SURVIVAL ANALYSIS
USING
KAPLAN-MEIER MODEL

HAMZA YUSUF ADAM

ABUBAKAR TAFAWA BALEWA UNIVERSITY, BAUCHI

IASC – AFRICAN MEMBER GROUP

Install and load required R package

❖ We'll use three R packages:

- *lubridate* for formatting dates and time
- *survival* for computing survival analyses
- *survminer* for summarizing and visualizing the results of survival analysis

❖ Install the packages

- install.packages(c("lubridate","survival", "survminer"))

❖ load the packages

- library(lubridate)
- library(survival)
- library(survminer)

Dealing with dates in R

- Data will often come with start and end dates rather than pre-calculated survival times. The first step is to make sure that these are formatted as dates in R

- Lets create small example with variables HIV_date and AIDS_date.

```
date_ex <-
    tibble(
      HIV_date = c("2015-04-20", "2010-02-10", "2017-10-30“,
                   “2018-12-04),
       AID_date = c("2016-04-15", "2018-07-04", "2019-10-31“,
                   “2022-01-01)
    )

date_ex
```

Create Date Output

A tibble: 4 x 2

	HIV_date	AIDS_date
	<chr>	<chr>
1	2015-04-20	2016-04-15
2	2010-02-10	2018-07-04
3	2017-10-30	2019-10-31
4	2018-12-04	2022-01-01

We observe that these are character variables and we need them to be as date. So, we format them.

We can use the package lubridate to format dates. Here, we use the ymd function

```
date_ex %>%
        mutate( HIV_date = ymd(HIV_date),
        AIDS_date = ymd(AIDS_date)
   )
```

Calculating survival times - lubridate

Using the lubridate package, the operator %--%, which is then converted to the number of elapsed seconds using as.duration and finally converted to years by dividing by dyears(1), which gives the seconds in a year. To give it in days or months, we divide by ddays(1) and dmonths(1) respectively.

```
date_ex %>%
   mutate(
    os_days =
     as.duration(HIV_date %--% AIDS_date) / ddays(1)
     )
```

Survival Time Output

- A tibble: 4 x 3

	HIV_date	AIDS_date	os_days
	<date>	<date>	<dbl>
1	2015-04-20	2016-04-15	361
2	2010-02-10	2018-07-04	3066
3	2017-10-30	2019-10-31	731
4	2018-12-04	2022-01-01	1124

Kaplan-Meier Survival Estimate

The Kaplan-Meier (KM) method is a non-parametric method used to estimate the survival probability from observed survival times (Kaplan and Meier, 1958).

This survival probability at time t_i, $S(t_i)$, is calculated as:

$$S(t_i) = S(t_{i-1})(1 - \frac{d_i}{n_i})$$

The estimated probability $S(t)$ is a step function that only that changes value at the time of each event.

Data set

❖We will use the lung cancer data available in the survival package

❖ The variables that we will use to demonstrate the method are:

- time: survival time in days
- status: censoring status and 1 = censored, 2 = dead
- sex: male =1, female = 2

head(cancer)

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

Estimating Survival Curves with Kaplan-Meier Method

❖ The Surv function from the survival package creates a survival object for use in a model formula

