Survival Analysis: Nonparametric Estimators

Samiran Sinha
Texas A&M University
sinha@stat.tamu.edu

August 20, 2021

Survival data with right censoring

- Also known as "time-to-event" data
- Contains *censored* data
  - We'll focus on *right-censored* data: censored values known to be at least as big as the recorded value

Terminology:

- T_i: time-to-event for subject i
- C_i: censoring time for subject i
- $V_i = \min(T_i, C_i)$: observed time for subject i
- $\Delta_i = I(T_i \leq C_i)$ (censoring indicator)
  - 1 if an actual event occurred at time V_i (i.e., $V_i = T_i$)
  - 0 if censored, $T_i > V_i$

Important points for right censored data

- We don't get to observe T.
- For every subject we observe (V, Δ).
- We want to make inference on the distribution of T based on n independent observations $\{(V_i, \Delta_i),\ i = 1, \ldots, n\}$.

Example 1

Take this small example. Suppose that the interest is in the distribution of time to death (in months) from HIV diagnosis.

Subject	Observed time (V)	Censoring indicator (Δ)
1	5	1
2	6	0
3	8	1
4	3	1
5	22	1

Survival function

Define the cumulative distribution function $F(t) = \text{pr}(T \leq t)$ and the survival function

$$S(t) = 1 - F(t) = \text{pr}(T > t).$$

How do we estimate $S(t)$? In survival analysis, usually the interest is in estimating $S(t)$, the probability of survival up to time t.

In our example, events are observed at times $3, 5, 8, 22$ months. Estimate S with a step function with jumps between these times.

Survival function

To begin, for our convenience we arrange the data in the ascending order of V.

Subject	Observed time (V)	Censoring indicator (Δ)
4	3	1
1	5	1
2	6	0
3	8	1
5	22	1

Note that

$$\begin{aligned} \widehat{S}(0) &= \widehat{\text{pr}}(T > 0) \\ &= \frac{\#\text{subjects surviving more than 0 months}}{\text{Total } \#\text{subjects}} \\ &= \frac{5}{5} = 1. \end{aligned}$$

Survival function estimate

We use $\widehat{S}(t)$ to denote an estimator of $S(t)$.

$$\widehat{S}(3) = \widehat{\text{pr}}(T > 3) = \widehat{\text{pr}}(\text{ don't die at } T = 3 | \text{survive at least 3 months}) \times \widehat{\text{pr}}(\text{survive at least 3 months})$$

$$= \left(1 - \frac{1}{5}\right) \times 1 = 0.80$$

$$\widehat{\text{pr}}(\text{ don't die at } T = 3 | \text{ survive at least 3 months})$$

$$= \left(1 - \frac{\#\text{ deaths at 3}}{\#\text{ subjects who survived at least 3 months}}\right) = \left(1 - \frac{1}{5}\right)$$

$$\widehat{\text{pr}}(T \geq 3) = \widehat{\text{pr}}(\text{survive at least 3 months})$$

$$= \frac{\#\text{subjects who survived at least 3 months}}{\text{Total number of subjects}} = 1$$

Survival function estimate

$$\widehat{\text{pr}}(T > 5) = \widehat{\text{pr}}(\text{ don't die at } T = 5 | \text{survive at least 5 months}) \times \widehat{\text{pr}}(\text{survive at least 5 months})$$

Note that

$$\widehat{\text{pr}}(\text{ don't die at } T = 5 | \text{survive at least5 months}) = \left(1 - \frac{\#\text{ deaths at 5}}{\#\text{subjects who survived at least 5 months}}\right) = \left(1 - \frac{1}{4}\right)$$

$$\widehat{\text{pr}}(T \geq 5) = \widehat{\text{pr}}(\text{ survive at least 5 months}) = \widehat{\text{pr}}(T > 5-) = \widehat{\text{pr}}(T > 3) = 0.8$$

Hence,

$$\widehat{\text{pr}}(T > 5) = 0.75 \times 0.80 = 0.6$$

Survival function estimate

Why

$$\widehat{\text{pr}}(T > 5-) = \widehat{\text{pr}}(T > 3)?$$

Consider

$$\begin{aligned}
\widehat{\text{pr}}(T > 3.00001) &= \widehat{\text{pr}}(T > 3.00001 \cap T \geq 3.00001) \\
&= \widehat{\text{pr}}(T > 3.00001 | T \geq 3.00001)\widehat{\text{pr}}(T \geq 3.00001) \\
&= \{1 - \widehat{\text{pr}}(\text{death at time } 3.00001 | T \geq 3.00001)\}\widehat{\text{pr}}(T > 3) \\
&= (1 - \frac{0}{4}) \times 0.8 \\
&= 0.8.
\end{aligned}$$

Survival function estimate

Next, consider

$$\begin{aligned}
\widehat{\text{pr}}(T > 3.00002) &= \widehat{\text{pr}}(T > 3.00002 \cap T \geq 3.00002) \\
&= \widehat{\text{pr}}(T > 3.00002 | T \geq 3.00002)\widehat{\text{pr}}(T \geq 3.00002) \\
&= \{1 - \widehat{\text{pr}}(\text{death at time } 3.00002 | T \geq 3.00002)\}\widehat{\text{pr}}(T > 3.00001) \\
&= (1 - \frac{0}{4}) \times 0.8 \\
&= 0.8.
\end{aligned}$$

Following the above procedure we can show $\widehat{\text{pr}}(T > 5-) = \widehat{\text{pr}}(T > 3)$.

Survival function estimate continued:

Similarly,

$$\widehat{S}(6) = \widehat{\text{pr}}(T > 6) = \left(1 - \frac{0}{3}\right) \times 0.60 = 0.60$$

$$\widehat{S}(8) = \widehat{\text{pr}}(T > 8) = \left(1 - \frac{1}{2}\right) \times 0.60 = 0.30$$

$$\widehat{S}(22) = \widehat{\text{pr}}(T > 22) = \left(1 - \frac{1}{1}\right) \times 0.30 = 0.00$$

Survival function estimate

$$\begin{aligned}\widehat{\text{pr}}(T > 6.00001) &= \widehat{\text{pr}}(T > 6.00001 \cap T > 6 \text{ or } T \geq 6.00001) \\ &= \widehat{\text{pr}}(T > 6.00001 | T > 6 \text{ or } T \geq 6.00001)\text{pr}(T > 6 \text{ or } T \geq 6.00001)\end{aligned}$$

Note that

$$\begin{aligned}&\widehat{\text{pr}}(\text{ don't die at } T = 6.00001 | \text{survive at least 6.00001 months}) \\ &= \left(1 - \frac{\#\text{ deaths at 6.00001}}{\#\text{subjects who survived at least 6.00001 months}}\right) = \left(1 - \frac{0}{2}\right)\end{aligned}$$

$$\begin{aligned}\widehat{\text{pr}}(T \geq 6.00001) &= \widehat{\text{pr}}(\text{survive at least 6.00001 months}) \\ &= \widehat{\text{pr}}(T > 6) = 0.6\end{aligned}$$

Hence,

$$\widehat{S}(6.00001) = \widehat{\text{pr}}(T > 6.00001) = 1 \times 0.6 = 0.6.$$

