Section 1.1

Statistics - The science of planning studies and experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data.

Data - Collections of observations (such as measurements, genders, survey responses, etc.)

Population - A complete collection of all individuals to be studied.

Census - The collection of data from every member of the population.

Sample - A subcollection of members selected from the population.

- Sample data must be collected in an appropriate way, such as through a process of random selection.

- If sample data are not collected in an appropriate way, the data may be completely useless!

Section 1.2

When analyzing data, we must consider the following factors:

- Context of the data --- *The context affects the kind of statistical analysis that should be used.*

- Source of the data --- *Be skeptical of studies from sources that may be biased.*

- Sampling method --- *The sampling method can introduce bias and influence the validity of the conclusions.*

- Conclusions --- *Claims must be justified by the statistical analysis.*

- Practical implications --- *Statistical significance can differ from practical significance.*

Common sense and practical considerations are very important when thinking statistically.

Read Pages 6 - 9.

Common problems when analyzing data...

- Misuse of graphs - bad scale, inconsistent scale, wrong type of graph, etc.

- Bad samples - self-selected (voluntary response) samples are often biased. People with strong opinions tend to participate.

- Correlation does not imply causality

- Reported results - reported observations may be biased; collect observations yourself

- Small samples

- Percentages - Will you give 110% to this course?

- Loaded questions

- Order of questions

- Nonresponse

- Missing data - data can be omitted intentionally or unintentionally; data can missed because of special factors

- Self-interest study - be wary of studies where the sponsor has something to gain

- Precision vs. accuracy

- Deliberate distortions

Section 1.3

Parameter - A numerical measurement describing some characteristic of a population.

Statistic - A numerical measurement describing some characteristic of a sample.

Quantitative (numerical) data - Numbers representing counts or measurements.

Categorical (qualitative) data - Names or labels that are not numbers representing counts or measurements.

Discrete data - Data that take on a finite or "countable" number of values.

Continuous data - Data that are numerical, but not discrete. Continuous data may take on *any* numerical values.

To better understand the differences between discrete and continuous data, see the paragraph following example 3 on page 18.

Levels of measurement

Nominal level - characterized by data that consist of names, labels, or categories

Ordinal level - characterized by data that can be ordered, but differences between values cannot be determined or are meaningless

Interval level - similar to ordinal level, except differences between data values are meaningful. An additional feature of this level is that there is no natural zero value.

Ratio level - interval level with a natural zero value, which represents none of a quantity

Section 1.4

Observational study - A study in which we observe and measure specific characteristics, but we do not attempt to modify the subjects being studied

Experiment - A study in which we apply some "treatment" and then proceed to observe its effects on the subjects of the experiment

A **simple random sample** of *n* subjects is selected in such a way that *every possible sample of size n* has the same chance of being chosen.

A **random sample** is selected in such a way that each *individual member* in the population has an equal chance of being selected.

A **probability sample** involves selecting members from a population in such a way that each member of the population has a known (but not necessarily the same) chance of being selected.

Every simple random sample is a random sample, but not vice versa.

Systematic sampling - select every *k*th element in the population

Convenience sampling - use results that are easy to get

Stratified sampling - divide the population into groups (strata) so that subjects in the same group share certain characteristics, then draw a sample from each group

Cluster sampling - divide the population into groups (clusters), then randomly select some clusters

These sampling methods do not result in simple random samples. We will often require that sample data be a simple random sample!

Types of studies...

Cross-sectional study - A study in which data are observed and collected at one point in time

Retrospective study - A study in which data are collected from the past by looking back in time through records, interviews, etc.

Prospective (longitudinal) study - A study in which data are collected in the future from groups called cohorts

Characteristics of good experimental design...

- **Randomization** - assign subjects to groups through a process of random selection

- **Replication** - repeat the experiment on more than one subject so that typical erratic behavior does not hide the effects of treatment

- **Blinding** - do not allow the subject to know whether he or she is receiving the treatment or placebo; this allows for us to determine whether the treatment is significantly different from the placebo effect

- **Controlling Effects of Variables** - plan the experiment so that you are able to distinguish among the effects of other factors (i.e. eliminate confounding)

Types of experimental design...

Completely Randomized - assign subjects to treatment groups by random selection

Randomized Block - form blocks of subjects with similar features, and within blocks, randomly assign subjects to treatment groups

Rigorously Controlled - *carefully* assign to different treatment groups subjects that are similar in ways important to the experiment

Matched Pairs - compare exactly two treatment groups by using subjects matched in "related" pairs

Experimental error...

Sampling Error - difference between a sample result and the actual population result. These errors result from chance fluctuations in the samples.

Nonsampling Error - errors arising when data is incorrectly collected, measured, recorded, or analyzed

Section 2.1

General characteristics of data...

1. **Center** - a representative value that indicates the "middle" of the data set

2. **Variation** - a measure of the amount by which the data values vary

3. **Distribution** - the nature or "shape" of the spread of data

4. **Outliers** - sample values that lie very far away from the vast majority of other values

5. **Time** - changing characteristics of data over time

Section 2.2

A **frequency distribution** is a table that shows how a data set is partitioned among all of several classes by listing all of the classes along with the number of data values in each class.

Example

{5.1,4.8,5.2,6.2,4.9,5.3,5.6,5.7,6.0,4.2,5.7,5.7,5.8,5.0,5.1}

Height (ft)	Frequency	Cumulative freq.
4.0-4.4	1	1
4.5-4.9	2	3
5.0-5.4	5	8
5.5-5.9	5	13
6.0-6.4	2	15

Lower class limits - smallest numbers that belong to the different classes

Upper class limits - largest numbers that belong to the different classes

Class boundaries - numbers used to separate classes; these numbers are the midpoints of the intervals from consecutive upper to lower class limits; class boundaries should not be data values

Class midpoints - values in the middle of the classes

Class width - difference between consecutive lower class limits or upper class limits

Constructing a Frequency Distribution...

Choose the number of classes, class limits, and class width in some convenient or intuitive way, or follow this procedure

1. Determine the number of classes---a convenient number between 5 and 20

2. Class width should be about (max - min)/(no. of classes)

3. Choose min data value, or convenient smaller value, for first lower class limit

4. Use the class width to find all lower class limits

5. Determine the upper class limits

Now create your list, but first use tally marks!

Example with some sunspot numbers

```
162.2 166.2 169.4 170.3 168.3 168.1 169.2 168.0 165.6 163.7 161.9 161.2
157.8 153.0 153.7 156.6 157.7 158.9 160.8 162.8 163.3 163.2 164.8 163.9
162.7 163.9 162.5 159.6 157.9 155.2 151.6 147.6 143.4 139.3 134.1 131.0
129.9 127.6 123.4 119.7 117.4 114.9 112.0 110.4 109.9 108.7 107.4 105.6
```

- **Relative frequency distributions** use percentages instead of frequency counts. The sum of the relative frequencies should be 100% (or very close to it).

- **Cumulative frequency distributions** accumulate the sums of the frequency counts at each class. The final sum must equal the number of data values.

A particular type of distribution that arises frequently in applications is the **normal distribution**.

We will study normal distributions in great detail, but for now...

A normal distribution is characterized by

- The frequencies start low, then increase to one or two high frequencies, then decrease to low frequencies
- The distribution is approximately symmetric

Section 2.3

A histogram is a graphical version of a frequency distribution.

More specifically, a **histogram** is a graph consisting of bars of equal width drawn next to each other without gaps. The horizontal axis represents the classes of quantitative data values and the vertical axis represents frequencies (or relative frequencies).

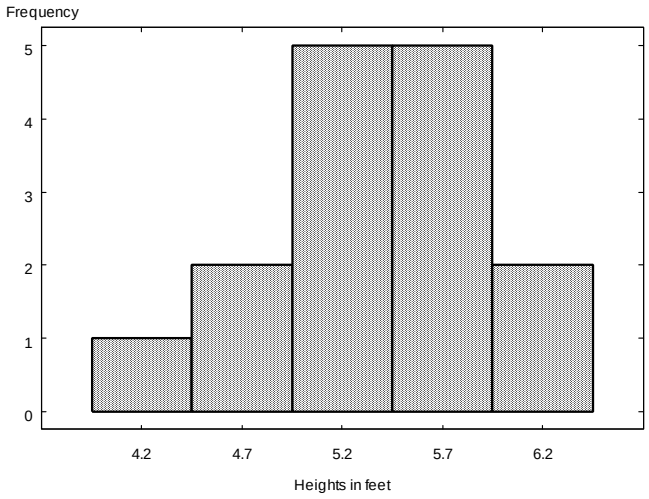
- For the horizontal scale of a histogram, use class boundaries or class midpoints.

