Maximum Likelihood Estimation With Right Censoring

Newcastle University

2018/19

MAS2603

Group 31

Academic Leader: Dr. Joe Matthews

NAME:	*STUDENT ID:*
Harry Ward	170067643
William Seed	170100881
Lingyu Tan	180538368
Huakai Zhang	170382074

Contents

1 Survival Analysis **4**
- 1.1 Survival Function 4
- 1.2 Hazard Function 4

2 Censoring **4**

3 Non-parametric Estimator **5**
- 3.1 Kaplan-Meier estimator 5
- 3.2 Actuarial Method 6

4 Weibull Distribution **7**
- 4.1 PDF, CDF, Hazard function, shape and scale parameters 7
- 4.2 Characteristics of the Weibull Distribution as the shape parameter varies 7

5 Maximum Likelihood estimation **8**

6 Maximum Likelihood Estimator of λ for Weibull Distribution when α is known. **9**

7 MLE of α and λ for Weibull distribution when both parameters are unknown **11**

8 Comparison between different types of distributions in survival analysis **12**
- 8.1 Log-normal Distribution 13
- 8.2 Gamma Distribution 13

9 Other Variables **14**

10 Conclusion **15**

11 References **16**

12 Group Contributions **17**

13 Group Meeting minutes **17**

A R code calculating MLE of α and λ for Weibull distribution when both parameters are unknown **18**

B R code plotting fitted curves for Weibull distribution **20**

C R code calculating MLE of μ and σ for Log-normal distribution when both parameters are unknown **22**

D R code calculating MLE of α and β for Gamma distribution when

Abstract

In this report we will introduce the idea of survival analysis and different concepts that stem from this new, important type of analysis, which is used in so many areas of research and development. We will also get a handle on different types of censoring, how it can impact our analysis and what we can do to work around it. Looking at different types of estimators both parametric and non-parametric will help us understand the data we're given and go on to interpret it correctly. Then we will turn our attention to a crucial distribution, the Weibull distribution and we will see how it relates specifically to survival analysis. The main focus of our report is maximum likelihood estimation with right censoring, we will calculate the MLE for the Weibull distribution, starting off with when α is known and then moving onto when α is unknown. This will allow us to make some clear plots that help us interpret the data very efficiently. We can then compare these parametric methods with the non-parametric methods using plots in R. Our focus will then shift from the Weibull distribution to see if there are any other distributions that may be more effective at modelling survival data.

1 Survival Analysis

We begin by introducing **survival analysis**, this is the analysis of time-to-event data. [1]

Examples of time-to-event data are: the time until a machine part breaks, the time until a terminally ill patient dies, the time until one wins the lottery.

Using survival analysis we can answer important questions such as: what proportion of the population will survive past a certain point? What rate will an event happen? What is the probability an event may happen to a subject right now?

1.1 Survival Function

The **survival function** is given by:

$$S(t) = P(T > t) = 1 - F(T)$$

where $f(t)$ is the probability density function and $F(t)$ is the cumulative distribution function of time-till-event data with time t. The CDF is a monotonically increasing function and so $S(t)$ is monotonically decreasing. The survival function is used in survival analysis far more then the PDF and CDF which is something new for survival analysis. The interpretation of the survival funtion is: the probability a subject will live up to a time t without an event occurring.

1.2 Hazard Function

The **hazard function** is given by:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)}$$

The interpretation of the hazard function is a little bit tricky, it is often thought of as $\Pr(T = t | T \geq t)$ however, time is continuous so this probability will in fact be 0. This is why we have the limit. Also, it does not always have to be less than one so the hazard function cannot be accurately described as a probability, therefore our interpretation is: the risk of an event occurring at this instant given a subject has survived up to this point without the event occuring.

2 Censoring

During a survival analysis, incomplete observations such as the loss of survive time will possibly happen, which we call **censoring**. Vaguely speaking, a censored observation contains only partial information of interest. According to which part of the data is unknown, we have three types of censoring.[2]

- **Right Censoring**: Observation ceases before the event of interest is realised, which means that the survival time is only known to exceed a certain value.
- **Left Censoring**: The event has already occurred before the observation, which means that the data point is below a certain value but it is unknown by how much.
- **Interval Censoring**: The event occurs within an interval. It often happens if the observation is discrete.

In survival analysis, the most typical censoring is right censoring. And according to the cause of censoring, we can classify right censoring into three types.[3]

- **Type I Censoring**: The observation is designed to end at a fixed time. In this case, every object that does not have an event is censored at the fixed time.
- **Type II Censoring**: The observation is designed to end when there have been a prespecified number of events.
- **Random Censoring**: Each subject has a censoring time that is statistically independent of their failure time.

In the dataset we use, the type of censoring is right censoring and random censoring.

3 Non-parametric Estimator

Non-parametric estimation is a statistical method that allows the functional form of a fit to data to be obtained in the absence of any guidance or constraints from theory. As a result, the procedures of non-parametric estimation have no meaningful associated parameters.[4] The reason for us using this method is because estimates and graphs obtained by non-parametric methods can be helpful to us to choose a distribution. There are two different methods to estimate the survival function which is called Kaplan-Meier estimator and actuarial method.

