Testing Random-Number Generators

Raj Jain
Washington University
Saint Louis, MO 63131
Jain@cse.wustl.edu

These slides are available on-line at:

http://www.cse.wustl.edu/~jain/cse574-06/

1. Chi-square test
2. Kolmogorov-Smirnov Test
3. Serial-correlation Test
4. Two-level tests
5. K-dimensional uniformity or k-distributivity
6. Serial Test
7. Spectral Test

Testing Random-Number Generators

Goal: To ensure that the random number generator produces a random stream.

- Plot histograms
- Plot quantile-quantile plot
- Use other *t*ests
- Passing a test is necessary but not sufficient
- Pass ≠ Good

  Fail ⇒ Bad
- New tests ⇒ Old generators fail the test
- Tests can be adapted for other distributions

Chi-Square Test

- Most commonly used test
- Can be used for any distribution
- Prepare a histogram of the observed data
- Compare observed frequencies with theoretical

k = Number of cells

o_i = Observed frequency for ith cell

e_i = Expected frequency

$$D = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- D=0 ⇒ Exact fit
- D has a chi-square distribution with k-1 degrees of freedom.

⇒ Compare D with $\chi^2_{[1-\alpha;\, k-1]}$ Pass with confidence α if D is less

Example 27.1

- 1000 random numbers with $x_0 = 1$
- $\chi^2_{[0.9;9]} = 14.68$
- Observed difference = 10.380
- Observed is Less $\Rightarrow$ Accept IID U(0, 1)

$$x_n = (125x_{n-1} + 1) \bmod (2^{12})$$

Cell	Obsrvd	Exptd	$\frac{(o-e)^2}{e}$
1	100	100.0	0.000
2	96	100.0	0.160
3	98	100.0	0.040
4	85	100.0	2.250
5	105	100.0	0.250
6	93	100.0	0.490
7	97	100.0	0.090
8	125	100.0	6.250
9	107	100.0	0.490
10	94	100.0	0.360
Total	1000	1000.0	10.380

Chi-Square for Other Distributions

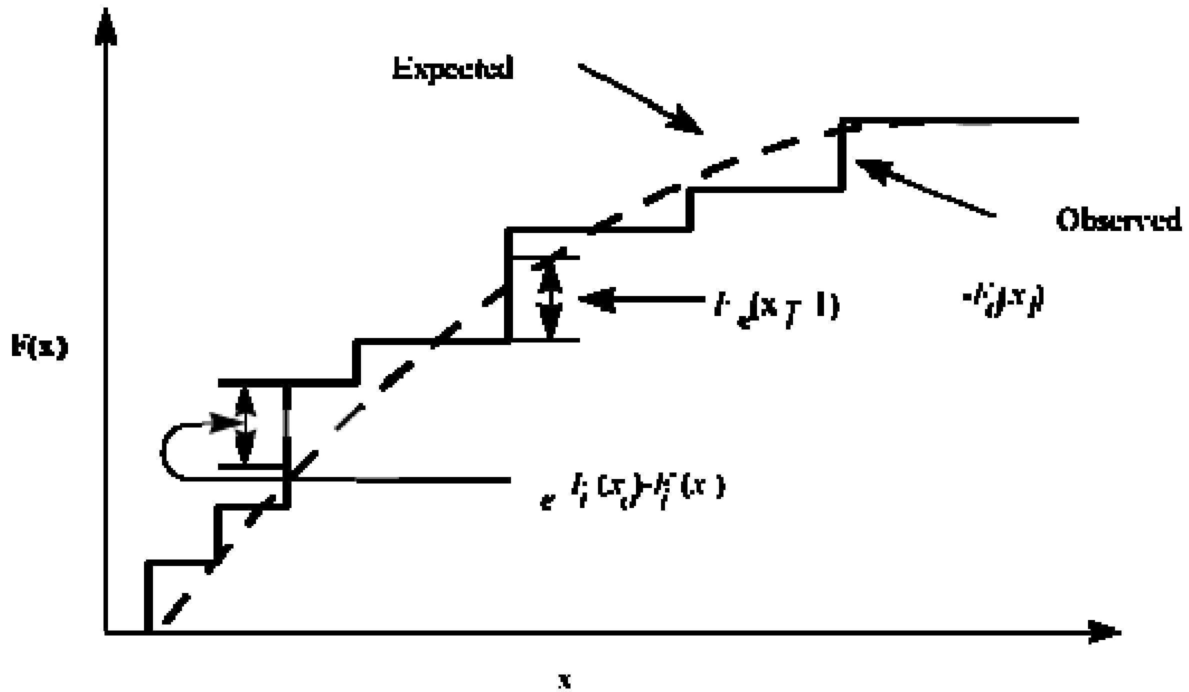
- Errors in cells with a small e_i affect the chi-square statistic more
- Best when e_i's are equal.

