Assignment_#1_Responses_Jack_Tan

JACKTANSNAKE

2021/1/25

```
# Load the survival package:
library(survival)
library(tidyverse)

# Read in Cell data:
Cell = read.csv('https://www.macalester.edu/~addona/Cell.csv')
Stroke = read.csv('https://www.macalester.edu/~addona/Stroke.csv')
```

1. Lifetime data commonly occur in the engineering environment. In this example, from a Canadian aluminum smelter, alumina is liquified in a steel-lined box (or cell) which is built to withstand extremely high temperatures. In the smelting process, aluminium is produced as a byproduct when the cell functions like a battery with molten alumina as the electrolyte. The cell needs to be replaced when the carbon lining cracks, allowing impurities into the process. The failure time data listed here (also in Cell.csv) represent days of service until replacement for 17 cells: 1540, 1415, 660, 999, 1193, 1006, 869, 1035, 797, 296, 775, 1424, 1169, 1500, 728, 670, 841.

(a) Find a non-parametric estimate of the chance that a cell lasts at most 1175 days.

$$P(X \leq 1175) = \frac{12}{17} = 0.7058824$$

(b) Fit an Exponential model, and a Weibull model, for these lifetimes, and use these two models to re-estimate the probability from (a).

- Exponetial model:

```
survreg(Surv(Time)~1, dist = "exponential", data = Cell)
```

```
## Call:
## survreg(formula = Surv(Time) ~ 1, data = Cell, dist = "exponential")
##
## Coefficients:
## (Intercept)
##    6.902861
##
## Scale fixed at 1
##
## Loglik(model)= -134.3   Loglik(intercept only)= -134.3
## n= 17
```

```r
lambda <- 1/exp(6.902861)
pexp(1175, rate = lambda)
```

```
## [1] 0.6929562
```

The exponential estimate would $P(X \leq 1175) = 0.6929562$.

- Weibull:

```r
survreg(Surv(Time)~1, dist = "weibull", data = Cell)
```

```
## Call:
## survreg(formula = Surv(Time) ~ 1, data = Cell, dist = "weibull")
##
## Coefficients:
## (Intercept)
##    7.011883
##
## Scale= 0.3012274
##
## Loglik(model)= -122.8   Loglik(intercept only)= -122.8
## n= 17
```

