Math 450 Homework 4
Geometric Optics, Probability modeling

due March 8, 2021

1. The parabolic solutions of the mirror equation can be derived directly using coordinate geometry if one postulates "Huygens' principle". Suppose a flash of light is emitted from the origin and reflected off a mirror so that all light rays travel straight up. Under Huygen's principle, all light rays travel at the same speed, and all light rays reflected off the mirror will reach a horizontal line above the focus at the same time. If the bottom of the mirror is 12 centimeters below the origin, use coordinate geometry to find an equation for the mirror.

   If a light ray bounces from the mirror at a point (x, y), origin to a line at height h, then a light ray bouncing off the bottom of the mirror and back up to the line will travel a distance $2 \times 12 + h = 24 + h$. Huygens' principle says that all other light rays must travel the same distance, so total distance travelled will be $\sqrt{x^2 + y^2} + (h - y) = 24 + h$. Now, using algebra,

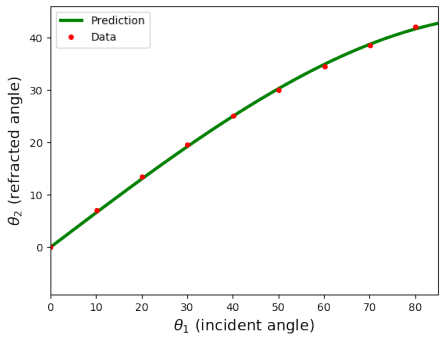
$$\sqrt{x^2 + y^2} = 24 - y \tag{1}$$
$$x^2 + y^2 = 24^2 - 48y + y^2 \tag{2}$$
$$y = (x^2 - 576)/48 \tag{3}$$

   The equation for the mirror is $y = (x^2 - 576)/48$.

2. Polish natural philosopher Vitello measured refraction angles experimentally in the 13th century. A close approximation of Vitello's data (in Exercise 4 in the "Geometric optics" lecture) relating angle of incidence to angle of refraction for light passing from air to glass is given below.

   (a) Use least squares to fit a curve of the form $\theta_2 = a_1\theta_1 + a_3\theta_1^3$ to Vitello's data.

   Recover $a_1 \approx 0.66$ and $a_2 \approx -2.2 \times 10^-5$. Visualizing the fit we find good agreement,



   (b) Solve Snell's Law (see lecture notes) for the angle of refraction θ_2 as a function of the angle of incidence θ_1, and calculate the McLaurin series this function to fourth order.

Snell's law states that

$$\frac{\sin\theta_1}{v_1} = \frac{\sin\theta_2}{v_2}.$$

Solving for θ_2 we find

$$\theta_2 = \sin^{-1}\left(\frac{v_2}{v_1}\sin\theta 1\right).$$

We then take the McLaurin series to fourth order and recover

$$\theta_2 = \frac{v_2}{v_1}\theta_1 + \frac{v_2(v_2^2 - v_1^2)}{6v_1^3}\theta_1^3 + O(\theta_1^5).$$

If we truncate at fourth order, we recover a polynomial which looks remarkably like the polynomial in (a) with $a_1 = v_2/v_1$ etc.

(c) Using your previous results, estimate the refractive index of glass relative to air. Explain your reasoning.

The refractive index of light passing from air to glass is given by

$$\frac{\sin(\theta_2)}{\sin(\theta_1)} = \frac{v_2}{v_1}$$

where θ_1 is the incidence angle in air, θ_2 is the refracted angle in glass, and v_1, v_2 are quantities from Snell's law (see (b)).
We can approximate v_2 and v_1 from the cubic fit in (a) using the truncated McLaurin series in (b):

$$a_1 = \frac{v_2}{v_1}$$
$$a_3 = \frac{v_2(v_2^2 - v_1^2)}{6v_1^3}$$

Notice that a_1 *is* the refractive index, v_2/v_1. **Therefore the refractive index from air to glass is approximately** 0.66**.**

3. Find the probability generating function for the distribution of the sum of the rolls of two tetrahedral (4-sided) dice.

$$\frac{x^8 + 2x^7 + 3x^6 + 4x^5 + 3x^4 + 2x^3 + x^2}{16}$$

4. The hypergeometric distribution for the probability of drawing exactly k white stones from an urn filled with K white stones and N black stones in m draws without replacing any stones is

$$P(k) = \frac{(N \text{ choose } n)(K \text{ choose } k)}{(N+K \text{ choose } n+k)}$$

where "x choose y" $= x!/(y!(x-y)!)$. Show that when the number of stones in an urn is taken to be large, but of fixed initial proportion, then the hypergeometric distribution for drawing a small number of stones from the jar is approximately a binomial distribution, and hence, approximately a Poisson distribution.

