Random Number Generation

PSEUDO-RANDOM NUMBER GENERATION

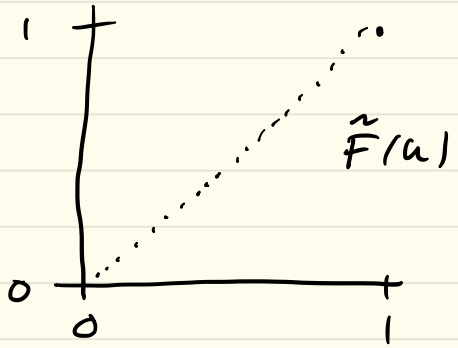
Some function when queried gives a number in $(0, 1)$

Desired Properties

1. UNIFORM
   - many many many samples should look like coming from $U(0,1)$

e.g. the "empirical CDF"

$\tilde{F}(u) = \#\text{samples} \leq u$



INDEPENDCE

- ? How would you test?
- more to come

PSEUDO-RANDOM NUMBER GENERATION (RNG)

IMPORTANT CONSIDERATIONS

- FAST
- PORTABLE BETWEEN COMPUTERS
- REPLICABLE
- CLOSELY APPROXIMATE IDEAL STATISTICAL PROPERTIES OF UNIFORMITY AND INDEPENDENCE
- "LONG CYCLE"

RNGs based on functions produce a sequence of numbers

$X_0\ X_1\ X_2\ \ldots\ X_j\ \ldots\ X_k\ \ldots\ X_{m-1}$

period of cycle

- "cryptographically secure"
  - Given long sequence of past samples, very hard to guess next sample

Common RNGs

Linear Congruential Method (LCM)

produces sequence $X_0, X_1, X_2, \ldots$ by

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

multiplier — increment — modulus

Values chosen for a, c, m, and X_0 have ENORMOUS impact on statistical properties and cycle length

Random numbers are in $[0, m-1]$, converted by

$$R_i = \frac{X_i}{m}$$

Tausworthe Generator

- an option in Möebius

operates on sequence of binary digits

$b_0\ b_1\ \dots\ b_k\ \dots$

numbers formed by grouping bits, e.g. every 64 to get 64 bit unsigned integers

$$b_n = c_{q-1} \cdot b_{n-1} \oplus c_{q-2} \cdot b_{n-2} \oplus \cdots \oplus c_0 \cdot b_{n-q}$$

c_i are binary coefficients

$\cdot$ is logical AND

Need good choices for c_i

COMBINED LINEAR CONGRUENTIAL GENERATOR

- technique to create longer period generator
- Use k different LCM

  j^{th} generator has m_j prime, a_j multiplier
- $X_{i,j}$ is i^{th} output from generator j

$$X_{i,j} \sim U[0, m_j - 1]$$

$$Y_i = \left(\sum_{j=1}^{k} (-1)^{j-1} X_{i,j} \right) \left(\text{mod } m_1 - 1 \right) \qquad R_i = \begin{cases} X_i / m_1, & X_i > 0 \\ (m_1 - 1)/m_1, & X_i = 0 \end{cases}$$

maximum period

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

CRYPTOGRAPHICALLY SECURE RNG

CRNG should pass the "next bit" test

Given the 1st k bits of a random sequence, no polynomial time algorithm can predit the next bit with probability of success > 50%

CRNG should pass "state compromise extensions"

- state of LCM is last number generated
- state of Mersienne Twister is 624 64-bit integers

IN EVENT THAT SOME OR ALL OF STATE IS COMPOMISED, SHOULD BE IMPOSSIBLE TO RECONSTRUCT STREAM OF RANDOM NUMBERS PRIOR TO REVOLATION

Random Number Streams

seed : starting point in sequence

$$x_0 \quad x_1 \quad x_2 \quad \cdots \quad x_j \quad \cdots \quad x_{m-1} \quad x_0 \quad x_1 \quad \cdots$$

start here and $R_1 = \frac{x_j}{m}$, $R_2 = \frac{x_{j+1}}{m}$ etc.

YOU CAN run multiple streams from the same mathematical sequence, separated by b

$$x_0 \quad x_1 \quad \cdots \quad x_{b-1} \quad x_b \quad x_{b+1} \quad \cdots \quad x_{2b-1} \quad x_{2b} \quad x_{2b+1} \quad \cdots$$

$\frac{x_0}{m}, \frac{x_1}{m}, \ldots$ $\frac{x_b}{m}, \frac{x_{b+1}}{m} \ldots$ $\frac{x_{2b}}{m}, \frac{x_{2b+1}}{m} \ldots$

? For really large period b, how do you FIND these starting points??

TESTS FOR RANDOM NUMBERS

Reminder of how statistical testing works....

$S = \{S_1, S_2, \ldots, S_n\}$ a set of DATA SAMPLES, from some (unknown) probability distribution

We wish to look for statistical support for some hypothesis about the data

H_0 : usually called null hypothesis

H_A : usually called alternative hypothesis

- Construct some statistic with known distribution using S
- Ask "What's the probability of seeing this statistic if H_0 is true?

"CONFIDENCE LEVEL" specifies decision

IF probability of observing statistic $< \alpha$

WE REJECT THE NULL HYPOTHESIS

α typically .1, .05, .01

LOGIC ....

" IF H_0 is true, then there is only a (10%, 5%, 1%) chance of data set S yielding up that statistic. So probably H_0 is not true"

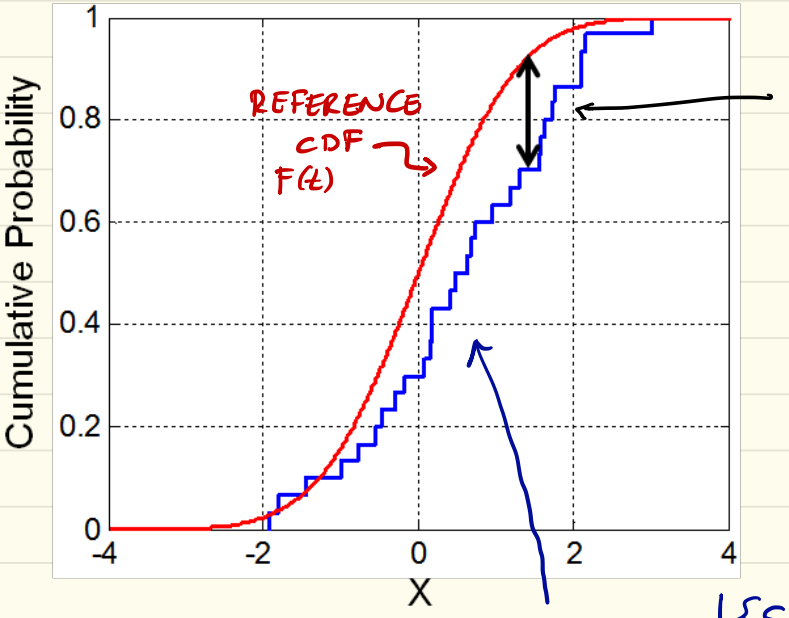
Called $1-\alpha$ confidence level

"WITH 99% confidence we reject H_0"

