Chapter 4 – Nonparametric Methods to Estimate Survival Functions

If during a study, one observes mixtures of censored observations, missing observations and completion times, the heuristic definitions given in Chapter 2 are inadequate to analyze such complex data. The methods given in this chapter are less efficient than parametric models, when one has chosen the correct model, but they are important when one has no base of experience from which to choose a distribution. We saw this case with the recent COVID-19 pandemic.

In many cases, individuals died without being diagnosed or were misdiagnosed with respiratory distress. In June 2021, the number of known death due to COVID-19 in the United States crossed 600,000. The actual number of death due to COVID-19 may be more than 900,000. In the influenza (***Spanish***) epidemic of 1918, the US had 675,000 deaths attributed to the disease but the US had less than one-third of its present population of 332 million. The US population at the time of my birth was approximately 147 million people. However, Alaska and Hawai'i were not yet states and their populations were not counted. The 1918 US population was only 103 million. More than 10% of the US population had the disease as of June 2021. The ***Spanish Influenza*** actually lasted through 1922 but most deaths ended by 1921. Most death estimates are computed by looking retrospectively at expected numbers of deaths compared to the observed number of deaths. The excess number of deaths is then attributed to the disease. However, as many families lived in quarantine for months, death rates due to cancer may have increased due to one's inability to obtain treatment whereas a decline in death rates may have occurred in workplace injuries.

The goal in this chapter is to find nonparametric estimates of the density, the survival and the hazard functions. The ***product limit estimator*** (PL) developed by Kaplan-Meier will be used to do so. An alternative method is to use a ***life table analysis*** (LTA).

Product Limit Estimates

Let $t_1, \ldots, t_n$ be the exact survival times of all n observations under investigation. Relabel these values using the notation from order statistics that we mentioned earlier (i.e., $t_{1:n} \leq t_{2:n} \leq \cdots \leq t_{n:n}$).

$$\hat{S}(t_{i:n}) = \frac{n-i}{n}.$$

The tacit assumption in life studies is that lifetime is non-negative and so $S(0) = 1$ with $t_{0:n} = 0$. One can connect these points using different methods. This definition provides estimates at the exact time points observed. If one modifies the definition as follows

$$\hat{S}(t) = \frac{n-i}{n} \text{ for } t_{i-1:n} < t \leq t_{i:n}$$

The expression will produce a non-increasing step function (descending staircase) for all time up to the last failure, $t_{n:n}$, for which $\hat{S}(t) = 0$ for all $t > t_{n:n}$. Example 4.1 in the text has ten survival times of lung cancer patients given as 4, 5, 6, 8, 8, 8, 10, 10, 11, 12. Notice that the data are provided already ordered. We have the following estimates:

$$\hat{S}(t) = 1 \text{ for } t < 4, \hat{S}(t) = \frac{9}{10} \text{ for } 4 \leq t < 5, \hat{S}(t) = \frac{8}{10} \text{ for } 5 \leq t < 6, \hat{S}(t) = \frac{7}{10} \text{ for } 6 \leq t < 8.$$

In the case of t=8, we have multiple ties for $\hat{S}(t) = \frac{4}{10}, 8 \leq t < 10$ and so on. The median survival time, m, is the point in time at $\hat{S}(m) = 1/2$. Since

$$\hat{S}(6) = \frac{7}{10} \text{and } \hat{S}(8) = \frac{4}{10}$$

We have the sample median, m, bounded between 6 and 8. The authors suggest using the method of linear interpolation, which suggest using a piecewise linear function between the successive points. Let us solve this using linear interpolation.

$$\frac{8-6}{0.4-0.7} = \frac{8-m}{0.4-0.5} \Rightarrow m = 7.\overline{3}$$

This technique works provided all observed times and failure or death times. As a prelude to a more general method, one can rewrite the survival function recursively

$$\hat{S}(t) = \prod_{t>t_{r:n}} \frac{n-r}{n-r+1} = \prod_{t>t_{r:n}} \left(1 - \frac{1}{n-r+1}\right)$$

Where r is the number of deaths before time t. When all the lifetimes are uncensored,

$$\hat{S}(t_{i:n}) = \hat{S}(t_{i-1:n}) \frac{n-i}{n-i+1}.$$

The variance of the product limit estimator is approximated by

$$Var\left[\hat{S}(t)\right] = \left[\hat{S}(t)\right]^2 \sum_{r} \frac{1}{(n-r)(n-r+1)}$$

Where the index of summation is over all r such $t > t_{r:n}$. The following formula allows one to estimate the mean by finding the area under the survival curve using the areas of the rectangles.

$$\hat{\mu} = t_{1:n} + \sum_{i=0}^{n} \hat{S}(t_{i:n})(t_{i+1:n} - t_{i:n})$$

A generalization of the ***Kaplan – Meier*** estimator can be expressed as

$$\hat{S}(t) = \prod_{i:t>t_{r:n}}^{r} \left(1 - \frac{d_i}{n_i}\right)$$

Let r be the distinct number of times at which death occurred, let $t_i, i = 1, \cdots, r$ be a time at which at least one death or failure occurred, d_i is the number of deaths that occurred at time t_i and n_i is the number of survivors up to time t_i. The times are not the order statistics in this case as one may encounter multiply censored data.