Re-expressing the hypergeometric formula with factorials,

$$P(k) = \frac{\frac{N!}{n!(N-n)!}\frac{K!}{k!(K-k)!}}{\frac{(N+K)!}{(n+k)!(N+K-n-k)!}}.$$

We need to consider this formula in the case of very large urns (N and K grow without bound) but the fraction of stones of each color is constant (so N/K is constant), while n and k are both small and fixed. In class, we saw that $a!/(a-b)!\ a^b$ as a becomes infinite if b is fixed. We can apply this formula directly above three times to get

$$P(k) = \frac{(n+k)!}{n!k!}\frac{N^n K^k}{(N+K)^{n+k}}.$$

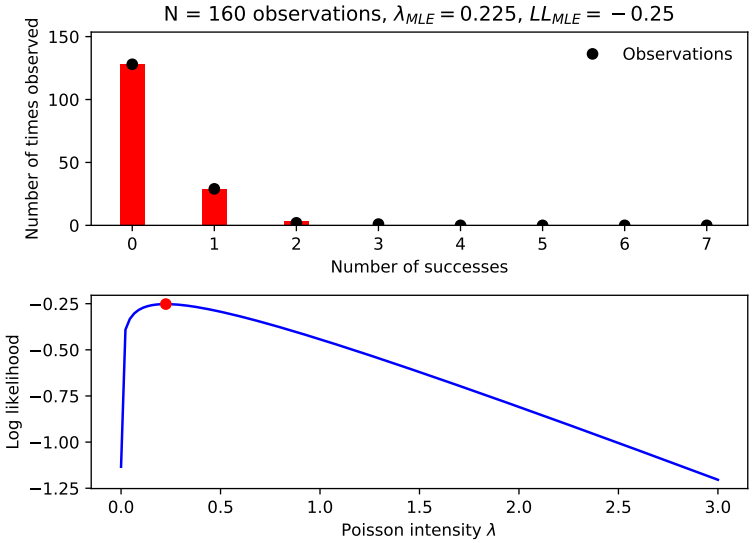
If we define $p = N/(N+K) = 1/(1+K/N)$, we know p will be fixed as N and K grow without bound because N/K is fixed. Then

$$P(k) = \frac{(n+k)!}{n!k!}p^n(1-p)^k,$$

which is a slightly less usual way to express the binomial distribution term for drawing n black stones and k white stones in $k+n$ draws.

5. Is the number of major hurricanes reaching Florida each year between 1850 and 2010 Poisson-distributed? Fit a Poisson distribution, report the best λ value, and discuss the fit.

   There are a few ways to get data to determine this. I suggest using the table of major hurrications (category 3 and greater), from which we can extract the years of the hurricanes, and count how many occur in each year. The fit seems good, but depending on your data, you might get a different intensity estimate and a different less-convincing fit. (we aren't a stats class, so we haven't delved into measures of goodness of fit) %



6. In 1975 observations, a subset of Penn State's female students were binned according to their height and asked to stand in order in front of Old Main (see below). Fit these data with a normal distribution.

We first construct the data from the figure.

Height (inches)	# of students
59”	2
60”	5
61”	7
62”	10
63”	16
64”	23
65”	20
66”	16
67”	9
68”	6
69”	6
70”	3
71”	1
72”	1

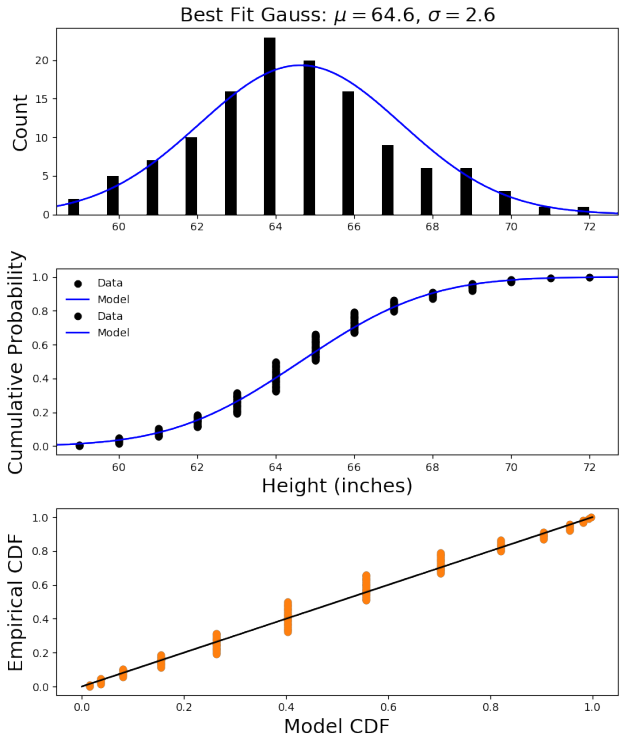
You may have counted slighly differently – that’s OK.

