24 SURVIVAL ANALYSIS

Objectives

- Simulate the fates of 25 individuals over a 10-day period.
- Calculate the Kaplan-Meier product limit estimate.
- Graphically analyze the Kaplan-Meier survival curve.
- Assess how sample size affects the Kaplan-Meier estimate.
- Assess how censorship affects the Kaplan-Meier estimate.

Suggested Preliminary Exercise: Life Tables and Survivorship Curves

INTRODUCTION

A population of black bears has been surveyed for 10 years, and ecologists note that the number of bears in the population has declined over this time frame. Why? Changes in numbers of individuals over time can be directly traced back to the population's birth, death, immigration, and emigration rates. The population may have declined because the birth rate dropped, the death rate increased, immigration dropped, or emigration increased. A combination of any or all of these factors may also be responsible for the decline. *Mortality* and its counterpart, *survival*, are keys to the demographic equation for all organisms. How do ecologists estimate these two important parameters? In this exercise we'll explore one method for estimating survival.

In your life table exercise, you tracked the fates of individuals over time, noting how many individuals in the cohort were still alive at each time step, and then calculated the survivorship schedule and survival probabilities from your data. Suppose we followed a cohort of 100 newborns over time, carefully noting when deaths occurred. We start with $S_0 = 100$, count individuals again at the next time step (S_1) and then at time step S_2. Suppose $S_1 = 40$ and $S_2 = 10$. The *survivorship schedule* (see Exercise 12, "Life Tables, Survivorship Curves, and Population Growth") tells us that the probability that an individual will survive *from birth to time x*. Thus, the probability of surviving to age 1 is $S_1/S_0 = 40/100 = 0.4$, and the probability of surviving from birth to age 2 is $S_2/S_0 = 10/100 = 0.1$. **Age-specific survival probabilities**, in contrast, tell us the probability that an individual will survive ***from one age to the next***—such as the probability that an individual alive in time S_1 will be alive at time S_2. In life table calculations, the age-specific survival probability is calculated as $g_x = l_{x+1}/l_x$. In our example, the probability that an individual of age S_1 will survive to age S_2 is 0.10/0.40 = 0.25. The life table "cohort"

analysis is one way of calculating survival. However, this method is not always possible to use, especially if the organisms of interest are long-lived. Fortunately, alternatives for estimating survival exist.

Kaplan-Meier Survival Analysis

When the research question can be posed as "how long does it take until death occurs?" the **Kaplan-Meier survival analysis**, also known as the Kaplan-Meier product limit estimate or the Kaplan-Meier survival curve, can be used to estimate survival. The Kaplan-Meier method (1958) involves tracking the fates of individuals over time and estimating how long it takes for death to occur. The method has been applied broadly to measure how long it takes for *any specific event* to occur—such as the time it takes until death, the time until a cancer patient recovers from a treatment, the time until an infection appears, the time until pollination occurs, and so on.

The Kaplan-Meier method is conceptually similar to life table calculations because you keep track of the number of individuals alive and the number of deaths that occur over intervals of time. Specifically, you count the number of individuals who die at a certain time and divide that number by the number of individuals that are "at risk" (alive and part of the study) at that time. If we do this for each time period in the study, we will be able to compute two survival probabilities: the *conditional survival probability* and the *unconditional survival probability*. We will describe how each is computed with a brief example.

Suppose you initiate a study on beetle mortality and track 20 individuals over 5 days, each day recording the number of deaths and the number of individuals still alive. Let's also suppose that some of your population decides to emigrate out of the population so you can no longer track their fates. The data you collect are:

	A	B	C
1	Day	Emigrants	Deaths
2	1	1	3
3	2	0	4
4	3	1	2
5	4	0	1
6	5	0	2

Now let

t be a particular time period, such as 1 day
d be the number of deaths at time t_i
n be the number of individuals at risk at the beginning of time t_i.

The **conditional survival probability,** P_c, is the probability of surviving to a specific time, given that you survived to the previous time (this is similar to the age-specific survival probabilities in the life table). P_c is computed as

$$P_c = 1 - \frac{d_i}{n_i} \qquad \text{Equation 1}$$

The term d_i / n_i gives the number of individuals that die in time step i divided by the number of individuals still alive *and* still in the population (the number at risk). This is the conditional mortality probability, or the probability that an individual will die during that time step. Since survival can be computed as 1 minus mortality, Equation 1 gives the conditional survival probability.

