The Weibull in R is actually parameterized a fair bit differently from the book. In R, the density for $x > 0$ is

$$f(x) = \frac{a}{b} \left(\frac{x}{b}\right)^{a-1} e^{-(x/b)^a}$$

This means that $a = \alpha$ in the book’s parameterization and $\frac{1}{b^a} = \lambda$ in the book’s parameterization. Thus to use $\alpha = 0.5$, $\lambda = 1.2$, this corresponds to $a = \text{shape} = 0.5$, $b = \text{scale} = (1/\lambda)^{1/\alpha} = (1/1.2)^{1/0.5}$. 
Adding legends to plots

For the homework, it would be good to add a legend to make your plots more readable.
Log-likelihoods

The problem is that we are trying to take logarithms of things that are already undefined. Instead, we need to manipulate the probabilities with logarithms first before trying to evaluate these large exponents and binomial coefficients.

The log-likelihood is

$$\log L(p) = \log \left[ \binom{10000}{9000} p^{9000}(1 - p)^{1000} \right]$$

$$= \log \left[ \binom{10000}{9000} \right] + 9000 \log(p) + 1000 \log(1 - p)$$

An important point to realize is that $\log \left[ \binom{10000}{9000} \right]$ doesn’t depend on $p$, so maximizing $\log L(p)$ is equivalent to maximizing $9000 \log(p) + 1000 \log(1 - p)$. This way, we don’t have to evaluate this very large binomial coefficient.
Log-Likelihoods

When the likelihood is multiplied by a constant that doesn’t depend on the parameter, we sometimes ignore the constant. Thus, we might write

$$L(p) \propto p^{9000}(1 - p)^{1000}$$

or even just drop the constant altogether. So sometimes, you’ll see

$$L(p) = p^{9000}(1 - p)^{1000}$$

even though this isn’t the probability of the data. The constant changes the scale on the $y$-axis, but doesn’t change the shape of the curve or the value on the $p$ (horizontal) axis where the maximum occurs.

Now we can plot and evaluate $9000 \log(p) + 1000 \log(1 - p)$ in R even though we can’t evaluate $p^{9000}(1 - p)^{1000}$ directly (even though they are mathematically equivalent).
The log-likelihood function

\[ 9000 \cdot \log(p) + 1000 \cdot \log(1 - p) \]
Maximizing the likelihood function

In some cases, we can maximize the likelihood function analytically, usually using calculus techniques. For the binomial case, we can take the derivative of the likelihood or log-likelihood function and set it equal to 0 to find the maximum.

\[
\frac{d}{dp} \log L(p) = \frac{d}{dp} \left\{ \log \binom{n}{k} + k \log p + (n - k) \log(1 - p) \right\} = 0
\]

\[
\Rightarrow 0 + \frac{k}{p} - \frac{n - k}{1 - p} = 0
\]

\[
\Rightarrow (1 - p)k = p(n - k)
\]

\[
\Rightarrow k - kp - np + kp = 0
\]

\[
\Rightarrow k = np
\]

\[
\Rightarrow p = \frac{k}{n}
\]
Maximizing the likelihood function

Since \( p = \frac{k}{n} \), the proportion of successes, maximizes \( \log L(p) \), and therefore the likelihood as well, the maximum likelihood estimator for \( p \) is \( \hat{p} = \frac{k}{n} \). We say \textit{estimator} for the general function that works for any data, and \textit{estimate} for a particular value like \( \hat{p} = 0.9 \).
Suppose you have 3 lightbulbs that last 700, 500, and 1100 hours. Assuming that their lifetimes are exponentially distributed with rate $\lambda$, what is the maximum likelihood estimate of $\lambda$?
Maximum likelihood estimation for two-parameter distributions

To use maximum likelihood for two-parameter families of distributions, such as the normal ($\mu$ and $\sigma^2$), the beta distribution, and the gamma distribution, you can write down the log-likelihood and then try to find the maximum for this surface. Graphically, the log-likelihood is plotted in a third dimension where the first two dimensions are the different parameter values.

