

# Survival Analysis

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Definitions . . . . .	2
1.2	Equation . . . . .	2
<b>2</b>	<b>Survival function, Hazard &amp; Hazard ratio</b>	<b>3</b>
2.1	The Survival function/Model . . . . .	3
2.2	Hazard . . . . .	3
2.3	Hazard Ratio . . . . .	3
2.4	Different models . . . . .	3
2.5	Kaplan-Meier survival model . . . . .	4
2.6	More examples of use . . . . .	4
<b>3</b>	<b>Censoring and truncation</b>	<b>5</b>
3.1	Definitions . . . . .	5
<b>4</b>	<b>Kaplan Meier Model in R</b>	<b>6</b>
4.1	Kaplan Meier Model with no X variable . . . . .	6
4.2	Summary . . . . .	6
4.3	Kaplan-Meier curve . . . . .	7
4.4	Kaplan-Meier model with an X variable . . . . .	8
4.5	Summary . . . . .	8
4.6	Kaplan-Meier curve with an X variable . . . . .	8
4.7	The LOG-RANK-TEST . . . . .	9
4.8	Exemple of plot with ggplot2 . . . . .	10
<b>5</b>	<b>Life tables</b>	<b>11</b>
5.1	Definitions . . . . .	11
<b>6</b>	<b>Kaplan-Meier Curves and Log-rank Test</b>	<b>12</b>
6.1	Calculating the steps . . . . .	12

# 1 Introduction

## 1.1 Definitions

The **Survival analysis** also known as time to event analysis is used to predict the time until an event occurs.

Most important concepts for this analysis:

- **Exposure** (clock starts), time zero of the analysis for this subject.
- **Event** (clock stops),
- **Survival time**, (difference between time of Event and time of Exposure).

A Survival time can be estimated in this examples:

- Exposure (Cancer diagnostic), Event (Death )
- Exposure (Marriage), Event (Divorce)

## 1.2 Equation

$Y$  = time-to-event

Where the  $Y$  = outcome, depends on *Time* and on the *Event* (0 = NO, 1 = YES) :

- 0 : the event didn't occurred
- 1 : the event occurred

## 2 Survival function, Hazard & Hazard ratio

### 2.1 The Survival function/Model

survival function  $S(t) = P(T > t)$  = Probability of Survival **beyond time t**.

### 2.2 Hazard

$Hazard(Haz) = P(T < t + d | T > t)$  = probability of dying in the next few seconds **given alive now**.

For the Exponential survival model, the hazard function correspond to the rate of the exponential curve.

### 2.3 Hazard Ratio

$Hazard(HR) = \frac{Haz, x=1}{Haz, x=0}$  = relative ratio,

At a given instant in time someone who is exposed is “relative ration” times more important to someone who is not.

### 2.4 Different models

Two type of functions/models to illustrate the decrease in survival probability.

$S(t)$  is the survival function

- **Kaplan-Meier survival model** (non-parametric)
  - Pros : Simple to interpret, can estimate  $S(t)$
  - Cons : No functional form (no mathematical function, because of steps), can not estimate hazard ratio
- **Exponential survival model** (parametric)
  - Pros : Can estimate the  $S(t)$ , and Hazard ratio
  - Cons : Not always realistic, because assumes constant hazard ( death is not constant )
- **Cox proportional Hazard model** (semi-parametric), sort of a combination of KM model and Exponential model
  - Pros : Haz can fluctuate with time, Can estimate Hazard ratio
  - Cons : Can not estimate  $S(t)$

## 2.5 Kaplan-Meier survival model

Also known as Product-Limit Method, or the life table method.

This is a non-parametric curve, explains the selected data. The ticks are censored data