⇒ Use an equi-probable histogram with variable cell sizes

- Combine adjoining cells so that the new cell probabilities are approximately equal.
- The number of degrees of freedom should be reduced to k-r-1 (in place of k-1), where r is the number of parameters estimated from the sample.
- Designed for discrete distributions and for large sample sizes only ⇒ Lower significance for finite sample sizes and continuous distributions
- If less than 5 observations, combine neighboring cells

Kolmogorov-Smirnov Test

- Developed by A. N. Kolmogorov and N. V. Smirnov
- Designed for continuous distributions
- Difference between the observed CDF (cumulative distribution function) $F_o(x)$ and the expected cdf $F_e(x)$ should be small.



Kolmogorov-Smirnov Test

- K^+ = maximum observed deviation below the expected cdf
- K^- = minimum observed deviation below the expected cdf

$$K^+ = \sqrt{n} \max_x \left(F_o(x) - F_e(x)\right)$$

$$K^- = \sqrt{n} \max_x \left(F_e(x) - F_o(x)\right)$$

- $K^+ < K_{[1-\alpha;n]}$ and $K^- < K_{[1-\alpha;n]} \Rightarrow$ Pass at α level of significance.
- Don't use max/min *of* $Fe(x_i)-F_o(x_i)$
- Use $F_e(x_{i+1})-F_o(x_i)$ for K^-
- For $U(0, 1)$: $F_e(x)=x$
- $F_o(x) = j/n$,
  where $x > x_1, x_2, \ldots, x_{j-1}$

$$K^+ = \sqrt{n} \max_j \left(\frac{j}{n} - x_j\right)$$

$$K^- = \sqrt{n} \max_j \left(x_j - \frac{j-1}{n}\right)$$

Example 27.2

30 Random numbers using a seed of x_0=15:

$$x_n = 3x_{n-1} \bmod 31$$

- The numbers are:

14, 11, 2, 6, 18, 23, 7, 21, 1, 3, 9, 27, 19, 26, 16, 17, 20, 29, 25, 13, 8, 24, 10, 30, 28, 22, 4, 12, 5, 15.

Example 27.2 (Cont)

The normalized numbers obtained by dividing the sequence by 31 are:

0.45161, 0.35484, 0.06452, 0.19355, 0.58065, 0.74194,
0.22581, 0.67742, 0.03226, 0.09677, 0.29032, 0.87097,
0.61290, 0.83871, 0.51613, 0.54839, 0.64516, 0.93548,
0.80645, 0.41935, 0.25806, 0.77419, 0.32258, 0.96774,
0.90323, 0.70968, 0.12903, 0.38710, 0.16129, 0.48387.

Example 27.2 (Cont)

- $K_{[0.9;n]}$ value for n = 30 and a = 0.1 is 1.0424

$$
\begin{aligned}
K^- &= \sqrt{n} \max_j \left(x_j - \frac{j-1}{n}\right) \\
&= \sqrt{30} \times 0.03026 \\
&= 0.1767
\end{aligned}
$$

$$
\begin{aligned}
K^+ &= \sqrt{n} \max_j \left(\frac{j}{n} - x_j\right) \\
&= \sqrt{30} \times 0.03026 \\
&= 0.1767
\end{aligned}
$$

- Observed<Table
  $\Rightarrow$ Pass

j	x_j	$\frac{j}{n} - x_j$	$x_j - \frac{j-1}{n}$
1	0.03226	0.00108	0.03226
2	0.06452	0.00215	0.03118
3	0.09677	0.00323	0.03011
4	0.12903	0.00430	0.02903
5	0.16129	0.00538	0.02796
6	0.19355	0.00645	0.02688
7	0.22581	0.00753	0.02581
8	0.25806	0.00860	0.02473
$\vdots$			
29	0.93548	0.03118	0.00215
30	0.96774	0.03226	0.00108
	Max	0.03226	0.03226

Chi-square vs. K-S Test

K-S test	Chi-Square Test
Small samples	Large Sample
Continuous distributions	Discrete distributions
Differences between observed and expected cumulative probabilities (CDFs)	Differences between observed and hypothesized probabilities (pdfs or pmfs).
Uses each observation in the sample without any grouping $\Rightarrow$makes a better use of the data	Groups observations into a small number of cells
Cell size is not a problem	Cell sizes affect the conclusion but no firm guidelines
Exact	Approximate

Serial-Correlation Test

- Nonzero covariance ⇒ Dependence. The inverse is not true
- R_k = Autocovariance at lag k = Cov[x_n, x_{n+k}]

$$R_k = \frac{1}{n-k} \sum_{i=1}^{n-k} (U_i - \frac{1}{2})(U_{i+k} - \frac{1}{2})$$

- For large n, R_k is normally distributed with a mean of zero and a variance of $1/[144(n-k)]$
- 100(1-α)% confidence interval for the autocovariance is:

$$R_k \mp z_{1-\alpha/2}/(12\sqrt{n-k})$$

For $k \geq 1$ Check if CI includes zero

- For $k = 0$, R_0= variance of the sequence Expected to be 1/12 for IID $U(0,1)$