Survival function estimate

$$
\begin{aligned}
&\widehat{\text{pr}}(T > 7) \\
=&\text{Probability of surviving more than 7 months} \\
=&\widehat{\text{pr}}(\text{ don't die at } T = 7 | \text{survive at least 7 months}) \\
&\times \widehat{\text{pr}}(\text{survive at least 7 months})
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\widehat{\text{pr}}(\text{don't die at } T = 7 | \text{survive at least 7 months}) \\
&= \left(1 - \frac{\#\text{ deaths at 7}}{\#\text{subjects who survived at least 7 months}}\right) = \left(1 - \frac{0}{2}\right)
\end{aligned}
$$

$$
\widehat{\text{pr}}(T \geq 7) = \widehat{\text{pr}}(\text{survive at least 7 months}) = \widehat{\text{pr}}(T > 7-) = \widehat{\text{pr}}(T > 6) = 0.6
$$

Hence,

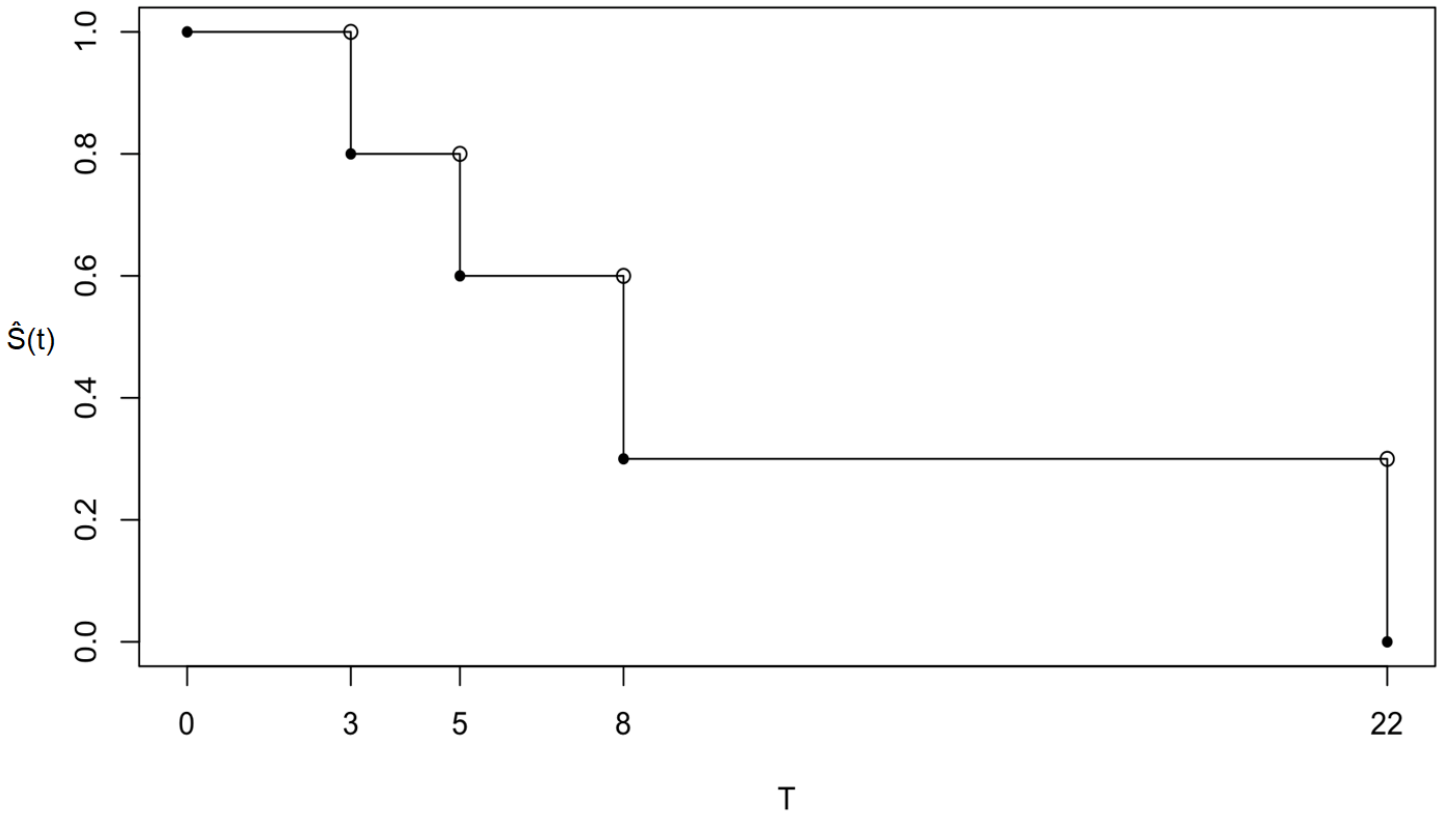
$$
\widehat{S}(7) = \widehat{\text{pr}}(T > 7) = 1 \times 0.6 = 0.6.
$$

To do a plot of the survival function

Code

```
myx=c(0, 3, 5, 6, 8, 22)
myy=c(1, 1,  0.8, 0.6, 0.6, 0.3, 0)
plot(stepfun(myx, myy))
```

Survival function estimate continued



Kaplan-Meier estimator

What we have done so far is called Kaplan-Meier estimation. Formally it is given by:

$$\widehat{S}(t) = \prod_{t(i) \leq t} \frac{n_i - d_i}{n_i} = \prod_{t(i) \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

where $\widehat{\lambda}_i = d_i/n_i$, estimator of the hazard

- $t_{(1)}, t_{(2)}, \ldots, t_{(m)}$ are the ordered unique event times
- n_i is number "at risk" at time $t_{(i)}$
- d_i is the number of actual "deaths" at time $t_{(i)}$ (does not include the censored events)

Kaplan-Meier estimator:

Subject	Observed time (V)	Censoring indicator (Δ)	n_i	d_i	$\widehat{\lambda}_i$	$\widehat{S}(t) = (1-\widehat{\lambda}_i)\widehat{S}(t-)$
4	3	1	5	1	0.2	$(1-0.2)\times 1 = 0.8$
1	5	1	4	1	0.25	$(1-0.25)\times 0.8 = 0.6$
2	6	0	3	0	0	$(1-0)\times 0.6 = 0.6$
3	8	1	2	1	0.5	$(1-0.5)\times 0.6 = 0.3$
5	22	1	1	1	1	$(1-1)\times 0.3 = 0$

Lung cancer data

Code

```
library(survival)
data(lung)
head(lung)
lung$SurvObj <- with(lung, Surv(time, status == 2))
head(lung)
km.as.one <- survfit(SurvObj ~ 1, data = lung, conf.type = "log-log")
plot(km.as.one)
# to obtain a nicer colored figure I use
plot(km.as.one, col=c("red", "blue", "blue"), lwd=2)
```

Kaplan-Meier estimator:

- Confidence interval for the Kaplan-Meier estimator can be calculated using different approaches, some options are: plain, log, log-log.
- The *log* is the default option for conf.type.