3.1 Kaplan-Meier estimator

Kaplan-Meier estimator is used to estimate a population survival curve from a survival sample. [5] For every individual is followed until death, the curve may be estimated simply by computing the fraction surviving at each time. However, in most studies not all individuals can follow up to the end, so these people's result have been censored (In the graph we just put a mark their to show that). Kaplan-Meier analysis allows estimation of survival over time because it only considered when a specific thing happens (which is died or lost follow in this case).

Here is our Kaplan-Meier survival estimator:

$$S(t_i) = \prod_{t_i \leqslant t} (1 - \frac{d_i}{n_i})$$

For practical computational purposes, the same results can be obtained more efficiently by using the Kaplan-Meier estimator where $S(t_i)$ is the estimated survival probability for any particular one of the t time periods; n_i is the number of subjects at risk at the beginning of time period t_i; and d_i is the number of subjects who ends their survey during time period t_i.

3.2 Actuarial Method

Actuarial method (also called Actuarial estimator) is used to deal with large data set. It can group the time into intervals. The process is similar to the formation of a frequency table and a histogram in elementary statistics. It can also gives an estimate for each probability separately and then multiplies the estimates together to estimate the survival probability.[2]

Compares to Kaplan-Meier estimator, we can find in actuarial method the most important bit is time variate, because it consider each survival percentage at every amount of time on x-axis, whereas in Kaplan-Meier we treated 'event' (for example when some one is died) as the most useful part. Although they have same explanatory variable and response variable, their processing ways are still different with each other.

These are two plots which indicates these two method, figure 1 is our Kaplan-Meier curve for cancer data and figure 2 is an example of an Acturial Method plot.

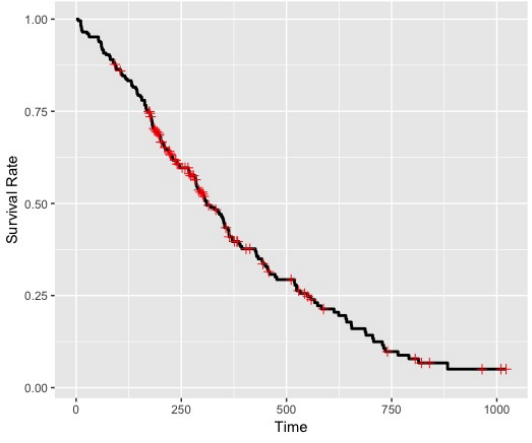


Figure 1: Kaplan-Meier curve for cancer data set

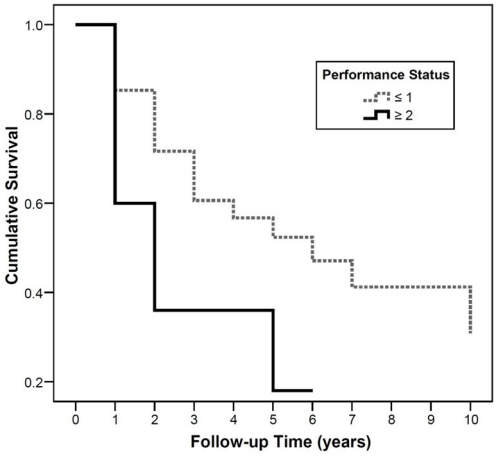


Figure 2: Actuarial method sample graph

4 Weibull Distribution

The **Weibull distribution** is a distribution used in survival analysis, reliability engineering, failure analysis, and a whole host of other areas. It is a continuous probability distribution for random variables which take values in $[0, \infty)$. The distribution is named after the Swedish mathematician Waloddi Weibull. [1]

It is a very useful and versatile distribution because of the fact the Weibull distribution can take on characteristics of other distributions depending on the shape parameter, α.

4.1 PDF, CDF, Hazard function, shape and scale parameters

Probability density function:

$$f(t) = \begin{cases} \alpha\lambda t^{\alpha-1} e^{-(\lambda t^{\alpha})} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

Cumulative distribution function:

$$F(t) = \begin{cases} 1 - e^{-(\lambda t^{\alpha})} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

Survival function:

$$S(t) = e^{-(\lambda t^{\alpha})}$$

Hazard function:

$$h(t) = \alpha\lambda t^{\alpha-1}$$

The parameter α ($\alpha > 0$) is called the shape parameter. The parameter λ ($\lambda > 0$) is called the scale parameter. The probabilty density function shown here can be denoted $W(\lambda, \alpha)$, and this PDF is the parameterization used in medical statistics.

4.2 Characteristics of the Weibull Distribution as the shape parameter varies

As stated earlier, the Weibull distribution can take on characteristics of other distributions as the shape parameter, α changes.

- When $\alpha = 1$, the distribution indicates that the failure rate is constant over time, and the Weibull Distribution reduces to an exponential distribution. Also, the hazard function for the Weibull distribution reduces simply to λ when $\alpha = 1$.
- When $\alpha < 1$, it indicates that the failure rate decreases over time.
- When $\alpha > 1$, it indicates that the failure rate increases with time. This happens if there is an 'aging' process. Also, when $\alpha = 2$, the Weibull distribution reduces to the Rayleigh distribution.

