# Introducing parmsurvfit Package - Simple Parametric Survival Analysis with R

*by Victor Wilson, Ashley Jacobson, Shannon Pileggi*

**Abstract** This article introduces the R package parmsurvfit, which executes basic parametric survival analysis techniques similar to those in Minitab. Among these are plotting hazard, cumulative hazard, survival, and density curves, computing survival probabilites, and computing summary statistics based on a specified parametric distribution. We describe appropriate usage of these functions, interpretation of output, and provide examples of how to utilize these functions in real-world datasets.

## Introduction

Survival analysis is a branch of statistics that primarily deals with analyzing the time until an event of interest occurs. This event could be a variety of different things such as death, development of disease, or first score of a soccer game. Observations in survival analysis may also be desrcibed as censored, which occurs when an observation's survival time is incomplete. The most common way that this occurs is through right censoring, which occurs when a subject does not experience the event of interest within the duration of the study. Right censoring can also occur if a subject drops out before the end of the study and does not experience the event of interest. Due to the inherent issue of censoring that is typically found in datasets involving survival analysis, computations and analyses can be difficult to carry out with many standard functions avaiable in R, as the majority of these do not account for censored data. The censored data here is of value and we cannot merely eliminate the observations which have censored data.

Some of the most popular techniques and statistics utilized when carrying out a survival analysis are computing what are known as the survival and hazard functions. The survival function is important because it gives the proability of surviving (also known as not experiencing the event of interest) beyond any given time $t$. Similarly, the hazard function is also useful to compute because it gives the conditional probability that the subject will experience the event in the next instance of time, given that they have survived up until the specified point in time. Other popular statistics that are utilized are median survial time, mean survival time, and percentiles of survival time. In this package, all of the functions that we developed utilize parametric methods of survival anlaysis, which assumes that the distribution of the survival times follows a known probability distribution.

Currently, R does have many survival packages that address non-parametic survival analysis, such as the **survival** package. Moreover, R does have some packages that aid in estimation for parametric survival analysis, including **fitdistrplus**. However, Minitab has very concise and easy to utilize functions for computing and displaying many parametric survival statistics and plots, but this same output is not readily available in any single one package in R, or in some cases not available at all. Thus, we decided to develop a package that emulates the output found in Minitab for parametric survival analysis, which contains all of these commonly utilized statistics and plots.

This paper describes the functions that the **parmsurvfit** package contains, how the data is formatted in order to utilize these functions, and what the output of these functions represent. There are four major groups of functions that we created: fitting the censored data, displaying plots (density, hazard, cumulative hazard, and survival), computing statistics (mean, median, survival probabilities), and assessing fit (qqplot, Anderson Darling statistic). The majority of this paper will be organized following these groups of functions.

## Fitting Right Censored Survival Data

As mentoined previously, this function is very similar to the function fitdistcens found in the fitdistrplus package, which computes the Maximum Likelihood Estimates (MLEs) for right-censored data. **Is the data organized differently than required for fitdistcens? Explain the output**

## Example

```
#> Fitting of the distribution ' logis ' on censored data by maximum likelihood
#> Parameters:
#>           estimate
#> location 16.741581
```

```
#> scale     2.798533
#> Fixed parameters:
#> data frame with 0 columns and 0 rows
```