Then assume that the heights are normally distributed with mean μ and standard deviation σ. We use the maximum likelihood estimator of these parameters as derived in class,

$$\mu_{MLE} = \frac{1}{n}\sum_{j=1}^{n} X_j \approx 64.6''$$

$$\sigma_{MLE} = \sqrt{\frac{z - \mu_{MLE})^2}{n}} \approx 2.6''$$

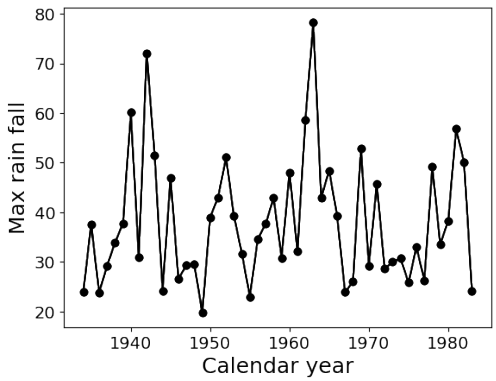
where $n = 125$ is the total number of students, and X_j is a vector of heights for each student (i.e. $X_1 = 59$, $X_2 = 59$, $X_3 = 60$, $X_4 = 60$, $X_5 = 59$, ...). We graph the results,



7. Exercise 6 in the "Horse kicks and height" lecture gives is a data set for maximum rainfalls in Brussels, Belgium over a 50 year period. Two possible models of this data are the normal distribution (a.k.a. the normal distribution) and Gumbel's distribution. Gumbel's distribution seems the better choice, a priori, because of its frequent use in risk analysis to describe extreme natural events, but we'd like to test this hypothesis ourselves.

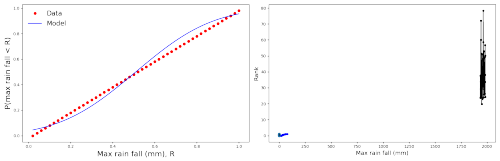
   (a) Plot the time series.

   Simply take the data and plot the max rain fall vs the year,

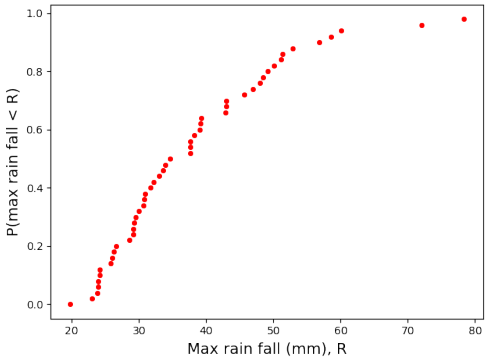


   (b) Plot the empirical cumulative distribution function (CDF) for the maximum daily rainfall.

   First sort the data, and then plot the rank (as position in the sorted vector) on the y-axis, against the max rain fall.



Then we have to normalize by dividing the rank by the total number of measurements, to obtain the *empirical* cumulative density function as required:



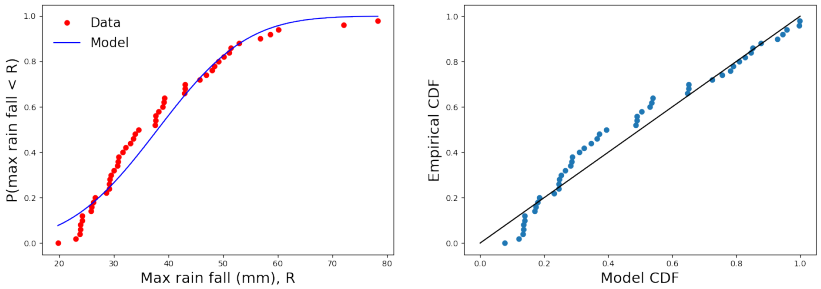
(c) Find the normal distribution under which the data are most likely.

