22.4UC2

Numerical Computation and Statistics in Engineering

Unit 7: Random Number Generation

Random Number Generation

In this unit we talk about random numbers. In particular, we will address the following questions:

- What are random numbers?
- How do we generate random numbers (for example for simulations)?
- How do we test numbers for being random?

Examples

The following sequences of numbers claim to be sequences of random numbers between 0 and 9:

- 0;
- 01314 51333 6;
- 85512 20802 54616 43738 80217
  68589 41082 01665 73927 40023
  55389 71274 82963 66439 15864
  36514 84192 82521 36924 80428
- 10308 06601 92086 26330 24354
  72132 74113 21424 45001 00143
  25014 16401 70207 22423 58514
  11622 00824 52112 34591 11327

Examples

3141592653	5897932384	6264338327	9502884197	1693993751	0582097494	4592307816	4062862089	9862803482	5342117067
9821480865	1328230664	7093844609	5505822317	2535940812	8481117450	2841027019	3852110555	9644622948	9549303819
6442881097	5665933446	1284756482	3378678316	5271201909	1456485669	2346034861	0454326648	2133936072	6024914127
3724587006	6063155881	7488152092	0962829254	0917153643	6789259036	0011330530	5488204665	2138414695	1941511609
4330572703	6575959195	3092186117	3819326117	9310511854	8074462379	9627495673	5188575272	4891227938	1830119491
2983367336	2440656643	0860213949	4639522473	7190702179	8609437027	7053921717	6293176752	3846748184	6766940513
2000568127	1452635608	2778577134	2757789609	1736371787	2146844090	1224953430	1465495853	7105079227	9689258923
5420199561	1212902196	0864034418	1598136297	7477130996	0518707211	3499999983	7297804995	1059731732	8160963185
9502445945	5346908302	6425223082	5334468503	5261931188	1710100031	3783875288	6587533208	3814206171	7766914730
3598253490	4287554687	3115956286	3882353787	5937519577	8185778053	2171226806	6130019278	7661119590	9216420198
9380952572	0106548586	3278865936	1533818279	6823030195	2035301852	9689957736	2259941389	1249721775	2834791315
1557485724	2454150695	9508295331	1686172785	5889075098	3817546374	6493931925	5060400927	7016711390	0984882401
2858361603	5637076601	0471018194	2955596198	9467678374	4944825537	9774726847	1040475346	4620804668	4259069491
2933136770	2898915210	4752162056	9660240580	3815019351	1253382430	0355876402	4749647326	3914199272	6042699227
9678235478	1636009341	7216412199	2458631503	0286182974	5557067498	3850549458	8586926995	6909272107	9750930295
5321165344	9872027559	6023648066	5499119881	8347977535	6636980742	6542527862	5518184175	7467289097	7772793800
0816470600	1614524919	2173217214	7723501414	4197356854	8161361157	3525521334	7574184946	8438523323	9073941433
3454776241	6862518983	5694855620	9921922218	4272550254	2568876717	9049460165	3466804988	6272327917	8608578438
3827967976	6814541009	5388378636	0950680064	2251252051	1739298489	6084128488	6269456042	4196528502	2210661186
3067442786	2203919494	5047123713	7869609563	6437191728	7467764657	5739624138	9086583264	5995813390	4780275900
9946576407	8951269468	3983525957	0982582262	0522489407	7267194782	6848260147	6990902640	1363944374	5530506820
3496252451	7493996514	3142980919	0659250937	2216964615	1570985838	7410597885	9597729754	9893016175	3928468138
2686838689	4277415599	1855925245	9539594310	4997252468	0845987273	6446958486	5383673622	2626099124	6080512438
8439045124	4136549762	7807977156	9143599770	0129616089	4416948685	5584840635	3422072225	8284886481	5845602850
6016842739	4522674676	7889525213	8522549954	6667278239	8645659611	6354886230	5774564980	3559363456	8174324112

Examples

The data were generated as follows:

- 0;
- my office phone number;
- random data from a normal distribution;
- random data from an exponential distribution;
- the first 2500 digits of π.

Random means ...

typical, representative, arbitrary, accidental, scrambled, aimless, patternless, senseless, causeless, independent, incidental, indeterminate, indescribable, undirected, uncontrolled, uncertain, unrelated, unordered, unpredictable.

Something is random if it is not possible to describe it more briefly than to reproduce it.