Because we started with a population of 20 individuals, the number at risk for death at the beginning of day 1 is 20. During that day, 3 individuals died, so the conditional mortality probability is 3/20 = 0.15, and the conditional survival probability is 1 – 0.15 = 0.85. Now let's consider day 2. At the beginning of day 2, there are only 16 individuals at risk. Three individuals died the previous time step, and one left the population through emigration. The individual that left the study is called a **censored observation**.

Individuals that die in the previous time step, as well as censored individuals, cannot be considered at risk, so on day 2 only 16 individuals are at risk. On day 2, 4 deaths occurred, so the conditional mortality probability is 4/16 = 0.25, and the conditional survival probability is 1 – 0.25 = 0.75. The rest of the computations are shown in Figure 1.

	A	B	C	D	E	F
1	Day	Emigrants	Deaths	# at risk	Deaths / at risk	P_c
2	1	1	3	20	= 3/ 20 = 0.15	1 - 0.15 = 0.85
3	2	0	4	= 20 - 3 - 1 = 16	= 4 / 16 = 0.25	1 - 0.25 = 0.75
4	3	1	2	= 16 - 4 - 0 = 12	= 2 / 12 = 0.16	1 - 0.16 = 0.84
5	4	0	1	= 12 - 2 - 1 = 9	= 1 / 9 = 0.11	1 - 0.11 = 0.89
6	5	0	2	= 9 - 1 = 8	= 2 / 8 = 0.25	1 - 0.25 = 0.75

Figure 1

The **unconditional survival probability** , P_u, is the probability of survival from the start of the study *to* a specific time (this is similar to the survivorship schedule in the life table). The unconditional probability is equal to the cumulative product of the conditional probabilities, which is why the Kaplan-Meier method is sometimes called the Kaplan-Meier product limit estimate. The equation can be expressed as

$$P_u = \prod_{j=1}^{i}\left(1-\frac{d_j}{n_j}\right) \qquad \text{Equation 2}$$

where the Π symbol means "multiply all of the individual conditional probabilities together." The computations are shown in Figure 2.

For day 1, the unconditional survival probability is the same as the conditional survival probability. P_u for day 2 gives the probability that an individual at the start of the study will survive through day 2. This is obtained by multiplying the conditional survival probability for day 1 by day 2, since both conditions must be met in order for an individual to be alive at the end of day 2.

	A	B	C	D	E	F	G
1	Day	Emigrants	Deaths	# at risk	Deaths / at risk	P_c	P_u
2	1	1	3	20	= 3/ 20 = 0.15	1 - 0.15 = 0.85	= 0.85
3	2	0	4	= 20 - 3 - 1 = 16	= 4 / 16 = 0.25	1 - 0.25 = 0.75	= 0.85 * 0.75 = .6375
4	3	1	2	= 16 - 4 - 0 = 12	= 2 / 12 = 0.16	1 - 0.16 = 0.84	= 0.85 * 0.75 * 0.84 = .54
5	4	0	1	= 12 - 2 - 1 = 9	= 1 / 9 = 0.11	1 - 0.11 = 0.89	= 0.85 * 0.75 * 0.84 * 0.89 = .48
6	5	0	2	= 9 - 1 = 8	= 2 / 8 = 0.25	1 - 0.25 = 0.75	= 0.85 * 0.75 * 0.84 * 0.89 * 0.75 = .36

Figure 2

Notice that P_u decreases with each day because the probability of living to a given period must decrease as ever-greater time periods are considered. Sometimes ecologists are interested in expressing P_u as a daily probability. To obtain a daily survival estimate, you would take the appropriate root. For example, P_u = 0.36 on day 5 in Figure 2. This gives the probability that an individual will survive through day 5. What would daily survival be to obtain P_u = 0.36 on day 5? A daily probability of x would have to yield 0.36 when multiplied by itself once for each day, so $x^5 = 0.36$. By taking the fifth root of 0.36, you could solve for x. The spreadsheet formula is **0.36^(1/5)**.

