Chapter 1
Univariate Survival Analysis

This chapter considers univariate survival analysis: there is a response variable but no predictors. In the analysis of "time to event" data, there are n individuals and the time until an event is recorded for each individual. Typical events are failure of a product or death of a person or reoccurrence of cancer after surgery, but other events such as first use of cigarettes or the time that baboons come down from trees (early in the morning) can also be modeled. The data is typically right skewed and censored data is often present.

Censoring occurs because of time and cost constraints. A product such as light bulbs may be tested for 1000 hours. Perhaps 30% fail in that time but the remaining 70% are still working. These are censored: they give partial information on the lifetime of the bulbs because it is known that about 70% last longer than 1000 hours. Handling censoring and time dependent covariates is what makes the analysis of time to event data different from other fields of statistics.

Reliability analysis is used in *engineering* to study the lifetime (time until failure) of manufactured products, while survival analysis is used in *actuarial sciences*, *statistics*, and *biostatistics* to study the lifetime (time until death) of humans, often after contracting a deadly disease. In the *social sciences*, the study of the time until the occurrence of an event is called the analysis of event time data or event history analysis. In *economics*, the study is called duration analysis or transition analysis. Hence reliability data = failure time data = lifetime data = survival data = event time data.

1.1 Functions Related to the Survival Function

In this text $\log(t) = \ln(t) = \log_e(t)$ while $\exp(t) = e^t$. One of the difficulties with survival analysis is that the response Y = survival time is usually not

observed, instead the censored response is observed. In this chapter the data will be right censored, and "right" will often be omitted. In the following definition, note that both $T \geq 0$ and $Y \geq 0$ are nonnegative.

Definition 1.1. Let $Y \geq 0$ be the time until an event occurs. Then Y is called the **survival time** or time until event. The survival time is **censored** if the event of interest has not been observed. Let Y_i be the ith survival time. Let Z_i be the time the ith observation (possibly an individual or machine) leaves the study for any reason other than the event of interest. Then Z_i is the time until the ith observation is censored. Then the **right censored survival time** T_i of the ith observation is $T_i = \min(Y_i, Z_i)$. Let $\delta_i = 0$ if T_i is (right) censored ($T_i = Z_i$) and let $\delta_i = 1$ if T_i is not censored ($T_i = Y_i$). Then the univariate survival analysis data is $(T_1, \delta_1), (T_2, \delta_2), ..., (T_n, \delta_n)$. Alternatively, the data is $T_1, T_2^*, T_3, ..., T_{n-1}^*, T_n$ where the * means that the case was (right) censored. Sometimes the asterisk * is replaced by a plus +, and Y_i, y_i or t_i can replace T_i.

In this chapter we will assume that the censoring mechanism is independent of the time to event: Y_i and Z_i are independent. Often censoring occurs because of cost and time constraints.

For example, in a study breast cancer patients who receive a lumpectomy, suppose the researchers want to keep track of 100 patients for five years after receiving a lumpectomy (tumor removal). The response is time until death after a lumpectomy. Patients who are lost to the study (move or eventually refuse to cooperate), and patients who are still alive after the study are censored. Perhaps 15% die, 5% move away and so leave the study, and 80% are still alive after 5 years. Then 85% of the cases are (right) censored. The actual study may take two years to recruit patients, follow each patient for 5 years, but end 5 years after the end of the two year recruitment period. So patients enter the study at different times, but the censored response is the time until death or censoring from the time the patient entered the study.

Definition 1.2. i) The **cumulative distribution function** (cdf) of Y is $F(t) = P(Y \leq t)$. Since $Y \geq 0$, $F(0) = 0, F(\infty) = 1$, and $F(t)$ is nondecreasing.

ii) The probability density function (**pdf**) of Y is $f(t) = F'(t)$.

iii) The **survival function** of Y is $S(t) = P(Y > t)$. $S(0) = 1, S(\infty) = 0$ and $S(t)$ is nonincreasing.

iv) The **hazard function** of Y is $h(t) = \dfrac{f(t)}{1 - F(t)}$ for $t > 0$ and $F(t) < 1$. Note that $h(t) \geq 0$ if $F(t) < 1$.

v) The **cumulative hazard function** of Y is $H(t) = \int_0^t h(u)du$ for $t > 0$. It is true that $H(0) = 0, H(\infty) = \infty$, and $H(t)$ is nondecreasing.

Assume $Y \geq 0$. Then $F(0) = 0$, $S(0) = 1$, and $H(0) = 1$. Note that $S(\infty) = 0$ implies that $H(\infty) = \infty$ where $\lim_{t\to\infty} H(t) = H(\infty)$. Memorize that $0 \leq F(t) \leq 1$, $0 \leq S(t) \leq 1$, $f(t) \geq 0$, $h(t) \geq 0$, and $H(t) \geq 0$.

Given one of $F(t), f(t), S(t), h(t)$ or $H(t)$, the following theorem shows how to find the other 4 quantities for $t > 0$. Each of these five quanities completely determines the distribution of the random variable. In reliability analysis, the reliability function $R(t) = S(t)$, and in economics, Mill's ratio $= 1/h(t)$. In actuarial sciences, $h(t)$ is the force of mortality.

Theorem 1.1.

A) $F(t) = \int_0^t f(u)du = 1 - S(t) = 1 - \exp[-H(t)] = 1 - \exp[-\int_0^t h(u)du]$.

B) $f(t) = F'(t) = -S'(t) = h(t)[1 - F(t)] = h(t)S(t) = h(t)\exp[-H(t)] = H'(t)\exp[-H(t)]$.

C) $S(t) = 1 - F(t) = 1 - \int_0^t f(u)du = \int_t^\infty f(u)du = \exp[-H(t)] = \exp[-\int_0^t h(u)du]$.

D)

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = \frac{F'(t)}{1 - F(t)} = \frac{-S'(t)}{S(t)} = \; - \; \frac{d}{dt}\log[S(t)] = H'(t).$$

E) $H(t) = \int_0^t h(u)du = -\log[S(t)] = -\log[1 - F(t)]$.

Tips: i) If $F(t) = 1 - \exp[G(t)]$ for $t > 0$, then $H(t) = -G(t)$ and $S(t) = \exp[G(t)]$.

ii) For $S(t) > 0$, note that $S(t) = \exp[\log(S(t))] = \exp[-H(t)]$. Finding $\exp[\log(S(t))]$ and setting $H(t) = -\log[S(t)]$ is easier than integrating $h(t)$.

Next an interpretation for the hazard function is given. If $P(B) > 0$, then

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} \quad \text{and} \quad P(A|B) = \frac{P(A)}{P(B)}$$

if $A \subseteq B$. Suppose the time until event is the time until death. Note that

$$P[t < Y < t + \Delta t | Y > t] = \frac{P[t < Y \leq t + \Delta t]}{P(Y > t)} = \frac{F(t + \Delta t) - F(t)}{1 - F(t)}.$$

So

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} P[t < Y \leq t + \Delta t | Y > t] = \lim_{\Delta t \to 0} \frac{\frac{F(t+\Delta t) - F(t)}{\Delta t}}{1 - F(t)}$$

$$= \frac{f(t)}{1 - F(t)} = h(t).$$

So for small Δt, it follows that $h(t)\Delta t \approx P[t < Y < t + \Delta t | Y > t] \approx$ P(person dies in interval $(t, t + \Delta t]$ given that the person has survived up to time t).

Larger $h(t)$ implies that the hazard of death is higher. The hazard function takes into account the *aging* of the observation (person or product).

For example, an 80 year old white male has about a 50% chance of living to 85 while a 100 year old white male has about a 50% chance of living to 101, although the percentage of white males living to 101 is tiny.

Example 1.1. Suppose $Y \sim EXP(\lambda)$ where $\lambda > 0$, then $h(t) = \lambda$ for $t > 0$, $f(t) = \lambda e^{-\lambda t}$ for $t > 0$, $F(t) = 1 - e^{-\lambda t}$ for $t > 0$, $S(t) = e^{-\lambda t}$ for $t > 0$, $H(t) = \lambda t$ for $t > 0$ and $E(Y) = 1/\lambda$. The **exponential distribution** can be a good model if failures are due to random shocks that follow a Poisson process (light bulbs, electrical components), but constant hazard means that a used product is as good as a new product: aging has no effect on the probability of failure of the product. The exponential distribution is the only distribution of a continuous random variable Y with a constant hazard function since $h(t)$ completely determines the distribution of the random variable Y. Derive $H(t), S(t), F(t)$, and $f(t)$ from the constant hazard function $h(t) = \lambda$ for $t > 0$ and some $\lambda > 0$.

Solution: $H(t) = \int_0^t h(u)du = \int_0^t \lambda du = \lambda t$ for $t > 0$.
$S(t) = e^{-H(t)} = e^{-\lambda t}$, for $t > 0$.
$F(t) = 1 - S(t) = 1 - e^{-\lambda t}$ for $t > 0$.
Finally, $f(t) = h(t)S(t) = \lambda e^{-\lambda t} = F'(t)$ for $t > 0$.

Suppose the observed survival times $T_1, ..., T_n$ are a censored data set from an exponential $EXP(\lambda)$ distribution. Let $T_i = Y_i^*$. Let $\delta_i = 0$ if the case is censored and let $\delta_i = 1$, otherwise. Let $r = \sum_{i=1}^n \delta_i$ = the number of uncensored cases. Then the maximum likelihood estimator (MLE) $\hat{\lambda} = r/\sum_{i=1}^n T_i$. So $\hat{\lambda} = r/\sum_{i=1}^n Y_i^*$. A 95% confidence interval (CI) for λ is $\hat{\lambda} \pm 1.96\hat{\lambda}/\sqrt{r}$. See Section 1.2.

Example 1.2. If $Y \sim$ Weibull(γ, λ) where $\gamma > 0$ and $\lambda > 0$, then $h(t) = \lambda\gamma t^{\gamma-1}$ for $t > 0$, $f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^\gamma)$ for $t > 0$, $F(t) = 1 - \exp(-\lambda t^\gamma)$ for $t > 0$, $S(t) = \exp(-\lambda t^\gamma)$ for $t > 0$, $H(t) = \lambda t^\gamma$ for $t > 0$. The Weibull$(\lambda, \gamma = 1)$ distribution is the EXP(λ) distribution. The hazard function can be increasing, decreasing or constant. Hence the **Weibull distribution** often fits reliability data well, and the Weibull distribution is an important distribution in reliability analysis. Derive $H(t), S(t), F(t)$, and $f(t)$ if $Y \sim$ Weibull(λ, γ).

Solution:

$$H(t) = \int_0^t h(u)du = \int_0^t \lambda\gamma u^{\gamma-1}du = \lambda\gamma\frac{u^\gamma}{\gamma}\Big|_0^t = \lambda t^\gamma \quad \text{for} \quad t > 0.$$

$S(t) = \exp[-H(t)] = \exp[-\lambda t^\gamma]$, for $t > 0$.
$F(t) = 1 - S(t) = 1 - \exp[-\lambda t^\gamma]$ for $t > 0$.
Finally, $f(t) = h(t)S(t) = \lambda\gamma t^{\gamma-1}\exp[-\lambda t^\gamma]$ for $t > 0$.

Recall from the central limit theorem that the sample mean $\overline{X} = \sum_{i=1}^{n} X_i/n$ is approximately normal for many distributions. For many distributions, $\min(X_1, ..., X_n)$ is approximately Weibull. Suppose a product is made of m components with iid failure times X_{im}. Suppose the product fails as soon as one of the components fails, eg a chain of links fails when the weakest link fails. Then often the failure time $Y_i = \min(X_{im}, ..., X_{im})$ is approximately Weibull.

Notation: The set $\{t : f(t) > 0\}$ is the support of Y. Often the support of Y is $(0, \infty) = t > 0$, and the formulas will omit the $t > 0$.

Theorem 1.2. $E(Y) = \int_0^\infty yf(y)dy = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt$ if $\lim_{t\to\infty} tS(t) = 0$.

1.2 Estimating the Survival Function

Notation: Let the indicator variable $I_A(Y_i) = 1$ if $Y_i \in A$ and $I_A(Y_i) = 0$ otherwise. Often write $I_{(t,\infty)}(Y_i)$ as $I(Y_i > t)$.

Definition 1.3. If none of the survival times are censored, then the **empirical survival function** $\hat{S}_E(t) =$ (number of individual with survival times $> t$)/(number of individuals) $= a/n$. So

$$\hat{S}_E(t) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i > t) = \hat{p}_t =$$

sample proportion of lifetimes $> t$.

Assume $Y_1, ..., Y_n$ are iid with $Y_i \geq 0$. Fix $t > 0$. Then $I(Y_i > t)$ are iid binomial(1,$p = P(Y_i > t)$). So $n\hat{S}_E(t) \sim$ binomial(n,$p = P(Y_i > t)$). Hence $E[n\hat{S}_E(t)] = nP(Y > t)$ and $V[n\hat{S}_E(t)] = nS(t)F(t)$. Thus $E[\hat{S}_E(t)] = S(t)$ and $V[\hat{S}_E(t)] = S(t)F(t)/n = [S(t)(1-S(t))]/n \leq 0.25/n$. Thus $SD[\hat{S}_E(t)] = \sqrt{V[\hat{S}_E(t)]} \leq 0.5/\sqrt{n}$. So need $n \approx 100$ for $SD[\hat{S}_E(t)] < 0.05$.

Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ be the observed ordered survival times (= lifetimes = death times). Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times. Let $d_i =$ number of deaths at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are **no ties**. If $m < n$ and some $d_i \geq 2$, then there are **ties**.

Then $\hat{S}_E(t)$ is a step function with $\hat{S}_E(0) = 1$ and $\hat{S}_E(t) = \hat{S}_E(t_{i-1})$ for $t_{i-1} \leq t < t_i$. Note that $\sum_{i=1}^{m} d_i = n$. Know how to compute and plot $\hat{S}_E(t)$ given the $t_{(i)}$ or given the t_i and d_i. Use a table like the one below. Let

$a_0 = n$ and $a_i = \sum_{k=1}^{n} I(T_i > t_i) = \#$ of cases $t_{(j)} > t_i$ for $i = 1, ..., m$. Then $\hat{S}_E(t_i) = a_i/n = \sum_{k=1}^{n} I(T_i > t_i)/n = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$.

t_i	d_i	$\hat{S}_E(t_i) = \hat{S}_E(t_{i-1}) - \frac{d_i}{n} = a_i/n$
$t_0 = 0$		$\hat{S}_E(0) = 1 = \frac{n}{n} = \frac{a_0}{n}$
t_1	d_1	$\hat{S}_E(t_1) = \hat{S}_E(t_0) - \frac{d_1}{n} = \frac{a_0 - d_1}{n} = \frac{a_1}{n}$
t_2	d_2	$\hat{S}_E(t_2) = \hat{S}_E(t_1) - \frac{d_2}{n} = \frac{a_1 - d_2}{n} = \frac{a_2}{n}$
$\vdots$	$\vdots$	$\vdots$
t_j	d_j	$\hat{S}_E(t_j) = \hat{S}_E(t_{j-1}) - \frac{d_j}{n} = \frac{a_{j-1} - d_j}{n} = \frac{a_j}{n}$
$\vdots$	$\vdots$	$\vdots$
t_{m-1}	d_{m-1}	$\hat{S}_E(t_{m-1}) = \hat{S}_E(t_{m-2}) - \frac{d_{m-1}}{n} = \frac{a_{m-2} - d_{m-1}}{n} = \frac{a_{m-1}}{n}$
t_m	d_m	$\hat{S}_E(t_m) = 0 = \hat{S}_E(t_{m-1}) - \frac{d_m}{n} = \frac{a_{m-1} - d_m}{n} = \frac{a_m}{n}$

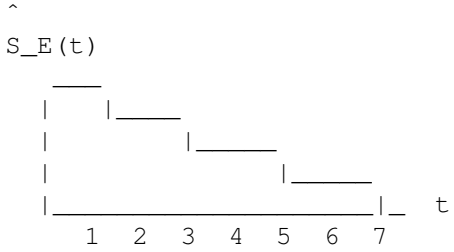
Let $\hat{S}(t)$ be the estimated survival function. Let $t(p)$ be the pth percentile of Y: $P(Y \leq t(p)) = F(t(p)) = p$ so $1 - p = S(t(p)) = P(Y > t(p))$. Then $\hat{t}(p)$, the estimated time when 100 p % have died, can be estimated from a graph of $\hat{S}(t)$ with "over" and "down" lines. a) Find $1-p$ on the vertical axis and draw a horizontal "over" line to $\hat{S}(t)$. Draw a vertical "down" line until it intersects the horizontal axis at $\hat{t}(p)$. Usually want $p = 0.5$ but sometimes $p = 0.25$ and $p = 0.75$ are used.

Example 1.3. Smith (2002, p. 68) gives steroid induced remission times for leukemia patients. The $t_{(j)}$, $t - i$ and d_i are given in the following table. The a_i and $\hat{S}_E(t)$ needed to be computed. Note that $a_i = \#$ of cases with $t_{(j)} > t_i$. For the following table, the 2nd column $t_{(j)}$ gives the 21 ordered survival times. The 3rd column t_i gives the distinct ordered survival times. Often just the number is given, so $t_1 = 1$ would be replaced by 1. The 4th column d_i tells how many events (remissions) occurred at time t_i and the last column computes $\hat{S}_E(t_i)$. A good check is that the 1st column entry divided by n is equal to $a_i/n = \hat{S}_E(t_i) =$ last column entry. A graph of the estimated survival function would be a step function with times 0, 1, ..., 23 on the horizontal axis and $\hat{S}_E(t)$ on the vertical axis. A convention is to draw vertical lines at the jumps (at the t_i). So the step function would be 1 on

(0,1), 19/21 on (1,2), ..., 1/21 on (22,23) and 0 for $t > 23$. The vertical lines connecting the steps are at $t = 1, 2, ..., 23$.

a_i	$t_{(j)}$	t_i	d_i	$\hat{S}_E(t_i) = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$
21		$t_0 = 0$		$\hat{S}_E(0) = 1 = 21/21$
	1			
19	1	$t_1 = 1$	2	$\hat{S}_E(1) = (21-2)/21 = 19/21$
	2			
17	2	$t_2 = 2$	2	$\hat{S}_E(2) = (19-2)/21 = 17/21$
16	3	$t_3 = 3$	1	$\hat{S}_E(3) = (17-1)/21 = 16/21$
	4			
14	4	$t_4 = 4$	2	$\hat{S}_E(4) = (16-2)/21 = 14/21$
	5			
12	5	$t_5 = 5$	2	$\hat{S}_E(5) = (14-2)/21 = 12/21$
	8			
	8			
	8			
8	8	$t_6 = 8$	4	$\hat{S}_E(8) = (12-4)/21 = 8/21$
	11			
6	11	$t_7 = 11$	2	$\hat{S}_E(11) = (8-2)/21 = 6/21$
	12			
4	12	$t_8 = 12$	2	$\hat{S}_E(12) = (6-2)/21 = 4/21$
3	15	$t_9 = 15$	1	$\hat{S}_E(15) = (4-1)/21 = 3/21$
2	17	$t_{10} = 17$	1	$\hat{S}_E(17) = (3-1)/21 = 2/21$
1	22	$t_{11} = 22$	1	$\hat{S}_E(22) = (2-1)/21 = 1/21$
0	23	$t_{12} = 23$	1	$\hat{S}_E(23) = (1-1)/21 = 0$ good check

Example 1.4. If $d_i = 1, 1, 1, 1$ and if $t_i = 1, 3, 5, 7$, then $a_1 = 3$, $a_2 = 2$ and $a_3 = 1$. Hence $\hat{S}_E(1) = 0.75, \hat{S}_E(3) = 0.5, \hat{S}_E(5) = 0.25$, and $\hat{S}_E(7) = 0$, and the estimated survival function is graphed as below.



Let $t_1 \leq t < t_m$. Then the **classical large sample 95% CI** for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\hat{S}_E(t_c) \pm 1.96\sqrt{\frac{\hat{S}_E(t_c)[1-\hat{S}_E(t_c)]}{n}} = \hat{S}_E(t_c) \pm 1.96 SE[\hat{S}_E(t_c)] = [L, U].$$

Use $[\max(0, L), \min(1, U)]$.

Let $0 < t$. Let

$$\tilde{p}_{t_c} = \frac{n\hat{S}_E(t_c) + 2}{n+4}.$$

Then the **plus four 95% CI** for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\tilde{p}_{t_c} \pm 1.96\sqrt{\frac{\tilde{p}_{t_c}[1-\tilde{p}_{t_c}]}{n+4}} = \tilde{p}_{t_c} \pm 1.96SE[\tilde{p}_{t_c}] = [L, U].$$

Use $[\max(0, L), \min(1, U)]$. For this CI, four imaginary T_i^* are added to the sample, with two of the $T_i^* > t_m > t_c$ and two $< t_1 < t_c$. See Agresti and Coull (1998).

The 95% large sample CI $\hat{S}_E(t_c) \pm 1.96SE[\tilde{p}_{t_c}]$ is also interesting.

Example 1.5. Let $n = 21$ and $\hat{S}_E(12) = 4/21$.

a) Find the 95% classical CI for $\hat{S}_E(12)$.

b) Find the 95% plus four CI for $\hat{S}_E(12)$.

Solution: a)

$$\frac{4}{21} + 1.96\sqrt{\frac{\frac{4}{21}(1-\frac{4}{21})}{21}} = \frac{4}{21} \pm 0.16795 = [0.0225, 0.3584].$$

b)

$$\tilde{p}_{12} = \frac{21\frac{4}{21} + 2}{21+4} = \frac{6}{25}.$$

So the 95% CI is

$$\frac{6}{25} + 1.96\sqrt{\frac{\frac{6}{25}(1-\frac{6}{25})}{25}} = \frac{6}{25} \pm 0.16742 = [0.0726, 0.4074].$$

Note that the CIs are not very short since $n = 21$ is small.

Let Y_i = time to event for ith person. $T_i = \min(Y_i, Z_i)$ where Z_i is the censoring time for the ith person (the time the ith person is lost to the study for any reason other than the time to event under study). The censored data is $y_1, y_2+, y_3, ..., y_{n-1}, y_n+$ where y_i means the time was uncensored and y_i+ means the time was censored. Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ be the ordered survival times (so if y_4+ is the smallest survival time, then $t_{(1)} = y_4+$). A status variable will be 1 if the time was uncensored and 0 if censored.

Let $[t_0, t_m) = I_1 \cup I_2 \cup \cdots \cup I_m = [t_0, t_1) \cup [t_1, t_2) \cdots \cup [t_{m-1}, t_m)$ where $t_0 \geq 0$ and $t_m = \infty$ is possible. It is possible that the 1st interval will have left endpoint $t_0 > 0$ and the last interval will have finite right endpoint $t_m < \infty$. Suppose that the following quantities are known: d_j = # deaths in I_j,
c_j = # of censored survival times in I_j, and
n_j = # at risk in I_j = # who were alive and not yet censored at the start of

I_j (at time t_{j-1}). Note that $n_1 = n$ and $n_j = n_{j-1} - d_{j-1} - c_{j-1}$ for $j > 1$. This equation shows how those at risk in the $(j-1)$th interval propagate to the jth interval.
Let $n'_j = n_j - \frac{c_j}{2}$ = average number at risk in I_j.

Definition 1.4. The **lifetable estimator** or actuarial method estimator of $S_Y(t)$ takes $\hat{S}_L(0) = 1$ and

$$\hat{S}_L(t_k) = \prod_{j=1}^{k} \frac{n'_j - d_j}{n'_j} = \prod_{j=1}^{k} \tilde{p}_j$$

for $k = 1, ..., m-1$. If $t_m = \infty$, $\hat{S}_L(t)$ is undefined for $t > t_{m-1}$. Suppose $t_m \neq \infty$. Then take $\hat{S}_L(t) = 0$ for $t \geq t_m$ if $c_m = 0$. If $c_m > 0$, then $\hat{S}_L(t)$ is undefined for $t \geq t_m$. (Some programs use $\hat{S}_L(t) = 0$ for $t \geq t_m$ if $t_m \neq \infty$.)

To graph $\hat{S}_L(t)$, use linear interpolation (connect the dots). If $n'_j = 0$, take $\tilde{p}_j = 0$. Note that

$$\hat{S}_L(t_k) = \hat{S}_L(t_{k-1}) \frac{n'_k - d_k}{n'_k} \quad \text{for} \quad \mathrm{k} = 1, ..., \mathrm{m} - 1.$$

The lifetable estimator is used to estimate $S_Y(t) = P(Y > t)$ when there is censoring. Also, the actual event or censoring times are unknown, but the number of event and censoring times in each interval I_j is known for $j = 1, ..., m$. Let $p_j = \mathrm{P}(\text{surviving through } I_j | \text{ alive at the start of } I_j) = P(Y > t_j | Y > t_{j-1}) = \frac{P(Y > t_j, Y > t_{j-1})}{P(Y > t_{j-1})} = \frac{S(t_j)}{S(t_{j-1})}$. Now $p_1 = S(t_1)/S(t_0) = S(t_1)$ since $S(0) = S(t_0) = 1$. Writing $S(t_k)$ as a telescoping product gives

$$S(t_k) = S(t_1) \frac{S(t_2)}{S(t_1)} \frac{S(t_3)}{S(t_2)} \cdots \frac{S(t_{k-1})}{S(t_{k-2})} \frac{S(t_k)}{S(t_{k-1})} = p_1 p_2 \cdots p_k = \prod_{j=1}^{k} p_j.$$

Let $\hat{p}_j = 1 -$ (number dying in I_j)/(number with potential to die in I_j). Then $\tilde{p}_j = 1 - d_j/n'_j$ is the estimate of p_j used by the lifetable estimator, assuming that the censoring is roughly uniform over each interval.

