Marathon Data

Sturdivant and Clark

2022-08-03

Motivational Video

Are we living in a time of records?

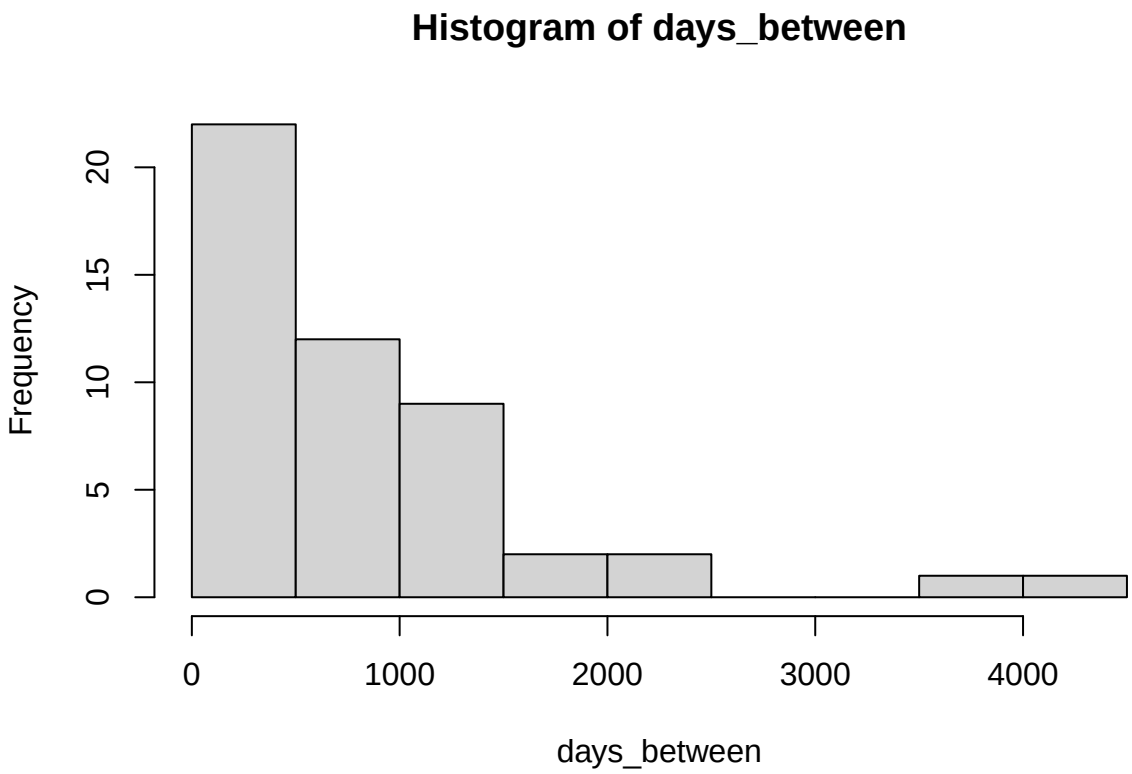
How can we address this question?

What would randomness look like?