$$Surv(lung\$time,\ lung\$status)[1:20]$$

❖ the $survfit$ function creates survival curves based on a formula

$$f1 = survfit(Surv(time,\ status) \sim 1, data = lung)$$

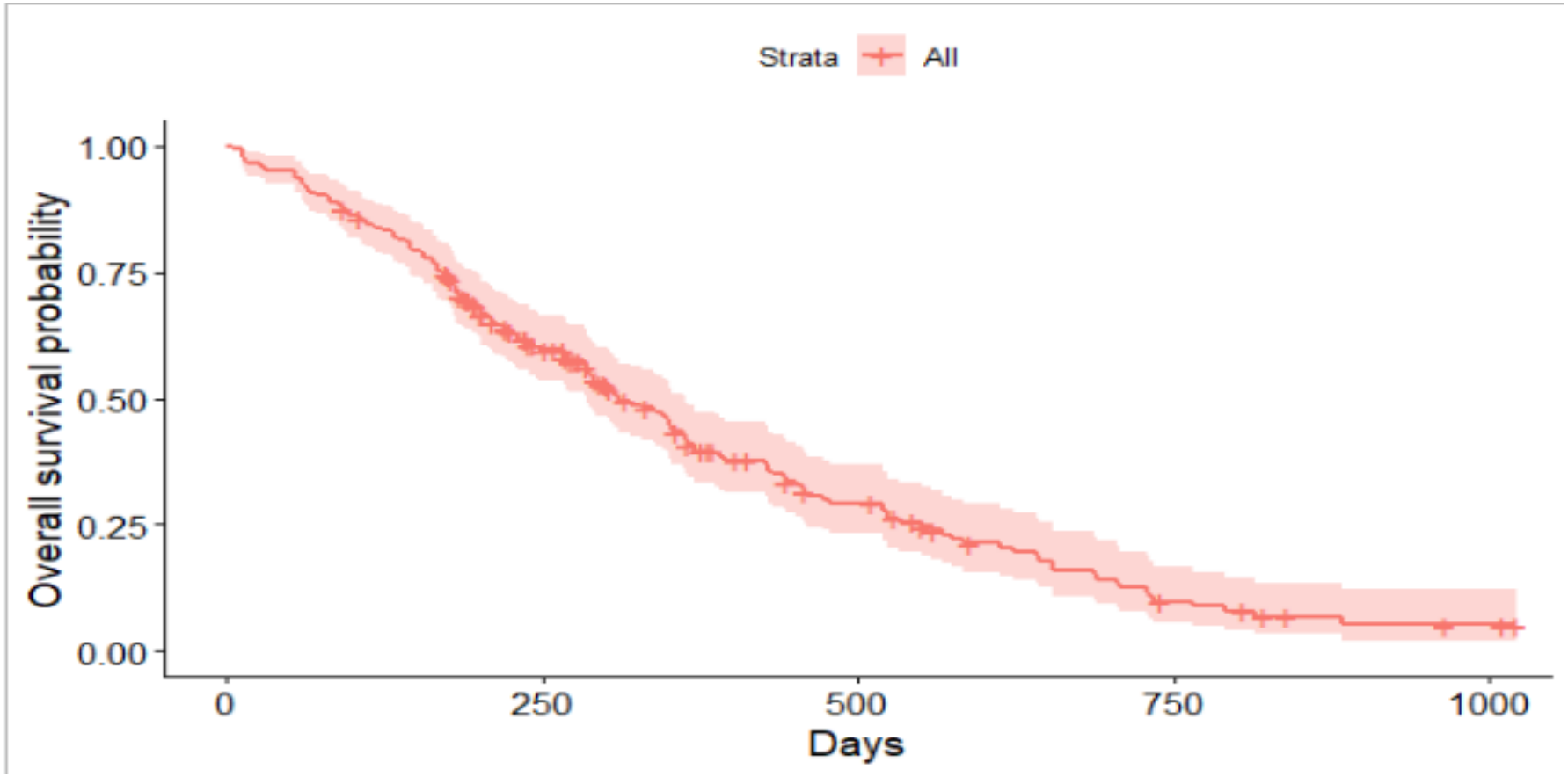
- You can create the survival curves by sex

