Survival analysis: Kaplan Meier Curves

Outline for today's class

- Motivating question(s) for survival analysis
- Terms associated with survival analysis
- Survival and hazard functions
- Kaplan-Meier curves key concepts
- Kaplan-Meier curves plotting
- Testing for statistical differences between survival curves

Who lives longer (or alternatively who fails faster)?

- Survival analysis is used to determine differences in time to an event (such as death) in two or more groups
- If you want to know if there is a difference in survival in patients randomized to drug A vs. B., how would you do that, what do you need to consider?

Theory–the basics

- What we estimate in survival analyses is differences between groups in the event rate.
- An **event** is a disease, death, relapse or anything else that can happen to someone. An event is also sometimes referred to as a **failure**.
- **Survival time** is the time variable.
- Survival time is measured from the time the person is followed (e.g. start of a study or t=0) for an event until they have the event, are **censored** due to some other event (e.g. death or loss to follow-up), or the observation period ends (e.g. end of the study, end of data collection round, etc.)

Censoring

- We often can't just measure survival as a case-fatality incidence rate especially for chronic conditions because of **censoring**.
- Survival analyses uses special techniques to account for censoring, which almost always happens in human studies.
- **Censoring** occurs when a person:

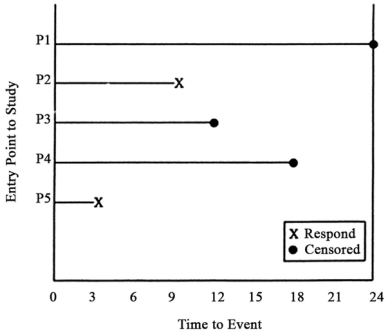


Figure 1:

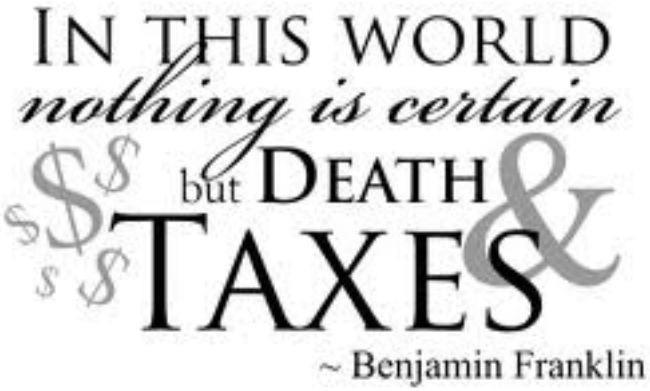


Figure 2:

  - does not experience the event of interest during the study (over the observation period)
  - is lost to follow-up
  - withdraws from the study
- Most survival time is **right censored** as in the figure on the previous slide because we don't have information on the full survival time. We use the observed survival time in survival analysis.
- **Left censoring** occurs when we don't know the lower bound of survival time. For example if our outcome is HIV positivity as measured by an HIV test, we don't actually know the survival time to the event (HIV exposure/infection) because it occurred sometime prior to our test measurement so the survival time we measure is greater than the actual survival time. We will not consider left censoring further here.

Survival Analysis Notation (from Kleinbaum and Klein Survival Analysis textbook, 2005)

- **T** is a random variable indicating a person's survival time
- **t** any value of **T**
- δ is a random variable indicating that a person in the dataset had the event (usually coded as 1) or was censored (usually coded as 0). Sometimes this is referred to as the person's status (had the event or did not have the event)

Survivor function

- **S(t)** is used to denote the survivor function. It gives the probability that a person's survival time (**T**) exceeds time **t**

- Ranges from 0 to ∞

Hazard function h(t)

- **h(t)** (also known as the hazard rate or λ) is used to denote the hazard function. It is related to survival and can be derived from S(t) (we will not cover this here)
- The hazard is a rate and not a probability like the survivor function.

- Ranges from 0 to ∞

- Kleinbaum says that "This mathematical formula is difficult to explain in practical terms."
- The numerator gives the probability that a person will have an event during t+dt given they have survived up to time t
- Gives the instantaneous potential for an event to occur *given* that an individual has survived to time t
- Focuses on failing rather than surviving and is sometimes called the **conditional failure rate**
- The units are probability per unit time making it a rate rather than a probability and difficult to interpret outright (ranges from 0 to ∞)

Different hazard models (Kleinbaum)

- Constant hazard (exponential model): h(t) is stable for healthy people. h(t)=λ no matter what the value of t is.
- Increasing Weibull: h(t) increases with time (cancer with low survival)
- Decreasing Weibull: h(t) decreases with time (persons recovering from surgery)
- Increasing and decreasing lognormal: h(t) increases and then decreases (TB patients)

courtesy of Kleinbaum courtesy of Kleinbaum

What are KM curves?

