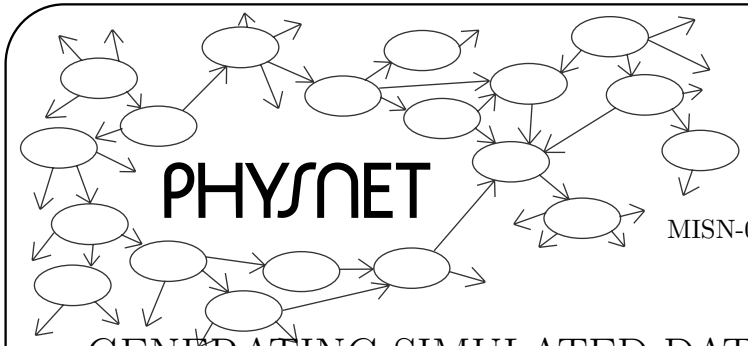


MISN-0-214

GENERATING SIMULATED DATA

0.2344	0.5597	0.2550	0.6221	0.1616
0.4280	0.8566	0.0764	0.5530	0.1217
0.9076	0.3092	0.8079	0.5903	0.1015
0.5172	0.0372	0.2945	0.6479	0.0307
0.6939	0.2494	0.1568	0.7431	0.0064
0.0979	0.7194	0.0607	0.0856	0.4124
0.3172	0.5215	0.8528	0.7764	0.7920
0.4308	0.4852	0.0076	0.0600	0.9685
0.8543	0.3570	0.1754	0.8245	0.7780
0.4719	0.9782	0.3910	0.7791	0.5908
0.9465	0.4233	0.2928	0.1144	0.1378
0.4059	0.6884	0.3114	0.9346	0.8325
0.4391	0.7984	0.2657	0.8618	0.8878
0.3493	0.4471	0.3466	0.6252	0.3092
0.9689	0.3645	0.6332	0.4285	0.4922
0.2296	0.9430	0.7017	0.5200	0.4428
0.2417	0.2791	0.4092	0.4410	0.5791
0.0784	0.6185	0.2990	0.3515	0.5278
0.9578	0.6237	0.2737	0.1892	0.1815
0.9014	0.5185	0.7840	0.8045	0.6843
0.8890	0.2831	0.8429	0.2754	0.3512
0.2090	0.1528	0.4877	0.8276	0.3824
0.8235	0.3080	0.4168	0.3307	0.7098
0.3303	0.7961	0.8809	0.1133	0.5768
0.6046	0.9054	0.7467	0.5013	0.1385
0.5996	0.9866	0.1023	0.8873	0.8824

Project PHYSNET Physics Bldg. Michigan State University East Lansing, MI

GENERATING SIMULATED DATA

by

Mike Brandl and Ray G. Van Ausdal

Title: **Generating Simulated Data**

Author: Mike Brandl, Michigan State University and Ray G. Van Ausdal, University of Pittsburgh at Johnstown.

Version: 5/9/2000 Evaluation: Stage 0

Length: 2 hr; 20 pages

Input Skills:

1. Vocabulary: Errors in measurement (MISN-0-371)). Random and pseudorandom numbers (MISN-0-354)).
2. Skills: None. It is useful to be able to generate a set of pseudorandom numbers between zero and one (MISN-0-354).

Output Skills (Knowledge):

K1. Vocabulary: uniform and non-uniform distributions; probability density function; probability in an interval; normalization; cumulative probability function; Lorentzian and Gaussian distributions.

K2. State the four steps to generate simulated data.

K3. Demonstrate graphically and mathematically how this method works.

Output Skills (Rule Application):

R1. Given a simple probability density function, form its cumulative probability function, and generate data that simulates the original distribution.

Output Skills (Problem Solving):

P1. Given a graph of the cumulative probability function vs. x, describe analytically, graphically, and verbally the distribution of data that would be generated.

