Continuous Random Variables

Dr.Mamatha.H.R

Professor

Department of Computer Science and Engineering

PES University

Bangalore

Course material created using various Internet resources and text book

Consider the following table of sales, divided into intervals of 1000 units each,

interval		
(0,1000]		
(1000,2000]		
(2000,3000]		
(3000,4000]		
(4000,5000]		
(5000,6000]		
(6000,7000]		

and the relative frequency of each interval.

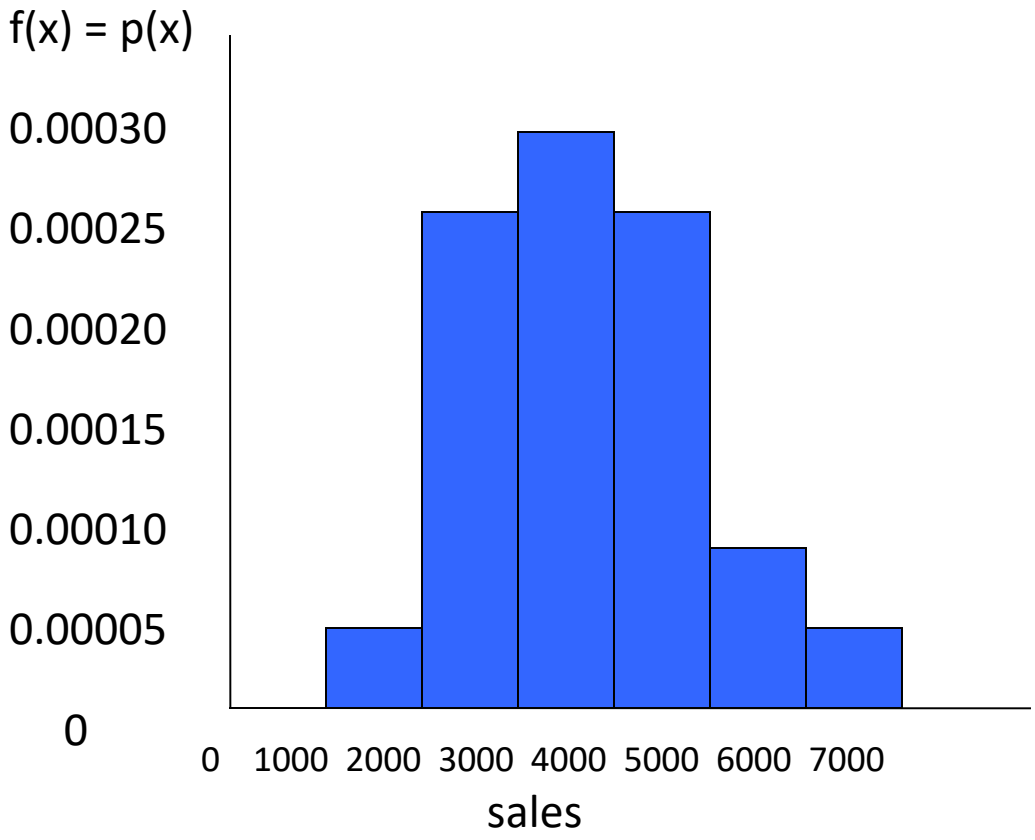
interval	relative freq.	
(0,1000]	0	
(1000,2000]	0.05	
(2000,3000]	0.25	
(3000,4000]	0.30	
(4000,5000]	0.25	
(5000,6000]	0.10	
(6000,7000]	0.05	
	1.00	

We're going to divide the relative frequencies by the width of the cells (which here is 1000). This will make the graph have an area of 1.

interval	relative freq.	$f(x) = \frac{\text{relative freq.}}{\text{cell width}}$
(0,1000]	0	0
(1000,2000]	0.05	0.00005
(2000,3000]	0.25	0.00025
(3000,4000]	0.30	0.00030
(4000,5000]	0.25	0.00025
(5000,6000]	0.10	0.00010
(6000,7000]	0.05	0.00005

