# BIM 105
# Probability and Statistics for Biomedical Engineers


## David M. Rocke

## Department of Biomedical Engineering

# Random Variables

- If we have a probabilistic experiment in which the outcome is or can be thought of as a number, then this is called a *random variable*.
- Formally, a random variable is a function of outcomes in a sample space that assigns a number to each outcome.
- This can be an integer, either bounded or unbounded, or a real number, including possibly negative numbers.
- We usually denote a random variable by a capital Latin letter like $X$, and a realized value of the random variable by a lower-case Latin letter like $x$ or $x_i$.

# Three Machines

- Machines 1, 2, and 3 are either up (operational) or down (not working).
- Each has independently a 10% chance of being down.
- There are eight elements in the sample space {UUU, UUD, UDU, UDD, DUU, DUD, DDU, DDD}
- We can compute the probability of each outcome since the chance of up is 0.9 and the chance of down is 0.1. For example, the chance of UDU is $(0.9)(0.1)(0.9) = 0.081$

| Outcome | Probability |
|---|---|
| UUU | (.9)(.9)(.9) = 0.729 |
| UUD | (.9)(.9)(.1) = 0.081 |
| UDU | (.9)(.1)(.9) = 0.081 |
| UDD | (.9)(.1)(.1) = 0.009 |
| DUU | (.1)(.9)(.9) = 0.081 |
| DUD | (.1)(.9)(.1) = 0.009 |
| DDU | (.1)(.1)(.9) = 0.009 |
| DDD | (.1)(.1)(.1) = 0.001 |

- A random variable we could assign to the sample space is the number X of machines that are up.
- The possible values for this random variable are {0, 1, 2, 3}
- We can compute the probabilities for each of the values from the probabilities of the items in the sample space.
- This is called a *discrete* probability distribution.
- Discrete because the possible values are listable.
- The probability distribution is also called a probability mass function.

| Outcome | Probability | X |
|---|---|---|
| UUU | (.9)(.9)(.9) = 0.729 | 3 |
| UUD | (.9)(.9)(.1) = 0.081 | 2 |
| UDU | (.9)(.1)(.9) = 0.081 | 2 |
| UDD | (.9)(.1)(.1) = 0.009 | 1 |
| DUU | (.1)(.9)(.9) = 0.081 | 2 |
| DUD | (.1)(.9)(.1) = 0.009 | 1 |
| DDU | (.1)(.1)(.9) = 0.009 | 1 |
| DDD | (.1)(.1)(.1) = 0.001 | 0 |

| x | P(X = x) |
|---|---|
| 0 | 0.001 |
| 1 | 3(0.009) = 0.027 |
| 2 | 3(0.081) = 0.243 |
| 3 | 0.729 |
| Total | 1.000 |

# Defects in a Silicon Wafer

- The number of defects could be 0, or it could be 1, 2, 3, …
- There is no specific upper limit to the number of defects.
- Suppose that 60.7% have no defects, 30.3% have 1 defect, 7.6% have 2 defects, 1.3% have 3 defects, 0.1% have 4 defects.
- More than 4 defects can occur, but it happens less than 1 time in 5,000, so for the moment we will ignore that possibility.

| x | P(X = x) |
|---|---|
| 0 | 0.607 |
| 1 | 0.303 |
| 2 | 0.076 |
| 3 | 0.013 |
| 4 | 0.001 |
| Total | 1.000 |

# Computer Disk Drive Failure

- The time (in hours) from the initial operation of a disk drive until it fails is a random variable.
- It can take on any positive value.
- To specify this completely, we need to know the probability that the failure time is less than $x$ hours for every value of $x > 0$.
- This requires the specification of the distribution function $F(x) = P(X \leq x)$, which must be defined for every value of $x$.

$X =$ failure time of the disk drive, a random variable

$$F(x) = P(X \le x) = 1 - e^{-\lambda x} \text{ , where } \lambda = 10^{-6}$$

There are 8,760 hours in a year.

$$\begin{aligned} F(8760) &= P(X \le 8760) \\ &= 1 - e^{-8760\lambda} \\ &= 1 - e^{-0.00876} \\ &= 0.00872 \\ F(8760 \times 5) &= F(43800) \\ &= 1 - e^{-43800\lambda} \\ &= 1 - e^{-0.0438} \\ &= 0.0429 \end{aligned}$$

The chance that this disk drive will run continuously for 5 years without a failure exceeds 95%. Is this consistent with your experience of computer disk drives? How would you measure this on a new disk drive design?

