

WILD 7960 - Analysis of Wildlife Populations

www.auburn.edu/~grandjb/wildpop

Lecture 03 – Kaplan-Meier and Known Fate Analyses

Readings:

[Pollock, K. H., S. R. Winterstein, C. M. Bunck, and P. D. Curtis. 1989. Survival analysis in telemetry studies: the staggered entry design. *Journal of Wildlife Management* 53:7-15.](#)

Other resources:

1. Kaplan-Meier and similar known-fate methods

Bunck, C. M., and K. H. Pollock. 1993. Estimating survival in radio-tagged birds. Pages 51-63. In J. D. Lebreton and P. M. North (eds.) *Marked individuals in the study of bird populations*. Birkhauser Verlag/Basel, Switzerland.

Cox, D. R. and D. Oakes. 1984. *Analysis of survival data*. Chapman and Hall, New York, N. Y. 201 pp.

Pollock, K. H., S. R. Winterstein, and M. J. Conroy. 1989. Estimation and analysis of survival distributions for radio-tagged animals. *Biometrics*. 45:99-109.

Sauer, J. R., and B. K. Williams. 1989. Generalized procedures for testing hypotheses about survival or recovery rates. *Journal Wildlife Management* 53:137-142.

White et al.

<http://www.cnr.colostate.edu/~gwhite/mark/mark.htm>

(see documentation section)

http://www.cnr.colostate.edu/class_info/fw663/Mark.html

Hosmer, D.W. and Lemeshow, S. 2000. *Applied logistic regression*. 2nd Ed. John Wiley & Sons, Inc. New York.

Kaplan Meier (Product Limit) Procedure

1. Developed in the late 50's for use in medicine and engineering
2. First suggested for wildlife survival studies in mid- to late-80's by Pollock (1984) and Pollock et. al (1989).
3. Early approaches assumed that survival of every animal was independent (i.e., constant survival probability over all animals and periods)
4. Used when relocation is certain or unrelated to mortality
 - a. Radio-tagged animals
 - b. Plants or sessile creatures

Terms and definitions

WILD 7960 - Analysis of Wildlife Populations

www.auburn.edu/~grandjb/wildpop

Interval survival rates – the proportion of animals surviving some discrete period of time

Product-limit – refers to the fact that survival rate during any interval can be calculated as the product of the survival rates during shorter intermediate intervals. Thus the survival rate for the interval can never be greater than the lowest estimate of survival during any intermediate period.

Survival function – equation describing the probability of an animal in population surviving t units of time from the beginning of a study.

At-risk – having some probability of detectable mortality. Thus, in radio-telemetry studies, transmitter failure or emigration beyond the study area boundaries is grounds for removal from the at-risk group

Right censor – to remove an animal from the at-risk group for some reason unrelated to mortality

Staggered Entry – the addition of additional subjects to the at risk group during the course of the study.

Study design

1. Individuals marked or uniquely identifiable
2. Subjects can be relocated without failure (except when censoring)
3. All individuals relocated at the end of each survival interval (e.g., at each a_j)
4. Assumptions
 - a. Animals in the population of interest have been sampled randomly – sex, age, location, etc.
 - b. Survival times are independent for different animals
 - c. Time of death is known exactly
 - d. Capturing, marking, and observing does not influence survival (but may condition to eliminate potential effects)
 - e. Censoring is random (i.e., unrelated to fate – either survival or mortality)
 - 1) May be useful to assume that marker failure is zero (e.g. studies of emigration time)
 - f. Identification of time origin is crucial
 - 1) in medical and engineering studies – treatment time
 - 2) in wildlife studies – may be no natural origin time (e.g., season or other) may be crucial for example studies in winter may have a different survival rate than studies in summer; survival rates may vary greatly after the start of a hunting season or other event
 - g. Staggered entry assumes that newly tagged animals have the same survival function as previously tagged
 - 1) condition data
 - 2) assumptions can be modeled

WILD 7960 - Analysis of Wildlife Populations

www.auburn.edu/~grandjb/wildpop

Basic K-M calculations

a_1, a_2, \dots, a_g – discrete time points when deaths (failures) occur
(alternatively use regular sampling periods e.g. days, weeks, etc.)

r_1, r_2, \dots, r_g – number of animals at risk at these time points

d_1, d_2, \dots, d_g – number of deaths at these time points

$$\hat{S}(a_i) = 1 - \frac{d_i}{r_i}$$

is the probability of surviving the interval from a_{i-1} to a_i and

$$\hat{S}(a_{i+1}) = \left(1 - \frac{d_i}{r_i}\right) \left(1 - \frac{d_{i+1}}{r_{i+1}}\right)$$

is the probability of surviving the interval from a_{i-1} to a_{i+1}

$$\text{Thus } \hat{S}(t) = \prod_{i=1}^m (1 - d_i / r_i)$$

is a short way of writing the product of the survival rates in all m intervals.