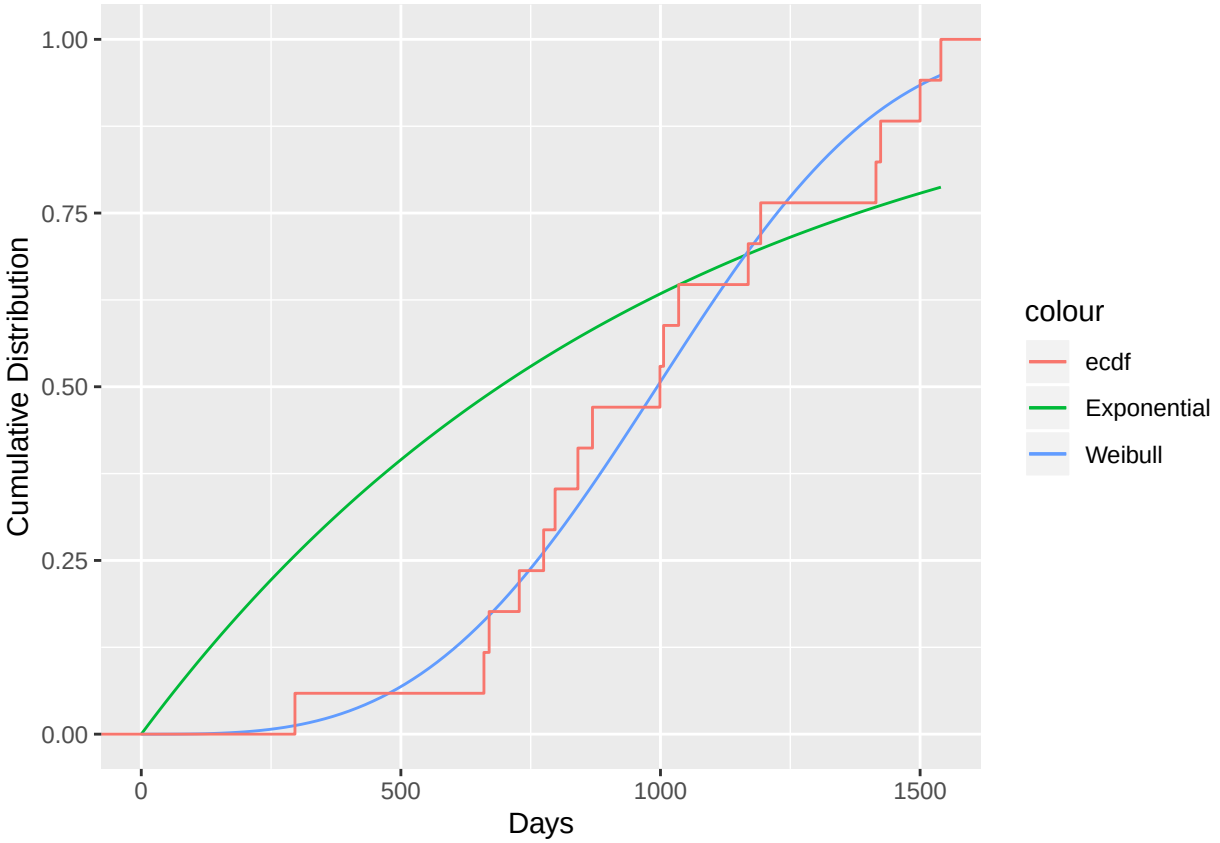
```r
a <- 1/0.3012274
b <- exp(7.011883)
pweibull(1175, shape = a, scale = b)
```

```
## [1] 0.7014677
```

The weibull estimate would be $P(X \leq 1175) = 0.7014677$.

(c) Plot the empirical cumulative distribution function (ecdf) for this data. Add to this plot the estimated cumulative distribution functions using the Exponential and Weibull models. Based on the graph you created, which of these 3 models would you not feel comfortable using?

```r
ggplot(data = data.frame(x = c(0, max(Cell$Time))), aes(x = x)) +
  stat_function(aes(color = "Exponential"), fun = pexp, args = list(rate = lambda)) +
  stat_function(aes(color = "Weibull"), fun = pweibull, args = list(shape = a, scale = b)) +
  stat_ecdf(data = Cell, aes(x = Time, color = "ecdf")) +
  labs(y = "Cumulative Distribution", x = "Days")
```



I don't feel comfortable using the Exponential one because the distribution does not capture the trend of the data well.

(d) Would you say that a comparison of your answers to (a) and (b) would have led you to the same conclusion you arrived at in (c)? If so, explain why, and if not, discuss what you take away from this realization.

I don't think that such a comparison would lead me to the same conclusion because the results were all around 0.7. However, the exponential curve has a different shape with all other curves, which leads to the conclusion that simply checking $P(X \leq k)$ is not enough to assure a parametric method successfully captures the trend of given data.

(e) Find the mean failure time of these cells. What does this value correspond to in your graph from (c)?

```
mean(Cell$Time)
```

```
## [1] 995.1176
```

This value corresponds to the area above the ecdf curve but below the y=1 horizontal line.

