CS 661, Spring 2008
Midterm Solutions

1. (a) The likelihood function $L(X_1, \ldots, X_n)$ is

$$L(X_1, \ldots, X_n; p) = \prod_{i=1}^{n} (1-p)^{X_i - 1} p = (1-p)^{\sum_{i=1}^{n} X_i - n} p^n.$$

To derive the MLE for p, we need to find the value of p that maximizes $L(X_1, \ldots, X_n; p)$. Equivalently, we can find the value of p that maximizes

$$\ln L(X_1, \ldots, X_n; p) = \left(\sum_{i=1}^{n} X_i - n \right) \ln(1-p) + n \ln(p).$$

Taking the derivative of $\ln L(X_1, \ldots, X_n; p)$ with respect to p and setting it equal to zero yields

$$\frac{d}{dp} \left[\left(\sum_{i=1}^{n} X_i - n \right) \ln(1-p) + n \ln(p) \right] = - \left(\sum_{i=1}^{n} X_i - n \right) \frac{1}{1-p} + \frac{n}{p} = 0,$$

which implies

$$- \left(\sum_{i=1}^{n} X_i - n \right) p + n(1-p) = 0.$$

Thus,

$$\left(\sum_{i=1}^{n} X_i \right) p = n,$$

so

$$p = \frac{n}{\sum_{i=1}^{n} X_i}.$$

Thus, the MLE of p is $\hat{p} = \frac{n}{\sum_{i=1}^{n} X_i}$.

(b) We first compute the first moment of a geometric random variable X as

$$\mathcal{E}[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1} p = p \sum_{k=1}^{\infty} k(1-p)^{k-1}.$$

Let $1 - p = q$ and note that $kq^{k-1} = \frac{d}{dq} q^k$. Therefore,

$$\mathcal{E}[X] = p \sum_{k=1}^{\infty} \frac{d}{dq} q^k = p \frac{d}{dq} \left(\sum_{k=1}^{\infty} q^k \right) = p \frac{d}{dq} \left(\frac{q}{1-q} \right) = \frac{p}{(1-q)^2} = \frac{1}{p}.$$

To obtain the method-of-moments estimator of p, we set the (theoretical) first moment $\mu = \mathcal{E}[X]$ equal to the first sample moment $\bar{X}_n = (1/n) \sum_{i=1}^{n} X_i$. Thus, we get the equation

$$\frac{1}{p} = \bar{X}_n,$$

and solving for p, we get the method-of-moments estimator of p to be $\hat{p} = 1/\bar{X}_n$, which is the same as the MLE.

2. There are several reasons supporting fitting an exponential distribution to the data:

   - The data are interarrival times to a website, and it is reasonable to assume that there are many different groups of users that are visiting the website. The time between visits from each individual group form an arrival stream to the website, and the superposition of the streams gives the total hits to the website. Thus, by the Palm-Khinchine Theorem, the superposition arrival stream is reasonably modeled as a Poisson process, which means the interarrival times are IID exponential.
   - The shape of the histogram looks like the density function of an exponential distribution.
   - The sample mean is reasonably close to the sample standard deviation, so the sample coefficient of variation is close to 1, which also supports an exponential distribution for the data since an exponential has a coefficient of variation of 1.
   - The K-S test has a p-value > 0.15, so we don't reject the null hypothesis at level 0.05 that the data comes from an exponential distribution since the p-value is greater than 0.05.

   The chi-square test has a p-value < 0.005, so the null hypothesis would be rejected, but the result doesn't make any sense since the number of degrees of freedom is 0.

3. Fitting a unif$(4.85, 6.66)$ to the data is completely inappropriate. A confidence interval is used to make inferences about the mean of a distribution; i.e., we are approximately 95% confident that the true mean of the distribution from which the data came lies somewhere between 4.85 and 6.66 since those are the endpoints of the 95% confidence interval that was constructed. Using this interval to fit the data completely misinterprets the meaning of a confidence interval.

4. In stage i, we need to generate a Weibull with shape parameter α_i and scale parameter β_i. The CDF of this distribution is $F_i(x) = 1 - \exp\{-(x/\beta_i)^{\alpha_i}\}$ for $x \geq 0$. We can generate a sample Y_i having this distribution by using inversion. Letting $U_i \sim \text{unif}[0, 1]$, we set $F_i(Y_i) = U_i$ and solve for Y_i:

$$1 - \exp\{-(Y_i/\beta_i)^{\alpha_i}\} = U_i,$$

   which yields $Y_i = \beta_i[-\ln(1 - U_i)]^{1/\alpha_i}$. (Alternatively, we can let $Y_i = \beta_i[-\ln(U_i)]^{1/\alpha_i}$.) After stage i, the worker continues with probability p_i, so we determine if the worker continues by generating $V_i \sim \text{unif}[0, 1]$, and the worker continues if $V_i \leq p_i$, and the worker stops if $V_i > p_i$. We need to make sure that all of the U_i and all the V_i are mutually independent. Putting this all together, we get the following algorithm:

0. Set $i = 1$ and $X = 0$.

1. Generate $U_i, V_i \sim \text{unif}[0,1]$, where U_i and V_i are independent and independent of all previous U_j and V_j, $j = 1, \ldots, i-1$.