External Resources (Optional):

1. Access to a computer is useful.

GENERATING SIMULATED DATA

by

Mike Brandl and Ray G. Van Ausdal

1. Overview

This module presents a method for generating random data which simulates the data that might be received from an actual experiment. Such simulated data could be used to test theoretical models in situations where experimental tests would be too expensive, time-consuming or dangerous. (For example, the failure modes of a nuclear reactor must be studied without the benefit of experiment.) Artificially generated data could also be used to test methods of data reduction (such as least squares fitting of a function).[1]

2. Describing Random Events

2a. Random Numbers: Uniform Distribution. A sequence of numbers is "random" if the value of any number in that sequence cannot be predicted from the values of the numbers preceding it.[2] If all of the possible values have the same probability of occurring in the sequence, the numbers have a uniform distribution. The throw of a single die is an example: all faces are equally likely.

2b. Non-uniform Distributions. If all the possible values do not have the same probability of occurring, the numbers have a non-uniform distribution. The probability of a certain value depends on the value itself. The throw of a pair of dice is an example: "snake eyes" are not as likely as sevens. Many physical situations involve non-uniform distributions. For example, repeated measurements of a pendulum period show that values near the mean value are more likely to occur than values far from the mean.[3]

2c. Specifying Non-Uniform Probabilities. For discrete variables, such as the number of radioactive decays occurring in a given time interval, we could assign a probability to each possible value of the variable. However, for continuous variables, like times or distances, there are an

[1] See "Least Squares Fitting of Experimental Data" (MISN-0-162).
[2] See "Generation of Random Numbers by Computer" (MISN-0-354).
[3] See "Errors in Measurement" (MISN-0-371).

infinite number of possible values contained in any finite interval, so the probability of any one particular value occurring is infinitesimal.

2d. The Probability Density Function. We can characterize the probability of occurrence of values of a continuous variable x by the probability density function $p(x)$, where

$p(x_0)dx =$ the (infinitesimal) probability that the value of the variable x lies somewhere in the interval between x_0 and $x_0 + dx$.

The quantity $p(x)$ is the probability per unit interval in the variable x.[4]

2e. The Probability Contained in a Finite Interval. The actual probability $P(x)$ $(x_1 \leq x \leq x_2)$ that a continuous random variable x will assume a value on the finite interval between x_1 and x_2 can be obtained by integrating the probability density function between those two limits:

$$p(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} p(x')dx'. \tag{1}$$

Since the random variable x is certain to assume a value somewhere between its lowest and highest values x_ℓ and x_μ, the integral of $p(x)$ over the total range of the variable must be equal to one:

$$\int_{x_l}^{x_u} p(x')dx' = 1. \tag{2}$$

That is, $p(x)$ is "normalized" to one.

2f. The Cumulative Probability Function. The cumulative probability function $P_c(x)$ is defined as

$$P_c(x) \equiv \int_{x_l}^{x} p(x')dx'. \tag{3}$$

$P_c(x_0)$ is the probability that the value of x lies somewhere between x_ℓ and x_0. $P_c(x_0)$ increases monotonically from zero at x_ℓ to one at x_u (see Fig. 1).

[4] This is sometimes called the "distribution function." A "uniform distribution" is a special case of the non-uniform distribution with $p(x) =$ constant.

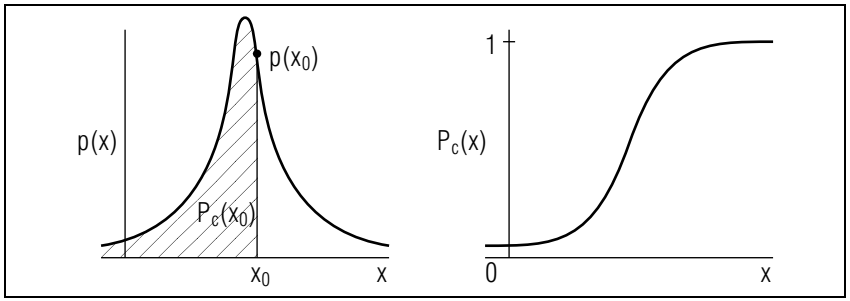


Figure 1. The cumulative probability function $P_c(x)$ is the integral of the probability density function $p(x)$.