TEST FOR STATISTICAL UNIFORMITY

One test is "one sample" Kolmogorov–Smirnov Test

– compare empirical CDF with that of the "reference distribution"



$$D = \max_t \left| B(t) - F(t) \right|$$

$$B(t) = \frac{|\{s_i \in S \mid s_i \le t\}|}{|S|}$$

NOTICE — IS LARGEST DEVIATION

TEST FOR STATISTICAL UNIFORMITY

USE K-S TABLE (look up) example

$n \backslash \alpha$	0.01	0.05	0.1
2	0.929	0.842	0.776
4	0.733	0.624	0.564
10	0.490	0.410	0.368
20	0.356	0.294	0.264
50	0.230	0.190	0.170
OVER 50	$\frac{1.63}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$

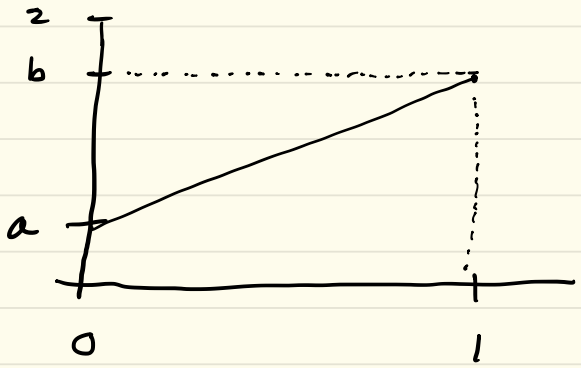
n is # samples

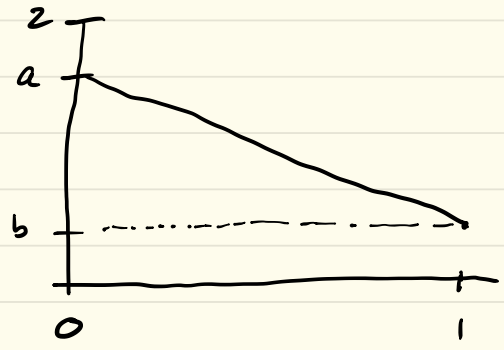
How to use:

1. Choose α for $1-\alpha$ confidence, e.g. 95%
2. Compute empirical distribution function, measure D
3. Find statistic threshold $d(n,\alpha)$ given n & α
4. Reject H_0 if $D > d(n,\alpha)$

Example

consider family of distributions "similar" to U





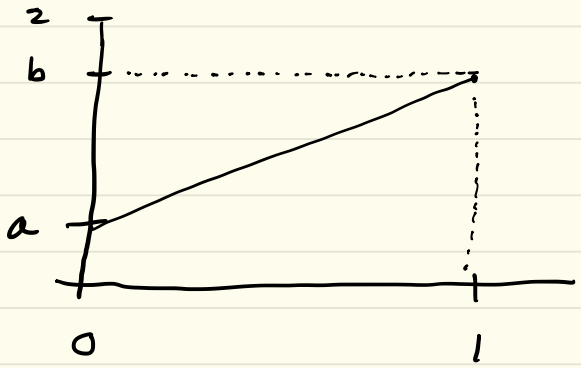
$$f(t) = \frac{a + (b-a)\cdot t}{N(a,b)}$$

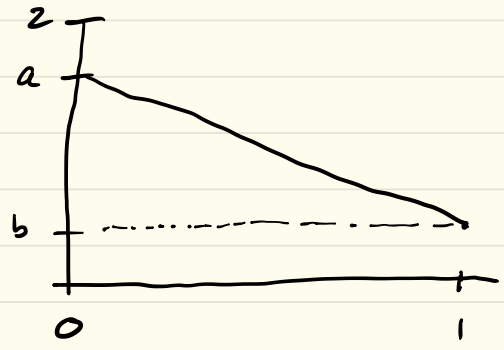
$N(a,b)$ = area under curve $= \min\{a,b\} + |b-a|/2.0$