- Used to look at differences in survival between two groups (such as drug treatment or placebo in a clinical trial)
- Uses longitudinal data
- y-axis = percent surviving
- x-axis = time

KM Survival Estimate formula

Data layout

general data layout needed for software to conduct data analysis

Example

- Let's do a KM curve in excel, look at some descriptive statistics, and then do it in R
- Descriptives: average hazard and median survival time

```
#install.packages("survival") #for survival analysis by group
#install.packages('ggfortify') #for survival analysis by group
#install.packages("survminer") #for pairwise diffs
library(survminer)#for pairwise diffs
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

```
## Loading required package: magrittr
```

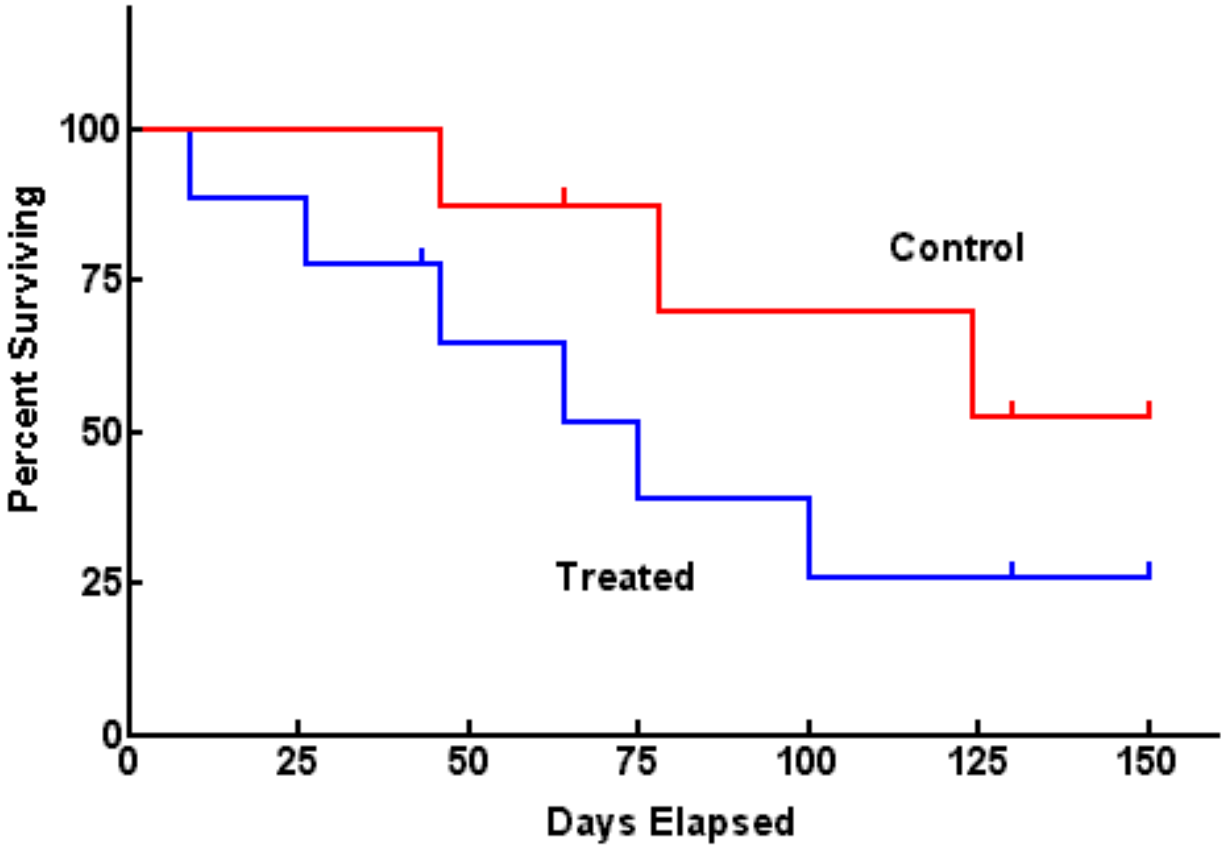

Percent Surviving
100
75
50
25
0
0
25
50
75
100
125
150
Days Elapsed
Control
Treated

Figure 3:

