Mathematics and Biostatistics

Chapter: [1]
Functions and Sequences

Section: [1.1]
Four Ways to Represent a Function

Functions arise whenever one quantity depends on another.

Consider the following situations:

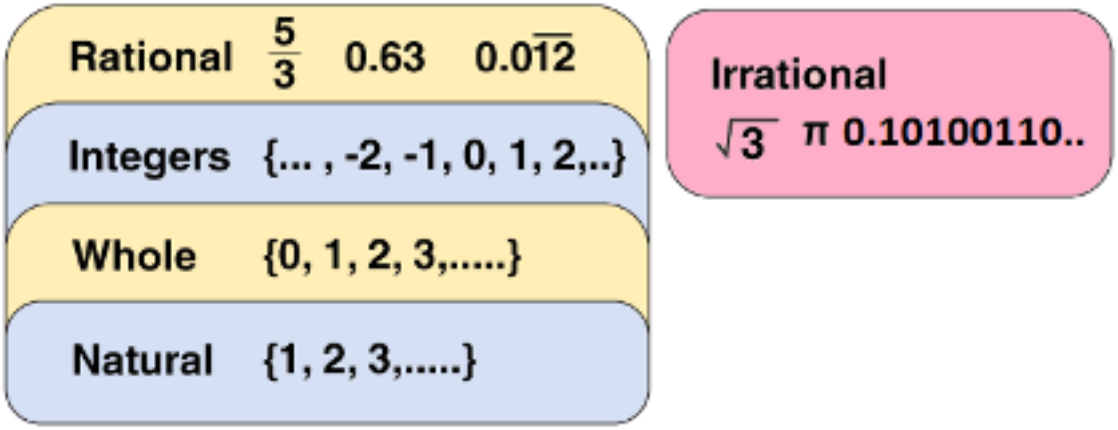
1. The area A of a circle depends on the radius r of the circle ($A = \pi r^2$).
2. The human population of the world P depends on the time t.

A function can be represented in several ways:

1. Table of values
2. Words
3. Formula
4. Graph

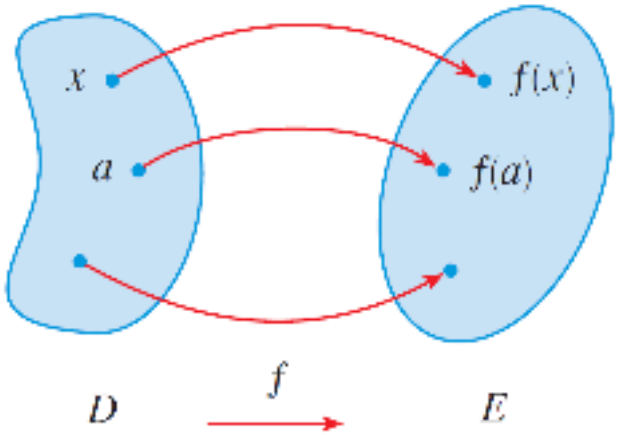
Real Numbers

- The set of all numbers.
- Denoted by $\mathbb{R} = (-\infty, \infty)$



Definition

A function f is a rule that assigns to each element x in a set D exactly one element, called $f(x)$, in a set E.



It's helpful to think of a function as a machine.

We can think of the domain as *the set of all possible inputs* and the range as *the set of all possible outputs*.

NOTES:

1. The sets D and E are sets of **real numbers**.
2. The set D is called the **domain** of the function.
3. The number $f(x)$ is the value of f at x.
4. The **range** of f is the set of all possible values of $f(x)$ as x varies throughout the domain.

NOTES (*Continue*):

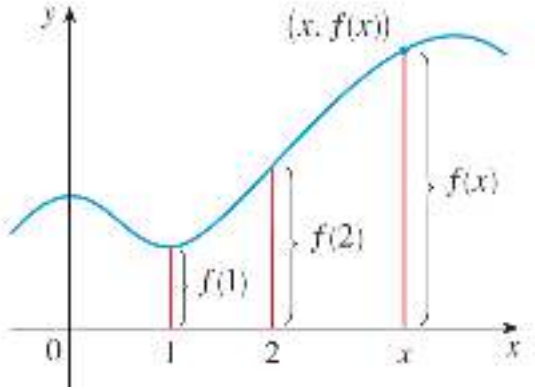
5. A symbol that represents an arbitrary number in the domain of a function f is called an **independent variable**.
6. A symbol that represents a number in the range of f is called a **dependent variable**.

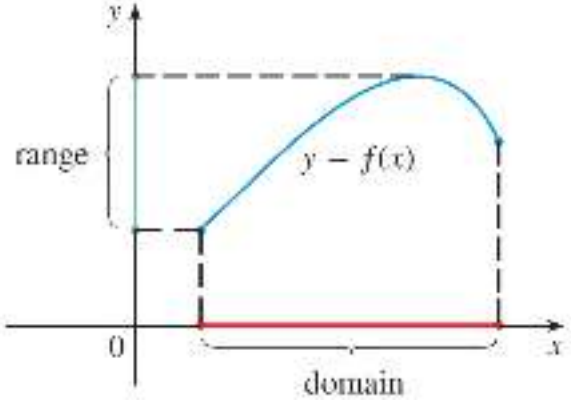
Examples

1. If $f(x) = 2x - 1$. Then $f(-1) = -3$. So, $-1 \in$ the domain of f, and $-3 \in$ the range of f.
2. If $g(x) = \frac{1}{x}$. Then $g(0) = \frac{1}{0}$ is not a real number. So, $0 \notin$ the domain of g.

Graphs of Functions

- The most common method for visualizing a function is its graph.
- If f is a function with domain D, then its graph is the set of ordered pairs $\{(x, f(x))|x \in D\}$.
- The graph of f consists of all points (x, y) in the coordinate plane such that $y = f(x)$ and x is in the domain of f.



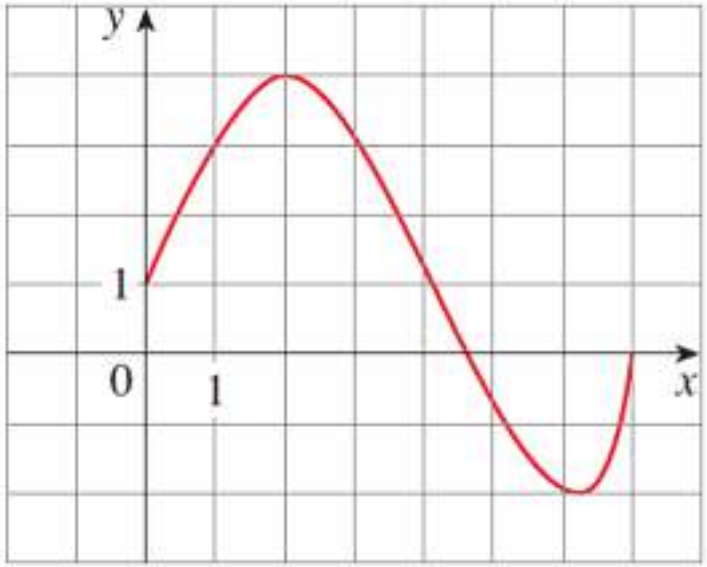


Example:

The graph of a function f is shown in the figure below.

a) Find the values of $f(1)$ and $f(5)$.

b) What are the domain and range of f?



Solution:

a) $f(1) = 3$ and $f(5) \approx -0.7$.

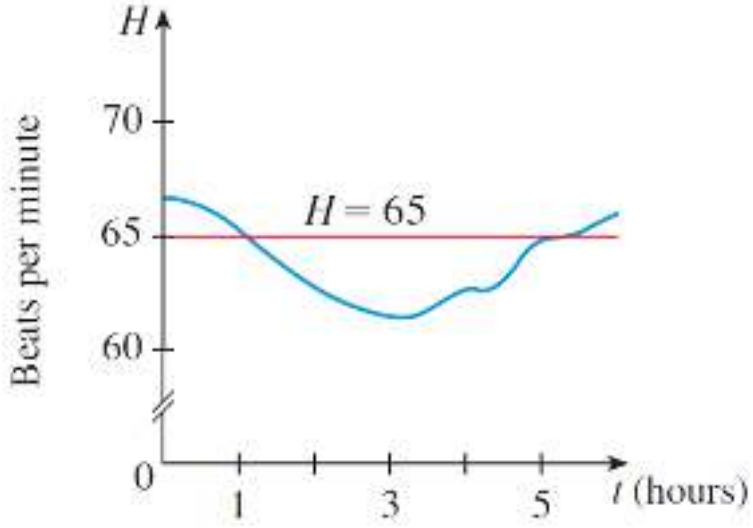
b) Domain $= [0,7]$ and range of $f = [-2,4]$

Example: (Antihypertension medication دواء خافض للضغط)

The figure below shows the effect of nifedipine tablets (antihypertension medication) on the heart rate $H(t)$ of a patient as a function of time.

a) Estimate the heart rate after two hours.

b) During what time period is the heart rate less than 65 beats/min?



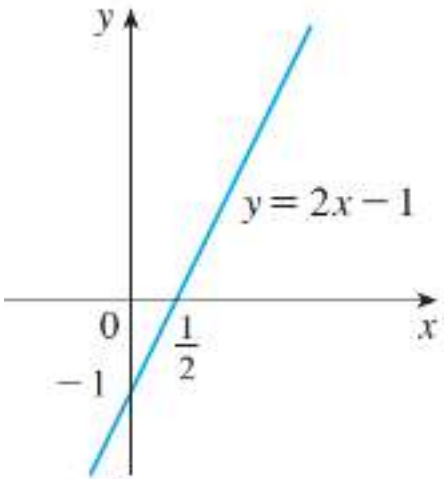
Solution:

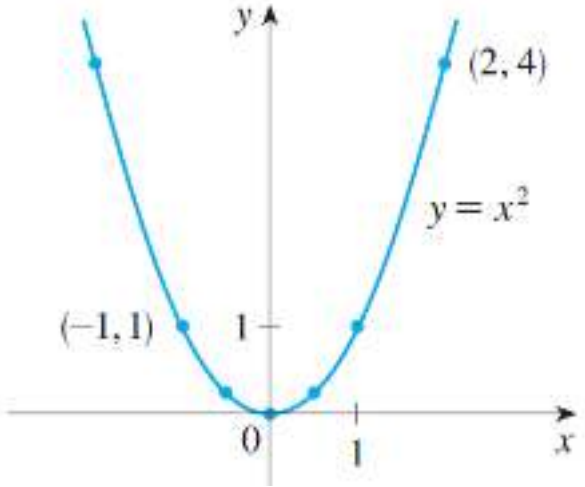
a) $H(2) \approx 62.5$ beats/min.

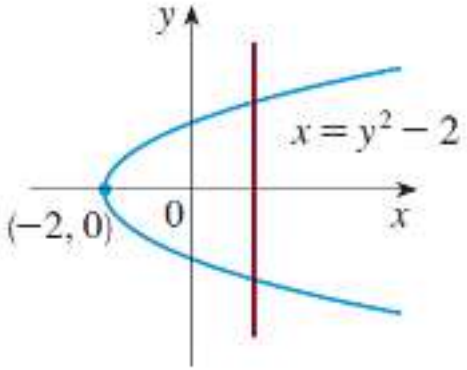
b) From 1 hour to 5 hours after the tablet is administered.

The Vertical Line Test

A curve in the xy −plane is the graph of a function of x if and only if no vertical line intersects the curve more than once.







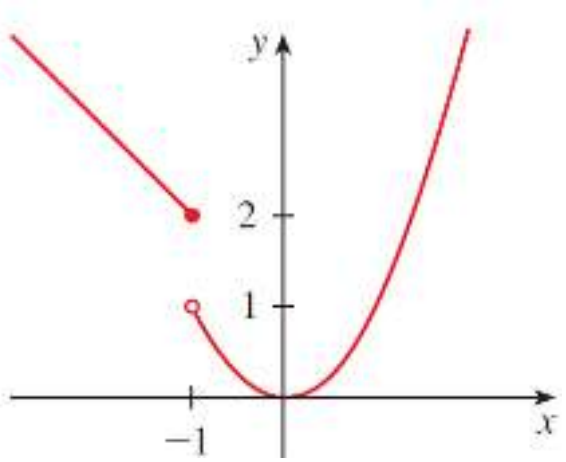
Piecewise Defined Functions

Functions that are defined by different formulas in different parts of their domains are called **piecewise defined functions**.

Example:

A function f is defined by $f(x) = \begin{cases} 1 - x & \text{if} \quad x \leq -1 \\ x^2 & \text{if} \quad x > -1 \end{cases}$.

Evaluate $f(-2)$, $f(-1)$, and $f(0)$ and sketch the graph.

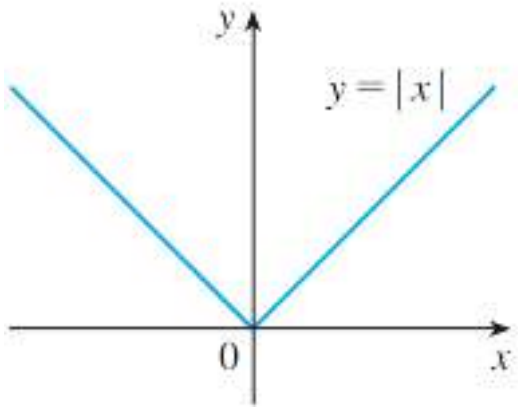


Solution

$$f(-2) = 1 - (-2) = 3$$
$$f(-1) = 1 - (-1) = 2$$
$$f(0) = 0^2 = 0$$

Absolute Value

- The absolute value of a number b, denoted by $|b|$, is the distance from b to 0 on the real number line.
- Distances are always positive or 0, so we have $|b| \geq 0$ for all $b \in \mathbb{R}$.
- In general, $|b| = \begin{cases} b & \text{if} \quad b \geq 0 \\ -b & \text{if} \quad b < 0 \end{cases}$.
- For example, $|-3| = 3$, $\left|\sqrt{2}\right| = \sqrt{2}$, and $|0| = 0$.
- The graph of the function $f(x) = |x|$ is shown.



Periodic Functions

- Many phenomena in the life sciences display a recurring type of behavior (نمط متكرر).
- For example:
  - ✓ Breathing (التنفس).
  - ✓ Beating of the heart (خفقان القلب).
  - ✓ Seasonal migration of butterflies (الهجرة الموسمية للفراشات).
- Such phenomena are referred to as **periodic** (دوريّة).

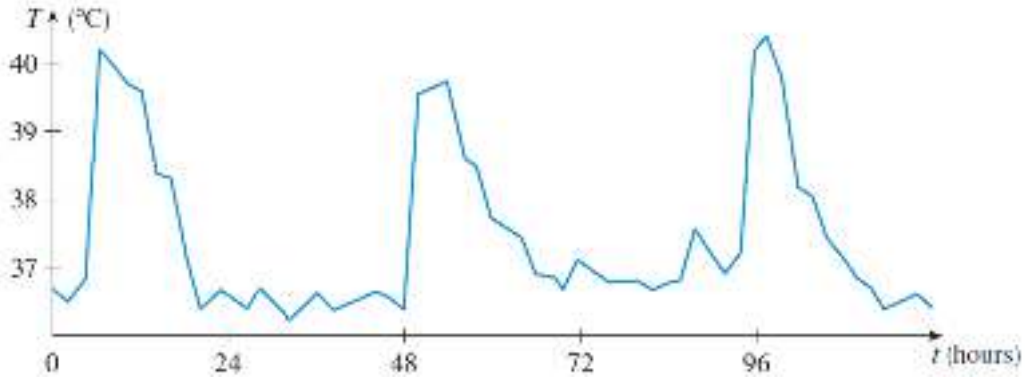
Definition

A function f is called **periodic** if there is a positive constant T such that $f(x+T) = f(x)$ for all values of x in the domain of f. The smallest value of T for which this is true is called the **period** of f.

Example

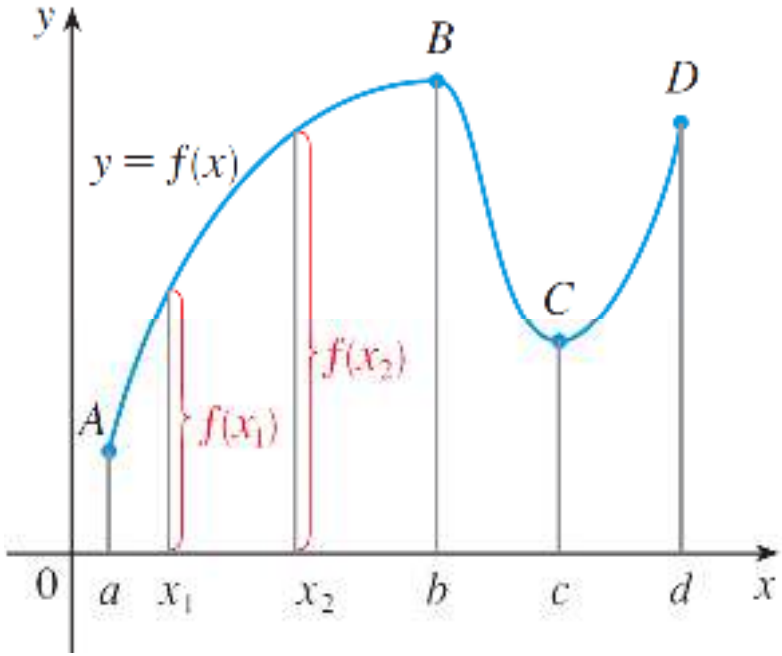
(Malarial Fever حمّى الملاريا)

The figure shows a typical temperature chart for a fever in humans induced by a species of malaria. Notice that the temperature approximately satisfies $T(t+48) = T(t)$. So, the temperature function has a period of about 48 hours.



Increasing and Decreasing Functions

- The graph shown in the figure rises from A to B, falls from B to C, and rises again from C to D.
- The function f is said to be increasing on the interval $[a, b]$, decreasing on $[b, c]$, and increasing again on $[c, d]$.



Definition

A function f is called **increasing** on an interval I if $f(x_1) < f(x_2)$ whenever $x_1 < x_2$ in I. It is called **decreasing** on I if $f(x_1) > f(x_2)$ whenever $x_1 < x_2$ in I.

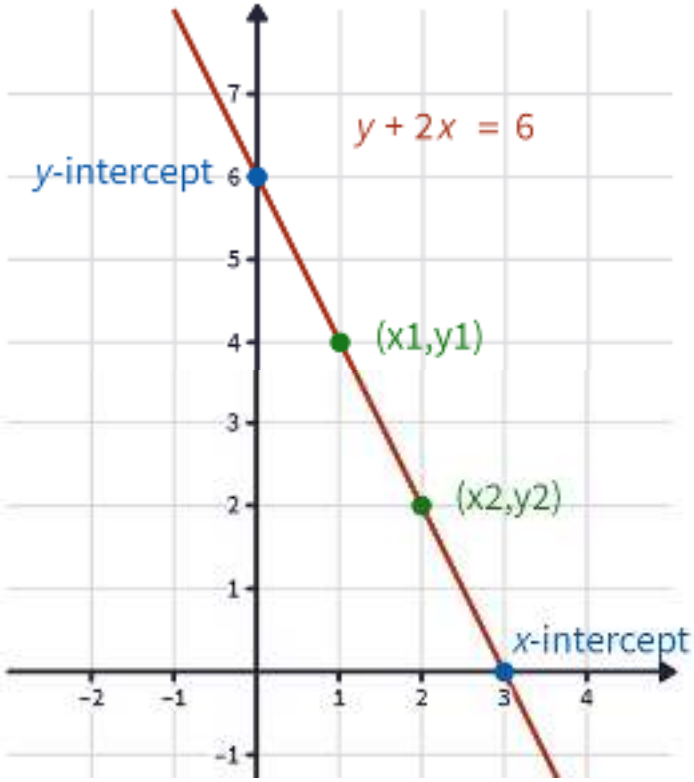
Mathematics and Biostatistics

Chapter: [1]
Functions and Sequences

Section: [1.2]
A Catalog of Essential Functions

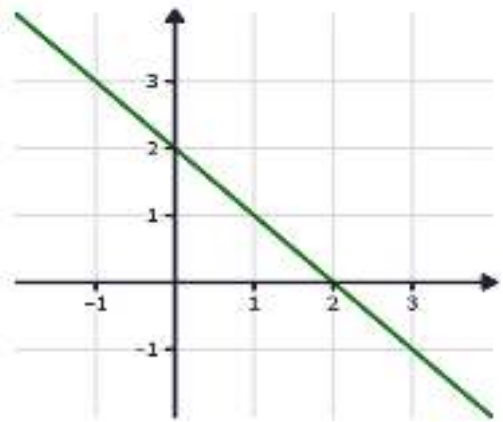
Linear Models

- When we say that y is a linear function of x, we mean that the graph of the function is a **line**.
- We can use the **slope-intercept** form of the equation of a line to write a formula for the function as $y = f(x) = mx + b$ where m is the **slope** of the line and b is the $\boldsymbol{y}$ **−intercept**.
- The y −intercept can be found by putting $x = 0$ in the equation of the line, and the x −intercept can be found by putting $y = 0$ in the equation of the line
- The slope $m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{4-2}{1-2} = \frac{2}{-1} = -2$.



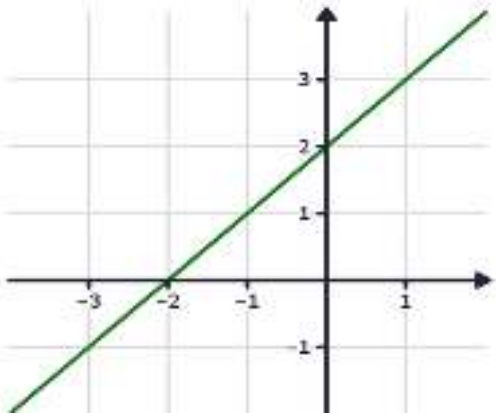
Linear Models

Note: There are four cases for slopes of lines



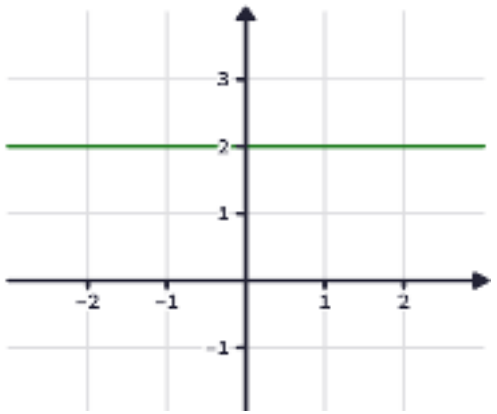
Negative Slope

$m < 0$



Positive Slope

$m > 0$



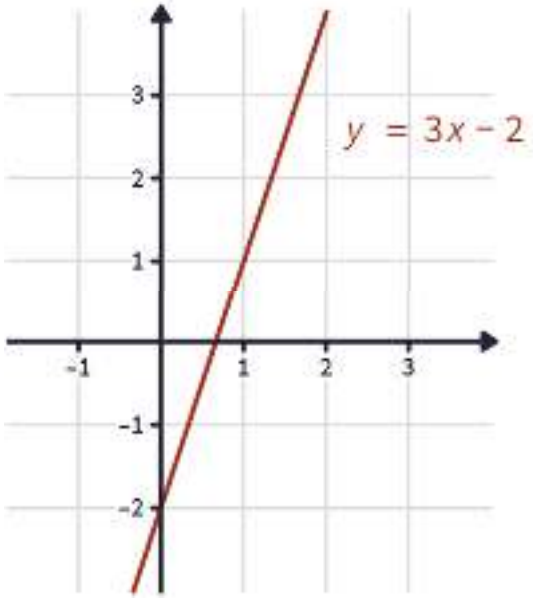
Zero Slope

$m = 0$

Undefined Slope

Linear Models

- A characteristic feature of linear functions is that **they increase or decrease at a constant rate**.
- If the quantities x and y are related by an equation $y = kx$ for some constant $k \neq 0$, we say that y **varies directly** as x, or y is **proportional** to x.
- The constant k is called the **constant of proportionality**.