Here is an example excerpt from a Product Manual, in this case for the Seagate Barracuda ES.2 Near-Line Serial ATA drive:

*The product shall achieve an Annualized Failure Rate - AFR - of 0.73% (Mean Time Between Failures - MTBF - of 1.2 Million hrs) when operated in an environment that ensures the HDA case temperatures do not exceed 40°C. Operation at case temperatures outside the specifications in Section 2.9 may increase the product Annualized Failure Rate (decrease MTBF). AFR and MTBF are population statistics that are not relevant to individual units.*

*AFR and MTBF specifications are based on the following assumptions for business critical storage system environments:*

- *8,760 power-on-hours per year.*
- *250 average motor start/stop cycles per year.*
- *Operations at nominal voltages.*
- *Systems will provide adequate cooling to ensure the case temperatures do not exceed 40°C. Temperatures outside the specifications in Section 2.9 will increase the product AFR and decrease MTBF.*

# Random Variables and Populations

- When a random variable is described, with the sample space, and the likelihood of each value in it, then this in effect describes a population.
- The population is of hypothetical repeated observations of the random variable.
- The distribution function tells us the probability of each outcome.

# Discrete Random Variables

- The set of possible values is discrete, meaning that it is finite, or that it can be listed even though it is infinite.
- A finite set of possible values might be
  - {0, 1, 2, 3}
  - {1/2, 1, 3/2, 2}
- An infinite set might be the set of even numbers
  - {0, 2, 4, 6, … }
- We must specify the set of possible values plus the probability mass function, which gives the probability of each outcome.
- Each probability must be between 0 and 1 and they must add up to 1.

