# 4
# Consecutive follow-up intervals

In the last chapter we touched on the difficulty of estimating the probability of failure during a fixed follow-up period when the observation times for some subjects are censored. A second problem with fixed follow-up periods is that it may be difficult to compare the results from different studies; a five-year probability of failure can only be compared with other five-year probabilities of failure, and so on. Finally, by ignoring *when* the failures took place, all information about possible changes in the probability of failure during follow-up is lost.

The way round these difficulties is to break down the total follow-up period into a number of shorter consecutive intervals of time. We shall refer to these intervals of time as *bands*. The experience of the cohort during each of these bands can then be used to build up the experience over any desired period of time. This is known as the *life table* or *actuarial* method. Instead of a single binary probability model there is now a sequence of binary models, one for each band. This sequence can be represented by a conditional probability tree.

## 4.1 A sequence of binary models

Consider an example in which a three-year follow-up interval has been divided into three one-year bands. The experience of a subject during the three years may now be described by a sequence of binary probability models, one for each year, as shown by the probability tree in Fig.4.1. The four possible outcomes for this subject, corresponding to the tips of the tree, are

1. failure during the first year;
2. failure during the second year;
3. failure during the third year;
4. survival for the full three-year period.

The parameter of the first binary model in the sequence is $\pi^1$, the probability of failure during the first year; the parameter of the second binary model is $\pi^2$, the probability of failure during the second year, given the subject has not failed before the start of this year, and so on. These are

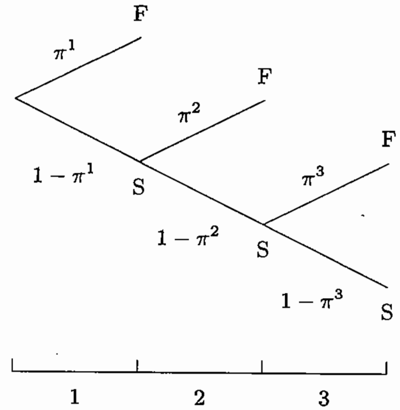


**Fig. 4.1.** A sequence of binary probability models.

all conditional probabilities — conditional on not having failed before the start of the year in question. The reason the probabilities are written with superscripts is that we have adopted the convention that a superscript is used to index *time*, and a subscript is used to index subjects or groups of subjects. It is important to distinguish these two situations, and using subscripts for both can be confusing.*

Suppose, for illustration, that the probability of failure is 0.3 in the first year; 0.2 in the second year, given the subject survives the first year without failure; and 0.1 in the third year, given the subject survives the first two years without failure. These illustrative values for the three conditional probabilities are shown on the conditional probability tree in Fig.4.2.

In this tree, the four final outcomes listed above correspond to the tips of the tree, and their probabilities can be calculated by multiplying conditional probabilities along the branches of the tree in the usual way. For example, the probability of the second outcome is made up from the probability that the subject survives the first year (0.7), multiplied by the probability that the subject fails during the second year (0.2). Using this rule, the four possible outcomes for any subject occur with probabilities:

$$
\begin{gathered}
0.3\\
0.7 \times 0.2\\
0.7 \times 0.8 \times 0.1\\
0.7 \times 0.8 \times 0.9
\end{gathered}
$$

*Note that $\pi^2$ does not refer to $\pi \times \pi$. To avoid confusion we shall always use brackets when taking powers; for example, the square of $\pi$ will be written $(\pi)^2$.



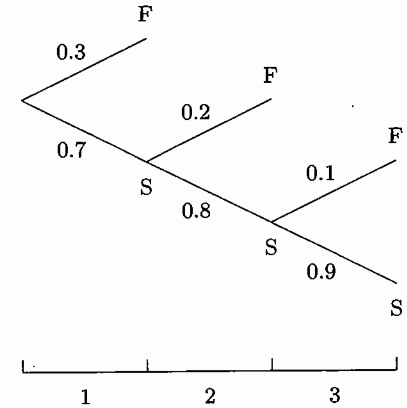


**Fig. 4.2.** Illustrative values for the conditional probabilities.

These probabilities work out to be 0.3, 0.14, 0.056, and 0.504, and these add to 1, as they should, since there are no other possible outcomes. The probability of failing at *some* stage is

$$0.3 + 0.14 + 0.056 = 0.496.$$

More conveniently this probability can be found by subtracting from 1 the probability of surviving the three years without failing, giving

$$1 - 0.504 = 0.496.$$

The probabilities of surviving one, two, and three years without failing are called the *cumulative survival probabilities* for the cohort. They are calculated by multiplying the conditional probabilities of surviving each year, and in this case are:

$$
\begin{gathered}
0.7\\
0.7 \times 0.8\\
0.7 \times 0.8 \times 0.9.
\end{gathered}
$$

which work out to be 0.7, 0.56, and 0.504.