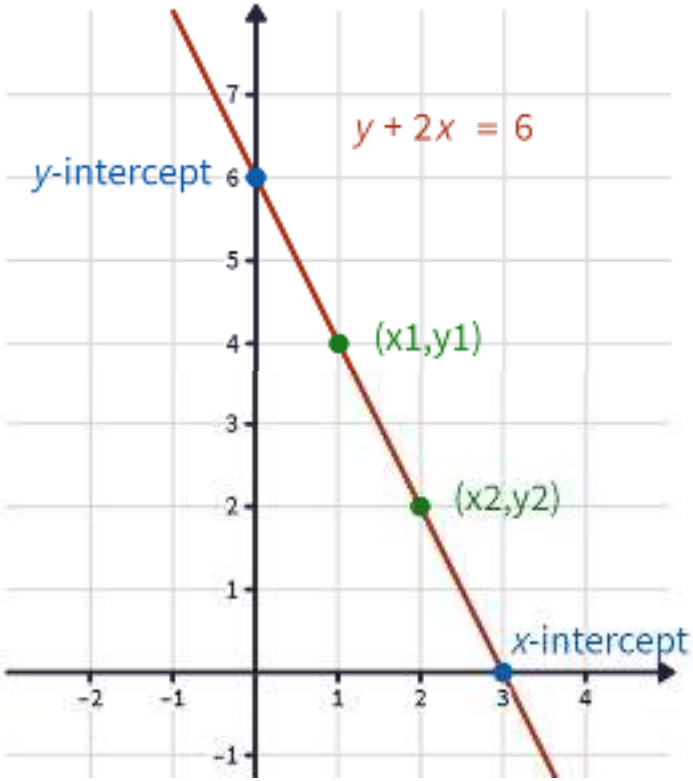
x	y
1.0	1.0
1.1	1.3
1.2	1.6
1.3	1.9
1.4	2.2
1.5	2.5

- The figure shows a graph of the linear function $f(x) = 3x - 2$ and a table of sample values. Notice that whenever x increases by 0.1, the value of $f(x)$ increases by 0.3. So $f(x)$ increases three times as fast as x.

Linear Models

- The figure shows a graph of the linear function $f(x) = -2x + 6$ and a table of sample values. Notice that whenever x increases by 1, the value of $f(x)$ decreases by 2.

x	y
0	6
1	4
2	2
3	0



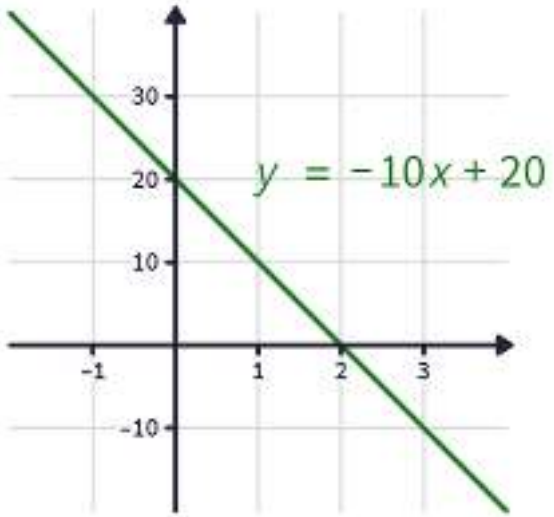
Linear Models

Example:

a) As dry air moves upward, it expands and cools. If the ground temperature is 20°C and the temperature at a height of 1 km is 10°C, express the temperature T (in °C) as a function of the height h (in kilometers), assuming that a linear model is appropriate.

b) Draw the graph of the function in part (a). What does the slope represent?

c) What is the temperature at a height of 2.5 km?



Solution $T(h) = m \cdot h + b$

$$T(0) = 20 \quad \Rightarrow m \cdot 0 + b = 20 \quad \Rightarrow b = 20$$

$$T(1) = 10 \quad \Rightarrow m \cdot 1 + 20 = 10 \Rightarrow m = -10$$

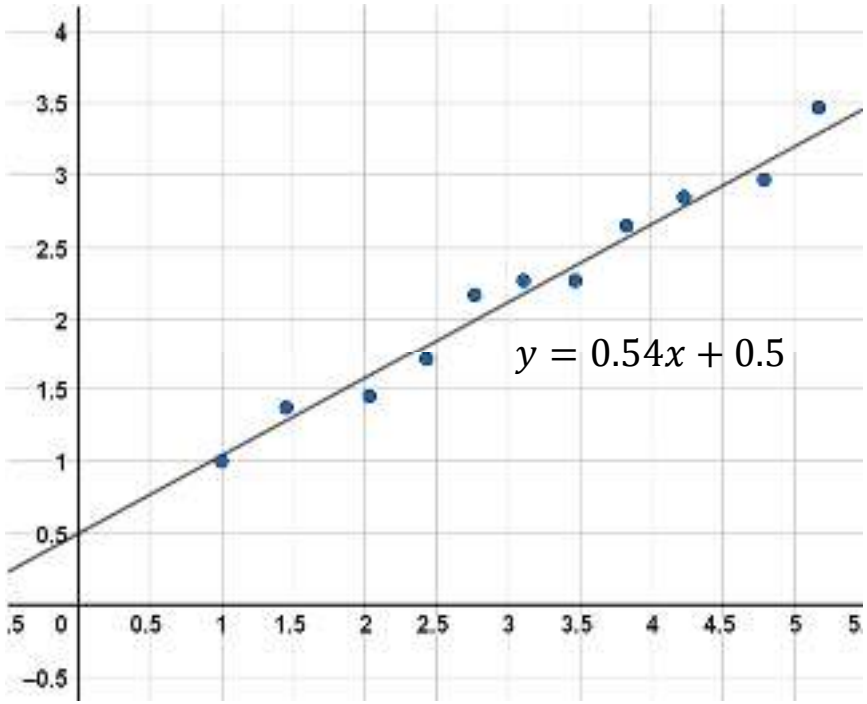
$$\therefore T(h) = -10h + 20$$

$$T(2.5) = -10(2.5) + 20 = -5°\text{C}$$

Linear Models

Notes:

- If there is no physical law or principle to help us formulate a model, we construct an empirical model (**نموذج تجريبي**), which is based entirely on collected data.
- We seek a curve that "**fits**" the data in the sense that it captures the basic trend of the data points.
- A better linear model is obtained by a procedure from statistics called **linear regression (الانحدار الخطي)** (*Section 11.3*).



Polynomials

A function P is called a **polynomial** if

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

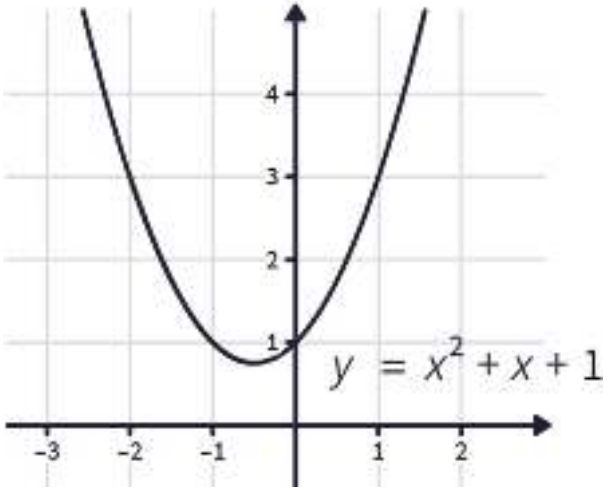
where n is a nonnegative integer and the numbers $a_n, a_{n-1}, \cdots, a_2, a_1, a_0$ are constants called the *coefficients of the polynomial*.

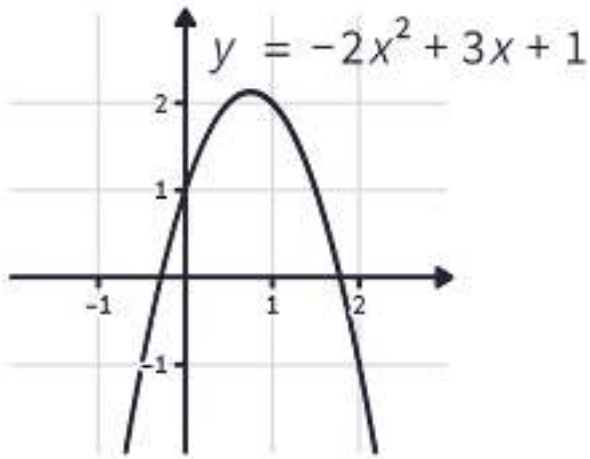
- The domain of any polynomial is $\mathbb{R} = (-\infty, \infty)$.
- If the leading coefficient $a_n \neq 0$, then the degree of the polynomial is n.
- For example, the function $P(x) = x^4 - 2x^2 + x - 5$ is a polynomial of degree 4.

Polynomials

- A polynomial of degree 1 is of the form $P(x) = mx + b$ and so it is a linear function.

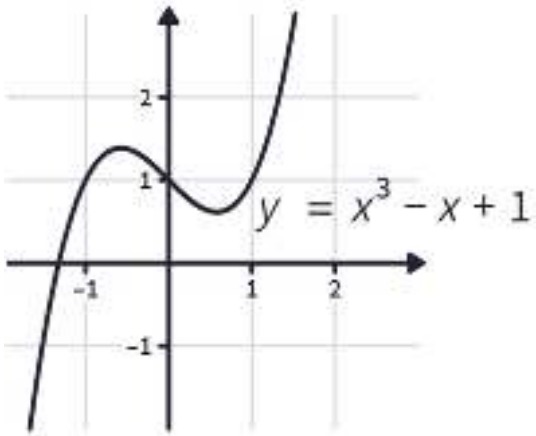
- A polynomial of degree 2 is of the form $P(x) = ax^2 + bx + c$ and is called a **quadratic function**.



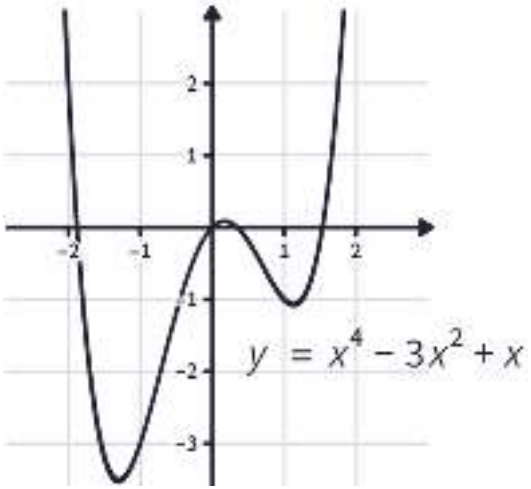


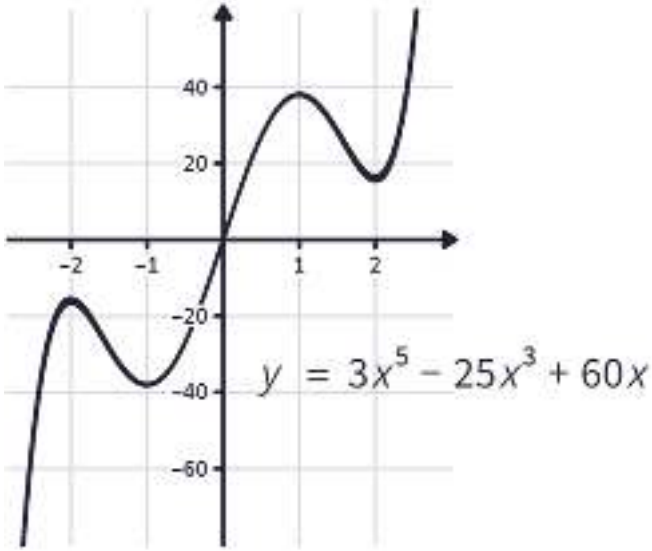
Polynomials

- A polynomial of degree 3 is of the form $P(x) = ax^3 + bx^2 + cx + d$ and is called a **cubic function**.



- **Other Examples:**

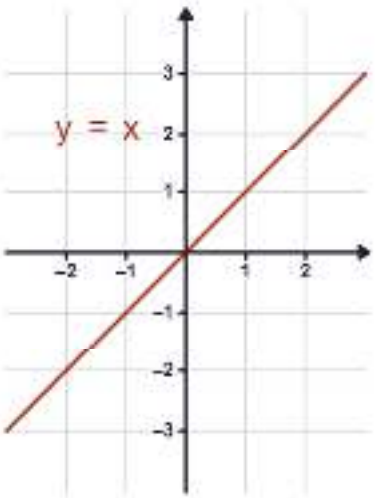




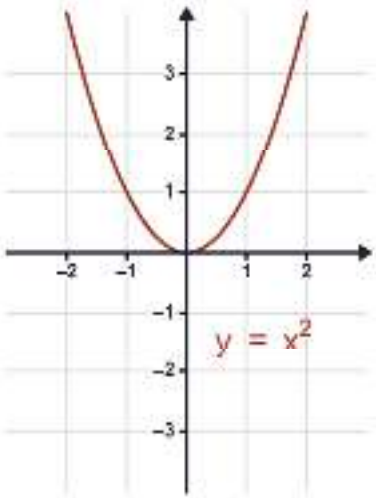
Power Functions

A function of the form $f(x) = x^p$, where p is a constant, is called a **power function**.

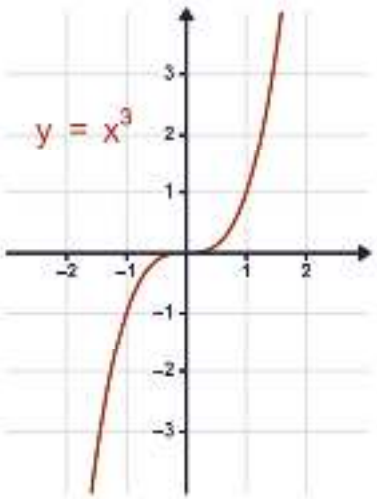
Case (1): $p = n$, where n is a **positive integer**, $n = 1,2,3, \cdots$.



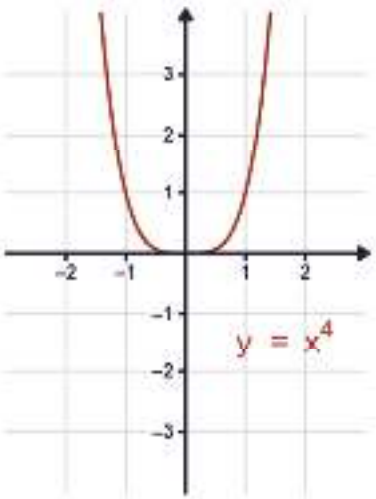
Odd Power



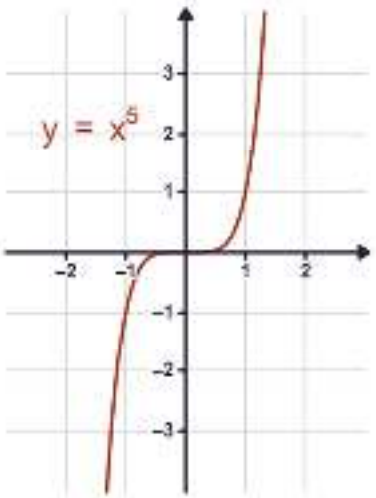
Even Power



Odd Power



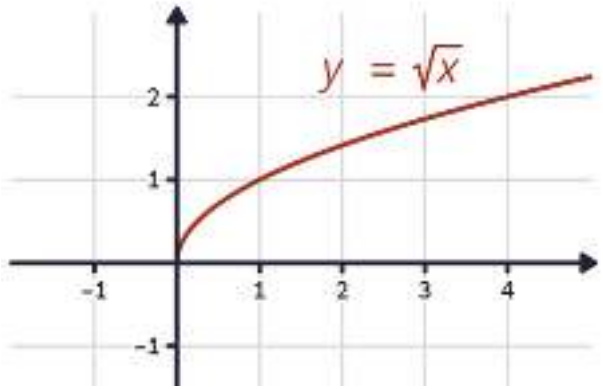
Even Power



Odd Power

Power Functions

Case (2): $p = \frac{1}{n}$, where n is a **positive integer**, $n = 1,2,3,\cdots$.

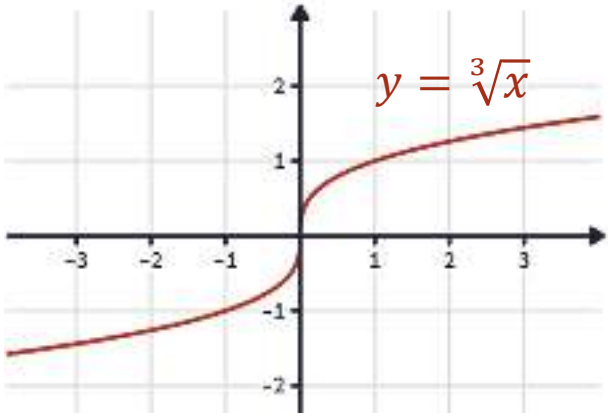


Domain $= [0, \infty)$

Range $= [0, \infty)$

Even Root

Avoid even roots of negative numbers



Domain $= \mathbb{R}$

Range $= \mathbb{R}$

Odd Root

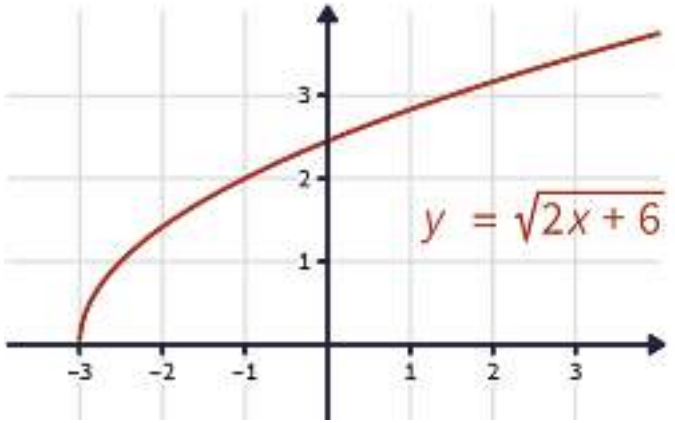
Power Functions

Example: Find the domain of the following functions.

$f(x) = \sqrt{2x+6}$

$$2x + 6 \geq 0$$
$$2x \geq -6$$
$$x \geq -3$$

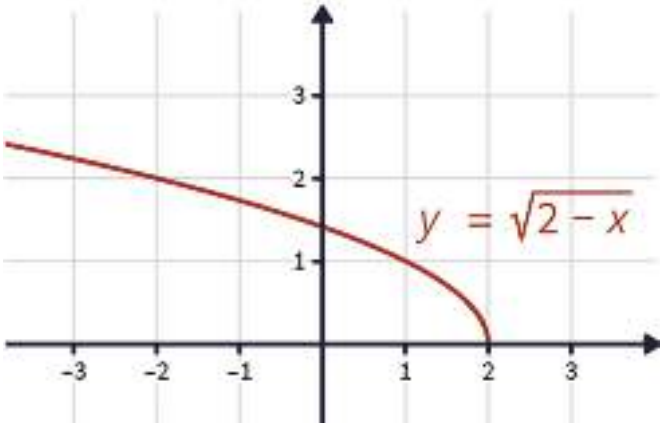
$x \in [-3, \infty)$



$g(x) = \sqrt{2-x}$

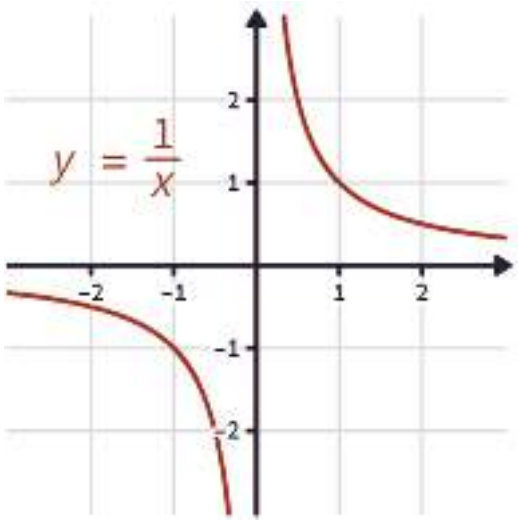
$$2 - x \geq 0$$
$$-x \geq -2$$
$$x \leq 2$$

$x \in (-\infty, 2]$



Power Functions

Case (3): $p = -1$. The graph of the reciprocal function $f(x) = x^{-1} = \frac{1}{x}$ is shown below.