```
library(survival) #for calculating KM values
library(ggfortify) #for KM curves
library(readxl) # for reading in excel file
library(ggplot2) # for plotting KM curve
library(tidyverse) # for various packages
```

```
## -- Attaching packages --------------------------------------------------------- tidyverse 1.2.1 --
```

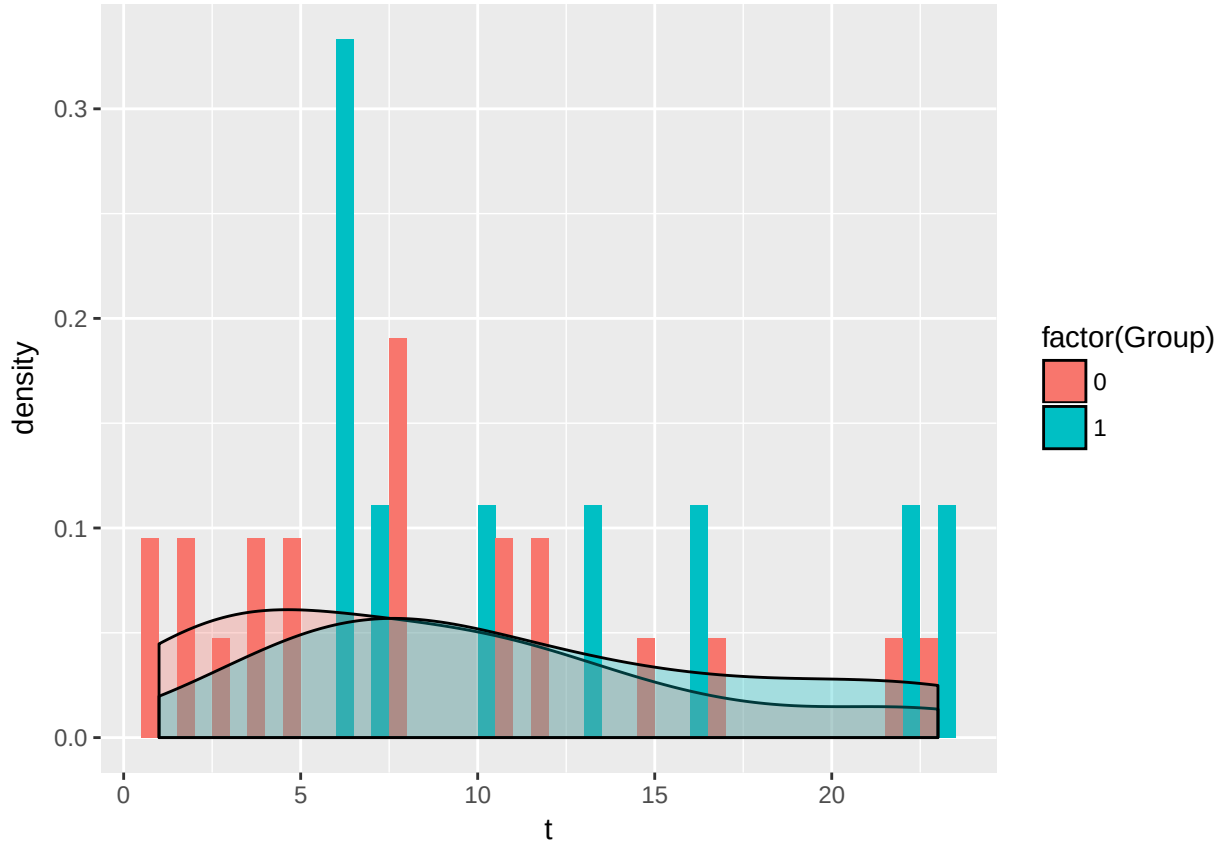
```
## v tibble  1.4.2     v purrr   0.2.4
## v tidyr   0.8.0     v dplyr   0.7.4
## v readr   1.1.1     v stringr 1.3.0
## v tibble  1.4.2     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
```

```
data <- read_excel("D:/Dropbox/@2018 Spring/ADA/Class 9 Kaplan Meier Curves/Surv_data_class9b.xlsx", sh
```

Let's look at survival time without censored events. Do we see any differences?