**Exercise 4.1.** In a three-year follow-up study the conditional probabilities of failure during the first, second, and third years are 0.05, 0.09, and 0.12 respectively. Draw a probability tree for the possible outcomes for a new subject, and label the branches of the tree with the appropriate conditional probabilities. Calculate the probability of each of the outcomes, and the probabilities of surviving

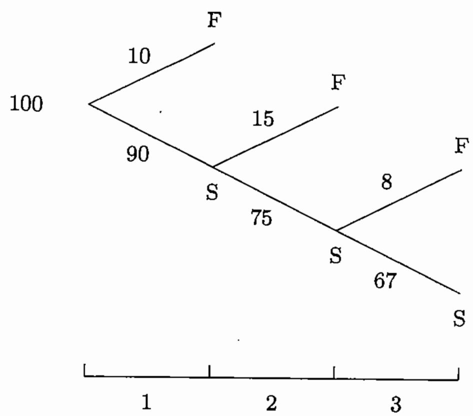


**Fig. 4.3.** Survival of 100 subjects through three time bands.

one, two, and three years without failing. Calculate also the probability of failing at some time during the three-year follow-up.

## 4.2 Estimating the conditional probabilities of failure

Suppose that 100 subjects join the cohort at the start of the three-year interval and that 10 fail during the first year, 15 during the second, and 8 during the third, leaving 67 who survive until the end of three years (see Fig.4.3). Assuming the same conditional probabilities of failure for each of the 100 subjects, these data can be used to estimate their most likely values.

Intuitively it seems sensible to use the experience of those subjects who are observed in each year to estimate the conditional probability of failure during that year. The most likely values of the three conditional probabilities would then be

$$\frac{10}{100}, \frac{15}{90}, \frac{8}{75},$$

but is this a legitimate thing to do? It corresponds to regarding the three-year follow-up study as equivalent to three separate and independent one-year follow-up studies in which the subjects come from the survivors of the previous year. In fact this is a legitimate thing to do because the likelihood for $\pi^1$, $\pi^2$, and $\pi^3$ is the same whether the data are regarded as coming from one three-year study or from three one-year studies. This may be shown algebraically as follows.



The probabilities of the four possible outcomes in the three-year study are

$$\begin{gathered}\pi^1\\(1-\pi^1)\pi^2\\(1-\pi^1)(1-\pi^2)\pi^3\\(1-\pi^1)(1-\pi^2)(1-\pi^3)\end{gathered}$$

A subject who fails during the first year therefore contributes

$$\log(\pi^1)$$

to the log likelihood. A subject who fails during the second year contributes

$$\log(1-\pi^1)+\log(\pi^2),$$

a subject who fails during the third year contributes

$$\log(1-\pi^1)+\log(1-\pi^2)+\log(\pi^3),$$

and a subject who survives all three years contributes

$$\log(1-\pi^1)+\log(1-\pi^2)+\log(1-\pi^3).$$

Multiplying these by the numbers of subjects with each outcome, that is 10, 15, 8, and 67 respectively, and adding, gives a total log likelihood of

$$\begin{gathered}10\log(\pi^1)+90\log(1-\pi^1)\\+15\log(\pi^2)+75\log(1-\pi^2)\\+8\log(\pi^3)+67\log(1-\pi^3).\end{gathered}$$

This is the same as the log likelihood obtained by regarding the data as from three separate and independent one-year studies; the first based on 10 failures and 90 survivors, the second on 15 failures and 75 survivors, and the third on 8 failures and 67 survivors.

**Exercise 4.2.** If we were to adopt the more restrictive model that $\pi^1, \pi^2, \pi^3$ are all equal with common value $\pi$, what would be the most likely value of $\pi$?

This exercise makes it clear that, in the analysis of such studies, the basic atom of data is not the subject, but the observation of one subject through one time band.