# The Cumulative Distribution Function

Let X be a discrete random variable

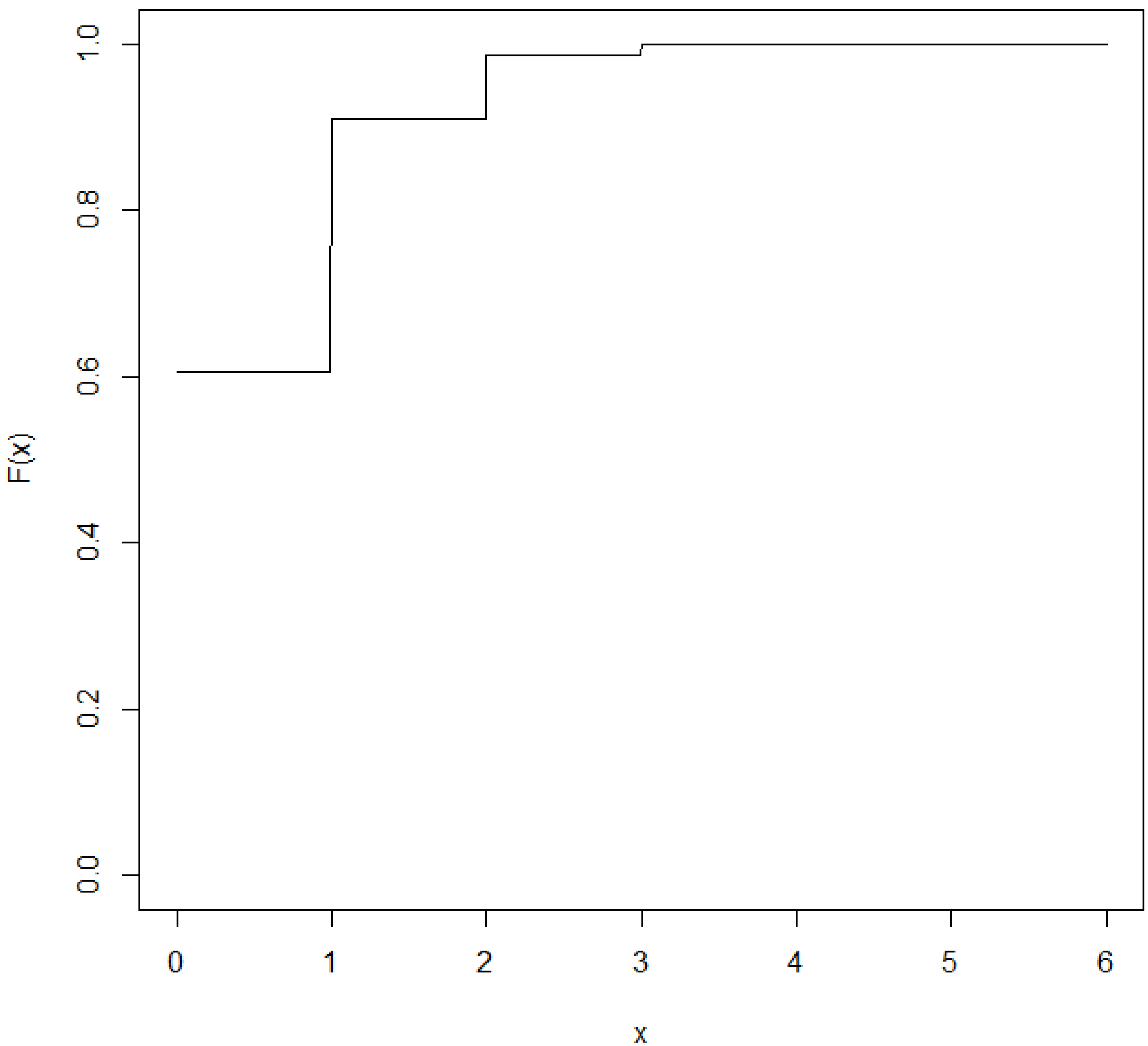
The probability mass function is

$f(x) = P(X = x)$

The cumulative distribution function (CDF) is

$F(x) = P(X \leq x)$

| x | f(x) = P(X = x) | F(x) = P(X ≤ x) |
|---|---|---|
| 0 | 0.607 | 0.607 |
| 1 | 0.303 | 0.910 |
| 2 | 0.076 | 0.986 |
| 3 | 0.013 | 0.999 |
| 4 | 0.001 | 1.000 |
| Total | 1.000 | |


1.0
0.8
0.6
0.4
0.2
0.0
F(x)
0
1
2
3
4
5
6
x

# The Cumulative Distribution Function

$$F(0) = 0.607$$

$$F(0.5) = 0.607$$

$$F(1) = 0.910$$

$$F(.9999999) = 0.607$$

$$F(1.0000001) = 0.910$$

| x | f(x) = P(X = x) | F(x) = P(X ≤ x) |
|---|---|---|
| 0 | 0.607 | 0.607 |
| 1 | 0.303 | 0.910 |
| 2 | 0.076 | 0.986 |
| 3 | 0.013 | 0.999 |
| 4 | 0.001 | 1.000 |
| Total | 1.000 | |

# The Mean of a Discrete Random Variable

- The *mean* of a discrete random variable $X$ is the weighted average of the possible values, each weighted by the probability of that outcome. It is usually denoted by $\mu = \mu_X$.
- The mean is also called the *expectation*, the *expected value*, or the *mathematical expectation* of the random variable.
- In spite of the name, the expected value does not even need to be a possible value of the random variable, much less the one that you "expect" to get.

# Mean Defect Number

$$\mu_X = \sum_x xP(x)$$

$$= 0.498$$

| x | f(x) = P(X = x) | xP(x) |
|---|---|---|
| 0 | 0.607 | 0 |
| 1 | 0.303 | 0.303 |
| 2 | 0.076 | 0.152 |
| 3 | 0.013 | 0.039 |
| 4 | 0.001 | 0.004 |
| **Total** | **1.000** | **0.498** |

# The Variance of a Discrete Random Variable

- The *variance* of a discrete random variable is the weighted average of the squared deviations of the possible values, each weighted by the probability of that outcome. It is usually denoted $\sigma^2 = \sigma_X^2$ or by $V(X)$.
- The square root of the variance is the *standard deviation*, denoted by $\sigma = \sigma_X$.
- These are population descriptions. We do not at this point have a sample.

