432 Class 16

https://thomaselove.github.io/432-2025

2025-03-06

## Today’s Agenda

* Testing the difference between two survival curves
  + log rank tests (Mantel-Haenszel)
  + Peto-Peto modification of the Gehan-Wilcoxon test
* Customizing a Kaplan-Meier plot
  + Plotting Cumulative Event Rates and Cumulative Hazards
* A larger example: Colon Cancer Trial
  + Survival Objects and Kaplan-Meier Estimates
  + Fitting a simple Cox Regression model

## Today’s R Setup

knitr::opts\_chunk$set(comment = NA)  
  
library(janitor)  
library(naniar)  
library(conflicted)  
library(here)  
library(survival)   
library(survminer)   
library(easystats)  
library(tidyverse)  
  
theme\_set(theme\_bw())   
conflicts\_prefer(dplyr::filter)

# Our Original Example (from Class 15)

## The aml\_432 data

The aml data provided in the **survival** package describe survival times (in months) for 23 subjects with acute myelogenous leukemia.

Question: Should the standard course of chemotherapy be extended (maintained) for additional cycles or not?

aml\_432 <- survival::aml |> tibble() |>  
 rename(maintain = x, censored = status, months = time) |>  
 arrange(desc(months)) |>  
 mutate(subject = as.character(row\_number()),  
 death = 1 - censored,  
 maintain =   
 fct\_recode(maintain, "Main" = "Maintained",   
 "No" = "Nonmaintained")) |>  
 relocate(subject)

## The aml\_432 codebook

| aml\_432 | Description |
| --- | --- |
| subject | subject identifying code |
| months | known months of survival in follow-up period |
| censored | 1 if subject’s follow-up was censored, else 0 |
| maintain | Chemotherapy was Maintained (Main or No) |
| death | 1-censored, so 1 if died, 0 if censored. |

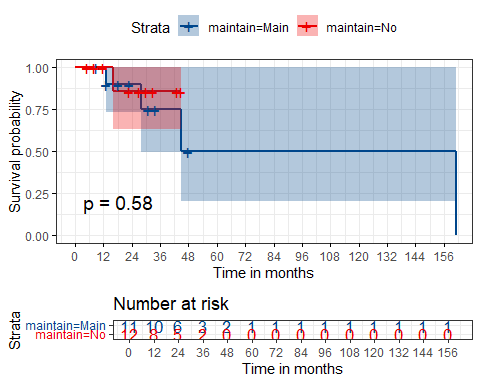
## Survival Object and K-M Estimator by Maintenance Group

surv\_aml <- Surv(time = aml\_432$months, event = aml\_432$death)  
  
km\_aml\_grp <- survfit(surv\_aml ~ aml\_432$maintain)  
print(km\_aml\_grp, print.rmean = TRUE)

Call: survfit(formula = surv\_aml ~ aml\_432$maintain)  
  
 n events rmean\* se(rmean) median 0.95LCL 0.95UCL  
aml\_432$maintain=Main 11 4 97.2 28.4 103 28 NA  
aml\_432$maintain=No 12 1 140.3 19.2 NA NA NA  
 \* restricted mean with upper limit = 161

## Kaplan-Meier Plot, by Group

ggsurvplot(km\_aml\_grp, data = aml\_432,  
 ggtheme = theme\_bw(), palette = "lancet",  
 conf.int = TRUE,  
 xlab = "Time in months",  
 break.time.by = 12,  
 risk.table = TRUE,  
 risk.table.height = 0.25,   
 risk.table.col = "strata",  
 pval = TRUE)



## Testing the difference between 2 survival curves

To obtain a significance test comparing these two survival curves, we turn to a log rank test, which tests the null hypothesis for all where the two exposures have survival functions and .

survdiff(surv\_aml ~ aml\_432$maintain)

Call:  
survdiff(formula = surv\_aml ~ aml\_432$maintain)  
  
