# Marathon Data

Sturdivant and Clark

2022-08-03

## Motivational Video

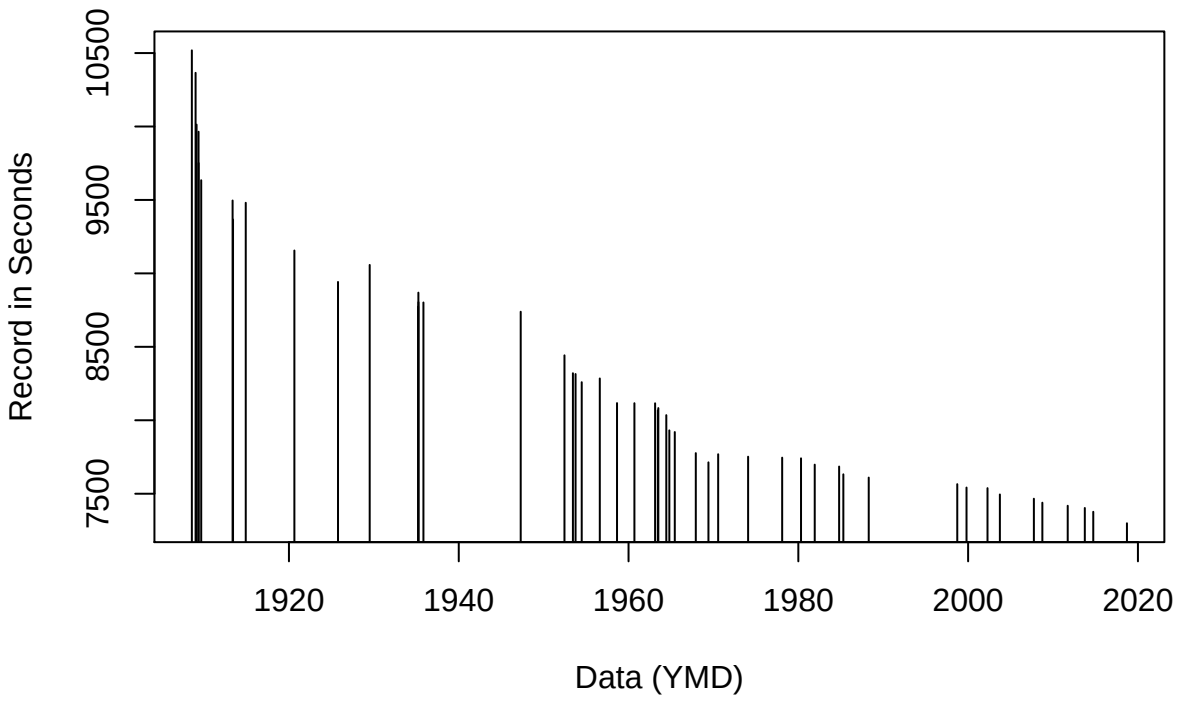
### Are we living in a time of records?

- Include NY Times article screenshot of headline.
- Brief summary of article's premise. ### How can we address this question? ### What would randomness look like?

