

Biased and incomplete samples

The previous chapters considered empirical likelihood based on observations from the distribution (or distributions) of interest. This chapter considers empirical likelihood inference in some nonstandard sampling settings. In biased sampling, the data are sampled from a distribution different from the one we want to study. In censoring, some of the observations are not completely observed, but are known only to belong to a set. The prototypical example is the time until an event. For an event that has not happened by time t , the value is known only to be in (t, ∞) . Truncation is a more severe distortion than censoring. Where censoring replaces a data value by a subset, truncation deletes that value from the sample if it would have been in a certain range. Truncation is an extreme form of biased sampling where certain data values are unobservable.

These incomplete sampling ideas are closely related. They have also been widely studied in varied settings. A lot is known about NPMLE's for incomplete sampling, while there is a comparatively small body of knowledge about the corresponding likelihood ratios.

6.1 Biased sampling

It is common in applied statistics for data to be sampled from a distribution other than the one for which inferences are to be drawn. Sometimes this is an undesirable feature, as with measuring equipment for which the chance of recording a value depends on what that value is. Other times it is an intentional device to gain more informative data, as in retrospective sampling of people with rare diseases, or importance sampling in simulations. Finally, there are settings like the sampling of families by independent sampling of children. There, averages over the sampled families are biased towards larger families, while averages over the sampled children are not biased.

A concrete and common example is length biased sampling. Some methods of sampling cotton fibers, sample them with probability proportional to their length. If one samples people waiting in a hospital room at a random time, those with longer waits, and presumably less serious ailments, are more likely to be in the sample. If one samples entries in an Internet log file, the longer sessions are over-represented.

Suppose that a random variable Y has distribution F_0 , but that we obtain a length biased sample. Let X be one of our observations. Then X has distribution

G_0 with CDF

$$G_0((-\infty, x]) = \frac{\int_0^x y dF_0(y)}{\int_0^\infty y dF_0(y)}.$$

This is a proper distribution if $E(Y)$ is positive and finite.

More generally, suppose that $Y \in \mathbb{R}^d$ has distribution F_0 and that $X \in \mathbb{R}^d$ has distribution G_0 , where for $A \subseteq \mathbb{R}^d$

$$G_0(A) = \frac{\int_A u(y) dF_0(y)}{\int u(y) dF_0(y)},$$

for a biasing function $u(y) \geq 0$, with $0 < \int u(y) dF_0(y) < \infty$. When $0 \leq u(y) \leq 1$, biased sampling has an acceptance sampling interpretation. The value Y is first sampled from F_0 , and then with probability $u(Y)$ it is accepted, while with probability $1 - u(Y)$ this value of Y is rejected. Sampling continues until the first time a Y is accepted. That first accepted Y is the observed value of X . If $u(y) \leq B$ for some $B > 0$ then u -biased sampling gives the same data distribution as v -biased sampling with $v(y) = u(y)/B$, and so the acceptance sampling interpretation carries over to any bounded biasing function u .

The nonparametric likelihood for F is

$$L(F) = \prod_{i=1}^n \frac{F(\{X_i\}) u(X_i)}{\int u(x) dF(x)}.$$

Suppose that $u(x) > 0$ for all x , and let $u_i = u(X_i)$. Then the NPMLE is easily shown to be

$$\hat{F} = C \sum_{i=1}^n \frac{\delta_{X_i}}{u_i}, \quad C^{-1} = \sum_{i=1}^n \frac{1}{u_i}. \quad (6.1)$$

The NPMLE weights each observation in inverse proportion to its sampling probability. In the acceptance sampling setting this has the natural interpretation that every accepted value X represents $1/u(X)$ sampled values of which on average one was accepted. This downweighting is familiar in survey sampling, where it includes stratified sampling and the more general Horvitz-Thompson estimator. It is also well known in Monte Carlo simulation, where the method of importance sampling samples from a distribution other than the nominal one. When $u(x) = c > 0$ for all x , then the NPMLE \hat{F} reduces to the usual NPMLE F_n .

For length biased sampling the NPMLE of the mean of F_0 is

$$\frac{\sum_{i=1}^n u_i^{-1} X_i}{\sum_{i=1}^n u_i^{-1}} = \left(\frac{1}{n} \sum_{i=1}^n X_i^{-1} \right)^{-1}.$$

This is the harmonic mean of the sample.

If $u(x) = 0$ is possible, then the NPMLE is not unique. Any mixture distribution $\alpha H + (1 - \alpha) \hat{F}$, where H puts all its probability on $\{x \mid u(x) = 0\}$, is also an NPMLE, for $0 \leq \alpha < 1$. If there is such a thing as a cotton fiber of zero

length, then F_0 could put probability $\alpha \in [0, 1)$ on such fibers and the distribution of the data would be the same for any value of α . The mean fiber length would be affected by α , and any value between 0 and the harmonic mean would be an NPMLE for $\int x dF(x)$. A pragmatic approach is to fix $\alpha = 0$ and consider any inferences to be on F restricted to the set $\{X \mid u(X) > 0\}$.

By considering the estimating equation

$$0 = \int m(x, \theta) dF(x) = \int \frac{m(x, \theta)}{u(x)} dG(x),$$

we find that we can work directly with biased data, simply by replacing the estimating function $m(x, \theta)$ by $\tilde{m}(x, \theta) \equiv m(x, \theta)/u(x)$. In particular, the NPMLE is the solution $\hat{\theta}$ to

$$\sum_{i=1}^n \tilde{m}(X_i, \theta) = 0,$$

and the profile empirical likelihood ratio function for θ is

$$\mathcal{R}(\theta) = \max \left\{ \prod_{i=1}^n n w_i \mid \sum_{i=1}^n w_i \tilde{m}(X_i, \theta) = 0, \sum_{i=1}^n w_i = 1, w_i \geq 0 \right\}.$$

Tests and confidence regions for θ depend on the distribution of $\tilde{m}(X, \theta)$ under the sample distribution G_0 .

Figure 6.1 displays the widths of 46 shrubs, as reported in Muttlak & McDonald (1990). These shrubs were obtained by transect sampling. Any shrub intersecting a line on the ground was sampled. The probability of a shrub entering the sample is thus proportional to its width. The top histogram shows the observed widths. The bottom histogram shows the data weighted inversely to its sampling probability. The height of each bar is proportional to the sum of $1/X_j$, summed over X_j in the corresponding interval.

The mean μ and variance σ^2 , of the shrub widths are defined by

$$\begin{aligned} 0 &= \int [x - \mu] dF(x), \quad \text{and} \\ 0 &= \int [(x - \mu)^2 - \theta] dF(x), \end{aligned}$$

and so, accounting for the bias, the NPMLE's are defined through

$$\begin{aligned} 0 &= \sum_{i=1}^n X_i^{-1} (X_i - \mu), \quad \text{and} \\ 0 &= \sum_{i=1}^n X_i^{-1} [(X_i - \mu)^2 - \sigma^2]. \end{aligned}$$

Figure 6.2 shows the profile empirical likelihood ratio function for the mean shrub width μ . Figure 6.3 shows the profile empirical likelihood ratio function

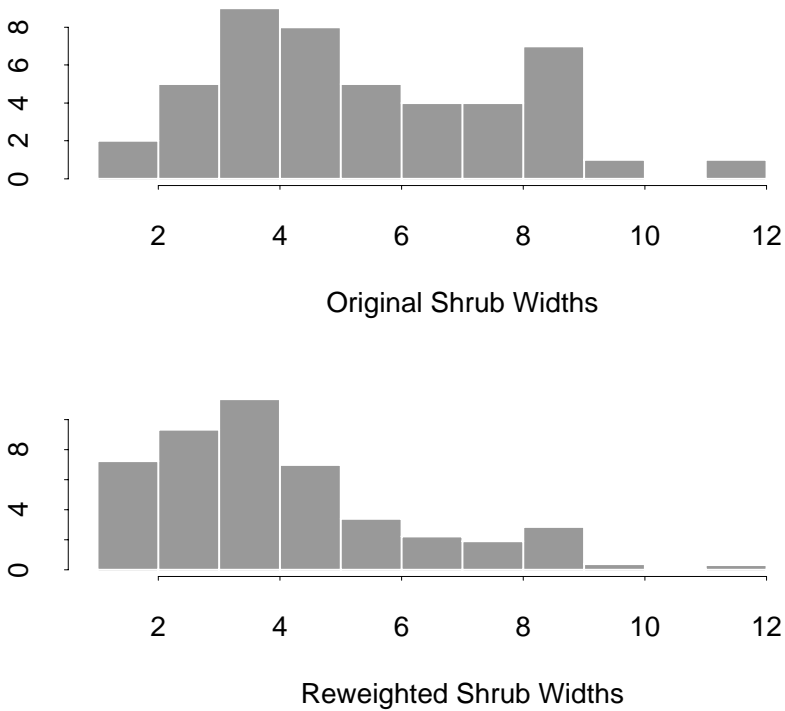


Figure 6.1 The top histogram shows the widths of 46 shrubs found by transect sampling. The bottom histogram has the same total area, and the same bins, but each shrub is weighted inversely to its width to correct for sampling bias.

for the standard deviation σ of shrub width. Taking account of the sampling bias makes a big difference in the inferences, reducing the mean shrub width μ and the standard deviation σ .