$$f2 = survfit(Surv(time,\ status) \sim sex,\ data = lung)$$

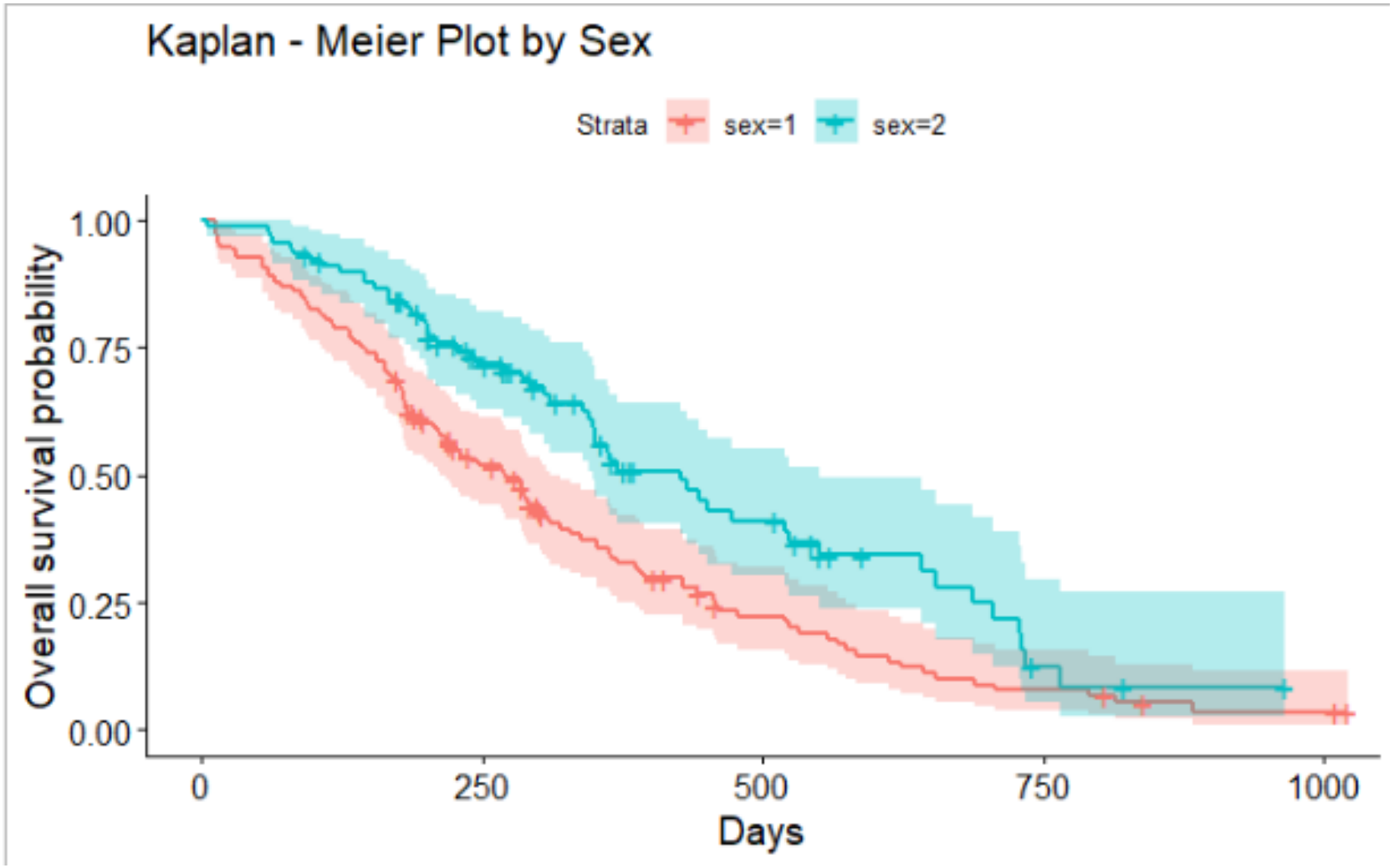
Kaplan-Meier Plot using $ggsurvplot$

The ggsurvplot function from the survminer package is built on ggplot2, and can be used to create Kaplan-Meier plots

```
ggsurvplot(
       fit = survfit(Surv(time, status) ~ 1, data = lung),
        xlab = "Days",
        ylab = "Overall survival probability")
```

