### Week Description Reading Material

<table>
<thead>
<tr>
<th>Week</th>
<th>Description</th>
<th>Reading Material</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Random Variables</td>
<td>Wallach et al., Appendix</td>
</tr>
<tr>
<td>Mar 1- Mar 5</td>
<td>• Why these are important in simulation</td>
<td>Wallach et al., Chapter 1</td>
</tr>
<tr>
<td></td>
<td>• Examples</td>
<td>Other References</td>
</tr>
<tr>
<td></td>
<td>• Distributions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Expectation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Working with Statistics in Simulation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>– Approximation of distributions from numerical outputs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>– Expected Values (Mean, Variance, Covariance – 2 methods)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>– Random Sampling, Monte Carlo Methods</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Bayesian Statistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Stochastic elements of dynamic models</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Two forms of models</td>
<td></td>
</tr>
</tbody>
</table>
Week 9 Objectives

1) To review some of the basic concepts of statistics that are useful in working with dynamic models

2) To expose students to random sampling and to analysis of simulation results using basic statistics

3) To introduce concepts of stochastic dynamic model components and why these are needed
Why Statistics?

• Deterministic, dynamic models thus far
• No model or measurement is perfect; they have uncertainties
• Need to understand uncertainty in model outputs if they are used for decision support policies, or control
• How can we effectively apply models if they are uncertain, and how do we know how uncertain they are?
Outline of Statistical Notions

• Random Variable
• Probability distributions of random variables
• Multiple random variables
• Expectations (mean, variance, covariance, correlation, …)
• Conditional distributions
• Bayesian statistics
• Empirical Examples
  – Random Sampling – Monte Carlo Methods
  – Approximating distributions using simulation results
  – Estimating Expectations (2 ways)
• Introduction of random variables in dynamic models
Random Variable

- Outcomes of an “experiment” involving some level of uncertainty. If the exact same experiment is repeated and the outcome is not the same every time, then the outcome is a random variable. It can be numerical outcomes, either integer or real numbers, or it may not be a number.
- Tossing a coin: outcomes = Heads or Tails
- Rolling a die: outcomes may be 1, 2, 3, 4, 5, or 6 – nothing else.
- Distance you hit a golf ball: any real number between 0 and 250 m
Random Variable

- In each example, the outcome is unknown in one sense, but it is known that there are limits to the outcomes that are possible (Range or Sample Space)
- The population relates to the number of experiments
- It is not true that Random Variable is equal to complete lack of knowledge about the outcome of an experiment
Probability, Density Function, Cumulative Distribution

• When is a finite number of possible outcomes, such as in a coin toss or die roll, the probability of a particular outcome is used (i.e., $P(\text{Heads}) = \text{probability that the coin toss is Heads}; P(2) = \text{probability that a die roll will result in a value of 2}$.

• These are examples of discrete random variables

• Density Function is $P(X=x)$, where we want to represent the probability that a random variable, $X$, is equal to a particular value, $x$. 
Probability, Density Function, Cumulative Distribution

• The Cumulative Density Function, \( F_X(x) \), is the probability that the Random Variable is less than or equal to a particular value, \( x \), or \( F_X(x) = P[X \leq x] \)

• For example, what is the probability that the Die-throw Random Variable is less than or equal to 4?

• The cumulative Distribution is represented by 0, 1/6, 2/6, 3/6, 4/6, 5/6, and 1 for die outcome values of 0, 1, 2, 3, 4, 5, and 6, respectively
Probability, Density Function, Cumulative Distribution

- For continuous variables, like the distance a golf ball is hit, there are infinite possible values of an experiment outcome, and it is not possible to refer to the probability of any particular value (it is 0).
- The continuous Density Function, \( f(x) \), is used, and \( f(x) \sim P(X \text{ falls between } x \text{ and } x+dx) \)
- The Cumulative Distribution for a continuous RV is:

\[
F_X(x) = \int_{-\infty}^{x} f_X(u)du
\]
Multiple Random Variables

- Consider two Random Variables, $X_1$ and $X_2$
- The joint density function is
  \[ f_{X_1,X_2}(x_1, x_2) \]
- The marginal density of $X_1$ is
  \[ f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2) \, dx_2 \]
- and Cumulative Distribution Function integrates over both
  \[ F_{X_1,X_2}(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{X_1,X_2}(u_1, u_2) \, du_1 du_2 \]
Expectation