Domain $= (-\infty, 0) \cup (0, \infty) = \mathbb{R} - \{0\}$

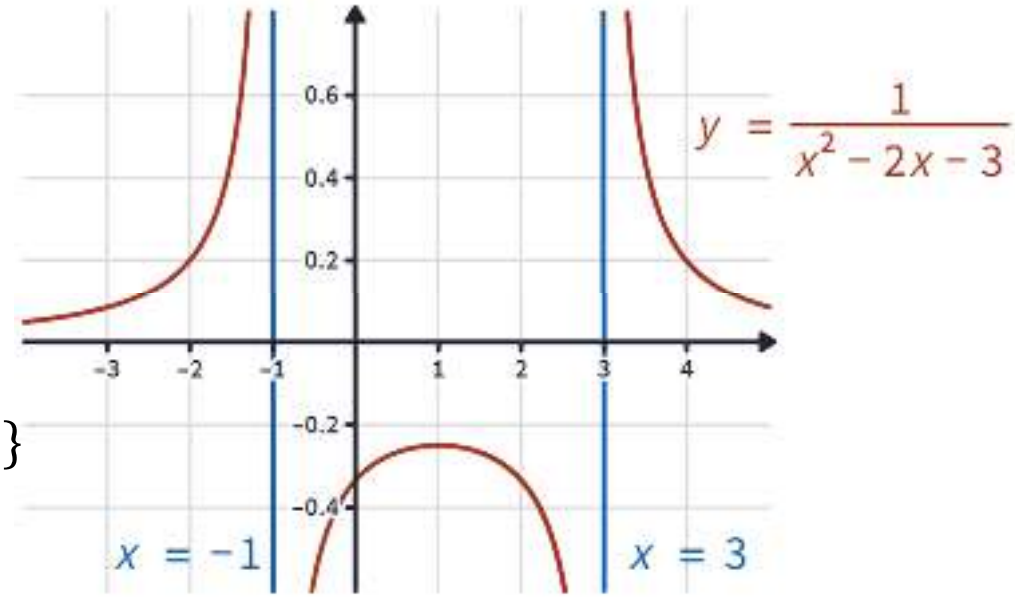
Range $= (-\infty, 0) \cup (0, \infty) = \mathbb{R} - \{0\}$

Rational Functions

- A **rational function** $f(x) = \frac{P(x)}{Q(x)}$ is a ratio of two polynomials $P(x)$ and $Q(x)$.
- The domain consists of all values of x such that $Q(x) \neq 0$.
- **For rational functions, avoid division by 0.**

Example: Find the domain of

$$f(x) = \frac{1}{x^2 - 2x - 3}$$



Solution

$$\begin{aligned} \text{domain} &= \mathbb{R} - \{x^2 - 2x - 3 = 0\} \\ &= \mathbb{R} - \{(x-3)(x+1) = 0\} \\ &= \mathbb{R} - \{-1, 3\} \end{aligned}$$

Algebraic Functions

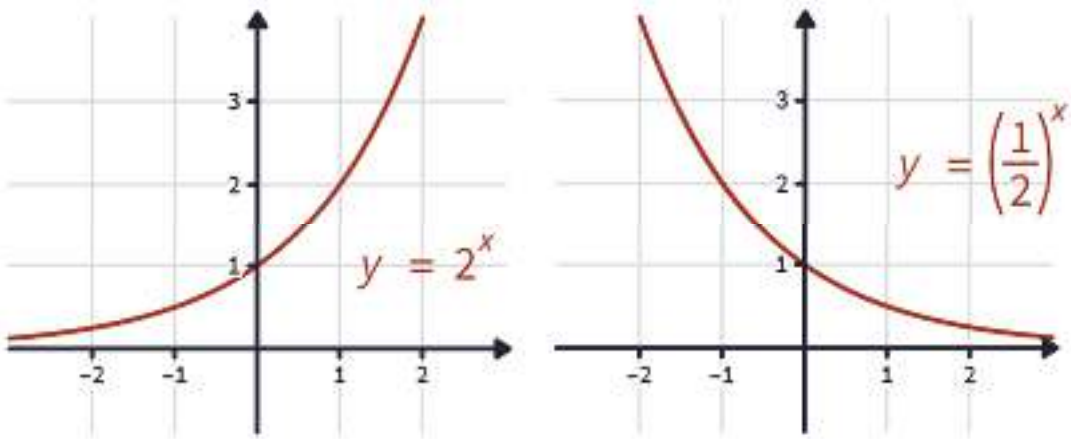
- A function f is called an **algebraic function** if it can be constructed using algebraic operations (such as addition, subtraction, multiplication, division, and taking roots) starting with polynomials.
- Any rational function is automatically an algebraic function.

Examples: $f(x) = \sqrt{x^2 + 1}$

$$g(x) = \frac{x^4 - 16x^2}{x + \sqrt{x}} + (x - 2)\sqrt[3]{x - 1}$$

Exponential Functions

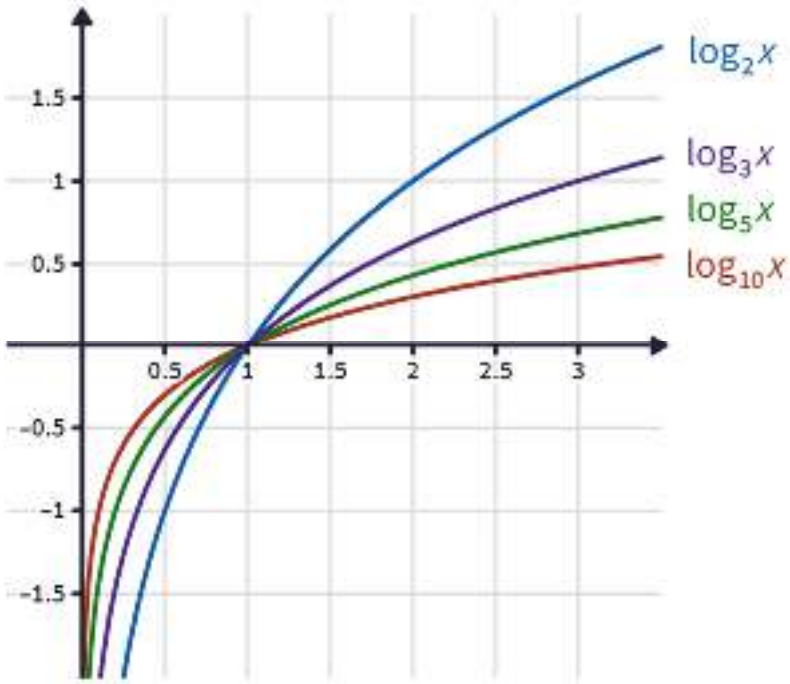
- The **exponential functions** are the functions of the form $f(x) = b^x$, where the base b is a positive constant.
- The graphs of $y = 2^x$ and $y = \left(\frac{1}{2}\right)^x$ are shown below.
- In both cases the domain is $\mathbb{R} = (-\infty, \infty)$ and the range is $(0, \infty)$.



Logarithmic Functions

- The **logarithmic functions** $f(x) = \log_b x$, where the base b is a positive constant, are the *inverse functions* of the exponential functions.

- The accompanying figure shows the graphs of four logarithmic functions with various bases.
- In each case the domain is $(0, \infty)$, the range is $\mathbb{R} = (-\infty, \infty)$, and the function increases slowly when $x > 1$.



Mathematics and Biostatistics

Chapter: [1]
Functions and Sequences

Section: [1.4]
Exponential Functions

Introduction

- We often hear people saying that something is "**growing exponentially**." What does that mean, exactly?
- We answer that question in this section by looking at examples of **exponential growth (نمو)** and **decay (اضمحلال)** as modeled by exponential functions.

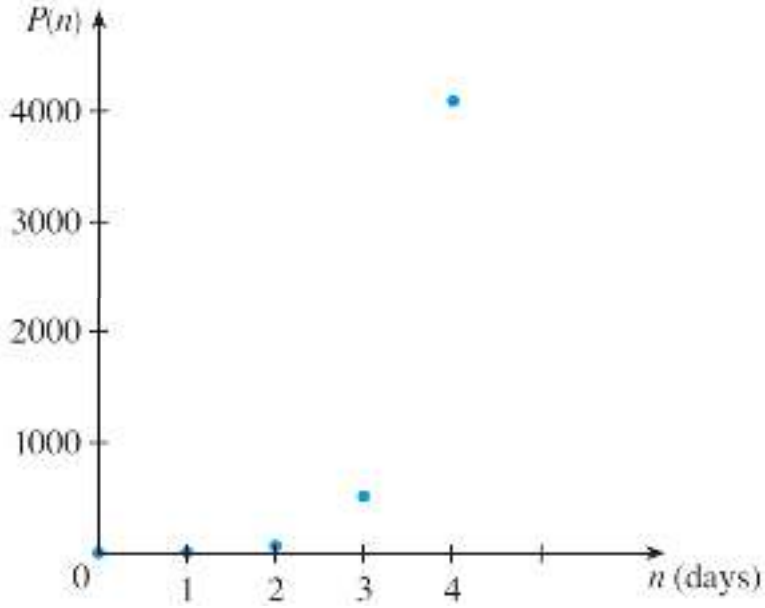
The Growth of Malarial Parasites

- Malaria kills more than a million people every year.
- To understand the mechanisms that regulate malarial growth, controlled experiments have been done on mice.
- Individual cells of a species of malaria reproduce (تتكاثر) every 24 hours.
- The parasites develop in red blood cells for a period of 24 hours and then they all burst at the same time, quickly reinvade new blood cells, and start the process again.

The Growth of Malarial Parasites

- Each infected blood cell produces **eight new parasites** when it bursts.
- If $P(n)$ is the number of parasites after n days, then:

$$P(0) = 1 = 8^0,$$
$$P(1) = 8 \times 1 = 8^1,$$
$$P(2) = 8 \times P(1) = 64 = 8^2,$$
$$P(3) = 8 \times P(2) = 512 = 8^3,$$
$$\vdots$$
$$P(n) = 8^n$$



Day n	$P(n)$
0	1
1	8
2	64
3	512
4	4,096
5	32,768

Exponential Functions

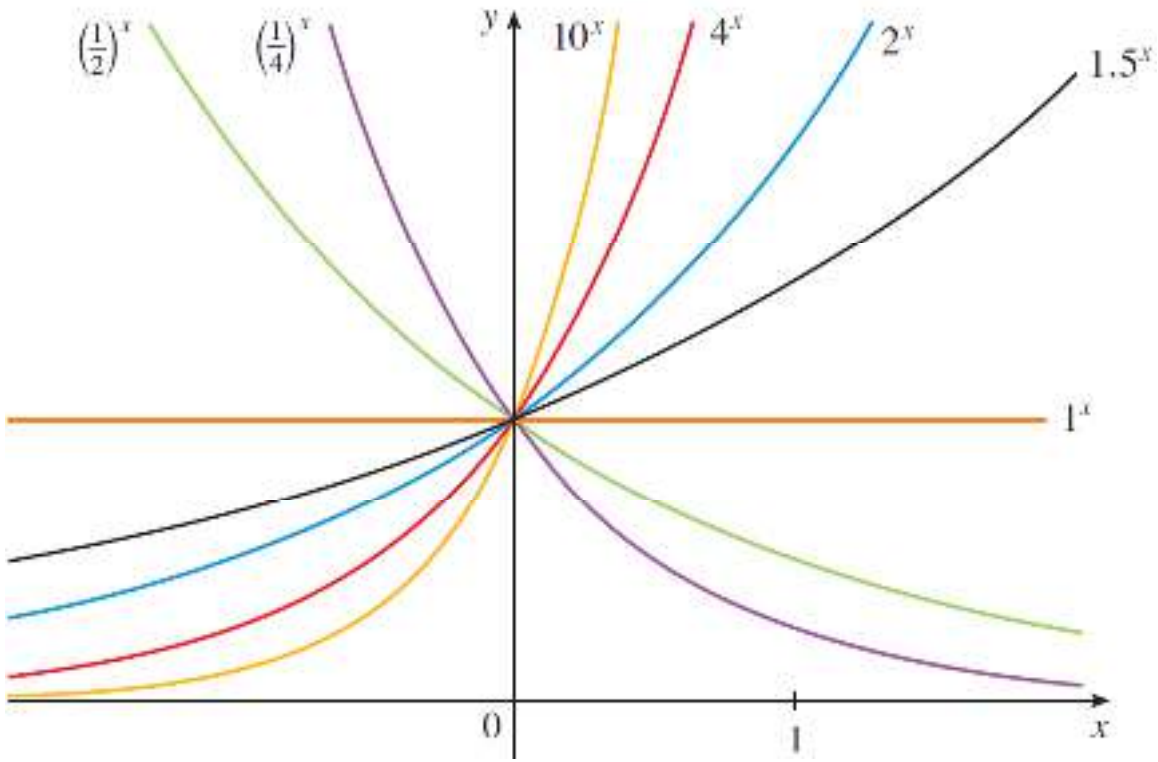
An **exponential function** is a function of the form $f(x) = b^x$, where b is a positive constant. Observe that if $b \neq 1$, then the exponential function has domain $\mathbb{R}$ and range $(0, \infty)$.

Notes

- If $x = n$ is a positive integer, then $b^n = \underbrace{b \cdot b \cdots \cdots b}_{n-\text{factors}}$.
- If $x = 0$, then $b^0 = 1$.
- If $x = -n$, then $b^{-n} = \frac{1}{b^n}$. **Example:** $2^{-3} = \frac{1}{2^3} = \frac{1}{8}$.
- If $x = \frac{p}{q}$, then $b^{p/q} = \sqrt[q]{x^p}$. **Example:** $16^{3/4} = \sqrt[4]{16^3} = 8$.

Exponential Functions

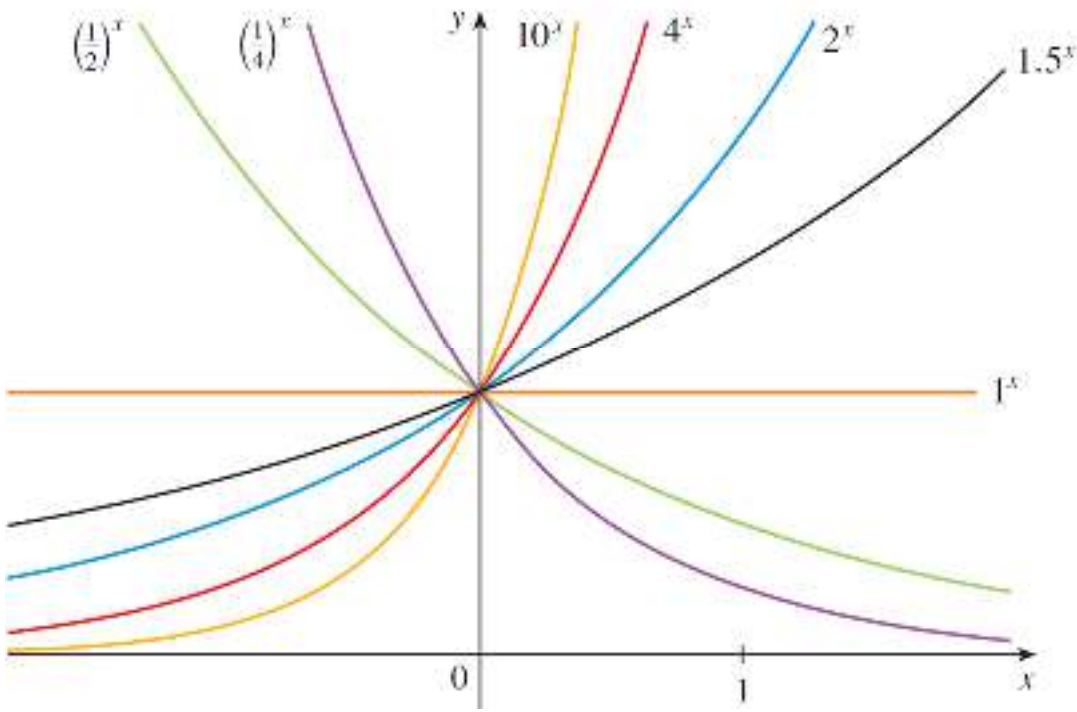
The graphs of members of the family of functions $y = b^x$ are shown below for various values of the base b.



Exponential Functions

The graphs of members of the family of functions $y = b^x$ are shown below for various values of the base b.

- All of these graphs pass through the point $(0,1)$ because $b^0 = 1$ for $b \neq 0$.
- As the base b gets larger, the exponential function grows more rapidly (for $x > 0$).
- If $0 < b < 1$, the exponential function decreases; if $b = 1$, it is a constant; and if $b > 1$, it increases.
- $\left(\frac{1}{b}\right)^x = \frac{1}{b^x} = b^{-x}$.



Laws of Exponents

If a and b are positive numbers and x and y are any real numbers, then

- $b^{x+y} = b^x \cdot b^y$
- $b^{x-y} = \frac{b^x}{b^y}$
- $(b^x)^y = b^{xy}$
- $(ab)^x = a^x \cdot b^x$

Example Use the *Law of Exponents* to rewrite and simplify the expression.

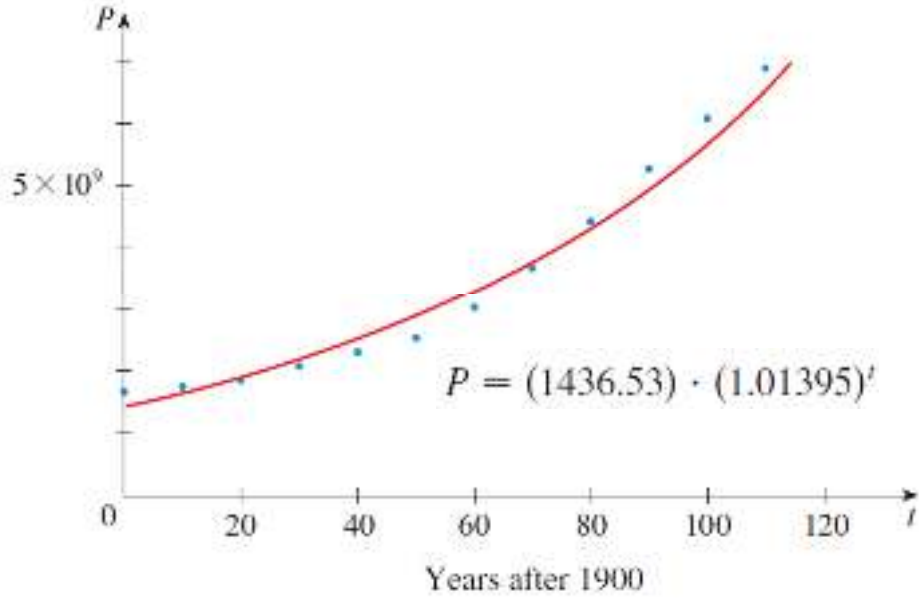
- $\frac{9^{-2}}{6^{-5}} = \frac{6^5}{9^2} = \frac{(2\times3)^5}{(3^2)^2} = \frac{2^5\times3^5}{3^4} = 2^5 \times 3^{5-4} = 32 \times 3 = 96.$
- $\frac{x^2\sqrt{x}}{\sqrt[3]{x^4}} = \frac{x^2 \cdot x^{1/2}}{x^{4/3}} = \frac{x^{2+1/2}}{x^{4/3}} = \frac{x^{5/2}}{x^{4/3}} = x^{\frac{5}{2}-\frac{4}{3}} = x^{\frac{7}{6}} = \sqrt[6]{x^7}.$

Exponential Growth

Example: World population growth

The table below shows data for the population of the world from 1900 to 2010. Use an exponential function to model the data.

t (years after 1900)	Population (millions)	t (years after 1900)	Population (millions)
0	1650	60	3040
10	1750	70	3710
20	1860	80	4450
30	2070	90	5280
40	2300	100	6080
50	2560	110	6870



Exponential Decay

Example:

The half-life of strontium-90, ^{90}Sr, is 25 years. This means that half of any given quantity of ^{90}Sr will disintegrate in 25 years.

a) If a sample of ^{90}Sr has a mass of 24 mg, find an expression for the mass $m(t)$ that remains after t years.

$$m(0) = 24$$

$$m(25) = \frac{1}{2} \cdot 24$$

$$m(50) = \frac{1}{2}\left(\frac{1}{2} \cdot 24\right) = \frac{1}{2^2} \cdot 24$$

$$m(75) = \frac{1}{2}\left(\frac{1}{2^2} \cdot 24\right) = \frac{1}{2^3} \cdot 24$$

$$m(100) = \frac{1}{2}\left(\frac{1}{2^3} \cdot 24\right) = \frac{1}{2^4} \cdot 24$$

$$\vdots$$

$$m(t) = \frac{1}{2^{t/25}} \cdot 24 = 24 \cdot \left(2^{-1/25}\right)^t$$

This is an exponential function with base $b = 2^{-1/25}$.

Exponential Decay

Example:

The half-life of strontium-90, ^{90}Sr, is 25 years. This means that half of any given quantity of ^{90}Sr will disintegrate in 25 years.

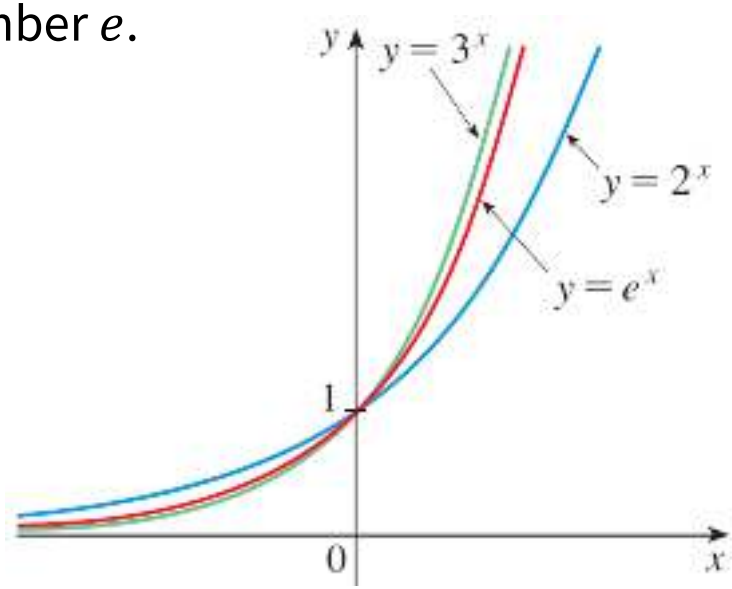
b) Find the mass remaining after 40 years, correct to the nearest milligram.

$$m(t) = 24 \cdot \left(2^{-1/25}\right)^t$$

$$m(40) = 24 \cdot 2^{-40/25} = 24 \cdot 2^{-8/5} = \frac{24}{\sqrt[5]{256}} \approx 7.9 \text{ mg}$$

The Number e

- Of all possible bases for an exponential function, there is one that is most convenient for the purposes of calculus, the number e.
- The number e lies between 2 and 3 and the graph of $y = e^x$ lies between the graphs of $y = 2^x$ and $y = 3^x$.
- The value of e, correct to five decimal places, is $e \approx 2.71828$.
- We call the function $f(x) = e^x$ the **natural exponential function**.