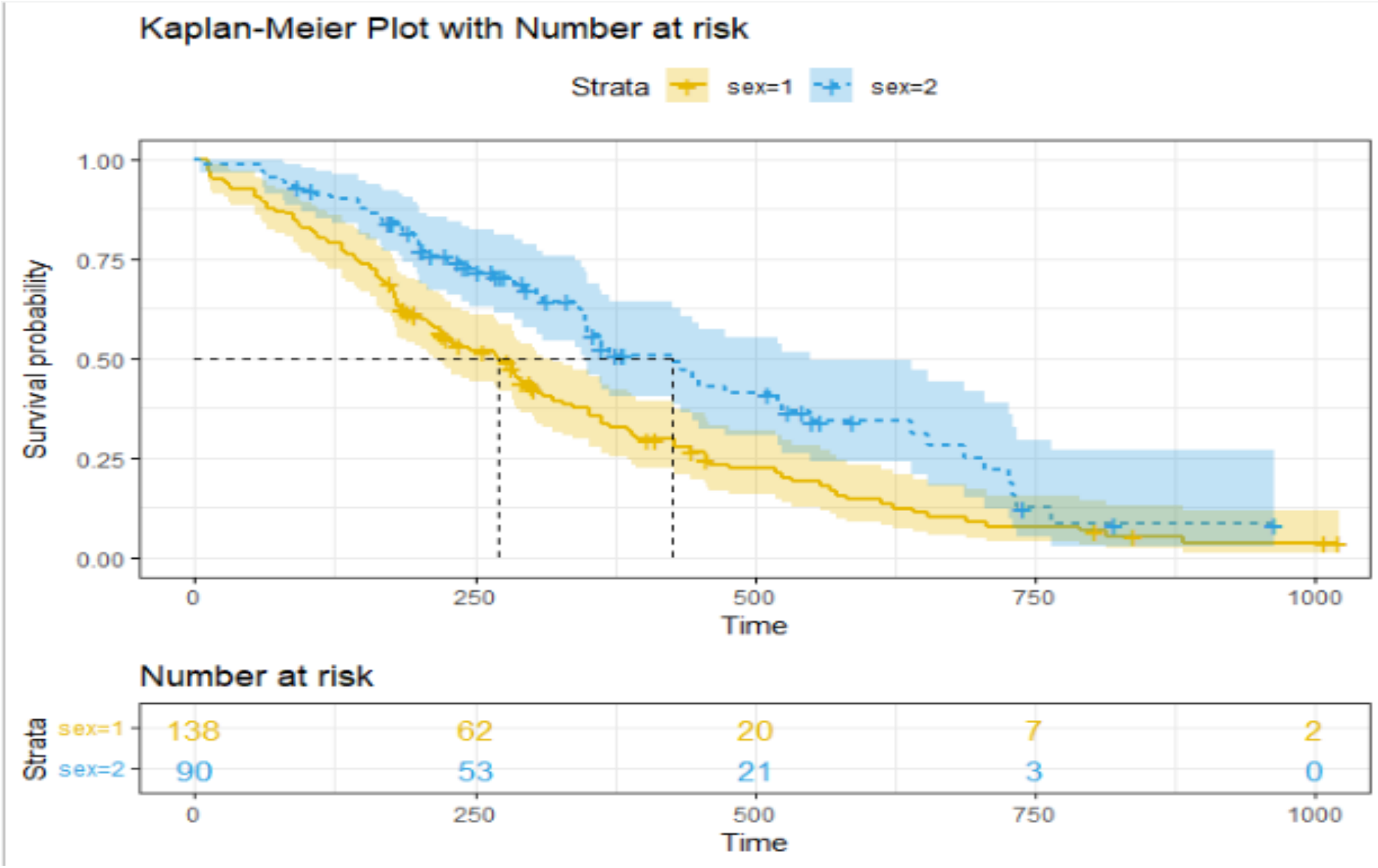

Strata
All
1.00
0.75
0.50
0.25
0.00
Overall survival probability
0
250
500
750
1000
Days

```
ggsurvplot(
       fit = survfit(Surv(time, status) ~ sex, data = lung),
       title = “Kaplan – Meier Plot by Sex”,
       xlab = "Days",
        ylab = "Overall survival probability")
```


Kaplan - Meier Plot by Sex
Strata
sex=1
sex=2
Overall survival probability
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
Days

A more detailed Curve

```
ggsurvplot(f2,
          pval = FALSE, conf.int = TRUE,
          title = “Kaplan-Meier Plot with Number at Risk”,
          risk.table = TRUE,              # Add risk table
          risk.table.col = "strata",   # Change risk table color by groups
          linetype = "strata",           # Change line type by groups
          surv.median.line = "hv",  # Specify median survival
          ggtheme = theme_bw(),  # Change ggplot2 theme
          palette = c("#E7B800", "#2E9FDF"))
```


Kaplan-Meier Plot with Number at risk
Strata
sex=1
sex=2
1.00
0.75
0.50
0.25
0.00
Survival probability
0
250
500
750
1000
Time
Number at risk
Strata
sex=1
sex=2
138
62
20
7
2
90
53
21
3
0
0
250
500
750
1000
Time

Estimating of $x - year$ Survival

One quantity of interest in a survival analysis is the probability of surviving beyond a certain number of years or months.

For example, to estimate the probability of surviving to 1 year, use $summary$ function with the $times$ argument.

```
summary(survfit(Surv(time, status) ~ sex, data = lung), times = 365.25)
```

Call: survfit(formula = Surv(time, status) ~ sex, data = lung)

sex=1

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
365.2500	35.0000	85.0000	0.3361	0.0434	0.2609	0.4329

sex=2

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
365.2500	30.0000	36.0000	0.5265	0.0597	0.4215	0.6576

Estimating Median Survival Time

Another quantity of interest analysis is the average survival time, which we quantity using the median.

The median survival time is obtained by:

$$summary(f2)\$table$$

summary(f2)$table

	records	n.max	n.start	events	rmean	se(rmean)	median	0.95LCL	0.95UCL
sex=1	138	138	138	112	326.0841	22.91156	270	212	310
sex=2	90	90	90	53	460.6473	34.68985	426	348	550

The median survival time for male group is 270 days compared to 426 days for female. This implies that female with lung cancer appears to have survival advantage over their male counterparts.

Comparing Survival Times between Groups

The significance tests between groups can be compared through the use of log-rank test. The test is the most common for comparing survival times and equally weights observation over the entire follow up period.

We obtained the p-value by

```
btwn_grp = survdiff(Surv(time, status) ~ sex, data = lung)
btwn_grp
```

Call:
survdiff(formula = Surv(time, status) ~ 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

References

- Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457-481.
- Clark, T., Bradburn, M., Love, S. *et al.* Survival Analysis Part I: Basic concepts and first analyses. *Br J Cancer* **89,** 232–238 (2003).
- Ilker Etikan, Suleiman Abubakar, and Rukyya Alkassim (2017) **The Kaplan Meier Estimate in Survival Analysis: Biometrics & Biostatistics International Journal** Volume 5 Issue 2 - 2017 *Department of Biostatistics, Near East University Faculty of Medicine, Cyprus (Article)*
- https://www.emilyzabor.com/tutorials/survival_analysis_in_r_tutorial.html#Packages