

CPE 345: Modeling and Simulation

Lecture 7

Today's topic

- Random Number Generation (chapter 7)
 - Properties of Random Numbers
 - Generating Pseudo-random Numbers
 - Techniques for Generating Random Numbers
 - Testing for Randomness

Properties of Random numbers

- What properties should random numbers satisfy?
- Uniformly distributed random numbers in $(0,1)$
 - A sequence of random numbers R_1, R_2, \dots must have two important statistical properties:

- **Uniformity**

- Each random number R_i is a sample drawn from an uniform distribution

$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{ow} \end{cases}$$

$$E(X) = \frac{1}{2}$$

$$\text{var}(X) = \frac{1}{12}$$

- **Independence**

- All samples $(R_i, i = 1, 2, \dots)$ are independent

$$P(a \leq R_i \leq b) = \frac{1}{b-a} \quad \text{for any sample } R_i, a, b \in (0,1), a \leq b$$

Important consequences

- If the interval $(0, 1)$ is divided into n classes, or subintervals of equal length, the number of observations in each interval is N/n ; $N =$ total number of observations
 - If n intervals of length l_1, l_2, \dots, l_n ; the number of observations in the interval i is N/l_i ($\sum_i l_i = 1$)
 - We have already used this property to generate arrivals and departures
- Observing a value in a particular interval is independent of the previous values drawn

Generating pseudo-random numbers

- Why pseudo-random?
 - Generating truly random numbers in a simulation is not possible
 - Natural processes generate truly random numbers
 - If random numbers are generated using a known method, then they can be replicated, so they are not truly random anymore
 - In practice: generate pseudo-random numbers (with uniform distribution) that satisfy the ideal properties of the random distribution.
 - Some errors/departure from randomness in generating pseudo-random numbers
 - The generated numbers may not be uniformly distributed
 - The generated numbers may be discrete valued instead of continuous valued
 - The mean may be too high or too low
 - The variance may be too high or too low
 - There may be dependence between the generated numbers
 - Autocorrelation
 - Numbers successively higher or lower
 - Several numbers above the mean, then several numbers below the mean

Pseudo-random number generator routines

- The routine should pass all tests for departure from uniformity and independence – will follow shortly
- Some other requirements:
 - Speed – usually very large number of random numbers should be generated in a typical simulation (hundreds of thousands)
 - Portability – between machines and languages – same results wherever it is executed
 - Long-cycle – the numbers are repeated after a certain cycle length.
 - Repeatability – You may want to run simulations in exactly the same conditions, or you may want to specifically choose different starting points
 - Seed selection

Techniques for generating random numbers

- **Linear Congruential Method**

- Produces a sequence of integers X_1, X_2, \dots, X_{m-1} , according to the following recursive relationship

$$X_{i+1} = (aX_i + c) \bmod m, \quad i = 0, 1, 2, \dots$$

$X_0 = \text{seed}$
 $a = \text{constant multiplier}$
 $c = \text{increment}$
 $m = \text{modulus}$

- If $c=0$: multiplicative congruential method, otherwise, mixed congruential method
- The choice of the parameters drastically affects the statistical properties and the cycle length
- Variants of this technique used in the computer generation of random numbers (rand() functions)

Examples: ex. 7.1

Example 7.1: $X_0 = 27$, $a = 17$, $c = 43$, $m=100$

X	Normalized X
27	0.27
2	0.2
77	0.77
52	0.52
27	0.27
2	0.02
77	0.77
52	0.52
27	0.27
2	0.02
77	0.77
52	0.52
27	0.27

Cycle length 4. Is this acceptable?
What happens for different seed selection?

