Weibull in R

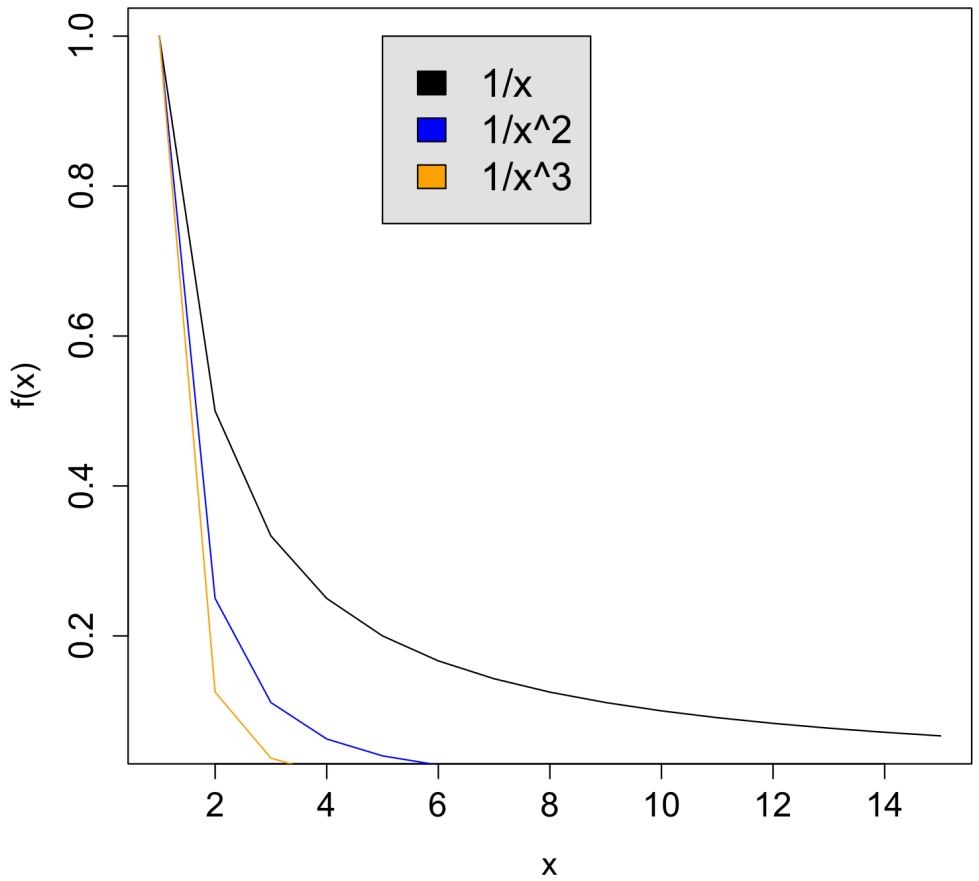
The Weibull in R is actually parameterized a fair bit differently from the book. In R, the density for $x > 0$ is

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

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 $a = \texttt{shape} = 0.5$, $b = \texttt{scale} = (1/\lambda)^{1/\alpha} = (1/1.2)^{1/0.5}$.

Adding legends to plots

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



Adding legends to plots

```
x <- 1:15
plot(x,1/x,xlab="x",type="l",ylab="f(x)",ylim=c(0,1),cex.lab=1.3,cex.axis=1.4)
points(x,1/x^2,type="l",col="blue",cex=3,pch=15) #pch ignored
points(x,1/x^3,type="l",col="orange",cex=3,pch=16) #because line
legend(5,1,legend=c("1/x","1/x^2","1/x^3"),fill=c("black","blue","orange"),
cex=1.6,bty="o",bg="grey90")
```

Adding points and lines to a plot

```
x <- 1:15
plot(x,1/x,xlab="x",type="l",ylab="f(x)",ylim=c(0,1),cex.lab=1.3,
cex.axis=1.4)
points(x,1/x^2,type="l",col="blue",cex=3,pch=15) #pch ignored
points(x,1/x^3,type="l",col="orange",cex=3,pch=16) #because line
points(x,1/x^2,col="blue",cex=3,pch=15) #pch ignored
points(x,1/x^3,col="orange",cex=3,pch=16) #because line
points(x,1/x,cex=3)
legend(5,1,legend=c("1/x","1/x^2","1/x^3"),fill=c("black","blue",
"orange"),
cex=1.6,bty="o",bg="grey90")
```

Plotting

Try copying and pasting the previous code into R and modifying bits and pieces. It might be easiest to copy and paste the code into a file called `plot.r` or something similar and type `> source(‘‘plot.r")`. The code should be in the same directory, or you can try an environment like R Studio to have the code nearby.

Try `help(plot)`, `help(plot.default)`, and `help(legend)` for more plotting options. You can also try to search for “R graphics gallery” online to see examples of R graphics with source code that you can modify.

Mean residual life

The mean residual life is the expected remaining lifetime given that someone has survived up to time x. This is

$$\text{mrl}(x) = E(X - x | X > x)$$

You can also think of this as the area under the survival curve to the right of x divided by $S(x)$. (See next slide)
For conditional expectations generally, we have $E(X|X > x) =$

$$\sum_{k>x} kP(X = k|X > x) = \frac{\sum_{k>x} kP(X = k, X > x)}{P(X > x)} = \frac{\sum_{k>x} kP(X = k)}{P(X > x)}$$

The analogous expression for a continuous random variable is

$$E(X|X > x) = \frac{\int_x^\infty uf(u)\, du}{P(X > x)}$$

Mean residual life

Note that $\text{mrl}(0) = E(X - 0|X > 0) = E(X)$ if X is a positive random variable. Generally,

$$E(X - x|X > x) = \frac{\int_x^\infty (t - x) f(t)\, dt}{P(X > x)}$$

You can use integration by parts to show that the numerator is $\int_x^\infty S(t)\, dt$ (I'm using t instead of u because I use u for integration by parts). Plugging in $x = 0$ gives us that

$$E(X) = \int_0^\infty S(x)\, dx$$

Integration by parts

You should verify for yourself that the integration by parts works out. Use the formula

$$\int u\, dv = uv - \int v\, du$$

where u and v are functions of t. Let

$$\begin{aligned} dv &= f(t)\, dt \\ v &= -S(t) \\ u &= t - x \\ du &= dt \end{aligned}$$

Lack of memory property for the exponential

The memoryless property of the exponential says that

$$P(X > x + z | X > x) = P(X > z)$$

If we apply this to the mean residual life, this means that

$$E(X - x | X > x) = E(X) = 1/\lambda$$

so the mean residual life is constant.

Quantiles

The pth quantile is the value x_p such that

$$S(x_p) = 1 - p$$

or equivalently that $F(x_p) = p$. This is implemented in R using functions such as `qexp()`, `qweibull`, etc.

As an example, the median of a distribution is the value x_m such that $F(x_m) = S(x_m) = 0.5$, and this is found in R using, for example `qexp(.5,rate=3)` (median of an exponential with rate 3). The 99th percentile is found using `qexp(.99,rate=3)`.

Quantiles

This can sometimes be found analytically. For example the pth quantile for an exponential satisifies

$$S(x_p) = 1 - p \Rightarrow e^{-\lambda x} = 1 - p \Rightarrow -\lambda x = \log(1-p) \Rightarrow x = \frac{-\log(1-p)}{\lambda}$$

For the median, this is $\frac{-log(1/2)}{\lambda}$. Recalling properties of logarithms, that $\log x^a = a \log x$, we have $x_m = \frac{\log 2}{\lambda}$ for the median of an exponential distribution.

Other distributions

We'll look at both parametric and nonparametric models for survival data, but parametric models are more common. Knowing which model best fits the data can help understand the data better than a non-parametric model (Table, p. 38)

Common models include:

- exponential
- Weibull
- gamma and generalized gamma
- lognormal
- log logistic
- exponential power
- normal
- Gompertz
- inverse Gaussian
- Pareto

Censoring and truncation

Chapter 2 gives an overview of some model-based techniques such as regression and models for competing risks. We'll move on to Chapter 3 to discuss censoring in more detail.