Know how to get the lifetable estimator and $SE(\hat{S}_L(t_i))$ from output.

```
 (left output)                        (right output)
interval survival survival SE or interval survival survival SE
0    50   1.00        0            0    50  0.7594    0.0524
50  100   0.7594    0.0524        50   100  0.5889    0.0608
100 200   0.5889    0.0608       100   200  0.5253    0.0602
```

Since $\hat{S}_L(0) = 1$, $\hat{S}_L(t)$ is for the left endpoint for the "left output", and for the right endpoint for the "right output." For both cases, $\hat{S}_L(50) = 0.7594$ and $SE(\hat{S}_L(50)) = 0.0524$.

A 95% CI for $S_Y(t_i)$ based on the lifetable estimator is

$$\hat{S}_L(t_i) \pm 1.96 \; SE[\hat{S}_L(t_i)].$$

Hence for the above output, a 95% CI for $S_Y(50)$ is $0.7594 \pm 1.96(0.0524) = [0.6567, 0.8621]$.

Know how to compute $\hat{S}_L(t)$ with a table like the one below. The first 4 entries need to be given but the last 3 columns may need to be filled in. On an exam you may be given a table with all but a few entries filled.

I_j, d_j, c_j, n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[t_0 = 0, t_1)$, d_1, c_1, n_1	$n_1 - \frac{c_1}{2}$	$\frac{n'_1 - d_1}{n'_1}$	$\hat{S}_L(t_0) = \hat{S}_L(0) = 1$
$[t_1, t_2)$, d_2, c_2, n_2	$n_2 - \frac{c_2}{2}$	$\frac{n'_2 - d_2}{n'_2}$	$\hat{S}_L(t_1) = \hat{S}_L(t_0)\frac{n'_1 - d_1}{n'_1}$
$[t_2, t_3)$, d_3, c_3, n_3	$n_3 - \frac{c_3}{2}$	$\frac{n'_3 - d_3}{n'_3}$	$\hat{S}_L(t_2) = \hat{S}_L(t_1)\frac{n'_2 - d_2}{n'_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[t_{k-1}, t_k)$, d_k, c_k, n_k	$n_k - \frac{c_k}{2}$	$\frac{n'_k - d_k}{n'_k}$	$\hat{S}_L(t_{k-1}) =$ $\hat{S}_L(t_{k-2})\frac{n'_{k-1} - d_{k-1}}{n'_{k-1}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[t_{m-2}, t_{m-1})$, d_{m-1}, c_{m-1}, n_{m-1}	$n_{m-1} - \frac{c_{m-1}}{2}$	$\frac{n'_{m-1} - d_{m-1}}{n'_{m-1}}$	$\hat{S}_L(t_{m-2}) =$ $\hat{S}_L(t_{m-3})\frac{n'_{m-2} - d_{m-2}}{n'_{m-2}}$
$[t_{m-1}, t_m = \infty)$, d_m, c_m, n_m	$n_m - \frac{c_m}{2}$	$\frac{n'_m - d_m}{n'_m}$	$\hat{S}_L(t_{m-1}) =$ $\hat{S}_L(t_{m-2})\frac{n'_{m-1} - d_{m-1}}{n'_{m-1}}$

Also get a 95% CI from output like that below. So the 95% CI for S(50) is [0.65666,0.86213] $\approx$ [0.6567,0.8621].

```
time survival SDF_LCL SDF_UCL
0     1.0     1.0     1.0
50    0.7594  0.65666 0.86213
```

Example 1.6. Allison (1995, p. 49-51) gives time until death after heart transplant for 68 patients. The 1st 5 columns are given, but the last 3 columns need to be computed. Use 4 digits in the computations.

I_j	t_j	d_j	c_j	n_j	$n'_j = n_j - c_j/2$	$\tilde{p}_j = \frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t_j) = \hat{S}_L(t_{j-1})\tilde{p}_j$
[0,50)	0	16	3	68	66.5	0.7594	$\hat{S}(0) = 1$
[50,100)	50	11	0	49	49	0.7755	$\hat{S}(50) = 0.7594$
[100,200)	100	14	2	38	37	0.8919	$\hat{S}(100) = 0.5889$
[200,400)	200	5	4	32	30	0.8333	$\hat{S}(0) = 0.5252$
[400,700)	400	2	6	23	20	0.90	$\hat{S}(400) = 0.4376$
[700,1000)	700	4	3	15	13.5	0.7037	$\hat{S}(700) = 0.7037$
[1000,1300)	1000	1	2	8	7	0.8571	$\hat{S}(1000) = 0.2771$
[1300,1600)	1300	1	3	5	3.5	0.7143	$\hat{S}(1300) = 0.2375$
[1600,∞)	1600	0	1	1	0.5	1.0	$\hat{S}(1600) = 0.1696$

Greenwood's formula is

$$SE[\hat{S}_L(t_j)] = \hat{S}_L(t_j)\sqrt{\sum_{i=1}^{j} \frac{1-\tilde{p}_i}{\tilde{p}_i n'_i}}$$

where $j = 1, ..., m-1$. The formula is best computed using software.

Now suppose the data is censored but the event or censoring times T_i are known with $Y_i^* = T_i = \min(Y_i, Z_i)$ where Y_i and Z_i are independent. Let $\delta_i = I(Y_i \leq Z_i)$ so $\delta_i = 1$ if T_i is uncensored and $\delta_i = 0$ if T_i is censored. Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ be the observed ordered survival times. Let $\gamma_j = 1$ if $t_{(j)}$ is uncensored and 0, otherwise. Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times corresponding to the $t_{(j)}$ with $\gamma_j = 1$. Let $d_i =$ number of events (deaths) at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are **no ties**. If $m < n$ and some $d_i \geq 2$, then there are **ties**. Let $n_i = \sum_{j=1}^{n} I(t_{(j)} \geq t_i) =$ # at risk at $t_i =$ # alive and not yet censored just before t_i (and just after t_{i-1}).

Example 1.7. Suppose $n = 6$, $Y_i \sim EXP(1), E(Y_i) = 1, Z_i \sim EXP(0.1)$, and $E(Z_i) = 10$. In the table below, Y_i and Z_i are not observed, $m = 5$, and the observed data is T_i and δ_i.

Y_i	0.2887	0.1796	1.1301	1.4165	0.2720	0.6667
Z_i	0.8967	1.6158	10.5266	1.0520	2.2329	4.2917
$T_i = Y_i^*$	0.2887	0.1796	1.1301	1.0520	0.2720	0.6667
δ_i	1	1	1	0	1	1
$t_{(j)}$	0.1796	0.2720	0.2887	0.6667	1.0522	1.1301
γ_j	1	1	1	1	0	1
t_i	0.1796	0.2720	0.2887	0.6667		1.1301

Consider intervals $I_1 = (0, t_1], I_2 = (t_1, t_2], ..., I_m = (t_{m-1}, t_m]$. Let n_k be the number at risk for interval I_k, $d_k =$ number of deaths in $I_k =$ number of deaths at t_k, and

$$\hat{p}_k = 1 - \frac{d_k}{n_k} = 1 - \frac{\text{number dying in I}_\text{k}}{\text{number with potential to die in I}_\text{k}} \approx \frac{S(t_k)}{S(t_{k-1})} \approx$$

P(survive in interval $(t_{k-1}, t_k)|$ alive at start of I_k). Then

$$\hat{S}_K(t_i) = \prod_{k=1}^{i} \hat{p}_k.$$

Note that individuals who die or are censored at time t_k are "at risk at t_k."

Definition 1.5. The **Kaplan Meier estimator = product limit estimator** of $S_Y(t_i) = P(Y > t_i)$ is $\hat{S}_K(0) = 1$ and

$$\hat{S}_K(t_i) = \prod_{k=1}^{i} (1 - \frac{d_k}{n_k}) = \hat{S}_K(t_{i-1})(1 - \frac{d_i}{n_i}).$$

$\hat{S}_K(t)$ is a step function with $\hat{S}_K(t) = \hat{S}_K(t_{i-1})$ for $t_{i-1} \leq t < t_i$ and $i = 1, ..., m$. If $t_{(n)}$ is uncensored then $t_m = t_{(n)}$ and $\hat{S}_K(t) = 0$ for $t > t_m$. If $t_{(n)}$ is censored, then $\hat{S}_K(t) = \hat{S}_K(t_m)$ for $t_m \leq t \leq t_{(n)}$, but $\hat{S}_K(t)$ is undefined for $t > t_{(n)}$.

Know how to compute and plot $\hat{S}_k(t_i)$ given the $t_{(j)}$ and γ_j or given the t_i, n_i and d_i. Use a table like the one below. Let $n_0 = n$. If $f_{i-1} =$ number of events (deaths) and number censored in time interval $[t_{i-1}, t_i)$, then $n_i = n_{i-1} - f_{i-1} =$ number of $t_{(j)} \geq t_i$.

t_i	n_i	d_i	$\hat{S}_K(t)$
$t_0 = 0$			$\hat{S}_K(0) = 1$
t_1	n_1	d_1	$\hat{S}_K(t_1) = \hat{S}_K(t_0)[1 - \frac{d_1}{n_1}]$
t_2	n_2	d_2	$\hat{S}_K(t_2) = \hat{S}_K(t_1)[1 - \frac{d_2}{n_2}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_j	n_j	d_j	$\hat{S}_K(t_j) = \hat{S}_K(t_{j-1})[1 - \frac{d_j}{n_j}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_{m-1}	n_{m-1}	d_{m-1}	$\hat{S}_K(t_{m-1}) = \hat{S}_K(t_{m-2})[1 - \frac{d_{m-1}}{n_{m-1}}]$
t_m	n_m	d_m	$\hat{S}_K(t_m) = 0 = \hat{S}_K(t_{m-1})[1 - \frac{d_m}{n_m}]$

Example 1.8. Modifying Smith (2002, p. 113) slightly, suppose that the ordered censored survival times in days until repair of $n = 13$ street lights is 36, 38, 38, 38+, 78 112, 112, 114+, 162+, 189, 198, 237, 489+.

f_j	$t_{(j)}$	γ_j	t_i	n_i	d_i	$\hat{S}(t)$
						$\hat{S}(0) = 1$
1	36	1	36	13	1	$\hat{S}(36) = 0.9231$
3	38	1	38	12	2	$\hat{S}(38) = 0.7692$
	38	1				
	38	0				
1	78	1	78	9	1	$\hat{S}(78) = 0.6837$
4	112	1	112	8	2	$\hat{S}(112) = 0.5128$
	112	1				
	114	0				
	162	0				
1	189	1	189	4	1	$\hat{S}(189) = 0.3846$
1	198	1	198	3	1	$\hat{S}(198) = 0.2564$
1	237	1	237	2	1	$\hat{S}(36) = 0.1282$
	489	0				

Know how to find a 95% CI for $S_Y(t_i)$ based on $\hat{S}_K(t_i)$ using output: the 95% CI is $\hat{S}_K(t_i) \pm 1.96\ SE[\hat{S}_K(t_i)]$. The R output below gives $t_i, n_i, d_i, \hat{S}_K(t_i)$, $SE(\hat{S}_K(t_i))$ and the 95% CI for $S_Y(36)$ is $[0.7782, 1]$.

```
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   36     13       1    0.923  0.0739       0.7782        1.000
```

In general, a 95% CI for $S_Y(t_i)$ is $\hat{S}(t_i) \pm 1.96\ SE[\hat{S}(t_i)]$. If the lower endpoint of the CI is negative, round it up to 0. If the upper endpoint of the CI is greater than 1, round it down to 1. **Do not use impossible values of** $S_Y(t)$. Note that $\hat{S}_K(36) \pm 1.96 SE[\hat{S}_K(36)] = 0.923 \pm 1.96(0.0793) = 0.923 \pm 0.1448 = [0.7782, 1.0678]$. So use [0.7782, 1] since $S(y) \in [0, 1]$.

Let $P(Y \leq t(p)) = F_Y(t(p)) = p$ for $0 < p < 1$. Be able to get $t(p)$ and 95% CIs for $t(p)$ from SAS output for $p = 0.25, 0.5, 0.75$. For the output below, the CI for $t(0.75)$ is not given. The 95% CI for $t(0.50) \approx 210$ is [63,1296]. The 95% CI for $t(0.25) \approx 63$ is [18,195].

```
Quartile estimates
Percent point estimate lower upper
75         .            220.0  .
50       210.00         63.00 1296.00
25        63.00         18.00 195.00
```

R plots the KM survival estimator along with the pointwise 95% CIs for $S_Y(t)$. If we guess a distribution for Y, say $Y \sim W$, with a formula for $S_W(t)$, then the guessed $S_W(t_i)$ can be added to the plot. If roughly 95% of the $S_W(t_i)$ fall within the bands, then $Y \sim W$ may be reasonable. For example, if $W \sim EXP(1)$, use $S_W(t) = \exp(-t)$. If $W \sim EXP(\lambda)$, then $S_W(t) = \exp(-\lambda t)$. Recall that $E(W) = 1/\lambda$.