6.2 Multiple biased samples

Now suppose that s samples are available, $X_{ij} \in \mathbb{R}^d$, for $i = 1, \dots, s$ and $j = 1, \dots, n_i$. All the observations are independent, but there are s different biases: $X_{ij} \sim G_{i0}$, where

$$G_{i0}(A) = \frac{\int_A u_i(x) dF_0(x)}{\int u_i(x) dF_0(x)}. \quad (6.2)$$

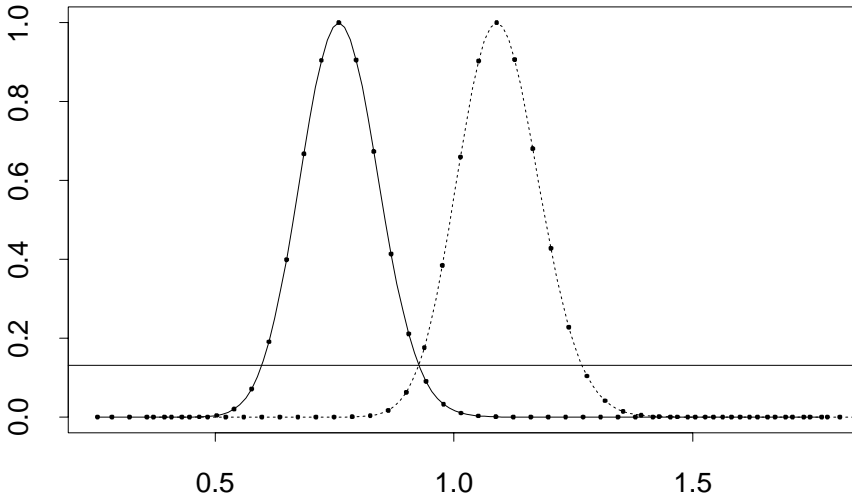


Figure 6.2 The solid curve shows the empirical likelihood ratio for the mean shrub width, after accounting for length biased sampling. The dotted curve does not account for length biased sampling. The horizontal reference line designates a 95% confidence level using an $F_{1,45}$ criterion.

The functions $u_i(x)$ are real valued and nonnegative with

$$0 < \nu_i(F_0) \equiv \int u_i(x)dF_0(x) < \infty. \tag{6.3}$$

Data of this kind could arise in s clinical trials with different enrollment criteria for subjects, or with measurements on the same underlying phenomenon from s different devices. They also arise in choice-based sampling in marketing. When studying the brand preferences of consumers, sample i might correspond to consumers whose brand preferences are known to belong in the i 'th in a list of s subsets of brands.

The NPMLE of F_0 is useful for the data fusion problem of combining these differently biased observations. We will assume that the domain of X is given by $\mathcal{X} = \{x | \sum_{i=1}^s u_i(x) > 0\}$. There can be no data points sampled from outside of this domain. Our inferences are implicitly on F'_0 where $F'_0(A) = F_0(\mathcal{X} \cap A)/F_0(\mathcal{X})$.

Let Z_1, \dots, Z_h be the distinct observations among the sample X_{ij} values, and let $n_{ik} = \#\{X_{ij} = Z_k | 1 \leq j \leq n_i\}$. Let F be a distribution putting probability

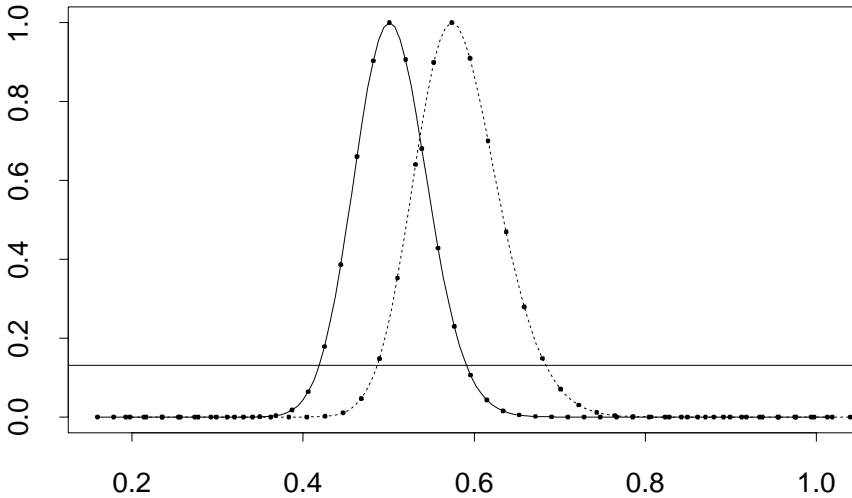


Figure 6.3 The solid curve shows the empirical likelihood ratio for the standard deviation of shrub width, after accounting for length biased sampling. The dotted curve does not account for length biased sampling. The horizontal reference line is for 95% confidence using an $F_{1,45}$ criterion.

$p_k \geq 0$ on Z_k , with $\sum_{k=1}^h p_k = 1$. Then the likelihood for F is

$$\prod_{k=1}^h \prod_{i=1}^s \left(\frac{u_i(Z_k) p_k}{\sum_{l=1}^h p_l u_i(Z_l)} \right)^{n_{ik}}$$

As in the single sample unbiased case of Chapter 2.3, it is possible to ignore ties and work with observation specific weights $w_{ij} \geq 0$ on X_{ij} . The weights generate p_k if

$$p_k = \sum_{i=1}^s \sum_{j=1}^{n_i} w_{ij} 1_{X_{ij}=Z_k}$$

The likelihood in terms of the weights w_{ij} is

$$\prod_{i=1}^s \prod_{j=1}^{n_i} \frac{u_i(X_{ij}) w_{ij}}{\sum_{l=1}^s \sum_{t=1}^{n_l} w_{lt} u_i(X_{lt})} = \prod_{i=1}^s \prod_{j=1}^{n_i} \frac{u_i(X_{ij}) w_{ij}}{\sum_{k=1}^h p_k u_i(Z_k)} \quad (6.4)$$

Notice that the denominator is unaffected by how the probability p_k is allocated among weights w_{ij} for $X_{ij} = Z_k$. It follows that for fixed p_k , the maximizing weights are $w_{ij} = p_k / m_k$ where m_k is the number of observations in the combined samples for which $X_{ij} = Z_k$. Interestingly, the weight w_{ij} does not depend on which sample contributed the value X_{ij} . For any p_k the likelihood in terms

of $w_{ij} = p_k/m_k$ is a constant multiple of the likelihood in terms of p_k , and this constant cancels when forming nonparametric likelihood ratios. The same argument goes through if $\sum_{k=1}^h p_k < 1$, except that then $\nu_i(F)$ cannot be written as a weighted sum of sample values.

The factors $u_i(X_{ij})$ in the numerator of (6.4) do not depend on w_{ij} and so they may be ignored. The log likelihood may be taken to be

$$\sum_{i=1}^s \sum_{j=1}^{n_i} \log(w_{ij}) - \sum_{k=1}^s n_k \log \left(\sum_{i=1}^s \sum_{j=1}^{n_i} w_{ij} u_i(X_{ij}) \right) \quad (6.5)$$

To find the NPMLE we must maximize the expression in (6.5) over $w_{ij} \geq 0$ subject to $\sum_i \sum_j w_{ij} = 1$.

For $s = 1$, the sampling probability of an observation is known, apart from a constant factor, and in Chapter 6.1 we saw that the NPMLE weights the data in inverse proportion to that probability. For $s \geq 2$, matters are more complicated. There is not always an NPMLE, and when an NPMLE exists it is not always unique. But under mild conditions a unique NPMLE exists:

Theorem 6.1 *A unique NPMLE exists if and only if for every proper subset $B \subset \{1, \dots, s\}$*

$$\left(\bigcup_{i \in B} \{X_{i1}, \dots, X_{in_i}\} \right) \cap \left(\bigcup_{i \notin B} \{X \mid u_i(X) > 0\} \right) \neq \emptyset.$$

Proof. Vardi (1985). \square

In words: there will not be a unique NPMLE if the s sample data sets can be partitioned into two subsets, where no observation in the first subset could possibly have been observed in the second subset. This does not mean that each pair of sampling distributions has to overlap. A pair that does not overlap might be bridged by a third sample. If even one of the samples, say sample i , has $u_i(X) > 0$ for all X , then a unique NPMLE will exist.

To illustrate why the NPMLE might not be unique, take $s = 2$, let $\mathcal{X}_i = \{x \mid u_i(x) > 0\}$ for $i = 1, 2$ and suppose that $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$. If the distribution of X_{ij} from G_{i0} is consistent with acceptance sampling from F_0 with acceptance probability $u_i(x)$ then it is equally consistent with an acceptance probability of $u_i(x)/100$. Such samples, even as $n_i \rightarrow \infty$, cannot help us determine the relative weight $F(\mathcal{X}_i)/(F(\mathcal{X}_1) + F(\mathcal{X}_2))$ that belongs on the domain of sample i . If however $\mathcal{X}_1 \cap \mathcal{X}_2 = Z$ and $\Pr(X_{ij} \in Z) > 0$ for both $i = 1, 2$, then we may be able to estimate $F(Z)/F(\mathcal{X}_i)$ from the X_{ij} and put these together to estimate $F(\mathcal{X}_i)/(F(\mathcal{X}_1) + F(\mathcal{X}_2))$. It is possible to have a degenerate NPMLE with $F(\mathcal{X}_i) = 1$ if sample i has no observations in Z .

For $s > 2$, if two samples intersect, then we can estimate the relative probabilities of their domains \mathcal{X}_i . If we can estimate the ratios $F(\mathcal{X}_i)/F(\mathcal{X}_{i'})$ and $F(\mathcal{X}_{i'})/F(\mathcal{X}_{i''})$ then we can estimate $F(\mathcal{X}_i)/F(\mathcal{X}_{i''})$. We can estimate the relative probabilities of other domains if there is a chain of ratios connecting them.