Censoring

Generally, censoring occurs when the value for a time is only known within an interval. We can distinguish right-censoring, where we know that $x_i > x$, left-censoring, where we know that $x_i < x$, and interval censoring, where we know that $a < x_i < b$.

Right-censoring is the kind of censoring we talked about the first week of class, where a patient drops out of study perhaps due to moving or unrelated death.

Left-censoring could occur if you wanted to model the time of onset of a disease to death, where the time of death is known but the time of onset is estimated. Another example might be length of pregnancy, where the exact start of the pregnancy is only estimated.

Interval censoring occurs when both left- and right-censoring occur.

Right censoring

We'll start looking at Type I censoring. In the simplest case, we think of the censoring time as fixed at the end of the study, with all patients entering the study at the same time.

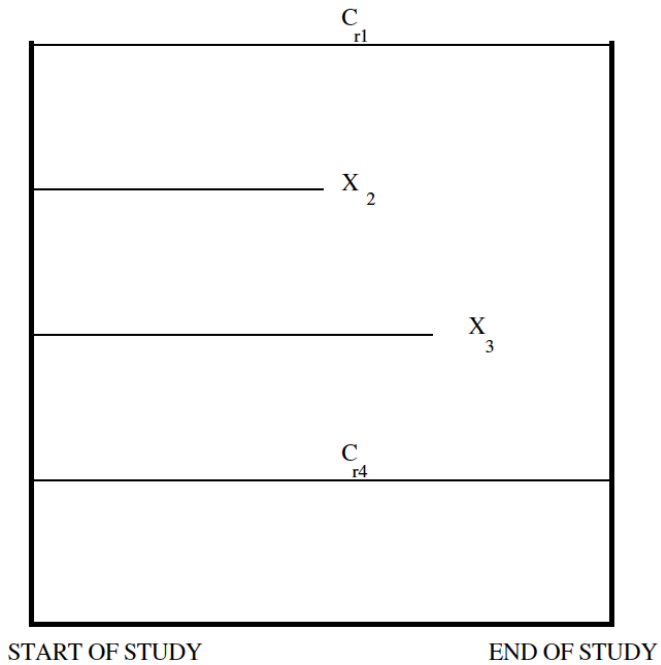


Figure 3.1 *Example of Type I censoring*

Right censoring

The notation used here is that the observed time is either X_i or C_{ri} depending on whether the ith observation is observed or censored. C_r is used to stand for right-censoring.

Type I censoring can be generalized so that different individuals have different start times

Right censoring

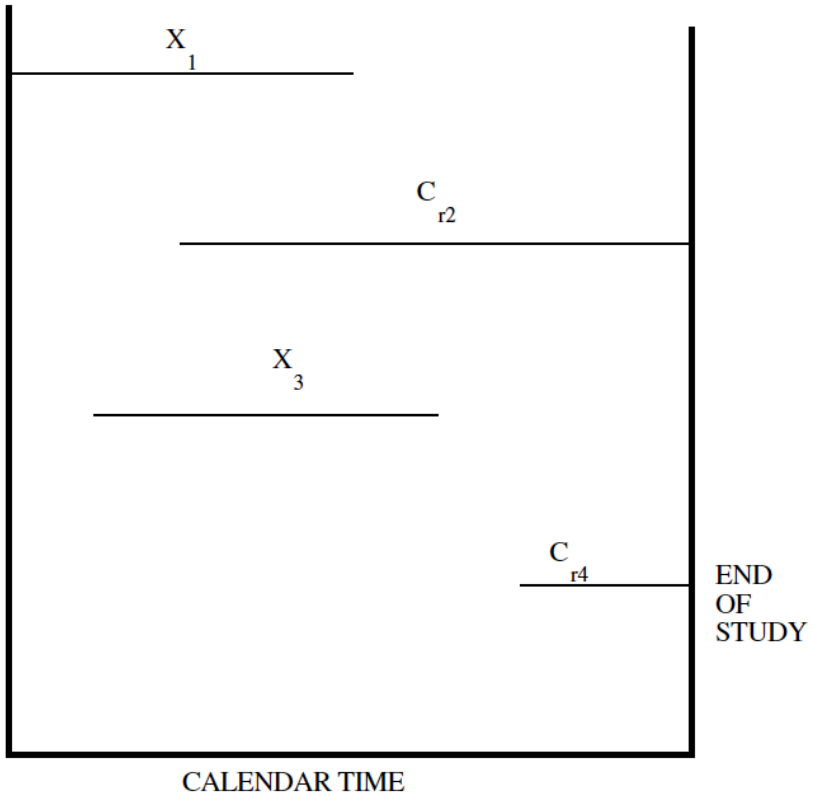


Figure 3.3 *Generalized Type I censoring when each individual has a different starting time*

Right censoring

This can be dealt with by treating each patient as starting at time 0 but with individualized fixed censoring times.

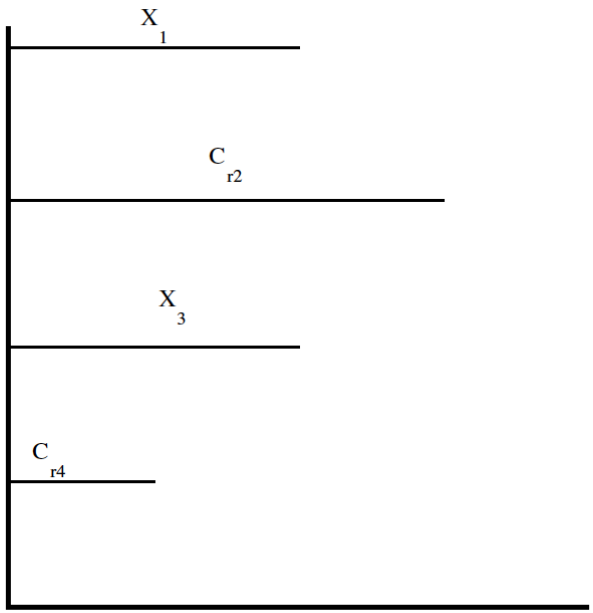


Figure 3.4 *Generalized Type I censoring for the four individuals in Figure 3.3 with each individuals starting time backed up to 0. $T_1 = X_1$ (death time for first individual) ($\delta_1 = 1$); $T_2 = C_{r2}$ (right censored time for second individual) ($\delta_2 = 0$); $T_3 = X_3$ (death time for third individual) ($\delta_3 = 1$); $T_4 = C_{r4}$ (right censored time for fourth individual) ($\delta_4 = 0$).*

Right censoring with two censoring times

Another complication is to have multiple planned fixed censoring events. In an animal study, only a subset of animals might be followed for a longer period of time in order to reduce costs of the study.

Type II right-censoring

In Type II right-censoring, a study continues until a fixed number r of individuals experiences the event, at which point the study is terminated. For example, if $r = n/2$, then this will tell you the median time to failure without having to wait for all items to fail. This can be used in animal studies (again to save money) or for testing equipment (median time to failure for lightbulbs).

In this type of censoring, if you think of the smallest failure times, $X_{(1)}$, $X_{(2)}$, etc., then these are called order statistics. Here we observe the r smallest failure times, but not the failure times of the remaining $n - r$ times. The smallest r failure times are the order statistics $X_{(1)}, \ldots, X_{(r)}$, and these can be used to estimate the distrbution of X without observing all data using maximum likelihood and theoretical results for order statistics.

Progressive type II right-censoring

To reduce costs, you can also generalize type II censoring so that after r items have failed, a subset of the remaining $n - r$ items is followed until r_1 of these fail. The process can be reiteratated through multiple rounds.

Competing risks censoring

Here accidental death or other events might remove a subject at a random time from a study, which is different from censoring due to the study ending. This generally requires independence between the random censoring time and time to failure.

Most studies have a combination of random (competing risks) censoring and type I censoring.

Representing censoring mathematically

Mathematically, it is convenient to think of the event as occurring at some time X, whether or not it is observed, and to let another random variable, say δ indicate whether or not censoring occurs. If censoring occurs at a fixed time, C_r, then let $T = \min(X, C_r)$. Then we observe (T, δ) instead of X. That is we observe a time T, and we know whether or not censoring has occurred, but we don't directly observe X. Here, you can let $\delta = 0$ if censoring occurs and otherwise $\delta = 1$ (or vice versa).