Kaplan-Meier Survival Curves

The results of the Kaplan-Meier analysis are often graphed; graphs are known as the Kaplan-Meier survival curves (Figure 3). Comparing the survival curves of two different populations can yield insightful information about the timing of deaths in

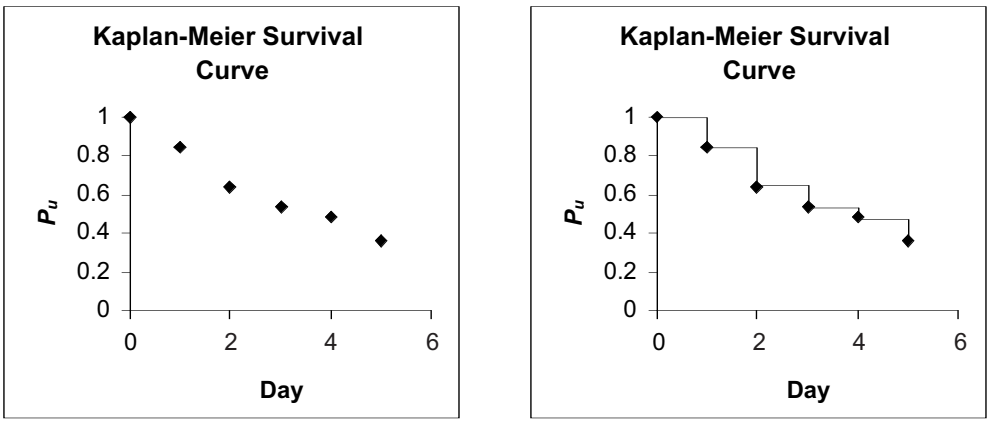

Figure 3 Kaplan-Meier survival curves for a hypothetical population. The unit time is plotted on the *x*-axis; P_u is plotted on the *y*-axis. In Kaplan-Meier curves, the raw data are plotted as in the graph on the left, then the data points are connected with horizontal and vertical bars as shown on the right. Large vertical steps downward indicate a large number of deaths in the given time period, while large horizontal steps indicate few deaths have occurred during an interval.

response to different environmental conditions. Often in the literature, you will see the survival curves for two different populations on the same graph so that you can compare the two easily.

PROCEDURES

The method outlined by Kaplan and Meier (1958) is one of the most referenced papers in the field of science, suggesting that is has played an important role in ecology and other sciences since its publication. The goal of this exercise is to set up a spreadsheet model of the Kaplan-Meier product limit estimate, and to learn how censored observations and sample size affect the survival probabilities. As always, save your work frequently to disk.

INSTRUCTIONS

ANNOTATION

A. Set up the model population.

1. Open a new spreadsheet and set up column headings as shown in Figure 4.

We'll track 25 individuals for 10 days and keep track of their fates over time. Row 10 will track Individual 1's fate, Row 11 will track Individual 2's fate, and so on to Row 34.

	A	B	C	D	E	F	G	H	I	J	K
1	***Survival Analysis***										
2											
3	**Model Inputs:**										
4	Survival =	0.9									
5	Total sample =	25									
6	Prob. of censor =	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
7											
8		**Day**									
9	**Individual**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

Figure 4

2. Set up a linear series from 1 to 25 in cells A10–A34.

In cell A10 enter the value 1.
In cell A11 enter the formula **=1+A10**. Copy this formula down to cell A34.

3. In cell B4, enter a value for the probability that an individual will survive each 24-hour period (daily survival).

Enter the value 0.9 in cell B4. In reality, you wouldn't know what this number is; you are using the Kaplan-Meier method to estimate this parameter.

4. Enter the number of individuals in the initial population in cell B5.

Enter the value 25 in cell B5.

5. In cells B6–K6, enter a value for the probability that an individual in the population will be censored on a given day.

Enter the value 0.1 in cells B6–K6.
This is the probability that an individual will leave the study on any given day so that its fate cannot be tracked over time. For now, we set that probability to 0.1 for all days. Later in the exercise you will change these values to determine how censored observations, and the time at which they occur, affect survival probability estimates.