Chapter 4.1 – Current Life Table Analysis

In the US, the National Center for Health Statistics publishes detailed decennial life tables after each decennial census. These complete life tables use one-year age groups. Between censuses, annual life tables are also published. The annual life tables are often seen in five-year age intervals and are called ***abridged life tables***. Tables 4.4 and 4.5 (book) are, respectively, a complete decennial life table for the total U.S. population for 1989—1991 and an abridged life table for the same population for 1998. The abridged table in Table 4.5 was constructed based on a complete life table.

Clinical life tables have the following columns:

1. ***Age interval*** $[x, x + t)$. This is the time interval between two exact ages x and $x + t$; t is the length of the interval. For example, the interval 20—21 includes the time interval from the 20th birthday up to the 21st birthday (but not including the 21st birthday).
2. ***Proportion of deaths*** *during the interval* $q_{x-u,x}$. The information is obtained from census data.

   For example, $q_{x,x+t}$, for age interval $[x - u, x)$ is the proportion for example 20—21 is the proportion of persons who died on or after their 20th birthday and before their 21st birthday. It is an estimate of the conditional probability of dying in the interval given the person is alive at age x. This column is usually calculated from data of the decennial census of population and deaths occurring in the given time interval. For example, the mortality rates in Table 4.4 are calculated from the data of the 1990 Census of Population and deaths occurring in the United States in the three years 1989—1991. This column is the foundation of the life table from which all of the other columns are derived. In Microsoft EXCEL files I added the life table calculations for men and women articulated in one-year increments by race/ethnicity based on the 2010 Census data. It is likely to be used for insurance purposes for the next decade due to

   a. The 2020 Decennial Census was impaired due to the pandemic, political intervention, etc.

b. As of June 22, 2021, more than 600,000 US citizens died from COVID-19 and almost 500,000 were over the age of 65. While this figure is large, there are more than 49 million US citizens over 65. A crude estimate of their mortality would be about 1%. If the pandemic recurs in fall 2021 and beyond, death tolls in specific groups will cause the life (mortality) tables to change but I would not expect new life tables to be based on pandemic numbers, if the outbreak is contained.

We will use the 2010 life tables for Hispanic women to illustrate the method.

3. ***Number living at beginning of age interval***. The initial value of ℓ_x, the size of the hypothetical population, is usually 100,000 or 1,000,000. The successive values are computed using the formula

$$\ell_x = \ell_{x-u} \times \left(1 - q_{x-u,x}\right)$$

where $q_{x-u,x}$ is the proportion, which died in the preceding interval and u is the unit by which ages are separated into bins such as yearly. The life tables are set in one-year increments.

4. The number in an interval is denoted by $d_{x-u,x}$ where u is the unit that separates age groups.

$$d_{x-u,x} = \ell_x q_{x-u,x} = \ell_x - \ell_{x+u}$$

provided subjects can only be removed due to death. We will later see situations in medical studies where subjects can be lost to follow up and other reasons. In the table below, you can see the values for the current life tables of Hispanic females in the 2010 Census for the first ten years. For example, the number of persons dying during age interval 7—8, is $d_{7,8} = 9$ or $\ell_{20} q_{19,20} = 99{,}385 \times 0.000093 = 9.2428$. The values agree to the nearest integer.

5. The term, $L_x - u, x$, is the total number of years lived in the interval. In many cases one may not the exact number of years ***(to several decimal places, i.e., days)*** lived. The approximation given below assumes a uniform distribution of deaths across the year. In early years, this may be liberal, in later years, it may be conservative and in the middle years fairly reasonable.

$$L_{x-u,x} = u \times \left(\ell_x + \frac{d_{x-u,x}}{2}\right).$$

In the table given below, you will see the first ten years of the life tables for Hispanic women. The number of deaths per 100,000 Hispanic women are given together with the number of Hispanic women living in each age group. The number of deaths and births are provided through state records.

