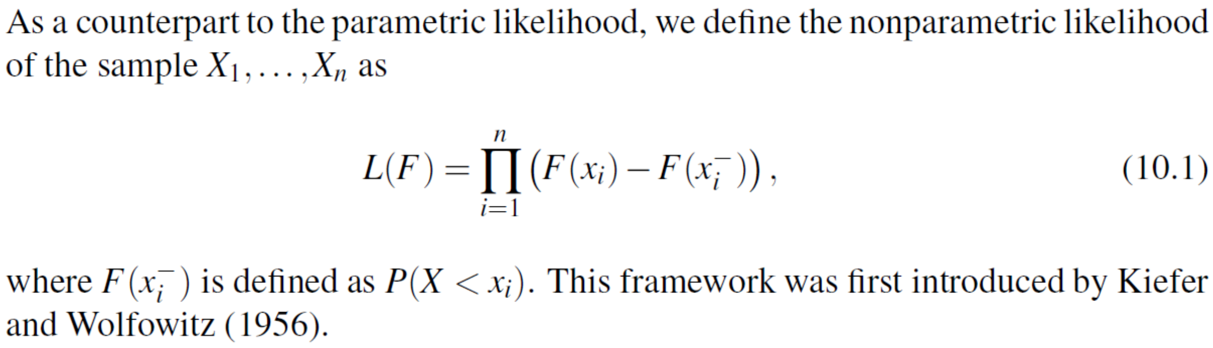
ISyE 6404 – Nonparametric Statistics

Lecture #5 – Estimating Distribution Functions (CDF) (Textbook page 183)

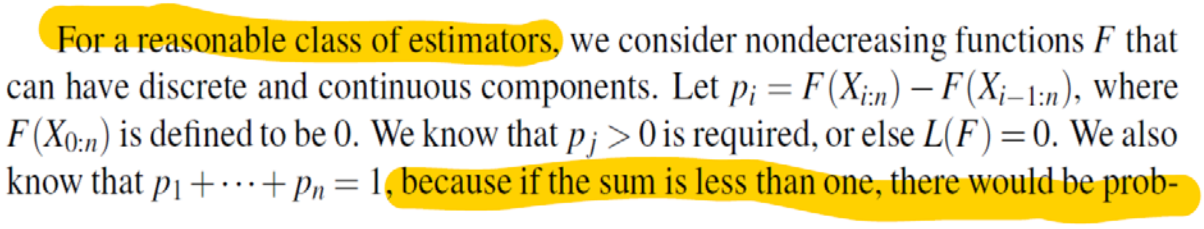
10.1 Nonparametric Maximum Likelihood (NPMLE)



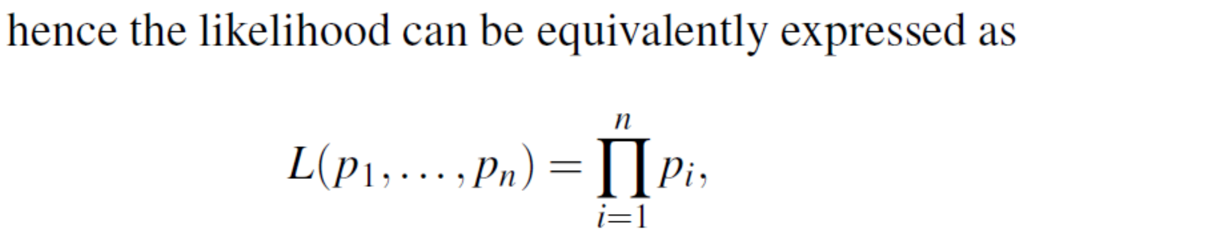
As a counterpart to the parametric likelihood, we define the nonparametric likelihood of the sample $X_1, \ldots, X_n$ as

$$L(F) = \prod_{i=1}^{n} \left(F(x_i) - F(x_i^-) \right), \tag{10.1}$$

where $F(x_i^-)$ is defined as $P(X < x_i)$. This framework was first introduced by Kiefer and Wolfowitz (1956).

