Introductory Statistics: A Problem-Solving Approach

by Stephen Kokoska

Chapter 6

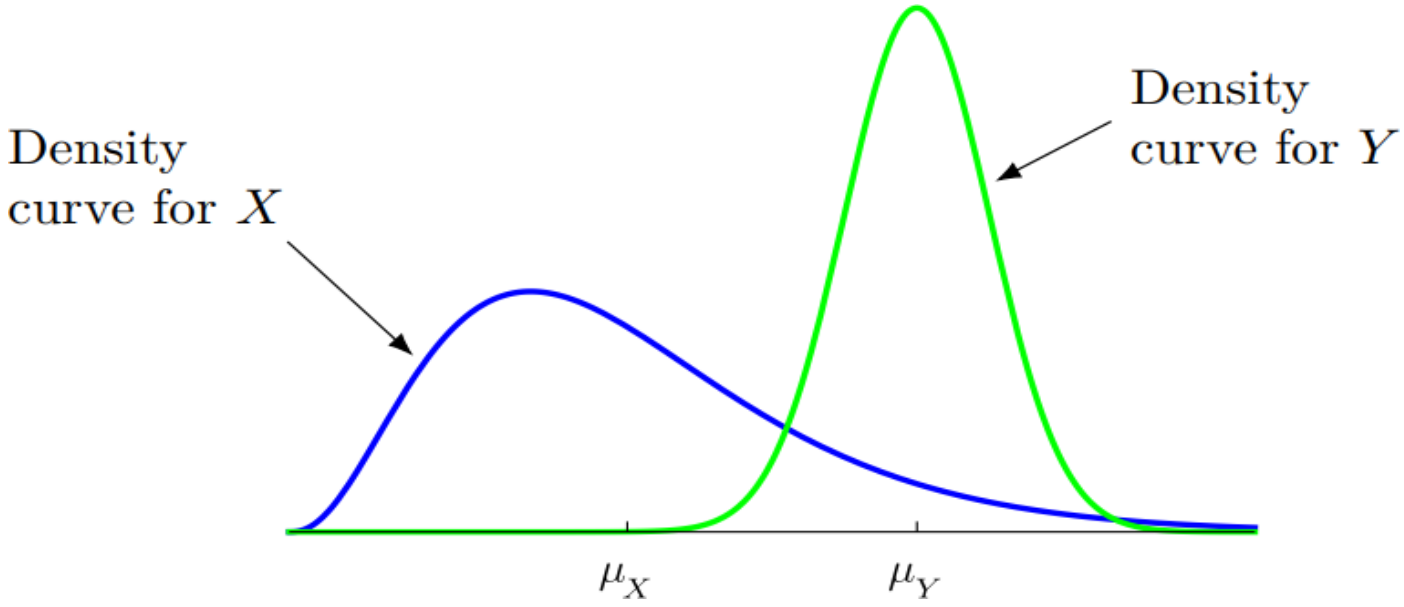
Continuous Probability Distributions

Probability Density Functions (pdf)

Definition

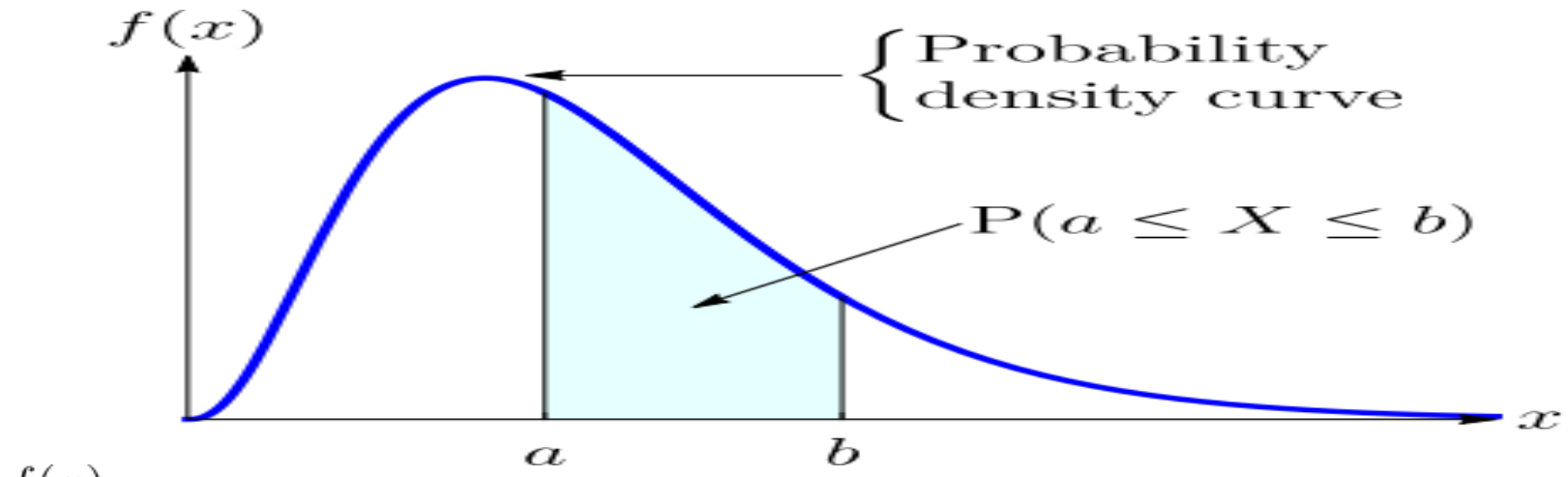
A **probability distribution for a continuous random variable X** is given by a smooth curve called a **density curve**, or **probability density function** (pdf). The curve is defined so that the probability that X takes on a value between a and b $(a < b)$ is the area under the curve between a and b.