Examples

As a working definition the following is also good:

A (long) sequence of numbers is random if the shortest computer program printing out the sequence has about the size (length) of the sequence.

A formal definition for *infinite* sequences of the following:

> A sequence is random if, for every block length n, the probability that a particular block of length n is observed is $1/n$.

For example for a sequence of numbers between 1 and 6 to be random it is not enough that each number occurs with probability $1/6$. We also require that each tuple of numbers

$$11, 12, 13, 14, 15, 16, 21, 22, 23, 24, 25, 26, 31, \dots, 66$$

occurs with probability $1/36$, and that each triple of numbers like $111, 123, 461, \dots$ occurs with probability $1/216$, etc.

Generating Pseudo-Random Numbers

We first consider random numbers for the uniform distribution on the interval $[0,1]$. Recall that

$$\begin{aligned} \mu &= \int_0^1 x\,dx = \tfrac{1}{2}x^2\Big|_0^1 = \tfrac{1}{2} \text{ and} \\ \sigma &= \int_0^1 x^2\,dx - \mu^2 = \tfrac{1}{3}x^3\Big|_0^1 - \tfrac{1}{4} = \tfrac{1}{12}\,. \end{aligned}$$

In practice it is hard to generate true random numbers, in particular on a computer. Such numbers will always have patterns (in particular repetitions), and the aim is to reduce such patterns.

Generating Pseudo-Random Numbers

Among others the following problems occur:

- The numbers may not be uniformly distributed.
- The values are usually discrete, and not continuous.
- The (sample) mean may differ from $\frac{1}{2}$.
- There may be cyclic variations, including
  - autocorrelation between numbers;
  - numbers successively lower or higher than adjacent numbers;
  - a couple of numbers below the mean followed by a couple of numbers above the mean.

We will have to apply tests to sequences of such numbers.

Generating Pseudo-Random Numbers

There are also some more practical issues, even if we accept that we can only generate pseudo-random numbers:

- The method (program) generating pseudo-random numbers should be fast.
- The pseudo-random numbers should have a long *cycle*. If 100,000 numbers are needed for a simulation the cycle length should be at least that big.
- The random numbers should be replicable (yes, indeed (!)). This feature is needed for debugging and comparison.
- The generated numbers should approximate the uniform distribution (or any other distribution for which we generate random numbers.)

The Linear Congruential Method

This method, suggested in 1951 by Lehmer, produces integers between 0 and $m-1$ according to the rule

$$x_{n+1} = (ax_n + c) \mod m\,.$$

These numbers are then used to represent the numbers

$$y_n = x_n/m$$

in the unit interval $[0,1]$. The value x_0 is called the *seed*, a the *constant multiplier*, and c the *increment*. The choice of values x_0, a, c, m has big impact on the quality of the random numbers. For m common values are $2^{31}-1$ and 2^{48}.

For m a power of 2 the period (cycle length) m can be achieved with $c \neq 0$ relatively prime to m (i.e., $\gcd(c,m)=1$) and $a \equiv 1 \mod 4$.

If $c=0$ then the largest possible period is $m/4$, which can be achieved with $a \equiv 3 \mod 8$ or $a \equiv 5 \mod 8$, and x_0 odd.

The Linear Congruential Method

Example. We consider the values $a = 13$, $c = 0$, $m = 2^6 = 64$, and $x_0 = 1, 2, 3, 4$.

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$x_0 = 1$	1	13	41	21	17	29	57	37	33	45	9	53	49	61	25	5	1
$x_0 = 2$	2	26	18	42	34	58	50	10	2								
$x_0 = 3$	3	39	59	63	51	23	43	47	35	7	27	31	19	55	11	15	3
$x_0 = 4$	4	52	36	20	4												

$\square$

These numbers show how vital a good choice of the parameters is. Also note the large gaps between the pseudo-random numbers. Even with the period length of 16 the gaps between the random numbers are quite large.

Tests for Random Numbers

There are a couple of tests we can perform on sequences of pseudo-random numbers to ensure that the sequences have the desired properties. Among them are the following:

- *Simple Frequency test.* This test uses the chi-square test to compare the (uniform) distribution against the set of numbers in the sequence.
- *Autocorrelation test.* This tests for correlation between the numbers and compares it to the expected correlation 0.
- *Runs test.* These simple tests test for ups and downs, or for values above and below the mean. These tests have a chi-square distribution.
- *Gap test.* This test counts the number of digits that appear between repetitions of a particular digit, and then uses the Kolmogorov-Smirnov test to compare with the expected size of gaps.
- *Poker test.* Treats numbers grouped together as a poker hand. These hands are compared to what is expected using the chi-square test.