In some cases, such as for the normal distribution, this can be done analytically by setting both partial derivatives to 0 and solving the system of equations. In other cases, numerical methods must be used. Another approach is to assume one of the parameters, reducing the problem to one parameter, and solving for the other parameter analytically. Then you can search over values of the first parameter.
Maximum-likelihood estimation for two-parameter distributions

It turns out that for the Weibull, the analytic approach doesn’t work, and so numerical methods are generally used. The simplest method is to use a grid. I.e., try all values of $\alpha$ in some interval and all values of $\lambda$ in some interval using some increments. If you think $0.1 < \alpha < 10$ for example, you could try all values in increments of .01 and all values of $\lambda$ from say, $1 < \lambda < 10$ in increments of .01. This would require evaluating the log-likelihood function 10 million times. Otherwise, you might be able to use faster numerical methods such as Newton-Raphson. Ordinarily, you’ll be able to let the software do this for you.
For survival analysis, we need likelihood functions that incorporate censoring. A general framework is to have separate densities and probabilities for cases of complete observations, censored observations, and truncated observations. Assuming that all observations are independent, we can write the likelihood as the product of densities and probabilities from all of these cases.
Likelihoods with censoring and truncation

In the most general set up, you can allow different types of functions:

\[ f(x) \] exact lifetimes/death times
\[ S(C_r) \] right-censored observations
\[ 1 - S(C_l) \] left-censored observations
\[ [S(L) - S(R)] \] interval-censored observations
\[ \frac{f(x)}{S(Y_L)} \] left-truncated observations
\[ \frac{f(x)}{1 - S(Y_R)} \] right-truncated observations
\[ \frac{f(x)}{S(Y_L) - S(Y_R)} \] interval-truncated observations
For censored (but not truncated) data, the overall likelihood is

\[ \prod_{i \in D} f(x_i) \prod_{i \in R} S(C_{ri}) \prod_{i \in L} S(C_{li}) \prod_{i \in I} [S(L_i) - S(R_i)] \]

where \( D \) is the set of death times, \( R \) is the set of right-censored observations, \( L \) is the set of left-censored observations, and \( I \) is the set of right-censored observations.
If you have truncated data, then replace each term with the analogous conditional density, for example replace $f(x)$ with $\frac{f(x)}{1-S(Y_R)}$ for right-truncated data (when you condition on observing only deaths).
The likelihood with right-censoring

When we’ve observed a right-censored time, \( C_r \), we’ve observed \( (T = C_r, \delta = 0) \), so the contribution to the likelihood for this observation is

\[
Pr[T = C_r, \delta = 0] = Pr[T = C_r|\delta = 0]Pr(\delta = 0) = 1 \cdot Pr(\delta = 0) = Pr(X > C_r) = S(C_r)
\]

When we’ve observed a (non-censored) death-time, the contribution to the likelihood is

\[
Pr[T, \delta = 1] = Pr[T = t|\delta = 1]P(\delta = 1) = \frac{Pr(T = t)}{P(\delta = 1)} \cdot P(\delta = 1) = Pr(T = t) = f(t)
\]

We can therefore write

\[
Pr(t, \delta) = [f(t)]^{\delta}[S(t)]^{1-\delta}
\]
The likelihood with right-censoring

The previous slide gave the likelihood of a single observation. The likelihood of a sample is the product over all observations (assuming that the observations are independent). Therefore

\[ L = \prod_{i=1}^{n} Pr(t_i, \delta_i) = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} = \prod_{i: \delta_i=1} f(t_i) \prod_{i: \delta_i=0} S(t_i) \]

which is of the form of the general likelihood function from a few slides ago. There are only two products instead of four because we only have one type of censoring.
Notation with the hazard function

Because \( h(t) = \frac{f(t)}{S(t)} \), and \( S(t) = e^{-H(t)} \), you can also write

\[
L = \prod_{i=1}^{n} [h(t_i)]^{\delta_i} e^{-H(t)}
\]

which expresses the likelihood in terms of the hazard and cumulative hazard functions.
Example with exponential and right-censoring