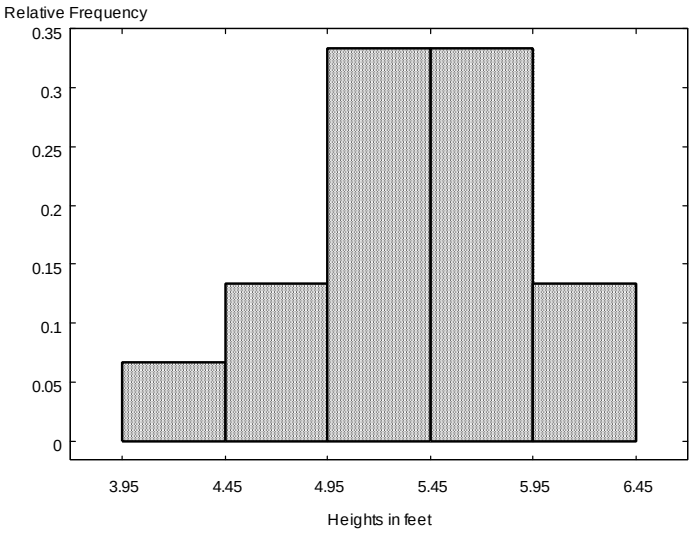
- For the vertical scale, use class frequencies.

Histograms of normally distributed data have a "bell shape," or may be traced by a "bell-shaped curve."

Example with heights from above...

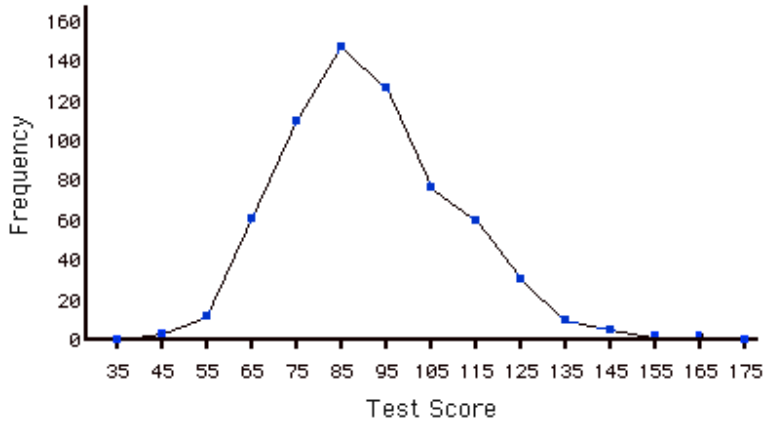
{5.1,4.8,5.2,6.2,4.9,5.3,5.6,5.7,6.0,4.2,5.7,5.7,5.8,5.0,5.1}

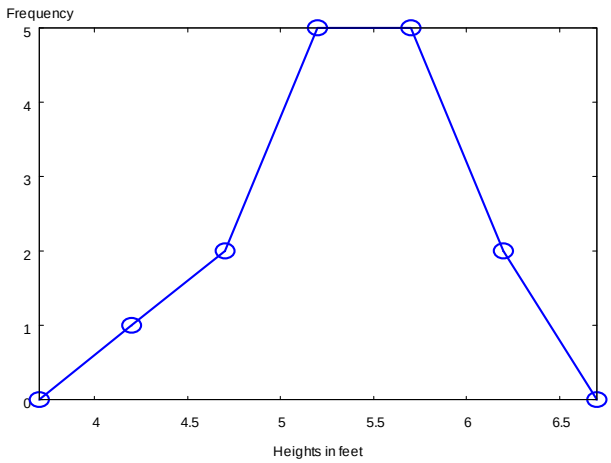




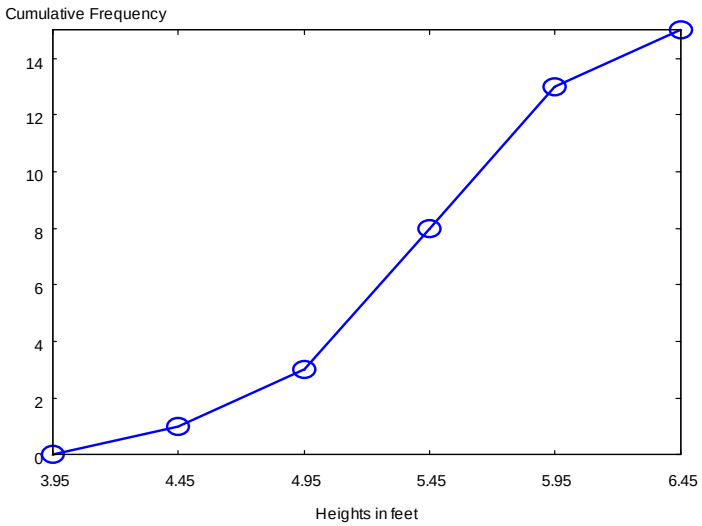
Section 2.4

A **frequency polygon** uses line segments connected to points directly above class midpoints. The height of each point corresponds to frequency or relative frequency. Additional classes with zero frequencies are typically added to each end.

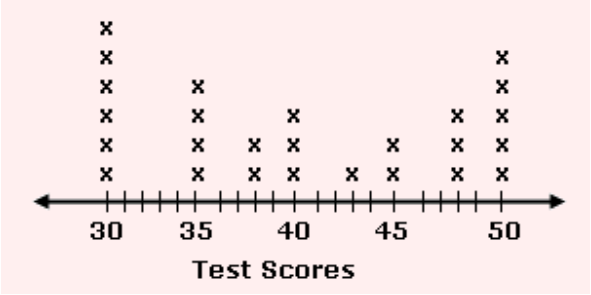




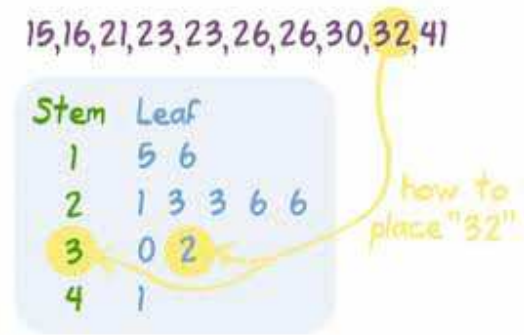
An **ogive** is essentially a cumulative frequency polygon. An ogive uses class boundaries along the horizontal axis and cumulative frequencies along the vertical axis.



A **dotplot** (or line plot) is a graph where each data value is plotted as a dot along a number line. Dots representing equal values are stacked.



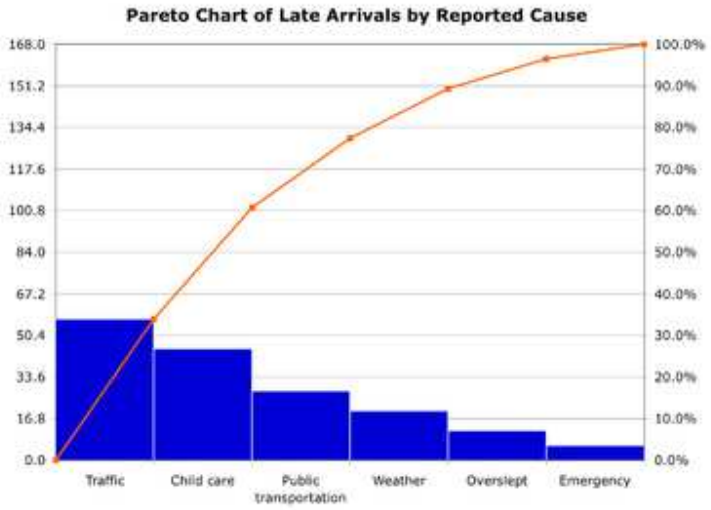
A **stem-and-leaf plot** (or stemplot) represents quantitative data by separating each value into two parts: the stem and the leaf.



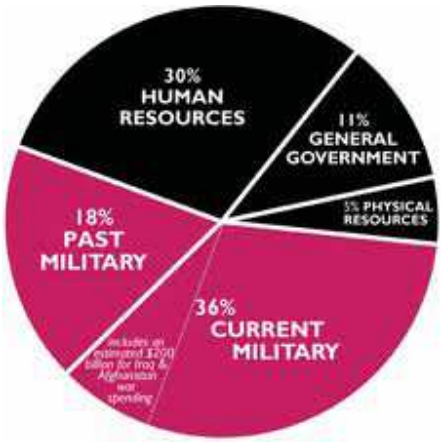
1	9
2	2 5 6 7 8 9
3	0 4 6 7
4	2 3 4 6 8 8 9
5	2 3 5 7 8
6	2

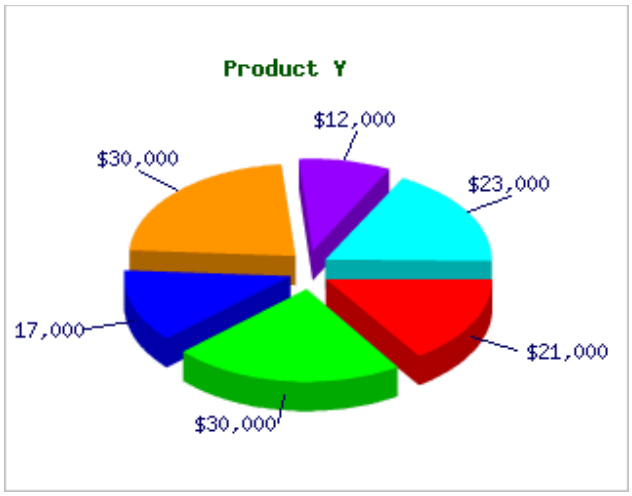
A **bar graph** is ....

