STARTED on 9/29/21 (midway through lecture)

NOT Missing but Incomplete

Partial data

HANDOUT #7: CENSORED DATA

where you have partial information about some of your data points.

1. Type I Censoring - Fixed Censoring Time
2. Type II Censoring - Fixed Number of Observed Failures
3. Random Censoring
4. Many other Types of Censoring
5. Form of Censored Data
6. Parametric Estimation When Data is Censored
7. Distribution-Free Estimation When Data is Censored
8. Example - Using SAS and R

- Usually time to event data

- Ex: when you only know the event was at least a particular #.

Supplemental Reading

- Problem 4.51, Problem 6.34, Problem 15.10 in Tamhane/Dunlop book

CENSORED DATA

In some situations, the observations from a population or the outcomes from a process are not a complete set of data. Some of the experimental units on whom we planned to make observations or take measurements are removed from the experiment (study) prior to the end of the experiment. There may be partial information about these experimental units but not complete information.

Suppose we have n randomly selected experimental units from a population whose times to occurrence of an event are to be observed:

- Time to failure of an electrical device
- Time to occurrence of a tumor in a laboratory animal injected with a toxic chemical
- Time until patient's blood pressure reaches a specified level after receiving a treatment.
- Amount of stress at which alloy specimen fractures

Several forms of censoring will now be defined.

Suppose n units are placed on test and we want to measure the variable T_i on unit $i = 1, \ldots, n$.

1. **No Censoring - Complete Data Set**

   If we observe $T_1, \ldots, T_n$ on all n units and hence we have a complete data set. Standard methods of estimation are then used to estimate population parameters and make inferences about the population.

2. **Right Censoring - Incomplete Data Set**

   If we observe the value of T_i for $i = 1, \cdots, m$ but only know that the remaining $n - m$ units have values of T_i greater than their recorded values then we have right censored data. We only have a lower bound on the value of T_i.

- Type I and Type II censoring are special cases of Right Censoring.

Ex: Someone leaves hospital before they die. Don't know true until death, but know that it is at least up until they left the hospital

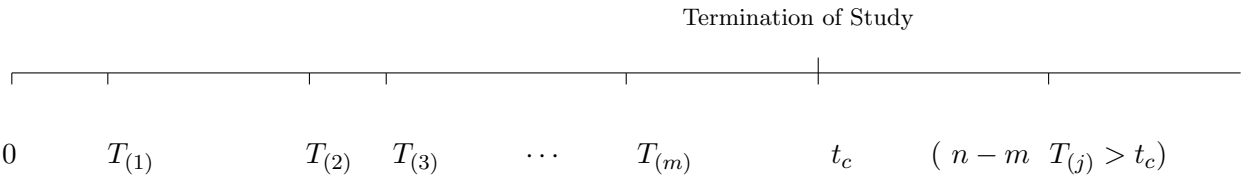
Type I Censoring - Stop Experiment at Time t_c

(The textbook refers to this type of censoring as Type II, see pages 235-236 in textbook)

Suppose n units are placed on test and $T_1,\ T_2,\ \ldots,\ T_n$ are their times to failure. Suppose the reliability study terminates at a preselected time, t_c. All experimental units which have not failed before time t_c have **Type I** censoring.

The experiment has n units on test and the experiment is terminated at a preselected time t_c. Suppose m units have failed before the experiment terminates. Then the ordered times to failure will satisfy:

$$T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(m)} \leq t_c < T_{(m+1)} \leq \cdots \leq T_{(n)}$$



Because the experiment is terminated at time t_c, we will only observe the times for m of the units. For the remaining $n-m$ units, we will only know that their times satisfy $T_j > t_c$.

In Type I censoring, the time to termination of the experiment, t_c is fixed but the number of observed times, m is random. Type II fixes this one.

A problem arises if t_c is too small in that nearly all the data values will be censored.

Type II Censoring - Stop Experiment When mth Unit Fails

(The textbook refers to this type of censoring as Type I)

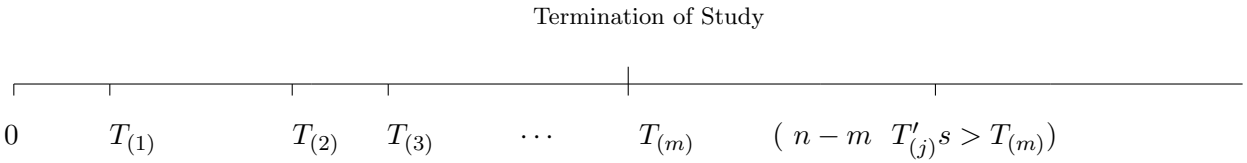
An alternatively experimental design has the experiment terminate when the mth (preselected) unit fails. This is referred to as **Type II censoring**.

The experiment has n units on test and the experiment is terminated when m units have failed.

The ordered times to failure will satisfy:

$$T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(m)} < T_{(m+1)} \leq \cdots \leq T_{(n)}$$

With the times: $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(m)}$ observed but the times for the remaining $n - m$ units unobserved because the experiment was terminated before these units failed.



The remaining $n - m$ units will not have their failure times recorded. However, we will have the information that $T_{(j)} > T_{(m)}$ for $j = m + 1, \ldots, n$.

Type II censoring has the number of observed times, m fixed but the time to termination of the experiment, $T_{(m)}$ random.

If the device under test is very reliable, and m is too large, then it may take a long time to terminate the study, i.e., for m units to fail.

3. **Left Censoring - Incomplete Data Set**

If we observe the value of T_i for $i = 1, \cdots, m$ but only know that the remaining $n - m$ units have values of T_i less than their recorded values then we have left censored data. We only have an upper bound on the value of T_i.

- For example, we place n units on test and want to record their time to failure. The units are inspected at specified inspection times and some of the units fail prior to the first inspection time. The failure times of these units would be unknown except we know that they are less than the time of the first inspection, left censoring.

  Note, if there are units still functioning after the last inspection time then we would have right censoring.

- A second example would be if a measuring device is not sensitive enough to measure observations below a known threshold. For example, if we were measuring ultrasonic noise, the measuring device may not record a value for an sound having frequency less than 20KHz. Thus, any sound having less than 20KHz would not be recorded but we would know that the sound would have a value less than 20KHz.

  For example, monitoring communication between bats. We know there are sound waves being transmitted by observing the behaviour of the bats but the device fails to record any sounds.

4. **Random Censoring - Incomplete Data** (Anything not Left or Right censoring)

A third type of censoring is Random Censoring in which individual experimental units fail to have their time to the event recorded due to :

- Patient in study just stops returning to medical center
- Machine applying stress to specimens breaks down before specimen fractures
- Patient is removed from study due to side effects of drug
- Operator in Lab stops working
- Emergency budget cuts cause study to be reduced in size resulting in some patients dropping out of study
- Laboratory equipment fails while recording values for a given experimental unit

In this situation, we observe $min\{T_i, C_i\}$, where T_i is the time until event occurs for unit i and C_i is the time at which unit i leaves the experiment(study). Note that Type I censoring is equivalent to random censoring with $C_i \equiv t_c$ for all n units. In most applications, T and C are taken to be independent.

5. **Interval Censoring - Incomplete Data Set**

- Each of the censored units would have their measured value as being within an interval and no specific value recorded.
- In a failure time study with n units on test, the units are observed at specified inspection times. Therefore, the actual time to failure is unknown. The failure times are recorded as being between two inspection times.

6. **Other Types of Censoring or Incomplete Data Set**

- Arbitrary Censoring - Combinations of left and right censoring and interval censoring with overlapping intervals
- Truncated Data - Censoring occurs when there is a bound on all observations for which an exact value is not recorded: lower bound for observations censored on the right, upper bounds for observations censored on the left, and both upper and lower bounds for observations that are interval censored. Truncated data arises when even the existence of a potential observation would be unknown if its value were to lie in a certain range of values, either below a specified point τ_L or above a specified value, τ_U.

  Example from *Statistical Methods for Reliability Data*, **Meeker & Escobar** Ultrasonic inspection is used to detect flaws in titanium alloys. Ultrasonic signal amplitude is positively correlated with crack size. Titanium grain boundaries reflect signals as well as flaws. Thus below a specified threshold, τ_L it is impossible to be sure whether a signal is a flaw or grain boundary.

  In a lab test of the inspection process, specimens with flaws of known size are inspected. The signal's amplitude is measured only when it is above τ_L. All specimens in which a signal is not recorded would have left-censored observations.

  However, in a production inspection process, a flaw is not detected when the signal's amplitude is below τ_L. Thus the number of flaws that are present with signal amplitude below τ_L is unknown. The signal amplitudes recorded above τ_L are known as left-truncated observations, or observations from a left-truncated distribution.

Form of Censored Data

The data will have the following structure in each of four situations:

- Let $T_1, \ldots, T_n$ be the times to the event with $T_1, \ldots, T_n$ *iid* having cdf F.

  The values of $T_1, \ldots, T_n$ may or may not be observed depending on whether or not they have been censored.

- Let $Y_1, \ldots, Y_n$ be the observed data with $Y_1, \ldots, Y_n$ *iid* having cdf H.

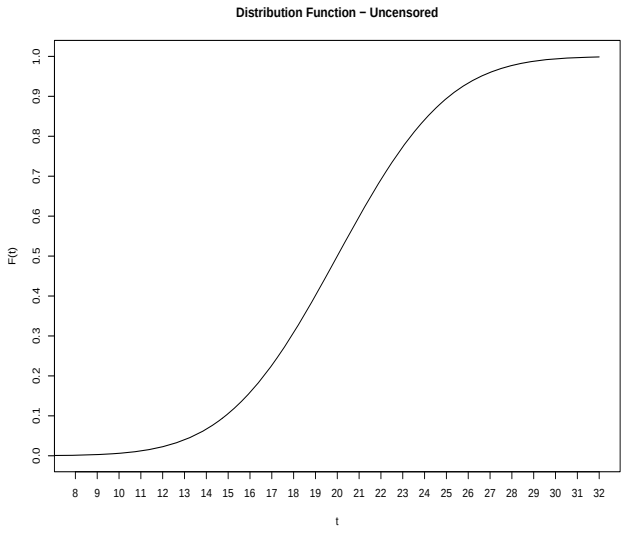
The form of Y_i and its cdf H will depend on the type of censoring in the experiment:

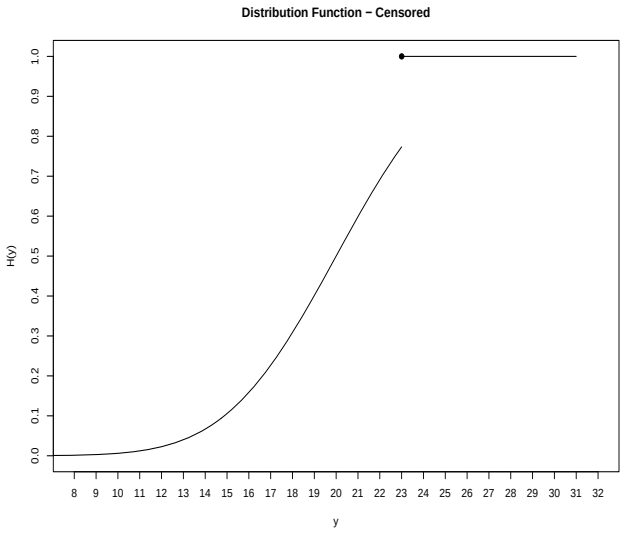
1. **No Censoring:** $Y_i = T_i$ for $i = 1, \ldots, n$ and $Y_1, \ldots, Y_n$ *iid* having cdf $H = F$.