### Displaying Plots

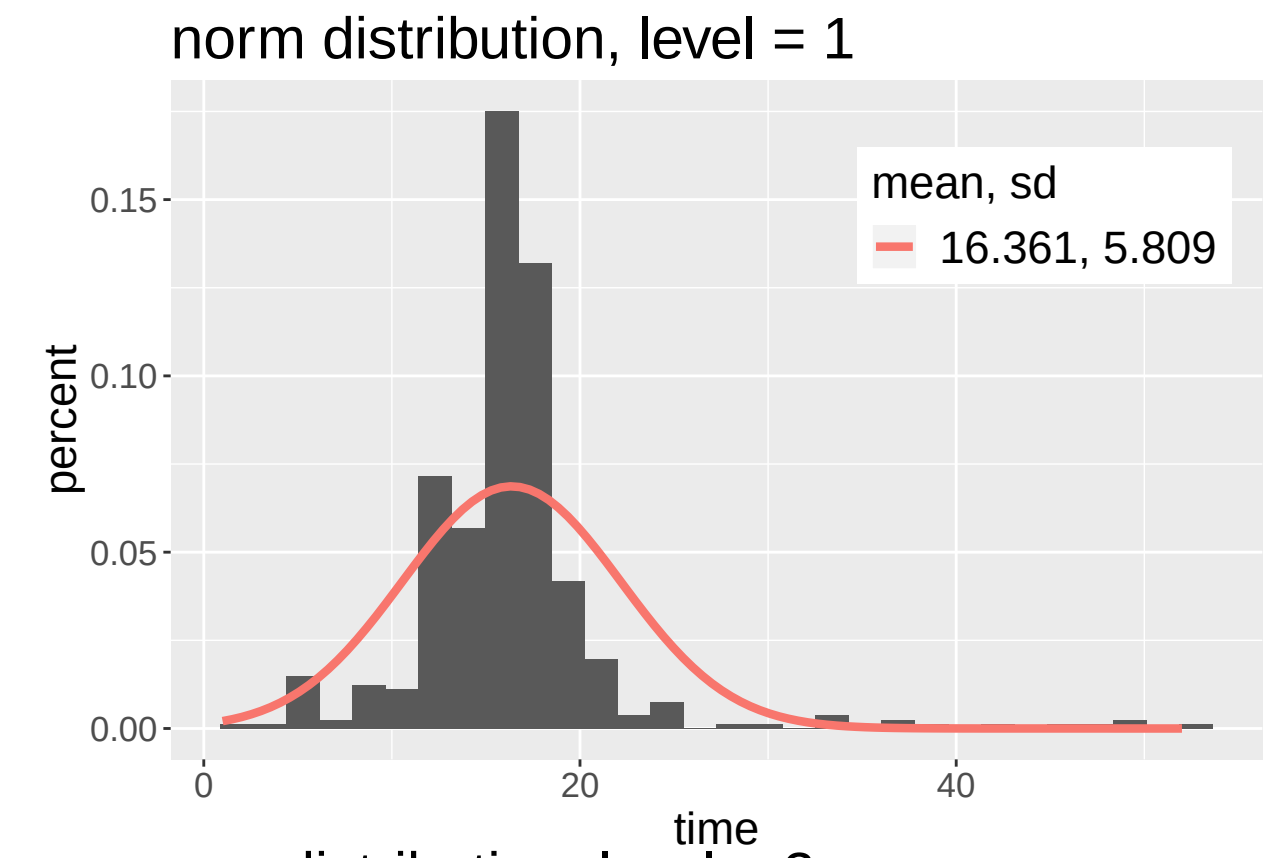
This section introduces an overview of the many types of plots that are available to be displayed via this package. Some of the most common plots used in Survival Analysis are survival plots, hazard plots, probability density plots, and cumulative hazard plots. We designed these functions with an intent to have the output displayed be very easy to read and interpret. Below is a list of each function and it's relationship to other functions, as well as the forumla used to compute each function.