Nelson Aalen estimator

- Note that the survival function $S(t)$ and the cumulative hazard function $\Lambda(t)$ are related via

$$S(t) = \exp\{-\Lambda(t)\}.$$

- Nelson Aalen estimator of $\Lambda(t)$ is

$$\widehat{\Lambda}(t) = \sum_{t_{(i)} \leq t} \frac{d_i}{n_i},$$

therefore, another estimator of $S(t)$ is then

$$\widehat{S}(t) = \exp\{-\widehat{\Lambda}(t)\} = \exp\{-\sum_{t_{(i)} \leq t} \frac{d_i}{n_i}\}.$$

- Note that the Kaplam-Meier estimator is

$$\widehat{S}(t) = \prod_{t(i) \leq t} \frac{n_i - d_i}{n_i} = \prod_{t(i) \leq t} \left(1 - \frac{d_i}{n_i}\right).$$

Nelson Aalen estimator

- Note that the approximate variance of $\widehat{\Lambda}(t)$ is

$$\sigma^2(t) = \sum_{t_{(i)} \leq t} \frac{d_i(n_i - d_i)}{n_i^2(n_i - 1)},$$

- Since for a large sample, $\widehat{\Lambda}(t)$ follows approximate normal distribution, the $(1-\alpha)100\%$ CI for $\Lambda(t)$ is $\widehat{\Lambda}(t) \pm Z_{1-\alpha/2}\sigma(t)$.
- Likewise the $(1-\alpha)100\%$ CI for the survival function $S(t)$ is

$$\exp\{-\widehat{\Lambda}(t) + Z_{1-\alpha/2}\sigma(t)\},\ \exp\{-\widehat{\Lambda}(t) - Z_{1-\alpha/2}\sigma(t)\}$$

Lung cancer data– Nelson-Aalen estimator

Plot of the survival function based on the Nelson -Aalen estimator

Code

```
library(survival)
data(lung)
head(lung)
```

	inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss	SurvObj
1	3	306	2	74	1	1	90	100	1175	NA	306
2	3	455	2	68	1	0	90	90	1225	15	455
3	3	1010	1	56	1	0	90	90	NA	15	1010+
4	5	210	2	57	1	1	90	60	1150	11	210
5	1	883	2	60	1	0	100	90	NA	0	883
6	12	1022	1	74	1	1	50	80	513	0	1022+

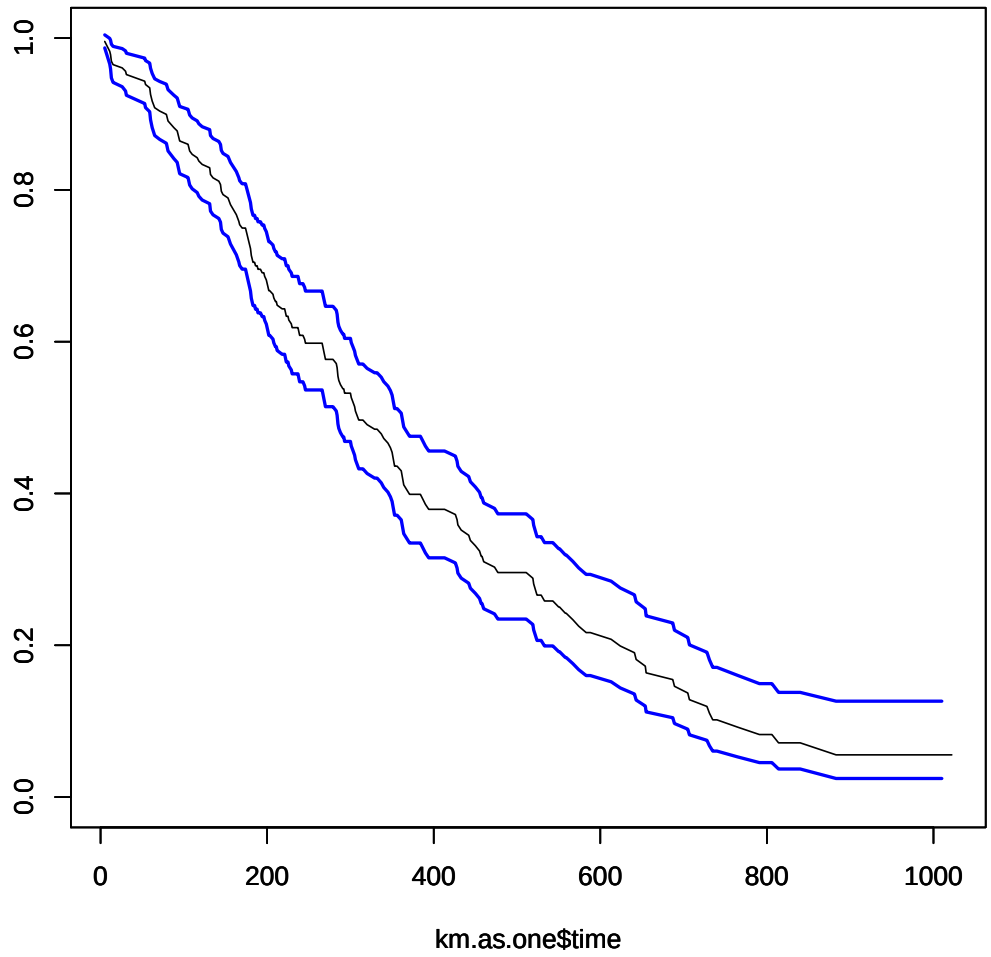
Lung cancer data– Nelson-Aalen estimator

Plot of the survival function based on the Nelson -Aalen estimator

Code

```
 lung$SurvObj <- with(lung, Surv(time, status == 2))
head(lung)
km.as.one <- survfit(SurvObj ~ 1, data = lung, conf.type = "log-log")
my.hazard=km.as.one$n.event/km.as.one$n.risk
cum.hazard=cumsum(my.hazard)
myvar=cumsum( km.as.one$n.event*(km.as.one$n.risk
-km.as.one$n.event)/(km.as.one$n.risk$^2$*(km.as.one$n.risk-1)) )
mysd=sqrt(myvar)
plot(km.as.one$time, exp(-cum.hazard), ylim=c(0, 1), ylab="", type="l")
par(new=T);
plot(km.as.one$time, exp(-cum.hazard-1.96*mysd), ylab="",col="blue",
lwd=2, ylim=c(0, 1), type="l")
par(new=T);
plot(km.as.one$time, exp(-cum.hazard+1.96*mysd), ylab="", col="blue",
lwd=2, ylim=c(0, 1), type="l")
```

Plot of the estimated survival function along with the 95% pointwise CI



Nelson-Aalen estimator of $\Lambda(t)$ and $S(t)$

Subject	Observed time (V)	Censoring indicator (Δ)	n_i	d_i	$\widehat{\lambda}_i$	$\sum_{t_{(i)} \leq t} \widehat{\lambda}_i$	$\widehat{S}(t) = \exp\{-\sum_{t_{(i)} \leq t} \widehat{\lambda}_i\}$
4	3	1	5	1	0.2	0.2	0.82
1	5	1	4	1	0.25	0.45	0.64
2	6	0	3	0	0	0.45	0.64
3	8	1	2	1	0.5	0.95	0.39
5	22	1	1	1	1	1.95	0.14

Compare these results with the Kaplan-Meier estimator

Mean time-to-event T

- Often we would like to estimate μ, the mean of T.
- If f denotes the density function of T, then $\mu = \int_0^\infty tf(t)dt$.
- The more useful formula is $\mu = \int_0^\infty S(t)dt$.
- The mean can be estimated by $\widehat{\mu} = \int_0^\infty \widehat{S}(t)dt$, usually the range of integration is taken as $(0, \tau)$ where τ largest observed time in the dataset, and $\widehat{S}(t)$ is the Kaplan-Meier estimator of $S(t)$.
- In other words, $\widehat{\mu}$ is the area under the estimated survival function.
- Let $0 = \tau_0 < \tau_1 < \cdots < \tau_k$ be the distinct time points (failure and censoring) of an observed data.
- Then $\widehat{\mu} = \sum_{i=1}^k \Delta\tau_i \widehat{S}(\tau_{i-1})$, where $\Delta\tau_i = \tau_i - \tau_{i-1}$.