Graph

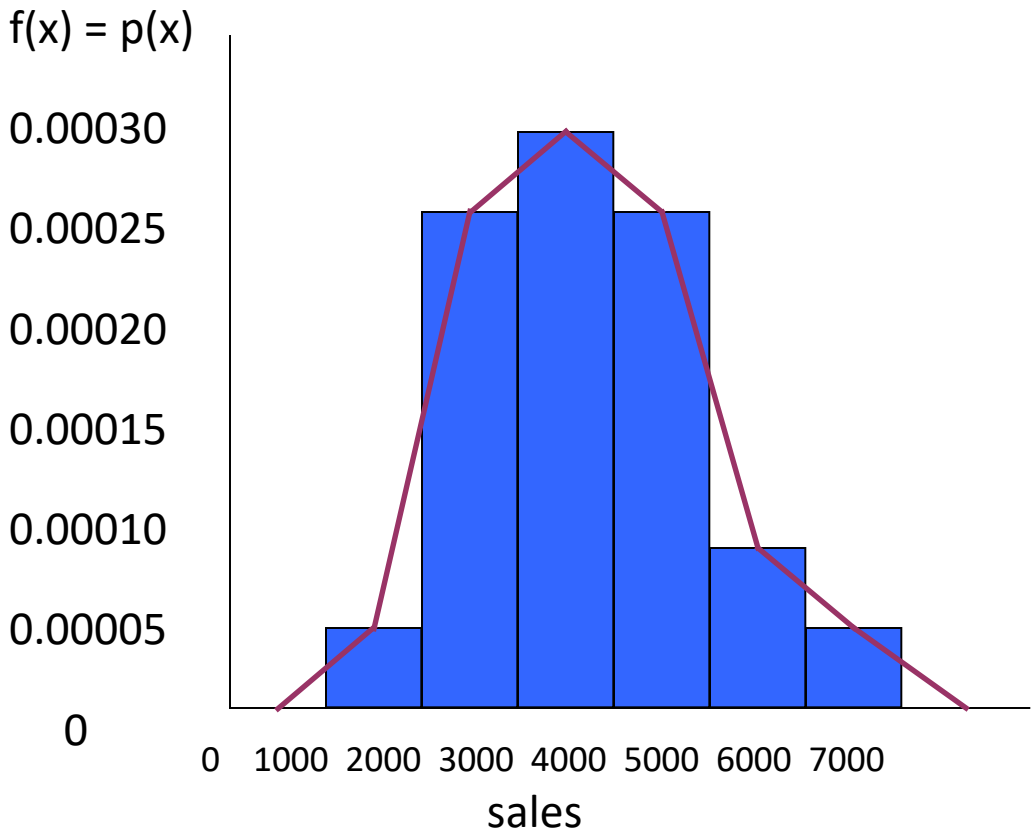
interval	$f(x) = \frac{\text{relative freq.}}{\text{cell width}}$
(0,1000]	0
(1000,2000]	0.00005
(2000,3000]	0.00025
(3000,4000]	0.00030
(4000,5000]	0.00025
(5000,6000]	0.00010
(6000,7000]	0.00005



The area of each bar is the frequency of the category, so the total area is 1.