Example 7.1. Different seed selection

$X_0 = 3$
 $a = 17$
 $c = 43$
 $m = 100$

Cycle length 20

- Too small in both cases
- Depends on the choice of the seed

X	Normalized X
3	0.03
94	0.94
41	0.41
40	0.4
23	0.23
34	0.34
21	0.21
0	0
43	0.43
74	0.74
1	0.01
60	0.6
63	0.63
14	0.14
81	0.81
20	0.2
83	0.83
54	0.54
61	0.61
80	0.8
3	0.03
94	0.94

Choosing the parameters

- Some observations related to example 7.1
- Two more important properties
 - Maximum density
 - The numbers generated in example 7.1 are actually discrete, not continuous. The set of values

$$I = \{0, 1/m, 2/m, 3/m, \dots, (m-1)/m\}$$

- The approximation for continuous is acceptable for m large
 - In practice, $m = 2^{31} - 1$ and $m = 2^{48}$ are commonly used
- Maximum period
 - Maximal period can be achieved by proper choice of the parameters
 - For $m = 2^b$, and $c \neq 0$, the longest possible period is $P = m = 2^b$, achieved if c is relatively prime to m (the greatest common factor of c and m is 1), and $a = 1 + 4k$, k integer.

Maximum period – cont.

- For $m = 2^b$, and $c = 0$, the longest possible period is $P = m/4 = 2^{b-2}$, achieved if the seed X_0 is odd, and $a = 3+8k$ or $a = 5+8k$, $k=0,1, \dots$
- For m a prime number and $c = 0$, the longest possible period is $P = m-1$, achieved if a has the property that the smallest integer k s.t. a^k-1 is divisible by m is $k = m-1$.
- Study example 7.4 in the book. The chosen parameters are actually used in generators:

$$a = 7^5 = 16807$$

$$m = 2^{31} - 1 = 2,147,483,647 \text{ (prime number)}$$

$$P = m - 1$$

Are these numbers sufficiently large?

Combining Linear Congruential Generators

- The period length for example 7.4. seems extremely large
- With PCs capable of hundreds of MIPS, simulations are getting bigger and 10^9 cycle period is too short.
- Solution: choose multiple generators and combine them
 - If $X_{i,1}, X_{i,2}, \dots, X_{i,k}$ are the i -th output for different generators, where the j -th generator has prime modulus m_j , and a_j is chosen such that $P_j = m_j - 1$

$$X_i = \left(\sum_{j=1}^k (-1)^{j-1} X_{i,j} \right) \bmod (m_1 - 1)$$

$$R_i = \begin{cases} \frac{X_i}{m_1}, & X_i > 0 \\ \frac{m_1 - 1}{m_1}, & X_i = 0 \end{cases}$$

$$P = \frac{(m_1 - 1)(m_2 - 1) \dots (m_k - 1)}{2^{k-1}}$$

Tests for Random Numbers

- **Uniformity tests**
 - **Frequency test** – Kolmogorov-Smirnov or χ^2 (chi-square) tests
 - Compare the distribution to the uniform distribution
- **Independence tests**
 - **Runs test** – looks for patterns of increasing/decreasing values
 - **Autocorrelation test** – tests for correlation between numbers
 - **Gap test** – counts the number of digits that appear between the repetition of a particular digit, than uses the Kolmogorov-Smirnov test to compare with the expected size of gaps
 - **Poker test** – Treats numbers grouped together as a poker hand. The hands obtained are compared to what is expected using a chi-square test.