Theorem 6.2 Suppose that F_0 is a distribution \mathbb{R}^d , that for $i = 1, \dots, s$, the functions u_i satisfy (6.3), and that distributions G_{i0} are defined by (6.2). Let \mathcal{G} be a graph on s vertices with an edge connecting vertices i and i' if and only if $\int u_i(x)u_{i'}(x)dF_0 > 0$. Then F_0 is uniquely determined as a function of G_{i0} for $i = 1, \dots, s$, if and only if the graph \mathcal{G} is connected. In that case the probability that a unique NPMLE exists tends to 1 as $\min_i n_i \rightarrow \infty$.

Proof. Gill, Vardi & Wellner (1988). \square

To maximize the nonparametric likelihood, suppose for a moment that we know the values $\nu_i = \int u_i(X)dF_0(X)$. Let $N = \sum_{i=1}^s n_i$, $\nu = (\nu_1, \dots, \nu_s)$, and $U_{ij} = (u_1(X_{ij}), \dots, u_s(X_{ij}))$. Then let $\ell(\nu)$ maximize $\sum_i \sum_j \log(w_{ij}) - \sum_i n_i \log(\nu_i)$ subject to

$$\sum_i \sum_j w_{ij} = 1, \quad \text{and}$$

$$\sum_i \sum_j w_{ij} U_{ij} = \nu.$$

The computation of $\ell(\nu)$ reduces to empirical likelihood maximization for a mean as described in Chapter 3.14, and the solution has

$$w_{ij}(\nu) = \frac{1}{N} \frac{1}{1 + \delta'(U_{ij} - \nu)}$$

where the Lagrange multiplier $\delta = \delta(\nu)$ satisfies

$$\sum_i \sum_j \frac{U_{ij} - \nu}{1 + \delta'(U_{ij} - \nu)} = 0.$$

To find the NPMLE we maximize

$$\ell(\nu) = \sum_i \sum_j \log(w_{ij}(\nu)) - \sum_i n_i \log(\nu_i)$$

over $\nu = (\nu_1, \dots, \nu_s)$.

To incorporate estimating equations, we must now impose the additional constraint

$$\sum_{i=1}^s \sum_{j=1}^{n_i} w_{ij} m(X_i, \theta) = 0.$$

Given ν , the solution is

$$w_{ij}(\nu, \theta) = \frac{1}{N} \frac{1}{1 + \lambda' m(X_{ij}, \theta) + \delta'(U_{ij} - \nu)}$$

where $\lambda(\nu, \theta)$ and $\delta(\nu, \theta)$ satisfy

$$\sum_i \sum_j \frac{U_{ij} - \nu}{1 + \lambda' m(X_{ij}, \theta) + \delta'(U_{ij} - \nu)} = 0$$

$$\sum_i \sum_j \frac{m(X_{ij}, \theta)}{1 + \lambda' m(X_{ij}, \theta) + \delta'(U_{ij} - \nu)} = 0.$$

The profile empirical likelihood ratio for θ is then the ratio

$$\mathcal{R}(\theta) = \frac{\max_{\nu} \prod_i \prod_j w_{ij}(\nu, \theta)}{\max_{\nu} \prod_i \prod_j w_{ij}(\nu)}.$$

Under mild conditions, the asymptotic distribution of $-2 \log(\mathcal{R}(\theta_0))$ is $\chi_{(p)}^2$ where p is the dimension of θ . See the references in Chapter 6.9.

6.3 Truncation and censoring

Truncation is an extreme version of biased sampling, where the bias function is

$$u(x) = \begin{cases} 1 & \text{if } x \in T \\ 0 & \text{if } x \notin T, \end{cases} \quad (6.6)$$

for some set T . Consider historical data on the heights of men drafted into an army, where a minimum height restriction H was in effect. For conclusions on the heights of draft-aged men, such data represent a sample truncated to $T = [H, \infty)$.

Censoring is a milder form of information loss than truncation. An observation is censored to the set C if instead of observing X we only observe the fact that $X \in C$. A censored point is known to have existed, whereas a truncated point produces no observation. In the example above, if we knew the number of draft-aged men rejected because of the height restriction, then their heights would be censored to the set $C = (0, H)$.

Suppose θ is a quantity that depends in part on how a truncated random variable X is distributed over values $X \notin T$. Then some assumptions are necessary to get an estimate of θ . The truncated data can never contain any $x \notin T$, but perhaps there is a way to extrapolate from $x \in T$ to $x \notin T$. In practice our extrapolation might introduce a systematic error that we can neither check nor correct. Yet it may be better to patch in a possibly flawed extrapolation than to ignore the truncation completely.

One way to extrapolate is to fit a parametric model for X with density or mass function $f(x; \theta)$. Our sample is from $X \sim f(x; \theta)$ conditional on $X \in T$. For such a model, the likelihood is

$$L_{\text{TRUN}}(\theta; X_1, \dots, X_n) = \prod_{i=1}^n \frac{f(X_i; \theta)}{\int_T dF(x; \theta)},$$

with estimating equations

$$\sum_{i=1}^n \left(\frac{\frac{\partial}{\partial \theta} f(X_i; \theta)}{f(X_i; \theta)} - \frac{\frac{\partial}{\partial \theta} \int_T dF(x; \theta)}{\int_T dF(x; \theta)} \right) = 0. \quad (6.7)$$

Both parametric and empirical likelihood inferences can be based on (6.7). Both inference methods give the same maximum likelihood estimate $\hat{\theta}$. Whether $\hat{\theta}$ is estimating the desired quantity can depend on how accurate the parametric model is. Empirical likelihood confidence regions will have the right asymptotic coverage for the quantity being estimated by $\hat{\theta}$ under very weak conditions, whereas parametric likelihood regions will have coverage levels sensitive to the parametric model used.

Now suppose that we have two parametric models $f_j(x; \theta_j)$ for $j = 1, 2$ with $\theta_j \in \mathbb{R}^{p_j}$, and that the quantity we are interested in can be written as $\tau_j(\theta_j)$ under model j . For example f_1 might be a normal distribution and f_2 might be a gamma distribution. The values $\tau_j(\hat{\theta}_j)$ will not in general agree with each other. The nature of the discrepancy can be investigated using an empirical likelihood confidence region for $(\tau_1(\theta_1), \tau_2(\theta_2))$ or for $\tau_1(\theta_1) - \tau_2(\theta_2)$. This will not indicate which, if either, of the parametric models provides a reliable extrapolation. But it does allow us to judge whether the extrapolated answer is sensitive to the extrapolation formula, without knowing which, if either, parametric model is right.

In applications with censored and truncated data, the nature of the censoring and truncation rules may vary from observation to observation. The general case has X_i truncated to T_i then censored to C_i , taking $C_i = \{X_i\}$ for uncensored data, and T_i equal to the domain of X_i , usually a subset of \mathbb{R}^d , for untruncated data. The sets T_i may be random. Apart from trivial exceptions, the set C_i has to be random, because it depends on X_i . We consider coarsening at random (CAR), in which T_i has been partitioned at random and independently of Y_i into a number of sets. The set that happens to contain X_i is observed as C_i .

Some of the most widely studied types of censoring are listed below. Of these, [Examples 6.1](#) and [6.2](#) will be considered at greater length. We will find that a form of conditional likelihood is most suitable for them.

Example 6.1 (Right censoring) Here, the distribution of the real-valued random variables X_i is of direct interest. For each X_i there is a $Y_i \in \mathbb{R}$. This Y_i may be random. If $X_i \leq Y_i$ we observe X_i , otherwise X_i is censored to (Y_i, ∞) . We say that X_i is right censored by Y_i . For example, X_i could be survival time after an operation, with Y_i the time from the operation to the end of the study.

Example 6.2 (Left truncation) The pair (X_i, Y_i) is observed if and only if $X_i \geq Y_i$. The Y_i may be random. We say that X_i is left truncated by Y_i . In astronomy, the brightness X of an object may be left truncated by some function $Y = h(Z)$ of its distance Z from Earth.

Example 6.3 (Left truncation and right censoring) The random variable X_i is right censored by Y_i , and left truncated by $Z_i < Y_i$. If $Z_i \leq X_i < Y_i$ then X_i is observed directly. If $Z_i < Y_i \leq X_i$ then X_i is censored to (Y_i, ∞) and if $Z_i > X_i$ then none of X_i , Y_i or Z_i are observed. For example, X_i could be the survival time after a transfusion, Y_i a corresponding censoring time, and Z_i the time between the transfusion and the beginning of a study of transfused patients.

Example 6.4 (Double censoring) The random variable $X_i \in \mathbb{R}$ is observed if $Z_i \leq X_i \leq Y_i$, is right censored to (Y_i, ∞) if $X_i > Y_i$, and is left censored to $(-\infty, Z_i)$ if $X_i < Z_i$. Here $Z_i \leq Y_i$ and either or both may be random.

Example 6.5 (Interval censoring) The random variable X_i is censored to the set $(Z_{i,k}, Z_{i,k+1}]$ for $Z_{i,1} < Z_{i,2} < \dots < Z_{i,K_i}$. For example, $Z_{i,k}$ could be times at which patients are studied or equipment is inspected, and X_i the time of some change in status. Interval-censored data are also known as current status data. The usual likelihood for data from a continuous parametric distribution is motivated by arguing that each component of each observation was interval censored to a small interval.

Suppose that X_1, \dots, X_n are sampled from a common distribution F , and are conditionally independent given right censoring times Y_1, \dots, Y_n . Let $Z = \min(X, Y)$ and let $\delta = 1_{X \leq Y}$ indicate an uncensored failure. By convention, X is not considered censored when $X = Y$. Similarly, if several observations are tied at the same value of Z , the censoring times are deemed to follow the failure times by an infinitesimal amount.