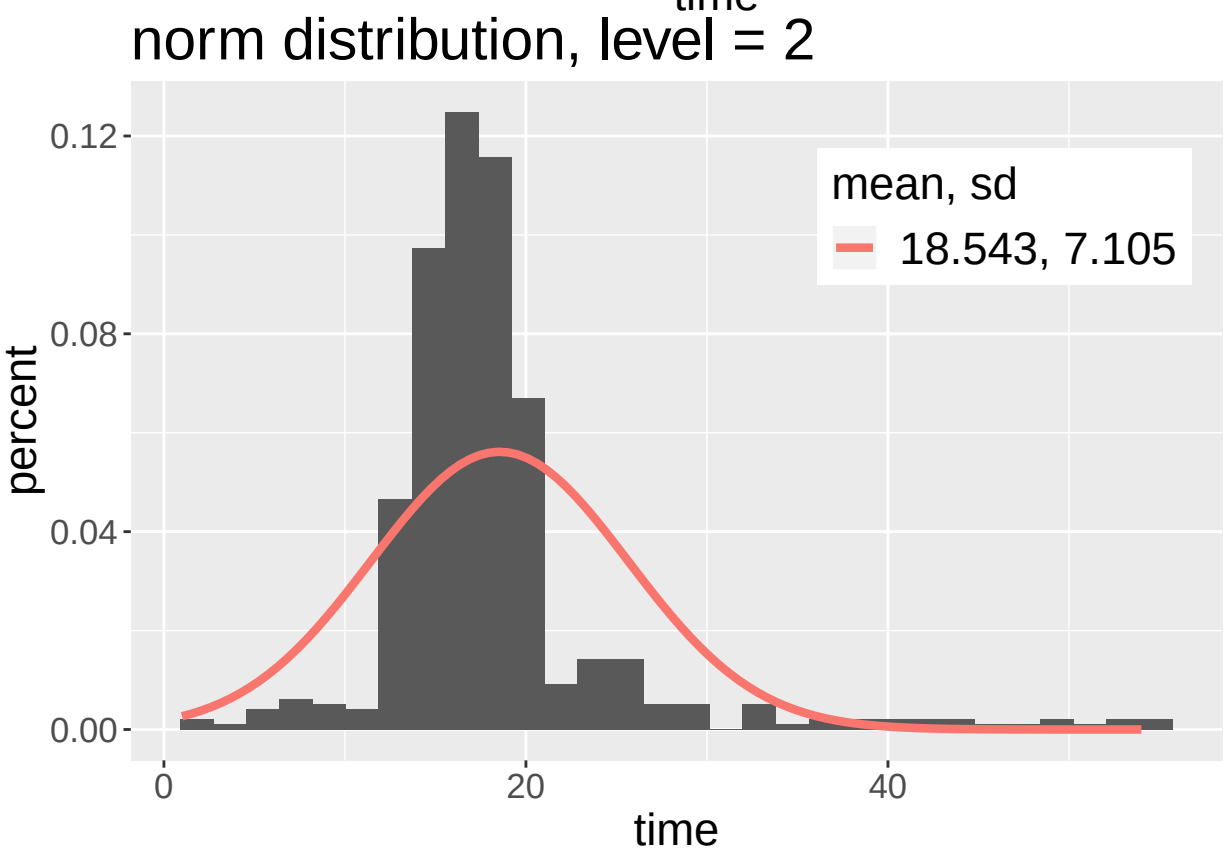
| Function | Relationships |
|---|---|
| PDF | $f(t) = \frac{d}{dt}F(t)$ |
| CDF | $F(t) = \int_0^t f(y)dy$ |
| Survival | $S(t) = 1 - F(t) = \exp[-H(t)] = \exp[-\int_0^t h(y)dy]$ |
| Hazard | $h(t) = f(t)/S(t) = -\frac{d}{dt}\ln[S(t)]$ |
| Cum. Haz. | $H(t) = \int_0^t h(y)dy = -\ln[S(t)]$ |