2. **Type I Censoring:** For all $T_i > t_c$, data is censored, therefore we have

$$Y_i = \begin{cases} T_i & \text{if } T_i \leq t_c \\ t_c & \text{if } T_i > t_c \end{cases}$$

   The cdf of Y_i, H has a jump at t_c of height p where $p = P[Y_i = t_c] = P[T_i > t_c] = 1 - F(t_c) \neq 0$. This occurs because the distribution of Y_i is truncated at t_c.





3. **Type II Censoring:** The first m times to the event are observed and then the experiment is terminated. With $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(n)}$, we observe $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(m)}$ and then experiment is then terminated. Thus, the largest $n-m$ times to event are not observed. Therefore we have

$$Y_{(i)} = \begin{cases} T_{(i)} & \text{for } i = 1, \ldots, m \\ T_{(m)} & \text{for } i = m+1, \ldots, n \end{cases}$$

Using the properties of the order statistics, the joint pdf of the Y_i is

$$f(y_1, y_2, \ldots, y_n) = \binom{n}{m} m! f(y_1) f(y_2) \cdots f(y_m) [S(y_m)]^{n-m}$$

where $S(y) = 1 - F(y)$ is the survival function.

4. **Random Censoring:** Let $C_1, \ldots, C_n$ be the times at which the n units were censored with $C_i's$ iid with cdf $G(\cdot)$. Let $T_1, \ldots, T_n$ be the times to the occurrence of the event for the n units with $T_i's$ iid with cdf $F(\cdot)$. We observe $(Y_1, \delta_1), (Y_2, \delta_2), \ldots, (Y_n, \delta_n)$ where $Y_i = min(T_i, C_i)$ and

$$\delta_i = I(T_i \leq C_i) = \begin{cases} 1 & \text{if Unit i is NOT censored} \\ 0 & \text{if Unit i IS censored} \end{cases}$$

With random censoring $Y_1, \ldots, Y_n$ are iid with cdf $H(\cdot)$.

The joint pdf of (Y, δ) is obtained using the independence of T and C:

If $\delta = 1$ (uncensored observation), then $T \leq C$ and $Y = T$, which yields

$$h(y, \delta) = f(y) P[C \geq y] = f(y)[1 - G(y)]$$

If $\delta = 0$ (censored observation), then $T \geq C$ and $Y = C$, which yields

$$h(y, \delta) = g(y) P[T \geq y] = g(y)[1 - F(y)] = g(y) S(y)$$

Using δ, we can combine the two situations into a single equation:

$$h(y, \delta) = [f(y)]^{\delta} [1 - G(y)]^{\delta} [g(y)]^{1-\delta} [S(y)]^{1-\delta}$$

For data having Type I, Type II, or Random Censoring our goal is to estimate the cdf F and parameters associated with F using the observed data $Y_1, \ldots, Y_n$.

STOP Wednesday 9/29/21

START Friday 10/1/21

Parametric Estimation

Suppose we know the family of distributions for which F is a member. For example, suppose the family is a logistic family of distributions with parameters $(\theta_1,\ \theta_2)$, that is, the family of distributions has pdf,

$$f(t;\theta_1,\ \theta_2) = \frac{1}{\theta_2}\frac{e^{-(t-\theta_1)/\theta_2}}{[1+e^{-(t-\theta_1)/\theta_2}]^2}$$

We want to estimate the parameters associated with F, θ_1 and θ_2. In order to use MLE techniques we need to specify the likelihood function for the data. The form of the likelihood will depend on the type of censoring:

1. **No Censoring:**

   Let $f(\cdot,\boldsymbol{\theta})$ be the pdf associated with $F(\cdot,\boldsymbol{\theta})$. The likelihood function for the observed data $Y_i's$ is given by

   $$L(\boldsymbol{\theta}) = f(Y_1, Y_2, \ldots, Y_n;\boldsymbol{\theta}) = \prod_{i=1}^{n} f(Y_i;\boldsymbol{\theta}) = \prod_{i=1}^{n} f(T_i;\boldsymbol{\theta})$$

   Select the value of $\boldsymbol{\theta}$ to maximize $L(\boldsymbol{\theta})$.

2. **Random Censoring**

   Assume T_i and C_i are independent with $Y_i = min(T_i, C_i)$ and F the cdf of $T_i's$, H the cdf of $Y_i's$, and G the cdf of $C_i's$.

   (2a.) If Y has a discrete distribution, then the joint pdf for the pair (Y_i, δ_i) is given by

   $$P[Y_i = y, \delta = 1] = P[Y_i = y, C_i \geq T_i] = P[T_i = y, C_i \geq y] = P[T_i = y]P[C_i \geq y] = f(y)[1-G(y)]$$

   $$P[Y_i = y, \delta = 0] = P[Y_i = y, T_i \geq C_i] = P[C_i = y, T_i \geq y] = P[C_i = y]P[T_i \geq y] = g(y)[1-F(y)]$$

   (2b.) For continuous Y, the joint pdf for the pair (Y_i, δ_i) is given by

   $$L((Y_i,\delta_i);\boldsymbol{\theta}) = \begin{cases} f(Y_i;\boldsymbol{\theta})P[C_i \geq T_i] & \text{if } \delta_i = 1 \text{ (uncensored)} \\ g(Y_i)P[C_i < T_i] & \text{if } \delta_i = 0 \text{ (censored)} \end{cases}$$

   $$L((Y_i,\delta_i);\boldsymbol{\theta}) = \begin{cases} f(Y_i;\boldsymbol{\theta})[1-G(Y_i)] & \text{if } \delta_i = 1 \text{ (uncensored)} \\ g(Y_i)[1-F(Y_i;\boldsymbol{\theta})] & \text{if } \delta_i = 0 \text{ (censored)} \end{cases}$$

   $$L((Y_i,\delta_i);\boldsymbol{\theta}) = [f(Y_i;\boldsymbol{\theta})]^{\delta_i}[S(Y_i;\boldsymbol{\theta})]^{(1-\delta_i)}[1-G(Y_i)]^{\delta_i}[g(Y_i)]^{(1-\delta_i)},$$

   where $S(Y_i;\boldsymbol{\theta}) = 1 - F(Y_i;\boldsymbol{\theta})$ is the survival function for T_i.

   In the likelihood for the full sample we will designate terms involving $[1-G(Y_i)]$ and $g(Y_i)$ as $K(Y_i,\delta_i)$ because they do not involve the unknown parameters $\boldsymbol{\theta}$. Hence, the likelihood function is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L(Y_i, \delta_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} [f(Y_i; \boldsymbol{\theta})]^{\delta_i} [S(Y_i; \boldsymbol{\theta})]^{(1-\delta_i)} \prod_{i=1}^{n} K(Y_i, \delta_i)$$

which yields

$$L(\boldsymbol{\theta}) = \prod_{i \epsilon U} f(T_i; \boldsymbol{\theta}) \prod_{i \epsilon C} S(C_i; \boldsymbol{\theta}) \left[\prod_{i=1}^{n} K(Y_i, \delta_i) \right]$$

where U is the set of indices for the uncensored units and C is the set of indices for the censored units.

3. **Type I Censoring**

The likelihood is obtained from the Random censoring case with $C_i \equiv t_c$ for all $i \epsilon C$:

$$L(\boldsymbol{\theta}) = \prod_{i \epsilon U} f(T_i; \boldsymbol{\theta}) \prod_{i \epsilon C} S(C_i; \boldsymbol{\theta}) \prod_{i=1}^{n} K(Y_i, \delta_i) = [S(t_c; \boldsymbol{\theta})]^{n-n_U} \prod_{i \epsilon U} f(T_i; \boldsymbol{\theta}) \left[\prod_{i=1}^{n} K(Y_i, \delta_i) \right]$$

4. **Type II Censoring**

Observe $T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(m)}$ with remaining $n - m$ units having $Y_i \equiv T_{(m)}$. Therefore, the likelihood is given by

$$\begin{aligned} L(\boldsymbol{\theta}) &= \binom{n}{m} \left[m! \prod_{i=1}^{m} f(Y_{(i)}; \boldsymbol{\theta}) \right] [S(Y_{(m)}, \boldsymbol{\theta})]^{n-m} \left[\prod_{i=1}^{n} K(Y_i, \delta_i) \right] \\ &= \frac{n!}{(n-m)!} \left[\prod_{i=1}^{m} f(T_{(i)}; \boldsymbol{\theta}) \right] [S(T_{(m)}, \boldsymbol{\theta})]^{n-m} \left[\prod_{i=1}^{n} K(Y_i, \delta_i) \right] \end{aligned}$$

Once we have established the likelihood function for the data, we then select $\hat{\boldsymbol{\theta}}$ to maximize $L(\boldsymbol{\theta})$ the likelihood function.