Hypothesis Testing

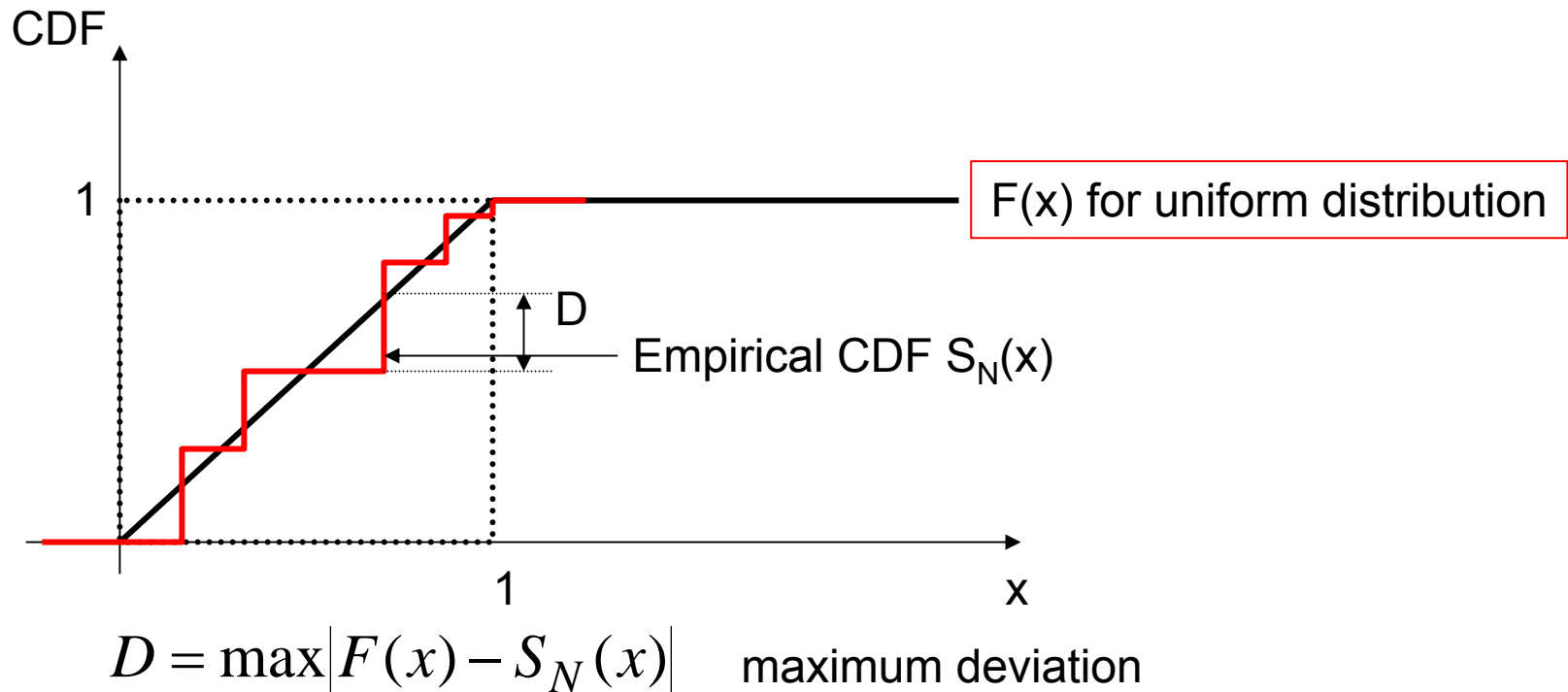
- Uniformity test
 - H_0 = numbers are uniformly distributed
 - H_1 = numbers are not uniformly distributed
 - Failure to reject the null hypothesis means that no evidence of non-uniformity has been found on the basis of this test. It does not mean that further testing is unnecessary.
- Independence test
 - H_0 = numbers are independently generated
 - H_1 = numbers are not independently generated
 - Failure to reject the null hypothesis means that no evidence of non-independence has been found on the basis of this test. It does not mean that further testing is unnecessary.
- For each test, a level of significance must be stated

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

- frequently, α is set to 0.01 or 0.05 (1%, 5% failure by chance)

Frequency Tests

- Kolmogorov-Smirnov test:
 - Compare empirical CDF (for N samples) with the expected CDF



For a given α and N random samples, find the critical value from Table A.8

Frequency Test – cont.

- Critical values for large number of samples:
 - For $N > 35$:

$D_{0.10}$	$D_{0.05}$	$D_{0.01}$
$1.22 / \sqrt{N}$	$1.36 / \sqrt{N}$	$1.63 / \sqrt{N}$

- How to compute empirical CDF

$$S_N(x) = \frac{\text{number of } R_1, R_2, \dots, R_N \text{ which are } \leq x}{N}$$

Kolmogorov-Smirnov test: step by step

- Rank the data from smallest to largest. $R(i)$ = the smallest i -th observation

$$R(1) \leq R(2) \leq \dots \leq R(N)$$

- Compute $D^+ = \max_{1 \leq i \leq N} \left\{ \frac{i}{N} - R(i) \right\}$

$$D^- = \max_{1 \leq i \leq N} \left\{ R(i) - \frac{i-1}{N} \right\}$$

- Compute $D = \max(D^+, D^-)$
- Determine a critical value D_α from Table A.8 for the specified significance level (α) and the given sample size N
- If $D \geq D_\alpha$ the null hypothesis is rejected;
- If $D \leq D_\alpha$ no difference is detected between the true distribution and the uniform distribution

