Fitting Of Distributions For Survival Data

BLADDER CANCER

Fathima Ayoob
18-Pst-021

CONTENTS

INTRODUCTION
1

ABOUT THE STUDY
2

METHODOLOGY
3

ANALYSIS
6

CONCLUSION
11

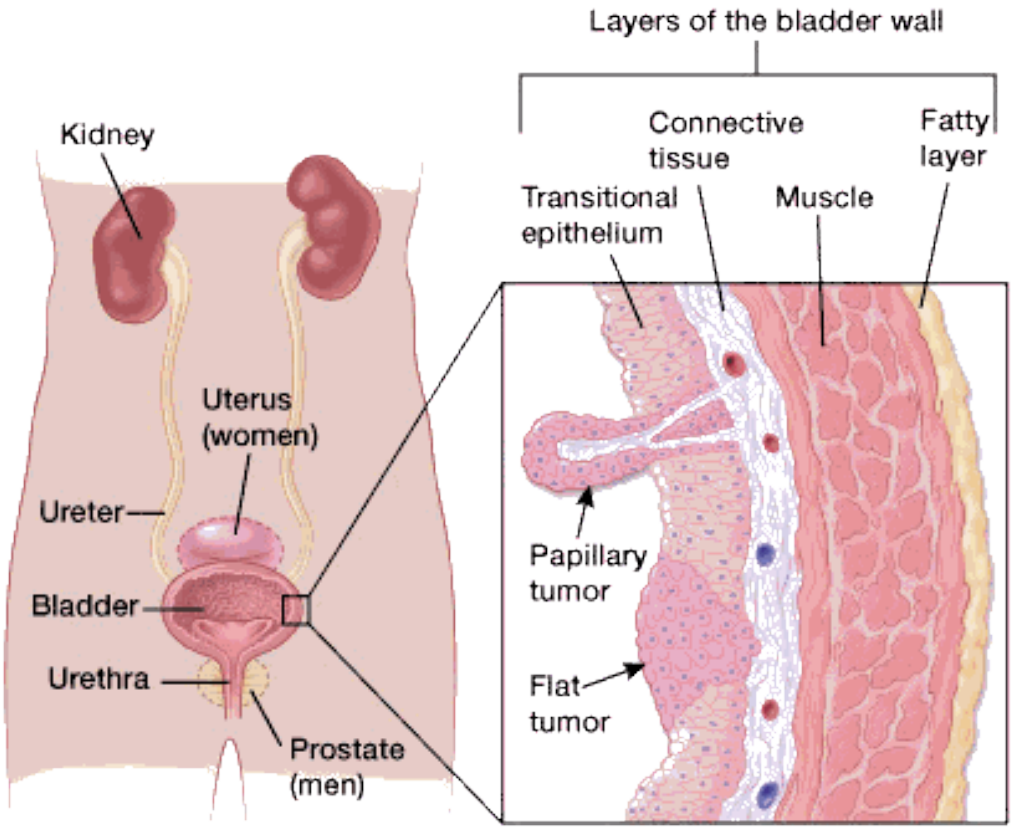
REFERENCES
12

INTRODUCTION

BLADDER CANCER

Bladder cancer starts when cells that make up the urinary bladder start to grow out of control. As more cancer cells develop, they can form a tumor and, with time, spread to other parts of the body.

The bladder is a hollow organ in the lower pelvis. It has flexible, muscular walls that can stretch to hold urine and squeeze to send it out of the body. The bladder's main job is to store urine. Urine is liquid waste made by the 2 kidneys and then carried to the bladder through 2 tubes called ureters. When you urinate, the muscles in the bladder contract, and urine is forced out of the bladder through a tube called the urethra.



ABOUT THE STUDY

OBJECTIVE

We obtained a secondary survival data and fit various possible survival distributions.
Then used graphical and statistical tests to compare the goodness of fit.

ABOUT THE DATA

The Bladder cancer data was obtained from M Pagano and K Gauvreau, "Principles of Biostatistics, 2nd Ed. Duxbury 2000. Chapter 21, exercise 9, page 512.
86 patients after surgery were assigned to placebo or chemotherapy (thiopeta).
Endpoint is time to recurrence in months. Data on the number of tumors removed at surgery was also collected.

The columns in the data were as follows:

Time: Denotes the survival time in months

Group: A categorical variable, which was assigned as 1 to patients who were given placebo and 0 for chemotherapy

Censor: A categorical variable which indicates whether the data with respect to the patient is censored = 0 or censor=1 indicates remission.

Number: Indicates number of tumors removed after surgery

METHODOLOGY

SURVIVAL ANALYSIS

Survival analysis corresponds to a set of statistical approaches used to investigate the time it takes for an event of interest to occur.

Survival time and type of events in cancer studies

There are different types of events, including:

- Relapse
- Progression
- Death

The time from 'response to treatment' (complete remission) to the occurrence of the event of interest is commonly called survival time (or time to event).

CENSORED DATA

Survival analysis focuses on the expected duration of time until occurrence of an event of interest (relapse or death). However, the event may not be observed for some individuals within the study time period, producing the so-called censored observations.

