406.311 Simulation
Fall 2007

**Midterm**
**Monday, October 29, 2007**
**(75 minutes; closed book)**

**Problem 1** (15 points)
True/False

____ A difference between input random variables $X$ and output random variables $Y$ is that the distribution of $Y$ is known.

____ The standard deviation is a *scale* parameter; that is, it reflects the scaling chosen for the random variable, such as inches or feet.

____ The normal family of distributions has two *shape* parameters.

____ A model is *valid* (or not) regardless of its intended purpose.

____ Program debugging is more closely related to verification than to validation.

____ Of two simulation models of the same system, the model with the greater detail is more valid.

____ The width of a confidence interval for a performance measure is proportional to the standard error ($\frac{s}{\sqrt{n}}$) of its estimator.

____ There is no value in performing a simulation study if you can't get good input data.

____ The method of moments estimates the parameters of a distribution by equating the $k$-th moment of distribution to the $k$-th moment of data.

____ Frequently, the number of alternatives is so large that a simulation run cannot be made for each of them. The process of determining which alternatives will and will not be evaluated is called validation.

____ The acceptance-rejection usually requires more than one random numbers to generate a random variate.

____ If the summary statistics show $\overline{x} < \tilde{x}$, we may think of the Gamma distribution as a good candidate.

____ The Lexis ratio of a binomial distribution is usually less than 1.

____ Maximum-likelihood estimation and the method of moments have the same purpose: to transform random numbers $U$ into random variates $X$.

____ The random variables $X$ and $(X - 6)/3$ have the same value for the kurtosis (the fourth standardized moment).

**Problem 2** (20 points)
Let the random variable $X$ have pdf

$$f(x) = \begin{cases} \frac{1}{2}x & \text{if } 0 < x < 1, \\ \frac{1}{2} & \text{if } 1 \le x \le \frac{5}{2}. \end{cases}$$

(a) Derive an algorithm to generate a random variate from $f(x)$, using the *composition* method, combined with the *inverse-transform* method. (*Hint*: You should explicitly show the distribution functions $F(x)$ and their inverses $F^{-1}(x)$ and describe all the required steps of the overall algorithm.)

(b) Derive an algorithm to generate a random variate from $f(x)$, using the *acceptance-rejection* method. (*Hint*: You should explicitly show the majorizing function $t(x)$ and its "normalized" density $r(x)$ and describe all the required steps of the overall algorithm.)

**Problem 3** (10 points)
Input modeling is viewed as having three steps:

1. Choosing a family of distributions,
2. Fitting a distribution to data, and
3. Validating (testing) the fitted distributions.

For each of the following concepts, state the step (1, 2, or 3) that is most closely related to the concept.

____ `Comparison` part obtained from the BestFit.

____ Histogram plot.

____ Anderson-Darling test.

____ Estimation of parameters using such methods as maximum likelihood.

____ The "physics" of the problem, such as non-negativity or continuity.

**Problem 4** (10 points)
Consider a weighing process of a package which is very likely to include some errors. We know from the historical data that the true weight $W$ of the contents of a package has a Gamma distribution with mean $12kg$ and standard deviation $1kg$. For a Monte Carlo simulation of this process, we model the measured weight as $M = WD$, where the relative error $D$ had a lognormal distribution with mean 1 and standard deviation 0.1. We want to estimate $\theta_1 = \Pr(M \geq 12)$ and $\theta_2 = \mathrm{E}(M)$ with $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively.

Of the above, which are the

(a) input variable(s)?

(b) output variable(s)?

(c) performance measure(s)?

(d) input-model distribution(s)?

(e) point estimator(s)?

**Problem 5** (10 points)
A set of discrete data was hypothesized to come from a *geometric distribution* with the following probability density function

$$g(x) \; = \; p\,(1-p)^x, \quad x \; = \; 0, 1, 2, \; \ldots$$

Derive the maximum likelihood estimator (MLE) $\hat{p}$ of $p$.

*Hint*: This task is more easily accomplished if, instead of working directly with $L(\beta)$, you work with its logarithm. In other words, define the *log-likelihood function* as $l(\beta) = \ln L(\beta)$. Since the logarithm function is strictly increasing. maximizing $L(\beta)$ is equivalent to maximizing $l(\beta)$.