## 4.3 A cohort life table

In cohorts where subjects are examined at yearly intervals, the data are often presented in the form of numbers of failures and censorings occurring each year. An example is given in Table 4.1, which refers to survival of a

**Table 4.1.** Survival by stage at diagnosis

| | Stage I | | | Stage II | | |
|---|---|---|---|---|---|---|
| Year | N | D | L | N | D | L |
| 1 | 110 | 5 | 5 | 234 | 24 | 3 |
| 2 | 100 | 7 | 7 | 207 | 27 | 11 |
| 3 | 86 | 7 | 7 | 169 | 31 | 9 |
| 4 | 72 | 3 | 8 | 129 | 17 | 7 |
| 5 | 61 | 0 | 7 | 105 | 7 | 13 |
| 6 | 54 | 2 | 10 | 85 | 6 | 6 |
| 7 | 42 | 3 | 6 | 73 | 5 | 6 |
| 8 | 33 | 0 | 5 | 62 | 3 | 10 |
| 9 | 28 | 0 | 4 | 49 | 2 | 13 |
| 10 | 24 | 1 | 8 | 34 | 4 | 6 |

group of women with cancer of the cervix diagnosed at either stage I or stage II. The women are examined annually, and censoring occurs if they cease attending the clinic; $N$ is the number alive and still under observation at the start of each time band, $D$ is the number who die during each band, and $L$ is the number censored during each band.

The estimation of survival experience of the stage I women over the first four years is shown in Fig.4.4. Of the 110 subjects who started the first year, 5 die and 5 are censored. The effective size of the cohort in the first year is taken to be 107.5 and the probability of a subject dying during the first year, given the subject was alive at the start of the year, is estimated to be $5/107.5 = 0.0465$. The conditional probability of surviving the year is estimated to be

$$1 - 0.0465 = 0.9535.$$

The calculations of failure and survival probabilities are shown in Fig.4.4. The cumulative survival probabilities are found by multiplying the conditional survival probabilities for each year. For example, the cumulative probability of surviving 3 years is

$$0.9535 \times 0.9275 \times 0.9152 = 0.8093.$$

**Exercise 4.3.** Using Table 4.1, draw a tree showing the survival experience for stage II women over the first four years, and calculate the conditional survival probabilities for each of these years.

A table of cumulative survival probabilities by year is called a *life table*, and a plot of the cumulative survival probabilities against years survived is called a *survival curve.* The survival curves for both stage I and stage II



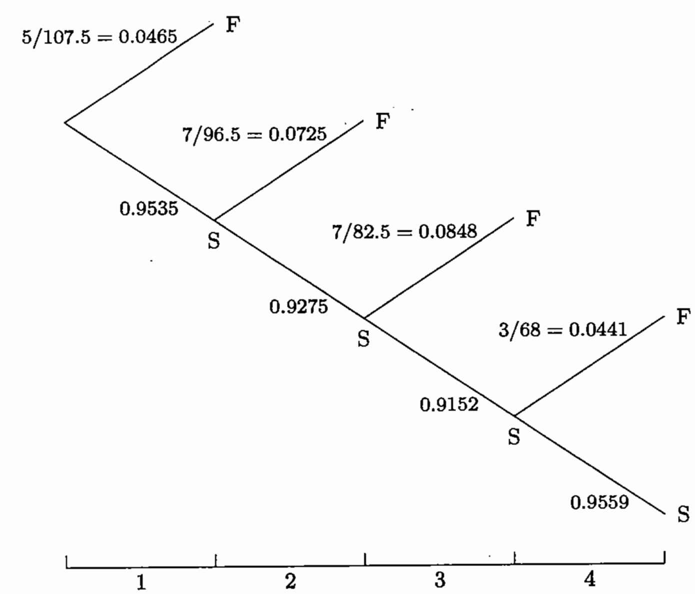


**Fig. 4.4.** Estimated conditional probabilities for stage 1 women.

women are shown in Fig.4.5. It is conventional to start survival curves at a probability of one for surviving at least zero years. These plots are useful for studying whether the probability of failure is changing with follow-up time, and for calculating survival probabilities for different periods of time.