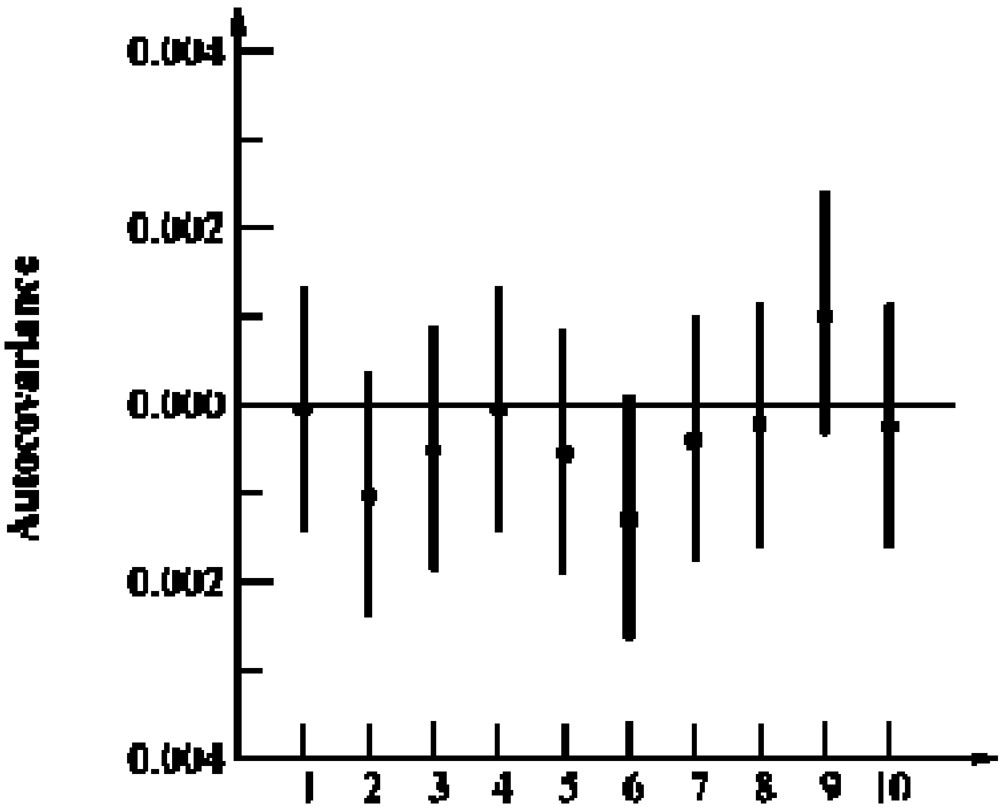
Example 27.3: Serial Correlation Test

$$x_n = 7^5 x_{n-1} \bmod (2^{31} - 1)$$

10,000 random numbers with x_0=1:

Lag k	Autocovariance R_k	St. Dev. of R_k	90% Confidence Interval Lower Limit	90% Confidence Interval Upper Limit
1	-0.000038	0.000833	-0.001409	0.001333
2	-0.001017	0.000833	-0.002388	0.000354
3	-0.000489	0.000833	-0.001860	0.000882
4	-0.000033	0.000834	-0.001404	0.001339
5	-0.000531	0.000834	-0.001902	0.000840
6	-0.001277	0.000834	-0.002648	0.000095
7	-0.000385	0.000834	-0.001757	0.000986
8	-0.000207	0.000834	-0.001579	0.001164
9	0.001031	0.000834	-0.000340	0.002403
10	-0.000224	0.000834	-0.001595	0.001148

Example 27.3 (Cont)



- All confidence intervals include zero ⇒ All covariances are statistically insignificant at 90% confidence.

Two-Level Tests

- If the sample size is too small, the test results may apply locally, but not globally to the complete cycle.
- Similarly, global test may not apply locally
- Use two-level tests

$\Rightarrow$ Use Chi-square test on n samples of size k each and then use a Chi-square test on the set of n Chi-square statistics so obtained

$\Rightarrow$ Chi-square on Chi-square test.