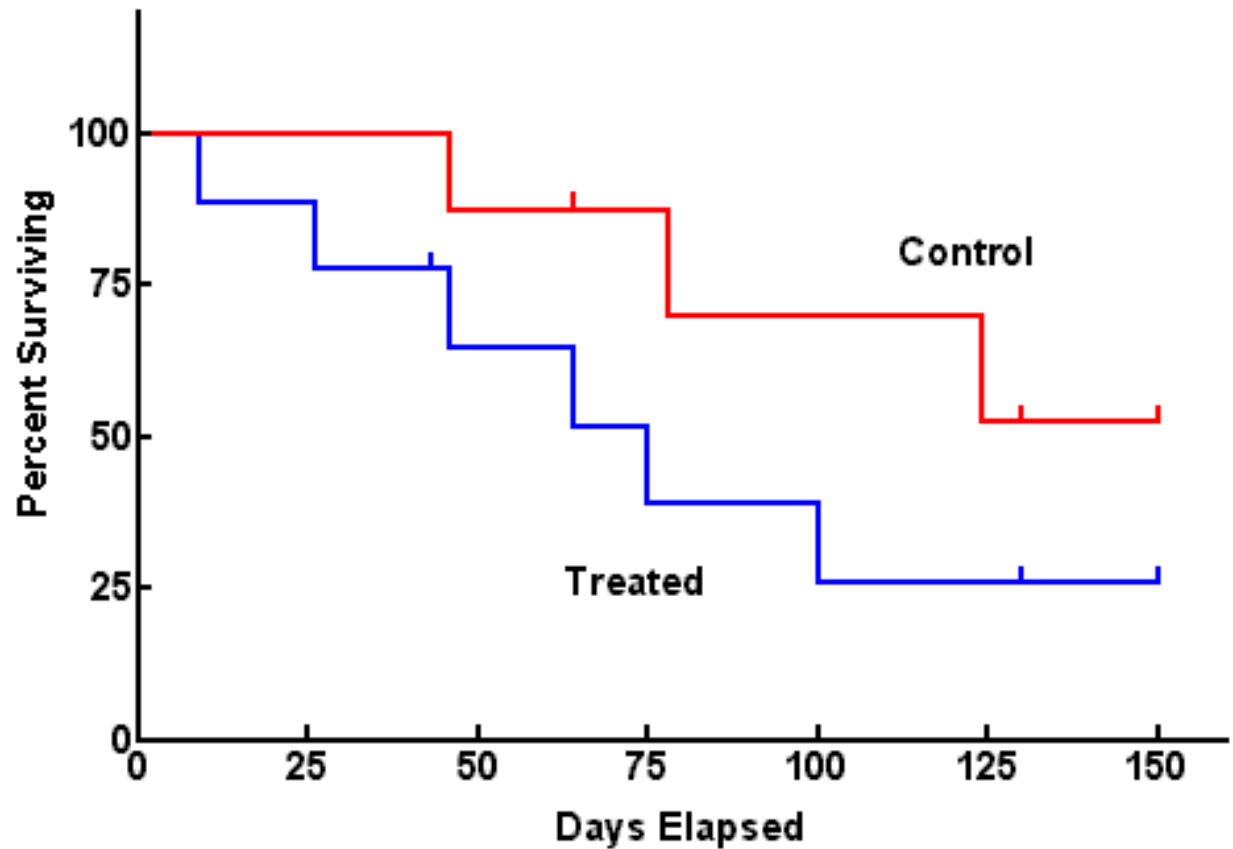


Figure 1: Kaplan Meier Curve

## 2.6 More examples of use

In Health, Time to:

- Death
- Device failure
- Re-admission

In industry, Time to:

- Component failure
- Business failure
- staff promotion

In marriage, Time to:

- Divorce
- Second child

## 3 Censoring and truncation

### 3.1 Definitions

**Censoring** : Censoring occurs when we don't know the exact time-to-event for an included observation. Lack of some part of the information.

There are 3 different types of censoring :

- Right censoring (time to event greater than a value), this type of censoring could sometimes be informative when the censoring is actually correlated to the expected event (consequence).
- Left censoring (time to event less than a value)
- Interval censoring ( time to event between two values)

An example of right censoring is a time-to-event that continues after the experiment ( patient that is still alive a the end of the study).

**Truncation** : Truncation occurs when observation are excluded by virtue of their time-to-event. Short or Long time events that where not measured.

There are 2 different types of truncation :

- left truncation ( short time-to-event values, small values that where not measured )
- right truncation ( long time-to-event values, large values that where not measured )

The whole data set can be truncated, whereas data points can be censored.