A **Pareto chart** is a bar graph for qualitative data, with the added requirement that the bars are arranged in descending order according to frequencies.

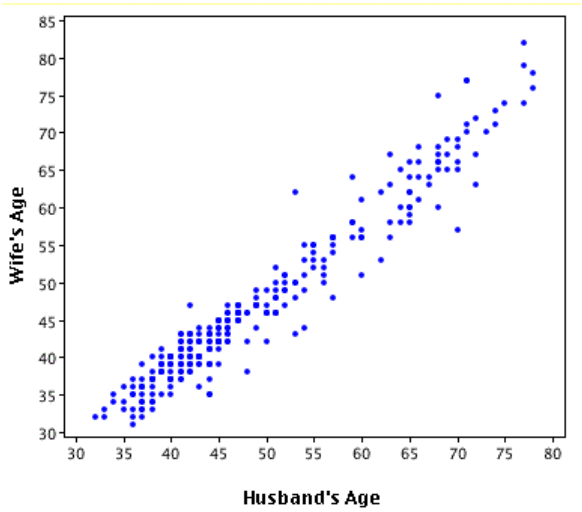


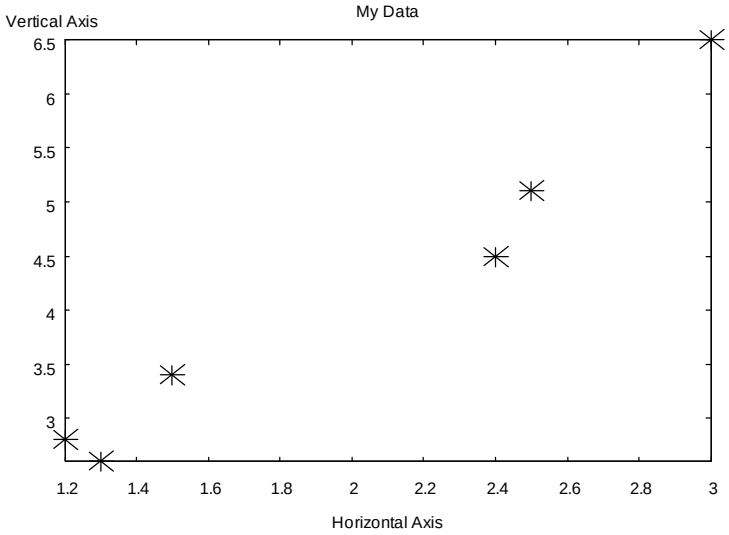
A **pie chart** (or circle graph) is a graph that depicts qualitative data as slices of a circle, where the size of each slice is proportional to the frequency count for the category.



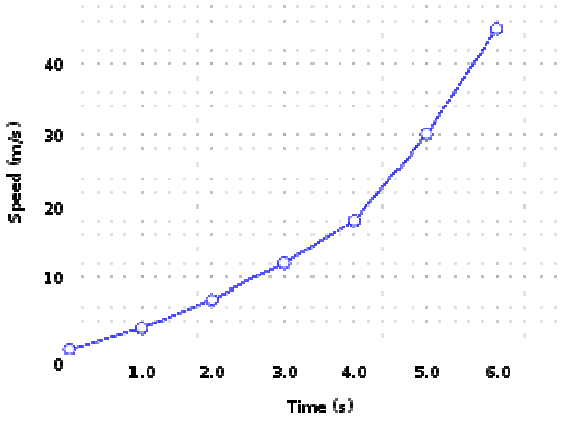


A **scatterplot** is a plot of ordered pairs of quantitative data with a horizontal axis and a vertical axis. The horizontal axis is associated with the first coordinate in the ordered pair, and the vertical axis is associated with the second coordinate.





A **time-series graph** is an example of a line graph. These graphs display quantitative data that have been collected at different points in time. Ordered pairs, in which the 1st coordinates represent time, are plotted and often connected by line segments.



Section 3.2

The **arithmetic mean** (or simply mean) of a data set is the measure of center found by adding the data values and then dividing the sum by the total number of values.

- The mean is a fairly reliable measure of center---when samples are selected from the same population, sample means tend to be more consistent than other measures of center.
- The mean takes every data value into account.
- The mean is particularly sensitive to extreme data values---it *is not a resistant* measure of center.

The **median** is the numerically middle data value if there are an odd number of values or the mean of the two middle values if there are an even number.

- The data must be arranged numerically to determine the median.
- The median *is a resistant* measure of center---it is not dramatically affected by extreme values.

The **mode** is the data value that occurs with the greatest frequency (assuming there is one).

- If two data values occur with same greatest frequency, there are two modes and the data set is **bimodal**.
- If there are more than two modes, the data set is **multimodal**.
- If there are no repeated data values, there is **no mode**.

The **midrange** is the measure of center found by computing the value midway between the minimum data value and the maximum: add the minimum and maximum and then divide by two.

- The midrange is rarely used because is a very sensitive to extreme values.

Warning: The word *average* is often used for the mean, but it can be used to refer to any of the measures of center.

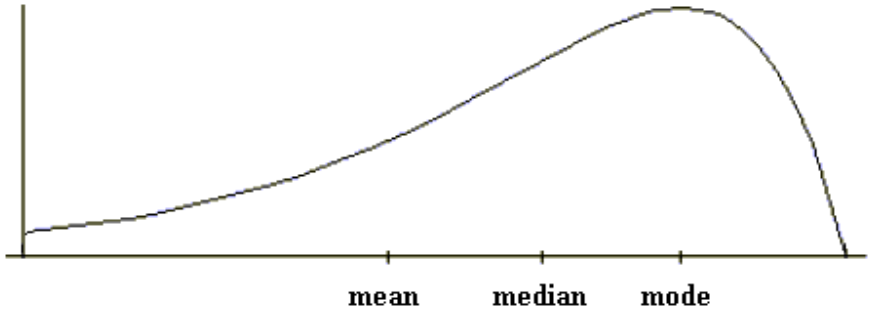
If certain data values "carry more weight" than others, it may be appropriate to compute a **weighted mean**. A weighted mean is computed by multiplying each data value by its associated "weight", adding, and then dividing by the total "weight".

Some examples of weighted means are grade point averages, expected values in probability, and means obtained from frequency distributions.

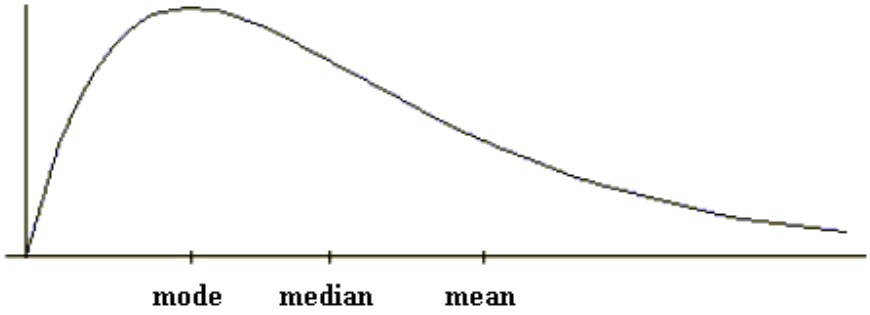
A comparison of the mean, median, and mode can provide information about how a distribution is **skewed**.

While the mean and median do not always characterize the shape of the distribution, the following are generally true...

Skewed left (negatively skewed)---mode greater than mean and median



Skewed right (positively skewed)---mode less than mean and median



Section 3.3

While we have studied several useful measures of center, these measures are not sufficient to describe the distribution of a data set. Another important quantity is the variation.

The **range** of a set of data values is the difference between the maximum and minimum values:

Range = Max value - Min value

The most commonly used measure of variation is the standard deviation, which measures the extent of the deviations from the mean.

The **sample standard deviation**, denoted by *s*, is obtained from either one of the following formulas:

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

or

$$s = \sqrt{\frac{n\ \sum(x^2) - (\sum x)^2}{n\ (n - 1)}}$$

- The standard deviation is a measure of variation of all values from the mean.
- The standard deviation is never negative, and it is zero only when all data values are all the same.
- Small *s* indicates little variation. Large *s* indicates great variation.
- The standard deviation is very sensitive to extreme values.