We use the maximum likelihood estimators where the X_i corresponds to the max rainfall for year i.

$$\mu_{MLE} = \frac{\sum_{i=1934}^{1983} X_i}{\sum_{i=1934}^{1983} 1} = 38.038$$

$$\sigma_{MLE} = \sqrt{\frac{\sum_{i=1934}^{1983} (X_i - \mu_{MLE})^2}{\sum_{i=1934}^{1983} 1}} \approx 12.788$$

To visualize how well the normal distribution under which the data are most likely describes the data, plot the CDF up against the data.



Not great. Not bad, but not great...

(d) Use numerical minimization to find the form of Gumbel's distribution under which the data are most likely.

The Gumbel distribution is the 2-parameter distribution with PDF

$$f(x) = \frac{1}{\beta} \exp\left(-\frac{x-\mu}{\beta} + e^{-\frac{x-\mu}{\beta}}\right)$$

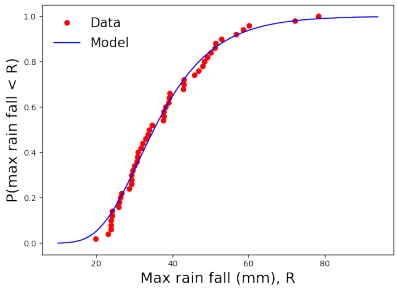
and CDF

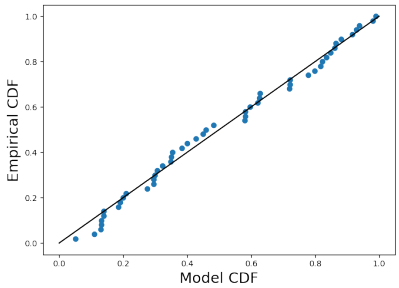
$$F(x) = \exp\left(\exp\left(-\frac{x-\mu}{\beta}\right)\right).$$

You're asked to use numerical minimization to find the most likely form of Gumbel's distribution. You should find

$$\mu \approx 31.28$$
$$\beta \approx 10.46$$

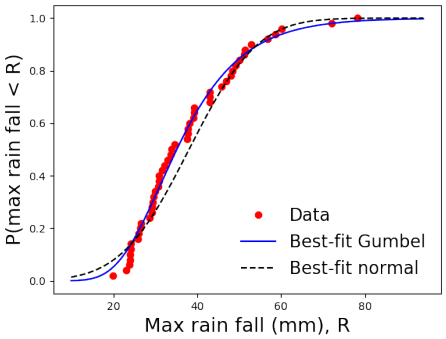
or thereabouts. The Gumbel CDF looks pretty good!





(e) Draw the empirical CDF, togethor with the best-fit normal CDF and the best-fit Gumbel CDF together on a single plot. Using this plot, discuss the quality of the distribution fits.

Comparing both fits,



Visually the Gumbel fit seems to do better for the bulk of the data, specifically the mid-range (most frequent) data. We can also attempt to quantify "better" in a few ways. One crude way is to check the sum-of-square differences between best-fit distributions and the data,

$$\text{Err} = \sum_{j=1935}^{1983} (X_j - F(X_j))^2,$$

where X_j is the max rainfall for year j and $F(X)$ is the CDF (normal or Gumbel). Find

$$\text{Err}_{Gaussian} \approx 3.39$$
$$\text{Err}_{Gumbel} \approx 0.04.$$

Pretty clear that the Gumbel distribution is a better fit!

8. Suppose a Poisson process is used to describe the pattern of cars passing by Lizzy as she waits to cross a one-way street, with λ being the frequency of cars per minute. Lizzy needs T seconds to cross the street. Find an expression for how long Lizzy should expect to wait before being able to cross.

Pick an arbitrary start time $t = 0$, and let n denote the n^{th} car crossing. The probability that the waiting time until the first car's arrival is $\leq t$ minutes is given by $1 - e^{-\lambda t}$. The probability that Lizzy can immediately cross is the probability that the waiting time is *greater* than $T/60$ minutes,

$$p = 1 - (1 - e^{-\lambda T/60}) = e^{-\lambda T/60}$$

where again λ is the frequency of cars per minute, and T is the number of seconds Lizzy needs to cross the street. The probability that Lizzy can cross after one car is $(1-p)p$, after two

cars is $(1-p)^2 p$, and after n cars is $(1-p)^n p$. Thus we have a probability distribution on which car arrival after which Lizzy can safely cross. But *when* does the n^{th} car arrival occur?