Example Let $T_1, \ldots, T_n$ be iid exponential: $F(t;\beta) = 1 - e^{-t/\beta}$

$$\Rightarrow f(t;\beta) = \frac{1}{\beta}e^{-t/\beta} \quad \text{and} \quad S(t;\beta) = 1 - F(t;\beta) = e^{-t/\beta}$$

Suppose we have **random censoring** and let n_U be the number of uncensored observations. The likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i\epsilon U} f(T_i;\boldsymbol{\theta}) \prod_{i\epsilon C} S(C_i;\boldsymbol{\theta}) \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\ &= \prod_{i\epsilon U} \frac{1}{\beta}e^{-T_i/\beta} \prod_{i\epsilon C} e^{-C_i/\beta} \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\ &= (\beta)^{-n_U} e^{-\frac{1}{\beta}\left(\sum_{i\epsilon U} T_i + \sum_{i\epsilon C} C_i\right)} \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\ &= (\beta)^{-n_U} e^{-\frac{1}{\beta}\sum_{i=1}^{n} Y_i} \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \end{aligned}$$

Thus, the log-likelihood is given by

$$l(\beta) = log(L(\beta)) = -n_U log(\beta) - \frac{1}{\beta}\sum_{i=1}^{n} Y_i + log\left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \quad \Rightarrow$$

$$\frac{d}{d\beta}l(\beta) = \frac{-n_U}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} Y_i + 0$$

$$\frac{d}{d\beta}l(\beta)|_{\beta=\hat{\beta}} = 0 \quad \Rightarrow \quad \hat{\beta} = \frac{1}{n_U}\sum_{i=1}^{n} Y_i \quad \Rightarrow$$

$$\hat{\beta} = \frac{\text{Total Time n Units were Operating}}{\text{Total Number of Units Which reached Event}}$$

The same estimator is obtained for Type I censoring.

For Type II censoring, the estimator becomes

$$\hat{\beta} = \frac{1}{n_U}\sum_{i=1}^{n} Y_i = \frac{1}{n_U}\left[\sum_{i=1}^{n_U} T_{(i)} + (n - n_U)T_{(n_U)}\right]$$

$\hat{\beta}$ is referred to as the Winsorized Mean - replace all extreme values with the largest(smallest) non-extreme value.

Note that the naive estimator, $\hat{\beta}^*$ would satisfy

$$\hat{\beta}^* \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i \;=\; \frac{1}{n}\sum_{i\epsilon U} T_i + \frac{1}{n}\sum_{i\epsilon C} C_i \;\leq\; \frac{1}{n}\sum_{i=1}^{n} T_i \;=\; \hat{\beta}, \text{ if no censoring occurred}$$

Thus, $\hat{\beta}^*$ would on the average underestimate β, because $\hat{\beta}$ is an unbiased estimator of β which would imply

$$E\left[\hat{\beta}^*\right] \;\leq\; E\left[\hat{\beta}\right] \;=\; \beta$$

Example Let $T_1, \ldots, T_n$ be iid with a Weibull distribution: $F(t;\gamma,\alpha) = 1 - e^{-(t/\alpha)^\gamma}$

$$\Rightarrow f(t;\gamma,\alpha) = \frac{\gamma}{\alpha}(t/\alpha)^{\gamma-1}e^{-(t/\alpha)^\gamma} \quad \text{and} \quad S(t;\gamma,\alpha) = 1 - F(t;\gamma,\alpha) = e^{-(t/\alpha)^\gamma}$$

Suppose we have **Type I censoring** and let $m = n_U$ be the number of uncensored observations. The likelihood function is given by

$$\begin{aligned}
L(\boldsymbol{\theta}) &= \prod_{i\epsilon U} f(T_i;\boldsymbol{\theta}) \prod_{i\epsilon C} S(C_i;\boldsymbol{\theta}) \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\
&= \\
&= \prod_{i\epsilon U} \frac{\gamma}{\alpha}(T_i/\alpha)^{\gamma-1}e^{-(T_i/\alpha)^\gamma} \prod_{i\epsilon C}\left[e^{-(t_c/\alpha)^\gamma}\right] \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\
&= \\
&= \prod_{i\epsilon U} \frac{\gamma}{\alpha}(T_i/\alpha)^{\gamma-1}e^{-(T_i/\alpha)^\gamma} \left[e^{-(t_c/\alpha)^\gamma}\right]^{n-m} \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\
&= \\
&= \left(\frac{\gamma}{\alpha}\right)^m \left(\frac{1}{\alpha^{\gamma-1}}\right)^m \left(\prod_{i\epsilon U} T_i^{\gamma-1}\right) \left(e^{-\frac{1}{\alpha^\gamma}\sum_{i\epsilon U} T_i^\gamma}\right) \left(e^{-(n-m)(t_c/\alpha)^\gamma}\right) \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right]
\end{aligned}$$

The log-likelihood function is given by

$$l(\gamma,\beta) = log(L(\gamma,\beta))$$

$$
\begin{aligned}
l(\gamma,\beta) &= log(L(\gamma,\beta)) \\
&= mlog(\gamma) - m\gamma log(\alpha) + (\gamma-1)\sum_{i\epsilon U} log(T_i) - \frac{1}{\alpha^\gamma}\sum_{i\epsilon U} T_i^\gamma - \frac{(n-m)}{\alpha^\gamma}t_c^\gamma \; + \; log\left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\
&= mlog(\gamma) - m\gamma log(\alpha) + (\gamma-1)\sum_{i\epsilon U} log(Y_i) - \frac{1}{\alpha^\gamma}\left(\sum_{i\epsilon U} T_i^\gamma + (n-m)t_c^\gamma\right) \; + \; log\left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\
&= mlog(\gamma) - m\gamma log(\alpha) + (\gamma-1)\sum_{i\epsilon U} log(Y_i) - \frac{1}{\alpha^\gamma}\left(\sum_{i=1}^{n} Y_i^\gamma\right) \; + \; log\left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right]
\end{aligned}
$$

Taking derivatives of $l(\gamma,\alpha)$ wrt γ and α yield,

$$
\begin{aligned}
\frac{d}{d\alpha}l(\gamma,\alpha) &= \frac{d}{d\alpha}\left(mlog(\gamma) - m\gamma log(\alpha) + (\gamma-1)\sum_{i\epsilon U} log(Y_i) - \frac{1}{\alpha^\gamma}\left(\sum_{i=1}^{n} Y_i^\gamma\right) \; + \; log\left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right]\right) \\
&= 0 - \frac{m\gamma}{\alpha} + 0 + \frac{\gamma}{\alpha^{\gamma+1}}\sum_{i=1}^{n} Y_i^\gamma + 0
\end{aligned}
$$

$$
\begin{aligned}
\frac{d}{d\gamma}l(\gamma,\alpha) &= \frac{d}{d\gamma}\left(mlog(\gamma) - m\gamma log(\alpha) + (\gamma-1)\sum_{i\epsilon U} log(Y_i) - \frac{1}{\alpha^\gamma}\left(\sum_{i=1}^{n} Y_i^\gamma\right) \; + \; log\left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right]\right) \\
&= \frac{m}{\gamma} - mlog(\alpha) + \sum_{i\epsilon U} log(Y_i) - \frac{1}{\alpha^\gamma}\sum_{i=1}^{n} log(Y_i)Y_i^\gamma + \frac{1}{\alpha^\gamma}log(\alpha)\left(\sum_{i=1}^{n} Y_i^\gamma\right) + 0
\end{aligned}
$$

Then setting the derivatives equal to 0 yield the following equations:

$$
\hat{\alpha} = \left(\frac{1}{m}\sum_{i=1}^{n} Y_i^{\hat{\gamma}}\right)^{1/\hat{\gamma}} \qquad (1)
$$

$$
0 = \frac{1}{\hat{\gamma}} + \frac{1}{m}\sum_{i\epsilon U} log(Y_i) - \frac{\sum_{i=1}^{n} Y_i^{\hat{\gamma}} log(Y_i)}{\sum_{i=1}^{n} Y_i^{\hat{\gamma}}} \qquad (2)
$$

Notice that equation (2) involves just $\hat{\gamma}$.

A numerical solution can be easily obtained and then the value of $\hat{\alpha}$ can be obtained from equation (1).

A numerical example will be considered next to illustrate the use of SAS and R in obtaining estimators of γ and α in the Weibull model when the data has censoring.

EXAMPLE The following example (slightly modified) is from *Statistical Analysis of Reliability Data* by M.J. Crowder, A.C. Kimber, R.L. Smith, and T.J. Sweeting. In an experiment to determine the strength of a braided cord after weathering, the strengths of 48 pieces of cord that had been weathered for a specified length of time were investigated. The company wanted to estimate the probability that the cord would have strength of at least 53, that is, estimate $S(53) = P[T > 53]$. Seven cords were damaged during the study which resulted in a decrease in their strength. Therefore, the study produced right censored strength values. The strengths of the remaining 41 cords were determined as shown below:

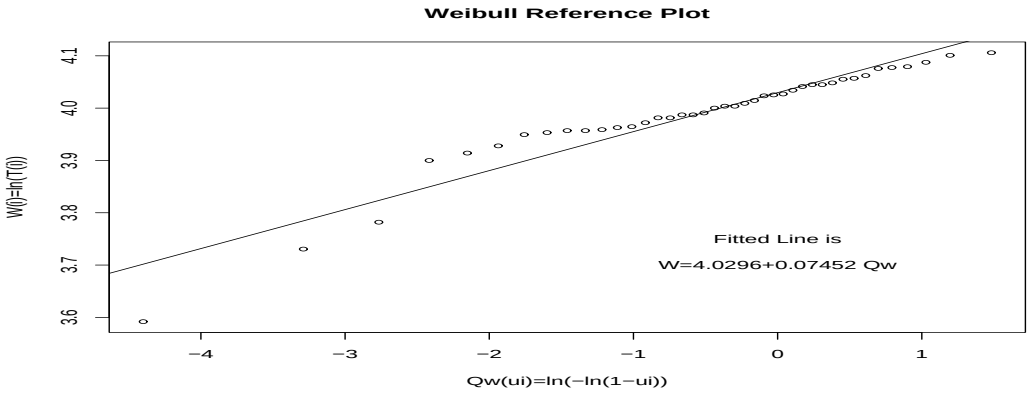
```
36.3  52.4  54.8  57.1  60.7  41.7  52.6  54.8  57.3
43.9  52.7  55.1  57.7  49.4  53.1  55.4  57.8
50.1  53.6  55.9  58.1  50.8  53.6  56.0  58.9
51.9  53.9  56.1  59.0  52.1  53.9  56.5  59.1
52.3  54.1  56.9  59.6  52.3  54.6  57.1  60.4
```

The 7 censored strength values from the damaged cords are given next:

```
26.8  29.6  33.4  35.0  40.0  41.9 42.5
```

The true strength values of the 7 cords, T_i are unobservable but we know $T_i > Y_i$, where Y_i are the observed values.