**Exercise 4.4.** Use Fig.4.5 to read off the five-year survival probabilities in each of the two groups.

## 4.4 The use of exact times of failure and censoring

In the calculations described above, the conditional probability of failure during each time band has been estimated by assuming, as in Chapter 3, that half the losses during the band occurred at the start and half at the end. If the individual times at which failure (or censoring) occur are known then it is possible to avoid this assumptions by choosing the bands so short that each failure occupies a band by itself. Such a choice of bands is shown in Fig.4.6 for the early follow-up experience of 50 subjects. The horizontal line represents follow-up time, failures are marked as •, and losses as ×. The bands are shown by vertical bars. Only the first few events are shown.

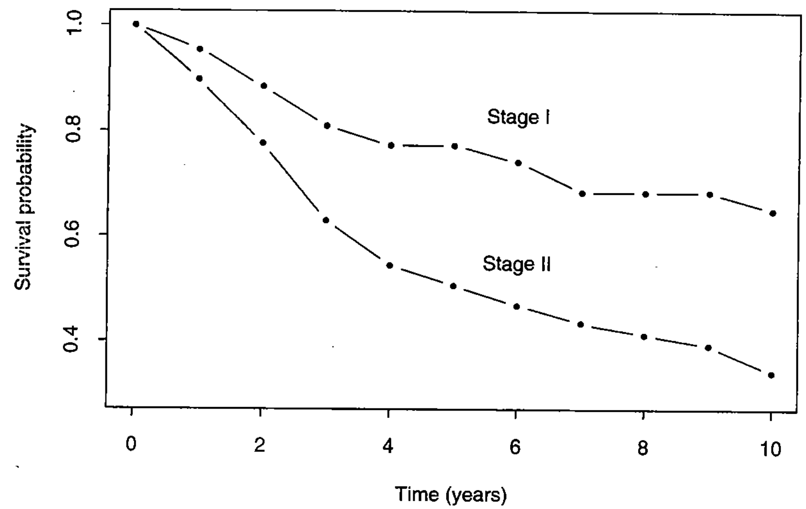


**Fig. 4.5.** Survival curves for Stage I and Stage II women.

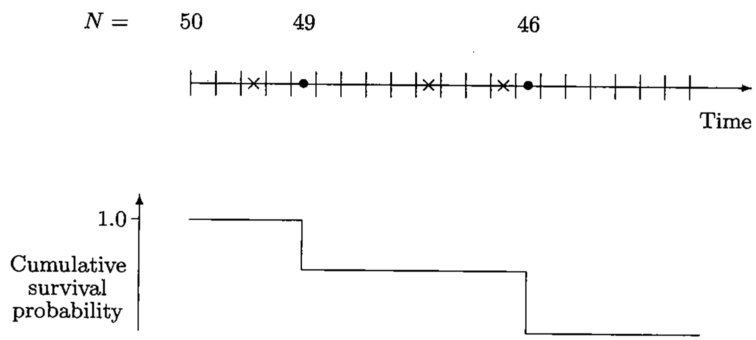


**Fig. 4.6.** Early follow-up experience of 50 subjects.

For bands in which there are no failures the estimated survival probability is 1. For bands which contain a failure the estimated survival probability is $1 - 1/N$ where $N$ is the number at risk just before the failure. Thus for the band which contains the first failure $N = 49$ and the estimated survival probability is $1 - 1/49 = 48/49$. The estimate of the cumulative survival probability up to the end of this band is

$$1 \times 1 \times \cdots \times 48/49 = 0.9796.$$



For the band which contains the second failure $N = 46$, so the estimated survival probability for this band is $1-1/46 = 45/46$. The cumulative probability of survival up to the end of the fourth band is therefore estimated at

$$1 \times \cdots \times 48/49 \times 1 \times \cdots \times 45/46 = 0.9583.$$

These calculations continue until there are no more bands which contain failures.

The bands containing each failure can be made so short that they refer to the actual time of failure. When this is done the cumulative survival probability over time takes the value 1 until the first failure, when it drops to 0.9796; then it stays at 0.9796 until the second failure when it drops to 0.9583, and so on. The plot of cumulative survival probability versus time survived takes the stepped shape shown in Fig.4.6, where the steps occur at the failure times.