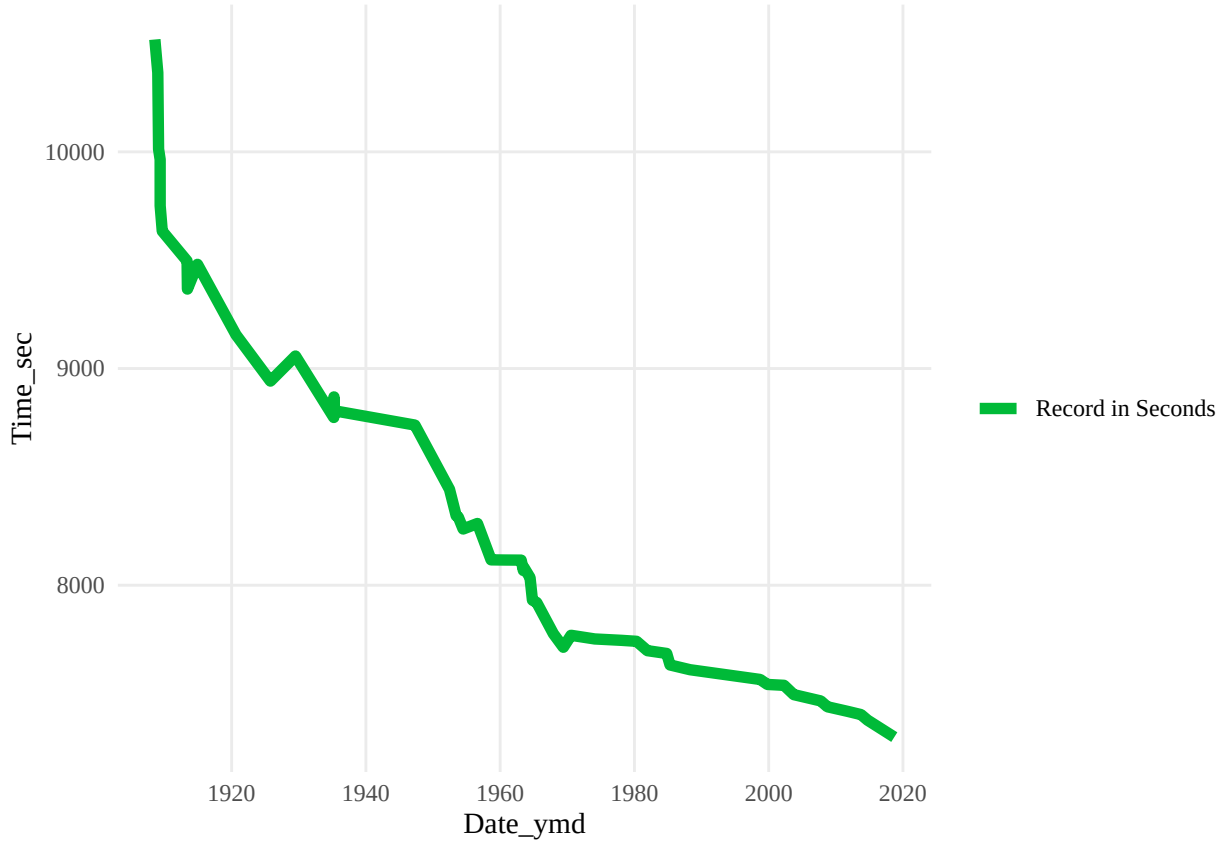
## Getting and Visualizing Data

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.