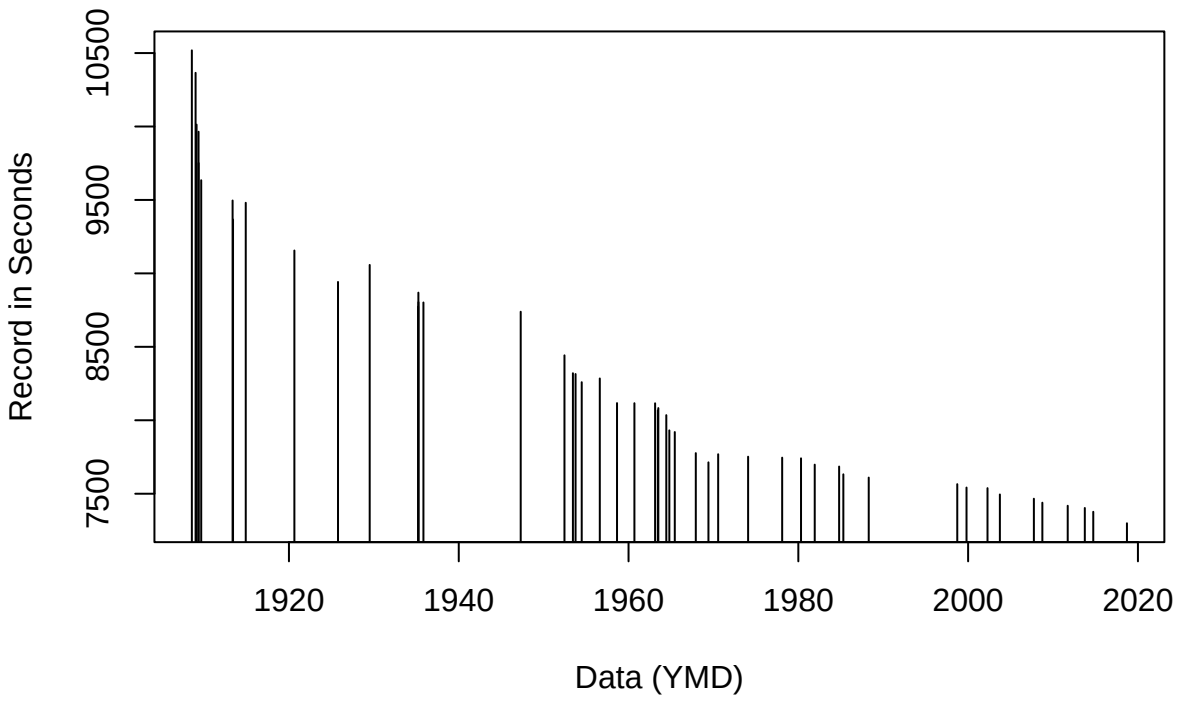
Getting and Visualizing Data

Data from Wikipedia (https://en.wikipedia.org/wiki/Marathon_world_record_progression) with the world records for men's marathon since 1908 are depicted graphically in several ways below.

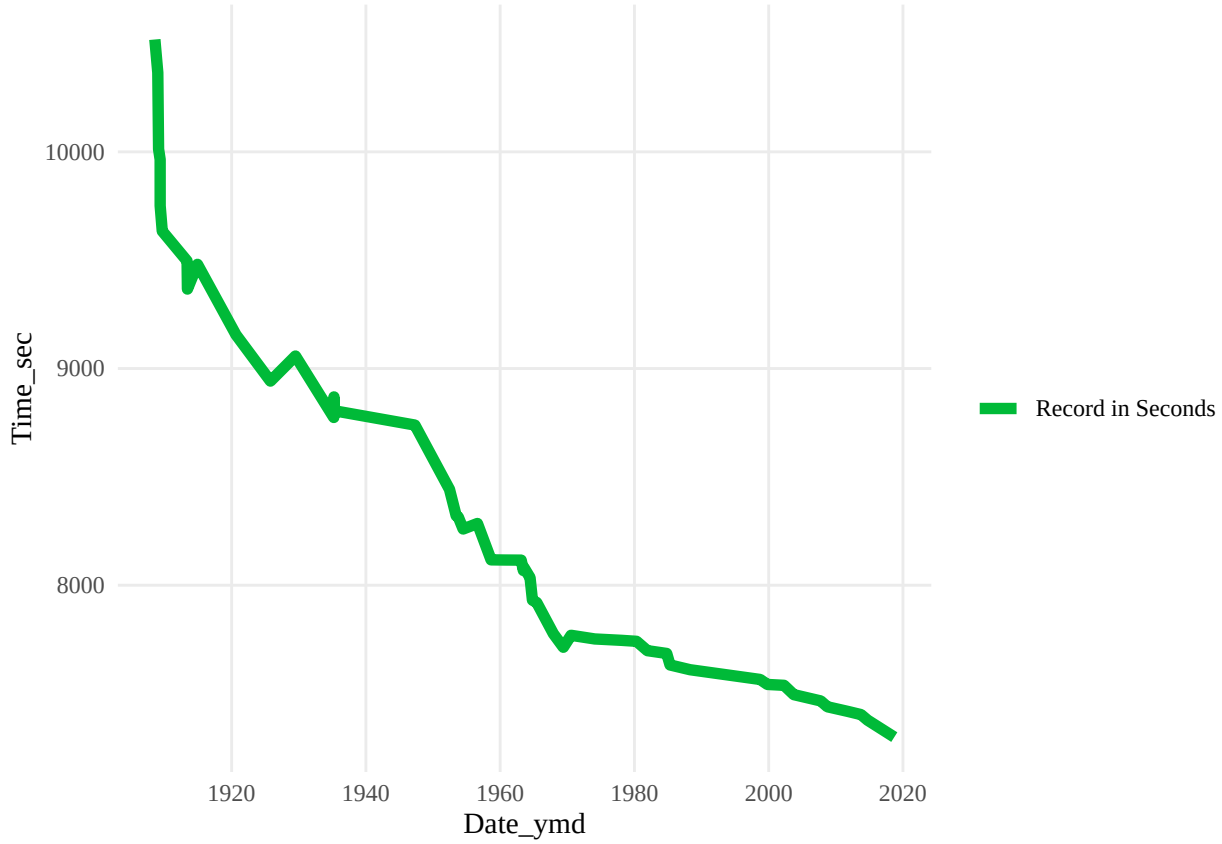
```
record_table_mod<-read_rds("record_table_mod.rds")
days_between = as.numeric(diff(record_table_mod$Date_ymd))
hist(days_between)
```