If we have exponential times $t_1, \ldots, t_n$ where $t_i$ has been censored if $\delta_i = 1$, then

$$L = \prod_{i=1}^{n} (\lambda e^{-\lambda t_i})^{\delta_i} \exp[-\lambda t_i (1 - \delta_i)]$$

$$= \lambda^r \exp \left[-\lambda \sum_{i=1}^{n} t_i \right]$$

where $r = \sum_{i=1}^{n} \delta_i$, the number of non-censored death times. This is very similar to the usual likelihood for the exponential except that instead of $\lambda^n$, we have $\lambda^r$ where $r \leq n$. 
The log-likelihood for the exponential example is

$$\log L = r \log \lambda - \lambda \sum_{i=1}^{n} t_i$$

the derivative is

$$\frac{r}{\lambda} - \sum_{i=1}^{n} t_i$$

Setting this equal to 0, we obtain

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^{n} t_i} = \frac{r}{nt}$$
Suppose survival times are assumed to be exponentially distributed and we have the following times (in months):

1.5, 2.4, 10.5, 12.5+, 15.1, 20.2+

Find the maximum likelihood estimate of $\lambda$. 
Example with exponential data and right-censoring

The main summaries needed for the data are the sum of the times (whether or not they are censored), and the number of non-censored observations. There are 6 observations and three are not censored, so 

\[ r = \sum_{i=1}^{n} \delta_i = 4. \]

The sum of the times is

\[ 1.5 + 2.4 + 10.5 + 12.5 + 15.1 + 20.2 = 60.2 \]

Therefore the maximum likelihood estimate (MLE) is

\[ \hat{\lambda} = \frac{4}{60.2} = 0.066 \]

This corresponds to a mean survival time of 15.02 months.
Example with exponential data and INCORRECTLY ignoring right-censoring

If we had (incorrectly) ignored censoring and treated those times as noncensored, we would have obtained

$$\hat{\lambda} = \frac{6}{60.2} = 0.0997$$

with a mean survival time of 10.03 months. If we had dropped the observations that were censored, we would have obtained

$$\hat{\lambda} = \frac{4}{29.5} = 0.136 \Rightarrow E(T) = 7.38 \text{ months}$$
This example is exercise 3.5 in the book (page 89):

Suppose the time to death has a log-logistic distribution with parameters \( \lambda \) and \( \alpha \). Based on the following left-censored sample, construct the likelihood function.

0.5, 1, 0.75, 0.25- 1.25-

where \( - \) denotes a left-censored observation.
log-logistic example

Here we only have one type of censoring: left censoring, so in terms of our general framework for setting up the likelihood we have

\[
L = \prod_{i \in D} f(x_i) \prod_{i \in L} (1 - S(C_i))
\]

There are three death times observed and two left-censored observations, so the first product has three terms and the second product has two terms. We can use the table on page 38 to get the density and survival functions.
The log-logistic density for $x > 0$ is

$$f(x) = \frac{\alpha x^{\alpha-1} \lambda}{[1 + \lambda x^\alpha]^2}$$

The survival function is

$$S(x) = \frac{1}{\lambda x^\alpha}$$

which means that

$$1 - S(x) = 1 - \frac{1}{1 + \lambda x^\alpha} = \frac{\lambda x^\alpha}{1 + \lambda x^\alpha}$$
The log-logistic function: density when $\lambda = 1$
The likelihood is therefore

\[
\prod_{i=1}^{3} \frac{\alpha x_i^{\alpha-1} \lambda}{[1 + \lambda x_i^{\alpha}]^2} \prod_{i=4}^{5} \frac{\lambda x_i^{\alpha}}{1 + \lambda x_i^{\alpha}}
\]
Using the data, we can write this as

\[
L = \frac{\alpha(0.5)^{\alpha-1}\lambda}{[1 + \lambda(0.5)^\alpha]^2} \cdot \frac{\alpha(1)^{\alpha-1}\lambda}{[1 + \lambda(1)^\alpha]^2} \cdot \frac{\alpha(0.75)^{\alpha-1}\lambda}{[1 + \lambda(0.75)^\alpha]^2} \cdot \frac{\lambda(0.25)^\alpha}{1 + \lambda(0.25)^\alpha} \cdot \frac{\lambda(1.25)^\alpha}{1 + \lambda(1.25)^\alpha}
\]
log-logistic example