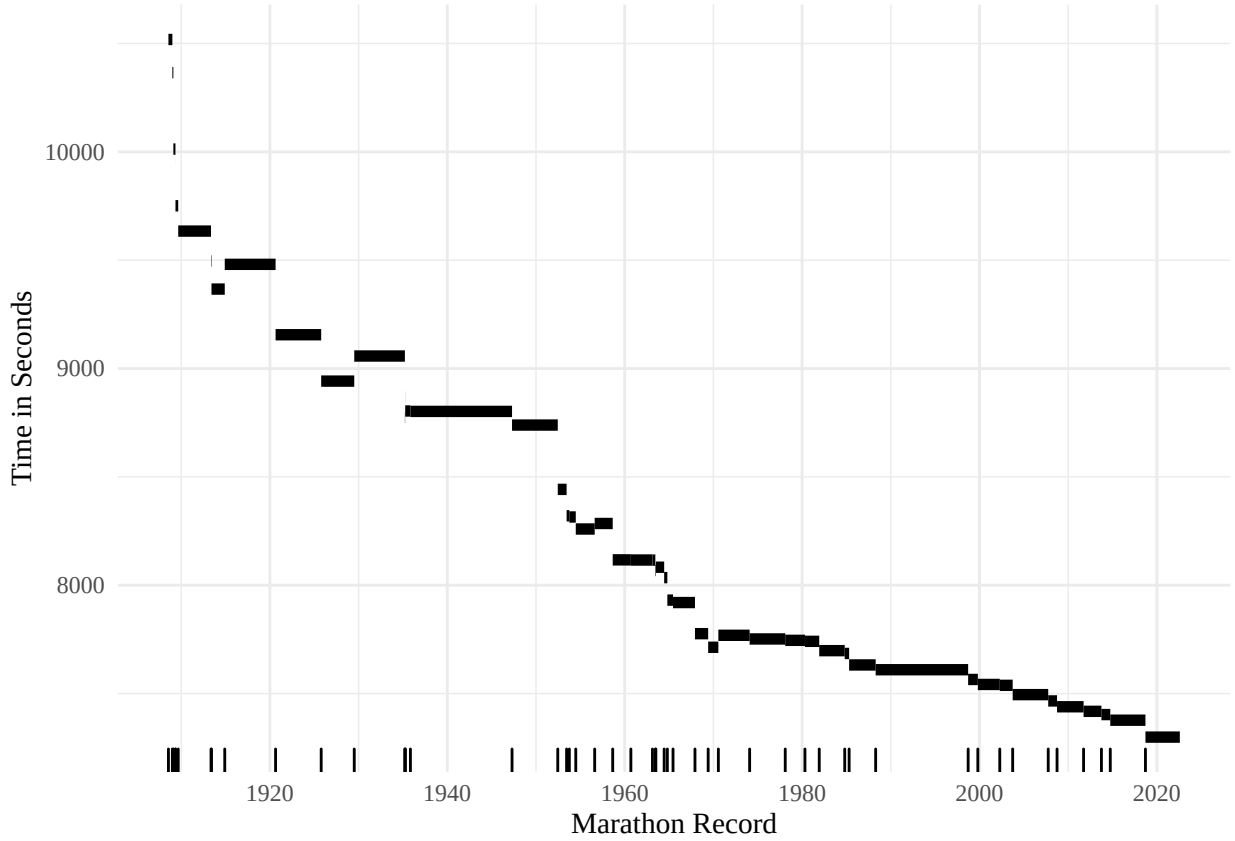
```
record_table_mod<-read_rds("record_table_mod.rds")

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? 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? What are possible variables you could model for this data? What type of data is each variable? Which would best help answer the question posed?

## 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 following 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 $\lambda$ is the rate: the average number of events per unit of time. Note the possible values of T are greater than 0 (positive only).

NOTE HERE: we guide them to the interarrival times... could have another module where they attempt to see if the Poisson distribution describes counts of the number of records in intervals - this would be interesting as it would potential more easily find periods with unusually high/low numbers.

Questions to explore:

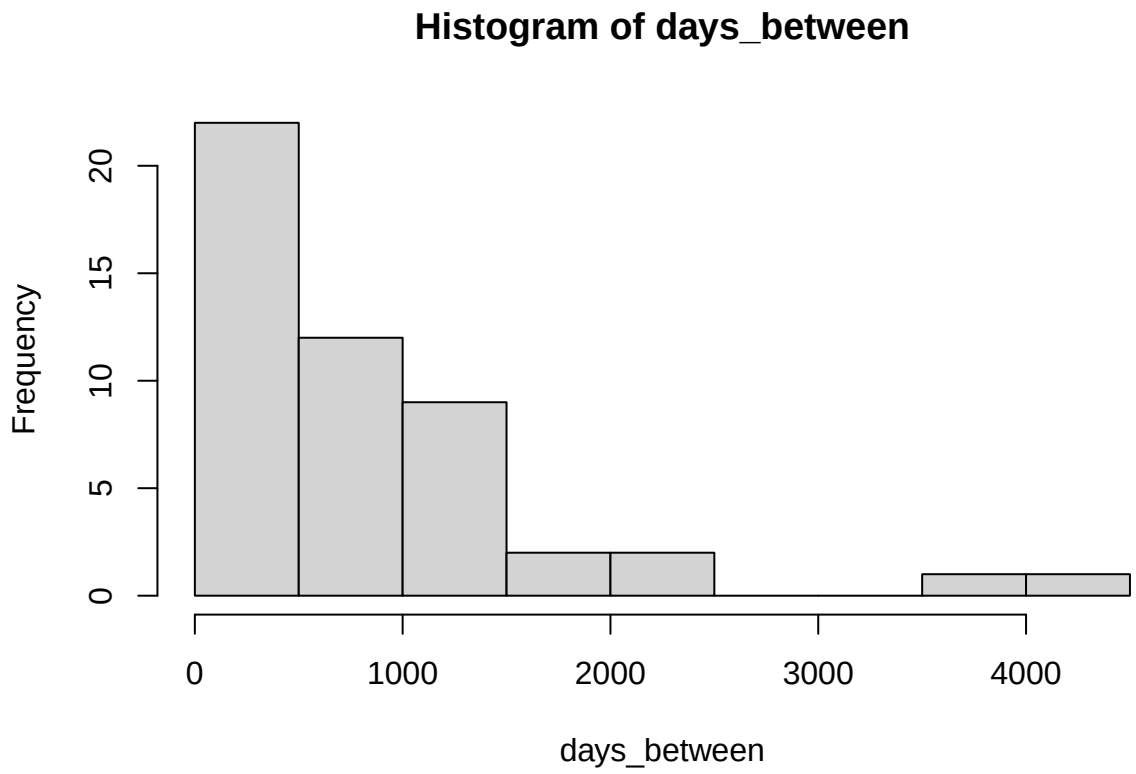
Are the assumptions of the Poisson process reasonable for the marathon record data? Explain.

How might you determine if the exponential model "fits" the marathon data?

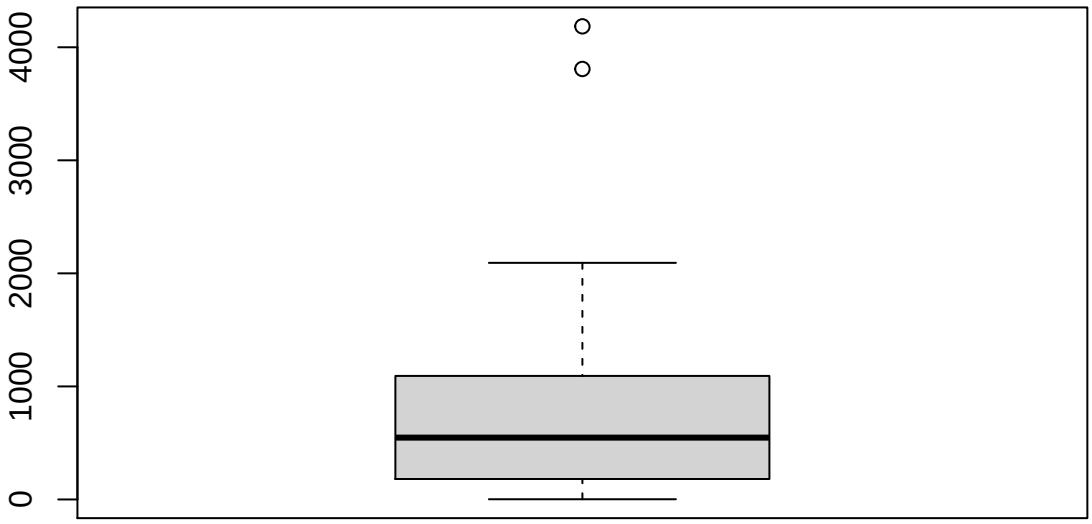
**Exponential Distribution for Interarrivals**

The variable "days_between" is the elapsed time (in days) between world records. Produce a plot (or plots) of this variable.