Note that in this notation, C_r is not random if the censoring time is fixed, even though we used a capital letter. (I'm following the notation in the book.)

Left censoring

Here we let C_l be a left-censoring time, meaning that the event occurred at some time $X < C_l$, where the exact value for X is unknown. Observations can again be represented as pairs (T, ε) where $\varepsilon = 1$ if the observation is not left-censored, and otherwise $\varepsilon = 0$, and $T = \max(X, C_l)$ (as opposed to using the minimum for right-censoring).

Examples of left-censoring include

- determining when a child learns a task after enrolling in a study, where some of the children already know the task when first enrolled.
- asking when a patient first had a certain symptom before some event, where the patient knows they've had the symptom since at least a certain number of months ago
- wanting to measure time of onset of disease to an event, where we only know the time of the first diagnosis

Truncation

Truncation seems similar to censoring, but is not quite the same. The idea is that certain individuals are observed only when their time comes within a certain window.

An example is a survival analysis where patients receive Social Security. Since there is a minimum age for Social Security recipients, the analysis only includes patients that are at least old enough to receive social security. In this example, the data is left-truncated and we must use conditional probabilties — probabilities of survival given that the patient is at least old enough to receive Social Security.

Right-truncation might occur if you only observe deaths, even though there might be cases where deaths did not occur, but you do not know how many of these occurred.

Mathematically, the main difference between censoring and truncation is that truncation requires using conditional probabilities.

Likelihoods for censored data

To model survival data, we consider the probability of the data (or joint density of the data) as a function of the parameters in the model. The idea is then to find the particular values of the parameters that best fit the data, meaning that they make the data more probable, or less surprising.

The technique is called **maximum likelihood**, which is worth reviewing

Maximum Likelihood

Suppose I flip a coin 100 times, and I observe heads 90 times. What is the best explanation for this data?

Maximum Likelihood

If the coin was fair, it would be very surprising to observe 90 heads out of 100. If the coin was biased, then it would not be as surprising. There is still some probability of observing tails, so we wouldn't think that the coin always came out heads. If there is some probability of the coin coming out heads, then $P(H) = .9$ is the best guess, and this is the maximum likelihood estimate.

We can think of a maximum likelihood estimate as an inference to the best explanation of the data.

Maximum Likelihood

For the coin flipping example, we could also consider all possible parameters p for the probability of heads, and compute the probability of the data as a function of p. Then we would have

$$L(p) = P(90 \text{ heads}) = \binom{100}{90} p^{90}(1-p)^{10}$$

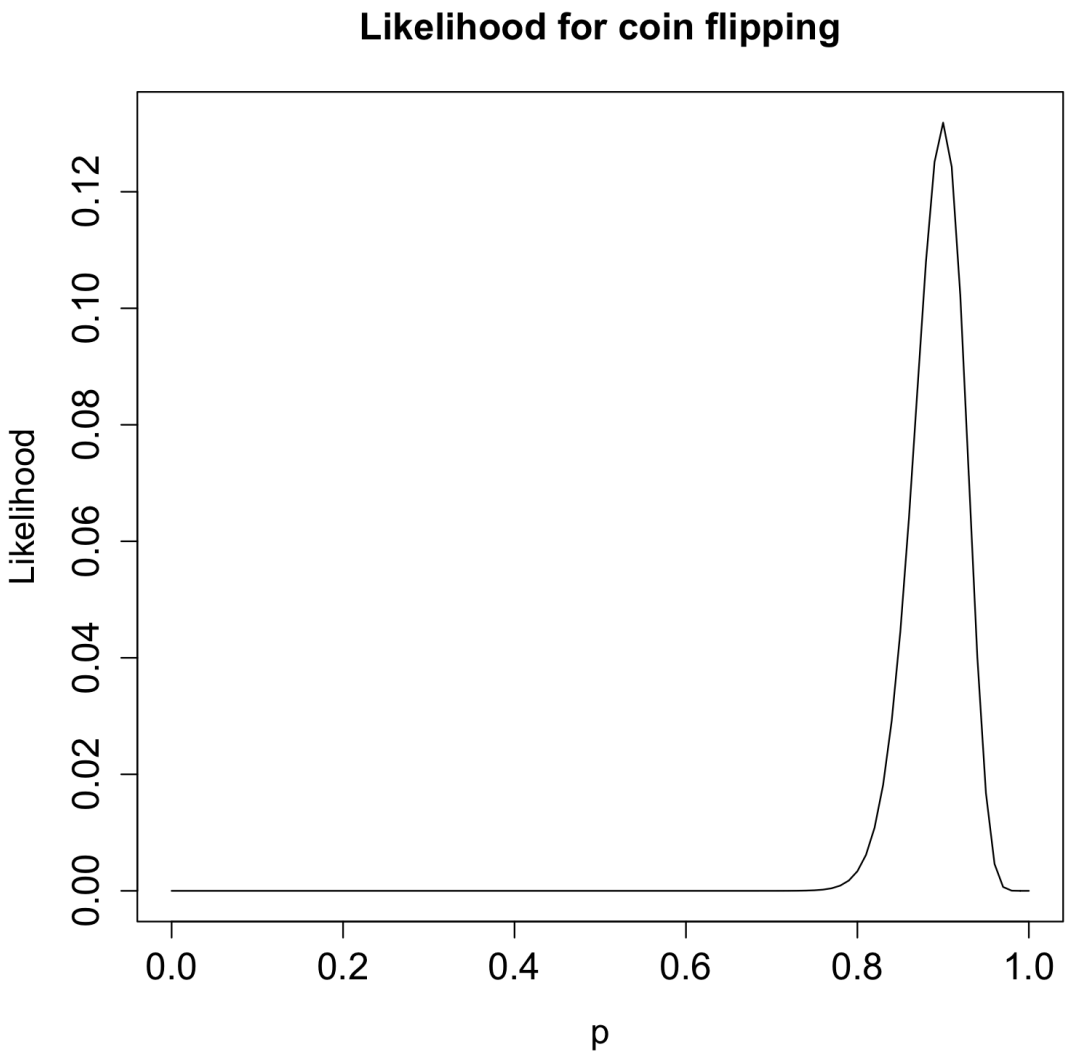
Recall that $\binom{a}{b} = \frac{a!}{b!(a-b)!}$ and that $a! = 1 \times 2 \times \cdots \times a$, with $0! = 1$. Also recall that $\binom{a}{b} = \binom{a}{a-b}$, so $\binom{100}{90} = \binom{100}{10}$.
If we plot this function, it will have a maximum at $p = 0.9$.

Plotting the likelihood in R

To plot the likelihood in R, compute the probability as a function of p. we might want to try all values of p from 0 to 1 in increments of 0.01, so

```
> p <- seq(0,1,.01)
> L <- choose(100,90)*p^90*(1-p)^10
> plot(p,L,type="l",xlab="p",ylab="Likelihood",cex.lab=1.3,
cex.axis=1.4,main="Likelihood for coin flipping",cex.main=1.4)
```

