Lecture 1

A Brief Review of Statistics

Variables and Types of Data

Qualitative variables are variables that have distinct categories according to some characteristic or attribute. For example, if subjects are classified according to gender (male or female), then the variable gender is qualitative.

Quantitative variables are variables that can be counted or measured. For example, the variable age is numerical, and people can be ranked in order according to the value of their ages.

Discrete variables assume values that can be counted. Discrete variables can be assigned values such as 0, 1, 2, 3 and are said to be countable. Examples of discrete variables are the number of children in a family, the number of students in a classroom, etc.

Continuous variables can assume an infinite number of values between any two specific values. They are obtained by measuring. They often include fractions and decimals. Temperature, for example, is a continuous variable, since the variable can assume an infinite number of values between any two given temperatures.

Variables and Types of Data *(cont.)*

In addition to being classified as qualitative or quantitative, variables can be classified by how they are categorized, counted, or measured. For example, can the data be organized into specific categories, such as area of residence (rural, suburban, or urban)? Can the data values be ranked, such as first place, second place, etc.?

The **nominal level of measurement** classifies data into mutually exclusive (nonoverlapping) categories in which no order or ranking can be imposed on the data. Classifying survey subjects as male or female is an example of nominal level measurement.

The **ordinal level of measurement** classifies data into categories that can be ranked; however, precise differences between the ranks do not exist. For instance, people can be classified according to their build (small, medium, or large).

The **interval level of measurement** ranks data, and precise differences between units of measure do exist; however, there is no meaningful zero. Temperature is an example of such a variable. There is not a meaningful zero. Zero in Celsius is 32 degrees Fahrenheit.

Variables and Types of Data *(cont.)*

The ratio level of measurement possesses all the characteristics of interval measurement, and there exists a true zero. In addition, true ratios exist when the same variable is measured on two different members of the population. Ratio scales have differences between units (1 inch, 1 pound, etc.) and a true zero. In addition, the ratio scale contains a true ratio between values. For example, if one person can lift 200 pounds and another can lift 100 pounds, then the ratio between them is 2 to 1.

What is probability?

Everybody is familiar with the phrase “the probability that a coin will land heads is 0.5”. But what exactly does this mean? Loosely speaking, probability is the chance of an event occurring. There are at least two different interpretations of probability. One is called the **frequentist i**nterpretation. In this view, probabilities represent long run frequencies of events. For example, the above statement means that, if we flip the coin many times, we expect it to land heads about half the time. The other interpretation is called the **Bayesian** interpretation of probability. In this view, probability is used to quantify our **uncertainty** about something; hence it is fundamentally related to information rather than repeated trials. In the Bayesian view (after the 18th century mathematician Thomas Bayes), the above statement means we believe the coin is equally likely to land heads or tails on the next toss.

One advantage of the Bayesian interpretation is that it can be used to model our uncertainty about events that do not have long term frequencies. For example, we might want to compute the probability that the polar ice cap will melt by 2020 CE. This event will happen zero or one times, but cannot happen repeatedly. Nevertheless, we ought to be able to quantify our uncertainty about this event; based on how probable we think this event is. In this cases, the idea of repeated trials of flipping a coin does not make sense, but the Bayesian interpretation is valid.

Probability experiments, outcomes and sample spaces

A **probability experiment or random experiment** is a chance process that leads to well-defined results called outcomes.

An **outcome** is the result of a single trial of a probability experiment.

For example, we have a trial when we flip a coin once, roll one die once, or the like. When a coin is tossed, there are two possible outcomes: head or tail. In the roll of a single die, there are six possible outcomes: 1, 2, 3, 4, 5, or 6. In any experiment, the set of all possible outcomes is called the *sample space.*

A **sample space** is the set of all possible outcomes of a probability experiment.

Examples of sample spaces:

- Toss one coin with outcomes: head, tail.
- Roll a die with outcomes: 1, 2, 3, 4, 5, 6.
- Answer a true/false question with outcomes: true, false.
- Toss two coins with outcomes: head-head, tail-tail, head-tail, tail-head
- Here is the sample space of rolling two dies:

	Die 2					
Die 1	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Classical probability

In many cases, one must find the probability of two or more outcomes. For this reason, it is necessary to distinguish between an outcome and an event.

An **event** consists of a set of outcomes of a probability experiment. An event can be one outcome or more than one outcome. For example, if a die is rolled and a 6 shows, this result is called an *outcome,* since it is a result of a single trial. An event with one outcome is called a **simple event.** The event of getting an odd number when a die is rolled is called a **compound event,** since it consists of three outcomes or three simple events. In general, a compound event consists of two or more outcomes or simple events.

According to the approach of **classical probability**, we can use sample spaces to determine the numerical probability that an event will happen. You do not actually have to perform the experiment to determine that probability. *Classical probability assumes that all outcomes in the sample space are equally likely to occur.* For example, when a single die is rolled, each outcome has the same probability of occurring. Since there are six outcomes, each outcome has a probability of 1/2. When a card is selected from an ordinary deck of 52 cards, you assume that the deck has been shuffled, and each card has the same probability of being selected. In this case, it is 1/52.

Classical probability *(cont.)*

Let’s find the probability that two of the three children in a family are girls if the family has three children. The sample space for the gender of the children for a family that has three children has eight outcomes, that is, BBB, BBG, BGB, GBB, GGG, GGB, GBG, and BGG. Since there are three ways to have two girls, namely, GGB, GBG, and BGG, $P(\text{two girls}) = 3/8$.

Basic rules of probability

Rule 1. The probability of any event *E, P(E),* is a number (either a fraction or decimal) between and including 0 and 1. This is denoted by $0 \leq P(E) \leq 1$.

Rule 2. If an event *E* cannot occur (i.e., the event contains no members in the sample space), its probability is 0. For example, the probability of getting 9 when a single dice is rolled is 0.

Rule 3. If an event *E* is certain, then the probability of *E* is 1. For example, the probability of getting a number less than 7 when a single dice is rolled is 1 because all outcomes—1, 2, 3, 4, 5, and 6—are less than 7, and the probability of the event consisting of all outcomes in the sample space is 1.

Rule 4. The sum of the probabilities of all the outcomes in the sample space is 1.

Complementary events

The **complement of an event** *E* is the set of outcomes in the sample space that are not included in the outcomes of event *E.* The complement of *E* is denoted by ¬*E* or $\bar{E}$ (read "*E* bar"). For example, the compliment of getting 4 when rolling a single die is the event consisting of getting 1 or 2 or 3 or 5 or 6.

The outcomes of an event and the outcomes of the complement make up the entire sample space. For example, if two coins are tossed, the sample space is HH, HT, TH, and TT. The complement of "getting all heads" is not "getting all tails," since the event "all heads" is HH, and the complement of HH is HT, TH, and TT. Hence, the complement of the event "all heads" is the event "getting at least one tail." Since the event and its complement make up the entire sample space, it follows that the sum of the probability of the event and the probability of its complement will equal 1. In other words:

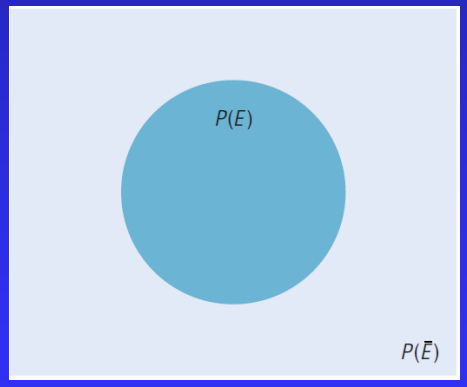
Rule 5. $P(E) = 1 - P(\neg E)$ or

$P(\neg E) = 1 - P(E)$ or

$P(E) + P(\neg E) = 1$

Venn diagram for probability and complement

In a Venn diagram, an event is represented by a figure consisting of all the outcomes (represented as points) that belong to the event. The probability of the event is the area of the figure. For example, the area inside the circle is *P(E)*. The area outside the circle is *p(¬E).* Since the probability of the sample space is 1, *P*(*S*) = 1, then *P*(*E*) = P(*S*) – P(¬E), i.e., *P*(*E*) = 1 – P(¬E). Here the sample space is the whole space.



Empirical Probability

The difference between classical and **empirical probability** is that classical probability assumes that certain outcomes are equally likely (such as the outcomes when a die is rolled), while empirical probability relies on actual experience to determine the likelihood of outcomes. In empirical probability, one might actually roll a given die 6000 times, observe the various frequencies, and use these frequencies to determine the probability of an outcome.

Formula for Empirical Probability: Given a frequency distribution, the probability of an event being in a given class is:

$$\mathrm{P(E)} = \frac{frequency\ for\ the\ class}{total\ frequency}$$

This probability is called *empirical probability* and is based on observation.

Law of large numbers. When a coin is tossed one time, it is common knowledge that the probability of getting a head is 1/2. But what happens when the coin is tossed 50 times? Will it come up heads 25 times? Not all the time. You should expect about 25 heads if the coin is fair. But due to chance variation, 25 heads will not occur most of the time. If the empirical probability of getting a head is computed by using a small number of trials, it is usually not exactly . However, as the number of trials increases, the empirical probability of getting a head will approach the theoretical probability of , if in fact the coin is fair (i.e., balanced). This phenomenon is an example of the **law of large numbers.**

The Addition Rules for Probability

The following notation is used:

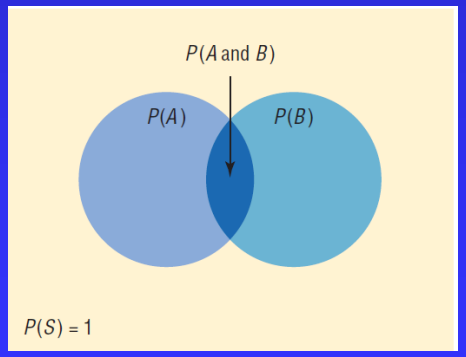
$P(A \cup B)$ denotes the union of the events A and B. An outcome belongs to $P(A \cup B)$ if it belongs A or to B. $P(A \cup B)$ is also denoted as $P(A \text{ or } B)$.

$P(A \cap B)$ denotes the intersection of the events A and B. In other words, an outcome belongs to $P(A, B)$ if it belongs to both A and B. $P(A \cap B)$ is also denoted as $P(A, B)$ or $P(A \text{ and } B)$.

Addition rule:

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

The rule can be verified by a Venn diagram:



The Addition Rules for Probability *(cont.)*

On the previous slide, the area in the intersection or overlapping part of both circles corresponds to $P(A \cap B)$. When the area of circle A is added to the area of circle B, the overlapping part is counted twice. It must, therefore, be subtracted once to get the correct area or probability.

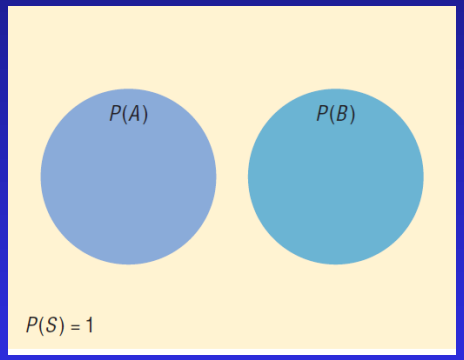
Venn diagrams were developed by mathematician John Venn (1834–1923) and are used in set theory and symbolic logic. They have been adapted to probability theory also. In set theory, the symbol represents the *union* of two sets, and A B corresponds to A or B. The symbol represents the *intersection* of two sets, and A B corresponds to A and B.

For example, let's find the probability of rolling a number greater than 3 or an even number when a single dice is rolled. Let's denote A = {a number greater than 3 is rolled} and B = {an even number is rolled}. We are looking for $P(A \cup B)$. Event A includes the following outcomes 4, 5, and 6. Therefore, P(A) = 3/6 = ½. Event B includes the following outcomes 2,4, and 6. Hence, P(B) = 3/6 = ½. Event $P(A \cap B)$ includes the outcomes 4 and 6 (this is the intersection of the set 4,5 and 6 with the set 2, 4 and 6). Therefore, $P(A \cap B) = 2/6 = 1/3$. Using the addition rule, we get: $P(A \cup B) = 1/2 + 1/2 - 1/3 = 2/3$.

Mutually exclusive events

Two events are **mutually exclusive events** if they cannot occur at the same time (i.e., they have no outcomes in common; or their intersection is empty). For example, getting 4 and getting 6 when a single dice is rolled are two mutually exclusive events because either 4 or 6 can be rolled but not both of them.

Here is a Venn diagram for two mutually exclusive events:



Since the events do not overlap (do not have outcomes in common), their intersection $A \cap B$) is empty. Therefore $P(A \cap B) = 0$. as a result, **the addition law for mutually exclusive events** is simpler:

$P(A \cup B) = P(A) + P(B)$ if A and B are mutually exclusive.

Consider the following problem: At a political rally, there are 20 Republicans, 13 Democrats, and 6 Independents. If person is selected at random, find the probability that he or she is either a Democrat or an Independent. Solution: *P*(Democrat or Independent) = *P*(Democrat) + *P*(Independent) = 13/39 + 6/39 = 19/39

The Multiplication Rule

In observing the outcomes of a random experiment, one is often interested in how the outcome of one event A is influenced by the outcome of another event B. For example, the fact that it rains today might affect the probability of raining tomorrow, and the fact that New York Yankees lost their game today has no effect on the probability of raining tomorrow.

Two events *A* and *B* are **independent events** if the fact that *A* occurs does not affect the probability of *B* occurring.

Examples of independent events:

- Rolling a die and getting a 6, and then rolling a second die and getting a 3.

- Drawing a card from a deck and getting a queen, replacing it, and drawing a second card and getting a queen.

The *multiplication rules* can be used to find the probability of two or more events that occur in sequence.

Multiplication rule: When two events are independent, the probability of both occurring is $P(A, B) = P(A) * P(B)$

Example: A coin is flipped and a die is rolled. The probability of getting a head on the coin and a 4 on the die is $P(\text{head and } 4) = P(\text{head}) * P(4) = ½ * 1/6 = 1/12$ since the two events are independent. Note that the sample space for the coin is H, T; and for the die it is 1, 2, 3, 4, 5, 6.