```
days_between = as.numeric(diff(record_table_mod$Date_ymd))
hist(days_between)
```




```
boxplot(days_between)
```




Questions:

- Does the plot seem consistent with an exponential distribution (provide link to information about exponential here or have them look it up)?
- Why would a distribution such as the normal distribution not be appropriate for this data?

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 $\lambda$

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 $\lambda$?

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 $\lambda$ extracted.

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

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

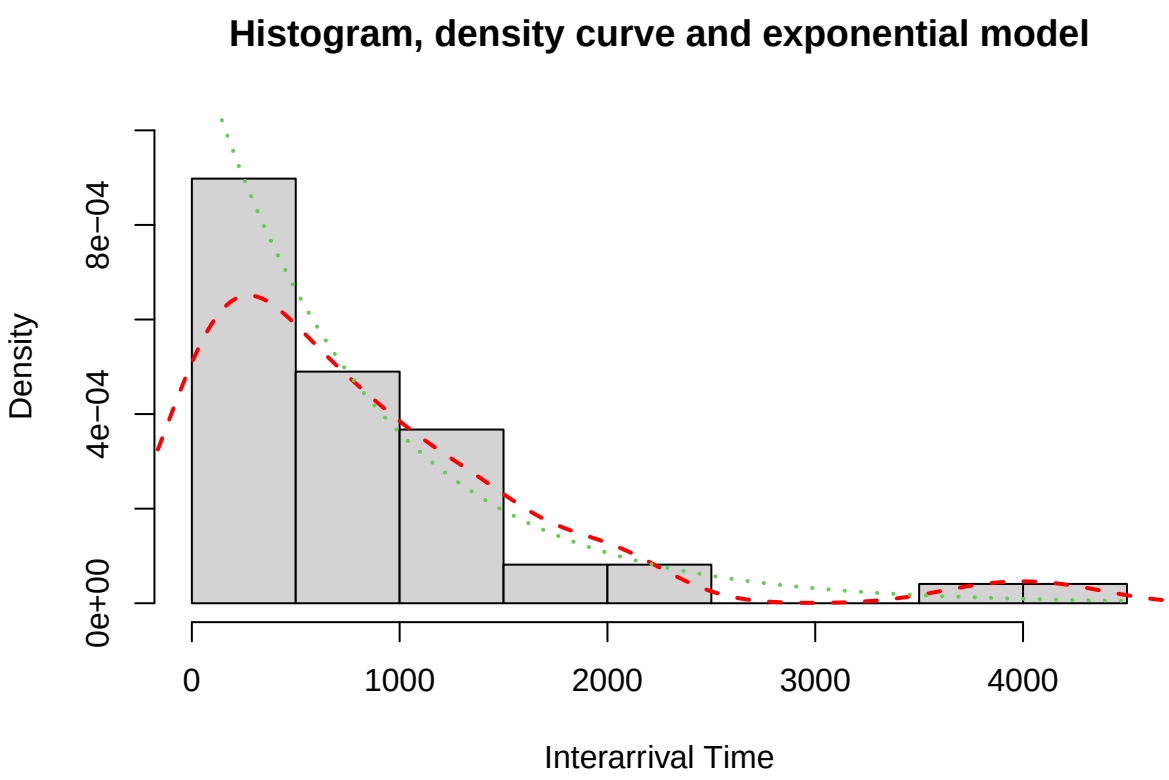
Questions:

- How does this compare to your estimate of the mean above?
- What are the units of the rate and the mean?

### Exponential Distribution Model Fit

Once we have an estimate of the rate parameter, we can examine the exponential model and see how well it fits (matches) our observed data. One way to do this is graphically. Below we show the histogram of the data we examined earlier. The red dashed line is the "density" curve of the data - sort of a smoothed or continuous curve instead of the binned histogram based on the data. Finally, the green dashed line is the plot of the exponential model.