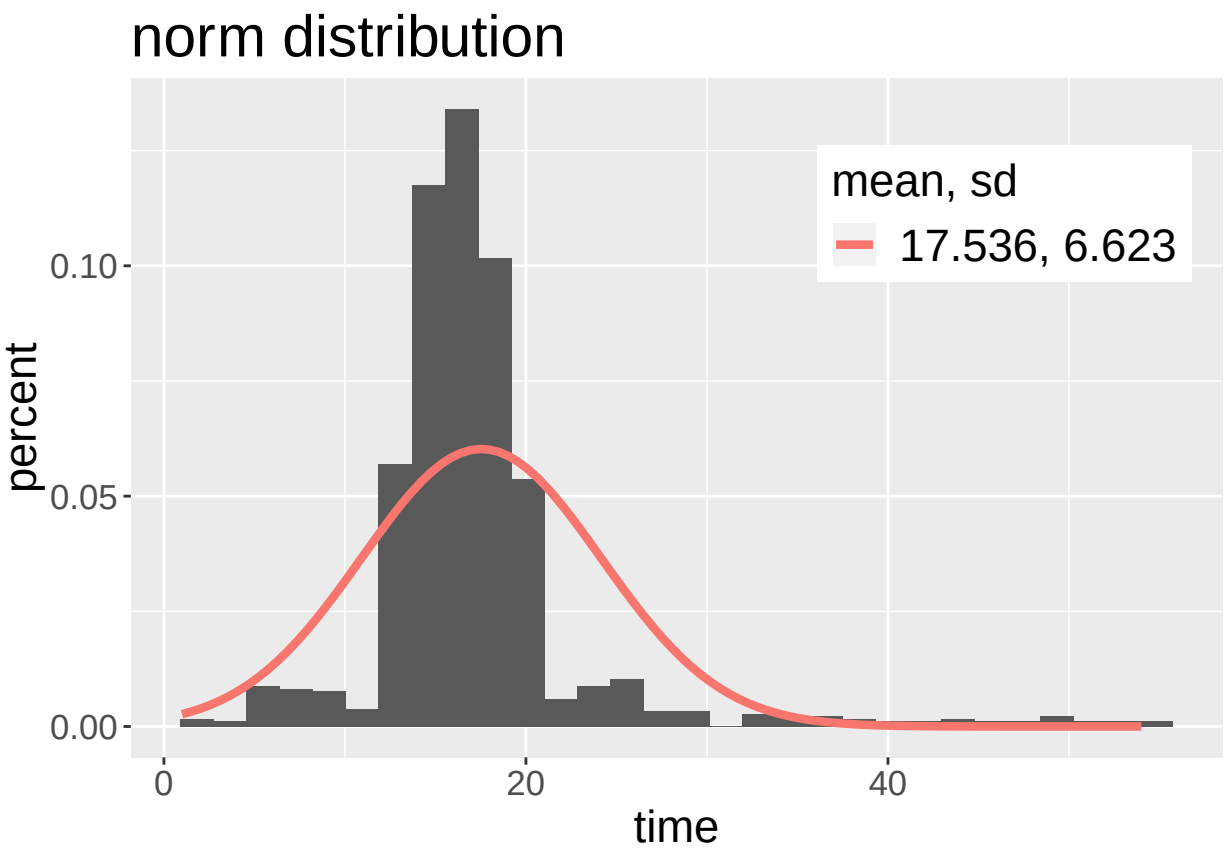
### Density Plots/histograms

The `plot_density` function creates a histogram of the data and overlays the density function of a fitted parametric distribution. Parameters estimates for the specified parametric distribution are provided as well. This function also supports the ability to plot seperate histograms and density functions for each level of a grouping variable. An example of this function is shown below:

```
plot_density(Firstdrink, "norm", time = "Age", by = "Gender")
```


norm distribution, level = 1
mean, sd
16.361, 5.809
percent
0.15
0.10
0.05
0.00
0
20
40
time