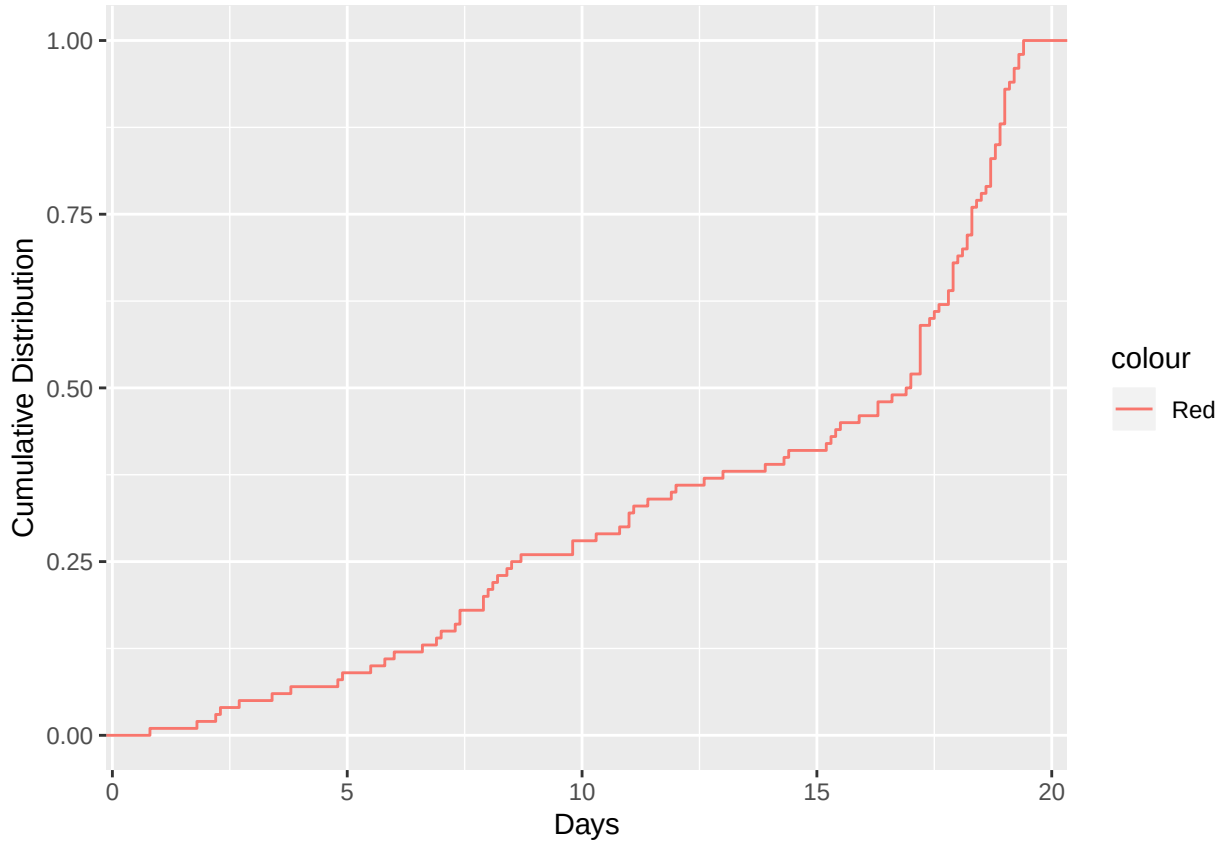
2. The data contained in Stroke.csv are from a study of the occurrence and risk factors for stroke in patients with no prior history of stroke. The variables are: T: Time until first stroke from baseline evaluation (in years) Censored: 1 = no stroke occurred during study , 0 = stroke occurred during study Age: Age of patient (in years) Sex: 0 = female , 1 = male SBP: systolic blood pressure DBP:

diastolic blood pressure BMI: body mass index Smoke: 1 = current smoker , 0 = other ACH: Alcoholic drinking status (1 = current drinker , 0 = other) HDL: high density lipoprotein cholesterol LDL: low density lipoprotein cholesterol DM: Diabetes status (1 = yes , 0 = no) ALBU: albuminuria status (1 = normal , 2 = elevated , 3 = very high)

If no stroke occurred during the study (i.e., if Censored = 1), the value of T represents a lower bound on the stroke-free time experienced by an individual. We have not yet seen how to handle such cases, thus, for now, we will ignore the Censored variable and assume that all values of T are exact stroke-free times.

(a) Plot the ecdf of the stroke-free time experienced by a patient.

```
ggplot(data = data.frame(x = c(0, max(Stroke$T))), aes(x = x)) +
  stat_ecdf(data = Stroke, aes(x = T, color = "Red"))+
  labs(y = "Cumulative Distribution", x = "Days")
```



(b) Add to the plot from (a) the estimated cumulative distribution functions using the Normal, and Weibull, models. Do these 2 models turn out to be very similar, or different, from each other?

