Exam 1

Victor Feagins

10/4/2021

## Question 1

Consider the mortality table based on 608 known deaths of Dall Mountain Sheep. The data are expressed per 1,000 sheep on the table below

Table 1:Dall Mountain Sheep Deaths In Intervals

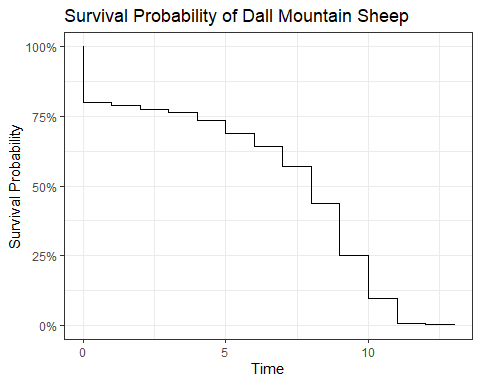
| Start\_Interval | End\_Interval | Number\_Of\_Survivors | Number\_of\_Deaths |
| --- | --- | --- | --- |
| 0 | 1 | 1,000 | 199 |
| 1 | 2 | 801 | 12 |
| 2 | 3 | 789 | 13 |
| 3 | 4 | 776 | 12 |
| 4 | 5 | 764 | 30 |
| 5 | 6 | 734 | 46 |
| 6 | 7 | 688 | 48 |
| 7 | 8 | 640 | 69 |
| 8 | 9 | 571 | 132 |
| 9 | 10 | 439 | 187 |
| 10 | 11 | 252 | 156 |
| 11 | 12 | 96 | 90 |
| 12 | 13 | 6 | 3 |
| 13 | 14 | 3 | 3 |

### Compute and plot the Survival Function

Surv.1 <- Surv(time = df.1.t$Start\_Interval,   
 time2 = df.1.t$End\_Interval,   
 event = rep(1, nrow(df.1.t)),  
 type = "interval")  
  
summary(fit.1 <- survfit(Surv.1~1))

## Call: survfit(formula = Surv.1 ~ 1)  
##   
## time n.risk n.event survival std.err lower 95% CI upper 95% CI  
## 0 1001.00 199.2 0.801 1.01e-02 0.77665 0.82611  
## 1 801.80 12.0 0.789 1.02e-02 0.76413 0.81467  
## 2 789.79 13.0 0.776 1.02e-02 0.75061 0.80225  
## 3 776.78 12.0 0.764 1.02e-02 0.73815 0.79075  
## 4 764.76 30.0 0.734 1.02e-02 0.70714 0.76188  
## 5 734.73 46.0 0.688 1.01e-02 0.65990 0.71730  
## 6 688.69 48.0 0.640 9.71e-03 0.61096 0.67042  
## 7 640.64 69.1 0.571 8.93e-03 0.54116 0.60248  
## 8 571.57 132.1 0.439 6.88e-03 0.40932 0.47083  
## 9 439.44 187.2 0.252 3.46e-03 0.22650 0.28037  
## 10 252.25 156.2 0.096 8.93e-04 0.07939 0.11609  
## 11 96.10 90.1 0.006 1.46e-05 0.00270 0.01331  
## 12 6.01 3.0 0.003 5.18e-06 0.00097 0.00928  
## 13 3.00 3.0 0.000 0.00e+00 NA NA

autoplot(fit.1, conf.int = F) +  
 labs(title = "Survival Probability of Dall Mountain Sheep",  
 x = "Time",  
 y = "Survival Probability")

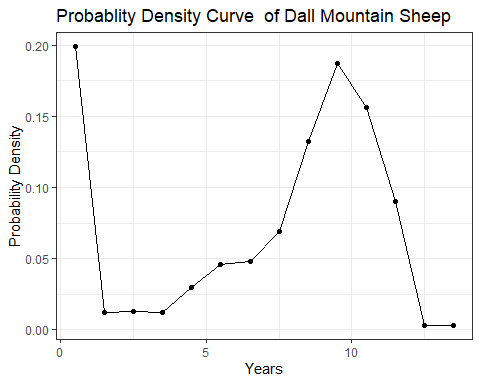


### Compute and plot the Probability Function

df.1 <- df.1 %>%   
 mutate("f(t)" = Number\_of\_Deaths/(max(Number\_Of\_Survivors) \* (End\_Interval - Start\_Interval)))

Table 2:Probability Density Function Calcuated at the Middle of the Interval

| Midtime | f(t) |
| --- | --- |
| 0.5 | 0.199 |
| 1.5 | 0.012 |
| 2.5 | 0.013 |
| 3.5 | 0.012 |
| 4.5 | 0.030 |
| 5.5 | 0.046 |
| 6.5 | 0.048 |
| 7.5 | 0.069 |
| 8.5 | 0.132 |
| 9.5 | 0.187 |
| 10.5 | 0.156 |
| 11.5 | 0.090 |
| 12.5 | 0.003 |
| 13.5 | 0.003 |