If $\lim_{t\to\infty} tS_Y(t) \to 0$, then $E(Y) = \int_0^\infty t f_Y(t)dt = \int_0^\infty S_Y(t)dt$. Hence an estimate of the mean $\hat{E}(Y)$ can be obtained from the area under $\hat{S}(t)$.

Greenwood's formula is

$$SE[\hat{S}_K(t_j)] = \hat{S}_K(t_j)\sqrt{\sum_{i=1}^{j} \frac{d_j}{n_j(n_j - d_j)}}$$

where $j = 1, ..., m-1$. The formula is best computed using software.

Definition 1.6. The **Nelson Aelen estimator** of $S_Y(t)$ is

$$\hat{S}_N(t_i) = \prod_{k=1}^{i} \exp\left(\frac{-d_k}{n_k}\right) = \exp\left(-\sum_{k=1}^{i} \frac{d_k}{n_k}\right) = \hat{S}_N(t_{i-1}) \exp\left(-\frac{d_i}{n_i}\right)$$

where $\hat{S}_N(0) = 1$ and t_i, d_i, and n_i are the same as for the Kaplan Meier estimator.

1.3 Estimating the (Cumulative) Hazard Function

Two important estimators of the cumulative hazard function use $\hat{H}(t_i) = -\log(\hat{S}(t_i))$.

Definition 1.7. The Kaplan Meier estimator of $H_Y(t)$ is $\hat{H}_K(0) = 0$,

$$\hat{H}_K(t_i) = -\log(\hat{S}_k(t_i) = -\sum_{k=1}^{i} \log\left(1 - \frac{d_k}{n_k}\right) = \hat{H}_K(t_{i-1}) - \log\left(1 - \frac{d_i}{n_i}\right).$$

Definition 1.8. The Nelson Aelen estimator of $H_Y(t)$ is $\hat{H}_N(0) = 0$,

$$\hat{H}_N(t_i) = \sum_{k=1}^{i} \frac{d_k}{n_k} = \hat{H}_K(t_{i-1}) + \frac{d_i}{n_i}.$$

Note that $\hat{S}_N(t_i) = \exp(-\hat{H}_n(t_i))$ and $\hat{H}_N(t_i) = -\log(\hat{S}_N(t_i))$.

A 95% CI for $H_Y(t_i)$ is $\hat{H}(t_i) \pm 1.96 SE[\hat{H}(t_i)] = [L, U]$. Use $[\max(0, L), U]$. Also,

$$SE[\hat{H}_N(t_i)] = \sqrt{\sum_{k=1}^{i} \frac{d_k}{n_k^2}} = \sqrt{SE[\hat{H}_N(t_{i-1})] + \frac{d_i}{n_i^2}}.$$

For the hazard function with $t_0 = 0$,

$$\hat{h}_K(t_i) = \hat{h}_N(t_i) = \frac{d_i}{n_i(t_{i+1} - t_i)}$$

for $i = 1, ..., m-1$.

Example 1.9.

t_i	n_i	d_i	$\hat{h}_K(t_i)$
10	18	1	$\frac{1}{18(19-10)} = 0.00617$
19	15	1	$\frac{1}{15(30-19)} = 0.00606$
30	13	1	

For the life table estimator with interval $I_j = [t_{j-1}, t_j)$, d_j, and n'_j,

$$\hat{h}_L(t) = \frac{d_j}{(n'_j - \frac{d_j}{2})(t_j - t_{j-1})}$$

$t_{j-1} \le t < t_j$ with $\hat{h}_L(t)$ undefined for the last interval $[t_{m-1}, t_m)$. Sometime $t^* = (t_{j-1} + t_j)/2$ is used.

Example 1.10.

I_j	t^*	d_j	n'_j	$\hat{h}_L(t^*)$
$[0, 50)$	25	16	66.5	$\frac{16}{(66.5 - 16/2)(50 - 0)} = 0.00547$
$[50, 100)$	75	11	49	$\frac{11}{(49 - 11/2)(100 - 50)} = 0.005058$

Example 1.11. The data is from Klein and Moeschberger (2002, pp. 2, 86). There were 21 children with acute leukemia in complete or partial remission induced by the drug Prednisone, and the children were given the drug over a six month period. Note that $t_0 = 0$, $t_{(j)}$ = time until relapse, and $n_j = \sum_j t_{(j)} \geq t_i$. See the following table for computations. Using that table, a 95% CI for $H_Y(13)$ is $\hat{H}_N(13) \pm 1.96 SE[\hat{H}_n(13)] = 0.3517 \pm 1.96\sqrt{0.0217} = 0.3517 \pm 0.2888 = [0.0630, 0.6404]$.

$t_{(j)}$	γ_j	t_i	n_i	d_i	$\hat{H}_N(t_i)$	$(SE[\hat{H}_N(t_i)])^2$
		$t_0 = 0$			0	
6	1	6	21	3	$0 + 3/21 = 0.1428$	$0 + 3/21^2 = 0.0068$
6	1					
6	1					
6	0					
7	1	1	17	1	$0.1428 + 1/17 = 0.2017$	$0.0068 + 1/17^2 = 0.0103$
9	0					
10	1	10	15	1	$0.2017 + 1/15 = 0.2683$	$0.0103 + 1/15^2 = 0.0147$
10	0					
11	0					
13	1	13	12	1	$0.2683 + 1/12 = 0.3517$	$0.0147 + 1/12^2 = 0.0217$
16	1	16	11	1	$0.3517 + 1/11 = 0.4426$	$0.0217 + 1/11^2 = 0.0299$
17	0					
19	0					
20	0					
22	1	22	7	1	$0.4426 + 1/7 = 0.5854$	$0.0299 + 1/7^2 = 0.2243$
23	1	23	6	1	$0.5854 + 1/6 = 0.7521$	$0.2244 + 1/6^2 = 0.2795$
25	0					
32	0					
32	0					
34	0					
35	0					

1.4 Maximum Likelihood Estimation

Definition 1.9. Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ be the pdf of a sample $\boldsymbol{Y}$ with parameter space Θ. If $\boldsymbol{Y} = \boldsymbol{y}$ is observed, then the **likelihood function** is $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y}) =$

$f(\boldsymbol{y}|\boldsymbol{\theta})$. For each sample point $\boldsymbol{y} = (y_1, ..., y_n)$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{y}) \in \Theta$ be a parameter value at which $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y})$ attains its maximum as a function of $\boldsymbol{\theta}$ with $\boldsymbol{y}$ held fixed. Then a maximum likelihood estimator (**MLE**) of the parameter $\boldsymbol{\theta}$ based on the sample $\boldsymbol{Y}$ is $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$.

The following remarks are important. I) It is crucial to observe that the likelihood function is a function of $\boldsymbol{\theta}$ (and that $y_1, ..., y_n$ act as fixed constants). Note that the pdf or pmf $f(\boldsymbol{y}|\boldsymbol{\theta})$ is a function of n variables while $L(\boldsymbol{\theta})$ is a function of k variables if $\boldsymbol{\theta}$ is a $1 \times k$ vector. Often $k = 1$ or $k = 2$ while n could be in the hundreds or thousands.

II) If $Y_1, ..., Y_n$ is an independent sample from a population with pdf or pmf $g(y|\boldsymbol{\theta})$, then the likelihood function

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|y_1, ..., y_n) = \prod_{i=1}^{n} g(y_i|\boldsymbol{\theta}). \tag{1.1}$$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} g_i(y_i|\boldsymbol{\theta})$$

if the Y_i are independent but have different pdfs or pmfs.

III) If the MLE $\hat{\boldsymbol{\theta}}$ exists, then $\hat{\boldsymbol{\theta}} \in \Theta$. Hence if $\hat{\boldsymbol{\theta}}$ is not in the parameter space Θ, then $\hat{\boldsymbol{\theta}}$ is not the MLE of $\boldsymbol{\theta}$.

Theorem 1.3: Invariance Principle. If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then $\tau(\hat{\boldsymbol{\theta}})$ is the MLE of $\tau(\boldsymbol{\theta})$ where τ is a function with domain Θ.

Really just need $\Theta \in dom(\tau)$ so $tau(\hat{\theta})$ is well defined: can't have $\log(-7.89)$ or $\sqrt{-1.57}$.

There are **four commonly used techniques** for finding the MLE.

- Potential candidates can be found by differentiating $\log L(\boldsymbol{\theta})$, the log likelihood.
- Potential candidates can be found by differentiating the likelihood $L(\boldsymbol{\theta})$.
- The MLE can sometimes be found by direct maximization of the likelihood
- $L(\boldsymbol{\theta})$.
- **Invariance Principle**: If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then $\tau(\hat{\boldsymbol{\theta}})$ is the MLE of $\tau(\boldsymbol{\theta})$.

The one parameter case can often be solved by hand with the following technique. To show that $\hat{\theta}$ is the MLE of θ is equivalent to showing that $\hat{\theta}$ is the global maximizer of $\log L(\theta)$ on Θ where Θ is an interval with endpoints a and b, not necessarily finite. Suppose that $\log L(\theta)$ is continuous on Θ. Show that $\log L(\theta)$ is differentiable on (a, b). Then show that $\hat{\theta}$ is the unique

solution to the equation $\frac{d}{d\theta} \log L(\theta) = 0$ and that the 2nd derivative evaluated at $\hat{\theta}$ is negative: $\frac{d^2}{d\theta^2} \log L(\theta)\Big|_{\hat{\theta}} < 0$. See Remark 1.1V below.

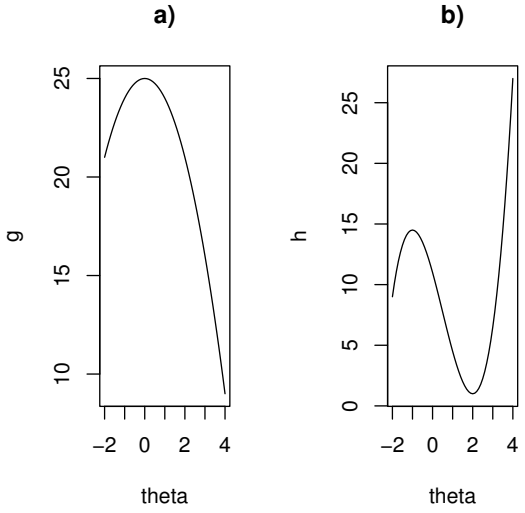


Fig. 1.1 The local max in a) is a global max, but the local max at $\theta = -1$ in b) is not the global max.

Remark 1.1. From calculus, recall the following facts. I) If the function g is continuous on an interval $[a, b]$ then both the max and min of g exist. Suppose that g is continuous on an interval $[a, b]$ and differentiable on (a, b). Solve $g'(\theta) \equiv 0$ and find the places where $g'(\theta)$ does not exist. These values are the **critical points**. Evaluate g at a, b, and the critical points. One of these values will be the min and one the max.

II) Assume g is continuous. Then g has a local max at the critical point θ_o if g is increasing for $\theta < \theta_o$ in a neighborhood of θ_o and if g is decreasing for $\theta > \theta_o$ in a neighborhood of θ_o (and θ_o is a global max if you can remove the phrase "in a neighborhood of θ_o"). The first derivative test is often used: if g is continuous at θ_o and if there exists some $\delta > 0$ such that $g'(\theta) > 0$ for all θ in $(\theta_o - \delta, \theta_o)$ and $g'(\theta) < 0$ for all θ in $(\theta_o, \theta_o + \delta)$, then g has a local max at θ_o.

III) If g is strictly concave ($\frac{d^2}{d\theta^2} g(\theta) < 0$ for all $\theta \in \Theta$), then any local max of g is a global max.

IV) Suppose $g'(\theta_o) = 0$. The 2nd derivative test states that if $\frac{d^2}{d\theta^2} g(\theta_o) < 0$, then g has a local max at θ_o.

V) If $g(\theta)$ is a continuous function on an interval with endpoints $a < b$ (not necessarily finite), differentiable on (a, b) and if the **critical point is unique**, then the critical point is a **global maximum** if it is a local maximum. To see this claim, note that if the critical point is not the global max then there would be a local minimum and the critical point would not be unique. Let $a = -2$ and $b = 4$. In Figure 1.1 a), the critical point for $g(\theta) = -\theta^2 + 25$ is at $\theta = 0$, is unique, and is both a local and global maximum. In Figure 1.1 b), $h(\theta) = \theta^3 - 1.5\theta^2 - 6\theta + 11$, the critical point $\theta = -1$ is not unique and is a local max but not a global max.

VI) If g is strictly convex ($\frac{d^2}{d\theta^2} g(\theta) > 0$ for all $\theta \in \Theta$), then any local min of g is a global min. If $g'(\theta_o) = 0$, then the 2nd derivative test states that if $\frac{d^2}{d\theta^2} g(\theta_o) > 0$, then θ_o is a local min.

VII) If $g(\theta)$ is a continuous function on an interval with endpoints $a < b$ (not necessarily finite), differentiable on (a, b) and if the **critical point is unique**, then the critical point is a **global minimum** if it is a local minimum. To see this claim, note that if the critical point is not the global min then there would be a local maximum and the critical point would not be unique.