2. Set $Y_i = \beta_i[-\ln(U_i)]^{1/\alpha_i}$ and let $X = X + Y_i$.

3. If ($V_i > p_i$ or $i = 4$), then return X and stop.
   Else, $i = i + 1$ and goto step 1.

An alternative (and more efficient) way of generating a sample of X is to use composition. Let I be a random variable denoting the total number of stages that are used, and note that I has the following probability mass function:

$$I = \begin{cases} 1 & \text{with probability } q_1 = 1 - p_1, \\ 2 & \text{with probability } q_2 = p_1(1 - p_2), \\ 3 & \text{with probability } q_3 = p_1 p_2 (1 - p_3), \\ 4 & \text{with probability } q_4 = p_1 p_2 p_3. \end{cases}$$

If $I = k$, then we return $X = Y_1 + Y_2 + \cdots + Y_k$, where each $Y_i \sim \text{Weibull}(\alpha_i, \beta_i)$, and $Y_1, Y_2, \ldots, Y_k$ are independent. Hence, the complete algorithm for generating a sample of X is as follows:

0. Generate $U_0 \sim \text{unif}[0,1]$.

1. If $U_0 \le q_1$, let $I = 1$.
   If $q_1 < U_0 \le q_1 + q_2$, let $I = 2$.
   If $q_1 + q_2 < U_0 \le q_1 + q_2 + q_3$, let $I = 3$.
   If $q_1 + q_2 + q_3 < U_0 \le q_1 + q_2 + q_3 + q_4$, let $I = 4$.

2. Let $Z = 0$.

3. For $j = 1, \ldots, I$, do
   - Generate $U_j \sim \text{unif}[0,1]$, independent of $U_0, U_1, \ldots, U_{j-1}$.
   - Let $Y_j = \beta_i[-\ln(U_i)]^{1/\alpha_i}$.
   - Let $Z = Z + Y_j$

4. Return $X = Z$.

5. First we need to compute the CDF F of X. Note that $F(x) = 0$ for $x < 2$. For $2 \le x < 3$,

$$F(x) \;=\; \int_2^x \left(\frac{3}{2}u - 3\right) du \;=\; \left[\frac{3}{4}u^2 - 3u\right]_{u=2}^{x} \;=\; \frac{3}{4}x^2 - 3x + 3,$$

and note that $F(3) = 3/4$. For $x \geq 3$,

$$\begin{aligned} F(x) &= \int_2^3 \left(\frac{3}{2}u - 3\right) du + \int_3^x e^{-4(u-3)} du \\ &= \frac{3}{4} + \left[-\frac{1}{4}e^{-4(u-3)}\right]_{u=3}^{x} = 1 - \frac{1}{4}e^{-4(x-3)}. \end{aligned}$$

Hence, the CDF F is

$$F(x) = \begin{cases} 0 & \text{if } x < 2, \\ \frac{3}{4}x^2 - 3x + 3 & \text{if } 2 \leq x < 3, \\ 1 - \frac{1}{4}e^{-4(x-3)} & \text{if } 3 \leq x. \end{cases}$$

To generate a sample of X using inversion, we set $F(X) = U$ and solve for X. For $0 \leq U < 3/4$, we solve for X in

$$\frac{3}{4}X^2 - 3X + 3 = U, \quad \text{or equivalently,} \quad \frac{3}{4}X^2 - 3X + 3 - U = 0,$$

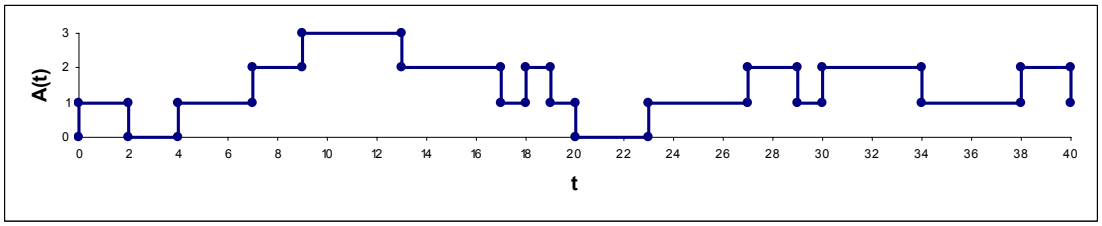
so $X = 2 \pm 2\sqrt{3U}/3$ by the quadratic formula. Since we need $2 \leq X < 3$ when $0 \leq U < 3/4$, we take $X = 2 + 2\sqrt{3U}/3$. For $3/4 \leq U < 1$, we solve for X in

$$1 - \frac{1}{4}e^{-4(X-3)} = U,$$

so $X = 3 - \ln(4(1-U))/4$. We cannot substitute U for $1-U$ in this case since we need $3/4 \leq U < 1$. Putting this all together, we get the following algorithm to generate samples of X:

1. Generate $U \sim \text{unif}[0, 1]$.
2. If $0 \leq U < 3/4$, return $X = 2 + 2\sqrt{3U}/3$.
3. If $3/4 \leq U < 1$, return $X = 3 - \ln(4(1-U))/4$.

6. (a) Below is a plot of $A(t)$ vs. t:



The times t at which $A(t)$ changes are 0, 2, 4, 7, 9, 13, 17, 18, 19, 20, 23, 27, 29, 30, 34, 38, 40.

(b) Calculating the area under the $A(t)$ plot from $t = 0$ to $t = 40$ and dividing by 40, we get

$$\bar{A}(40) = \frac{1}{40}[2 + 3 + 4 + 12 + 8 + 1 + 2 + 1 + 4 + 4 + 1 + 8 + 4 + 4] = \frac{29}{20}$$