Variance of the mean estimator

- Suppose that D =total number of failures in the dataset.
- Ordered failure times: $v_1^* < \cdots < v_D^*$

$$\text{Var}(\widehat{\mu}) = \sum_{i=1}^{D} \{ \int_{v_i^*}^{\tau} \widehat{S}(u) du \}^2 \times \frac{d_i}{n_i(n_i - d_i)}$$

- $100(1-\alpha)\%$ CI is $\widehat{\mu} \pm Z_{1-\alpha/2} \sqrt{\text{Var}(\widehat{\mu})}$.
- Besides this analytical formula for the standard error for a large sample size, the standard error of this estimator $\widehat{\mu}$ can also be calculated by the bootstrap method.

Percentile estimation

- pth percentile estimation: $\widehat{q}_p = \inf\{t : \widehat{S}(t) \leq (1-p)\}$, the smallest time at which the survival function is less than or equal to $(1-p)$
- median estimation (50th percentile): $\widehat{m} = \inf\{t : \widehat{S}(t) \leq 0.5\}$, the smallest time at which the survival function is less than or equal to 0.5
- 25th percentile estimation: $\widehat{q}_{0.25} = \inf\{t : \widehat{S}(t) \leq 0.75\}$, the smallest time at which the survival function is less than or equal to 0.75
- 75th percentile estimation: $\widehat{q}_{0.75} = \inf\{t : \widehat{S}(t) \leq 0.25\}$, the smallest time at which the survival function is less than or equal to 0.25

Mean time estimation

Code

```
library(survival)
data(lung)
head(lung)
lung$SurvObj <- with(lung, Surv(time, status == 2))
head(lung)
km.as.one <- survfit(SurvObj ~ 1, data = lung, conf.type = "log-log")
plot(km.as.one)
print(km.as.one, print.rmean=TRUE) #by default observed maximum time is
# considered to be tau
print(km.as.one, print.rmean=TRUE, rmean=1200) # here the upper
# limit is specified as 1200
```

- The above code produces standard error of $\widehat{\mu}$, and we can use it to construct a CI.
- The following code produces the median and its 95% CI.
- We can also estimate other percentiles of the distribution of T along with their CI.

Code

```
quantile(km.as.one, prob=c(0.25, 0.5, 0.75), conf.int=TRUE)
```

Some basic quantities

- Suppose that $f(t)$ is the density function of the time-to-event T.
- The survival function $S(t) = \text{pr}(T > t) = \int_t^\infty f(u)du$. Thus we can obtain the survival function from the density function.
- On the other hand, $f(t) = -dS(t)/dt$, hence the density can be obtained from the survival function.
- For a discrete valued T with mass points, $t_1 < t_2 < \cdots < t_k$, the survival function is $S(t) = \text{pr}(T > t) = \sum_{t_j > t} p_j$, where $p_j = \text{pr}(T = t_j)$.

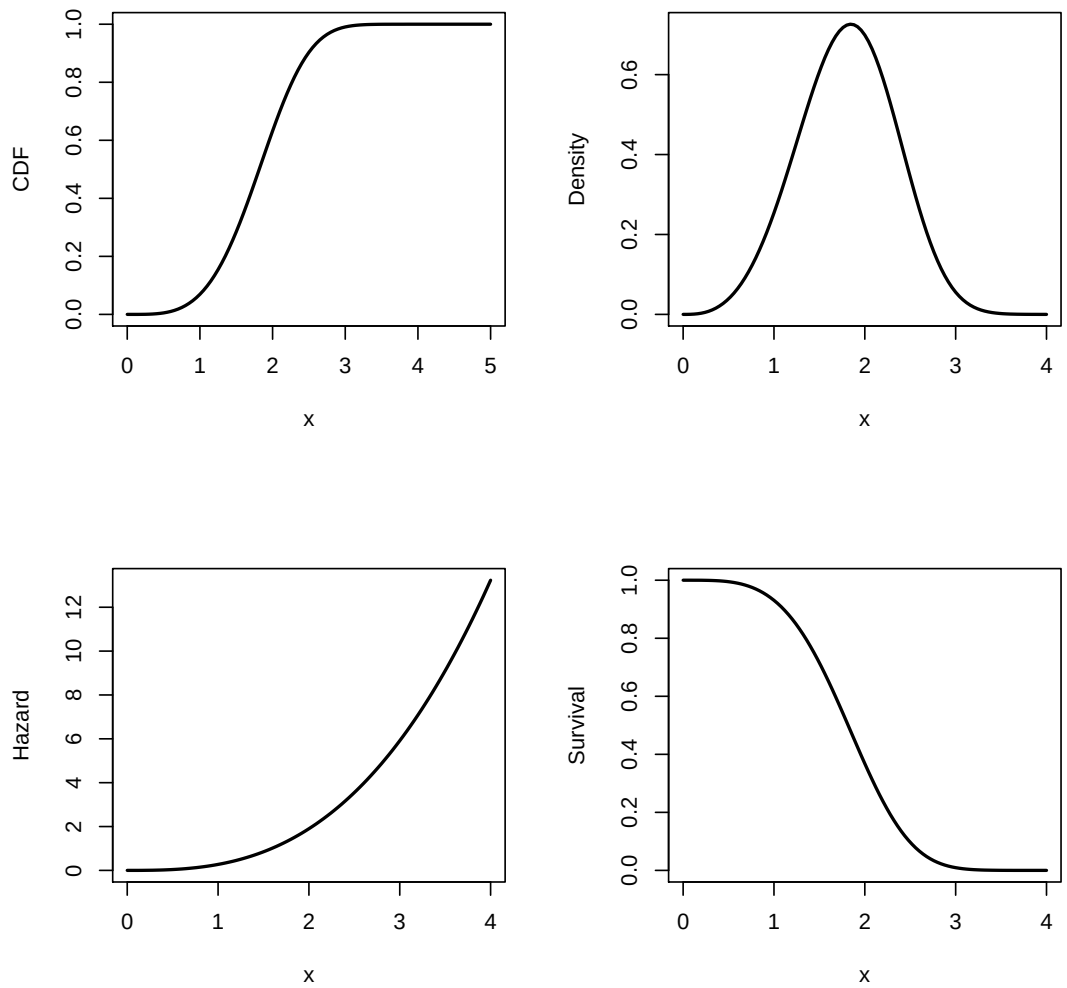
Plot of CDF/density/hazard/survival function

- Suppose that T follows Weibull distribution with the shape and scale parameters $\alpha = 3.8$ and $\lambda = 2$. Then the mean of T is $\lambda\Gamma(1 + 1/\alpha) = 2\Gamma(1 + 1/3.8) = 1.8075$.