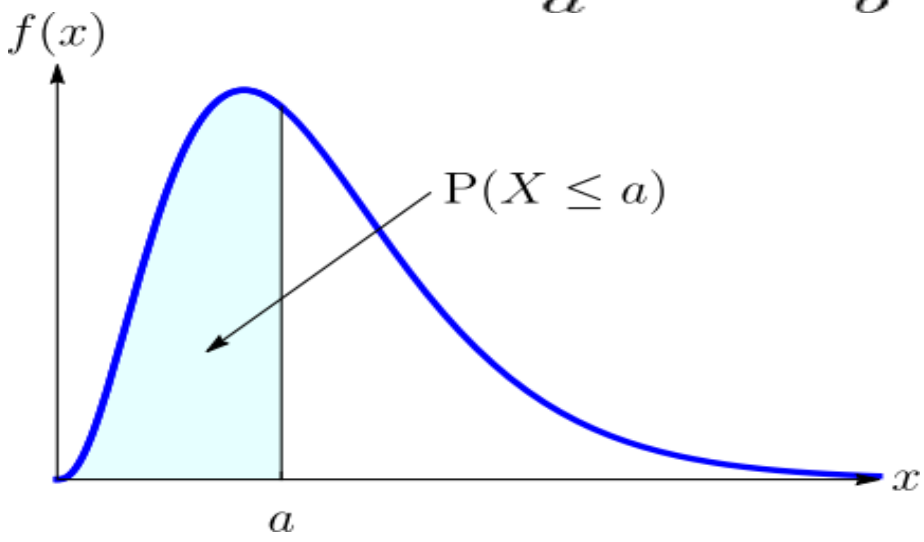
Examples of Density Curves

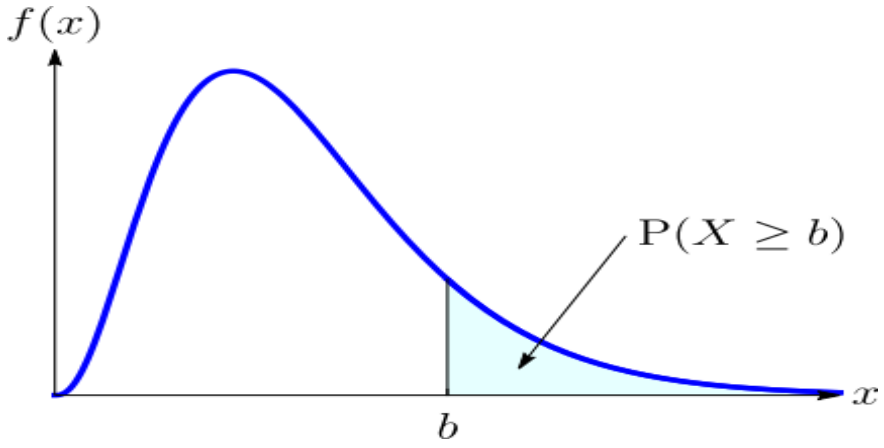


Probability Density Functions (pdf)

Probability in a continuous world is area under a curve. The following figures illustrate the correspondence between the probability of an event (defined in terms of a continuous random variable) and the area under the density curve.







Uniform Distribution

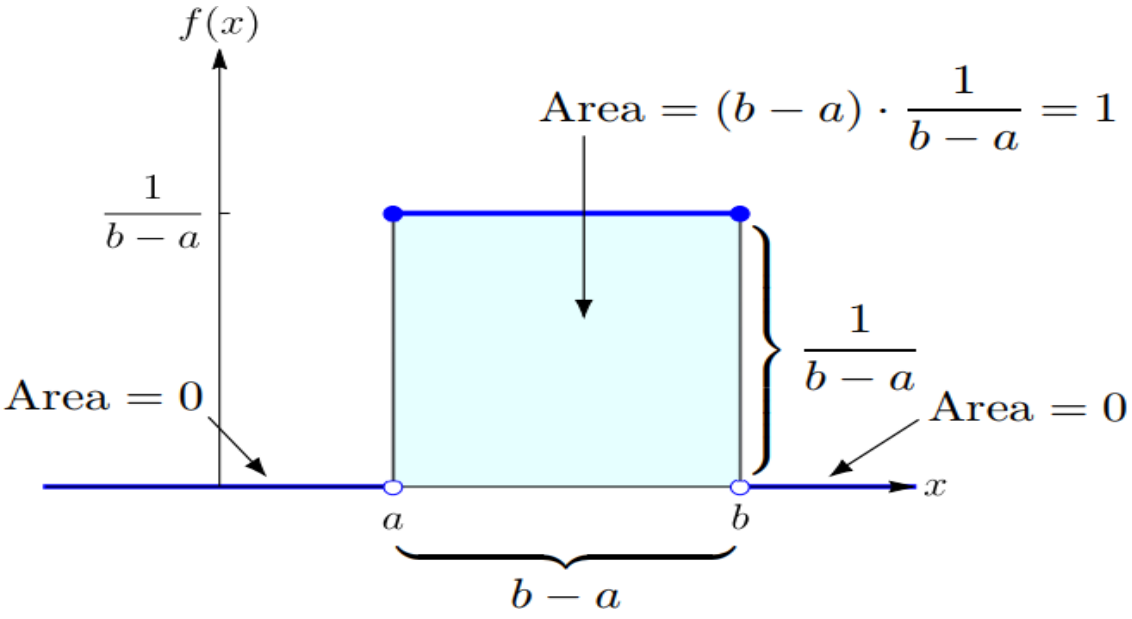
Definition X~U(a, b)

The random variable X has a **uniform distribution** on the interval $[a, b]$ if

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases} \qquad -\infty < a < b < \infty \tag{6.2}$$

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12} \tag{6.3}$$

Area Under Uniform Probability Density Function = 1



Example 6.1 Reef Dives

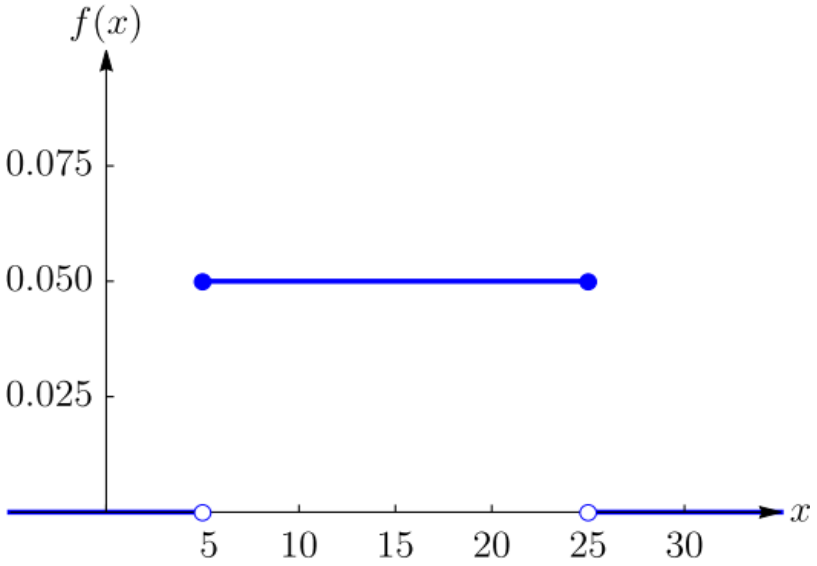
Bonaire, an island in the Dutch Caribbean, is considered one of the top 10 diving destinations in the world. A careful examination of boat records has shown that the time it takes to reach a randomly selected dive site has a uniform distribution between 5 and 25 minutes. Suppose a dive site is selected at random.

(a) Carefully sketch a graph of the probability density function.
(b) Find the probability that it takes at most 10 minutes to reach the dive site.
(c) Find the probability it takes between 10 and 20 minutes to reach the dive site.
(d) Find the mean time it takes to reach a dive site, as well as the variance and standard deviation.