Current Life Table – Hispanic women by year
based on the 2010 Decennial Census for the first ten years

Age	$q_{x-u,x}$	ℓ_x	$d_{x-u,x}$	$L_{x-u,x}$	T_x	E_x
0	0.004974	100,000	497	99,561	8,313,014	83.1
1	0.000386	99,503	38	99,483	8,213,453	82.5
2	0.000239	99,464	24	99,452	8,113,970	81.6
3	0.000178	99,440	18	99,432	8,014,518	80.6
4	0.000151	99,423	15	99,415	7,915,086	79.6
5	0.000124	99,408	12	99,402	7,815,671	78.6
6	0.000105	99,395	10	99,390	7,716,269	77.6
7	0.000093	99,385	9	99,380	7,616,879	76.6
8	0.000089	99,376	9	99,371	7,517,499	75.6
9	0.000091	99,367	9	99,362	7,418,128	74.7
10	0.000099	99,358	10	99,353	7,318,765	73.7

$$L_{7,8} = 1 \times \left(99{,}376 + \frac{9}{2}\right) = 99380.5 \text{ or } 99{,}380.$$

The term, T_x is the total number of years lived beyond age x. This term is given by

$$T_x = \sum_{j>x} L_{x,x+j}$$

and is more readily computed recursively as $T_x = L_{x-u,x} + T_{x+u}$. For example, among Hispanic women aged 8, $T_7 = L_{7,8} + T_8 = 99{,}380 + 7{,}517{,}499 = 7{,}616{,}879$.

6. The a***verage remaining lifetime*** or average number, E_x, of years of life remaining at the beginning of age, x, is

$$E_x = \frac{T_x}{\ell_x}$$

This is also known as the ***life expectancy*** at a given age, or the ***mean residual life***.

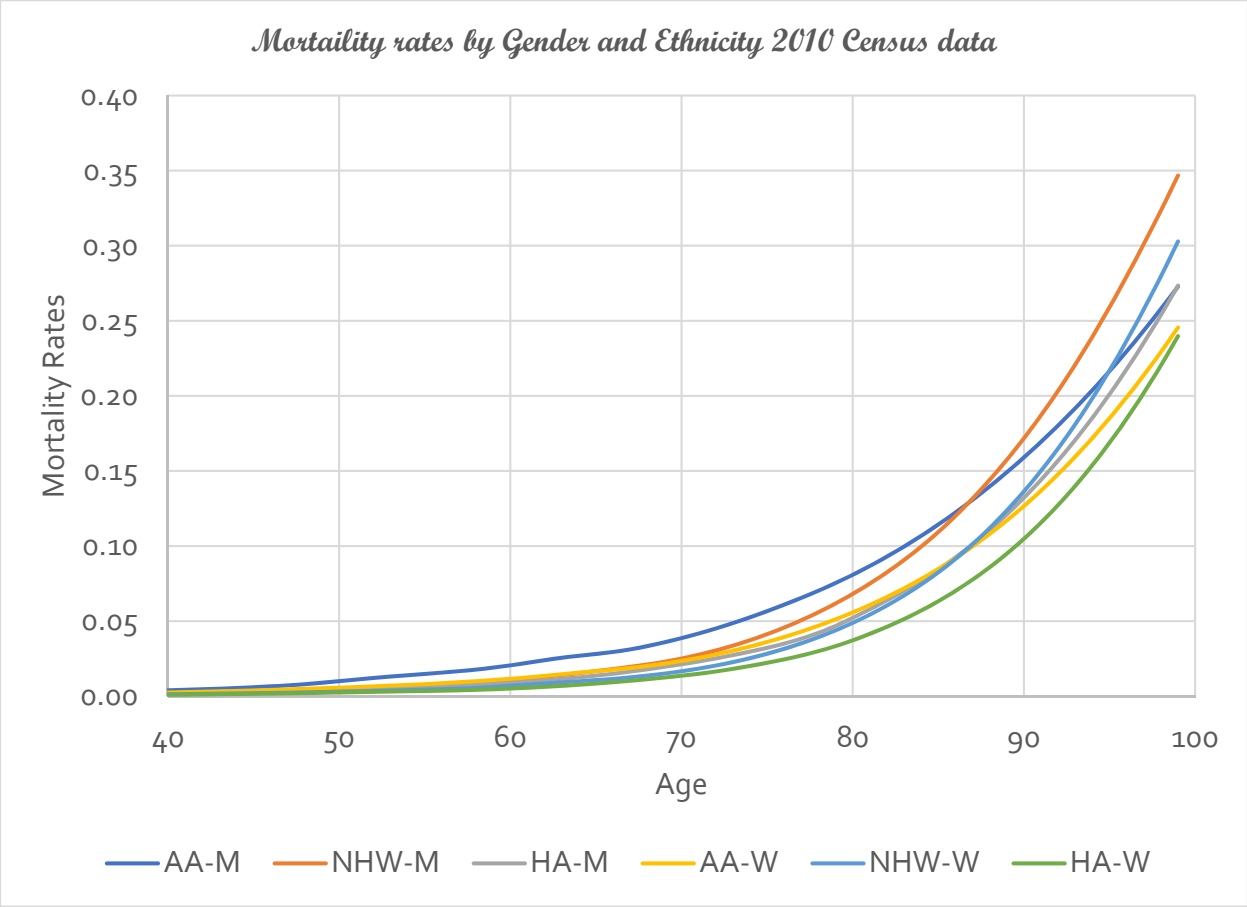
$$E_8 = \frac{T_8}{\ell_8} = \frac{7{,}517{,}499}{99{,}376} = 75.647$$

The number is the expected number of years that a Hispanic girl can expect to live on her eighth birthday. The average lifetime of this cohort is predicted to be 83.647 years.