The likelihood function



Maximum Likelihood

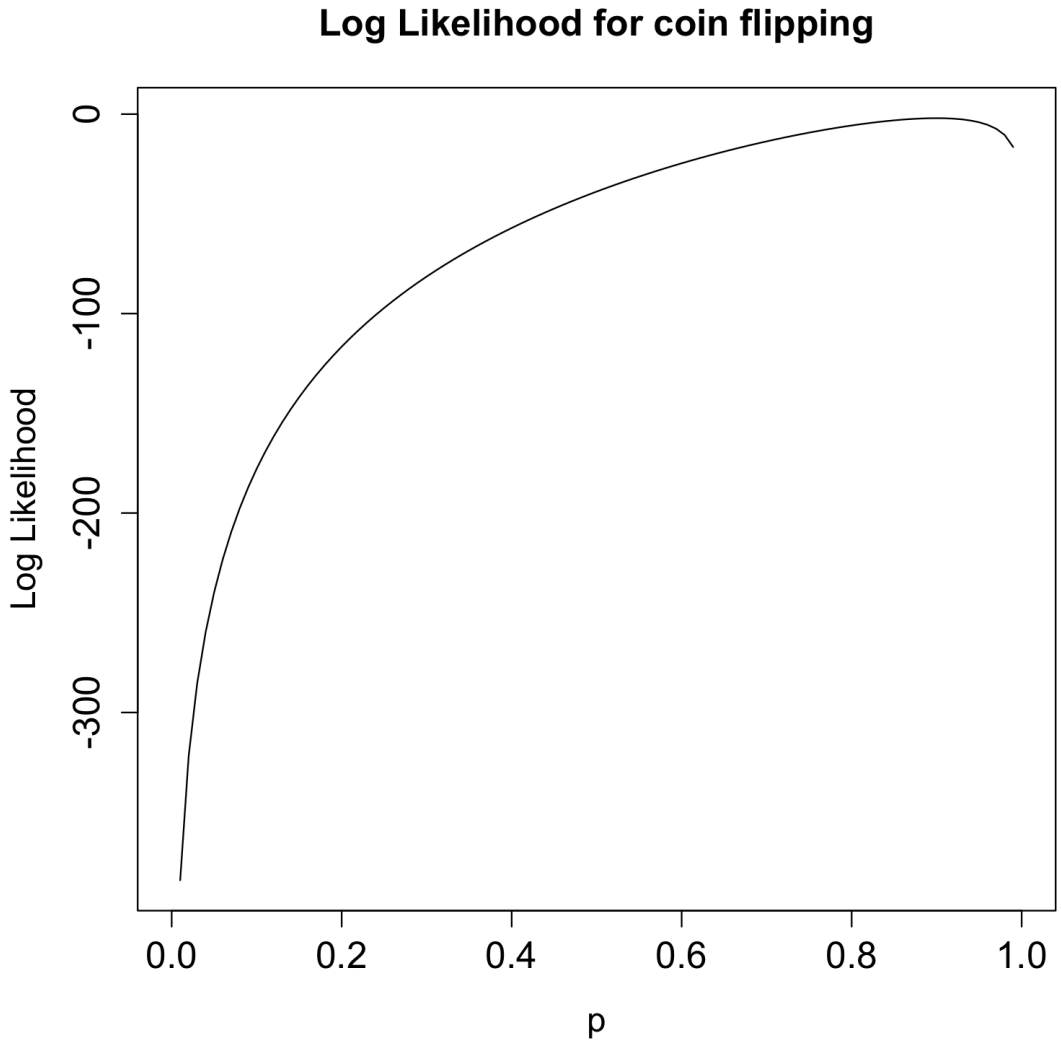
In many cases, the likelihood function must be maximized numerically in the computer. We often work with the (natural) logarithm of the likelihood rather than the likelihood function itself. There are two main reasons for this:

Maximum Likelihood

We often work with the log-likelihood because

- Probabilities (or density values) of any particular observed data are typically very small for large numbers of observations. We are not so much interested in the absolute probabilities of the data as the relative probabilities for different parameters. The probability of observing exactly 9000 heads in 10000 flips is very small even for $p = .9$, but is relatively speaking much larger for $p = .9$ than for $p = .5$. Logarithms help numerically in the computer to deal with very small numbers. You can easily work with cases that have probabilities (or densities) less than 10^{-10000}, which would be too small to represent directly in the computer. But their logarithms can be represented in the computer.
- Mathematically, probabilities and likelihoods are often products, and since the log of a product is the sum of the logs, the logarithm is often easier to work with mathematically. To maximize a function, you can maximize the log of the function rather than the function itself, and maximizing a sum is easier than maximizing a product

The log-likelihood function



Maximum Likelihood

To illustrate the point about numerical likelihoods, consider the coin flipping example with 10,000 coin flips, 9000 of which are heads. If I try this in R, I get

```
> L <- choose(10000,9000)*p^9000*(1-p)^1000
> L
  [1] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN
 [19] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN
 [37] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN
 [55] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN
 [73] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN
 [91] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN
> L <- p^9000*(1-p)^1000
> L
  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Maximum Likelihood

The numbers are too big for R to work with directly. Taking logarithms directly also doesn't help