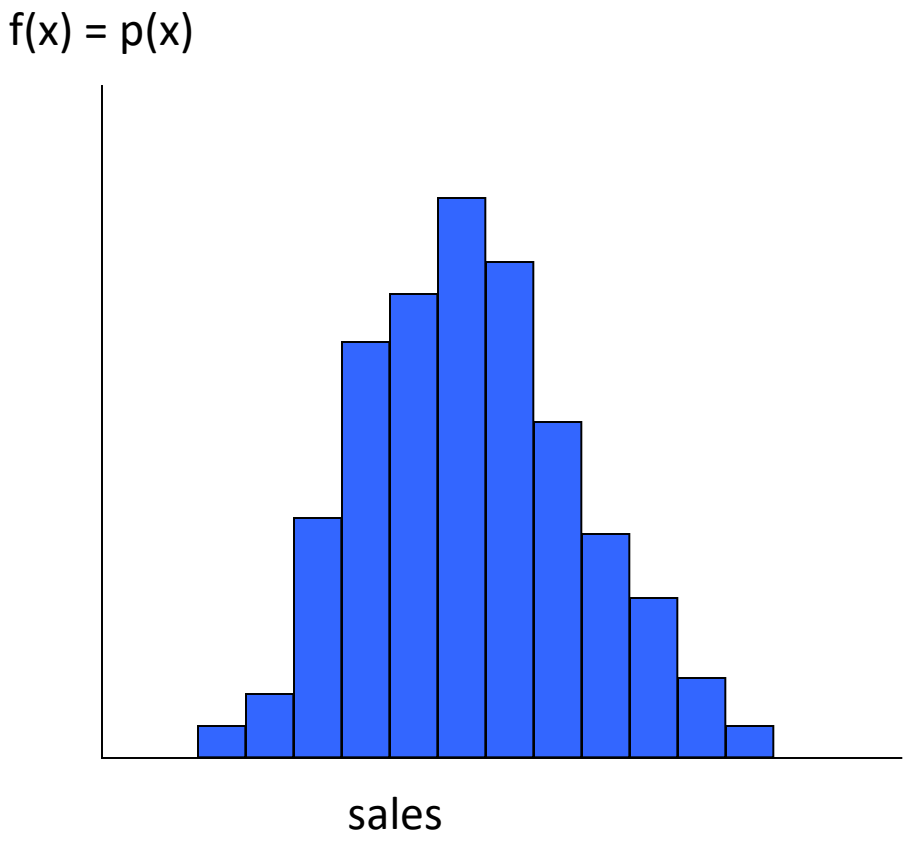
Graph

interval	$f(x) = \frac{\text{relative freq.}}{\text{cell width}}$
(0,1000]	0
(1000,2000]	0.00005
(2000,3000]	0.00025
(3000,4000]	0.00030
(4000,5000]	0.00025
(5000,6000]	0.00010
(6000,7000]	0.00005



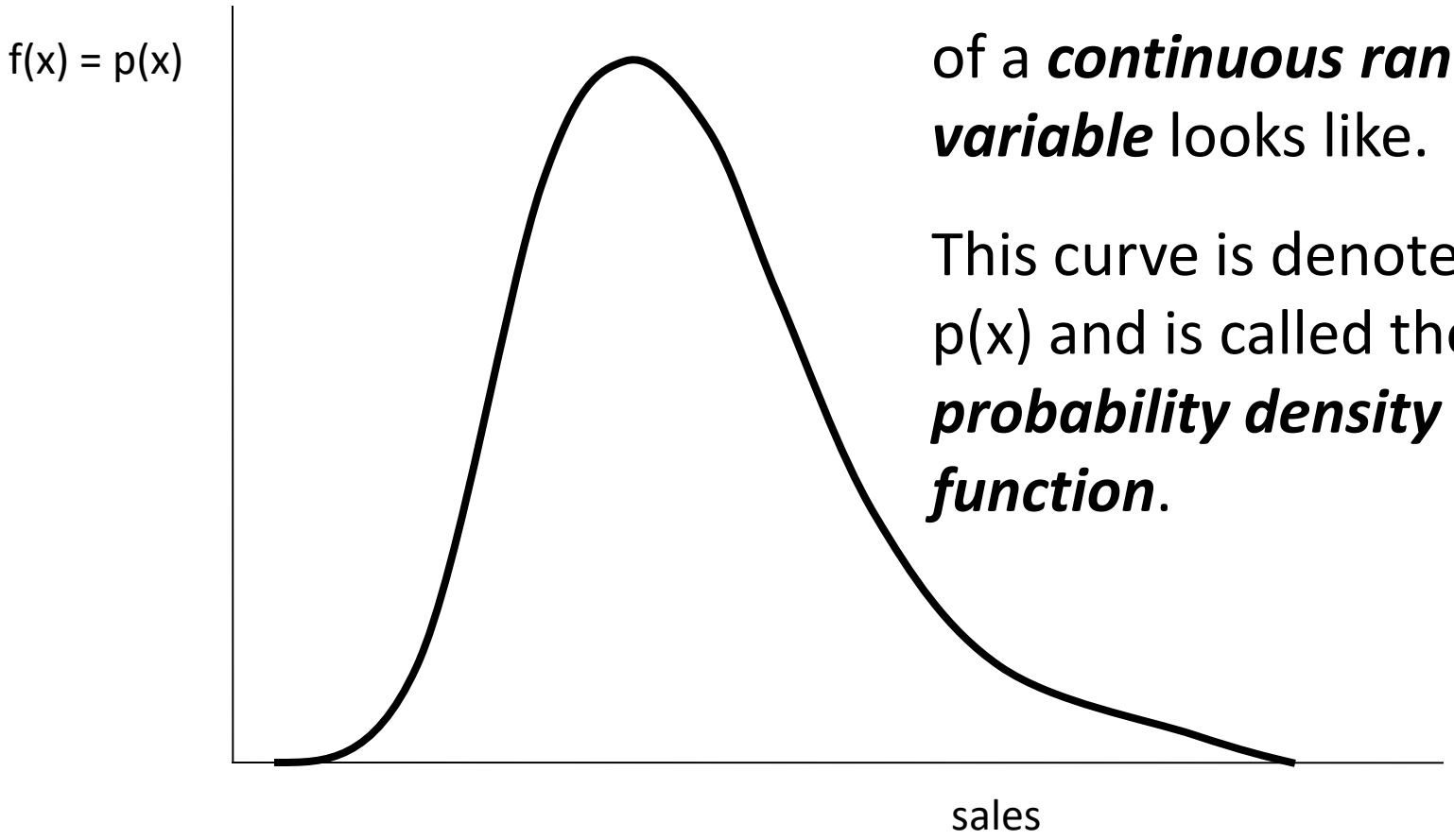
Here is the frequency polygon.