Censoring may arise in the following ways:

- a patient has not (yet) experienced the event of interest, such as relapse or death, within the study time period;
- a patient is lost to follow-up during the study period;
- a patient experiences a different event that makes further follow-up impossible.

This type of censoring, named right censoring, is handled in survival analysis.

SURVIVAL FUNCTION

Assume that T is a continuous random variable with probability density function (p.d.f.) f(t) and cumulative distribution function (c.d.f.) F(t) = Pr{T < t}, giving the probability that the event has occurred by duration t.
The Survival function is given by,
S(t) = Pr{T ≥ t} = 1 − F(t)
which gives the probability of being alive just before duration t, or more generally, the probability that the event of interest has not occurred by duration t.

HAZARD FUNCTION

An alternative characterization of the distribution of T is given by the hazard function, or instantaneous rate of occurrence of the event, defined as

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt | T \geq t\}}{dt}.$$

The numerator of this expression is the conditional probability that the event will occur in the interval [t, t+dt) given that it has not occurred before, and the denominator is the width of the interval. Dividing one by the other we obtain a rate of event occurrence per unit of time. Taking the limit as the width of the interval goes down to zero, we obtain an instantaneous rate of occurrence.

The difference between a hazard function and a survival function can be summarized as follows:

- The hazard function focuses on failing.
- The survival function focuses on surviving.

SURVIVAL DISTRIBUTIONS

Survival times are subject to random variations and like any random variables form a distribution , here known as survival distribution.

Some of the most important distributions are as follows:

EXPONENTIAL DISTRIBUTION

$$\begin{aligned} h(t) &= \lambda, \\ S(t) &= \exp(-\lambda t), \\ f(t) &= \lambda \exp(-\lambda t). \end{aligned}$$

WEIBULL DISTRIBUTION

$$\begin{aligned} h(t) &= \lambda p(\lambda p)^{p-1}, \\ S(t) &= \exp(-(\lambda t)^p), \\ f(t) &= \lambda p(\lambda p)^{p-1} \exp(-(\lambda t)^p). \end{aligned}$$

GAMMA DISTRIBUTION

$$f(t) = \frac{\lambda(\lambda t)^{k-1} e^{-\lambda t}}{\Gamma(k)},$$

$$S(t) = 1 - I_k(\lambda t),$$

LOGNORMAL DISTRIBUTION

$$\begin{aligned} h(t) &= \frac{\alpha}{\sqrt{2\pi t}} \exp\left(\frac{-\alpha^2(\ln(\lambda t))^2}{2}\right)\left(1 - \Phi(\alpha \ln(\lambda t))\right)^{-1}, \\ S(t) &= 1 - \Phi(\alpha \ln(\lambda t)), \\ f(t) &= \frac{\alpha}{\sqrt{2\pi t}} \exp\left(\frac{-\alpha^2(\ln(\lambda t))^2}{2}\right). \end{aligned}$$