This method of estimating the cumulative survival probabilities is called the *Kaplan–Meier* method, after the authors of the paper which showed that this procedure yields the most likely value of the survival curve. It is widely used in clinical follow-up studies for which individual failure times are known. If the failure times are measured exactly the failures will all occur at separate times, but if they are measured to the nearest month (for example) then there may be several failures at the same time. In this case the probability of failure is estimated by dividing the number of failures at that failure time by the total number of subjects at risk just before the failure time. If losses also occur at this time then, by convention, they are included in the number at risk.

## 4.5 An example of the Kaplan–Meier method

Table 4.2 shows the time from diagnosis to death from melanoma, or loss to follow-up, for 50 subjects. Times are in complete months so that subjects dying during the first month are recorded as surviving one month, and so on. For two subjects diagnosis took place at death, so the time was recorded as zero.

Note that probabilities of failure are estimated only for times at which failures occurred. The first of these is at time zero; the number at risk is 50, with 2 failures, so the probability of failure at this time point is $2/50 = 0.04$, and the survival probability is $1 - 0.04 = 0.96$. The next time at which a failure occurs is one month; the number at risk is 48, with one failure, so the probability of failure at this time point is $1/48 = 0.0208$ and the probability of surviving is $1 - 0.0208 = 0.9792$. The next time at which a failure occurs is at 2 months, when there are two failures. The probability of failure is $2/47 = 0.0426$, and the survival probability is $1 - 0.0426 = 0.9574$. At three months there is one failure and one loss to follow-up. In fact this loss was a death from a cause other than melanoma, but when estimating survival

**Table 4.2.** Cumulative survival probabilities from the Kaplan–Meier method. Non-melanoma deaths (*) are counted as losses.

| Month | $N$ | $D$ | $L$ | Conditional probability of death | Conditional probability of survival | Cumulative prob. of survival |
|---|---|---|---|---|---|---|
| 0 | 50 | 2 | | 0.0400 | 0.9600 | 0.9600 |
| 1 | 48 | 1 | | 0.0208 | 0.9792 | 0.9400 |
| 2 | 47 | 2 | | 0.0426 | 0.9574 | 0.9000 |
| 3 | 45 | 1 | 1* | 0.0222 | 0.9778 | 0.8800 |
| 8 | 43 | 1 | | 0.0233 | 0.9767 | 0.8595 |
| 10 | 42 | 1 | | 0.0238 | 0.9762 | 0.8391 |
| 12 | 41 | 1 | 1* | 0.0244 | 0.9756 | 0.8186 |
| 13 | 39 | 1 | | 0.0256 | 0.9744 | 0.7976 |
| 15 | 38 | 1 | | 0.0263 | 0.9737 | 0.7766 |
| 18 | 37 | | 1* | | | |
| 19 | 36 | 1 | | 0.0278 | 0.9722 | 0.7551 |
| 21 | 35 | | 1 | | | |
| 27 | 34 | | 2 | | | |
| 30 | 32 | | 1 | | | |
| 33 | 31 | 1 | 1 | 0.0323 | 0.9677 | 0.7307 |
| 34 | 29 | 1 | | 0.0345 | 0.9655 | 0.7055 |
| 38 | 28 | | 1 | | | |
| 40 | 27 | | 1 | | | |
| 41 | 26 | 1 | | 0.0385 | 0.9615 | 0.6784 |
| 43 | 25 | | 1 | | | |
| 44 | 24 | | 1 | | | |
| 46 | 23 | | 1 | | | |
| 54 | 22 | | 1 | | | |
| 55 | 21 | 1 | | 0.0476 | 0.9524 | 0.6461 |
| 56 | 20 | 1 | | 0.0500 | 0.9500 | 0.6138 |
| 57 | 19 | | 2 | | | |
| 60 | 17 | | 1* | | | |

probabilities from melanoma alone it is counted as a loss to follow-up. (We return to a fuller discussion of this point in Chapter 7.) The number at risk was 45, with one failure, so the probability of failure is $1/45 = 0.022$ and the probability of survival is $1 - 0.022 = 0.9778$, and so on. A plot of the cumulative survival probability against time is shown in Fig.4.7.