If we make the intervals 500 units instead of 1000, the graph would probably look something like this:



The height of the bars increases and decreases more gradually.

If we made the intervals infinitesimally small, the bars and the frequency polygon would become smooth, looking something like this:



This what the distribution of a ***continuous random variable*** looks like.

This curve is denoted f(x) or p(x) and is called the ***probability density function***.

Continuous Random Variables

A random variable is **continuous** if its probabilities are given by areas under a curve.

The curve is called a **probability density function** (pdf) for the random variable. Sometimes the pdf is called the **probability distribution**.

The function $f(x)$ is the probability density function of X.

Let X be a continuous random variable with probability density function $f(x)$. Then

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Computing Probabilities

Let *X* be a continuous random variable with probability density function *f*(*x*). Let *a* and *b* be any two numbers, with *a* < *b*. Then

$$P(a \le X \le b) = P(a \le X < b) = P(a < X \le b) = \int_a^b f(x)dx.$$

In addition,

$$P(X \le a) = P(X < a) = \int_{-\infty}^{a} f(x)dx$$

$$P(X \ge a) = P(X > a) = \int_a^{\infty} f(x)dx.$$

pmf versus pdf

- For a discrete random variable, we had a probability mass function (pmf).
- The pmf looked like a bunch of spikes, and probabilities were represented by the heights of the spikes.
- For a continuous random variable, we have a probability density function (pdf).
- The pdf looks like a curve, and probabilities are represented by areas under the curve.

Pr(a < X < b)

A continuous random variable has an infinite number of possible values & the probability of any one particular value is zero.

If X is a continuous random variable, which of the following probabilities is largest? (Hint: This is a trick question.)

- 1. $\Pr(a < X < b)$
- 2. $\Pr(a \leq X < b)$
- 3. $\Pr(a < X \leq b)$
- 4. $\Pr(a \leq X \leq b)$

They're all equal.

They differ only in whether they include the individual values a and b, and any one particular value has zero probability!

Properties of probability density functions (pdfs)

- 1. $f(x) \geq 0$ for values of x
- This means that when we draw the pdf curve, while it may be on the left side of the vertical axis (have negative values of x), it can not go below the horizontal axis, where f would be negative.
- $Pr(-\infty < X < \infty) = 1$
- The total area under the pdf curve, which corresponds to the total probability, is 1.

A hole is drilled in a sheet-metal component, and then a shaft is inserted through the hole. The shaft clearance is equal to the difference between the radius of the hole and the radius of the shaft. Let the random variable X denote the clearance, in millimeters. The probability density function of X is

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

Components with clearances larger than 0.8 mm must be scrapped. What proportion of components are scrapped?