When comparing data sets, you should compare sample standard deviations only when sample means are approximately the same.

The **population standard deviation**, denoted by σ, should be used when measuring variation in the population and is obtained from the following formula:

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$$

where μ is the population mean.

Warning: Calculators and software normally are capable of computing both standard deviations. Be sure you know which one you're computing!

The **sample or population variance** is the square of the corresponding standard deviation.

Interpretations and rules of thumb...

- For many data sets, the vast majority of sample values lie within two standard deviations of the mean.

- Minimum "usual" value ≈ mean - 2 × (standard deviation)

- Maximum "usual" value ≈ mean + 2 × (standard deviation)

- $s \approx \frac{\text{Range}}{4}$

For data sets that are approximately normally distributed...

- About 68% of all values lie within 1 standard deviation of the mean, i.e. between $\bar{x} - s$ and $\bar{x} + s$.

- About 95% of all values lie within 2 standard deviations of the mean, i.e. between $\bar{x} - 2s$ and $\bar{x} + 2s$.

- About 99.7% of all values lie within 3 standard deviations of the mean, i.e. between $\bar{x} - 3s$ and $\bar{x} + 3s$.

Chebyshev's Theorem

For **any** data set, the fraction of values lying within *K* standard deviations of the mean is always at least $1-1/K^2$.

The sample standard deviation is an unbiased estimator of the population standard deviation---sample standard deviations obtained from different samples within a single population tend to target the population standard deviation.

Using the mean absolute deviation (page 102), or dividing by the sample size instead of *(n-1),* results in a biased estimator.

Another way to think about the denominator of the formula for *s* is that (*n-1*) is the number of degrees of freedom.

When comparing variation in significantly different data sets, such as those involving different units of measurement or with very different means, the **coefficient of variation** should be used. The **CV** is a standard deviation relative to the mean.

$$CV = \frac{s}{\bar{x}} * 100\% \text{ (Sample CV)}$$

$$CV = \frac{\sigma}{\mu} * 100\% \text{ (Population CV)}$$

Section 3.4

Measures of relative standing show location relative to other values in the data set.

The ***z* score** associated with a given data value is the number of standard deviations that value is from the mean.

$$z = \frac{x - \bar{x}}{s} \quad \text{(sample)}$$

$$z = \frac{x - \mu}{\sigma} \quad \text{(population)}$$

Notice that the *z* score is positive if the data value is greater than the mean and negative if the data value is less than the mean.

For reasons we will see later, most of the time we will round our *z* scores to the nearest hundredth.

Based on Chebyshev's Theorem and our observations about normal distributions, we have the following rules of thumb:

- Most ordinary data values lie have *z* scores between -2 and 2.

- Unusually small data values have *z* scores less than -2.

- Unusually big data values have *z* scores greater than 2.

Percentiles are measures of location that divide a data set into 100 groups with about 1% of the values in each group.

percentile of x = 100 * (# of values < x) / (total # of values)

(Round to the nearest whole number.)

To find the value associated with a given percentile:

- Arrange the data in ascending order.
- Solve the formula above for the # of values < x.
- Call your solution L.
- If L is a whole number, the value is the mean of the L-th data value and the (L+1)-th data value.
- If L is not a whole number, round it up to the nearest whole number. The value is the L-th data value.

Quartiles are measures of location that divide a set of data into four groups with about 25% of the values in each group.

The quartiles can be computed as the 25th, 50th, and 75th percentiles.

It is often easier to compute the quartiles as medians.

- The second quartile is the median.
- The first quartile is the median of the lower half (ignoring the actual data value at the location of the median).
- The third quartile is the median of the upper half (ignoring the actual data value at the location of the median).

There is not complete agreement among statisticians and authors on how to define and compute the quartiles.

The **interquartile range** (IQR) is the difference of the third and first quartiles.

Remember that about 50% of the data values lie between the first and third quartiles.

For a set of data, the **5-number summary** consists of the minimum value, the 1st quartile, the median, the 3rd quartile, and the maximum value.

A **boxplot** (box-and-whisker plot) is a type of graph showing the 5-number summary above a number line.

Constructing a boxplot...

Outliers are data values that lie more than 1.5 times the IQR above the 3rd quartile or below the 1st quartile. Compute the limits on the outliers as follows:

Any outliers are

less than Q_1 - 1.5 * IQR

or

greater than Q_3 + 1.5 * IQR.

In a **modified boxplot**, outliers are indicated with asterisks and whiskers are drawn only to the extreme values that are not outliers.

Section 4.2

A **probability experiment** (or random experiment) is a process involving chance in which observations are made and recorded.

In a probability experiment, the individual "things" that are observed are called the **outcomes**.

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

There is not necessarily only one sample space for a probability experiment, but there is usually only one best sample space.

An **event** is any subset of the sample space. In other words, an event is any collection of outcomes of the probability experiment.

An individual outcome is sometimes called a **simple event**.

In a probability experiment, the probability of an event is a number, between 0 and 1, that measures the likelihood of the event.

The probability of an event *A* is written *P(A)*. For any event A,

$$0 \leq P(A) \leq 1.$$

If the event *A* cannot occur, then *P(A)=0*. Sometimes these are called impossible events.

If the event *A* is certain to occur, then *P(A)=1*. These are called certain events.

There are several approaches to assigning the probability of an event...

1. An empirical (experimental) probability is assigned by counting observations:

$$P(A) = \frac{number\ of\ times\ A\ occurred}{number\ of\ trials}$$

2. A theoretical (classical) probability is assigned by assuming equally likely outcomes and counting numbers of outcomes:

$$P(A) = \frac{number\ of\ outcomes\ in\ A}{number\ of\ outcomes\ in\ the\ sample\ space}$$

The **Law of Large Numbers** says that an empirical probability will get closer and closer to the corresponding theoretical probability as the number of trials increases.

If theoretical probabilities are estimated by empirical probabilities, a large number of trials should be used.

3. A geometric probability is a classical probability assigned by measuring and comparing length, area, volume, etc.

4. A subjective probability is a probability estimate based on knowledge of relevant circumstances.

Even if all outcomes are not equally likely, a theoretical probability can often be founding by counting outcomes. We simply give the appropriate "weights" to each outcome, much in the spirit of a weighted mean.

Important fact: The sum of the probabilities of ALL **outcomes** must always equal 1. In other words,

The probability of the sample space is 1.

The **complement** of the event *A*, denoted by $\bar{A}$, is the set of all outcomes in the sample space that are NOT in *A*.

$$P(A) + P(\bar{A}) = 1$$

$$P(\bar{A}) = 1 - P(A)$$

$$P(A) = 1 - P(\bar{A})$$

The **odds in favor** of the event *A* are given by $\frac{P(A)}{P(\bar{A})}$.

The **odds against** *A* are given by $\frac{P(\bar{A})}{P(A)}$.

These odds can also be computed by counting outcomes.

Section 4.3

The **union** of the two events *A* and *B*, written $A \cup B$, is the event that *A* occurs or *B* occurs or both occur. Unions are often expressed in words by saying "*A* or *B*."

The **intersection** of the two events *A* and *B*, written $A \cap B$, is the event that both *A* and *B* occur. Intersections are often expressed in words by saying "*A* and *B*."

The sets *A* and *B* are said to be **disjoint** if $A \cap B$ is an impossible event, i.e. the intersection is the empty set.

The following very important probability rule relates unions and intersections of events: For any events *A* and *B*,

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

If A and B are disjoint, then $P(A \cap B) = 0$, and the rule becomes much simpler: For disjoint events *A* and *B*,

$$P(A \cup B) = P(A) + P(B).$$

Section 4.4

Up to now, we have been thinking of events as being associated with single, 1-stage, experiments. It is often convenient to think of certain complicated experiments as a sequences of single-stage experiments.

For example...

