Midterm
CS 661: System Simulation — Spring 2008
Prof. Marvin K. Nakayama

Print family (or last) name: _____

Print given (or first) name: _____

I have read and understand all of the instructions below, and I will obey the Academic Honor Code.

_____

Signature and Date

Instructions

- This exam has 12 pages in total, numbered 1 to 12. Make sure your exam has all the pages.
- This is a closed book, closed-note exam.
- For all problems, follow these instructions:
  - Show your work and give reasons.
  - For any problem that asks you to provide an algorithm, be sure to give a step-by-step algorithm. Do not explain how to implement your algorithm in Arena or Excel, but rather, you should provide an algorithm (i.e., pseudo-code) that can be implemented in any language. Also, for any random variates needed by your algorithm, be sure to explicitly provide the steps needed to generate the random variates. Also, be sure to explicitly specify all constants or parameter values used in your algorithm. You may assume that you have at your disposal a good unif[0,1] random number generator.
  - In any problem for which you use an approximation based on the central limit theorem, you can use a quantile point from the normal distribution. Any confidence interval you provide should be at a 95% confidence level.
  - You can use the backs of the sheets if you need more space.
  - The last two pages of the exam have some possibly useful reference material.

Problem	1	2	3	4	5	6	Total
Points							

1. **[20 points]** Recall that a geometric random variable X with parameter p, $0 < p < 1$, has probability mass function

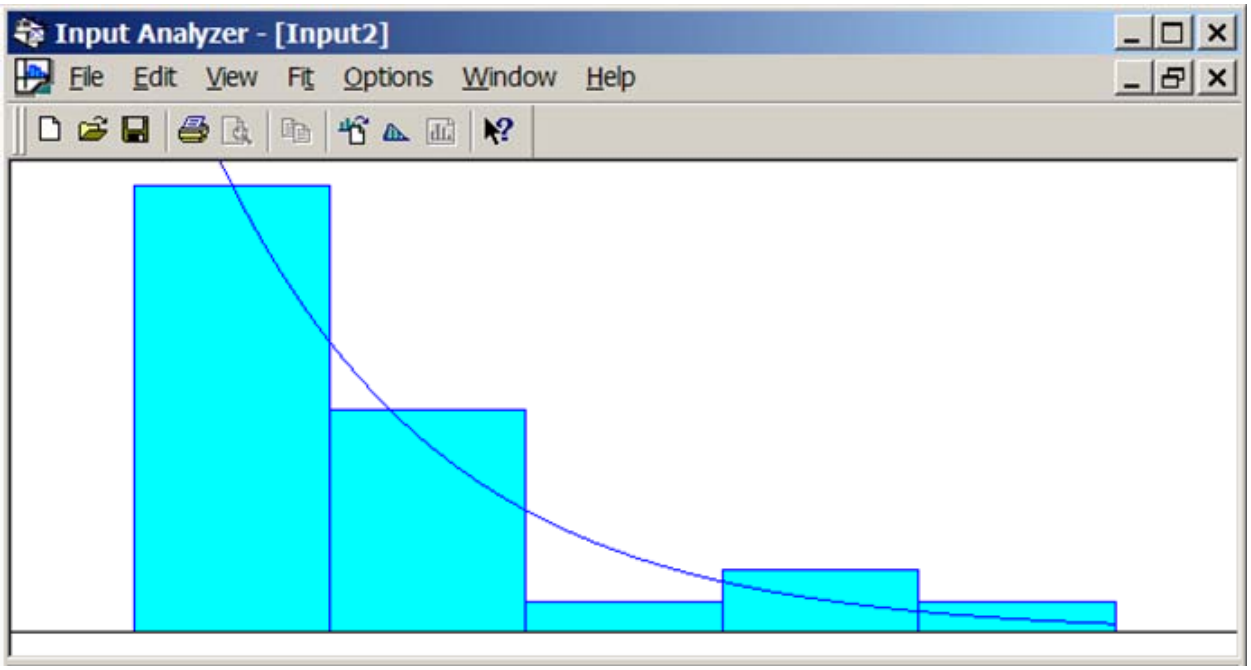
$$\mathcal{P}(X = x) = \begin{cases} (1-p)^{x-1}p, & \text{for } x = 1, 2, 3, \ldots \\ 0 & \text{otherwise} \end{cases}.$$

Suppose that $X_1, X_2, \ldots, X_n$ are independent observations that we have decided come from a geometric distribution.

(a) **[10 points]** Derive the maximum likelihood estimator (MLE) of the parameter p.

(b) **[10 points]** Derive the method-of-moments estimator of the parameter p. Be sure to derive all necessary moments.