```
data2<-data[which(data$d==1),]#delete censored events
ggplot(data2, aes(t, fill=factor(Group))) + geom_histogram(aes(y=..density..), binwidth=1, bins=30, pos
```

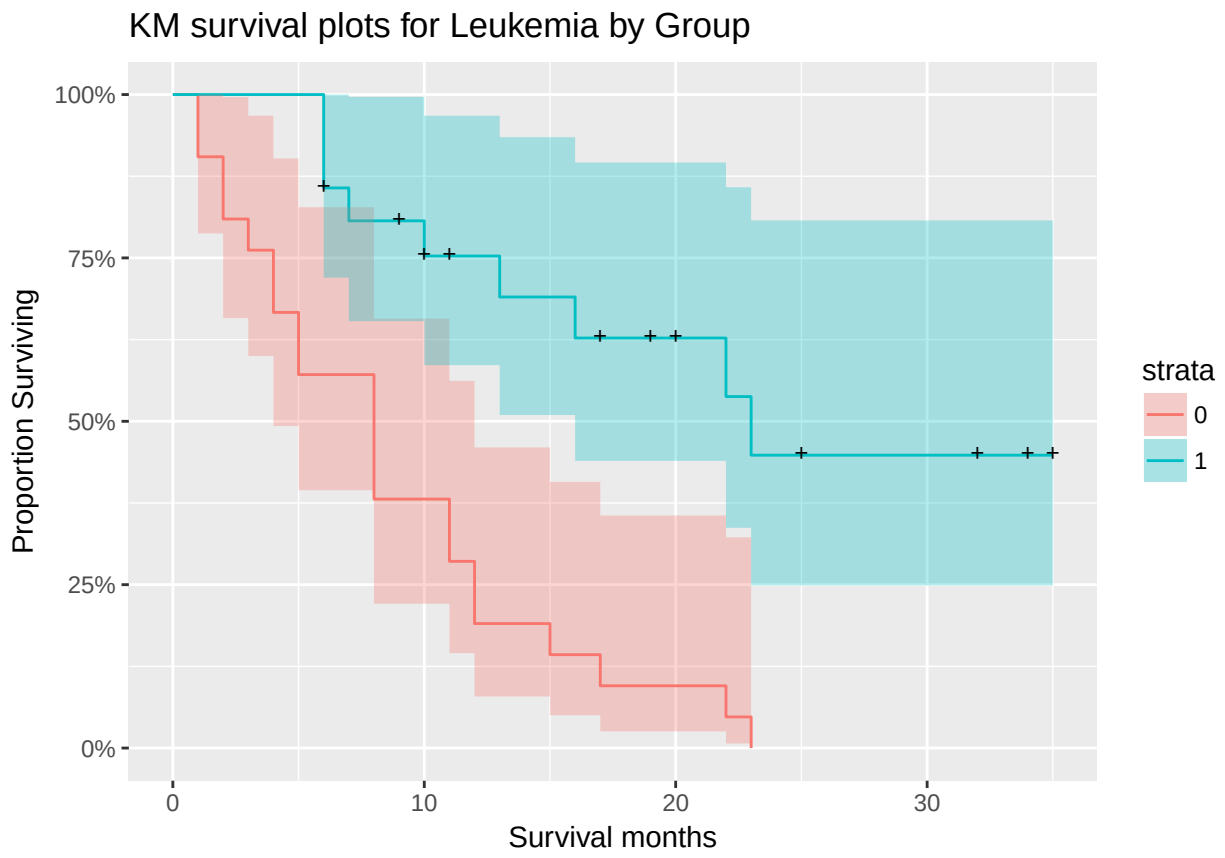


##Let's do a KM curve in excel and then code it in R

```r
leukemia.surv <- survfit(Surv(t, d) ~ Group, data) #calculates KM survivor function values for plotting
summary(leukemia.surv) #get KM values
```