# Variance of the Defect Number

$$\sigma_X^2 = \sum_x (x-\mu_X)^2 P(x)$$

$$= 0.491996$$

$$= \sum_x x^2 P(x) - \mu_X^2$$

$$= 0.740 - (0.498)^2 = 0.491996$$

$$\sigma_X = \sqrt{0.491996} = 0.701424$$

| x | P(X = x) | xP(x) | $(x-\mu_X)^2P(x)$ | $x^2P(x)$ |
|---|---|---|---|---|
| 0 | 0.607 | 0 | $(-.498)^2(.607) = 0.150538$ | 0 |
| 1 | 0.303 | 0.303 | $(0.502)^2(0.303) = 0.076357$ | 0.303 |
| 2 | 0.076 | 0.152 | $(1.502)^2(.076) = 0.171456$ | 0.304 |
| 3 | 0.013 | 0.039 | $(2.502)^2(0.013) = 0.081380$ | 0.117 |
| 4 | 0.001 | 0.004 | $(3.502)^2(0.001) = 0.012264$ | 0.016 |
| Total | 1.000 | 0.498 | 0.491996 | 0.740 |

# Continuous Random Variables

- A continuous random variable is one where the possible values consist of a range of real numbers.
- Common values for the range include $(-\infty, \infty)$, and $[0, \infty)$, but it could also be a range of values like $[0, 1]$.
- For every value $x$ in the range, we need to know $F_X(x) = P(X \leq x)$. This is called the cumulative distribution function or CDF.
- There is another function $f_X(x)$, the probability density function (PDF), which is like a histogram with very small bar widths.
- For any interval $[a, b]$, the area under the curve of $f_X(x)$ is the probability that the random variable falls in that interval.
- Thus $f_X(x)$ is the derivative of $F_X(x)$.