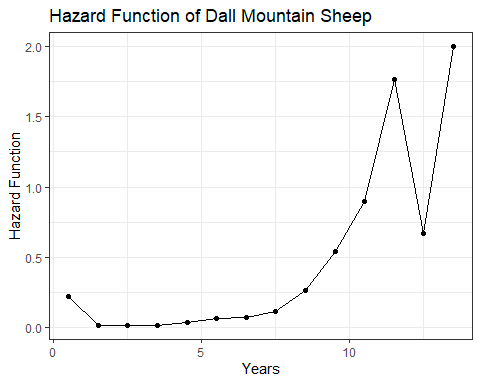


### Compute and plot the Hazard Function

df.1 <- df.1 %>%   
 mutate(Deaths\_Per\_Time = Number\_of\_Deaths/(End\_Interval - Start\_Interval),  
 "h(t)" = Deaths\_Per\_Time/(Number\_Of\_Survivors - Number\_of\_Deaths/2))

Table 3:Hazard Function Calcuated at the Middle of the Interval

| Midtime | h(t) |
| --- | --- |
| 0.5 | 0.22098834 |
| 1.5 | 0.01509434 |
| 2.5 | 0.01661342 |
| 3.5 | 0.01558442 |
| 4.5 | 0.04005340 |
| 5.5 | 0.06469761 |
| 6.5 | 0.07228916 |
| 7.5 | 0.11395541 |
| 8.5 | 0.26138614 |
| 9.5 | 0.54124457 |
| 10.5 | 0.89655172 |
| 11.5 | 1.76470588 |
| 12.5 | 0.66666667 |
| 13.5 | 2.00000000 |



### Discussing the Hazard Function

Looking at the hazard function we see that the first value is high possible indicating infant mortality. After the sheep lives past this point the hazard function increases over time typical behavior for life span which according to the [National Park Service](https://www.nps.gov/articles/about-dall-sheep.htm) is 12 to 16 years. No sheep in our sample lived to 16 years I suspect this is due to how the sample was collected. Is is unlikely that humans were recording when the sheep were born and more likely they caught young sheep and kept up with them in the wild. It might be possible that the sheep dying at 13 years might be 16 years old. This hypothesis would explain the low initial hazard function because death at births are not captured.

### Mean and Median Survival Times

print(fit.1, print.rmean = TRUE)

## Call: survfit(formula = Surv.1 ~ 1)  
##   
## records n events rmean\* se(rmean) median 0.95LCL 0.95UCL  
## [1,] 1000 1001 1001 6.56 0.122 8 8 8  
## \* restricted mean with upper limit = 13

We see that the mean survival time is 6.559 years and the median 8 years

## Question 2

Consider the remission data of 40 female patients diagnosed with breast cancer in a certain hospital during the years 1989–93. The data include: the age of the patient as of last contact (follow up/death), stage of cancer at time of diagnosis, date of last contact, status (alive/dead), the number of days lived, the number of years lived. The data are given on the following page.

Table 4:First 10 Female Patients diagnosed with Breast Cancer

| ID | Age | Stage | Diagnosis | Last | Status | Days | Years |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 39 | 1 | 02/01/89 | 10/23/92 | A | 1,360 | 3.723 |
| 2 | 55 | 1 | 03/22/89 | 02/12/95 | A | 2,153 | 5.895 |
| 3 | 56 | 2 | 04/16/89 | 09/05/89 | D | 142 | 0.389 |
| 4 | 63 | 1 | 05/23/89 | 12/20/92 | D | 1,307 | 3.578 |
| 5 | 62 | 2 | 06/12/89 | 12/28/95 | A | 2,390 | 6.543 |
| 6 | 42 | 2 | 09/05/89 | 12/17/90 | A | 468 | 1.281 |
| 7 | 45 | 1 | 10/05/89 | 08/04/95 | A | 2,129 | 5.829 |
| 8 | 38 | 2 | 11/30/89 | 11/10/91 | D | 710 | 1.944 |
| 9 | 53 | 2 | 01/07/90 | 10/25/90 | D | 291 | 0.797 |
| 10 | 55 | 1 | 02/03/90 | 01/31/91 | D | 362 | 0.991 |

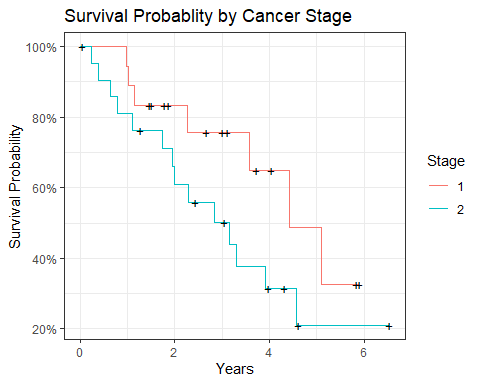
Compute and plot the PL estimates of S(t) at every time to death for the Stage 1 and 2 groups.