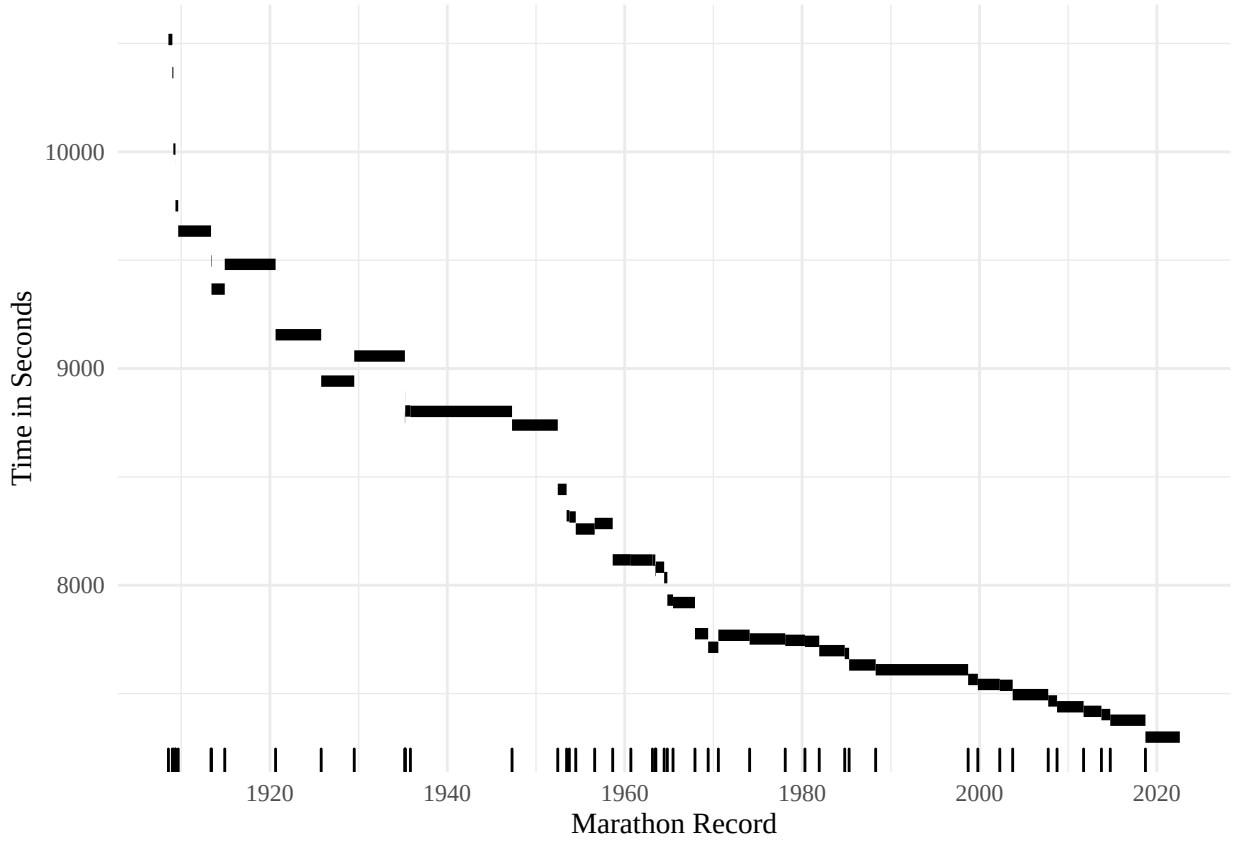

```
plot(record_table_mod$Date_ymd, record_table_mod$Time_t, type = "h",
     xlab = "Data (YMD)", ylab = "Record in Seconds")
```




```
record_table_mod %>% ggplot() +
  geom_line(aes(x=Date_ymd, y=Time_sec, color = "Record in Seconds"),size=2) +
  scale_color_manual(name="",
                     values = c("Record in Seconds"="#00ba38")) +
  theme(panel.grid.minor = element_blank())
```




```
ggplot() +
  geom_segment(data=record_table_mod, aes(x=Date_ymd,xend=end,y=Time_sec,yend=Time_sec),size=2) +
  ylab("Time in Seconds")+
  xlab("Marathon Record")+
  geom_rug(data=record_table_mod,aes(x=Date_ymd,y=Time_sec,),sides="b")
```