3. Data Generation: the Method

3a. A Four Step Process. The basic method used to generate data whose distribution is described by a given probability density function is suprisingly simple. There are four steps:[5]

(i) Given the probability density function $p(x)$, form the cumulative probability function, $P_c(x)$.

(ii) Generate a pseudorandom number r_i which is a member of a sequence of numbers uniformly distributed on the interval zero to one.[6]

(iii) Find the value x_i such that $P_c(x_i) = r_i$ (see Fig. 2). Symbolically, this process is written $x_i = {P_c}^{-1}(r_i)$.

(iv) Repeat steps (ii) and (iii) to obtain the desired set of values x_i distributed according to the probability density function $p(x)$.

3b. The Basis: A Graphical Demonstration. Figure 3 demonstrates graphically how this method changes a uniform distribution of random numbers into a specific non-uniform distribution. The cumulative probability function $P_c(x)$ is plotted in Fig. 3a. A number of evenly spaced points are placed on the vertical axis, representing the uniformly distributed random numbers r_i. Inversions of each of these r_i are indicated graphically. The resulting points on the x-axis represent the numbers x_i

[5] In some cases, not all of these steps can be done simply. Approximations, numerical methods or graphical methods might be required.

[6] You may choose to use tables or calculator or computer methods (MISN-0-354). The cover of this module is a short table of pseudorandom numbers.

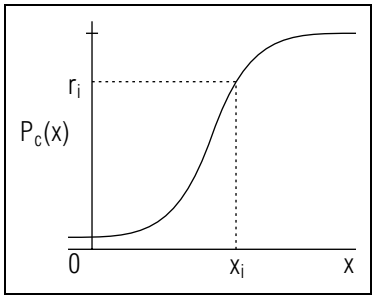


Figure 2. The process of finding $x_i = P_c^{-1}(r_i)$ is indicated graphically.

corresponding to the r_i's.[7] The spacing of points on the x-axis is not uniform, the points being closest together near the value at which the probability density function (Fig. 3b.) has its maximum. A point on the vertical axis chosen at random will be most likely to correspond to a point in the densely packed region of the x-axis.

3c. The Basis (Mathematical, Demonstration). We can also show mathematically that the data generated by this method will be distributed

[7] A second graphical method involves constructing histograms which can be compared to $p(x)$. See Sect. 5 of "Error Analysis of 50 Readings of 5 Second Intervals" (MISN-0-373).

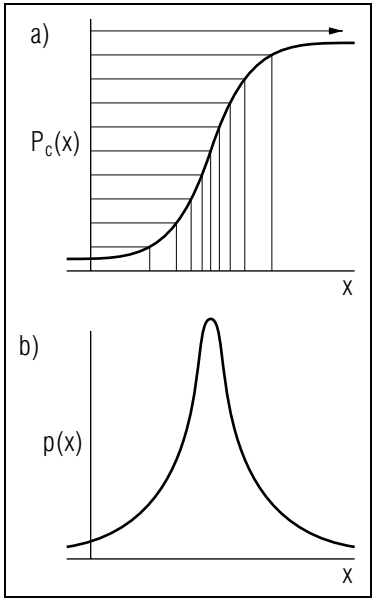


Figure 3. Uniformly distributed numbers r_i on the vertical axis translate into non-uniformly distributed x_i on the horizontal axis. 18The region of highest density corresponds to the peak of $p(x)$.

