CMPT 815
Computer Systems and Performance Evaluation
Assignment One

SOLUTIONS

1. A web server has "session" interarrival times that are independent and identically distributed, following an exponential distribution with parameter λ = 5 sessions/minute.

   (a) (*2 marks*) What is the mean interarrival time, in seconds?

   $1/\lambda$ = (1/5 minutes)(60 seconds/minute) = 12 seconds

   (b) (*2 marks*) Suppose that we observe the system at some randomly chosen time T, and find that 6 seconds have elapsed since the last arrival. What is the mean time from T until the next arrival?

   Since the exponential distribution is memoryless, the expected time until the next arrival does not depend on how long it has been since the previous arrival. Thus, the expected time until the next arrival is just the mean interarrival time of 12 seconds.

   (c) (*2 marks*) Give the mean number of session arrivals in a randomly chosen period of duration 2 minutes.

   (5 sessions/minute) x (2 minutes) = 10 sessions

   (d) (*2 marks*) Give the mean number of session arrivals in a period of duration 3 minutes, **conditional** on the **first of these arrivals** occurring exactly 18 seconds into the period.

   1 + (3 - 0.3 minutes)(5 sessions/minute) = 14.5 sessions

2. Consider the following random variables:

   W: exponential with parameter λ;
   X: Erlang with 3 exponential stages, each with parameter γ;
   Y: Lognormal with parameters μ =1.0, and σ;[1]
   Z: Pareto with minimum value k, and shape parameter of 1.25.

   Suppose that each of these random variables has mean value 12.

---

[1] I.e., lnY is normally distributed with mean value μ and variance σ^2. Note that the mean value of the lognormal distribution with parameters μ, σ is given by $e^{\mu + (\sigma^2/2)}$, and Excel, for example, has a built-in function for the lognormal cumulative distribution function.

(a) (*4 marks*) What are the values of λ, γ, σ, and k?

$E[W] = 1/\lambda = 12$ so $\lambda = 1/12$
$E[X] = 3/\gamma = 12$ so $\gamma = 3/12 = 0.25$
$E[Y] = e^{1+(\sigma^2/2)} = 12$ so σ = square root of $2(\ln(12) - 1) \approx 1.723$
$E[Z] = 1.25k/(1.25 - 1) = 12$ so $k = 2.4$

(b) (*4 marks*) For each random variable, give the probability that it is less than the mean.

$F_W(12) = 1 - e^{-12/12} \approx 0.632$
$F_X(12) = 1 - (1 + (0.25)(12) + ((0.25)(12))^2/2)e^{-(0.25)(12)} \approx 0.577$
$F_Y(12)$ = (in Excel) LOGNORMDIST(12, 1, 1.723) $\approx$ 0.806
$F_Z(12) = 1 - (2.4/12)^{1.25} \approx 0.866$

(c) (*4 marks*) For each random variable, give the probability that it exceeds 240.

$1 - F_W(240) = e^{-240/12} \approx 2.06 \times 10^{-9}$
$1 - F_X(240) = (1 + (0.25)(240) + ((0.25)(240))^2/2)e^{-(0.25)(240)} \approx 1.63 \times 10^{-23}$
$1 - F_Y(240)$ = (in Excel) 1 - LOGNORMDIST(240, 1, 1.723) $\approx 4.66 \times 10^{-3}$
$1 - F_Z(240) = (2.4/240)^{1.25} \approx 3.16 \times 10^{-3}$

3. (*4 marks*) Consider a content provider with a large collection of 10,000,000 available items (for example, user-generated videos). Suppose that the popularities of these items can be modeled as following a (bounded) Zipf(1.25) distribution. If the most popular item is accessed at a rate of 1 request/second, what percentage of requests (in total) are to the "cold" items with access rate less than 1 request/day? Repeat, but now assuming a Zipf(0.85) distribution.