Example: 2010 Census – Current Life Tables

In the empirical mortality graphs given below, the hazard functions are shown for people aged 40 or more. In the graphs below you will see the mortality rates, $q_{x,x+1}$, for each year by gender and ethnicity. Notice that the gender-ethnic group with the lowest mortality rate belongs to Hispanic women for all age groups beyond 40. Notice that beyond 85, the gender-ethnic group with the highest

mortality rate falls to Non-Hispanic White (NHW) men. This phenomenon is known as the ***Hispanic paradox***, wherein people with the limited access to health-care often have longer life outcomes as do Hispanic Men, Hispanic Women and African-American Women.



One can individualize these findings. As a Non-Hispanic White male aged 73, the life tables suggest that my mean residual life was 11.6 years on my last birthday. The US data in 2010 was not articulated for Asian Americans who may have the greatest longevity.

Current Life Table – Non-Hispanic men by year based on the 2010 Decennial Census aged 70-75

Age	$q_{x-u,x}$	ℓ_x	$d_{x-u,x}$	$L_{x-u,x}$	T_x	E_x
70	0.025211	72,808	1,836	71,890	987,754	13.6
71	0.027672	70,973	1,964	69,991	915,864	12.9
72	0.030548	69,009	2,108	67,955	845,873	12.3
73	0.033841	66,901	2,264	65,769	777,918	11.6
74	0.037505	64,637	2,424	63,425	712,150	11.0
75	0.041549	62,212	2,585	60,920	648,725	10.4

The World Fact book (produced by the CIA) is an excellent reference for longevity, birthrates and death rates by country. Population pyramids are given to show the age distribution of the populace of each country. One can also see that Monaco has the longest life expectancy for a newborn child at 89.40 years. The US now ranks 46th among nations worldwide in life expectancy. By contrast, Afghanistan has a life expectancy of only 53.25 years. The other 30 countries with the smallest life expectancy are in Africa.

Chapter 4.2 – Clinical Life Tables

Joseph Berkson is often recognized as a major pioneer in the development of clinical life tables. The current life table method requires large amounts of information that are usually complied from birth, death and census records. In clinical trials, one is often left with much smaller samples. Smaller sizes often eliminate the usefulness of current life tables in clinical trials. A statistician may be requested to estimate a five-year survival rate for a particular disease such as breast cancer. In this setting, patients are entered into a study upon diagnosis. The current clinical study may have patients, who have been in the study for one year, two years, etc. Likewise, they may have been in different stages of cancer progression when they entered the study. Thus, the methods of analysis and the collection of data differ from those obtained in census records.

In clinical life tables, the survival or death times are separated into

1. ***intervals or bins*** denoted by $[t_i, t_{i+1})$ with lower and upper bounds. These values may be days, weeks, months, etc. depending upon the disease under investigation. The open end of the right interval implies that deaths at time, t_{i+1}, are counted in the interval $[t_{i+1}, t_{i+2})$
2. The ***midpoint*** of the interval is denoted by

$$t_{m_i} = \frac{t_i + t_{i+1}}{2}$$

These values are essential in plotting empirical density and hazard functions.

3. The ***width*** of the interval or bin is denoted by

$$b_i = t_{i+1} - t_i$$

These values are essential in computing empirical density and hazard functions. The width of the last interval is theoretically unbounded but an upper bound is often used.

4. The ***number lost*** to follow up in the interval, $[t_i, t_{i+1})$, is denoted by l_i. These are patients who leave the study voluntarily and their survival information is unknown. The times are treated as censored observations as the patients have not died.
5. The ***number of patients withdrawn*** from the study in the interval, $[t_i, t_{i+1})$, is denoted by w_i. The survival time is listed as the length of time in the study. Patients may be withdrawn from a treatment plan if they have an adverse reaction to a medication, a secondary complication unrelated to the cancer diagnosis. These times are treated as censored observations.
6. The ***number of patients dying*** from the study in the interval, $[t_i, t_{i+1})$, is denoted by d_i. The survival times of these patients is the length of their time in the study. These times are uncensored.
7. The ***number of patients entering*** the interval, $[t_i, t_{i+1})$, is denoted by n_i'. The first entry n_1' is the total sample size of the clinical study and successive values are computed iteratively (recursively) as

$$n_i' = n_{i-1}' - l_{i-1} - w_{i-1} - d_{i-1}.$$

8. The ***number of patients exposed*** to risk during the interval, $[t_i, t_{i+1})$, is denoted by n_i. The value adjusts for patients who are lost or withdrawn during the time interval and computed as

$$n_i = n_i' - \frac{1}{2}(l_{i-1} + w_{i-1}).$$

9. The ***conditional proportion dying*** during the interval, $[t_i, t_{i+1})$, is denoted by $\hat{q}_i$ and computed as the ratio of the number of deaths in the interval to the number of patients exposed to risk in the interval.