Notice that $b = 2.0 - a$, and that $U(0,1)$ has $a = b = 1.0$

Example

consider family of distributions "similar" to U

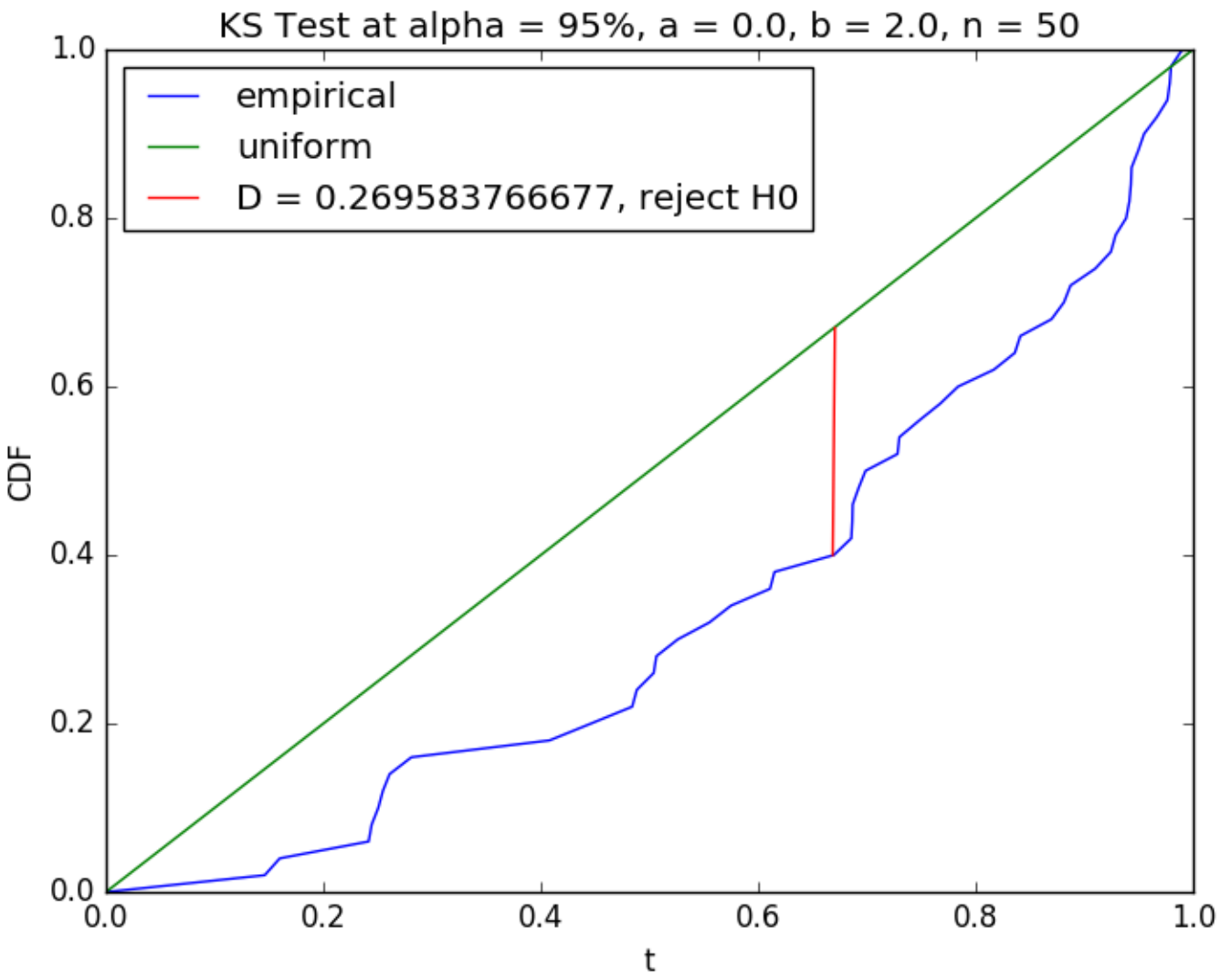


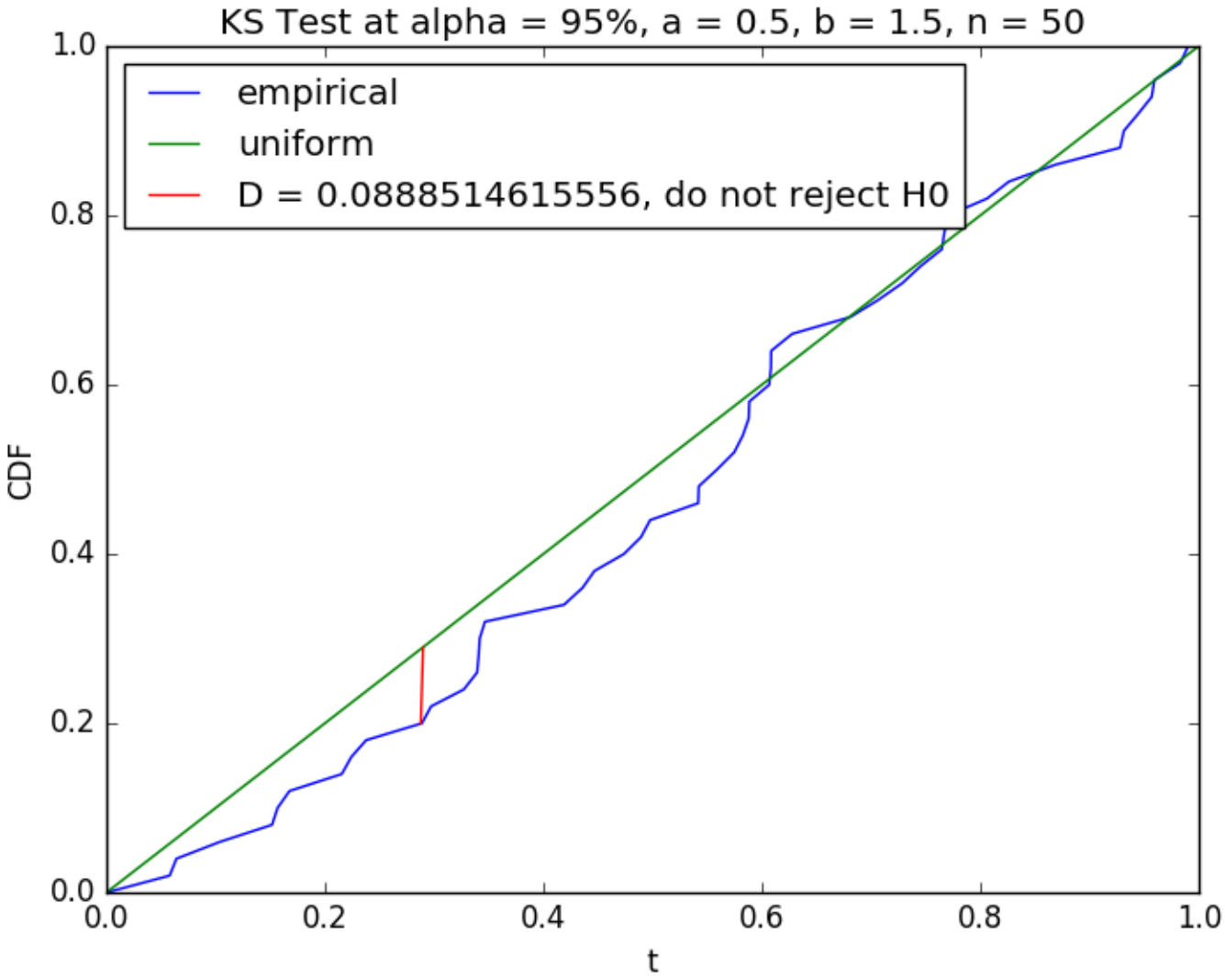