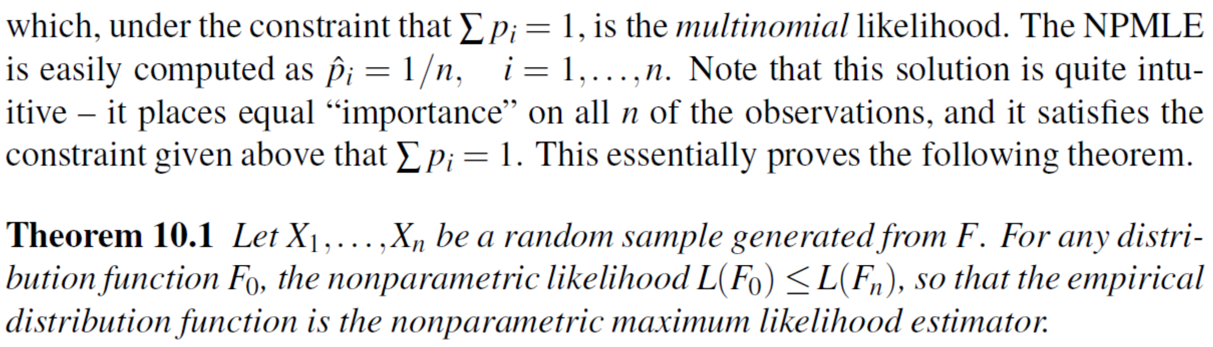


For a reasonable class of estimators, we consider nondecreasing functions F that can have discrete and continuous components. Let $p_i = F(X_{i:n}) - F(X_{i-1:n})$, where $F(X_{0:n})$ is defined to be 0. We know that $p_j > 0$ is required, or else $L(F) = 0$. We also know that $p_1 + \cdots + p_n = 1$, because if the sum is less than one, there would be prob-



hence the likelihood can be equivalently expressed as

$$L(p_1, \ldots, p_n) = \prod_{i=1}^{n} p_i,$$



which, under the constraint that $\sum p_i = 1$, is the *multinomial* likelihood. The NPMLE is easily computed as $\hat{p}_i = 1/n, \quad i = 1, \ldots, n$. Note that this solution is quite intuitive – it places equal "importance" on all n of the observations, and it satisfies the constraint given above that $\sum p_i = 1$. This essentially proves the following theorem.

Theorem 10.1 *Let $X_1, \ldots, X_n$ be a random sample generated from F. For any distribution function F_0, the nonparametric likelihood $L(F_0) \leq L(F_n)$, so that the empirical distribution function is the nonparametric maximum likelihood estimator.*

The following provides the definition of empirical distribution function.

Let $X_1, X_2, \dots, X_n$ be a sample from a population with continuous CDF F. In Chapter 3, we defined the *empirical (cumulative) distribution function* (EDF) based on a random sample as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \le x).$$

10.3 Kaplan-Meier Estimator

The nonparametric likelihood can be generalized to all sorts of observed data sets beyond a simple i.i.d. sample. The most commonly observed phenomenon outside the i.i.d. case involves *censoring*. To describe censoring, we will consider $X \geq 0$, because most problems involving censoring consist of lifetime measurements (e.g., time until failure).

Definition 10.1 *Suppose X is a lifetime measurement. X is* **right censored** *at time t if we know the failure time occurred after time t, but the actual time is unknown. X is* **left censored** *at time t if we know the failure time occurred before time t, but the actual time is unknown.*

Definition 10.2 **Type-I censoring** *occurs when n items on test are stopped at a fixed time t_0, at which time all surviving test items are taken off test and are right censored.*

Type I censoring is a common problem in drug treatment experiments based on human trials; if a patient receiving an experimental drug is known to survive up to a time t but leaves the study (and humans are known to leave such clinical trials much more frequently than lab mice) the lifetime is right censored.

Suppose we have a sample of possibly right-censored values. We will assume the random variables represent lifetimes (or "occurrence times"). The sample is summarized as $\{(X_i, \delta_i), \quad i = 1, \dots, n\}$, where X_i is a time measurement, and δ_i equals 1 if the X_i represents the lifetime, and equals 0 if X_i is a (right) censoring time. If $\delta_i = 1$, X_i contributes $dF(x_i) \equiv F(x_i) - F(x_i^-)$ to the likelihood (as it does in the i.i.d. case). If $\delta_i = 0$, we know only that the lifetime surpassed time X_i, so this event contributes $1 - F(x_i)$ to the likelihood. Then

$$L(F) = \prod_{i=1}^{n} (1 - F(x_i))^{1-\delta_i} \left(dF(x_i)\right)^{\delta_i}. \tag{10.2}$$

The argument about the NPMLE has changed from (10.1). In this case, no probability mass need be assigned to a value X_i for which $\delta_i = 0$, because in that case, $dF(X_i)$ does not appear in the likelihood. Furthermore, the accumulated probability mass of the NPMLE on the observed data does not necessarily sum to one, because if the largest value of X_i is a censored observation, the term $S(X_i) = 1 - F(X_i)$ will only be positive if probability mass is assigned to a point or interval to the right of X_i.