Mathematics and Biostatistics

Chapter: [1]

Functions and Sequences

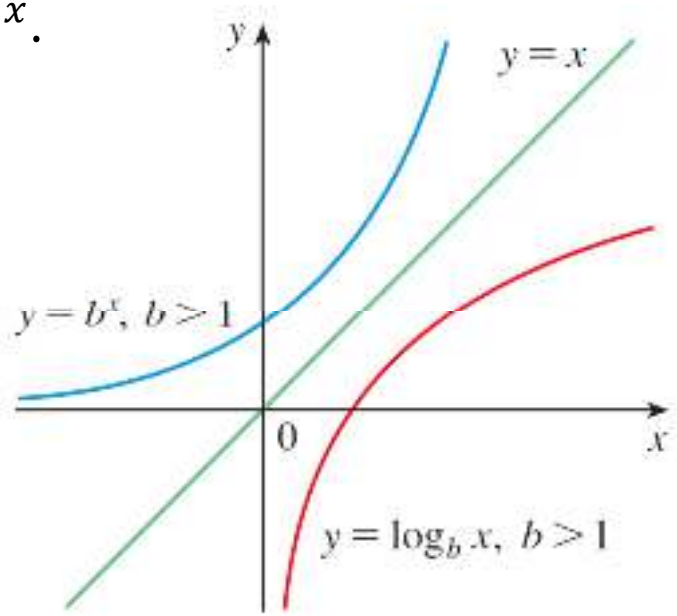
Section: [1.5]

Logarithms; Semilog and Log-Log Plots

Logarithmic Functions

- If $b > 0$ and $b \neq 0$, the logarithmic function with base b is denoted by $\log_b x$, and it is the **inverse** of the exponential function b^x.
- The domain of $\log_b x$ is $(0, \infty)$, and its range is $\mathbb{R}$.
- $\log_b x = y \iff x = b^y$
- **The cancellation equations:**

$$\log_b(b^x) = x \quad \text{for every } x \in \mathbb{R}$$
$$b^{\log_b x} = x \quad \text{for every } x > 0$$



- $\log_b 1 = 0$ and $\log_b b = 1$

Laws of Logarithms

If x and y are **positive numbers**, then

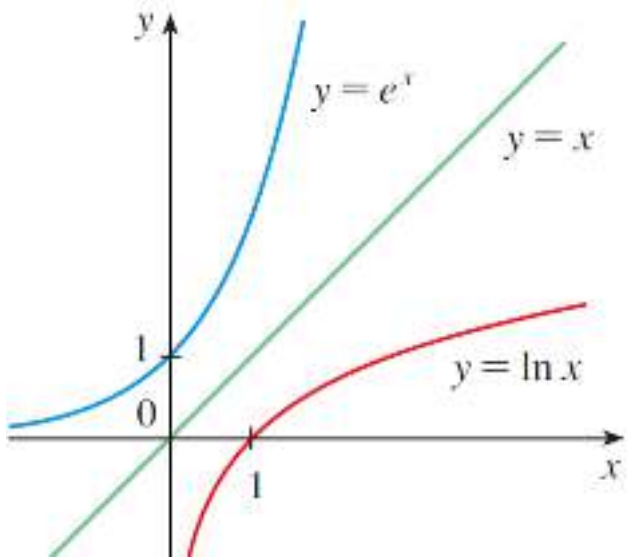
- $\log_b(xy) = \log_b x + \log_b y$
- $\log_b\left(\frac{x}{y}\right) = \log_b x - \log_b y$
- $\log_b(x^r) = r \log_b x$, where r is any real number.

Example Use the laws of logarithms to evaluate $\log_2 80 - \log_2 5$.

$$\log_2 80 - \log_2 5 = \log_2\left(\frac{80}{5}\right) = \log_2 16 = \log_2(2^4) = 4\log_2 2 = 4$$

Natural Logarithms

- Of all possible bases b for logarithms, the most convenient choice of a base is the number e, which was defined in Section 1.4.
- The logarithm with base e is called the **natural logarithm** and has a special notation $\boldsymbol{\log_e x = \ln x}$.
- $\ln x = y \iff x = e^y$.
- $\ln e = 1$ and $\ln 1 = 0$.
- $\ln(e^x) = x$ for $x \in \mathbb{R}$, and $e^{\ln x} = x$ for $x > 0$.



Natural Logarithms

Example Find x if $\ln x = 5$.

$$\ln x = 5 \Rightarrow x = e^5$$

Example Solve the equation $e^{5-3x} = 10$.

$$\begin{aligned} e^{5-3x} = 10 &\Rightarrow \ln(e^{5-3x}) = \ln 10 \\ &\Rightarrow 5 - 3x = \ln 10 \\ &\Rightarrow -3x = -5 + \ln 10 \Rightarrow x = \frac{1}{3}(5 - \ln 10) \end{aligned}$$

Example Express $\ln a + \frac{1}{2}\ln b$ as a single logarithm.

$$\ln a + \frac{1}{2}\ln b = \ln a + \ln\left(b^{1/2}\right) = \ln a + \ln\sqrt{b} = \ln\left(a\sqrt{b}\right)$$

Change of Base Formula

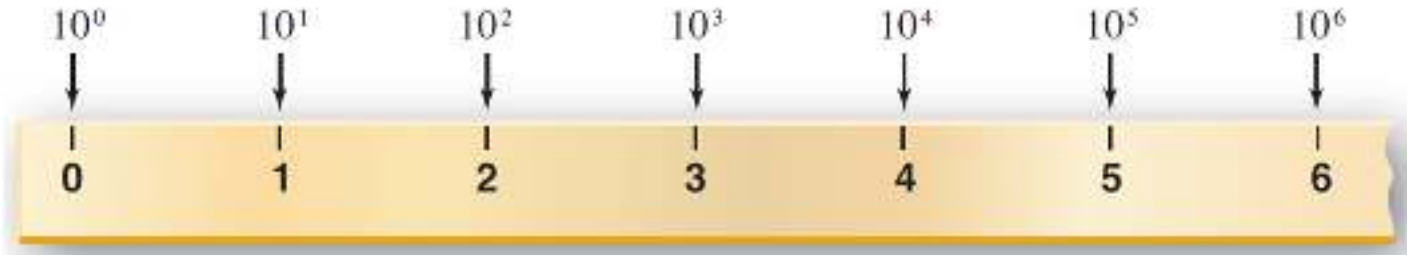
For any positive number b $(b \neq 1)$, we have $\log_b a = \dfrac{\ln a}{\ln b}$

Example Evaluate $\log_8 5$ correct to six decimal places.

$$\log_8 5 = \frac{\ln 5}{\ln 8} \approx 0.773976$$

Semilog Plots

- We've seen that the exponential function $y = b^x$ $(b > 0)$ increases so rapidly that it's sometimes difficult to represent data points conveniently on a single plot.
- On the other hand, we have just seen that the logarithmic functions, increase very slowly.
- For that reason, **logarithmic scales** are often used when real-world quantities involve a huge disparity in size.
- In such cases, the equidistant marks on a logarithmic scale represent consecutive **powers of 10**.



Semilog Plots

- In biology it's common to use a **Semilog plot** *to see whether data points are appropriately modeled by an exponential function*.
- This means that instead of plotting the points (x, y), we plot the points $(x, \log_{10} x)$.
- In other words, we use a logarithmic scale on the vertical axis.
- If we start with an exponential function of the form $y = a \cdot b^x$ and take logarithms of both sides, we get

$$y = a \cdot b^x$$

$$\log_{10} y = \log_{10}(a \cdot b^x)$$

$$\log_{10} y = \log_{10} a + \log_{10}(b^x)$$

$$\underbrace{\log_{10} y}_{Y} = \underbrace{\log_{10} a}_{B} + x\,\underbrace{\log_{10} b}_{M}$$

$$Y = B + Mx$$

which is the equation of a line with slope M and Y −intercept B.

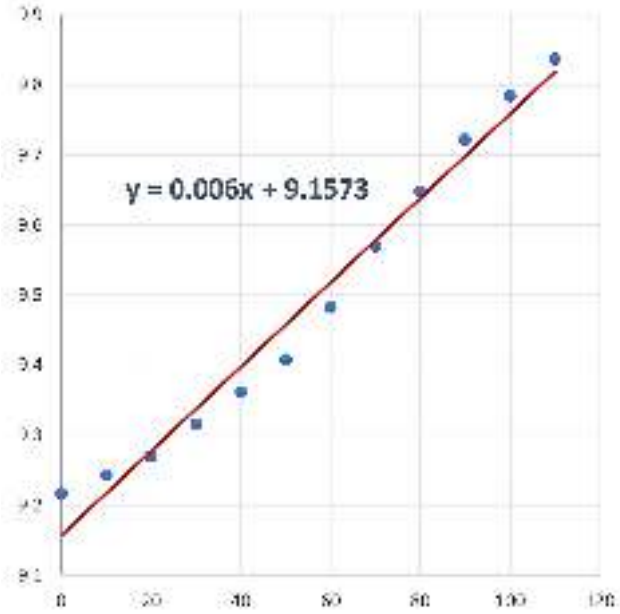
Semilog Plots

- So if we obtain experimental data that we suspect might possibly be exponential, then we could graph a Semilog scatter plot and see if it is approximately linear.
- If so, we could then obtain an exponential model for our original data.

Example: In Section 1.4, we presented data for the population of the world from 1900 to 2010. We see that the data points in the figure lie very nearly on a straight line and, using linear regression, we get the equation

$$\log_{10} P(t) = 9.1573 + 0.006\,t$$

x	y	log(y)
0	1650000000	9.2174839
10	1750000000	9.2430380
20	1860000000	9.2695129
30	2070000000	9.3159703
40	2300000000	9.3617278
50	2560000000	9.4082400
60	3040000000	9.4828736
70	3710000000	9.5693739
80	4450000000	9.6483600
90	5280000000	9.7226339
100	6080000000	9.7839036
110	6870000000	9.8369567



Semilog Plots

Note: In the previous example, applying the exponential function with base 10 to both sides of this equation, we obtain an equation for the population as follows.

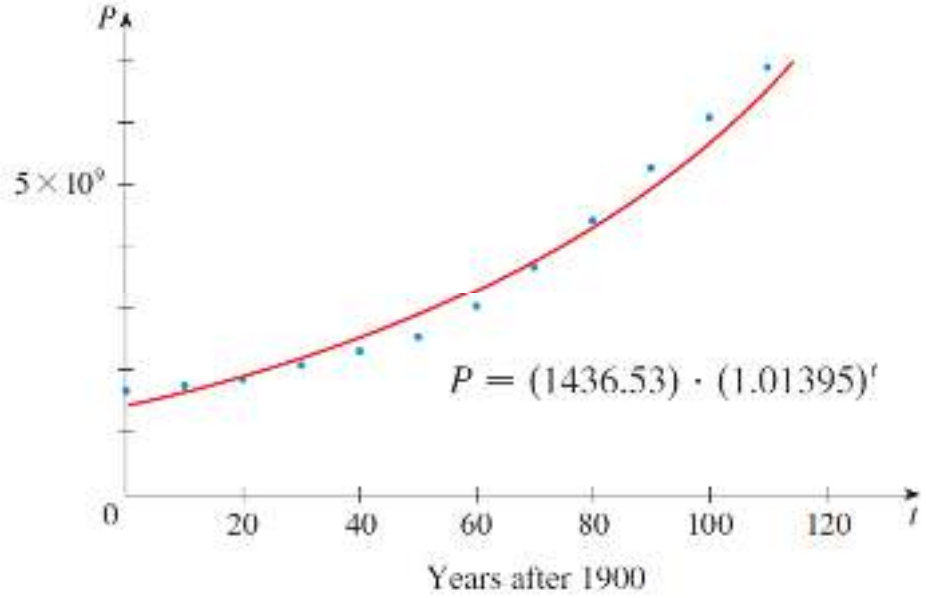
$$\log_{10} P(t) = 9.1573 + 0.006\,t$$

$$P(t) = 10^{9.1573+0.006\,t}$$

$$= 10^{9.1573} \cdot 10^{0.006\,t}$$

$$= 1436481377 \cdot (10^{0.006})^{t}$$

$$= 1436481377 \cdot (1.013911386\,)^{t}$$



Semilog Plots

Example:

x	y	$Y = \log_{10} y$
-2	0.1	$\log_{10}(10^{-1}) = -1$
-1	1	$\log_{10} 1 = 0$
0	10	$\log_{10} 10 = 1$
1	100	$\log_{10}(10^{2}) = 2$

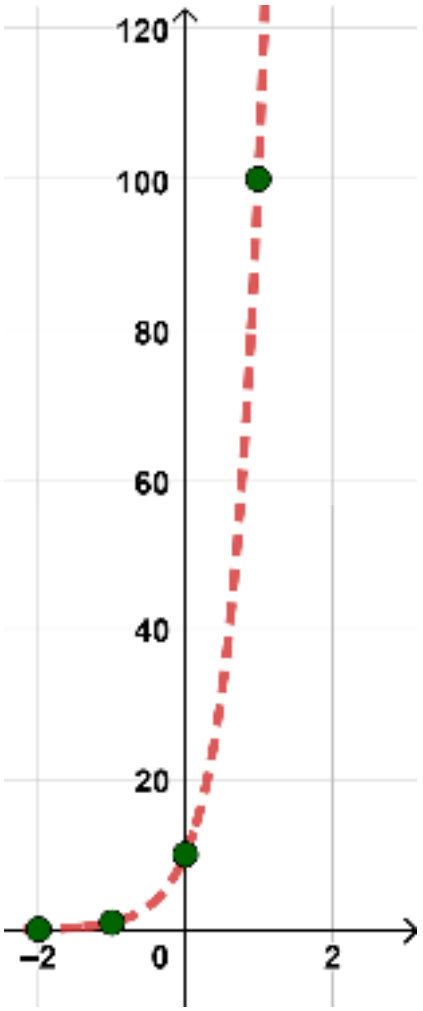
$$Y = x + 1$$

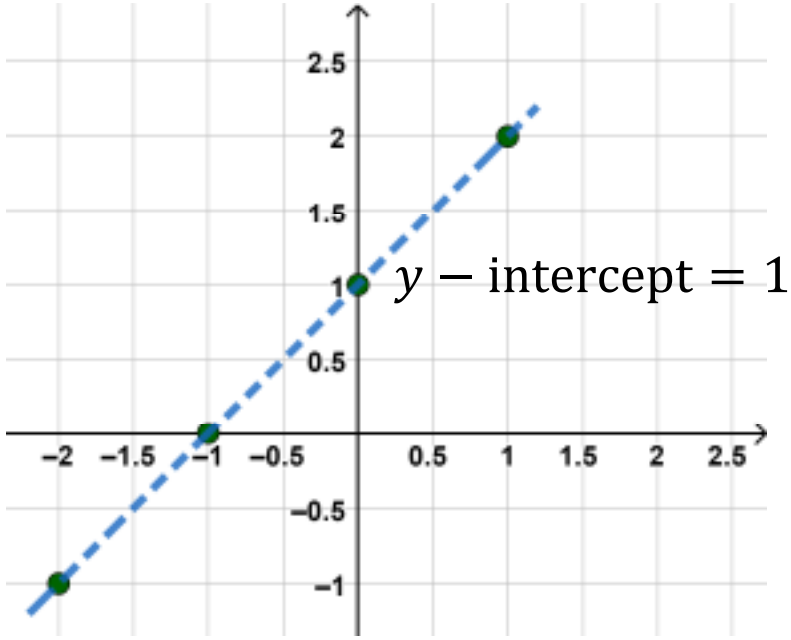
$$\log_{10} y = x + 1$$

$$y = 10^{x+1}$$

$$y = 10 \cdot 10^{x}$$

Exponential





$y - \text{intercept} = 1$

$$\text{slope } m = \frac{1 - 0}{0 - (-1)} = 1$$

$$Y = x + 1$$

Log-Log Plots

- If we use logarithmic scales on both the horizontal and vertical axes, the resulting graph is called a **log-log plot**.
- It is used when we suspect that a **power function** might be a good model for our data.
- If we start with a power function $y = C\,x^p$ and take logarithms of both sides, we get

$$y = C \cdot x^p$$

$$\log_{10} y = \log_{10}(C \cdot x^p)$$

$$\log_{10} y = \log_{10} C + \log_{10}(x^p)$$

$$\underbrace{\log_{10} y}_{Y} = \underbrace{\log_{10} C}_{A} + p\,\underbrace{\log_{10} x}_{X}$$

$$Y = A + p\,X$$

which is the equation of a line with slope p and Y –intercept A. So, the points $(\log_{10} x\,, \log_{10} y)$ lie on a straight line.

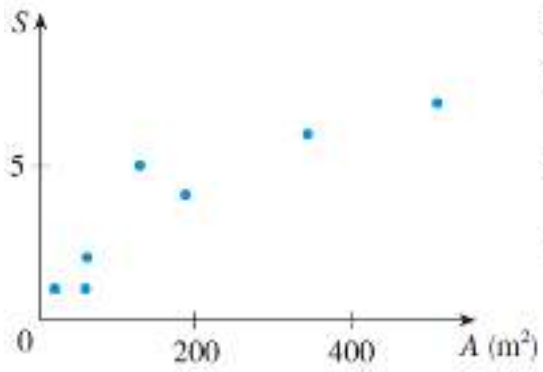
Log-Log Plots

Example: The following table gives the areas of several caves in central Mexico and the number of bat species that live in each cave.

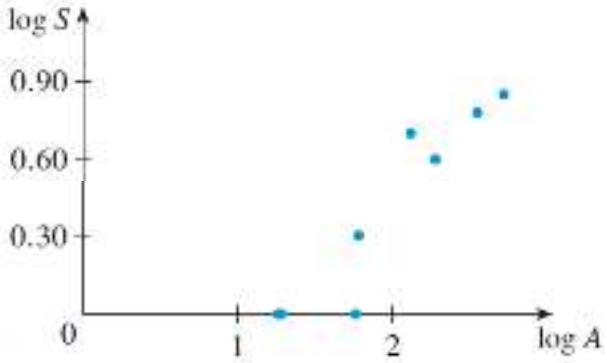
a) Make a scatter plot and a log-log plot of the data.

Cave	Area (m^2)	Number of Species
La Escondida	18	1
El Escorpion	19	1
El Tigre	58	1
Misión Imposible	60	2
San Martin	128	5
El Arenal	187	4
La Ciudad	344	6
Virgen	511	7

Let A denote the surface area of a cave and S the number of bat species in the cave.



Scatter



Log-Log

log A	log S
1.26	0
1.28	0
1.76	0
1.78	0.30
2.11	0.70
2.27	0.60
2.54	0.78
2.71	0.85

Log-Log Plots

Cave	Area (m^2)	Number of Species
La Escondida	18	1
El Escorpion	19	1
El Tigre	58	1
Misión Imposible	60	2
San Martin	128	5
El Arenal	187	4
La Ciudad	344	6
Virgen	511	7

Example: The following table gives the areas of several caves in central Mexico and the number of bat species that live in each cave.

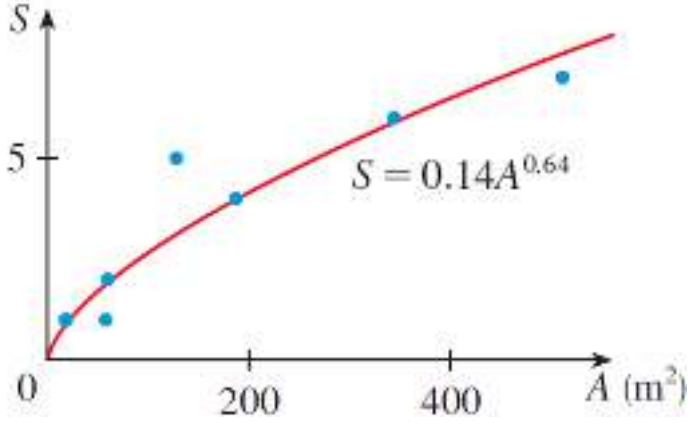
b) Is a power model appropriate? If so, find an expression for it.

It appears that $\log_{10} S$ is approximately a linear function of $\log_{10} A$. With a computer, we get the linear model

$$\log_{10} S = 0.64 \log_{10} A - 0.86$$

$$S = 10^{0.64 \log_{10} A - 0.86} = 10^{\log_{10}(A^{0.64})} \cdot 10^{-0.86}$$

$$S = 0.14\, A^{0.64}$$



Summary: Linear, Exponential, or Power Model?

To determine whether a linear, exponential, or power model is appropriate, we make a **scatter plot**, a **semilog plot**, and a **log-log plot**.

- If the *scatter plot* of the data lies approximately on a line, then a linear model is appropriate.
- If the *semilog plot* of the data lies approximately on a line, then an exponential model is appropriate.
- If the *log-log plot* of the data lies approximately on a line, then a power model is appropriate.

Mathematics and Biostatistics

Chapter: [11]
Descriptive Statistics

Section: [11.1]
Numerical Descriptions of Data

Data Consist of *information* coming from *observations*, *counts*, *measurements*, or *responses*.

- Each data element is called an **individual**.
- The property of an individual that is measured is called a **variable**.

Statistics The science of **collecting**, **organizing**, **analyzing**, and **interpreting** data in order to *make decisions*.

Population The collection of **all** outcomes, responses, measurements, or counts that are of interest.

Sample A **subset**, or part, of the population.

Parameter

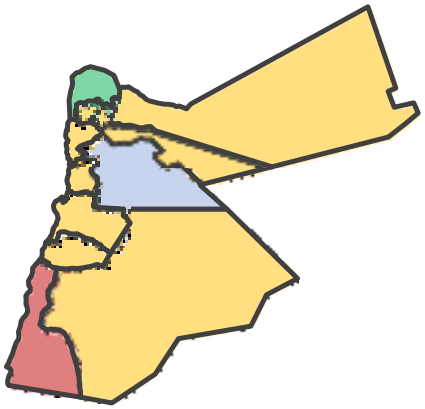
A numerical description of a population characteristic.

Average age of all people in JORDAN. (μ)

Statistic

A numerical description of a sample characteristic.

Average age of people from a sample of three cities. ($\bar{x}$)



Branches of Statistics

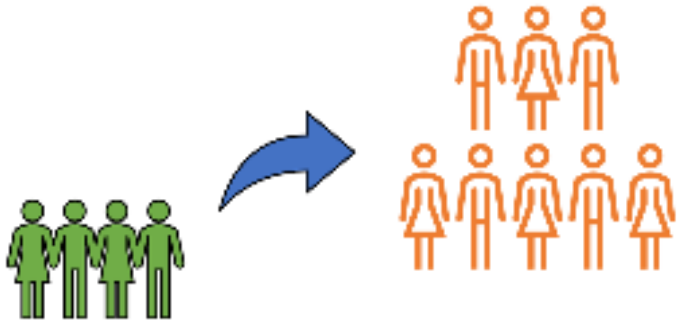
Descriptive Statistics

Involves organizing, summarizing, and displaying data.

Tables, charts, averages.

Inferential Statistics

Involves using sample data to draw conclusions about a population.



Types of Data

[1] Qualitative (Categorical)

Consists of all attributes, labels, or nonnumerical entries.

Blood Type (Class)
O
A
B
AB

Example: blood type, major, place of birth, eye color, ID number.

Note: The mathematical computations between data are not meaningful.

Two types of qualitative data:

1. ***Nominal:*** No order. For example: blood type, hair color.
2. ***Ordinal:*** Can be ordered. For example: best five football team in the world.

Types of Data

[2] Quantitative (Numerical)

Numerical measurements or counts.

Example: ages, temperature, weights, study-year.