5. Save your work.

B. Simulate fates of individuals over time.

1. In cells B10–B34, enter a formula to assign a fate to each individual for day 1.

In cell B10 enter the formula **=IF(RAND()<$B$6,"C",IF(RAND()>$B$4,"D",1))**. Copy your formula down to row 34.
The formula in B10 will assign a fate to individual 1 on day 1. The individual will be either alive (1), censored (C), or dead (D). The formula contains two **IF** functions and a **RAND** function, so it is a nested formula. Remember that the **IF** function consists of three parts separated by commas. In the first part of the function, you specify a criteria. If the criteria is true, the spreadsheet will do or carry out whatever you specify in the second portion of the function. If the criterion is false, the spreadsheet will carry out what you specify in the third portion of the function. Let's review the B10 formula carefully.

The criterion is that a random number (the **RAND()** portion of the formula) is less than the value in cell B6 (the probability of being censored on day 1). If the criterion is true, the individual is censored and the spreadsheet will return the letter *C*. If the criterion is false, the individual is not censored, and the second **IF** function will be computed.

The second **IF** function tells the spreadsheet to evaluate whether a random number between 0 and 1 is greater than the value in cell B4—the true (but unknown to you, the researcher) daily survival probability. If the random number is greater than the survival probability, the individual will die (the spreadsheet will return the letter D). If the random number is less than the value in cell B4, the spreadsheet will return the number 1, indicating that the individual survived that day. When you copy your formula down for the 25 individuals in the population, you should see that some individuals die and some become censored. Press F9, the calculate key, to generate new fates for individuals in the population.

2. In cell C10, enter a formula to assign a fate to individual 1 for day 2.

In cell C10 enter the formula **=IF(OR(B10="D",B10="C",B10=""),"",IF(RAND()<$C$6,"C",IF(RAND()>$B$4,"D",1)))**. Don't be intimidated by the length of this formula.
If the individual in cell C10 died or was censored on day 1, we want to return a blank cell (i.e., two double quotes). If the individual survived day 1, then we want to know what happened on day 2. The formula in cell C10 is another nested **IF** function. There

are multiple criteria, however, in the first **IF** function, and these criteria are given with an **OR** function. The **OR** function is used to evaluate whether the value in cell B10 is "D" or "C" or "". If any one of those three conditions is *true*, the spreadsheet will return a blank, or "". If none of the conditions is true, the individual must have survived day 1, and the second **IF** function is computed; it has the same form as the formula in cell B10, with the spreadsheet again returning a value of "C," "D," or the number 1.

3. Select cell C10, and copy its formula across to cell K10. Modify the formula in each cell to reflect the probability of censorship for the appropriate day.

Double-check your formulae. They should read as follows:

- In cell D10,
  =IF(OR(C10="D",C10="C",C10=""),"",IF(RAND()<$D$6,"C",IF(RAND()>$B$4,"D",1)))
- In cell E10,
  =IF(OR(D10="D",D10="C",D10=""),"",IF(RAND()<$E$6,"C",IF(RAND()>$B$4,"D",1)))
- In cell F10,
  =IF(OR(E10="D",E10="C",E10=""),"",IF(RAND()<$F$6,"C",IF(RAND()>$B$4,"D",1)))

and so on. Your spreadsheet should now resemble Figure 5, although the fates of your individuals will likely be different than that shown.

4. Select cells C10–K10, and copy the formula down to row 34.

5. Save your work.

	A	B	C	D	E	F	G	H	I	J	K
8		Day									
9	**Individual**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
10	1	1	D								
11	2	1	1	C							
12	3	C									
13	4	1	1	1	1	1	1	1	1	C	
14	5	1	1	1	1	D					

Figure 5

C. Compute survival probabilities.

The first calculations in the Kaplan-Meier estimate involve counting the number of individuals at risk (still alive) during each day, and to count the number of deaths that occur each day.

1. Set up new headings as shown in Figure 6.

	A	B	C	D	E	F	G	H	I	J	K
35	# at risk										
36	# deaths										
37	# censored										
38	Conditional P_c										
39	Unconditional P_u										
40	Expected survival										
41	Daily survival										

Figure 6

2. In cell B35, enter the number of at-risk individuals in the population on day 1.

Enter 25 in cell B35.
The number at risk on day 1 is 25 because we started with a sample size of 25.