$X =$ failure time of the disk drive, a random variable

Possible values are in $[0, \infty)$

$$F(x) = P(X \le x) = 1 - e^{-\lambda x} \text{ , where } \lambda = 10^{-6}$$

$$f(x) = \lambda e^{-\lambda x}$$

$$f(x) = F'(x)$$

$$F(a) = \int_0^a f(x)dx$$

$$F(b) - F(a) = P(a \le x \le b) = \int_a^b f(x)dx$$

$$P(x = a) = 0$$

$$P(a \le x \le b) = P(a \le x < b) = P(a < x \le b) = P(a < x < b)$$

If the possible values of $x$ are $[a,b]$ then

$$\int_a^b f(x)dx = 1$$

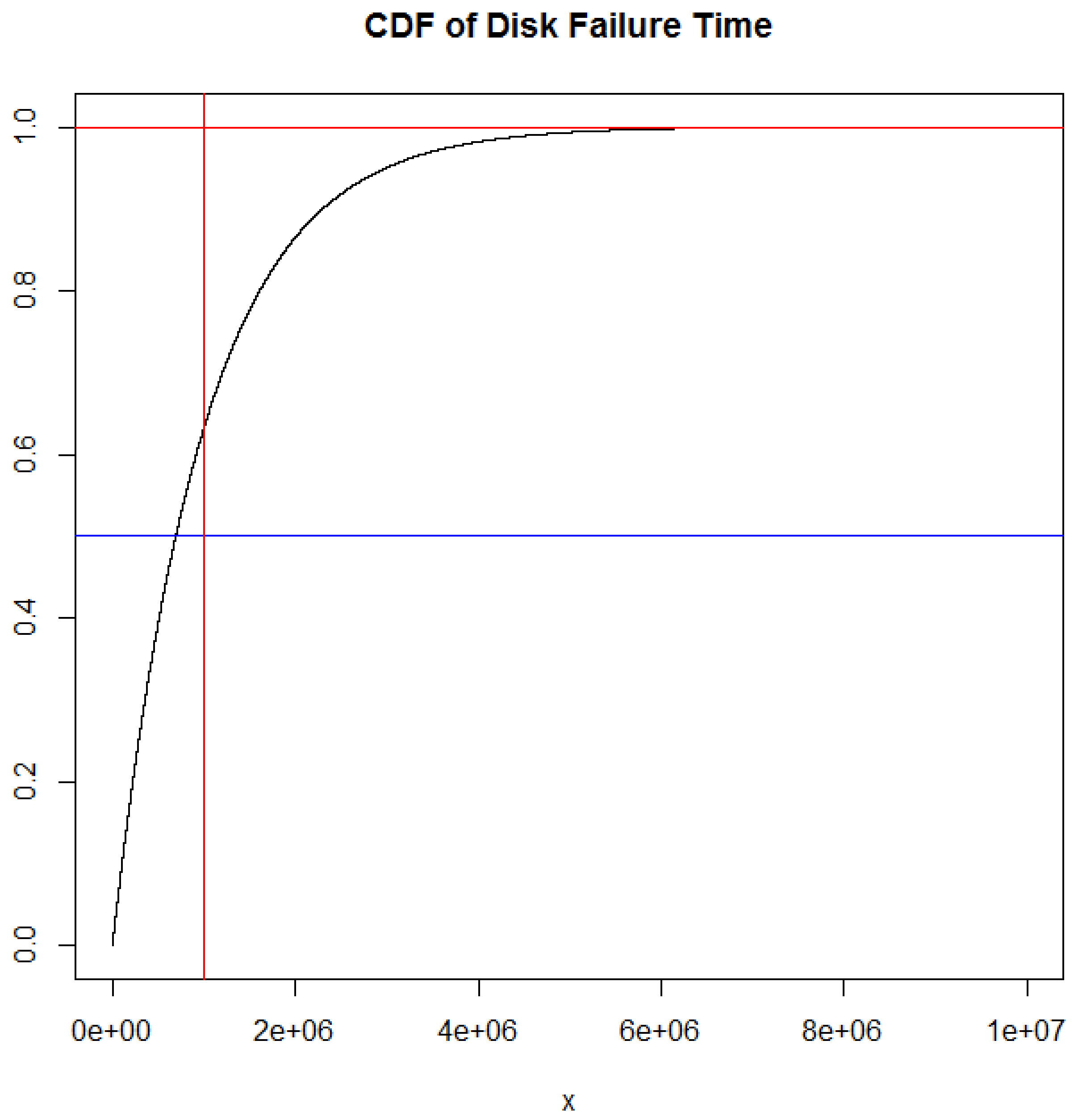
$$F(a) = 0$$

$$F(b) = 1$$

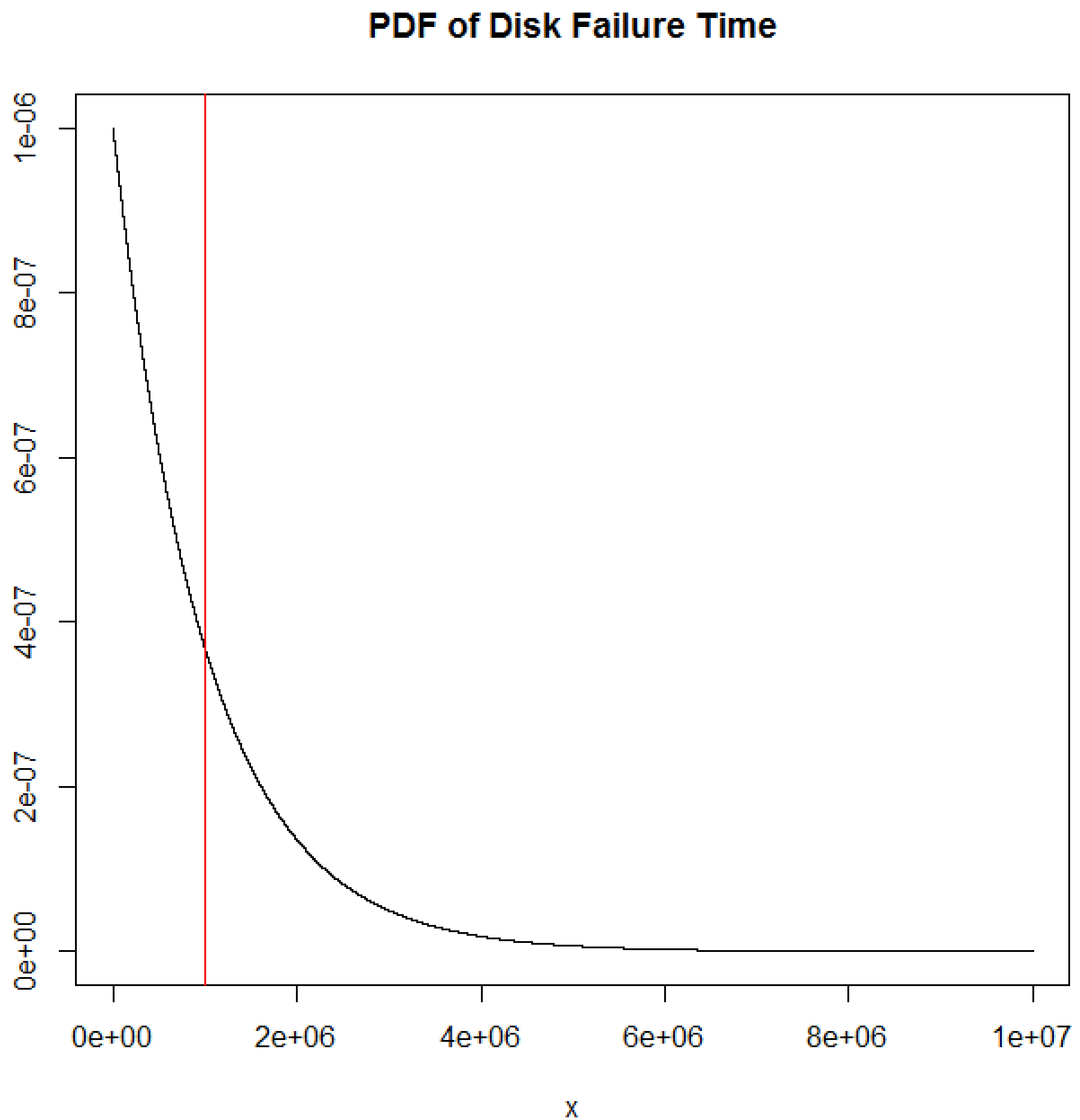
If $b = \infty$, then

$$\lim_{x\to\infty} F(x) = 1$$

$$F(x) = \int_a^x f(t)dt$$



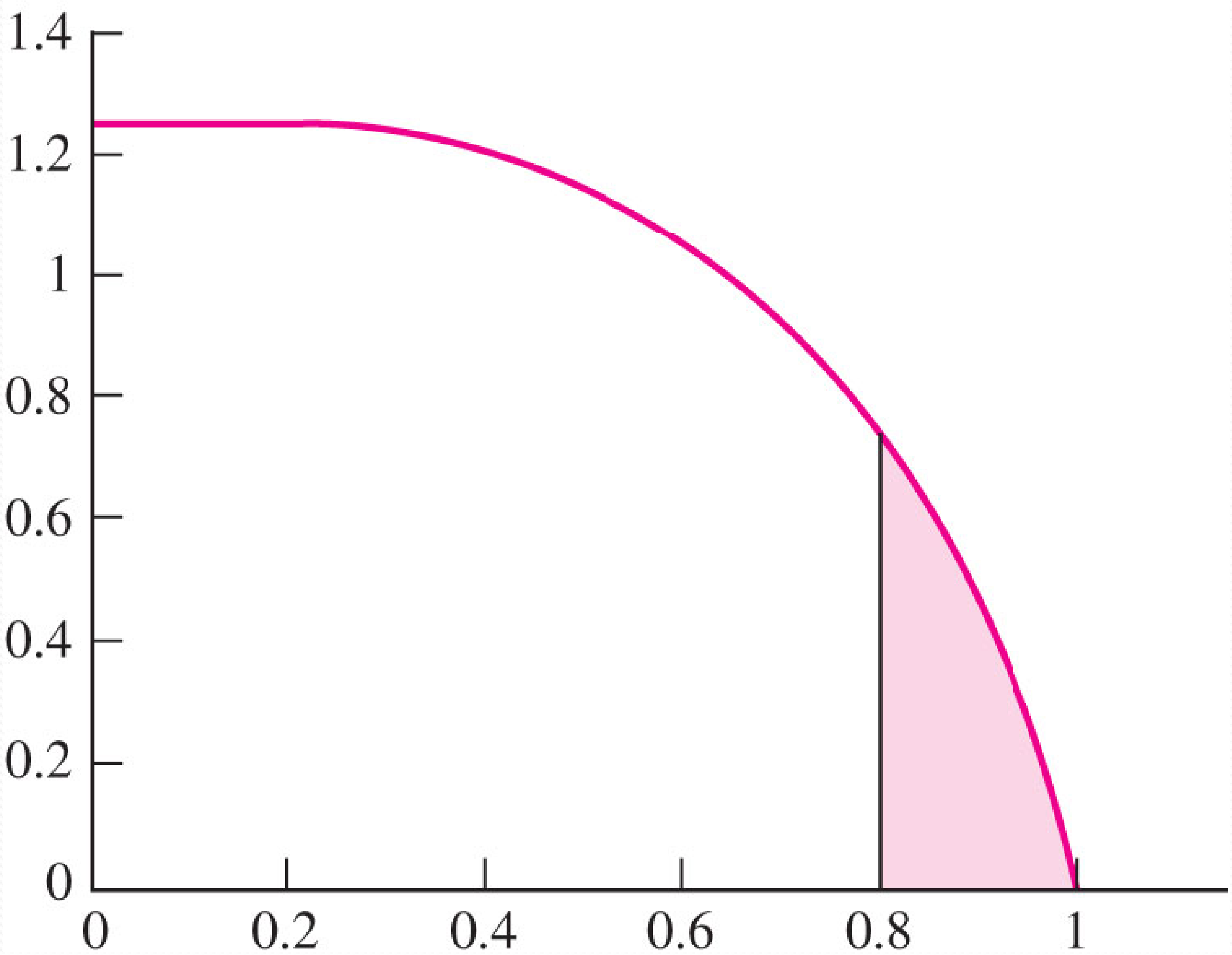
CDF of Disk Failure Time
0.0
0.2
0.4
0.6
0.8
1.0
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
x


