2) o su logaritmo (3.3) como si la muestra fuese completa, sólo que {k_j , j=11, ..., 28} se sustituye por sus esperanzas. Como en el caso multinomial la estimación máximo-verosímil de π_j es,

(3.2.1) $$k_j / N\ ,$$

la estimación máximo-verosímil se obtiene mediante iteraciones de la siguiente ecuación,

(3.2.2) $$\pi_j^{(p+1)} = \frac{1}{N}\sum g_{ij}\frac{\pi_j^{(p)}}{|S_i|}$$

El sobreíndice p designa a la iteración p-ésima.

Estimados los valores de las probabilidades, {π_j , j=11,..., 28}, se pueden obtener estimaciones

de las frecuencias $\{k_j\}$ utilizando (3.2.2).

En el Apéndice se muestra el programa con que se realizó el cómputo de las frecuencias que corresponden a los tiempos de duración del estado larval. Se probaron varios valores iniciales para estudiar el comportamiento del algoritmo. Todos ellos produjeron resultados prácticamente iguales, si bien se observaron diferencias en el número de iteraciones necesarias para lograr la estabilidad de los mismos. El procedimiento más rápido fue el que adjudicó idénticos valores a las probabilidades iniciales. Un problema observado fue que si en alguna de las iteraciones una probabilidad tomaba el valor cero, no conseguía salir de él en las sucesivas iteraciones. Para evitar este enquistamiento, toda vez que la probabilidad tomaba el valor nulo se lo sustituía por un valor pequeño –0.001- para que se pudiera continuar con las iteraciones.

4. IMPUTACIÓN DE LAS VARIABLES EXPLICATIVAS A LOS TIEMPOS DE DURACIÓN INDIVIDUALES

Además de registrar la fecha de la visita en que la larva se había encontrado convertida en pupa o crisálida, los biólogos consignaron, para cada larva, el valor medio asumido por ciertas variables climáticas durante el curso de su vida larval. Las variables estudiadas fueron:
Viento: en km/h, a 50 cm. de altura.

Humedad relativa: en %.

Heliofanía: horas de exposición solar efectiva diaria.

Lluvia: en mm totales caídos durante el período de vida de la larva.

Fotoperíodo: horas transcurridas desde la salida hasta la puesta del sol cada día.

Temperatura media: en °C, promedio de las mediciones diarias de temperatura registradas a las 9, 15 y 21 horas del día.

Los valores de las variables climáticas pueden indexarse con el subíndice "i", que identifica a las larvas. A manera de ejemplo, considérese la temperatura media que se designa con "X". Los valores observados de X son $X_{(i)}$, i=1, ..., 138. El dato que se desearía tener es, sin embargo, X_j, el valor de la temperatura media asociada con el tiempo exacto j-ésimo. Pero éste, en razón de la censura, es desconocido.

A partir del conocimiento de $X_{(i)}$ y S_i , i=1, ...138 se imputa X_j de la siguiente manera.

Considérese que, dado que j pertenece a S_i ,

$$P(X_{(i)} \Rightarrow X_j) = \frac{\pi_j}{|S_i|} \tag{4.1}$$

El valor a imputar a X_j es:

$$X_j = \sum X_{(i)} \frac{\pi_j}{|S_i|} g_{ij} / k_j \ , \tag{4.2}$$

habiendo sido definida g_{ij} en (3.1.3). Los $\{\pi_j$, j=11, ..., 28$\}$ se estiman a través de los valores obtenidos según (3.2.2). La división por k_j da a X_j la misma escala de medida que $X_{(i)}$.

5. RESULTADO DE LAS IMPUTACIONES

Los datos fueron imputados según los procedimientos descritos en las secciones 3 y 4. Antes de la aplicación de los algoritmos, el conjunto de 138 larvas se dividió en grupos de acuerdo con los valores de variables explicativas. Las larvas se clasificaron por tres variables dicotómicas a partir de las variables Temperatura media, Fotoperíodo y Viento, tomando como punto de corte el valor de la mediana respectiva. De los ocho estratos así construidos, dos no registraron datos, por lo cual se aplicó el algoritmo a los 6 restantes.

La distribución de frecuencias de la duración del período larval correspondiente a cada uno de los grupos se presenta en la Tabla 5.1.