Let $\mathcal{X} = (X_1, \dots, X_n)$ and $\mathcal{Y} = (Y_1, \dots, Y_n)$. The likelihood for F and G from right-censored data is the product of a marginal and conditional likelihood

$$L(F, G; \mathcal{X}, \mathcal{Y}) = L(F, G; \mathcal{Y}) \times L(F, G; \mathcal{X} \mid \mathcal{Y}) \quad (6.8)$$

where $L(F, G; \mathcal{Y}) = G(Y_1, \dots, Y_n)$ and

$$\begin{aligned} L(F, G; \mathcal{X} \mid \mathcal{Y}) &= \prod_{i:\delta_i=1} F(\{X_i\}) \prod_{i:\delta_i=0} F((Y_i, \infty)) \\ &= \prod_{i=1}^n F(\{Z_i\})^{\delta_i} F((Z_i, \infty))^{1-\delta_i}. \end{aligned} \quad (6.9)$$

Any factor of 0^0 in (6.9) is understood to be 1. These likelihoods are nonparametric, but are easily modified if F or G are known to belong to parametric families.

It is usual to base inferences for F on the conditional likelihood (6.9). That conditional likelihood does not depend on G , and it can be computed from the Z_i and δ_i without knowing the Y_i from uncensored X_i . Because the marginal likelihood of the Y_i does not depend on F , using the conditional likelihood does not lead to a loss of information on F . In the absence of a known functional relationship between F and G , the conditional likelihood (6.9) gives the same likelihood ratio

function for F as the full likelihood (6.8). We did not need to assume that the Y_i were independent in order to settle on the conditional likelihood. Of course strong dependence in Y_i could make the conditional likelihood very uninformative.

For left-truncated data, we never observe $X < Y$, but if Y is independent of X then we may reasonably hope to learn the distribution of X , or at least that part of it larger than the left end point of the Y distribution. A conditional likelihood approach is also applicable to left-truncated data, though the derivation is more complicated.

Suppose that X and Y are independent from distributions F and G , respectively, and that independent (X, Y) pairs are truncated to the set $\{(x, y) \mid x \geq y\}$. The likelihood is then

$$L(F, G; \mathcal{X}, \mathcal{Y}) = \alpha^{-n} \prod_{i=1}^n 1_{X_i \geq Y_i} F(\{X_i\}) G(\{Y_i\}) \quad (6.10)$$

where

$$\alpha = \iint_{x \geq y} dF(x) dG(y) = \int G((-\infty, u]) dF(u) = \int F([u, \infty)) dG(u)$$

is the probability that $X \geq Y$.

The likelihood (6.10) may be factored into a product of marginal and conditional likelihoods by either of

$$\begin{aligned} L(F, G; \mathcal{X}, \mathcal{Y}) &= L(F, G; \mathcal{X}) \times L(F, G; \mathcal{Y} \mid \mathcal{X}) \\ &= L(F, G; \mathcal{Y}) \times L(F, G; \mathcal{X} \mid \mathcal{Y}) \end{aligned}$$

where

$$\begin{aligned} L(F, G; \mathcal{X}) &= \alpha^{-n} \prod_{i=1}^n G((-\infty, X_i]) F(\{X_i\}) \\ L(F, G; \mathcal{Y} \mid \mathcal{X}) &= \prod_{i=1}^n 1_{X_i \geq Y_i} \frac{G(\{Y_i\})}{G((-\infty, X_i])} \\ L(F, G; \mathcal{Y}) &= \alpha^{-n} \prod_{i=1}^n F([Y_i, \infty)) G(\{Y_i\}) \\ L(F, G; \mathcal{X} \mid \mathcal{Y}) &= \prod_{i=1}^n 1_{X_i \geq Y_i} \frac{F(\{X_i\})}{F([Y_i, \infty))}. \end{aligned}$$

Suppose that interest centers on F . The conditional likelihood based on X_i given Y_i depends on F but not on G , and hence is available for inference. Unlike the case of right censoring, the marginal distribution of Y_i depends on F , so that even without a known link between F and G , there may be an information loss from using the conditional likelihood. The conditional distribution of Y_i given X_i does not involve F , suggesting that the marginal distribution of X_i has all the information on F , but the marginal distribution of X_i involves G .

If G is known, then the full likelihood (6.10) for F may be written

$$L(F) = \prod_{i=1}^n \frac{1_{X_i \geq Y_i} F(\{X_i\}) G(\{Y_i\})}{\int G((-\infty, u]) dF(u)} \\ \propto \prod_{i=1}^n \frac{F(\{X_i\})}{\int G((-\infty, u]) dF(u)}. \quad (6.11)$$

The likelihood (6.11) is proportional to the likelihood of biased sampling with $u(x) = G((-\infty, u])$ and so the NPMLE for F with G known is given by (6.1) with $u_i = G((-\infty, X_i])$. The following theorem lends support to the use of conditional likelihood for F when the distribution G is not known.

Theorem 6.3 *Suppose that X_i is observed with independent left truncation by Y_i . In the joint NPMLE (\hat{F}, \hat{G}) of (F, G) , the distribution \hat{F} is the maximizer of the conditional likelihood $L(F, G; \mathcal{X} \mid \mathcal{Y})$ above.*

Proof. Wang (1987) and Keiding & Gill (1990). \square

6.4 NPMLE's for censored and truncated data

For $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ independently sampled from F and censored, by coarsening at random, to C_i , including $C_i = \{X_i\}$ for uncensored data, the conditional (on C_i) likelihood may be written

$$L_c(F) = \prod_{i=1}^n \int_{C_i} dF(x) = \prod_{i=1}^n F(C_i), \quad (6.12)$$

and for truncated and censored data the conditional likelihood is

$$L_c(F) = \prod_{i=1}^n \frac{\int_{C_i} dF(x)}{\int_{T_i} dF(x)} = \prod_{i=1}^n \frac{F(C_i)}{F(T_i)}. \quad (6.13)$$

The censored data likelihood (6.12) does not always have a unique maximum. There are 2^n disjoint sets of the form

$$E_j = \bigcap_{i=1}^n D_{ij}$$

where each D_{ij} is either C_i or $C_i^c = \mathcal{X} - C_i$. The union of these E_j is \mathcal{X} . Letting $w_j = F(E_j)$, $M = 2^n$, and $H_{ij} = 1_{E_j \subset C_i}$ we can write the conditional likelihood from (6.12) as

$$L_c(F) = \prod_{i=1}^n \left(\sum_{j=1}^M H_{ij} w_j \right). \quad (6.14)$$

Thus F is determined only up to the values of w_j .

Theorem 6.4 *There is a unique set of weights $w_j \geq 0$ with $\sum_{j=1}^M w_j = 1$ that maximize (6.14).*

Proof. Let $w = (w_1, \dots, w_M)'$ belong to the closed, bounded, and convex set $S = \{w \mid w_j \geq 0, \sum_{j=1}^M w_j = 1\}$. After identifying $F(E_j)$ with w_j , the function $L_c(w)$ is a continuous function on the compact set S , so it attains a maximum there. This maximum is nonzero, and so $\ell(w) = \log(L_c)$ also attains a finite maximum ℓ_m on S . It remains to show that this maximum is unique.

Suppose to the contrary that $\ell(u) = \ell(v) = \ell_m$ and that $u_{j'} > v_{j'}$ for some j' . If there is no i' with $H_{i'j'} = 1$, then taking $w_{j'} = 0$ and $w_j = u_j/(1 - u_{j'})$ for $j \neq j'$ we find $\ell(w) > \ell(u)$ contradicting the maximality of $\ell(u)$. So let i' satisfy $H_{i'j'} = 1$ and put $w = (u + v)/2$. Convexity of S implies that $w \in S$. Now $\ell(w) - \ell_m = \ell(w) - (\ell(u) + \ell(v))/2$ is a sum of n nonnegative terms, one for each i . The term for i' is strictly positive, contradicting the maximality of ℓ_m .

□

A censored-data NPMLE is not necessarily a good estimator of F . See Chapter 6.9 for an example with bivariate censoring and for a remedy.

It is not practical to keep track of 2^n probability weights. For uncensored data, at most n of the E_j are nonempty. As the next two theorems show, a great simplification occurs when X_i are real values and C_i are all intervals, with or without truncation.

Theorem 6.5 *Let $C_i = [L_i, R_i]$, and let $E_j = [p_j, q_j]$ for $j = 1, \dots, m$ be the set of intervals with endpoints taken from $U = \cup_{i=1}^n \{L_i, R_i\}$ and that contain no interior points from U . Then there are uniquely determined probabilities $w_j \geq 0$ on E_j with $\sum_{j=1}^m w_j = 1$ such that F maximizes L if and only if $F(E_j) = w_j$.*

Proof. Peto (1973). □

Theorem 6.6 *Let $X_i \in \mathbb{R}$ be truncated to $T_i \subseteq \mathbb{R}$ and then censored to C_i , a finite union of disjoint closed intervals $[L_{ik}, R_{ik}]$, $k = 1, \dots, K_i$. Let $E_j = [p_j, q_j]$ for $j = 1, \dots, m$ be the set of intervals with endpoints taken from $U = \cup_{i=1}^n \cup_{k=1}^{K_i} \{L_{ik}, R_{ik}\}$ and that contain no interior points from U . Let $D = \cup_{j=1}^m E_j$. Then:*

1. Any NPMLE F has $F(D) = 1$, unless $T_i \cap D = C_i \cap D$ for all i .
2. The likelihood depends on F only through $w_j = F(E_j)$, $j = 1, \dots, m$.
3. The likelihood is

$$L_c(F) = \prod_{i=1}^n \frac{\sum_{j=1}^m H_{ij} w_j}{\sum_{j=1}^m K_{ij} w_j}$$

where $H_{ij} = 1_{E_j \subseteq C_i}$ and $K_{ij} = 1_{E_j \subseteq T_i} \geq H_{ij}$.