PDF of Disk Failure Time
1e-06
8e-07
6e-07
4e-07
2e-07
0e+00
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
x

# Example 2.41

- A hole is drilled in a sheet-metal component and a shaft inserted through the hole. The clearance in mm is a random variable X with density shown below. If the clearance is large than 0.8mm it must be scrapped. What is the chance of this happening?

$$f(x) = \begin{cases} 1.25(1-x^4) & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$


1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

$$f(x) = \begin{cases} 1.25(1 - x^4) & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$1 - F(0.8) = P(X > .8) = \int_{0.8}^{\infty} f(x)dx$$

$$= \int_{0.8}^{1} 1.25(1 - x^4)dx$$

$$= 1.25x - \frac{1.25}{5}x^5 \Big|_{0.8}^{1}$$

$$= 1.25 - 0.25 - (1.25)(0.8) + 0.25(0.8)^5$$

$$= 1 - 1 + 0.08192$$

$$= 0.08192$$

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

$$= \int_{0}^{x} f(t)dt \text{ for } 0 < x < 1$$

$$= \int_{0}^{x} 1.25(1 - t^4)dt$$

$$= 1.25(t - \frac{t^5}{5})\Big|_0^x$$

$$= 1.25x - \frac{x^5}{4}$$

$$F(0) = 0$$

$$F(1) = 1.25 - 0.25 = 1$$

Find the probability that the clearance is less than 0.5mm

$$F(0.5) = (1.25)(0.5) - (0.25)(0.5)^5 = \frac{5}{8} - \frac{1}{128}$$

$$= \frac{79}{128} = 0.6172$$

The mean and the variance of a continuous random variable with set of possible values $R$

$$\mu_X = \int_R x f(x) dx$$

$$\sigma_X^2 = V(X) = \int_R (x - \mu_X)^2 f(x) dx$$

$$= \int_R x^2 f(x) dx - 2\mu_X \int_R x f(x) dx + \int_R \mu_X^2 f(x) dx$$

$$= \int_R x^2 f(x) dx - 2\mu_X^2 + \mu_X^2$$

$$= \int_R x^2 f(x) dx - \mu_X^2$$

For the clearance random variable

$$\mu_X = \int_0^1 x[1.25(1-x^4)]dx$$

$$= \int_0^1 1.25(x - x^5)dx$$

$$= 1.25\left(\frac{x^2}{2} - \frac{x^6}{6}\right)\Bigg|_0^1$$

$$= 1.25/3 = 0.4167$$

$$\sigma_X^2 = \int_0^1 x^2[1.25(1-x^4)]dx - \mu_X^2$$

$$= \int_0^1 1.25(x^2 - x^6)dx - \mu_X^2$$

$$= 1.25\left(\frac{x^3}{3} - \frac{x^7}{7}\right)\Bigg|_0^1 - (0.4167)^2$$

$$= 0.0645$$

$$\sigma_X = 0.254$$

$$X = \text{ failure time of the disk drive, a random variable}$$

Possible values are in $[0, \infty)$

$$F(x) = P(X < x) = 1 - e^{-\lambda x} \text{ , where } \lambda = 10^{-6}$$

$$f(x) = \lambda e^{-\lambda x}$$

$$\mu_X = \int_0^\infty \lambda x e^{-\lambda x} dx$$

$$= -(x + \lambda^{-1}) e^{-\lambda x} \Big|_0^\infty \quad (e^{-\lambda x} \text{ falls faster than x rises})$$