ANALYSIS

CURVES

EXPONENTIAL DISTRIBUTION

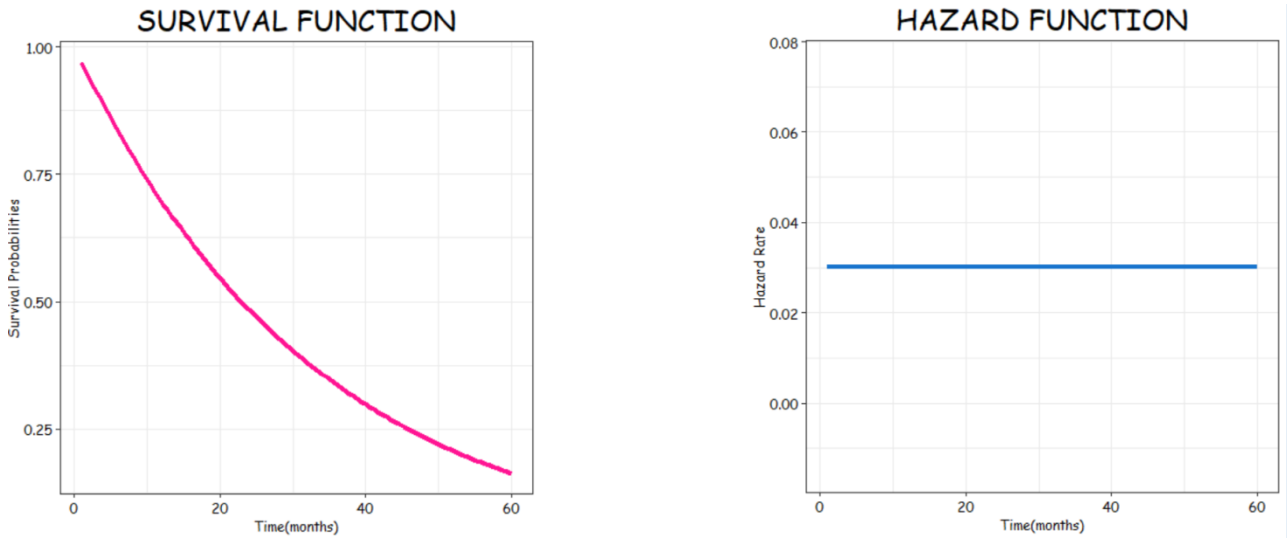


Fig 1

WEIBULL DISTRIBUTION

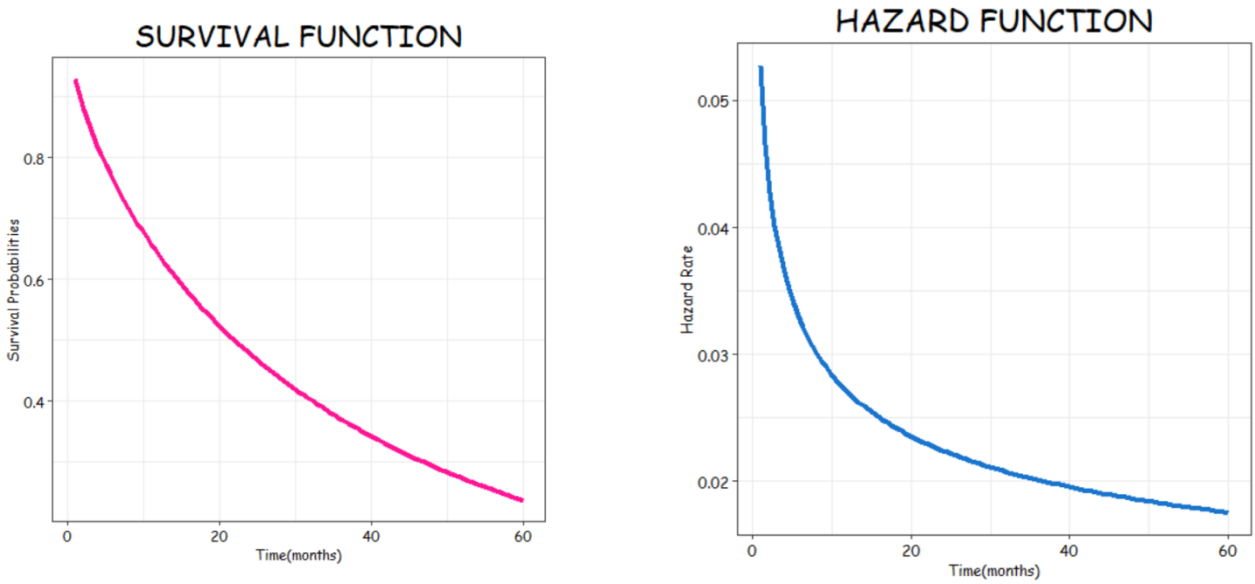


Fig 2