Two types of quantitative data:

1. ***Discrete:*** the data that takes certain values like study-year: 1, 2, 3, 4, 5.
2. ***Continuous:*** the data that takes any value in a range or interval like weights.

Measures of Central Tendency

- Because numerical data can display many more patterns than categorical data, many different summary statistics can be obtained for numerical variables.
- One way to make sense of such data is to find a typical number in the "**center**" of the data. Any such number is called a *measure of central tendency* (مقياس نزعة مركزية).
- One of the most common measures of central tendency is the **average** (or the *mean*).

Definition Let $x_1, x_2, \cdots, x_n$ be n data points. The **mean**, or **average**, denoted by $\bar{x}$, is the sum of the values of x divided by n:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Example The lengths of 10 eggs from cuckoo birds were measured, giving the data below (in mm). Calculate the mean egg length.

$$19 \quad 19 \quad 20 \quad 22 \quad 22 \quad 22 \quad 23 \quad 24 \quad 25 \quad 35$$

$$\bar{x} = \frac{19 + 19 + 20 + 22 + 22 + 22 + 23 + 24 + 25 + 35}{10} = \frac{231}{10} = 23.1$$

- Another measure of central tendency is the **median**, which is *the middle number of an ordered list of numbers*.

Definition Let $x_1, x_2, \cdots, x_n$ be n data points written in **ascending** (*increasing*) order.

a) If n is **odd**, the median is the *middle number*.

b) If n is **even**, the median is the *average of the two middle numbers*.

Example Calculate the median of each data set:

a) 312, 257, 421, 289, 526, 374, 497.

Sort first: 257, 289, 312, 374, 421, 497, 526. **Median** $= \mathbf{374.}$

b) 7, 8, 9, 10, 11, 12, 13, 13, 14, 17, 17, 45.

Median $= \frac{\mathbf{12+13}}{\mathbf{2}} = \mathbf{12.5}$

- The *mode* of a data set is a third measure of central tendency, but it is usually less informative than the mean or median.

Definition The **mode** of a data set is the element that appears most often in the data set.

Example Calculate the mode of each data set:

a) 77, 82, 74, 81, 79, 84, 74, 78. **Mode:** 74.

b) 18, 19, 27, 22, 29, 19, 25, 21, 22, 30. **Two Modes:** 19 and 22.

c) 5, 9, 2, 4, 3, 8. **No Mode**.

Measures of Spread

- **Measures of spread (الانتشار)** (also called **measures of dispersion (التشتت)**) describe the *spread*, or *variability*, of the data around a central value.
- For example, each of the following data sets has mean 72, but it is clear that the first set of data has more variability than the second.

Data set 1:	50	58	78	81	93
Data set 2:	72	71	72	72	73

- The most important measures of spread in statistics are the **standard deviation (الانحراف المعياري)** and **variance (التباين)**.

Measures of Spread

Definition Let $x_1, x_2, \cdots, x_n$ be n data points, and let $\bar{x}$ be their mean. The **standard deviation** of the data is

$$\text{s.d.} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x-\bar{x})^2} = \sqrt{\frac{1}{n}\left(\sum_{k=1}^{n}(x^2) - \frac{1}{n}\left(\sum_{k=1}^{n} x\right)^2\right)} \quad \text{(for population)}$$

The **variance** is $(\text{s.d.})^2$, the square of the standard deviation.

- Both the variance and the standard deviation **cannot** *be negative*.
- The quantity $x - \bar{x}$ in the above formula is called the **deviation of the x value from the mean**. The *sum* of the deviations of the x values from the mean is always *zero*.

Measures of Spread

Example Find the variance and standard deviation of the data **4, 1, 7, 3, 5**.

- Mean: $\bar{x} = \frac{4+1+7+3+5}{5} = 4.$
- $\text{s.d.} = \sqrt{\frac{1}{5} \times 20} = \sqrt{4} = 2$
- Variance $= 2^2 = 4$

x	$x - \bar{x}$	$(x - \bar{x})^2$
4	0	0
1	–3	9
7	3	9
3	–1	1
5	1	1
Total	**0**	**20**

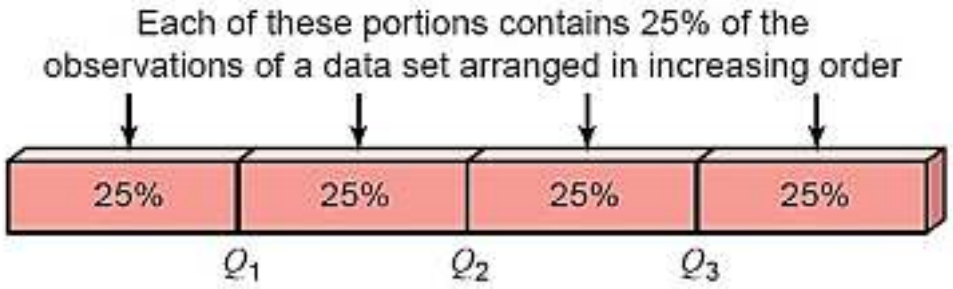
Example Find the variance and standard deviation of the data **6, 12, 6, 4**.

- $\text{s.d.} = \sqrt{\frac{1}{4} \times \left(232 - \frac{1}{4} \times 28^2\right)} = \sqrt{9} = 3$
- Variance $= 3^2 = 9$

x	x^2
6	36
12	144
6	36
4	16
28	**232**

The Five-Number Summary

- Another simple indicator of the *spread* of data is the *location* of the **minimum** and **maximum** values.
- Other indicators of *spread* are the **quartiles (الربيعيات)**.
- The *median of the lower half* of the data is called the **first quartile, Q_1**. The *median of the upper half* of the data is called the **third quartile, Q_3**.
- The median of the data is also called the **second quartile, Q_2**.

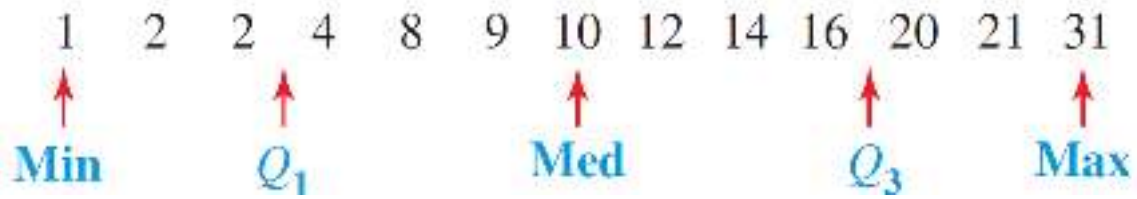


The Five-Number Summary

Definition The **five-number summary** for a data set is the set of these five numbers, written in the indicated order:

Minimum **Q_1** **Median** **Q_3** **Maximum**

Example Find the five-number summary of the data set



$$Q_1 = \frac{2+4}{2} = 3 \qquad Q_3 = \frac{16+20}{2} = 18$$

The Five-Number Summary

Example Find the five-number summary of the data set

- A simple indicator of spread is the **range**, which is the difference between the maximum and minimum values:

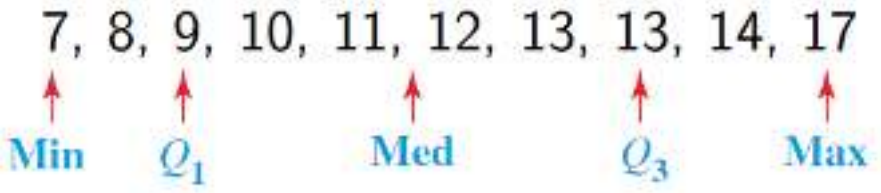
Range = Maximum – Minimum

- This can be compared to the spread of the middle of the data as measured by the **interquartile range (IQR)**, which is the difference between the third and the first quartiles:

$IQR = Q_3 - Q_1$

The Five-Number Summary

Example Find the range and the IQR of the data set

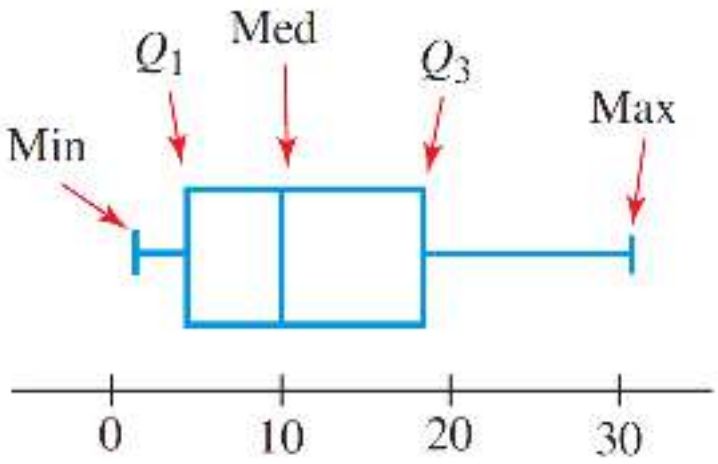


1) Range = 17 – 7 = 10.
2) IQR = $Q_3 - Q_1$ = 13 – 9 = 4.

The Five-Number Summary

The Box Plot

- A **box plot** (also called a **box-and-whisker plot**) is a method for graphically displaying the five-number summary.
- The plot consists of a rectangle whose left- and right-hand sides correspond to Q_1 and Q_3, respectively.
- The box is divided by a vertical line segment at the location of the median.



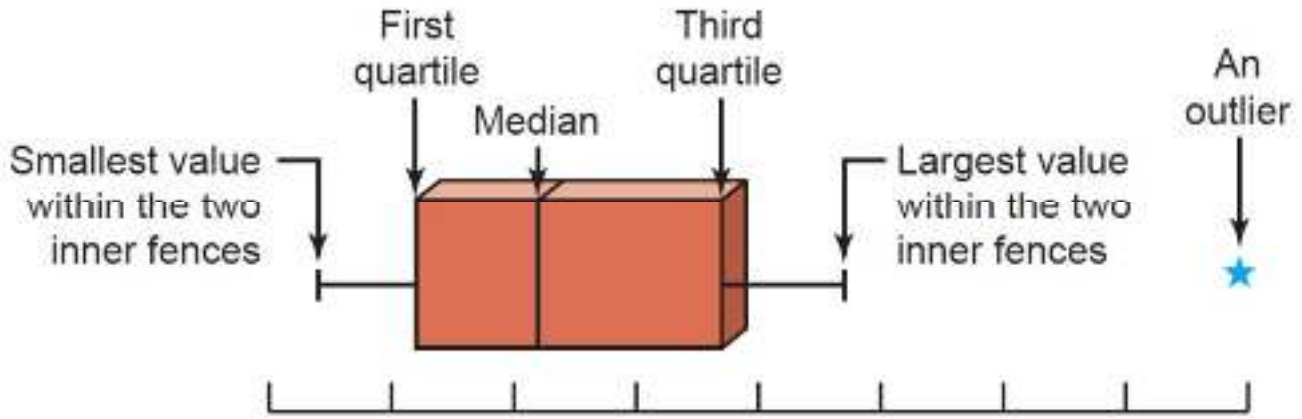
- The whiskers are horizontal line segments that extend from both edges of the box, to the location of the maximum and minimum values.

Outliers

- **Outlier** is a data entry that is far away from the other entries in the data set.
- One way to identify outliers is to use the interquartile range as follows:
  1) Find the IQR = $Q_3 - Q_1$.
  2) Find the **lower** and **upper fences** of the data using:
     - Lower Fence = $Q_1 - (1.5 \times \text{IQR})$
     - Upper Fence = $Q_3 + (1.5 \times \text{IQR})$
  3) Any data entry x is an outlier if $x >$ Upper Fence or $x <$ Lower Fence.

Outliers

Another common **box plot** convention is to extend the whiskers to the largest and smallest data points that are not outliers. Outliers are then identified with single points. We refer to this as an outlier box plot.



Outliers

Example Are there any outliers for the following data set?

6, 7, 8, 9, 10, 15, 16, 16, 20, 20, 23, 33, 50, 58, 104.

1. The data is **sorted**.
2. $Q_1 = 9$ and $Q_3 = 33$. So, IQR = 33 – 9 = 24.
3. Lower Fence = $Q_1 - (1.5 \times \text{IQR}) = 9 - 36 = -27$.
4. Upper Fence = $Q_3 + (1.5 \times \text{IQR}) = 33 + 36 = 69$.
5. Since 104 > 69 then the data entry 104 is an outlier.

Mathematics and Biostatistics

Chapter: [11]
Descriptive Statistics

Section: [11.2]
Graphical Descriptions of Data

Frequency Distribution

- **Frequency Distribution** is a table that shows *classes* or intervals of data entries with a *count* of the number of entries in each class called *frequencies*.

- **Example:** The following is the blood type of 12 students.

O	AB	A	A
B	O	A	O
AB	B	B	O

Class	Frequency (F)
O	4
A	3
B	3
AB	2
Total	**12**

Frequency Distribution

- **Relative frequency of a class** $= \frac{\text{Class Frequency}}{\text{Data Size}} = \frac{\text{F}}{n}$.
- **Cumulative frequency of a class** $=$ the sum of frequencies of that class and all previous classes.

- **Example:** The following is the blood type of 12 students.

O	AB	A	A
B	O	A	O
AB	B	B	O

Class	Frequency (F)	Relative Frequency	Cumulative Frequency
O	4	$\frac{4}{12} = 0.333$	4
A	3	$\frac{3}{12} = 0.250$	$4 + 3 = 7$
B	3	$\frac{3}{12} = 0.250$	$3 + 7 = 10$
AB	2	$\frac{2}{12} = 0.167$	$2 + 10 = 12$
Total	**12**	**1.000**	

Displaying Categorical (Qualitative) Data

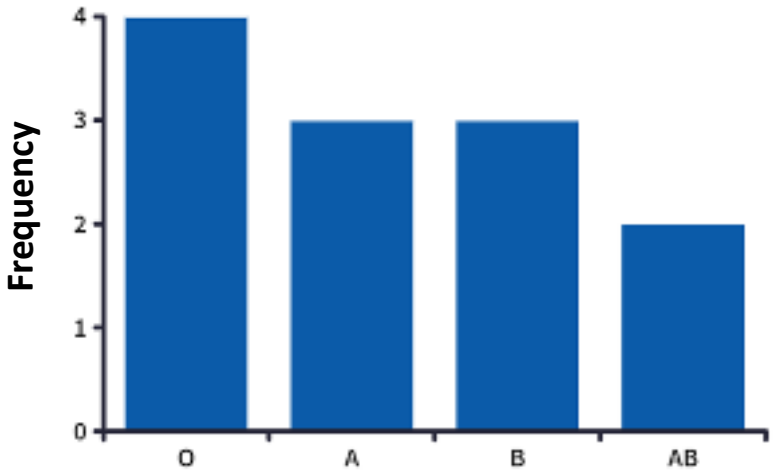
- Two simple visual summaries are often used for categorical data: *bar graphs* and *pie charts*.
- A **bar graph** consists of vertical bars, one bar for each category.
  - ✓ The height of each bar is proportional to the number of individuals in that category.
  - ✓ So the vertical axis has a numerical scale corresponding to the number (or sometimes the proportion) of individuals in each category.
  - ✓ The labels on the horizontal axis describe the categories.

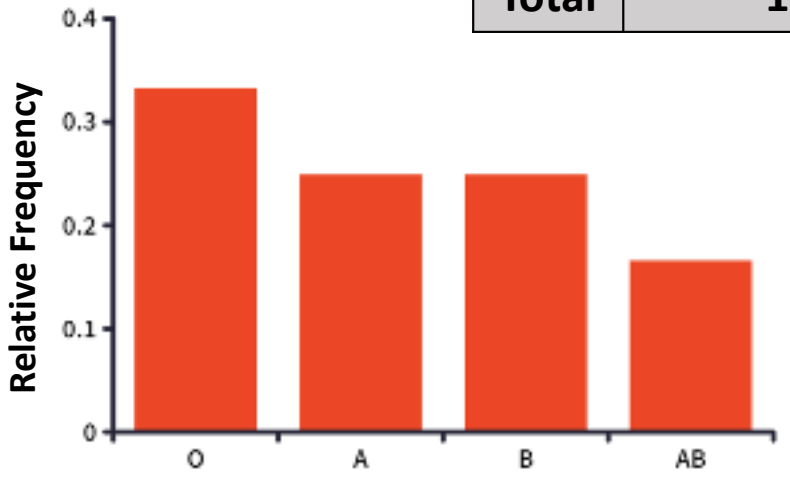
Displaying Categorical (Qualitative) Data

- **Example:** The following is the blood type of 12 students.

O	AB	A	A
B	O	A	O
AB	B	B	O

Class	Frequency (F)
O	4
A	3
B	3
AB	2
Total	**12**





Displaying Categorical (Qualitative) Data

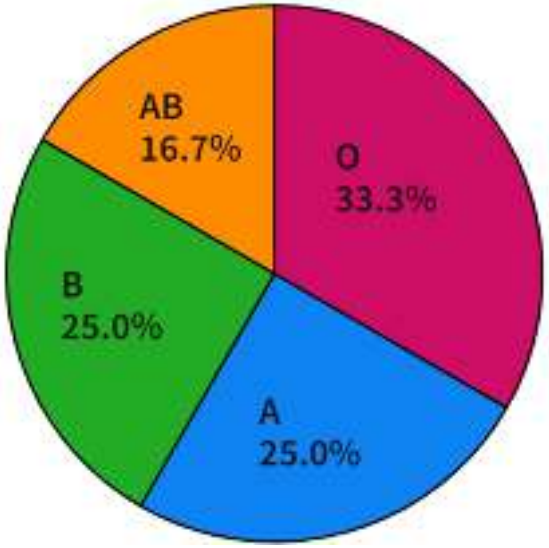
- **Pie Chart** is a circle that is divided into *sectors* that represent classes.
  - ✓ The area of each sector is *proportional* to the frequency (or relative frequency) of each class.
  - ✓ Find the **central angle** θ for each sector using the corresponding frequency or relative frequency of that class, where

$$\theta = \frac{\text{Frequency}}{\text{Data Size}} \times 360^\circ = \text{Relative Frequency} \times 360^\circ$$

Displaying Categorical (Qualitative) Data

- **Example:** The following is the blood type of 12 students.

O	AB	A	A
B	O	A	O
AB	B	B	O



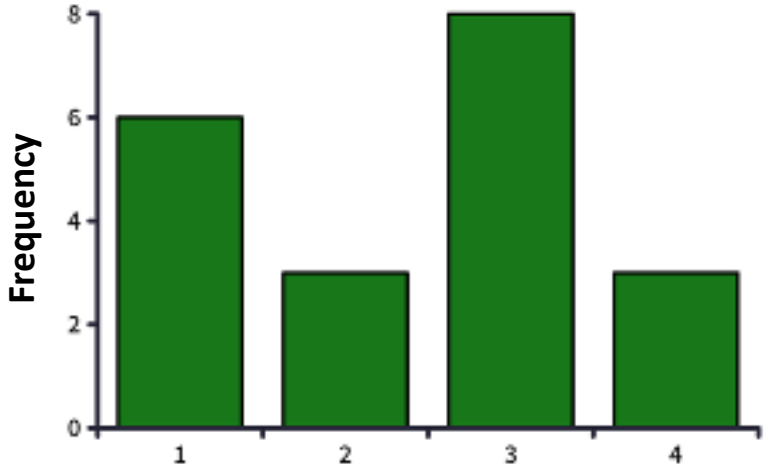
Class	Frequency (F)	Relative Frequency	θ
O	4	0.333	$\frac{4}{12} \times 360 = 120$
A	3	0.250	$\frac{3}{12} \times 360 = 90$
B	3	0.250	$\frac{3}{12} \times 360 = 90$
AB	2	0.167	$\frac{2}{12} \times 360 = 60$
Total	**12**	**1.000**	$\mathbf{360^\circ}$

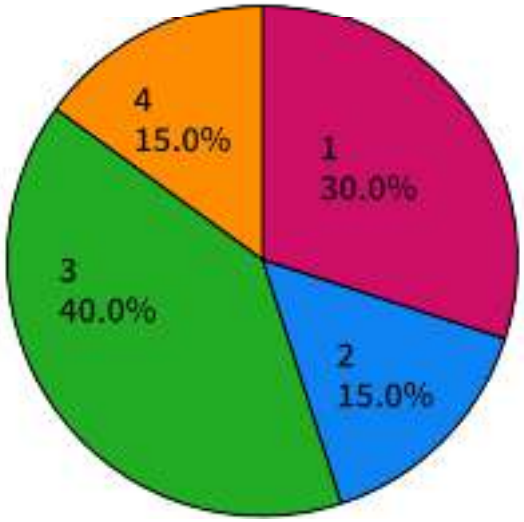
Displaying Numerical Discrete Data

Example: Summarize the following data using a frequency distribution, then draw bar graph and pie chart for these data.

1	1	4	3	4
3	1	3	2	1
3	3	1	2	3
1	3	4	3	2

Value	Frequency (F)	Relative Frequency	Cumulative Frequency	θ
1	6	0.30	6	108
2	3	0.15	9	54
3	8	0.40	17	144
4	3	0.15	20	54
Total	**20**	**1.000**		**360**





Displaying Numerical Continuous Data: Histogram

The goal is to group the observations according to *intervals* and recording the frequencies of the intervals.

Example: The table below shows regular frequency distribution of continuous data.

Class	F
1-5	5
6-10	8
11-15	6
16-20	8
21-25	5
26-30	4

Notes:

- The classes do not *overlap*.
- Each class has a
  - ✓ *lower class limit*, which is the least number that can belong to the class. The lower class limits are 1, 6, 11, 16, 21, and 26.
  - ✓ *an upper class limit*, which is the greatest number that can belong to the class. The upper class limits are 5, 10, 15, 20, 25, and 30.
- *Class Length* = Upper Limit of a Class − Lower Limit of that Class. In our example, the class length is $5 - 1 = 4$.

Displaying Numerical Continuous Data: Histogram

Class	F
1-5	5
6-10	8
11-15	6
16-20	8
21-25	5
26-30	4

Notes (*continue*):

- *Class Width* = distance between lower (or upper) limits of consecutive classes. For example, the class width in the frequency distribution shown is $6 - 1 = 5$.
- **Class Length = Class Width − 1**.
- Range = Maximum Data − Minimum Data.

Displaying Numerical Continuous Data: Histogram

Example: For the following set of data, construct a frequency distribution that has *seven classes*.

100	180	200	90	271
85	65	230	150	120
130	80	230	126	132
112	90	341	170	190

STEP 1: Decide on the number of classes to include in the frequency distribution. The number of classes should be between 5 and 20; otherwise, it may be difficult to detect any patterns.

- The number of classes $= 7$ which is stated in the problem.

STEP 2: Determine the minimum and maximum data, then evaluate the range.

- Min $= 65$ and Max $= 341$. So, the range $= 341 - 65 = 276$.

Displaying Numerical Continuous Data: Histogram

Example: For the following set of data, construct a frequency distribution that has *seven classes*.

