OntarioTech Engineering & Applied Science

SOFE 4820U: Modelling and Simulation Winter 2024

Dr. Anwar Abdalbari

Week 3 : Randomness and Generation Continued

Chapter outlines

- Random number simulation
- Issues with Linear Congruential Generator
- Testing for uniformity and independence

Linear Congruential Generator (LCG)

- One of the most common methods for generating random numbers is the Linear Congruence Generator (LCG).
- LCM produces a sequence of integers X1, X2, …. between 0 and m-1 by following the relationship:

$$x_{k+1} = (a * x_k + c) \, mod \, m$$

- where
  - **x0** is the initial value (seed or non-negative integers)
  - **a** is the multiplier (non-negative integers)
  - **c** is the increment (non-negative integers)
  - **m** is the modulo (non-negative integers)
- LCG has the following characteristics:
  - It is cyclical with a period that is approximately equal to m
  - The generated numbers are discretized

Linear Congruential Generator (LCG)

- Notice that, due to modulo, Xi will be re-generated after a finite number of recursions.

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

- Cycle Length (Period): for a sequence X1, X2, …., cycle length is the number of generation until X1 is generated again.

- Ex: for the following sequence:
  5, 13, 21, 10, 2, 14, 5, 13, 21, 10, 2, 14, 5, 13, 21, …
  the cycle is 6.

Generating Pseudo-Random Numbers with LCM

- We can generate numbers from random integers X1, X2,….. of the LCM method by:

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

$$R_i = \frac{X_i}{m}, i = 1,2, \dots.$$

Example: $x_0 = 27, a = 17, c = 43, and\ m = 100.$

$$X_1 = (17 \times 27 + 43) \bmod 100 = 502 \bmod 100 = 2; \quad R_1 = 2/100 = 0.02$$
$$X_2 = (17 \times 2 + 43) \bmod 100 = 77 \bmod 100 = 77; \quad R_2 = 77/100 = 0.77$$
$$X_3 = (17 \times 77 + 43) \bmod 100 = 1352 \bmod 100 = 52; \quad R_3 = 52/100 = 0.52$$
$$X_4 = (17 \times 52 + 43) \bmod 100 = 927 \bmod 100 = 27; \quad R_4 = 27/100 = 0.27$$
$$X_5 = (17 \times 27 + 43) \bmod 100 = 502 \bmod 100 = 2; \quad R_5 = 2/100 = 0.02$$
$$\vdots \qquad\qquad \vdots$$

Generating Pseudo-Random Numbers with LCM

In the previous example:

- The period (cycle length) is 4 ($X1 = X5$).
- How do we increase the cycle length?

Uniformity and Independence

For *R1, R2, ….* To better imitate uniformity and independence, the integers $X_1, X_2, \ldots$ must have two properties:

- **Maximum Density**: The value m should be as large as possible to generate R_i from many possible values between 0 and 1.
- **Maximum Period**: by proper choice of the parameters $X_0, a, c, and\ m,$ we need to prevent cycling and have the largest possible period.
    - **Cycling**: Recurrence of the same sequence of generated numbers.
    - It would be best if X_i is generated once all numbers between *0 and m-1* are generated for once.

A Reasonable Choice for LCM Parameters

- A reasonable choice for LCM parameters that enable R1, R2, …. To imitate uniformity and independence is as follows: (Lewis et al., 1969):
  - $a = 7^5 = 16807$
  - $c = 0$
  - $m = 2^{31} - 1 = 2{,}147{,}483{,}647$
  - $x_0 \ can \ be \ anything.$
- With that setting, the maximum period is achieved (period is $m - 1 = 2^{31} - 2$).
- Random numbers can be generated by:

$$R_i = \frac{X_i}{m+1}, i = 1,2,,,,$$

- Check with $\mathrm{a} = 7^5 = 16807, c = 0\ , m = 2^{31} - 1$
- The period (cycle length) is $2^{31} - 2 = 2{,}147{,}483{,}646$