Code

```
par(mfrow=c(2, 2))
curve(pweibull(x, 3.8, scale=2), from=0, to=5, lwd=2, ylab="CDF")
# plot CDF
curve(dweibull(x, 3.8, scale=2), from=0, to=4, lwd=2, ylab="Density")
# plot of the density function
curve(dweibull(x, 3.8, scale=2)/(1-pweibull(x, 3.8, scale=2)),
from=0, to=4, lwd=2, ylab="Hazard") # plot hazard
curve(1-pweibull(x, 3.8, scale=2), from=0, to=4, lwd=2, ylab="Survival")
# plot of the survival function
```

Plot of different aspects of the Weibull distribution



Some basic quantities

- The hazard function $\lambda(t)$ is the instantaneous failure rate, or probability that a subject of age t experiences failure at the next instant. Mathematically,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\text{pr}(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

- The hazard is related with the density and survival function through $\lambda(t) = f(t)/S(t)$.
- Also, $\lambda(t) = -d\log\{S(t)\}/dt$.
- Another related quantity is cumulative hazard, $\Lambda(t) = \int_0^t \lambda(u)du$. Some people may use h and H to denote hazard and cumulative hazard functions.
- Note that $S(t) = \exp\{-\Lambda(t)\}$. Thus, knowing one of hazard, cumulative hazard, density, and survival function, is equivalent to knowing other three.

Some basic relations in mathematical terms

Suppose that T is an absolutely continuous positive valued random variable with the density function f, CDF F, survival function S, hazard function λ, and the cumulative hazard function Λ. Then the following relations hold.

- $F(t) = \int_0^t f(u)du$
- $S(t) = 1 - F(t) = \int_t^\infty f(u)du$
- $\Lambda(t) = \int_0^t \lambda(u)du$
- $S(t) = \exp\{-\Lambda(t)\}$
- $\lambda(t) = -d\log\{S(t)\}/dt$
- $\lambda(t) = f(t)/S(t)$

A simple example

- Suppose that the hazard function of T is $\lambda(t) = \lambda_0$ a constant. What is its survival function?
- The cumulative hazard is $\Lambda(t) = \int_0^t \lambda(u)du = \lambda_0 t$. So, the survival function is

$$S(t) = \exp\{-\Lambda(t)\} = \exp(-\lambda_0 t).$$

- If the random variable T has a constant hazard, we call it exponential random variable.

Important information regarding hazards

- The hazard function $\lambda(t)$ is a non-negative quantity.
- The cumulative hazard is $\Lambda(t)$ is a non-negative and non-decreasing function.

Code

```
curve(exp(-2*x), from=0, to=4, lwd=2) # plot of the survival
# function with lambda=2
curve(2*x, from=0, to=4, lwd=2) # plot of the cumulative hazard
#for lambda=2
# plot of the above two curves in the same figure
curve(exp(-2*x), from=0, to=4, lwd=2, ylim=c(0, 8), ylab="", lty=4)
par(new=T)
curve(2*x, from=0, to=4, lwd=2, ylim=c(0, 8), ylab="", axes=F)
```

Logrank Test

- This test is used to test the difference between two survival curves. This test is most suitable to detect a difference between groups when the risk (hazard) of the event in one group is consistently greater than the risk in the other group. The test may not detect a difference when survival curves cross, a likely scenario in medical sciences with a surgical intervention. Therefore, to get a clearer idea it is always recommended to plot the survival curves besides conducting the hypothesis test.
- Suppose that we have time-to-event data from two groups.
- D : total number of failures counting both datasets
- Ordered failure times of the combined data $v_1^* < \cdots < v_D^*$
- $d_{i,1}(d_{i,2})$: the number of failures in group 1 (group 2) at time v_i^*
- $d_i = d_{i,1} + d_{i,2}$: the total number of failures from both groups at time v_i^*
- $n_{i,1}(n_{i,2})$: the number of subjects at risk in group 1 (group 2) at time v_i^*
- $n_i = n_{i,1} + n_{i,2}$: the total number of subjects from both groups at risk at time v_i^*

Logrank Test

- $H_0 : \lambda_1(t) = \lambda_2(t)$ for all $t \leq \tau$
- $H_a : \lambda_1(t) \neq \lambda_2(t)$ for at least one t
- Consider the statistic

$$T = \sum_{i=1}^{D} n_{i,1} \left(\frac{d_{i,1}}{n_{i,1}} - \frac{d_i}{n_i} \right) = \sum_{i=1}^{D} \left(d_{i,1} - \frac{n_{i,1} d_i}{n_i} \right)$$

- If the null hypothesis is true, then, an estimator of the expected hazard rate in the 1st group under H_0 is the pooled sample estimator of the hazard rate d_i/n_i at time v_i^*. Using only data from the 1st group sample, the estimator of the hazard rate is $d_{i,1}/n_{i,1}$. If null hypothesis holds, then we would expect the difference between d_i/n_i and $d_{i,1}/n_{i,1}$ will be small for every i.

Logrank Test

$$\text{Var}(T) = \sum_{i=1}^{D} \frac{n_{i,1}}{n_i} \left(1 - \frac{n_{i,1}}{n_i}\right) \left(\frac{n_i - d_i}{n_i - 1}\right) d_i$$

- The test statistic is $T^2/\text{Var}(T)$, and under H_0 it follows the χ^2 distribution with degrees of freedom 1.

Log-rank test

- We wanted to test if the hazard of failure is the same for both genders.
- $H_0 : \lambda_1(t) = \lambda_2(t)$ for all $t \leq 1022$ versus $H_0 : \lambda_1(t) \neq \lambda_2(t)$ for at least one t