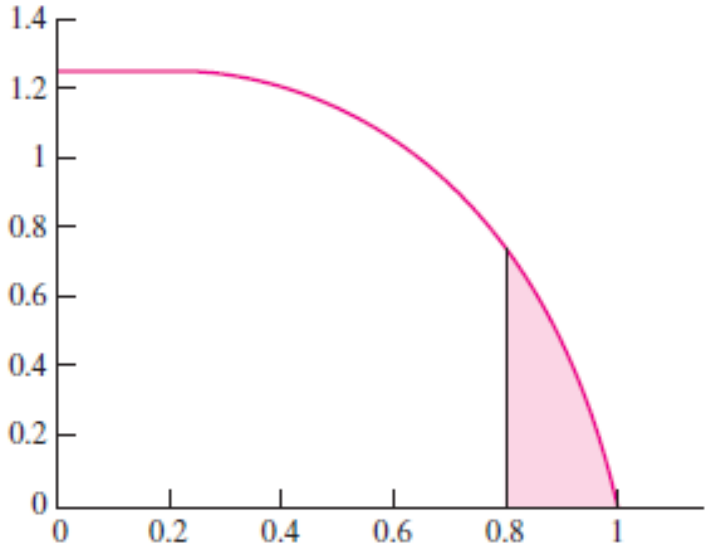


FIGURE 2.13 Graph of the probability density function of X, the clearance of a shaft. The area shaded is equal to $P(X > 0.8)$.

This area is given by

$$
\begin{aligned}
P(X > 0.8) &= \int_{0.8}^{\infty} f(x)\,dx \\
&= \int_{0.8}^{1} 1.25(1 - x^4)\,dx \\
&= 1.25\left(x - \frac{x^5}{5}\right)\Big|_{0.8}^{1} \\
&= 0.0819
\end{aligned}
$$

More on Continuous Random Variables

Let *X* be a continuous random variable with probability density function *f*(*x*). The **cumulative distribution function** of *X* is the function

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

The mean of *X* is given by

$$\mu_X = \int_{-\infty}^{\infty} xf(x)dx.$$

The variance of *X* is given by

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x-\mu_X)^2 f(x)dx$$

$$= \int_{-\infty}^{\infty} x^2 f(x)dx - \mu_X^2.$$

Find the cumulative distribution function $F(x)$ and plot it.

Solution

The probability density function of X is given by $f(t) = 0$ if $t \leq 0$, $f(t) = 1.25(1 - t^4)$ if $0 < t < 1$, and $f(t) = 0$ if $t \geq 1$. The cumulative distribution function is given by $F(x) = \int_{-\infty}^{x} f(t)\,dt$. Since $f(t)$ is defined separately on three different intervals, the computation of the cumulative distribution function involves three separate cases.

If $x \leq 0$:

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t)\,dt \\ &= \int_{-\infty}^{x} 0\,dt \\ &= 0 \end{aligned}$$

If $0 < x < 1$:

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t)\,dt \\ &= \int_{-\infty}^{0} f(t)\,dt + \int_{0}^{x} f(t)\,dt \\ &= \int_{-\infty}^{0} 0\,dt + \int_{0}^{x} 1.25(1-t^4)\,dt \\ &= 0 + 1.25\left(t - \frac{t^5}{5}\right)\Bigg|_{0}^{x} \\ &= 1.25\left(x - \frac{x^5}{5}\right) \end{aligned}$$

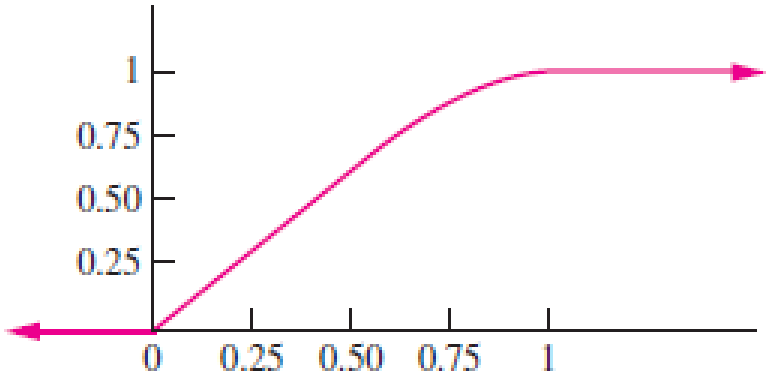
If $x > 1$:

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t)\,dt \\ &= \int_{-\infty}^{0} f(t)\,dt + \int_{0}^{1} f(t)\,dt + \int_{1}^{x} f(t)\,dt \\ &= \int_{-\infty}^{0} 0\,dt + \int_{0}^{1} 1.25(1-t^4)\,dt + \int_{1}^{x} 0\,dt \\ &= 0 + 1.25\left(t - \frac{t^5}{5}\right)\Bigg|_{0}^{1} + 0 \\ &= 0 + 1 + 0 \\ &= 1 \end{aligned}$$

Therefore

$$F(x) = \begin{cases} 0 & x \le 0 \\ 1.25\left(x - \dfrac{x^5}{5}\right) & 0 < x < 1 \\ 1 & x \ge 1 \end{cases}$$

A plot of $F(x)$ is presented here.