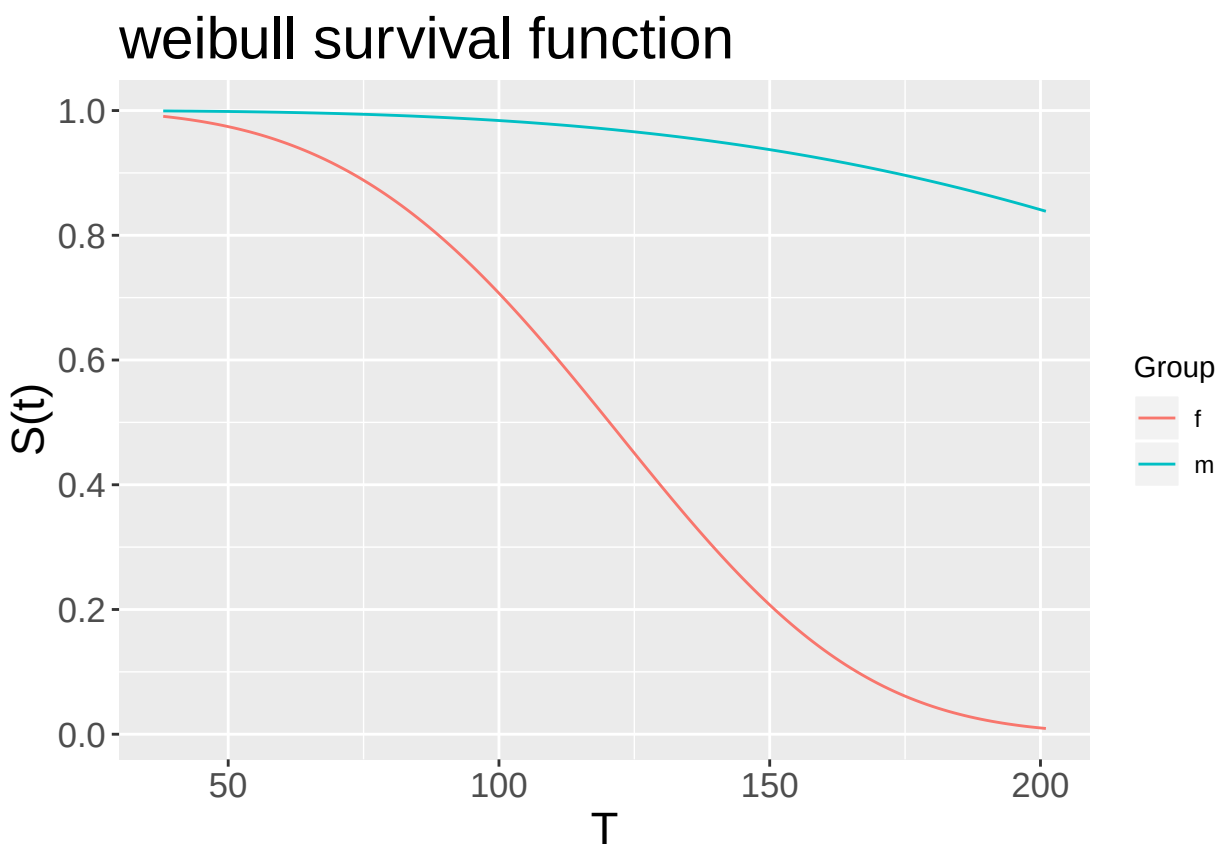

norm distribution, level = 2
mean, sd
18.543, 7.105
percent
0.12
0.08
0.04
0.00
0
20
40
time




We ran the `plot_density` function, utilizing the 'Firstdrink.txt' dataset available in our package. This dataset contains data on the age of first consumption of alcholic beverage for 1000 indiviuals. As seen above, a seperate histogram and density plot was created for males and females.

### Survival Plots

Survival plots are used to estimate the proportion of subjects that survive beyond a specified time **t**. We were motivated to create the function `plot_surv` in an attempt to create hazard plots that are easy to produce, when dealing with data set up for Survival Analysis. This function plots the survival curve of right censored data given that it follows a specified parametric distribution. Some examples of the distributions that this function supports are the Weibull, Log-Normal, Exponential, Normal, and Logistic distributions. This function also provides the option to plot by a grouping variable, which if specified, displays separate curves for each group of the specified variable.




In this example, we fit a Weibull distribution to the rats dataset available in the **survival** package, grouping by the "sex" variable. The rats dataset contains 300 observations, with 3 rats each being selected from 100 litters and 1 rat in each litter being adminstered a drug. The event of interest in this study was whether or not a rat developed a tumor following the beginning of the study. For each rat, the litter number (1-100), type of treatment received (coded as 1 = drug, 0 = control), time until development of tumor or last follow-up (measured in days), final event status (1 = tumor, 0 = censored), and sex were recorded. Below is an excerpt of the data frame.

```
library(parmsurvfit)
data(rats)
head(rats)
```

```
#>   litter rx time status sex
#> 1      1  1  101      0   f
#> 2      1  0   49      1   f
#> 3      1  0  104      0   f
#> 4      2  1   91      0   m
#> 5      2  0  104      0   m
#> 6      2  0  102      0   m
```

For example, the rat represented in line 2 came from Litter 1, did not receive the drug, experienced the development of a tumor at 49 days, and was a female rat.