Tips: a) $\exp(a) = e^a$ and $\log(y) = \ln(y) = \log_e(y)$ is the **natural logarithm**.
b) $\log(a^b) = b \log(a)$ and $\log(e^b) = b$.
c) $\log(\prod_{i=1}^n a_i) = \sum_{i=1}^n \log(a_i)$.
d) $\log L(\theta) = \log(\prod_{i=1}^n f(y_i|\theta)) = \sum_{i=1}^n \log(f(y_i|\theta))$.
e) If t is a differentiable function and $t(\theta) \neq 0$, then $\frac{d}{d\theta} \log(|t(\theta)|) = \frac{t'(\theta)}{t(\theta)}$ where $t'(\theta) = \frac{d}{d\theta} t(\theta)$. In particular, $\frac{d}{d\theta} \log(\theta) = 1/\theta$.
f) Any additive term that does not depend on θ is treated as a constant with respect to θ and hence has derivative 0 with respect to θ.

With censoring and truncation, the likelihood function changes. Often $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|y_1, ..., y_n) = \prod_{i=1}^n L(\boldsymbol{\theta}|y_i)$. Note that $1 - F(w) = S(w)$.

a) For iid individual data, $L(\boldsymbol{\theta}|y_i) = f(y_i)$ if Y has pdf $f(y)$.

b) For iid individual data, $L(\boldsymbol{\theta}|y_i) = p(x_i)$ if Y has pmf $p(y)$.

c) If it is only known that y_i is in some interval $(c_{j-1}, c_j]$, then $L(\boldsymbol{\theta}|y_i) = P(y_i \in (c_{j-1}, c_j]) = F(c_j) - F(c_{j-1})$.
The endpoints can be open or closed if Y is from a continuous distribution.

d) If y_i is right censored at u_i, then the interval is $[u_i, \infty)$, and $L(\boldsymbol{\theta}|y_i) = 1 - F(u_i)$.

e) For grouped data from the table below, $L(\boldsymbol{\theta}) = \prod_{j=1}^{m} [F(c_j) - F(c_{j-1})]^{n_j}$.

interval	number
$(c_0, c_1]$	n_1
$(c_1, c_2]$	n_2
$(c_2, c_3]$	n_3
$\vdots$	$\vdots$
$(c_{m-2}, c_{m-1}]$	n_{m-1}
$(c_{m-1}, c_m]$	n_m

f) If y_i is left truncated at d_i, then $L(\boldsymbol{\theta}|y_i) = \dfrac{f(y_i)}{1 - F(d_i)}$.

g) If y_i is left truncated at d_i and right censored at u_i, then $L(\boldsymbol{\theta}|y_i) = \dfrac{1 - F(u_i)}{1 - F(d_i)}$.

h) If the data are left truncated at d with $n - k$ uncensored cases y_i and k cases right censored at u, then $L(\boldsymbol{\theta}) = \dfrac{[\prod_{i=1}^{n-k} f(y_i)][1 - F(u)]^k}{[1 - F(d)]^n}$.

i) (**Rare**, the interval is $(0, d]$.): If y_i is censored below at d, $L(\boldsymbol{\theta}|y_i) = F(d)$.

j) (**Rare**): If y_i is truncated above at u, $L(\boldsymbol{\theta}|x_i) = \dfrac{f(y_i)}{F(u)}$.

Note that left truncated = truncated below = truncated, and right censored = censored above = censored are often used.

1.5 Simulations for KM Confidence Intervals

Section 1.2 described confidence intervals for the Kaplan Meier estimator. We will describe another CI, and two more CIs are easy to compute with R. Then we will simulate the four CIs.

The Agresti and Coull (1998) plus four 95% CI adds two successes (deaths) and two failures (survives) to the data set from a binomial distribution, and then computes the classical binomial 95% CI from the modified data set. For $t \in [t_1, t_m]$, Olive (2010, problem 16.45) modifies this procedure by adding two artificial deaths just before time t_1 and two artificial censored observations after the largest death time t_m. Then the classical 95% CI for the Kaplan Meier estimator is computed from the modified data set.

Hence

$$\tilde{S}_K(t_i) = \left(1 - \frac{1}{n+4}\right)\left(1 - \frac{1}{n+3}\right)\prod_{k=1}^{i}\left(1 - \frac{d_k}{n_k + 2}\right)$$

for $i = 1, ..., m$ where the first two terms are due to the two artificial deaths at the just before t_1 and $n_k + 2$ is used in the product due to the two artificial cases censored at time t_m. Also $[SE(\tilde{S}_K(t_i)]^2 =$

$$[\tilde{S}_K(t_i)]^2 \left(\sum_{k=1}^{i} \frac{d_k}{(n_k+2)(n_k+2-d_k)} + \frac{1}{(n+4)(n+4-1)} + \frac{1}{(n+3)(n+3-1)} \right)$$

for $i = 1, ..., m-1$.

If the CI is initially [L,U], then the CI $[\max(0, L), \min(1, U)]$ is used. In addition to the classical Kaplan Meier CI, there is a log CI that uses $log(\hat{S})$ and a log–log CI that uses $\log(-\log(\hat{S}))$ that are easy to compute with software.

Simulations were done in *R*. The function *kmsim2* simulates the classical, log, log–log, and plus four CIs for the Kaplan Meier estimator and is in the collection of *R* functions *survpack*. See Yang (2016) for a bigger simulation. The plus four CI worked well for $S(t_{(1)})$ and $S(t_{(n)})$.

The program *kmsim2* computes censored data $T = \min(Y, Z)$ where $Y \sim EXP(1)$. Then a 95% CI is made for $S_Y(t_{(j)})$ for each of the n $t_{(j)}$. This is done for runs=5000 data sets and the program computes the proportion of times the CI contains $S_Y(t_{(j)}) = \exp(-t_{(j)})$. The average scaled CI lengths (the average of $\sqrt{n}$ CI length) are also computed. The ccov is the proportion for the classical $\hat{S} \pm 1.96SE(\hat{S})$ interval while p4cov is for the plus 4 CI. The lcov is based on a CI that uses $\log(\hat{S})$ and llcov is based on a CI that uses $\log(-\log(\hat{S}))$. The three classical CIs are not made if the last case is censored so NA is given. The plus four CI seems to be good at $t_{(1)}$ and $t_{(n)}$. With 5000 runs, coverage between 0.94 and 0.96 would not give much evidence that the coverage is different from the nominal covarage of 0.95.

```
library(survival)
kmsim2(n=10,runs=5000)
$ccov
 [1] 0.8852 0.9604 0.9736 0.9720 0.9666 0.9544 0.9380
     0.9062 0.8404  NA

$lcov
 [1] 0.8772 0.9470 0.9564 0.9618 0.9632 0.9670 0.9702
     0.9800 0.9828  NA

$llcov
 [1] 0.7694 0.8886 0.9130 0.9222 0.9242 0.9230 0.9258
     0.9246 0.9208  NA

$p4cov
 [1] 0.9978 0.9082 0.9090 0.9132 0.9200 0.9236 0.9330
     0.9410 0.9550 0.9734
```

```
$clen
 [1] 0.8213907 1.3221304 1.7054981 1.8938355 1.9760212
     1.9803150 1.9032412 1.5986898 1.0969514       NA

$llen
 [1] 0.7698268 1.2214843 1.5940815 1.9111395 2.0769800
     2.1522128 2.1692379 2.1519330 2.2099754        NA

$lllen
 [1] 1.471560 1.679038 1.776042 1.826047 1.832765
     1.791306 1.692973 1.526673 1.264845       NA

$p4len
 [1] 1.327469 1.471418 1.569004 1.632534 1.665829
     1.669772 1.641687 1.578521 1.470567 1.189487
```

The above output is for $n = 10$ with 5000 runs. The table below summarizes the CI coverages and scaled lengths for t_1, t_3, t_{n-2}, and t_{n-1}.

Table 1.1 Simulated CI Coverages and Scaled Lengths

n	t_i	cov/len	clas	log	loglog	plus4
10	t_1	cov	0.885	0.877	0.769	0.998
		len	0.821	0.770	1.471	1.327
10	t_3	cov	0.974	0.956	0.913	0.909
		len	1.705	1.594	1.776	1.569
10	t_{n-2}	cov	0.906	0.980	0.925	0.941
		len	1.599	2.512	1.527	1.579
10	t_{n-1}	cov	0.840	0.983	0.921	0.955
		len	1.097	2.210	1.265	1.470

1.6 Summary

Let $Y \geq 0$ be a nonnegative random variable.

Then the **cumulative distribution function** (cdf) $F(t) = P(Y \leq t)$. Since $Y \geq 0$, $F(0) = 0, F(\infty) = 1$, and $F(t)$ is nondecreasing.

The probability density function (**pdf**) $f(t) = F'(t)$.

The **survival function** $S(t) = P(Y > t)$. $S(0) = 1, S(\infty) = 0$ and $S(t)$ is nonincreasing.

The **hazard function** $h(t) = \dfrac{f(t)}{1 - F(t)}$ for $t > 0$ and $F(t) < 1$. Note that $h(t) \geq 0$ if $F(t) < 1$.

The **cumulative hazard function** $H(t) = \int_0^t h(u)du$ for $t > 0$. It is true that $H(0) = 0, H(\infty) = \infty$, and $H(t)$ is nondecreasing.

1) Given one of $F(t), f(t), S(t), h(t)$ or $H(t)$, be able to find the other 4 quantities for $t > 0$.

A) $F(t) = \int_0^t f(u)du = 1 - S(t) = 1 - \exp[-H(t)] = 1 - \exp[-\int_0^t h(u)du]$.

B) $f(t) = F'(t) = -S'(t) = h(t)[1 - F(t)] = h(t)S(t) = h(t)\exp[-H(t)] = H'(t)\exp[-H(t)]$.

C) $S(t) = 1 - F(t) = 1 - \int_0^t f(u)du = \int_t^\infty f(u)du = \exp[-H(t)] = \exp[-\int_0^t h(u)du]$.

D)

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = \frac{F'(t)}{1 - F(t)} = \frac{-S'(t)}{S(t)} = - \frac{d}{dt}\log[S(t)] = H'(t).$$

E) $H(t) = \int_0^t h(u)du = -\log[S(t)] = -\log[1 - F(t)]$.

Tip: if $F(t) = 1 - \exp[G(t)]$ for $t > 0$, then $H(t) = -G(t)$ and $S(t) = \exp[G(t)]$.

Tip: For $S(t) > 0$, note that $S(t) = \exp[\log(S(t))] = \exp[-H(t)]$. Finding $\exp[\log(S(t))]$ and setting $H(t) = -\log[S(t)]$is easier than integrating $h(t)$.

Know that if $Y \sim EXP(\lambda)$ where $\lambda > 0$, then $h(t) = \lambda$ for $t > 0$, $f(t) = \lambda e^{-\lambda t}$ for $t > 0$, $F(t) = 1 - e^{-\lambda t}$ for $t > 0$, $S(t) = e^{-\lambda t}$ for $t > 0$, $H(t) = \lambda t$ for $t > 0$ and $E(T) = 1/\lambda$. The **exponential distribution** can be a good model if failures are due to random shocks that follow a Poisson process, but constant hazard means that a used product is as good as a new product.

2) Suppose the observed survival times $T_1, ..., T_n$ are a censored data set from an exponential (λ) distribution. Let $T_i = Y_i^*$. Let $\delta_i = 0$ if the case is censored and let $\delta_i = 1$, otherwise. Let $r = \sum_{i=1}^n \delta_i$ = the number of uncensored cases. Then the MLE $\hat{\lambda} = r/\sum_{i=1}^n T_i$. So $\hat{\lambda} = r/\sum_{i=1}^n Y_i^*$. A 95% CI for λ is $\hat{\lambda} \pm 1.96\hat{\lambda}/\sqrt{r}$.

Know that if $Y \sim$ Weibull(λ, γ) where $\lambda > 0$ and $\gamma > 0$, then $h(t) = \lambda\gamma t^{\gamma-1}$ for $t > 0$, $f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^\gamma)$ for $t > 0$, $F(t) = 1 - \exp(-\lambda t^\gamma)$ for $t > 0$, $S(t) = \exp(-\lambda t^\gamma)$ for $t > 0$, $H(t) = \lambda t^\gamma$ for $t > 0$. The Weibull$(\lambda, \gamma = 1)$ distribution is the EXP(λ) distribution. The hazard function can be increasing, decreasing or constant. Hence the **Weibull distribution** often fits reliability data well, and the Weibull distribution is the most important distribution in reliability analysis.

3) Let $\hat{S}(t)$ be the estimated survival function. Let $t(p)$ be the pth percentile of Y: $P(Y \leq t(p)) = F(t(p)) = p$ so $1 - p = S(t(p)) = P(Y > t(p))$. Then $\hat{t}(p)$, the estimated time when 100 p % have died, can be estimated from a graph of $\hat{S}(t)$ with "over" and "down" lines. a) Find $1 - p$ on the vertical axis and draw a horizontal "over" line to $\hat{S}(t)$. Draw a vertical "down" line until it intersects the horizontal axis at $\hat{t}(p)$. Usually want $p = 0.5$ but sometimes $p = 0.25$ and $p = 0.75$ are used.