```
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)
```




Questions:

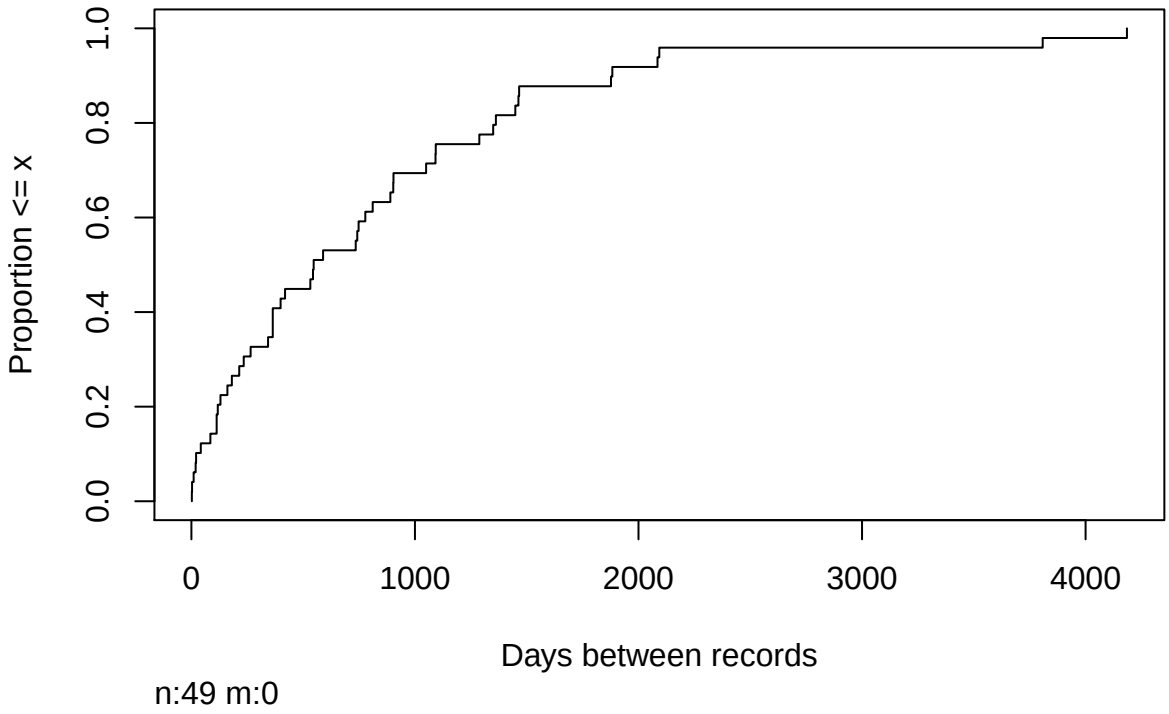
- Does the exponential model seem reasonable?
- What issues are there with using the provided graphs to assess model fit?

NOTE: here it might be fun to have an interactive graph where they can change the rate parameter and see the impact, and also maybe plot a different distribution (normal for example) and compare.

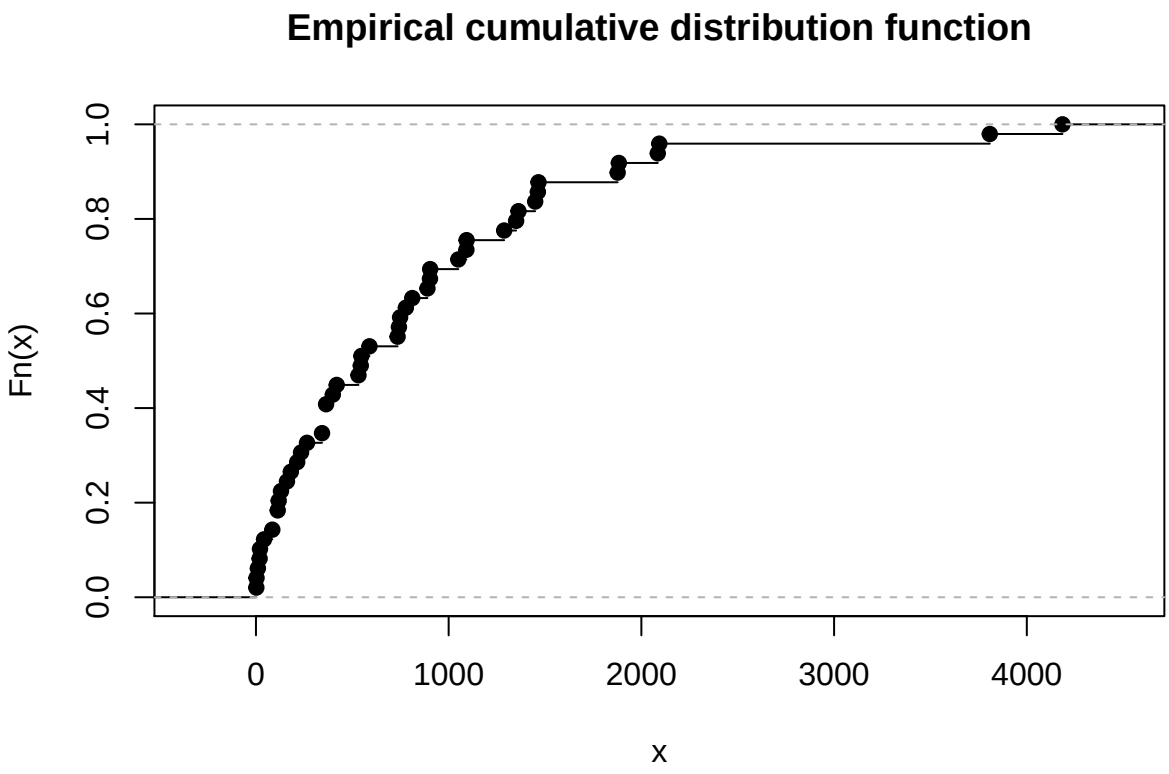
The histogram/density are not always the best for assessing model fit. Another common approach is to consider the cumulative distribution function (CDF). This function gives the P(T<t) for all values of t. There is an "empirical" version of this that is based solely on the actual data known as the ecdf ("Empirical CDF"). The plot below (NOTE: two versions here, we can just have one) shows this function for our data. (NOTE2: here we might have link to a page on ECDF and also CDF?)