Censoring and staggered entry

1. When individuals are right censored, the number in the at risk group r_i is reduced at time a_i .
2. Likewise when individuals are entered into the at risk sample the value r_i is increased at time. Alternatively, newly marked individuals can be "conditioned" to allow for marking or handling effects by delaying their addition to the at risk group for some predetermined time period.

Estimating standard deviation and SE

$$\text{var}(\hat{S}) = \hat{S}_j \sum_1^j \frac{d_j}{r_j(r_j - d_j)}$$

or, a simpler formula that behaves better at smaller and larger values of \hat{S} is

$$\text{var}(\hat{S}) = \frac{\hat{S}_j^2(1 - S_j)}{r_j}$$

The standard deviation and the SE = $\sqrt{\text{var}(\hat{S})}$.

Confidence intervals

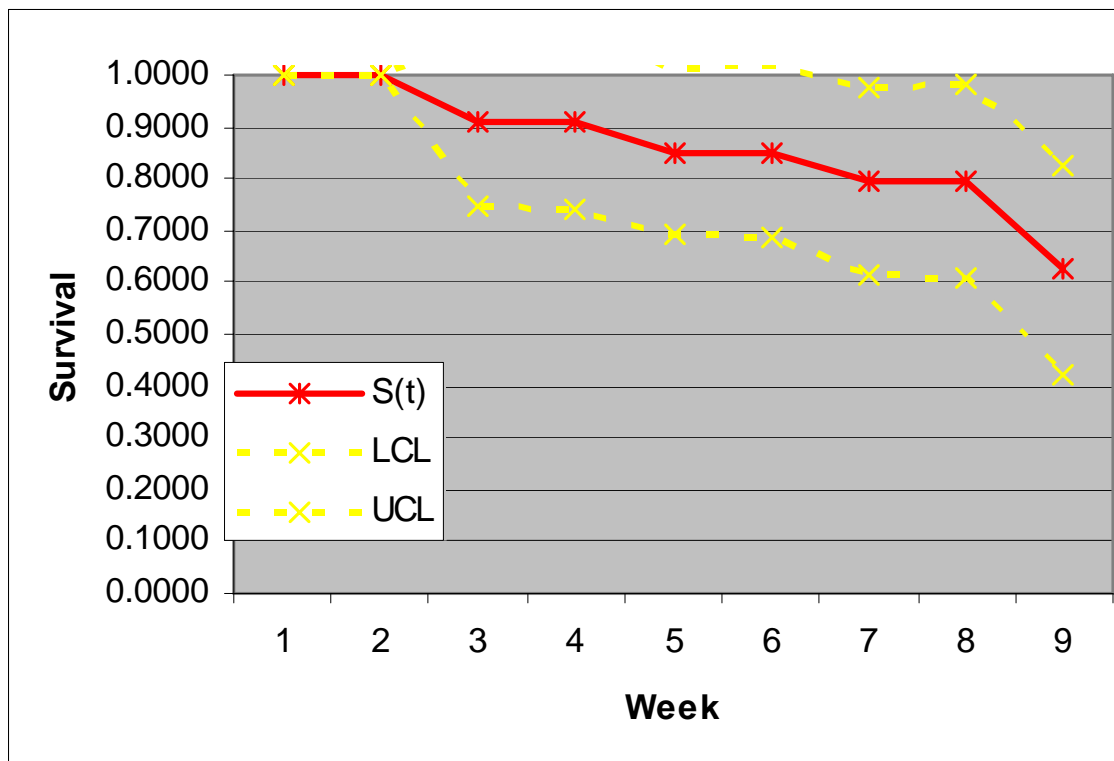
Approximate 95%: $\hat{S} \pm 1.96(\text{var}(\hat{S}))$