• Expectation of a random variable

\[ E(X) = \sum_{All \ x} xP(x) \quad \text{or} \quad E(X) = \int_{-\infty}^{\infty} xf_X(x)dx \]

• Expectation of a function of a RV

\[ E[g(X)] = \sum_{All \ x} g(x)P(x) \quad \text{or} \quad E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx \]

• and \( E(x) \) can be thought of as the average or mean value, or the CENTROID of the distribution
Variance

\[ \text{var} = E[(X - E(X))^2] \]

\[ = \int_{-\infty}^{\infty} [x - E(X)]^2 f_X(x) dx \]

Standard Deviation, \( \sigma \)

\[ \sigma = \sqrt{\text{var}} \]

Covariance

\[ \text{cov} = E[(X_1 - E(X_1))(X_2 - E(X_2))] \]
Correlation Coefficient

\[ \rho(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sigma_{X_1} \sigma_{X_2}} \]

and \( -1 \leq \rho(X_1, X_2) \leq 1 \)
Uniform Density

Uniform Distribution
Min=8, Max=15

\[ f(x), \text{Density Function} \]

\[ 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3 \]

\[ 0, 5, 10, 15, 20, 25 \]
Normal Density Function

\[ f_X(x) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
Cumulative Normal Distribution

![Cumulative Normal Distribution Graph](image-url)
Empirical Distribution Approximation

- Random samples from a population are assumed to occur with equal probability $=1/n$, where $n=$ number of samples
- Rank samples from lowest to highest value
- Assign each a probability of $1/n$
- Compute cumulative probability, adding the successive $1/n$ values. The values should start at $1/n$ and end at 1.00
- Plot cumulative probability vs. ranked sample values
- This creates cumulative P
Example

<table>
<thead>
<tr>
<th>Random Sample</th>
<th>1/n</th>
<th>Cum P</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.43427E-06</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>5.94743E-05</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>0.000298321</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>0.000445209</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>0.000596605</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>0.000910535</td>
<td>0.01</td>
<td>0.06</td>
</tr>
<tr>
<td>0.000995323</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td>0.002146959</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>0.01162529</td>
<td>0.01</td>
<td>0.09</td>
</tr>
<tr>
<td>0.016066432</td>
<td>0.01</td>
<td>0.1</td>
</tr>
<tr>
<td>0.023208857</td>
<td>0.01</td>
<td>0.11</td>
</tr>
<tr>
<td>0.027589202</td>
<td>0.01</td>
<td>0.12</td>
</tr>
<tr>
<td>0.028065741</td>
<td>0.01</td>
<td>0.13</td>
</tr>
<tr>
<td>skip</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.592556</td>
<td>0.01</td>
<td>0.96</td>
</tr>
<tr>
<td>5.70823694</td>
<td>0.01</td>
<td>0.97</td>
</tr>
<tr>
<td>5.818862915</td>
<td>0.01</td>
<td>0.98</td>
</tr>
<tr>
<td>5.983276367</td>
<td>0.01</td>
<td>0.99</td>
</tr>
<tr>
<td>7.79838562</td>
<td>0.01</td>
<td>1</td>
</tr>
</tbody>
</table>

Empirical Cumulative Distribution

![Empirical Cumulative Distribution Graph](image-url)
Questions

• How do you create a density function from the previous example?
• How is the mean computed?
• How is the variance computed?
• Some Likelihood methods give probability values that are NOT 1/n, how would you compute mean and variance for these data?
Estimation of properties of random variables

• Mean can be estimated by assuming that each sample collected has equal likelihood of being selected

• From the formula to estimate $E(X)$, if $P(X)$ is the same for every one of $n$ samples, then $P(X=x) = 1/n$, and the mean can be estimated by:

$$E(X) = \sum_{All \ x} xP(x) \approx \sum_{i=1}^{n} x_i \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$
Variance is similar

\[ \text{var} = E[(X - E(X))^2] \approx \sum_{i=1}^{n} (x_i - \bar{x})^2 P(x_i) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

But, for an unbiased estimate, divide by (n-1)
Random Sampling

• Take one or more observations of a random variable
• Each observation:
  – is assumed to be independent
  – creates an observation from a population that has a defined probability distribution
• Collecting $n$ samples allows one to estimate properties of the population, such as mean and variance
• Sampling is useful in stochastic simulation; outputs characterize uncertainty in models
Pseudo Random Number