As an example for reading this plot, when x = 1000 that is 1000 days between world records we see (reading the y value) that around 70 percent of the times between records were less than this value.

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




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

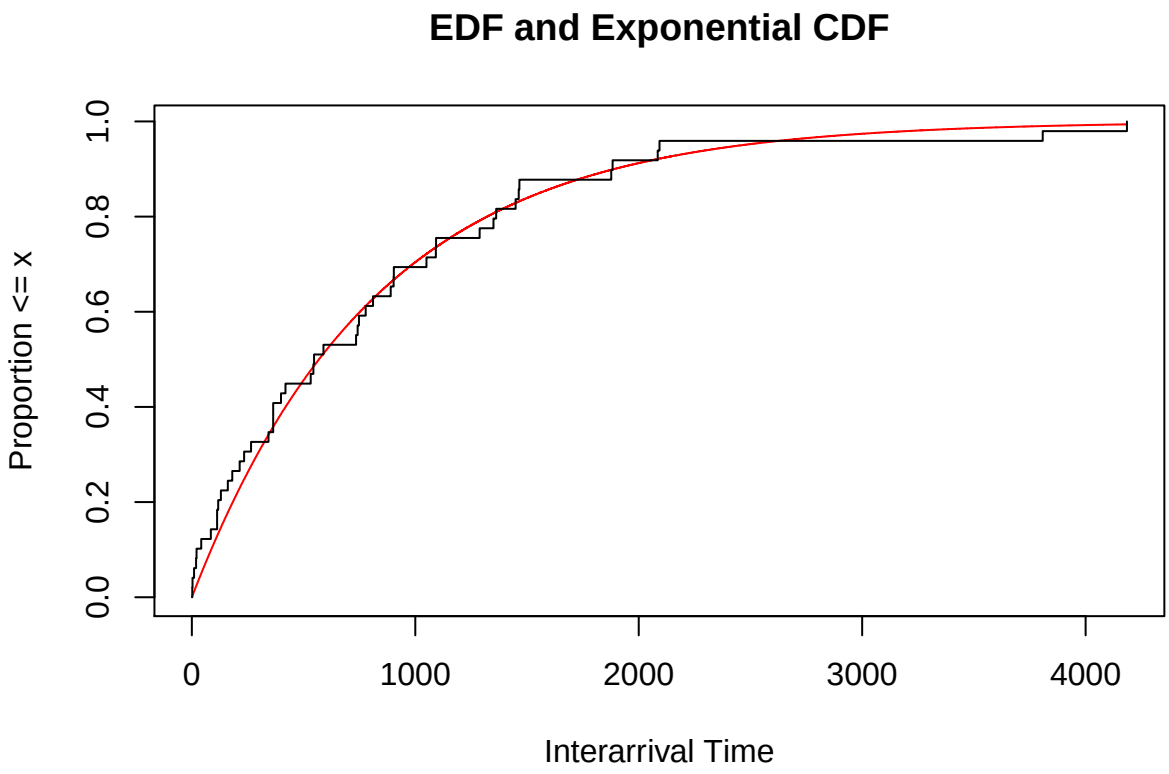



Questions:

- What percentage of the times between records were less than 4000 days?
- What percentage of the times between records were less than 0 days?
- What is the maximum value for y on this graph?
- What is (approximately) the median number of days - the value for which half the times were less?

The ECDF provides, in essence, the same information as the CDF but for the data. To see if our model fits well, we can compare the exponential CDF to the ECDF as in the graph below.