Tests for Random Numbers

We will treat some of these test with the following set of data, created using the C++ random number generator:

Our level of significance will always be $\alpha = 0.05$.

Tests for Random Numbers

Frequency Test

The first test is a frequency test. We sub-divide the interval $[0,1]$ into $n=10$ intervals of equal length $[0,0.1)$, ... , $[0.9,1]$ and count the number O_i of observed numbers in each interval i. The statistics

$$\chi^2 = \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i}$$

has a chi-square distribution with $10-1$ degrees of freedom. Here E_i is the expected number in each interval, which is $E_i = \frac{50}{10} = 5$.

We test the null hypothesis that the data are from a uniform distribution. The hypothesis is accepted if $\chi^2 < \chi^2_{0.05,9} = 16.919$.

Tests for Random Numbers

We find the following data

	1	2	3	4	5	6	7	8	9	10
O_i	1	4	9	8	2	3	8	5	5	5
$\frac{(O_i - E_i)^2}{E_i}$	3.2	0.2	3.2	1.8	1.8	0.8	1.8	0	0	0

and $\chi^2 = 12.8$. Thus, the hypothesis is accepted.

This frequency test is usually applied if there is enough data so that each $E_i \geq 5$.

Tests for Random Numbers

Runs Test

The first runs test is about ups and downs, i.e., whether the next number in the list of random numbers is above or below the current number. The following sequence is obtained from our data set.

$$
\begin{array}{ccccccccccccccc}
+ & - & - & + & - & + & + & - & - & + & - & + & - & + & - \\
+ & - & + & - & - & + & - & - & + & - & + & - & + & - & - \\
- & + & + & + & - & + & - & - & + & - & + & - & - & + & - \\
+ & - & - & + & & & & & & & & & & &
\end{array}
$$

A consecutive sequence of $+$'s or $-$'s is called a run. If the sequence is truly random the mean and variance of the number of runs are given by

$$\mu = \frac{2N-1}{3} \qquad \text{and} \qquad \sigma = \sqrt{\frac{16N-29}{90}},$$

Tests for Random Numbers

where N is the total number of random numbers. If N is sufficiently large this distribution of runs is approximately normal, and we can use standard hypothesis testing.

We test the null hypothesis that the data come from a uniform normal distribution at significance level $\alpha = 0.05$.

We observed 37 runs, and calculate

$$z = \frac{37 - \mu}{\sigma} = \frac{37 - 33}{\sqrt{8.566}} = 1.3666\,.$$

For the two-tailed test the hypothesis is rejected if $z < -z_{0.05/2}$ of if $z > z_{0.05/2}$. The value for $z_{0.05/2}$ is 1.960, and the null hypothesis that the data is from a uniform distribution is accepted.

Tests for Random Numbers

A different runs test tests for values above and below the mean. For this test we get the following data

$$
\begin{array}{ccccccccccccccc}
+ & + & + & - & + & - & + & + & + & - & + & - & + & + & + \\
- & - & - & + & - & - & + & + & - & + & - & + & - & + & - \\
- & - & - & + & + & + & + & + & - & - & - & + & - & - & + \\
- & + & - & - & + & & & & & & & & & &
\end{array}
$$

If n_1 and n_2 are the numbers of observations below and above the mean (here 24 and 26, and b the total number of runs (here 29). Then b is approximately normally distributed with mean and standard deviation

$$\mu = \frac{2n_1 n_2}{N} + \frac{1}{2} \qquad \text{and} \qquad \sigma = \sqrt{\frac{2n_1 n_2(2n_1 n_2 - N)}{N^2(N-1)}},$$

Tests for Random Numbers

where N is again the number of random numbers. The distribution

$$z = \frac{b - \mu}{\sigma}$$

is approximately a standard normal distribution, and we find

$$z = \frac{29 - 25.45}{\sqrt{12.2049}} = 1.0161\,.$$

Since $-z_{0.01/2} \leq 1.0161 \leq z_{0.01/2}$ the null hypothesis is accepted.