 N Observed Expected (O-E)^2/E (O-E)^2/V  
aml\_432$maintain=Main 11 4 3.47 0.0804 0.304  
aml\_432$maintain=No 12 1 1.53 0.1827 0.304  
  
 Chisq= 0.3 on 1 degrees of freedom, p= 0.6

## Alternative log rank tests

An alternative is the *Peto and Peto modification of the Gehan-Wilcoxon test*, which results from adding rho=1 to the survdiff function (rho=0, the default, yields the log rank test.)

survdiff(surv\_aml ~ aml\_432$maintain, rho = 1)

Call:  
survdiff(formula = surv\_aml ~ aml\_432$maintain, rho = 1)  
  
 N Observed Expected (O-E)^2/E (O-E)^2/V  
aml\_432$maintain=Main 11 3.262 2.8 0.0751 0.275  
aml\_432$maintain=No 12 0.941 1.4 0.1504 0.275  
  
 Chisq= 0.3 on 1 degrees of freedom, p= 0.6

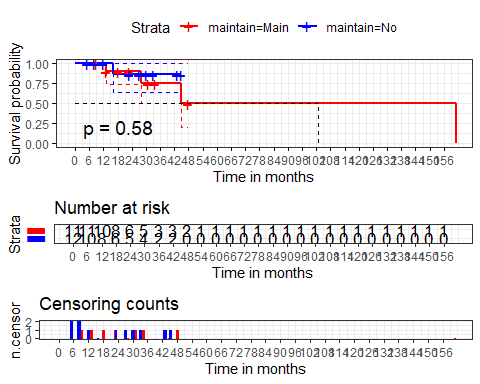
## Alternative log rank tests

* As compared to the log rank test, this Peto-Peto modification (and others using rho > 0) give greater weight to the left hand (earlier) side of the survival curves.
* To obtain chi-square tests that give greater weight to the right hand (later) side of the survival curves than the log rank test, use rho < 0.

The log rank test generalizes to permit survival comparisons across more than two groups.

## A Highly Customized K-M Plot

ggsurvplot(km\_aml\_grp,   
 data = aml\_432,   
 palette = c("red", "blue"),  
 risk.table = TRUE,   
 pval = TRUE,   
 conf.int = TRUE,   
 xlab = "Time in months",   
 break.time.by = 6,   
 ggtheme = theme\_bw(),  
 risk.table.y.text.col = T,  
 risk.table.height = 0.25,   
 risk.table.y.text = FALSE,  
 ncensor.plot = TRUE,  
 ncensor.plot.title = "Censoring counts",  
 ncensor.plot.height = 0.25,  
 conf.int.style = "step",  
 surv.median.line = "hv")



## Customizing the K-M Plot Further

See <https://rpkgs.datanovia.com/survminer/> or <https://github.com/kassambara/survminer/> for many more options.

Also, consider [this YouTube Video from Frank Harrell](https://www.youtube.com/watch?v=EoIB_Obddrk) entitled “[Survival Curves: Showing More by Showing Less](https://www.youtube.com/watch?v=EoIB_Obddrk)” which highlights the value of interactive approaches.

# A Larger Example

## Colon Cancer Trial

The colon data set from the **survival** package shows data from one of the first successful trials of adjuvant chemotherapy for colon cancer. While the data includes two records per person, one for recurrence and one for death, we’ve filtered to look only at recurrence.

col\_432 <- tibble(survival::colon) |>  
 filter(etype == 1) |>  
 select(id, rx, time, status, age, node4, surg, extent) |>  
 mutate(extent = factor(extent), id = as.character(id))  
  
dim(col\_432)

[1] 929 8

col\_432 |> n\_miss()

[1] 0

## Variables we will use today

| Variable | Description |
| --- | --- |
| id | subject identifying code |
| rx | Treatment, either Obs(ervation), Lev(amisole) or Lev(amisole)+5-FU |
| time | days until recurrence |
| status | censoring status (1 = censored, 0 = recurrence) |

Levamisole is a low-toxicity compound previously used to treat worm infestations in animals; 5-FU is a moderately toxic (as these things go) chemotherapy agent.