2. **[15 points]** Suppose that we have collected data of times between hits to a particular website. We collected 25 of these times, and used Arena's Input Analyzer to fit an exponential distribution to the data. The results from the Input Analyzer are below.




```
Distribution Summary

Distribution: Exponential
Expression: EXPO(5.54)
Square Error: 0.007447

Chi Square Test
  Number of intervals = 2
  Degrees of freedom  = 0
  Test Statistic      = 0.218
  Corresponding p-value < 0.005

Kolmogorov-Smirnov Test
  Test Statistic = 0.0891
  Corresponding p-value > 0.15

Data Summary

Number of Data Points = 25
Min Data Value        = 0.0723
Max Data Value        = 23.7
Sample Mean           = 5.54
Sample Std Dev        = 6.01

Histogram Summary

Histogram Range     = 0 to 24
Number of Intervals = 5
```

Comment on the appropriateness of fitting an exponential distribution to the data. Be sure to discuss the output from the Input Analyzer.

3. **[15 points]** Suppose that we are building a simulation model, and one of the activities that we are modeling is the time to complete a job. We collected 50 job-completion times, and from these data we computed the following sample statistics:

Statistic	Value
Sample Size n	50
Sample Average $\bar{X}(n)$	5.76
Sample Variance $S^2(n)$	10.69
Sample Standard Deviation $S(n)$	3.27

From these data, we construct a 95% confidence interval as $\bar{X}(n) \pm 1.96S(n)/\sqrt{n}$, which is $(4.85, 6.66)$. Based on this, we decide to model the distribution of the job-completion time as a unif$(4.85, 6.66)$. Comment on the appropriateness of this approach.

4. **[15 points]** To complete a job, a worker must go through up to $k = 4$ stages in sequence. For each $i = 1, \ldots, 4$, the time to complete stage i is a Weibull random variable with shape parameter α_i and scale parameter β_i, where the times in the different stages are independent. After completing stage i, the worker will go to stage $i+1$ with probability p_i, $i = 1, \ldots, k-1$; i.e., after completing stage i, the worker will stop working with probability $(1 - p_i)$. The event that a worker stops at stage i is independent of all of the times to complete the previous stages. Let X denote the total time the worker spends on the job. Give an algorithm for generating samples of X. Be sure to provide all the details of your algorithm.

5. **[15 points]** Let X be a random variable with density function

$$f(x) = \begin{cases} 0 & \text{if } x < 2, \\ \frac{3}{2}x - 3 & \text{if } 2 \le x < 3, \\ e^{-4(x-3)} & \text{if } x \ge 3. \end{cases}$$

Give an algorithm to generate observations of X using the inverse-transform method.

6. **[20 points]** Consider a single-server queueing system. Suppose that initially, the system is empty and the server is idle. A customer who arrives and finds the server idle enters service immediately. A customer who arrives and finds the server busy joins the end of a single queue. Upon completing service for a customer, the server chooses a customer from the queue (if any) in a first-in, first-out (FIFO) manner. Assume that **the first customer arrives at time** 0, and the sequence of interarrival times (in minutes) is

$$4,\ 3,\ 2,\ 9,\ 5,\ 4,\ 3,\ 8,\ 3,\ 6, \ldots .$$

The sequence of service times (in minutes) is

$$2,\ 9,\ 4,\ 2,\ 1,\ 6,\ 5,\ 6,\ 3,\ 2, \ldots ;$$

i.e., the first customer, who arrives at time 0, has a service time of 2 minutes; the second customer has a service time of 9 minutes; and so on. For this problem, you are to carry out a simulation by hand up to time 40.

(a) **[15 points]** Define $A(t)$ to be the total number of customers in the system (in service and in the queue) at time t. Make a plot of $A(t)$ vs. t between times 0 and 40 for this realization. Be sure to label the horizontal axis to show the times of that $A(t)$ changes values.

(b) **[5 points]** From your plot in part (a), compute $\bar{A}(40)$ = the time-average total number of customers in the system during the time interval $[0, 40]$.