100	180	200	90	271
85	65	230	150	120
130	80	230	326	132
112	90	341	170	190

- The number of classes $= 7$.
- Range $= 276$.

STEP 3: Find the class width using the formula

$$\text{Class Width} = \frac{\text{Range}}{\text{Number of Classes}}$$

and your answer should be *rounded up* if it is *not an integer*.

- $\text{Class Width} = \frac{276}{7} \approx 39.43 \rightarrow 40.$

STEP 4: Determine the class length.

- Class Length $=$ Class Width $- 1 = 40 - 1 = 39.$

Displaying Numerical Continuous Data: Histogram

Example: For the following set of data, construct a frequency distribution that has *seven classes*.

100	180	200	90	271
85	65	230	150	120
130	80	230	126	132
112	90	341	170	190

- The number of classes $= 7$.
- Range $= 276$.
- Class Width $= 40$.
- Class Length $= 39$.

STEP 5: Find the class limits. You can use the minimum data entry as the lower limit of the first class. To find the remaining lower limits, add the class width to the lower limit of the preceding class. Then find the upper limit of the first class.

STEP 6: Find the frequency F for each class.

Class	F
65-104	6
105-144	5
145-184	3
185-224	2
225-264	2
265-304	1
305-344	1

Displaying Numerical Continuous Data: Histogram

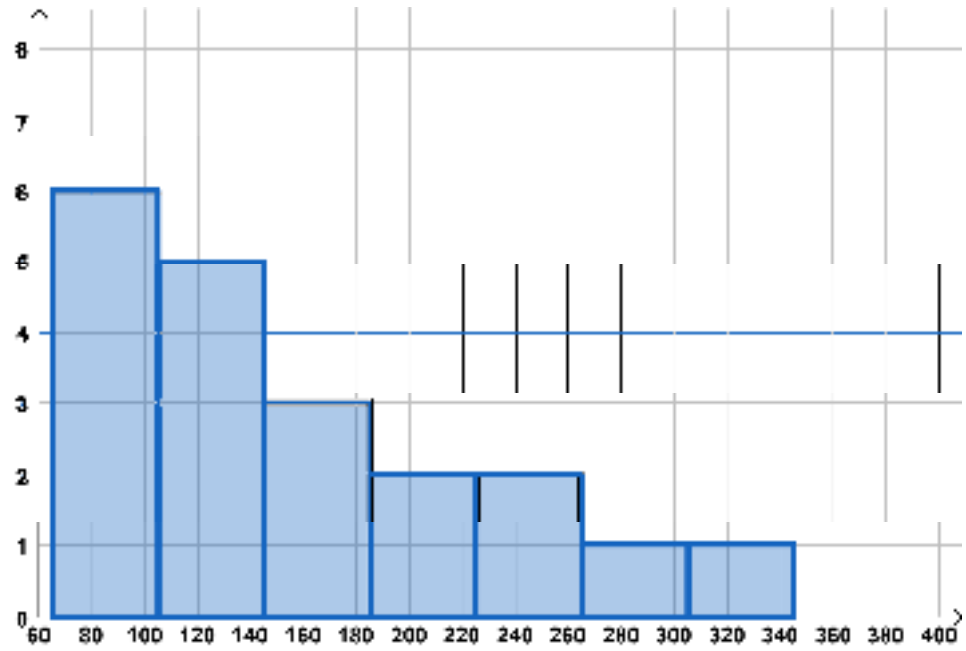
Example: For the following set of data, construct a frequency distribution that has *seven classes*.

100	180	200	90	271
85	65	230	150	120
130	80	230	126	132
112	90	341	170	190

Class	F	Relative Frequency	Cumulative Frequency
65-104	6	6/20 = 0.30	6
105-144	5	5/20 = 0.25	11
145-184	3	3/20 = 0.15	14
185-224	2	2/20 = 0.10	16
225-264	2	2/20 = 0.10	18
265-304	1	1/20 = 0.05	19
305-344	1	1/20 = 0.05	20
Total	**20**	**1**	

Displaying Numerical Continuous Data: Histogram

Histogram is a *bar graph* that represents the frequency distribution of a data set.



A histogram has the following properties:

- The horizontal scale is quantitative and measures the data entries.
- The vertical scale measures the frequencies of the classes.
- Consecutive bars must touch.

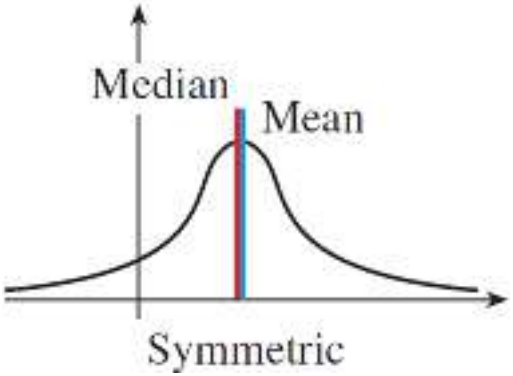
Displaying Numerical Continuous Data: Histogram

- Because consecutive bars of a histogram must touch, bars must begin and end at **class boundaries** instead of *class limits*.
- Class boundaries are the numbers that separate classes without forming gaps between them.
- For data that are *integers*, **subtract 0.5** from each *lower limit* to find the **lower class boundaries**.
- To find the **upper class boundaries**, **add 0.5** to each *upper limit*.
- The upper boundary of a class will equal the lower boundary of the next higher class.

Class	F	Class Boundaries
65-104	6	64.5-104.5
105-144	5	104.5-144.5
145-184	3	144.5-184.5
185-224	2	184.5-224.5
225-264	2	224.5-264.5
265-304	1	264.5-304.5
305-344	1	304.5-344.5

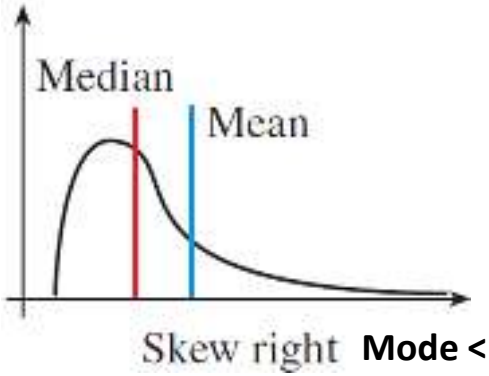
Displaying Numerical Continuous Data: Histogram

The histogram allows us to determine whether the data are **symmetric (Bell Shaped, Normal)** *around the mean* or whether the data are **skewed**.



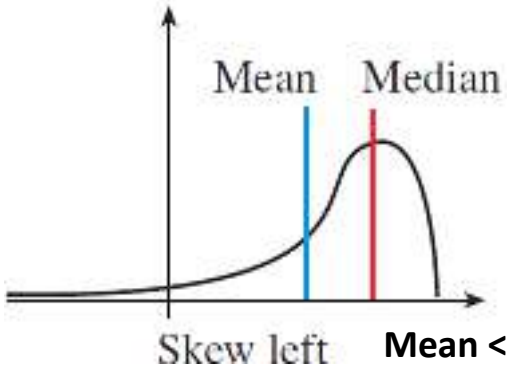
Mean = Median = Mode

If the histogram has a **long 'tail'** on the **right**, we say that the data are *skewed to the right*.



Mode < Median < Mean

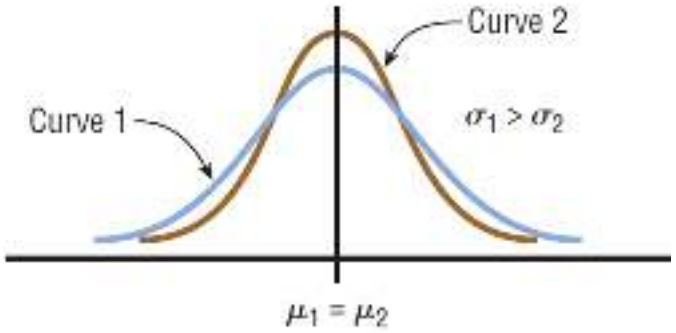
If there is a **long tail** to the **left**, the data are *skewed to the left*.



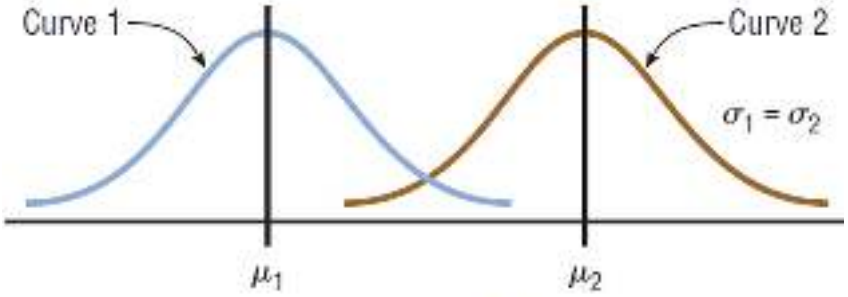
Mean < Median < Mode

Normal (Bell-Shaped) Distribution

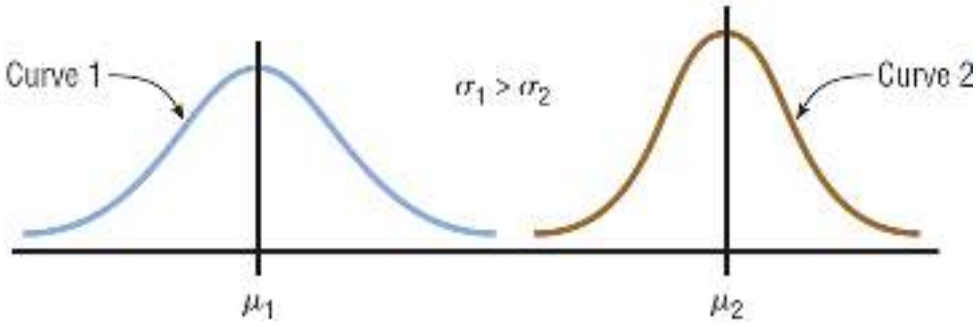
- All normal curves are symmetric around their centers and have the same "bell" shape.
- The *shape* and *position* of a normal distribution curve depend on two parameters, the **mean (μ)** and the **standard deviation (σ)**.
- The **standard normal distribution** is a normal distribution with a *mean* of **0** and a *standard deviation* of **1**.



(a) Same means but different standard deviations



(b) Different means but same standard deviations

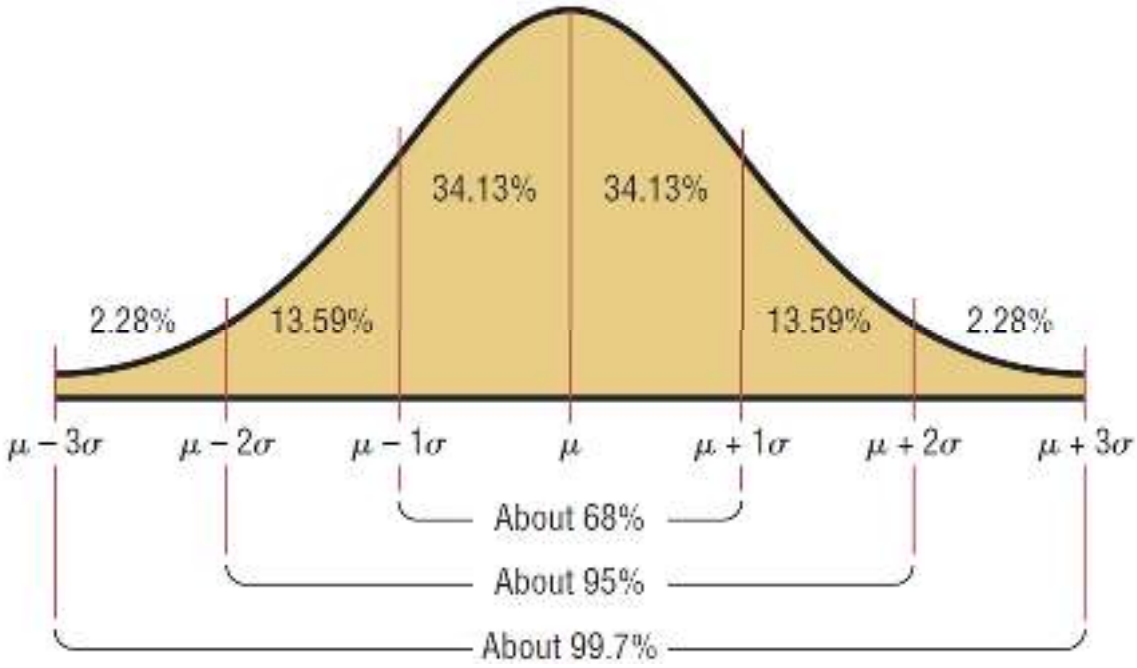


(c) Different means and different standard deviations

Areas Under a Normal Distribution Curve (*Empirical Rule*)

The **area** under the part of a normal curve that lies

- within **1 standard** deviation of the mean is approximately 0.68, or **68%**;
- within **2 standard deviations**, about 0.95, or **95%**;
- and **within 3 standard deviations**, about 0.997, or **99.7%**.



Areas Under a Normal Distribution Curve (*Empirical Rule*)

Example: The age distribution of a sample of 5000 persons is bell shaped with a **mean** of **40** years and a **standard deviation** of **12** years. Determine the approximate percentage of people who are 16 to 64 years old.

$$16 = \mu - k \cdot \sigma$$
$$16 = 40 - 12k$$
$$k = 2$$

$$64 = \mu + k \cdot \sigma$$
$$64 = 40 + 12k$$
$$k = 2$$

- Since $k = 2$ then the interval $[16,64]$ contains approximately 95% of the observations.
- Note that $0.95 \times 5000 = 4750$ persons.

Mathematics and Biostatistics

Chapter: [11]
Descriptive Statistics

Section: [11.3]
Relationships between Variables

Introduction

- Often more than one variable is *measured* for each individual in a sample with the objective being *to determine if there are any* ***relationships*** *among the variables*.
- In this section we begin to study these issues by exploring how relationships among variables in a data set can be *described*.
- We consider **three different kinds** of possible relationships:
  - ✓ relationships between two categorical variables,
  - ✓ between a categorical variable and a numerical variable,
  - ✓ and between two numerical variables.

Two Categorical Variables

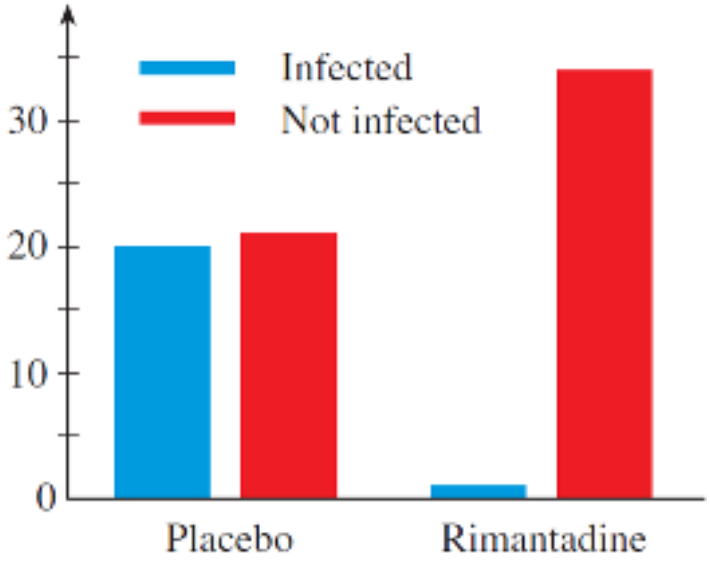
Example: Rimantadine is an antiviral drug that can be administered to prevent infection by influenza.

- A study divided a sample of 76 children into two groups and administered rimantadine to one group and a placebo to the other.
- The researchers then determined whether each child was infected with influenza during the flu season.
- These data can most easily be summarized using a **contingency table**.

	Infected	Not infected	Row total
Placebo	20	21	41
Rimantadine	1	34	35
Column total	21	55	76

Two Categorical Variables

Example: Rimantadine is an antiviral drug that can be administered to prevent infection by influenza.



	Infected	Not infected	Row total
Placebo	20	21	41
Rimantadine	1	34	35
Column total	21	55	76

1. How many infected children are there?
   Answer: 21
2. How many children take Rimantadine?
   Answer: 35
3. How many children took Rimantadine and were not infected by influenza?
   Answer: 34

Two Categorical Variables

- The table in the previous example suggests that “infection status” is contingent on “drug status” since almost none of the individuals receiving rimantadine became infected (only 1 out of 35) whereas approximately half of the individuals receiving the placebo became infected (20 out of 41).
- Do the data provide *good evidence* that, in the population at large, rimantadine protects against influenza? In Chapter 13 we will study how to answer this question. Doing so requires the use of **inferential statistics**.

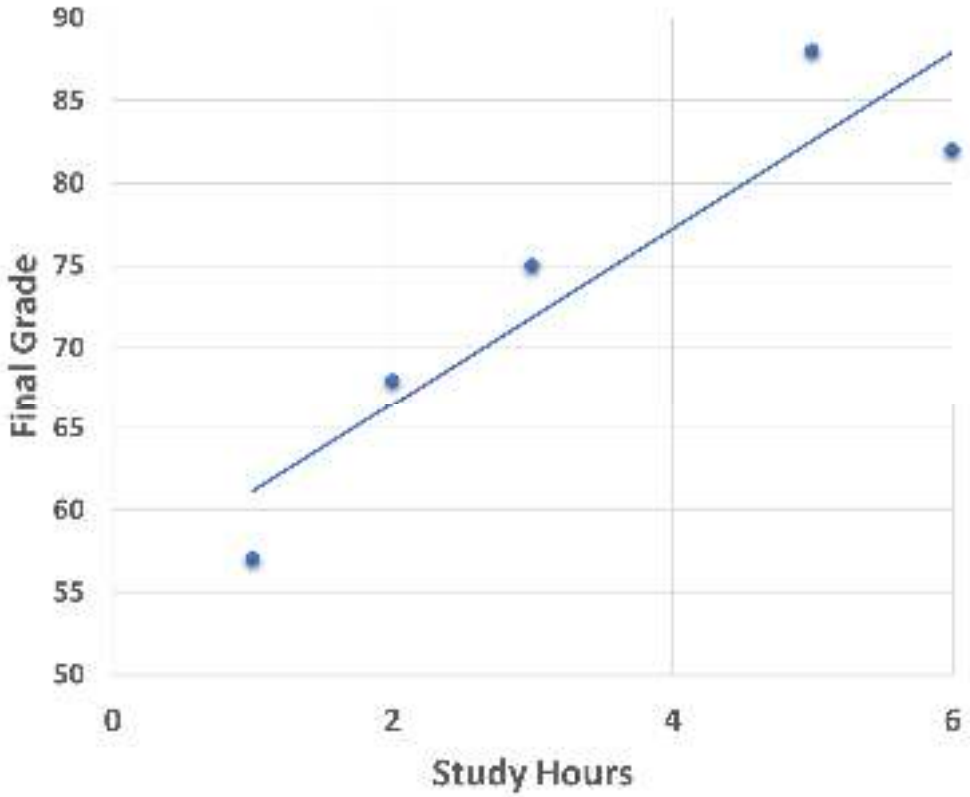
Two Numerical Variables

- When two numerical variables (label the two variables y and x) are measured on each individual, the possible relationships between them can be quite complicated.
- The best way to start is by constructing a **scatter plot**.
- A scatter plot is then a plot of y against x.
- Each individual in the data set is represented by a single point in the plane whose x −coordinate is its value of the x −variable, and whose y −coordinate is its value of the y −variable.

Two Numerical Variables

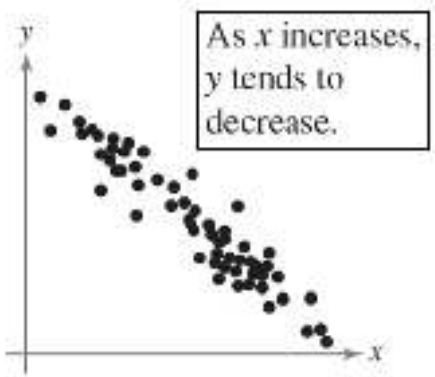
Example: The following table shows the study hours of 6-students and their final grades in a course.

Hours (x)	Grade (y)
6	82
1	57
5	88
2	68
3	75

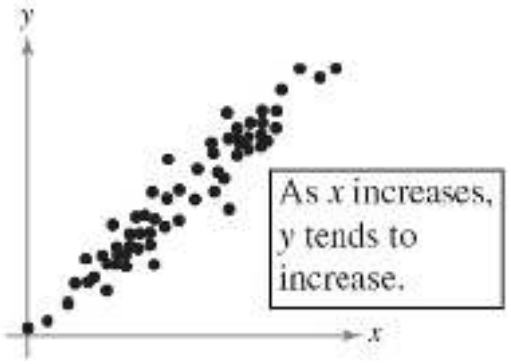


Two Numerical Variables

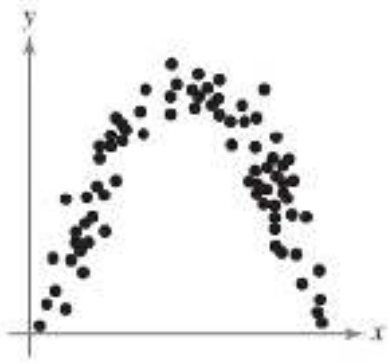
- We can see, from the scatter plot of the previous example, that the students that have higher hours of studying also tend to have higher final grades.
- **Correlation** is a relationship between two variables x and y, where x is the independent variable and y is the dependent variable.
- Scatter plots help to reveal correlations between numerical variables.



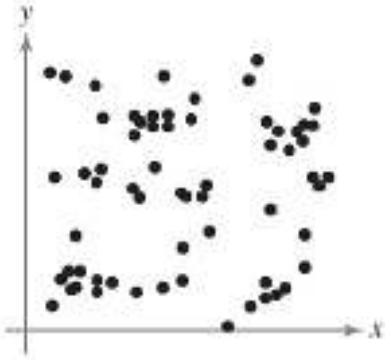
Negative Linear Correlation



Positive Linear Correlation



Nonlinear Correlation



No Correlation

Two Numerical Variables

- One of the most important kinds of relationships studied in statistics is a *linear* one.
- For instance, the data in the study hours and grades example suggest that the relationship between study hours and grades is approximately linear.
- One way to quantify this more formally is to draw a line through the scatter plot that, in some sense, best fits the data. But how do we find the equation of this line?
- The most common approach to fitting a line to data is called the **least-squares** fit or **linear regression**.

Linear Regression

Definition: Let $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ be the values of two numerical variables measured on n individuals. The regression line $y = ax + b$ for the data has a and b given by

$$a = \frac{\overline{xy} - (\bar{x})(\bar{y})}{\overline{x^2} - (\bar{x})^2}$$

$$b = \bar{y} - a \cdot \bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the average values of x and y, $\overline{x^2}$ is the average of the square of the x values, and $\overline{xy}$ is the average of the product of the $x-$ and $y-$values.

