CS 661, Spring 2010
Midterm Solutions

1. (a) The likelihood function $L_n(\theta)$ of an IID set of samples $X_1, \ldots, X_n$ is the product of the probability mass function evaluated at the n points, so

$$L_n(\theta) = \prod_{i=1}^{n} p(X_i) = \prod_{i=1}^{n} \binom{m}{X_i} \theta^{X_i}(1-\theta)^{m-X_i}.$$

Taking the log gives

$$\mathcal{L}_n(\theta) = \ln(L_n(\theta)) = \sum_{i=1}^{n} \left[\ln \binom{m}{X_i} + X_i \ln\theta + (m - X_i)\ln(1-\theta) \right].$$

We then compute the derivative of $\mathcal{L}_n(\theta)$ with respect to θ to get

$$\frac{d}{d\theta}\mathcal{L}_n(\theta) = \sum_{i=1}^{n} \left[\frac{X_i}{\theta} - \frac{m - X_i}{1-\theta} \right] = \left(\frac{1}{\theta(1-\theta)} \sum_{i=1}^{n} X_i \right) - \frac{mn\theta}{\theta(1-\theta)}$$

Then equating this to 0 and solving for θ yields

$$\hat{\theta} = \frac{1}{mn} \sum_{i=1}^{n} X_i$$

as the MLE for θ.

(b) For a gamma(α, β) distribution, the population mean is $\mu_1 = \alpha\beta$ and the population variance is $\alpha\beta^2$, so the population second moment is $\mu_2 = \alpha\beta^2 + \alpha^2\beta^2$. Let $M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ be the kth sample moment, $k = 1, 2$. Then set $\mu_1 = M_1$ and $\mu_2 = M_2$, giving us the equations $\alpha\beta = M_1$ and $\alpha\beta^2 + \alpha^2\beta^2 = M_2$, and we want to solve for α and β in these two equations. The first equation yields $\alpha = M_1/\beta$, and substituting this into the second equation gives $M_1\beta + M_1^2 = M_2$, so $\beta = (M_2 - M_1^2)/M_1$. Then substituting this into the first equation gives $\alpha = M_1^2/(M_2 - M_1^2)$. Thus, the method-of-moments estimators of α and β are $\hat{\alpha} = M_1^2/(M_2 - M_1^2)$ and $\hat{\beta} = (M_2 - M_1^2)/M_1$.

2. Define $Y_i = I\{X_i \le x\}$, and then $Y_1, \ldots, Y_n$ are IID since $X_1, \ldots, X_n$ are. Also, $\hat{F}_n(x)$ is the sample average of the Y_i. Since each Y_i is a Bernoulli random variable with $\mathcal{P}(Y_i = 1) = \mathcal{P}(X_i \le x) = F(x)$ and $\mathcal{P}(Y_i = 0) = P(X_i > x) = 1 - F(x)$, we have that $\mathcal{E}[Y_i] = F(x)$ and $\mathcal{V}[Y_i] = F(x)(1 - F(x))$. Also, $0 < \mathcal{V}[Y_i] < 1 < \infty$ since we assumed that x is chosen such that $0 < F(x) < 1$. Thus, for large n, the central limit theorem (CLT) implies that

$$\frac{\sqrt{n}}{\sqrt{\mathcal{V}[Y_i]}}(\hat{F}_n(x) - \mathcal{E}[Y_i]) \overset{\mathcal{D}}{\approx} \mathcal{N}(0, 1),$$

or equivalently that $\hat{F}_n(x) \overset{\mathcal{D}}{\approx} \mathcal{N}(\alpha_n, \beta_n^2)$ with $\alpha_n = \mathcal{E}[Y_i] = F(x)$ and $\beta_n^2 = \mathcal{V}[Y_i]/n = F(x)(1 - F(x))/n$.

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

   - The data are interarrival times of new phone calls to a switch, and it is reasonable to assume that there are many different phones that are arrive at the switch. The time between successive call initiations from each individual phone form an arrival stream to the switch, and the superposition of the streams gives the total calls routed through the switch. 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.

4. (a) We have that G_1 is an Erlang distribution with shape parameter k and scale parameter β. Recall that if Z has this distribution, then Z can be expressed as $Z = A_1 + A_2 + \cdots + A_k$, where the A_i are IID exponentials with mean β. Thus, we can generate a sample from G_1 as follows:

   > 1. Generate IID $U_1, U_2, \ldots, U_k \sim \text{unif}[0, 1]$.
   > 2. Let $A_i = -\beta \ln(U_i)$ for each $i = 1, \ldots, k$.
   > 3. Return $Z = A_1 + A_2 + \cdots + A_k$.