Some possibly useful facts

- The CDF F of random variable X is defined as $F(z) = P(X \leq z)$. If X is a continuous random variable with density function f, then $F(z) = \int_{-\infty}^{z} f(x)dx$. If X is a discrete random variable with probability mass function p and support $\{x_1, x_2, x_3, \ldots\}$, then $F(z) = \sum_{x_i \leq z} p(x_i)$.
- The expectation or mean μ of a random variable X is $\mu = \mathcal{E}[X] = \sum_{x_i} x_i p(x_i)$ when X is a discrete random variable with probability mass function $p(\cdot)$ and support $\{x_1, x_2, \ldots\}$, and $\mu = \int_{-\infty}^{\infty} x f(x)\, dx$ when X is a continuous random variable with density function f.
- The variance of a random variable X is $\mathcal{V}(X) = \mathcal{E}[(X-\mu)^2] = \mathcal{E}[X^2] - \mu^2$.
- The covariance of random variables X and Y with means μ_X and μ_Y is $\mathcal{C}(X,Y) = \mathcal{E}[(X-\mu_X)(Y-\mu_Y)] = \mathcal{E}[XY] - \mu_X\mu_Y$.
- For IID data $X_1, X_2, \ldots, X_n$, the sample mean is $\bar{X}(n) = \frac{1}{n}\sum_{k=1}^{n} X_i$, and the sample variance is $S^2(n) = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}(n))^2$.
- The normal distribution with mean μ and variance $\sigma^2 > 0$ is denoted by $\mathcal{N}(\mu, \sigma^2)$ and has density function $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$. Its CDF cannot be evaluated in closed form.
- If $Z_1, Z_2, \ldots, Z_n$, $n \geq 1$, are IID $\mathcal{N}(0,1)$ random variables, then $X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ has a chi-squared distribution with n degrees of freedom, which is denoted by χ_n^2.
- If $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi_n^2$ with Z and U independent, then $T = Z/\sqrt{U/n}$ has a Student-t distribution with n degrees of freedom, which is denoted by t_n.
- If $X_1, X_2, \ldots, X_n$ are IID $\mathcal{N}(\mu, \sigma^2)$, then $(n-1)S^2(n)/\sigma^2 \sim \chi_{n-1}^2$ and $(\bar{X}(n) - \mu)/\sqrt{S^2(n)/n} \sim t_{n-1}$.
- The uniform distribution over the interval $[a, b]$ has density function $f(x) = 1/(b-a)$ for $a \leq x \leq b$, and $f(x) = 0$ otherwise. Its CDF is $F(x) = (x-a)/(b-a)$ for $a \leq x \leq b$, and $F(x) = 0$ otherwise. Its mean is $(a+b)/2$ and its variance is $(b-a)^2/12$.
- The exponential distribution with parameter $\lambda > 0$ has density function $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, and $f(x) = 0$ for $x < 0$. Its CDF is $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$, and $F(x) = 0$ for $x < 0$. Its mean is $1/\lambda$ and its variance is $1/\lambda^2$.
- The gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$ has density function $f(x) = \beta^{-\alpha} x^{\alpha-1} \exp\{-x/\beta\}/\Gamma(\alpha)$ for $x \geq 0$, and $f(x) = 0$ for $x < 0$, where $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt$. In general, its CDF cannot be evaluated in closed form. Its mean is $\alpha\beta$ and its variance is $\alpha\beta^2$.
- The Weibull distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$ has density function $f(x) = \alpha\beta^{-\alpha} x^{\alpha-1} \exp\{-(x/\beta)^{\alpha}\}$ for $x \geq 0$, and $f(x) = 0$ for $x < 0$. Its CDF is $F(x) = 1 - \exp\{-(x/\beta)^{\alpha}\}$ for $x \geq 0$, and $F(x) = 0$ for $x < 0$. Its mean is $(\beta/\alpha)\Gamma(1/\alpha)$ and its variance is $(\beta^2/\alpha)[2\Gamma(2/\alpha) - (1/\alpha)(\Gamma(1/\alpha))^2]$.
- Law of large numbers: if $X_1, X_2, \ldots, X_n$ is a sequence of IID random variables with finite mean μ, then $\bar{X}(n) \approx \mu$ for large n.
- Central limit theorem: if $X_1, X_2, \ldots, X_n$ is a sequence of IID random variables with finite mean μ and positive, finite variance σ^2, then $\frac{\sqrt{n}}{\sigma}(\bar{X}(n) - \mu) \stackrel{\mathcal{D}}{\approx} \mathcal{N}(0,1)$ for large n, where $\stackrel{\mathcal{D}}{\approx}$ means approximately equal to in distribution. Also, $\frac{\sqrt{n}}{S(n)}(\bar{X}(n) - \mu) \stackrel{\mathcal{D}}{\approx} \mathcal{N}(0,1)$ for large n.
- Confidence intervals: if $X_1, X_2, \ldots, X_n$ is a sequence of IID random variables with finite mean μ and positive, finite variance σ^2, then an approximate $100(1-\alpha)\%$ confidence interval for μ is $[\bar{X}(n) \pm z_{1-\alpha/2} S(n)/\sqrt{n}]$, where $z_{1-\alpha/2}$ is the constant such that $\mathcal{P}(\mathcal{N}(0,1) > z_{1-\alpha/2}) = \alpha/2$. For example, $\mathcal{P}(\mathcal{N}(0,1) > 1.65) = 0.05$, $\mathcal{P}(\mathcal{N}(0,1) > 1.96) = 0.025$, and $\mathcal{P}(\mathcal{N}(0,1) > 2.60) = 0.005$.