5 Maximum Likelihood estimation

Now that we have covered the background material we can move onto the main focus of our project and that is **maximum likelihood estimation** (MLE). Initially we will find the MLE for different cases of a Weibull distribution but we will eventually move on to explore other distributions such as the Gamma distribution.

In the following sections where we move into MLEs, we will be using a data set throughout, named "cancer", which we found using the package "survival" in R, this data set shows how long patients survived whilst having lung cancer, their gender, and whether their data was censored or uncensored. All of this information for the cancer data will be of much importance into our research into maximum likelihood estimation with right censoring.

As stated in the previous section, the Weibull distribution is a two parameter distribution, which makes it harder to find the maximum likelihood estimates for α and λ, which we discovered quickly. For that reason we will first consider the case where α is known, and then move onto situations when both parameters are unknown, when numerical methods are needed to aid our research.

Our first obstacle to overcome to find the MLE is to alter the likelihood function to account for censoring.

If a patient died at t_i, it's contribution to the likelihood function is the density at that duration, this can be written as the product of the survival and hazard functions.[9]

$$L_i = f(t_i) = h(t_i)S(t_i)$$

If the patient is still alive at t_i, all we know under censoring is that the lifetime exceeds t_i. Therefore the probability of this event is,

$$L_i = S(t_i)$$

This becomes the contribution of a censored observation to the likelihood. Both types of contribution share the survival function, because in both cases the patient

lived up to time t_i. A death multiplies this contribution by the hazard function, but a censored observation doesn't.

We can write the two contributions in a single expression. For this to happen, let δ_i be a death indicator, taking the value one if the patient i died and the value zero otherwise. Then the likelihood function can be written as,

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i)$$

6 Maximum Likelihood Estimator of λ for Weibull Distribution when α is known.

Using the hazard function and survival function for the Weibull distribution we can substitute these into the formula for the likelihood.

$$L = \prod_{i=1}^{n} (\alpha \lambda t_i^{\alpha-1})^{\delta_i} e^{-(\lambda t_i^{\alpha})}$$

Here we take logs of both sides and turn this into a log-likelihood which is denoted by ℓ.

$$\begin{aligned}
\ell &= \sum_{i=1}^{n} \ln[(\alpha \lambda t_i^{\alpha-1})^{\delta_i} e^{-(\lambda t_i^{\alpha})}] \\
&= \sum_{i=1}^{n} \delta_i [\ln \alpha + \ln \lambda + (\alpha - 1) \ln t_i] - \lambda \sum_{i=1}^{n} t_i^{\alpha} \\
&= \ln \alpha \sum_{i=1}^{n} \delta_i + \ln \lambda \sum_{i=1}^{n} \delta_i + (\alpha - 1) \sum_{i=1}^{n} \ln t_i - \lambda \sum_{i=1}^{n} t_i^{\alpha}
\end{aligned}$$

Now, we differentiate with respect to λ.

$$\frac{\mathrm{d}\ell}{\mathrm{d}\lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} t_i^{\alpha}$$

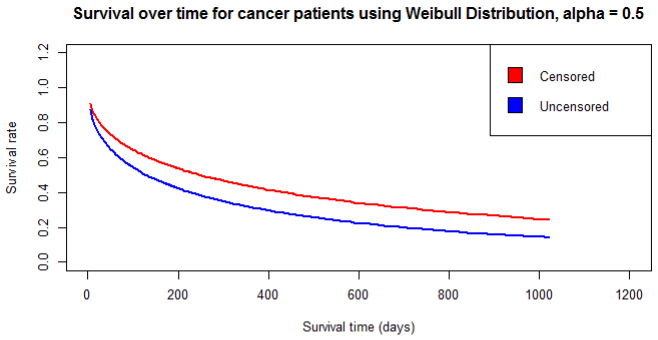
Now, we set $\frac{\mathrm{d}\ell}{\mathrm{d}\lambda}$ equal to zero and replace λ with $\hat{\lambda}$. We then solve for $\hat{\lambda}$ to find the maximum.

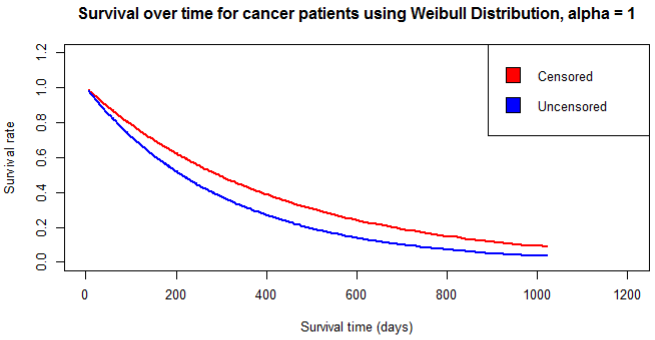
$$\begin{aligned}
0 &= \frac{1}{\hat{\lambda}} \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} t_i^{\alpha} \\
\hat{\lambda} &= \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i^{\alpha}}
\end{aligned}$$