We can simplify the likelihood as

\[
L = \prod_{i=1}^{3} \frac{\alpha x_i^{\alpha-1} \lambda}{[1 + \lambda x_i^{\alpha}]^2} \prod_{i=4}^{5} \frac{\lambda x_i^{\alpha}}{1 + \lambda x_i^{\alpha}}
\]

\[
= \frac{\alpha^3 \lambda^5 x_4 x_5 \left( \prod_{i=1}^{5} x_i \right)^{\alpha-1}}{\prod_{i=1}^{5} (1 + \lambda x_i^{\alpha}) \prod_{i=1}^{3} 1 + \lambda x_i^{\alpha}}
\]

\[
\log L = 3 \log \alpha + 5 \log \lambda + \sum_{i \in L} \log(x_i) + (\alpha - 1) \sum_{i=1}^{n} \log x_i
\]

\[- \sum_{i=1}^{n} \log(1 + \lambda x_i^{\alpha}) - \sum_{i \in D} \log(1 + \lambda x_i^{\alpha})
\]
We’ll look at evaluating the log-logistic likelihood in this example in R. First, we’ll look at how to write your own functions in R.

An example of a function would be to add 1 to a variable.

```r
> f <- function(x) {
+   return(x+1)
+ }
> f(3)
[1] 4
> f(c(2,3))
[1] 3 4
```

This function takes x as an input returns the input plus 1. Note that f() can also take a vector or a matrix as input, in which case it adds 1 to every element.
Functions can also have more than one argument. For example

```r
> function poisbinDiff <- function(x,n,p) {
+   value1 <- ppois(x,lambda=n*p)
+   value2 <- pbinom(x,n,p)
+   return(abs(value1-value2)/value2)
+ }
```

What does this function do?
The previous functions considers an experiment with $X$ successes and computes $P(X \leq x)$ for two models: binomial and Poisson. In many cases, the Poisson is a good approximation to the binomial with $\lambda = np$, so the function computes the difference in probabilities for the two models, and divides by the probability under the binomial. This returns the relative error using the Poisson to approximate the binomial.

The point of using functions is to reduce the tedium of writing several lines instead of writing one line to do several steps. This is particularly useful if you want to call a sequence of steps many times with different values.
Writing a likelihood function in R

To get R to numerically compute a likelihood value for you, you can write a similar user-defined function. Recall that the likelihood for exponential data (without censoring) is

$$L = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

You can write the likelihood function as

```r
> L <- function(x,lambda) {
+     value <- lambda^n * exp(-lambda * x)
+     return(value)
+ }
```

where $x = \sum_{i=1}^{n} x_i$. 
Writing the log-logistic likelihood function in R

The log-logistic function is a little more complicated and uses two parameters, but the idea is the same. We’ll write the function in R in a way that depends on the data and doesn’t generalize very well. (You’d have to write a new function for new data).

```r
> Like <- function(alpha,lambda) {
  value <- 1
  value <- value*alpha^3*lambda^5*(0.5*.75)^(alpha-1)*
  + (1.25*.25)^alpha #the plus here just indicates a line break
  value <- value/(1+lambda*(.5)^alpha)^2
  value <- value/(1+lambda)^2
  value <- value/(1+lambda*(.75)^alpha)^2
  value <- value/(1+lambda*(1.25)^alpha)
  value <- value/(1+lambda*(.25)^alpha)
  return(value)
}
```
The log-logistic likelihood for example data
Finding the maximum likelihood estimate by grid search

Although computing all values over a grid might not be the most efficient way to find the MLE, it is a brute force solution that can work for difficult problems. In this case, you can evaluate the Like() function for different parameters of $\alpha$ and $\lambda$. I tried for values between 0 and 10 for both $\alpha$ and $\lambda$ in increments of 0.1. This requires 100 values for $\alpha$ and, independently, 100 values for $\lambda$, meaning that the likelihood is computed 10000 times.

Doing this for all of these values requires some sort of loop, but then you can find the best parameter values up to the level of precision tried. For these values, I obtain $\left( \hat{\alpha}, \hat{\lambda} \right) = (2.6, 5.0)$, which gives a likelihood of 0.03625.
Find the maximum likelihood estimate by grid search

Although the grid search is inefficient, it gives you a nice plot which gives you some idea of how peaked the likelihood function is and how it depends on the parameters. In this case, the likelihood changes more rapidly as $\lambda$ changes than as $\alpha$ changes. This can be confirmed with the likelihood function.