## 4 Kaplan Meier Model in R

```
rm(list = ls())

library(tidyverse)
library(survival)

time    <- c(2,4,6,8,11,15,16,18,18,20,22,22,25,27,28,32,32,34,34)
death   <- c(1,1,0,1,1,1,1,0,0,1,0,1,1,1,1,0,1,0,0) # Censoring or not, 1 = died, 0 = censored
over40  <- c(1,1,0,1,1,1,1,0,0,1,0,1,0,1,1,0,1,1,0)  # Is over 40 or not, 1 = YES, 0 = NO

df <- tibble(time,death,over40)
```

### 4.1 Kaplan Meier Model with no X variable

### 4.2 Summary

```
# We use ~1 when there is no X variable ( additional categorical variable )
km.model <- survfit(Surv(time = time, event = death) ~ 1,
                    type = "kaplan-meier") # Kaplan-Meier is the default value

km.model
```

```
## Call: survfit(formula = Surv(time = time, event = death) ~ 1, type = "kaplan-meier")
##
##      n events median 0.95LCL 0.95UCL
## [1,] 19      12      25      16      NA
```

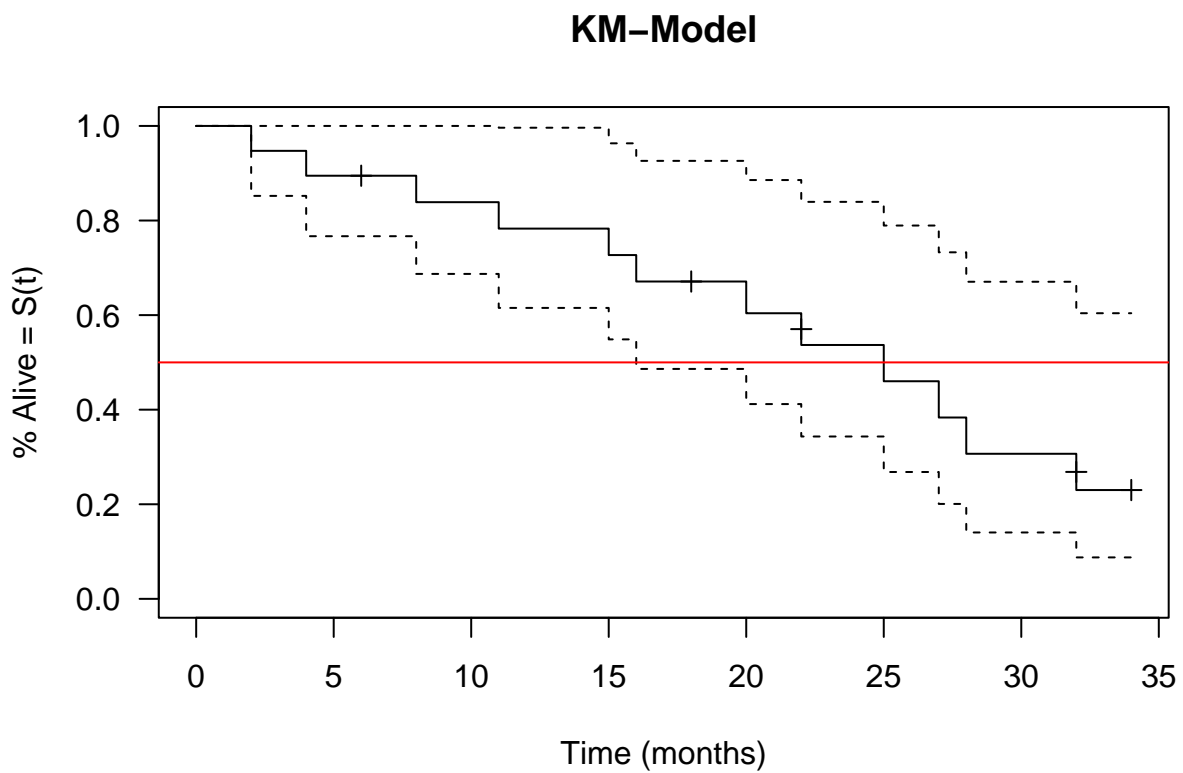
To visualize the survival at specific time, and confidence intervals.

```
summary(km.model)
```

```
## Call: survfit(formula = Surv(time = time, event = death) ~ 1, type = "kaplan-meier")
##
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   ---- -
##      2      19        1   0.947  0.0512   0.8521      1.000
##      4      18        1   0.895  0.0704   0.7669      1.000
##      8      16        1   0.839  0.0854   0.6871      1.000
##     11      15        1   0.783  0.0963   0.6152      0.996
##     15      14        1   0.727  0.1044   0.5487      0.963
##     16      13        1   0.671  0.1103   0.4862      0.926
##     20      10        1   0.604  0.1179   0.4119      0.886
##     22       9        1   0.537  0.1224   0.3433      0.839
##     25       7        1   0.460  0.1267   0.2682      0.789
##     27       6        1   0.383  0.1267   0.2007      0.733
##     28       5        1   0.307  0.1224   0.1404      0.671
##     32       4        1   0.230  0.1133   0.0876      0.604
```