Questions to explore:

Are there issues with the data? What is our response variable? What type of data is it? Describe what each plot tells us about the data. Which plot(s) do you find most helpful (and why)? Does the time between marathon records appear random? Are there historical "times of records"?
When (if ever) would you predict a sub 2 hour marathon? Is there a limit to the fastest marathon in the future? If so, how fast?

Developing a Model

Poisson Process

"A Poisson Process is a model for a series of discrete events where the average time between events is known, but the exact timing of events is random" meeting the folling criteria: https://towardsdatascience.com/the-poisson-distribution-and-poisson-process-explained-4e2cb17d459

- Events are independent of each other. The occurrence of one event does not affect the probability another event will occur.
- The average rate (events per time period) is constant.
- Two events cannot occur at the same time.

The time between events (known as the interarrival times) follow an exponential distribution defined as:

$$P(T > t) = e^{-\lambda t}$$

Where T is the random variable of the time until the next event, t is a specific time for the next event, and λ is the rate: the average number of events per unit of time.

Questions to explore:

Are the assumptions of the Poisson process reasonable for the marathon record data? Explain. How might you determine if the exponential model "fits" the marathon data?

Exponential Distribution for Interarrivals

There are several steps one might use to determine if a model, such as exponential interarrivals, is appropriate. The exponential distribution has certain attributes, for example:

$E(T) = 1/\lambda$ $SD(T) = 1/\lambda$

In other words, the mean of the distibution is the same as its standard deviation and both are equal to the one over the rate λ

Activity: the variable "days_between" has the time between world records. Compute the mean and standard deviation for this variable. Are they reasonably similar? What is your best estimate of λ?