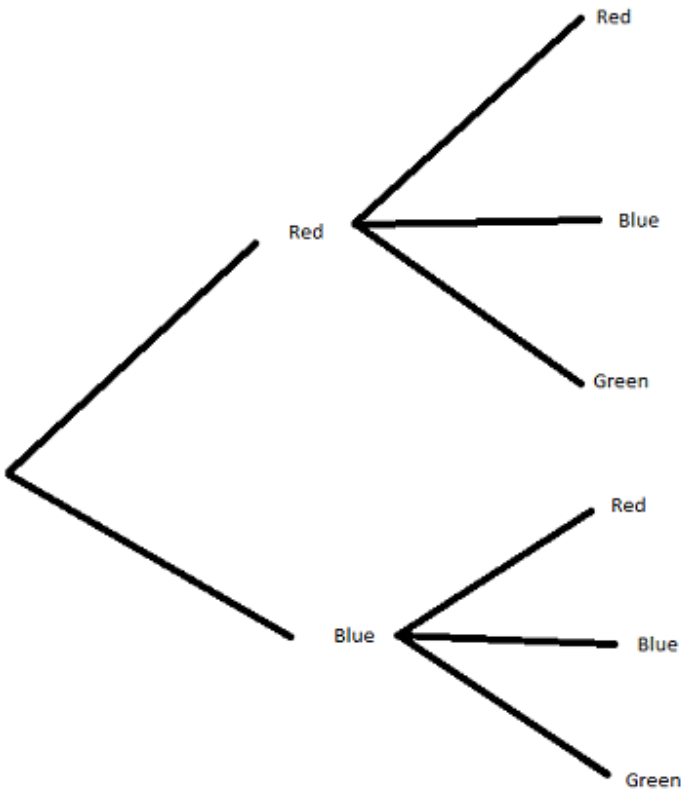
Jar 1 contains 4 red marbles and 7 blue marbles. Jar 2 contains 1 red marble, 3 blue marbles, and 4 green marbles. A single marble is selected at random from each jar.

This complicated probability experiment is best thought about in two stages:

Stage 1 - Select a marble from Jar 1

Stage 2 - Select a marble from Jar 2

Here is the tree diagram for the two-stage experiment. Determine the probability of each branch.



Multiplication Rule: For a tree diagram associated with a multi-stage experiment, the probability of a given path is the product of the probabilities along the branches.

Suppose a marble is selected from Jar 1 and placed into Jar 2. Then a marble is selected from Jar 2. How does the probability tree diagram change?

In some situations, probabilities of events change when more information is obtained. For example...

What is the probability that you will carry an umbrella on any given day?

What is the probability that you will carry an umbrella on any given day, if you know it is raining that day?

The second probability is called a **conditional probability**.

Suppose *A* and *B* are events, *P(A|B)* represents the probability of event *A* occurring after it is assumed that event *B* has already occurred. *P(A|B)* is "the probability of *A* given *B*."

For any events A and B,

$$P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A).$$

Two events *A* and *B* are **independent** if the occurrence of one does not affect the probability of the other.

For example...

Select two letters at random from the word MISSISSIPPI. Let *A* be the event of selecting the letter S and let *B* be the event of selecting the letter P.

If the selections are made with replacement, *A* and *B* are independent.

If the selections are made without replacement, *A* and *B* are dependent.

If the size of a sample is no more than 5% of the population size, we will sometimes approximate probabilities by treating events as being independent, even if selections are made without replacement.

Section 4.5

Recall...

For any events A and B,

$$P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A).$$

This formula can be rewritten to say

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

Important ideas to keep in mind:

- In general, $P(A|B) \neq P(B|A)$.

- *A* and *B* are independent if and only if $P(A|B) = P(A)$ **and** $P(B|A) = P(B)$.

Section 4.6

The Fundamental Principle of Counting

If there are n ways to do a first task and m ways to do a second task, then there are $n \times m$ ways to do the combination of tasks.

This generalizes to more tasks, and it follows that...

A collection of n different items can be arranged in $n!$ different ways (counting different orders as different ways).

Permutations (when items are different)

If the following requirements are met:

1. There are n different items available.
2. We select r of the n items (without replacement).
3. We consider rearrangements of the same items to be different (i.e. order matters).

The number of permutations of r items selected from n different items is

$$_nP_r = \frac{n!}{(n-r)!}$$

Permutations (when some items are identical)

If the following requirements are met:

1. There are n items available, and n_1 of them are alike, n_2 of them are alike,..., n_k of them are alike.
2. We select all of the n items (without replacement).
3. We consider rearrangements of distinct items to be different (i.e. order of different items matters).

The number of permutations of all n items selected without replacement is

$$\frac{n!}{n_1!\,n_2!\,\ldots\,n_k!}$$

Combinations

If the following requirements are met:

1. There are n different items available.
2. We select r of the n items (without replacement).
3. We consider rearrangements of the same items to be the same (i.e. order does not matter).

The number of combinations of r items selected from n different items is

$$ {}_nC_r = \frac{n!}{(n-r)!\,r!} $$

Section 5.2

A **random variable** is a variable (often represented by x) that has a single numerical value, determined by chance, for each outcome of a procedure.

A **probability distribution** is a description that gives the probability for each value of a random variable. A probability distribution is often expressed in graph, table, or formula.

Examples...

1. Roll a fair six-sided die. Let $x =$ the number rolled.
2. Randomly select a single quiz from those returned. Let $x =$ score on the quiz.
3. Randomly select a full-term, newborn baby from St. James Hospital. Let $x =$ the weight (in lbs) of the baby.

A **discrete random variable** is a variable from a discrete data set.

A **continuous random variable** is a random variable from a continuous data set.

Requirements for a Probability Distribution

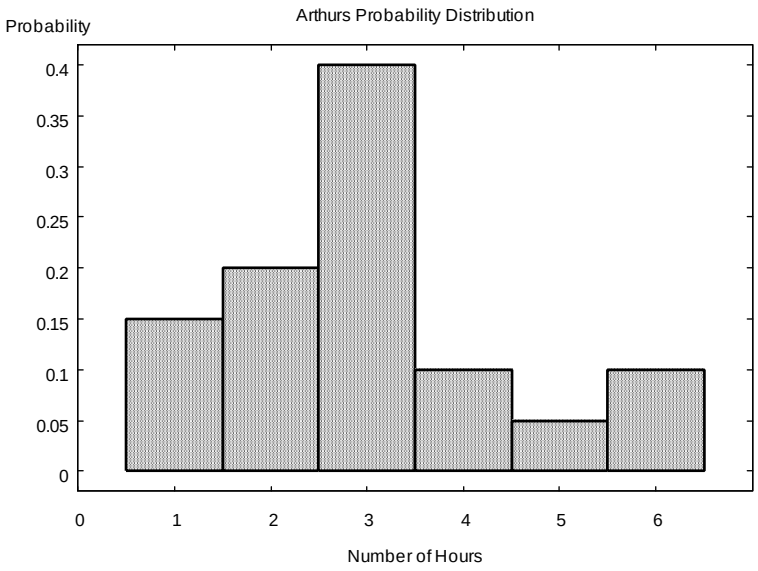
1. $\sum P(x) = 1,$ or at least very close to 1 if you are rounding.
2. $0 \leq P(x) \leq 1$

Example: The following table shows the probabilities assigned by Arthur to the number of hours spent on homework on any given night. Show that the table defines a probability distribution.

Hours	Probability
1	0.15
2	0.20
3	0.40
4	0.10
5	0.05
6	0.10

There are a number of ways to graph a probability distribution. We will normally use a **probability histogram**.

Here is the probability histogram for Arthur's probability distribution.



Mean, Variance, and Standard Deviation

(for Discrete Random Variables)

- Mean for a probability distribution (Expected Value)

$$\mu = \sum [x \times P(x)]$$

- Variance for a probability distribution

$$\sigma^2 = \sum [(x - \mu)^2 \times P(x)]$$

or

$$\sigma^2 = \sum [x^2 \times P(x)] - \mu^2$$

- Standard deviation for a probability distribution

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

Recall that we used the standard deviation to identify the minimum and maximum "usual" values:

minimum "usual" value = $\mu - 2\sigma$

maximum "usual" value = $\mu + 2\sigma$

Rare Event Rule for Inferential Statistics

If, under a given assumption, the probability of a particular *observed* event is extremely small, we conclude that the assumption is probably not correct.

Using probabilities to determine when results are unusual...

- x is an unusually high number if the probability of x or more is 5% or less
- x is an unusually low number if the probability of x or fewer is 5% or less

Section 5.3

A **binomial probability distribution** results from a procedure that meets all of the following:

1. The procedure has a fixed number of trials.
2. The trials must be independent.
3. Each trial must have outcomes that can be classified in exactly two categories: *success* or *failure*.
4. The probability of a success remains the same for all trials.

If a procedure satisfies the four requirements, the distribution of the random variable x, where x is the number of successes, is called a **binomial distribution**.

In some cases, the trials are not technically independent, but may be treated as such. Recall our earlier guideline:

If the size of a sample is no more than 5% of the population size, we will sometimes approximate probabilities by treating events as being independent, even if selections are made without replacement.

Notation for Binomial Distributions

- Probability of success = p (Single trial)
- Probability of failure = $q = 1 - p$ (Single trial)
- Number of trials = n
- Number of successes in n trials = x
- Probability of x successes in n trials = $P(x)$

By using the appropriate counting techniques, it can be shown that

$$P(x) = {}_nC_x \cdot p^x \cdot q^{n-x}$$

Section 5.4

Since the formulas from above for mean, variance, and standard deviation apply for any discrete probability distributions, they certainly apply for binomial distributions.