3. In cell B36, enter a formula to count the number of deaths on day 1.

In cell B36 enter the formula **=COUNTIF(B10:B34,"D")**.
The number of deaths on day 1 is the number of D's that appear for the 25 individuals.

4. In cell B37, enter a formula to count the number of censored observations on day 1.

In cell B37 enter the formula **=COUNTIF(B10:B34,"C")**.
The number of censored observations on day 1 is the number of C's that appear for the 25 individuals.

5. In cell B38, enter a formula to compute the *conditional* probability of survival, P_c.

In cell B38, enter the formula **=1-(B36/B35)**.
This is the spreadsheet version of Equation 1:

$$P_c = 1 - \frac{d_i}{n_i}$$

The conditional probability of survival is the probability of survival to a particular time period, *given that you survived to the previous time*. This probability is easy to calculate if you know the number of deaths at a specific time and the number of individuals at risk at that same time. The number of deaths divided by the number at risk gives the conditional probability of mortality, so 1 minus that value is the conditional probability of survival.

6. In cell B39, enter a formula to compute the *unconditional* probability of survival, P_u.

In cell B39 we used the formula **=PRODUCT($B$38:B38)**.
The unconditional probability of survival is the probability of surviving to a particular time. It is calculated in Equation 2 as the cumulative product of the conditional probabilities:

$$P_u = \prod_{j=1}^{i} \left(1 - \frac{d_j}{n_j}\right)$$

7. In cell B40, enter a formula to compute the expected P_u for day 1, given the survival parameter in cell B4.

In cell B40 enter the formula **=$B$4^B9**.
The ^ symbol means raises the value in cell C4 (the survival probability) to the number of days under consideration.

8. In cell B41, enter a formula to compute the actual daily survival for each P_c.

In cell B41 enter the formula **=B39^(1/B9)** to obtain the daily survival estimate for day 1.
Remember that the P_c gives the probability of surviving to a specific time period. To convert the P_c to daily survival probabilities, take the appropriate root. For example, take the third root of P_c for day 3, the seventh root of P_c for day 7, and so on, to obtain the daily survival estimate. To obtain roots in spreadsheets, use the exponent as a fraction.

9. In cell C35, compute the number of individuals at risk for day 2.

In cell C35 enter the formula **=B35-(B36+B37)**.
Remember that the number of individuals at risk are those currently alive and not censored.

10. Select your formulae from steps 3–8 and copy them across to column K.

Your spreadsheet should now look something like Figure 7, but (with the exception of Row 40) your numbers will likely be different.

	A	B	C	D	E	F	G	H	I	J	K
35	# at risk	25	20	14	12	10	6	5	5	5	4
36	# deaths	3	2	0	1	1	0	0	0	1	0
37	# censored	2	4	2	1	3	1	0	0	0	0
38	Conditional P_c	0.88	0.9	1	0.917	0.9	1	1	1	0.8	1
39	Unconditional P_u	0.88	0.792	0.792	0.726	0.653	0.653	0.653	0.653	0.523	0.523
40	Expected survival	0.9	0.81	0.729	0.656	0.59	0.531	0.478	0.43	0.387	0.349

Figure 7

11. Save your work.

D. Create graphs.

1. Graph P_c, P_u, and expected P_u as a function of time. Interpret your graph.

Use the line graph option and label your axes fully.