Let X be the time it takes to reach a dive site. The random variable X has a uniform distribution between the times 5 and 25. X ~ U(5, 25)

Its probability density function is

$$f(x)=\begin{cases}\dfrac{1}{25-5}=0.05 & \text{if } 5\le x\le 25\\ 0 & \text{otherwise}\end{cases}$$



Example 6.1 Reef Dives

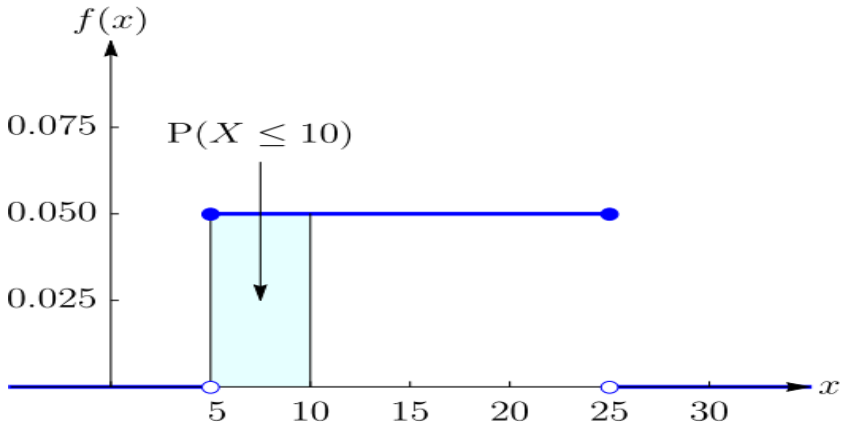
(b) Find the probability that it takes at most 10 minutes to reach the dive site.

$P(X \le 10) = P(5 \le X \le 10)$

$= \text{area of rectangle}$

$= 5(0.05) = 0.25$

The probability it takes at most 10 minutes to reach the dive site is 0.25.

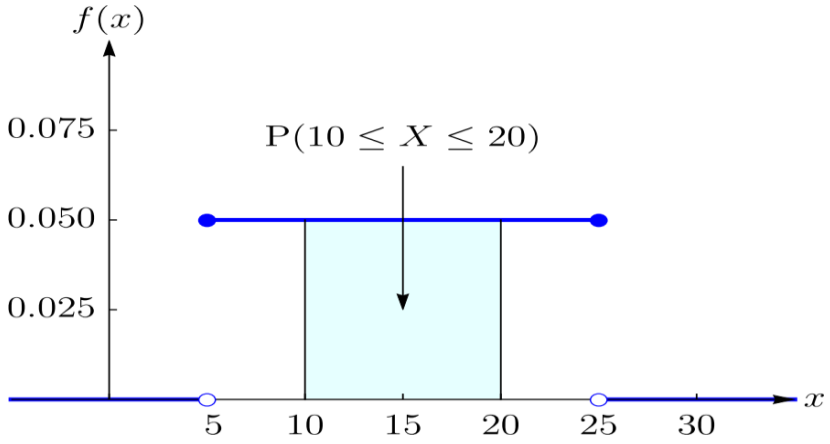


(c) Find the probability it takes between 10 and 20 minutes to reach the dive site

$P(10 \le X \le 20) = \text{area of rectangle}$

$= (10)(0.05) = 0.50$

Thus, the probability that it takes between 10 and 20 minutes to reach a dive site is 0.50.



(d) $\mu = \dfrac{5+25}{2} = 15$ $\sigma^2 = \dfrac{(25-5)^2}{12} \approx 33.33$

Normal Probability Distribution

The Normal Probability Distribution

$X \sim N(\mu, \sigma^2)$

Suppose X is a normal random variable with mean μ and variance σ^2. The probability density function is given by

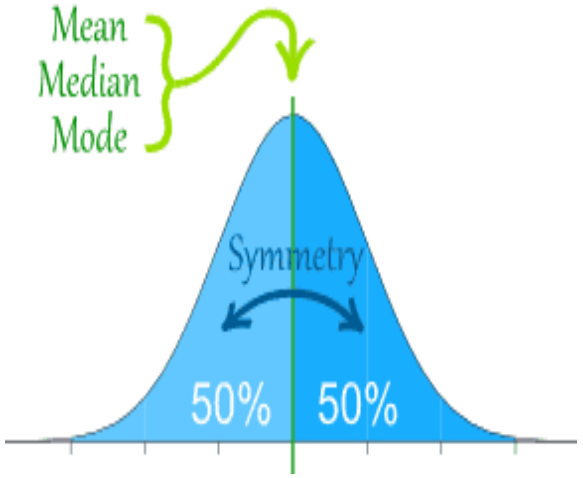
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad (6.5)$$

and

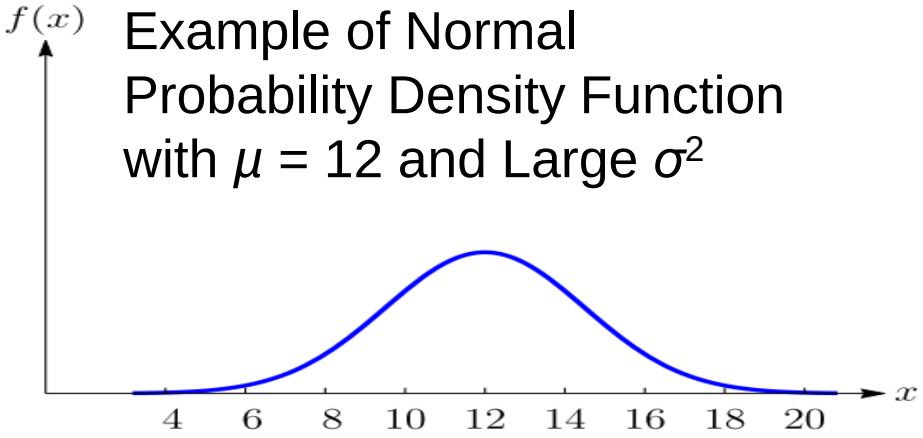
$$-\infty < x < \infty \qquad -\infty < \mu < \infty \qquad \sigma^2 > 0 \qquad (6.6)$$

Properties:

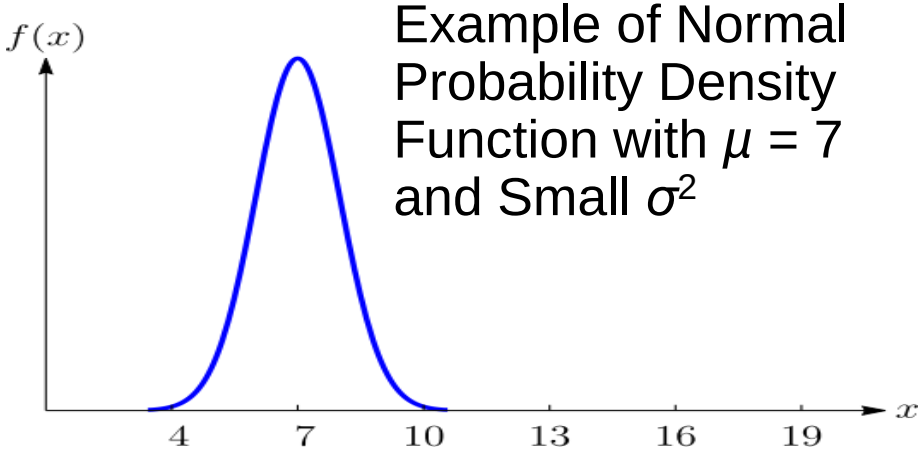
- Very common and the most important distribution in statistics.
- Used to model many natural phenomena and extensively in inference.
- Completely characterized by its mean μ and variance σ^2.
- The mean, mode and median are all equal.
- The curve is symmetric at the center (around μ). Exactly half of the values are to the left and to the right of center.
- The total area under the curve is 1.
- It is often called a "Bell Curve“ because it looks like a bell.