- Mean for a probability distribution (Expected Value)

$$\mu = \sum [x \times P(x)]$$

- Variance for a probability distribution

$$\sigma^2 = \sum [(x - \mu)^2 \times P(x)]$$

or

$$\sigma^2 = \sum [x^2 \times P(x)] - \mu^2$$

- Standard deviation for a probability distribution

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

However, for binomial distributions, these formulas can be greatly simplified:

- $\mu = np$
- $\sigma^2 = npq$
- $\sigma = \sqrt{npq}$

Recall, once again, the rules of thumb for identifying the minimum and maximum "usual" values:

$$\text{minimum "usual" value} = \mu - 2\sigma$$
$$\text{maximum "usual" value} = \mu + 2\sigma$$

Section 5.5

The Poisson distribution is another common discrete probability distribution. This distribution is often used for describing behavior such as radioactive decay and patients arriving at an ER.

The **Poisson distribution** is a discrete probability distribution that applies to occurrences of some event over a specified interval (of time, distance, etc). The random variable x is the number of occurrences of the event in the interval. The probability of the random variable x is given by

$$P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!},$$

where μ is the mean number of occurrences in the interval.

Requirements for the Poisson Distribution

1. The random variable x is the number of occurrences of an event over some interval.
2. The occurrences must be random.
3. The occurrences must be independent of each other.
4. The occurrences must by uniformly distributed over the interval.

In a Poisson distribution,

- mean = μ
- $\sigma^2 = \mu$
- $\sigma = \sqrt{\mu}$

Section 6.1

If a continuous random variable has a distribution whose graph is a symmetric, bell-shaped curve described by an equation of the form

$$y = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma \cdot \sqrt{2\pi}},$$

we say that the random variable has a **normal distribution**.

Notice that the normal distribution is determined by two parameters: the mean and the standard deviation.

Sections 6.2 & 6.3

The graph of a continuous probability distribution is called a **density curve**.

Every density curve must satisfy the following requirements:

1. The total area under the curve must equal 1.
2. The y-coordinate of every point on the curve must be greater than or equal to zero. (The curve cannot fall below the x-axis.

Because the total area under a density curve is equal to 1, there is a correspondence between area and probability.

Uniform Distribution

A continuous random variable has a **uniform distribution** if its values are spread evenly over the range of possible values (i.e. all possible values are equally likely). The graph of a uniform distribution is a horizontal line segment.

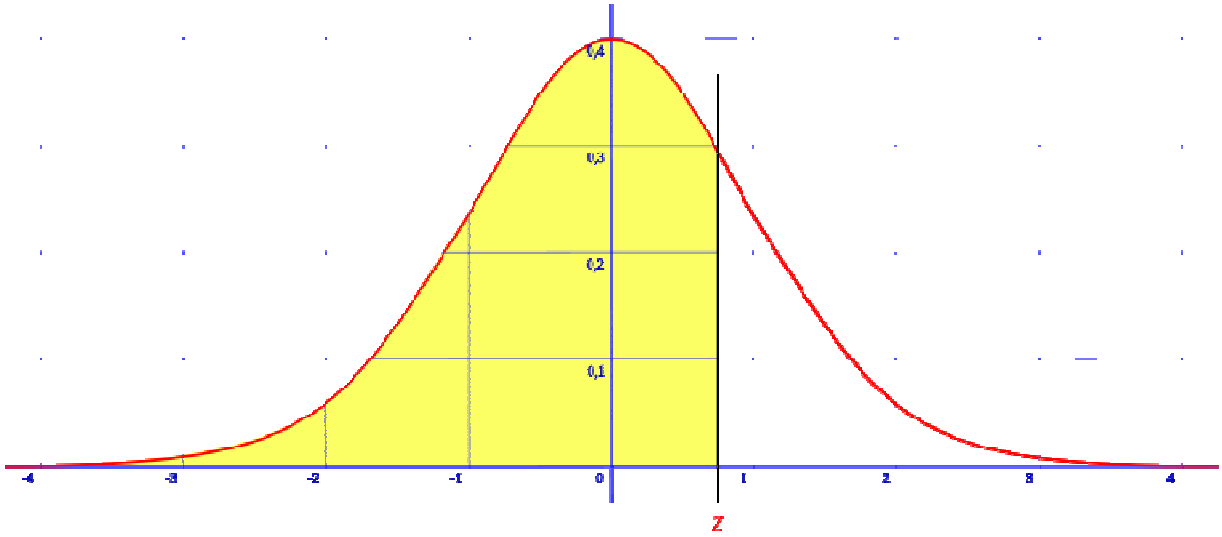
Standard Normal Distribution

The **standard normal distribution** is the normal distribution with mean zero ($\mu = 0$) and standard deviation one ($\sigma = 1$).

If x is a random variable in a normal distribution, then $z = \frac{x-\mu}{\sigma}$ is the corresponding random variable in the standard normal distribution.

Notice that if $\mu = 0$ and $\sigma = 1$, then $z = x$, and your random variable is already *normalized*.

For any given z-score, the probability of a score less than or equal to z is the area under the standard normal curve to the left of z.



Computing the area under the curve is a calculus problem. We'll find area by using our calculators or a table.

For example...

- $P(z < 1) = 0.8413$
- $P(z < 2.09) = 0.9817$
- $P(z < -1.24) = 0.1075$
- $P(-1.23 < z < 1.98) = 0.8668$
- $P(z > 1.3) = 1 - P(z < 1.3) = 0.09680$

For each of these, shade the appropriate region under the standard normal curve.

Examples

1. Temperatures are normally distributed with mean 0 and standard deviation 1. In a sample of 500 temperature measurements, about how many lie between -1.38 and 0.87?
2. Birth weights in Norway are normally distributed with mean 3570 grams and standard deviation 500 grams. What percent of newborns have weight less than 3100 grams? In a year, a certain hospital delivers 600 babies. How many weigh less than 2900 grams?
3. Men's heights are normally distributed with mean 69.0 in and standard deviation 2.8 in. What percent of men are shorter than 65 in or taller than 72 in?

In some problems we'd like to find the z-score associated with a certain probability. This is the inverse of the type of problem we did above.

Example

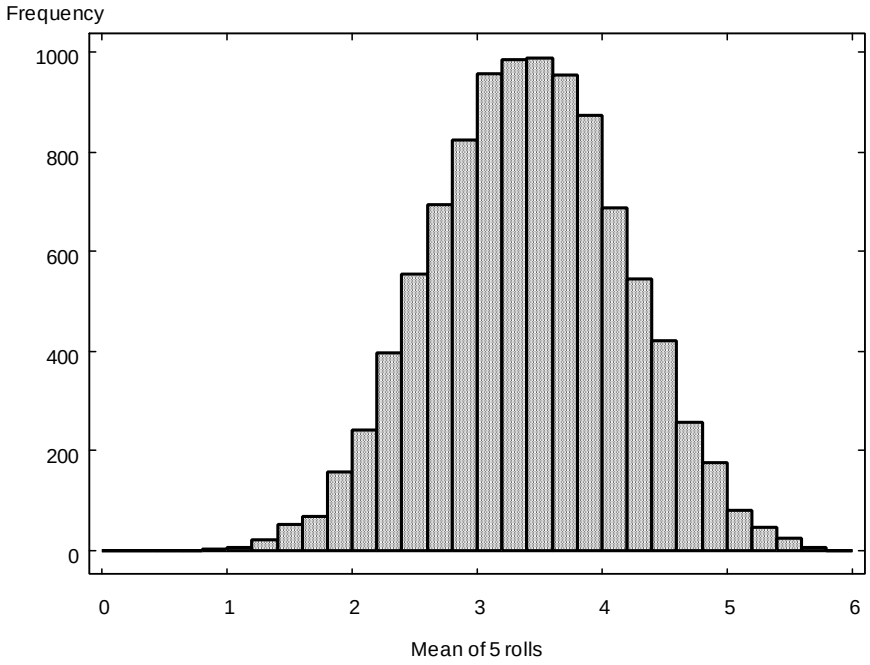
Refer to Example 3 above. What height should a doorway be so that 95% of men can walk through without bending down?

Section 6.4

The **sampling distribution of a statistic** is the distribution of all values of the sample statistic when all samples of size n are taken from the same population.

Here is an example to help us understand the *sampling distribution of the mean*:

1. Roll a die 5 times.
2. Let the random variable x be the mean value of the 5 rolls. Find x.
3. Repeat the 5-roll experiment many times and keep a record of the values of x.
4. The distribution of the variable x is the sampling distribution of the mean.
5. The sampling distribution could be represented by a table, probability histogram, or formula.