Dependent events

When the outcome or occurrence of the first event affects the outcome or occurrence of the second event in such a way that the probability is changed, the events are said to be **dependent events.**

For example, suppose a card is drawn from a deck and *not* replaced, and then a second card is drawn. What is the probability of selecting an ace on the first card and a king on the second card? In this case, the two events are dependent. The probability of selecting an ace on the first draw is 4/52 since there are 4 aces and 52 cards . If that card is *not* replaced (returned back to the stack), the probability of selecting a king on the second card is 4/51, since there are 4 kings and only 51 cards remaining. The outcome of the first draw has affected the outcome of the second draw. On the other hand, if a card was drawn from a deck and *not* replaced, and then a second card was drawn, then the probability of selecting an ace on the first card and a king on the second card would be two independent events. The probability of getting an ace on the first card would be 4/52 and the probability of getting a king on the second card would be 4/52.

Another examples of dependent events: being a lifeguard and getting a suntan; having high grades and getting a scholarship; parking in a no-parking zone and getting a parking ticket.

Conditional probability

The **conditional probability** of an event *B* in relationship to an event *A* is the probability that event *B* occurs after event *A* has already occurred. The notation for conditional probability is $P(B \mid A)$. This notation does not mean that *B* is divided by *A*; rather, it means the probability that event *B* occurs given that event *A* has already occurred. In the card example without replacement, $P(B \mid A)$ is the probability that the second card is a king given that the first card is an ace, and it is equal to 4 since the first card was *not* replaced.

When two events are dependent, the probability of both occurring is

$P(A, B) = P(A) * P(B \mid A)$

Example: At a university, there were 5 burglaries reported in 2003, 16 in 2004, and 32 in 2005. If a researcher wishes to select at random two burglaries to further investigate, find the probability that both will have occurred in 2004. Solution: In this case, the events are dependent since the researcher wishes to investigate two distinct cases. Hence the first case is selected and not replaced.

P(C1, C2) = P(C1) * P(C2| C1) = 16/53 * 15/52. Initially, there are 16 burglaries in 2004 and 53 total cases. After a selecting one burglary in 2004, there are15 remaining burglaries in 2004 and a total of 52 cases.

Conditional probability *(cont.)*

The conditional probability of an event *B* in relationship to an event *A* was defined as the probability that event *B* occurs after event *A* has already occurred.

The probability that the second event *B* occurs given that the first event *A* has occurred can be found by dividing the probability that both events occurred by the probability that the first event has occurred. The formula is:

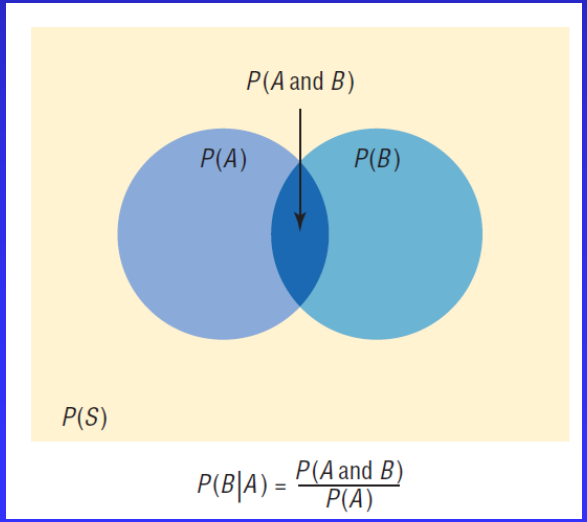
$P(B \mid A) = P(A, B) / P(A)$

If two events are independent, then P(B | A) = P(B) since B does not depend on A. If we substitute P(B) for $P(B \mid A)$ in the formula above, we get the multiplication rule for independent events: P(A, B) = P(A) * P(B).

Example: The probability that Sam parks in a no-parking zone *and* gets a parking ticket is 0.06, and the probability that Sam cannot find a legal parking space and has to park in the no-parkingzone is 0.20. On Tuesday, Sam arrives at school and has to park in a no-parking zone. Find the probability that he will get a parking ticket. Solution: Let *N* ={parking in a no-parking zone} and *T* = {getting a ticket}. Then, P(T | N) = P(T, N) / P(N) = 0.06 / 0.20 = 0.30

Venn diagram for conditional probability

Here is a Venn diagram for conditional probability. The reasoning here is that if you assume *A* has occurred, then *A* becomes the sample space for the next calculation and P(A) is the denominator of the probability fraction . The numerator *P*(*A* and *B*) represents the probability of the part of *B* that is contained in *A.* Hence, *P*(*A* and *B*) becomes the numerator of the probability fraction.



Measures of central tendency

Mean and median

The word *average* is often used to describe a set of data using one central value. The word average, however is ambiguous, since several different methods can be used to obtain an average. Loosely stated, the average means the center of the distribution or the most typical case. Measures of average are also called *measures of central tendency* and include the *mean, median, mode,* and *midrange.*

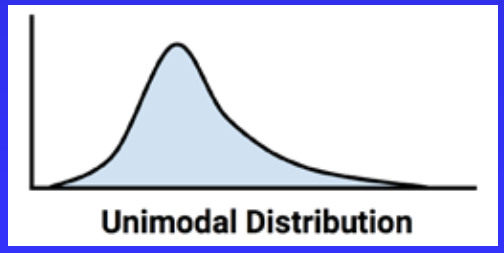
The **mean**, also known as the arithmetic average, is found by adding the values of the data and dividing by the total number of values. For example, the mean of 3, 2, 6, 5, and 4 is found by adding 3+2+6+5+4=20 and dividing by 5; hence, the mean of the data is 20/5=4.

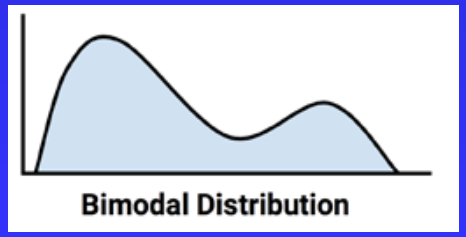
The **median** is the halfway point in a data set. Before you can find this point, the data must be arranged in ascending or increasing order. When the data set is ordered, it is called a data array. The median either will be a specific value in the data set or will fall between two values. For example, to find the median of 177 153 122 141 189 155 162 165 149 157 240, you first need to arrange data in ascending order: 122, 141, 149, 153, 155, 157, 162, 165, 177, 189, 240. Since there are an odd number of data values, the middle data value, 157, is the median.

Mode

Let's find the median of 684, 764, 656, 702, 856, 1133, 1132, 1303. After arranging data values in ascending order, we get: 684, 764, 656, 702, 856, 1133, 1132, 1303. The number of data values is even and there are two middle values: 764 and 856. The median is one half of their sum, i.e., (764+856)/2=810.

The third measure of average is called the **mode**. The mode is the value that occurs most often in the data set. It is sometimes said to be the most typical case. A data set that has only one value that occurs with the greatest frequency is said to be **unimodal**. If a data set has two values that occur with the same greatest frequency, both values are considered to be the mode and the data set is said to be **bimodal**. If a data set has more than two values that occur with the same greatest frequency, each value is used as the mode, and the data set is said to be **multimodal**. On the diagram below, data values are plotted on the x-axis and their frequencies are plotted on the y-axis.





Mode *(cont.)*

Let's find the mode of 18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10. It is helpful to arrange the data in order: 10, 10, 10, 11.3, 12.4, 14.0, 18.0, 34.5. Obviously, the mode is 10 with frequency 3. There is no mode in the data set 10, 10, 10, 11.3, 12.4, 14.0, 18.0, 34.5 because each value occurs only once. The data set 104 104 104 104 104 107 109 109 109 110 109 111 112 111 109 is bimodal with modes 104 and 109.

The **midrange** is defined as the sum of the lowest and highest values in the data set, divided by 2. The symbol MR is used for the midrange. MR = (lowest value + highest value) / 2.

Data description

Histograms

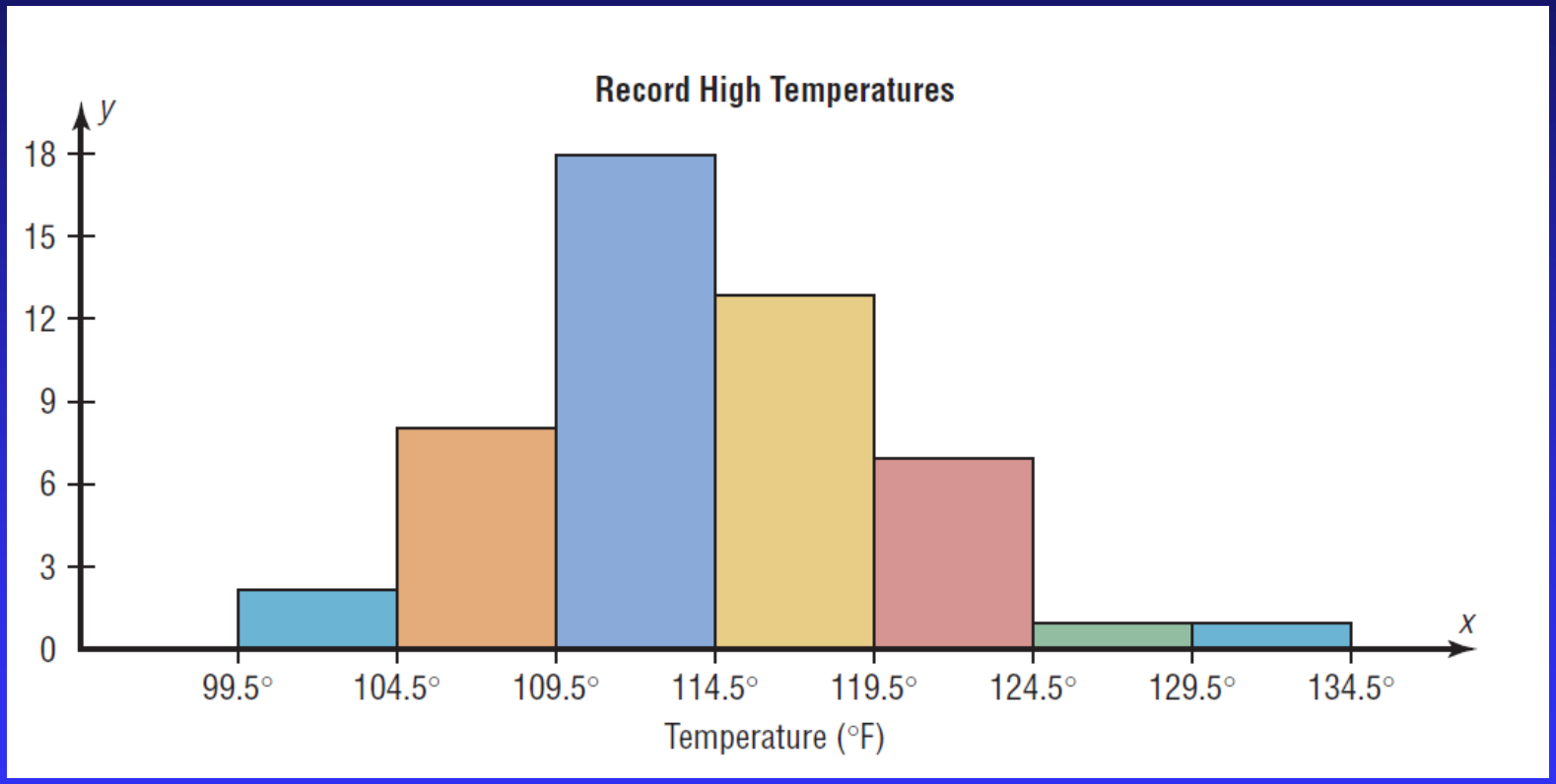
The **histogram** is a graph that displays the data by using contiguous vertical bars (unless the frequency of a data class is 0) of various heights to represent the frequencies of data classes.

Example: the data below shows the record high temperatures for each of the 50 states. The record temperatures are grouped into 7 classes. For example 2 states have record high temperatures between 99.5 and 104.5

Class boundaries	Frequency
99.5–104.5	2
104.5–109.5	8
109.5–114.5	18
114.5–119.5	13
119.5–124.5	7
124.5–129.5	1
129.5–134.5	1

This is the so called frequency distribution. It consists of *classes* and their corresponding *frequencies.* Each data value is placed into a quantitative or qualitative category called a **class.** The **frequency** of a class then is the number of data values contained in a specific class. For example, the frequencies of the class 104.5-109.5 is 8.

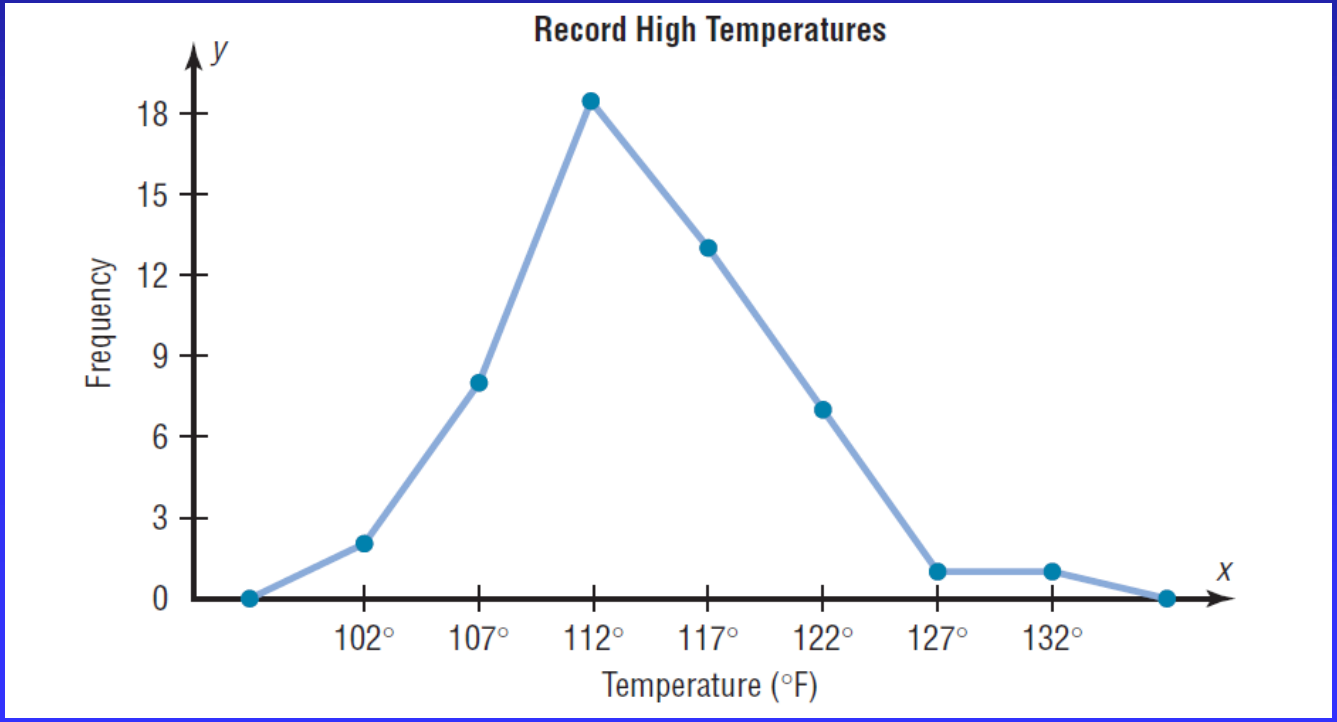
Histograms *(cont.)*



The frequencies are represented on the *y* axis and the class boundaries on the *x* axis

Frequency polygon

The **frequency polygon** is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points. In our example, the *x* axis is labeled with the midpoint of each class. The frequencies are represented on the *y* axis.