   (b) We will use inversion to generate a sample from CDF G_2. First, we need to compute the CDF G_2. For $0 \le x \le 4$, we get

   $$G_2(x) = \int_0^x \frac{3}{32} u \, du = \frac{3}{64} x^2.$$

   For $x > 4$, we get

   $$G_2(x) = \int_0^4 \frac{3}{32} u \, du + \int_4^x \frac{1}{4} \alpha e^{-\alpha u} \, du = \frac{3}{4} + \frac{1}{4}(1 - e^{-\alpha x}) = 1 - \frac{1}{4} e^{-\alpha x}.$$

   Thus, the CDF is

   $$G_2(x) = \begin{cases} 0 & \text{if } x < 0, \\ 3x^2/64 & \text{if } 0 \le x \le 4, \\ 1 - e^{-\alpha x}/4 & \text{if } x > 4, \end{cases}$$

which we now need to invert. Note that $G_2(4) = 3/4$. For $0 \le U \le 3/4$, we set $3x^2/64 = U$, and solve for x, giving $x = 8\sqrt{U/3}$. For $3/4 < U \le 1$, we set $1 - e^{-\alpha x}/4 = U$, and solve for x, giving $x = -\ln[4(1-U)]/\alpha$. Hence, our algorithm for generating a sample X from distribution G_2 is as follows:

1. Generate $U \sim \text{unif}[0,1]$.
2. If $0 \le U \le 3/4$, then return $X = 8\sqrt{U/3}$, and stop.
3. If $3/4 < U \le 1$, then return $X = -\ln[4(1-U)]/\alpha$ and stop.

(c) First we derive $\mathcal{P}(X \le x \mid Y \le b)$. Note that $\mathcal{P}(X \le \infty, Y \le y) = \mathcal{P}(Y \le y)$ since $\mathcal{P}(X \le \infty) = 1$. Thus,

$$\begin{aligned}\mathcal{P}(X \le x \mid Y \le b) &= \frac{\mathcal{P}(X \le x, Y \le b)}{\mathcal{P}(Y \le b)} = \frac{\sum_{i=1}^2 p_i F_i(x,b)}{\mathcal{P}(X \le \infty, Y \le b)} \\ &= \frac{\sum_{i=1}^2 p_i G_i(x) H_i(b)}{\sum_{j=1}^2 p_j G_j(\infty) H_j(b)} = \frac{\sum_{i=1}^2 p_i G_i(x) H_i(b)}{\sum_{j=1}^2 p_j H_j(b)}\end{aligned}$$

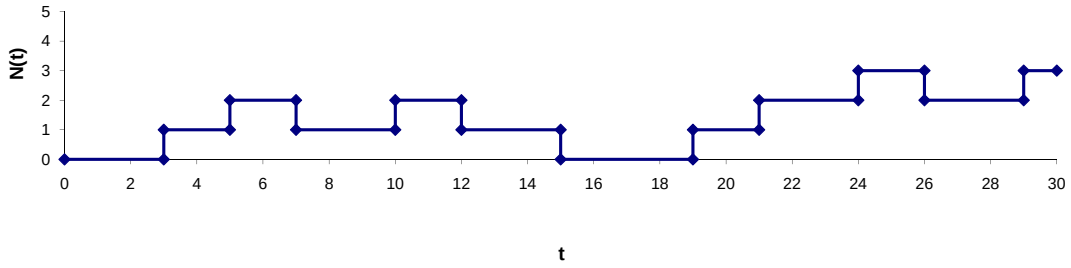
since $G_j(\infty) = 1$. Define $q_i = p_i H_i(b)/(\sum_{j=1}^2 p_j H_j(b))$, and note that $q_i \ge 0$ and $\sum_{i=1}^2 q_i = 1$. Hence,

$$\mathcal{P}(X \le x \mid Y \le b) = \sum_{i=1}^2 q_i G_i(x),$$

which is mixture of the distributions G_1 and G_2. We gave algorithms for generating samples from G_1 and G_2 in the previous parts. Also, $H_i(b) = 1 - \exp\{-\lambda_i b\}$, so $q_i = p_i(1 - \exp\{-\lambda_i b\})/\sum_{j=1}^k p_j(1 - \exp\{-\lambda_j b\}$. Thus, we get the following algorithm to generate a sample of X given $Y \le b$.

1. Generate $U_0 \sim \text{unif}[0,1]$.
2. If $U_0 < q_1$, then generate U_1 and U_2 as IID unif$[0,1]$ independent of U_0, and return $X = -\beta \ln(U_1) - \beta \ln(U_2)$, and stop.
3. If $U_0 > q_1$, then generate $U_3 \sim \text{unif}[0,1]$ independent of U_0.
4. If $0 \le U_3 \le 3/4$, then return $X = 8\sqrt{U/3}$, and stop.
5. If $3/4 < U_3 \le 1$, then return $X = -\ln[4(1-U_3)]/\alpha$ and stop.

5. (a) Below is a plot of $N(t)$ vs. t:



The times t at which $N(t)$ changes are 0, 3, 5, 7, 10, 12, 15, 19, 21, 24, 26, 29.

(b) Calculating the area under the $N(t)$ plot from $t = 0$ to $t = 30$ and dividing by 30, we get 39/30.