$$\hat{q}_i = \frac{d_i}{n_i}.$$

10. The ***conditional proportion surviving*** during the interval, $[t_i, t_{i+1})$, is denoted by $\hat{p}_i$ and is the complement of the ***conditional proportion dying***. $\hat{p}_i = 1 - \hat{q}_i$.
11. The ***cumulative proportion surviving*** the interval, $[t_i, t_{i+1})$ is estimated and denoted by $\hat{S}(t_i)$ and is known as the cumulative survival. The value is computed recursively as

$$\hat{S}(t_i) = \hat{p}_{i-1} \times \hat{S}(t_{i-1}).$$

This value is similar to the one computed using the ***Kaplan-Meier*** estimator.

12. The ***probability density function*** is estimated at the midpoints, t_{m_i}, of the intervals using the relationship that the derivative of the survival function is the negative of the density function so

$$\hat{f}(t_{m_i}) = -\frac{\hat{S}(t_{i+1}) - \hat{S}(t_i)}{t_{i+1} - t_i} = -\frac{\hat{S}(t_{i+1}) - \hat{S}(t_i)}{b_i} = \frac{\hat{q}_{i-1}}{b_i} \times \hat{S}(t_i)$$

13. The ***hazard function*** is estimated at the midpoints, t_{m_i}, of the intervals as

$$\hat{h}(t_{m_i}) = \frac{d_i}{b_i\left(n_i - \frac{1}{2}d_i\right)} = \frac{2\hat{q}_i}{b_i(1+\hat{p}_i)}$$

The formula comes from the ***average number of deaths*** in an interval divided by the ***average of survivors***. Recall that the hazard function is the ratio of the density function divided by the survival function.

$$\hat{h}(t_{m_i}) = \frac{\hat{f}(t_{m_i})}{\hat{S}(t_{m_i})} = \frac{2\hat{f}(t_{m_i})}{\hat{S}(t_i) + \hat{S}(t_{i+1})}$$

An ***alternative estimate*** is given by

$$\hat{h}(t_{m_i}) = -\frac{\ln(\hat{p}_i)}{b_i}.$$

An expression for the ***variance*** of each estimators of the survival, density and hazard functions is given as

a. $Var[\hat{S}(t_i)] \cong [\hat{S}(t_i)]^2 \times \sum_{j=1}^{i-1} \frac{\hat{q}_i}{n_j \hat{p}_j}$.

b. $Var[\hat{f}(t_{m_i})] \cong \frac{[\hat{S}(t_i)\hat{q}_i]^2}{b_i} \times \left[\sum_{j=1}^{i-1}\left(\frac{\hat{q}_i}{n_j\hat{p}_j} + \frac{\hat{p}_i}{n_j\hat{q}_j}\right)\right]$.

c. $Var[\hat{h}(t_{m_i})] \cong \frac{[\hat{h}(t_{m_i})]^2}{n_j\hat{q}_i}\left\{1 - \left[\frac{\hat{h}(t_{m_i})b_i}{2}\right]^2\right\}$.

The ***median survival time*** can also be estimated as well. Let j be the interval for which $[t_j, t_{j+1})$ $\hat{S}(t_{j+1}) < 0.5$ and $\hat{S}(t_j) \geq 0.5$. The median estimated by interpolation which uses the point slope form for the equation of straight line as: $y - y_0 = m(x - x_0)$

$$\frac{1}{2} - S(t_j) = \frac{S(t_{j+1}) - S(t_j)}{t_{j+1} - t_j}(\hat{t}_m - t_j)$$

Now solving for $\hat{t}_m$ we have

$$t_m = t_j + b_j \frac{1/2 - S(t_j)}{S(t_{j+1}) - S(t_j)}$$

Then by substitution the median survival is estimated by

$$\hat{t}_m = t_j + \frac{\hat{S}(t_j) - 0.5}{\hat{f}(t_{m_i})}.$$

One can also obtain estimates of the median of the residual life function using these methods. The median life t_{mr} would be the point wherein $S(t_j) = \frac{1}{2}S(t_i)$ Let j be interval for which this is true. In practice what we do is an interpolation. So let j be the interval such

$$S(t_j) \geq \frac{1}{2}S(t_i) \text{ and } S(t_{j+1}) < \frac{1}{2}S(t_i)$$

Using the same interpolation argument given above, we have

$$t_{mr}(i) = t_j - t_i + b_j \frac{S(t_j) - 0.5 * S(t_i)}{S(t_j) - S(t_i)}.$$

The estimate, $\hat{S}(t_j)$, is the estimated proportion surviving beyond the interval that contains a 50% reduction in survival probability.