A Weibull reference distribution plot for the 41 uncensored strengths is displayed here:



From the plot, it would appear that the Weibull model is adequate.

The following SAS program will be used to obtain the MLE's of α and γ:

```
*weib_mle_censored.sas in Files/SASCode;

option ls=75 ps=55 nocenter nodate;
title 'Strength of Braided Cord';
data cords;
input S C @@;
label S  = 'Strength of Cord' C ='Censoring (1=Yes)';
cards;
36.3 0 52.4 0 54.8 0 57.1 0 60.7 0 41.7 0 52.6 0 54.8 0 57.3 0
43.9 0 52.7 0 55.1 0 57.7 0 49.4 0 53.1 0 55.4 0 57.8 0
50.1 0 53.6 0 55.9 0 58.1 0 50.8 0 53.6 0 56.0 0 58.9 0
51.9 0 53.9 0 56.1 0 59.0 0 52.1 0 53.9 0 56.5 0 59.1 0
52.3 0 54.1 0 56.9 0 59.6 0 52.3 0 54.6 0 57.1 0 60.4 0
26.8 1 29.6 1 33.4 1 35.0 1 40.0 1 41.9 1 42.5 1
run;
proc lifereg data=cords;
 model  S*C(1) = /dist=weibull covb;
run;

OUTPUT from SAS Program:

The LIFEREG Procedure in SAS

Number of Observations                48
Noncensored Values                    41
Right Censored Values                  7
Name of Distribution             Weibull

                          Standard   95% Confidence     Chi-
Parameter     DF Estimate    Error       Limits       Square Pr > ChiSq
Intercept      1   4.0257   0.0100   4.0061   4.0454  161324     <.0001
Scale          1   0.0615   0.0077   0.0481   0.0786
Weibull Scale  1  56.0223   0.5615  54.9325  57.1337
Weibull Shape  1  16.2591   2.0363  12.7202  20.7827
```

The estimators of γ and α are obtained by recalling, $F(t) = 1 - e^{-(t/\alpha)^\gamma}$.

Thus, α is the Weibull Scale parameter and γ is the Weibull Shape parameter.

$$\hat{\gamma} = \text{Weibull Shape} = 16.2591; \quad \hat{\alpha} = \text{Weibull Scale} = 56.0223$$

With $S(t) = e^{-(t/\alpha)^\gamma}$, the estimate of $S(53)$ would be given by:

$$\widehat{S}(53) = e^{-(53/\hat{\alpha})^{\hat{\gamma}}} = e^{-(53/56.0223)^{16.2591}} = .6664$$

Thus, we would estimate that approximately 67% of the cords would have strength of at least 53

From the uncensored data, we have 28 of the 41 values are greater than 53, that is, 68.3%. The two values are very close.

The estimated standard error of the estimator (see STAT 611) is obtained from the inverse of the information matrix: $\widehat{SE}(\hat{S}(53)) = 0.046$

We can use the following R code to obtain the estimates of α and γ for a Weibull model with censored data:

```
 R Code in eCampus under the name cords_censoredMLE.R
#Estimate parameters using censoring
 library(MASS)
 library(survival)
st = c(36.3,  52.4,  54.8,  57.1,  60.7,  41.7,  52.6,  54.8, 57.3,
       43.9,  52.7,  55.1,  57.7,  49.4,  53.1,  55.4,  57.8, 50.1,
       53.6,  55.9,  58.1,  50.8,  53.6,  56.0,  58.9,  51.9, 53.9,
       56.1,  59.0,  52.1,  53.9,  56.5,  59.1,  52.3,  54.1, 56.9,
       59.6,  52.3,  54.6,  57.1,  60.4,
       26.8,  29.6,  33.4,  35.0,  40.0,  41.9,  42.5)
stcens = c(rep(1,41),rep(0,7))
cords = survreg(Surv(st, stcens) ~ 1, dist='weibull')
summary(cords)

#Estimate parameters ignoring censoring
fitdistr(st,"weibull",lower=c(0,0))

OUTPUT from R Code:
Call:
survreg(formula = Surv(st, stcens) ~ 1, dist = "weibull")
             Value Std. Error     z         p
(Intercept)  4.03     0.0100 401.7  0.00e+00
Log(scale)  -2.79     0.1252 -22.3 7.82e-110

Scale= 0.0615

# MLE's ignoring censoring:

     shape        scale
   9.4232284   54.5099326
 ( 1.1920947) ( 0.8643678)
```

From the above R function we must reinterpret the parameters to obtain:

$$\hat{\gamma} = 1/\text{Scale} = 1/.0615 = 16.2602, \qquad \hat{\alpha} = e^{\text{Intercept}} = e^{4.03} = 56.2609$$

and

$$\widehat{S}(53) = e^{-(53/\hat{\alpha})^{\hat{\gamma}}} = e^{-(53/56.2609)^{16.2602}} = .6847$$

Recall $\widehat{S}(53) = .6664$ using SAS code. The difference is possibly due to round-off error or precision differences in the calculations.

If the censoring of seven values had been ignored, the mle's based on 48 uncensored values are

$$\hat{\gamma} = 9.4232, \qquad \hat{\alpha} = 54.5099, \qquad \text{and} \qquad \widehat{S}(53) = e^{-(53/\hat{\alpha})^{\hat{\gamma}}} = e^{-(53/54.5099)^{9.4232}} = .4642$$

This displays a substantial change in the estimated proportion of cords that would have strength greater than 53 units, 46.4% ignoring censoring and 66.7% taking censoring into account.

The following example will further illustrate the importance of taking into account whether or not a data value is censored. Daily rainfall in millimeters was recorded over a 47 year period in Sydney Australia. The greatest amount of rain falling in a 24 hour period was recorded for each of the 47 years. There was a problem with the instruments that recorded the rainfall and any value greater than 2000 was considered to be a very inaccurate measurement of the true rainfall. Thus, we have right censored data and all values greater than 2000 will be set at 2000 prior to the data being analyzed. A Weibull model was fit to three sets of data: Data Set 1: original data values; Data Set 2: All values greater than 2000 are replaced by 2000; and Data Set 3: All values greater than 2000 are deleted. The following estimates of α and γ in the Weibull model along with the estimated quantiles for both tails of the distribution illustrate the importance of correctly taking into account censored values. The next pages contains the three data sets and a graph of the three versions of the Weibull pdf.

```
Original Data Set:  47 Maximum Daily Rainfall Values
 1468 3830  909 1781 2675  955 1565 1800  909 1397 2002 1717 1872 1849  701
  580  841  556 1331 2718 1359  719  994 1106  475  978 1227  584 1544 1737
 1188  808  846 1715 2543 1859 1372 1389  962  850  452  747 2649 1138 1334
  681 1564

Censored   Values: Rainfall larger than 2000 replaced with 2000 and designated as Censored
   1468 2000  909 1781 2000  955 1565 1800  909 1397 2000 1717 1872 1849  701
    580  841  556 1331 2000 1359  719  994 1106  475  978 1227  584 1544 1737
   1188  808  846 1715 2000 1859 1372 1389  962  850  452  747 2000 1138 1334
    681 1564

Delete Censored Rainfalls:   All maximum rainfall values greater than 2000  are Deleted
   1468  909 1781  955 1565 1800  909 1397 1717 1872 1849  701  580  841  556
   1331 1359  719  994 1106  475  978 1227  584 1544 1737 1188  808  846 1715
   1859 1372 1389  962  850  452  747 1138 1334  681 1564

MLEs for Two Parameters and Various Quantiles in a Weibull Model:

MLE for Weibull Model Parameters :

Original Data:              shape =  2.1103  ( 0.2246)       scale = 1537.2333  ( 112.1696)
Censored Data:              shape =  2.5966  ( 0.3303)       scale = 1472.6650  (  89.1911)
Censored values deleted:    shape =  3.0409  ( 0.3811)       scale = 1314.0159  (  71.2643)

Estimates of 9 Quantiles Using the 3 Methods:

                  Q(.01)   Q(.05)    Q(.10)    Q(.25)    Q(.5)     Q(.75)    Q(.9)     Q(.95)   Q(.99)
Original Data     173.80   376.25    529.20    851.80    1292.15   1794.57   2282.34   2585.46  3169.78
WithCensored      250.44   469.16    619.04    911.43    1278.81   1670.07   2030.49   2247.60  2651.75
DeleteCensored    289.48   494.77    626.91    872.30    1164.81   1463.02   1728.68   1884.94  2171.25
```

Impact of Censoring

Original Data Set: 47 Maximum Daily Rainfall Values

1468 3830 909 1781 2675 955 1565 1800 909 1397 2002 1717 1872 1849 701
580 841 556 1331 2718 1359 719 994 1106 475 978 1227 584 1544 1737
1188 808 846 1715 2543 1859 1372 1389 962 850 452 747 2649 1138 1334
681 1564

Censored Values: Rainfall larger than 2000 replaced with 2000 and designated as Censored

1468 2000 909 1781 2000 955 1565 1800 909 1397 2000 1717 1872 1849 701
580 841 556 1331 2000 1359 719 994 1106 475 978 1227 584 1544 1737
1188 808 846 1715 2000 1859 1372 1389 962 850 452 747 2000 1138 1334
681 1564

Delete Censored Rainfalls: All maximum rainfall values greater than 2000 are Deleted

1468 909 1781 955 1565 1800 909 1397 1717 1872 1849 701 580 841 556
1331 1359 719 994 1106 475 978 1227 584 1544 1737 1188 808 846 1715
1859 1372 1389 962 850 452 747 1138 1334 681 1564

MLEs for Two Parameters and Various Quantiles in a Weibull Model

MLE for Weibull Model :

Original Data: shape = 2.1103 (0.2246) scale = 1537.2333 (112.1696)

Censored Data: shape = 2.5966 (0.3303) scale = 1472.6650 (89.1911)

Censored values deleted: shape = 3.0409 (0.3811) scale = 1314.0159 (71.2643)

	Q(.01)	Q(.05)	Q(.10)	Q(.25)	Q(.5)	Q(.75)	Q(.9)	Q(.95)	Q(.99)
Original Data	173.80	376.25	529.20	851.80	1292.15	1794.57	2282.34	2585.46	3169.78
WithCensored	250.44	469.16	619.04	911.43	1278.81	1670.07	2030.49	2247.60	2651.75
DeleteCensored	289.48	494.77	626.91	872.30	1164.81	1463.02	1728.68	1884.94	2171.25

Distribution-Free Estimators

Don't know which model fits our data (in previous example weibull was good model of our distribution.