Example (7.6)

- Example (7.6) Kolmogorov – Smirnov Test
 - Five numbers are generated: 0.44; 0.81; 0.14; 0.05; 0.93
 - Significance level $\alpha = 0.05$

R(i)	0.05	0.14	0.44	0.81	0.93
i/N	0.20	0.40	0.60	0.80	1.00
i/N-R(i)	0.15	0.26	0.16	-	0.07
R(i)-(i-1)/N	0.05	-	0.04	0.21	0.13

$D_{0.05} = 0.565 \Rightarrow$ The distribution cannot be distinguished from uniform

Chi-square test

- Uses the statistic

$$\chi_0^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad \text{Chi-square distributed with } n-1 \text{ degrees of freedom}$$

- O_i = the observed number in the i -th class
- E_i = the expected number in the i -th class
- n = the number of classes
- Compare with critical value $\chi_{\alpha, n-1}^2$ ([critical value table](#))
 - If $\chi_0^2 > \chi_{\alpha, n-1}^2$, reject null hypothesis
 - If $\chi_0^2 \leq \chi_{\alpha, n-1}^2$, indistinguishable from the uniform distribution
- For equally spaced classes and N samples $E_i = \frac{N}{n}$
- The test is valid for large number of samples $N \geq 50$
 - Recommended that N , and n are chosen such that each $E_i \geq 5$

Study Example 7.7 in your text book

Independence Tests

- Sometimes the generators pass the Kolmogorov-Smirnov and chi-square tests for uniformity, but the numbers generated are not independent → **need independence tests**

- **Run Tests**

- **Run = succession of similar events**

- Example: coin flipping: H T T H H T T T H T

- Six runs: length 1, 2, 2, 3, 1, 1

- For sequence of random numbers, we can define up runs and down runs (successive numbers are increasing or decreasing)

$-0.87, +0.15, +0.23, +0.45, -0.69, -0.32, -0.30, +0.19, 0.24.$

$- + + + - - - +$

$-0.87, +0.15, -0.23, +0.05, -0.69, +0.32, -0.40, +0.19, 0.24.$

$- + - + - + - +$

**Are these
reasonable
generators?**

Runs test

- The number of runs

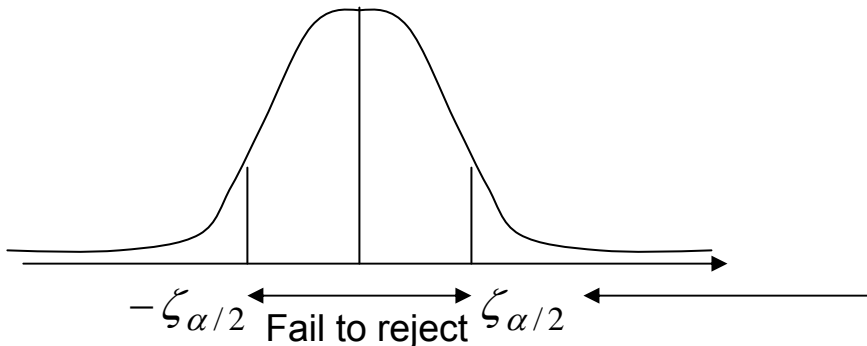
- a = total number of runs in a truly random sequence
- For $N > 20$, distribution of a assumed Gaussian (normal) with

$$\mu_a = \frac{2N - 1}{3}$$

$$\sigma_a^2 = \frac{16N - 29}{90}$$

- Standardized normal test statistic: test for independence

$$Z_0 = \frac{a - \mu_a}{\sigma_a} \rightarrow \text{Gaussian with zero mean and unit variance}$$
$$Z_0 \sim N(0, 1)$$



Find from Gaussian distribution tables
For significance $\alpha/2$

Determine critical values for significance $\alpha/2$

- Significance level $\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$

$$\alpha = P(Z_0 \geq \zeta_{\alpha/2}) + P(Z_0 \leq -\zeta_{\alpha/2}) = 2P(Z_0 \geq \zeta_{\alpha/2})$$

$$\frac{\alpha}{2} = P(Z_0 \geq \zeta_{\alpha/2}) = 1 - \phi(\zeta_{\alpha/2}) \Rightarrow \text{determine } \zeta_{\alpha/2}$$

tabulated



Runs above and below the mean

- For example, if mean is 0.5

0.1 0.2 0.22 0.14 0.33 0.18 0.63 0.58 0.53 0.76 0.9 0.82
0.27 ...