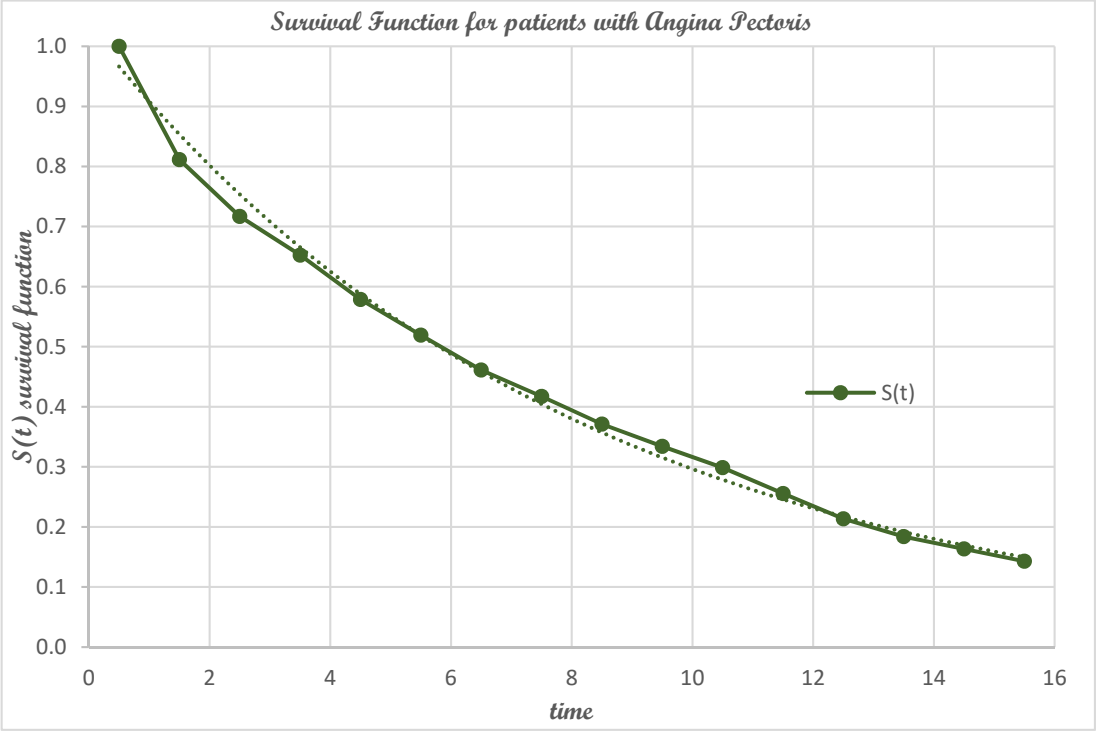
Example: Angina Pectoris – In the study of 2,418 patients with Angina Pectoris the life table is given below. In this table the number exposed in each interval is computed as are the conditional probabilities and the survival function.

beg	end	mid	Width	Lost to Follow Up	With drawn	Number Dying	Number Entering	Number Exposed	CPD	CPS	S(t)
0	1	0.5	1	0	0	456	2418	2,418.0	0.1886	0.8114	1.0000
1	2	1.5	1	39	0	226	1962	1,942.5	0.1163	0.8837	0.8114
2	3	2.5	1	22	0	152	1697	1,686.0	0.0902	0.9098	0.7170
3	4	3.5	1	23	0	171	1523	1,511.5	0.1131	0.8869	0.6524
4	5	4.5	1	24	0	135	1329	1,317.0	0.1025	0.8975	0.5786
5	6	5.5	1	107	0	125	1170	1,116.5	0.1120	0.8880	0.5193
6	7	6.5	1	133	0	83	938	871.5	0.0952	0.9048	0.4611
7	8	7.5	1	102	0	74	722	671.0	0.1103	0.8897	0.4172
8	9	8.5	1	68	0	51	546	512.0	0.0996	0.9004	0.3712
9	10	9.5	1	64	0	42	427	395.0	0.1063	0.8937	0.3342
10	11	10.5	1	45	0	43	321	298.5	0.1441	0.8559	0.2987
11	12	11.5	1	53	0	34	233	206.5	0.1646	0.8354	0.2557
12	13	12.5	1	33	0	18	146	129.5	0.1390	0.8610	0.2136
13	14	13.5	1	27	0	9	95	81.5	0.1104	0.8896	0.1839
14	15	14.5	1	23	0	6	59	47.5	0.1263	0.8737	0.1636
15	16	15.5	1	0	0	0	30	30.0	-	1.0000	0.1429