Suppose it is **unknown** to which family of distributions the cdf of T_i belongs. We want to use the data to estimate the survival function: $S(t) = Pr[T > t]$ and parameters associated with the survival function, such as, μ, $\tilde{\mu}$, and σ.

- **No Censoring**

1. Estimator of μ based on $\hat{S}(t)$

   With no censoring, $\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > t)$ = proportion of $T_i's$ greater than t.

   $$E[T] = \mu = \int_0^\infty t dF(t) = tF(t)|_0^\infty - \int_0^\infty F(t)dt = \int_0^\infty [1 - F(t)]dt = \int_0^\infty S(t)dt \Rightarrow$$

   $$\hat{\mu} = \int_0^\infty \hat{S}(t)dt = \frac{1}{n}\sum_{i=1}^{n}\int_0^\infty I(T_i \geq t)dt = \frac{1}{n}\sum_{i=1}^{n}\int_0^{T_i} dt = \frac{1}{n}\sum_{i=1}^{n} T_i = \bar{T}$$

2. Estimator of median, M based on $\hat{S}(t)$

   Recall, $S(t) = 1 - F(t)$, so we have

   $$Q(u) \;=\; inf(t : F(t) \geq u) \;=\; inf(t : 1 - S(t) \geq u)\; inf(t : S(t) \leq 1 - u)$$

   Therefore, the median is defined by

   $$M = Q(.5) \;=\; inf(t : F(t) \geq .5) \;=\; inf(t : S(t) \leq .5)$$

   Thus, we can estimate the median by

   $$\hat{M} = inf(t : \hat{S}(t) \leq .5)$$

   In general the pth quantile is estimated by

   $$\widehat{Q}(p) = inf(t : \hat{S}(t) \leq 1 - p)$$

STOP 10/1/21

Kaplan-Meier Product Limit Estimator of S(t)

Suppose we have n homogeneous units placed on test.

Let $t_1 < t_2 < \cdots < t_k$ be the times at which units fail

Let d_j units fail at time t_j for $j = 1, 2, \ldots, k$

Let m_j be the number of units that are censored in the interval $[t_j,\ t_{j+1})$ for $j = 0, 1, \ldots, k$,
with $t_o = 0$ and $t_{k+1} = \infty$

Let $n_j = (m_j + d_j) + (m_{j+1} + d_{j+1}) + \cdots + (m_k + d_k)$ be number of units at risk for failure just prior to time t_j → e.g. the # of units still around at time t_j

Note: $n_1 = n - m_0,\ n_2 = n - (d_1 + m_1)$

```
     mo        d1    m1              d2       m2       d3       m3          d4
|-------------|-------------------|---------------|------------------|-----
0             t1                  t2              t3                  t4
```

The Kaplan-Meier product limit estimator of $S(t) = P[T > t]$, where T is the time to failure, is given by

$$\text{For } t \in [t_i,\ t_{i+1}), \quad \widehat{S}(t) = \prod_{j=1}^{i} \frac{n_j - d_j}{n_j}$$

The estimator $\widehat{S}(t)$ is undefined for $t > t_k$

An heuristic justification of the estimator is as follows.

For $t \in [0,\ t_1)$, $S(t) = P[T > t] = P[\text{unit survives beyond time t}] \Rightarrow$

$\widehat{S}(t) = 1$ because there are no failures in $[0,\ t_1)$

For $t \in [t_1,\ t_2)$,

$S(t) = P[T > t] = P[\text{unit survives in } [t_1,\ t] \mid \text{unit survives in } [0,\ t_1)]\ P[\text{unit survives in } [0,\ t_1)] \Rightarrow$

$$\widehat{S}(t) = \left(\tfrac{n_1 - d_1}{n_1}\right) \cdot 1$$

For $t \in [t_2,\ t_3)$,

$S(t) = P[T > t] = P[\text{unit survives in } [t_2,\ t] \mid \text{survives in } [t_1,\ t_2)]\ P[\text{unit survives in } [t_1,\ t_2)] \Rightarrow$

$$\widehat{S}(t) = \left(\tfrac{n_2 - d_2}{n_2}\right)\left(\tfrac{n_1 - d_1}{n_1}\right)$$

For $t \in [t_3,\ t_4)$,

$S(t) = P[T > t] = P[\text{unit survives in } [t_3,\ t] \mid \text{survives in } [t_2,\ t_3)]\ P[\text{unit survives in } [t_2,\ t_3)] \Rightarrow$

$$\widehat{S}(t) = \left(\tfrac{n_3 - d_3}{n_3}\right)\left[\left(\tfrac{n_2 - d_2}{n_2}\right)\left(\tfrac{n_1 - d_1}{n_1}\right)\right]$$

From the Kaplan-Meier estimator we can then obtain estimators of the mean and median similarly as was done for uncensored data:

$$\hat{\mu} = \int_0^\infty \hat{S}(t)dt \Rightarrow$$

$$\hat{\mu} = \quad \text{area under} \quad \hat{S}(\cdot) \quad \text{curve} \quad = \quad \sum_{i=1}^{k} [T_{(i)} - T_{(i-1)}] S(T_{(i-1)})$$

where $T_{(i)}$ are the observed times to the event for the k uncensored units, with $T_{(0)} = 0, S(0) = 1$.

The estimated standard error of $\widehat{S}(t)$ can be obtained from the Greenwood Formula:

Case 1: No Censoring

$\widehat{S}(t) = \frac{n_t}{n}$, where n_t is the number of n units still working at time t

$Var(\widehat{S}(t)) = \frac{S(t)[1-S(t)]}{n} \quad \Rightarrow \quad \widehat{SE}[\widehat{S}(t)] = \sqrt{\frac{\hat{S}(t)[1-\hat{S}(t)]}{n}}$

Case 2: Censoring with Tied Observations

$$\widehat{SE}[\widehat{S}(t)] = \widehat{S}(t) \sqrt{\sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}}$$

The courses STAT 645 - 646 will extensively cover the topic of censored data and survival analysis. The following books are good references for the analysis of censored data and survival analysis.

- *The Statistical Models and Methods for Lifetime Data* by Lawless
- *Survival Analysis Using S* by Tableman and Kim
- *Statistical Methods for Reliability Data* by Meeker and Escobar

The Kaplin-Meier PL Estimator will now be obtained for the data in the cord strength example using the following SAS code:

cord_censored_KM.sas

```
option ls=75 ps=55 nocenter nodate;
title 'Strength of Braided Cord'
data cords;
input S C @@;
label S  = 'Strength of Cord' C ='Censoring (1=Yes)';
cards;
36.3 0 52.4 0 54.8 0 57.1 0 60.7 0 41.7 0 52.6 0 54.8 0 57.3 0
43.9 0 52.7 0 55.1 0 57.7 0 49.4 0 53.1 0 55.4 0 57.8 0
50.1 0 53.6 0 55.9 0 58.1 0 50.8 0 53.6 0 56.0 0 58.9 0
51.9 0 53.9 0 56.1 0 59.0 0 52.1 0 53.9 0 56.5 0 59.1 0
52.3 0 54.1 0 56.9 0 59.6 0 52.3 0 54.6 0 57.1 0 60.4 0
26.8 1 29.6 1 33.4 1 35.0 1 40.0 1 41.9 1 42.5 1
run;
proc lifetest data=cords outsurv=a plots=(s);
 time S*C(1) ;
 proc print data=a;
 run;
```

OUTPUT FROM LIFETEST:

The LIFETEST Procedure

Product-Limit Survival Estimates

S	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	48
26.8000*	.	.	.	0	47
29.6000*	.	.	.	0	46
33.4000*	.	.	.	0	45
35.0000*	.	.	.	0	44
36.3000	0.9773	0.0227	0.0225	1	43
40.0000*	.	.	.	1	42
41.7000	0.9540	0.0460	0.0318	2	41
41.9000*	.	.	.	2	40
42.5000*	.	.	.	2	39
43.9000	0.9295	0.0705	0.0393	3	38
49.4000	0.9051	0.0949	0.0452	4	37
50.1000	0.8806	0.1194	0.0502	5	36
50.8000	0.8562	0.1438	0.0544	6	35
51.9000	0.8317	0.1683	0.0581	7	34
52.1000	0.8072	0.1928	0.0613	8	33
52.3000	.	.	.	9	32
52.3000	0.7583	0.2417	0.0667	10	31
52.4000	0.7338	0.2662	0.0688	11	30
52.6000	0.7094	0.2906	0.0708	12	29
52.7000	0.6849	0.3151	0.0724	13	28
53.1000	0.6605	0.3395	0.0739	14	27
53.6000	.	.	.	15	26
53.6000	0.6115	0.3885	0.0761	16	25
53.9000	.	.	.	17	24
53.9000	0.5626	0.4374	0.0774	18	23
54.1000	0.5382	0.4618	0.0778	19	22
54.6000	0.5137	0.4863	0.0781	20	21
54.8000	.	.	.	21	20
54.8000	0.4648	0.5352	0.0779	22	19
55.1000	0.4403	0.5597	0.0776	23	18
55.4000	0.4158	0.5842	0.0770	24	17
55.9000	0.3914	0.6086	0.0763	25	16
56.0000	0.3669	0.6331	0.0753	26	15
56.1000	0.3425	0.6575	0.0742	27	14
56.5000	0.3180	0.6820	0.0728	28	13
56.9000	0.2935	0.7065	0.0712	29	12
57.1000	.	.	.	30	11
57.1000	0.2446	0.7554	0.0672	31	10
57.3000	0.2202	0.7798	0.0648	32	9
57.7000	0.1957	0.8043	0.0620	33	8
57.8000	0.1712	0.8288	0.0589	34	7
58.1000	0.1468	0.8532	0.0553	35	6
58.9000	0.1223	0.8777	0.0512	36	5
59.0000	0.0978	0.9022	0.0465	37	4
59.1000	0.0734	0.9266	0.0408	38	3
59.6000	0.0489	0.9511	0.0337	39	2
60.4000	0.0245	0.9755	0.0242	40	1
60.7000	0	1.0000	0	41	0