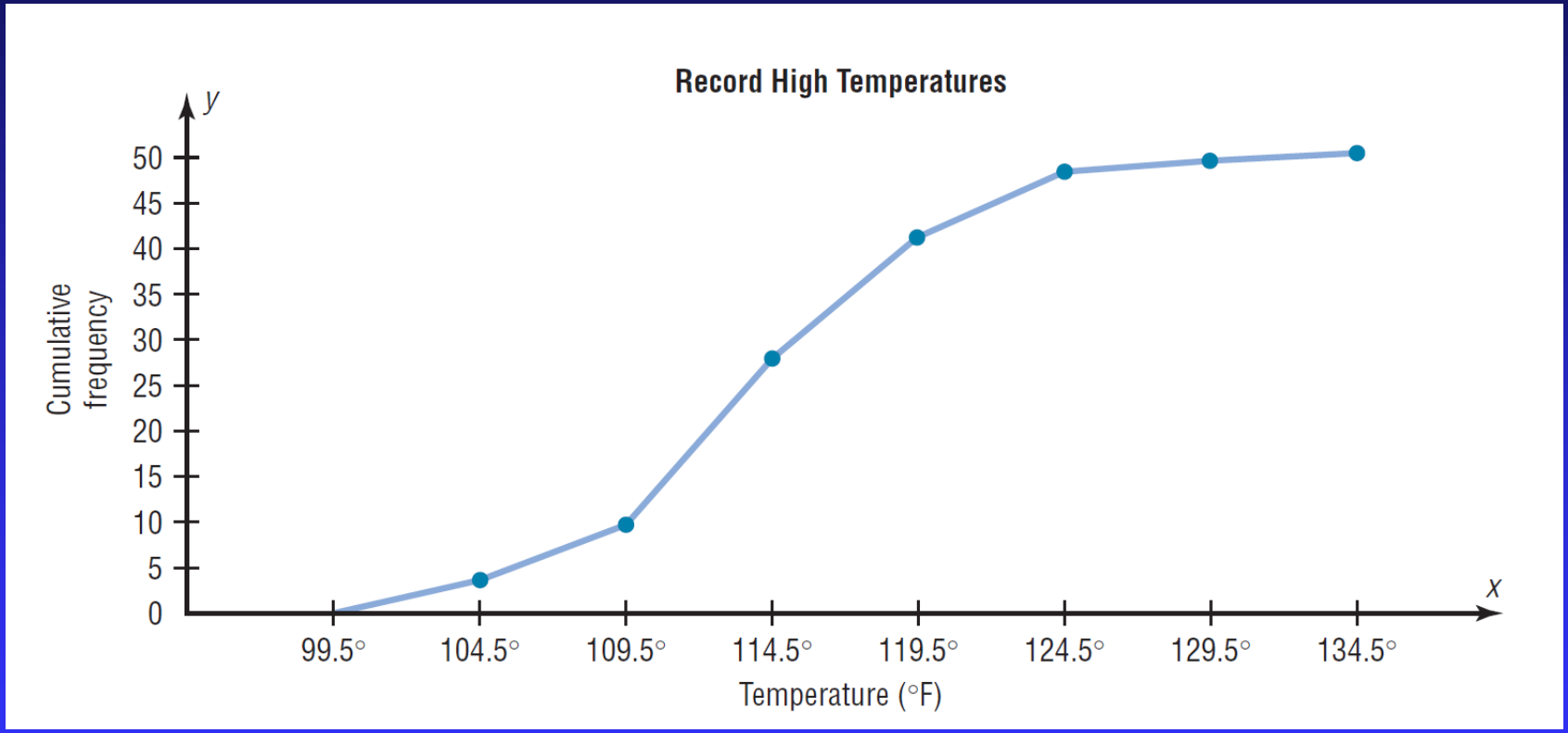
Ogive

The *cumulative frequency graph,* or *ogive,* shows the **cumulative frequency** for the upper boundary of each class. The cumulative frequency is the sum of the frequencies accumulated up to the upper boundary of a class. Here is the table of the cumulative frequencies:

	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

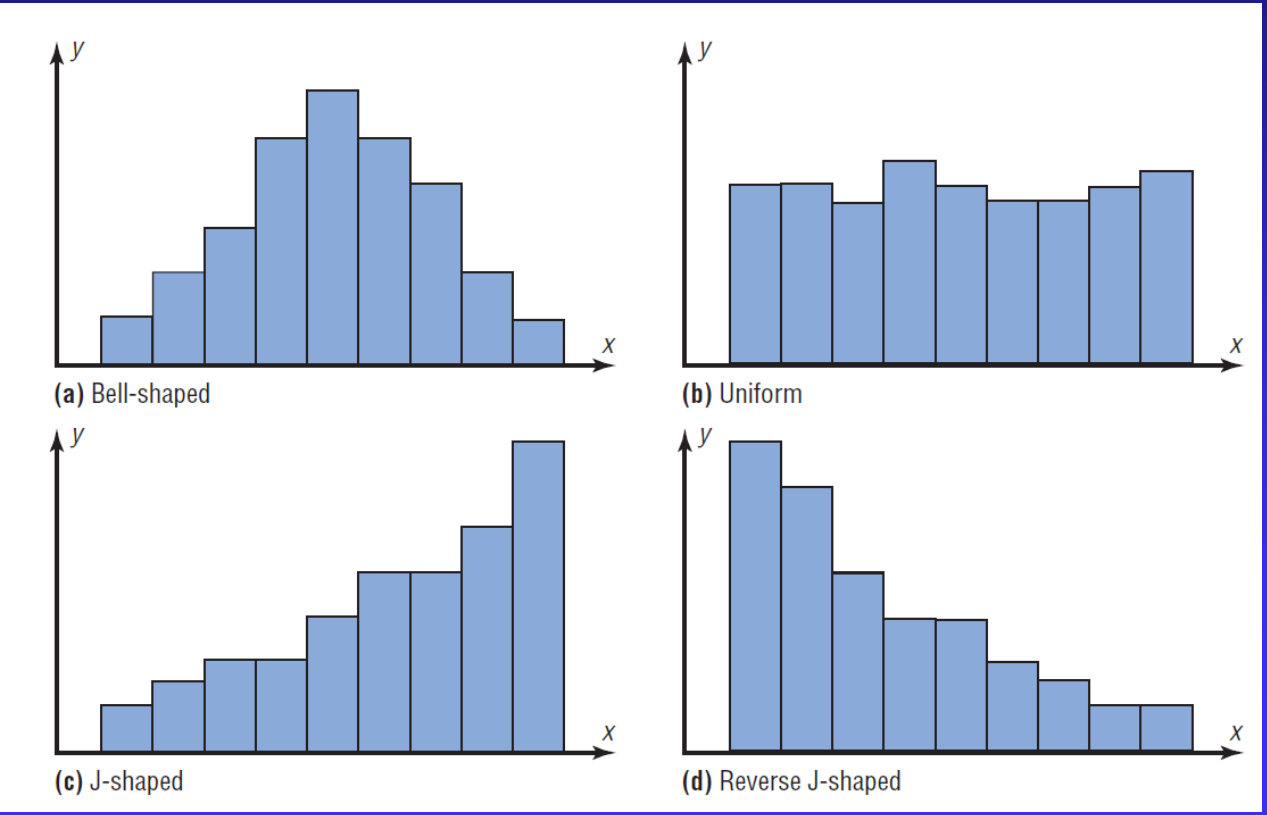
The ogive is shown on the next slide. The cumulative frequencies are plotted for each upper class boundary.

Ogive *(cont.)*

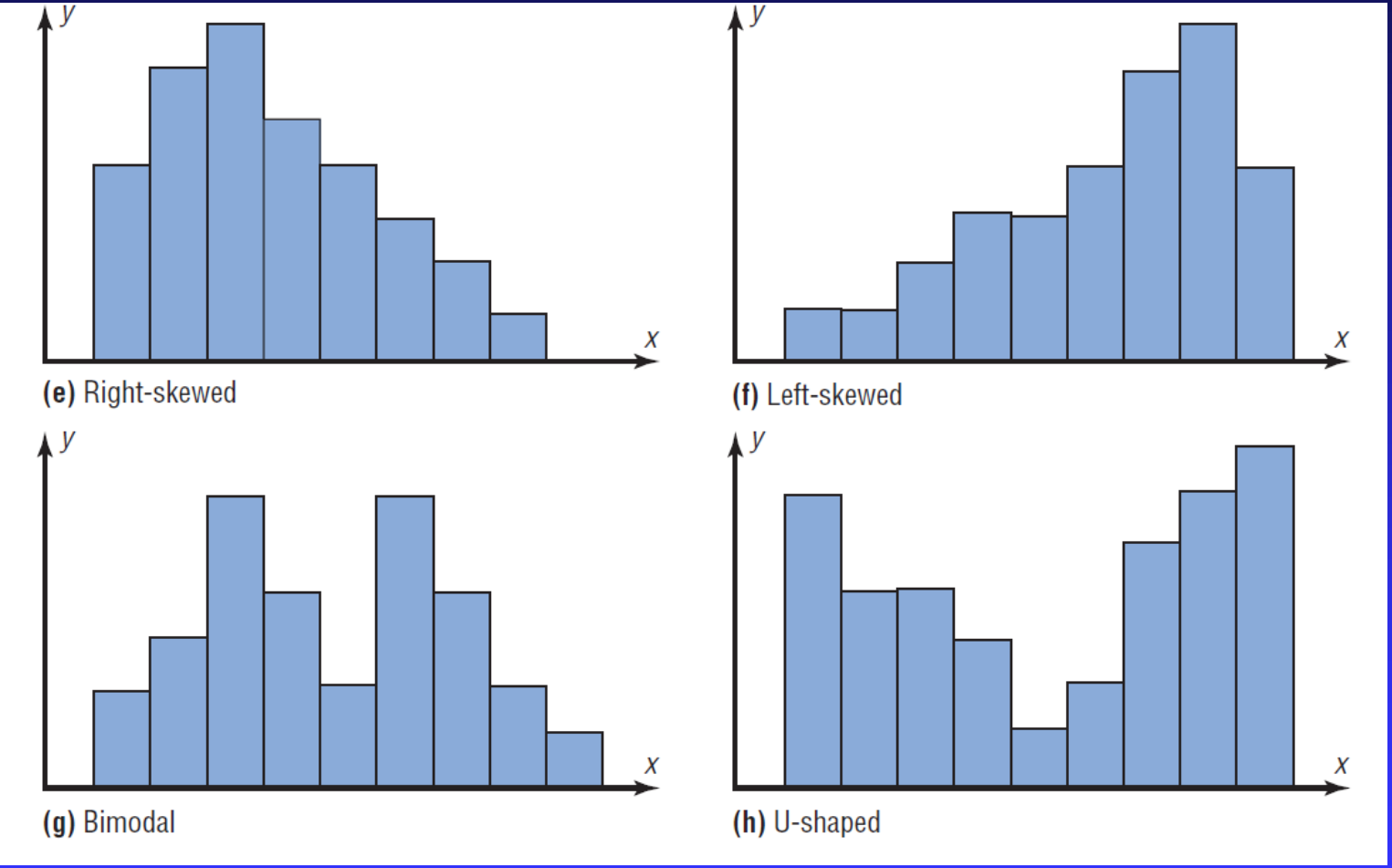

Record High Temperatures
y
50
45
40
35
30
25
20
15
10
5
0
Cumulative frequency
x
99.5°
104.5°
109.5°
114.5°
119.5°
124.5°
129.5°
134.5°
Temperature (°F)

Distribution shapes

A distribution can have many shapes, and one method of analyzing a distribution is to draw a histogram or frequency polygon for the distribution.