In the graphs that follow, you see nonparametric estimates of the survival, density and hazard functions for the Angina Pectoris patients. If one wants to use a parametric distribution, these graphs help one identify which distribution might be best. In this case, I added an exponential function to model the decay and the exponential survival function looks like a good fit. The graphs in the text show

a descending step function in a staircase. The linear point-to-point extension provides a better sense of the graph. The use of counts as opposed to actual survival times, one gets a raw sketch of the survival function. One can see that some distributions would not fit well any given data.

The graphical information from one survival curve may not agree with that from a density curve or a hazard function. As can be seen in the curves for the Angina Pectoris patients.

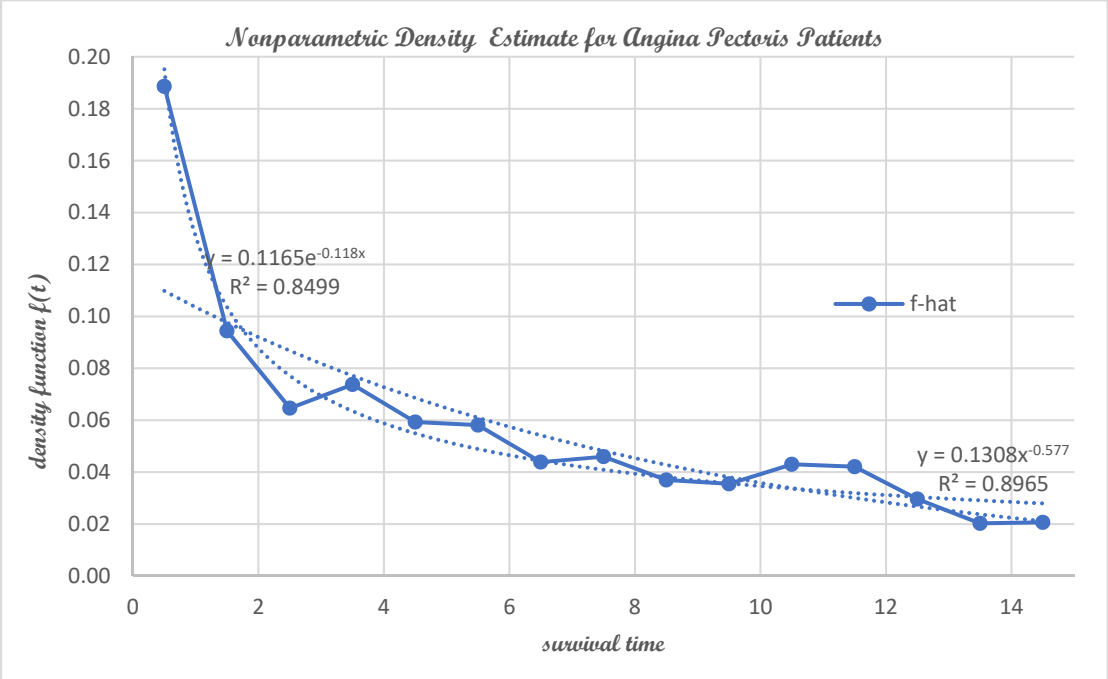


When one graphs the density function, the curve that appears to be exponential but a ***power function*** provides a fit that appears better with a higher R^2. The power function for the density suggests that a ***Pareto*** distribution would be more appropriate. The Pareto density function is given by