Code

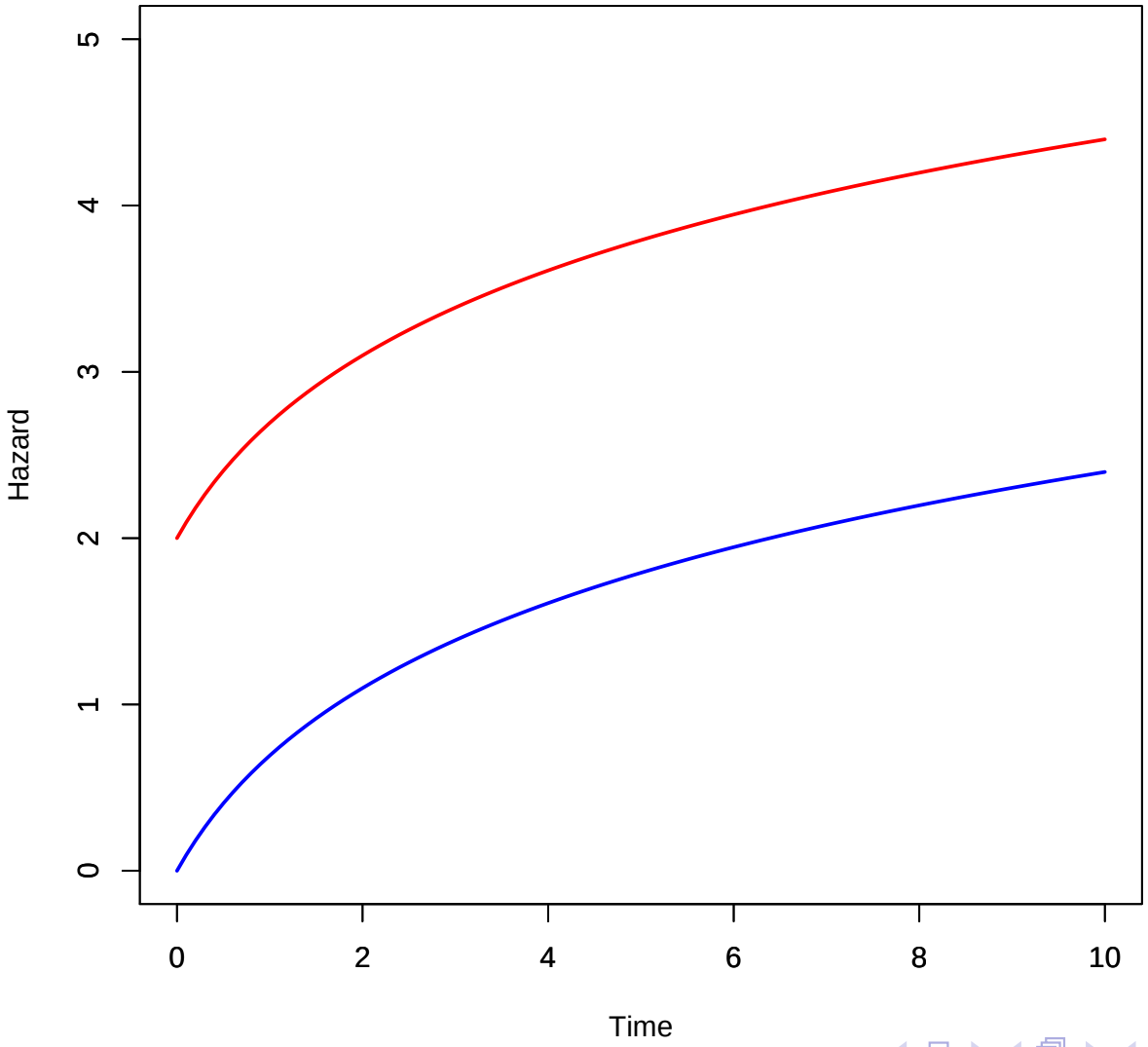
```
 mystatus=lung$status-1
 out <- survdiff(Surv(time, mystatus)~sex, data = lung)
 out
#Call:
#survdiff(formula = Surv(time, mystatus) ~ sex, data = lung)
#
#        N Observed Expected (O-E)^2/E (O-E)^2/V
#sex=1 138      112     91.6      4.55      10.3
#sex=2  90       53     73.4      5.68      10.3
#
#Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

When logrank test works best

Code

```
 myt=seq(0, 10, 0.1)
 lambda1=log(1+myt)
 lambda2=log(1+myt)+2
 # plot of two hazard functions
 pdf("class_note_fig1.pdf")
 plot(myt, lambda2, type="l", ylim=c(0, 5), ylab="Hazard",
 xlab="Time", lwd=2, col="red")
 par(new=T)
 plot(myt, lambda1, type="l", ylim=c(0, 5), ylab="",
 lwd=2, col="blue", xlab="")
dev.off()
```

Plot of two hazards that are in a constant difference

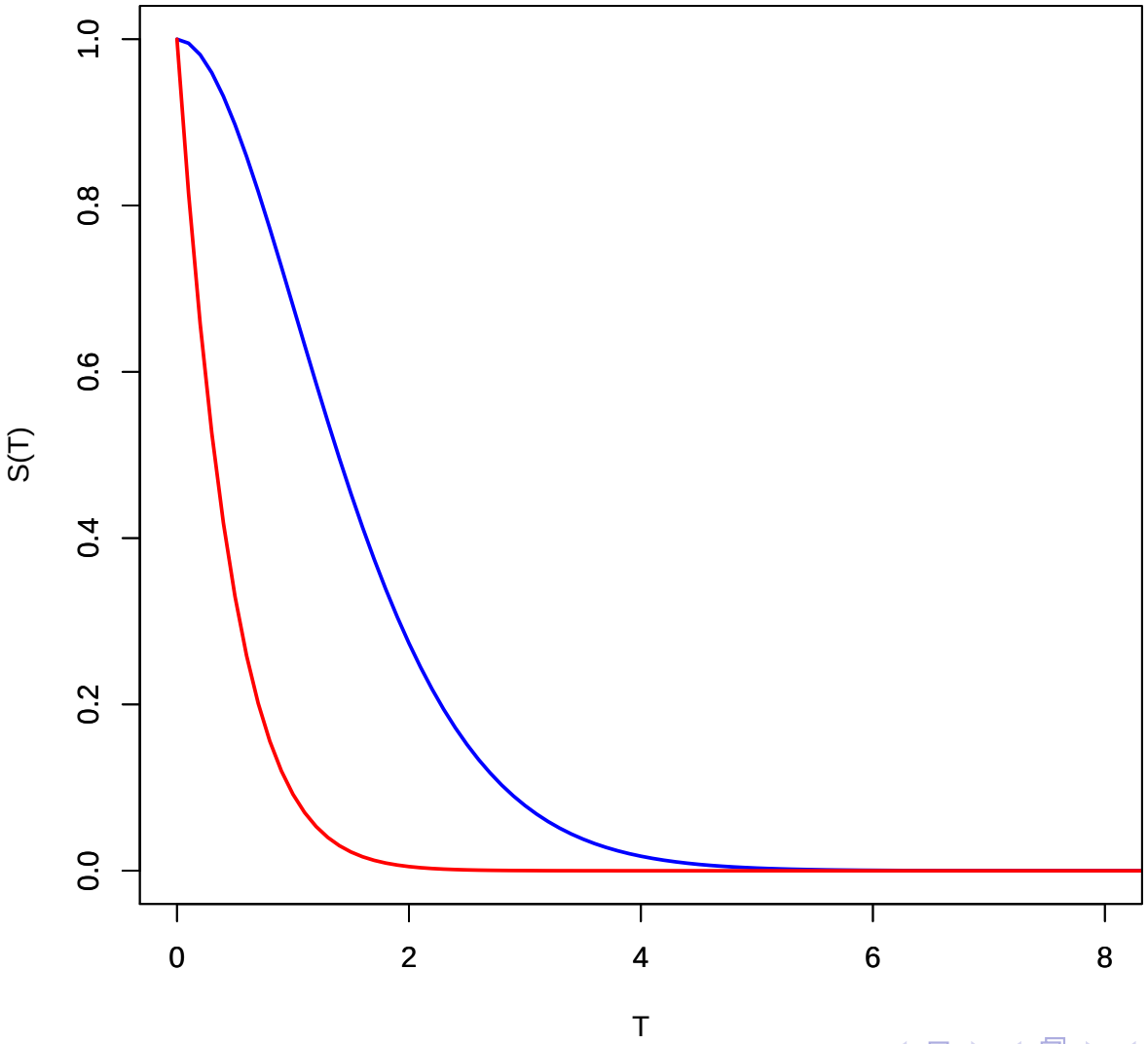


When logrank test works best

Code