- In terms of runs up and down:

+ + - + - + - - + + - (*)

- If we define the runs as above and below the mean:

- - - - - + + + + + - (**)

- Same test can be run with the runs defined as in (**)

- n1 = number above of mean
- n2 = number below the mean
- b = number of runs – Gaussian r.v.

$$\mu_b = \frac{2n_1n_2}{N} + \frac{1}{2}$$

$$\sigma_b^2 = \frac{2n_1n_2(2n_1n_2 - N)}{N^2(N-1)}$$

Auto-correlation test

- Computation of the auto-correlation between every m (the lag) numbers starting with i
 - autocorrelation ρ_{im} \rightarrow between $R_i, R_{i+m}, R_{i+2m}, R_{i+(M+1)m}$, M is the largest value s.t. $i+(M+1)m \leq N$
- Nonzero autocorrelation \rightarrow lack of independence
 - $H_0: \rho_{im} = 0$
 - $H_1: \rho_{im} \neq 0$
- For large values of M , the distribution of the estimator $\hat{\rho}_{i,m}$ is approx. normal (Gaussian).

The statistic $z_0 = \frac{\hat{\rho}_{i,m}}{\sigma_{\hat{\rho}_{i,m}}}$ is normally distributed with mean 0 and unit variance, for large M

Auto-correlation test: step-by-step

- Compute

$$\hat{\rho}_{i,m} = \frac{1}{M+1} \left[\sum_{k=0}^M R_{i+km} R_{i+(k+1)m} \right] - 0.25$$

$$\sigma_{\hat{\rho}_{i,m}} = \frac{\sqrt{13M+7}}{12(M+1)}$$

[Schmidt & Taylor – 1970]

- Compute statistic

$$z_0 = \frac{\hat{\rho}_{i,m}}{\sigma_{\hat{\rho}_{i,m}}}$$

- Do not reject the null hypothesis (independence) if

$$-\zeta_{\alpha/2} \leq z_0 \leq \zeta_{\alpha/2}$$

α = level of significance

Gap test

- Gap = interval between the recurrences of the same digit
- Frequency of the gaps
 - The observed frequencies of various gap sizes compared to the theoretical frequency – Kolmogorov-Smirnov test
 - Theoretical frequency distribution

$$P(\text{gap} \leq x) = F(x) = 0.1 \sum_{n=0}^x (0.9)^n$$

Note: Probability of occurrence for certain digit is 0.1.

Poker test

- Frequency with which certain digits are repeated in a series of numbers
- Example
0.255, 0.577, 0.414, 0.828, 0.909, 0.303, 0.001
- Pair of like digits generated
- For three digits: three possibilities
 - All different
 - All equal
 - One pair of like digits

$$P(\text{exactly one pair}) = \binom{3}{2} (0.1)(0.9) = 0.27$$

no. of possibilities

Given a fixed digit, this digit is the same

Given a fixed digit, this digit different

Poker test: cont.

$$\begin{aligned} P(\text{three different digits}) &= P(\text{second different from first})P(\text{third different from first and second}) = \\ &= (0.9)(0.8) = 0.72 \end{aligned}$$

$$\begin{aligned} P(\text{three like digits}) &= P(\text{second digit same as first})P(\text{third digit same as first and second}) = \\ &= (0.1)(0.1) = 0.01 \end{aligned}$$

Poker test:

Measure observed frequency for the three cases

Compute expected frequency E_i (probabilities*1000)

Perform chi-square test

Example 7.14.

| Combination 1 | Obs. Freq. | Expected Freq. | $(O_i - E_i)^2 / E_i$ |
|--------------------|------------|----------------|-----------------------|
| 3 different digits | 680 | 720 | 2.22 |
| 3 like digits | 31 | 10 | 44.10 |
| Exactly 1 pair | 289 | 270 | 1.33 |

$$47.65 > \chi_{0.05,2}^2 = 5.99$$

Homework

- Problem 7 page 286, chapter 7