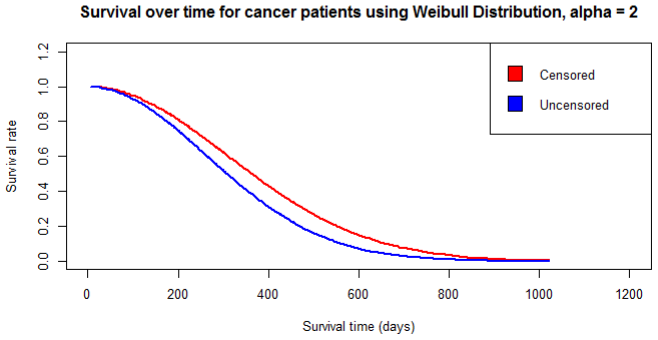
When using the cancer data, we can estimate what $\hat{\lambda}$ is by using different values of α, using the survival times of patients and by using the amount of patients who were uncensored.

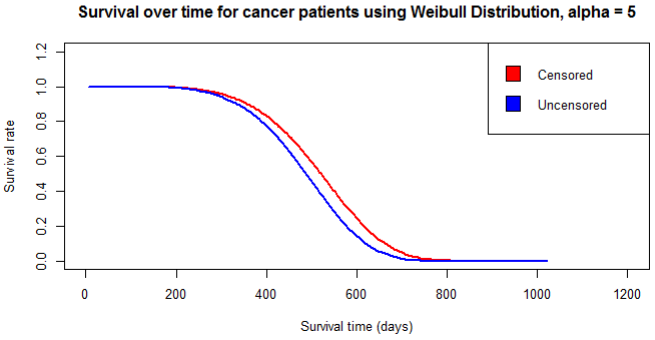
α	$\hat{\lambda}$ when censored	$\hat{\lambda}$ when uncensored
0.5	0.044	0.061
1	0.0024	0.0034
2	0.0000053	0.0000073
5	1.81×10^{-14}	2.51×10^{-14}

From the table, we can see how $\hat{\lambda}$ changes as we vary α, it is clear to see that as α increases, $\hat{\lambda}$ decreases rapidly, the reason for this is from the equation for $\hat{\lambda}$, in the denominator for the equation, we have the sum of all times each to the power alpha, this will give massive numbers because our survival times from the cancer data range from 0 to just over 1000. Also, for the same value of α, the censored value for $\hat{\lambda}$ is lower than the uncensored value. This is further shown in the graphs below.









The graphs above show simulations of the Weibull survival function with varying shape parameter and scale parameter, $\hat{\lambda}$, calculated from the formula for $\hat{\lambda}$ in this section. From the graphs, there are a few characteristics of the survival curve which are easy to see, as time tends to infinity, the survival rate tends to zero. However the time taken to reach this survival time depends on the parameters which we use and calculate, for example as α increases, the maximum survival time drops. Also, you can see the difference between censored data and data which assumes everything is

uncensored. The censored survival curve is always above the uncensored line, meaning that when the data is censored, the assumption is that more people will be alive at any point in time, compared to when all the data is assumed to be uncensored. The reason for this is because when we treat all observations as uncensored, we assume that the time given is when they died, however we know from censoring that they could have survived for a longer time than this, therefore the censored times will always be longer than or equal to the uncensored times, leading to an increased rate of survival at any point in time.

7 MLE of α and λ for Weibull distribution when both parameters are unknown

From the previous section, we already know the log-likelihood functions for Weibull distribution:

$$\ell = \ln\alpha \sum_{i=1}^{n} \delta_i + \ln\lambda \sum_{i=1}^{n} \delta_i + (\alpha - 1) \sum_{i=1}^{n} \ln t_i - \lambda \sum_{i=1}^{n} t_i^{\alpha} \quad \text{(General form)}$$

$$\ell = n\ln\alpha + n\ln\lambda + (\alpha - 1) \sum_{i=1}^{n} \ln t_i - \lambda \sum_{i=1}^{n} t_i^{\alpha} \quad \text{(Ignore censoring)}$$

In order to find the MLE of α and λ when both of them are unknown, we need to use some numerical optimization methods to maximize these two log-likelihood functions.

In the R {Stats} package, we found a function called nlm, that is, non-linear minimization. We can use it to find the value of α and λ when $-\ell$ reaches its minimum(i.e. when ℓ reaches its maximum).

Using the R code in Appendix A, we can calculate the MLE of α and λ as below:

Case	α	λ
Censored	1.313612382	0.000359415
Uncensored	1.4624512710	0.0002006867

Table 1: MLE of α and λ for Weibull distribution

To test the accuracy of our estimators, we need to introduce a concept: the empirical distribution function (ecdf), which is defined below:[6]

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(X_i<x)}$$

where X_i is a sample from population X with common distribution function $F(x)$, $I_{(X_i<x)}$ is the indicator function. Informally speaking, ecdf is just like cumulative

distribution function (cdf); however, cdf is for a kind of distribution, ecdf is based on a specific dataset.