GAMMA DISTRIBUTION

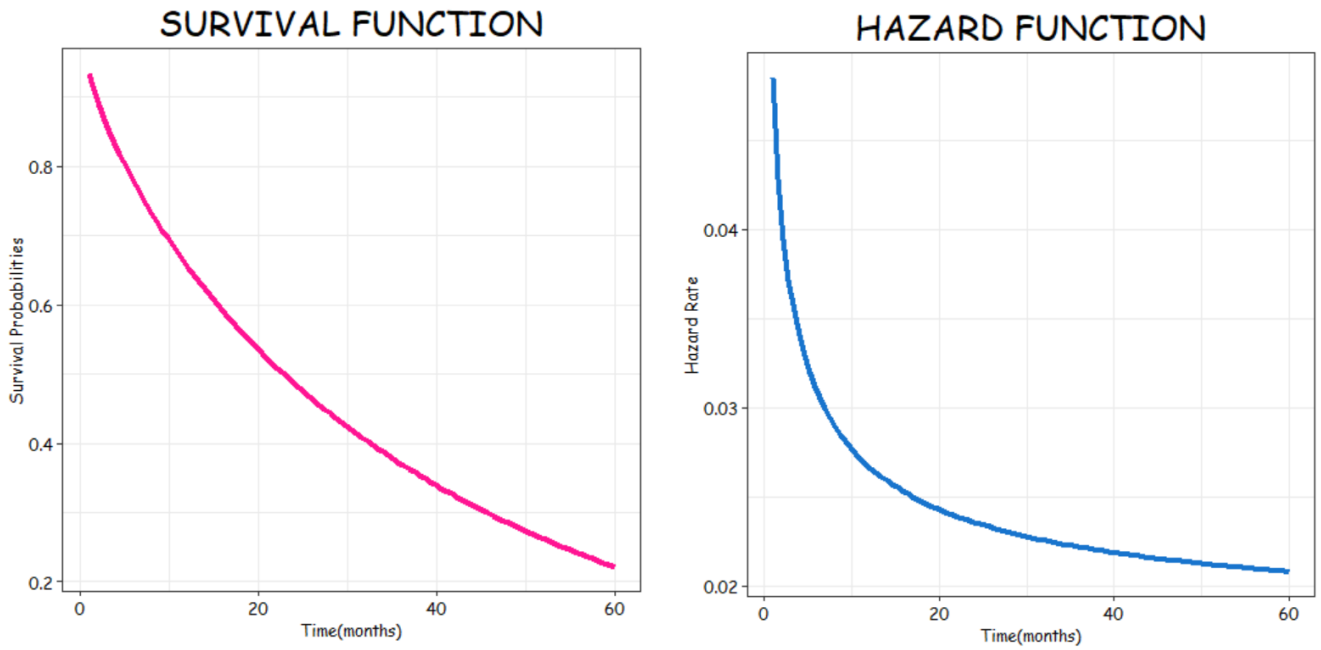


Fig 3

LOGNORMAL DISTRIBUTION

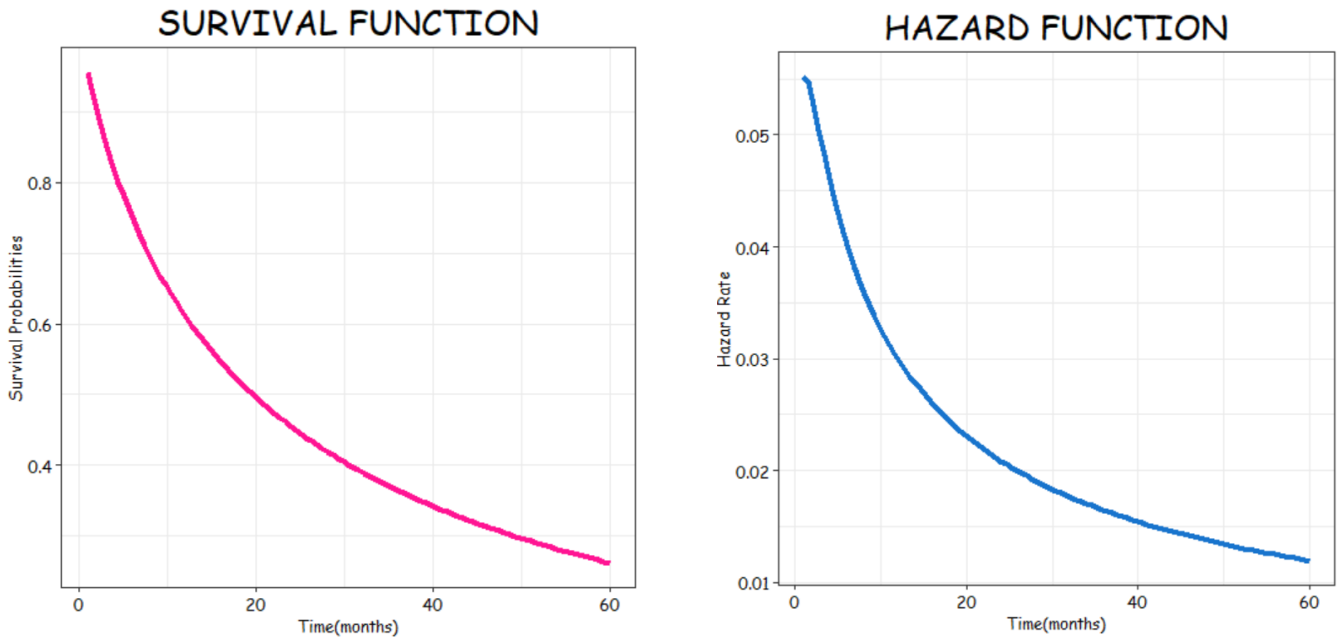


Fig 4