The Poker Test

We test for patterns in individual numbers. We generated random numbers of two digits each. If the numbers are truly random then $\frac{10}{100}100\%$ will have twice the same digit, while $\frac{10 \cdot 9}{100}100\%$ will have different digits. We use again the chi-square test with random variable

$$\chi^2 = \sum_{i=1,2} \frac{(O_i - E_i)^2}{E_i}$$

where O_i are the observed frequencies, and E_i the expected frequencies. With our data of 50 random numbers we find

	O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
Two different digits	47	45	0.0889
Two like digits	3	5	0.8

Since $\chi^2 = 0.8889 < \chi_{0.05,2} = 5.991$ the null hypothesis is accepted.

Random Variate Generation

Next we want to generate random numbers for more general probability distributions. The following result gives in theory a method to generate random numbers for *any* probability distribution:

> Let x be a (continuous) random variable with cumulative distribution $F(x)$. Then the random variable $y = F(x)$ is a uniform random variable over the interval $0 \le y \le 1$.

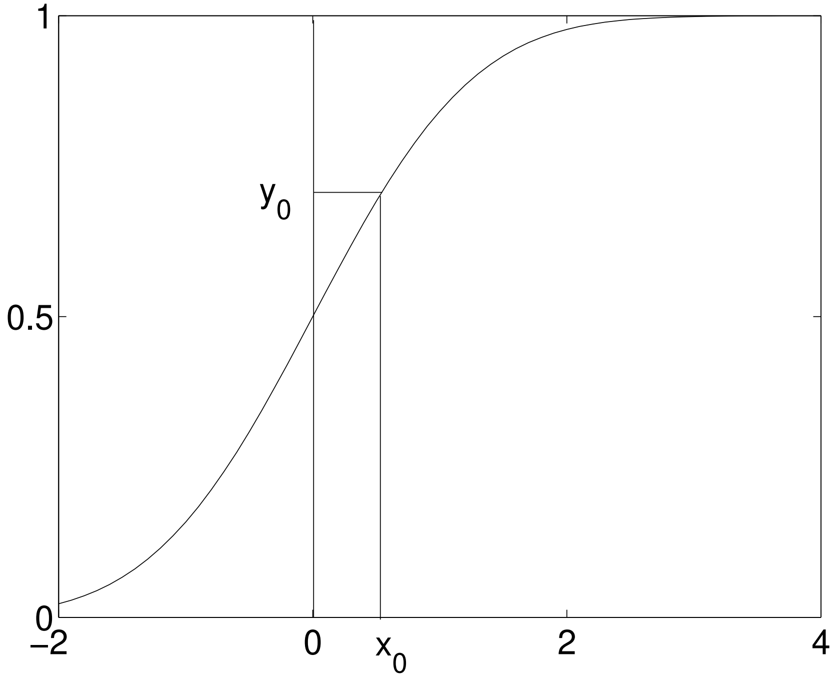
Indeed, y takes values between 0 and 1. Moreover, for every fixed y_0,

$$P(y \le y_0) = P(x \le x_0) = F(x_0) = y_0 \,,$$

(as shown on the next graph) so that the cumulative distribution for y is

$$\begin{cases} y & \text{if } 0 \le y \le 1, \\ 0 & \text{else.} \end{cases}$$

Random Variate Generation



Taking the derivative with respect to y shows that the density function is indeed

$$\begin{cases} 1 & \text{for } 0 \leq y \leq 1 \\ 0 & \text{else.} \end{cases}$$

Random Variate Generation

This result yields the following algorithm for generating random numbers for a probability distribution with density function $\rho(x)$ and cumulative distribution $F(x) = \int_{-\infty}^{x} \rho(x)\, dt$:

- Draw uniformly random numbers r_i between 0 and 1.
- Substitute into $r_i = F(x_i)$ and solve for x_i.

In practice the last step is often difficult or even impossible and we have to look for approximation of the function

$$x_i = F^{-1}(r_i)\,.$$

Examples

Example. We look at the exponential distribution with density function $\rho(x)$

$$\begin{cases} \frac{e^{-x/2}}{2} & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}$$

and cumulative distribution (for $x \geq 0$)

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} \rho(t)\, dt = \int_0^x \frac{e^{-x/2}}{2}\, dt = -e^{-t/2}\Big|_0^t \\ &= 1 - e^{-x/2}\,. \end{aligned}$$