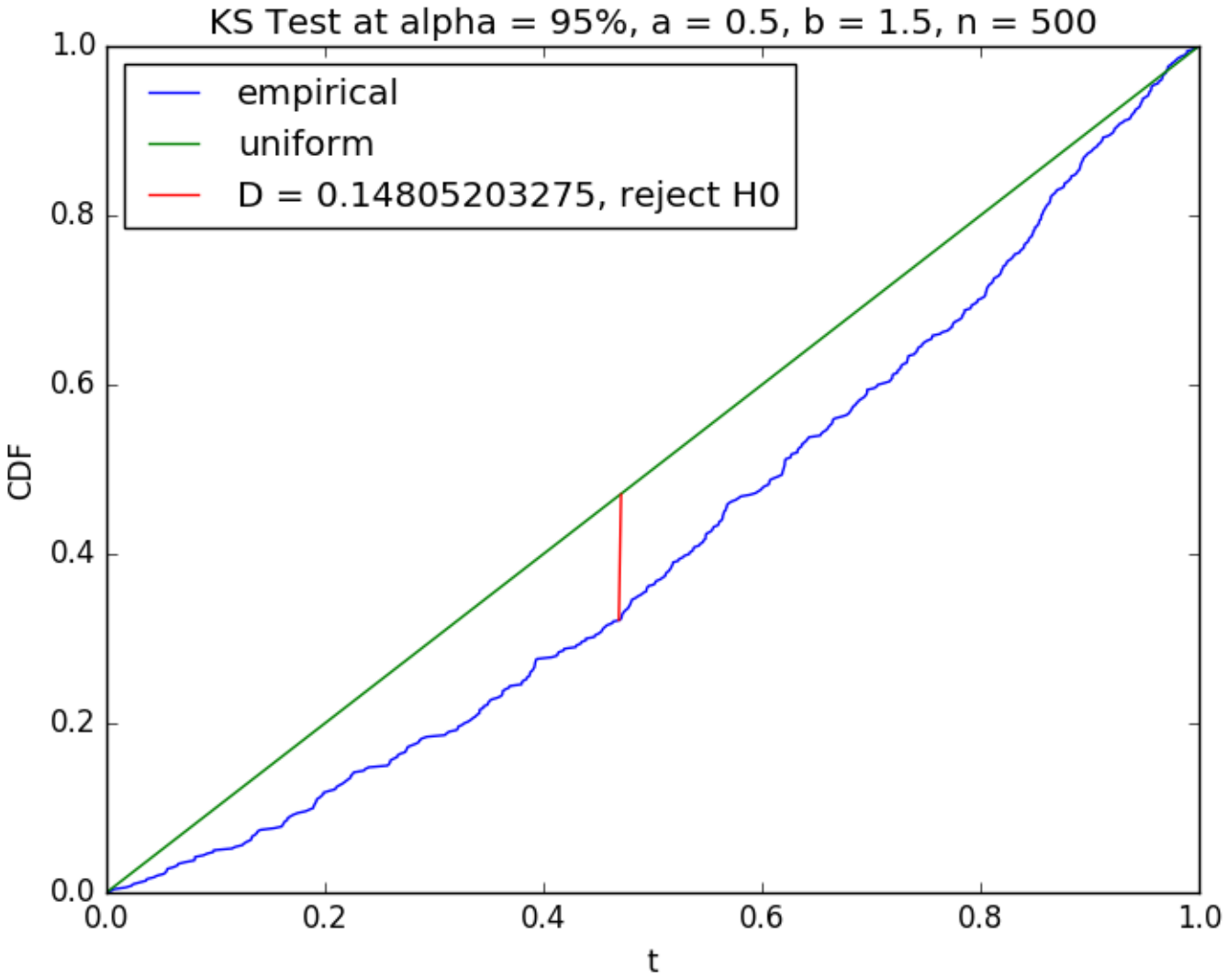
$$f(t) = \frac{a + (b-a)\cdot t}{N(a,b)}$$

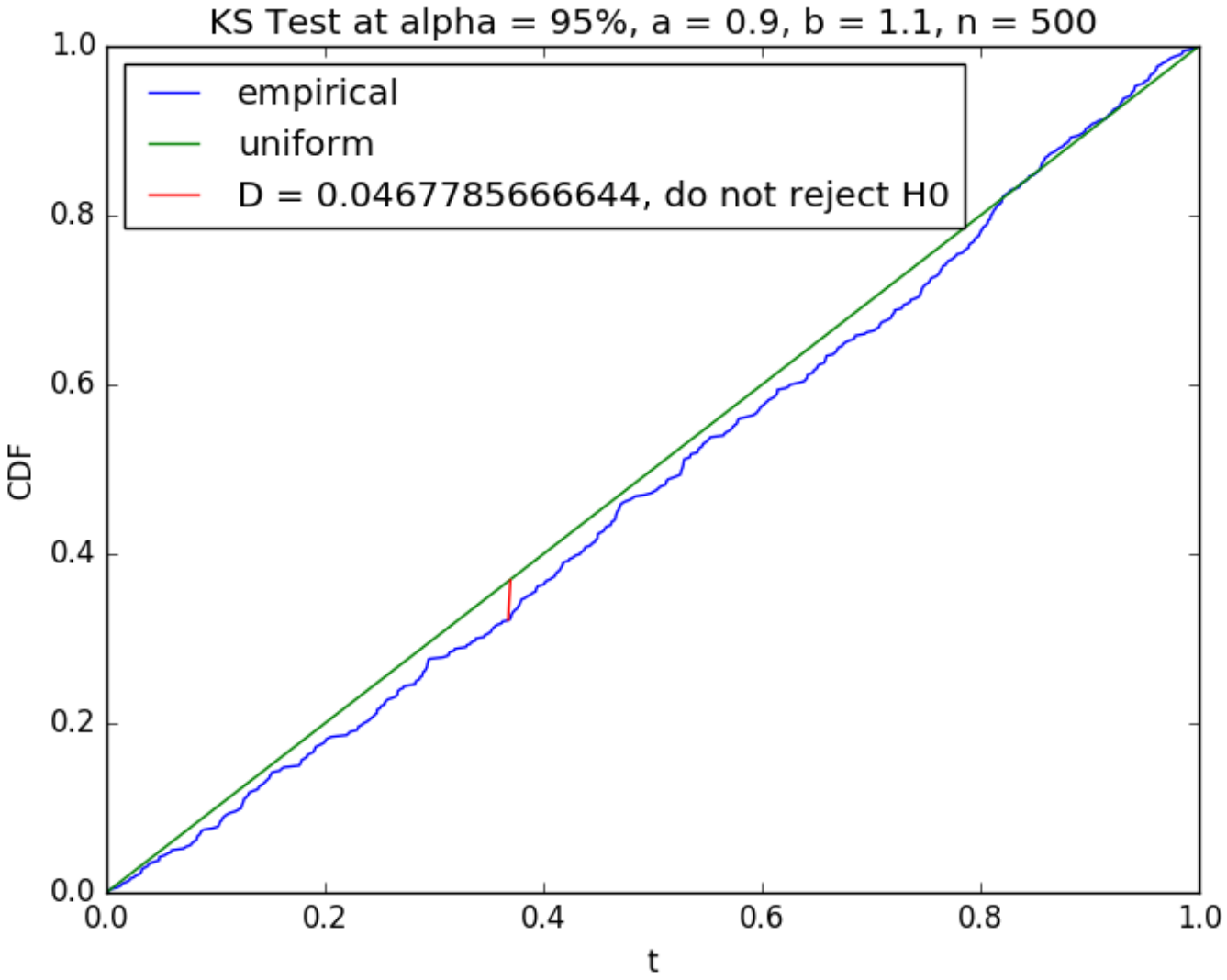
$N(a,b)$ = area under curve $= \min\{a,b\} + |b-a|/2.0$