Issues with LCG

- Statistical properties
  - Uniformity, independence
  - Maximum density leaves no large gaps in [0,1]
  - Cycling
  - Must have a proper choice of a, c, m and X_0
- Cycle length
  - Typically m=2.1 billion = $2^{31} - 1$ or more

Efficiency of LCGs

- Speed and efficiency are aided by using modulus m which is either a power of 2 or close to power of 2.
- Most digital computers use binary representation of numbers, modulo or remaindering operation can be conducted efficiently when modulo is power of 2.
- In the remaindering operation of $aX + c$, only the b rightmost binary digits are considered.

Combined LCGs

- As computing power has increased, the complexity of the systems that we are able to simulate has also increased.
- One approach is to combine two or more multiplicative congruential generators in such a way that the combined generators have good statistical properties and a longer period.

Testing for uniformity and independence

- Since RNGs are completely deterministic, we need to test them to see if they appear to be random and "Independent and Identically Distributed" IID uniform on [0,1].
- There are two types of tests:
    - Frequency Test
        - Kolmogorov-Simirnov test
        - Chi-squares Test
    - Autocorrelation test

Hypotheses for testing of uniformity

- In testing for uniformity, the hypotheses are as follows:
  - $H_0: R_i \sim U/[0,1]$
  - $H_1: R_i \nsim U/[0,1]$

- The null hypothesis, H_0 reads the numbers are distributed uniformly in the interval [0,1].
- Failure to reject the null hypothesis means that no evidence of non-uniformity has been detected on the basis of this test.

Hypotheses for testing and independence

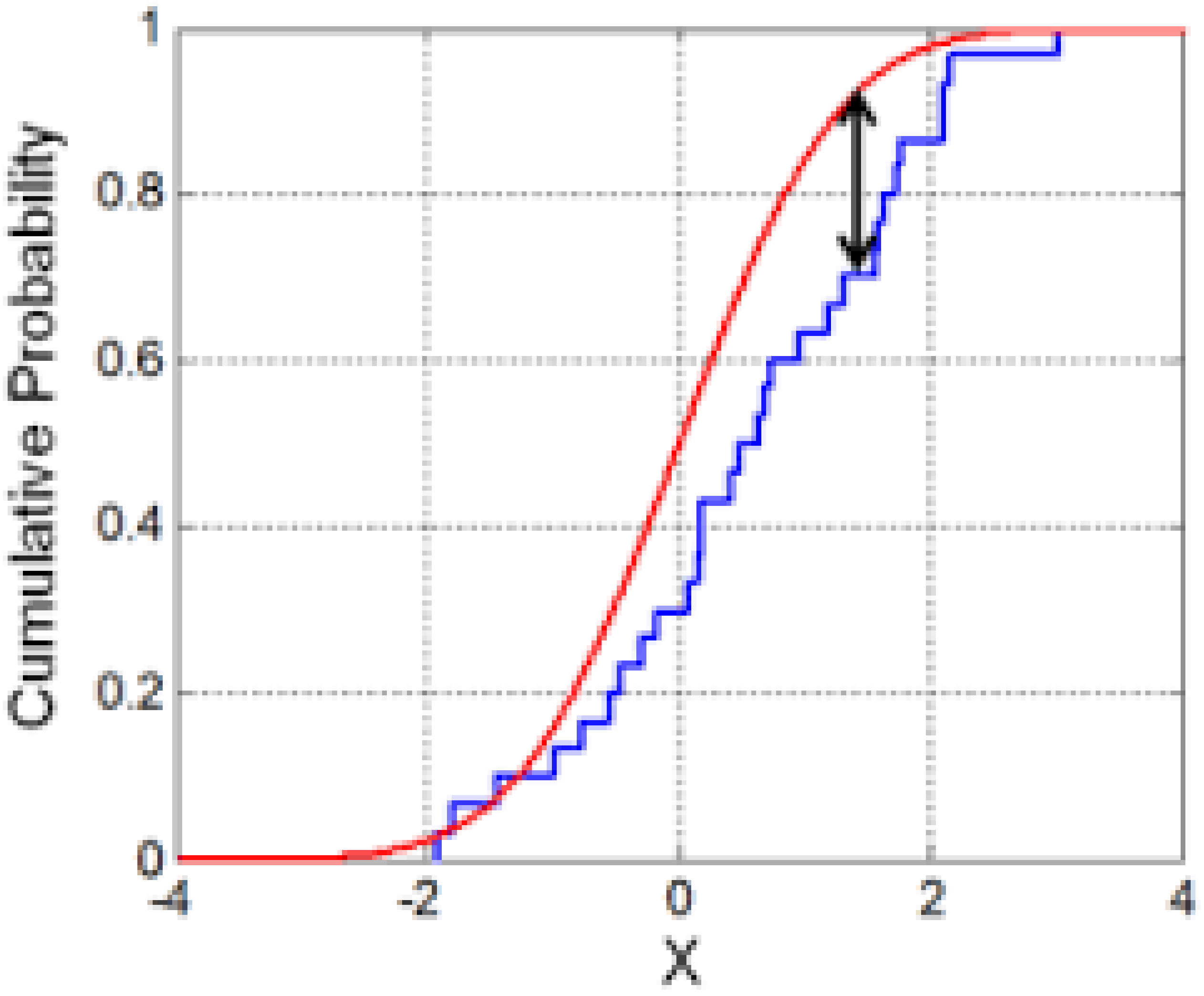
- In testing for independence, the hypotheses are as follows:
  - $H_0 \colon R_i \sim independently$ ➔ the data follow a specific distribution
  - $H_1 \colon R_i \nsim independently$ ➔ the data do not follow the specified distribution