- Exponetial model:

```
survreg(Surv(T)~1, dist = "gaussian", data = Stroke)
```

```
## Call:
```

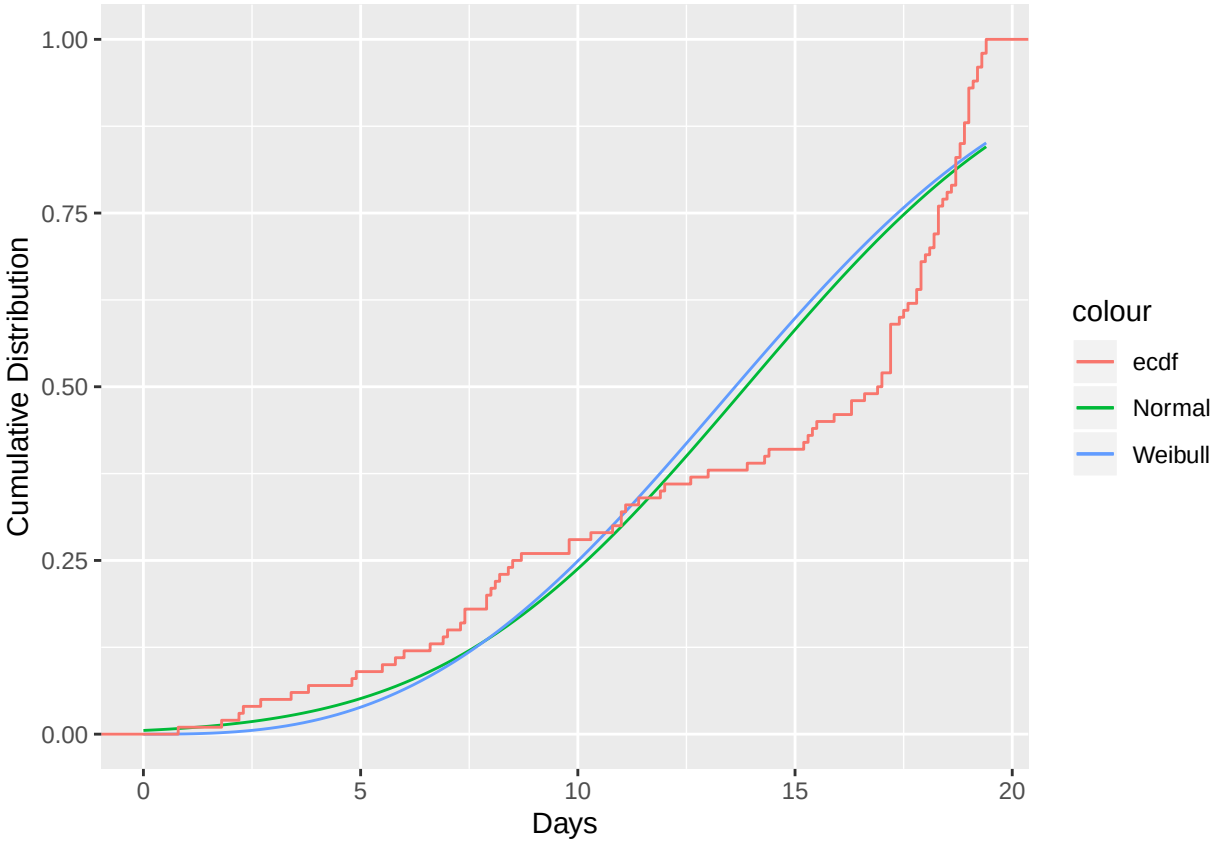
```
## survreg(formula = Surv(T) ~ 1, data = Stroke, dist = "gaussian")
##
## Coefficients:
## (Intercept)
##      13.873
##
## Scale= 5.434277
##
## Loglik(model)= -311.2   Loglik(intercept only)= -311.2
## n= 100
```

```
survreg(Surv(T)~1, dist = "weibull", data = Stroke)
```

```
## Call:
## survreg(formula = Surv(T) ~ 1, data = Stroke, dist = "weibull")
##
## Coefficients:
## (Intercept)
##    2.739948
##
## Scale= 0.3500541
##
## Loglik(model)= -314.2   Loglik(intercept only)= -314.2
## n= 100
```

```
m = 13.873
var = 5.434277
a <- 1/0.3500541
b <- exp(2.739948)
```

```
ggplot(data = data.frame(x = c(0, max(Stroke$T))), aes(x = x)) +
  stat_function(aes(color = "Normal"), fun = pnorm, args = list(mean = m, sd = var)) +
  stat_function(aes(color = "Weibull"), fun = pweibull, args = list(shape = a, scale = b)) +
  stat_ecdf(data = Stroke, aes(x = T, color = "ecdf")) +
  labs(y = "Cumulative Distribution", x = "Days")
```