according to $p(x)$. Using the method of Sect. 3, we effectively started with a probability density function $p(x)$, and found a set of random variables x_i, each given by solving the following equation for x_i:

$$r_i = P_c(x_i) = \int_{x_l}^{x_i} p(x')dx'. \tag{4}$$

We must show that the resulting probability density function for the variables x_i (let us call it $p_1(x)$) is the same as $p(x)$. Since for each r_i there corresponds an x_i, on a one to one basis, the (infinitesimal) probability that x_i has a value in the interval between x_0 and $x_0 + dx$ is equal to the probability that r_i has its value in the corresponding interval r_0 to $r_0 + dr$. That is,

$$p_1(x_o)dx = p_2(r_o)dr, \tag{5}$$

or, in general,

$$p_1(x) = p_2(r)\frac{dr}{dx}, \tag{6}$$

Since r has a uniform distribution on the interval zero to one, $p_2(r) = K$, a constant. The normalization conditon, Equation (2), demands that $K = 1$. Equation (6) becomes

$$p_1(x) = \frac{dr}{dx} = \frac{d}{dx}\int_{x_l}^{x} p(x')dx' = p(x). \tag{7}$$

That is, $p_1(x) = p(x)$, as needed.

4. The Lorentzian Distribution

4a. Typical Uses for the Lorentzian. The Lorentzian distribution is often used in quantum mechanics to describe resonance phenomena. For instance, a proton and a pion are most likely to interact in a collision when their total energy is close to the mass-energy of a delta particle. The probability of an interaction as a function of energy, $p(E)$, is Lorentzian in form. The distribution of photon energies within a given spectral line is also Lorentzian.

4b. The Lorentzian Function. The Lorentzian probability density function is

$$P_L(x) = \frac{1}{\pi}\frac{\Gamma/2}{(x-\mu)^2 + (\Gamma/2)^2}, \tag{8}$$

where μ is the mean value of the variable x and Γ is the "full width at half maximum" of the peak (see Figure 4).

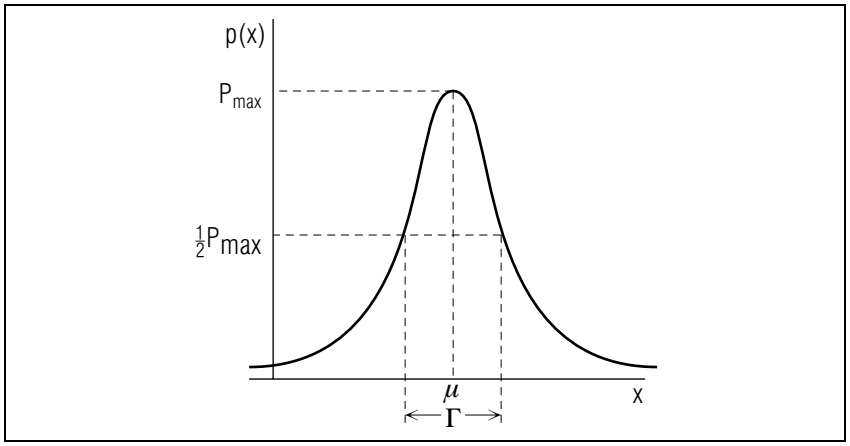


Figure 4. The Lorentzian distribution is peaked symmetrically about the mean value μ. Its shape depends upon the value for Γ.

4c. $P_c(x)$ for the Lorentzian. The cumulative probability function for the Lorentzian distribution is

$$P_c(x) = \int_{-\infty}^{x} \frac{1}{\pi}\frac{\Gamma/2}{(x'-\mu)^2 + (\Gamma/2)^2}dx', \tag{9}$$

since the values of x may range from $-\infty$ to $+\infty$. Performing the integration yields

$$P_c(x) = \frac{1}{\pi}\tan^{-1}\left(\frac{x-\mu}{\Gamma/2}\right) + \frac{1}{2}. \tag{10}$$

4d. Generating Lorentzian Data. Our process involves repeatedly setting P_c(x) equal to some random number r_i between zero and one, and solving for the corresponding x_i. We can solve for x_i algebraically, yielding

$$x_i = \frac{\Gamma}{2}\tan\left[\pi\left(r_i - \frac{1}{2}\right)\right] + \mu. \tag{11}$$

4e. Sample Data. We now use random numbers r_i to generate Lorentzian data x_i. Table I lists data generated by this method for two different cases. The first is for $\mu = 1/2$, $\Gamma = 0.05$ (a "narrow" distribution). The second is for $\mu = 1/2$, $\Gamma = 0.5$ (a "wide" distribution).