The **indicator function** $I_A(x) \equiv I(x \in A) = 1$ if $x \in A$ and 0, otherwise. Sometimes an indicator function such as $I_{(0,\infty)}(y)$ will be denoted by $I(y > 0)$.

If none of the survival times are censored, then the **empirical survival function** = (number of individual with survival times $> t$)/(number of individuals) $= a/n =$

$$\hat{S}_E(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > t) = \hat{p}_t = \text{sample proportion of lifetimes} > t.$$

Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ be the observed ordered survival times (= lifetimes = death times). Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times. Let d_i = number of deaths at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are **no ties**. If $m < n$ and some $d_i \geq 2$, then there are **ties**.

$\hat{S}_E(t)$ is a step function with $\hat{S}_E(0) = 1$ and $\hat{S}_E(t) = \hat{S}_E(t_{i-1})$ for $t_{i-1} \leq t < t_i$. Note that $\sum_{i=1}^{m} d_i = n$.

4) Know how to compute and plot $\hat{S}_E(t)$ given the $t_{(i)}$ or given the t_i and d_i. Use a table like the one below. Let $a_0 = n$ and $a_i = \sum_{k=1}^{n} I(T_i > t_i) = \#$ of cases $t_{(j)} > t_i$ for $i = 1, ..., m$. Then $\hat{S}_E(t_i) = a_i/n = \sum_{k=1}^{n} I(T_i > t_i)/n = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$.

t_i	d_i	$\hat{S}_E(t_i) = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$
$t_0 = 0$		$\hat{S}_E(0) = 1 = \frac{n}{n} = \frac{a_0}{n}$
t_1	d_1	$\hat{S}_E(t_1) = \hat{S}_E(t_0) - \frac{d_1}{n} = \frac{a_0 - d_1}{n} = \frac{a_1}{n}$
t_2	d_2	$\hat{S}_E(t_2) = \hat{S}_E(t_1) - \frac{d_2}{n} = \frac{a_1 - d_2}{n} = \frac{a_2}{n}$
$\vdots$	$\vdots$	$\vdots$
t_j	d_j	$\hat{S}_E(t_j) = \hat{S}_E(t_{j-1}) - \frac{d_j}{n} = \frac{a_{j-1} - d_j}{n} = \frac{a_j}{n}$
$\vdots$	$\vdots$	$\vdots$
t_{m-1}	d_{m-1}	$\hat{S}_E(t_{m-1}) = \hat{S}_E(t_{m-2}) - \frac{d_{m-1}}{n} = \frac{a_{m-2} - d_{m-1}}{n} = \frac{a_{m-1}}{n}$
t_m	d_m	$\hat{S}_E(t_m) = 0 = \hat{S}_E(t_{m-1}) - \frac{d_m}{n} = \frac{a_{m-1} - d_m}{n} = \frac{a_m}{n}$

5) Let $t_1 \leq t < t_m$. Then the **classical large sample 95% CI** for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\hat{S}_E(t_c) \pm 1.96\sqrt{\frac{\hat{S}_E(t_c)[1 - \hat{S}_E(t_c)]}{n}} = \hat{S}_E(t_c) \pm 1.96 SE[\hat{S}_E(t_c)].$$

6) Let $0 < t$. Let

$$\tilde{p}_{t_c} = \frac{n\hat{S}_E(t_c) + 2}{n + 4}.$$

Then the **plus four 95% CI** for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\tilde{p}_{t_c} \pm 1.96\sqrt{\frac{\tilde{p}_{t_c}[1 - \tilde{p}_{t_c}]}{n + 4}} = \tilde{p}_{t_c} \pm 1.96 SE[\tilde{p}_{t_c}].$$

Let Y_i = time to event for ith person. $T_i = \min(Y_i, Z_i)$ where Z_i is the censoring time for the ith person (the time the ith person is lost to the study for any reason other than the time to event under study). The censored data is $y_1, y_2+, y_3, ..., y_{n-1}, y_n+$ where y_i means the time was uncensored and y_i+ means the time was censored. $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ are the ordered survival times (so if y_4+ is the smallest survival time, then $t_{(1)} = y_4+$). A status variable will be 1 if the time was uncensored and 0 if censored.

Let $[0, \infty) = I_1 \cup I_2 \cup \cdots \cup I_m = [t_0, t_1) \cup [t_1, t_2) \cdots \cup [t_{m-1}, t_m)$ where $t_o = 0$ and $t_m = \infty$. It is possible that the 1st interval will have left endpoint > 0 ($t_0 > 0$) and the last interval will have finite right endpoint ($t_m < \infty$). Suppose that the following quantities are known: d_j = # deaths in I_j,
c_j = # of censored survival times in I_j,

$n_j = \#$ at risk in $I_j = \#$ who were alive and not yet censored at the start of I_j (at time t_{j-1}).
Let $n'_j = n_j - \frac{c_j}{2}$ = average number at risk in I_j.

7) The **lifetable estimator** or actuarial method estimator of $S_Y(t)$ takes $\hat{S}_L(0) = 1$ and

$$\hat{S}_L(t_k) = \prod_{j=1}^{k} \frac{n'_j - d_j}{n'_j} = \prod_{j=1}^{k} \tilde{p}_j$$

for $k = 1, ..., m-1$. If $t_m = \infty$, $\hat{S}_L(t)$ is undefined for $t > t_{m-1}$. Suppose $t_m \neq \infty$. Then take $\hat{S}_L(t) = 0$ for $t \geq t_m$ if $c_m = 0$. If $c_m > 0$, then $\hat{S}_L(t)$ is undefined for $t \geq t_m$. **To graph** $\hat{S}_L(t)$, use linear interpolation (connect the dots). If $n'_j = 0$, take $\tilde{p}_j = 0$. Note that

$$\hat{S}_L(t_k) = \hat{S}_L(t_{k-1}) \frac{n'_k - d_k}{n'_k} \text{ for } \mathrm{k} = 1, ..., \mathrm{m} - 1.$$

8) Know how to get the lifetable estimator and $SE(\hat{S}_L(t_i))$ from output.

```
(left output)                              (right output)
interval survival survival SE or interval survival survival SE
0    50   1.00        0          0    50   0.7594   0.0524
50   100  0.7594    0.0524       50   100  0.5889   0.0608
100  200  0.5889    0.0608      100   200  0.5253   0.0602
```

Since $\hat{S}_L(0) = 1$, $\hat{S}_L(t)$ is for the left endpoint for the "left output," and for the right endpoint for the "right output." For both cases, $\hat{S}_L(50) = 0.7594$ and $SE(\hat{S}_L(50)) = 0.0524$.

9) A 95% CI for $S_Y(t_i)$ based on the lifetable estimator is

$$\hat{S}_L(t_i) \pm 1.96 \; SE[\hat{S}_L(t_i)].$$

10) Know how to compute $\hat{S}_L(t)$ with a table like the one below. The first 4 columns need to be given but the last 3 columns may need to be filled in. On an exam you may be given a table with all but a few entries filled.

I_j, d_j, c_j, n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[t_0 = 0, t_1)$, d_1, c_1, n_1	$n_1 - \frac{c_1}{2}$	$\frac{n'_1 - d_1}{n'_1}$	$\hat{S}_L(t_o) = \hat{S}_L(0) = 1$
$[t_1, t_2)$, d_2, c_2, n_2	$n_2 - \frac{c_2}{2}$	$\frac{n'_2 - d_2}{n'_2}$	$\hat{S}_L(t_1) = \hat{S}_L(t_0)\frac{n'_1 - d_1}{n'_1}$
$[t_2, t_3)$, d_3, c_3, n_3	$n_3 - \frac{c_3}{2}$	$\frac{n'_3 - d_3}{n'_3}$	$\hat{S}_L(t_2) = \hat{S}_L(t_1)\frac{n'_2 - d_2}{n'_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[t_{k-1}, t_k)$, d_k, c_k, n_k	$n_k - \frac{c_k}{2}$	$\frac{n'_k - d_k}{n'_k}$	$\hat{S}_L(t_{k-1}) =$ $\hat{S}_L(t_{k-2})\frac{n'_{k-1} - d_{k-1}}{n'_{k-1}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[t_{m-2}, t_{m-1})$, d_{m-1}, c_{m-1}, n_{m-1}	$n_{m-1} - \frac{c_{m-1}}{2}$	$\frac{n'_{m-1} - d_{m-1}}{n'_{m-1}}$	$\hat{S}_L(t_{m-2}) =$ $\hat{S}_L(t_{m-3})\frac{n'_{m-2} - d_{m-2}}{n'_{m-2}}$
$[t_{m-1}, t_m = \infty)$, d_m, c_m, n_m	$n_m - \frac{c_m}{2}$	$\frac{n'_m - d_m}{n'_m}$	$\hat{S}_L(t_{m-1}) =$ $\hat{S}_L(t_{m-2})\frac{n'_{m-1} - d_{m-1}}{n'_{m-1}}$

11) Also get a 95% CI from output like that below. So the 95% CI for S(50) is (0.65666,0.86213).

```
time survival SDF_LCL SDF_UCL
0    1.0      1.0     1.0
50   0.7594   0.65666 0.86213
```

Let $Y_i^* = T_i = \min(Y_i, Z_i)$ where Y_i and Z_i are independent. Let $\delta_i = I(Y_i \leq Z_i)$ so $\delta_i = 1$ if T_i is uncensored and $\delta_i = 0$ if T_i is censored. Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ be the observed ordered survival times. Let $\gamma_j = 1$ if $t_{(j)}$ is uncensored and 0, otherwise. Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times corresponding to the $t_{(j)}$ with $\gamma_j = 1$. Let $d_i =$ number of deaths at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are **no ties**. If $m < n$ and some $d_i \geq 2$, then there are **ties**.

12) Let $n_i = \sum_{j=1}^{n} I(t_{(j)} \geq t_i) = \#$ at risk at $t_i = \#$ alive and not yet censored just before t_i. Let $d_i = \#$ of events (deaths) at t_i. The **Kaplan Meier estimator** = **product limit estimator** of $S_Y(t_i) = P(Y > t_i)$ is $\hat{S}_K(0) = 1$ and $\hat{S}_K(t_i) = \prod_{k=1}^{i}(1 - \frac{d_k}{n_k}) = \hat{S}_K(t_{i-1})(1 - \frac{d_i}{n_i})$. $\hat{S}_K(t)$ is a step function with $\hat{S}_K(t) = \hat{S}_K(t_{i-1})$ for $t_{i-1} \leq t < t_i$ and $i = 1, ..., m$. If $t_{(n)}$ is uncensored then $t_m = t_{(n)}$ and $\hat{S}_K(t) = 0$ for $t > t_m$. If $t_{(n)}$ is censored, then $\hat{S}_K(t) = \hat{S}_K(t_m)$ for $t_m \leq t \leq t_{(n)}$, but $\hat{S}_K(t)$ is undefined for $t > t_{(n)}$.

13) Know how to compute and plot $\hat{S}_k(t_i)$ given the $t_{(j)}$ and γ_j or given the t_i, n_i and d_i. Use a table like the one below.

t_i	n_i	d_i	$\hat{S}_K(t)$
$t_0 = 0$			$\hat{S}_K(0) = 1$
t_1	n_1	d_1	$\hat{S}_K(t_1) = \hat{S}_K(t_0)[1 - \frac{d_1}{n_1}]$
t_2	n_2	d_2	$\hat{S}_K(t_2) = \hat{S}_K(t_1)[1 - \frac{d_2}{n_2}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_j	n_j	d_j	$\hat{S}_K(t_j) = \hat{S}_K(t_{j-1})[1 - \frac{d_j}{n_j}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_{m-1}	n_{m-1}	d_{m-1}	$\hat{S}_K(t_{m-1}) = \hat{S}_K(t_{m-2})[1 - \frac{d_{m-1}}{n_{m-1}}]$
t_m	n_m	d_m	$\hat{S}_K(t_m) = 0 = \hat{S}_K(t_{m-1})[1 - \frac{d_m}{n_m}]$

14) Know how to find a 95% CI for $S_Y(t_i)$ based on $\hat{S}_K(t_i)$ using output: the 95% CI is $\hat{S}_K(t_i) \pm 1.96\ SE[\hat{S}_K(t_i)]$. The R output below gives $t_i, n_i, d_i, \hat{S}_K(t_i), SE(\hat{S}_K(t_i))$ and the 95% CI for $S_Y(36)$ is $(0.7782, 1)$.

```
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   36     13       1    0.923  0.0739       0.7782        1.000
```

15) In general, a 95% CI for $S_Y(t_i)$ is $\hat{S}(t_i) \pm 1.96\ SE[\hat{S}(t_i)]$. If the lower endpoint of the CI is negative, round it up to 0. If the upper endpoint of the CI is greater than 1, round it down to 1. **Do not use impossible values of** $S_Y(t)$.