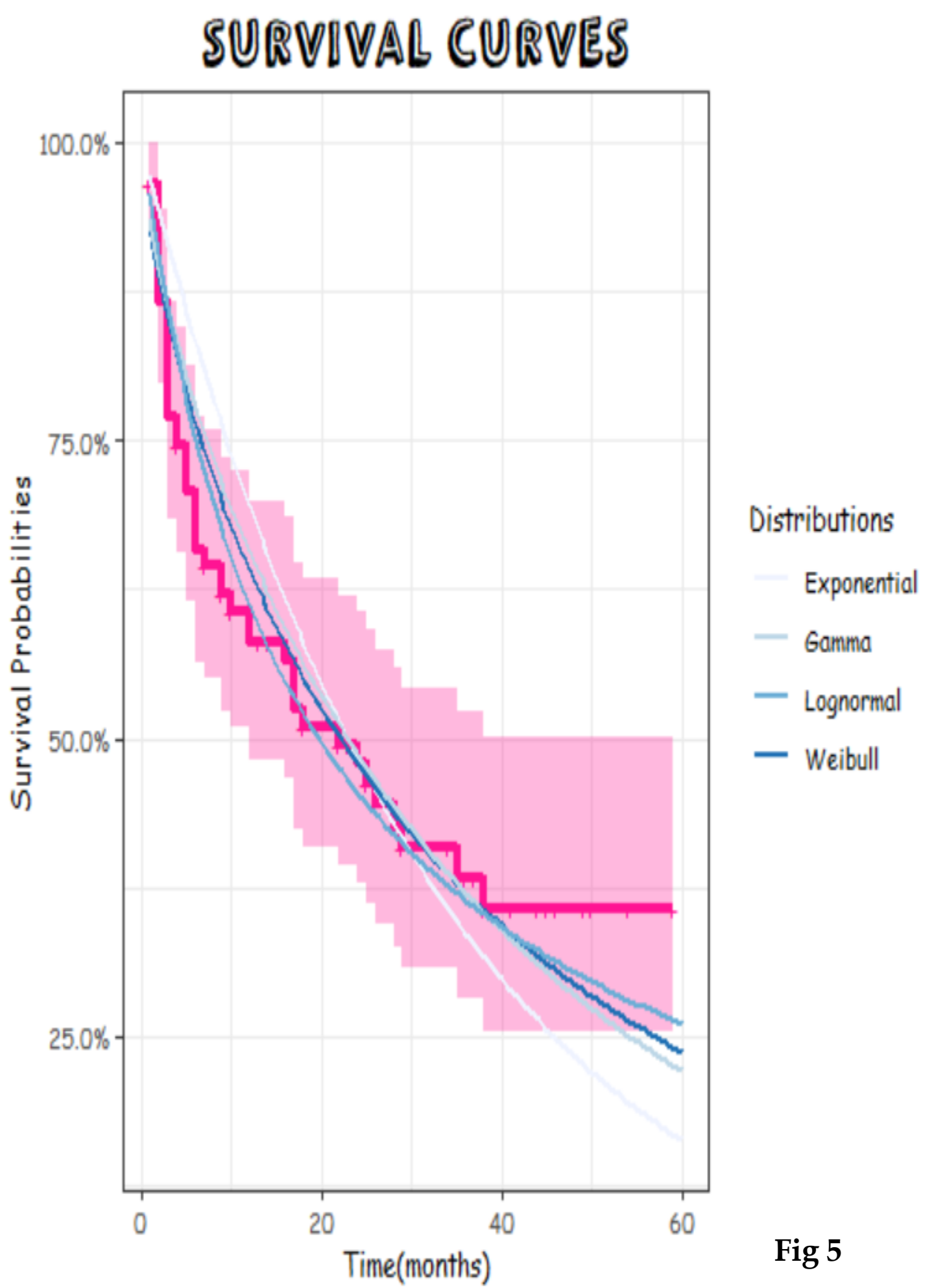

SURVIVAL CURVES
100.0%
75.0%
50.0%
25.0%
Survival Probabilities
0
20
40
60
Time(months)
Distributions
Exponential
Gamma
Lognormal
Weibull