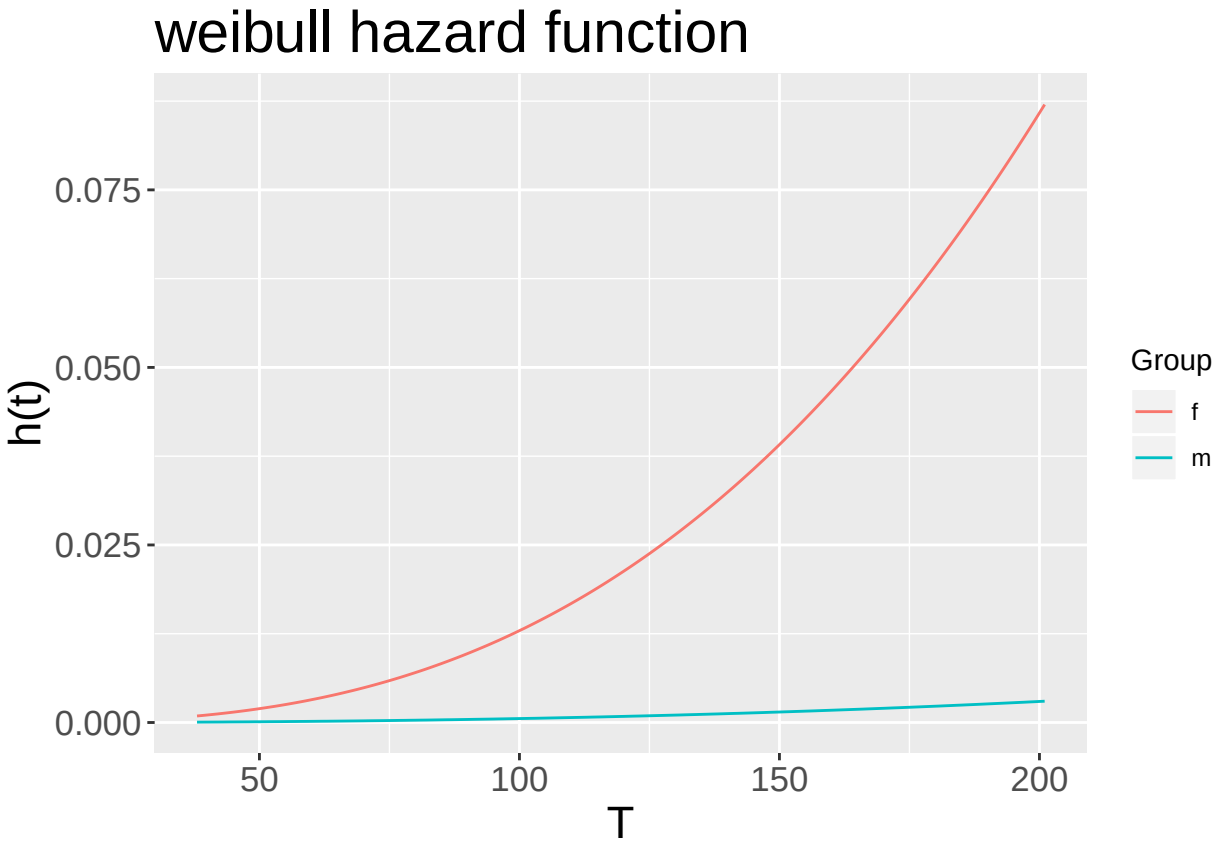
As seen in the plot above, two different survival curves were plotted. The blue line represents the estimated survival curve for male rats, while the red line represents the estimated survival curve for female rats. From this plot, we see that the survival curve for male rats is consistently above the survival curve for female rats throughout all points in time. Due to this, we can conclude that female rats tend to experience the event of interest much more quickly than male rats. This can also be interpreted as male rats tend to survive longer than female rats in this study.

### Hazard plots

Hazard plots, on the other hand, are used to display the conditional risk that a subject will experience the event of interest in the next instant of time, given that the subject has survived beyond a certain amount of time. Essentially, the hazard function attempts to assess the risk that an individual who has

not yet experienced the event in the very next small amount of time. For example, if we observe that a rat has survived for 75 days already, the hazard function would estimate the risk that the rat will die in the next short instant of time, based on the fact that it has already survived 75 days. We created the `plot_haz` function in order to easily plot hazard functions given that it follows a specified parametric distribution, with the option to include a grouping variable.