### 4.3 Kaplan-Meier curve

```
plot(  
  km.model,      # used model  
  conf.int = T,  # include confidence intervals  
  xlab = "Time (months)",  
  ylab = "% Alive = S(t)",  
  main = "KM-Model",  
  las = 1,      # rotates the values on the y axis for better readability  
  mark.time = T # adds the censored values to the graph as a tick  
)  
  
abline(h = 0.5, col = "red")
```



## 4.4 Kaplan-Meier model with an X variable

## 4.5 Summary

```
# We use ~1 when there is no X variable ( additional categorical variable )
km.model2 <- survfit(Surv(time = time, event = death) ~ over40,
                    type = "kaplan-meier") # Kaplan-Meier is the default value

km.model2
```

```
## Call: survfit(formula = Surv(time = time, event = death) ~ over40,
##      type = "kaplan-meier")
##
##              n events median 0.95LCL 0.95UCL
## over40=0  7         1    NA      25      NA
## over40=1 12        11    18      11      NA
```

To visualize the survival at specific time, and confidence intervals.

```
summary(km.model2)
```

```
## Call: survfit(formula = Surv(time = time, event = death) ~ over40,
##      type = "kaplan-meier")
##
##              over40=0
##      time      n.risk      n.event      survival      std.err lower 95% CI
##      25.000         3.000         1.000         0.667         0.272         0.300
## upper 95% CI
##      1.000
##
##              over40=1
## time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    2    12     1  0.9167  0.0798    0.7729    1.000
##    4    11     1  0.8333  0.1076    0.6470    1.000
##    8    10     1  0.7500  0.1250    0.5410    1.000
##   11     9     1  0.6667  0.1361    0.4468    0.995
##   15     8     1  0.5833  0.1423    0.3616    0.941
##   16     7     1  0.5000  0.1443    0.2840    0.880
##   20     6     1  0.4167  0.1423    0.2133    0.814
##   22     5     1  0.3333  0.1361    0.1498    0.742
##   27     4     1  0.2500  0.1250    0.0938    0.666
##   28     3     1  0.1667  0.1076    0.0470    0.591
##   32     2     1  0.0833  0.0798    0.0128    0.544
```

## 4.6 Kaplan-Meier curve with an X variable

```
plot(
  km.model2,      # used model
  conf.int = F,   # include confidence intervals
```

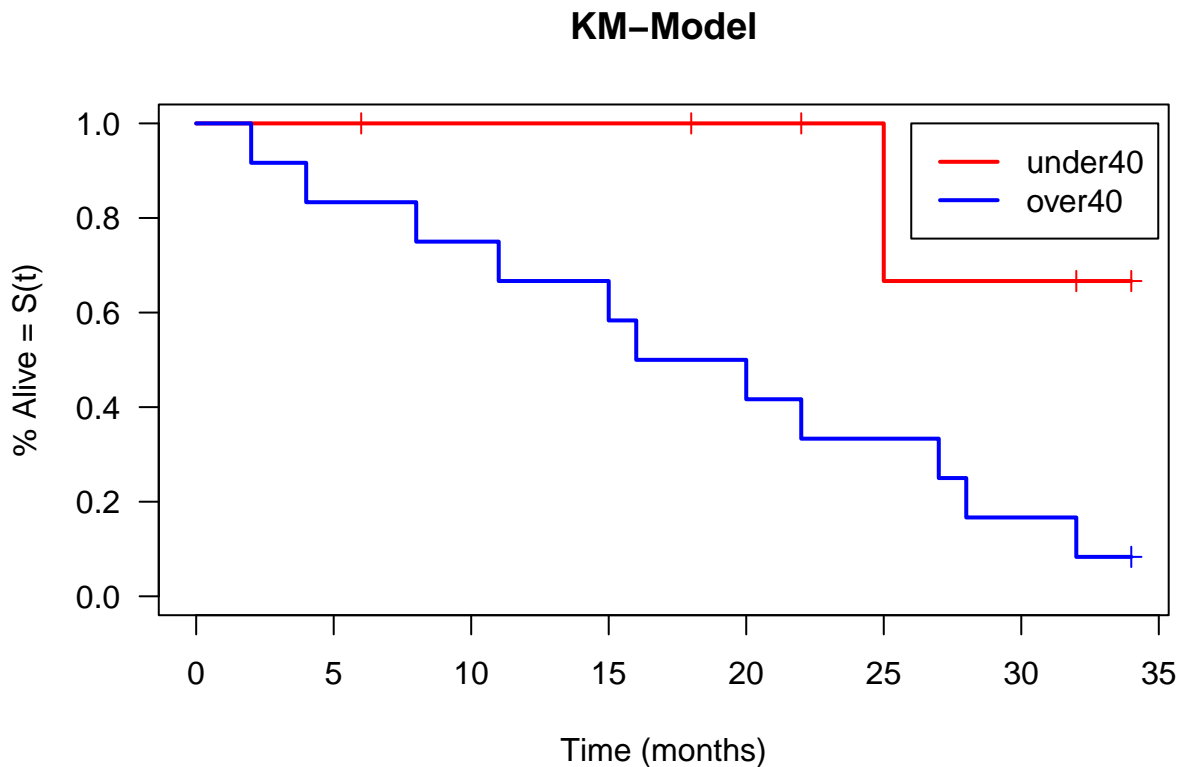


```

xlab = "Time (months)",
ylab = "% Alive = S(t)",
main = "KM-Model",
col = c("red","blue"), # add colors to the plots
lwd = 2, # makes the line a bit more thick
las = 1, # rotatates the values on the y axis for better readability
mark.time = T # adds the censored values to the graph as a tick
)

legend(26, # x coordinate for the box
1, # y coordinate for the box
legend = c("under40","over40"), # Names for the legend
lty = 1, # linetype
lwd = 2, # linewidth
col = c("red","blue"),
bty = "", # boxtype shape
cex = 1 # boxfont size
)