Table I. Simulated "Lorentzian" data depend on Γ, the "width" of the distribution.

x_i Data "Narrow"	x_i Data "Wide"	Original $r_i^\dagger$
0.48792	0.37915	.35667
0.48465	0.34646	.32469
0.46490	0.14899	.19700
0.53481	0.84810	.80175
0.45384	0.03837	.15799
0.57377	1.23767	.89599
0.49331	0.43307	.41673
0.52238	0.72379	.73241
0.67149	2.21486	.95392
0.35731	-0.92688	.05521
Mean=0.504	Mean=0.544	

† Pseudorandom numbers generated by the power residue method.

Note that just as in a real experiment, you do not get exactly the "right" answer. Neither data set gives a mean value of exactly 1/2 and neither is even symmetric about $x = 1/2$. That is, in this case, more data points lie below that value than above. A very large amount of data would need to be generated to show the true Lorentzian character of the distribution.[8]

5. The Gaussian Distribution

5a. Typical Uses for the Gaussian. Random errors of measurement will cause a set of repeated measurements of the same quantity to be distributed according to the Gaussian distribution.[8] For example, ten measurements of a pendulum period will not give exactly the same result because of these random errors.

5b. The Gaussian Function. A Gaussian distribution probability density function is

$$p(x) = \frac{1}{\sigma(2\pi)^{1/2}}\, e^{-\dfrac{(x-\mu)^2}{2\sigma}}, \tag{12}$$

[8] Also called the "bell-shaped curve."

where μ is the mean value of the variable x, and σ is the standard deviation of the measurements. (If many repeated measurements are made, about 68% of them will lie in the range $\mu \pm \sigma$.)

5c. $P_c(x)$ and $P_c^{-1}(x)$ for the Gaussian. The expression

$$P_c(x) = \int_{-\infty}^{x} \frac{1}{\sigma(2\pi)^{1/2}}\, e^{-\dfrac{(x-\mu)^2}{2\sigma}}\, dx, \tag{13}$$

cannot be integrated analytically. However, algebraic formulas for $P_c(x)$ and $P_c^{-1}(x)$ do exist. The function we need is

$$x = P_c^{-1}(r) \simeq \sigma \left[t - \frac{c_o + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3} \right] + \mu, \tag{14}$$

where $t = \left[\ell n\,(1-r)^{-2}\right]^{1/2}$

$c_o = 2.515517$; $c_1 = 0.802853$; $c_2 = 0.010328$

$d_1 = 1.432788$; $d_2 = 0.189269$; $d_3 = 0.001308$

Clearly, if much data must be generated a computer should be used.

Acknowledgments

Preparation of this module was supported in part by the National Science Foundation, Division of Science Education Development and Research, through Grant #SED 74-20088 to Michigan State University.

PROBLEM SUPPLEMENT

1. The values of a variable x are distributed according to the probability density function $p(x) = C$, a constant, in the interval $|x| \leq 1$ and $p(x) = 0$ for $1 < |x|$.

   a. Use the normalization condition to find the value of C.

   b. Sketch $p(x)$ vs. x.

   c. What is the probability that x will occur in the range $-1 \leq x \leq 0$?

   d. Integrate $p(x)$ to find $P_c(x)$.

   e. Sketch $P_c(x)$ vs. x.

   f. Write the equation $x_i = P_c^{-1}(r_i)$.

   g. Generate 20 data points x_i. Examine the results and describe their distribution qualitatively.

   h. What fraction of your generated data points lie in the range $-1 \leq x \leq 0$? How should the answer compare with your answer to part (c)?