- $\ln(ab) = \ln(a) + \ln(b)$ and $\ln(a^b) = b\ln(a)$.
- Parameter estimation: Suppose $X_1, \ldots, X_n$ are IID with distribution $F_{\theta_1,\ldots,\theta_d}$, where $\theta_1, \ldots, \theta_d$ are unknown and to be estimated.
  - MLE: If $F_{\theta_1,\ldots,\theta_d}$ is a continuous distribution with density function $f_{\theta_1,\ldots,\theta_d}$, then compute the likelihood $L(X_1, \ldots, X_n) = \prod_{i=1}^n f_{\theta_1,\ldots,\theta_d}(X_i)$. If $F_{\theta_1,\ldots,\theta_d}$ is a discrete distribution with probability mass function $p_{\theta_1,\ldots,\theta_d}$, then compute the likelihood $L(X_1, \ldots, X_n) = \prod_{i=1}^n p_{\theta_1,\ldots,\theta_d}(X_i)$. Then choose the values of $\theta_1, \ldots, \theta_d$ that maximize $L(X_1, \ldots, X_n)$, or equivalently that maximize $\ln L(X_1, \ldots, X_n)$.
  - Method-of-moments: Let $\mu_j = \mathcal{E}[X^j]$ be the jth (theoretical) moment of $F_{\theta_1,\ldots,\theta_d}$, and let $M_j = \frac{1}{n}\sum_{i=1}^n X_i^j$ be the jth sample moment of the data. Then set $\mu_j = M_j$ for $j = 1, 2, \ldots, d$, and solve for $\theta_1, \ldots, \theta_d$.
- Goodness-of-fit tests: Suppose $X_1, \ldots, X_n$ are IID from some hypothesized distribution F.
  - Chi-squared test: Divide samples into k bins. Let O_i be the number of observations in the ith bin, and let E_i be the expected number of observations in the ith bin. Compute test statistic $\chi^2 = \sum_{i=1}^k (O_i - E_i)^2/E_i$, which has a chi-squared distribution with ν degrees of freedom, where $\nu =$ number of bins $-$ (1 + number of estimated parameters).
  - Kolmogorov-Smirnov (K-S) test: Compute empirical CDF $\hat{F}_n(x) = (\#\text{ observations} \le x)/n$. Then compute $D_n = \max_{x\in\Re} |\hat{F}_n(x) - F(x)|$. We obtain the test statistic by centering and scaling D_n, where the centering and scaling depend on the hypothesized F.
- LCG: $Z_i = (aZ_{i-1} + c) (\text{mod } m)$ and $U_i = Z_i/m$, for constants a, c and m.
- Let $U_1, \ldots, U_n$ be n successive samples from random number generator. Null hypothesis is H_0 : samples are IID unif$[0, 1]$. Fix level α of test, and define z_p such that $\mathcal{P}(Z \le z_p) = p$.
  - Runs-up/down test: Let R_n is the total number of runs, $\mu_n = \mathcal{E}[R_n] = (2n-1)/3$ and $\sigma_n^2 = \mathcal{V}[R_n] = (16n - 29)/90$. Reject H_0 at level α if $R_n \notin [\mu_n - z_{\alpha/2}\sigma_n,\ \mu_n + z_{1-\alpha/2}\sigma_n]$.
  - Runs-above/below-mean test: Let R_n is the total number of runs, $\mu_n = \mathcal{E}[R_n] = (n + 1)/2$ and $\sigma_n^2 = \mathcal{V}[R_n] = (n - 1)/4$. Reject H_0 at level α if $R_n \notin [\mu_n - z_{\alpha/2}\sigma_n,\ \mu_n + z_{1-\alpha/2}\sigma_n]$.
- Inverse-transform method to generate sample X having CDF F. Generate $U \sim$ unif$[0, 1]$, and return $X = F^{-1}(U)$.
- Convolution method: Suppose $X = Y_1 + Y_2 + \cdots + Y_n$, where $Y_1, Y_2, \ldots, Y_n$ are independent, and $Y_i \sim F_i$. We generate a sample of X by generating $Y_1 \sim F_1$, $Y_2 \sim F_2$, $\ldots$, $Y_n \sim F_n$, where $Y_1, \ldots, Y_n$ are generated independently, and returning $X = Y_1 + \cdots + Y_n$.
- Composition method: Suppose we can write CDF F as $F(x) = p_1F_1(x) + p_2F_2(x) + \cdots + p_dF_d(x)$ for all x, where $p_i > 0$ for each i and $p_1 + \cdots + p_d = 1$. We then generate a sample $X \sim F$ as follows. First generate I having pmf $(p_1, p_2, \ldots, p_d)$. If $I = j$, then independently generate $Y_j \sim F_j$, and return $X = Y_j$.