Solution:

```
mean(days_between)
```

```
## [1] 821.0408
```

```
sd(days_between)
```

```
## [1] 886.2897
```

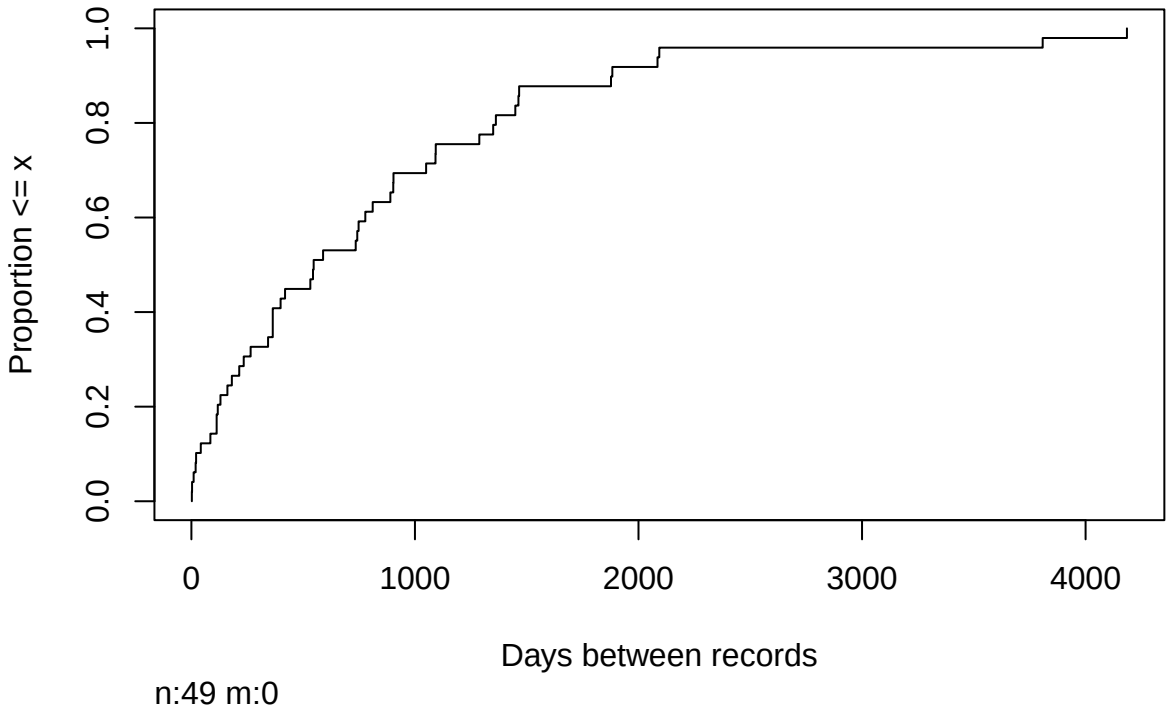
In the R package "MASS" is a function "fitdistr" which we can use to determine the "best" exponential distribution for a given set of data. The command is below and the estimate of the rate λ extracted. How does this compare to your estimate above?

```
expfit=fitdistr(days_between,"exponential")
exprate<-expfit$estimate
exprate
```