- Similarly, *K-S* on *K-S*
- Can also use this to find a ``nonrandom'' segment of an otherwise random sequence.

k-Distributivity

- k-Dimensional Uniformity
- Chi-square ⇒ uniformity in one dimension
  ⇒ Given two real numbers a_1 and b_1 between 0 and 1 such that $b_1 > a_1$

$$P(a_1 \le u_n < b_1) = b_1 - a_1 \quad \forall b_1 > a_1$$

- This is known as 1-distributivity property of u_n.
- The 2-distributivity is a generalization of this property in two dimensions:

$$P(a_1 \le u_{n-1} < b_1 \text{ and } a_2 \le u_n < b_2)$$

$$= (b_1 - a_1)(b_2 - a_2)$$

For all choices of a_1, b_1, a_2, b_2 in [0, 1], $b_1 > a_1$ and $b_2 > a_2$

k-Distributivity (Cont)

- k-distributed if:

$$P(a_1 \le u_n < b_1, \ldots, a_k \le u_{n+k-1} < b_k)$$

$$(b_1 - a_1) \cdots (b_k - a_k)$$

- For all choices of a_i, b_i in [0, 1], with $b_i > a_i$, i=1, 2, ..., k.
- k-distributed sequence is always (k-1)-distributed. The inverse is not true.
- Two tests:

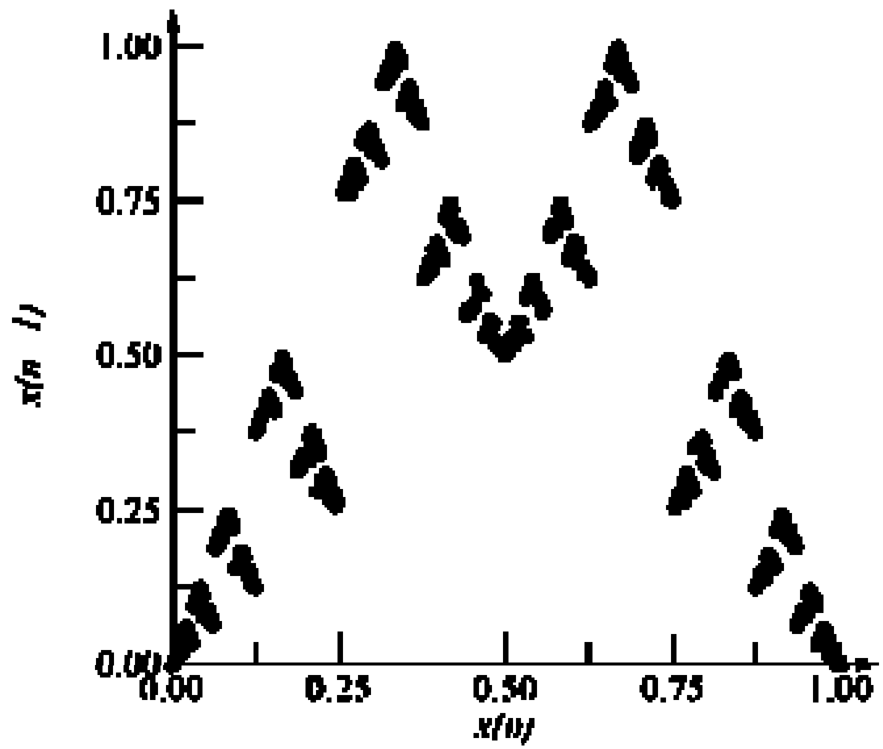
1. Serial test
2. Spectral test
3. Visual test for 2-dimensions: Plot successive overlapping pairs of numbers

Example 27.4

- *T*ausworthe sequence generated by:

  $$x^{15} + x + 1$$

- The sequence is *k*-distributed for *k* up to $\lceil /l \rceil$, that is, *k*=1.
- In two dimensions: Successive overlapping pairs (x_n, x_{n+1})

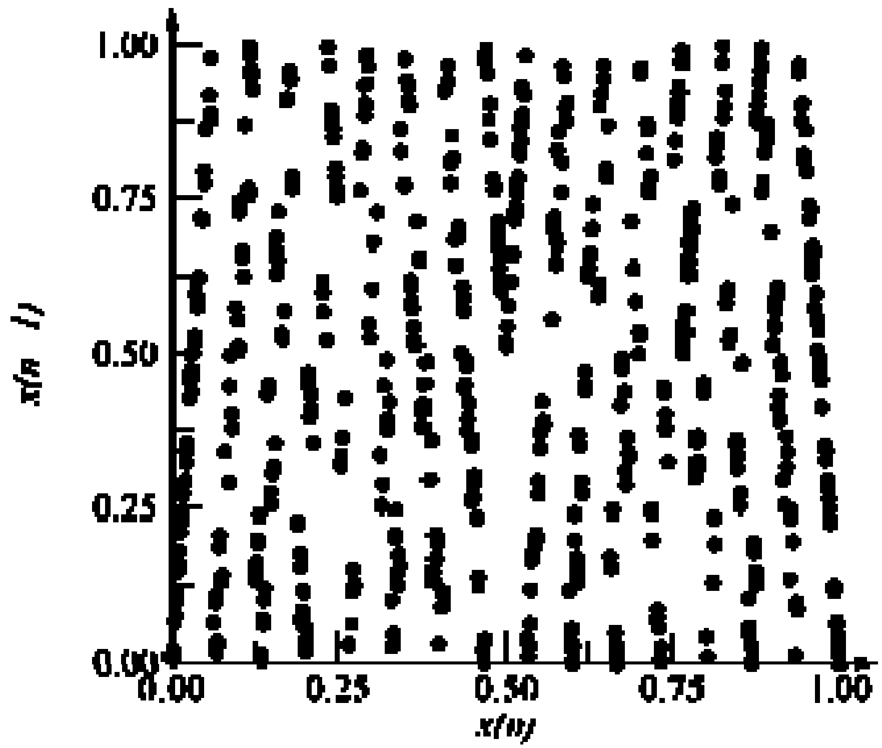


Example 27.5

- Consider the polynomial:

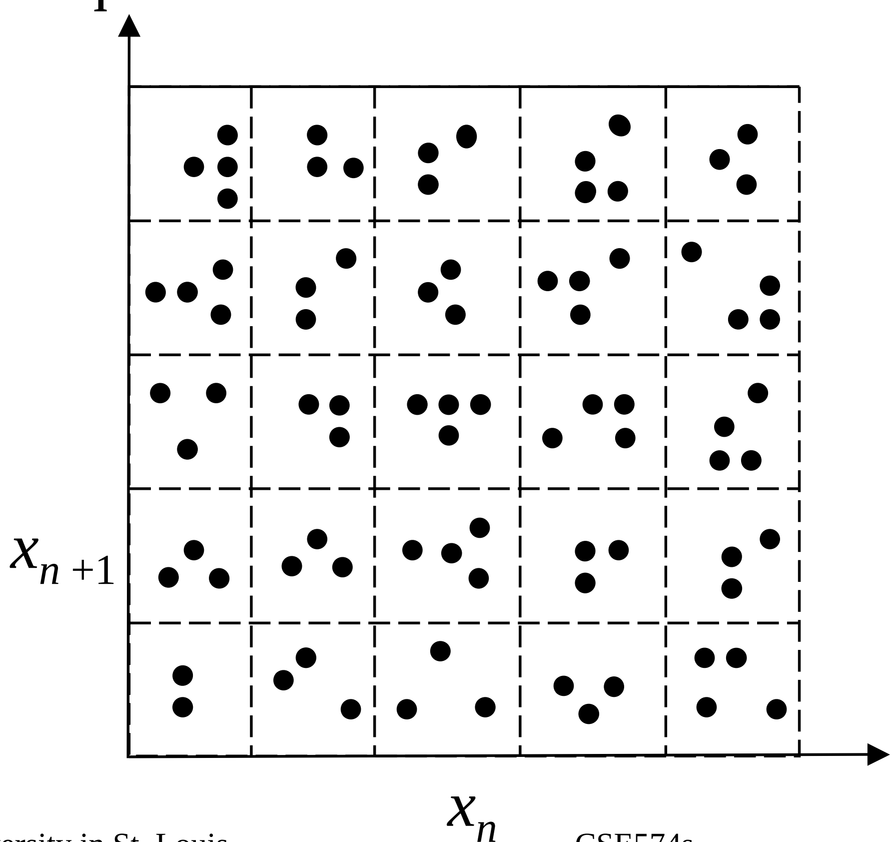
  $x^{15} + x^4 + 1$

- Better 2-distributivity than Example 27.4



Serial Test

- Goal: To test for uniformity in two dimensions or higher.
- In two dimensions, divide the space between 0 and 1 into K^2 cells of equal area



Serial Test (Cont)

- Given $\{x_1, x_2, \ldots, x_n\}$, use $n/2$ non-overlapping pairs (x_1, x_2), (x_3, x_4), … and count the points in each of the K^2 cells.
- Expected= $n/(2K^2)$ points in each cell.
- Use chi-square test to find the deviation of the actual counts from the expected counts.
- The degrees of freedom in this case are K^2-1.
- For k-dimensions: use k-tuples of non-overlapping values.
- k-tuples must be non-overlapping.
- Overlapping $\Rightarrow$ number of points in the cells are not independent chi-square test cannot be used
- In visual check one can use overlapping or non-overlapping.
- In the spectral test overlapping tuples are used.
- Given n numbers, there are n-1 overlapping pairs, n/2 non-overlapping pairs.