4. There are unique NPMLE weights w_j , unless

- (a) $H_{ij} = H_{ij'}$, for some $1 \leq j < j' \leq m$, and all $i = 1, \dots, n$, or
- (b) There is a subset R with $C_i \cap D \subset R$ or $C_i \cap D \subset R^c$ for all i .

Proof. The results above are collected from Turnbull (1976). \square

If 4(a) above happens, the non-uniqueness is that the sum $w_j + w_{j'}$ is determined but not w_j itself. Condition 4(b) is like the graph condition in [Theorem 6.2](#). When it happens, $F(R)$ and $F(R^c)$ cannot be determined.

Another simplification can be achieved if we restrict attention to those distributions with a given set of support points. In the sieved NPMLE, we take a list of points $x_1, \dots, x_{n'} \in \mathcal{X}$, including at least one element from each C_i . If there are indices $j \neq j'$ with $x_{j'} \in C_i$ whenever $x_j \in C_i$ then remove any one such x_j from the list. We repeat this removal process until no more points can be removed, and relabel the remaining points as x_1, \dots, x_m . The sieved likelihood, for censored but not truncated data, is $L_{\text{SIEVE}}(F) = \prod_i F(C_i) = \prod_i \sum_{j=1}^m 1_{x_j \in C_i} F(\{x_j\})$.

Theorem 6.7 *Given points x_1, \dots, x_m , there is a unique sieved-NPMLE maximizing $L_{\text{SIEVE}}(F)$.*

Proof. van der Laan (1995, Chapter 3.3). \square

6.5 Product-limit estimators

For real-valued X subject to right censoring or left truncation, there is an explicit closed form for the NPMLE. These NPMLE's are more conveniently derived through the hazard function, defined below, than through the cumulative distribution function.

The survival function is $S(t) = F([t, \infty)) = 1 - F((-\infty, t])$. It is widely used in medical applications, where it is natural to consider the fraction of subjects surviving past time t . For continuous distribution functions with density f , the hazard function is defined as

$$\lambda(t) = \lim_{\varepsilon \rightarrow 0^+} \frac{1}{\varepsilon} \Pr(X \leq t + \varepsilon \mid X \geq t) = \frac{f(t)}{S(t+)} = \frac{f(t)}{S(t)}.$$

The product $\lambda(t)dt$ gives the probability of failure before time $t + dt$, conditional on surviving at least to time t . For continuously distributed data $\lambda(t) = -d \log(S(t))/dt$, and so $S(t) = \exp(-\int_0^t \lambda(u)du)$.

For discrete distributions with $F(\{t_j\}) > 0$, for a finite (or countably infinite) number of t_j , the hazard function is defined as

$$\lambda_j = \Pr(X = t_j \mid X \geq t_j) = \frac{F(\{t_j\})}{S(t_j)}.$$

For discrete distributions $S(t) = \prod_{t_j \leq t} (1 - \lambda_j)$, $F(\{t_i\}) = \lambda_j \prod_{t_j < t_i} (1 - \lambda_j)$, and $F((-\infty, t]) = 1 - \prod_{t_j \leq t} (1 - \lambda_j)$.

The cumulative hazard is defined as

$$\Lambda(t) = \int_0^t \frac{dF(u)}{F((-\infty, u))},$$

which simplifies to $\int_0^t \lambda(u) du$ for continuous distributions, to $\sum_{t_j \leq t} \lambda_{t_j}$ for discrete distributions, and to a sum of discrete and continuous hazards for distributions with discrete and continuous parts.

For right-censored data, the NPMLE \hat{F} must have positive probability at each observed failure time. Let the observed failure times be $t_1 < t_2 < \dots < t_k$, suppose that $t_1 > 0 \equiv t_0$, and define $t_{k+1} = \infty$. Let $d_j \geq 1$ be the number of failures at t_j and suppose that m_j observations were censored in the interval $[t_j, t_{j+1})$. The NPMLE puts 0 probability inside the interval (t_i, t_{i+1}) for $i < k$, because moving such probability to t_{i+1} would increase at least one factor in L_c and would not decrease any of them. The number

$$r_j = (d_i + m_i) + \dots + (d_k + m_k)$$

denotes the number of subjects at risk of failure just prior to t_j .

Let $\lambda_j = \Pr(X = t_j \mid X \geq t_j)$ denote the hazard probabilities of the distribution F . The conditional likelihood given by (6.9) or (6.12) may be written

$$L_c(F) = \prod_{j=1}^k \lambda_j^{d_j} (1 - \lambda_j)^{r_j - d_j},$$

and so the NPMLE has $\hat{\lambda}_j = d_j/r_j$. The CDF of the NPMLE may be written

$$F((-\infty, t]) = 1 - \prod_{j|t_j \leq t} \frac{r_j - d_j}{r_j} \quad (6.15)$$

This is the celebrated Kaplan-Meier product-limit estimator.

If the largest observed failure time is greater than the largest observed censoring time, then the NPMLE is unique. Otherwise $F((-\infty, t_k]) < 1$ and any distribution that satisfies (6.15) for $t \leq t_k$ is also an NPMLE. A common convention to force uniqueness is to place probability $1 - F((-\infty, t_k])$ on the largest observed censoring time when that time is larger than t_k .

There is also a product-limit estimator for left truncation of X by an independent Y . Let F and G be distributions of X and Y , respectively. Let a_G and b_G be the smallest and largest observable values of Y . Formally $a_G = \inf\{y \mid G((y, \infty)) < 1\}$, and $b_G = \sup\{y \mid G((y, \infty)) > 0\}$. The values a_F and b_F are defined similarly. If $a_G > a_F$, then the lower end of the F distribution cannot be observed. We assume that either $a_G \leq a_F$, or that we are satisfied with inferences on $\Pr(X \leq t)/\Pr(X \geq a_G)$. We also assume that $b_G \leq b_F$. If $b_G > b_F$, then the upper end of the G distribution cannot be observed.

The assumption of independence between X and Y is often reasonable, but is not to be made lightly. In astronomy this assumption follows from a simplifying idea that, at very large scales, space is the same everywhere and in every direction (the cosmological principle). This independence is thought to be nearly, though perhaps not exactly, correct. We suppose that G is unknown, and so we use the conditional likelihood based on the distribution of X given Y .

It is convenient to work in terms of ordered observations $X_{(1)} \leq X_{(2)} \leq \dots \leq$

$X_{(n)}$. Let $Y_{(i)}$ denote their concomitants, that is $(X_{(i)}, Y_{(i)})$ for $i = 1, \dots, n$ are the points (X_i, Y_i) , $i = 1, \dots, n$, after sorting on X_i . Of course, the $Y_{(i)}$ are not necessarily in increasing order. To simplify the derivation, we suppose that there are no i and j for which $X_i = Y_j$. Candidates for the NPMLE \hat{F} put nonnegative probability on every observed value $X_{(i)}$ and put no probability anywhere else. For such distributions, $F((Y_{(i)}, \infty)) = F((Z_i, \infty))$, where $Z_i = \max\{X_{(j)} \mid X_{(j)} < Y_{(i)}\}$. The conditional likelihood is

$$L_c(F) = \prod_{i=1}^n \frac{F(\{X_{(i)}\})}{F((Z_i, \infty))} = \prod_{i=1}^n \frac{F(\{X_{(i)}\})}{F((X_{(i)}, \infty))^{K_i}},$$

where

$$\begin{aligned} K_i &= \#\{j \mid Z_j = X_{(i)}\} \\ &= \#\{j \mid X_{(i)} < Y_{(j)} \leq X_{(i+1)}\}, \end{aligned}$$

with $X_{(n+1)} = \infty$.

Writing in terms of $\lambda_i = F(\{X_{(i)}\})/F((X_{(i)}, \infty))$,

$$\begin{aligned} L_c(F) &= \prod_{i=1}^n \frac{\lambda_i \prod_{j=1}^{i-1} (1 - \lambda_j)}{\left[\prod_{j=1}^i (1 - \lambda_j)\right]^{K_i}} \\ &= \prod_{i=1}^n \lambda_i (1 - \lambda_i)^{n-i - \sum_{j=i}^n K_j}. \end{aligned}$$

This is maximized by values

$$\begin{aligned} \hat{\lambda}_i &= \frac{1}{n - i + 1 - \sum_{j=i}^n K_j} \\ &= \frac{1}{\#\{j \mid X_{(j)} \geq X_{(i)}\} - \#\{j \mid Y_{(j)} > X_{(i)}\}} \\ &= \frac{1}{\#\{j \mid Y_{(j)} < X_{(i)} \leq X_{(j)}\}}, \end{aligned}$$

so that

$$\hat{F}((-\infty, t]) = 1 - \prod_{i=1}^n \left(1 - \frac{1_{X_i \leq t}}{\sum_{l=1}^n 1_{Y_l < X_i \leq X_l}}\right). \quad (6.16)$$

Equation (6.16) is known as the Lynden-Bell estimator. The Lynden-Bell estimator can be degenerate: $\hat{F}((-\infty, X_{(i)}]) = 1$ for some $i < n$ is possible.

6.6 EL for right censoring

Table 6.1 presents the AML data. Most of the values indicate the time until relapse of a patient whose leukemia has gone into remission. Those values with a + sign designate right-censored times.