Fig 5

QQ PLOTS

EXPONENTIAL DISTRIBUTION

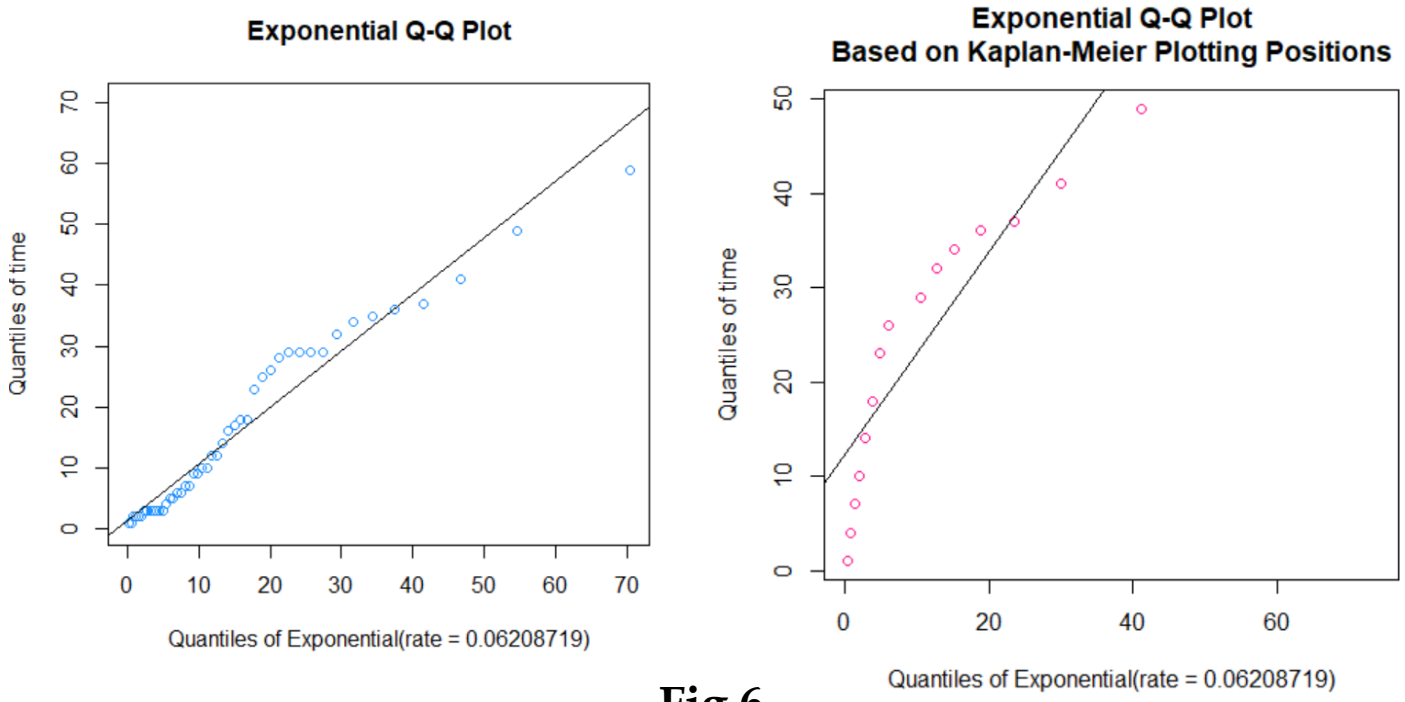


Fig 6

WEIBULL DISTRIBUTION

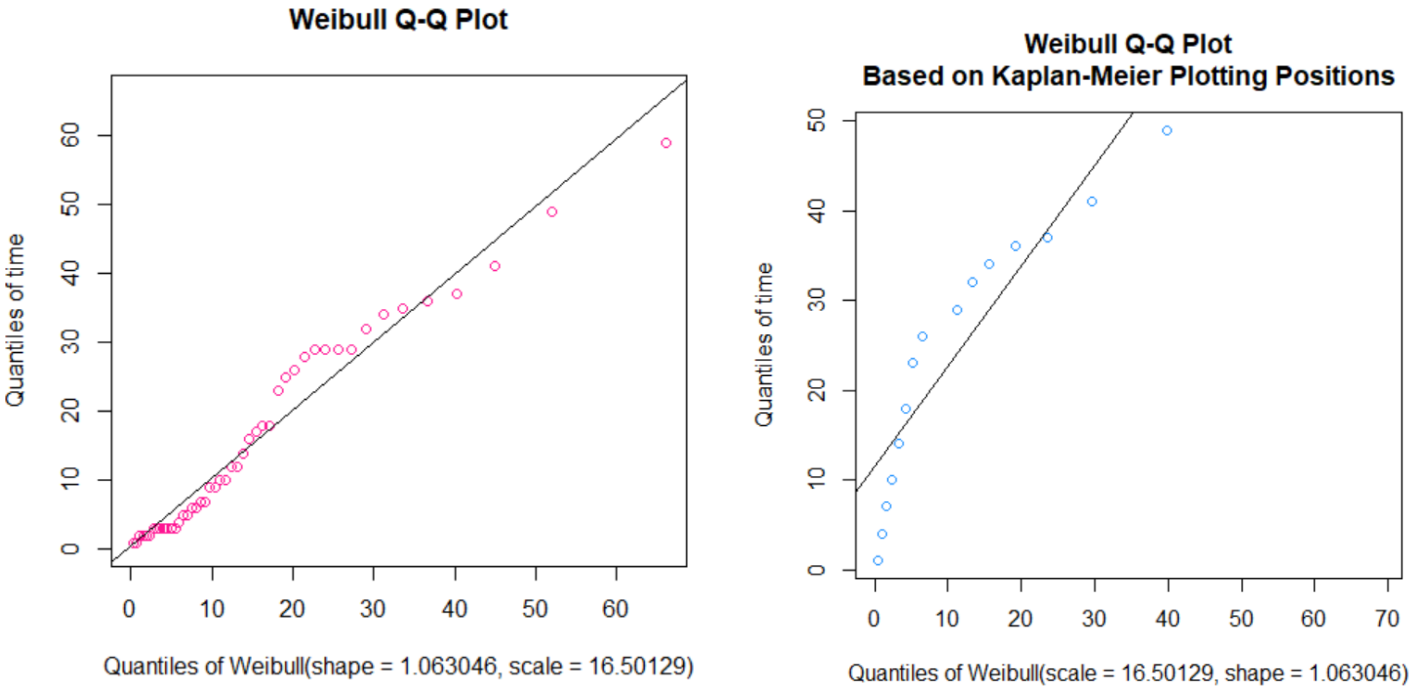


Fig 7

GAMMA DISTRIBUTION

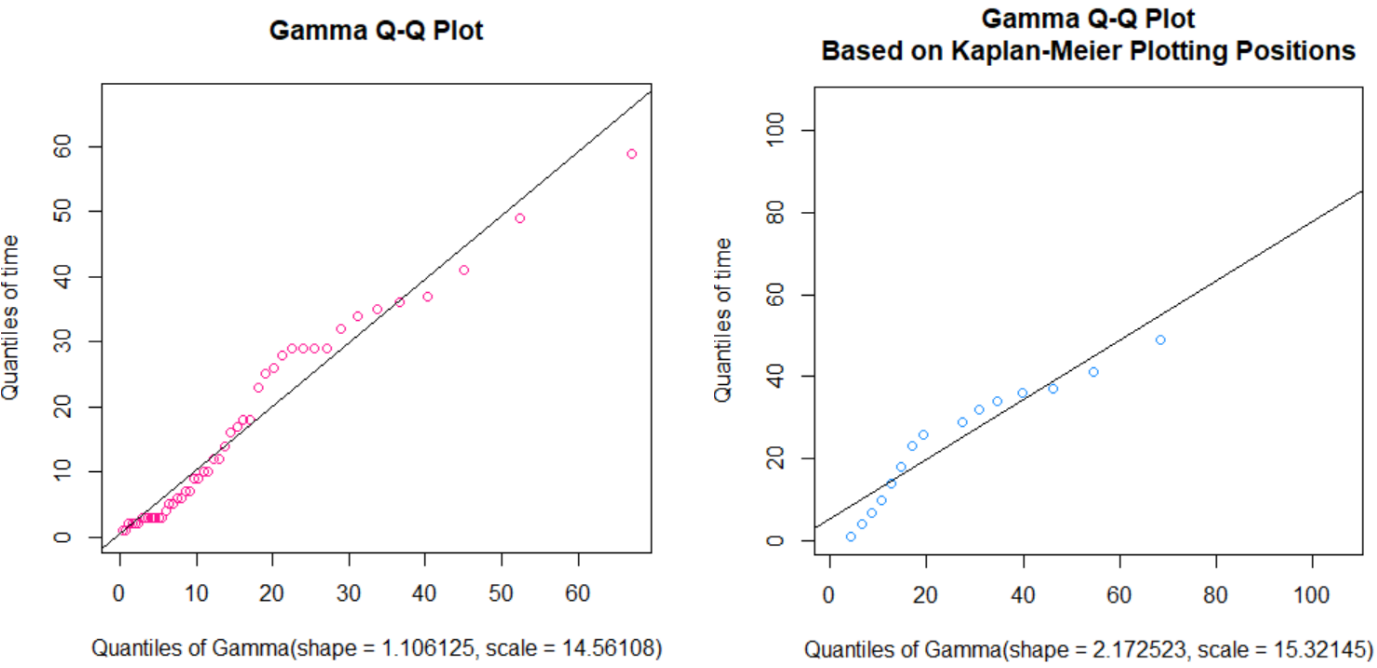


Fig 8

LOGNORMAL DISTRIBUTION