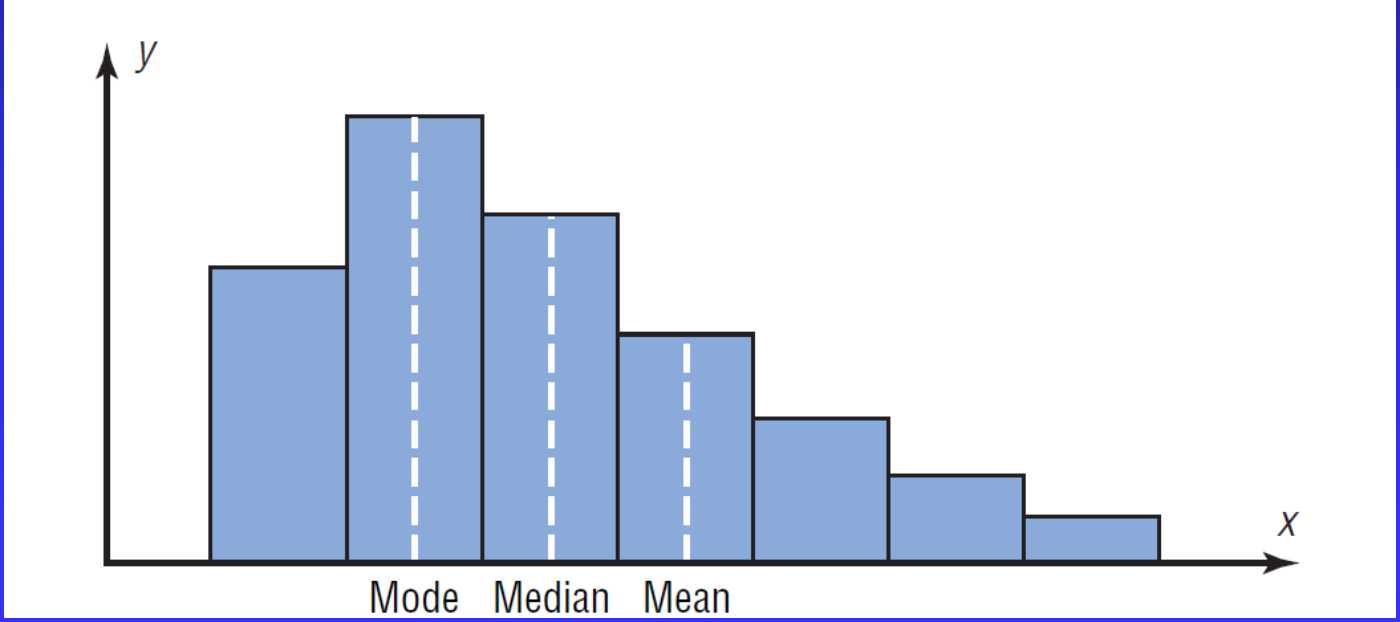


Distribution shapes *(cont.)*



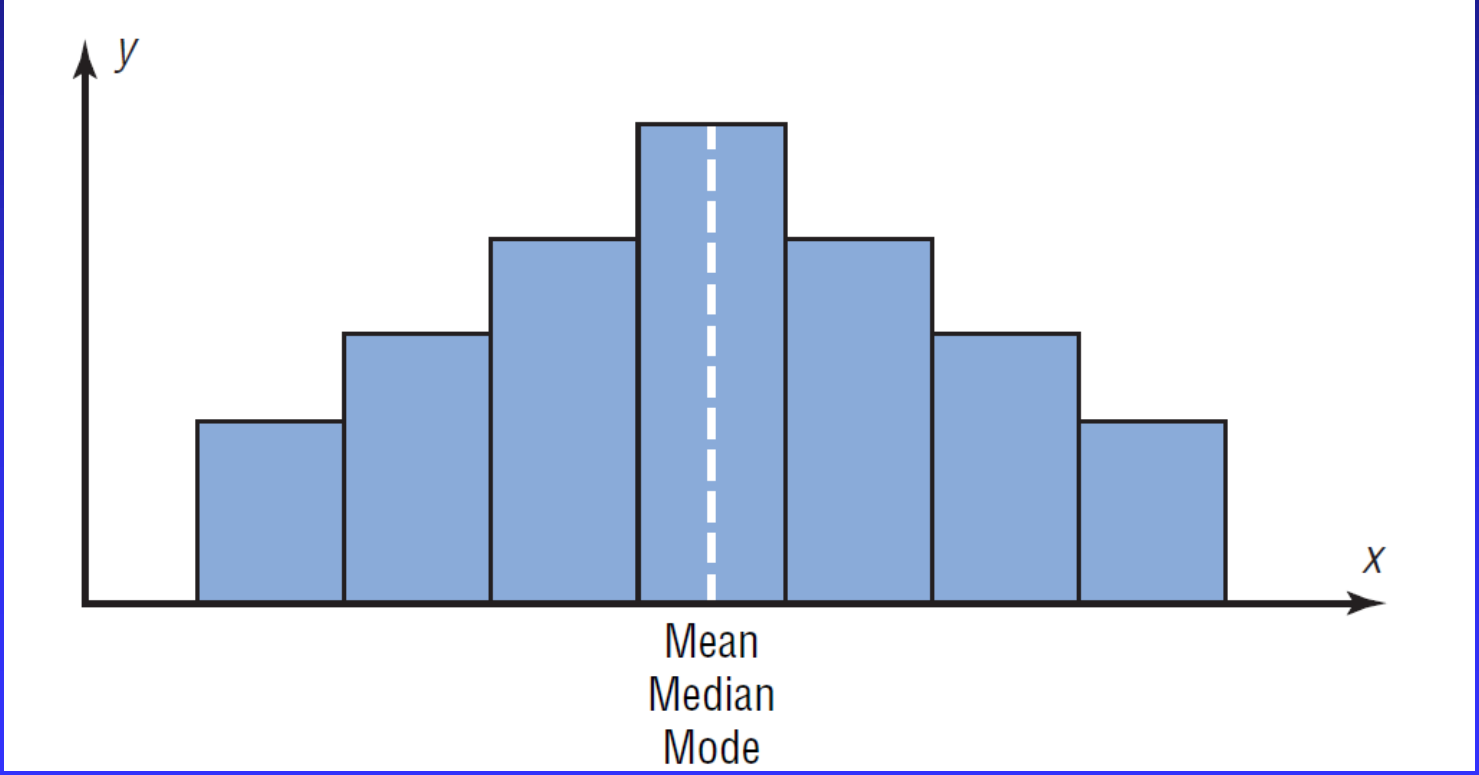
Right-skewed distribution

In a **positively skewed** or **right-skewed distribution,** the majority of the data values fall to the left of the mean and cluster at the lower end of the distribution; the “tail” is to the right. Also, the mean is to the right of the median, and the mode is to the left of the median. As an example, a negatively skewed distribution results if the majority of students score very low on an instructor’s examination. These scores will tend to cluster to the left of the distribution.



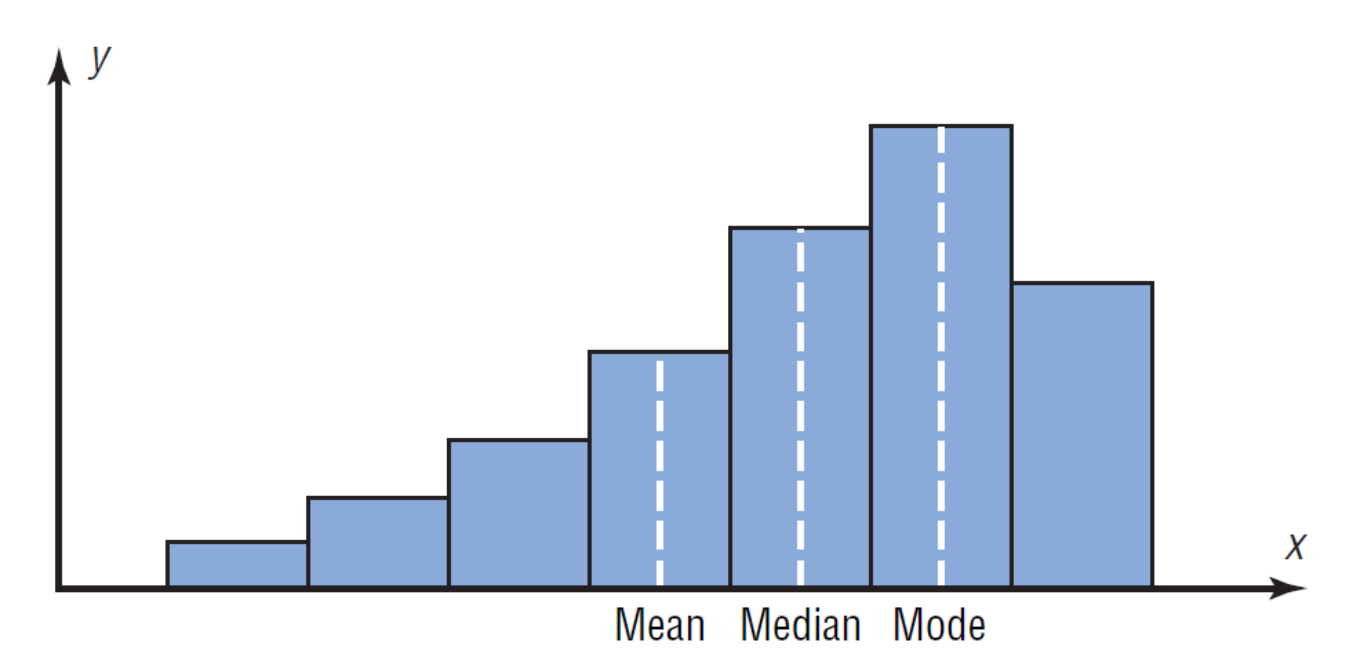
Symmetric distribution

In a **symmetric distribution,** the data values are evenly distributed on both sides of the mean. In addition, when the distribution is unimodal, the mean, median, and mode are the same and are at the center of the distribution.



Left-skewed distribution

When the majority of the data values fall to the right of the mean and cluster at the upper end of the distribution, with the tail to the left, the distribution is said to be **negatively skewed** or **left-skewed.** Also, the mean is to the left of the median, and the mode is to the right of the median. As an example, a negatively skewed distribution results if the majority of students score very high on an instructor's examination. These scores will tend to cluster to the right of the distribution.



Measures of Data variability

Data variability

Knowing the average of a data set is not enough to describe the data set entirely. Even though a shoe store owner knows that the average size of a man's shoe is size 10, she would not be in business very long if she ordered only size 10 shoes.

In addition to knowing the average, you must know how the data values are dispersed. That is, do the data values cluster around the mean, or are they spread more evenly throughout the distribution? The measures that determine the spread of the data values are called *measures of variation,* or *measures of dispersion.* These measures include the *range, variance,* and *standard deviation.* In general, **variability** is the statistical term for the degree to which data values differ from one another.

The range

The **range** is the highest value minus the lowest value, i.e.,

R = highest value - lowest value

The following data represents the salaries for the staff of some company:

	Staff Salary
Owner	$100,000
Manager	40,000
Sales representative	30,000
Workers	25,000
	15,000
	8,000

The range is R = $100,000 - $15,000 = $85,000.

Variance and standard deviation

The **variance** is the average of the squares of the distance each value is from the mean. The symbol for the population variance is σ^2 (σ is the Greek lowercase letter sigma).

The formula for the population variance is

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

where

X = individual value
μ = population mean
N = population size

The **standard deviation** is the square root of the variance. The symbol for the population standard deviation is σ.

The corresponding formula for the population standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

Variance and standard deviation *(cont.)*

Variance is always non-negative. If the variance is small, then the data values tend to be very close to the mean and to each other. If the variance is high, the data values are spread out around the mean and from each other. This information is useful in comparing two (or more) data sets to determine which is more (most) variable.

The reason to use the squared distances instead of the actual distances in the formula for the variance is that the actual distances could be negative or positive, with negative distances canceling out positive distances. For example, the following data values are spread around the zero: -11, -10, -9, -8, -7, -6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5 , 6, 7, 8, 9, 10, 11. The sum of the actual distances from zero (the mean) does not reflect the data spread because it is zero.

By taking the squared distances, we ignore the sign of the distances and work with squared distance magnitudes.

Variance and standard deviation *(cont.)*

The standard deviation has the same unit of measurement as the raw data. Therefore, we can compare the standard deviation to deviations from the mean. For example, a data value could be two standard deviations from the mean. In computing the variance, the distances are squared, and therefore, the units of the resultant numbers are the squares of the units of the original raw data. Finding the square root of the variance puts the standard deviation in the same units as the raw data.

The measures of variance and standard deviation are used to determine the consistency of a variable. For example, in the manufacture of fittings, such as nuts and bolts, the variation in the diameters must be small, or the parts will not fit together.

Variance and standard deviation *(cont.)*

As an example, lets find the variance and standard deviation for the following data values: 35, 45, 30, 35, 40, 25.

First, we need to find the mean:

$$\mu = \frac{\Sigma X}{N} = \frac{35 + 45 + 30 + 35 + 40 + 25}{6} = \frac{210}{6} = 35$$

Then, we subtract the mean from each value, and place the result in column B of the following table.

A X	B $X - \mu$	C $(X - \mu)^2$
35	0	0
45	10	100
30	−5	25
35	0	0
40	5	25
25	−10	100

Then, we square each result and place the squares in column C. Finally, we find the sum of the squares in column C and divide by N to find the variance:

$$\Sigma(X - \mu)^2 = 0 + 100 + 25 + 0 + 25 + 100 = 250$$

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} = \frac{250}{6} = 41.7$$

The standard deviation is the square root of the variance:

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} = \sqrt{41.7} = 6.5$$

Population vs. sample

The previous formulas for variance and standard deviation are called **population *variance and population standard deviation***. It is important to know the difference between population and a sample.

A ***population*** consists of ALL objects that are being studied.

Since it is impractical in most cases to study the whole population due to its prohibitive size (for example, the entire US population), experiments and observations are limited to samples from the population:

A ***sample*** is a group (a subset) of object selected from a population. If the subjects of a sample are properly selected, most of the time they should possess the same or similar characteristics as the subjects in the population.

Measures obtained by using all the data values in the population are called ***parameters***. On the other hand, measures obtained by using the data values from samples are called ***statistics***. As an example, the average of the sales from a sample of representatives is a *statistic,* whereas the average of sales obtained from the entire population is a *parameter.*

Population vs. sample *(cont.)*

In most cases the purpose of calculating a statistic is to estimate the corresponding parameter. For example, we might want to estimate the average of the whole population by calculating the average of a representative sample.

The formulas for the population variance and standard deviation does not give the best estimate of the population variance if they are applied to samples. For example, when the population is large and the sample is small (usually less than 30), the variance computed by this formula usually underestimates the population variance. Therefore, instead of dividing by n, researchers prefer to find the variance of the sample by dividing by n -1, giving a slightly larger value and an *unbiased* estimate of the population variance.

By definition, an estimate is unbiased if its expected value equals its true value (if you do not know what an expected value is, you van discard this sentence).

The formulas for sample variance and sample standard deviation provided on the next slide are better estimators of the population variance and standard deviation.