Surv.2 <- Surv(df.2$Years, df.2$Status.l)  
fit.2 <- survfit(Surv.2 ~ Stage, data = df.2 , robust = TRUE)  
summary(fit.2)

## Call: survfit(formula = Surv.2 ~ Stage, data = df.2, robust = TRUE)  
##   
## Stage=1   
## time n.risk n.event survival std.err lower 95% CI upper 95% CI  
## 0.991 18 1 0.944 0.0540 0.844 1.00  
## 1.021 17 1 0.889 0.0741 0.755 1.00  
## 1.142 16 1 0.833 0.0878 0.678 1.00  
## 2.264 11 1 0.758 0.1077 0.573 1.00  
## 3.578 7 1 0.649 0.1362 0.430 0.98  
## 4.416 4 1 0.487 0.1738 0.242 0.98  
## 5.098 3 1 0.325 0.1760 0.112 0.94  
##   
## Stage=2   
## time n.risk n.event survival std.err lower 95% CI upper 95% CI  
## 0.246 21 1 0.952 0.0465 0.8655 1.000  
## 0.389 20 1 0.905 0.0641 0.7875 1.000  
## 0.652 19 1 0.857 0.0764 0.7198 1.000  
## 0.797 18 1 0.810 0.0857 0.6579 0.996  
## 1.098 17 1 0.762 0.0929 0.5999 0.968  
## 1.736 15 1 0.711 0.0997 0.5403 0.936  
## 1.944 14 1 0.660 0.1047 0.4839 0.901  
## 2.004 13 1 0.610 0.1083 0.4303 0.863  
## 2.292 12 1 0.559 0.1105 0.3792 0.823  
## 2.839 10 1 0.503 0.1127 0.3241 0.780  
## 3.157 8 1 0.440 0.1148 0.2638 0.734  
## 3.307 7 1 0.377 0.1143 0.2082 0.683  
## 3.912 6 1 0.314 0.1112 0.1571 0.629  
## 4.578 3 1 0.210 0.1132 0.0727 0.604

## Comparing Survival Functions

autoplot(fit.2, conf.int = FALSE) +  
 labs(title = "Survival Probablity by Cancer Stage",  
 x = "Years",  
 y = "Survival Probability",  
 colour = "Stage")



survdiff(Surv.2 ~ Stage, data = df.2)

## Call:  
## survdiff(formula = Surv.2 ~ Stage, data = df.2)  
##   
## N Observed Expected (O-E)^2/E (O-E)^2/V  
## Stage=1 19 7 10.5 1.15 2.33  
## Stage=2 21 14 10.5 1.14 2.33  
##   
## Chisq= 2.3 on 1 degrees of freedom, p= 0.1

Doing a log rank test we see the curves are significantly different at the .1 level but not at the .05 level. Looking at the curves we see that stage 2 had a smaller survival curve throughout time which makes practical sense since stage 2 cancer is more intense then stage 1.

## Compute the survival function and the variance

To calculate Survival and variance at a particular point we will linearly interpolate the points closest to the point at interest that have had events.

For the variance of S(t) We can use Greenwood’s Formula

Where is the number at risk right before and is the number of deaths occurred at time

fit.df.2 <- broom::tidy(fit.2)  
fit.df.2 <- fit.df.2 %>%   
 group\_by(strata) %>%   
 mutate(variance = estimate^2 \* cumsum(n.event/(n.risk \* (n.risk - n.event)))) %>% #Estimating the variance   
 ungroup()