- This null hypothesis H_0 reads that the numbers are independent,
- Failure to reject the null hypothesis means that no evidence of dependence has been detected based on this test.
- This does not imply that further testing of the generator for independence is unnecessary.

Uniformity Test

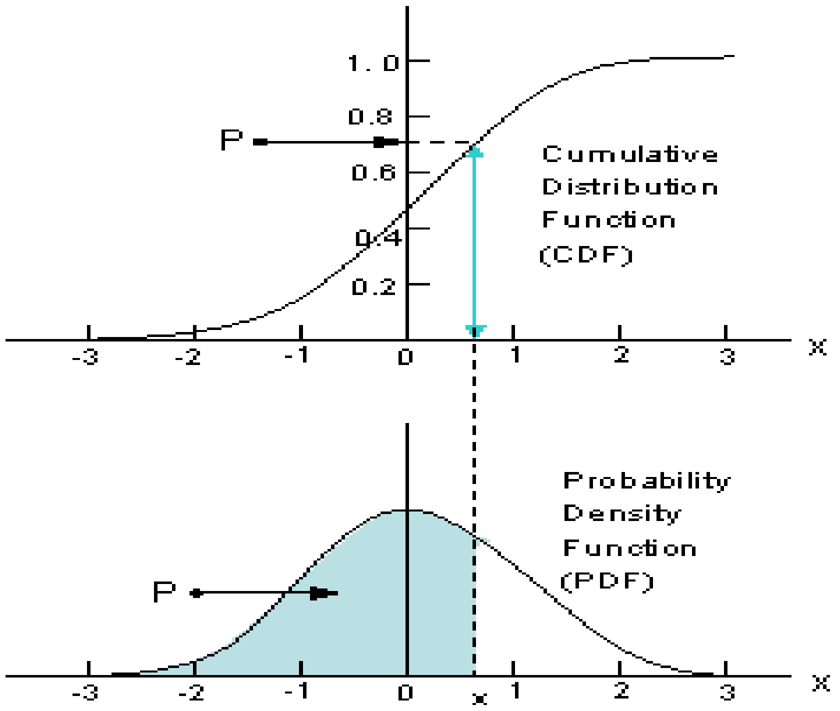
- Two methods of testing uniformity:
  - Kolmogorov—Smirnov test
  - Chi-square test
- Both tests measure the degree of agreement between the distribution of a sample of generated random numbers and the uniform distribution.
- Based on the null hypothesis of no significant difference between sample distribution and theoretical distribution.