> Like(2.6,5)
[1] 0.0362532
> Like(2.82,5)
[1] 0.03553457
> Like(2.6,5.5)
[1] 0.03604236

Increasing $\alpha$ by 10% from the (approximate) MLE lowers the likelihood more than increasing $\lambda$ by 10%.
I used a slow, brute force method to generate the likelihood surface with a resolution of 10000 points (100 values for each parameter). It took some trial and error to determine reasonable bounds for the plot. Here is code that generates it:

```r
> plot(c(0,7),c(0,100),type="n",xlab="alpha",ylab="lambda",cex.axis=1.3,cex.lab=1.3)
> for(i in 1:100) {
+ for(j in 1:100) {
+ if(Like(i/15,j) < 10^-5) points(i/15,j,col="grey95",pch=15)
+ else if(Like(i/15,j) < 10^-3) points(i/15,j,col="grey75",pch=15)
+ else if(Like(i/15,j) < 10^-2) points(i/15,j,col="grey55",pch=15)
+ else if(Like(i/15,j) < 2*10^{-2}) points(i/15,j,col="grey35",pch=15)
+ else if(Like(i/15,j) < 4*10^{-2}) points(i/15,j,col="red",pch=15)
+ }
+ }
```
Loops in R

You should be able to try to copy and paste the previous code without problems. The code uses for loops, so these should be explained if you haven’t seen them before.

The idea behind a for loop is to execute a bit of code repeatedly, as many times as specified in the loop. For loops are natural ways to implement summation signs. For example, $\sum_{i=1}^{10} i^2$ can be evaluated in R as

```r
> sum <- 0
> for(i in 1:10) {
+   sum <- sum + i^2
+ }
> sum
[1] 385
```

For loops are also useful for entering in the values of vectors or matrices one by one.
I plotted the likelihood rather than the log-likelihood. For this data set, there were only 5 observations, so we didn’t run into numerical problems with the likelihood. Using a grid search, it mattered very little whether we used the likelihood or log likelihood. However, many of the likelihoods are less than $10^{-6}$ with only five observations. With 100 observations, you could easily have likelihoods around $10^{-100}$, so you might need to use logarithms for larger sample sizes.

It would be good practice to plot the log-likelihood surface rather than the likelihood surface. As in the one-dimensional case, the log-likelihood tends to look flatter than the likelihood, although this will partly depend on how you choose your color scheme.
Heatmap approach

An easier approach is to use a built-in function such as `image()`. The idea here is again to use color to encode the likelihood for each combination of parameters. Here is code that accomplishes this assuming that the object `likes` has 3 columns: horizontal axis value, vertical axis value, and likelihood.

```
> image(likes2,axes=F)
> axis(1,labels=c(0,2,4,6,8,10),at=c(0,.2,.4,.6,.8,1.0))
> axis(2,labels=c(0,2,4,6,8,10),at=c(0,.2,.4,.6,.8,1.0))
> mtext(side=1,expression(alpha),cex=1.3,at=.5,line=3)
> mtext(side=2,expression(lambda),cex=1.3,at=.5,line=3)
```

The axis are scaled to be between 0 and 1 by default, so I specified no axes, and then used the `axis()` command to have customized axes.
Heatmap approach
There are two ways to encode a matrix of likelihood values. One is a matrix where the $ij$th component is the likelihood for $\alpha = \alpha_i$ and $\lambda = \lambda_j$. The second is the previous approach where the values of $\alpha$ and $\lambda$ are given in separate columns and the third column is the likelihood. This first approach is used by `image()`. The second approach might be used by other plotting functions in R.
Matrix of log-likelihoods (parameter values from 1 to 10, not 0 to 1)

e.g., image(log(likes2),col=topo.colors(24))
If you don’t want to assume a model for survival times, you can instead use nonparametric methods. We’ll begin assuming we have right-censored data.

The idea is that instead of estimating a smooth curve from a family of functions for the survival function, we’ll use the observed times as giving the best estimates of surviving for that length of time. We therefore think about the survival function directly instead of working through the likelihood using a density function.
Empirical Cumulative Distribution Function (ECDF)

The approach is related to the empirical distribution function that is used in other parts of nonparametric statistics. Mathematically, the ECDF can be written as

$$\hat{F}_n(x) = \text{(proportion of observations} \leq x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$$

where $I(x_i \leq x) = 1$ if $x_i \leq x$ and is otherwise 0. The function is plotted as a step function where vertical shifts occur at distinct values observed in the data.