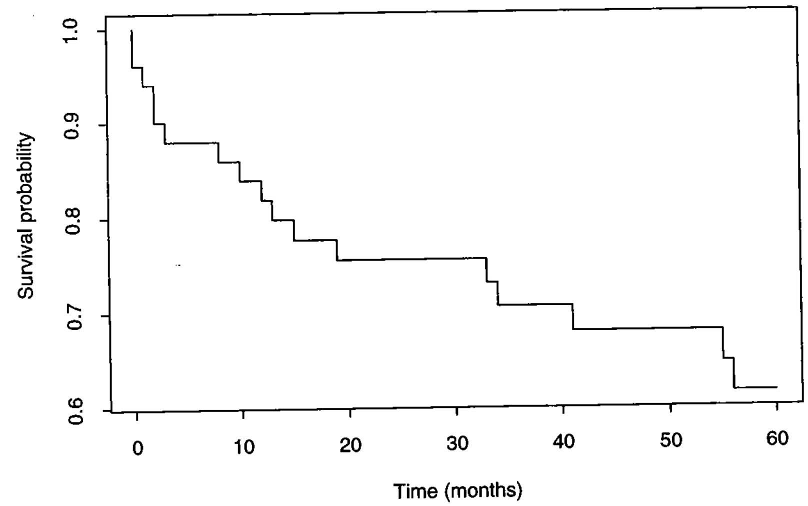


**Fig. 4.7.** Cumulative survival probability by the Kaplan–Meier method.

**Solutions to the exercises**

**4.1** See Fig.4.8. The probabilities of failure during the first, second and third years are

$$0.05 \quad 0.95 \times 0.09 = 0.0855 \quad 0.95 \times 0.91 \times 0.12 = 0.1037.$$

The probability of surviving three years is

$$0.95 \times 0.91 \times 0.88 = 0.7608.$$

The survival probabilities for the three years are

$$0.95 \quad 0.8645 \quad 0.7608.$$

The probability of failure at some time during the three years is

$$0.05 + 0.0855 + 0.1037 = 0.2392$$

or

$$1 - 0.7608 = 0.2392.$$

**4.2** The overall log likelihood is

$$33\log(\pi) + 232\log(1-\pi),$$

which is equivalent to observing 33 failures in 265 subjects. The most likely value of $\pi$ is, therefore $33/265 = 0.125$.

**4.3** See Fig.4.9.

**4.4** The five year survival probabilities from Fig.4.5 are 0.78 (Stage I) and 0.51 (Stage II).

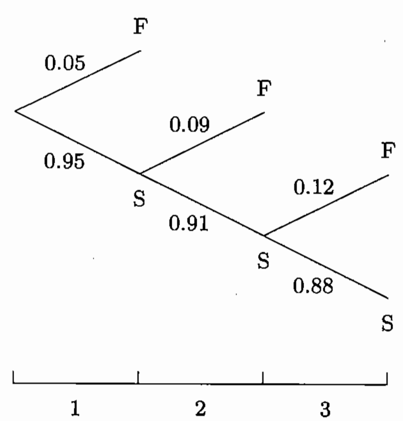


**Fig. 4.8.** Solution to exercise 4.1.



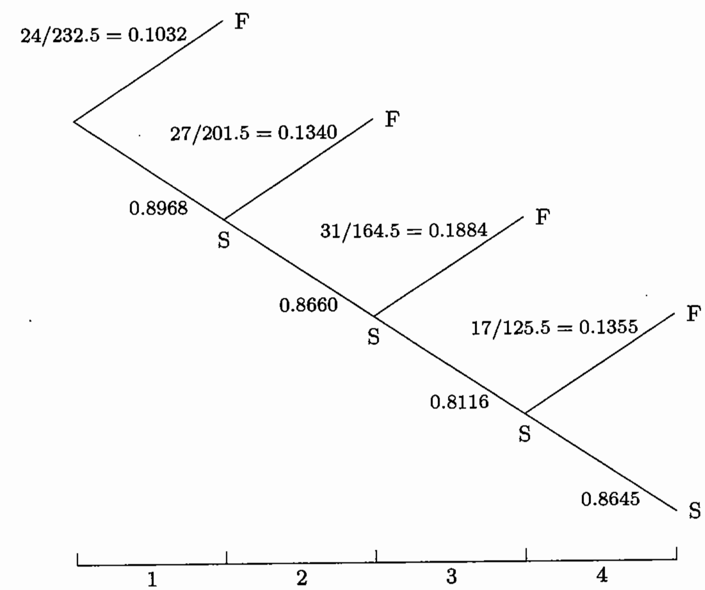


**Fig. 4.9.** Estimated conditional probabilities for stage II women.