- Pseudo random number mimics an experiment where one samples the interval \([0,1]\) at random, with the number sampled equally likely to be any number in this range.
- Pseudo means that the numbers are not really random because the computer has an algorithm to compute it.
- The algorithm generates these numbers in a way that one could not detect any pattern at all.
Monte Carlo Sampling

• Inverse transformation method
  – Cumulative probability distribution
  – Pseudo random number [0,1] assumed to be cumulative probability
  – Solve for the value of the variable for this particular random number

• This is the “random” sample

• This is repeated n times to produce n random samples from the population with the probability distribution given
Inverse Sampling Method

\[ F_X(x) = P[X \leq x] = \int_{-\infty}^{x} f_X(x) \, dx \]

With \( f_X(x) \) the probability density and \( F_X(x) \) the cumulative probability of the random variable \( X \)
Example
Uniform Probability Distribution

\[ f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x \leq b \\ 0 & \text{else} \end{cases} \]

and

\[ F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{(x-a)}{(b-a)} & \text{if } a < x \leq b \\ 1.0 & \text{if } x > b \end{cases} \]

Set r.n. = F(x) and solve for x
If r.n. = 0.71, then the sample of x is 5.84
Example

Uniform Probability Distribution

\[ f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x \leq b \\ 0 & \text{else} \end{cases} \]

\[ F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a < x \leq b \\ 1 & \text{if } x > b \end{cases} \]

Set r.n. = F(x) and solve for x
If r.n. = 0.71, then the sample of x is 5.84
Example
Normal Distribution

Mean = 5.0
Std Dev = 1.0
r.n. = 0.71
Random x sample = 5.55
Bayesian Statistics

Joint probability

\[ P(x \cap y) = P(x, y) \]

Wilks, D.S. 2006. Statistical Methods in the Atmospheric Sciences
Bayesian Statistics

Conditional probability

\[ P(x \mid y) = \frac{P(x \cap y)}{P(y)} \]

\[ P(y \mid x) = \frac{P(x \cap y)}{P(x)} \]
Bayesian Statistics

Law of total probability

\[ P(\theta) = \sum_{i=1}^{4} P(\theta \cap y_i) \]

\[ P(\theta \mid y) = \frac{P(\theta \cap y)}{P(y)} \]

\[ P(\theta) = \sum_{i=1}^{4} P(\theta \mid y_i)P(y_i) \]
Bayesian Statistics

Bayes’ theorem

\[ P(\theta \cap y_i) = P(\theta, y_i) = P(\theta | y_i)P(y_i) = P(y_i | \theta)P(\theta) \]

\[ P(\theta | y_i) = \frac{P(y_i | \theta)P(\theta)}{P(y_i)} \]

\[ P(\theta | y_i) = \frac{\sum_{j=1}^{4} P(y_j | \theta)P(\theta)}{P(y_i)} \]
Bayesian Statistics

- Useful in Parameter Estimation
- Parameter is assumed to be a random variable
  - Imperfect knowledge of system
  - Variability in system; not all individuals in a system have the same characteristics
  - Inability to precisely measure a parameter
  - Model simplification
Bayesian Statistics

• Knowledge of parameter is represented by a random variable distribution

• Before measurements, knowledge is represented by a prior distribution, \( \pi(\theta) \)

• Measurements provide additional information about the random variable

• Posterior distribution is

\[
f_{\theta|Y}(\theta|Y = y) = \frac{f_{Y|\theta}(y|\theta)\pi(\theta)}{m_Y(y)}
\]
Bayesian Statistics

• Bayes’ Equation

\[ P(\theta|y) = \frac{P(y|\theta)P(\theta)}{\int_{\theta} P(y|\theta)P(\theta)} \]

• Very useful in parameter estimation
• Denominator is the marginal distribution of the random variable \( \theta \)
• \( P(\theta|y) \) is the probability of a particular random variable value given that a measurement was made with value \( y \)
• The posterior distribution is dependent on the measurements used to estimate it
Discussion
Homework

1. A sample of data were obtained, next page. Using Excel,
   - Compute the mean, variance, and standard deviation for these data
   - Draw the cumulative distribution of this random variable
   - Compare this distribution with a normal distribution that has the estimated mean and standard deviation, graphically
   - Do you think that the sample was taken from a population that was normally distributed?