```
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)
```

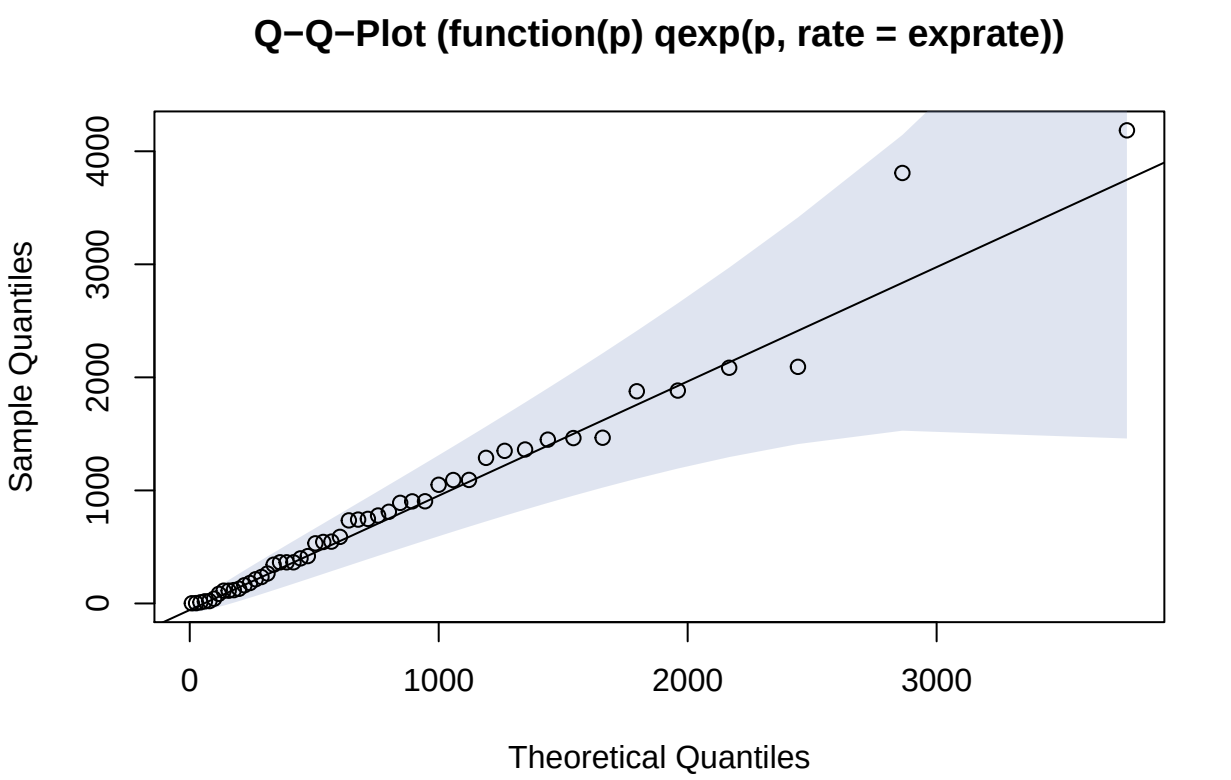



Questions:

- Does the model seem reasonable? (NOTE: again, might be nice to have ability to change parameter and distribution here)

Another plot that is sometimes used to assess model fit is known as the "QQ plot". It plots the theoretical quantiles (percentiles) of the proposed distribution and the observed data quantiles. If the model is appropriate, these values should be reasonably close so the plotted points should fall along the line y = x. This plot is shown below.

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




Questions:

- Based on the QQ plot is the model reasonable?
- Which values (if any) appear to fit least? Are the sample quantile higher or lower than the model suggests? Looking back at the plots of the original data, when did these occur?

Finally, there are formal tests of "goodness of fit". We will not discuss details (give them a link or two here) but they essentially do things like compare how close the ECDF is to the CDF. The key output of the test is the p-value. A small p-value (such as below 0.05) indicates that the model might not fit the data well. Three such tests are below, the Kolmogorov-Smirnov (KS), Anderson-Darling (AD), and Cramer-von Mises tests.

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