```



## 4.7 The LOG-RANK-TEST

The logrank test assesses whether the KM survival curves from two subpopulation are significantly different. Comparing the survival curves to see if they are different.

$H_0$ : survival in two groups is **the same**.  $H_1$ : survival in the two groups is **not the same**.

```

survdif(Surv(time,death)~ over40) # This can work with also more than 2 levels ( )

```

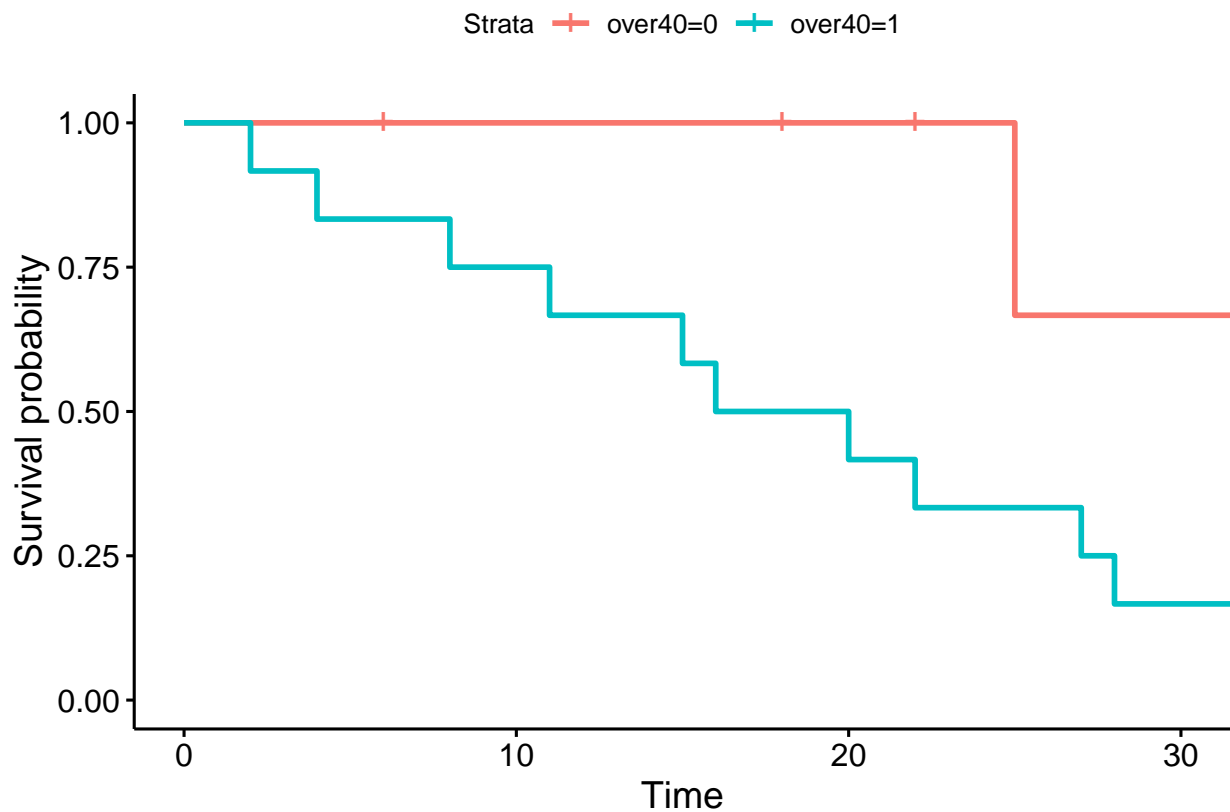
```
## Call:
## survdiff(formula = Surv(time, death) ~ over40)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## over40=0  7         1   4.93      3.13    5.36
## over40=1 12        11   7.07      2.18    5.36
##
##  Chisq= 5.4  on 1 degrees of freedom, p= 0.02
```