Normal Probability Distribution

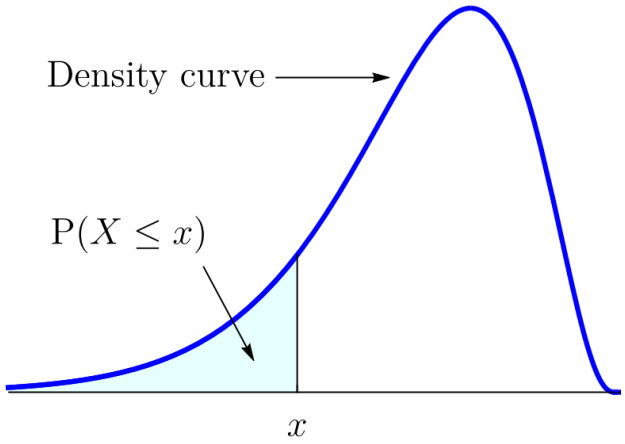


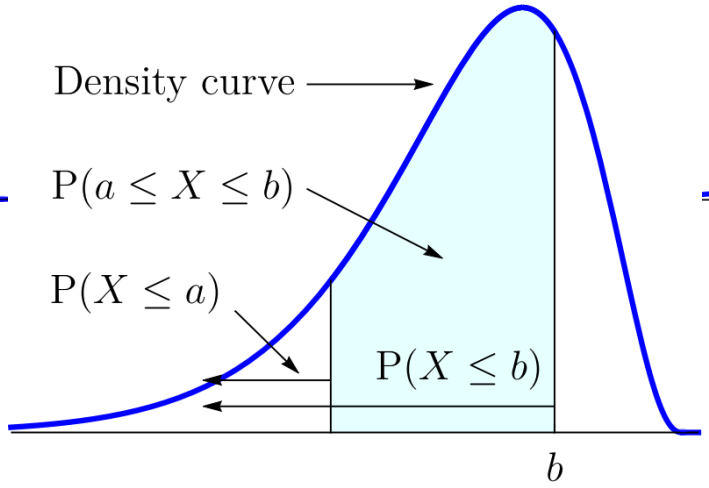
Example of Normal Probability Density Function with $\mu = 12$ and Large σ^2

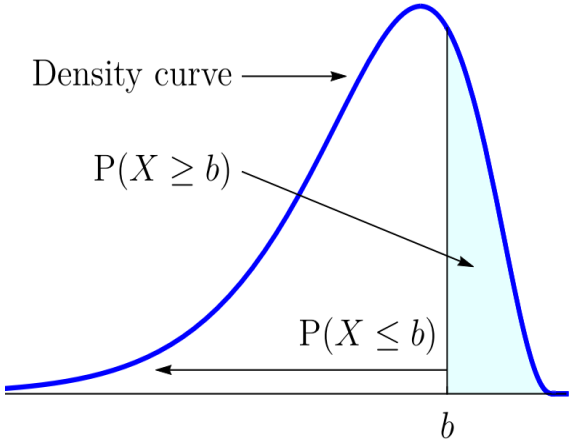


Example of Normal Probability Density Function with $\mu = 7$ and Small σ^2

Finding the required shaded area:







Use the Complement Rule

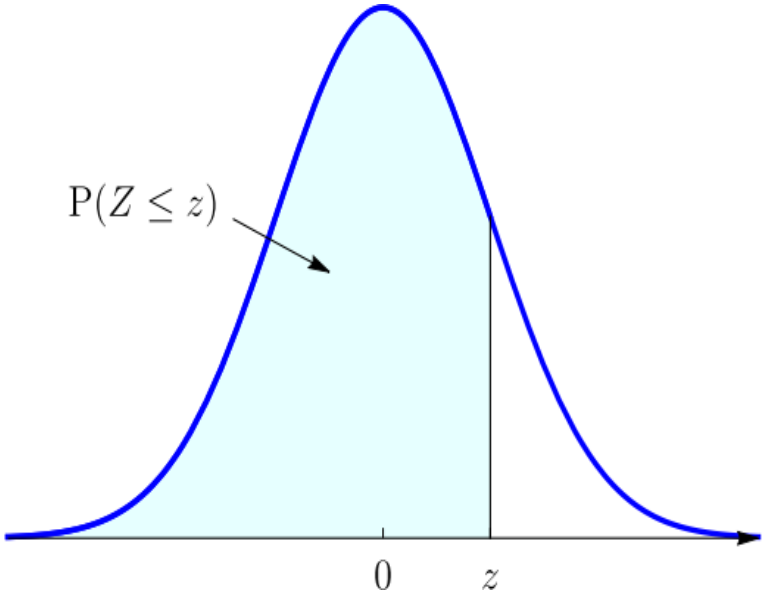
Standard Normal Distribution

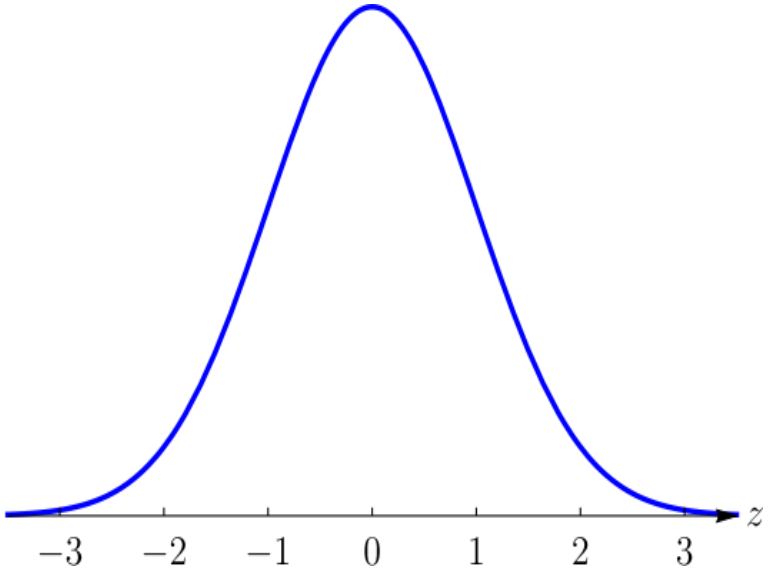
The Standard Normal Random Variable

Z~N(0, 1)

The normal distribution with $\mu = 0$ and $\sigma^2 = 1$ (and $\sigma = 1$) is called the **standard normal distribution**. A random variable that has a standard normal distribution is called a **standard normal random variable**, usually denoted Z. The probability density function for Z is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad -\infty < z < \infty \tag{6.7}$$