Let p_i be the probability mass assigned to $X_{i:n}$. This new notation allows for positive probability mass (call it p_{n+1}) that can be assigned to some arbitrary point or interval after the last observation $X_{n:n}$. Let $\tilde{\delta}_i$ be the censoring indicator associated with $X_{i:n}$. Note that even though $X_{1:n} < \cdots < X_{n:n}$ are ordered, the set $(\tilde{\delta}_1,\ldots,\tilde{\delta}_n)$ is not necessarily so ($\tilde{\delta}_i$ is called a *concomitant*).

If $\tilde{\delta}_i = 1$, the likelihood is clearly maximized by setting probability mass (say p_i) on $X_{i:n}$. If $\tilde{\delta}_i = 0$, some mass will be assigned to the right of $X_{i:n}$, which has interval probability $p_{i+1} + \cdots + p_{n+1}$. The likelihood based on censored data is expressed

$$L(p_1,\ldots,p_{n+1}) = \prod_{i=1}^{n} p_i^{\tilde{\delta}_i} \left(\sum_{j=i+1}^{n+1} p_j \right)^{1-\tilde{\delta}_i}.$$

Instead of maximizing the likelihood in terms of $(p_1,\ldots,p_{n+1})$, it will prove to be much easier using the transformation

$$\lambda_i = \frac{p_i}{\sum_{j=i}^{n+1} p_j}.$$

This is a convenient one-to-one mapping where

$$\sum_{j=i}^{n+1} p_j = \prod_{j=1}^{i-1}(1-\lambda_j), \quad p_i = \lambda_i \prod_{j=1}^{i-1}(1-\lambda_j).$$

The likelihood simplifies to

$$L(\lambda_1,\ldots,\lambda_{n+1}) \quad = \quad \prod_{i=1}^{n} \left(\frac{\lambda_i}{1-\lambda_i} \right)^{\tilde{\delta}_i} (1-\lambda_i)^{n-i+1}.$$

As a function of $(\lambda_1,\ldots,\lambda_{n+1})$, L is maximized at $\hat{\lambda}_i = \tilde{\delta}_i/(n-i+1)$, $i = 1,\ldots,n+1$.

Equivalently,

$$\hat{p}_i = \frac{\tilde{\delta}_i}{n-i+1}\prod_{j=1}^{i-1}\left(1-\frac{\tilde{\delta}_j}{n-j+1}\right).$$

The NPMLE of the distribution function (denoted $F_{KM}(x)$) can be expressed as a sum in p_i. For example, at the observed order statistics, we see that

$$\begin{aligned} S_{KM}(x_{i:n}) &\equiv 1-F_{KM}(x_{i:n}) = \prod_{j=1}^{i}\left(1-\frac{1}{n-j+1}\right)^{\tilde{\delta}_j} \qquad (10.3)\\ &= \prod_{j=1}^{i}\left(1-\frac{\tilde{\delta}_j}{n-j+1}\right). \end{aligned}$$

This is the *Kaplan–Meier* nonparametric estimator, developed by Kaplan and Meier (1958) for censored lifetime data analysis. It's been one of the most influential developments in the past century; their paper is the most cited paper in statistics (Stigler, 1994). E. L. Kaplan and Paul Meier never actually met during this time, but they both

EXAMPLE 10.1

Muenchow (1986) tested whether male or female flowers (of *Western White Clematis*), were equally attractive to insects. The data in the Table 10.1 represent waiting times (in minutes), which includes censored data. In R, use the functions

`Surv(), survfit()`

in `survival` package. `Surv()` is a function to create an object for lifetime data. `survfit()` function finds a Kaplan-Meier estimate of a survival curve.