#### 4.8 Exemple of plot with ggplot2

```
km.model2 <- survfit(Surv(time = time, event = death) ~ over40,
                     type = "kaplan-meier") # Kaplan-Meier is the default value

p <- survminer::ggsurvplot(fit = km.model2, data = df)

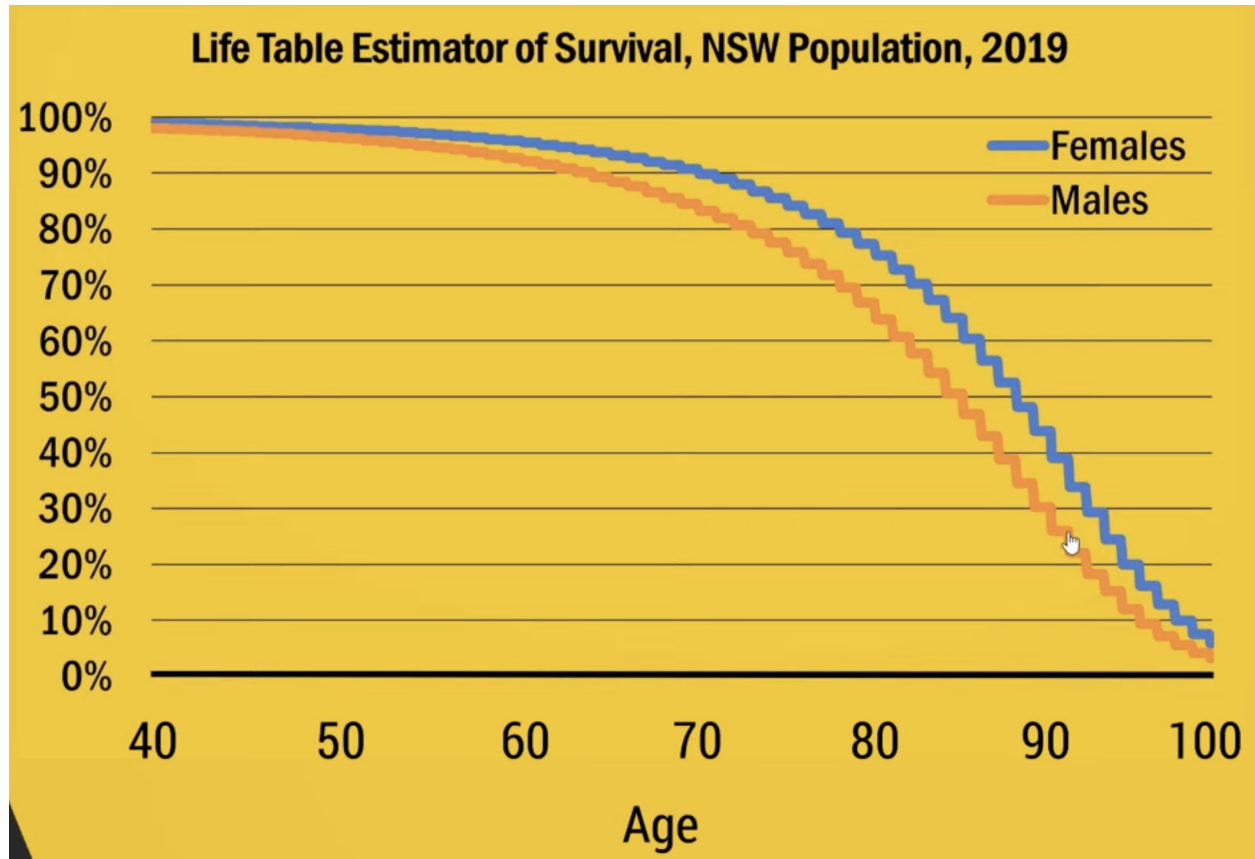
p
```