```
 myt=seq(0, 10, 0.1)
 lambda1=log(1+myt)
 lambda2=log(1+myt)+2
 %lapply( integrate(function(x) log(1+x),  lower=0, upper=myt), myt)
 f1=function(t)exp(-integrate( function(x)log(1+x),
 lower=0, upper=t)$value )
 mysrv1= lapply(myt, f1)
f2=function(t)exp(-integrate( function(x)2+log(1+x), lower=0,
upper=t)$value )
 mysrv2= lapply(myt, f2)
pdf("class_note_fig2.pdf")
plot(myt, mysrv1, type="l", lwd=2, col="blue", xlim=c(0, 8),
ylab="S(T)", xlab="T")
par(new=T)
plot(myt, mysrv2, type="l", lwd=2, col="red", xlim=c(0, 8),
 ylab="", xlab="")
dev.off()
```

Plot of the two survival curves



Generalization of logrank Test

- Consider the statistic (that includes weights W_i)

$$T = \sum_{i=1}^{D} \sum_{i=1}^{D} W_i \left(d_{i,1} - \frac{n_{i,1} d_i}{n_i} \right)$$

- If $W_i = n_i$ (Gehan-Breslow), then the difference between the two hazards where more observations are available is given more importance than the time point with a fewer observations. However, this test statistic may yield a misleading results when the censoring patterns are different in the two groups.
- A similar test is Tarone-Ware's test where $W_i = \sqrt{n_i}$.

Generalization of logrank Test

- Following test is known as Fleming and Harrington test

$$T = \sum_{i=1}^{D} \sum_{i=1}^{D} W_i \left(d_{i,1} - \frac{n_{i,1} d_i}{n_i} \right)$$

- Here $W_i = \{\widehat{S}(v_{i-1}^*)\}^p \{1 - \widehat{S}(v_{i-1}^*)\}^q$, $p, q \geq 0$, and $\widehat{S}(t)$ denotes the Kaplan-Meier estimator of the survival function based on the combined data.
- When $p = q = 0$ we get the logrank test.

Generalization of logrank Test

When $q = 0$ and $p > 0$, these weights give the most weight to early departures between the hazard rates, whereas, when $p = 0$ and $q > 0$, these tests give most weight to departures which occur late in time. By an appropriate choice of p and q, one can construct tests which have the most power against alternatives which have the 2 hazard rates differing over any desired region.

Revisit the lung cancer data

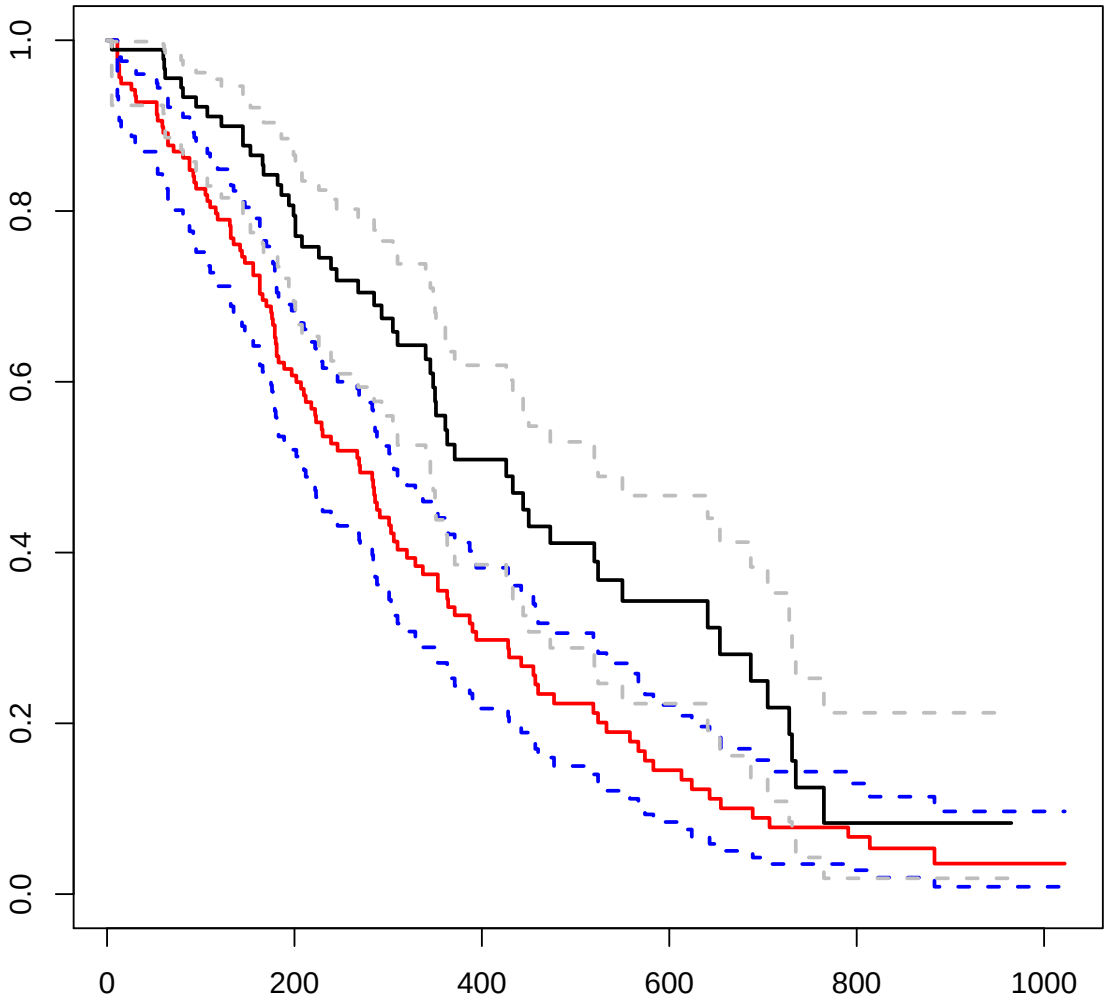
Code

```
library(survival)
 data(lung)
 head(lung)
 lungsex1=lung[lung$sex==1, ]
 lungsex2=lung[lung$sex==2, ]

 lungsex1$SurvObj <- with(lungsex1, Surv(time, status == 2))
 km.as.one <- survfit(SurvObj ~ 1, data = lungsex1,
 conf.type = "log-log")
 plot(km.as.one)

 lungsex2$SurvObj <- with(lungsex2, Surv(time, status == 2))
 km.as.two <- survfit(SurvObj ~ 1, data = lungsex2,
 conf.type = "log-log")
 plot(km.as.two)
```

Code