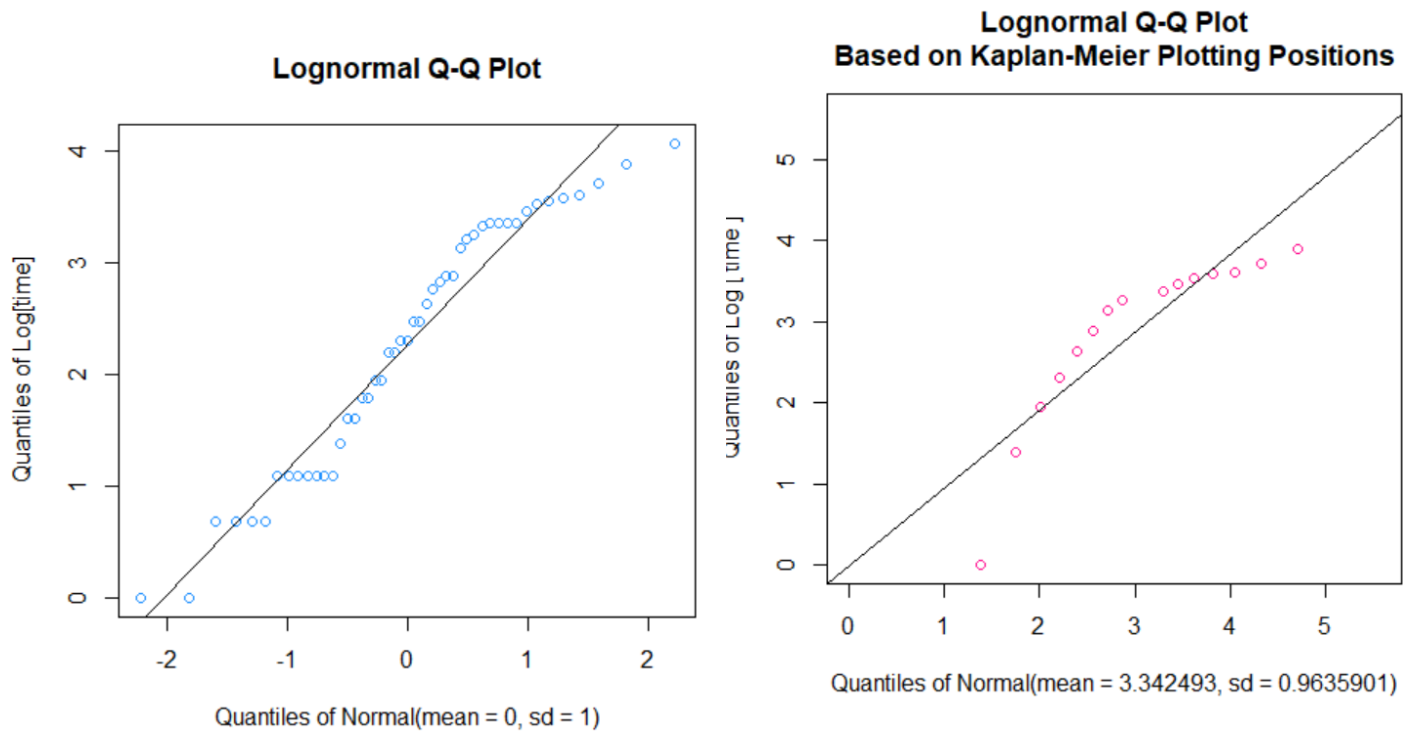


Fig 9

CONCLUSION

CURVES

Figures 1,2,3 and 4 shows the survival curves and hazard curves of all the distributions.
From figure 5 we can say that Gamma and Lognormal distributions almost fit the Kaplan Meir curve.