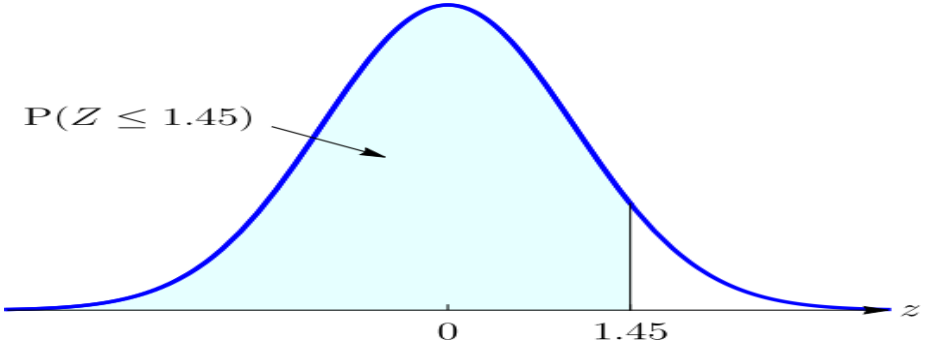




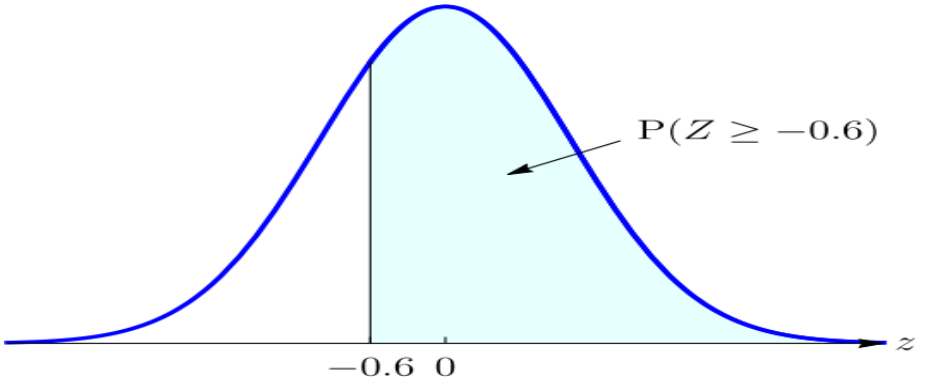
Standard Normal Distribution

Example 6.4: Use Table 3 in the Appendix to find each of the following probabilities associated with the standard normal distribution.

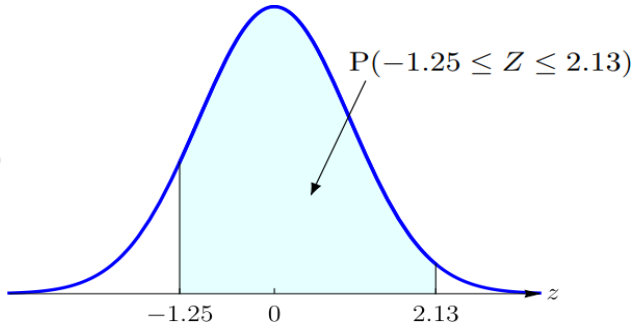
(a) $P(Z \le 1.45) = 0.9265$



(b) $P(Z \ge -0.6) = 1 - P(Z < -0.6)$
$= 1 - P(Z \le -0.6)$
$= 1 - 0.2743 = 0.7257$



(c) $P(-1.25 \le Z \le 2.13) = P(Z \le 2.13) - P(Z < -1.25)$
$= P(Z \le 2.13) - P(Z \le 21.25)$
$= 0.9834 - 0.1056 = 0.8778$

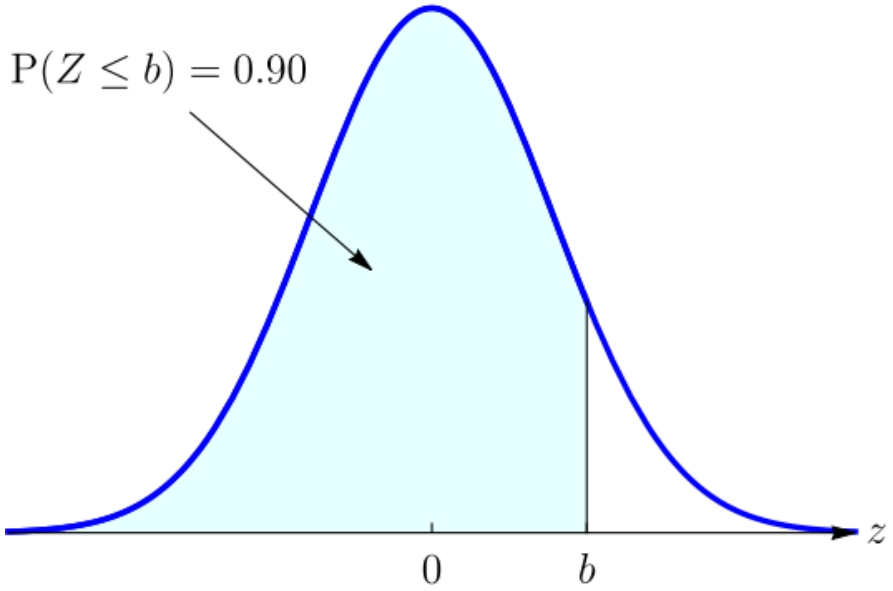


Standard Normal Distribution

(d) Find the value b such that $P(Z \leq b) = 0.90$

The bottom 90% of observations are at or less than how much?

- Search the body of Table 3 in the Appendix to find a cumulative probability as close to 0.90 as possible.
- Read the row and column entries to find b.
- In the body of Table 3, the closest cumulative probability to 0.90 is 0.8997. This corresponds to 1.28.



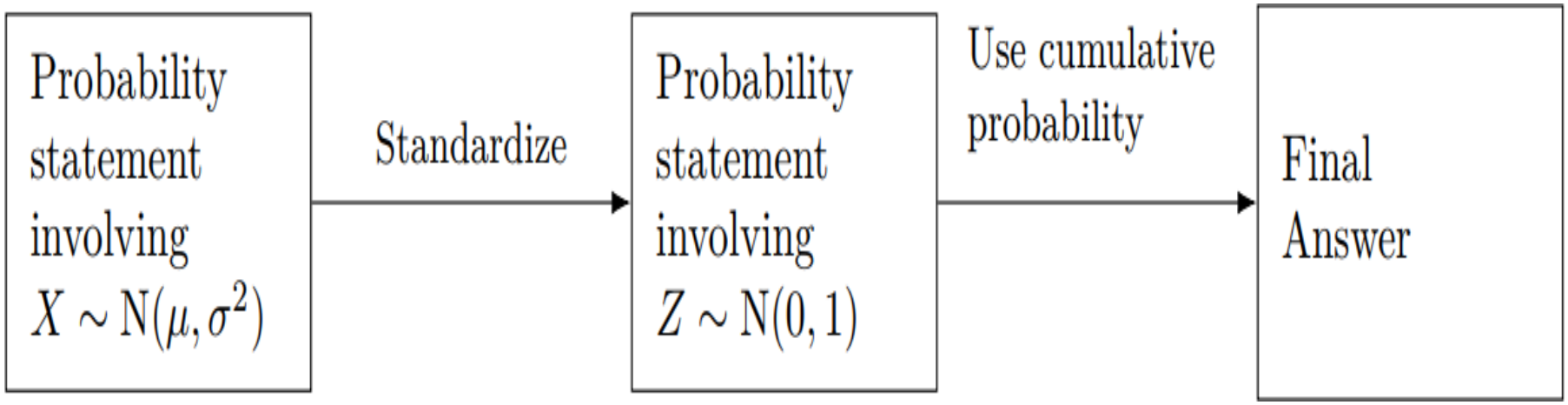
Standardization Rule

Standardization Rule

If X is a normal random variable with mean μ and variance σ^2, then a standard normal random variable is given by