The weibull model and the normal model look almost identical to each other, but they are still pretty off from the ecdf.

(c) If given the choice between using the ecdf model or the Normal/Weibull, why might we trust the ecdf model more?

Because the ecdf model does not cast any assumption about the distribution of the data. In this case, the weibull/ normal distribution assumption of the data might just be wrong, which is why we would trust ecdf more.

(d) Estimate the mean and median stroke-free time assuming the: (i) ecdf model,

(ii) Normal model, and (iii) Weibull model.

- (i) ecdf model:

```
mean(Stroke$T)
```

```
## [1] 13.873
```

```
median(Stroke$T)
```

```
## [1] 16.95
```

In the ecdf model, the mean and median stroke-free time is 13.873 abd 16.95.

- (ii) Normal model

```
S <- function(x) 1 - pnorm(x, mean = m, sd = var)
integrate(S, lower = 0, upper = Inf)
```

```
## 13.88223 with absolute error < 5.9e-05
```

```
qnorm(0.5, mean = m, sd = var)
```

```
## [1] 13.873
```

In the normal model, the mean and the median stroke-free time is 13.88223 and 13.873.

- (iii) Weibull model

```
S <- function(x) 1 - pweibull(x, shape = a, scale = b)
integrate(S, lower = 0, upper = Inf)
```

```
## 13.80045 with absolute error < 7e-06
```

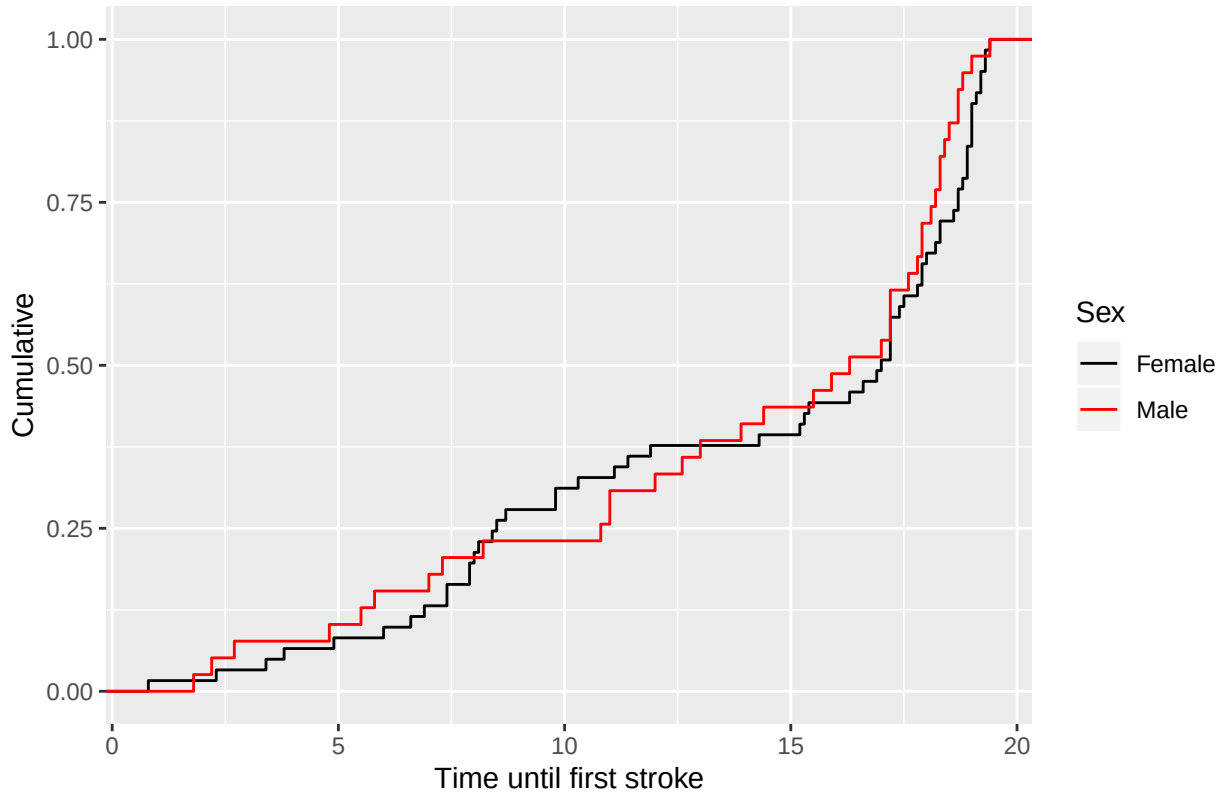
```
qweibull(0.5, shape = a, scale = b)
```

```
## [1] 13.62149
```