Notice that $b = 2.0 - a$, and that $U(0,1)$ has $a = b = 1.0$


KS Test at alpha = 95%, a = 0.0, b = 2.0, n = 50
empirical
uniform
D = 0.269583766677, reject H0
CDF
t
0.0
0.2
0.4
0.6
0.8
1.0


KS Test at alpha = 95%, a = 0.5, b = 1.5, n = 50
empirical
uniform
D = 0.0888514615556, do not reject H0
CDF
t
0.0
0.2
0.4
0.6
0.8
1.0


KS Test at alpha = 95%, a = 0.5, b = 1.5, n = 500
empirical
uniform
D = 0.14805203275, reject H0
CDF
t
0.0
0.2
0.4
0.6
0.8
1.0

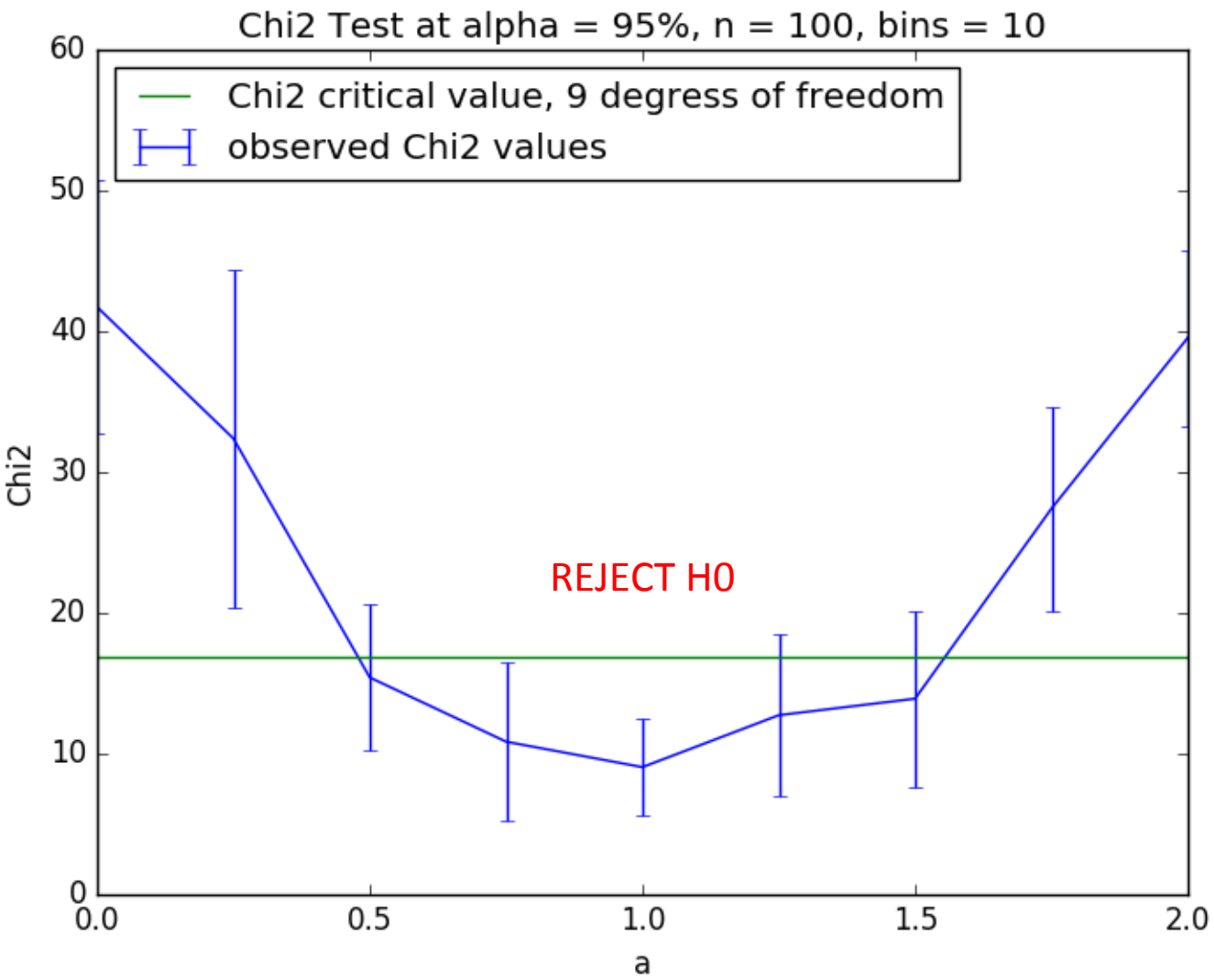

KS Test at alpha = 95%, a = 0.9, b = 1.1, n = 500
empirical
uniform
D = 0.0467785666644, do not reject H0
CDF
t
0.0
0.2
0.4
0.6
0.8
1.0