Sample variance and sample standard deviation

The formula for the sample variance, denoted by s^2, is

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$$

where

$\overline{X}$ = sample mean

n = sample size

The standard deviation of a sample (denoted by s) is

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}}$$

where

X = individual value

$\overline{X}$ = sample mean

n = sample size

Coefficient of variation

If two samples have the same units of measure, we can compare the spread of the samples by using the variance and standard deviation. For example, suppose that we want to compare the miles driven for different makes of second-hand cars. Suppose that that for a specific year, the standard deviation for Buicks was 422 miles and the standard deviation for Cadillacs was 350 miles. Then, we can conclude that the variation in mileage was greater in the Buicks. However, if we want to compare the standard deviations of two different variables, such as the number of sales per salesperson over a 3-month period and the commissions made by these salespeople?

The coefficient of variation is a statistic that allows you to compare standard deviations when the units are different:

The **coefficient of variation,** denoted by CVar, is the standard deviation divided by the mean. The result is expressed as a percentage.

For samples,

$$\text{CVar} = \frac{s}{\bar{X}} \cdot 100\%$$

For populations,

$$\text{CVar} = \frac{\sigma}{\mu} \cdot 100\%$$

Coefficient of variation *(cont.)*

For the example above, suppose that the mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Using the coefficient of variation we can compare the two variations.

CVar = (5/87)*100% = 5.7% sales

CVar = (773/5225)*100% = 14.8% commissions

Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

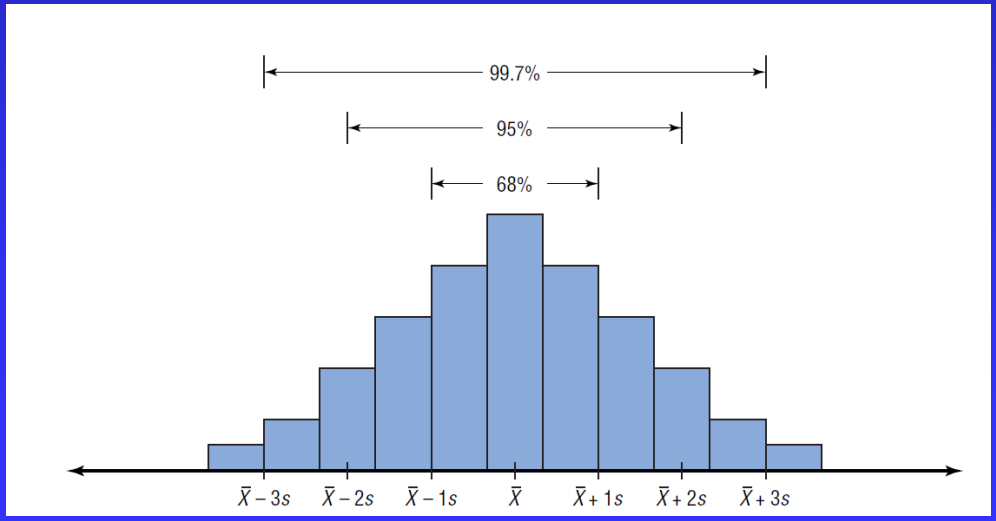
The Empirical (Normal) rule

Empirical rule: When a distribution is *bell-shaped* (or what is called *normal*), the following statements are true.

- Approximately 68% of the data values will fall within 1 standard deviation of the mean.

- Approximately 95% of the data values will fall within 2 standard deviations of the mean.

- Approximately 99.7% of the data values will fall within 3 standard deviations of the mean.



Measures of position

Measures of position

In addition to measures of central tendency and measures of variation, there are measures of position or location. The most frequently used measures of position are standard scores, percentiles, deciles, and quartiles.

The measures of position are used to locate the relative position of a data value in the data set. For example, if a value is located at the 80th percentile, it means that 80% of the values fall below it in the distribution and 20% of the values fall above it. The *median* is the value that corresponds to the 50th percentile, since one-half of the values fall below it and one half of the values fall above it.

A ***z* score** or **standard score** for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation. The symbol for a standard score is *z.* The formula is:

z = (value – mean) / standard deviation

A standard score or *z* score tells how many standard deviations a data value is above or below the mean for a specific distribution of values. If a standard score is zero, then the data value is the same as the mean.

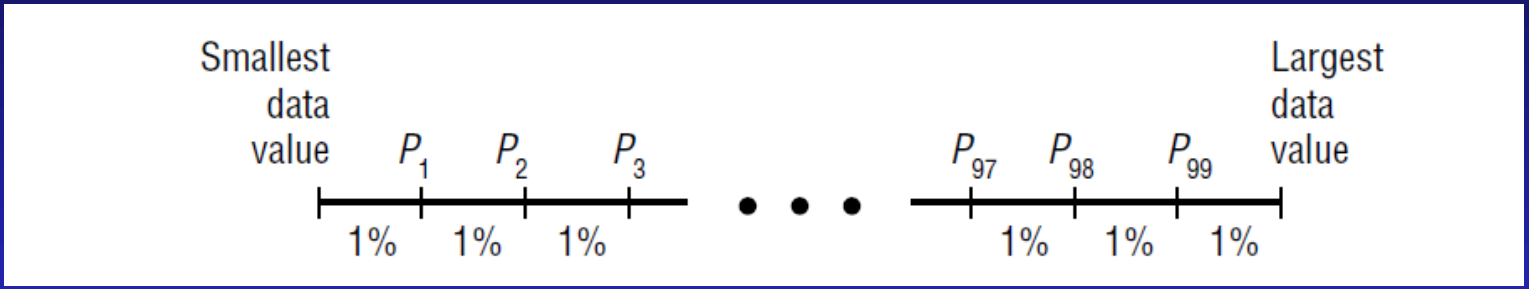
Standard score

Standard scores allow us to compare apples and oranges to some extent. For example, suppose that a student scored 65 on a music test and 30 on an English exam. It is impossible to compare the two scores directly because exams might differ in the number of questions, the value of each question, etc. However, we can impose the same standard on both tests and compare the score within the same standard. This standard is based on the concept of standard score, i.e., we can measure how far (in terms of standard deviations) is each score (65 and 30) from the mean of each exam. For the sake of illustration, suppose that the music test has a mean of 60 and a standard deviation of 10, and the English exam has a mean of 25 and a standard deviation of 5. Then the student's *z* score for the music exam is (65 – 60) / 10 = 0.5. His *z* score for the English exam is (30 – 25) / 5 = 1. Since the *z* score for the English is larger, his relative position in the English class is higher than his relative position in the music class.

In general, if the *z* score is positive, the score is above the mean. If the *z* score is 0, the score is the same as the mean. And if the *z* score is negative, the score is below the mean.

Percentiles

Percentiles divide the data set into 100 equal groups. For example, the 80th percentile is the value below which 80 percent of the observations may be found

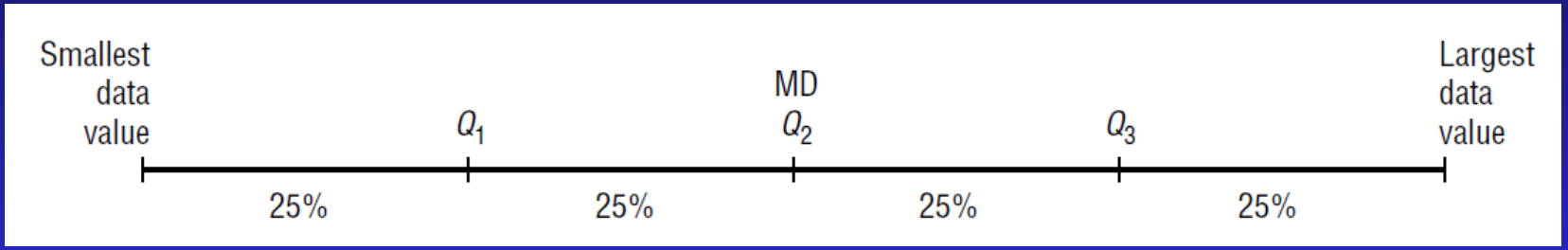


Percentiles, for example, are used to compare an individual's test score with the national norm. Tests such as the National Educational Development Test (NEDT) are taken by students in ninth or tenth grade. A student's scores are compared with those of other students locally and nationally by using percentile ranks.

Percentiles are not the same as percentages. That is, if a student gets 72 correct answers out of a possible 100, she obtains a percentage score of 72. There is no indication of her position with respect to the rest of the class. She could have scored the highest, the lowest, or somewhere in between. On the other hand, if a raw score of 72 corresponds to the 64th percentile, then she did better than 64% of the students in her class.

Quartiles

Quartiles divide the distribution into four groups. The quartiles are usually denoted by *Q*1, *Q*2, *Q*3. *Q*1 is the same as the 25th percentile; *Q*2 is the same as the 50th percentile, or the median; *Q*3 corresponds to the 75th percentile.



To find the quartiles you first need to arrange the data in order from lowest to highest. Then, find the median of the data values. This is the value for *Q*2. Find the median of the data values that fall below *Q*2. This is the value for *Q*1. Finally, find the median of the data values that fall above *Q*2. This is the value for *Q*3.

Example: suppose the data set is 15, 13, 6, 5, 12, 50, 22, 18. After arranging the data in order, we get 5, 6, 12, 13, 15, 18, 22, 50. Since the number of values is even, the median is (13 + 15)/2 = 14. Then, we need to find the median of the data values less than 14. These are 5, 6, 12, 13. Q1 = (6+12)/2 = 9. Finally, we need to find the median of the data values greater than 14. These are: 15, 18, 22, 50. Q3 = (18 + 22) / 2 = 20.

Outliers

In addition to dividing the data set into four groups, quartiles can be used as a rough measurement of variability. The **interquartile range (IQR)** is defined as the difference between *Q*1 and *Q*3 and is the range of the middle 50% of the data. The interquartile range is used to identify outliers.

An **outlier** is an extremely high or an extremely low data value when compared with the rest of the data values.

An outlier can strongly affect the mean and standard deviation of a variable. For example, suppose you have mistakenly recorded an extremely high data value. This value would then make the mean and standard deviation of the variable much larger than they really were.

There are no hard-and-fast rules on what to do with outliers, nor is there complete agreement among statisticians on ways to identify them. One procedure for identifying outliers recommends to search them among the values value that that smaller than *Q*1 - 1.5(IQR) or larger than *Q*3 + 1.5(IQR).

Let's find the outliers for the data set data set is 15, 13, 6, 5, 12, 50, 22, 18. We have already found that Q1 = 9 and Q3 = 20. Then IQR = 20 – 9 = 11. We need to search for outliers below *Q*1 - 1.5(IQR) = 9 – 1.5*11= - 7.5 and above *Q*3 + 1.5(IQR) = 20 + 1.5811 = 36.5. The value 50 is outside this interval. Therefore, it can be considered an outlier.

Exploratory data analysis (EDA)

Exploratory data analysis was developed by John Tukey and presented in his book *Exploratory Data Analysis* (Addison-Wesley, 1977).

The purpose of traditional analysis is to confirm various conjectures about the nature of the data.

The measure of central tendency used in EDA is the *median.* The measure of variation used in EDA is the *interquartile range* $Q3$ $Q1$. In EDA the data are represented graphically using a *boxplot* (sometimes called a box-and-whisker plot).

A **boxplot** can be used to graphically represent the data set. These plots involve five specific values:

1. The lowest value of the data set (i.e., minimum)

2. $Q1$

3. The median

4. $Q3$

5. The highest value of the data set (i.e., maximum)

These values are called a **five-number summary** of the data set.

Boxplot

A **boxplot** is a graph of a data set obtained by drawing a horizontal line from the minimum data value to $Q1$, drawing a horizontal line from $Q3$ to the maximum data value, and drawing a box whose vertical sides pass through $Q1$ and $Q3$ with a vertical line inside the box passing through the median or $Q2$.

Procedure for constructing a boxplot

1. Find the five-number summary for the data values, that is, the maximum and minimum data values, $Q1$ and $Q3$, and the median.

2. Draw a horizontal axis with a scale such that it includes the maximum and minimum data values.

3. Draw a box whose vertical sides go through $Q1$ and $Q3$, and draw a vertical line though the median.

4. Draw a line from the minimum data value to the left side of the box and a line from the maximum data value to the right side of the box

Boxplot *(cont.)*

Let's construct a boxplot for the following data 89, 47, 164, 296, 30, 215, 138, 78, 48, 39.

First, we need to arrange the data in order:

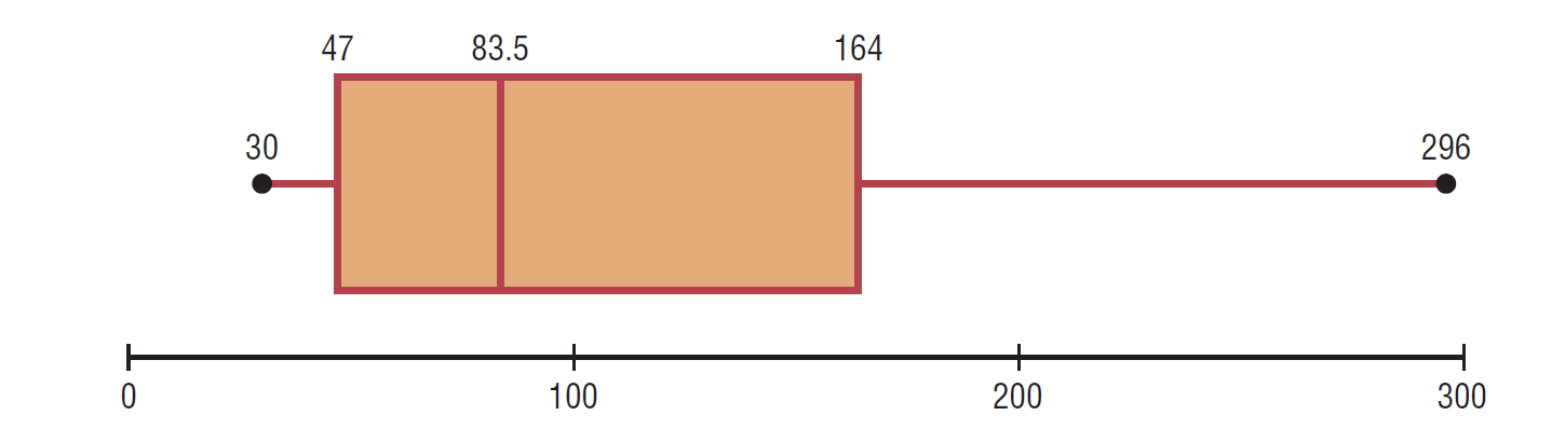
30, 39, 47, 48, 78, 89, 138, 164, 215, 296

Then, we find the median, Q2 = (78+89)/2=83.5

Q1=47

Q3=164

The boxplot is:



The distribution is somewhat positively skewed.