```
##
##  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 = 0.33215, p-value = 0.5517
```

```
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 = 3.2605, p-value = 0.1402
```

Questions:

- Find the p-values for each test:
- Based on the p-values does there appear to be evidence that the exponential model is a poor fit?

## Using and Interpreting the Model

Questions:

- Once we have the model, we can do things like compute probabilities of events. Compute probabilities that the next world record time is in the next year, next 2 years, next month.

```
pexp(365, rate = exprate)  # probability next world record occurs in the next year
```

```
## [1] 0.3588922
```

```
pexp(365*2, rate = exprate)  # probability next world record occurs in the next two years
```

```
## [1] 0.5889808
```

```
pexp(365/12, rate = exprate)  # probability next world record occurs in the next month
```

```
## [1] 0.03636865
```

- When do you think the 2 hour barrier will be broken? Can you answer this question directly with the model? If not, how might you do so?

### Communicating Results

Questions:

- Based on the modeling effort, how do you respond to the original question: have there been "times of records" or are these times still random?
- What are shortcomings of the model produced? What are other aspects of the data or other questions that you would like to answer that this model cannot? What ideas do you have for answering these questions? (NOTE: the 2 hour barrier question above is an example...)

## ADVANCED: Memoryless Property

One interesting property of the exponential distribution is that it is *memoryless*. What this means is that

$$P(X > x + t | X > t) = P(X > x)$$

If $X$ represents the time until the next world record, what does this mean in context of the marathon problem?

Let's prove this property of the exponential. Start by writing out the conditional probability. Recall that the definition of conditional probability is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Use this to express $P(X > x + t | X > t)$ as a fraction

Now we have to think a bit. Looking at our numerator, why can we simplify it to $P(X > x + t)$?

Recall that the **CDF** of a distribution is $F(a) = P(X < a)$. Therefore, can we express the numerator and the denominator as functions of the CDF?

Finally, use the fact that the CDF of the exponential is $F(t) = 1 - e^{-\lambda t}$ to prove that the numerator and the denominator are equal to $e^{-\lambda t}$.

Why does this prove the memoryless property?

Let's return to our problem. Do we think the memoryless assumption is valid? Why or why not?