Chi-square test

- create n equi-length partitions

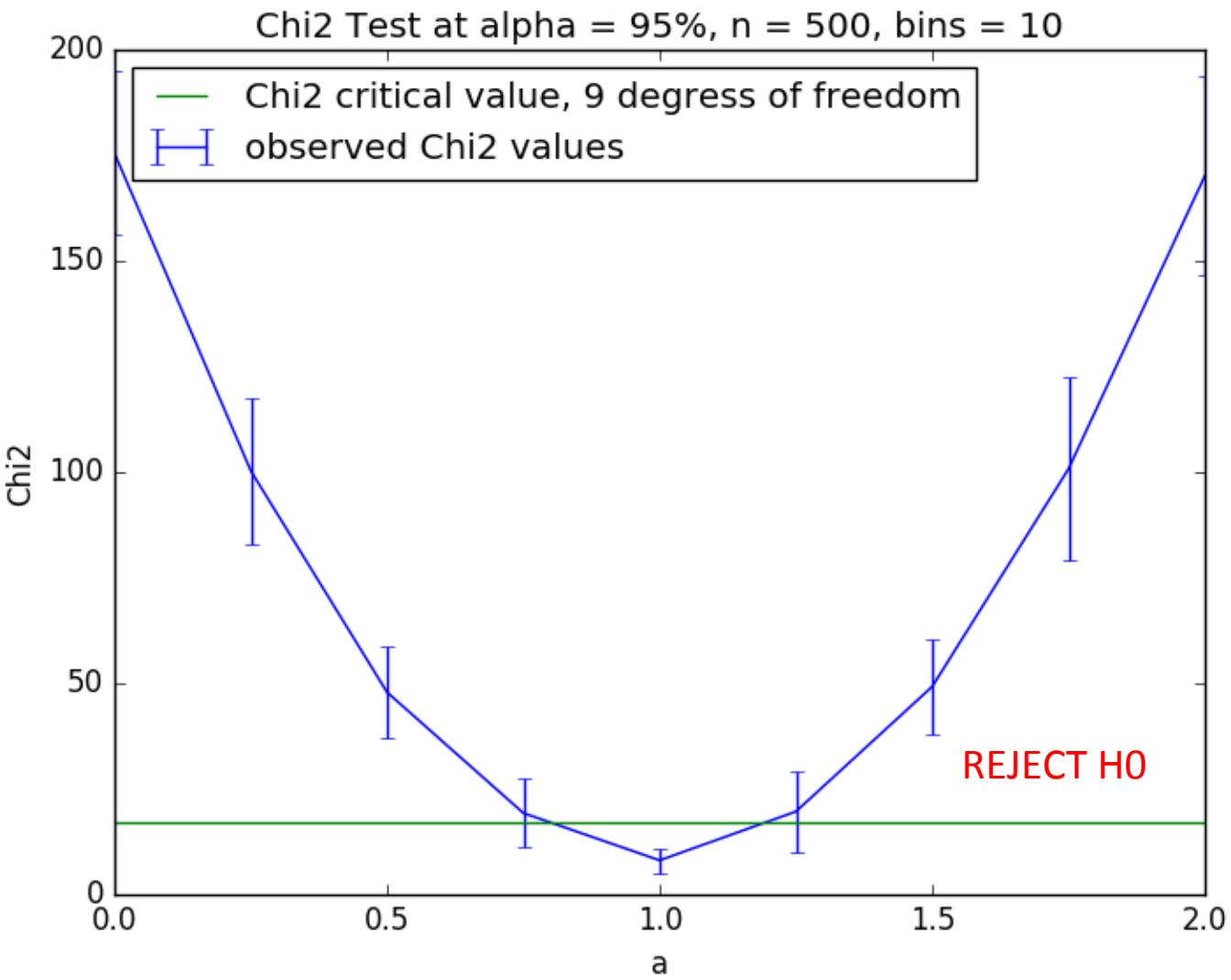
$$[0, \tfrac{1}{n}), [\tfrac{1}{n}, \tfrac{2}{n}) \cdots [\tfrac{n-1}{n}, 1)$$

- Bin each $S_i \in \{S_1, S_2, \ldots, S_N\}$
- Under the assumption of uniformity, bins will get approximately the same number of samples, N/n
- Let O_i be the OBSERVED number of samples in $[\tfrac{i-1}{n}, \tfrac{i}{n})$
- compute $$\chi_0^2 = \sum_{i=1}^{n} \frac{(O_i - N/n)^2}{N/n} \qquad N \geq 50$$
- said to have $n-1$ degrees of freedom
- compare against critical value of χ_0^2 distribution "TOO LARGE" taken to mean uniformity is unlikely