$$= 0 - (-\lambda^{-1})$$

$$= \lambda^{-1}$$

Uses integration by parts, but can be checked as follows

$$\frac{d}{dx}\left[-xe^{-\lambda x} - \lambda^{-1} e^{-\lambda x}\right] = -e^{-\lambda x} + \lambda x e^{-\lambda x} + e^{-\lambda x}$$

$$= \lambda x e^{-\lambda x}$$

$X =$ failure time of the disk drive, a random variable

Possible values are in $[0, \infty)$

$$F(x) = P(X < x) = 1 - e^{-\lambda x} \text{ , where } \lambda = 10^{-6}$$

$$f(x) = \lambda e^{-\lambda x}$$

$$\mu_X = \lambda^{-1}$$

$$\int_0^\infty \lambda x^2 e^{-\lambda x} dx = -\lambda^{-2}(u^2 - 2u + 2)e^u \Big|_0^{-\infty} \quad \text{using } u = -\lambda x$$

$$= 2\lambda^{-2}$$

$$\sigma_X^2 = 2\lambda^{-2} - (\lambda^{-1})^2$$

$$= \lambda^{-2}$$

$$\sigma_X = \lambda^{-1}$$

Transformations and Combinations of Random Variables

$$Y = X + a$$
$$\mu_Y = \mu_X + a$$
$$\sigma_Y^2 = \sigma_X^2$$

$$Y = bX$$
$$\mu_Y = b\mu_X$$
$$\sigma_Y^2 = b^2\sigma_X^2$$
$$\sigma_Y = b\sigma_X$$

$$Y = a + bX$$
$$\mu_Y = a + b\mu_X$$
$$\sigma_Y^2 = b^2\sigma_X^2$$
$$\sigma_Y = b\sigma_X$$

# This is important!

Transformations and Combinations of Random Variables

$Y = X_1 + X_2$

$\mu_Y = \mu_{X_1} + \mu_{X_2}$

$\sigma_Y^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2$ if $X_1, X_2$ are statistically independent

$X_1, X_2, ..., X_n$ are random variables

$Y = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$

$\mu_Y = c_1 \mu_{X_1} + c_2 \mu_{X_2} + \cdots + c_n \mu_{X_n}$

if the $\{X_i\}$ are all statistically independent, then

$\sigma_Y^2 = c_1^2 \sigma_{X_1}^2 + c_2^2 \sigma_{X_2}^2 + \cdots + c_n^2 \sigma_{X_n}^2$

$X_1, X_2, ..., X_n$ are random variables

$$Y = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$$

$$\mu_Y = c_1 \mu_{X_1} + c_2 \mu_{X_2} + \cdots + c_n \mu_{X_n}$$

if the $\{X_i\}$ are all statistically independent, then

$$\sigma_Y^2 = c_1^2 \sigma_{X_1}^2 + c_2^2 \sigma_{X_2}^2 + \cdots + c_n^2 \sigma_{X_n}^2$$

$$\mu_{X_i} = \mu$$

$$\sigma_{X_i}^2 = \sigma^2$$

$$\bar{X} = (X_1 + X_2 + \cdots + X_n) / n = n^{-1} X_1 + n^{-1} X_2 + \cdots + n^{-1} X_n$$

$$E(\bar{X}) = n^{-1}\mu + n^{-1}\mu + \cdots + n^{-1}\mu = n(n^{-1}\mu) = \mu$$

$$V(\bar{X}) = n^{-2}\sigma^2 + n^{-2}\sigma^2 + \cdots + n^{-2}\sigma^2 = n(n^{-2}\sigma^2) = \sigma^2 / n$$

# Behavior of the Sample Mean

Suppose we have $n$ independent data points $x_1, x_2, \ldots, x_n$

each with the same mean $\mu$ and variance $\sigma^2$

and if $\bar{x}$ is the sample mean of the $n$ data points then

The expected value (mean) of $\bar{x}$ is

$$E(\bar{x}) = \mu$$

The variance of $\bar{x}$ is

$$V(\bar{x}) = \sigma^2 / n$$

and the standard deviation of $\bar{x}$ is

$$SD(\bar{x}) = \sigma / \sqrt{n}$$