**INSERT HAZARD FUNCTION FORMULA**

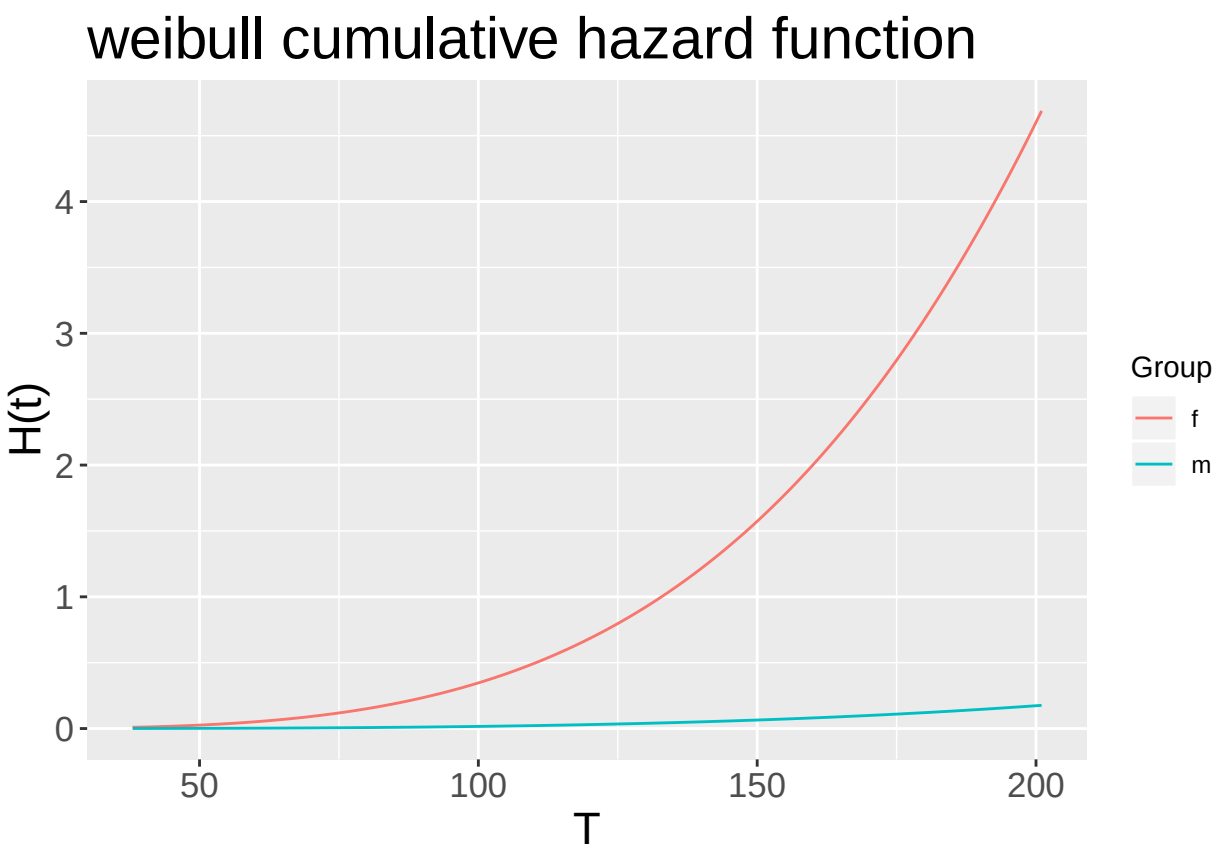



From this plot above, also using the rats dataset available in the survival package, we can see that as female rats continue to survive, their risk of experiencing the event of interest in the next instant of time dramatically increases. Contrastly, male rats do not seem to have a greater risk of experiencing the event of interest as they survive longer. This is demonstrated by the blue line being mostly flat across all points of time.

### Cumulative Hazard Plots

While hazard plots are usually useful in assessing a subject's risk of experiencing the event of interest in the next moment of time, these plots can be difficult to read and understand at times. Sometimes, the changes in hazard are very subtle, making it difficult to describe periods of increasing and decreasing risk. In order to accurately assess how hazard rates change over time, we investigate the accumulation of hazard rates over time, known as cumulative hazard. The cumulative hazard function, denoted **H(t)**, is the accumulated risk of experiencing an event up to time **t**. Since the cumulative hazard function is an accumulation of rates, it is important to note that this function is non-decreasing and is hardly ever remains constant by nature. We developed the function `plot_cumhaz` in order to easily display cumulative hazard plots, given that the data follows a specified parametric distribution. The functionality of this function is nearly identical to that of `plot_haz`, with the only distinction being that it plots cumulative hazard curves instead of hazard curves.