Well first, with probability p, Lizzy crosses before any cars come at all. With probability $1-p$ she has to wait until at least the first car. We saw in class that these arrivals are Poisson distributed, that is, the probability that there are $n-1$ arrivals in time t is given by

$$g_{n-1}(t) = \frac{(\lambda t)^{n-1} e^{-\lambda t}}{(n-1)!}.$$

We can re-interpret that as a probability density function on the time of the n^{th} arrival at time t *after we normalize it*,

$$q_n(t) = \frac{(\lambda t)^{n-1} e^{-\lambda t}}{(n-1)!} \left(\int_0^\infty \frac{(\lambda t)^{n-1} e^{-\lambda t}}{(n-1)!}\, dt \right)^{-1} = \frac{\lambda (\lambda t)^{n-1} e^{-\lambda t}}{(n-1)!}.$$

This, incidentally, is the Erlang distribution.

We don't know at *which* arrival Lizzy can cross, we just know the probabilities. Let's consider the two cases. With probability p, Lizzy doesn't have to wait; with probability $1-p$, she does – to get at that probability, $c(t)$, we combine the probability that it's after the n^{th} car arrival and the pdf on the time of the n^{th} car arrival,

$$\begin{aligned}
c(t) &= p \times 0 + \sum_{n=1}^{\infty} (1-p)^n p q_n(t) \\
&= \sum_{n=0}^{\infty} (1-p)^n p \left(\frac{\lambda (\lambda t)^n e^{-\lambda t}}{n!} \right) \\
&= \lambda p e^{-\lambda t} \sum_{n=0}^{\infty} \frac{((1-p)\lambda t)^n}{n!} \\
&= \lambda p e^{-\lambda t} e^{(1-p)\lambda t} \\
c(t) &= \lambda p (1-p) e^{-\lambda p t}
\end{aligned}$$

Note that $\int_0^\infty c(t)\, dt = 1-p$ and not 1. This makes sense since it's to be added to the 0 wait time, which occurs with probability p.

To find the expected crossing time, we compute the expected value,

$$\text{Expected value } = \int_0^\infty t c(t)\, dt = \frac{1-p}{\lambda p}.$$

Recall that we showed that $p = e^{-\lambda T/60}$. Thus Lizzy can expect to wait $(e^{\lambda T/60}) - 1)/\lambda$ minutes before crossing the street.

This is where I expected you to get – **but that's not quite right.** And here's how we can tell: $\lim_{\lambda \to 0} (e^{\lambda T/60}) - 1)/\lambda = T/60$, but it should be zero! If *no* cars come by, you should be able to cross immediately.

Why is this? Because we've left something out: up to the time of crossing, the waiting time is **truncated** and can't extend past T seconds! Actually correcting the calculations is onerous. We can explore this problem further via simulation (homework 5).

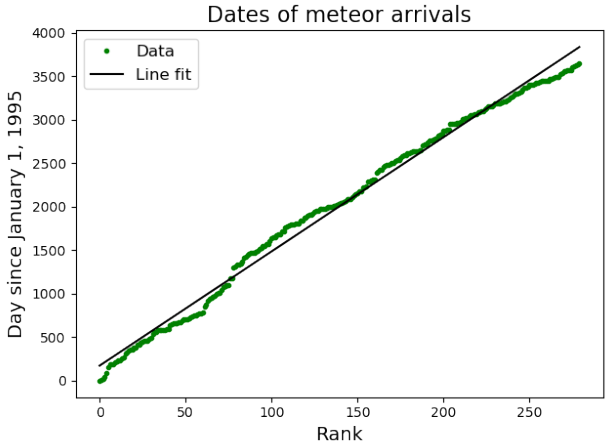
This was a deceptively difficult problem. I wanted to see some creativity and attempt at modeling. This one will be graded very lightly.

9. Here's your chance to do your own version of the meteor model.

a. Use the script from the website (in the Bolides section of the "Shooting stars" lecture() to retrieve a data set of dates and times of observations of meteor arrivals.