Maintained	9	13	13+	18	23	28+
	31	34	45+	48	161+	
Non-Maintained	5	5	8	8	12	16+
	23	27	30	33	43	45

Table 6.1 Shown are the number of weeks until relapse for patients whose acute myelogenous leukemia (AML) has gone into remission. One group of patients received maintenance chemotherapy, the other did not. Source: Embury et al. (1977).

The standard 95% confidence interval for $S(t)$ is $\hat{S}(t) \pm 1.96(\widehat{\text{Var}}(\hat{S}(t)))^{1/2}$, using Greenwood's formula

$$\widehat{\text{Var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{j|t_j \leq t} \frac{d_j}{r_j(r_j - d_j)}. \quad (6.17)$$

These intervals are based on a central limit theorem. They do not respect range restrictions, in that they can extend outside of the interval $[0, 1]$. When $S(t)$ takes an extreme value like 0.99 or 0.01, then symmetric intervals do not seem as natural as intervals that extend a greater distance towards $1/2$ than away from $1/2$. These standard intervals also tend to have poor coverage accuracy for moderate n .

Let t be a fixed time point for which $S(t)$ is of interest. Define the profile empirical likelihood function

$$\mathcal{R}(s, t) = \max \left\{ \prod_{j=1}^k \frac{\lambda_j^{d_j} (1 - \lambda_j)^{r_j - d_j}}{\hat{\lambda}_j^{d_j} (1 - \hat{\lambda}_j)^{r_j - d_j}} \mid 0 \leq \lambda_j \leq 1, \prod_{t_j \leq t} (1 - \lambda_j) = s \right\},$$

for $S(t)$. A Lagrange multiplier argument shows that the maximizing λ_j satisfy

$$\lambda_j = \frac{d_j}{r_j + \gamma 1_{t_j \leq t}}, \quad (6.18)$$

for a multiplier γ satisfying

$$\sum_{j|t_j \leq t} \log \left(\frac{r_j - d_j - \gamma}{r_j + \gamma} \right) - \log(s) = 0.$$

Theorem 6.8 shows that the empirical likelihood ratio may be used to construct pointwise confidence intervals for $S(t)$.

Theorem 6.8 For $i = 1, \dots, n$, let $X_i, Y_i \in \mathbb{R}$ be independent random variables with $X_i \sim F$ and $Y_i \sim G$. Let $(Z_i, \delta_i) = (\min(X_i, Y_i), 1_{X_i \leq Y_i})$, $i = 1, \dots, n$

be observed. Assume that $G((-\infty, t)) < 1$ and $0 < S(t) < 1$. Then

$$-2 \log \mathcal{R}(S(t), t) \rightarrow \chi_{(1)}^2,$$

in distribution as $n \rightarrow \infty$.

Proof. Thomas & Grunkemeier (1975), Li (1995b), and Murphy (1995). \square

The top plot in [Figure 6.4](#) shows the empirical likelihood function for $S(20)$, the probability that remission lasts for at least 20 weeks, in each of the two groups in the AML data. The likelihood curves overlap considerably. The survival difference is apparently not very large and the sample size is also small.

The middle plot in [Figure 6.4](#) shows the empirical likelihood ratio curve for the difference $\Delta = S_M(20) - S_N(20)$ in survival probabilities between the maintained (subscript M) and non-maintained (subscript N) groups. The probability of going 20 weeks or more in remission could reasonably be larger for either group. This likelihood ratio is defined as

$$\mathcal{R}(\Delta) = \frac{\max_{\theta_1 - \theta_2 = \Delta} \mathcal{L}(\theta_1, \theta_2)}{\max_{\theta_1, \theta_2} \mathcal{L}(\theta_1, \theta_2)}$$

where

$$\mathcal{L}(\theta_1, \theta_2) = \max \left\{ \prod_{i=1}^n (nw_i)^{\delta_i} \left(\sum_{j|t_j > t_i} nw_j \right)^{1-\delta_i} \mid w_i \geq 0, \sum_{i=1}^n w_i = 1, \right. \\ \left. \sum_{i=1}^n w_i M_i (1_{Y_i \geq 20} - \theta_1) = \sum_{i=1}^n w_i (1 - M_i) (1_{Y_i \geq 20} - \theta_2) = 0, \right\}$$

where M_i is 1 for the maintained group and 0 for the non-maintained group.

The bottom plot in [Figure 6.4](#) shows the empirical likelihood curve for the difference in medians between the two groups. This plot was computed by first computing

$$\mathcal{L}(\theta_1, \theta_2) = \max \left\{ \prod_{i=1}^n (nw_i)^{\delta_i} \left(\sum_{j|t_j > t_i} nw_j \right)^{1-\delta_i} \mid w_i \geq 0, \sum_{i=1}^n w_i = 1, \right. \\ \left. \sum_{i=1}^n w_i M_i (1_{Y_i \leq \theta_1} - 1/2) = \sum_{i=1}^n w_i (1 - M_i) (1_{Y_i \leq \theta_2} - 1/2) = 0, \right\}$$

on a fine grid of (θ_1, θ_2) values, then taking

$$\mathcal{R}(\Delta) = \frac{\max_{\theta_1 - \theta_2 = \Delta} \mathcal{L}(\theta_1, \theta_2)}{\max_{\theta_1, \theta_2} \mathcal{L}(\theta_1, \theta_2)},$$

as before. To maximize over θ_1 with $\theta_1 - \theta_2$ fixed at Δ is to profile out a variable that does not enter the estimating equations in a smooth way. There are few results of this kind, but the present case is covered by [Theorem 10.1](#). The median duration of remission does not differ significantly between these two groups.

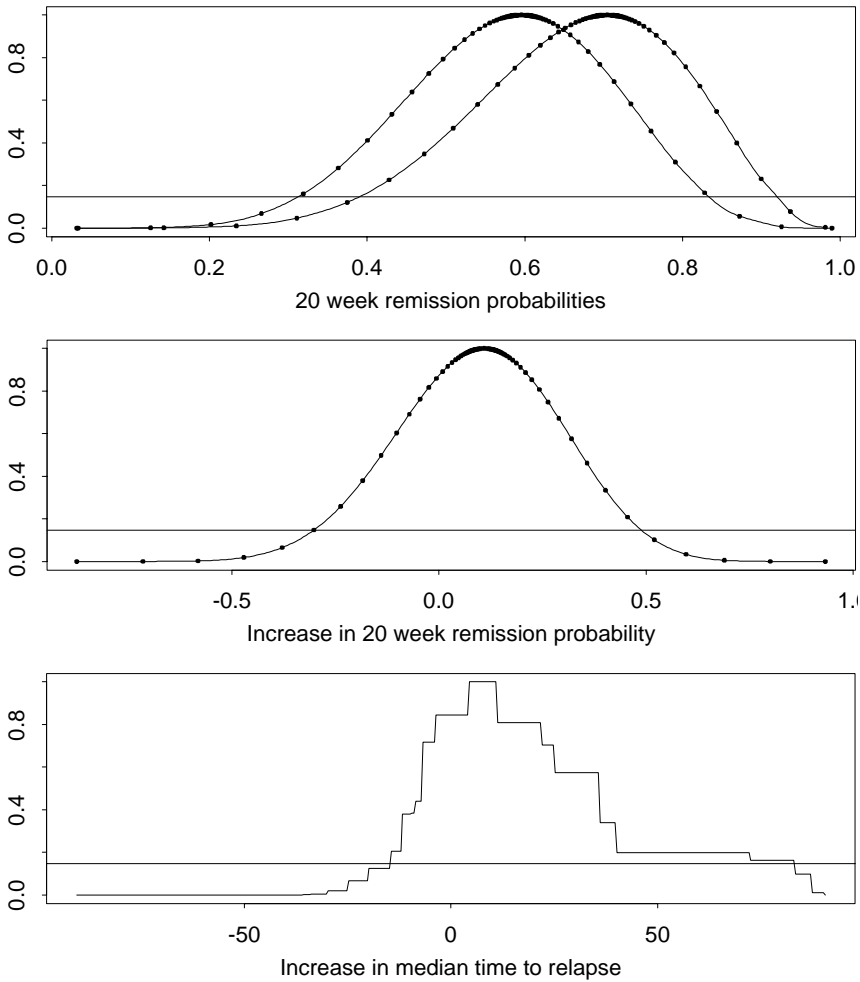


Figure 6.4 The top figure shows empirical likelihood curves for the probability that remission lasts at least 20 weeks. There is one curve for each of the treatment groups. The curve for the maintained group peaks at 0.7045, lying slightly to the right of the curve for the non-maintained group which peaks at 0.5956. The middle figure shows the empirical likelihood ratio function for the difference between the probabilities of remission lasting at least 20 weeks. This curve peaks at 0.1088. The bottom curve is for the difference in median times to relapse between the two groups. Each plot has a horizontal reference line at the approximate 95% confidence level using a $\chi^2_{(1)}$ criterion.

6.7 Proportional hazards

The most convenient way to incorporate predictor variables into survival time likelihoods is through the hazard function. Suppose that $X \in \mathbb{R}$ is a survival time, $Y \in \mathbb{R}$ is a corresponding censoring time, and $U \in \mathbb{R}^d$ is a vector of predictors. Let $Z = \min(X, Y)$ and $\delta = 1_{X \leq Y}$ as before. Cox's proportional hazards model has survival functions

$$S(X_i | U_i = u_i) = S_0(X_i) \exp(u_i' \beta),$$

for a baseline survival function S_0 , and a vector of parameters β . If S_0 is a continuous survival function, then this model has hazard function

$$\lambda(X_i | U_i = u_i) = \lambda_0(X_i) \exp(u_i' \beta).$$

The exponential model for covariates keeps the hazard function nonnegative. It also means that the effect of changing U_i is to make a proportional increase or decrease in the hazard rate, for all times. The baseline survival distribution corresponds to a random variable X with covariate vector $U = 0$. To make the baseline correspond to a default value U_0 , it is only necessary to replace each U_i by $U_i - U_0$.