$$Z = \frac{X - \mu}{\sigma} \qquad (6.8)$$

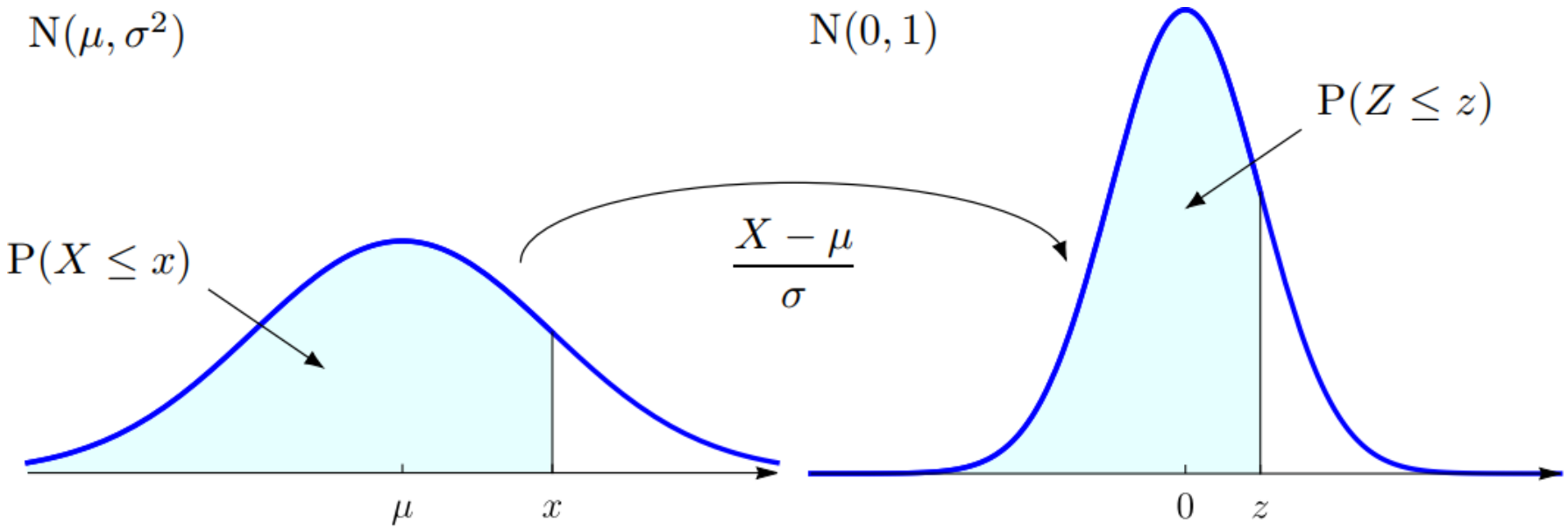
Computing Normal Probabilities:



Standardization Rule

An Illustration of Standardization:

The Areas of the Shaded Regions Are Equal



Example 6.6

Seat pitch (leg room) on a passenger airline is the distance from the back of one seat to the front of the one directly behind it. The greater the leg room, the more comfortable the seat and the less likely you are to travel with your knees against your chest. Some passengers pay a premium for extra leg room rather than sit in a regular coach seat. Suppose the leg room for all coach seats is normally distributed, with mean 31 in. and standard deviation 0.5 in. For a randomly selected coach seat, find the probability that

a. the seat pitch is between 30.5 and 32 in. (considered barely comfortable).
b. a randomly selected coach seat is constricted (any seat pitch less than 30 in. is considered constricted).

$$X \sim N(31, 0.5)$$

TomCarpenter/Shutterstock

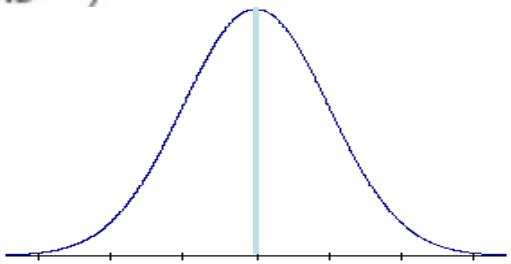
$$P(30.5 \le X \le 32)$$

$$= P\left(\frac{30.5-31}{0.5} \le \frac{X-31}{0.5} \le \frac{32-31}{0.5}\right)$$

$$= P(-1.00 \le Z \le 2.00)$$

$$= P(Z \le 2.00) - P(Z \le -1.00)$$

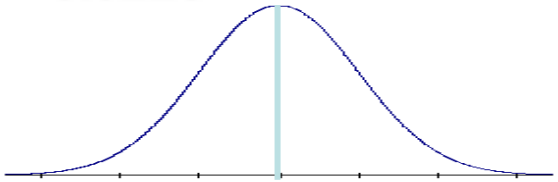
$$= 0.9772 - 0.1587 = 0.8185$$



$$P(X < 30) = P\left(\frac{X-31}{0.5} < \frac{30-31}{0.5}\right)$$

$$= P(Z < -2.00)$$

$$= 0.0228$$



Finding Observed Value Under Normal Curve (Inverse Approach)

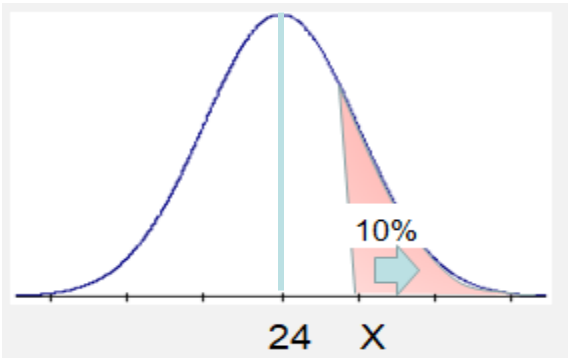
Example: The hourly wages of company employees are normally distributed with the mean of \$24 and a standard deviation of \$3.50.

a. How much the top 10% of employees at least make in hourly wages?

b. How much the bottom 10% of employees at most make in hourly wages?

a. $$P(X \geq b) = P\left(\frac{X-24}{3.5} \geq \frac{b-24}{3.5}\right)$$

$$= P\left(Z \geq \frac{b-24}{3.5}\right) = 1 - P\left(Z \leq \frac{b-24}{3.5}\right)$$



$$1 - P\left(Z \leq \frac{b-24}{3.5}\right) = 0.10 \Rightarrow P\left(Z \leq \frac{b-24}{3.5}\right) = 0.90$$

The top 10% of employees make at least \$28.48 in hourly wages.