In the weibull model, the mean and the median stroke-free time is 13.80045 and 13.62149.

(e) Compare the ecdf of T for males and females. One way (of many) to do this is: `ggplot(data=Stroke,aes(x=T,color=fa + stat_ecdf() + scale_color_manual(labels=c("Female","Male"),values=c("black","red")) + ggtitle("") + ylab("Cumulative") + xlab("Time until first stroke") + guides(color=guide_legend(ti`

```
ggplot(data=Stroke,aes(x=T,color=factor(Sex))) + stat_ecdf() +
scale_color_manual(labels=c("Female","Male"),values=c("black","red")) +
ggtitle("") + ylab("Cumulative") + xlab("Time until first stroke") +
guides(color=guide_legend(title="Sex"))
```



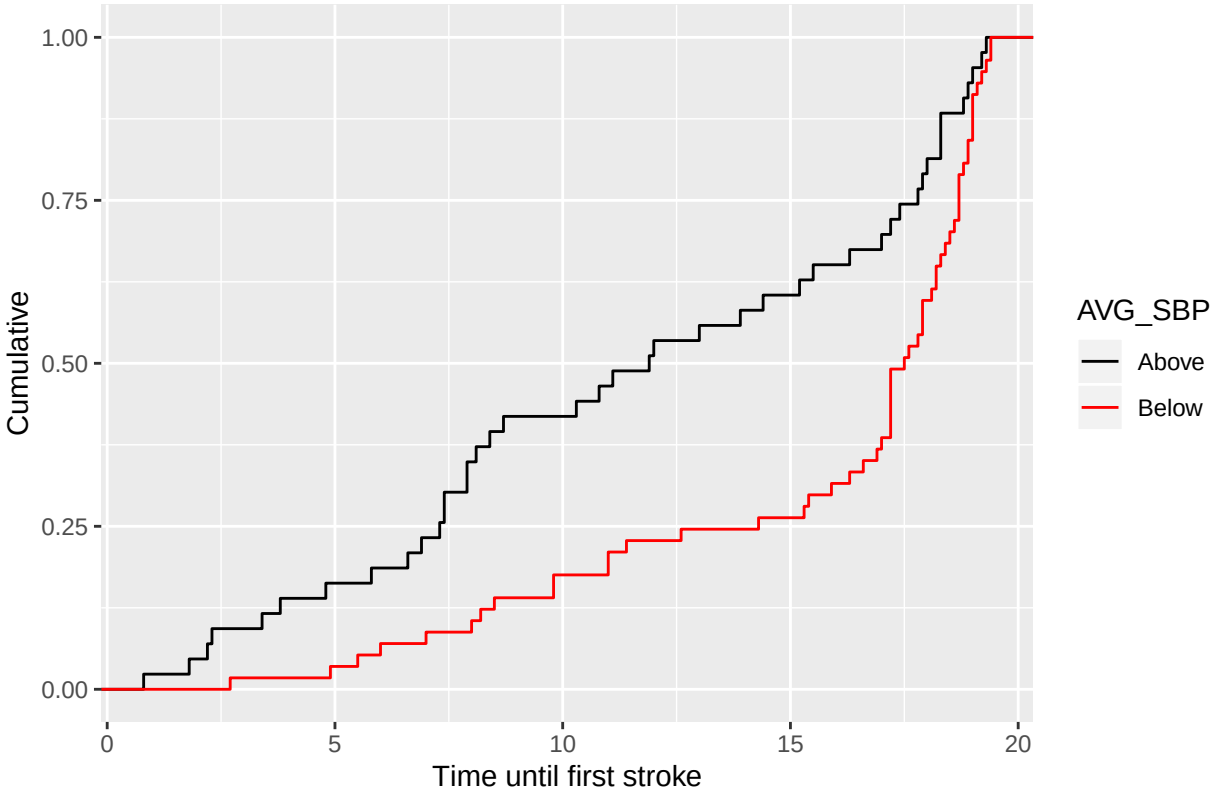
- Do you think Sex is a good explanatory variable for modeling time to first stroke? If so, which group has a shorter time to stroke?