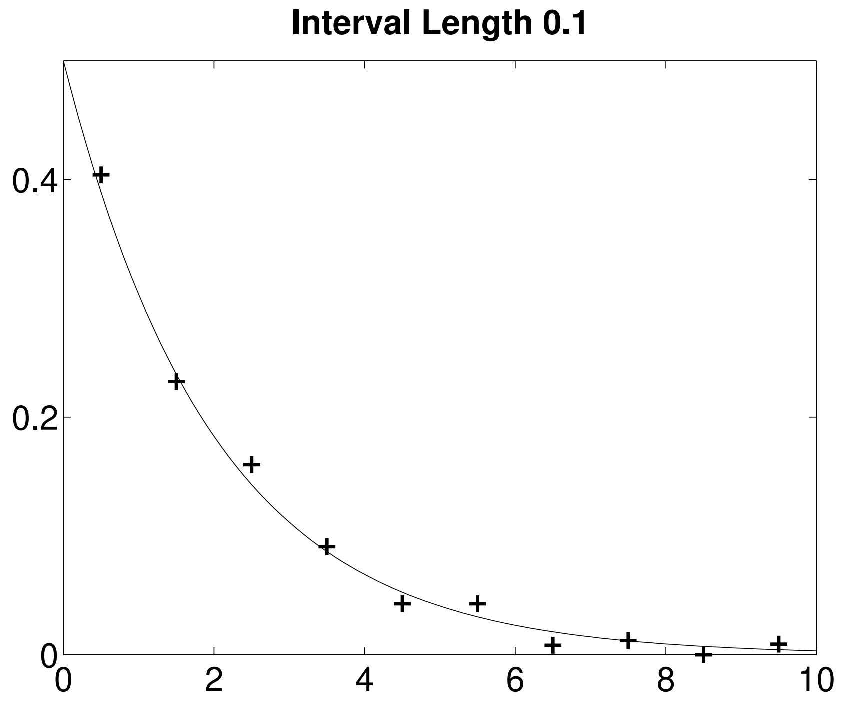
If we let $r = 1 - e^{-x/2}$ and solve for x then

$$x = -2\ln(1-r)\,.$$

Examples

Thus, drawing uniformly random numbers r_i between 0 and 1 gives random numbers x_i for the exponential probability distribution. Note that in practice $1 - r$ is replaced by r.

The graph shows a frequency plot of 1000 random numbers r_i drawn using the C++ random number generator. Here we used the inverse transform to get random numbers for the exponential distribution.
The interval length is 1.



$\square$

Examples

Example. We look at the uniform distribution on the interval $[a, b]$. We generated uniformly random numbers r on the unit interval $[0, 1]$. To get random numbers for the uniform distribution on $[a, b]$ recall that the density function for this distribution is

$$\begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{else} \end{cases}$$

so that the cumulative distribution is 0 if $x < a$, 1 if $x > b$, and

$$\begin{aligned} F(x) &= P(t \le x) = \int_{-\infty}^{x} \rho(t)\, dt = \int_{a}^{x} \frac{1}{b-a}\, dt \\ &= \frac{x-a}{b-a}\,. \end{aligned}$$

So we solve $r = \frac{x-a}{b-a}$ for x and get $x = a + (b-a)r$. $\square$

Normal Distribution

The above mentioned technique works well if the cumulative distribution has a closed form and if we can calculate the inverse of this function, i.e., solve for x. However, for many important distributions, notably the normal distribution, this is *not* possible. In practice we have to *approximate* the function

$$x = F^{-1}(r) \,.$$

A simple and reasonably good approximation for the inverse of the cumulative distribution of the standard normal distribution is the function

$$x = F^{-1}(r) = \frac{r^{0.135} - (1-r)^{0.135}}{0.1975} \,.$$

The approximation gives at least one correct decimal of the inverse for r taking values between 0.00135 and 0.9986.

Normal Distribution

Another (much better) approximation is

$$x = s + \frac{\sum_{i=0}^{4} p_i s^i}{\sum_{i=0}^{4} q_i s^i}$$

where $s = \sqrt{-2\ln(1-r)}$ and p_i, q_i are the following values:

$$\begin{array}{lcllcl}
p_0 & = & -0.322232431088 & q_0 & = & 0.099348462606 \\
p_1 & = & -1 & q_1 & = & 0.588581570495 \\
p_2 & = & -0.342242088547 & q_2 & = & 0.531103462366 \\
p_3 & = & -0.0204231210245 & q_3 & = & 0.10353775285 \\
p_4 & = & -0.0000453642210148 & q_4 & = & 0.0038560700634
\end{array}$$

This approximation is accurate up to about 6 digits for $0.5 < r < 1$.
Symmetry of the normal distribution allows to extend this to the interval $0 < r < 0.5$ using the transformation $r \mapsto 1 - r$ and $x \mapsto -x$.