## 5 Life tables

### 5.1 Definitions

Life tables illustrate the pattern of survivorship of a population by considering the probability of death at each consecutive age.



Input :

- population by age group
- deaths in each age group

Output :

- the survival functions at each age
- conditional life expectancy at each age ( probability of being alive)
- median, mean quantile survival

## 6 Kaplan-Meier Curves and Log-rank Test

### 6.1 Calculating the steps

It's a **non-parametric** estimator of survival. Doesn't have an estimator ( like the mean or the standard deviation for an normal distribution ). A Kaplan-Meier curve looks more like steps than a curve, and doesn't use parameters, it represents the data.

Wombat	Survival Time	Died? (1=died)	Time	n	d	Calculation	Survival, S(t)
F	4	1	$0 \leq t < 4$	10	0	<b>1</b>	<b>1</b>
G	6	1	$4 \leq t < 6$	10	1	$9/10 \times 1$	<b>0.9</b>
J	8	0	$6 \leq t < 8$	9	1	$8/9 \times 0.9$	<b>0.8</b>
H	11	1	$8 \leq t < 11$	8	0	$8/8 \times 0.8$	<b>0.8</b>
A	15	1	$11 \leq t < 15$	7	1	$6/7 \times 0.8$	<b>0.686</b>
E	15	1	$15 \leq t < 20$	6	2	$4/6 \times 0.686$	<b>0.457</b>
C	20	0	$20 \leq t < 25$	4	1	$3/4 \times 0.457$	<b>0.342</b>
D	25	1	$25 \leq t < 31$	2	1	$1/2 \times 0.342$	<b>0.171</b>
B	31	0					

Figure 2: Survival Table

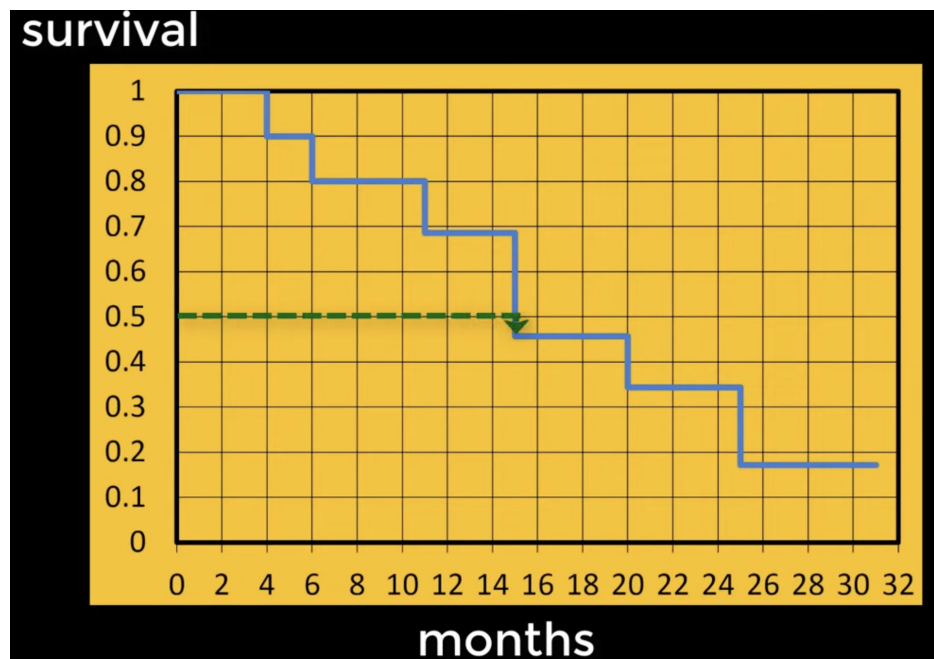


Figure 3: Survival Curve