```
> log(choose(10000,9000))
[1] Inf
> L <- log(p^9000*(1-p)^1000)
> L
  [1] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -
 [16] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -
 [31] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -
 [46] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -
 [61] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -
 [76] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -
 [91] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
```

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 evaulate these large exponents and binomial coefficients.
The log-likelihood is

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

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 is 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

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

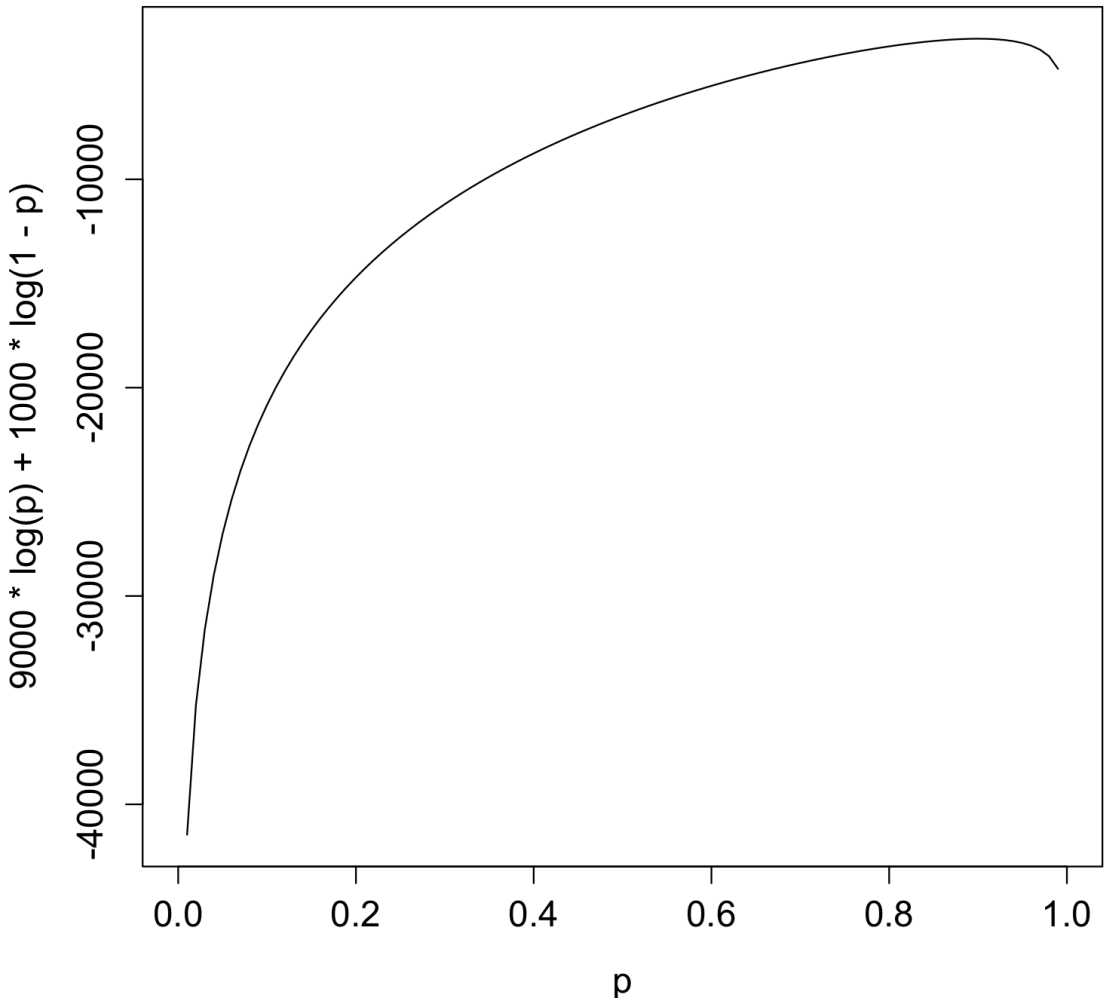
or even just drop the constant altogether. So sometimes, you'll see

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

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



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.

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

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

$$\Rightarrow (1-p)k = p(n-k)$$

$$\Rightarrow k - kp - np + kp = 0$$

$$\Rightarrow k = np$$

$$\Rightarrow p = \frac{k}{n}$$

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 $\widehat{p} = \frac{k}{n}$. We say *estimator* for the general function that works for any data, and *estimate* for a particular value like $\widehat{p} = 0.9$.

Maximum likelihood for the exponential

Suppose you have 3 lightbulbs that last 700, 500, and 1100 hours. Assuming that their lifetimes are exponentially distributed with rate λ, what is the maximum liklihood estimate of λ?

Likelihoods with censoring and truncation

For survival analysis, we need likelihood functions that incorporate censoring. A general framework is to have seperate 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:

$$
\begin{aligned}
f(x) &\text{ exact lifetimes/death times} \\
S(C_r) &\text{ right-censored observations} \\
1 - S(C_l) &\text{ left-censored observations} \\
[S(L) - S(R)] &\text{ interval-censored observations} \\
\frac{f(x)}{S(Y_L)} &\text{ left-truncated observations} \\
\frac{(f(x)}{1 - S(Y_R)} &\text{ right-truncated observations} \\
\frac{f(x)}{S(Y_L) - S(Y_R)} &\text{ interval-truncated observations}
\end{aligned}
$$

Likelihoods with censoring and truncation

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 interval-censored observations.

Likelihoods for truncated data

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

$$\begin{aligned} Pr[T = C_r, \delta = 0] &= Pr[T = C_r | \delta = 0] Pr(\delta = 0) = 1 \cdot Pr(\delta = 0) = Pr(X > C_r) \\ &= S(C_r) \end{aligned}$$

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

$$\begin{aligned} Pr[T, \delta = 1] &= Pr[T = t | \delta = 1] P(\delta = 1) = \frac{Pr(T = t)}{P(\delta = 1)} \cdot P(\delta = 1) = Pr(T = t) \\ &= f(t) \end{aligned}$$

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

Becuase $h(t) = \frac{f(t)}{S(t)}$, 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 λ^n, we have λ_r where $r \leq n$.

log-likelihood for exponential example

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

$$\widehat{\lambda} = \frac{r}{\sum_{i=1}^{n} t_i} = \frac{r}{n\bar{t}}$$

Example with exponential data and right-censoring

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 esimate of λ.

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 o 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

$$\widehat{\lambda} = \frac{4}{60.2} = 0.066$$

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

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

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

$$\widehat{\lambda} = \frac{4}{29.5} = 0.136 \Rightarrow E(T) = 7.38 \text{ months}$$

Constructing the likelihood function: log-logistic example

This example is exercise 3.5 in the book (page 89):

Suppose the time to death has a log-logistic distribution with parameters λ and α. Based on the following left-censored sample, construct the likelihood function.

$$0.5, 1, 0.75, 0.25\text{-}\ 1.25\text{-}$$

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

$$L = \prod_{i \in D} f(x_i) \prod_{i \in L} (1 - S(C_l))$$

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.

log-logistic example

The log-logistic density for $x > 0$ is

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

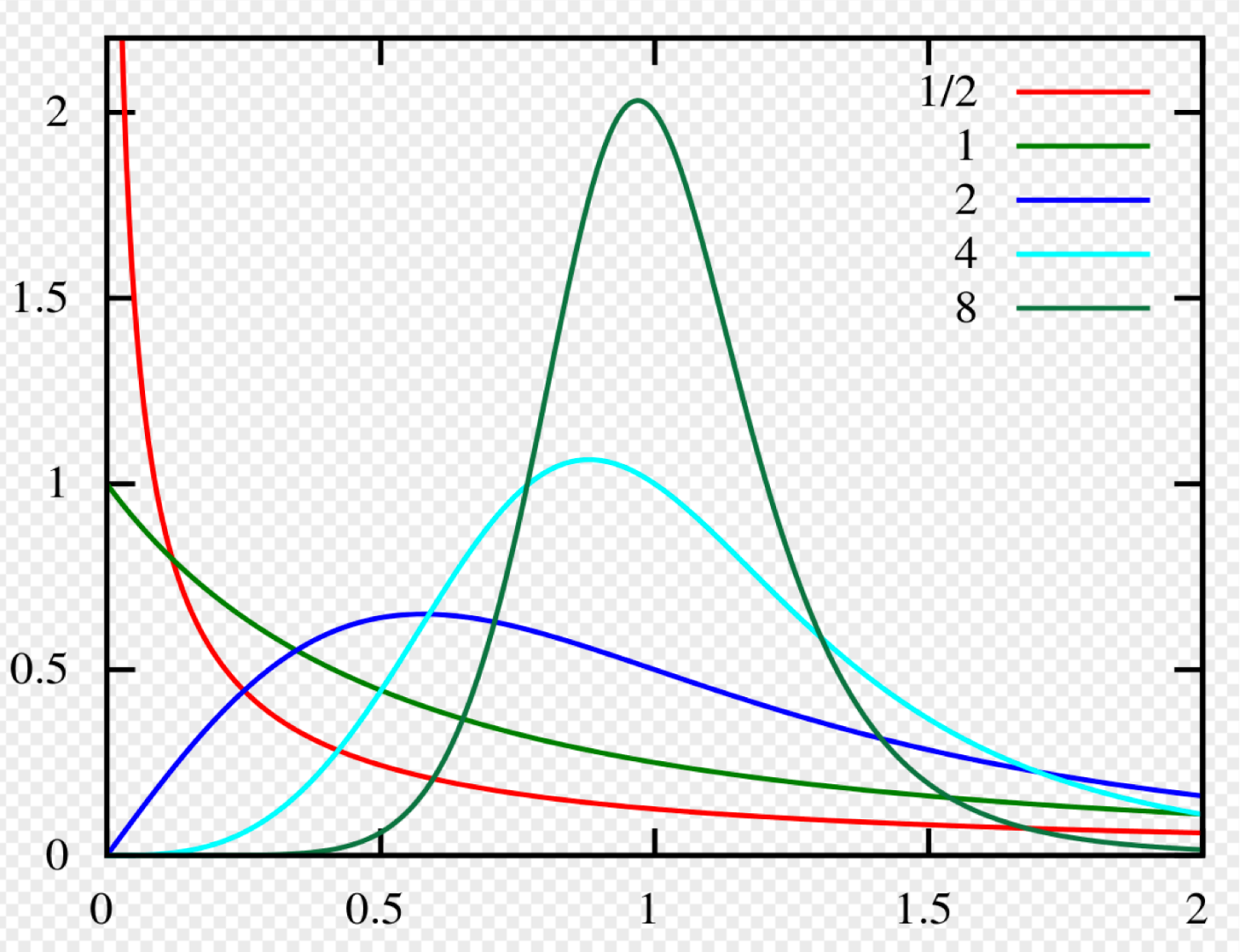
The survival function is

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

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 $\lambda = 1$



log-logistic example

The likelihood is therefore

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

log-logistic example

Using the data, we can write this as

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

log-logisitic example

We can simplify the likelihood as

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

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

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

$$- \sum_{i=1}^{n} \log(1+\lambda x_i^{\alpha}) - \sum_{i\in D} \log(1+\lambda x_i^{\alpha})$$

log-logistic likelihood in R

We'll look at evaluating the log-logistic likeilhood 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.