$$f(t) = \alpha \frac{\theta^{\alpha}}{t^{\alpha+1}} I_{(\theta,\infty)}(t).$$

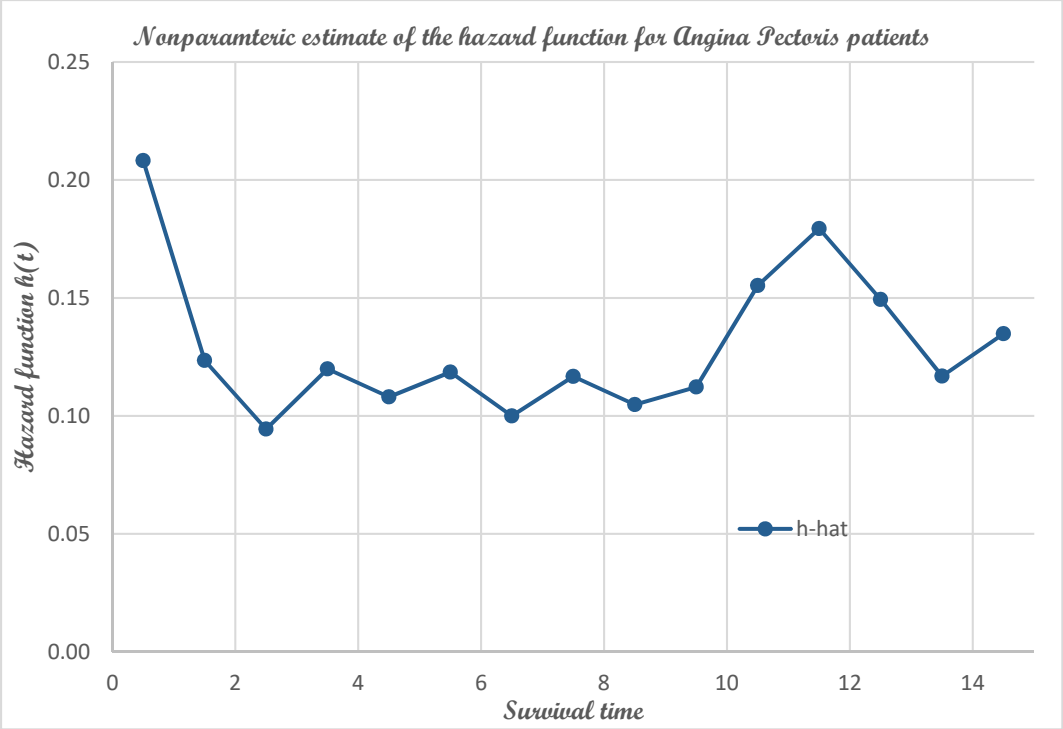
The exponential distribution does not fit as well. The importance of the discrimination often occurs in the tails distributions. Several distributions may provide very good fits to the same data set for values from the 10th to the 90th percentiles but differ in estimates in the tails of the distributions. In such cases, one needs a significant amount of data or a theoretical reason for the mechanism of death or part failure. For example, gears in a transmission can deteriorate in many different ways such as a gear surface may spall, brinell, pit, and so on but part failure more likely will come from a material fracture precipitated by one of the components of deterioration. Likewise, in health studies, as people age organs may not deteriorate at the same rate but often deterioration in one organ can lead to

deterioration in others. One may well know the fracture mechanics of failure or cardiovascular deterioration due to ischemia.



If one has the actual death and censoring times, one can attempt to fit a parametric distribution to the data and discriminate between the ***Pareto*** and the ***Exponential*** as candidates for the distribution of time to death.

The estimated hazard function, $\hat{h}(t)$, if adequate data are available, often gives one the most discriminating look at the mechanism of failure or death. In the nonparametric curve given below one can see the bath tub curve phenomenon. There appears to be an early period wherein people die at twice the rate of subsequent intervals. In the time period from 1.5 to 9.5 years the hazard function appears to oscillate about a constant, which would be indicative or exponential distribution. There is an increase in the hazard function beyond the ten-year mark at which the curve nearly doubles but then decreases after the twelve-year mark. One can use the ***Pareto*** distribution with its initially decreasing hazard rate and then the ***exponential*** model in which the failure is constant. One needs to be cautious about overfitting such a curve. Some would attempt to build piecewise functions but in doing so one may over fit the hazard function to the particular example. The piecewise function may produce an excellent fit, which reduces confidence intervals and leads one to over confident conclusions.



In the Pareto distribution, the hazard function is:

$$h(t) = \frac{\alpha}{t}, t > \theta.$$

SAS Program – Angina Pectoris

```
Title 'Survival of Males with Angina Pectoris';
DATA Males;
KEEP Freq Years Censored;
RETAIN Years -.5;
INPUT fail withdraw @@;
Years + 1;
Censored=0;
Freq=fail;
output;
Censored=1;
Freq=withdraw;
output;
datalines;
456   0 226  39  152 22 171 23 135 24 125 107
 83 133  74 102   51 68  42 64  43 45  34  53
 18  33   9  27    6 23   0 30
;
ods graphics on;
PROC LIFETEST Data=Males METHOD=LT intervals=(0 to 15 by 1)
plots=(S,LS,LLS,H,P);
time Years*Censored(1);
freq Freq;
run;
ods graphics off;
```

The SAS code above produces all the output with excellent graphics. The frequency (`freq`), command is important to preclude putting 0.5 in 456 times, and so on. The Years are created incrementally through a counter: `Years + 1.` The first year is one and each subsequent entry simply increases the previous year by one. If you use this feature you must enter the data in order. Each year is actually changed to its midpoint using the command `RETAIN Years -.5.` Notice that the command, `METHOD=LT`, invokes the use of the Life Table technique that we just discussed. The plots command produces a graph of the survival function, `S`, the log-survival function, `LS`, the log-log of the survival function, `LLS`, the hazard, `H`, and the density function, `P`.