For example, if your data are 1.5, 2.1, 5.2, 6.7, then $\hat{F}(3) = \hat{F}(4) = 0.5$ because 50% of your observations are less than or equal to both 3 and 4. $\hat{F}(x)$ then jumps to 0.75 at $x = 5.2$. 

Two ECDFs

ECDFs for men and women normal body temperature

Temperature

CDF

Men

Women
Nonparametric survival curve estimation

For survival analysis, we instead want an empirical estimator of the survival function, so we want the number of observations greater than a certain value, but we also need to account for censoring.

We also need to allow for ties in the times of events, including for non-censored events. For this, we’ll use the notation that \( t_i \) is the \( i \)th distinct death time, so that

\[
    t_1 < t_2 < \cdots < t_D
\]

with \( d_i \) deaths occurring at time \( t_i \). If only one person died at time \( t_i \), then \( d_i = 1 \), and if two people died at time \( t_i \), then \( d_i = 2 \), etc.
For notation, we also let $Y_i$ be the number of individuals who are at risk at time $t_i$ (i.e., individuals who are alive and haven't dropped out of the study for whatever reason).

The quantity $\frac{d_i}{Y_i}$ is the proportion of people at risk at time $t_i$ who died at time $t_i$. 
Kaplan-Meier estimator of the survival function

Kaplan and Meier proposed an estimator of the survival function as

\[
\hat{S}(t) = \begin{cases} 
1 & t < t_1 \\
\prod_{t_i \leq t} \left[ 1 - \frac{d_i}{Y_i} \right] & t \geq t_1 
\end{cases}
\]

Recall that \( t_1 \) is the earliest observed death.
First lets consider an example with no censoring. Suppose we have the following death times (in months):

\[ 8, 10, 15, 15, 30 \]

For this data, we have:

\[ t_1 = 8, t_2 = 10, t_3 = 15, t_4 = 30 \quad d_1 = 1, d_2 = 1, d_3 = 2, d_4 = 1 \]

\[ Y_1 = 5, Y_2 = 4, Y_3 = 3, Y_4 = 1 \]

The estimator says that the probability of surviving any quantity of time less than \( t_1 = 8 \) months is 1, since no one has died sooner than 8 months.
We have that $\hat{S}(7.99) = 1$. What is $\hat{S}(8.0)$?

For this case $t \geq t_1 = 1$, so we go to the second case in the definition. Then we need the product over all $t_i \leq 8.0$. Since there is only one of these, we have

$$\hat{S}(8.0) = 1 - \frac{d_1}{Y_1} = 1 - \frac{1}{5} = 0.80$$

The Kaplan-Meier estimate for surviving more than 8 months is simply the number of people in the study who did, in fact, survive more than 8 months.
Kaplan Meier estimator of the survival function

Note that if we want something like $\hat{S}(9)$, which is a time in between the observed death times, then since there was only one time less or equal to 9, we get the same estimate as for $\hat{S}(8)$. The Kaplan-Meier estimate of the survival function is flat in between observed death times (even if there is censoring in between those times and the number of subjects changes).

Consequently, the Kaplan-Meier estimate looks like a step function, with jumps in the steps occurring at observed death times.
Kaplan-Meier estimator of the survival function

To continue the example,

\[
\hat{S}(10) = \prod_{t_i \leq 10} \left[ 1 - \frac{d_i}{Y_i} \right]
\]

\[
= \left[ 1 - \frac{d_1}{Y_1} \right] \left[ 1 - \frac{d_2}{Y_2} \right]
\]

\[
= \left[ 1 - \frac{1}{5} \right] \left[ 1 - \frac{1}{4} \right]
\]

\[
= \frac{4}{5} \cdot \frac{3}{4} = \frac{3}{5}
\]

You can see that the final answer is the number of people who were alive after 10 months, which is fairly intuitive. You can also see that there was cancellation in the product.
The estimated survival function won’t change until $t = 15$. So now we have

$$\hat{S}(15) = \prod_{t_i \leq 15} \left[ 1 - \frac{d_i}{Y_i} \right]$$

$$= \left[ 1 - \frac{d_1}{Y_1} \right] \left[ 1 - \frac{d_2}{Y_2} \right] \left[ 1 - \frac{d_3}{Y_3} \right]$$

$$= \left[ 1 - \frac{1}{5} \right] \left[ 1 - \frac{1}{4} \right] \left[ 1 - \frac{2}{3} \right]$$

$$= \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{5}$$

Again, the probability is the proportion of people still alive after time $t$. 
Kaplan-Meier estimator of the survival function

At first it might seem odd that the K-M function, which is a product (the K-M estimator is also called the Product-Limit Estimator), is doing essentially what the ECDF function is doing with a sum. One way of interpreting the K-M function is that $1 - d_i/Y_i$ is the probability of not dying at time $t_i$.