Then we are going to verify if the estimators are efficient using the two criterions below:

- The censored curve is supposed to fit the Kaplan-Meier curve.
- The uncensored curve is supposed to fit the empirical distribution function(ecdf) of survival time. (All the ecdf we discuss in this article have been flipped vertically)

Here is the figure including Kaplan-Meier curve, ecdf, censored and uncensored Weibull distribution curve: (For R code, see Appendix B)

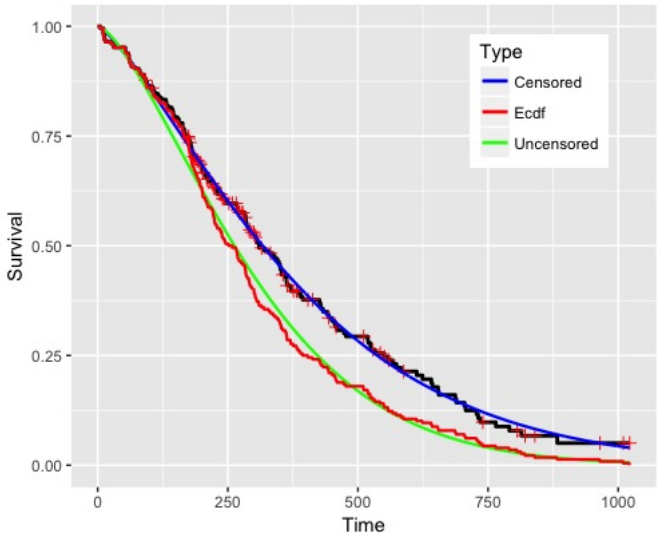


Figure 3: Fitting result using Weibull distribution

As we can see, the censored curve fits the Kaplan-Meier curve well, and the uncensored curve also fits the ecdf well. Therefore, Weibull distribution behaves perfectly in this survival analysis, and the MLE of α and λ obtained work very well.

8 Comparison between different types of distributions in survival analysis

Besides Weibull distribution, there are also some other distributions can be use to estimate survival analysis as well. Generalized Gamma distribution is an extensive family that contains nearly all of the most commonly distributions to itself, including exponential, Weibull,log-normal and Gamma. We will compare Weibull distribution to log-normal and Gamma distribution by using R. Our R code is in Appendix C and D.

8.1 Log-normal Distribution

Log-normal distribution is a continuous probability distribution of a random variable which logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then Y has a normal distribution which can write as $ln(x) \sim (\mu, \sigma^2)$.[11]

We already know that our log-normal CDF is $\frac{1}{2} + \frac{1}{2}erf\left[\frac{ln\left(x\right)-\mu}{\sigma\sqrt{2}}\right]^2$, so we use this to plot our survival data in R (Appendix C).If we only consider censoring, we have a result of (μ=5.65, σ=1.08).In figure 3 the blue line is our log-normal curve and the black line is the original Kaplan-Meier curve. We can see that log-normal distribution is not very fit our result because it looks very different. In figure 4 we then ignore censoring we have a result of (μ=5.42, σ=0.91) with green line represent ignore censoring log-normal curve and red curve represent our ECDF. From this two plots we can see that no matter whether we consider censoring, log-normal curve doesn't feat our data curve. So after this comparison we can conclude that Weibull is far more better than log-normal in dealing with this survival data.

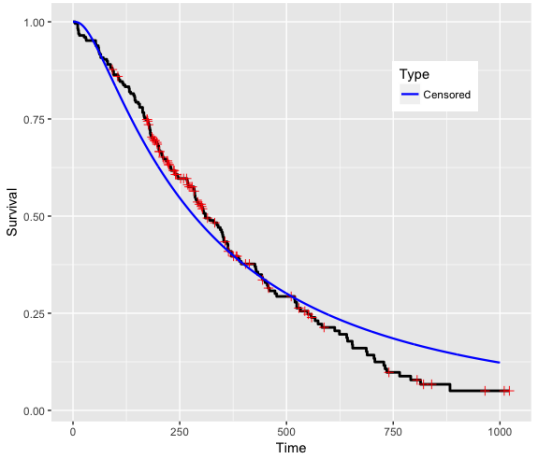


Figure 4: Log-normal result when consider censoring

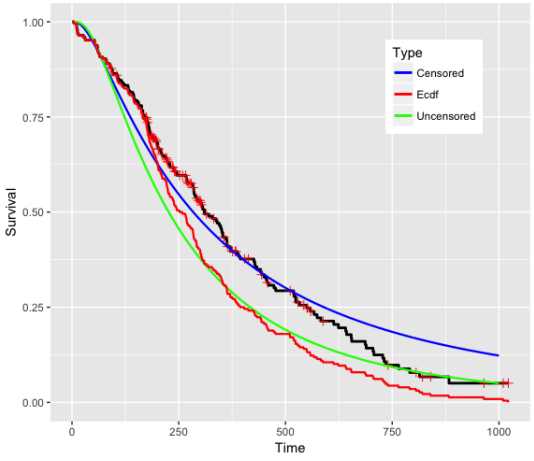


Figure 5: Log-normal result when ignore censoring

8.2 Gamma Distribution

In statistics, the Gamma distribution is a two-parameter family of continuous probability distributions With a shape parameter α and a scale parameter β. So we write $x \sim Ga(\alpha, \beta)$.[10].