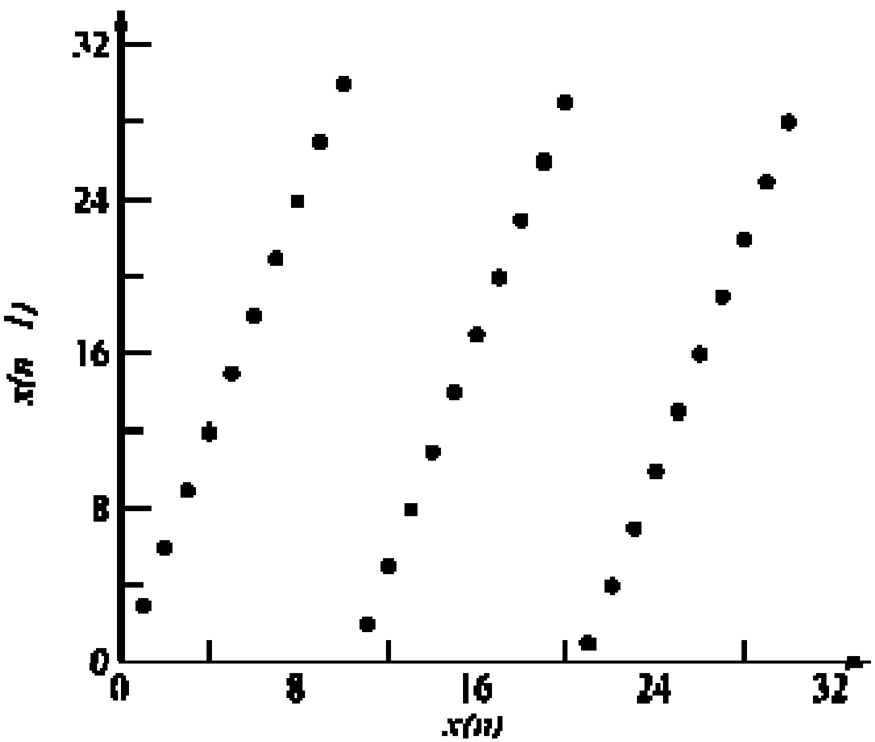
Spectral Test

- Goal: To determine how densely the k-tuples $\{x_1, x_2, \ldots, x_k\}$can fill up the *k*-dimensional hyperspace.
- The *k*-tuples from an LCG fall on a finite number of parallel hyper-planes.
- Successive pairs would lie on a finite number of lines
- In three dimensions, successive triplets lie on a finite number of planes.

Example 27.6: Spectral Test

$$x_n = 3x_{n-1} \bmod 31$$

Plot of overlapping pairs



- All points lie on three straight lines.

$$x_n = 3x_{n-1}$$
$$x_n = 3x_{n-1} - 31$$
$$x_n = 3x_{n-1} - 62$$

- Or:

$$x_n = 3x_{n-1} - 31k \quad k = 0, 1, 2$$

Example 27.6 (Cont)

- In three dimensions, the points (x_n, x_{n-1}, x_{n-2}) for the above generator would lie on five planes given by:

$$x_n = 2x_{n-1} + 3x_{n-2} - 31k \quad k = 0, 1, \ldots, 4$$

Obtained by adding the following to equation

$$x_{n-1} = 3x_{n-2} - 31k_1 \quad k_1 = 0, 1, 2$$

Note that $k+k_1$ will be an integer between 0 and 4.

Spectral Test (More)

- Marsaglia (1968): Successive k-tuples obtained from an LCG fall on, at most, $(k!m)^{1/k}$ parallel hyper-planes, where m is the modulus used in the LCG.
- Example: $m = 2^{32}$, fewer than 2,953 hyper-planes will contain all 3-tuples, fewer than 566 hyper-planes will contain all 4-tuples, and fewer than 41 hyper-planes will contain all 10-tuples. Thus, this is a weakness of LCGs.
- Spectral Test: Determine the max distance between adjacent hyper-planes.
- Larger distance $\Rightarrow$ worse generator
- In some cases, it can be done by complete enumeration

Example 27.7

- Compare the following two generators:

$$x_n = 3x_{n-1} \bmod 31$$

$$x_n = 13x_{n-1} \bmod 31$$

- Using a seed of x_0=15, first generator:

14, 11, 2, 6, 18, 23, 7, 21, 1, 3, 9, 27, 19, 26, 16, 17, 20, 29, 25, 13, 8, 24, 10, 30, 28, 22, 4, 12, 5, 15, 14.

- Using the same seed in the second generator:

9, 24, 2, 26, 28, 23, 20, 12, 1, 13, 14, 27, 10, 6, 16, 22, 7, 29, 5, 3, 8, 11, 19, 30, 18, 17, 4, 21, 25, 15, 9.