I don't think that `Sex` is a good explanatory variable for modeling time to first stroke, since we can see that the ecdf of `Male` and `Female` look similar and mostly interwined with each other, which tells us that we can't really tell whose time until first stroke is longer from `Sex`.

- Repeat this by comparing the stroke-free time for individuals with below average systolic blood pressure (SBP) vs. above average SBP.

```
Stroke_new <- Stroke %>%
  mutate(AVG_SBP = ifelse(SBP > mean(Stroke$SBP), "Above", "below"))
```

```
ggplot(data=Stroke_new,aes(x=T,color=factor(AVG_SBP))) + stat_ecdf() +
scale_color_manual(labels=c("Above","Below"),values=c("black","red")) +
ggtitle("") + ylab("Cumulative") + xlab("Time until first stroke") +
guides(color=guide_legend(title="AVG_SBP"))
```



3. Answer the following conceptual short-answer questions:

(a) The advantage of performing a parametric analysis is that, if the distribution choice is correct, estimates will typically have lower variability, and the procedure will thus have higher power. What is an advantage of performing a non-parametric analysis?

An advantage of a non-parametric analysis is that it does cast any assumption on the distribution of the data, and thus it technically has infinite flexibility to fit any kind of data. Further, if we could attain a large enough sample size, then the estimated median of a non-parametric analysis would better fit the truth.

(b) We have said that the rate of change of the cumulative distribution function (CDF) tells us how the density function looks. Thus, the density function is the of the CDF.

Derivative

4. Answer the following True/False questions:

(a) True/False: To obtain an estimate of mean survival from a survival curve, we follow the value 0.5 on the y-axis until we hit the curve, then drop a perpendicular down to the x-axis and read off the corresponding x-value.

False

(b) True/False: To obtain an estimate of mean survival from a survival curve, we find the area under the curve.

True

(c) True/False: Depending on its parameters, the shape of a Normal can change.

False

(d) True/False: Depending on its parameter, the shape of an Exponential can change.

False

(e) True/False: Depending on its parameters, the shape of a Weibull can change.

True

(f) True/False: With appropriate choices of the parameters, a Weibull can look like an Exponential, or a Normal.

True

(g) True/False: It is possible for a survival curve to increase in a given interval.

False

(h) True/False: It is possible for a survival curve to be greater than 1.

False

(i) True/False: It is possible for a density function to be greater than 1.

False

(j) True/False: Obtaining a large enough sample size can make up for an incorrect parametric assumption.

False