```
> 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 in R

```
Functions can also have more than one argument.  For example
> 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?

functions in R

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

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

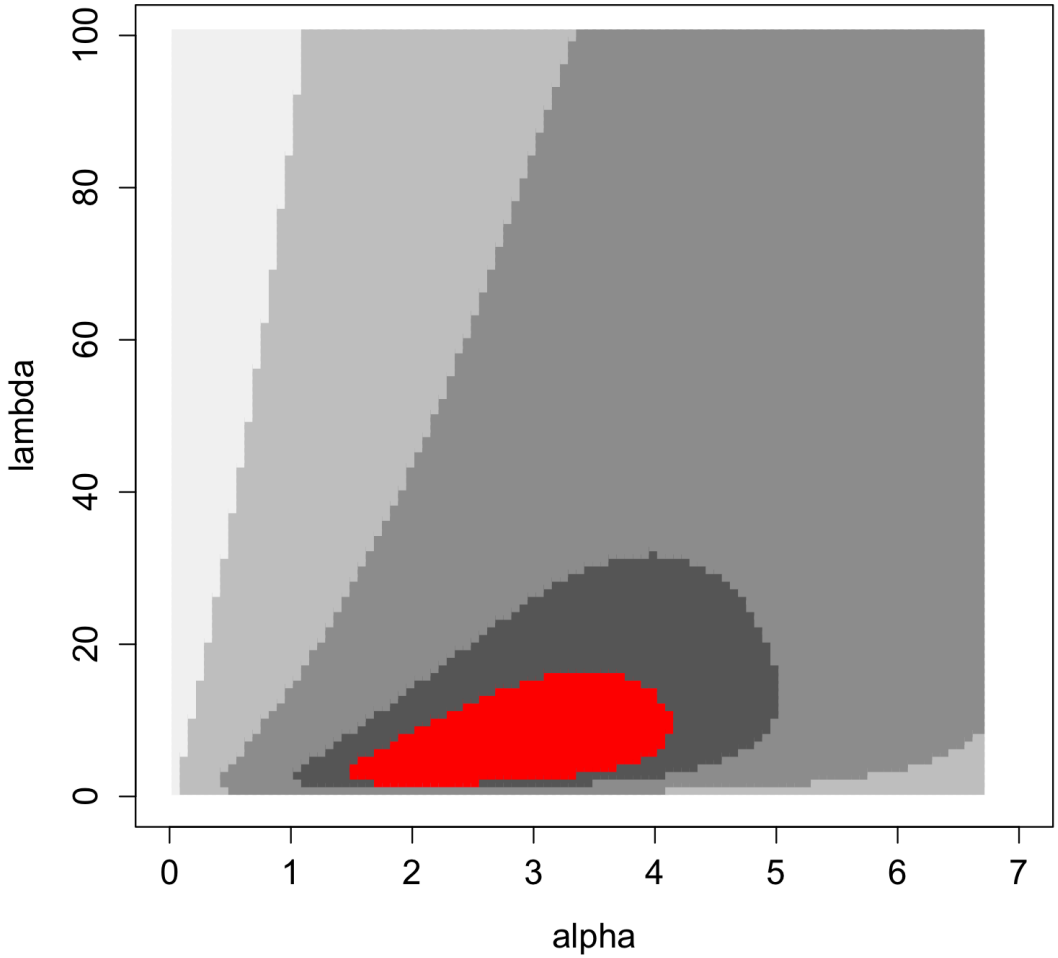
where $\mathtt{x} = \sum_{i=1}^{n} x_i$.

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).

```
> 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 α and λ. I tried for values between 0 and 10 for both α and λ in increments of 0.1. This requires 100 values for α and, independently, 100 values for λ, 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 obtaine $(\widehat{\alpha}, \widehat{\lambda}) = (2.6, 5.0)$, 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 λ changes than as α 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 α by 10% from the (approximate) MLE lowers the likelihood more than increasing λ by 10%.

Generating the likelihood surface

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

```
> 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) < .001) points(i/15,j,col="grey75",pch=15)
+ else if(Like(i/15,j) < .01) points(i/15,j,col="grey55",pch=15)
+ else if(Like(i/15,j) < .02) points(i/15,j,col="grey35",pch=15)
+ else if(Like(i/15,j) < .04) 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

```
> 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.

Likelihood versus log-likelihood

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 lieklihood 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 the likelihood, although this will partly depend on how you choose your color scheme.