Q-Q PLOTS

In figures 6,7,8 and 9 the left hand side plots for the entire data while the right hand side plots for the censored data.
Again from these plots we can say that Gamma and Lognormal distributions almost fit the data.

GOOD OF FIT TEST

Shapiro-Francia Goodness of Fit test was applied on Gamma and Lognormal distributions.

H0: Data follows Gamma Distribution.
H1: True cdf does not equal the Gamma Distribution.
P-value:0.0692978
Since p-value is greater than 0.05 we accept the null hypothesis

H0: Data follows Lognormal Distribution
H1: True cdf does not equal the Lognormal Distribution.
P-value: 0.002814913
Since the p-value is less than 0.05 we reject the null hypothesis
Hence we conclude that Gamma distribution best fits the survival data.

REFERENCES

http://www.stat.rice.edu/~sneeley/STAT553/Datasets/survivaldata.txt
https://www.rdocumentation.org/packages/EnvStats/versions/2.3.1/topics/qqPlotCensored
https://www.rdocumentation.org/packages/EnvStats/versions/2.3.1/topics/distChooseCensored
http://www.sthda.com/english/wiki/survival-analysis-basics
https://www.cancer.org/cancer/bladder-cancer/about/what-is-bladder-cancer.html
https://rdrr.io/cran/ggfortify/man/autoplot.survfit.html

DATA

time	group	censor	number
0	1	0	1
1	1	0	2
4	1	0	1
7	1	0	1
10	1	0	1
6	1	1	1
14	1	0	1
18	1	0	1
5	1	1	2
12	1	1	1
23	1	0	2
10	1	1	2
3	1	1	1
3	1	1	1
7	1	1	2
3	1	1	1
26	1	0	2
1	1	1	1
2	1	1	2
25	1	1	2
29	1	0	2
29	1	0	2
29	1	0	1

CODES

The following libraries were used:
gridExtra,Kmsurv,ggplot2,survminer,ggfortify,survival,flexsurv
extrafont and EvtStat

KM curve

```
fit <- survfit(Surv(time, censor) ~ 1, data = dat1)
autoplot(fit)
```

Survival and Hazard Curve

```
fit_exp <- flexsurvreg(Surv(time, censor) ~ 1,dist = "exponential",
data =dat2 )
```

#functions

```
expo <- function(x) {1-pexp(x,0.03023)}
hexpo<- function(x) {hexp(x,0.03023)}
```

#survival curve

```
exo_s=ggplot(data.frame(x = c(1, 60)), aes(x = x))+
stat_function(fun = expo, size=1.5,col="deeppink")+
theme_bw()+
ggtitle("SURVIVAL FUNCTION")+
xlab("Time(months)") + ylab("Survival Probabilities") +
theme(plot.title=element_text(color="black",size=20,
family="Comic Sans MS",hjust=0.5),
text=element_text(size = 10, family="Comic Sans MS"),
        axis.text.x=element_text(colour="black", size = 10),
        axis.text.y=element_text(colour="black", size = 10))
exo_s
```

#hazard curve

```
exo_h=ggplot(data.frame(x = c(1, 60)), aes(x = x))+
stat_function(fun = hexpo, size=1.5,col="dodgerblue3")+
theme_bw()+
ggtitle("HAZARD FUNCTION")+
xlab("Time(months)") + ylab("Hazard Rate") +
theme(plot.title = element_text(color="black",size=20,
family="Comic Sans MS",hjust=0.5),
text=element_text(size = 10, family="Comic Sans MS"),
        axis.text.x=element_text(colour="black", size = 10),
        axis.text.y=element_text(colour="black", size = 10))
exo_h
```

The same format is applied for all distributions

QQPlots

#for all the data

```
with(data,qqPlot(time,dist="lnorm",add.line = TRUE,points.col =
"dodgerblue",main="Lognormal Q-Q Plot"))
```

#for censored data

```
with(data, qqPlotCensored(time,censor,dist="lnorm",
censoring.side = "right",estimate.params = TRUE,prob.method =
"kaplan-meier", points.col = "deeppink", add.line = TRUE,
main =paste("Lognormal Q-Q Plot",
"Based on Kaplan-Meier Plotting Positions", sep = "\n")))
```

The same format is applied for all distributions

Goodness Of Fit Test

```
distChooseCensored(time,censor,method="sf",
censoring.side="right",choices = c("gamma","lnorm"))
```