% of requests to cold items = sum from i = n* to 10,000,000 { $1/i^\alpha$ }
-- x 100%
sum from i = 1 to 10,000,000 { $1/i^\alpha$ }

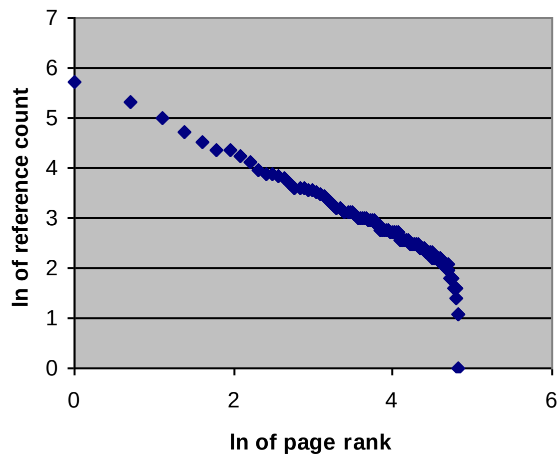
where n* = the smallest i such that $i^\alpha > 60 \times 60 \times 24$
For $\alpha = 1.25$, percentage of requests to cold items $\approx$ 7.53 %.
For $\alpha = 0.85$, percentage of requests to cold items $\approx$ 36.8 %.

4. (*6 marks*) On the class web site, there is a file giving some artificially-generated data on page reference counts for a hypothetical Web server. The data was generated by sampling from a (bounded) Zipf(α) distribution, with 125 pages and a total of 3000 references. Give a graph with the natural logarithm (base e) of the page popularity rank (based on the observed page reference counts) on the x-axis and the natural logarithm of the page access frequency on the y-axis, and from your graph estimate the value of α used when generating the data. (**Note:** the reference count for the *i*'th page in the file was incremented whenever the value *i* was sampled from the Zipf(α) distribution. However, sometimes by chance a page *j* ended up with more references

than a page *k*, for *j* > *k*, and so when making your graph you will need to sort the data so that pages are ordered according to the actual reference counts.)

A value of α = 0.8 was used when generating the data. The value of α can be determined from the negative of the slope (i.e, [ln(refs to page (a)) - ln(refs to page (b))]/[ln(rank of page (b)) - ln(rank of page (a))]).

Note that the graph tails off somewhat for large page ranks, even though the data was generated exactly according to a Zipf(0.8) distribution. This is simply because of unpopular pages that happen to receive even less references than they should, by chance, within the finite (length 3000) sequence of references.



5. On the class web site, there are two traces of client request arrival times to a media server. Each trace file contains one client arrival time per line, formatted as hh:mm:ss, in the order the arrivals occurred. The sep27h12-16 trace records the arrivals that occurred between 12:00 noon and 4:00pm on Sept. 27^{th}, 2000. The sep27 trace records all arrivals that occurred during a full 24-hour period from 3:00am on Sept. 27^{th} to 3:00am on Sept. 28^{th}.

   (a) (*6 marks*) For each trace, give the mean interarrival time and the squared coefficient of variation of interarrival times. (The squared coefficient of variation is defined as the variance divided by the square of the mean; for the exponential distribution, for example, it is equal to 1.)

   Sep27h12-16 trace: mean interarrival time is approximately 40.7 seconds; squared coefficient of variation is approximately 1.38.

   Sep27 trace: mean interarrival time is approximately 42.4 seconds; squared coefficient of variation is approximately 9.56.