```
 # to obtain a nicer colored figure I use
pdf("STAT645_lung_two_plots_together.pdf")
plot(km.as.one, col=c("red", "blue", "blue"), lwd=2,
 ylim=c(0, 1), xlim=range(lung$time), ylab="")
 par(new=T)
 plot(km.as.two, col=c("black", "grey", "grey"), lwd=2,
ylim=c(0, 1), xlim=range(lung$time), axes=F)
dev.off()
```

Plot of the two survival curves for male and female



Nonparametric tests

Want to test $H_0 : \lambda_1(t) = \lambda_2(t)$ for all $t \leq 1022$ versus $H_0 : \lambda_1(t) \neq \lambda_2(t)$ for at least one t

Code

```
 mystatus=lung$status-1
 out <- survdiff(Surv(time, mystatus)~sex, data = lung, rho=0.0)
 out
#Call:
#survdiff(formula = Surv(time, mystatus) ~ sex, data = lung, rho = 0)
#
#        N Observed Expected (O-E)^2/E (O-E)^2/V
#sex=1 138      112     91.6      4.55      10.3
#sex=2  90       53     73.4      5.68      10.3
#
# Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

Code

```
out <- survdiff(Surv(time, mystatus)~sex, data = lung, rho=1)
out
#Call:
#survdiff(formula = Surv(time, mystatus) ~ sex, data = lung, rho = 1)
#
#        N Observed Expected (O-E)^2/E (O-E)^2/V
#sex=1 138     70.4     55.6      3.95      12.7
#sex=2  90     28.7     43.5      5.04      12.7
#
#Chisq= 12.7  on 1 degrees of freedom, p= 4e-04
```

Kidney disease data

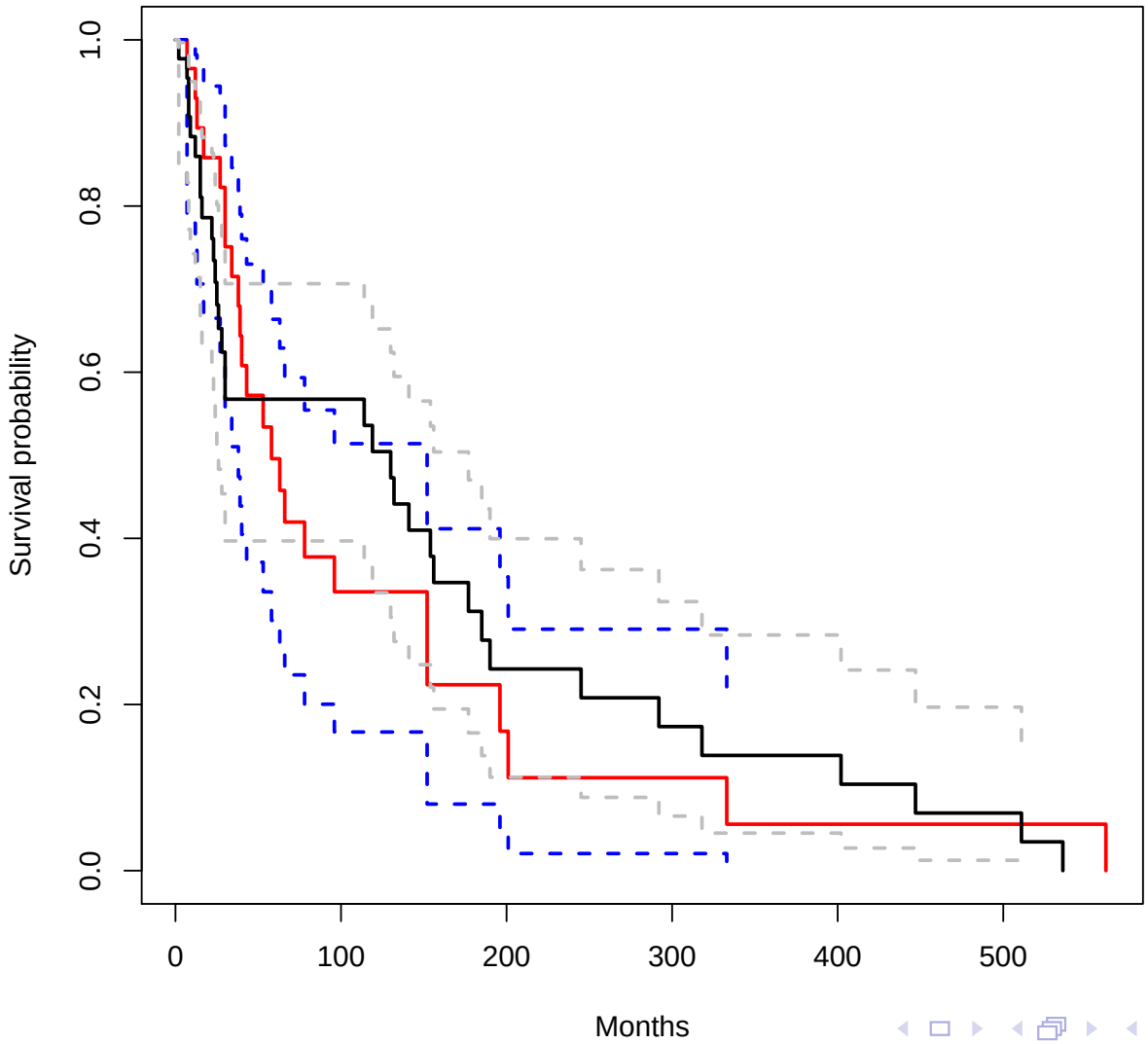
Code

```
library(survival)
 data(kidney)
 head(kidney)

 index=(1:nrow(kidney))[kidney$disease=="AN"| kidney$disease=="PKD"]
 # PKD: Polycystic kidney disease, AN: Analgesic nephropathy
 kd1=kidney[index,  ]
 kd2=kidney[-index, ]
 myd=rep(0, nrow(kidney))
 myd[index]=1
```

Code

```
 kd1$SurvObj <- with(kd1, Surv(time, status == 1))
 km.as.one <- survfit(SurvObj ~ 1, data = kd1, conf.type = "log-log")
 plot(km.as.one)

 kd2$SurvObj <- with(kd2, Surv(time, status == 1))
 km.as.two <- survfit(SurvObj ~ 1, data = kd2, conf.type = "log-log")
 plot(km.as.two)
pdf("STAT645_kidney_surv_plot_two_groups.pdf")
 # to obtain a nicer colored figure I use
 plot(km.as.one, col=c("red", "blue", "blue"), lwd=2, ylim=c(0, 1),
 xlim=range(kidney$time), ylab="")
 par(new=T)
 plot(km.as.two, col=c("black", "grey", "grey"), lwd=2, ylim=c(0, 1),
 xlim=range(kidney$time), axes=F, xlab="Months", ylab="Survival probability
dev.off()
```

Plot of the two survival curves PKD or AN and the other categories



Comment on the kidney disease data

The figure shows that there is not much of difference between the two survival functions at early time, however, there are some difference in the middle and and at a later time they cross each other. The log-rank test failed to find any significant difference between the two survival function (in terms of two hazards). The Fleming and Harrington tests, with $q > 0$, put more weight on the later time and compare the survival curves. Although there are some evidence of difference, they survival functions cross each other leading to barely significant p-value at the 5% level.

Different tests

Code

```
library(survMisc)
out<- survfit(Surv(time, status)~myd, data = kidney)
comp(ten(out), p=0, q=0)
comp(ten(out), p=3, q=3)
```