Two Numerical Variables

Example: The following table shows the study hours of 6-students and their final grades in a course. Find the regression line equation.

	x	y	x^2	xy
	6	82	36	492
	1	57	1	57
	5	88	25	440
	2	68	4	136
	3	75	9	225
Total	**17**	**370**	**75**	**1350**
Average	**3.4**	**74**	**15**	**270**

$$a = \frac{\overline{xy} - (\bar{x})(\bar{y})}{\overline{x^2} - (\bar{x})^2}$$
$$= \frac{270 - (3.4)(74)}{15 - (3.4)^2} \approx 5.349$$

$$b = \bar{y} - a \cdot \bar{x}$$
$$= 74 - (5.349)(3.4) \approx 55.814$$

$$\therefore \boldsymbol{y = 5.349\,x + 55.814}$$

Two Numerical Variables

Note: The regression line can be used to predict the value of y for a given value of x.

- For instance, in the previous example, if a student study for 4-hours, what is his predicted final grade?

$$\begin{aligned} y &= 5.349 \times 4 + 55.814 \\ &= 77.21 \end{aligned}$$

Mathematics and Biostatistics

Chapter: [12]
Probability

Section: [12.1]
Principles of Counting

Definitions

Probability Experiment — An action, or trial, through which specific results (counts, measurements, or responses) are obtained.

Outcome — The result of a single trial in a probability experiment.

Sample Space — The set of all possible outcomes of a probability experiment, and denoted by S or Ω.

Event — Consists of one or more outcomes and is a subset of the sample space, and denoted by capital letters $A, B, C, \cdots$

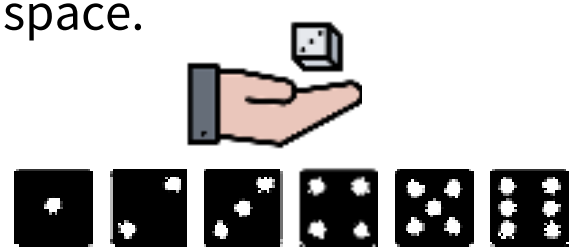
Some Examples

Example A probability experiment consists of tossing (flipping) a coin. Determine the number of outcomes and identify the sample space.

$$S = \{H, T\}$$

Example A probability experiment consists of rolling a six-sided die. Determine

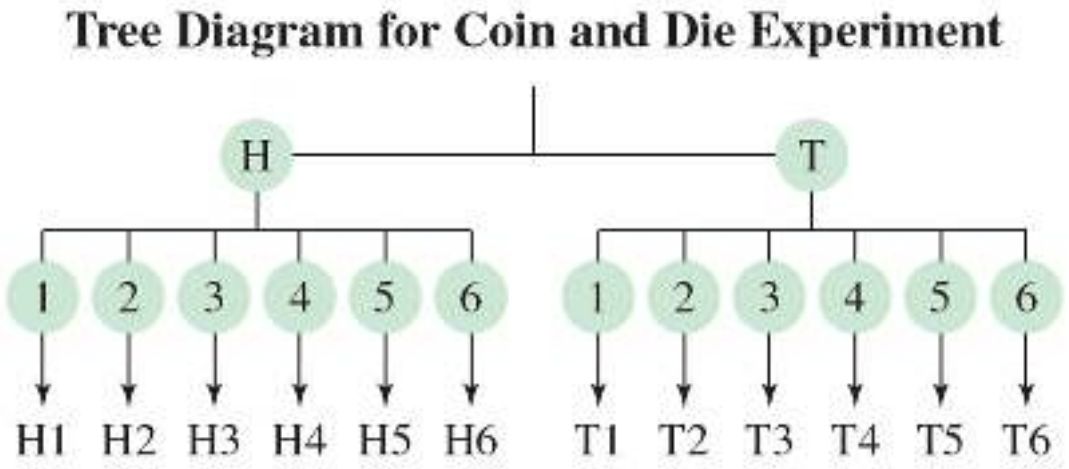
1. the number of outcomes and identify the sample space.
   $S = \{1,2,3,4,5,6\}$
2. the event that an odd appears.
   $A = \{1, 3, 5\}$
3. the event that a number greater than 5 appears.
   $B = \{6\}$ **Simple Event**
4. the event that a number less than or equal to 6 appears.
   $C = \{1,2,3,4,5,6\}$ **Sure Event**
5. the event that a number divisible by 10 appears.
   $D = \{\ \ \} = \Phi$ **Impossible Event**



Some Examples

Example Suppose you roll a die and flip a coin. Taken together, how many possible outcomes are there?

$$S = \{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}$$



Fundamental Principle of Counting

If one event can occur in m ways and a second event can occur in n ways, then the number of ways the two events can occur in sequence is $m \times n$. This rule can be extended to any number of events occurring in sequence.

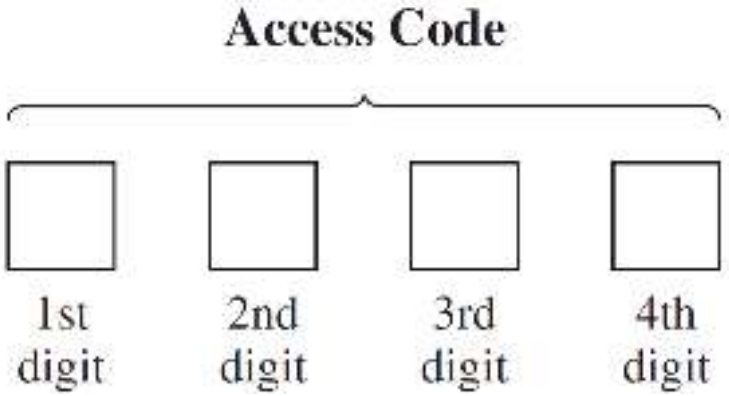
Example You are purchasing a new car. The possible manufacturers, car sizes, and colors are listed in the table. How many different ways can you select one manufacturer, one car size, and one color?

Manufacturer	Car size	Color
Ford	compact	white (W)
GM	midsize	red (R)
Honda		black (B)
		green (G)

Solution: There are three choices of manufacturers, two choices of car sizes, and four choices of colors. Using the Fundamental Counting Principle, there are $3 \times 2 \times 4 = 24$ ways.

Fundamental Principle of Counting

Example The access code for a car's security system consists of four digits. Each digit can be any number from 0 through 9. How many access codes are possible when



1. each digit can be used only once and not repeated?

Solution: There are $10 \times 9 \times 8 \times 7 = 5040$ ways.

2. each digit can be repeated?

Solution: There are $10 \times 10 \times 10 \times 10 = 10^4 = 10000$ ways.

3. each digit can be repeated but the first digit cannot be 0 or 1?

Solution: There are $8 \times 10 \times 10 \times 10 = 8000$ ways.

n −Factorial

Definition If n is a positive integer, then the expression $\boldsymbol{n!}$ is read as $\boldsymbol{n}$ **factorial** and is defined as follows.

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1$$

Examples

- $3! = 3 \times 2 \times 1 = 6$
- $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$
- $1! = 1$

Properties

- $0! = 1$
- $n! = n(n-1)! = n(n-1)(n-2)! = \cdots$

Example

$$\frac{6!}{4!\,2!} = \frac{6 \cdot 5 \cdot 4!}{4! \cdot 2} = 15$$

Permutations

Rule [1] Suppose that n distinct objects are to be **drawn sequentially**, or **ordered in a row**. Then, the number of ways to arrange n distinct objects in a row is $n!$.

Example In how many ways you can arrange the letters A, B, and C?

$3! = 6$

Example A witness reported that a car seen speeding away from the scene of the crime had a number plate that began with V or W, the digits were 4, 7 and 8 and the end letters were A, C, E. He could not however remember the order of the digits or the end letters. How many cars would need to be checked to be sure of including the suspect car?

$2! \times 3! \times 3! = 72$

Permutations

Rule [2] The number of different arrangements of n objects in which n_1 objects of them are the same, n_2 objects of them are the same, and so on, is

$$\frac{n!}{n_1! \cdot n_2! \cdots n_k!}$$

Example In how many ways you can arrange the letters A, A, B, and C?

$$\frac{4!}{2! \cdot 1! \cdot 1!} = 12$$

Example How many ways can we arrange the letters of the word **MISSISSIPPI**?

The word MISSISSIPPI contains 11 letters such that there are 4 I's, 4 S's, 2 P's, and 1 M. So, the total number of ways is

$$\frac{11!}{4! \cdot 4! \cdot 2! \cdot 1!} = 34650$$

Permutations

Rule [3] The number of ways of selecting m objects from n distinct objects where **order is important** is

$$P_m^n = \frac{n!}{(n-m)!}$$

Example In how many ways can we choose 3 numbers from the numbers 1, 2, 3, 4, and 5?

$$P_3^5 = \frac{5!}{(5-3)!} = \frac{5!}{2!} = \frac{5 \cdot 4 \cdot 3 \cdot 2!}{2!} = 60$$

Example A school musical director can select 2 musical plays to present next year. One will be presented in the fall, and one will be presented in the spring. If she has 9 to pick from, how many different possibilities are there?

$$P_2^9 = \frac{9!}{7!} = \frac{9 \cdot 8 \cdot 7!}{7!} = 72$$

Combinations

The Rule The number of ways of selecting m objects from n distinct objects where **order is NOT important** is

$$C_m^n = \frac{n!}{(n-m)!\ m!}$$

Example How many ways can an adviser choose 4 students from a class of 12 if they are all assigned the same task?

$$C_4^{12} = \frac{12!}{(12-4)!\,4!} = \frac{12!}{8!\ 4!} = \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8!}{8! \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 495$$

Example In a club there are 7 women and 5 men. A committee of 3 women and 2 men is to be chosen. How many different possibilities are there?

$$C_3^7 \cdot C_2^5 = \frac{7!}{4! \cdot 3!} \times \frac{5!}{3! \cdot 2!} = 35 \times 10 = 350$$

Mathematics and Biostatistics

Chapter: [12]
Probability

Section: [12.2]
What Is Probability?

Definitions and Concepts

Probability as a general concept can be defined as the **chance** of an event occurring.

Types of Probability

- Classical (Theoretical) Probability
- Empirical (Statistical) Probability
- Subjective Probability

Notation The probability that event E will occur is written as $P(E)$ and is read as "the probability of event E."

Classical Probability is used when each outcome in a sample space is equally likely to occur. The classical probability for an event E is given by

$$P(E) = \frac{\text{Number of outcomes in event } E}{\text{Total number of outcomes in sample space}}$$
$$= \frac{n(E)}{n(S)}$$

Definitions and Concepts

Example You roll a six-sided die. Find the probability of each event.

1. Event A: rolling a 3.

   Solution: Since $S = \{1,2,3,4,5,6\}$ and $A = \{3\}$ then $P(A) = \frac{1}{6}$.

2. Event B: rolling a 7.

   Solution: Since $B = \{\quad\} = \phi$ then $P(B) = \frac{0}{6} = 0$.

3. Event C: rolling a number less than 5.

   Solution: Since $C = \{1,2,3,4\}$ then $P(C) = \frac{4}{6} = \frac{2}{3}$.

Example If we twice flip a balanced coin, what is the probability of getting at least one head?

Solution: Since $S = \{HH, HT, TH, TT\}$ and $A = \{HH, HT, TH\}$ then $P(A) = \frac{3}{4} = 0.75$.

Definitions and Concepts

Empirical Probability is based on observations obtained from probability experiments. The empirical probability of an event E is the relative frequency of event E.

$$P(E) = \frac{\text{Frequency of event } E}{\text{Total frequency}} = \frac{f}{n}$$

Example The blood type of students are given by the frequency table shown. If one student is randomly selected, what is the probability that this person's blood type is:

Blood Type	Freq.
O	8
A	5
B	6
AB	1
Total	**20**

1. AB?

   Solution: $P(\text{AB}) = \frac{1}{20} = 0.05$

2. either A or B?

   Solution: $P(\text{A or B}) = \frac{5+6}{20} = \frac{11}{20} = 0.55$

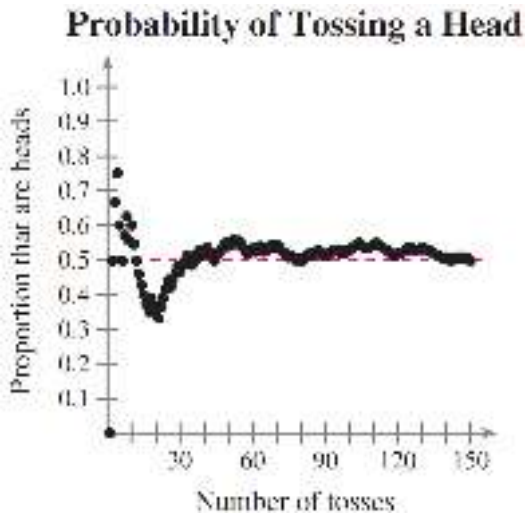
3. not O?

   Solution: $P(\text{not O}) = \frac{5+6+1}{20} = \frac{12}{20} = 0.60$

Definitions and Concepts

Law of Large Numbers

As an experiment is repeated over and over, the empirical probability of an event approaches the theoretical (actual) probability of the event.



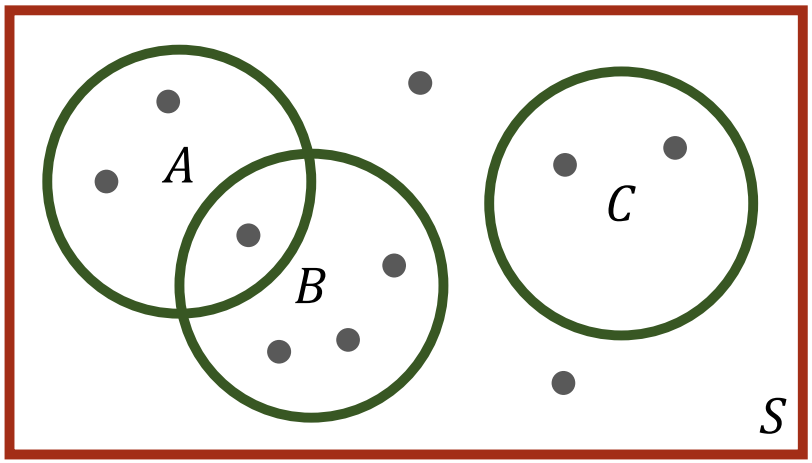
Subjective Probability

Subjective probabilities result from intuition (الحدس), educated guesses (تخمينات مدروسة), and estimates (تقديرات). For example:

1. Given a patient's health and extent of injuries, a doctor may feel that the patient has a 90% chance of a full recovery.
2. A business analyst may predict that the chance of the employees of a certain company going on strike (إضراب) is 0.25.

Venn Diagrams

- Provide a pictorial (شكلي) description of the sample space.
- To construct a Venn diagram, we first draw a rectangle to represent the sample space.
- Regions within the rectangle are then used to represent events, as shown in the figure.



Rules of Probability

General Rules

If S is the sample space of a probability experiment, and E is an event ($E \subset S$), then:

- $P(S) = 1$
- $0 \leq P(E) \leq 1$
- $P(\phi) = 0$

Notes

1. If $S = \{e_1, e_2, \cdots, e_n\}$ then $P(e_1) + P(e_2) + \cdots + P(e_n) = 1$.
2. If $S = \{e_1, e_2, \cdots, e_n\}$ where $P(e_1) = P(e_2) = \cdots = P(e_n)$ then the elementary events $e_1, e_2, \cdots, e_n$ are called equally likely and $P(e_i) = \frac{1}{n}$ for all $i = 1,2, \cdots, n$.

Example

If $S = \{e_1, e_2, e_3, e_4\}$ such that $P(e_1) = 0.3, P(e_2) = 0.4, P(e_3) = 0.2$. Find $P(e_4)$.

Solution:

$$P(e_1) + P(e_2) + P(e_3) + P(e_4) = 1$$

$$0.3 + 0.4 + 0.2 + P(e_4) = 1 \Rightarrow P(e_4) = 0.1$$

Rules of Probability

Example A stack (حزمة) contains eight tickets numbered 1, 1, 2, 2, 2, 3, 3, 3. One ticket will be drawn at random and its number will be noted.

1. List the sample space and assign probabilities to the elementary outcomes.

   Solution: $S = \{1,2,3\}$ where $P(1) = \frac{2}{8} = \frac{1}{4}, P(2) = P(3) = \frac{3}{8}$.

2. What is the probability of drawing an odd-numbered ticket?

   Solution: $P(\text{Odd}) = P(1) + P(3) = \frac{2}{8} + \frac{3}{8} = \frac{5}{8}$.

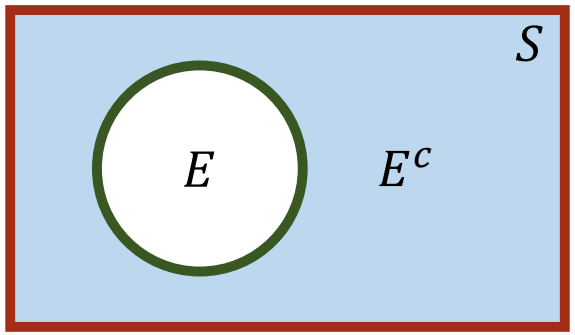
Example The probabilities of males and females in a statistics class is given by the ratio 3: 7. If we randomly select one student, find the probability that this student is male?

Solution: $P(\text{male}) = \frac{3}{3+7} = \frac{3}{10} = 0.3$

Rules of Probability

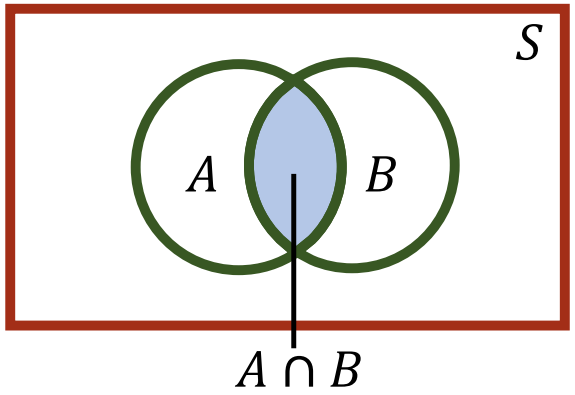
Complement of event E

- The complement of an event E, denoted by E^c, is the event consisting of all outcomes that are **not in E**.
- The occurrence of E^c means that the event E **does not occur**.
- $P(E^c) = 1 - P(E)$



Intersection of two events A and B

- The intersection of two events A and B, denoted by $A \cap B$, is the event consisting of all outcomes that are in both A and B.
- The occurrence of $A \cap B$ means that **both** A and B occur.

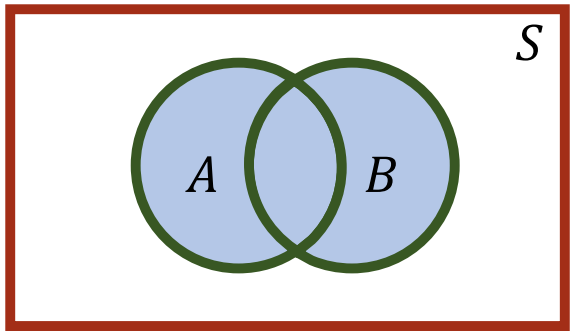


- If $A \cap B = \phi$ then A and B are called **disjoint** or **mutually exclusive**, and they cannot occur at the same time.

Rules of Probability

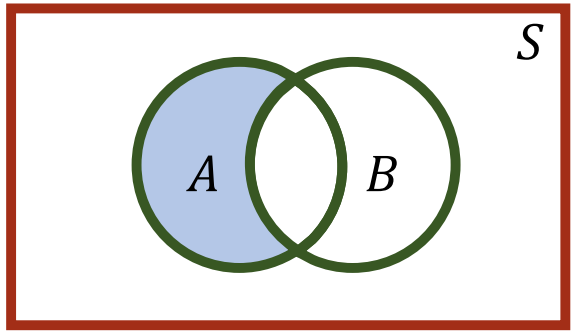
Union of two events A and B

- The union of two events A and B, denoted by $A \cup B$, is the event consisting of all outcomes that are either in A or B or in both.
- The occurrence of $A \cup B$ means that **at least** A or B occur.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If $A \cap B = \phi$ then $P(A \cup B) = P(A) + P(B)$.



Difference of two events A and B

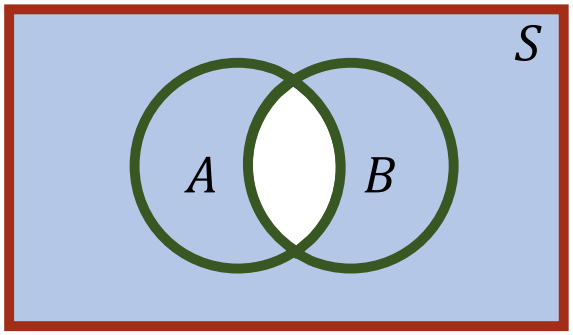
- The difference $A - B = A \cap B^c$ is the set of all outcomes that are in A but not in B.
- The occurrence of $A - B$ means that A occurs but B does not occur.
- $P(A - B) = P(A) - P(A \cap B)$
- In general, $P(A - B) \neq P(B - A)$



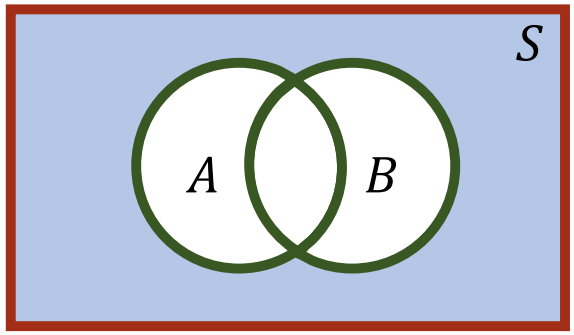
Rules of Probability

De Morgan Laws

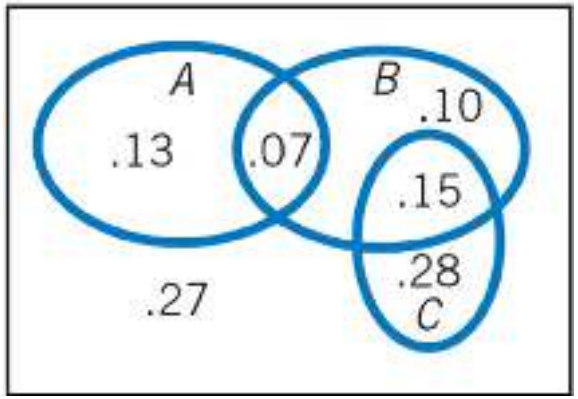
- $(A \cap B)^c = A^c \cup B^c$



- $(A \cup B)^c = A^c \cap B^c$



Example The given Venn diagram shows three events A, B, and C and also the probabilities of the various intersections. Then



- $P(A) = 0.13 + 0.07 = 0.20.$
- $P(B \cap C^c) = 0.10 + 0.07 = 0.17$
- $P(A \cup C) = 0.20 + 0.43 = 0.63$
- $P(A^c \cap B^c \cap C^c) = P\big((A \cup B \cup C)^c\big) = 0.27$

Rules of Probability

Example If A and B are two events such that $P(A) = \frac{19}{30}$, $P(B) = \frac{2}{5}$, and $P(A \cup B) = \frac{4}{5}$. Find $P(A \cap B)$.