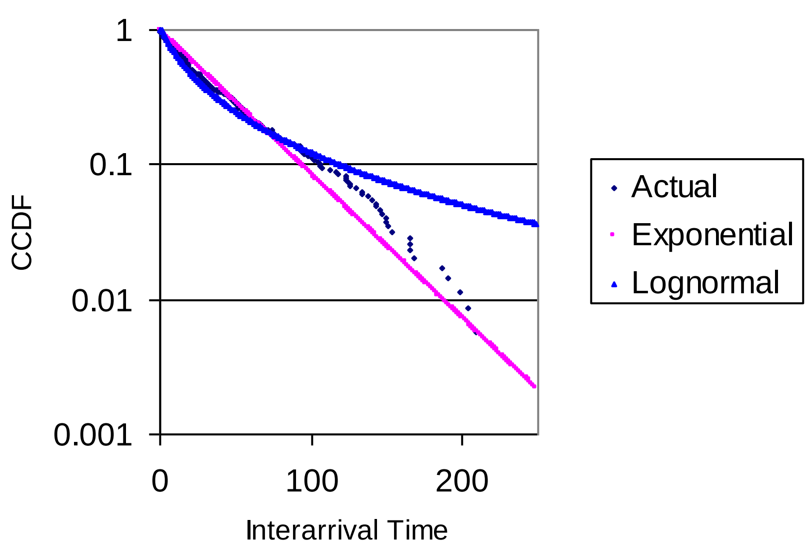
   (For finding the mean interarrival time, for example, calculate the sum divided by the total number of interarrival times.)

   (b) (*6 marks*) For the sep27h12-16 trace, create a graph with a base 10 log scale and maximum value 1 on the y-axis, and a linear scale from 1 to 250 seconds on the x-axis. On your graph plot: (i) the complementary cumulative distribution function (CCDF, 1-F(x)) of the request interarrival times, (ii) the CCDF of an exponential distribution with the same mean interarrival time as in the trace, and (iii) the CCDF of a lognormal distribution with parameters μ, σ chosen using *maximum likelihood* estimation. (Given n data points $x_1, x_2, \ldots, x_n$, the maximum likelihood estimators for the lognormal distribution parameters are $\mu = \frac{\sum \ln(x_k)}{n}, \sigma = \sqrt{\frac{\sum(\ln(x_k)-\mu)^2}{n}}$. When applying these formulas with the trace file data, treat any interarrival times that were recorded as 0 seconds in the trace file as interarrival times of 0.5 seconds instead.) Use symbols and spacings for the plotted points, and lengths of the x and y axes, that make the shapes of the curves easy to see and compare. Which one of the exponential and lognormal distributions provides a better fit to the data?

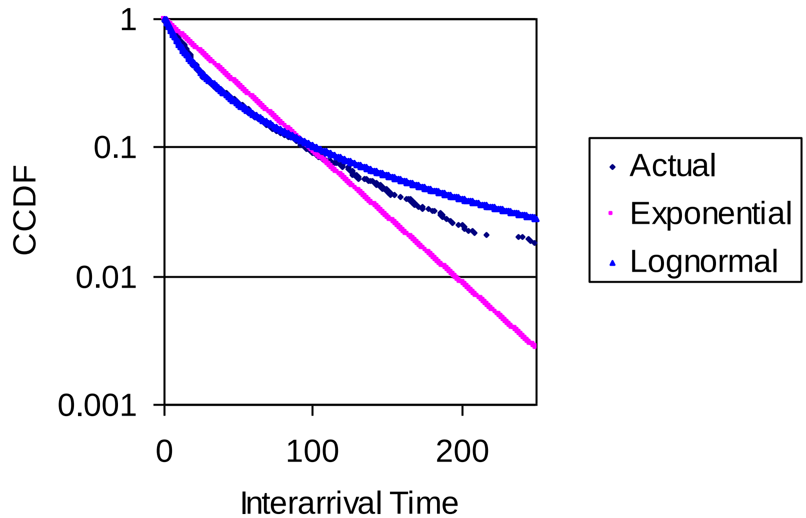
(c) (*6 marks*) Repeat part (b), but with the sep27 trace instead.

For the sep27h12-16 trace, the exponential distribution provides a better fit to the observed interarrival times. In contrast, for the full day sep27 trace, the heavy-tailed lognormal distribution provides a much better fit. This is, however, an artifact of including together data from both very light load early morning hours, and heavy load evening hours (i.e., non-stationarity).

A graph for the sep27h12-16 trace is as follows:



A graph for the full day trace is as follows:



1
0.1
0.01
0.001
CCDF
0
100
200
Interarrival Time
Actual
Exponential
Lognormal