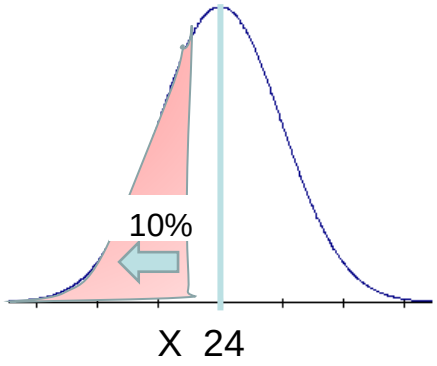
$$1.28 = \frac{b-24}{3.5}$$

$b = 24 + 1.28(3.5) = 28.48$

$$P(X \leq b) = P\left(\frac{X-24}{3.5} \leq \frac{b-24}{3.5}\right)$$

$$-1.28 = \frac{b-24}{3.5}$$

$$= P\left(Z \leq \frac{b-24}{3.5}\right) = 0.10$$



Checking the Normality Assumption

- Many statistical techniques are valid only if the observations are from a normal distribution.
- If an inference procedure requires normality, and the population distribution is not normal, then the conclusions are worthless.
- Therefore, it seems reasonable to check for normality, to make sure there is no evidence to refute this assumption.

Methods

(1) Graphs

(2) Backward Empirical Rule

(3) IQR/*s*

(4) Normal probability plot

(1) Graphs

- Construct a histogram, stem-and-leaf plot, and/or dot plot.
- Examine the shape of the distribution for indications that the distribution is not bell-shaped and symmetric.
- In a random sample, the sample distribution should be similar to the population distribution.

Checking the Normality Assumption

(2) Backward Empirical Rule

- Find $\bar{x}$ and s, and the intervals $(\bar{x} - ks, \bar{x} + ks)$, for $k = 1, 2, 3$.
- Compute the actual proportion of observations in each interval.
- If the actual proportions are close to 0.68, 0.95, 0.997, then normality seems reasonable. Otherwise, there is evidence to suggest that the shape of the distribution is not normal.

(3) IQR/*s*

- Find the interquartile range (IQR) divided by the sample standard deviation (s).
- This ratio should be close to 1.3 if the distribution is approximately normal. Otherwise, there is evidence to suggest that the shape of the distribution is not normal. Consider a standard normal random variable, $Z(\mu = 0, \sigma = 1)$. The Q_1 for Z is -0.6745 and the Q_3 is 0.6745. The IQR divided by s is $[0.6745-(-0.6745)]/1 = 1.349$.

$P(Z \le -0.6745) = 0.25$, and
$P(Z \le 0.6745) = 0.75$

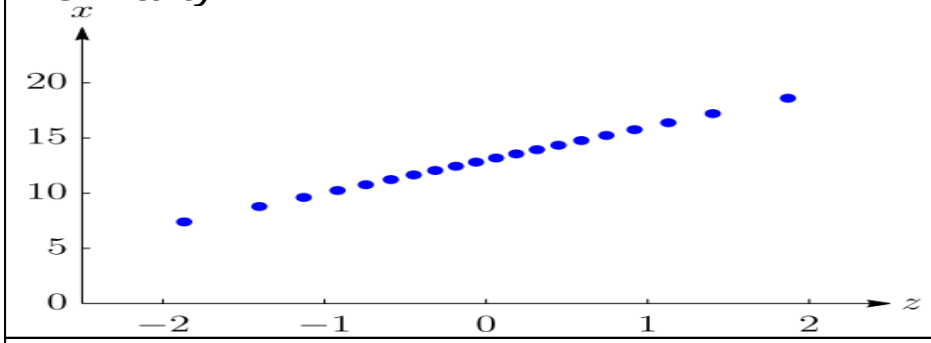
(4) Normal Probability Plot

- A scatter plot of each observation versus its corresponding standardized normal score.
- The points will fall along a straight line if the sample distribution is normal.

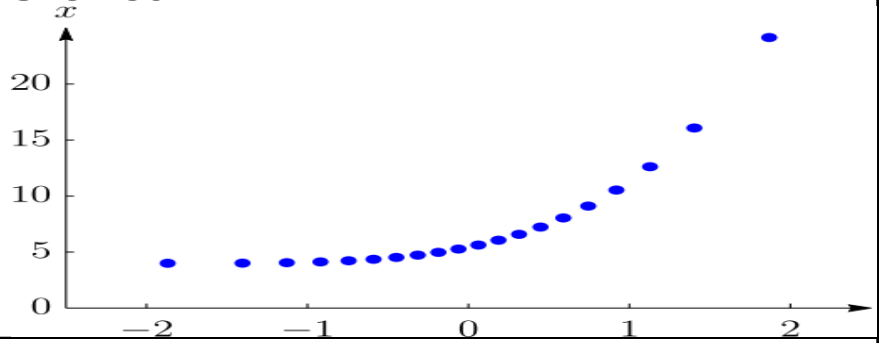
Constructing a Normal Probability Plot

- Suppose $x_1, x_2, \ldots, x_n$ is a set of observations.
- Order the observations from smallest to largest, and let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ represent the set of ordered observations.
- Find the standardized normal scores for a sample of size *n* in Appendix Table 4, $z_1, z_2, \ldots, z_n$.
- Plot the ordered pairs $(z_i, x_{(i)})$.

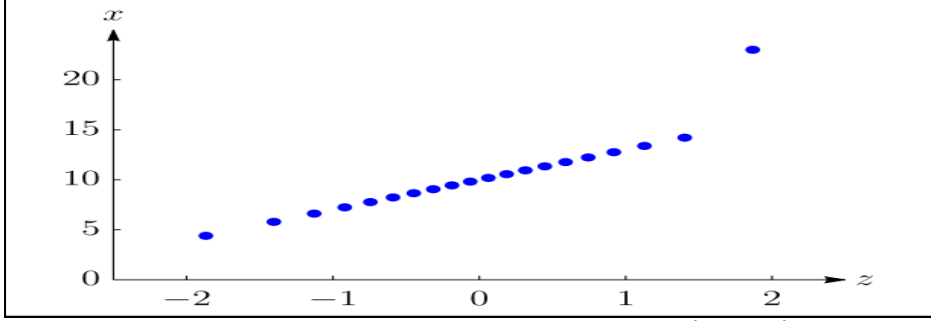
Here the points lie along an approximate straight line. There is no evidence of non-normality.