Information obtained from a Boxplot

1. If the median is near the center of the box, the distribution is approximately symmetric.

If the median falls to the left of the center of the box, the distribution is positively skewed.

If the median falls to the right of the center, the distribution is negatively skewed.

2. If the lines are about the same length, the distribution is approximately symmetric.

If the right line is larger than the left line, the distribution is positively skewed.

If the left line is larger than the right line, the distribution is negatively skewed.

Discrete Probability Distributions

Probability Distributions

Loosely speaking, a **random variable** is a variable whose values are determined by chance.

For example, if a die is rolled, a random variable *X* can be used to represent the outcomes. Then the value that *X* can assume is 1, 2, 3, 4, 5, or 6.

A Discrete variable has a finite number of possible values or an infinite number of values that can be counted. On the other hand, continuous random variables are obtained from data that can be measured rather than counted.

A **discrete probability distribution** consists of the values a discrete random variable can assume and the corresponding probabilities of the values. The probabilities are determined theoretically or by observation.

For example, here is a graphical representation of the probability distribution for the sample space for tossing three coins:

Number of heads X	0	1	2	3
Probability $P(X)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Mean, Variance, Standard Deviation, and Expectation

The mean, variance, and standard deviation for a probability distribution are computed differently from the mean, variance, and standard deviation for samples. For example, we cannot use the formula for the mean of a sample population if N is infinite (but still countable):

$$\mu = \frac{\Sigma X}{N}$$

Here is the formula for the mean (the average) for a discrete probability distribution:

Formula for the Mean of a Probability Distribution

The mean of a random variable with a discrete probability distribution is

$$\mu = X_1 \cdot P(X_1) + X_2 \cdot P(X_2) + X_3 \cdot P(X_3) + \cdots + X_n \cdot P(X_n)$$
$$= \Sigma X \cdot P(X)$$

where $X_1, X_2, X_3, \ldots, X_n$ are the outcomes and $P(X_1), P(X_2), P(X_3), \ldots, P(X_n)$ are the corresponding probabilities.

Note: $\Sigma X \cdot P(X)$ means to sum the products.

Mean, Variance, Standard Deviation, and Expectation *(cont.)*

As an example, let's find the mean of the number of spots that appear when a die is tossed. We will use the following graphical representation of the discrete distribution:

Outcome X	1	2	3	4	5	6
Probability $P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Hence, the mean $\mu = 1 * \frac{1}{6} + 2 * \frac{1}{6} + 3 * \frac{1}{6} + 4 * \frac{1}{6} + 5 * \frac{1}{6} + 6 * \frac{1}{6} = 3.5$

Another example, let's find the mean of the number of heads that occur if three coins are tossed. Here is the graphical representation:

Number of heads X	0	1	2	3
Probability $P(X)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

The mean $\mu = 0 * \frac{1}{8} + 1 * \frac{3}{8} + 2 * \frac{3}{8} + 3 * \frac{1}{8} = 1.5$

Mean, Variance, Standard Deviation, and Expectation *(cont.)*

To find the variance for the random variable of a probability distribution, subtract the theoretical mean of the random variable from each outcome and square the difference. Then multiply each difference by its corresponding probability and add the products. The formula is:

$$\sigma^2 = \Sigma[(X - \mu)^2 \cdot P(X)]$$

Finding the variance by using this formula is somewhat tedious. So for simplified manual computations, a shortcut formula can be used. This formula is algebraically equivalent to the longer one:

Formula for the Variance of a Probability Distribution

Find the variance of a probability distribution by multiplying the square of each outcome by its corresponding probability, summing those products, and subtracting the square of the mean. The formula for the variance of a probability distribution is

$$\sigma^2 = \Sigma[X^2 \cdot P(X)] - \mu^2$$

The standard deviation of a probability distribution is

$$\sigma = \sqrt{\sigma^2} \quad \text{or} \quad \sqrt{\Sigma[X^2 \cdot P(X)] - \mu^2}$$

Mean, Variance, Standard Deviation, and Expectation *(cont.)*

Let's find the variance and the standard deviation of the number of spots that appear when a die is tossed.

$$\sigma^2 = 1^2 * \frac{1}{6} + 2^2 * \frac{1}{6} + 3^2 * \frac{1}{6} + 4^2 * \frac{1}{6} + 5^2 * \frac{1}{6} + 6^2 * \frac{1}{6} - 3.5^2 = 2.9$$

$$\sigma = \sqrt{\sigma^2} = 1.7$$

Another concept related to the mean for a probability distribution is that of expected value or expectation.

The **expected value** of a discrete random variable of a probability distribution is the theoretical average of the variable. The formula is

$$\mu = E(X) = \Sigma X \cdot P(X)$$

The symbol $E(X)$ is used for the expected value.

The formula for the expected value is the same as the formula for the theoretical mean. The expected value, then, is the theoretical mean of the probability distribution. That is, $E(X) = \mu$.

Mean, Variance, Standard Deviation, and Expectation *(cont.)*

An example: A financial adviser suggests that you select one of two types of bonds in which to invest $5000. Bond *X* pays a return of 4% and has a default rate of 2%. Bond *Y* has a return 2.5% and a default rate of 1%. Which bond is a better investment on average? To answer the question we need to find the expected rate of return and decide which bond would be a better investment.

The return on bond *X* is $5000 * 4% = %200. The expected return then is:

$200 * 0.98 + (-$5000) * 0.02 = $96

Here, 0.98 is the probability of positive return and 0.02 is the probability of default, in which case the lost is -$5000. In other words, we get a positive return of $200 with probability 0.98 and we lose $5000 with probability 0.02.

The return on bond *Y* is $5000 * 2.5% = $125. The expected return then is:

$125 * 0.99 + (-$5000) * 0.01 = $73.75

In other words, we get $125 with probability 0.99 and we lose $5000 with probability 0.01.

Hence, bond *X* would be a better investment since the expected return is higher.

The Binomial Distribution

Many types of probability problems have only two outcomes or can be reduced to two outcomes. For example, when a coin is tossed, it can land heads or tails. When a baby is born, it will be either male or female. In a basketball game, a team either wins or loses. A true/false item can be answered in only two ways, true or false. Situations like these are called *binomial experiments.*

A **binomial experiment** is a probability experiment that satisfies the following four requirements:

1. There must be a fixed number of trials.
2. Each trial can have only two outcomes or outcomes that can be reduced to two outcomes. These outcomes can be considered as either success or failure.
3. The outcomes of each trial must be independent of one another.
4. The probability of a success must remain the same for each trial.

The outcomes of a binomial experiment and the corresponding probabilities of these outcomes are called a **binomial distribution.** In binomial experiments, the outcomes are usually classified as successes or failures.

The following notation is commonly used for binomial experiments and the binomial distribution:

The Binomial Distribution *(cont.)*

$P(S)$	The symbol for the probability of success
$P(F)$	The symbol for the probability of failure
p	The numerical probability of a success
q	The numerical probability of a failure

$P(S) = p$ and $P(F) = 1 - p = q$

n	The number of trials
X	The number of successes in n trials

Note that $0 \leq X \leq n$ and $X = 0, 1, 2, 3, \ldots$ $P(X) = \frac{n!}{(n-X)!X!} \cdot p^X \cdot q^{n-X}$

In a binomial experiment, the probability of exactly X successes in n trials is:

$$P(X) = \frac{n!}{(n-X)!X!} \cdot p^X \cdot q^{n-X}$$

The Binomial Distribution *(cont.)*

A coin is tossed 3 times. Let's find the probability of getting exactly two heads. This is binomial distribution because: we have three trials; there are only two outcomes for each trial, heads or tails; the outcomes are independent of one another (the outcome of one toss in no way affects the outcome of another toss); and the probability of a success (heads) is ½ in each case. Therefore n = 3, X = 2, p = ½ , q = ½ . Substituting in the formula yields:

P(2 heads) = $\frac{3!}{(3-2)!*2!} * \left(\frac{1}{2}\right)^2 * \left(\frac{1}{2}\right)^{(3-2)} = \frac{3}{8}$

The mean, variance, and standard deviation of a variable that has the *binomial distribution* can be found by using the following formulas.

Mean: μ = n*p

Variance: σ^2 =n*p*q

Standard deviation: σ =$\sqrt{n * p * q}$

The Poisson Distribution

A discrete probability distribution that is useful when *n* is large and *p* is small and when the independent variables occur over a period of time is called the **Poisson distribution.** In addition to being used for the stated conditions (i.e., *n* is large, *p* is small, and the variables occur over a period of time), the Poisson distribution can be used when a density of items is distributed over a given area or volume, such as the number of plants growing per acre or the number of defects in a given length of videotape.

The probability of *X* occurrences in an interval of time, volume, area, etc., for a variable where λ (Greek letter lambda) is the mean number of occurrences per unit (time, volume, area, etc.) is:

$$P(X;\lambda) = \frac{e^{-\lambda}\lambda^{X}}{X!} \qquad \text{where } X = 0, 1, 2, \ldots$$

The letter *e* is a constant approximately equal to 2.7183

The Poisson Distribution *(cont.)*

Example: there are 200 typographical errors randomly distributed in a 500-page manuscript. Let's find the probability that a given page contains exactly 3 errors.

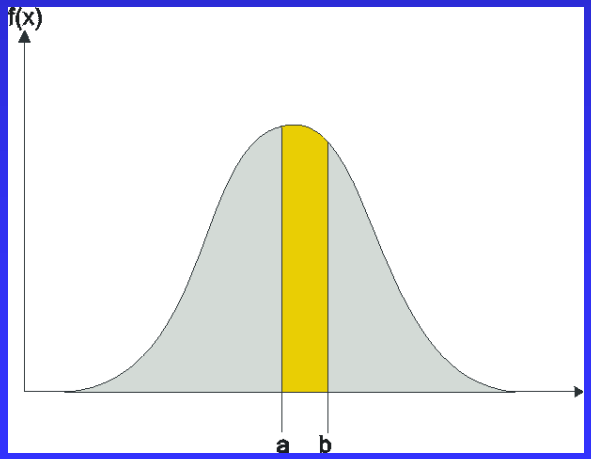
We need first to find the mean number λ of errors. Since there are 200 errors distributed over 500 pages, each page has an average of $\lambda = \frac{200}{500} = 0.4$ or 0.4 error per page. Since $X = 3$, substituting into the formula yields:

$$P(X; \lambda) = \frac{e^{-\lambda}\lambda^{x}}{X!} = \frac{2.7183^{-0.4}0.4^{3}}{3!} = 0.0072$$

Continuous distributions

Probability density function

Since continuous random variables can assume a continuum of values, we cannot explicitly specify the probability of each value as we do with discrete distributions. Instead, continuous distributions are described by probability density functions. **A probability density function** (**pdf**), of a continuous random variable describes the relative likelihood for this random variable to fall within a particular range of values. This is graphically depicted by the area under the density function curve. For example, if f(x) is a pdf, the probability that the value of the random variable x is between a and b, $P(a \leq x \leq b)$, equals the area painted in yellow on the figure below:



The Normal Distribution

The mathematical equation for the pdf for a **normal distribution** is:

$$y = \frac{e^{-(X-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$

where

$e \approx 2.718$ (≈ means “is approximately equal to”)

$p \approx 3.14$

μ = population mean

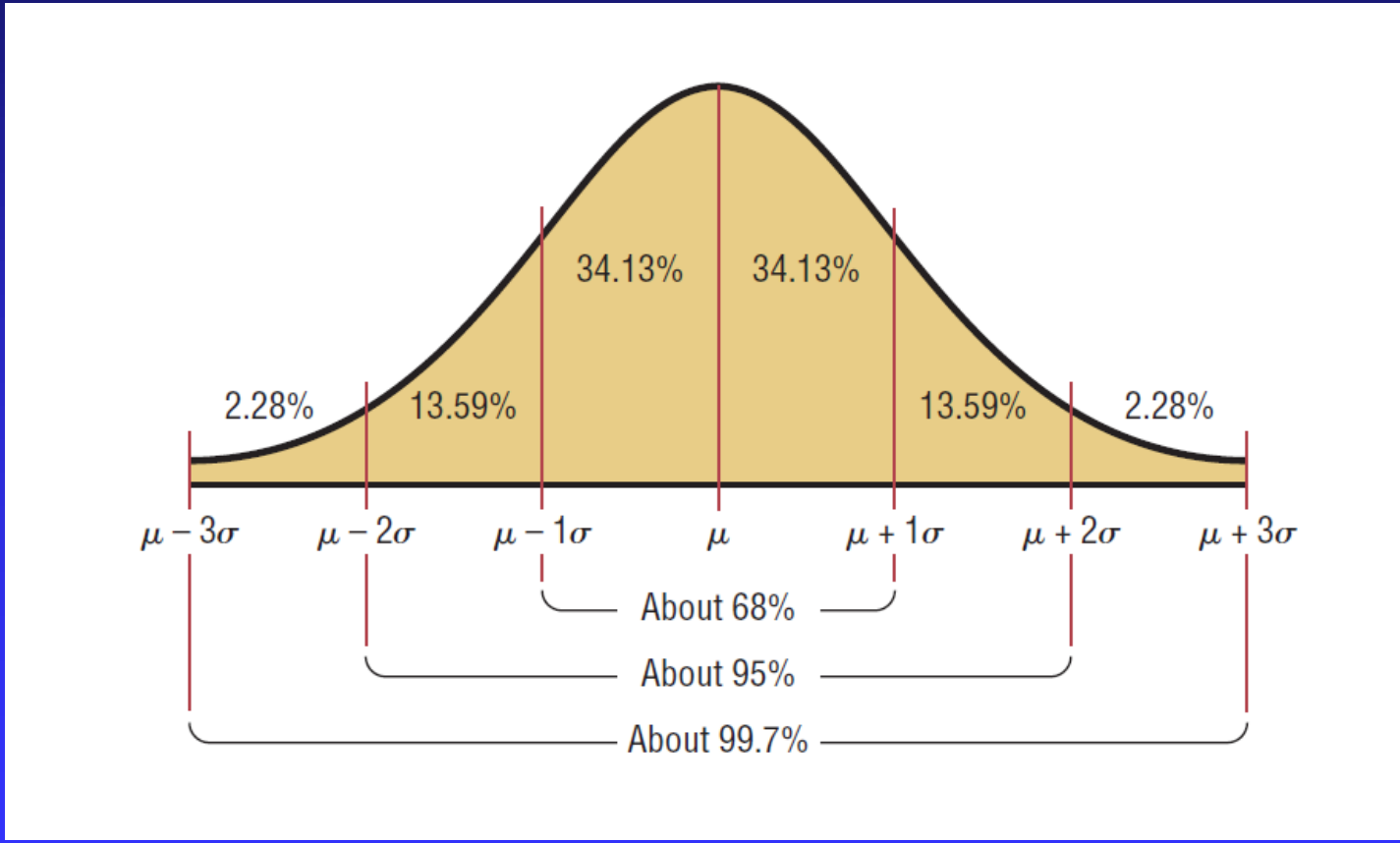
σ = population standard deviation

The formula looks quite complicated. In practice, it is important to remember that **normal distribution** is a continuous, symmetric, bell-shaped distribution of a variable.