Two important properties of the *sampling distribution of the mean*:

1. The sample means *target* the population mean. (The mean of the sample means is the population mean.)
2. The distribution of sample means is approximately normal.

Two important properties of the *sampling distribution of the variance*:

1. The sample variances *target* the population variance.
2. The distribution of sample variances tends to be skewed right.

Two important properties of the *sampling distribution of a sample proportion*:

1. The sample proportions *target* the population proportion.
2. The distribution of sample proportions is approximately normal.

See summary table on Page 281.

Biased and Unbiased Estimators

These statistics are unbiased estimators--they target the value of the corresponding population parameter.

- Mean
- Variance
- Proportion

These statistics are biased estimators--they do not target the value of the corresponding population parameter.

- Median
- Range
- Standard deviation (It's close though!)

Section 6.5

Central Limit Theorem

1. For a population with any distribution, if $n > 30$, then the sample means have a distribution that can be approximated by a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.
2. If $n \leq 30$ and the original population has a normal distribution, then the sample means have a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

There are exceptions to the rule in (1). Some populations may require a larger sample size.

Some things to keep in mind...

- As the sample size increases, the distribution of sample means gets closer and closer to a normal distribution.
- The mean of sample means is the same as the mean of the original population.
- As the sample size increases, the distribution of sample means gets narrower (because the standard deviation of sample means gets smaller).

Notation for the Sampling Distribution of the Mean

If all possible random samples of size n are selected from a population with mean μ and standard deviation σ, then the mean of the sample means is denoted by $\mu_{\bar{x}}$. Note that $\mu_{\bar{x}} = \mu$.

The standard deviation of sample means is denoted by $\sigma_{\bar{x}}$. Note that $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. In this context, $\sigma_{\bar{x}}$ is often called the **standard error of the mean**.

Central Limit Theorem for Proportions

Let p be the probability of success and q be the probability of failure. The sampling distribution of proportions for samples of size n is approximately normal with mean $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$.

Example

The new Endeavor SUV has been recalled because 5% of the cars experience brake failure. A random sample of 200 cars is obtained. What is the probability that the proportion of defective cars in the sample is less than 4%?

Section 7.2

A **confidence interval** is an interval of values used to estimate the true value of a population parameter.

Our first confidence intervals will be used to estimate a population proportion.

For example...

A Pew Research Center poll found the 70% of 1501 randomly selected U.S. adults believe in global warming. A 95% confidence interval estimate of the population proportion p is $0.677 < p < 0.723$.

Correct interpretations of this confidence interval

- We are 95% confident that the interval (0.677,0.723) actually contains the true value of the population proportion p.
- If we were to select many different samples of size 1501 and construct the corresponding confidence intervals, 95% of them would contain the true population proportion p.
- The process of computing the confidence interval will result in intervals that contain the true population proportion 95% of the time.

Incorrect interpretations of this confidence interval

- 95% of sample proportions fall between 0.677 and 0.723.
- There is a 95% chance that the true population proportion will fall between 0.677 and 0.723.

Computing population proportion confidence intervals...

Because sample proportions are normally distributed, we can find a confidence interval by computing the z-scores that bound a certain area (the confidence level) under the standard normal curve.

The **confidence level** is the probability $1 - \alpha$ that the confidence interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times.

The number $z_{\alpha/2}$ is the z-score with the property that it separates an area of $\alpha/2$ in the right tail of the standard normal distribution.

Confidence Interval for Population Proportion

It turns out that the confidence interval at the level $1 - \alpha$ is given by

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}},$$

where p is the population proportion, $\hat{p}$ is the sample proportion, $\hat{q} = 1 - \hat{p}$, and n is the number of sample values.

The number $E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$ is called the margin of error.

Determining Sample Size

Suppose we want to find the sample size required to estimate p with a desired margin of error.

When $\hat{p}$ is known: $n = \frac{[z_{\alpha/2}]^2 \hat{p}\hat{q}}{E^2}$

When $\hat{p}$ is not known: $n = \frac{[z_{\alpha/2}]^2 (0.25)}{E^2}$

Section 7.3 (Part 2)

To construct a confidence interval for the population mean μ:

Assuming...

1. The sample is a simple random sample.
2. The value of the population standard deviation σ is known.
3. Either the population is normally distributed or $n > 30$.

The confidence interval is

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

where $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ is the margin of error.

The correct interpretations of this confidence interval are similar to those given in Section 7.2.

Sample size required to estimate a population mean:

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{E}\right)^2$$

If σ is not known, there are several ways to approximate it.

- Using the range rule of thumb, $\sigma \approx$ range/4.
- Collect sample values and use $\sigma \approx s$.
- Estimate σ using the results of another study or survey.

Section 7.3 (Part 1)

When estimating a population mean when σ is not known, we use the Student's t-Distribution.

If a population has a normal distribution, then the distribution of

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

is a **Student t-distribution** for all samples of size n.

To construct a confidence interval for the population mean μ, when σ is not known:

Assuming...

1. The sample is a simple random sample.
2. Either the population is normally distributed or $n > 30$.

The confidence interval is

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}},$$

where $E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ is the margin of error and $t_{\alpha/2}$ is obtained using $n - 1$ degrees of freedom.

Important Properties of Student's t-Distribution:

1. The distribution is different for different sample sizes.
2. The distribution has the same general symmetric bell shape as the standard normal distribution, but it reflects the greater variability that is expected with small samples.
3. The distribution has a mean of $t = 0$.
4. The standard deviation of the distribution varies with sample size, but it is greater than 1.
5. As the sample size gets larger, the distribution gets closer to the standard normal distribution.

Section 7.4

When estimating a population standard deviation σ or population variance σ^2, we use a χ^2 (chi-square) distribution.

Warning: The chi-square distribution is not a symmetric distribution, so confidence interval estimates for the standard deviation and variance are not centered on the point estimate.

If a population has a normal distribution, then the distribution of

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

is a **chi-square distribution** for all samples of size n.

To construct a confidence interval for the population variance σ^2:

Assuming...

1. The sample is a simple random sample.
2. The population is normally distributed.

The confidence interval is

$$\frac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2},$$

where χ_R^2 and χ_L^2 are the right- and left-tailed critical numbers obtained using $n-1$ degrees of freedom.

If a confidence interval estimate of σ is required, it is obtained by taking square roots of the bounds given above.

Important Properties of Chi-Square Distribution:

1. The distribution is not symmetric. However, as the number of degrees of freedom increases, the distribution becomes more symmetric.
2. The values of chi-square can be zero or positive, but they cannot be negative.
3. As the sample size gets larger, the distribution gets closer to a normal distribution.

Section 8.1

In statistics, a **hypothesis** is a claim or statement about a property of a population.

A **hypothesis test** is a procedure for testing the validity of a claim about a property of a population.

Section 8.2-8.5

Null and Alternative Hypotheses...

The **null hypothesis**, H_0, is a statement that the value of a population parameter is *equal to* some claimed value.

- $H_0: p = 0.5$

The **alternative hypothesis**, H_1, is the statement that the parameter value somehow differs from that claimed by the null hypothesis.

- $H_1: p \neq 0.5$
- $H_1: p > 0.5$
- $H_1: p < 0.5$

Identifying the null and alternative hypotheses

1. Identify the specific claim to be tested and express it in symbolic form.
2. Write the symbolic form that must be true if the original claim is false.
3. Using the two symbolic expressions obtained so far
    - H_1 is the symbolic expression that does not contain the equality.
    - H_0 is the symbolic expression that the parameter equals the fixed value being considered.

See exercises 5-8 on pages 396 & 397.

Computing Test Statistics...

A **test statistic** is a value used in making a decision about the null hypothesis.

- Test statistic for proportion: $z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$
- Test statistic for mean (normal): $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$
- Test statistic for mean (Student t): $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$
- Test statistic for standard deviation: $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$

Tools for Assessing the Test Statistic...

The **critical region** is the set of values of the test statistic that cause us to reject the null hypothesis.

The **significance level**, α, is the probability that the test statistic will fall in the critical region when the null hypothesis is actually true. If the test statistic falls in the critical region, we reject the null hypothesis. α is the probability of making the mistake of rejecting the null hypothesis when it is true. The most common choice for α is 0.05 (corresponding to a confidence level of 95%).

A **critical value** is any value that separates the critical region from the values of the test statistic that lead to rejection of the null hypothesis. Critical values depend on the nature of the null hypothesis, the sampling distribution that applies, and the significance level.

The ***P*-value** is the probability of getting a value of the test statistic that is at least as extreme as the one representing the sample data.

- Critical region in left tail: P-value = area to the left of the test statistic
- Critical region in right tail: P-value = area to the right of the test statistic
- Critical region in two tails: P-value = twice the area beyond the test statistic

A **one-tailed test** indicates that the null hypothesis should be rejected when the test statistic is in the critical region *on one side* of the mean, depending on the direction of the inequality in the alternative hypothesis.