In dealing with Gamma, we do it in R (Appendix D) and finally we have our result. In consider censoring we have (α=1.42 , β=0.0036), when we consider censoring, Gamma curve seems quite fit our Kaplan-Meier curve except the tail. Then ignore censoring and estimators are (α=1.82, β=0.006). Overall from these two plots, we

determine that Gamma distribution fits very well if we ignore censoring, compares to Weibull distribution, Gamma is as good as Weibull. However if we consider censoring, we will see that Gamma distribution fits alright except the tail. And if we compares it with Weibull, it is good enough but not as good as Weibull.

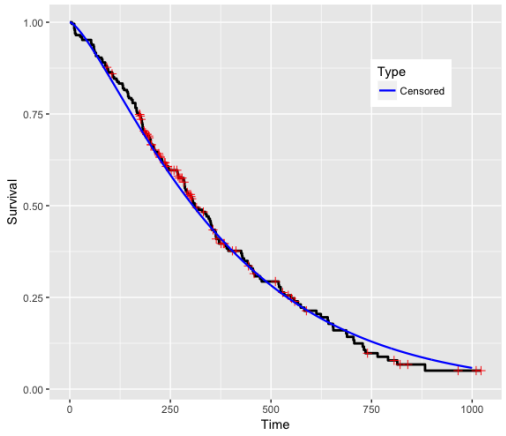


Figure 6: Gamma result when consider censoring

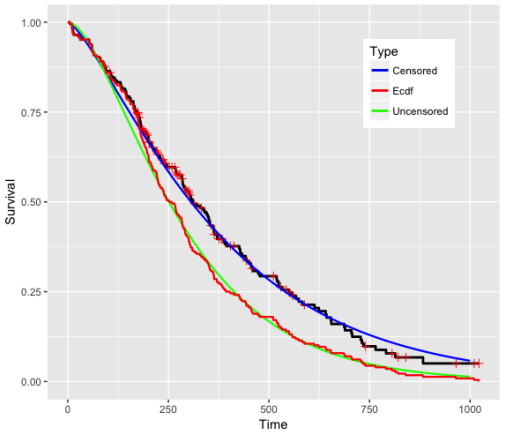


Figure 7: Gamma result when ignore censoring

9 Other Variables

There are many things that may impact a groups survival rate, the initial one we may consider is sex. Alternatively we could alter the variables ourselves, so we could have two groups of people on different treatments and see how each of the treatments perform. This type of analysis is vital in the field of medical study and has helped develop many drug treatments. Using R we can plot a Kaplan-Meier curve for males against females (Figure 8), on the same plot we have the p-value of the **log-rank test**, simply put this tells us if there is a significant difference between the two variables. The null hypothesis is that there is no difference in survival between the two groups. However, the log-rank test only works for categorical variables, so we can compare male vs female, treatment A vs treatment B but we cannot compare quantitative variables such as age, weight, height or gene expression. The p-value of the male vs female plot is 0.0013 which is significant even at the 0.5% level. This tells us that there is a significant difference between males and females with regards to survival rate of this specific type of cancer.

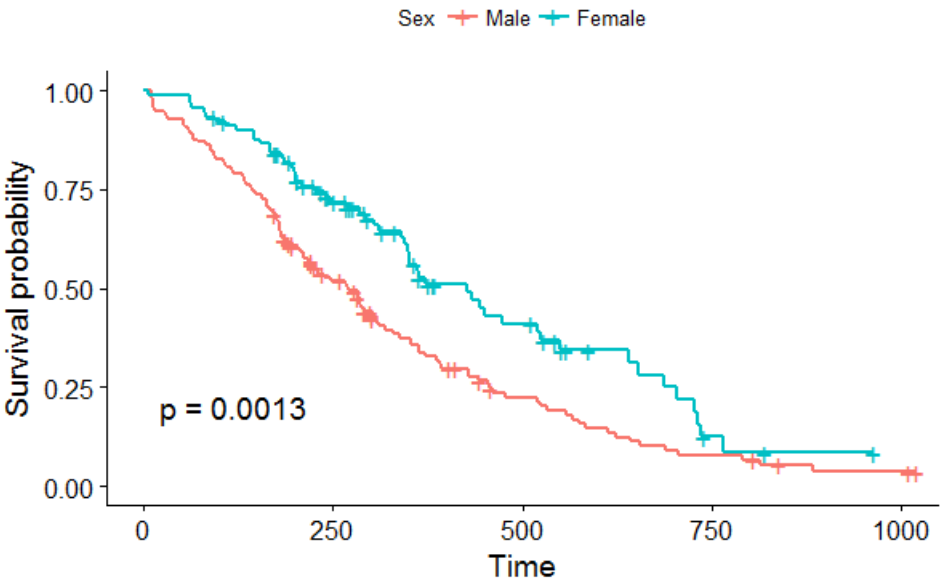


Figure 8: Male vs Female Kaplan-Meier plot

When we do have quantitative variables we must use another method such as the **Cox proportional hazards regression analysis**, which will work for both types of variables.[1] It has the additional benefit of comparing several different variables that effect survival rate at the same time. This is incredibly important as we do not want to be in a position where we're not sure which variable is causing the difference in survival of a group. For example, we may be comparing treatment A vs treatment B and we see a significant difference but the people who had treatment A were considerably older than those with treatment B, this could be the reason Treatment B seemed to perform better. This type of analysis will account for that.