```
## Call: survfit(formula = Surv(t, d) ~ Group, data = data)
##
##                 Group=0
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     21       2   0.9048  0.0641      0.78754        1.000
##     2     19       2   0.8095  0.0857      0.65785        0.996
##     3     17       1   0.7619  0.0929      0.59988        0.968
##     4     16       2   0.6667  0.1029      0.49268        0.902
##     5     14       2   0.5714  0.1080      0.39455        0.828
##     8     12       4   0.3810  0.1060      0.22085        0.657
##    11      8       2   0.2857  0.0986      0.14529        0.562
##    12      6       2   0.1905  0.0857      0.07887        0.460
##    15      4       1   0.1429  0.0764      0.05011        0.407
##    17      3       1   0.0952  0.0641      0.02549        0.356
##    22      2       1   0.0476  0.0465      0.00703        0.322
##    23      1       1   0.0000     NaN           NA           NA
##
##                 Group=1
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     6     21       3    0.857  0.0764        0.720        1.000
##     7     17       1    0.807  0.0869        0.653        0.996
##    10     15       1    0.753  0.0963        0.586        0.968
##    13     12       1    0.690  0.1068        0.510        0.935
##    16     11       1    0.627  0.1141        0.439        0.896
##    22      7       1    0.538  0.1282        0.337        0.858
##    23      6       1    0.448  0.1346        0.249        0.807
```

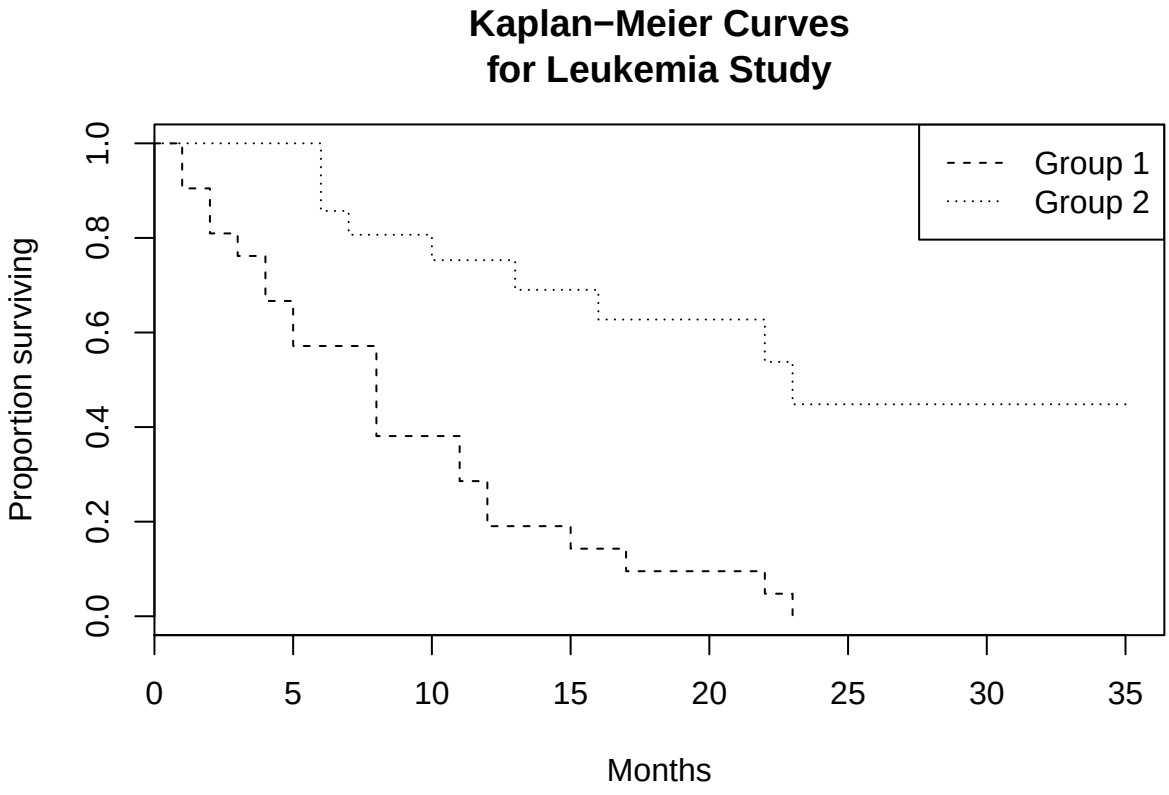
```r
autoplot(leukemia.surv) + labs(x="Survival months", y="Proportion Surviving", title="KM survival plots
```



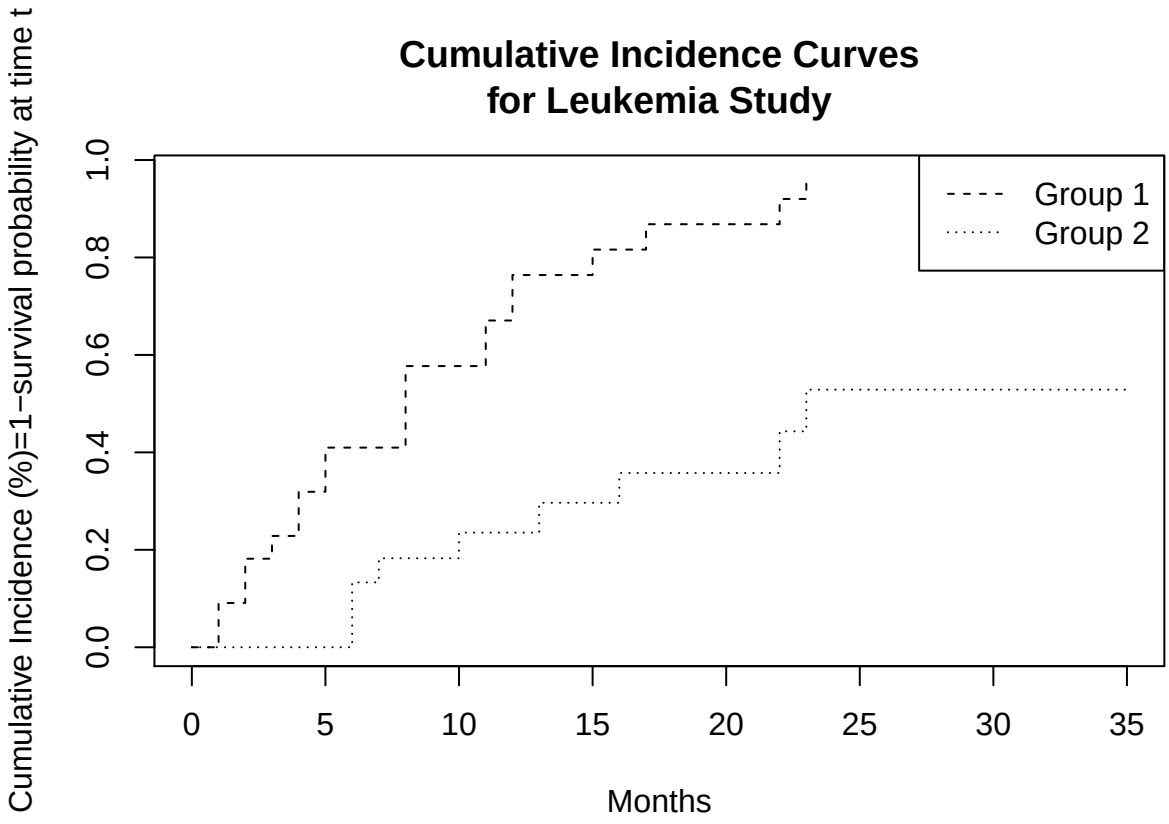