## Variables we won’t use today

| Variable | Description |
| --- | --- |
| age | age in years |
| node4 | 1 if more than 4 positive lymph nodes, else 0 |
| surg | time from surgery to registration (0=short, 1=long) |
| extent | Extent of local spread (1=submucosa, 2=muscle, 3=serosa, 4=contiguous structures) |

* The study is originally described in [Laurie (1989)](https://pubmed.ncbi.nlm.nih.gov/2778478/). The main report is found in [Moertel (1990)](https://pubmed.ncbi.nlm.nih.gov/2300087/).

## col\_432 Survival object, and data

Create Survival Object, S

col\_432$S <- Surv(time = col\_432$time, event = (col\_432$status == 0))  
  
col\_432

# A tibble: 929 × 9  
 id rx time status age node4 surg extent S  
 <chr> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <Surv>  
 1 1 Lev+5FU 968 1 43 1 0 3 968+  
 2 2 Lev+5FU 3087 0 63 0 0 3 3087   
 3 3 Obs 542 1 71 1 0 2 542+  
 4 4 Lev+5FU 245 1 66 1 1 3 245+  
 5 5 Obs 523 1 69 1 1 3 523+  
 6 6 Lev+5FU 904 1 57 1 0 3 904+  
 7 7 Lev 229 1 77 1 1 3 229+  
 8 8 Obs 3192 0 54 0 0 3 3192   
 9 9 Lev 3173 0 46 0 0 3 3173   
10 10 Lev+5FU 3308 0 68 0 1 3 3308   
# ℹ 919 more rows

## Kaplan-Meier Survival Estimates

Days of Survival, by Treatment Group

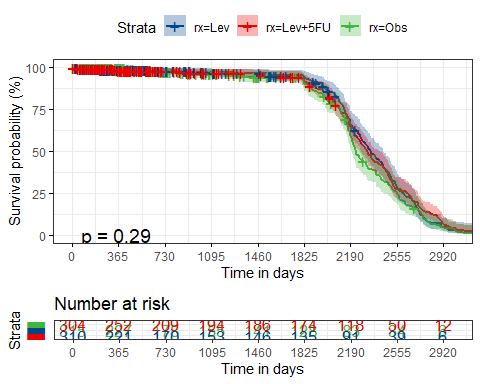
km\_col\_rx <- survfit(col\_432$S ~ col\_432$rx)  
  
print(km\_col\_rx, print.rmean = TRUE)

Call: survfit(formula = col\_432$S ~ col\_432$rx)  
  
 n events rmean\* se(rmean) median 0.95LCL 0.95UCL  
col\_432$rx=Obs 315 138 2257 36.4 2232 2200 2324  
col\_432$rx=Lev 310 138 2331 35.2 2349 2274 2449  
col\_432$rx=Lev+5FU 304 185 2320 33.2 2318 2255 2422  
 \* restricted mean with upper limit = 3329

## Kaplan-Meier Plot, by Treatment

* There is a problem with this plot.
* See if you can spot it (on the next slide.)

ggsurvplot(km\_col\_rx, data = col\_432,  
 ggtheme = theme\_bw(), palette = "lancet",  
 conf.int = TRUE,  
 fun = "pct", # plots survival probability as percentage  
 xlab = "Time in days",  
 break.time.by = 365,  
 risk.table = TRUE,  
 risk.table.height = 0.25,   
 risk.table.col = "strata",  
 risk.table.y.text = FALSE,  
 risk.table.y.text.col = TRUE,  
 pval = TRUE)



## Comparing three survival curves

survdiff(col\_432$S ~ col\_432$rx)

Call:  
survdiff(formula = col\_432$S ~ col\_432$rx)  
  
 N Observed Expected (O-E)^2/E (O-E)^2/V  
col\_432$rx=Obs 315 138 123 1.797 2.479  
col\_432$rx=Lev 310 138 144 0.247 0.364  
col\_432$rx=Lev+5FU 304 185 194 0.410 0.717  
  
 Chisq= 2.5 on 2 degrees of freedom, p= 0.3

This is the rho = 0 result: the Mantel-Haenszel test.