2. The actual lifetimes t of unstable nuclei of a given species are distributed according to the probability density function

$$p(t) = \frac{1}{\tau} e^{-t/\tau} \text{ for } 0 \leq t \leq \infty.$$

   τ is the mean lifetime of that species. Assume that $\tau = 1$ second for this example.

   a. Sketch $p(t)$ vs. t.

   b. What is the probability that a nucleus will decay during the first second?

   c. Find $P_c(t)$, the probability of decaying by time t.

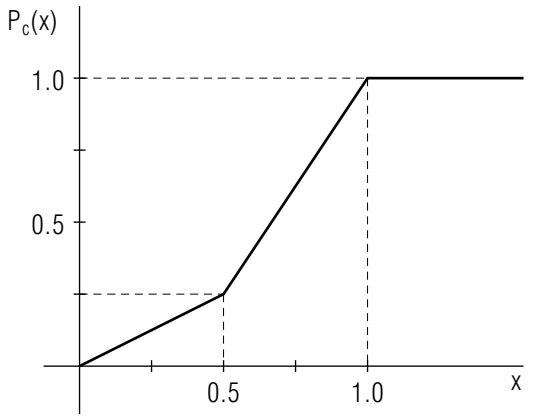
   d. Sketch $P_c(t)$ vs. t.

   e. Write the equation $t_i = P_c^{-1}(r_i)$.

   f. Generate 20 lifetimes t_i.

   g. What fraction of your nuclei have decayed during the first second? How does this answer compare with part (b)?

3. The cumulative probability function for a set of variables x is given by this graph:



   a. Describe the distribution of the resulting data generated in the interval $0 \leq x \leq 0.5$. Use graphical methods suggested by Section 4a.

   b. Repeat for the interval $0.5 \leq x \leq 1$.

   c. How do the distributions of part (a) and (b) differ?

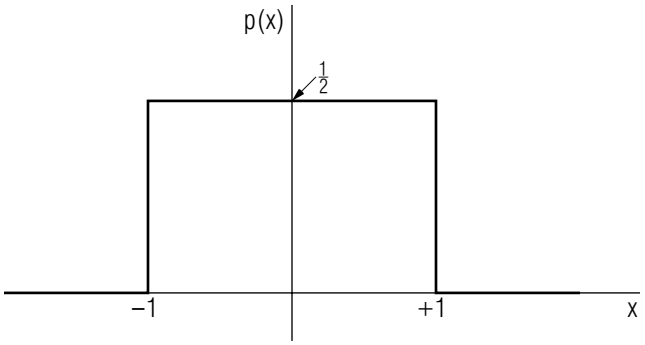
   d. Graph the probability density function for this data.

   e. Write the equation for the normalized probability function in four regions: $x < 0$, $0 \leq x \leq 0.5$, $0.5 \leq x \leq 1.0$ and $x > 1.0$.

Brief Answers:

1. a. $C = 1/2$

b.



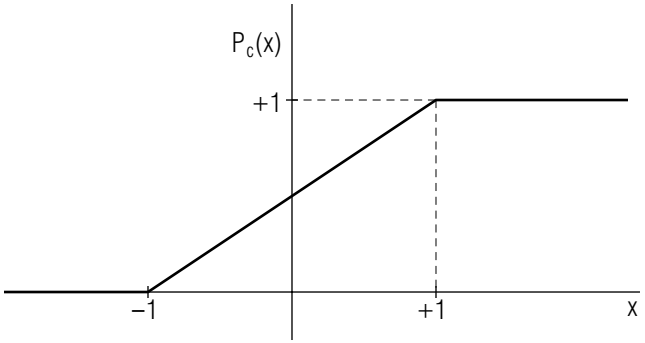
c. 1/2 or 50%

d. $P_c(x) = 0$ for $x < -1$

$P_c(x) = \frac{1}{2}(x+1)$ for $-1 < x < +1$

$P_c(x) = 1$ for $+1 < x$

e.