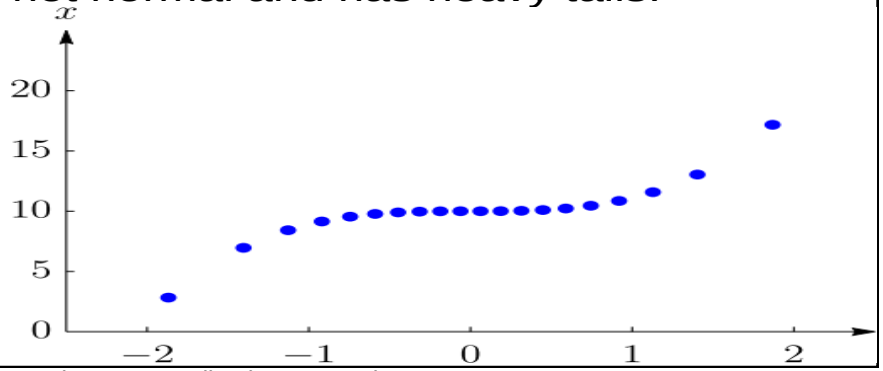
The curved graph suggests that the distribution is not normal and is positive skewed.



The plot suggests that the distribution is not normal and that the data set contains an outlier.



The plot suggests that the distribution is not normal and has heavy tails.



Example 6.9: Chemotherapy Protocol

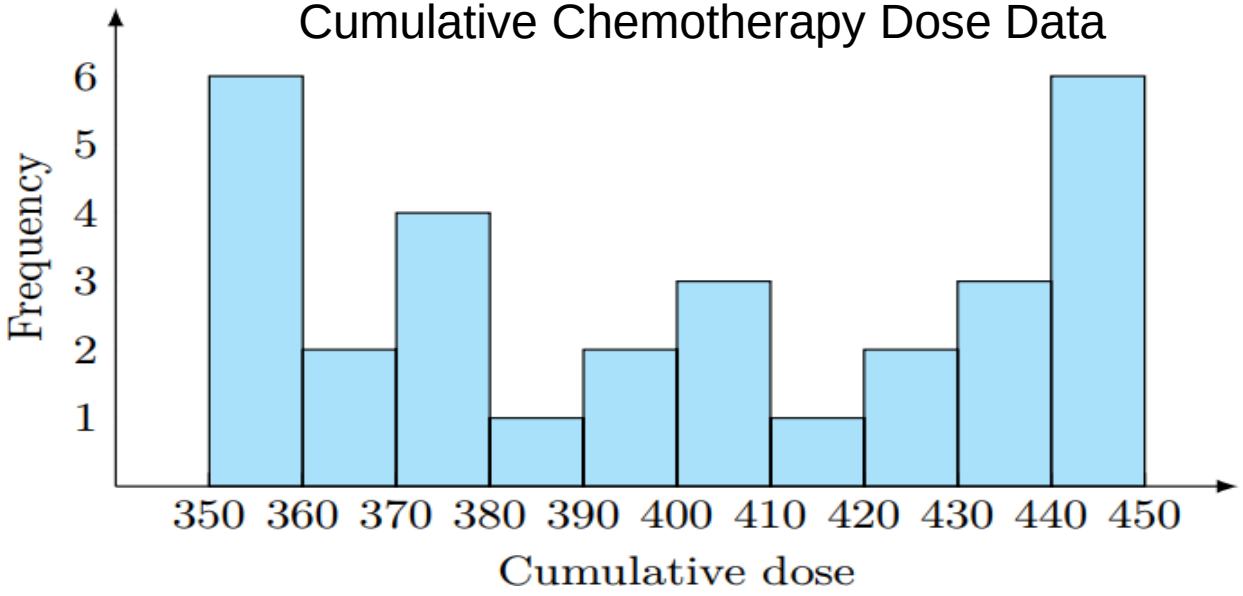
A certain protocol for chemotherapy states that the total dose for patients younger than age 12 is no greater than 450 mg/m^2 within six months. A random sample of 30 patients undergoing this form of chemotherapy was obtained, and their medical records were examined to determine the total dose of the drug they received over the previous six months (Source: National Cancer Institute clinical trials).

Is there any evidence to suggest that the distribution of six-month total dosage is not normally distributed?

350	351	352	353	354	358	361	364	371	376
377	378	387	396	399	402	406	408	412	424
427	430	432	437	440	441	443	446	447	449

Although the graph seems symmetric, it is not bell-shaped. Most of the data are concentrated in the tails of the distribution. This suggests that the data are not from a normal distribution.

Frequency Histogram for the Cumulative Chemotherapy Dose Data



Example 6.9: Chemotherapy Protocol

(2) Backward Empirical Rule $\bar{x} = 399.03$ and $s = 34.94$

Interval	Frequency	Proportion
$(\overline{x} - s, \overline{x} + s) = (364.09, 433.97)$	15	0.50
$(\overline{x} - 2s, \overline{x} + 2s) = (329.15, 468.91)$	30	1.00
$(\overline{x} - 3s, \overline{x} + 3s) = (294.21, 503.85)$	30	1.00

The first two proportions of observations contained in each of the given intervals (0.5 and 1.0) are significantly different from 0.68 and 0.95. This suggests the population of total chemotherapy doses is not normal.

(3) IQR/*s*

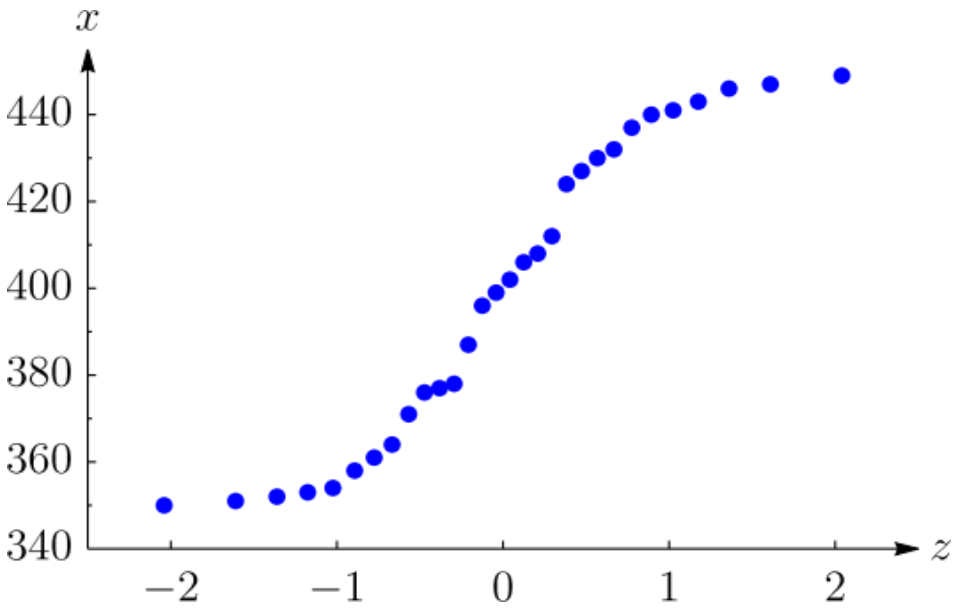
IQR/*s* = (432 – 364)/34.94 = 1.9452

This ratio is significantly different from 1.3, so there is evidence to suggest the underlying population is not normal.

Example 6.9: Chemotherapy Protocol

(4) Normal Probability Plot

The points do not lie along a straight line. Each tail is flat, which makes the graph look “S-shaped.” This suggests that the underlying population is not normal.



All four methods indicate that this sample did not come from a normal population.