**Problem 6** (10 points)
Suppose we have determined to use the *empirical distribution*, as opposed to fitting a theoretical distribution to data, for a certain input variable in a simulation modeling. Suppose also that we have grouped data into a frequency histogram with the following five adjacent intervals:

| Interval | Frequency |
|---|---|
| $[0, 2)$ | 10 |
| $[2, 4)$ | 30 |
| $[4, 6)$ | 50 |
| $[6, 8)$ | 25 |
| $[8, 10)$ | 6 |
| | 121 |

(a) (6 points) Plot the cumulative distribution $G(x)$. Show your calculations for $G(0), G(2), \cdots, G(10)$.

(b) (4 points) Determine the value of $G(5.3)$ using the linear interpolation. Show your calculation.

**Problem 7** (15 points)
You have collected the following data:

| | | | | | | |
|---|---|---|---|---|---|---|
| 4.40 | 28.62 | 28.32 | 62.54 | 14.13 | average | 17.78 |
| 7.14 | 6.04 | 11.32 | 82.19 | 42.14 | std dev | 18.42 |
| 5.08 | 22.25 | 16.80 | 33.43 | 17.07 | median | 11.74 |
| 3.49 | 0.87 | 42.75 | 15.46 | 3.51 | coeff of variation | 1.04 |
| 2.45 | 11.96 | 26.27 | 46.66 | 4.09 | max | 82.19 |
| 9.27 | 5.11 | 18.12 | 13.28 | 5.29 | min | 0.78 |
| 7.14 | 7.54 | 60.64 | 1.26 | 19.34 | | |
| 2.12 | 14.24 | 7.73 | 61.23 | 14.09 | | |
| 8.00 | 15.77 | 0.78 | 5.45 | 23.55 | | |
| 11.53 | 19.72 | 11.51 | 2.63 | 4.60 | | |

(a) (3 points) A reasonable hypothesis is that it comes from an exponential distribution. Why?

(b) (3 points) Which exponential distribution is it most likely to have come from?

(c) (9 points) Perform the $\chi^2$ goodness-of-fit test to see the data could reasonably come from an exponential distribution with $\beta = 15$. Use $\alpha = 0.05$. Note that $\chi^2_{9,\,0.95} = 17$.

| Interval | Observed |
|---|---|
| (0, 3] | 6 |
| (3, 6] | 9 |
| (6, 9] | 6 |
| (9, 12] | 5 |
| (12, 15] | 4 |
| (15, 20] | 7 |
| (20, 30] | 5 |
| (30, 45] | 3 |
| (45, 65] | 4 |
| (65, 90] | 1 |

**Problem 8** (10 points)
Short answers.

(a) (3 points) Consider the following EXCEL spreadsheet segment.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | 10 | 5 | 1 | 1 |
| 2 | 20 | 5 | 5 | 1 |
| 3 | 30 | 6 | 1 | 1 |
| 4 | 40 | 8 | 4 | 2 |
| 5 | 50 | 6 | 0 | 3 |

What will be the value returned by `=VLOOKUP(30,$A$1:$D$5,2)`?

What will be the value returned by `=VLOOKUP(35,$A$1:$D$5,2)`?

What will be the value returned by `=VLOOKUP(40,$A$1:$D$5,3,FALSE)`?

(b) Consider the following EXCEL function which is used to implement a lookup table.

```
=VLOOKUP(RAND(),$C$4:$E$7,2)
```

What is the meaning of the third argument (which is `2` in this case) of the `VLOOKUP` function.

(c) If $U_i$ is uniformly distributed between 0 and 1, then we can get that

$$Y = \frac{1}{12}\sum_{i=1}^{12} U_i$$

is approximately normally distributed under the assumption that the central limit theorem holds for $n = 12$. What are the mean and variance of $Y$?

**Bonus Problem** (up to 20 points)
Use the *method of moments* to estimate the two parameters, $\alpha$ and $\beta$, of the Gamma distribution which has the following pdf:

$$f(x) = \frac{x^{\alpha-1}\, e^{-x/\beta}}{\beta^{\alpha}\,\Gamma(\alpha)}, \quad x > 0, \quad \alpha, \beta > 0.$$

Assume that we know that the mean and variance of a Gamma random variable $X$ are:

$$E(X) = \alpha\beta \quad \text{and} \quad V(X) = \alpha\beta^2,$$

*Hint*:
Recall that $V(X) = E(X^2) - \{E(X)\}^2$.

*Note*: Your bonus points are calculated as $\left(\frac{100 - x}{100}\right) \times 20$, when you solve this problem right, where $x$ denotes your original score.