Implicit that you've done this through the rest of the problem!

b. Make a rank-arrival plot of the data from 1995 to 2005, where the independent variable is the rank of the arrival time (the first meteor is rank 1, the second to arrive is 2, ...) against the actual arrival time. If that curve is a perfectly straight line, that means the meteors are arriving at a constant rate with a regular period in between. If the curve is a little irregular but still mostly straight, that means the time between arrivals has some randomness, but the average arrival rate is constant. If this plot was more of steps or a zig-zag, that would mean meteors were arriving in clusters. If the plot is concave or convex, that would suggest a changing rate of arrival.

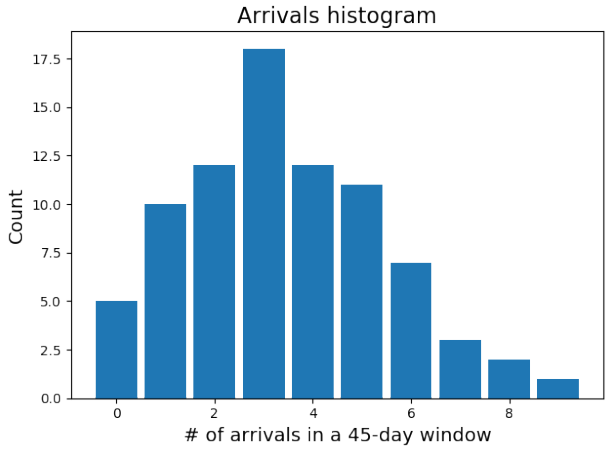


c. Based on the total number of meteor arrivals from 1995 to 2005, what was the average arrival rate (per day)?

Based on the total, the average arrival rate is

$$\frac{\text{num hits}}{\text{duration}} \approx 0.0767 \text{ per day.}$$

d. Plot a histogram of the number of meteors arriving every 45 days over the same 10 years.



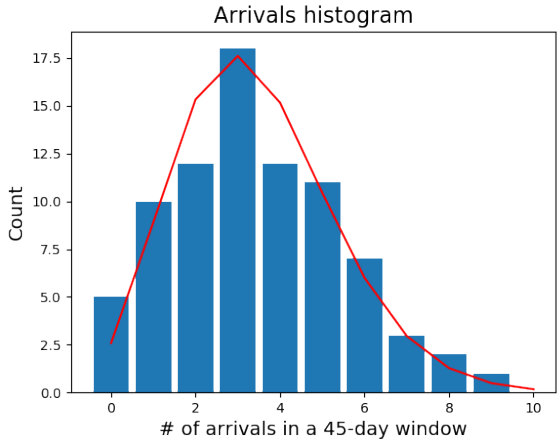
e. Fit a Poisson distribution to this histogram and estimate the Poisson distribution's intensity parameter λ. Plot your your histogram and your fit Poisson distribution on top of each other so we can compare them.

Use the maximum likelihood estimate for the Poisson intensity,

$$\lambda_{MLE} = \frac{\sum_{n=0}^{9} nX_n}{\sum_{n=0}^{9} X_n}$$

where the X_n are the number of times there are n arrivals in a 45-day window. Obtain

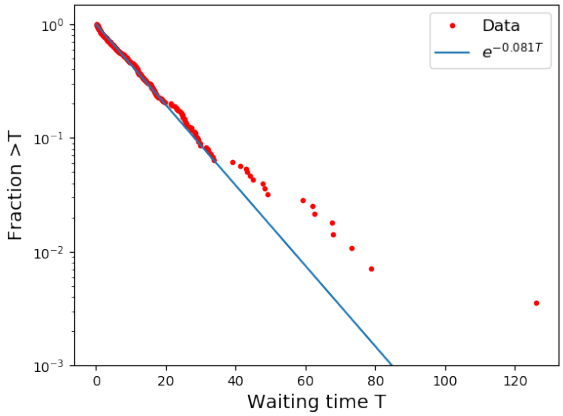
$$\lambda_{MLE} = \frac{\sum_{n=0}^{9} nX_n}{\sum_{n=0}^{9} X_n} \approx 3.44 \text{ per 45-day window.}$$



f. If the meteors are really arriving according to a Poisson process, the times between meteor arrivals should be exponentially distributed. Are they?

To address this problem, follow the argument in the notes showing that thewaiting times between arrivals from 2006-2016 are exponentially distributed.

You'll end up with a graph of $P(\Delta t > T)$ vs waiting time T as



Notice that the exponential fit does well with most of the data, but doesn't do well for large waiting times (I neglected those 17 values in the curve fitting here). Maybe because those large waiting times are rare, so the sampling isn't sufficient to resolve that part. Anyway, yes the exponential distribution does a reasonable job explaining the data.