Solution:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\frac{4}{5} = \frac{19}{30} + \frac{2}{5} - P(A \cap B)$$

$$P(A \cap B) = \frac{19}{30} + \frac{2}{5} - \frac{4}{5} = \frac{19}{30} + \frac{12}{30} - \frac{24}{30} = \frac{7}{30}$$

Example Given that $P(A^c) = \frac{2}{3}$, $P(B) = \frac{1}{2}$, and $P(A \cap B) = \frac{1}{12}$. Find $P(A^c \cap B^c)$.

Solution: Note that $P(A^c \cap B^c) = P\big((A \cup B)^c\big) = 1 - P(A \cup B)$.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \left(1 - \frac{2}{3}\right) + \frac{1}{2} - \frac{1}{12} = \frac{9}{12} = \frac{3}{4}$$

$$P(A^c \cap B^c) = 1 - \frac{3}{4} = \frac{1}{4}$$

Mathematics and Biostatistics

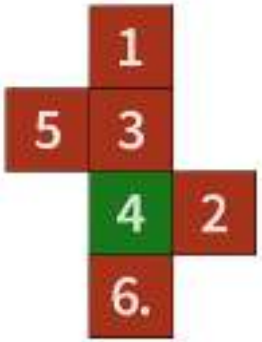
Chapter: [12]
Probability

Section: [12.3]
Conditional Probability

Definitions and Concepts

Conditional Probability is the probability of an event B occurring, given that another event A has already occurred, and is denoted by $P(B|A)$ and is read as "probability of B, given A."

Example A fair die is rolled.



$$P(4) = \frac{1}{6}$$

$$P(4|\text{Even Apears}) = \frac{1}{3}$$

Definitions and Concepts

Example The table shows the results of a study in which researchers examined a child's IQ and the presence of a specific gene in the child. Find the probability that:

	Gene Present	Gene Not Present	Total
High IQ	**31**	**19**	50
Normal IQ	**39**	**11**	50
Total	70	30	100

1. a child has a high IQ, given that the child has the gene.

   Solution: $P(\text{High IQ}|\text{Gene Present}) = \frac{31}{70}$.

2. a child does not have the gene

   Solution: $P(\text{Gene Not Present}) = \frac{30}{100} = 0.30$.

3. a child does not have the gene, given that the child has a normal IQ.

   Solution: $P(\text{Genet Not Present}|\text{Normal IQ}) = \frac{11}{50} = 0.22$.

Definitions and Concepts

The Rule Consider any two events A and B with $P(A) > 0$. The conditional probability of B given A is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Example A fair die is rolled. Let A be the event that an even number appears, and B is the event that the number 4 appears. Then

$$A = \{2,4,6\}$$
$$B = \{4\}$$
$$A \cap B = \{4\}$$
$$\therefore P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/6}{3/6} = \frac{1}{3}$$

Definitions and Concepts

The Complement Rule for Conditional Probability $P(B^c|A) = 1 - P(B|A)$

The Multiplication Rule $P(A \cap B) = P(B|A)P(A)$
$= P(A|B)P(B)$

Example There are 25 pens in a container on your desk. Among them, 20 will write well but 5 have defective ink cartridges. You will select 2 pens to take to a business appointment. Calculate the probability that:

1. Both pens are defective.

   Solution:

$$P(D_1 \cap D_2) = P(D_1) \cdot P(D_2|D_1) = \frac{5}{25} \times \frac{4}{24} = \frac{1}{30} = \frac{C_2^5}{C_2^{25}}$$

Definitions and Concepts

Example (*Continue*) There are 25 pens in a container on your desk. Among them, 20 will write well but 5 have defective ink cartridges. You will select 2 pens to take to a business appointment. Calculate the probability that:

2. One pen is defective but the other will write well.

Solution:

$$P(D_1 \cap G_2) + P(G_1 \cap D_2) = P(D_1) \cdot P(G_2|D_1) + P(G_1) \cdot P(D_2|G)$$

$$= \frac{5}{25} \times \frac{20}{24} + \frac{20}{25} \times \frac{5}{24} = \frac{1}{3}$$

$$= \frac{C_1^5 \cdot C_1^{20}}{C_2^{25}}$$

INDEPENDENT AND DEPENDENT EVENTS

Definition Two events are **independent** when the occurrence of one of the events does not affect the probability of the occurrence of the other event.

- Two events A and B are independent when

$$P(A|B) = P(A) \text{ or}$$
$$P(B|A) = P(B) \text{ or}$$
$$P(A \cap B) = P(A) \cdot P(B)$$

- Events that are *not independent* are **dependent**.

Example A coin is flipped and a die is rolled. Find the probability of getting a head on the coin and a 4 on the die.

Solution:

$$P(H \cap 4) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$$

INDEPENDENT AND DEPENDENT EVENTS

Results If A and B are independent events, then the following events are also independent.

- A^c and B. So, $P(A^c \cap B) = P(A^c) \cdot P(B)$
- A and B^c. So, $P(A \cap B^c) = P(A) \cdot P(B^c)$
- A^c and B^c. So, $P(A^c \cap B^c) = P(A^c) \cdot P(B^c)$

Example Two men, A and B are shooting a target. The probability that A hits the target is $P(A) = \frac{1}{3}$, and the probability that B shoots the target is $P(B) = \frac{1}{5}$, one independently of the other. Find the probability that:

1. Both men hit the target.

   Solution: $P(A \cap B) = P(A) \cdot P(B) = \frac{1}{3} \times \frac{1}{5} = \frac{1}{15}$

2. At least one of them hits the target.

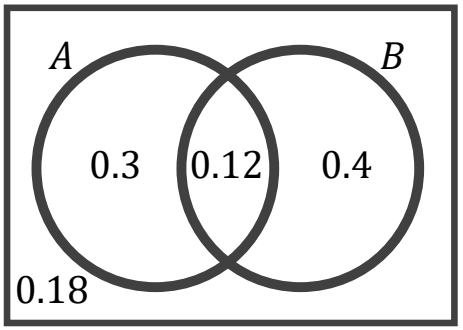
   Solution: $P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{3} + \frac{1}{5} - \frac{1}{15} = \frac{7}{15}$

3. Exactly one of them hits the target.

   Solution: $P(A^c \cap B) + P(A \cap B^c) = \frac{2}{3} \times \frac{1}{5} + \frac{1}{3} \times \frac{4}{5} = \frac{2}{5}$

INDEPENDENT AND DEPENDENT EVENTS

Example The given Venn diagram shows the events A and B and also the probability of their intersection.



1. Find $P(B^c|A)$.

   Solution:

   $P(B^c|A) = 1 - P(B|A) = 1 - 0.12 = 0.88$

2. Are A and B independent?

   Solution: No, since

   $$P(A) = 0.3 + 0.12 = 0.42$$
   $$P(B) = 0.4 + 0.12 = 0.52$$
   $$P(A \cap B) = 0.12$$
   $$\text{but } P(A \cap B) \neq P(A) \cdot P(B)$$

Mathematics and Biostatistics

Chapter: [12]
Probability

Section: [12.4]
Discrete Random Variables

RANDOM VARIABLES

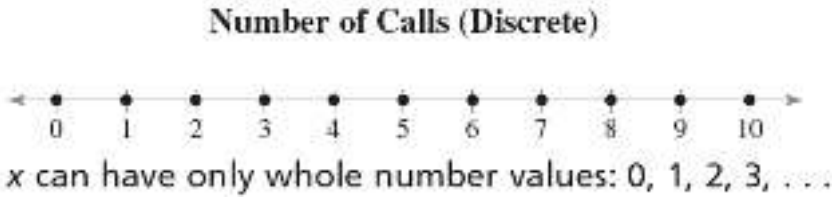
Random Variable

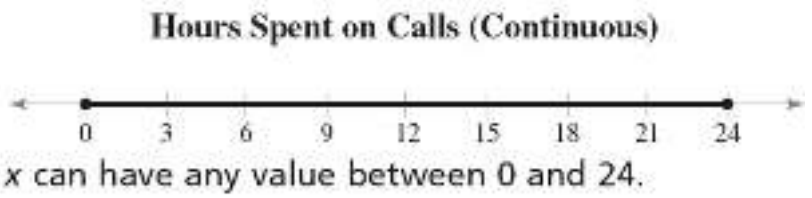
A **random variable** x represents a numerical value associated with each outcome of a probability experiment.

- The word *random* indicates that x is determined by chance.
- There are two types of random variables: **discrete** and **continuous**.

Types of Random Variables

- A random variable is **discrete** when it has a finite or countable number of possible outcomes that can be listed.
- A random variable is **continuous** when it has an uncountable number of possible outcomes, represented by an interval on a number line.





DISCRETE PROBABILITY DISTRIBUTIONS

Definition

- A **discrete probability distribution** lists each possible value the random variable can assume, together with its probability.
- A discrete probability distribution must satisfy these conditions:
  1. The probability of each value of the discrete random variable is between 0 and 1, inclusive $[0 \leq P(x) \leq 1]$.
  2. The sum of all the probabilities is 1.

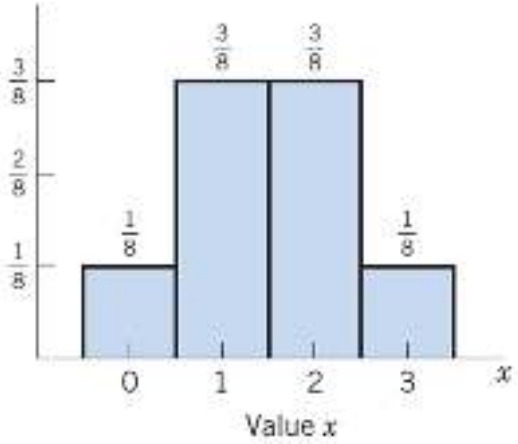
Example

If X represents the number of heads obtained in three tosses of a fair coin, find the probability distribution of X.

Solution:

$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$

$X = x$	0	1	2	3
$P(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$



DISCRETE PROBABILITY DISTRIBUTIONS

Example The discrete random variable X has the probability distribution shown. Find:

$X = x$	-3	-2	-1	0	1
$P(x)$	0.10	0.25	0.30	0.15	c

- The value of c.
  Solution: $0.10 + 0.25 + 0.30 + 0.15 + c = 1 \Rightarrow c = 0.20$
- $P(-3 \leq X < 0)$
  Solution: $P(-3 \leq X < 0) = P(-3) + P(-2) + P(-1) = 0.65$
- $P(X > -1)$
  Solution: $P(X > -1) = P(0) + P(1) = 0.35$
- $P(-1 < X < 1)$
  Solution: $P(-1 < X < 1) = P(0) = 0.15$

The Expected Value (Mean) and the Variance

Formulas The mean and the variance of a **discrete** random variable X are defined as follows.

- The expected value (mean) of X is given by $\mu_X = E(X) = \sum xP(x)$.
- The variance of X is

$$\sigma_X^2 = \text{Var}(X) = \sum (x - \mu)^2 P(x)$$

$$= E(X^2) - [E(X)]^2$$

  where $E(X^2) = \sum x^2 P(x)$.
- The standard deviation of X is $\sigma_X = \sqrt{\text{Variance}}$.

The Expected Value (Mean) and the Variance

Example In a family with two children, let X be the number of girls. Find the mean and the variance of X.

$X = x$	0	1	2
$P(x)$	0.25	0.50	0.25

Solution:

$$E(X) = \sum xP(x) = (0)(0.25) + (1)(0.50) + (2)(0.25)$$

$$= 1$$

$$\sigma_X^2 = E(X^2) - \big(E(X)\big)^2$$

$$= (0)^2(0.25) + (1)^2(0.50) + (2)^2(0.25) - (1)^2$$

$$= 0.5$$

The Expected Value (Mean) and the Variance

Example A random variable X has probability distribution given. Find the mean and the variance of X.

x	$P(x)$	$xP(x)$	$x^2P(x)$
-2	0.30	-0.60	1.20
-1	0.10	-0.10	0.10
0	0.15	0	0
1	0.40	0.40	0.40
2	0.05	0.10	0.20
		-0.20	1.90

Solution:

$$E(X) = \sum xP(x) = -0.20$$

$$\sigma_X^2 = E(X^2) - \big(E(X)\big)^2 = 1.90 - (-0.20)^2$$
$$= 1.86$$

The Expected Value (Mean) and the Variance

Notes

If a and b are constants, then:

- $E(aX + b) = aE(X) + b$
- $\text{Var}(aX + b) = a^2\text{Var}(X)$

Example

Let X be a random variable with mean $E(X) = 1$ and variance $\text{Var}(X) = 0.5$. Then

- $E(2X - 3) = 2E(X) - 3 = 2(1) - 3 = -1.$
- $\text{Var}(4 - 5X) = (-5)^2\text{Var}(X) = 25 \times 0.5 = 12.5$

Mathematics and Biostatistics

Chapter: [12]
Probability

Section: [12.5]
Continuous Random Variables
(*Standard Normal Distribution*)

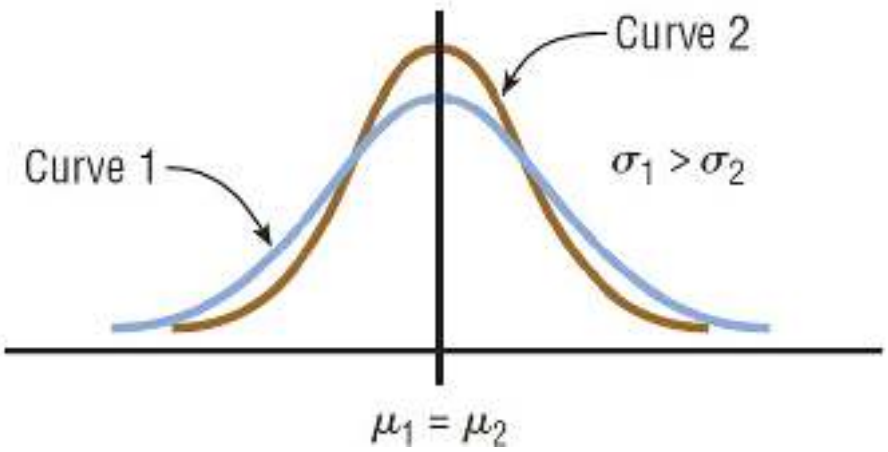
Normal Probability Distributions

Definition A **normal distribution** is a *continuous*, *symmetric*, *bell-shaped* distribution of a variable.

The Curve The mathematical equation for a normal distribution curve with mean μ and standard deviation σ is

$$P(x) = \frac{1}{\sigma\sqrt{\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad ; \quad x \in (-\infty, \infty)$$

- A normal distribution can have any mean and any positive standard deviation.
- These two parameters, μ and σ, determine the shape of the normal curve.
- The mean gives the location of the line of symmetry, and the standard deviation describes how much the data are spread out.



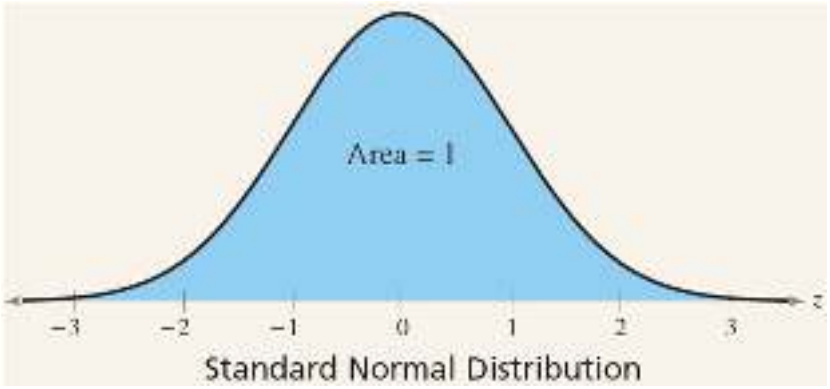
Normal Probability Distributions

Properties

- A normal distribution curve is **bell-shaped**.
- The mean, median, and mode are *equal* and are located at the center of the distribution.
- A normal distribution curve is **unimodal** (it has only one mode).
- The curve is **symmetric** about the *mean*, which is equivalent to saying that its shape is the same on both sides of a vertical line passing through the center.
- The curve is **continuous**; that is, there are no *gaps* or *holes*.
- The curve **never touches** the x −axis.
- The **total area** under a normal distribution curve is equal to **1**.

The Standard Normal Distribution

Definition The **standard normal distribution** is a normal distribution with a mean of 0 and a standard deviation of 1. In this case, the random variable is denoted by Z.

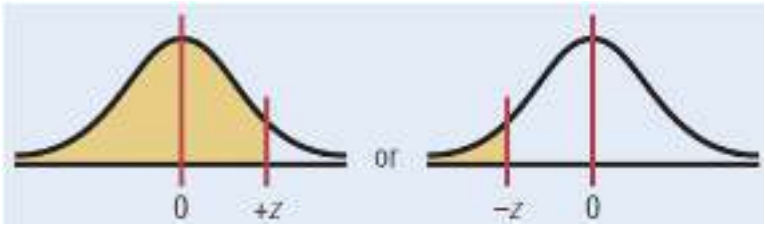


Using the Standard Normal Table

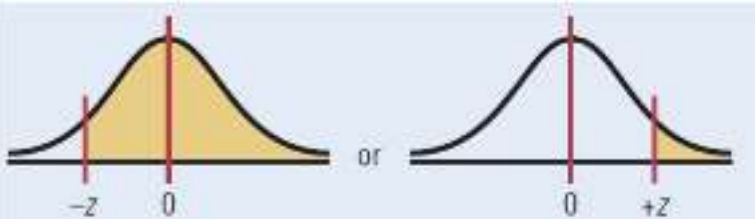
Standard Normal Table The table lists the **cumulative area** under the standard normal curve to the left of z from -3.49 to 3.49.

Rules

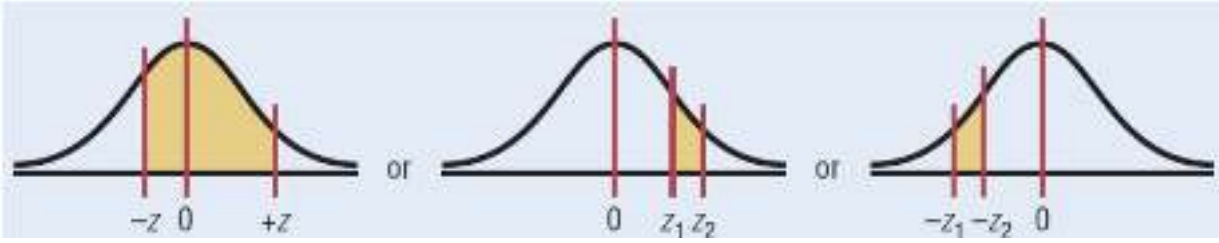
- $P(Z \leq c)$ is evaluated directly from the table.



- $P(Z \geq c) = 1 - P(Z < c) = P(Z \leq -c)$.



- $P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)$.

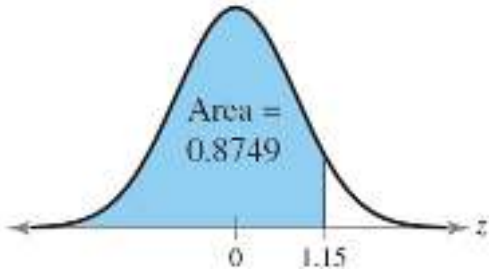


Using the Standard Normal Table

Example Find $P(Z \leq 1.15)$.

z	.00	.01	.02	.03	.04	.05	.06
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279

Solution: $P(Z \leq 1.15) = 0.8749$

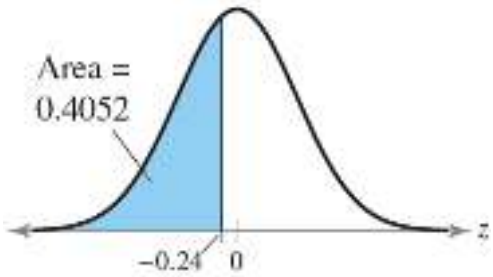


Using the Standard Normal Table

Example Find $P(Z \leq -0.24)$.

z	.09	.08	.07	.06	.05	.04	.03
−3.4	.0002	.0003	.0003	.0003	.0003	.0003	.0003
−3.3	.0003	.0004	.0004	.0004	.0004	.0004	.0004
−3.2	.0005	.0005	.0005	.0006	.0006	.0006	.0006
−0.5	.2776	.2810	.2843	.2877	.2912	.2946	.2981
−0.4	.3121	.3156	.3192	.3228	.3264	.3300	.3336
−0.3	.3483	.3520	.3557	.3594	.3632	.3669	.3707
−0.2	.3859	.3897	.3936	.3974	.4013	.4052	.4090
−0.1	.4247	.4286	.4325	.4364	.4404	.4443	.4483
−0.0	.4641	.4681	.4721	.4761	.4801	.4840	.4880

Solution: $P(Z \leq -0.24) = 0.4052$



Using the Standard Normal Table

Example Find $P(Z > 1.06)$.

Solution: $P(Z > 1.06) = 1 - P(Z \leq 1.06) = 1 - 0.8554 = 0.1446$

Example Find $P(Z > -1.84)$.

Solution: $P(Z > -1.84) = P(Z < 1.84) = 0.9671$

Example Find $P(0.21 < Z \leq 1.07)$.

Solution:

$$\begin{aligned} P(0.21 < Z \leq 1.07) &= P(Z \leq 1.07) - P(Z \leq 0.21) \\ &= 0.8577 - 0.5832 \\ &= 0.2745 \end{aligned}$$

Using the Standard Normal Table

Example Find c if $P(Z < c) = 0.2709$.

Solution: From the table, $c = -0.61$

Example Find c if $P(Z \geq c) = 0.1038$.

Solution:
$$P(Z \geq c) = 0.1038 \Rightarrow 1 - P(Z < c) = 0.1038$$
$$\Rightarrow P(Z < c) = 0.8962$$
$$\Rightarrow c = 1.96$$

Example Find c if $P(-0.60 < Z < c) = 0.4991$.

Solution:
$$P(-0.60 < Z < c) = 0.4991$$
$$P(Z < c) - P(Z < -0.60) = 0.4991$$
$$P(Z < c) - 0.2743 = 0.4991$$
$$P(Z < c) = 0.7734 \Rightarrow c = 0.75$$