10 Conclusion

Overall, we found that Weibull distribution was very well suited to survival analysis, which was seen throughout in our plots comparing the Weibull survival curves to Kaplan-Meier curves for example. We were successfully able to find the MLE for multiple distributions throughout the project as well, looking into different situations for each distribution, for example, when censoring was occurring or when we ignored censoring and treated everything as though it was uncensored.

Censoring was a large part of this project, and through calculating maximum likelihood estimates in different situations, and creating plots from distributions used in survival analysis, we were able to see and understand the differences in real life situations when we observed censoring. We saw the importance of taking censoring into account, because it gave a more realistic view into how long patients were likely to survive, which is vital for people and companies who are involved with survival times.

11 References

References

[1] Collett, D. *Modelling Survival Data in Medical Research.* Third ed. 2015. Print. Chapman & Hall/CRC Texts in Statistical Science Ser.

[2] Miller, Rupert G., Gail. Gong, and Alvaro. Muñoz. *Survival Analysis.* New York: Wiley, 1981. Print. Wiley Ser. in Probability and Mathematical Statistics.

[3] Parmar, Mahesh K.B., and David. Machin. *Survival Analysis : A Practical Approach.* Chichester, Eng. ; New York: J. Wiley, 1995. Print.

[4] Nonparametric Estimation (2019, May 6). Nonparametric Estimation. In Wolfram Mathoworld, Retrieved May 12, 2019, from `http://mathworld.wolfram.com/NonparametricEstimation.html`

[5] Le, Chap T. *Applied Survival Analysis.* New York: J. Wiley, 1997. Print. Wiley Ser. in Probability and Mathematical Statistics. Applied Probability and Statistics.

[6] Hu, Gaoge, Shesheng Gao, Yongmin Zhong, and Chengfan Gu. "Asymptotic Properties of Random Weighted Empirical Distribution Function." *Communications in Statistics: Theory and Methods* 44.18 (2015): 3812-824. Web.

[7] Anon, (2019). [online] Available at: `https://vassarstats.net/survival.html`

[8] Publishing, R. (2019). Weibull Distribution: Characteristics of the Weibull Distribution. [online] Weibull.com. Available at: `https://www.weibull.com/hotwire/issue14/relbasics14.htm`

[9] Data.princeton.edu. (2019). GR's Website. [online] Available at: `https://data.princeton.edu/wws509/notes/c7s2`

[10] (2019, May 8). Wikidata. Retrieved 23:11, May 12, 2019 `https://www.wikidata.org/w/index.php?title=Q117806&oldid=936906316.`

[11] Wikipedia contributors. (2019, April 19). Log-normal distribution. In Wikipedia, The Free Encyclopedia. Retrieved 23:09, May 12, 2019 from `https://en.wikipedia.org/w/index.php?title=Log-normal_distribution&oldid=893127110`

12 Group Contributions

- Will Seed: In this report, Will wrote the abstract, section 1, Survival Analysis and section 9, Other Variables .
- Huakai Zhang: In this report, Huakai wrote section 3, 8, 10.
- Lingyu Tan: In this report, Lingyu wrote section 2, 7 and all the appendices.
- Harry Ward: In this report, Harry wrote the section 4, 5, 6 and the conclusion.

13 Group Meeting minutes

Time	Present	Contents
06/03/2019 10:45-12:30	Everyone	How to plot survival function in and find maximum likelihood function for Weibull distribution
12/03/2019 10:45-12:00	Everyone	Derive the MLE for censor data and find some information for how to use R to estimate Weibull distribution
20/03/2019 11:00-13:10	Everyone	Working out the data. Plotting data "Cancer" into Kaplan-Meier Curve and Weibull distribution. Also we worked for the MLE of λ and α
01/05/2019 14:30-16:00	Everyone	Trying to work out the PowerPoint and the final project.
07/05/2019 10:00-11:00	Everyone	Working together to present our presentation to see how it goes.