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable S

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	57.1000	55.9000	58.9000
50	54.8000	53.6000	56.1000
25	52.4000	50.8000	53.9000

Mean	Standard Error
54.1824	0.7316

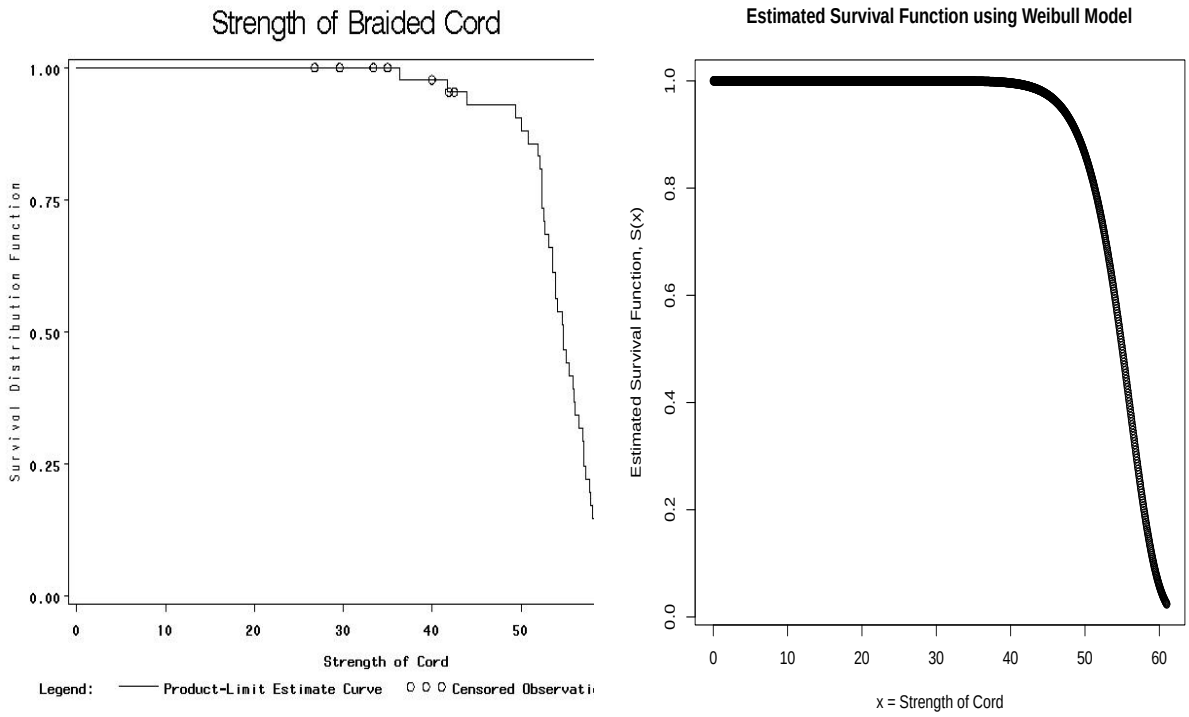
Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
48	41	7	14.58

Strength of Braided Cord

Obs	S	_CENSOR_	SURVIVAL	SDF_LCL	SDF_UCL
1	0.0	.	1.00000	1.00000	1.00000
2	26.8	0	1.00000	.	.
3	29.6	0	1.00000	.	.
4	33.4	0	1.00000	.	.
5	35.0	0	1.00000	.	.
6	36.3	1	0.97727	0.84941	0.99677
7	40.0	0	0.97727	.	.
8	41.7	1	0.95400	0.82832	0.98830
9	41.9	0	0.95400	.	.
10	42.5	0	0.95400	.	.
11	43.9	1	0.92954	0.79702	0.97675
12	49.4	1	0.90508	0.76630	0.96332
13	50.1	1	0.88062	0.73639	0.94855
14	50.8	1	0.85616	0.70725	0.93274
15	51.9	1	0.83170	0.67882	0.91607
16	52.1	1	0.80723	0.65102	0.89867
17	52.3	1	0.75831	0.59709	0.86206
18	52.4	1	0.73385	0.57087	0.84298
19	52.6	1	0.70939	0.54510	0.82342
20	52.7	1	0.68493	0.51975	0.80344
21	53.1	1	0.66046	0.49480	0.78305
22	53.6	1	0.61154	0.44606	0.74114
23	53.9	1	0.56262	0.39877	0.69781
24	54.1	1	0.53816	0.37565	0.67563
25	54.6	1	0.51369	0.35288	0.65313
26	54.8	1	0.46477	0.30837	0.60712
27	55.1	1	0.44031	0.28665	0.58362
28	55.4	1	0.41585	0.26528	0.55979
29	55.9	1	0.39139	0.24428	0.53562
30	56.0	1	0.36692	0.22366	0.51109
31	56.1	1	0.34246	0.20344	0.48620
32	56.5	1	0.31800	0.18363	0.46094
33	56.9	1	0.29354	0.16426	0.43528
34	57.1	1	0.24462	0.12696	0.38266
35	57.3	1	0.22015	0.10911	0.35563
36	57.7	1	0.19569	0.09188	0.32806
37	57.8	1	0.17123	0.07533	0.29988
38	58.1	1	0.14677	0.05959	0.27101
39	58.9	1	0.12231	0.04479	0.24133
40	59.0	1	0.09785	0.03115	0.21067
41	59.1	1	0.07338	0.01900	0.17881
42	59.6	1	0.04892	0.00889	0.14542
43	60.4	1	0.02446	0.00193	0.11054
44	60.7	1	0.00000	.	.

A plot of the Kaplan-Meier Estimator of S(t) and the MLE of S(t) based on a Weibull model are given below.



A comparison of the KM estimator to the values obtained from the Weibull model are displayed next:

$$\hat{S}_W(53.1) = e^{-(53.1/56.0223)^{16.2591}} = .6508 \qquad \hat{S}_{KM}(53.1) = .6605$$

$$\begin{aligned}
\hat{\mu}_{KM} &= \sum_{i=1}^{n_\nu} \left[T_{(i)} - T_{(i-1)}\right] S(T_{(i-1)} \\
&= (36.3-0)(1.0) + (41.7-36.3)(.97727) + (43.9-41.7)(.954) + \cdots (60.7-60.4)(.02446) \\
&= 54.1824
\end{aligned}$$

$$\hat{\mu}_W = \hat{\alpha}\Gamma\left(\frac{1+\hat{\gamma}}{\hat{\gamma}}\right) = \Gamma\left(\frac{17.2602}{16.2602}\right) = 54.4630$$

$$\hat{Q}_{KM}(.5) = \hat{S}_{K-M}(.5) = 54.8000$$

$$\hat{Q}_W(.5) = \hat{\alpha}(-log(.5))^{1/\hat{\gamma}} = (56.2609)(-log(.5))^{1/16.2602} = 55.0069$$

The estimates based on the Kaplan-Meier distribution free procedure and the MLE's based on a Weibull model are very close when the correct model is fit using the MLE of the parameters.

An analysis of the strength of the cords can also be analyzed using the following R code:

censoredcords_KM.R

```
 library(MASS)
 library(survival)


st = c(36.3,  52.4,  54.8,  57.1,  60.7,  41.7,  52.6,  54.8, 57.3,
       43.9,  52.7,  55.1,  57.7,  49.4,  53.1,  55.4,  57.8, 50.1,
       53.6,  55.9,  58.1,  50.8,  53.6,  56.0,  58.9,  51.9, 53.9,
       56.1,  59.0,  52.1,  53.9,  56.5,  59.1,  52.3,  54.1, 56.9,
       59.6,  52.3,  54.6,  57.1,  60.4,
       26.8,  29.6,  33.4,  35.0,  40.0,  41.9,  42.5)

stcens = c(rep(1,41),rep(0,7))

Surv(st, stcens)

cords.surv <- survfit(Surv(st, stcens) ~ 1,conf.type="log-log")
summary(cords.surv)
print(cords.surv,print.rmean=TRUE)


plot(cords.surv,conf.int=FALSE,log=FALSE,
main="Kaplan-Meier Estimator of Survival Function",xlab="Strength of Cord",
ylab="Survival Function")
```

Estimator from R Code:

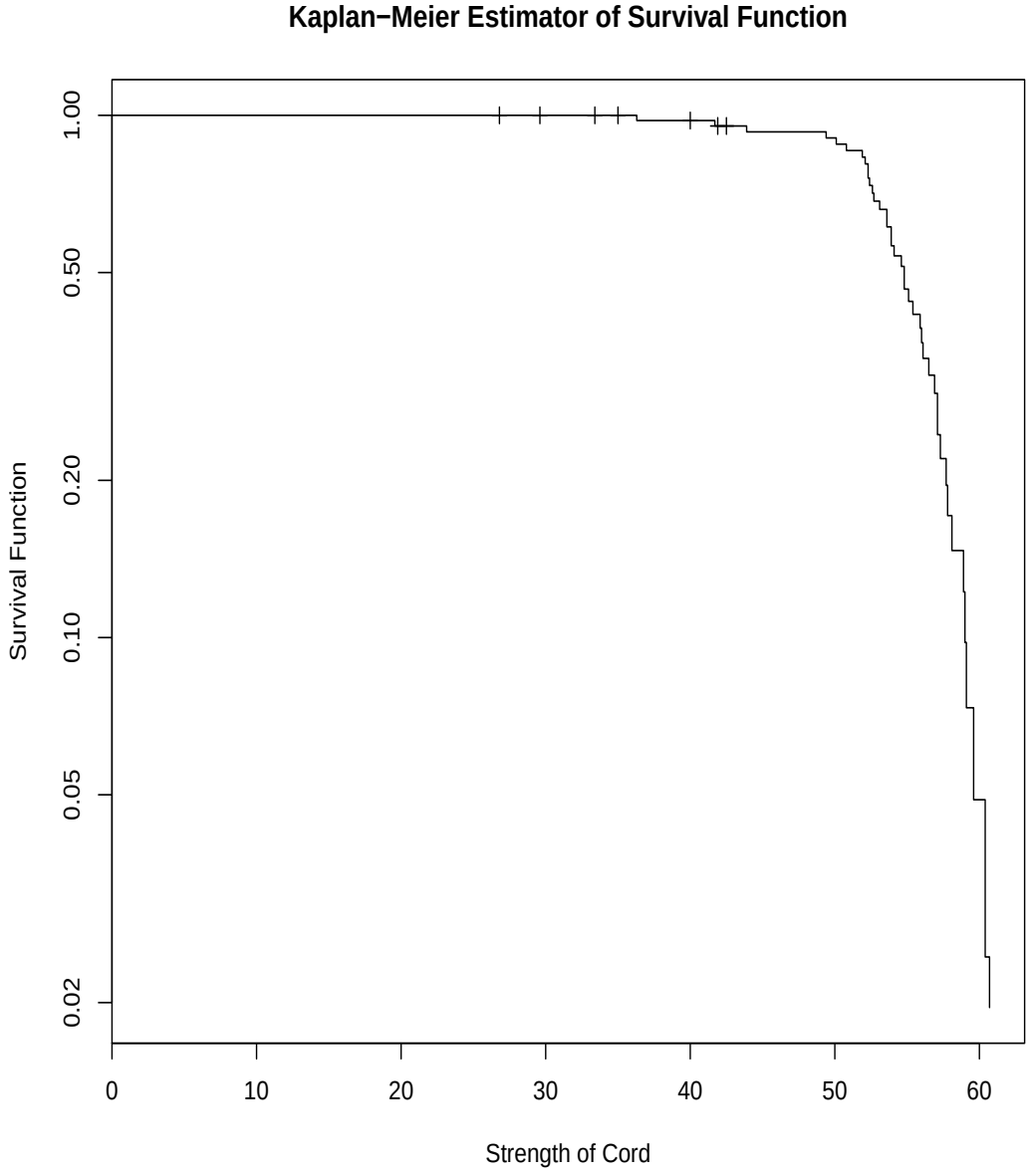
```
Call: survfit(formula = Surv(st, stcens) ~ 1, conf.type = "log-log")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
 36.3     44       1   0.9773  0.0225      0.84941        0.997
 41.7     42       1   0.9540  0.0318      0.82832        0.988
 43.9     39       1   0.9295  0.0393      0.79702        0.977
 49.4     38       1   0.9051  0.0452      0.76630        0.963
 50.1     37       1   0.8806  0.0502      0.73639        0.949
 50.8     36       1   0.8562  0.0544      0.70725        0.933
 51.9     35       1   0.8317  0.0581      0.67882        0.916
 52.1     34       1   0.8072  0.0613      0.65102        0.899
 52.3     33       2   0.7583  0.0667      0.59709        0.862
 52.4     31       1   0.7338  0.0688      0.57087        0.843
 52.6     30       1   0.7094  0.0708      0.54510        0.823
 52.7     29       1   0.6849  0.0724      0.51975        0.803
 53.1     28       1   0.6605  0.0739      0.49480        0.783
 53.6     27       2   0.6115  0.0761      0.44606        0.741
 53.9     25       2   0.5626  0.0774      0.39877        0.698
 54.1     23       1   0.5382  0.0778      0.37565        0.676
 54.6     22       1   0.5137  0.0781      0.35288        0.653
 54.8     21       2   0.4648  0.0779      0.30837        0.607
 55.1     19       1   0.4403  0.0776      0.28665        0.584
 55.4     18       1   0.4158  0.0770      0.26528        0.560
 55.9     17       1   0.3914  0.0763      0.24428        0.536
 56.0     16       1   0.3669  0.0753      0.22366        0.511
 56.1     15       1   0.3425  0.0742      0.20344        0.486
 56.5     14       1   0.3180  0.0728      0.18363        0.461
 56.9     13       1   0.2935  0.0712      0.16426        0.435
 57.1     12       2   0.2446  0.0672      0.12696        0.383
 57.3     10       1   0.2202  0.0648      0.10911        0.356
 57.7      9       1   0.1957  0.0620      0.09188        0.328
 57.8      8       1   0.1712  0.0589      0.07533        0.300
 58.1      7       1   0.1468  0.0553      0.05959        0.271
 58.9      6       1   0.1223  0.0512      0.04479        0.241
 59.0      5       1   0.0978  0.0465      0.03115        0.211
 59.1      4       1   0.0734  0.0408      0.01900        0.179
 59.6      3       1   0.0489  0.0337      0.00889        0.145
 60.4      2       1   0.0245  0.0242      0.00193        0.111
 60.7      1       1   0.0000     NaN           NA           NA

Call: survfit(formula = Surv(st, stcens) ~ 1, conf.type = "log-log")

records   n.max    n.start  events  *rmean  *se(rmean)  median  0.95LCL    0.95UCL
 48.000  48.000    48.000   41.000   54.182   0.723     54.800   53.100     56.100
* restricted mean with upper limit =  60.7
```

A graph of the survival is produced next:


Kaplan−Meier Estimator of Survival Function
Survival Function
1.00
0.50
0.20
0.10
0.05
0.02
0
10
20
30
40
50
60
Strength of Cord

Comparing Two Survival Curves from Censored Data

EXAMPLE from *The Statistical Analysis of Failure Time Data*

The following table gives the times from insult with the carcinogen DMBA to mortality from vaginal cancer in rats. Two groups were distinguished by a pretreatment regime. Four times to mortality were randomly censored and are denoted by an *. For these rats, we can see that their times to mortality exceed 216, 244, 204, and 344 days, respectively, but we do not know the times exactly. The random censoring occurred because the four rats died of causes unrelated to the application of the carcinogen and they were free of tumor at death, or they may simply not have developed tumor at the time of data analysis. There are two separate groups of survival times data in this example.

	Days to Vaginal Cancer Mortality in Rats									
Group 1	143	164	188	188	190	192	206	209	213	216
	220	227	230	234	246	265	304	216*	244*	
Group 2	142	156	163	198	205	232	232	233	233	233
	233	239	240	261	280	280	296	296	323	204*
	344*									

The following SAS program was used to obtain a Kaplan-Meier estimator for each of the two groups of rats.

```
 eCampus  -  kmest_rat.sas;


option ls=75 ps=55 nocenter nodate;
title 'Cancer Treatment-Estimated S(t)';
data cancer;
input T ST G @@;
LGT=log(T);
label ST = 'Censoring Indicator';
label T  = 'Time to Death';
label G  = 'Treatment Group';
cards;
143 1 1 164 1 1 188 1 1 188 1 1 190 1 1 192 1 1 206 1 1
209 1 1 213 1 1 216 1 1 220 1 1 227 1 1 230 1 1 234 1 1
246 1 1 265 1 1 304 1 1 216 0 1 244 0 1
142 1 2 156 1 2 163 1 2 198 1 2 205 1 2 232 1 2 232 1 2
233 1 2 233 1 2 233 1 2 233 1 2 239 1 2 240 1 2 261 1 2
280 1 2 280 1 2 296 1 2 296 1 2 323 1 2 204 0 2 344 0 2
run;
proc lifetest data=cancer outsurv=a plots=(s);
 time  T*ST(0);
 strata G;
run;
proc print data=a;
run;
```

OUTPUT FROM SAS:

Cancer Treatment-Estimated S(t)

The LIFETEST Procedure

Stratum 1: G = 1

Product-Limit Survival Estimates

T	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000	1.0000	0	0	0	19
143.000	0.9474	0.0526	0.0512	1	18
164.000	0.8947	0.1053	0.0704	2	17
188.000	.	.	.	3	16
188.000	0.7895	0.2105	0.0935	4	15
190.000	0.7368	0.2632	0.1010	5	14
192.000	0.6842	0.3158	0.1066	6	13
206.000	0.6316	0.3684	0.1107	7	12
209.000	0.5789	0.4211	0.1133	8	11
213.000	0.5263	0.4737	0.1145	9	10
216.000	0.4737	0.5263	0.1145	10	9
216.000*	.	.	.	10	8
220.000	0.4145	0.5855	0.1145	11	7
227.000	0.3553	0.6447	0.1124	12	6
230.000	0.2961	0.7039	0.1082	13	5
234.000	0.2368	0.7632	0.1015	14	4
244.000*	.	.	.	14	3
246.000	0.1579	0.8421	0.0934	15	2
265.000	0.0789	0.9211	0.0728	16	1
304.000	0	1.0000	0	17	0

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable T

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	234.000	216.000	265.000
50	216.000	192.000	234.000
25	190.000	188.000	216.000

Mean	Standard Error
218.757	9.403

Stratum 2: G = 2

Product-Limit Survival Estimates

T	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000	1.0000	0	0	0	21
142.000	0.9524	0.0476	0.0465	1	20
156.000	0.9048	0.0952	0.0641	2	19
163.000	0.8571	0.1429	0.0764	3	18
198.000	0.8095	0.1905	0.0857	4	17
204.000*	.	.	.	4	16
205.000	0.7589	0.2411	0.0941	5	15
232.000	.	.	.	6	14
232.000	0.6577	0.3423	0.1053	7	13
233.000	.	.	.	8	12
233.000	.	.	.	9	11
233.000	.	.	.	10	10
233.000	0.4554	0.5446	0.1114	11	9
239.000	0.4048	0.5952	0.1099	12	8
240.000	0.3542	0.6458	0.1072	13	7
261.000	0.3036	0.6964	0.1031	14	6
280.000	.	.	.	15	5
280.000	0.2024	0.7976	0.0902	16	4
296.000	.	.	.	17	3
296.000	0.1012	0.8988	0.0678	18	2
323.000	0.0506	0.9494	0.0493	19	1
344.000*	.	.	.	19	0

NOTE: The marked survival times are censored observations.