Kolmogorov-Smirnov Test



Probability Density Function & Cumulative Distribution Function for normal distribution

The CDF is the probability that random variable values less than or equal to x whereas the PDF is the probability that a random variable, say X, will take a value exactly equal to x.



Relations Between Two Different Typical Representations of a Population

Kolmogorov—Smirnov test

- A statistical hypothesis test.
- Kolmogorov—Smirnov test can be used to compare actual data to normal distribution
  - The cumulative probabilities of values in the data are compared with the cumulative probabilities in the theoretical normal distribution.
- Null-hypothesis: The sample is taken from a normal distribution
- The critical value of D_{a} is found from K-S table values for one sample test (default = 0.565), where a is the level of significance.
- Acceptance Criteria: if the calculated value is less than the critical value, then we accept the null hypothesis $(D < D_{\mathrm{a}})$.
- Reject Criteria: if the calculated value is greater than the table's value, then reject the null hypothesis.

K-S Test

- The K-S test is defined by:
  - H_0: The data follow a specific distribution
  - H_a:The data do not follow the specify distribution

- Test statics: the Kolmogorov-Smirnov test statistics are defined as:
- $D = \max_{1 \le Y \le N}[F(Y_i) - \frac{(i-1)}{N}, \frac{i}{N} - F(Y_i)]$
- where ***F*** is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution (i.e., no discrete distributions such as the binomial or Poisson),
- $Y_i$is N *ordered* data points

Example

- Consider the sequence of 5 numbers
  - 0.15, 0.94, 0.05, 0.51, and 0.29
  - Given a = 0.05
  - Critical Value D_a = 0.565
- Null Hypothesis: Whether the hypothesis of the uniformity can be rejected.
  - $D = \max_{1 \le Y \le N} [F(Y) - \frac{(i-1)}{N}, \frac{i}{N} - F(Y_i)]$

i	1	2	3	4	5
F(Yi)	0.05	0.15	0.29	0.51	0.94
i/N	0.2	0.4	0.6	0.8	1
i/N – F(Yi)	0.15	0.25	0.31	0.29	0.06
i-1/N	0	0.20	0.40	0.60	0.8
F(Yi) – [i-1/N]	0.05	-0.05	-0.11	-0.09	0.14

Example Continued

- $D = \max_{1 \leq Y \leq N} [F(Y\ ) - \frac{(i-1)}{N}, \frac{i}{N} - F(Y_i)]$
- $F(Y_i) - [\frac{(i-1)}{N}] = 0.14$
- $\frac{i}{N} - F(Y\ ) = 0.31$
- D = 0.31, D_a = 0.565
- If D< D_a , the Null Hypothesis is accepted
- 0.31 < 0.565