Tabla 5.1**. Distribución de frecuencias esperadas de la duración del período larval clasificadas por estrato***

Día	Estratos						
	I	II	III	IV	V	VI	TOTAL
11	0	0	0	0	0	0	**0**
12	0	0	2	0	7	0	**9**
13	0	0	1	0	0	0	**1**
14	0	0	3	0	12	0	**15**
15	25	0	3	0	0	0	**28**
16	0	0	0	2	3	1	**6**
17	2	0	0	2	0	2	**6**
18	4	11	0	4	0	2	**21**
19	2	0	0	4	0	3	**9**
20	0	1	0	3	0	10	**14**
21	0	0	0	0	0	1	**1**
22	0	0	0	0	0	0	**0**
23	0	0	0	0	0	6	**6**
24	0	0	0	0	0	0	**0**
25	0	0	0	0	0	10	**10**
26	0	0	0	0	1	6	**7**
27	0	0	0	0	0	2	**2**
28	0	0	0	0	0	3	**3**
TOTAL	**33**	**12**	**9**	**15**	**23**	**46**	**138**

* Las frecuencias esperadas fueron redondeadas a números enteros

La aplicación de los algoritmos a conjuntos más homogéneos condujo a frecuencias totales mejor suavizada que las que se hallaron cuando los algoritmos se aplicaron a los datos sin previa estratificación.

Por el procedimiento de imputación empleado, los datos quedan identificados por los tiempos indexados con el subíndice j, y se pierde la identificación de la larva.

Los datos imputados no se pueden comparar con los "verdaderos", ya que los datos no censurados son escasos.

Sin embargo, de los 138 casos, 115 poseen una leve censura en el sentido que la inexactitud es, como máximo, de dos días.

Se toma, entonces, al conjunto de los datos levemente censurados como patrón de comparación de los datos imputados.

En las figuras que siguen se presentan las distribuciones de frecuencias de los datos levemente censurados y los imputados.

Figura 5.1
Distribución de los tiempos transcurridos hasta el evento
(Datos levemente censurados, N=115)

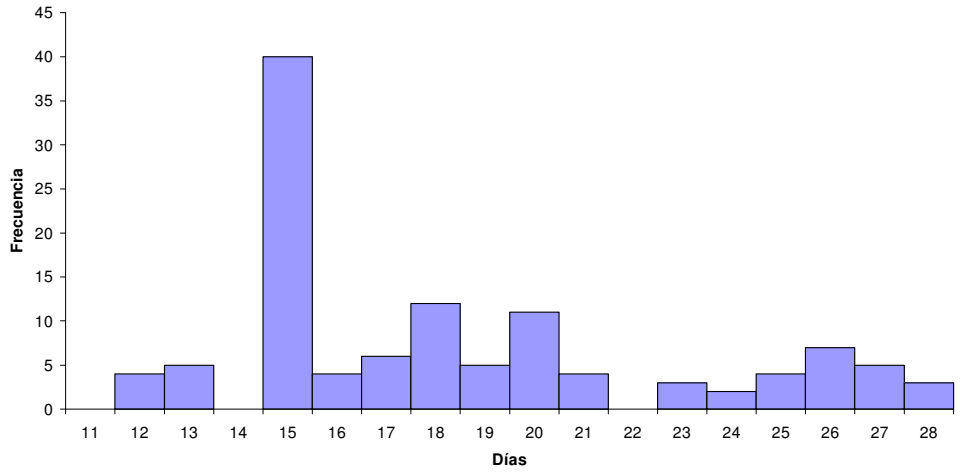
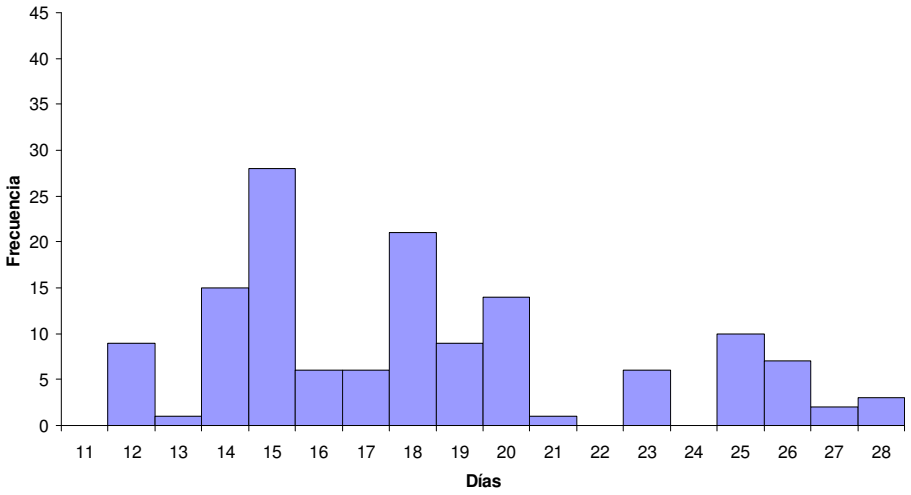


Figura 5.2
Distribución de los tiempos transcurridos hasta el evento
(Datos Imputados, N=138)



En la Tabla 5.2 se presentan medidas descriptivas de las variables estudiadas en las dos situaciones consideradas: datos levemente censurados y datos imputados.