## Emphasizing left-hand-side of curves

survdiff(col\_432$S ~ col\_432$rx, rho = 1)

Call:  
survdiff(formula = col\_432$S ~ col\_432$rx, rho = 1)  
  
 N Observed Expected (O-E)^2/E (O-E)^2/V  
col\_432$rx=Obs 315 75.9 67.0 1.205 2.445  
col\_432$rx=Lev 310 68.0 73.9 0.466 0.999  
col\_432$rx=Lev+5FU 304 92.4 95.5 0.101 0.253  
  
 Chisq= 2.6 on 2 degrees of freedom, p= 0.3

This is the rho = 1 result: the Peto and Peto modification of the Gehan-Wilcoxon test.

## Emphasizing the right-hand-side

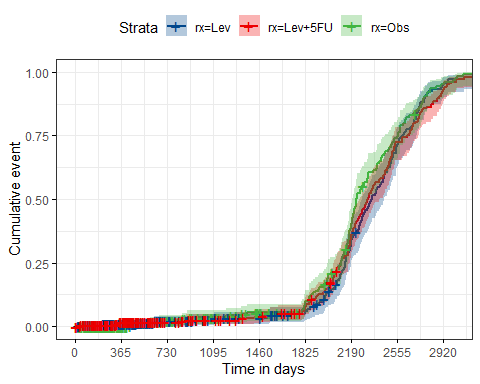
survdiff(col\_432$S ~ col\_432$rx, rho = -1)

Call:  
survdiff(formula = col\_432$S ~ col\_432$rx, rho = -1)  
  
 N Observed Expected (O-E)^2/E (O-E)^2/V  
col\_432$rx=Obs 315 578 465 27.9 2.416  
col\_432$rx=Lev 310 1221 1449 35.8 2.973  
col\_432$rx=Lev+5FU 304 1195 1082 12.0 0.693  
  
 Chisq= 4.5 on 2 degrees of freedom, p= 0.1

## Plotting the Cumulative Event Rate

Add fun = "event" to the plot…

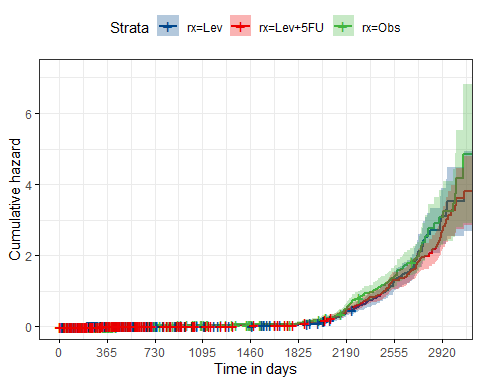
ggsurvplot(km\_col\_rx,   
 data = col\_432,   
 fun = "event",  
 ggtheme = theme\_bw(),   
 palette = "lancet",  
 xlab = "Time in days",  
 break.time.by = 365,  
 conf.int = TRUE)



## Cumulative Hazard Function

Add fun = "cumhaz" to the plot (f(y) = -log(y))

ggsurvplot(km\_col\_rx,   
 data = col\_432,   
 fun = "cumhaz",  
 ggtheme = theme\_bw(),   
 palette = "lancet",  
 xlab = "Time in days",  
 break.time.by = 365,  
 conf.int = TRUE)



## Cox Regression Model

This is a Cox proportional hazards regression model.

fit <- coxph( Surv(time, status == 0) ~ rx, data = col\_432)  
fit

Call:  
coxph(formula = Surv(time, status == 0) ~ rx, data = col\_432)  
  
 coef exp(coef) se(coef) z p  
rxLev -0.1587 0.8533 0.1212 -1.310 0.190  
rxLev+5FU -0.1635 0.8492 0.1130 -1.447 0.148  
  
Likelihood ratio test=2.44 on 2 df, p=0.2947  
n= 929, number of events= 461

* The key coefficients are the exponentiated ones, which describe relative hazards (hazard ratios.)