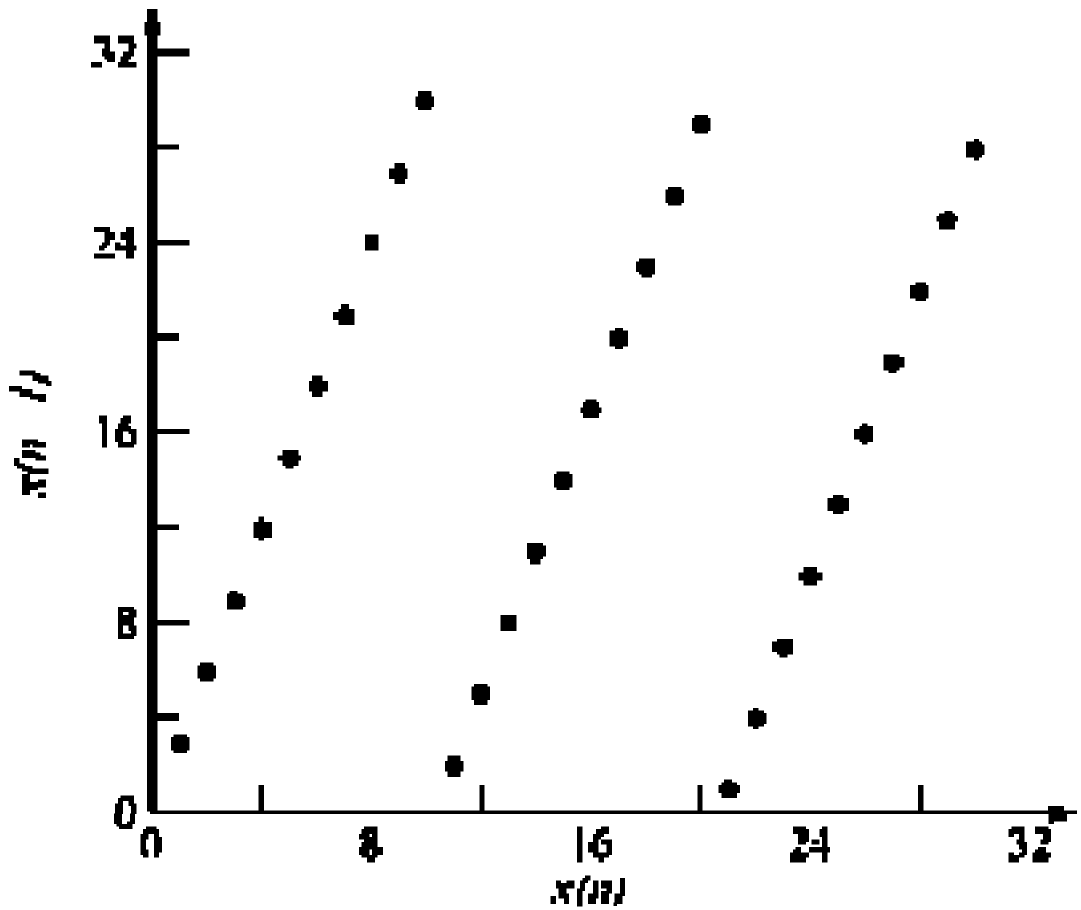
Example 27.7 (Cont)

- Every number between 1 and 30 occurs once and only once

$\Rightarrow$ Both sequences will pass the chi-square test for uniformity

Example 27.7 (Cont)

- First Generator:



Example 27.7 (Cont)

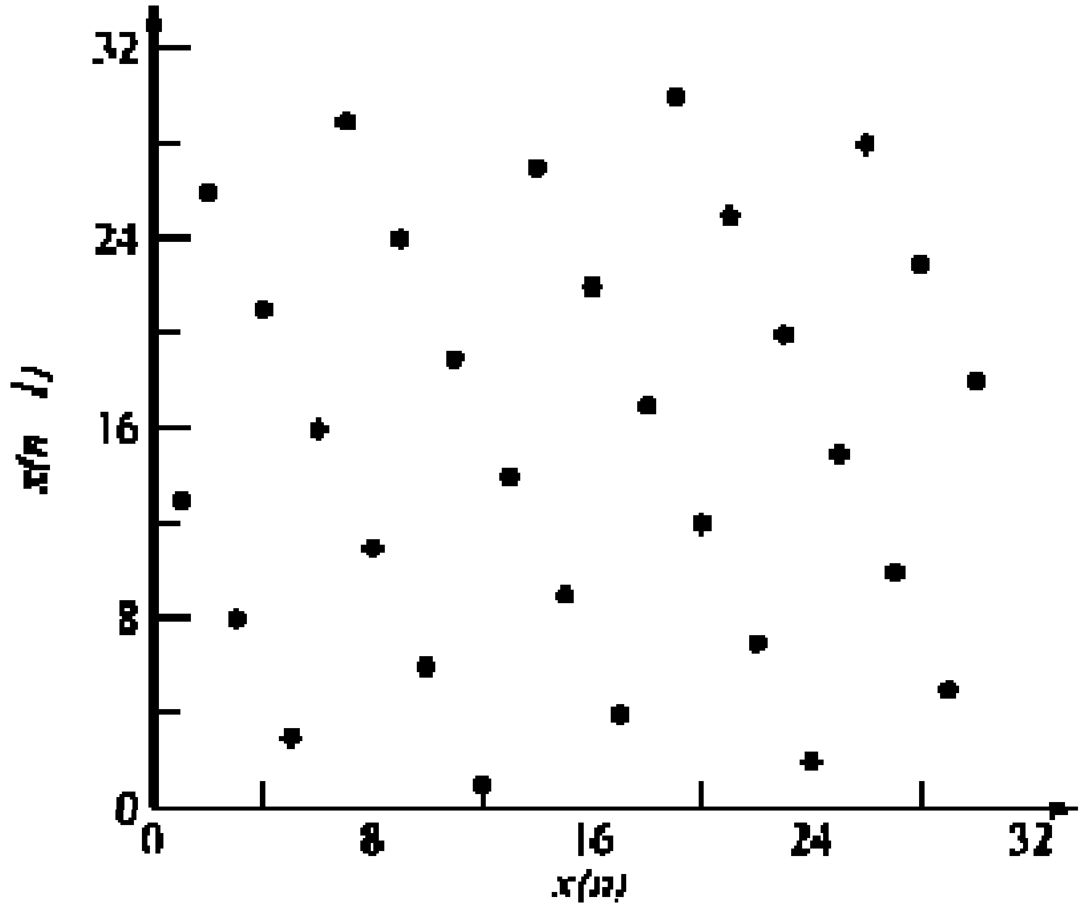
- Three straight lines of positive slope or ten lines of negative slope
- Since the distance between the lines of positive slope is more, consider only the lines with positive slope.