16) Let $P(Y \leq t(p)) = p$ for $0 < p < 1$. Be able to get $t(p)$ and 95% CIs for $t(p)$ from SAS output for $p = 0.25, 0.5, 0.75$. For the output below, the CI for $t(0.75)$ is not given. The 95% CI for $t(0.50) \approx 210$ is (63,1296). The 95% CI for $t(0.25) \approx 63$ is (18,195).

```
Quartile estimates
Percent point estimate lower upper
75          .           220.0  .
50        210.00        63.00 1296.00
25         63.00        18.00 195.00
```

17) R plots the KM survival estimator along with the pointwise 95% CIs for $S_Y(t)$. If we guess a distribution for Y, say $Y \sim W$, with a formula for $S_W(t)$, then the guessed $S_W(t_i)$ can be added to the plot. If roughly 95% of the $S_W(t_i)$ fall within the bands, then $Y \sim W$ may be reasonable. For example, if $W \sim EXP(1)$, use $S_W(t) = \exp(-t)$. If $W \sim EXP(\lambda)$, then $S_W(t) = \exp(-\lambda t)$. Recall that $E(W) = 1/\lambda$.

18) If $\lim_{t\to\infty} tS_Y(t) \to 0$, then $E(Y) = \int_0^\infty t f_Y(t)dt = \int_0^\infty S_Y(t)dt$. Hence an estimate of the mean $\hat{E}(Y)$ can be obtained from the area under $\hat{S}(t)$.

19) Let $\boldsymbol{Y} = (Y_1, ..., Y_n)$. If $\boldsymbol{y} = (y_1, ..., y_n)$ is the data then the **likelihood function** $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{y})$. For each sample point $\boldsymbol{y} = (y_1, ..., y_n)$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ be a parameter value at which $L(\boldsymbol{\theta}|\boldsymbol{y})$ attains its maximum as a function of $\boldsymbol{\theta}$ with $\boldsymbol{y}$ held fixed. Then a maximum likelihood estimator (**MLE**) of the parameter $\boldsymbol{\theta}$ based on the sample $\boldsymbol{Y}$ is $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$. Note: it is crucial to observe that the likelihood function is a function of $\boldsymbol{\theta}$ (and that $y_1, ..., y_n$ act as fixed constants). Often $\boldsymbol{\theta} = \theta$ is a scalar.

20) If the MLE $\hat{\boldsymbol{\theta}}$ exists, then $\hat{\boldsymbol{\theta}} \in \Theta$. If the MLE $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_k)$, then the MLE of θ_i is $\hat{\theta}_i$, the MLE of (θ_1, θ_5) is $(\hat{\theta}_1, \hat{\theta}_5)$, etc.

21) **Invariance Principle:** If $\hat{\boldsymbol{\theta}}$ is the MLE of θ, then $\tau(\hat{\boldsymbol{\theta}})$ is the MLE of $\tau(\boldsymbol{\theta})$. Here τ is a function of $\boldsymbol{\theta}$ with domain Θ.

22) For **individual data**, $Y_1, ..., Y_n$ are iid, usually with pdf $f(y)$ or pmf $p(y)$. Let $\boldsymbol{y} = (y_1, ..., y_n)$ be the observed data. Then the **likelihood function** $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^n g(y_i)$ where $g(y)$ is $f(y)$ or $p(y)$. The **log likelihood function** $\log(L(\boldsymbol{\theta})) = \sum_{i=1}^n \log(g(x_i))$. Usually use 22) to find the MLE.

23) For this class, assume that the maximum likelihood estimator (MLE) is a solution to $\dfrac{\partial}{\partial\theta_i} \log L(\boldsymbol{\theta}) \stackrel{set}{=} 0$ for $i = 1, ..., k$ where usually $k = 1$ or 2. (We will not use second derivatives to show that the MLE was the global max.)

Tips: a) $\exp(a) = e^a$. b) $\log(a^b) = b\log(a)$ and $\log(e^b) = b$. c) $\log(\prod_{i=1}^n a_i) = \sum_{i=1}^n \log(a_i)$.
d) Often $\log[L(\theta)] = \log(\prod_{i=1}^n f(x_i|\theta)) = \sum_{i=1}^n \log(f(x_i|\theta))$. e) If t is a differentiable function and $t(\theta) \neq 0$, then $\frac{d}{d\theta} \ln(|t(\theta)|) = \frac{t'(\theta)}{t(\theta)}$ where $t'(\theta) = \frac{d}{d\theta} t(\theta)$. In particular, $\frac{d}{d\theta} \ln(\theta) = 1/\theta$. f) Anything that does not depend on θ is treated as a constant with respect to θ and hence has derivative 0 with respect to θ.

24) For small n, if given $\boldsymbol{y}$ it can be easier to plug in the y_i to find the MLE. Sometimes you will solve for the MLE as a statistic, then plug $\boldsymbol{x}$ into the statistic.

25) Let $g(\boldsymbol{x}|\boldsymbol{\theta})$ be the pmf or pdf of a sample $\boldsymbol{Y}$. If $\boldsymbol{Y} = \boldsymbol{y}$ is observed, then **the likelihood function** $L(\boldsymbol{\theta}) = g(\boldsymbol{y}|\boldsymbol{\theta})$.

26) Often $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|y_1, ..., y_n) = \prod_{i=1}^n L(\boldsymbol{\theta}|y_i)$. Note that $1 - F(w) = S(w)$.

a) For iid individual data, $L(\boldsymbol{\theta}|y_i) = f(y_i)$ if Y has pdf $f(y)$.

b) For iid individual data, $L(\boldsymbol{\theta}|x_i) = p(y_i)$ if Y has pmf $p(y)$.

c) If it is only known that y_i is in some interval $(c_{j-1}, c_j]$, then $L(\boldsymbol{\theta}|y_i) = P(y_i \in (c_{j-1}, c_j]) = F(c_j) - F(c_{j-1})$.
The endpoints can be open or closed if Y is from a continuous distribution.

d) If y_i is right censored at u_i, then the interval is $[u_i, \infty)$, and $L(\boldsymbol{\theta}|y_i) = 1 - F(u_i)$.

e) For grouped data from the table below, $L(\boldsymbol{\theta}) = \prod_{j=1}^{m} [F(c_j) - F(c_{j-1})]^{n_j}$.

interval	number
$(c_0, c_1]$	n_1
$(c_1, c_2]$	n_2
$(c_2, c_3]$	n_3
$\vdots$	$\vdots$
$(c_{m-2}, c_{m-1}]$	n_{m-1}
$(c_{m-1}, c_m]$	n_m

f) If y_i is left truncated at d_i, then $L(\boldsymbol{\theta}|y_i) = \dfrac{f(y_i)}{1 - F(d_i)}$.

g) If y_i is left truncated at d_i and right censored at u_i, then $L(\boldsymbol{\theta}|y_i) = \dfrac{1 - F(u_i)}{1 - F(d_i)}$.

h) If the data are left truncated at d with $n - k$ uncensored cases y_i and k cases right censored at u, then $L(\boldsymbol{\theta}) = \dfrac{[\prod_{i=1}^{n-k} f(y_i)][1 - F(u)]^k}{[1 - F(d)]^n}$.

i) (**Rare**, the interval is $(0, d]$.): If y_i is censored below at d, $L(\boldsymbol{\theta}|y_i) = F(d)$.

j) (**Rare**): If y_i is truncated above at u, $L(\boldsymbol{\theta}|y_i) = \dfrac{f(y_i)}{F(u)}$.

Note that left truncated = truncated below = truncated, and right censored = censored above = censored are often used.

1.7 Complements

For some of the MLE rules, see Klugman et al. (2008) and Kellison and London (2011). Olive (2014, pp. 145-147) gives a correct proof of the invariance principle (most "proofs" in the literature are not valid).

Important papers include Aalen (1978), Kaplan and Meier (1958), Nelson (1969, 1972). For Greenwood's formula, see Kaplan and Meier (1958). Advanced works use theory from counting processes and martingales.

1.8 Problems

Problems with an asterisk * are especially important.

1.1. Suppose $H(t) = \frac{\lambda}{\theta}[e^{\theta t} - 1]$ for $t > 0$ where $\lambda > 0$ and $\theta > 0$. Find $a)h(t)$, $b)S(t)$, c) $F(t)$ and d) $f(t)$ for $t > 0$.

1.2. Suppose $T \sim \text{EXP}(\lambda)$. Show $P(T > t + s|T > s) = P(T > t)$ for any $t > 0$ and $s > 0$. This property is known as the memoryless property and implies that the future survival of the product does not depend on the past if the lifetime T of the product is exponential.

1.3. Suppose $F(t) = 1 - \exp[-at - (bt)^2]$ where $a > 0$, $b > 0$ and $t > 0$. Find $a)S(t)$, $b)f(t)$, c) $h(t)$ and d) $H(t)$ for $t > 0$.

1.4. Suppose $F(t) = 1 - \exp[-at - (ct)^3]$ where $a > 0$, $c > 0$ and $t > 0$. Find the following quantities for $t > 0$.

a) $S(t)$

b)$f(t)$

c) $h(t)$

d) $H(t)$

1.5. Suppose $H(t) = \alpha + \beta t^2$ for $t > 0$ where $\alpha > 0$ and $\beta > 0$.

a) Find $h(t)$.

b) Find $S(t)$.

c) Find $F(t)$.

1.6. Suppose

$$F(t) = 1 - \exp\left(\frac{-t^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and $t > 0$. Find the following quantities for $t > 0$.

a) $S(t)$

b)$f(t)$

c) $h(t)$

d) $H(t)$

1.7. Eleven death times from Collett (2003b, p. 16) are given below. The patients had malignant bone tumours.

```
11  13  13  13  13  13  14  14  15  15  17
```

a) Following Example 1.3, make a table with headers $t_{(j)}, t_i, d_i, \hat{S}_E(t) = \sum(T_i > t)/n$.

b) Plot $\hat{S}_E(t)$.

c) Find the 95% classical CI for $S(13)$ based on $\hat{S}_E(t)$.

d) Find the 95% plus four CI for $S(13)$ based on $\hat{S}_E(t)$.

1.8. Find the 95% classical CI for $S_Y(32)$ if $n = 9$ and $\hat{S}_E(32) = 4/9$.

1.9. Find the 95% plus four CI for $S_Y(32)$ if $n = 9$ and $\hat{S}_E(32) = 4/9$.

1.10. Find the 95% plus four CI for $S_Y(32)$ if $n = 9$ and $\hat{S}_E(32) = 6/9$.

1.11. Find the 95% classical CI for $S_Y(32)$ if $n = 9$ and $\hat{S}_E(32) = 6/9$.

1.12. Survival times for nine electrical components are given below.
8, 8, 23, 32, 32, 46, 57, 88, 109
Compute the empirical survival function $\hat{S}_E(t_i)$ by filling in the table below. Then plot the function.

$t_{(j)}$	t_i	d_i	$\hat{S}_E(t)$
	$t_0 = 0$		$\hat{S}_E(0) = 1 = \frac{9}{9}$
8			
8	8	2	$\hat{S}_E(8) =$
23			$\hat{S}_E(23) =$
32			
32			$\hat{S}_E(32) =$
46			$\hat{S}_E(46) =$
57			$\hat{S}_E(57) =$
88			$\hat{S}_E(88) =$
109			$\hat{S}_E(109) =$

1.13. The Klein and Moeschberger (1997, p. 141-142) data set consists of information from 927 1st born children to mothers who chose to breast feed their child. The event was time in weeks until weaned (instead of death). Complete the following table used to produce the lifetable estimator (on a separate sheet of paper).

I_j	d_j	c_j	n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[0, 2)$	77	2	927	926	0.9168	1.0000
$[2, 3)$	71	3	848	846.5	0.9161	0.9168
$[3, 5)$	119	6	774	771	0.8457	0.8399
$[5, 7)$	75	9	649	644.5	0.8836	0.7103
$[7, 11)$	109	7	565	561.5	0.8059	0.6276
$[11, 17)$	148	5	449	446.5	0.6685	0.5058
$[17, 25)$	107	3	296			0.3381
$[25, 37)$	74	0	186			
$[37, 53)$	85	0	112			
$[53, \infty)$	27	0	27			

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
9	11	1	0.909	0.0867	0.7392	1.000
13	10	1	0.818	0.1163	0.5903	1.000
18	8	1	0.716	0.1397	0.4422	0.990
23	7	1	0.614	0.1526	0.3145	0.913
31	5	1	0.491	0.1642	0.1691	0.813
34	4	1	0.368	0.1627	0.0494	0.687
48	2	1	0.184	0.1535	0.0000	0.485

1.14. The length of times of remission (time until relapse) in acute myelogeneous leukemia under maintenance chemotherapy for 11 patients is $9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+$. See Miller (1981, p. 49). From the output above what is the 95% CI for $S_Y(34)$?