Table 10.1 Waiting Times for Insects to Visit Flowers

Male Flowers			Female Flowers		
1	9	27	1	19	57
1	9	27	2	23	59
2	9	30	4	23	67
2	11	31	4	26	71
4	11	35	5	28	75
4	14	36	6	29	75*
5	14	40	7	29	78*
5	14	43	7	29	81
6	16	54	8	30	90*
6	16	61	8	32	94*
6	17	68	8	35	96
7	17	69	9	35	96*
7	18	70	14	37	100*
8	19	83	15	39	102*
8	19	95	18	43	105*
8	19	102*	18	56	
		104*			

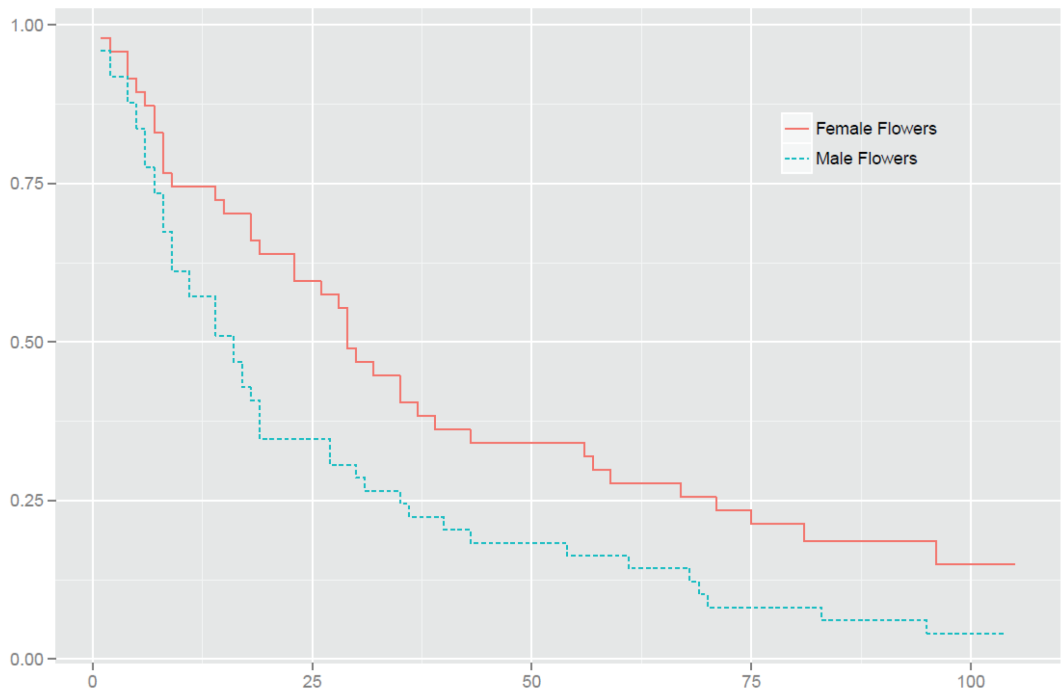


Figure 10.2 Kaplan-Meier estimator for Waiting Times (solid line for male flowers, dashed line for female flowers).

10.4 Confidence Interval for F

For "complete samples" (non-censored data),

Like all estimators, $\hat{F}(x)$ is only as good as its measurement of uncertainty. Confidence intervals can be constructed for $F(x)$ just as they are for regular parameters, but a typical inference procedure refers to a *pointwise* confidence interval about $F(x)$ where x is fixed.

A simple, approximate $1-\alpha$ confidence interval can be constructed using a normal approximation

$$\hat{F}(x) \pm z_{1-\alpha/2}\hat{\sigma}_{\hat{F}},$$

where $\hat{\sigma}_{\hat{F}}$ is our estimate of the standard deviation of $\hat{F}(x)$. If we have an i.i.d. sample, $\hat{F} = F_n$, and $\sigma^2_{F_n} = F(x)[1-F(x)]/n$, so that

$$\hat{\sigma}^2_{\hat{F}} = F_n(x)[1-F_n(x)]/n.$$

Right-Censoring Case:

In the case of right censoring, a confidence interval can be based on the Kaplan-Meier estimator, but the variance of $F_{KM}(x)$ does not have a simple form. Greenwood's formula (Greenwood, 1926), originally concocted for grouped data, can be applied to construct a $1-\alpha$ confidence interval for the survival function $(S = 1-F)$ under right censoring:

$$S_{KM}(t_i) \pm z_{\alpha/2}\hat{\sigma}_{KM}(t_i),$$

where

$$\hat{\sigma}^2_{KM}(t_i) = \hat{\sigma}^2(S_{KM}(t_i)) = S_{KM}(t_i)^2 \sum_{t_j \le t_i} \frac{d_j}{m_j(m_j - d_j)}.$$