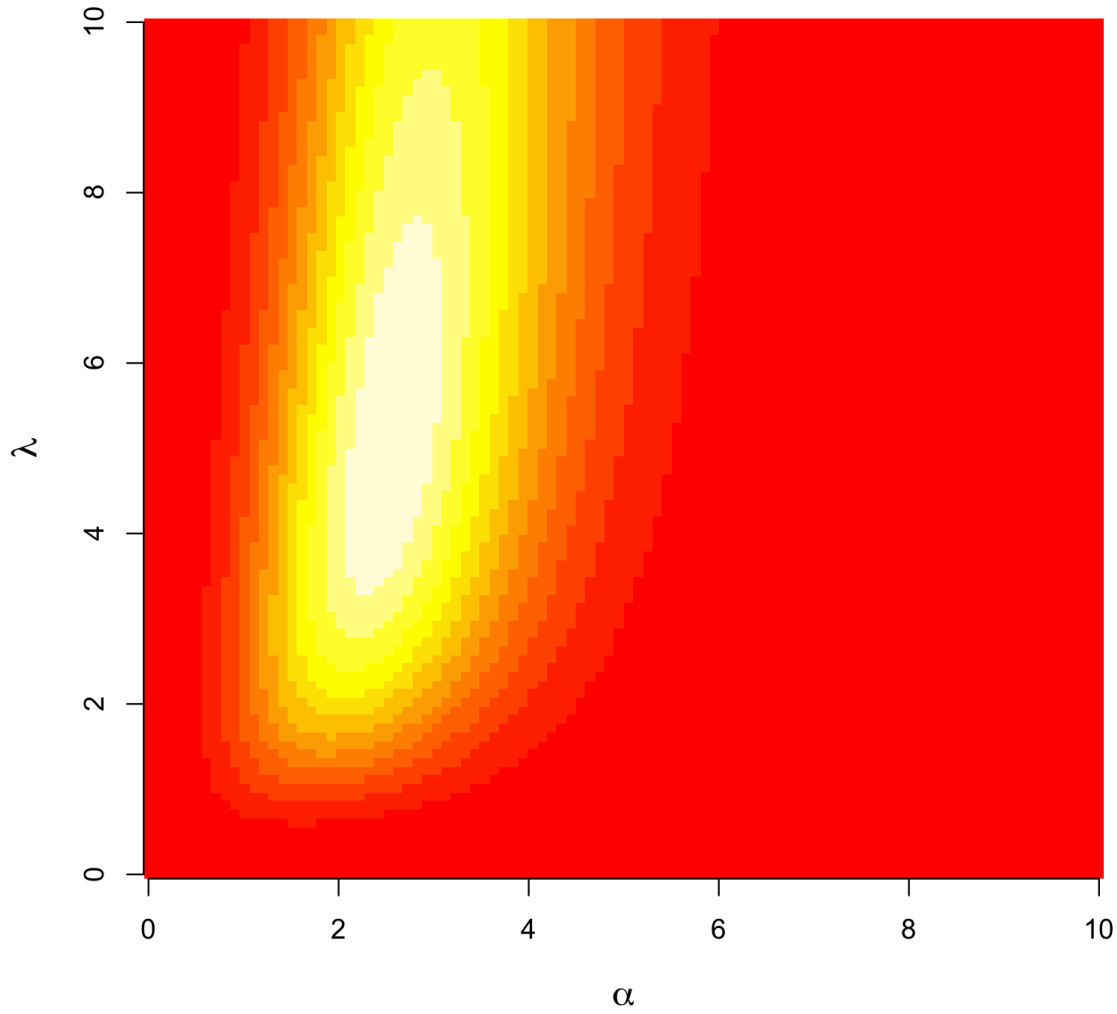
Heatmap approach

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 likeilhood.

```
> 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

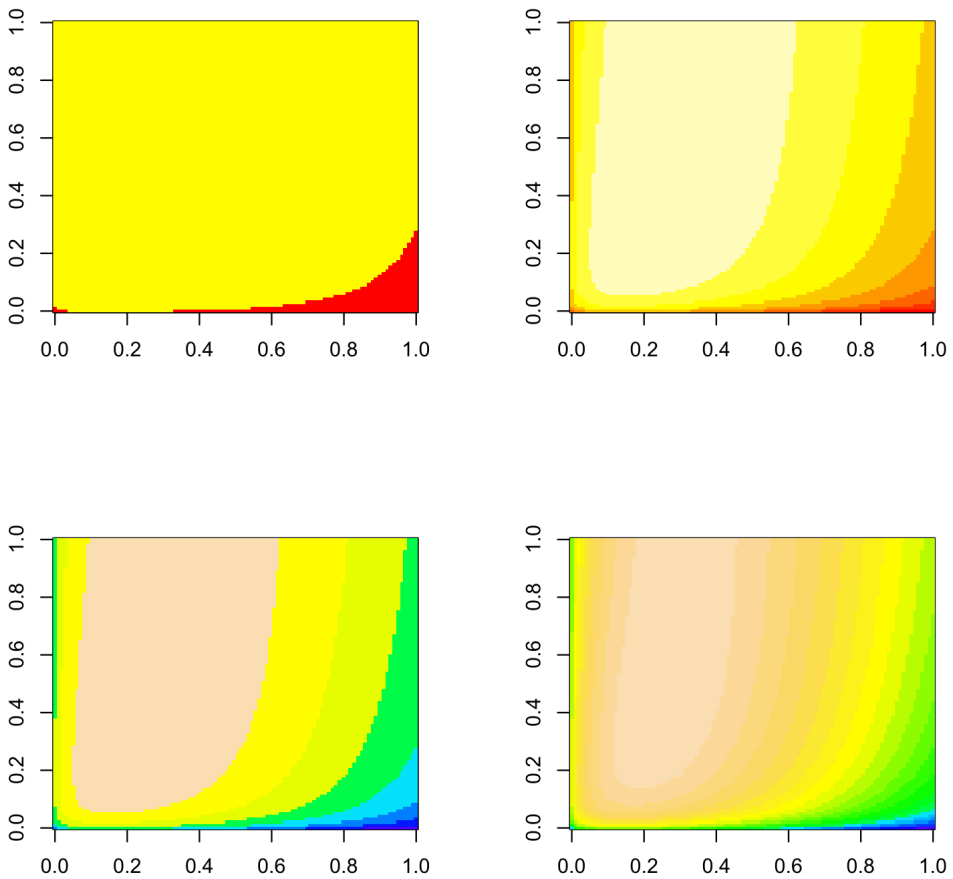


Matrix of likelihood values

There are two ways to encode a matrix of likelihood values. One is a matrix where the ijth component is the likelihood for $\alpha = \alpha_i$ and $\lambda = \lambda_j$. The second is the previous approach where the values of α and λ 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))`



Chapter 4: Nonparametric estimation

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)

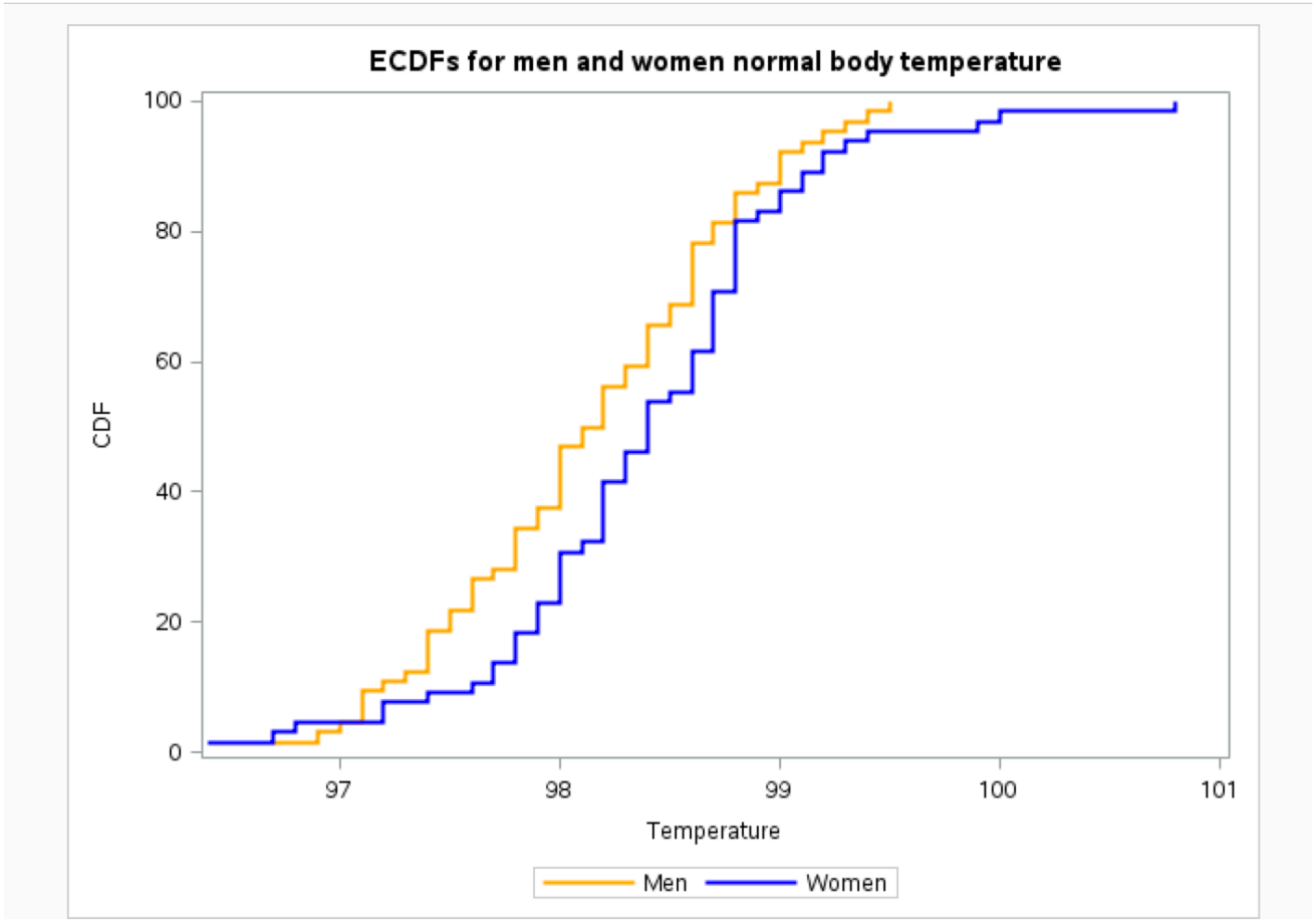
The approach is related to the empirical distribution function that is used in other parts of nonparameteric statistics. Mathematically, the ECDF can be written as

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

where $I(x_i \leq x) = 1$ if $x_i \leq x$ and is otherwise 0. The function is plotted as a ste p 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 $\widehat{F}(3) = \widehat{F}(4) =$ 0.5 because 50% of your observations are less than or equal to both 3 and 4. $\widehat{F}\ (x)$ then jumps to 0.75 at $x = 5.2$.

Two ECDFs


ECDFs for men and women normal body temperature
CDF
100
80
60
40
20
0
97
98
99
100
101
Temperature
Men
Women

Nonparametric survival curve estimation

For survival analysis, we instead want an empirical estimor 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 ith 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.

Nonparametric survival curve estimation

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

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

Recall that t_1 is the earliest observed death.

Kaplan-Meier estimator of the survival function

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.

Kaplan-Meier estimator of the survival function

We have that $\widehat{S}(7.99) = 1$. What is $\widehat{S}(8.0)$?
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

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

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 $\widehat{S}(9)$, 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 $\widehat{S}(8)$. 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 surivival function

To continue the example,

$$\begin{aligned}\widehat{S}(10) &= \prod_{t_i \leq 10} \left[1 - \frac{d_i}{Y_i}\right] \\ &= \left[1 - \frac{d_1}{Y_1}\right]\left[1 - \frac{d_2}{Y_2}\right] \\ &= \left[1 - \frac{1}{5}\right]\left[1 - \frac{1}{4}\right] \\ &= \frac{4}{5} \cdot \frac{3}{4} = \frac{3}{5}\end{aligned}$$

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.

Kaplan-Meier estimator of the survival function

The estimated survival function won't change until $t = 15$. So now we have

$$\begin{aligned}
\widehat{S}(15) &= \prod_{t_i \leq 15} \left[1 - \frac{d_i}{Y_i}\right] \\
&= \left[1 - \frac{d_1}{Y_1}\right]\left[1 - \frac{d_2}{Y_2}\right]\left[1 - \frac{d_3}{Y_3}\right] \\
&= \left[1 - \frac{1}{5}\right]\left[1 - \frac{1}{4}\right]\left[1 - \frac{2}{3}\right] \\
&= \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{5}
\end{aligned}$$

Again, the probabilty 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 $\widehat{S}(t)$, which is denoted $\widehat{V}[\widehat{S}(t)]$. The standard error is the square root of the estimated variance. This allows us to put confidence limits on $\widehat{S}(t)$.

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

$$\widehat{V}[\widehat{S}(t)] = \widehat{S}(t)^2 \sum_{t_l \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$$

Kaplan-Meier example with censoring

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 receall that t_i refers to the ith 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

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

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

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

$$\widehat{S}(15.1) = (0.5)\left[1 - \frac{1}{3}\right] = 0.167$$

Comparison to MLE

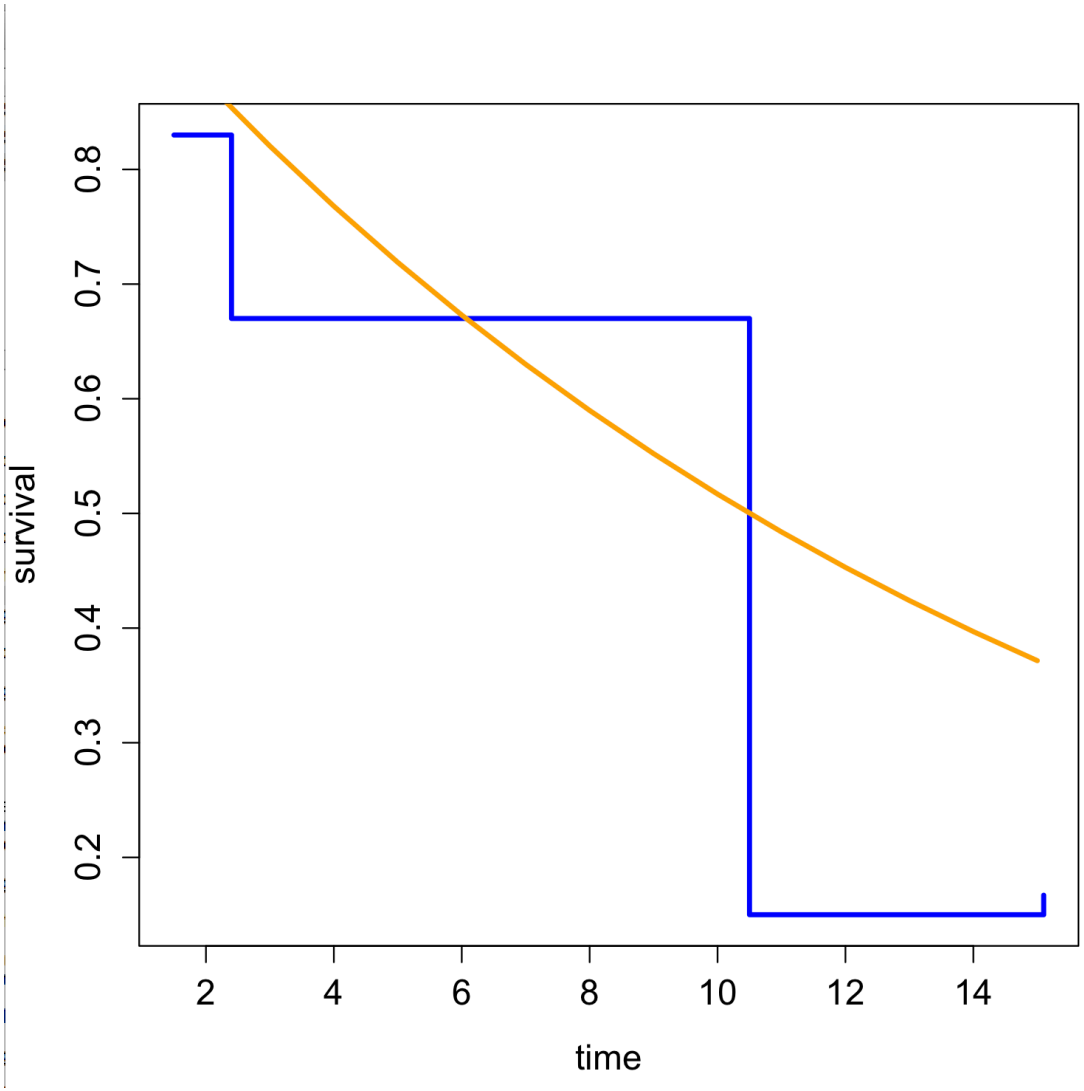
It is interesting to compare to the MLE that we obtained earlier under the exponential model. For the exponential model, we obtained $\widehat{\lambda} = 0.066$. The estimate survival function at the observed death times are