Summary of the Properties of the Theoretical Normal Distribution

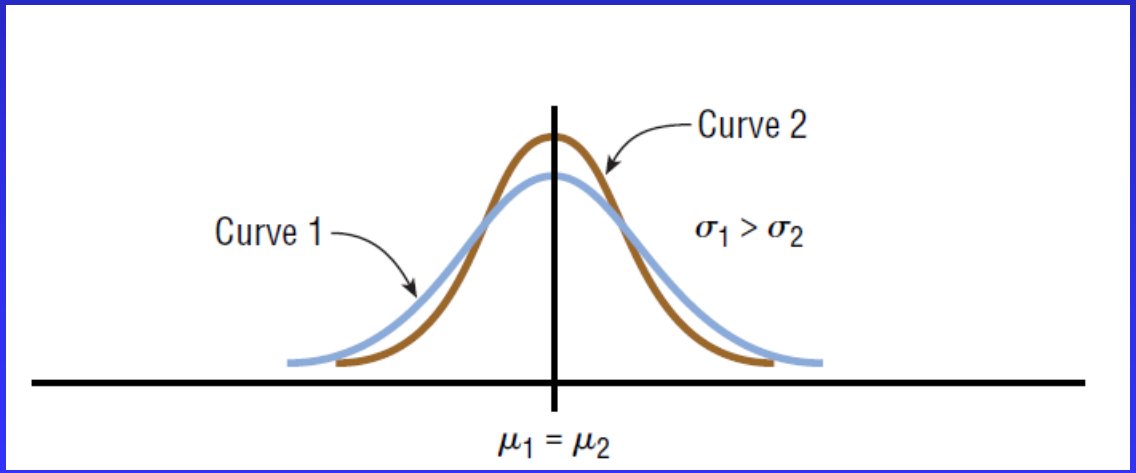
1. A normal distribution curve is bell-shaped.
2. The mean, median, and mode are equal and are located at the center of the distribution.
3. A normal distribution curve is unimodal (i.e., it has only one mode).
4. The curve is symmetric about the mean, which is equivalent to saying that its shape is the same on both sides of a vertical line passing through the center.
5. The curve is continuous; that is, there are no gaps or holes. For each value of *X*, there is a corresponding value of *Y.*
6. The curve never touches the *x* axis. Theoretically, no matter how far in either direction the curve extends, it never meets the *x* axis—but it gets increasingly closer.
7. The total area under a normal distribution curve is equal to 1.00, or 100%. This fact may seem unusual, since the curve never touches the *x* axis, but one can prove it mathematically by using calculus. (The proof is beyond the scope of this textbook.)
8. 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%.

The Normal Distribution *(cont.)*



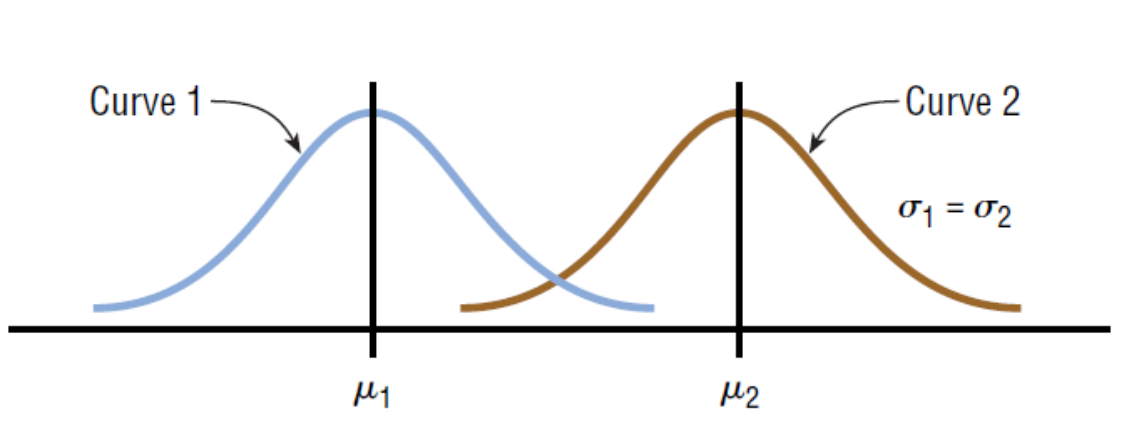
The Normal Distribution *(cont.)*

The shape and position of a normal distribution curve depend on two parameters, the *mean* and the *standard deviation.* Each normally distributed variable has its own normal distribution curve, which depends on the values of the variable's mean and standard deviation. The figure below shows two normal distributions with the same mean values but different standard deviations. The larger the standard deviation, the more dispersed, or spread out, the distribution is.



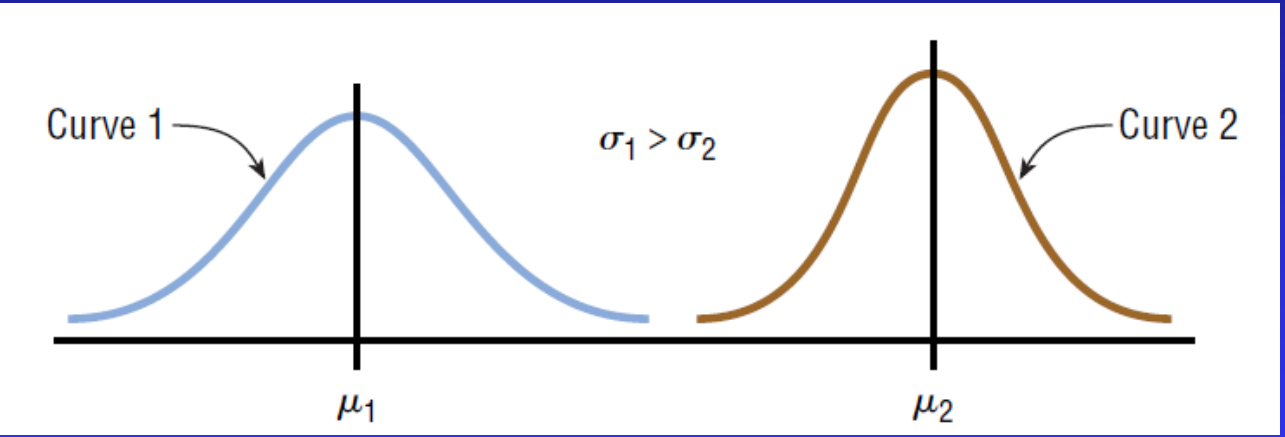
The Normal Distribution *(cont.)*

The next figure shows two normal distributions with the same standard deviation but with different means. These curves have the same shapes but are located at different positions on the *x* axis.



The Normal Distribution *(cont.)*

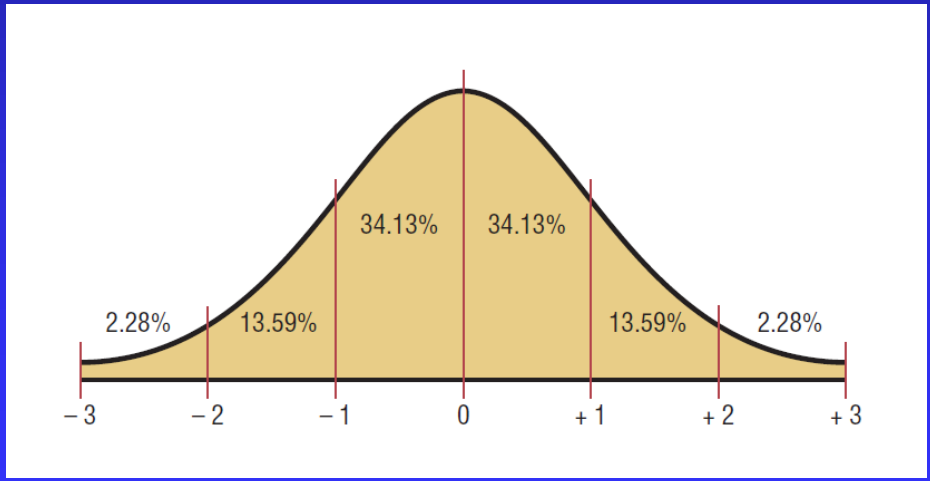
The figure below shows two normal distributions with different means and different standard deviations.



The Standard Normal Distribution

Since each normally distributed variable has its own mean and standard deviation, as stated earlier, the shape and location of these curves will vary. In practical applications, then, you would have to have a table of areas under the curve for each variable. In addition, the comparison between two different normal distributions might not be straightforward. To simplify the situation, statisticians use what is called the *standard normal distribution.*

The **standard normal distribution** is a normal distribution with a mean of 0 and a standard deviation of 1. This simplifies calculations.



For example, the formula of the pdf of standard normal distribution is:

$$y = \frac{e^{-z^2/2}}{\sqrt{2\pi}}$$

The Standard Normal Distribution *(cont.)*

All normally distributed variables can be transformed into the standard normally distributed variable by using the formula for the standard score:

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \quad \text{or} \quad z = \frac{X - \mu}{\sigma}$$

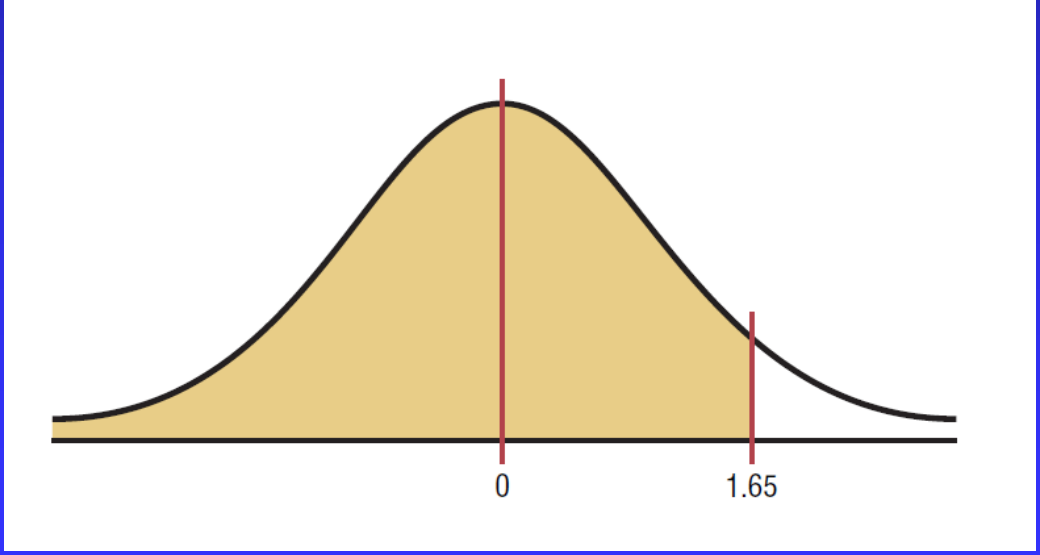
The ***z* value** is actually the number of standard deviations that a particular *X* value is away from the mean.

The probability that a normally distributed random variable falls within a given range can be computed using the pdf. Remember that probabilities are depicted as areas below the curve.

For example, if we want to find $P(x \leq 1.65)$, we need to find the are below the curve which is to the left of 1.65.