WILD 7960 - Analysis of Wildlife Populations

www.auburn.edu/~grandjb/wildpop

Table 3-2 from Pollock et al. 1989: Bobwhite survival in fall of 1986

Week	No. risk	No. deaths	S_i	Survival	No. censored	No. added	Var(S)	LCL	UCL
1	7	0	1.0000	1.0000	1	0	0.0000	1.0000	1.0000
2	6	0	1.0000	1.0000	0	5	0.0000	1.0000	1.0000
3	11	1	0.9091	0.9091	0	0	0.0826	0.7471	1.0711
4	10	0	1.0000	0.9091	0	6	0.0867	0.7392	1.0790
5	16	1	0.9375	0.8523	0	0	0.0819	0.6918	1.0128
6	15	0	1.0000	0.8523	0	0	0.0846	0.6865	1.0180
7	15	1	0.9333	0.7955	0	0	0.0929	0.6134	0.9775
8	14	0	1.0000	0.7955	0	0	0.0961	0.6070	0.9839
9	14	3	0.7857	0.6250	0	0	0.1023	0.4245	0.8255
Total		6		0.6250	1	11			



WILD 7960 - Analysis of Wildlife Populations

www.auburn.edu/~grandjb/wildpop

Known Fate Models

Known fate models use MLE for Kaplan-Meier-like problems. Basically, these are binomial probability models used to describe data that meet the assumption that subjects can be relocated and fate determined without failure. In other words the probability associated with recapture or recovery is assumed to be 1.0 and thus is not estimated. However, there is much greater flexibility in model selection and the procedure is based on MLE.

We estimated the survival rate (probability) of an animal via Kaplan-Meier method by,

$$\hat{S}(t) = \prod_{i=1}^m (1 - d_i / r_i),$$

where d_i was the number dying and r_i was the number at risk during i^{th} interval of some longer time period t .

If we think of this estimator in a slightly different fashion:

$$\hat{S}_i = \frac{y_i}{n_i} \text{ or } \hat{S}(t) = \prod_{i=1}^m \left(\frac{y_i}{n_i} \right)$$

where y_i is the number surviving and n_i is the number at risk, it should be fairly that the likelihood for this equation is:

$$\mathcal{L}(\theta | n_i, y_i) = \prod_{i=1}^m S_i^{y_i} (1 - S_i)^{n_i - y_i},$$

where θ is the survival model for the m periods, n_i is the number of individuals at risk during each period, y_i is the number surviving each period and S_i is the MLE of survival during each period.

This also can be written as the product of the probability of the individual observations:

$$\mathcal{L}(\theta | n_i, y_i) = \prod_{i=1}^m \prod_{j=1}^{n_i} S_{ij}^f (1 - S_{ij})^{1-f}.$$

or the sum of the log-likelihoods.

$$\ln(\mathcal{L}(\theta | n_i, y_i)) = \sum_{i=1}^m \sum_{j=1}^{n_i} f \ln(S_{ij}) + (1 - f) \ln(1 - S_{ij}),$$

where S_{ij} is the probability of individual i surviving the j^{th} interval.

WILD 7960 - Analysis of Wildlife Populations

www.auburn.edu/~grandjb/wildpop

Link functions and covariates

Link functions are used to examine the effects of covariates on survival. In effect, they provide the linkage between observations (data) and the survival estimate. These covariates may be discrete or continuous. For example we could examine the effect of week, age, or weight on weekly survival rates.

Recall that the likelihood for the n individuals over the m weeks in our known fate analysis was

$$\mathcal{L}(\theta | n_i, y_i) = \prod_{i=1}^m \prod_{j=1}^n S_{ij}^f (1 - S_{ij})^{1-f}$$

In order to estimate the survival rate as a function of some variable (e.g., age or weight) MARK uses a *link* function. Frequently this is done by way of the logit link:

$$S_i = \frac{e^{(\beta_0 + X_{1i}\beta_1)}}{1 + e^{(\beta_0 + X_{1i}\beta_1)}} \quad (1)$$

where X_{1i} is the first covariate (data) in the i^{th} interval. Other functions could be used, but the logit is the most familiar and is the equivalent of logistic regression. Since the data in the link can be coded to represent time periods or other discrete groups (more on how we do this later), we can drop the product of the likelihoods and substituting the logit link into the likelihood:

$$\mathcal{L}(\theta | n_i, y_i) = \prod_{j=1}^n \left(\frac{e^{(\beta_0 + X_{1i}\beta_1)}}{1 + e^{(\beta_0 + X_{1i}\beta_1)}} \right)^f \left(1 - \left(\frac{e^{(\beta_0 + X_{1i}\beta_1)}}{1 + e^{(\beta_0 + X_{1i}\beta_1)}} \right) \right)^{1-f}$$

(ignoring the binomial coefficient), thus the log-likelihood is:

$$\mathcal{L}(\theta | n_i, y_i) = \sum_{j=1}^n \left[f \ln \left(\frac{e^{(\beta_0 + X_{1i}\beta_1)}}{1 + e^{(\beta_0 + X_{1i}\beta_1)}} \right) + (1 - f) \ln \left(1 - \left(\frac{e^{(\beta_0 + X_{1i}\beta_1)}}{1 + e^{(\beta_0 + X_{1i}\beta_1)}} \right) \right) \right] \quad (2)$$

where n is the number of individuals in the sample.