1.15. The Lindsey (2004, p. 280) data set is for survival times for 110 women with stage 1 cervical cancer studied over a 10 year period. Use the life table estimator to compute the estimated survival function $\hat{S}_L(t_i)$ by filling in the table below. Then plot the function.

I_j	d_j	c_j	n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[0,1)$	5	5	110	107.5	0.9535	1.0000
$[1,2)$	7	7	100	96.5	0.9275	0.9535
$[2,3)$	7	7	86	82.5	0.9152	0.8843
$[3,4)$	3	8	72	68	0.9559	0.8093
$[4,5)$	0	7	61	57.5	1.0	0.7736
$[5,6)$	2	10	54	49	0.9591	0.7736
$[6,7)$	3	6	42	39	0.9230	0.7420
$[7,8)$	0	5	33			
$[8,9)$	0	4	28			
$[9,10)$	1	8	24			
$[10,\infty)$	15	0	15			

1.16. Survival times for 13 women with tumors from breast cancer that were negatively stained with HPA are given below.
23, 47, 69, 70+, 71+, 100+, 101+, 148, 181, 198+, 208+, 212+, 224+
See Collett (2003b, p. 6). Compute the Kaplan Meier survival function $\hat{S}_K(t_i)$ by filling in the table below. Then plot the function.

$t_{(j)}$	γ_j	t_i	n_i	d_i	$\hat{S}_K(t)$
		$t_0 = 0$			$\hat{S}_K(0) = 1$
23	1	23	13	1	$\hat{S}_K(23) =$
47	1	47			$\hat{S}_K(47) =$
69	1	69			$\hat{S}_K(69) =$
70	0				
71	0				
100	0				
101	0				
148	1	148			$\hat{S}_K(148) =$
181	1	181			$\hat{S}_K(181) =$
198	0				
208	0				
212	0				
224	0				

1.17. The Lindsey (2004, p. 280) data is for survival times for 234 women with stage 2 cervical cancer studied over a 10 year period. Use the life table estimator to compute the estimated survival function $\hat{S}_L(t_i)$ by filling in the table below. Show what you multiply to find $\hat{S}_L(t_i)$. Then plot the function.

I_j	d_j	c_j	n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[0,1)$	24	3	234	232.5	0.8968	1.0000
$[1,2)$	27	11	207	201.5	0.8660	0.8968
$[2,3)$	31	9	169	164.5	0.8116	0.7766
$[3,4)$	17	7	129	125.5	0.8645	0.6302
$[4,5)$	7	13	105	98.5	0.9289	0.5448
$[5,6)$	6	6	85	82	0.9268	0.5061
$[6,7)$	5	6	73	70	0.9286	0.4691
$[7,8)$	3	10	62			
$[8,9)$	2	13	49			
$[9,10)$	4	6	34			
$[10,\infty)$	24	0	24			

1.18. Times (in weeks) until relapse below are for 12 patients with acute myelogeneous leukemia who reached a state of remission after chemotherapy. See Miller (1981, p. 49).
5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45
Compute the Kaplan Meier survival function $\hat{S}_K(t_i)$ by filling in the table below. Show what you multiply to find $\hat{S}_k(t_i)$. Then plot the function.

$t_{(j)}$	γ_j	t_i	n_i	d_i	$\hat{S}_K(t)$
		$t_0 = 0$			$\hat{S}_K(0) = 1$
5	1	5	12	2	$\hat{S}_K(5) =$
5	1				
8	1	8			$\hat{S}_K(8) =$
8	1				
12	1	12			$\hat{S}_K(12) =$
16	0				
23	1	23			$\hat{S}_K(23) =$
27	1	27			$\hat{S}_K(27) =$
30	1	30			$\hat{S}_K(30) =$
33	1	33			$\hat{S}_K(33) =$
43	1	43			$\hat{S}_K(43) =$
45	1	45			$\hat{S}_K(45) =$

1.19. Suppose the random variable Y has probability density function (pdf) $f(y)$ where $f(y) = 0$ for $y < 0$ and the expected value $E(Y)$ exists. One way to get a new pdf $g(y)$ is to use

$$g(y) = \frac{yf(y)}{E(Y)}.$$

See Cox (1962, p. 65). Show $\int_0^\infty g(y)dy = 1$.

SAS Problems

SAS is a statistical software package that will be used in this course. You will need a disk. There are SAS manuals and books at the library, but they are not needed in this course. To use SAS on windows (PC), use the following steps.

i) Click the lower left icon to see programs in the icons Window. You can click on the desktop icon to escape. If your computer does not have SAS, go to another computer. If you click on something and can't get out of the information window, there is a Windows key that looks like 4 rectangles and is on the lower left of the keyboard near the Ctrl key. This Windows key can get you back to icons Windows.

ii) Use the homework link or (http://parker.ad.siu.edu/Olive/survhw.txt) to copy and paste the program for Problem 1.20 into *SAS*. Highlight the program for problem 1.20. Hit Ctrl–c. Click the lower left icon to see programs. Double click the `SAS 9.4` icon. The editor window is the lower window. Click on that window,then hit Ctrl–v to paste in the program. Then run>submit. Output will appear in a few minutes.

(You can copy and paste the program from (http://parker.ad.siu.edu/Olive/M473hw.txt) or (http://parker.ad.siu.edu/Olive/regsas.txt) problem 16.36. The *ls* stands for linesize so *l* is a lowercase *L*, not the number one.)

If you were not successful, look at the *log window* for hints on errors. A single typo can cause failure. Reopen your file in *Word* or *Notepad* and make corrections. Occasionally you can not find your error. Then find your instructor or wait a few hours and reenter the program. *Word* seems to make better looking tables, and copying from *Notepad* to *Word* can completely ruin the table.

iii) To copy and paste relevant output into *Word,* click on the output window and use the top menu commands "Edit>Select All" and then the menu commands "Edit>Copy".

(In *Notepad* use the commands "Edit>Paste". Then use the mouse to highlight the relevant output (**the table and statistics for the table**). Then use the commands "Edit>Copy".)

Finally, in *Word*, use the commands "Edit>Paste".

iv) This point explains the SAS commands. The semicolon ";" is used to end SAS commands and the "options ls = 70;" command makes the output readable. (An "*" can be used to insert comments into the SAS program. Try putting an * before the options command and see what it does to the output.) The next step is to get the data into SAS. The command "data heart;" gives the name "heart" to the data set. The command "input time status number;" says the first entry is the censored variable time, the 2nd variable status (0 if censored 1 if uncensored) and the third variable number (= number of deaths or number of cases censored, depending on status). The command "cards;" means that the data is entered below. Then the data in entered and the isolated semicolon indicates that the last case has been entered. The next 4 lines make perform the lifetable estimates for S(t) and

the corresponding confidence intervals. Also plots of the estimated survival and hazard functions are given. The command "run;" tells SAS to execute the program.

You may want to save your SAS output as the file **hw1d20.doc** It may be easier to save output from each problem as a *Word* document, but you may get an extra page printed whenever you use the printer.

1.20. The following problem gets the lifetable estimator using SAS. The data is on 68 patients that received heart transplants at about the time when getting a heart transplant was new. The following problem gets the lifetable estimator using SAS. See Allison (1995, p. 49-50).

a) Do i) through iii) above, and look at iv).

b) From the 1st page of output, *Number Failed* $= d_i$, *Number Censored* $= c_i$, *Effective Sample Size* $= n_i'$, *Survival* $= \hat{S}_L(t_{i-1}) =$ estimated survival for the left endpoint of the interval and *Survival Standard Error* $= SE[\hat{S}_L(t_{i-1})]$.

What is $SE[\hat{S}_L(200)]$?

c) From the 2nd page of output, *SDF_LCL SDF_UCL* gives a 95% CI for $S(t_{i-1})$.

What is the 95% CI for $S(200)$ using output?

d) Compute the 95% CI for $S(200)$ using the formula and $SE[\hat{S}_L(200)]$.

e) The SAS program (with plots(s,h)) plots both the survival and the hazard function (scroll down!). From the 2nd page of output, plot MIDPOINT vs HAZARD (so the first point is (25,0.0055)) **by hand**. Connect the dots to make an estimated hazard function. Notice that the estimated hazard function decreases sharply to about 200 days after surgery and then is fairly stable.

1.21. This problem examines the Allison (1995, p. 31-34) myelomatosis data (a cancer causing tumors in the bone marrow) with SAS using the Kaplan Meier product limit estimator. Obtain the SAS program for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt). Obtain the output from the program in the same manner as i) through iv) above Problem 1.20.

a) The output should be roughly 3 pages and a graph. Include this output in *Word*.

b) From the summary statistics of the first page of output, about when do 50% of the patients die?

c) From the first page of output (perhaps), what is the 95% CI for the time when 50% of the patients die?

d) From the 3rd page of output (perhaps), what is the 95% CI for $S_Y(13)$. This is the log log transformed CI, so will differ from the CI in e).

e) Make the CI using $\hat{S}_K(13)$ and $SE(\hat{S}_K(13))$ obtained from the 1st page of output (perhaps). If the interval is (L, U), use $[\max(0, L), \min(U, 1)]$ as the final interval.

f) From the plot of $\hat{S}_K(t)$ for the KM estimator, briefly explain survival for days 0–250 and for days 250–2250.

1.22. This Miller (1981, p. 49-50) data set is on remission times in weeks for leukemia patients. Twenty patients received treatment A and 20 received treatment B. The predictor *group* was 0 for A and 1 for B.

a) Obtain the SAS program for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt). Obtain the output from the program in the same manner as i) through iv) above Problem 1.20.

b) Do a 4 step test for $Ho : \beta = 0$.

c) Do a 4 step PLRT for $Ho : \boldsymbol{\beta} = \mathbf{0}$ (for $\beta = 0$). (The PLRT is better than the Wald test in b).)

R Problems

R is the free version of *Splus.* Click on the *Rgui* icon to get into *R.* Then typing *q()* gets you out of *R.*

Use the command *source("G:/survdata.txt")* **to download the data. See Preface or Section 5.1.** For the following problems, the *R* command can be copied and pasted from (http://parker.ad.siu.edu/Olive/survhw.txt) into *R.*

1.23. Miller (1981, p. 49) gives the length of times of remission (time until relapse) in acute myelogeneous leukemia under maintenance chemotherapy for 11 patients is
$9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+$.

a) Following Example 1.3, make a table with headers $t_{(j)}$, γ_j, t_i, n_i, d_i and $\hat{S}_K(t_i)$. Then compute the Kaplan Meier estimator. (You can check it with the *R* output obtained in b).)

b) Get into *R.* Copy and paste commands for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt) into *R.* Hit `Enter` and a plot should appear. Copy and paste the *R* output with header (time ... upper 95% CI) into *Word.* Following the *R* handout, click on the plot and hold down the *Ctrl* and *c* buttons simultaneously. Then in the *Word* Edit menu, select "paste."

Include this output with the homework. The center step function is the Kaplan Meier estimator $\hat{S}_K(t)$ while the lower and upper limits correspond to the confidence interval for $S_Y(t)$.

c) Write down the 95% CI for $S_Y(23)$ and then verify the CI by computing $\hat{S}_K(23) \pm 1.96SE(\hat{S}_K(23))$.

1.24. Copy and paste commands for parts a) and b) for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt) into *R.*

The commands make the KM estimator for censored data $T = \min(Y, Z)$ where $Y \sim EXP(1)$. The KM estimator attempts to estimate $S_Y(t) = \exp(-t)$. The points in the plot are $S_Y(t_{(j)}) = \exp(-t_{(j)})$, and the points

should be within the confidence intervals roughly 95% of the time (actually, if you make many plots the points should be in the intervals about 95% of the time, but for a given plot you could get a "bad data set" and then the rather more than 5% of the points are outside of the intervals).

a) Copy and paste the commands for a) and hit `Enter`. Then copy and paste the plot into *Word*.

b) Copy and paste the commands for b) and hit `Enter`. Then copy and paste the plot into *Word*.

c) As the sample size increases from $n = 20$ to $n = 200$, the CIs should become more narrow. Can you see this in the two plots? Are about 95% of the plotted points within the CIs?

1.25. Go to (http://parker.ad.siu.edu/Olive/survhw.txt) and copy and paste the source command near the top of the file into *R*.

Type the command `kmsim2(n=10)`, hit `Enter` and include the output in *Word*.

This program computes censored data $T = \min(Y, Z)$ where $Y \sim EXP(1)$. Then a 95% CI is made for $S_Y(t_{(j)})$ for each of the $n = 10$ $t_{(j)}$. This is done for 100 data sets and the program counts how many times the CI contains $S_Y(t_{(j)}) = \exp(-t_{(j)})$. The scaled lengths are also computed. The ccov is the count for the classical $\hat{S} \pm 1.96SE(\hat{S})$ interval while p4cov is for the plus 4 CI. The lcov is based on a CI that uses $\log(\hat{S})$ and llcov is based on a CI that uses $\log(-\log(\hat{S}))$. The 1st 3 CIs are not made if the last case is censored so NA is given. The plus 4 CI seems to be good at $t_{(1)}$ and $t_{(n)}$.