Empirical Distributions

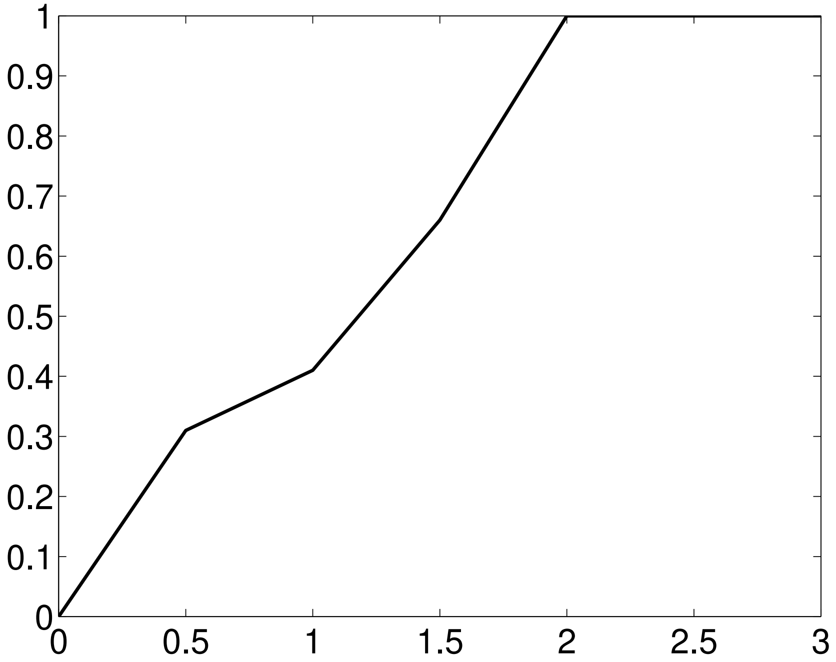
We treat empirical distributions by way of example.

Suppose that data is collected about repair times (in hours):

	Frequency	Rel. Freq.	Cumm. Freq.
$0 \leq x \leq 0.5$	31	0.31	0.31
$0.5 < x \leq 1.0$	10	0.1	0.41
$1.0 < x \leq 1.5$	25	0.25	0.66
$1.5 < x \leq 2$	34	0.34	1

The true cumulative probability function is unknown to us.

Empirical Distributions



The inverse transform technique applies here directly.

If r_1, the random number from the uniform distribution, is 0.83, then we can read off the x_1 value from the graph as 1.75. If we need more random numbers we will use a more systematic approach.

Empirical Distributions

If $0 \le x \le 0.5$ the cumulative probability function is

$$y = \frac{0.31 - 0}{0.5 - 0} x = 0.62x \qquad \text{or} \qquad x = 1.61y\,.$$

Here y takes values between 0 $(x = 0)$ and 0.31 $(x = 0.5)$.

If $0.5 < x \le 1$ the cumulative function has slope

$$\frac{0.41 - 0.31}{1.0 - 0.5} = 0.2$$

and the intercept is found by solving the equation

$$0.41 = 0.2 \cdot 1.0 + i$$

using the fact that the point $(1.0, 0.41)$ is on the line. Thus, the intercept is 0.21 and the cumulative function has the equation

$$y = 0.2x + 0.21$$

Empirical Distributions

and y takes values between 0.31 and 0.41. The inverse is the function

$$\begin{aligned} x &= \frac{1}{0.2}(y - 0.21) \\ &= 5y - 1.05 \end{aligned}$$

In general, if we have a line through (x_0, y_0) and (x_1, y_1) then the slope of the line is

$$\frac{y_1 - y_0}{x_1 - x_0}$$

and the intercept is found by solving for example

$$y_0 = \frac{y_1 - y_0}{x_1 - x_0} x_0 + i$$

so that

$$i = \frac{x_1 y_0 - x_0 y_1}{x_1 - x_0} .$$

Empirical Distributions

This way we get the following table describing the inverse of the cumulative distribution:

$$
\begin{array}{lrcl}
0.00 \le y \le 0.31 & x & = & 1.61y \\
0.31 < y \le 0.41 & x & = & 5y - 1.05 \\
0.41 < y \le 0.66 & x & = & 2y + 0.18 \\
0.66 < y \le 1.00 & x & = & 1.47y + 0.53
\end{array}
$$

Empirical Distributions