Definition

Let X be a continuous random variable with probability density function $f(x)$. Then the mean of X is given by

$$\mu_X = \int_{-\infty}^{\infty} x f(x)\, dx \tag{2.35}$$

The mean of X is sometimes called the expectation, or expected value, of X and may also be denoted by $E(X)$ or by μ.

Definition

Let X be a continuous random variable with probability density function $f(x)$. Then

- The variance of X is given by

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)\, dx \tag{2.36}$$

- An alternate formula for the variance is given by

$$\sigma_X^2 = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu_X^2 \tag{2.37}$$

- The variance of X may also be denoted by $V(X)$ or by σ^2.
- The standard deviation is the square root of the variance: $\sigma_X = \sqrt{\sigma_X^2}$.

Find the mean clearance and the variance of the clearance.

Solution

Using Equation (2.35), the mean clearance is given by

$$\mu_X = \int_{-\infty}^{\infty} x f(x)\, dx$$

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

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

$$= 0.4167$$

$$\sigma_X^2 = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu_X^2$$

$$= \int_0^1 x^2[1.25(1 - x^4)]\, dx - (0.4167)^2$$

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

$$= 0.0645$$

- **Chebyshev's Inequality**
- The mean of a random variable is a measure of the center of its distribution, and the standard deviation is a measure of the spread.
- Chebyshev's inequality relates the mean and the standard deviation by providing a bound on the probability that a random variable takes on a value that differs from its mean by more than a given multiple of its standard deviation.
- Specifically, the probability that a random variable differs from its mean by k standard deviations or more is never greater than $1/k^2$.

Chebyshev's Inequality

Let X be a random variable with mean μ_X and standard deviation σ_X. Then

$$P(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}$$

Only the case $k > 1$ is useful. When $k \leq 1$ the right hand $1/k^2 \geq 1$ and the inequality is trivial as all probabilities are ≤ 1.

The length of a rivet manufactured by a certain process has mean $\mu X = 50$ *mm* and standard deviation $\sigma X = 0.45$ *mm. What is the largest possible value for the* probability that the length of the rivet is outside the interval 49.1–50.9 mm?

The length of a rivet manufactured by a certain process has mean $\mu_X = 50$ mm and standard deviation $\sigma_X = 0.45$ mm. What is the largest possible value for the probability that the length of the rivet is outside the interval 49.1–50.9 mm?

Solution

Let X denote the length of a randomly sampled rivet. We must find $P(X \leq 49.1 \text{ or } X \geq 50.9)$. Now

$$P(X \leq 49.1 \text{ or } X \geq 50.9) = P(|X - 50| \geq 0.9) = P(|X - \mu_X| \geq 2\sigma_X)$$

Applying Chebyshev's inequality with $k = 2$, we conclude that

$$P(X \leq 49.1 \text{ or } X \geq 50.9) \leq \frac{1}{4}$$

Assume that the probability density function for X, the length of a rivet in Example 2.46, is

$$f_X(x) = \begin{cases} [477 - 471(x - 50)^2]/640 & 49 \le x \le 51 \\ 0 & \text{otherwise} \end{cases}$$

It can be verified that $\mu_X = 50$ and $\sigma_X = 0.45$. Compute the probability that the length of the rivet is outside the interval 49.1–50.9 mm. How close is this probability to the Chebyshev bound of 1/4?