## Hazard Ratios and the Cox Model

A hazard ratio compares the rate at which our event (like death, or recurrence) occurs in one group compared to another group over time, essentially indicating the relative risk of the event happening in one group compared to another.

* A hazard ratio of 1 indicates no difference between the groups
* A hazard ratio > 1 indicates a higher risk in the named group as compared to the baseline.
* A hazard ratio < 1 indicates a lower risk in the named group as compared to the baseline.

## The fit model

* Our model fit looks at the impact of three treatments, with Observation as the baseline treatment, on days until recurrence.

exp(fit$coefficients)

rxLev rxLev+5FU   
0.8532688 0.8492052

## Interpreting our fit (1/2)

* Note that we’re looking at the exponentiated coefficients here in order to describe relative hazards (hazard ratios.)

model\_parameters(fit, exponentiate = TRUE, pretty\_names = FALSE,   
 ci = 0.90, digits = 3)

Parameter | Coefficient | SE | 90% CI | z | p  
-----------------------------------------------------------------  
rxLev | 0.853 | 0.103 | [0.699, 1.041] | -1.310 | 0.190  
rxLev+5FU | 0.849 | 0.096 | [0.705, 1.023] | -1.447 | 0.148

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald z-distribution approximation.

The fit model estimates the hazard ratio for rxLev to be 0.853, with 90% CI (0.699, 1.041) which implies that the hazard of recurrence for a subject receiving Levamisole is estimated to be 85.3% as large as the hazard of recurrence for a subject who received Observation.

## Interpreting our fit (2/2)

model\_parameters(fit, exponentiate = TRUE, pretty\_names = FALSE,   
 ci = 0.90, digits = 3)

Parameter | Coefficient | SE | 90% CI | z | p  
-----------------------------------------------------------------  
rxLev | 0.853 | 0.103 | [0.699, 1.041] | -1.310 | 0.190  
rxLev+5FU | 0.849 | 0.096 | [0.705, 1.023] | -1.447 | 0.148

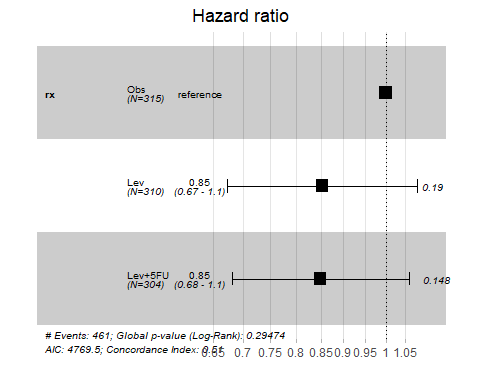
Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald z-distribution approximation.

The fit model estimates the hazard ratio for rxLev+5FU to be 0.849, with 90% CI (0.705, 1.023) which implies that the hazard of recurrence for a subject receiving Levamisole and 5-FU is estimated to be 84.9% as large as the hazard of recurrence for a subject who received Observation.

* A forest plot can be used to graph this result nicely. (See next slide.)

## Forest Plot for Cox model

ggforest(fit, data = col\_432)



## Model Performance for fit

model\_performance(fit)

Response residuals not available to calculate mean square error. (R)MSE  
 is probably not reliable.

Warning: Can't calculate weighted residuals from model.

# Indices of model performance  
  
AIC | AICc | BIC | Nagelkerke's R2 | RMSE | Sigma  
----------------------------------------------------------------  
4769.501 | 4769.514 | 4779.169 | 0.003 | 0.699 | 0.000

## More on Cox Models

Coming in Classes 22 and 23…

### Save our tibble to an R data set

write\_rds(col\_432, here("c16/data/col\_432.Rds"))

I’ve placed these data on [our 432 data page](https://github.com/THOMASELOVE/432-data).