It is important to remember these are *pointwise* confidence intervals, based on fixed values of t in $F(t)$. Simultaneous confidence bands are a more recent phenomenon and apply as a confidence statement for F across all values of t for which $0 < F(t) < 1$. Nair (1984) showed that the confidence bands by Hall and Wellner (1980) work well in various settings, even though they are based on large-sample approximations. An approximate $1-\alpha$ confidence band for $S(t)$, for values of t less than the largest observed failure time, is

$$S_{KM}(t) \pm \sqrt{-\frac{1}{2n}\ln\left(\frac{\alpha}{2}\right)} S_{KM}(t)\left(1+\hat{\sigma}^2_{KM}(t)\right).$$

10.5 Plug-in Principle

With an i.i.d. sample, the EDF serves not only as an estimator for the underlying distribution of the data, but through the EDF, any particular parameter θ of the distribution can also be estimated. Suppose the parameter has a particular functional relationship with the distribution function F:

$$\theta = \theta(F).$$

Examples are easy to construct. The population mean, for example, can be expressed

$$\mu = \mu(F) = \int_{-\infty}^{\infty} x dF(x)$$

and variance is

$$\sigma^2 = \sigma^2(F) = \int_{-\infty}^{\infty} (x-\mu)^2 dF(x).$$

10.6 Semi-Parametric Inference

The *proportional hazards* model for lifetime data relates two populations according to a common underlying hazard rate. Suppose $r_0(t)$ is a baseline hazard rate, where $r(t) = f(t)/(1-F(t))$. In reliability theory, $r(t)$ is called the *failure rate*. For some covariate x that is observed along with the lifetime, the positive function of $\Psi(x)$ describes how the level of x can change the failure rate (and thus the lifetime distribution):

$$r(t;x) = r_0(t)\Psi(x).$$

This is termed a *semi-parametric model* because $r_0(t)$ is usually left unspecified (and thus a candidate for nonparametric estimation) where as $\Psi(x)$ is a known positive function, at least up to some possibly unknown parameters. Recall that the CDF is related to the failure rate as

$$\int_{-\infty}^{x} r(u)du \equiv R(u) = -\ln S(x),$$

where $S(x) = 1 - F(x)$ is called the survivor function. $R(t)$ is called the *cumulative failure rate* in reliability and life testing. In this case, $S_0(t)$ is the baseline survivor function, and relates to the lifetime affected by $\Psi(x)$ as

$$S(t;x) = S_0(t)^{\Psi(x)}.$$

The most commonly used proportional hazards model used in survival analysis is called the *Cox Model* (named after Sir David Cox), which has the form

$$r(t;x) = r_0(t)e^{x'\beta}.$$

With this model, the (vector) parameter β is left unspecified and must be estimated. Suppose the baseline hazard function of two different populations are related

by proportional hazards as $r_1(t) = r_0(t)\lambda$ and $r_2(t) = r_0(t)\theta$. Then if T_1 and T_2 represent lifetimes from these two populations,

$$P(T_1 < T_2) = \frac{\lambda}{\lambda + \theta}.$$

The probability does not depend at all on the underlying baseline hazard (or survivor) function. With this convenient set-up, nonparametric estimation of $S(t)$ is possible through maximizing the nonparametric likelihood. Suppose n possibly right-censored observations $(x_1,\ldots,x_n)$ from $F=1-S$ are observed. Let ξ_i represent the number of observations at risk just before time x_i. Then, if δ_i=1 indicates the lifetime was observed at x_i,

$$L(\beta)=\prod_{i=1}^{n}\left(\frac{e^{x_i'\beta}}{\sum_{j\in\xi_i}e^{x_j'\beta}}\right)^{\delta_i}.$$

Related to the proportional hazard model, is the *accelerated lifetime model* used in engineering. In this case, the baseline survivor function $S_0(t)$ can represent the lifetime of a test product under usage conditions. In an accelerated life test, and additional stress is put on the test unit, such as high or low temperature, high voltage, high humidity, etc. This stress is characterized through the function $\Psi(x)$ and the survivor function of the stressed test item is

$$S(t;x)=S_0(t\Psi(x)).$$

Accelerated life testing is an important tool in product development, especially for electronics manufacturers who produce gadgets that are expected to last several years on test. By increasing the voltage in a particular way, as one example, the lifetimes can be shortened to hours. The key is how much faith the manufacturer has on the known acceleration function $\Psi(x)$.

In R, the `survival` package offers the function `coxph`, which computes Cox proportional hazards estimator for input data, much in the same way the `survfit` computes the Kaplan-Meier estimator.