```
leukemia.surv #Median survival
```

```
## Call: survfit(formula = Surv(t, d) ~ Group, data = data)
##
##          n events median 0.95LCL 0.95UCL
## Group=0 21     21      8       4      12
## Group=1 21      9     23      16      NA
```

```
plot(leukemia.surv, lty = 2:3, xlab="Months", ylab="Proportion surviving") #using base plot to calculat
legend("topright", c("Group 1", "Group 2"), lty = 2:3)
title("Kaplan-Meier Curves\nfor Leukemia Study")
```




```
lsurv2 <- survfit(Surv(t, d) ~ Group, data, type='fleming')
plot(lsurv2, lty=2:3, fun="event",
    xlab="Months", ylab="Cumulative Incidence (%)=1-survival probability at time t") #plot %failing at
legend("topright", c("Group 1", "Group 2"), lty = 2:3)
title("Cumulative Incidence Curves\nfor Leukemia Study")
```



Testing statistical differences between survival curves-The log rank test (most common)-Go to excel sheet

- Used to determined statistical differences between two survival curves
- Large sample chi-square test
- Approximate and exact formulas *exact approximate*)

Calculate exact log rank in R

```
survdiff(Surv(t, d) ~ Group,data=data)
```

```
## Call:
## survdiff(formula = Surv(t, d) ~ Group, data = data)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## Group=0 21       21     10.7      9.77      16.8
## Group=1 21        9     19.3      5.46      16.8
##
##  Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

Code for pairwise differences (for when you have more than 2 groups)

```
pairwise_survdiff(Surv(t, d) ~ Group,data=data)
```

```
##
##  Pairwise comparisons using Log-Rank test
##
## data:  data and Group
##
##   0
## 1 4.2e-05
##
## P value adjustment method: BH
```

Other tests (decribed but not covered in detail here)

- **Wilcoxon test** (called Breslow in SPSS), **Tarone-Ware test**, **Flemington-Harrington test**
  - places different weights on failures
  - For example, Wilcoxon and Tarone-Ware place more weight on early failures where the number at risk is larger
- **Stratified log rank test**- provides a way to test for differences between survival curves while controlling for another variable