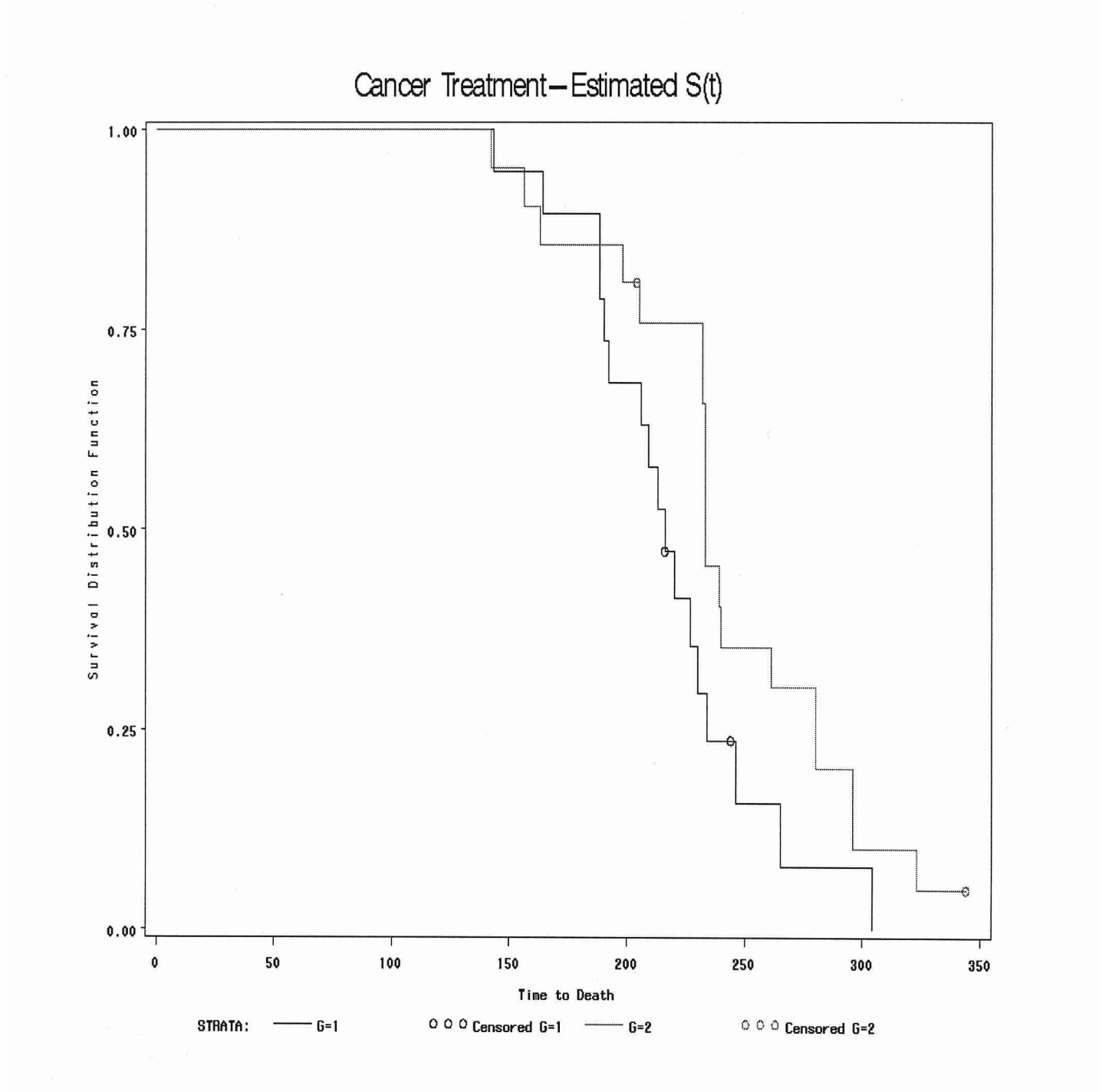
Summary Statistics for Time Variable T

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	280.000	233.000	296.000
50	233.000	232.000	280.000
25	232.000	163.000	233.000

Mean	Standard Error
240.795	11.206

A plot of the Kaplan-Meier Estimators of S(t) for the two groups is given below.



The following R code will produce similar results to what was achieved using the SAS code:

```
carcinogen_2G_KM.R

library(survival)

T = c( 143,164,188,188,190,192,206,209,213,216,
       220,227,230,234,246,265,304,216,244,
       142,156,163,198,205,232,232,233,233,233,
       233,239,240,261,280,280,296,296,323,204,
       344)

ST = c(rep(1,17),rep(0,2),rep(1,19),rep(0,2))

G = c(rep(1,19),rep(2,21))

out = cbind(T,ST,G)

Surv(T, ST)

carcin <- survfit(Surv(T, ST) ~ G)
summary(carcin)
print(carcin, print.rmean=TRUE,rmean="individual")

par(lab=c(15,20,4))
plot(carcin,ylab="Survival Function",xlab="Time to Death",mark.time=TRUE,
main="Cancer Treatment - Estimated S(t)",lty=1:2 )
legend(25,.8,c("Group 1","Group 2"),lty=1:2,lwd=2)
text(216,.4737,"+")
text(244,.2368,"+")
text(204,.8095,"+")
text(344,.0506,"+")
```

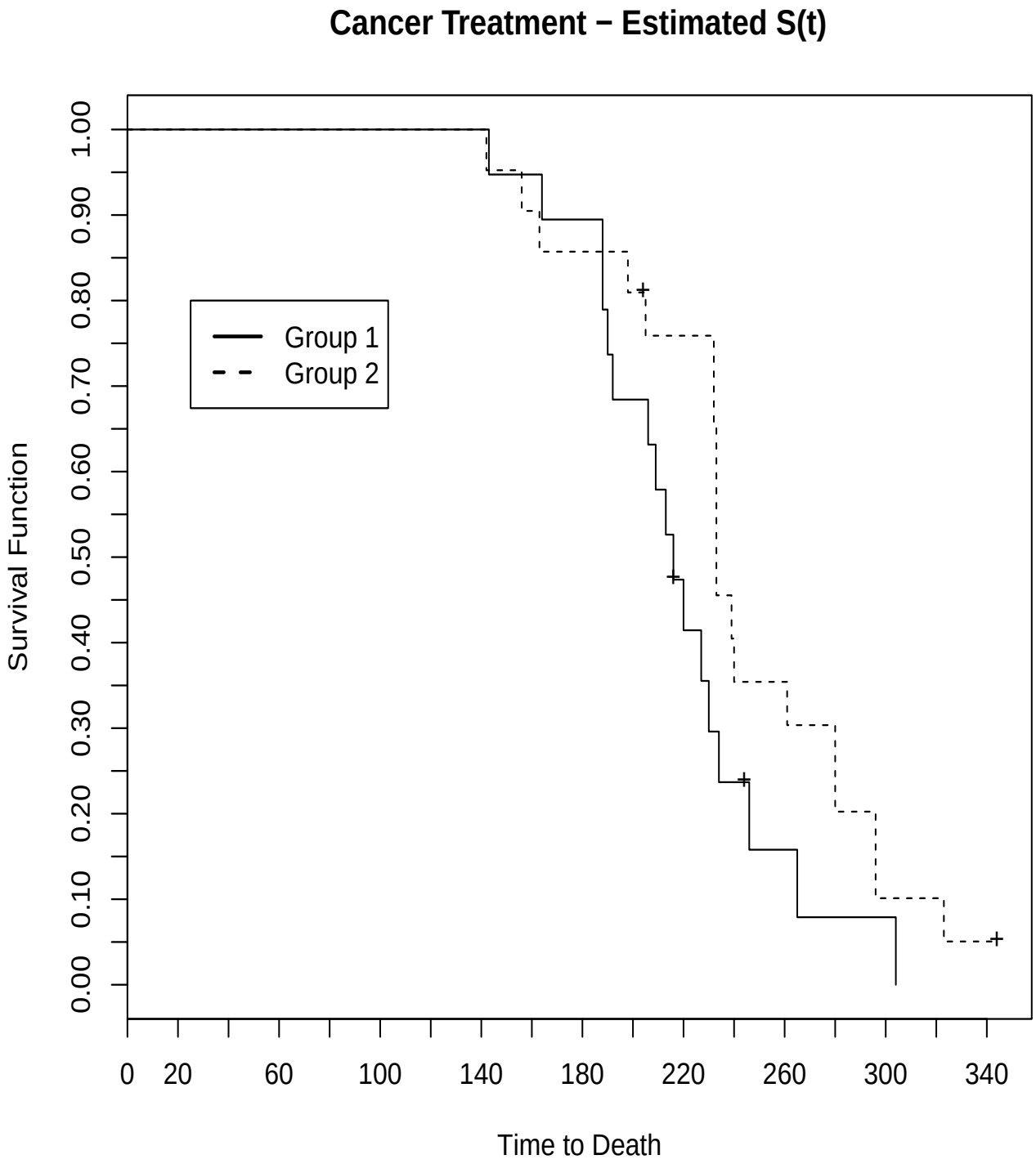
OUTPUT FROM R:

```
    records n.max n.start events *rmean *se(rmean) median 0.95LCL 0.95UCL
G=1      19    19      19     17    219       9.12    216     206     265
G=2      21    21      21     19    242      11.35    233     232     280
   * restricted mean with variable upper limit


Call: survfit(formula = Surv(T, ST) ~ G, conf.type = "log-log")

                G=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  143     19       1    0.947  0.0512      0.68119        0.992
  164     18       1    0.895  0.0704      0.64079        0.973
  188     17       2    0.789  0.0935      0.53191        0.915
  190     15       1    0.737  0.1010      0.47893        0.881
  192     14       1    0.684  0.1066      0.42794        0.844
  206     13       1    0.632  0.1107      0.37899        0.804
  209     12       1    0.579  0.1133      0.33208        0.763
  213     11       1    0.526  0.1145      0.28720        0.719
  216     10       1    0.474  0.1145      0.24438        0.673
  220      8       1    0.414  0.1145      0.19616        0.621
  227      7       1    0.355  0.1124      0.15191        0.566
  230      6       1    0.296  0.1082      0.11168        0.509
  234      5       1    0.237  0.1015      0.07578        0.447
  246      3       1    0.158  0.0934      0.03143        0.374
  265      2       1    0.079  0.0728      0.00567        0.288
  304      1       1    0.000     NaN           NA           NA

                G=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  142     21       1   0.9524  0.0465      0.70721        0.993
  156     20       1   0.9048  0.0641      0.67005        0.975
  163     19       1   0.8571  0.0764      0.61972        0.952
  198     18       1   0.8095  0.0857      0.56891        0.924
  205     16       1   0.7589  0.0941      0.51394        0.892
  232     15       2   0.6577  0.1053      0.41232        0.820
  233     13       4   0.4554  0.1114      0.23531        0.652
  239      9       1   0.4048  0.1099      0.19615        0.605
  240      8       1   0.3542  0.1072      0.15914        0.556
  261      7       1   0.3036  0.1031      0.12446        0.506
  280      6       2   0.2024  0.0902      0.06327        0.397
  296      4       2   0.1012  0.0678      0.01719        0.275
  323      2       1   0.0506  0.0493      0.00349        0.207
```


Cancer Treatment − Estimated S(t)
Survival Function
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
Group 1
Group 2
0
20
60
100
140
180
220
260
300
340
Time to Death