$$\begin{aligned} x_n &= 3x_{n-1} \\ x_n &= 3x_{n-1} - 31 \\ x_n &= 3x_{n-1} - 62 \end{aligned}$$

- Distance between two parallel lines $y=ax+c_1$ and $y=ax+c_2$ is given by $|c_2 - c_1|/\sqrt{1+a^2}$
- The distance between the above lines is $31/\sqrt{10}$ or 9.80.

Example 27.7 (Cont)

- Second Generator:



Example 27.7 (Cont)

- All points fall on seven straight lines of positive slope or six straight lines of negative slope.
- Considering lines with negative slopes:

$$x_n = -\frac{5}{2}x_{n-1} + k\frac{31}{2} \quad k = 0, 1, \ldots, 5$$

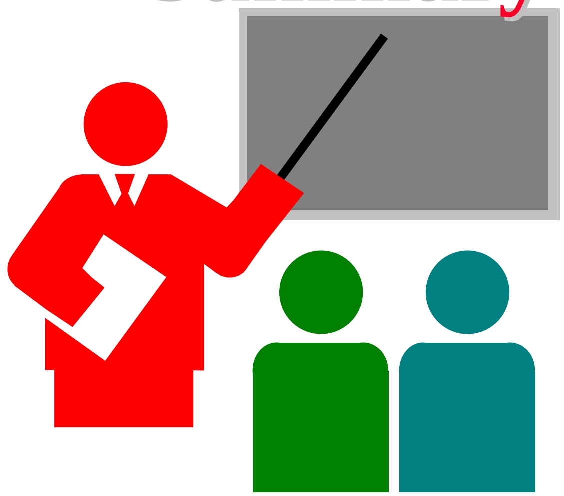
- The distance between lines is: $(31/2)/\sqrt{(1+(5/2)^2)}$ or 5.76.
- The second generator has a smaller maximum distance and, hence, the second generator has a better 2-distributivity.
- The set with a larger distance may **not** always be the set with fewer lines.

Example 27.7 (Cont)

- Either overlapping or non-overlapping *k*-tuples can be used.
  - With overlapping *k*-tuples, we have k times as many points, which makes the graph visually more complete.The number of hyper-planes and the distance between them are the same with either choice.
- With serial test, only non-overlapping *k*-tuples should be used.
- For generators with a large *m* and for higher dimensions, finding the maximum distance becomes quite complex.

  See Knuth (1981)

Summary



1. Chi-square test is a one-dimensional test
   Designed for discrete distributions and large sample sizes
2. K-S test is designed for continuous variables
3. Serial correlation test for independence
4. Two level tests find local non-uniformity
5. k-dimensional uniformity = k-distributivity
   tested by spectral test or serial test

Homework

- Submit detailed answer to Exercise 27.3. Print 10,000th number also.

Exercise 27.1

Generate 10,000 numbers using a seed of x_0=1 in the following generator:

$$x_n = 7^5 x_{n-1} \bmod (2^{31} - 1)$$

Classify the numbers into ten equal size cells and test for uniformity using the chi-square test at 90% confidence.

Exercise 27.2

Generate 15 numbers using a seed of x_0=1 in the following generator:

$$x_n = (5x_{n-1} + 1) \bmod 16$$

Perform a *K-S* test and check whether the sequence passes the test at a 95% confidence level.

Exercise 27.3

Generate 10,000 numbers using a seed of x_0=1 in the following LCG:

$$x_n = 48271 x_{n-1} \bmod (2^{31} - 1)$$

Perform the serial correlation test of randomness at 90% confidence and report the result.

Exercise 27.4

Using the spectral test, compare the following two generators

$$x_n = 7x_{n-1} \bmod 13$$

$$x_n = 11x_{n-1} \bmod 13$$

Which generator has a better 2-distributivity?