Taking the product over times $t_1, \ldots, t_k$ means the probability that you don’t die at time $t_1$, that you don’t die at time $t_2$ given that you don’t die at time $t_1$, and ... and that you don’t die at time $t_k$ that you haven’t died at any previous times.

The conditional probabilities come into play because $Y_i$ is being reduced as $i$ increases, so we are working with a reduced sample space. The product therefore gives the proportion of people in the sample who didn’t die up until and including time $t$. 
Kaplan-Meier estimator

If we didn’t have censoring, then we could just use the ECDF and subtract it from 1 to get the estimated survival function. What’s brilliant about the K-M approach is that generalizes to allow censoring in a way that wouldn’t be clear how to do with the ECDF.

To work with the K-M estimator, it helpful to visualize all the terms in a table. We can also compute the estimated variance of \( \hat{S}(t) \), which is denoted \( \hat{V}[\hat{S}(t)] \). The standard error is the square root of the estimated variance. This allows us to put confidence limits on \( \hat{S}(t) \).

One formula (there are others that are not equivalent) for the estimated variance is:

\[
\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}
\]
Now let’s try an example with censoring. We’ll use the example that we used for the exponential:

\[ 1.5, 2.4, 10.5, 12.5^+, 15.1, 20.2^+ \]

In this case there are no ties, but recall that \( t_i \) refers to the \( i \)th death time.
Kaplan-Meier example with censoring

Consequently, we have

\[ t_1 = 1.5, \ t_2 = 2.4, \ t_3 = 10.5, \ t_4 = 15.1, \quad d_1 = d_2 = d_3 = d_4 = 1 \]

\[ Y_1 = 6, \ Y_2 = 5, \ Y_3 = 4, \ Y_4 = 2, \ Y_5 = 1 \]

Following the formula we have

\[ \hat{S}(1.5) = \left[ 1 - \frac{1}{6} \right] = 0.83 \]

\[ \hat{S}(2.4) = \left[ 1 - \frac{1}{6} \right] \left[ 1 - \frac{1}{5} \right] = 0.67 \]

\[ \hat{S}(10.5) = \left[ 1 - \frac{1}{6} \right] \left[ 1 - \frac{1}{5} \right] \left[ 1 - \frac{1}{4} \right] = 0.5 \]

\[ \hat{S}(15.1) = (0.5) \left[ 1 - \frac{1}{3} \right] = 0.167 \]
Comparison to MLE

It is interesting to compare to the MLE that we obtained earlier under the exponential model. For the exponential model, we obtained $\hat{\lambda} = 0.066$. The estimate survival function at the observed death times are

\[
> 1 - \text{pexp}(1.5, \text{rate} = .066)
\]

[1] 0.9057427

\[
> 1 - \text{pexp}(2.4, \text{rate} = .066)
\]

[1] 0.8535083

\[
> 1 - \text{pexp}(10.5, \text{rate} = .066)
\]

[1] 0.5000736

\[
> 1 - \text{pexp}(15.1, \text{rate} = .066)
\]

[1] 0.3691324
The exponential model predicted higher survival probabilities at the observed death times than Kaplan-Meier except that they both estimate $\hat{S}(10.5)$ to be 0.5 (or very close for the exponential model. Note that the Kaplan Meier estimate still has an estimate of 50% survival for, say 12.3 months, whereas the exponential model estimates 44% for this time. As another example, $\hat{S}(10.0) = 0.67$ for Kaplan-Meier but 0.51 for the exponential model. The exponential model seems to be roughly interpolating between the values obtained by K-M.
K-M versus exponential
TABLE 1.1
Remission duration of 6-MP versus placebo in children with acute leukemia