Solution

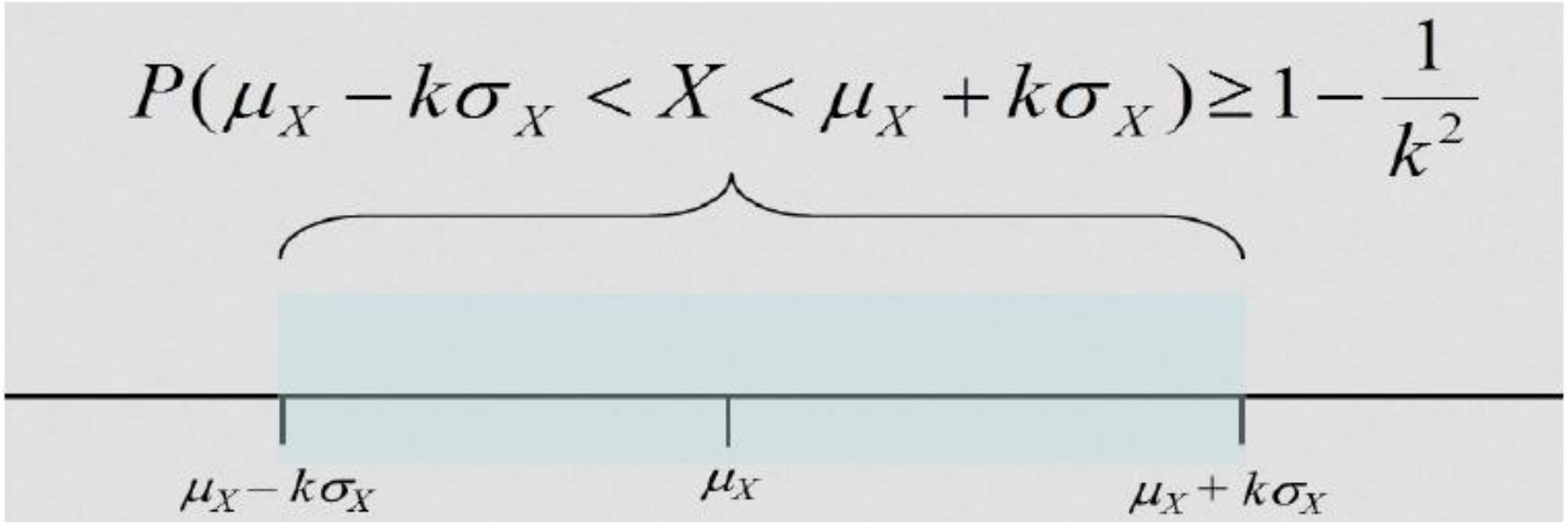
$$
\begin{aligned}
P(X \le 49.1 \text{ or } X \ge 50.9) &= 1 - P(49.1 < X < 50.9) \\
&= 1 - \int_{49.1}^{50.9} \frac{477 - 471(x - 50)^2}{640}\, dx \\
&= 1 - \frac{477x - 157(x - 50)^3}{640}\Bigg|_{49.1}^{50.9} \\
&= 0.01610
\end{aligned}
$$

The actual probability is much smaller than the Chebyshev bound of 1/4.

- Because the Chebyshev bound is generally much larger than the actual probability, it should only be used when the distribution of the random variable is unknown.
- When the distribution is known, then the probability density function or probability mass function should be used to compute probabilities.

Statement of Chebyshev's Inequality

Chebyshev's inequality states that at least $1-1/K^2$ of data from a sample must fall within K standard deviations from the mean, where K is any positive real number greater than one.



To illustrate the inequality, we will look at it for a few values of K:

For K = 2 we have $1-1/K^2 = 1 - 1/4 = 3/4 = 75\%$. So Chebyshev's inequality says that at least 75% of the data values of any distribution must be within two standard deviations of the mean.

For K = 3 we have $1 – 1/K^2 = 1 - 1/9 = 8/9 = 89\%$. So Chebyshev's inequality says that at least 89% of the data values of any distribution must be within three standard deviations of the mean.

The Population Median and Percentiles

- the sample median is the point that divides the sample in half.

The population median is defined analogously.

In terms of the probability density function,the median is the point at which half the area under the curve is to the left, and half the area is to the right.

- Thus if *X is a continuous random variable with probability density* function *f (x), the median of X is the point* x_m *that solves the equation*
- $P(X \leq x_m) = \int_{-\infty}^{x_m} f(x)\,dx = 0.5.$

Definition

Let X be a continuous random variable with probability mass function $f(x)$ and cumulative distribution function $F(x)$.

- The median of X is the point x_m that solves the equation $F(x_m) = P(X \leq x_m) = \int_{-\infty}^{x_m} f(x)\,dx = 0.5$.
- If p is any number between 0 and 100, the pth percentile is the point x_p that solves the equation $F(x_p) = P(X \leq x_p) = \int_{-\infty}^{x_p} f(x)\,dx = p/100$.
- The median is the 50th percentile.