a is parameter from distribution studied using KS test

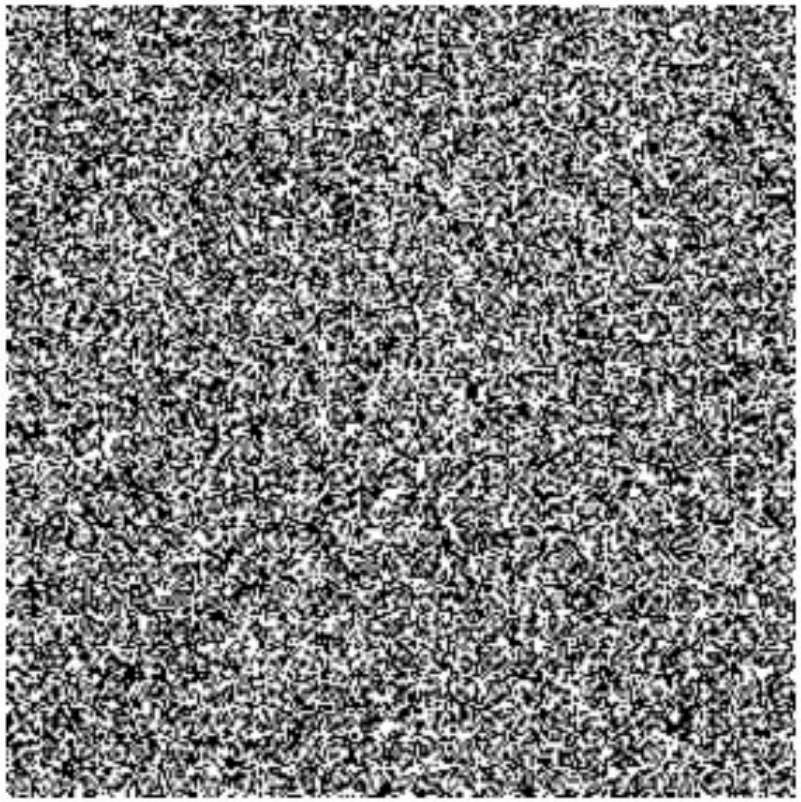
a = 1 is perfectly uniform



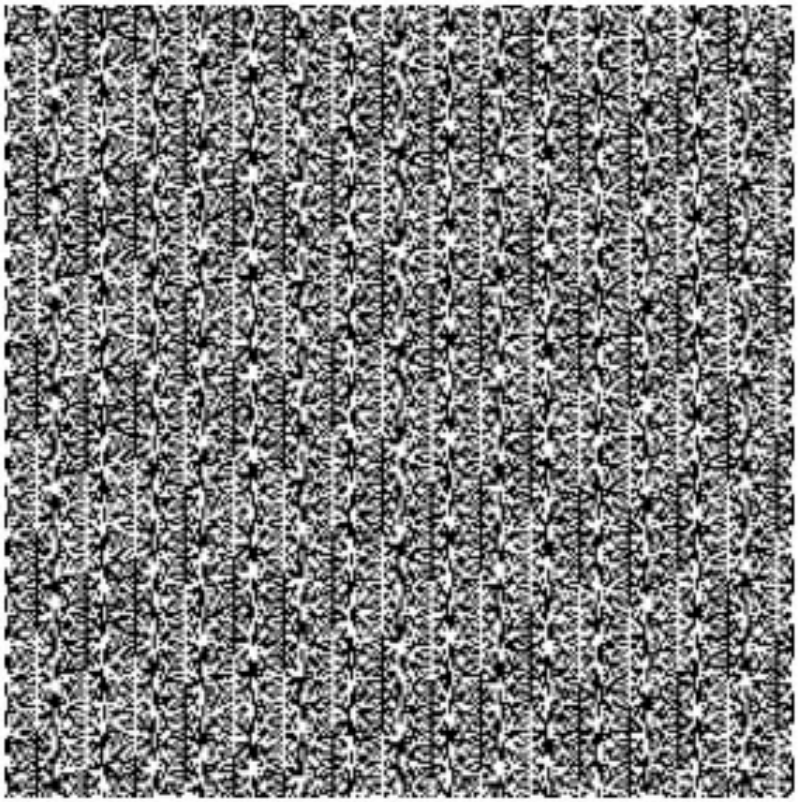
a is parameter from distribution studied using KS test

a = 1 is perfectly uniform

How can you tell whether your RNG produces independent numbers?



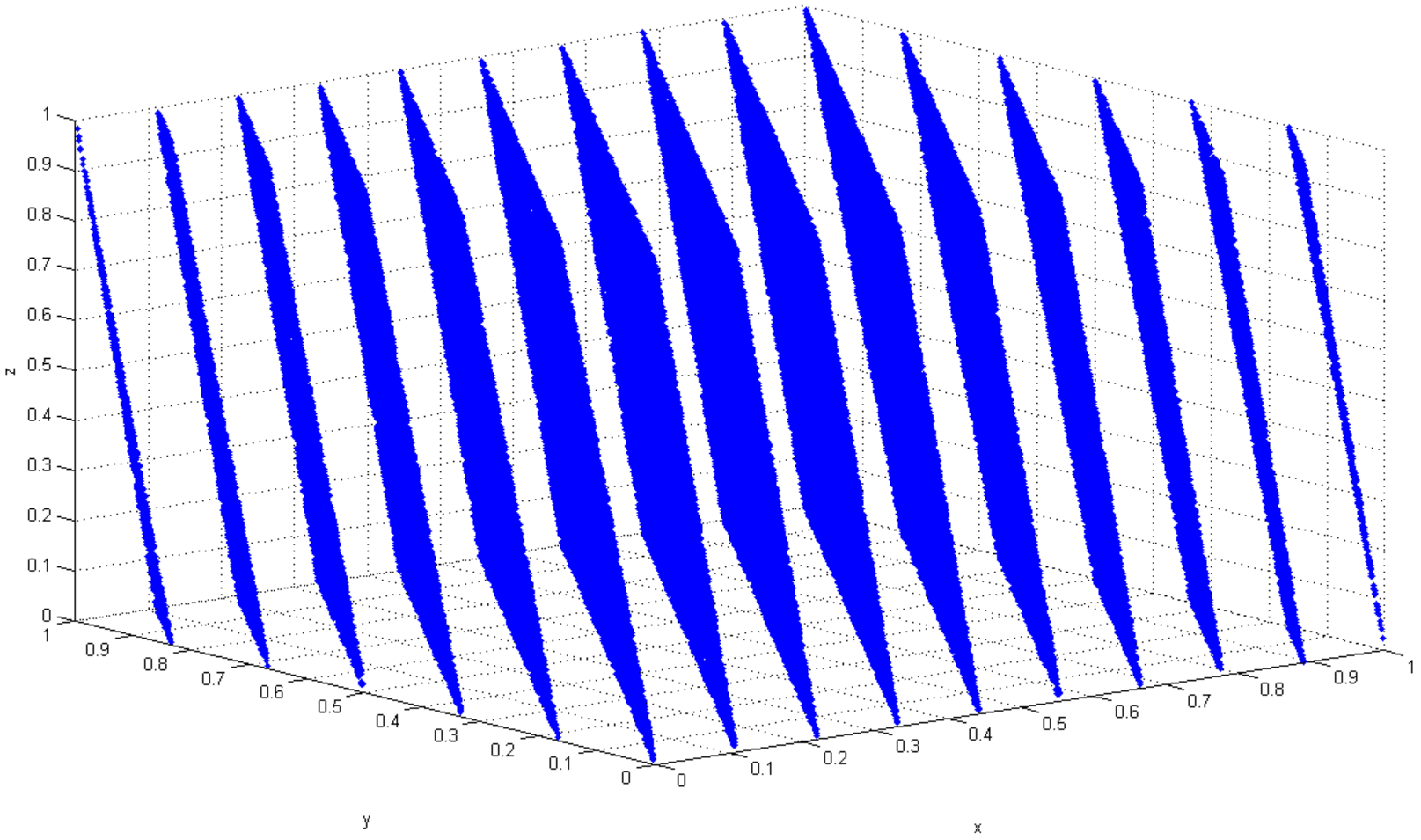
RANDOM.ORG



PHP rand() on Microsoft Windows

Test: plot (u_n, u_{n+1})

Patterns suggest lack of independence



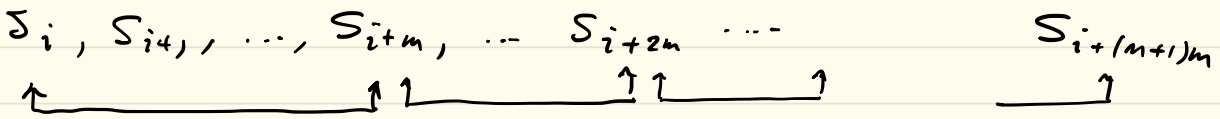
Plot of (u_{n-1}, u_n , u_{n+1}) using (old) rand() generation from early Unix