```
> 1-pexp(1.5,rate=.066)
[1] 0.9057427
> 1-pexp(2.4,rate=.066)
[1] 0.8535083
> 1-pexp(10.5,rate=.066)
[1] 0.5000736
> 1-pexp(15.1,rate=.066)
[1] 0.3691324
```

K-M versus exponential

The exponential model predicted higher survival probabilities at the observed death times than Kaplan-Meier except that they both estimate $\widehat{S}(10.5)$ 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, $\widehat{S}(10.0) = 0.67$ 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



Example with lots of censoring

TABLE 1.1
Remission duration of 6-MP versus placebo in children with acute leukemia

Pair	*Remission Status at Randomization*	*Time to Relapse for Placebo Patients*	*Time to Relapse for 6-MP Patients*
1	Partial Remission	1	10
2	Complete Remission	22	7
3	Complete Remission	3	32^+
4	Complete Remission	12	23
5	Complete Remission	8	22
6	Partial Remission	17	6
7	Complete Remission	2	16
8	Complete Remission	11	34^+
9	Complete Remission	8	32^+
10	Complete Remission	12	25^+
11	Complete Remission	2	11^+
12	Partial Remission	5	20^+
13	Complete Remission	4	19^+
14	Complete Remission	15	6
15	Complete Remission	8	17^+
16	Partial Remission	23	35^+
17	Partial Remission	5	6
18	Complete Remission	11	13
19	Complete Remission	4	9^+
20	Complete Remission	1	6^+
21	Complete Remission	8	10^+

$^+$Censored observation

K-M table

TABLE 4.1A

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

Time t_i	*Number of events* d_i	*Number at risk* Y_i	*Product-Limit Estimator* $\hat{S}(t) = \prod_{t_i \le t}[1 - \frac{d_i}{Y_i}]$	$\sum_{t_i \le t} \frac{d_i}{Y_i(Y_i - d_i)}$	$\hat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{Y_i(Y_i - d_i)}$.
6	3	21	$[1 - \frac{3}{21}] = 0.857$	$\frac{3}{21\times18} = 0.0079$	$0.857^2 \times 0.0079 = 0.0058$
7	1	17	$[0.857](1 - \frac{1}{17}) = 0.807$	$0.0079 + \frac{1}{17\times16} = 0.0116$	$0.807^2 \times 0.0116 = 0.0076$
10	1	15	$[0.807](1 - \frac{1}{15}) = 0.753$	$0.0116 + \frac{1}{15\times14} = 0.0164$	$0.753^2 \times 0.0164 = 0.0093$
13	1	12	$[0.753](1 - \frac{1}{12}) = 0.690$	$0.0164 + \frac{1}{12\times11} = 0.0240$	$0.690^2 \times 0.0240 = 0.0114$
16	1	11	$[0.690](1 - \frac{1}{11}) = 0.628$	$0.0240 + \frac{1}{11\times10} = 0.0330$	$0.628^2 \times 0.0330 = 0.0130$
22	1	7	$[0.628](1 - \frac{1}{7}) = 0.538$	$0.0330 + \frac{1}{7\times6} = 0.0569$	$0.538^2 \times 0.0569 = 0.0164$
23	1	6	$[0.538](1 - \frac{1}{6}) = 0.448$	$0.0569 + \frac{1}{6\times5} = 0.0902$	$0.448^2 \times 0.0902 = 0.0181$