<table>
<thead>
<tr>
<th>Pair</th>
<th>Remission Status at Randomization</th>
<th>Time to Relapse for Placebo Patients</th>
<th>Time to Relapse for 6-MP Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Partial Remission</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Complete Remission</td>
<td>22</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>Complete Remission</td>
<td>3</td>
<td>32+</td>
</tr>
<tr>
<td>4</td>
<td>Complete Remission</td>
<td>12</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>Complete Remission</td>
<td>8</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>Partial Remission</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>Complete Remission</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>Complete Remission</td>
<td>11</td>
<td>34+</td>
</tr>
<tr>
<td>9</td>
<td>Complete Remission</td>
<td>8</td>
<td>32+</td>
</tr>
<tr>
<td>10</td>
<td>Complete Remission</td>
<td>12</td>
<td>25+</td>
</tr>
<tr>
<td>11</td>
<td>Complete Remission</td>
<td>2</td>
<td>11+</td>
</tr>
<tr>
<td>12</td>
<td>Partial Remission</td>
<td>5</td>
<td>20+</td>
</tr>
<tr>
<td>13</td>
<td>Complete Remission</td>
<td>4</td>
<td>19+</td>
</tr>
<tr>
<td>14</td>
<td>Complete Remission</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>Complete Remission</td>
<td>8</td>
<td>17+</td>
</tr>
<tr>
<td>16</td>
<td>Partial Remission</td>
<td>23</td>
<td>35+</td>
</tr>
<tr>
<td>17</td>
<td>Partial Remission</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>Complete Remission</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>19</td>
<td>Complete Remission</td>
<td>4</td>
<td>9+</td>
</tr>
<tr>
<td>20</td>
<td>Complete Remission</td>
<td>1</td>
<td>6+</td>
</tr>
<tr>
<td>21</td>
<td>Complete Remission</td>
<td>8</td>
<td>10+</td>
</tr>
</tbody>
</table>

+Censored observation.
### TABLE 4.1A

*Construction of the Product-Limit Estimator and its Estimated Variance for the 6-MP Group*

<table>
<thead>
<tr>
<th>Time $t_i$</th>
<th>Number of events $d_i$</th>
<th>Number at risk $Y_i$</th>
<th>Product-Limit Estimator $\hat{S}(t) = \prod_{t_i \leq t} [1 - \frac{d_i}{Y_i}]$</th>
<th>$\sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i-d_i)}$</th>
<th>$\hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i-d_i)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>3</td>
<td>21</td>
<td>$[1 - \frac{3}{21}] = 0.857$</td>
<td>$\frac{3}{21 \times 18} = 0.0079$</td>
<td>$0.857^2 \times 0.0079 = 0.0058$</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>17</td>
<td>$[0.857](1 - \frac{1}{17}) = 0.807$</td>
<td>$0.0079 + \frac{1}{17 \times 16} = 0.0116$</td>
<td>$0.807^2 \times 0.0116 = 0.0076$</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>15</td>
<td>$[0.807](1 - \frac{1}{15}) = 0.753$</td>
<td>$0.0116 + \frac{1}{15 \times 14} = 0.0164$</td>
<td>$0.753^2 \times 0.0164 = 0.0093$</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>12</td>
<td>$[0.753](1 - \frac{1}{12}) = 0.690$</td>
<td>$0.0164 + \frac{1}{12 \times 11} = 0.0240$</td>
<td>$0.690^2 \times 0.0240 = 0.0114$</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>11</td>
<td>$[0.690](1 - \frac{1}{11}) = 0.628$</td>
<td>$0.0240 + \frac{1}{11 \times 10} = 0.0330$</td>
<td>$0.628^2 \times 0.0330 = 0.0130$</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>7</td>
<td>$[0.628](1 - \frac{1}{7}) = 0.538$</td>
<td>$0.0330 + \frac{1}{7 \times 6} = 0.0569$</td>
<td>$0.538^2 \times 0.0569 = 0.0164$</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>6</td>
<td>$[0.538](1 - \frac{1}{6}) = 0.448$</td>
<td>$0.0569 + \frac{1}{6 \times 5} = 0.0902$</td>
<td>$0.448^2 \times 0.0902 = 0.0181$</td>
</tr>
</tbody>
</table>