2. Generate 100 samples from:
   - A normal distribution with mean of 4,000 and standard deviation of 1,000
   - An Exponential distribution with parameter 12

3. Use the data for X1 and X2 on the next page, Problem 3,
   - to estimate the mean, variance, covariance for both X1 and X2
   - Draw probability density and cumulative distribution function for each of X1 and X2
## 50 Random Sample Data Values

### Problem 1

<table>
<thead>
<tr>
<th>Value 1</th>
<th>Value 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.866</td>
<td>14.865</td>
</tr>
<tr>
<td>6.551</td>
<td>14.375</td>
</tr>
<tr>
<td>8.326</td>
<td>10.871</td>
</tr>
<tr>
<td>13.263</td>
<td>13.934</td>
</tr>
<tr>
<td>15.351</td>
<td>12.614</td>
</tr>
<tr>
<td>9.245</td>
<td>11.326</td>
</tr>
<tr>
<td>12.026</td>
<td>13.958</td>
</tr>
<tr>
<td>13.287</td>
<td>8.772</td>
</tr>
<tr>
<td>14.694</td>
<td>12.519</td>
</tr>
<tr>
<td>7.996</td>
<td>9.763</td>
</tr>
<tr>
<td>9.621</td>
<td>15.968</td>
</tr>
<tr>
<td>12.574</td>
<td>9.099</td>
</tr>
<tr>
<td>8.780</td>
<td>13.880</td>
</tr>
<tr>
<td>13.399</td>
<td>14.118</td>
</tr>
<tr>
<td>12.614</td>
<td>10.487</td>
</tr>
<tr>
<td>9.700</td>
<td>11.939</td>
</tr>
<tr>
<td>10.993</td>
<td>16.379</td>
</tr>
<tr>
<td>12.254</td>
<td>13.025</td>
</tr>
<tr>
<td>20.107</td>
<td>15.635</td>
</tr>
<tr>
<td>12.397</td>
<td>24.321</td>
</tr>
<tr>
<td>11.865</td>
<td>12.808</td>
</tr>
<tr>
<td>12.041</td>
<td>11.436</td>
</tr>
<tr>
<td>9.414</td>
<td>15.355</td>
</tr>
<tr>
<td>14.631</td>
<td>11.593</td>
</tr>
<tr>
<td>11.147</td>
<td>8.319</td>
</tr>
</tbody>
</table>

### Problem 3

<table>
<thead>
<tr>
<th>X1</th>
<th>X2</th>
<th>P(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2</td>
<td>5.110</td>
<td>0.0129</td>
</tr>
<tr>
<td>9.7</td>
<td>1.437</td>
<td>0.0835</td>
</tr>
<tr>
<td>2.9</td>
<td>2.446</td>
<td>0.0051</td>
</tr>
<tr>
<td>11.5</td>
<td>5.265</td>
<td>0.0740</td>
</tr>
<tr>
<td>5</td>
<td>5.859</td>
<td>0.0209</td>
</tr>
<tr>
<td>18.6</td>
<td>10.803</td>
<td>0.0014</td>
</tr>
<tr>
<td>17.7</td>
<td>0.846</td>
<td>0.0031</td>
</tr>
<tr>
<td>12.7</td>
<td>8.980</td>
<td>0.0559</td>
</tr>
<tr>
<td>8.9</td>
<td>9.599</td>
<td>0.0784</td>
</tr>
<tr>
<td>10.4</td>
<td>3.458</td>
<td>0.0831</td>
</tr>
<tr>
<td>8.3</td>
<td>5.945</td>
<td>0.0714</td>
</tr>
<tr>
<td>16.2</td>
<td>6.266</td>
<td>0.0099</td>
</tr>
<tr>
<td>10.9</td>
<td>1.045</td>
<td>0.0802</td>
</tr>
<tr>
<td>15</td>
<td>10.925</td>
<td>0.0209</td>
</tr>
<tr>
<td>13.9</td>
<td>8.773</td>
<td>0.0360</td>
</tr>
<tr>
<td>7.9</td>
<td>10.201</td>
<td>0.0656</td>
</tr>
<tr>
<td>11.7</td>
<td>10.037</td>
<td>0.0714</td>
</tr>
<tr>
<td>16.6</td>
<td>4.636</td>
<td>0.0075</td>
</tr>
<tr>
<td>13.8</td>
<td>8.066</td>
<td>0.0376</td>
</tr>
<tr>
<td>7.2</td>
<td>0.644</td>
<td>0.0543</td>
</tr>
<tr>
<td>9.2</td>
<td>0.276</td>
<td>0.0809</td>
</tr>
<tr>
<td>13.3</td>
<td>10.948</td>
<td>0.0458</td>
</tr>
</tbody>
</table>