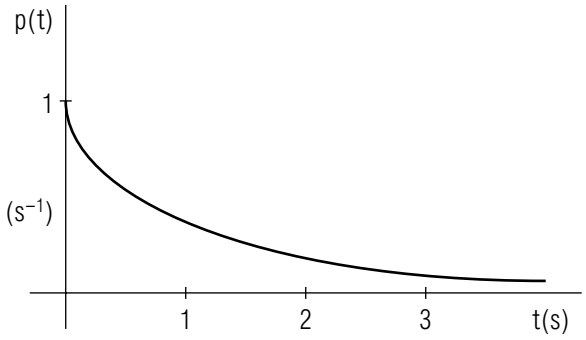


f. $x_i = 2r_i - 1$

g. The x_i depend upon the r_i generated. The distribution should look relatively uniform on the interval -1 to $+1$. No points should lie outside that interval.

h. The fraction depends upon the r_i that occurred. Any value is possible, although something near 50% is most likely.

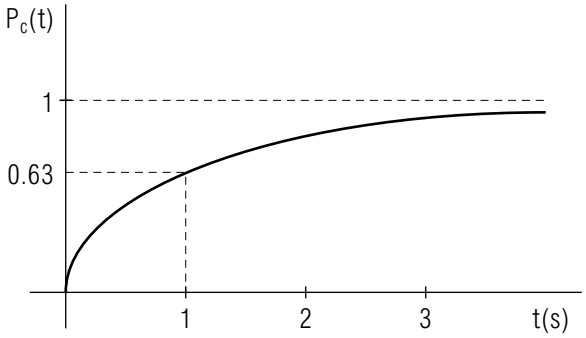
2. a.



b. $(1 - e^{-1}) = 0.63$ or 63%

c. $P_c(t) = 1 - e^{(-t/1\,\mathrm{s})}$

d.



e. $t_i = (1\,\mathrm{s})\ell n\,(1 - r_i)^{-1}$

f. The t_i depend upon the r_i generated.

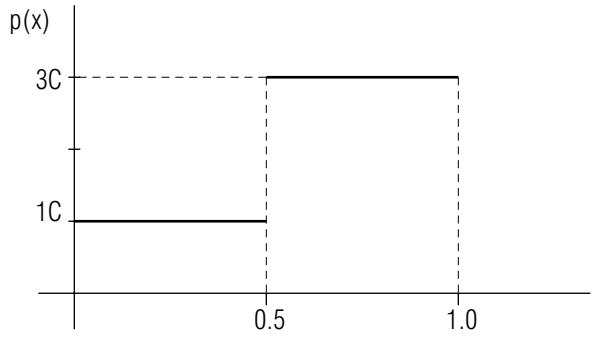
g. The fraction depends upon the r_i that occurred. Any value is possible, although something near 63% is most likely. Very long lifetimes should be infrequent.

3. a. A uniform distribution over the interval.

b. A uniform distribution over the interval.

c. The density (and therefore $p(x)$) for the interval $0.5 < x \leq 1$ is three times that of the interval $0 \leq x \leq 0.5$.

d.



e. The normalization requirement gives $C = 1/2$.

$p(x) = 0$	for			x	$<$	0.0
$p(x) = 1/2$	for	0.0	$\leq$	x	$\leq$	0.5
$p(x) = 3/2$	for	0.5	$<$	x	$\leq$	1.0
$p(x) = 0$	for	1	$<$	x	$>$	1.0

SPECIAL ASSISTANCE SUPPLEMENT

Hints for Problem Supplement:

1. a. Normalization means $\int_{x_\ell}^{x_\mu} p(x)dx = 1$.

   For this problem, this reduces to $\int_{-1}^{+1} C dx = 1$.

   c. The probability in a finite interval is $\int_{x_1}^{x_2} p(x)dx$.

   d. $P_c(x) = \int_{x_\ell}^{x} p(x)dx$. The integral must be evaluated three times, for x in the three regions $x < -1$, $-1 \leq x \leq +1$, and $x > +1$. In each case, $x_\ell = -\infty$.

   f. Solve the equation $P_c(x) = r$ for the variable x.

   g. Generate 20 random numbers r_i (or use tables), and plug each into the equation of part (f).