Tabla 5.2**. Comparación de estadísticas descriptivas de la duración del estado larval y de las variables explicativas obtenidas a partir de los datos levemente censurados y los obtenidos por imputación.**

Variable	Datos	Media	Mediana	Modo	D. Est.	Mínimo	Máximo
Tiempo	**Lev.Cens.**	18.37	17.00	15.00	4.41	12.00	28.00
	Imputados	18.17	18.00	15.00	4.26	12.00	28.00
Fotoperíodo	**Lev.Cens.**	14.48	14.70	14.90	0.51	13.50	15.00
	Imputados	14.50	14.58	14.90	0.45	13.50	15.00
Heliofanía	**Lev.Cens.**	8.73	8.60	8.00	0.78	7.80	10.30
	Imputados	8.72	8.71	8.00	0.71	7.92	9.93
Humedad	**Lev.Cens.**	73.74	74.50	71.60	3.41	66.10	81.00
	Imputados	74.11	74.70	71.60	3.12	66.38	81.00
Lluvia	**Lev.Cens.**	56.66	6.00	86.90	24.60	2.40	86.90
	Imputados	57.23	53.29	86.90	19.88	5.34	86.90
Temperatura	**Lev.Cens.**	22.02	22.40	25.50	3.24	16.60	25.50
	Imputados	21.94	22.00	25.50	3.03	16.80	25.50
Viento	**Lev.Cens.**	5.05	5.10	5.30	0.56	4.40	6.80
	Imputados	5.06	5.15	5.30	0.56	4.40	6.77

En función de las medidas descriptivas presentadas parecería que existe una aceptable correspondencia entre ambos conjuntos de datos.

Concluida la tarea de "completar" los datos censurados, se aplicarán los mismos métodos de análisis de supervivencia diseñados para datos no censurados. Este es el tema central de una segunda parte de este trabajo.

BIBLIOGRAFÍA

COLLETT, D. (1994): *Modelling Survival Data in Medical Research*. Chapman & Hall, London.

Cox, D. R. & Oakes, D. (1984): *Analysis of Survival Data*. Chapman & Hall, London.

DEMPSTER, A.P.; LAIRD, N.M. & RUBIN, D.B. (1977): "Maximun Likelihood for Incomplete via the EM Algorithm". *Journal of the Royal Statistical Society* Nº1.

SAS Release 6.11. Statistical Analysis Software.

APÉNDICE

Cómputo de frecuencias por el algoritmo EM e imputación de las variables explicativas para uno de los estratos.

```
data datos;
  infile 'a:\estrato1.dat';
  input anio mes vien hum hel fot lluv tmed;
```

```
proc iml;

m={0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0,
   0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0,
   0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0,
   0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0};

 ek={1.83 1.83 1.83 1.83 1.83 1.83 1.83 1.83 1.83 1.83 1.83 1.83  1.83 1.83 1.83 1.83 1.83 1.83};
 r =j(33,18,0);
 k= j(33,18,0);
 pi=j(33,18,0);
 cont=0;

 do until (max(abs(a))<0.0001);
 cont=cont+1;
 do i=1 to 33;
    do j=1 to 18;
    if m[i,j]=0 then pi[i,j]=0; else do; pi[i,j]=ek[j]/33; end;
    end;
 end;
 s=pi[,+];

 do i= 1 to 33;
        do j=1 to 18;
        if pi[i,j]<>0 then k[i,j]=pi[i,j]/s[i];
        end;
 end;

 ek=k[+,];

 do l=1 to 18;
```

```
            if ek[l]<0.0001 then ek[l]=0.001;
end;
a=pi-r;
r=pi;
end;
print cont;

prever=log(s);
ver=prever[+];
* print s prever;
print ver;

use datos;
read all var _all_ into a;
close datos;

t=a[,8];
v=a[,3];
hu=a[,4];
he=a[,5];
f=a[,6];
ll=a[,7];

temp=(t`*k)/ek;
vien=(v`*k)/ek;
hum=(hu`*k)/ek;
hel=(he`*k)/ek;
fot=(f`*k)/ek;
lluv=(ll`*k)/ek;

res=ek//temp//vien//hum//hel//fot//lluv;
result=res`;
print 'ek     temp     vien     hum     hel     fot     lluv';

print result;
```