A R code calculating MLE of α and λ for Weibull distribution when both parameters are unknown

```
if(!exists("cancer")) library(survival)

data("cancer")

timex <- cancer$time

#------------Censored or not------------#
# Considering Censoring
# statusx <- cancer$status - 1

# Ignoring Censoring
statusx <- rep(1, length(timex))
#---------------------------------------#

ll = function(alph){
  sum = 0
  for (i in 1 : 228)
  {
    sum = sum - (statusx[i]*(log(alph[1]) + log(alph[2]) +
    (alph[1]-1) * log(timex[i])) - alph[2]*(timex[i]^alph[1]))
  }
  return(sum)
}

nlm(ll, alph <- c(2,2), hessian = TRUE)
```

The censored results are below:

```
$minimum
[1] 1153.853

$estimate
[1] 1.313612382 0.000359415

$gradient
[1]       -4.050041 -14069.405308

$hessian
             [,1]       [,2]
[1,]     6184.405    2771687
[2,] 2771686.619  800886503

$code
[1] 3
```

```
$iterations
[1] 84
```

Comments:

By changing Part "Censored or not", we can obtain the MLE of α and λ when censored and uncensored in the output object "$estimate", in which the first value is for α and the second is for λ.

B R code plotting fitted curves for Weibull distribution

```
if(!exists("ggplot")) library(ggplot2)
if(!exists("survfit")) library(survival)
if(!exists("ggsurv")) library(GGally)

data(cancer)

sf.cancer <- survfit(Surv(time, status) ~ 1, data = cancer)

# Considering censoring
survfun1 = function(x)
{
  lambda = 0.000359415
  alpha = 1.313612382
  y <- exp(-lambda*((x)^alpha))
  d <- data.frame(x = x, y = y)
  return(d)
}
x1 <- seq(0, 1022, 1)
d1 <- survfun1(x1)

# Ignoring censoring
survfun2 = function(x)
{
  lambda <- 0.0002006867
  alpha <- 1.4624512710
  y <- exp(-lambda*((x)^alpha))
  d <- data.frame(x = x, y = y)
  return(d)
}
x2 <- seq(0, 1022, 1)
d2 <- survfun2(x2)

# Ecdf for survival time
survfun3 = function(x)
{
  res <- ecdf(cancer$time)
  y <- 1 - res(x)
  d <- data.frame(x = x, y = y)
  return(d)
}
x3 <- seq(0, 1022, 1)
d3 <- survfun3(x3)

# Plot figure
```

```
p <- ggsurv(sf.cancer, CI = T, surv.col = 'black', cens.col = 'red',
size.est = 1, size.ci = 0) +
geom_line(data = d1, aes(x, y, colour = "Censored"), size = 0.8) +
geom_line(data = d2, aes(x, y, colour = "Uncensored"), size = 0.8) +
geom_line(data = d3, aes(x, y, colour = "Ecdf"), size = 0.8) +
scale_color_manual(name = "Type", values = c("Censored" = "blue",
"Uncensored" = "green", "Ecdf" = "red")) +
theme(legend.position = c(0.8, 0.8))
```

Comments:

The figures have already been given in the article. The code above is for Weibull distribution. To obtain the figures for other distributions, we only need to change the function "survfun1" and "survfun2" using corresponding distributions, and all other lines are almost the same. Therefore, considering the length of the appendices, we decided not to put that for Log-normal and Gamma distributions on.

C R code calculating MLE of μ and σ for Log-normal distribution when both parameters are unknown

```
if(!exists("survreg")) library(survival)

x <- cancer$time
r <- length(x)

xs <- sort(x)

#-----------Censored or not-----------#

# Considering censoring  /  is.optional(lung$status - 1)
statusx <- lung$status

# Ignoring censoring
# statusx <- rep(1, r)

#-------------------------------------#

dat.weibull <- data.frame(xs, statusx)

names(dat.weibull) <- c("time", "status")

out.weibull <- survreg(Surv(time, status) ~ 1, dist = "lognormal",
    data = dat.weibull)

meanlog <- out.weibull$coef
sdlog <- out.weibull$scale

parms <- c(meanlog, sdlog)
names(parms) <- c("meanlog", "sdlog")

list(mles = parms)
```

The censored results are below:

```
$mles
 meanlog     sdlog
5.648811  1.084365
```

Comments:

By changing Part "Censored or not", we can obtain the MLE of μ and σ when censored and uncensored in the output object"mles" listed.

D R code calculating MLE of α and β for Gamma distribution when both parameters are unknown

```
if(!exists("cancer")) library(survival)
if(!exists("flexsurvreg")) library(flexsurv)

x <- cancer$time
r <- length(x)

xs <- sort(x)

#-----------Censored or not-----------#

# Considering censoring  /  is.optional(lung$status - 1)
statusx <- lung$status

# Ignoring censoring
# statusx <- rep(1, r)

#-------------------------------------#

dat.weibull <- data.frame(xs, statusx)

names(dat.weibull) <- c("time", "status")

flexsurvreg(Surv(time, status) ~ 1, dist = "gamma", data = dat.weibull)
```

The censored results are below:

```
Call:
flexsurvreg(formula = Surv(time, status) ~ 1, data = dat.weibull,
    dist = "gamma")

Estimates:
       est        L95%       U95%       se
shape  1.422406   1.174317   1.722907   0.139096
rate   0.003614   0.002795   0.004674   0.000474

N = 228,  Events: 165,  Censored: 63
Total time at risk: 69593
Log-likelihood = -1156.364, df = 2
AIC = 2316.727
```

Comments:

By changing Part "Censored or not", we can obtain the MLE of α and β when censored and uncensored in the column "est" under the output object "Estimates".