2. a. The probability in a finite interval is $\int_{t_1}^{t_2} p(t)dt$ with $t_1 = 0$, $t_2 = 1\,\text{s}$.
   $\int e^{-ax}dx = \dfrac{1}{a}e^{-ax}$.

   c. $P_c(t) = \int_0^t p(t)dt$

   e. Solve the equation $P_c(t) = r$ for t.

   f. Generate 20 random numbers r_i (or use tables), and plug each into the equation of part (e).

3. a. Use equally spaced lines on the vertical axis, (e.g., use 20).

   b. Use equally spaced lines on the vertical axis, (e.g., use 20). Compare the spacing on the horizontal axis.

   c. Compare the density of lines on the horizontal axis for the two intervals.

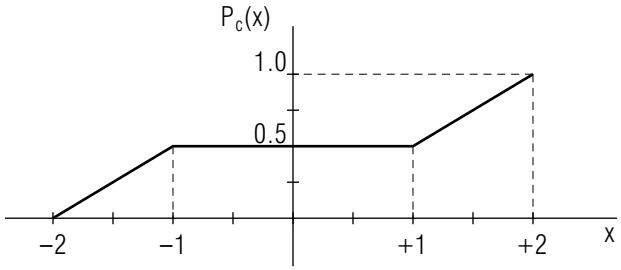
   d. Make sure your graph includes a numerical comparison of the densities in each interval.

   e. Normalization requires $\int_{x_\ell}^{x_u} p(x)dx = 1$. Does this area under your $p(x)$ vs. x curve equal one?

MODEL EXAM

1. Given: $p(x) = kx$ for: $0 < x < 2$; $p(x) = 0$ otherwise
   a. Use the normalization condition to find k.
   b. Sketch $p(x)$ vs. x.
   c. What is the probability that x will occur in the range $0 < x < +1$?
   d. Find $P_c(x)$ and sketch $P_c(x)$ vs. x.
   e. Generate 20 data points x_i. Describe their distribution qualitatively.

2. The cumulative probability function for a set of variables x is given in this sketch:

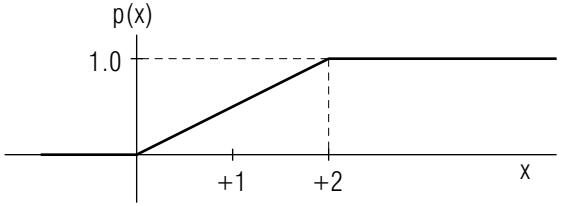


   Describe qualitatively the nature of the distribution of the data generated.

3. Define or describe the "cumulative probability function."

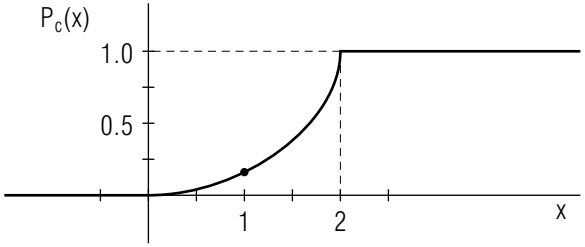
Brief Answers:

1. a. $k = 1/2$
   b.



   c. $P(0 \le x \le 1) = 1/4 \int_0^1 \frac{1}{2} x dx = \frac{1}{4} x^2 |_0^1 = \frac{1}{4}$

   d. $P_c(x) = \frac{1}{4} x^2$ for: $0 \le x \le 2$



   e. The density becomes increasingly higher as x increases from 0 to 2. There are no points outside that interval.

2. Data points fall either in the interval $-2 \le x \le -1$ or $1 \le x \le 2$. They are uniformly distributed on each of those intervals, with the same density on each.

3. The cumulative probability function $P_c(x)$ is the probability that the variable x_i will occur somewhere in the interval from x_ℓ to x.