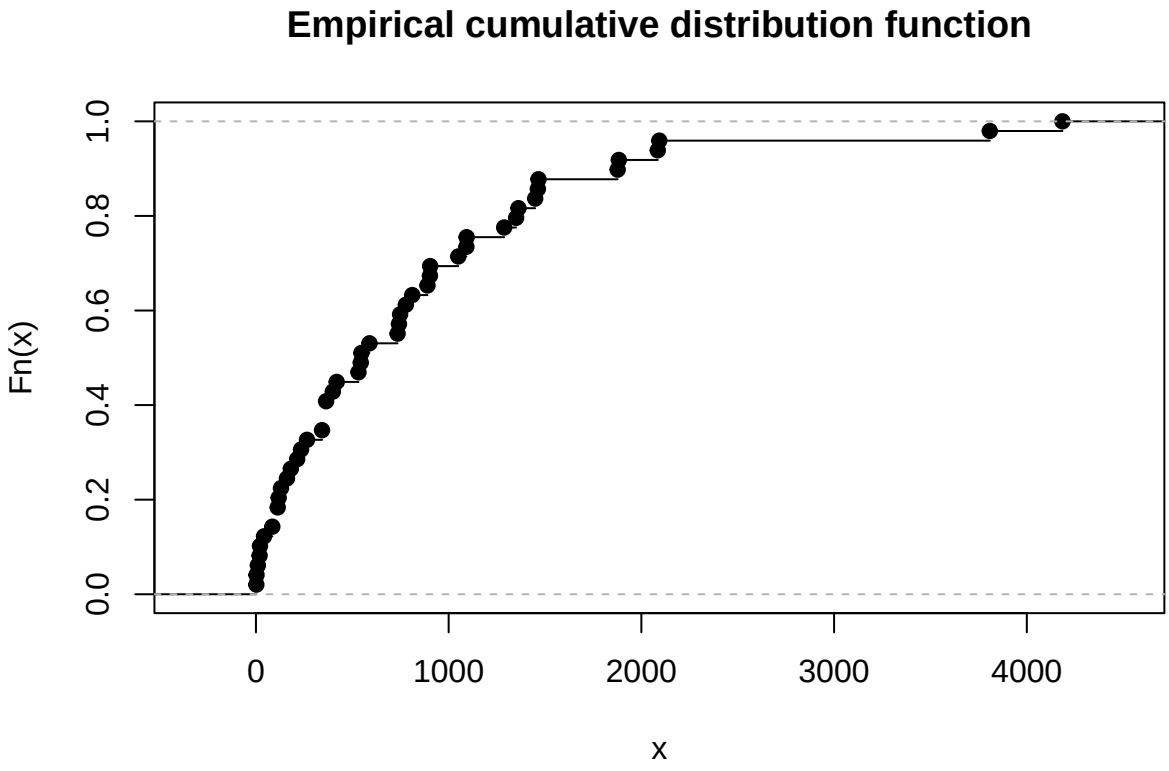
```
##        rate
## 0.001217966
```

Exponential Distribution Model Fit

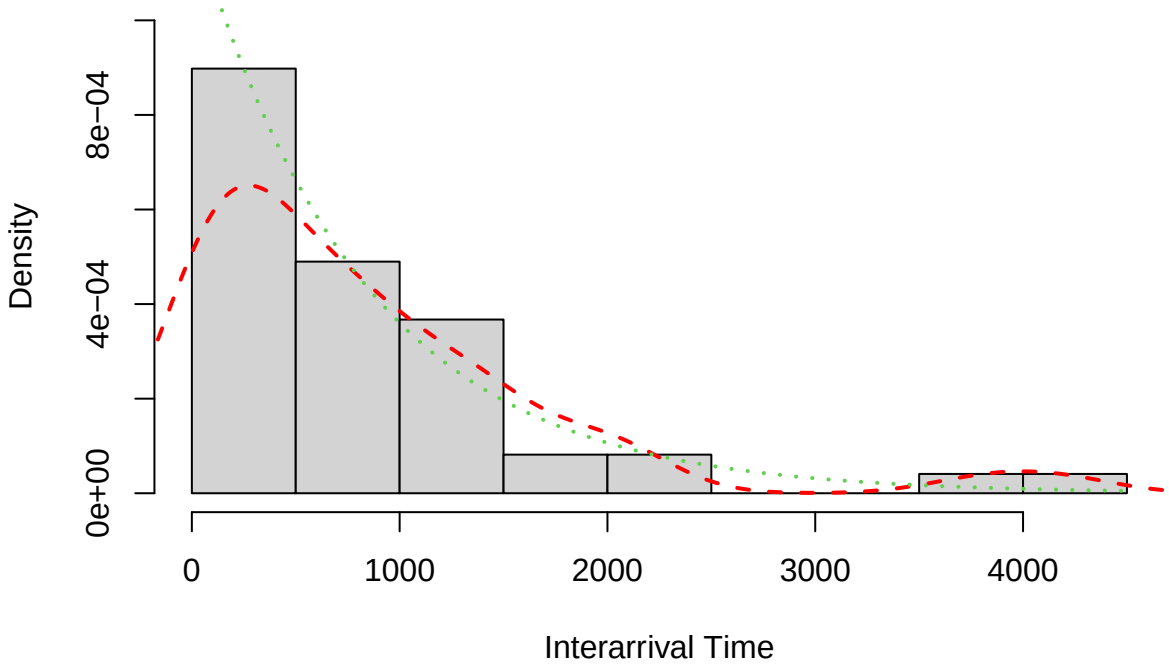
```
Ecdf(days_between,xlab='Days between records')
```