TEST FOR INDEPENDENCE

Given sequence of generated random numbers

$$S_1, S_2, \ldots S_k, \ldots S_N \qquad \text{with } \bar{S} = \frac{1}{N}\sum_{i=1}^{N} S_i$$

we can test for autocorrelation between every m numbers



k-lag autocorrelation

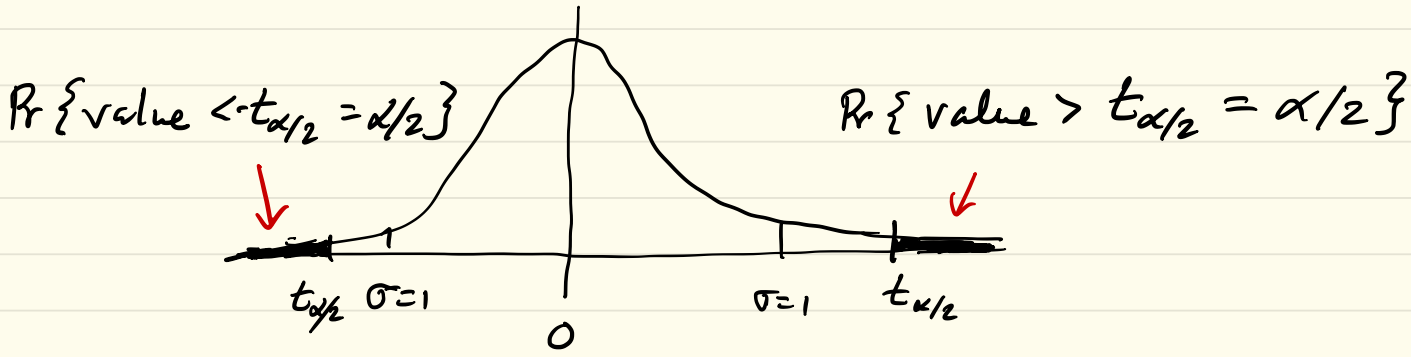
$$\tilde{\gamma}_k = \frac{\sum_{i=1}^{N-k} (S_i - \bar{S})(S_{i+k} - \bar{S})}{\sum_{i=1}^{N} (S_i - \bar{S})}$$

When $\{S_i, S_{i+m}, S_{i+2m}, \ldots\}$ are independent and set size is large, then

$\tilde{\gamma}_k$ is approximately normal, mean = 0, $\sigma = 1$

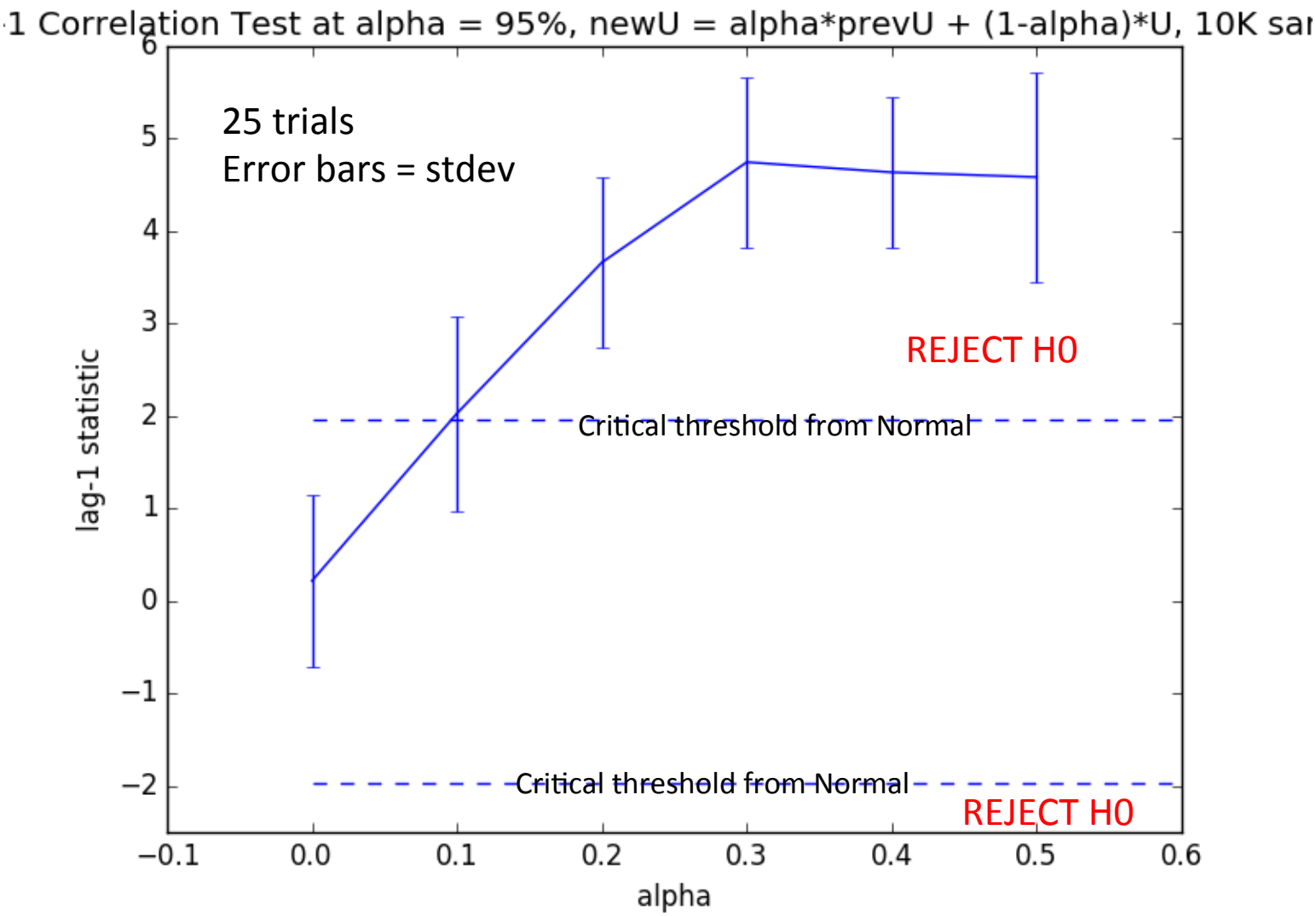
TESTING FOR RANDOMNESS, TYPICALLY ONLY LAG-1 TEST

Subject $\tilde{\gamma}_1$ to a "two tail" test



a positive value above $t_{\alpha/2}$ or a negative value below $-t_{\alpha/2}$ is evidence of non-independence

look up $t_{\alpha/2}$ values from standard normal table.



$$U_n = \alpha * U_{n-1} + (1-\alpha) * U$$