**Insert brief interpretation**




### Computing Survival Probabilities and Summary Statistics

While viewing plots such as those explained above are very useful in survival analysis, they only tell half of the story. In order to carry out a complete survival analysis, we must also compute statistics in order to supplement our plots. Some of the most common statistics utilized in parametric survival analysis are survival probabilities and typical summary statistics such as the mean, median, standard deviation, and percentiles of survival time.

### Computing Survival Probabilites

Being able to compute survival probabilites is especially of interest because it estimates the probability that a subject will not have experienced the event of interest beyond a specified time $t$. We developed the function `surv_prob` to compute probability of survival beyond time $t$, given that the data follows a specified parametric distribution. The first argument of this function is a data frame to be refrenced, containing a time column and censor column. The second argument is a string name for a distribution that has a corresponding density function and a distribution function. Some examples of the distributions that can be used here are `norm`, `lnorm`, `exp`, `weibull`, `logis`, `llogis`, and `gompertz`. The third argument is the time at which suvival is to be computed, which is input as a scalar quantity. The 4th argument is the string name of the time column of the dataframe. This argument defaults to `Time`. Similarly, the final argument is the string name of the censor column of the dataframe, which defaults to `Censor`. The censor column must be a numeric indicator variable where complete times correspond to a value of 1, and censored times correspond to a value of 0. An example of this function is shown below.

```
library(survival)
library(parmsurvfit)
data("rats")
surv_prob(rats, "lnorm", 110, time="time", censor="status")
```

```
#> P(T > 110) = 0.7948027
```

As seen in the output from the function above, utilizing the rats dataset available in the **survival** package and fiting a log-normal parametric distribution to the data, the estimated probability that a rat survives beyond 110 days is 0.7948, or roughly 80%.

### Computing Summary Statistics

Another useful form of output that we believed would be useful to also have in R is a table of summary statistics. Summmary statistics that are typically included are the mean, standard deviation, median, and IQR. The surv_summary function that we developed estimates various summary statistics, including mean, median, standard deviation, and percentiles of survival time given that the data follows a specified parametric distribution. This function also supports the option to provide seperate summary statistics for each level of a grouping variable, if desired.

```
library(parmsurvfit)
library(survival)
data("rats")
surv_summary(rats, "lnorm", time="time", censor = "status", by="sex")
```

```
#>
#>
#> For level = f
#> meanlog      4.876678
#> sdlog        0.487993
#> Log Liklihood    -247.0791
#> AIC      498.1581
#> BIC      504.1794
#> Mean     147.7833
#> StDev        76.63145
#> First Quantile   94.39914
#> Median       131.1941
#> Third Quantile   182.3311
#>
#> For level = m
#> meanlog      6.201628
#> sdlog        0.7477529
#> Log Liklihood    -18.39156
#> AIC      40.78312
#> BIC      46.80439
#> Mean     652.7505
#> StDev        564.981
#> First Quantile   298.0544
#> Median       493.5518
#> Third Quantile   817.2781
```

As seen above, after specifying the grouping variable of sex, two seperate tables were produced. We can see that the mean log survival time for female rats was smaller than the mean log survival time for male rats. The standard deviation of log survival time for female rats was also much smaller than that of male rats.

### Summary

This file is only a basic article template. For full details of *The R Journal* style and information on how to prepare your article for submission, see the Instructions for Authors.

*Victor Wilson*
*California Polytechnic State University, San Luis Obispo - Statistics Department*

victorjw26@yahoo.com

*Ashley Jacobson*
*California Polytechnic State University, San Luis Obispo - Statistics Department*

ashleypjacobson@gmail.com

*Shannon Pileggi*
*California Polytechnic State University, San Luis Obispo - Statistics Department*

spileggi@calpoly.edu