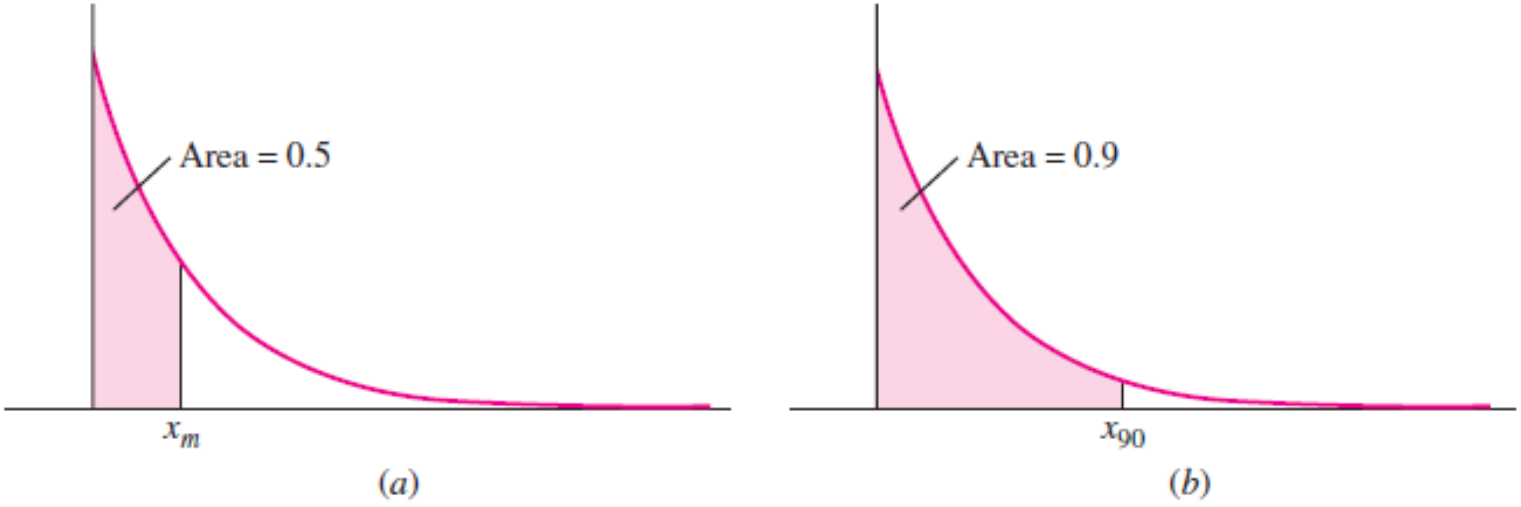


FIGURE 2.14 *(a)* Half of the population values are less than the median x_m. *(b)* Ninety percent of the population values are less than the 90th percentile x_{90}.

A certain radioactive mass emits alpha particles from time to time. The time between emissions, in seconds, is random, with probability density function

$$f(x) = \begin{cases} 0.1e^{-0.1x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Find the median time between emissions. Find the 60th percentile of the times.

Solution

The median x_m is the solution to $\int_{-\infty}^{x_m} f(x)\,dx = 0.5$. We therefore must solve

$$\int_0^{x_m} 0.1e^{-0.1x}\,dx = 0.5$$

$$-e^{-0.1x}\Big|_0^{x_m} = 0.5$$

$$1 - e^{-0.1x_m} = 0.5$$

$$e^{-0.1x_m} = 0.5$$

$$-0.1x_m = \ln 0.5$$

$$0.1x_m = 0.6931$$

$$x_m = 6.931$$

Half of the times between emissions are less than 6.931 s, and half are greater.

The 60th percentile x_{60} is the solution to $\int_{-\infty}^{x_{60}} f(x)\,dx = 0.6$. We proceed as before, substituting x_{60} for x_m, and 0.6 for 0.5. We obtain

$$1 - e^{-0.1x_{60}} = 0.6$$
$$e^{-0.1x_{60}} = 0.4$$
$$-0.1x_{60} = \ln 0.4$$

$$0.1x_{60} = 0.9163$$
$$x_{60} = 9.163$$

Sixty percent of the times between emissions are less than 9.163 s, and 40% are greater.