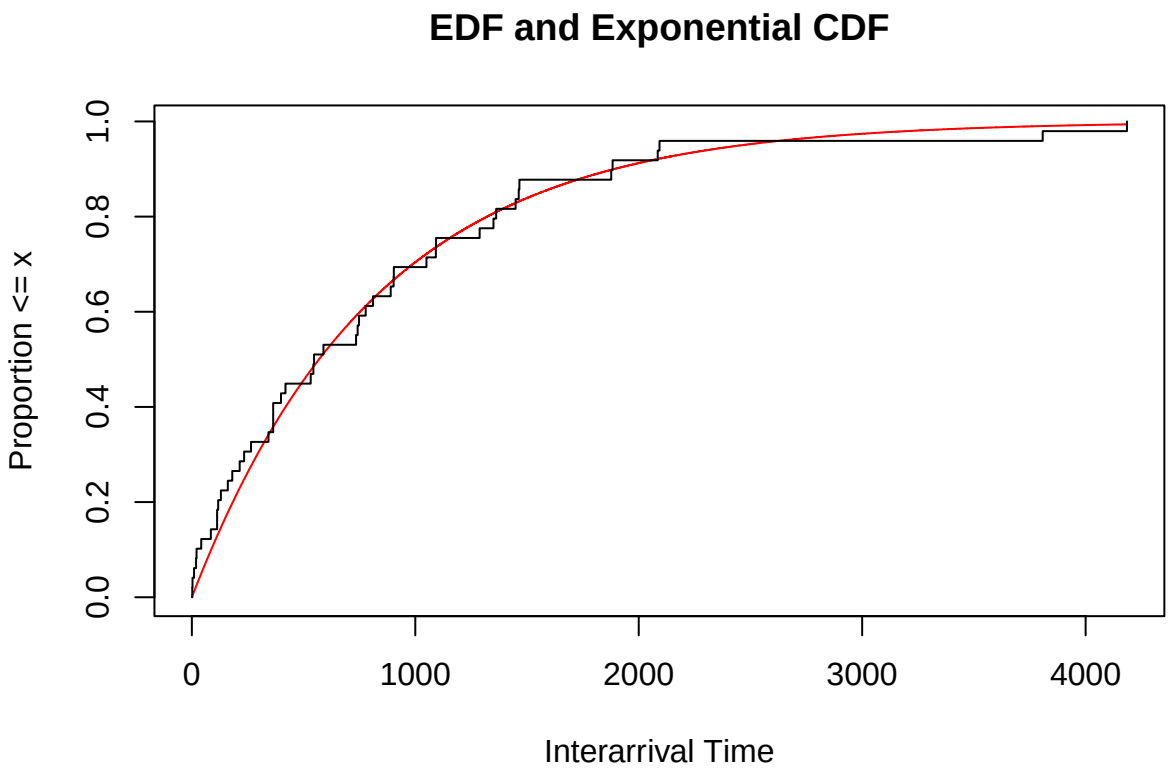

```
plot(ecdf(days_between),main="Empirical cumulative distribution function")
```