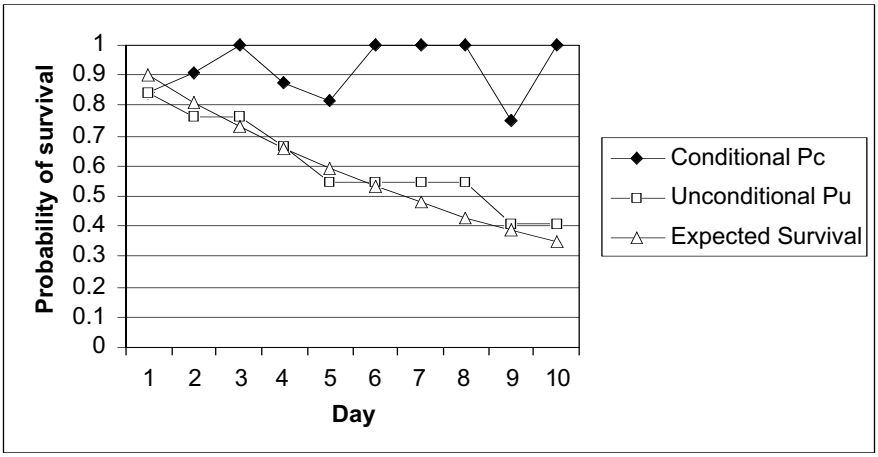

Figure 8

Your graph will look different than the Kaplan-Meier survival curve because the points are connected differently. However, the graphs are interpreted the same way. Note that the expected P_u is a straight line because we set the daily survival probability as a constant over time. Sharp drops in the P_u line indicate more mortality on a given day, and shallow drops in a line indicate fewer deaths occurring during a particular interval. Figure 8 shows few (no) deaths actually occurred from Day 5 to Day 8.

2. Press F9 to generate a new simulation. How do your results appear to change with each new simulation?

Your results should vary from simulation to simulation. This is due to the random number function changing the data set, and it is also due to the fact that our population consists of only 25 individuals (so there is some demographic stochasticity in this model). In order to fully understand how P_c and P_u "behave" over the 10-day period, we need to run several simulations, and track our results. We will do that in the next step.

E. Track 100 simulations.

1. Set up new headings as shown in Figure 9, but extend the trials to 100 (cell M109) and the days to 10 (cell W9).

	M	N	O	P	Q	R
9	**Trial**	**Day 1**	**Day 2**	**Day 3**	**Day 4**	**Day 5**
10	1					
11	2					
12	3					
13	4					
14	5					

Figure 9

2. Record a macro to track P_u for 100 trials, logging your results in cells N10–W109.

Open up the macro function as described in Exercise 2 or your user's manual. Once you have assigned a shortcut and the macro is in Record mode, perform the following steps:

- Select cells B39–K39. Copy.
- Select cell N9. Open Edit | Find.
- Leave the Find What box empty, and search by columns. Select Find Next, then Close. Your cursor should move down to cell N10.

- Open Edit | Paste Special and select the Paste Values option. Click OK.
- Select Tools | Macro | Stop Recording.

Run your macro until 100 trials have been computed.

3. Use the **AVERAGE** function in cells N110–W110 and **STDEV** function in cells N111–W111 to compute the average P_u and standard deviation for the 100 trials.

Your formula for day 1 should be **=AVERAGE(N10:N109)**. This gives the average unconditional probability that an individual will survive past day 1. The standard deviation is computed as **=STDEV(N10:N109)**. You will want to divide this number by 2 for graphing purposes in the next step.

4. Graph the average P_u for each day.

Use the column graph option. Your graph should resemble Figure 10 (without the error bars).

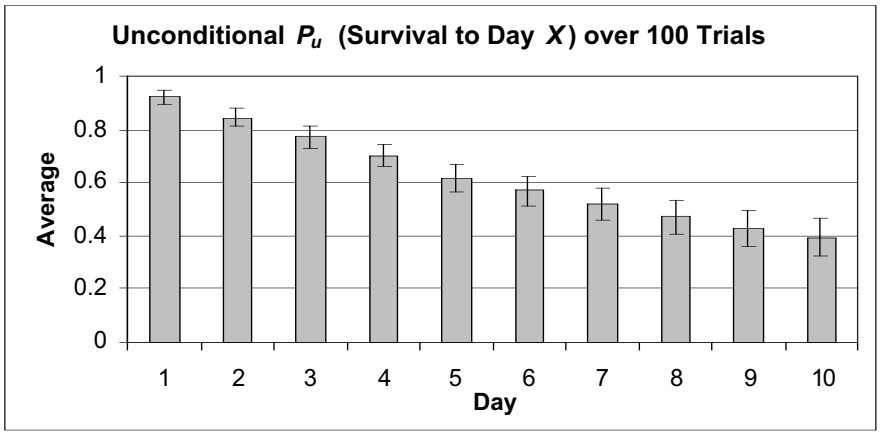

Figure 10

5. Add error bars to your graph. First, divide each standard deviation by 2 in cells N112–W112.

To add error bars, click on the columns in the graph to select them. Then go to Format | Selected Data Series | Y Error Bars. Select the Custom option. Click on the Display Both option. Place your cursor in the box labeled +, then use your mouse to select the standard deviations for your 100 trials divided by 2 (cells N112–W112). Do the same for the box labeled –. Click OK and your graph should be updated.