The data can be organized through two sequences of events unfolding in time. The first sequence specifies the time of the next failure to occur, and the second sequence specifies which of the subjects currently under study is the one to fail at that time. With S_0 completely unknown, it is reasonable that only the second sequence contains information on β . Suppose that there are no ties among the Z_i . For $j = 1, \dots, k = \sum_i \delta_i$, let the item labeled (j) be the one with the j 'th largest of the observed failure times $\{Z_i | \delta_i = 1\}$. Let $R_j = \{i | X_{(j-1)} < Z_i \leq X_{(j)}\}$ be the set of individuals at risk of failure, just prior to time $X_{(j)}$, taking $X_{(0)} = 0$. The partial likelihood is

$$L_P(\beta) = \prod_{j=1}^k \frac{\exp(u_{(j)}' \beta)}{\sum_{i \in R_j} \exp(u_i' \beta)},$$

after canceling $\lambda_0(X_{(j)}) \Delta t$ from the numerator and denominator in each of the k factors. This partial likelihood can be extended, with some difficulty, to take account of ties in Z_i , for observations not necessarily having tied U_i and δ_i .

The partial likelihood behaves like an ordinary parametric likelihood. Maximizing it provides consistent asymptotically normal estimates of β , under mild assumptions, and the profile likelihood formed by maximizing over S_0 can be used to construct confidence regions for β .

6.8 Further empirical likelihood ratio results

Asymptotic χ^2 distributions have been obtained for numerous truncation and censoring settings. Some theorems are quoted here. Some more are described in Chapter 6.9.

For left-truncated data of [Example 6.2](#) let $L_c(F)$ be the conditional likelihood $L(F, G; \mathcal{Y} | \mathcal{X})$ given on page 138. Define

$$\mathcal{R}_t(p) = \frac{\max\{L_c(F) \mid F((-\infty, t]) = p\}}{\max_F\{L_c(F)\}}.$$

Theorem 6.9 *Let $X \sim F_0$ and $Y \sim G_0$ be independent from continuous distributions. Assume that $\inf\{x \mid F_0(x) > 0\} > \inf\{x \mid G_0(x) > 0\}$ and that $F_0(t) = p_0 > 0$. Then $-2 \log \mathcal{R}_t(p_0) \rightarrow \chi_{(1)}^2$ as $n \rightarrow \infty$.*

Proof. Li (1995a, Theorem 1). \square

Suppose that $(X_i, Y_i, Z_i) \in \mathbb{R}^3$ are independent and identically distributed, with X doubly censored, by Y_i on the right and by X_i on the left, as in [Example 6.4](#). Let $F_X, F_Y,$ and F_Z be the distributions of $X, Y,$ and $Z,$ respectively. If $Z < X \leq Y,$ let $U = X$ and $\delta = 0,$ if $X > Y,$ let $U = Y$ and $\delta = 1,$ and if $X \leq Z,$ let $U = Z$ and $\delta = -1.$

Then the conditional likelihood function for F_X is

$$L_c(F) = \prod_{i=1}^n F(\{U_i\})^{\delta_i=0} F((U_i, \infty))^{\delta_i=1} F([0, U_i])^{\delta_i=-1},$$

using x^A as a shorthand for $x^{1A}.$ Let $\theta = \int q(x)dF_X(x),$ for a known function $q,$ and define

$$\mathcal{R}(\theta) = \frac{\max\{L_c(F) \mid \int q(x)dF(x) = \theta\}}{\max_F\{L_c(F)\}}.$$

Theorem 6.10 *Let \mathcal{R} and θ be as described above. Suppose that $F_X, F_Y,$ and F_Z are continuous distributions, that $F_X([A, B]) = 1,$ for some $0 \leq A < B < \infty,$ that $\Pr(Z < u \leq Y) > 0$ for all $u \in [A, B],$ that $F_Z([0, B]) = 1,$ and that $F_Y([0, A]) = 0.$ Suppose that q is a left continuous function of bounded variation on $[A, B]$ with $\int q^2(x)dF_X(x) - (\int q(x)dF_X(x))^2 > 0.$ Then $-2 \log \mathcal{R}(\theta_0) \rightarrow \chi_{(1)}^2.$*

Proof. Murphy & van der Vaart (1997, [Theorem 2.1](#)). \square

Now consider current status data as in [Example 6.5](#), with only one observation time $Z,$ but with a Cox model covariate $U \in \mathbb{R}.$ The event time is $X,$ and we observe Z and $\delta = 1_{X_i \leq Z_i}.$ Suppose that conditionally on $U,$ the hazard function of X is $\lambda(t) \exp(\theta U)$ for a parameter $\theta \in \mathbb{R}.$ The conditional likelihood in terms of the cumulative hazard of X is

$$L(\Lambda, \theta) = \prod_{i=1}^n \left(1 - \exp(-e^{\theta U_i} \Lambda(Z_i))\right)^{\delta_i} \left(\exp(-e^{\theta U_i} \Lambda(Z_i))\right)^{1-\delta_i}.$$

Now define

$$\mathcal{R}(\theta) = \frac{\max_{\Lambda}\{L(\Lambda, \theta)\}}{\max_{\Lambda, \eta}\{L(\Lambda, \eta)\}},$$

where Λ is maximized over the space of nondecreasing functions continuous from the right with limits from the left and taking values in $[0, C]$ for a known bound C , and θ is maximized over a parameter set Θ .

Theorem 6.11 *For the current status setting above let the observation time Z have a continuous positive density function on $[A, B]$ for some $0 < A < B < \infty$, where the true cumulative hazard Λ_0 of X satisfies $\Lambda_0(A-) > 0$ and $\Lambda_0(B) < C$. Let Λ_0 be differentiable on $[A, B]$ with a derivative everywhere above some $\epsilon > 0$, let U be bounded with $E(\text{Var}(U | Z)) > 0$, and assume that the true value θ_0 is interior to Θ . Finally, assume that $\phi(Z)$ has a bounded derivative on $[A, B]$ where*

$$\phi(Z) = \frac{E(U\psi(U, X, \delta)^2 | Z)}{E(\psi(U, X, \delta)^2 | Z)}, \quad \text{and}$$

$$\psi(U, X, \delta) = e^{\theta_0 U} \left[\delta \frac{\exp(-\exp(\theta_0 U))\Lambda(X)}{1 - \exp(-\exp(\theta_0 U))\Lambda(X)} - (1 - \delta) \right].$$

Then $-2 \log \mathcal{R}(\theta_0) \rightarrow \chi_{(1)}^2$ as $n \rightarrow \infty$.

Proof. Murphy & van der Vaart (1997, Theorem 2.2). \square

A form of cumulative hazard estimating equation can be used to define parameters in survival analysis. There, the parameter θ solves $\int q(x, \theta) d\Lambda(x) = C$ for some constant C .

Theorem 6.12 *For right-censored data as described in the conditions of Theorem 6.8, suppose that $q(x)$ is a left continuous function with*

$$0 < \int \frac{|q(x, \theta)|^m}{F([x, \infty))G([x, \infty))} d\Lambda(x) < \infty, \quad m = 1, 2.$$

Then $-2 \log \mathcal{R}(\theta_0) \rightarrow \infty$ where $\mathcal{R}(\theta)$ is an empirical likelihood ratio defined through the hazard function at observed failure times, and maximized subject to $\int g(x, \theta) d\Lambda(x) = C$.

Proof. Pan & Zhou (2000, Theorem 5). \square

6.9 Bibliographic notes

Biased sampling

Bratley, Fox & Schrage (1987) is a standard reference on Monte Carlo that includes importance sampling. Cochran (1977) and Lohr (1998) are standard references on finite population sampling. Cox (1967) proposed the harmonic mean for length biased data. The shrub data are from Muttlak & McDonald (1990). Jones (1991) uses them to investigate kernel density estimation from length biased samples.

The problem of constructing an NPMLE from one length biased and one unbiased sample was considered by Vardi (1982). Cosslett (1981) proposes an NPMLE for choice-based sample data. Vardi (1985) and Gill et al. (1988) consider multiple samples each with their own biasing conditions. Vardi (1985) gives an algorithm for computing the NPMLE, and shows that the NPMLE \hat{F} is a sufficient statistic for the unknown F_0 .

Empirical likelihood for a combination with one biased and one unbiased sample was considered by Qin (1993). Qin proves an ELT for the mean with one biased and one unbiased sample and states that the result holds more generally.

Qin & Zhang (1997) study a problem in which the bias functions in multiple biased sampling contain parameters. In group i the data are a biased sample defined through $u_i(x, \eta_i)$, incorporating unknown parameter vectors η_i . They were motivated by case-control studies, where samples are taken of people with and without a rare condition. Empirical likelihood methods can be used to draw inferences on the η_i , as well as to test the goodness of fit of the parametric specification of the bias function. Fokianos, Peng & Qin (1999) use this idea to test the goodness of a logistic link function. Qin (1999) considers three samples, $X_i \sim F$, $Y_i \sim G$, and $Z_i \sim \lambda F + (1 - \lambda)G$ for an unknown mixture proportion λ . The distributions F and G have densities f and g that are assumed to satisfy $\log(g(x)/f(x)) = \beta_0 + x\beta_1$. Qin (1999) finds asymptotic chisquared distributions for likelihood ratio tests of λ and β_1 .