Obviously, more than one covariate can be included and the linear equation extended. When the link function is employed, the β s are the estimated parameters; like the survival rates can be estimated for each occasion or constrained to be equal where appropriate using the data. The survival rates or real parameters of interest are calculated from the β s as in equation 1 above.

This concept has application to every estimator we will examine. The important concept here is that the survival probabilities are replaced by the link function submodel of the covariate(s). Conceivably every animal has a different survival probability that is related to the value of the covariates X_{ij} .

Example

Consider this survival data set for a fictitious critter. Let's say it's a radio-marked mouse. All of the mice are captured and radio-marked on the same night, and weights are determined once, at the time of capture. These mice are classified as belonging to two discrete groups (e.g., gender). At the end of the first time period (0), #5 is discovered to have died. By the

WILD 7960 - Analysis of Wildlife Populations

www.auburn.edu/~grandjb/wildpop

end of the second period mouse #4 also is dead.

Obs	Individual (i)	Fate (Survived) (y _i)	Time (x _{1i})	Group (x _{1i})	Mass at capture (g) (x _{3i})
1	1	1	0	0	26
2	2	1	0	0	28
3	3	1	0	1	25
4	4	1	0	1	22
5	5	0	0	1	20
6	1	1	1	0	26
7	2	1	1	0	28
8	3	1	1	1	25
9	4	0	1	1	22

Model 1: Survival rates vary between times

Let, β_0 equal the survival rate of individuals in group 0 during time 0; let β_1 equal the difference in survival between time periods 0 and 1. Thus, for the individual 1 in time period 1:

$$S_1 = \frac{e^{(\beta_0 + X_{11}\beta_1)}}{1 + e^{(\beta_0 + X_{11}\beta_1)}} = \frac{e^{(\beta_0 + 0\beta_1)}}{1 + e^{(\beta_0 + 0\beta_1)}}$$

In the table below the column labeled $\beta_0 + x_{1i}\beta_1$ is the linear equation, this linear equation is called the **logit**, for this model, \hat{S} is from equation 1, and $\ln(\mathcal{L})$ is from equation 2. The $\Sigma \ln(\mathcal{L})$ for the observations is -8.62 based on initial values for $\beta_i = 0.5$.

Obs	Individual (i)	Fate (Survived) (y _i)	Time (x _{1i})	$\beta_0 + x_{1i}\beta_1$	\hat{S}	$\ln(\mathcal{L})$
1	1	1	0	0.50	0.378	-0.97
2	2	1	0	0.50	0.378	-0.97
3	3	1	0	0.50	0.378	-0.97
4	4	1	0	0.50	0.378	-0.97
5	5	0	0	0.50	0.378	-0.47
6	1	1	1	1.00	0.269	-1.31
7	2	1	1	1.00	0.269	-1.31
8	3	1	1	1.00	0.269	-1.31
9	4	0	1	1.00	0.269	-0.31
Σ					-8.62	

Using the solver, we find the MLE for $\beta_0 = -1.39$ and $\beta_1 = 0.29$ as in the table below. Note that because of the model specification the \hat{S} are equal for all individuals in each time period.

WILD 7960 - Analysis of Wildlife Populations

www.auburn.edu/~grandjb/wildpop

Obs	Individual (i)	Fate		Time (x_{1i})	$\beta_0 + x_{1i}\beta_1$	\hat{S}	$\ln(\mathcal{L})$
		(Survived) (y_i)					
1	1	1		0	1.39	0.800	-0.22
2	2	1		0	1.39	0.800	-0.22
3	3	1		0	1.39	0.800	-0.22
4	4	1		0	1.39	0.800	-0.22
5	5	0		0	1.39	0.800	-1.61
6	1	1		1	1.10	0.750	-0.29
7	2	1		1	1.10	0.750	-0.29
8	3	1		1	1.10	0.750	-0.29
9	4	0		1	1.10	0.750	-1.39
Σ							-4.75

Model 2: Survival varies by group and time.