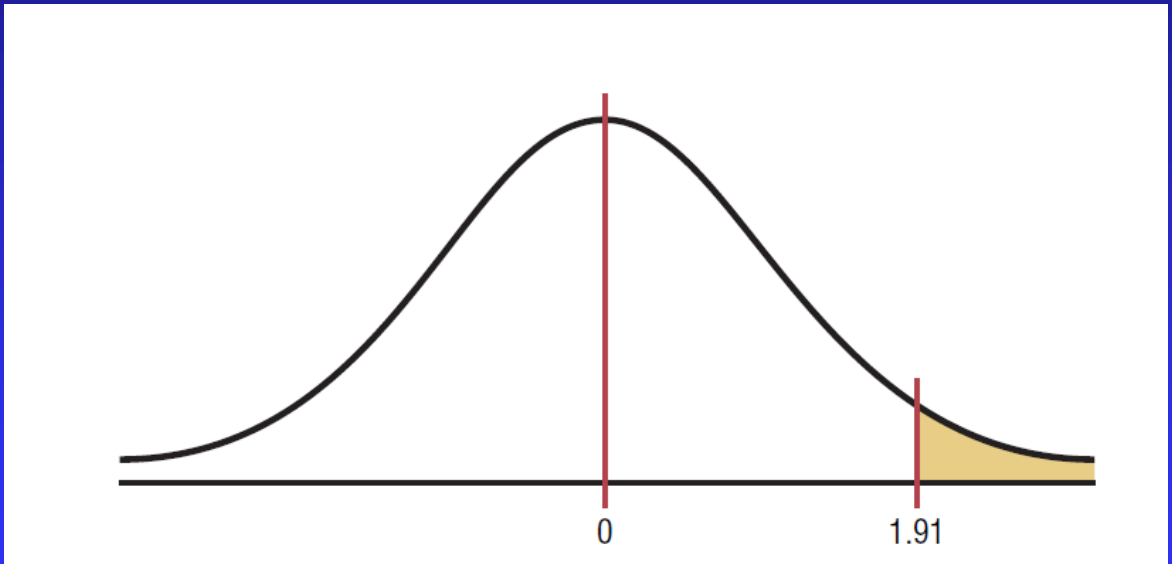
The Standard Normal Distribution *(cont.)*

This can be done by using tables, calculators, or statistical packages. For example, you can use the R function called *pnorm()* with an attribute 1.65. *pnorm(1.65)* returns 0.9505. This means that the probability that x is equal to or smaller than 1.65 is 95.05%. We will learn how to run R commands in a separate lecture. The thing to remember is that the probability of 95.05% equals the area below the pdf curve and to the left of 1.65:



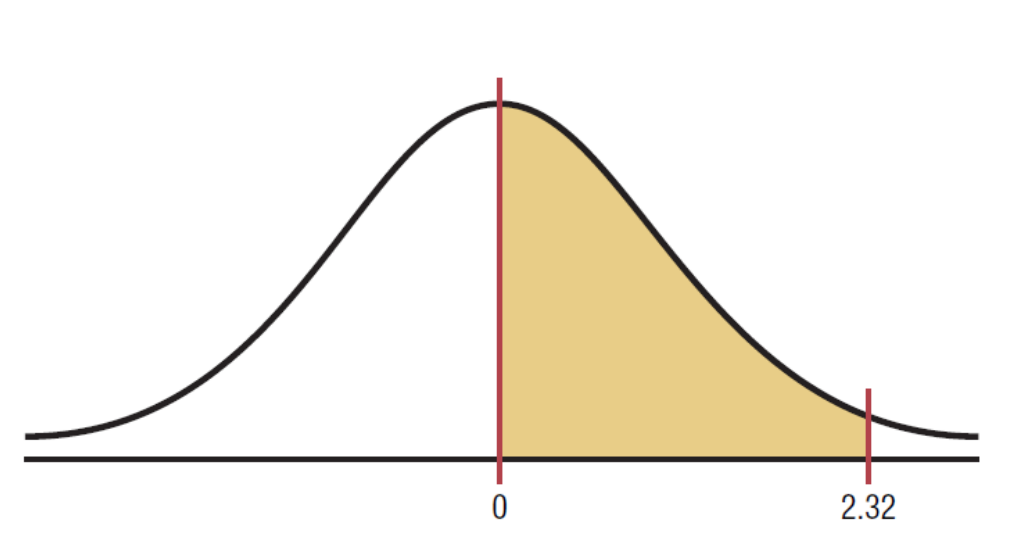
The Standard Normal Distribution *(cont.)*

Similarly, the probability that X is greater than 1.91, $P(x > 1.91)$, corresponds to the area below the curve and to the right of 1.91:



The Standard Normal Distribution *(cont.)*

Finally, the probability that x falls in the interval [0, 2.32] is given by the area below the pdf curve and between 0 and 2.32:



Confidence intervals

One aspect of inferential statistics is **estimation,** the process of estimating the value of a parameter from information obtained from a sample. For example, we can use a sample to estimate the population mean, proportion, variance, and standard deviation.

There are two types of estimates:

- point estimate: specific numerical value estimate of a parameter. For example, the best point estimate of the population mean μ is the sample mean $\bar{X}$.
- interval estimate: an interval or a range of values used to estimate the population parameter. This estimate (derived from a specific sample) may or may not contain the value of the parameter being estimated.

Intervals constructed in this way are called *confidence intervals.* Three common confidence intervals are used: the 90, the 95, and the 99% confidence intervals. For example, approximately 95% of the sample means taken from a population and same sample size will fall within 1.96 standard errors of the population mean, that is,

$$\mu \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

Confidence intervals

Stated in another way:

$$\bar{X} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

The value used for the 95% confidence interval is called *confidence level*. For a 99% confidence interval, the value 2.58 is used instead of 1.96 in the formula.

Formula for the confidence interval for the mean for specific alpha:

$$\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

For a 90% confidence interval, $z_{\alpha/2} = 1.65$; for a 95% confidence interval, $z_{\alpha/2} = 1.96$; and for a 99% confidence interval, $z_{\alpha/2} = 2.58$.

The term $z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$ is called the *maximum error of the estimate* (also called the *margin of error*).

Hypothesis testing

Researchers are often interested in answering types of questions, such as: a medical researcher is interested in finding out whether a new medication will change patients' heart rate as a side effect; a chemist invents an additive to increase the life of an automobile battery and wants to know weather the additive really increases the battery lifetime. These types of questions can be addressed through statistical **hypothesis testing,** which is a decision-making process for evaluating claims about a population. In hypothesis testing, the researcher must define the population under study, state the particular hypotheses that will be investigated, give the significance level, select a sample from the population, collect the data, perform the calculations required for the statistical test, and reach a conclusion.

A **statistical hypothesis** is a conjecture about a population parameter. This conjecture may or may not be true. There are two types of statistical hypotheses for each situation:

- The **null hypothesis,** symbolized by H_0, is a statistical hypothesis that states that there is no difference between a parameter and a specific value, or that there is no difference between two parameters.
- The **alternative hypothesis,** symbolized by H_1, is a statistical hypothesis that states the existence of a difference between a parameter and a specific value, or states that there is a difference between two parameters.

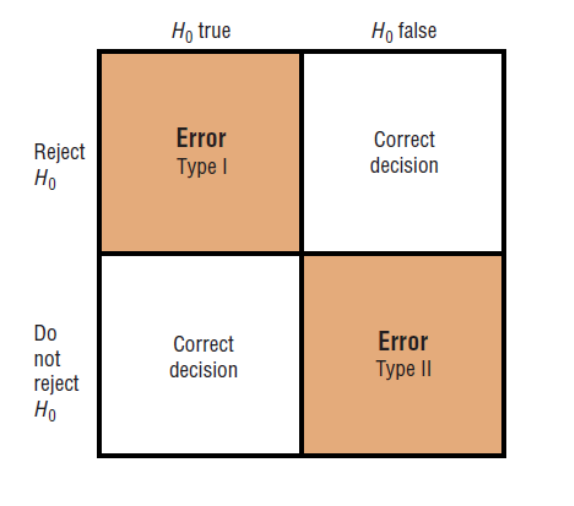
Hypothesis testing

For example, a researcher is particularly concerned with the pulse rate of the patients who take a new medication. Will the pulse rate increase, decrease, or remain unchanged after a patient takes the medication? Since the researcher knows that the mean pulse rate for the population under study is 82 beats per minute, the hypotheses for this situation are H_0: $\mu = 82$ and H_1: $\mu \neq 82$. The null hypothesis specifies that the mean will remain unchanged, and the alternative hypothesis states that it will be different. This test is called a *two-tailed test* (a term that will be formally defined later), since the possible side effects of the medicine could be to raise or lower the pulse rate.

In the second example, the chemists wants to determine if the additive increases the life of an automobile battery. If the mean lifetime of the automobile battery without the additive is 36 months, then her hypotheses are H_0: $\mu = 36$ and H_1: $\mu > 36$ In this situation, the chemist is interested only in increasing the lifetime of the batteries, so her alternative hypothesis is that the mean is greater than 36 months. The null hypothesis is that the mean is equal to 36 months. This test is called *right-tailed,* since the interest is in an increase only.

Hypothesis testing

- The researcher performs a **statistical test** that uses the data obtained from a sample to make a decision about whether the null hypothesis should be rejected. The numerical value obtained from a statistical test is called the **test value.**
- In the hypothesis-testing situation, there are four possible outcomes. In reality, the null hypothesis may or may not be true, and a decision is made to reject or not reject it on the basis of the data obtained from a sample. Notice that there are two possibilities for a correct decision and two possibilities for an incorrect decision. A **type I error** occurs if you reject the null hypothesis when it is true. A **type II error** occurs if you do not reject the null hypothesis when it is false.

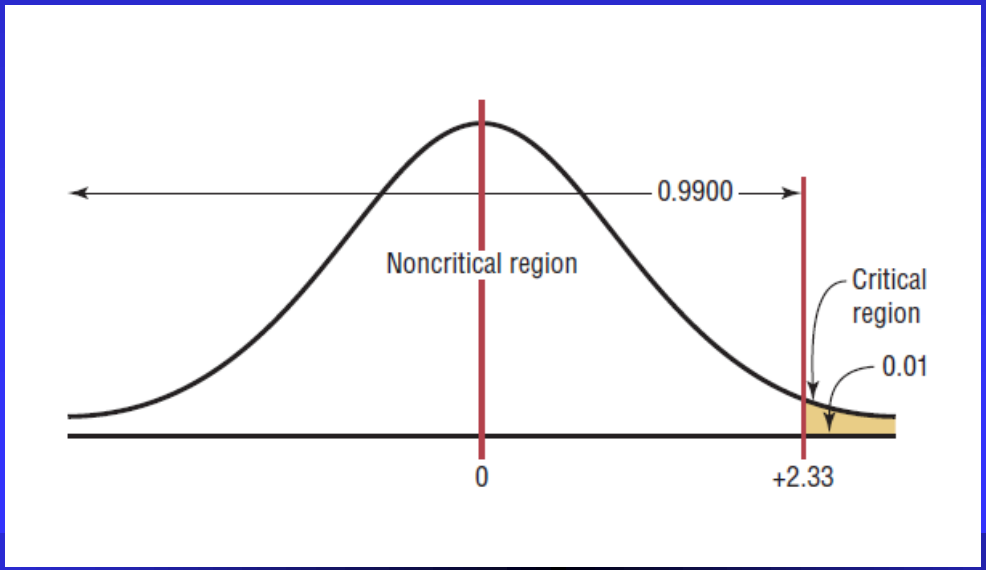


Hypothesis testing

- The **level of significance** is the maximum probability of committing a type I error. This probability is symbolized by α. That is, *P*(type I error) = α.
- The **critical value** separates the critical region from the noncritical region.
- The **critical** or **rejection region** is the range of values of the test value that indicates that there is a significant difference and that the null hypothesis should be rejected.
- The **noncritical** or **non-rejection region** is the range of values of the test value that indicates that the difference was probably due to chance and that the null hypothesis should not be rejected.
- The critical value can be on the right side of the mean or on the left side of the mean for a one-tailed test. Its location depends on the inequality sign of the alternative hypothesis. For example, in the satiation where the chemist is interested in increasing the average battery, the alternative hypothesis is H_1: μ > 36 . Since the inequality sign is > , the null hypothesis will be rejected only when the sample mean is significantly greater than 36. Hence, the critical value must be on the right side of the mean.

Hypothesis testing

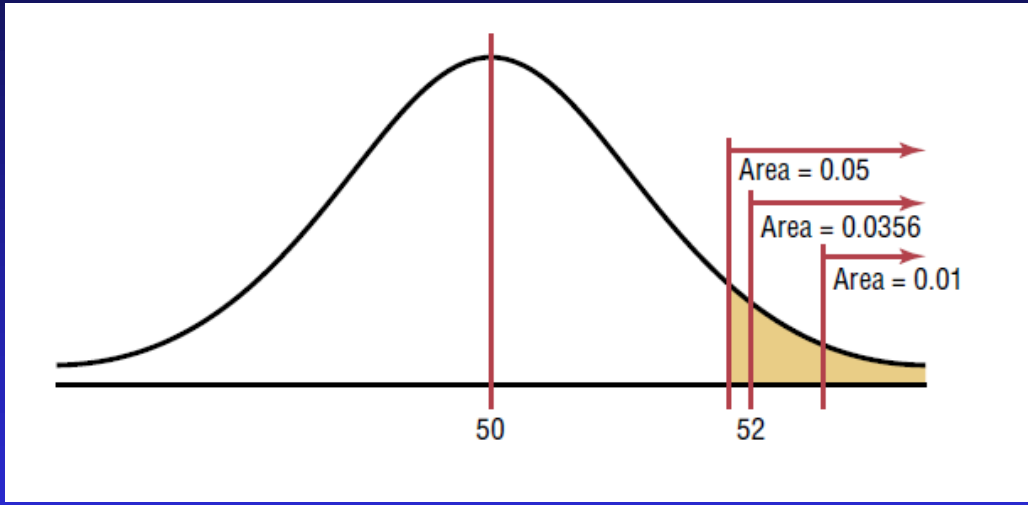
- A **one-tailed test** indicates that the null hypothesis should be rejected when the test value is in the critical region on one side of the mean. A one-tailed test is either a **right-tailed test** or **left-tailed test,** depending on the direction of the inequality of the alternative hypothesis.
- To obtain the critical value, the researcher must choose an alpha level for the battery life. Suppose the researcher choses α = 0.01. Then the researcher must find a value such that 1% of the area falls to the right of the value and 99% falls to the left of the value, as shown in the figure below:



Hypothesis testing

- Statisticians usually test hypotheses at the common a levels of 0.05 or 0.01 and sometimes at 0.10. The choice of the level depends on the seriousness of the type I error. Besides listing an a value, many computer statistical packages give a *P*-value for hypothesis tests.
- The ***P*-value** (or probability value) is the probability of getting a sample statistic (such as the mean) or a more extreme sample statistic in the direction of the alternative hypothesis when the null hypothesis is true. A small p-value suggests that a surprising outcome has occurred and discredits the null hypothesis.
- In other words, the *P*-value is the actual area under the standard normal distribution curve (or other curve, depending on what statistical test is being used) representing the probability of a particular sample statistic or a more extreme sample statistic occurring if the null hypothesis is true.
- For example, suppose that an alternative hypothesis is H_1: $\mu > 50$ and the mean of a sample is 52. If the computer printed a *P*-value of 0.0356 for a statistical test, then the probability of getting a sample mean of 52 or greater is 0.0356 if the true population mean is 50 (for the given sample size and standard deviation). The relationship between the *P*-value and the a value can be explained in this manner. For $P = 0.0356$, the null hypothesis would be rejected at $\alpha = 0.05$ but not at $\alpha = 0.01$.

Hypothesis testing



- In summary, if the *P*-value is less than α, reject the null hypothesis. If the *P*-value is greater than α, do not reject the null hypothesis.