Qin (1998) considers empirical likelihood inference in upgraded mixture models. This setting combines a “good sample” of directly observed data $Z \sim F$, with a “bad sample” of data from a density $p(x) = \int p(x|z)dF$ for a conditional density $p(x|z)$. These models originate with Hasminskii & Ibragimov (1993). The name is from van der Vaart & Wellner (1992), who develop a discrete consistent estimator of F . An example from Vardi & Zhang (1992) has $X = ZU$ for $U \sim U(0, 1)$ independently of Z . The KDD CUP 2000 data mining competition (Kohavi, Brodley, Frasca, Mason & Zheng 2001) featured Internet log entries, including some complete sessions of length Z and some clipped sessions that only included the first $\lceil UZ \rceil$ entries for a uniform U . The winning entry of Rafal Kustra, Jorge Picazo, and Bogdan Popescu showed that clipping produced an artifact wherein the rate at which visitors left a site appeared to increase with the duration of their session, although the true departure rate declined with duration. Qin (1998) shows how to use empirical likelihood inferences on upgraded mixture models, and how to incorporate parameterized data distortions $p(x | z; \theta)$.

Censoring and truncation

Peto (1973) considers NPMLE’s for general patterns of interval censored real values, and proves [Theorem 6.5](#). The proof of [Theorem 6.4](#) is adapted from the argument in Peto (1973).

An example of a bad NPMLE arises for bivariate failure times $X_i = (X_{i1}, X_{i2})$ where X_{ij} is subject to right censoring by Y_{ij} . If X_{i1} is censored but X_{i2} is not,

then $C_i = \{(x_1, X_{i2}) | Y_{i1} < x_1 < \infty\}$ is a ray of infinite length in the plane. If X_i have a continuous distribution F then there will never be an $X_{i'} \in C_i$ for $i' \neq i$. As Tsai, Leurgans & Crowley (1986) note, such additional points are necessary to properly distribute probability within C_i , and without them the NPMLE is not even consistent. van der Laan (1996) proposes a way to fix this problem in which a ray like C_i is replaced by a thin strip, with a width that decreases to 0 as $n \rightarrow \infty$.

Turnbull (1976) is a definitive reference on NPMLE's for combinations of censored and truncated real-valued data. Turnbull (1976)'s description of censoring is a form of coarsening at random. Heitjan & Rubin (1991) define and illustrate coarsening at random and show that under coarsening at random, the conditional likelihood is proportional to the full likelihood. Coarsening at random includes information loss due to missing data components or rounding.

Turnbull (1976) provides a self-consistency algorithm for finding the NPMLE. This algorithm is an example of the EM algorithm (see Dempster, Laird & Rubin (1977) and Baum (1972)). Efron (1967) used self-consistency to derive the Kaplan-Meier estimator. There is as yet no ELT for the general setting Turnbull considers.

The survival, hazard, and cumulative hazard functions are defined in Fleming & Harrington (1991), as well as in Kalbfleisch & Prentice (1980), who use a survivor function $\Pr(X \geq t)$ instead of the survival function $\Pr(X > t)$.

Kaplan-Meier estimator

Kaplan & Meier (1958) introduced the product-limit estimator for right-censored survival times, using an NPMLE argument. Similar estimators had previously been used by actuaries. The variance estimate of the Kaplan-Meier estimate is from Greenwood (1926). The NPMLE derivation is based on Kalbfleisch & Prentice (1980), who also present a derivation of Greenwood's formula.

Thomas & Grunkemeier (1975) gave a heuristic proof of [Theorem 6.8](#). This was later made rigorous by Li (1995b) and by Murphy (1995). Murphy (1995) proves an ELT for inferences on the cumulative hazard function. [Equation \(6.18\)](#) was obtained by Thomas & Grunkemeier (1975) and independently by Cox & Oakes (1984, Chapter 4.3) who use it to derive Greenwood's formula [\(6.17\)](#) from the curvature of the censored data empirical log likelihood.

Adimari (1997) considers empirical likelihood inferences for the mean of a distribution under independent right censoring. He finds an asymptotic chisquared distribution for $2n \sum_{i=1}^n \tilde{p}_i \log(1 + \lambda'(T_i - \mu))$ where \tilde{p}_i is the Kaplan-Meier probability of the observed failure time T_i and λ satisfies $\sum_{i=1}^n \tilde{p}_i (T_i - \mu) / (1 + \lambda'(T_i - \mu)) = 0$.

Pan & Zhou (2000) prove [Theorem 6.12](#). They also establish a chisquared calibration for parameters $\int q_n(x) d\Lambda(x)$ where q_n is a data-dependent function. Such parameters often arise where a data-based estimate of one quantity is plugged into an equation for another.

The AML data come from Embury et al. (1977). They are reproduced in Miller, Gong & Munoz (1981).

Lynden-Bell and astronomy

Efron & Petrosian (1994) explore some data where objects that are either too bright or too dim are truncated. The dim objects are not visible, while the bright ones are possibly not the sort of object of interest. They introduce a nonparametric estimate of the distribution function of brightness for such doubly truncated data and provide a bootstrap-based test of the cosmological principle.

Keiding & Gill (1990) and Woodroffe (1985) provide a detailed analysis of left-truncated sampling. The NPMLE in this case was found by Lynden-Bell (1971) and is known as the Lynden-Bell estimator. Lynden-Bell considered the more general setting in which an (X, Y) pair was observed with probability $u(X, Y)$, allowing a model in which the probability of observing an object of given brightness decreases smoothly from 1 to 0 as its distance from Earth increases. Lynden-Bell (1971) gives a small data set of 40 3CR quasars. The NPMLE is degenerate, putting positive weight on only three of the quasars. Woodroffe (1985) describes conditions leading to this degeneracy and conditions in which the probability of a degenerate NPMLE vanishes as $n \rightarrow \infty$. Lynden-Bell (1971) also implements a fix in which the histogram of an intermediate quantity is replaced by the nearest unimodal one.

Wang (1987) proves [Theorem 6.3](#). Keiding & Gill (1990) provide another proof and add the caveat that the maximizer \tilde{F} of the conditional likelihood is not a component of the joint NPMLE (\hat{F}, \hat{G}) in cases where \tilde{F} is degenerate. Li (1995a) proves [Theorem 6.9](#). Li, Qin & Tiwari (1997) consider the case where there is a known parametric family of distributions for G , but not for F . They use the marginal distribution of the X_i because in this setting it can have more information than the conditional distribution of the X_i given the Y_i . They also show how to get empirical likelihood ratio confidence regions for the probability α that an observation is not truncated.

Other

The proportional hazards model in Chapter 6.7 was proposed by Cox (1972). The partial likelihood argument for it is due to Cox (1975). Bailey (1984) considered the joint likelihood for β and S_0 taking jumps at observed failure times. He showed that the estimate of β obtained by maximizing the likelihood over β and S_0 is asymptotically equivalent to the one obtained by maximizing the partial likelihood. The resulting estimate of the cumulative hazard is equivalent to the one in Tsiatis (1981). Confidence regions for β or for the cumulative hazard (at finitely many points) can be obtained from the curvature of the log likelihood. Bailey (1984) remarks that the presence of a large number of nuisance parameters does not lead to difficulty. Murphy & van der Vaart (2000) consider the problem of infinite dimensional nuisance parameters more generally.

Murphy & van der Vaart (1997) prove [Theorems 6.10](#) and [6.11](#). They also establish χ^2 limits for some frailty models incorporating random effects into the proportional hazards framework.

In this chapter, survival analysis was viewed as analysis of life times that might be missing or partially observed. The modern treatment of survival analysis treats each subject's data as a counting process observed over a time window. The number of deaths for an individual is a counting process that starts at 0 and may increase to 1 in the time window of observation. A second counting process takes the value 1 if the individual is at risk of failure and 0 otherwise, whether the reason be failure or censoring. For a more comprehensive treatment of survival analysis, using the theory of counting processes, see Fleming & Harrington (1991) and Andersen, Borgan, Gill & Keiding (1993), with a very accessible applied presentation in Therneau & Grambsch (2000). Counting process models extend naturally to handle competing risks from different causes of death, events such as infections which can recur for individuals, and transitions between states such as cancer and remission.

The dual likelihood of Mykland (1995) is an extension of empirical likelihood to martingales. Dual likelihood inferences should cover many or most of the counting process examples, though this is outside the scope of the present text.

6.10 Exercises

Exercise 6.1 Suppose that X_1, \dots, X_n are IID with the exponential probability density function $f(x; \theta) = \theta \exp(-\theta x) 1_{x>0}$. Thus θ is the failure rate per unit time and $1/\theta = E(X)$. Suppose that Y_1, \dots, Y_n are censoring times independent of X_1, \dots, X_n , and let the observations be $Z_i = \min(X_i, Y_i)$ and $\delta_i = 1_{X_i < Y_i}$. Write an expression for the parametric conditional likelihood of X_1, \dots, X_n given Y_1, \dots, Y_n , in terms of Z_i and δ_i . Find the conditional MLE $\hat{\theta}$. Is this quantity interpretable, in the case where X_i are not exponentially distributed?

Exercise 6.2 Suppose that F puts weight $w_i \geq 0$ on x_i and that G puts weight $v_j \geq 0$ on y_j , where $\sum_{i=1}^n w_i = \sum_{j=1}^n v_j = 1$. Let $u_{ij} = 1_{x_i > y_j}$, and define $\alpha = \Pr(X > Y) = \sum_{i=1}^n \sum_{j=1}^n w_i v_j u_{ij}$. Show by Lagrange multipliers that

$$w_i = \left(\sum_{j=1}^n \frac{u_{ij}}{\sum_{k=1}^n w_k u_{kj}} \right)^{-1}$$

for the NPMLE in the Lynden-Bell setup.