Table 5:Variance Estimate of Cancer Patients

| time | estimate | strata | variance |
| --- | --- | --- | --- |
| 0.049 | 1.0000000 | Stage=1 | 0.000000000 |
| 0.991 | 0.9444444 | Stage=1 | 0.002914952 |
| 1.021 | 0.8888889 | Stage=1 | 0.005486968 |
| 1.142 | 0.8333333 | Stage=1 | 0.007716049 |
| 1.470 | 0.8333333 | Stage=1 | 0.007716049 |
| 1.514 | 0.8333333 | Stage=1 | 0.007716049 |
| 1.780 | 0.8333333 | Stage=1 | 0.007716049 |
| 1.873 | 0.8333333 | Stage=1 | 0.007716049 |
| 2.264 | 0.7575758 | Stage=1 | 0.011594364 |
| 2.678 | 0.7575758 | Stage=1 | 0.011594364 |
| 3.012 | 0.7575758 | Stage=1 | 0.011594364 |
| 3.107 | 0.7575758 | Stage=1 | 0.011594364 |
| 3.578 | 0.6493506 | Stage=1 | 0.018557743 |
| 3.723 | 0.6493506 | Stage=1 | 0.018557743 |
| 4.041 | 0.6493506 | Stage=1 | 0.018557743 |
| 4.416 | 0.4870130 | Stage=1 | 0.030203868 |
| 5.098 | 0.3246753 | Stage=1 | 0.030992952 |
| 5.829 | 0.3246753 | Stage=1 | 0.030992952 |
| 5.895 | 0.3246753 | Stage=1 | 0.030992952 |
| 0.246 | 0.9523810 | Stage=2 | 0.002159594 |
| 0.389 | 0.9047619 | Stage=2 | 0.004103229 |
| 0.652 | 0.8571429 | Stage=2 | 0.005830904 |
| 0.797 | 0.8095238 | Stage=2 | 0.007342620 |
| 1.098 | 0.7619048 | Stage=2 | 0.008638376 |
| 1.281 | 0.7619048 | Stage=2 | 0.008638376 |
| 1.736 | 0.7111111 | Stage=2 | 0.009932981 |
| 1.944 | 0.6603175 | Stage=2 | 0.010960371 |
| 2.004 | 0.6095238 | Stage=2 | 0.011720549 |
| 2.292 | 0.5587302 | Stage=2 | 0.012213512 |
| 2.437 | 0.5587302 | Stage=2 | 0.012213512 |
| 2.839 | 0.5028571 | Stage=2 | 0.012702559 |
| 3.047 | 0.5028571 | Stage=2 | 0.012702559 |
| 3.157 | 0.4400000 | Stage=2 | 0.013182540 |
| 3.307 | 0.3771429 | Stage=2 | 0.013071720 |
| 3.912 | 0.3142857 | Stage=2 | 0.012370100 |
| 3.984 | 0.3142857 | Stage=2 | 0.012370100 |
| 4.318 | 0.3142857 | Stage=2 | 0.012370100 |
| 4.578 | 0.2095238 | Stage=2 | 0.012814527 |
| 4.605 | 0.2095238 | Stage=2 | 0.012814527 |
| 6.543 | 0.2095238 | Stage=2 | 0.012814527 |

### At S(2) for stage 1 patients

s.2<- sliced.2 %$%   
 approx(time, estimate, 2)$y #Does linear interpolation for s(2)  
v.2 <- sliced.2 %$%  
 approx(time, variance, 2)$y #Does linear interpolation for var(s(2))

Using the linear interpolation at we get a value of 0.7754 and a variance of 0.0107

### At S(3) for stage 2 patients

s.3<- sliced.3 %$%   
 approx(time, estimate, 3)$y #Does linear interpolation for s(2)  
v.3 <- sliced.3 %$%  
 approx(time, variance, 3)$y #Does linear interpolation for var(s(2))

Using the linear interpolation at we get a value of 0.471 and a variance of 0.0129.

### What is the Median Remission Times

print(fit.2)

## Call: survfit(formula = Surv.2 ~ Stage, data = df.2, robust = TRUE)  
##   
## n events median 0.95LCL 0.95UCL  
## Stage=1 19 7 4.42 3.58 NA  
## Stage=2 21 14 3.16 1.94 NA

We see that for that the median survival time of Stage 1 is 4.42 years and the median survival time of Stage 2 is 3.16

## Question 3

Using the data in problem, Q2, is there a difference in the survival functions by stage of cancer assessed at the time of diagnosis?

In [Question 2](#Q2) we examined the differences between the 2 survival functions of cancer patients at different stages of cancer. Conceptually we would imagine that the stage 2 cancer patients would have a greater risk of death since stage 2 cancer is greater intensity then stage 1.

## Call: survfit(formula = Surv.2 ~ Stage, data = df.2, robust = TRUE)  
##   
## n events median 0.95LCL 0.95UCL  
## Stage=1 19 7 4.42 3.58 NA  
## Stage=2 21 14 3.16 1.94 NA

When simply looking at the descriptive statistics of we see that the stage 1 had 7 deaths while stage 2 had 14 deaths. With comparable sample sizes. The stage 1 median is also less then the median of stage 2 patients. We examine the survival curves [plots](#Compare) we see that stage 2 patients have a lower survival curve through out the time domain. When doing the statistical test to see if the groups are different we have significance at the .1 level. In conclusion it is fair to say that stage 1 cancer patients survive longer then stage 2 patients.