Let, β_0 equal the survival rate of individuals in group 0 during time 0 ignoring the effect of mass on survival. Let β_1 equal the difference in survival between time periods 0 and 1; let β_2 equal the difference in survival between groups 0 and 1. Thus, it follows from Equation 1 that for individual 1:

$$\begin{aligned}
 S_1 &= \frac{e^{(\beta_0 + X_{11}\beta_1 + X_{21}\beta_2)}}{1 + e^{(\beta_0 + X_{11}\beta_1 + X_{21}\beta_2)}} \\
 &= \frac{e^{(\beta_0 + 0\beta_1 + 0\beta_2)}}{1 + e^{(\beta_0 + 0\beta_1 + 0\beta_2)}}
 \end{aligned}$$

In the table below the column labeled $\beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2$ is the linear equation for this model, \hat{S} is from equation 1, and $\ln(\mathcal{L})$ is from equation 2. The $\Sigma \ln(\mathcal{L})$ for the observations is -9.42 based on initial values for $\beta_i = 0.5$.

Obs	Individual (i)	Fate		Time (x_{1i})	Group (x_{2i})	$\beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2$	\hat{S}	$\ln(\mathcal{L})$
		(Survived) (y_i)						
1	1	1		0	0	0.50	0.378	-0.97
2	2	1		0	0	0.50	0.378	-0.97
3	3	1		0	1	1.00	0.269	-1.31
4	4	1		0	1	1.00	0.269	-1.31
5	5	0		0	1	1.00	0.269	-0.31
6	1	1		1	0	1.00	0.269	-1.31
7	2	1		1	0	1.00	0.269	-1.31
8	3	1		1	1	1.50	0.182	-1.70
9	4	0		1	1	1.50	0.182	-0.20

WILD 7960 - Analysis of Wildlife Populations

www.auburn.edu/~grandjb/wildpop

Σ **-9.42**

Using the solver, we find and $\ln(\mathcal{L})$ is maximized when $\beta_0 = -16.12$, $\beta_1 = 0.6932$, and $\beta_2 = 15.43$. Note that because of the model specification the \hat{S} are equal within groups and time periods.

Obs	Individual (i)	Fate (Survived) (y _i)	Time (x _{1i})	Group (x _{2i})	$\beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2$	\hat{S}	$\ln(\mathcal{L})$
1	1	1	0	0	16.12	1.000	0.00
2	2	1	0	0	16.12	1.000	0.00
3	3	1	0	0	16.12	0.667	-0.41
4	4	1	0	1	0.69	0.667	-0.41
5	5	0	0	1	0.69	0.667	-1.10
6	1	1	1	0	15.43	1.000	0.00
7	2	1	1	0	15.43	1.000	0.00
8	3	1	1	0	0.00	0.500	-0.69
9	4	0	1	1	0.00	0.500	-0.69
Σ							-3.30

Model 3: all survival probabilities equal (null).

Let, β_0 equal the survival rate of all individuals. Thus, for each individual:

$$S_1 = \frac{e^{(\beta_0)}}{1 + e^{(\beta_0)}}$$

In the table below the column labeled β_0 is the logit for this model, \hat{S} is from equation 1, and $\ln(\mathcal{L})$ is from equation 2. The $\Sigma \ln(\mathcal{L})$ for the observations is based on the MLE for β_0 (-1.25).

Obs	Individual (i)	Fate (Survived) (y _i)	Time (x _{1i})	Group (x _{2i})	β_0	\hat{S}	$\ln(\mathcal{L})$	Obs
1	1	1	0	0	0.5	1.25	0.778	-0.25
2	2	1	0	0	0.5	1.25	0.778	-0.25
3	3	1	0	1	0.5	1.25	0.778	-0.25
4	4	1	0	1	0.5	1.25	0.778	-0.25
5	5	0	0	1	0.5	1.25	0.778	-1.50
6	1	1	1	0	0.5	1.25	0.778	-0.25
7	2	1	1	0	0.5	1.25	0.778	-0.25
8	3	1	1	1	0.5	1.25	0.778	-0.25
9	4	0	1	1	0.5	1.25	0.778	-1.50
Σ							-4.77	

WILD 7960 - Analysis of Wildlife Populations

www.auburn.edu/~grandjb/wildpop

Model Selection

Model name	K	$\ln(\mathcal{L})$	AIC_c	ΔAIC_c	$\mathcal{L}(\theta)$	w_i
S_t	2	-4.75	15.50	3.39	0.18	0.15
S_{t+g}	3	-3.30	17.40	5.29	0.07	0.06
$S.$	1	-4.77	12.11	0.00	1.00	0.80
Σ						1.00