6. Save your work.

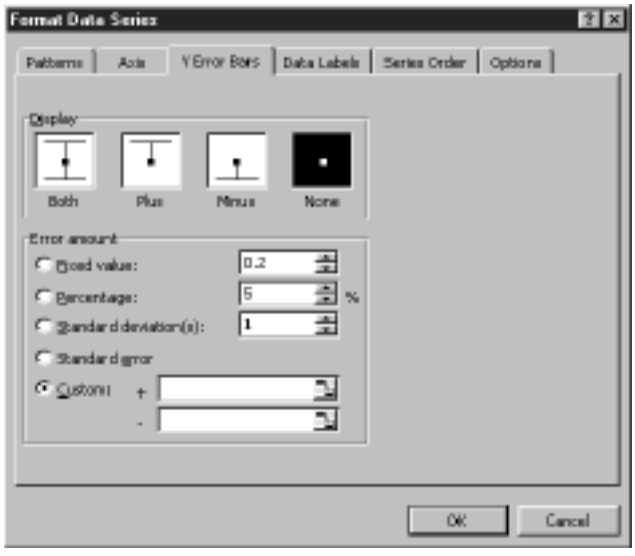

Figure 11

QUESTIONS

1. Interpret the Kaplan-Meier conditional and unconditional probabilities graph (e.g., Figure 8). What do long stretches of slightly sloping or horizontal lines indicate? What do steeply sloping vertical drops indicate?
2. What level of daily survival is needed to ensure that the population will persist for 10 days? Set up your spreadsheet as shown. Enter the expected P_u's for each level of daily survival (given in cells A45–A53). For example, cell B45 should compute P_u for day 1 when the daily survival is 0.1. Under what conditions is a population likely to persist for at least 10 days? Graph your results.

	A	B	C	D	E	F	G	H	I	J	K
44	Daily Survival	Expected P_u									
45	0.1										
46	0.2										
47	0.3										
48	0.4										
49	0.5										
50	0.6										
51	0.7										
52	0.8										
53	0.9										

3. The Kaplan-Meier estimate is often used because "uncooperative" individuals can be taken out of the picture. For example, individuals that fly away from your study plot are censored observations and can be subtracted from your "at risk" population. Compare your model results to a population where censored observations are absent (cells B6–K6 = 0). Erase your macro results (cells N10–W109), then run your macro again under the new conditions. Compare the average P_u and the standard deviations of the trials.
4. Under some conditions, censored observations may occur early in the study, and under some conditions censored observations may occur late in the study. For example, dispersal of individuals out of your study population may occur early or late in the study, depending on the time of year your study is being conducted. Compare how early censorship and late censorship affect P_c and P_u. Set cell B6 = 0.5 to assess early censorship (the remaining cells should be 0). Then set cell K6 = 0.5 (the remaining censorship probabilities should be 0). Describe your results in terms of P_u and its standard deviation.
5. In the spreadsheet model, we simulated the fate of individual's death or survival by linking a random number to a daily survival probability in cell B4. Thus we assumed that for each day, an individual had the same probability of surviving as any other day. What happens to the Kaplan-Meier estimates when survival probabilities vary over the course of the study? Modify your model to include this change and discuss your results in graphical form. For example, establish different daily survival probabilities in cells B4–K4, and adjust the formulae in cells B10–K34 so that the daily survival probability reflects your new entries in cells B4–K4.

*6. (Advanced) How does sample size affect both P_c and P_u? Modify your model and compare results when the sample size is increased from 25 to 50 individuals.

LITERATURE CITED

Kaplan, E. L. and P. Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistics Association* 53: 457–481.