K-S limitations

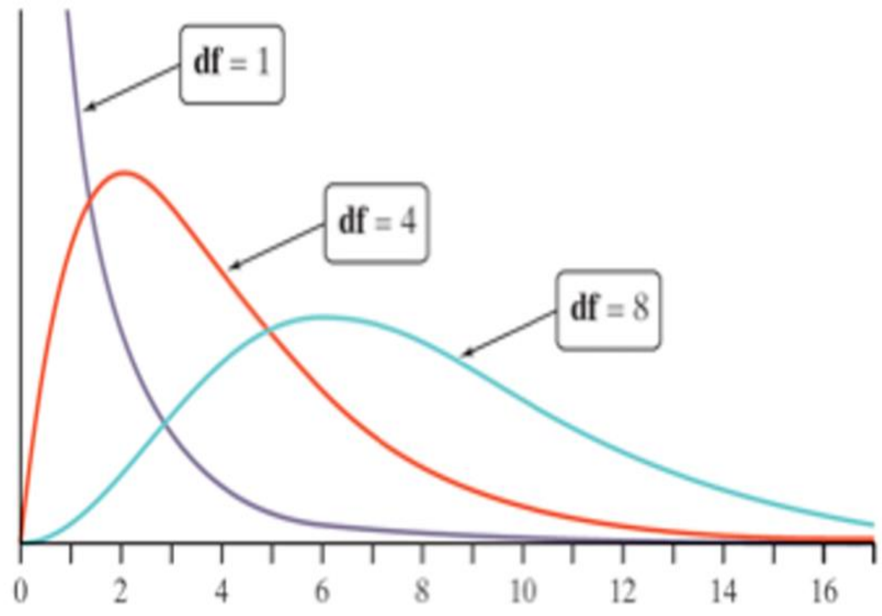
- It only applies to continuous distribution.
- It tends to be more sensitive near the center of the distribution than at the tails.
- It typically determined by simulation

Chi-squares Test

- A chi-square (χ^2) statistic is a measure of the difference between the **observed** and **expected** frequencies of the outcomes of a set of events or variables.
- Chi-square is useful for analyzing such differences in categorical variables, especially those nominal in nature.
- χ^2 depends on the size of the difference between actual and observed values, the degrees of freedom, and the sample size.
- χ^2 can be used to test whether two variables are related or independent from one another.
- It can also be used to test the goodness-of-fit between an observed distribution and a theoretical distribution of frequencies.

The Chi-Square Distributions

- The chi-square distributions are a family of distributions that take only positive values and are skewed to the right.
- A particular chi-square distribution is specified by giving its degrees of freedom.
- The chi-square test for a two-way table with r rows and c columns uses critical values from the chi-square distribution with (r – 1)(c – 1) degrees of freedom.
- The P-value is the area under the density curve of this chi-square distribution to the right of the value of the test statistic



df=(#rows-1)*(#columns-1)

Example:
df=(2-1)*(2-1)=1

Chi-squares Test

- The Formula for Chi-Square is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

$c =$ Degrees of freedom

$O =$ Observed value(s)

$E =$ Expected value(s)

Example

- A school principal would like to know which days of the week students are most likely to be absent. The principal expects that students will be absent equally during the 5-day school week. The principal selects a random sample of 100 teachers asking them which day of the week they had the highest number of student absences.
- The observed and expected results are shown in the table below. Based on these results, do the days for the highest number of absences occur with equal frequencies? Use 5% significance level.

	Monday	Tuesday	Wednesday	Thursday	Friday
Observed Absences	23	16	14	19	28
Expected Absences	20	20	20	20	20

Example Continued

- c= n-1 = 5-1 =4

$$\chi^2_c = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

c = Degrees of freedom

O = Observed value(s)

E = Expected value(s)

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

Example Continued

	Monday	Tuesday	Wednesday	Thursday	Friday
Observed Absences	23	16	14	19	28
Expected Absences	20	20	20	20	20

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

$c =$ Degrees of freedom

$O =$ Observed value(s)

$E =$ Expected value(s)

$$X_c^2 = \frac{3^2}{20} + \frac{(-4)^2}{20} + \frac{(-6)^2}{20} + \frac{(-1)^2}{20} + \frac{8^2}{20} = \frac{126}{20} = 6.3$$

The calculated chi-square value is 6.3 smaller than the critical value (9.49)

As you see, the null hypothesis is not rejected. Therefore, we fail to reject the null hypothesis at the 5% significance level.

Decision: Accept the null hypothesis that the data is the days of the highest number of absences occur with relatively equal frequencies.

References

- Bossel, H., 2013. *Modeling and simulation*. Springer-Verlag.
- Carson, John S. "Introduction to modeling and simulation." *Proceedings of the Winter Simulation Conference, 2005.*. IEEE, 2005.