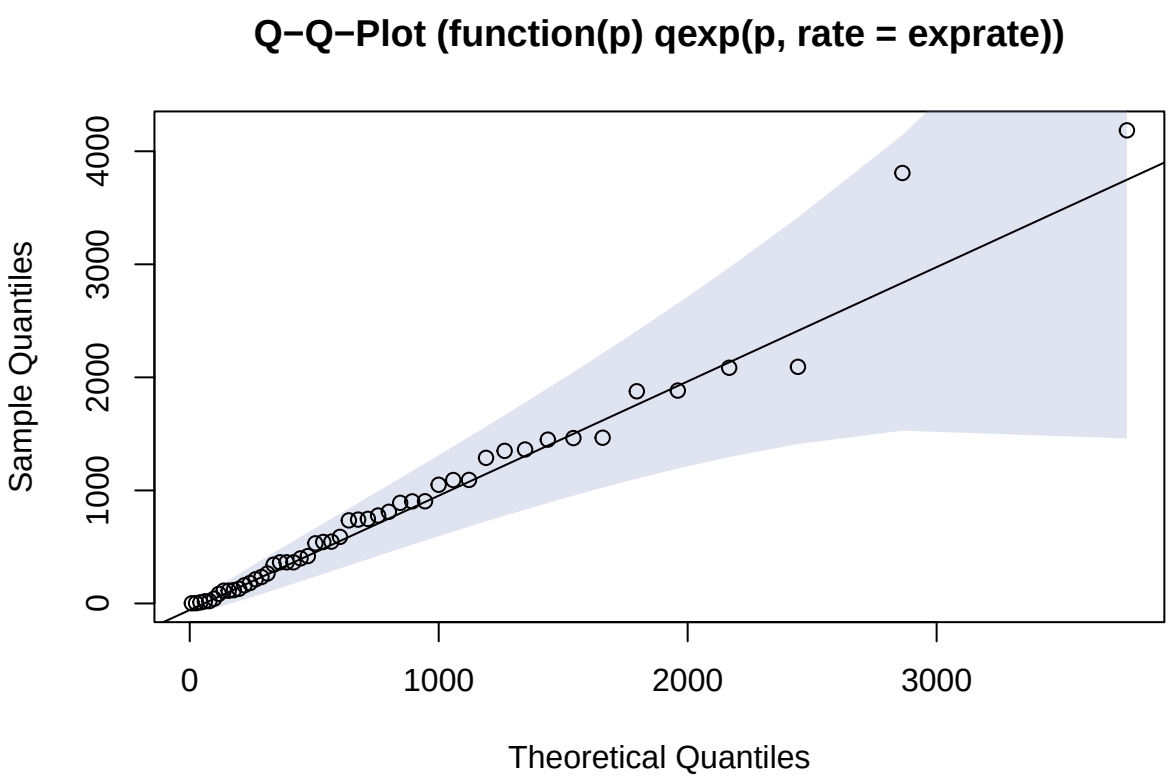

```
x=density(days_between)
hist(days_between,main="Histogram, density curve and exponential model",xlab="Interarrival Time",freq=F
lines(x,col="red",lty = 2, lwd = 2)
curve(dexp(x, rate = exprate),col=3, lty = 3, lwd = 2, add = TRUE)
```




```
x=seq(0,max(days_between),0.1)
plot(x,pexp(x,rate=exprate),type="l",col="red", main="EDF and Exponential CDF",xlab="Interarrival Time"
Ecdf(days_between,xlab='Interarrival Times',subtitle=FALSE,add=TRUE)
```




```
PlotQQ(days_between, function(p) qexp(p, rate=exprate))
```



/2022−08−15

```
# K-S Test
ks.test(days_between,pexp,rate=exprate)
```

```
## Warning in ks.test(days_between, pexp, rate = exprate): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  days_between
## D = 0.078053, p-value = 0.9264
## alternative hypothesis: two-sided
```

```
cvm.test(days_between,"pexp",rate=exprate,estimated = TRUE)
```

```
##
##  Cramer-von Mises test of goodness-of-fit
##  Braun's adjustment using 7 groups
##  Null hypothesis: exponential distribution
##  with parameter rate = 0.0012179662449355
##  Parameters assumed to have been estimated from data
##
## data:  days_between
## omega2max = 0.19159, p-value = 0.9074
```

```
ad.test(days_between,"pexp",rate=exprate,estimated = TRUE)
```

```
##
##  Anderson-Darling test of goodness-of-fit
##  Braun's adjustment using 7 groups
##  Null hypothesis: exponential distribution
##  with parameter rate = 0.0012179662449355
##  Parameters assumed to have been estimated from data
##
## data:  days_between
## Anmax = 2.8982, p-value = 0.2051
```

Using and Interpreting the Model