In a **two-tailed test**, the null hypothesis should be rejected when the test statistic is in the critical region *on either side* of the mean.

When performing a hypothesis test, our conclusion will always be one of the following:

1. Reject the null hypothesis.
2. Fail to reject the null hypothesis.

Errors in Hypothesis Tests...

- **Type I error:** The mistake of rejecting the null hypothesis when it is actually true. The symbol α is typically used to represent the probability of a type I error.
- **Type II error:** The mistake of failing to reject the null hypothesis when it is actually false. The symbol β is typically used to represent the probability of a type II error.

There are several methods for performing a hypothesis test. We will look at three methods.

Traditional Method

1. Identify the null and alternative hypotheses.
2. Select the significance level α based on the seriousness of a type I error. Make α small if the consequences of rejecting a true H_0 are severe. The values of 0.05 and 0.01 are very common.
3. Determine the appropriate test statistic and its sampling distribution.
4. Find the critical values and the critical region. Draw a graph illustrating the important values.
5. Reject H_o if the test statistic is in the critical region.
6. State your conclusion.

P-Value Method

1. Identify the null and alternative hypotheses.
2. Select the significance level α based on the seriousness of a type I error. Make α small if the consequences of rejecting a true H_0 are severe. The values of 0.05 and 0.01 are very common.
3. Determine the appropriate test statistic and its sampling distribution.
4. Find the P-value. Draw a graph illustrating the important values.
5. Reject H_o if the P-value is less than or equal to α.
6. State your conclusion.

Confidence Interval Method

1. Identify the null and alternative hypotheses.
2. Select the significance level α based on the seriousness of a type I error. Make α small if the consequences of rejecting a true H_0 are severe. The values of 0.05 and 0.01 are very common.
3. For a two-tailed test, construct a confidence interval with confidence level $1 - \alpha$. For a one-tailed test, construct a confidence interval with confidence level $1 - 2\alpha$.
4. Reject H_o if the population parameter has a value that is not included in the confidence interval.
5. State your conclusion.

Example

A researcher wishes to test the claim that the average cost of tuition and fees at a four-year public college is greater than $5700. She selects a random sample of 36 four-year public colleges and finds the mean to be $5950. The population standard deviation is $659. Is there evidence to support the claim at the level $\alpha = 0.05$?

Example

A telephone company representative estimates that 40% of its customers have call-waiting service. To test this hypothesis, she selected a sample of 100 customers and found that 37% had call waiting. At the level $\alpha = 0.01$, is there enough evidence to reject the claim?

Example

A medical investigation claims that the average number of infections per week at a certain hospital is 16.3. A random sample of 10 weeks had a mean number of 17.7 infections. The sample standard deviation is 1.8. Is there enough evidence to reject the investigator's claim at the level $\alpha = 0.05$?

Example

A cigarette manufacturer wishes to test the claim that the variance of the nicotine contents of its cigarettes is 0.644. Nicotine content is measured in milligrams, and assume that it is normally distributed. A random sample of 20 cigarettes has a standard deviation of 1.00 mg. Is there enough evidence to reject the manufacturer's claim at the level $\alpha = 0.05$?

Example

A hospital administrator believes that the standard deviation in the number of people using outpatient surgery per day is greater than 8. Assuming the numbers of people are normally distributed, a random sample of 15 days is selected. The data are shown below. Is there enough evidence to reject the administrator's claim at the level $\alpha = 0.10$?

25 30 5 15 18 42 16 9

10 12 12 38 8 14 27

Section 9.2

In this section, we

1. Test a claim about two population proportions where $H_0: p_1 = p_2$
2. Construct a confidence interval estimate for the difference between two population proportions: $p_1 - p_2$

For the assumptions and formulas, see the table on pages 442-443.

Example

Labor statistics indicate the 77% of cashiers and servers are women. A random sample of cashiers and servers in a large metropolitan area found that 112 of 150 cashiers and 150 of 200 servers were women. At the level $\alpha = 0.05$, is there sufficient evidence to conclude that a difference exists between the proportions of servers and cashiers who are women?

Example

In a sample of 200 men, 130 said they used seat belts. In a sample of 300 women, 63 said they used seat belts. At the level $\alpha = 0.01$, test the claim that men are more safety-conscious than women.

Section 9.3 (Part 1)

Two samples are **independent** if the sample values from one population are not related to or naturally associated with the sample values from the other population.

Two samples are **dependent**, consisting of **matched pairs**, if the sample values are paired, matched, or somehow inherently related.

In part 1, we have independent samples with σ_1 and σ_2 unknown and not assumed equal. The goal is to

1. Test a claim about two population means where $H_0: \mu_1 = \mu_2$
2. Construct a confidence interval estimate for the difference between two population means: $\mu_1 - \mu_2$

For the assumptions and formulas, see the table on pages 455-456. In part 1, we use an **unpooled** estimate for σ^2.

Example

The results of a study of words spoken in a day by men and women are given in the following table.

Men	**Women**
$n_1 = 186$	$n_2 = 210$
$\bar{x}_1 = 15668.5$	$\bar{x}_2 = 16215.0$
$s_1 = 8632.5$	$s_2 = 7301.2$

At the level $\alpha = 0.01$, test the claim that men talk less than women.

Section 9.3 (Part 2)

In part 2, we have independent samples with σ_1 and σ_2 unknown and ***assumed equal***.

We use a **pooled** estimate for σ^2. See page 460.

Section 9.4

When two data sets consist of dependent samples that are matched pairs, the techniques of Section 9.3 should not be used.

There are no exact procedures for dealing with dependent samples, but we will use the t-test approximation methods described on pages 468-469.

Procedure:

1. Verify that the requirements described on page 468 are satisfied.
2. Find the difference d for each matched pair.
3. Find the mean $\bar{d}$ and the standard deviation s_d of the differences.
4. Use the t-test techniques from Chapters 7 & 8 to find a confidence interval estimate for $\bar{d}$ or to test a hypothesis concerning $\bar{d}$.

Example

As part of the National Health and Nutrition Examination Survey, the Department of Health and Human Services obtained self-reported heights and measured heights for males aged 12-16. The measurements are in inches.

Reported	68	71	63	70	71	60	65	64	54	63	66	72
Measured	67.9	69.9	64.9	68.3	70.3	60.6	64.5	67	55.6	74.2	65.0	70.8

What do these data suggest about young males' abilities to self-report?

Section 9.5

For two normally distributed populations with equal variances, the sampling distribution of the test statistic $F = s_1^2/s_2^2$ is the ***F* distribution**.

1. The *F* distribution is not symmetric.
2. Values in the *F* distribution cannot be negative.
3. The shape of the distribution depends on two different degrees of freedom.
4. Large values of *F* are evidence against $\sigma_1^2 = \sigma_2^2$.

In this section, we test the hypothesis that two population variances are equal.

For the requirements and formulas, see the table on pages 477-478. The populations must be normally distributed!

Example

At a certain urban hospital, the standard deviation in the waiting times to see an ER doctor for a non-life-threatening emergency is 32 minutes. At a second hospital, the standard deviation is 28 minutes. The sample sizes are 16 and 18, respectively. Assume that the populations of waiting times are normally distributed. Does the evidence suggest that there is more variation in the waiting times at the first hospital?

Section 10.2

A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variable.

Scatterplots are often used to identify correlation.

The **linear correlation coefficient** r measures the strength of the linear correlation between paired quantitative x- and y-values in a sample.

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2}\sqrt{n \sum y^2 - (\sum y)^2}}$$

See the requirements on pages 498-499.

Properties of the Linear Correlation Coefficient

- The value of r is always between -1 and 1 (inclusive).
- The value of r is independent of scale, invariant under transformation, and not affected by the choice of x or y.
- r measures the strength of the *linear* relationship.
- r is sensitive to outliers.

We will use our calculators (or computers) to compute the linear correlation coefficient and an equation for the "best" linear equation that describes the relationship between the variables.

Example

The following data were collected when studying a voltage-controlled amplifier. All values are in volts.

Input Voltage	1.32	1.75	2.23	2.54	3.01	3.37	3.79	4.25	4.62	5.04	5.50
Output Voltage	14	13	12	11	10	9	8	7	6	5	4

In addition to computing the linear correlation coefficient, we can use a P-value to determine whether a linear correlation exists or not (at a certain significance level).

If the computed P-value is less than the significance level, we conclude that there is a linear correlation between the variables. Otherwise, there is not sufficient evidence to support the conclusion of a linear relationship.

The value of r^2 is the proportion of the variation in y that is explained by the linear relationship between x and y.

Section 10.3

Given a collection of paired sample data, the **regression equation**

$$\hat{y} = ax + b$$

algebraically describes the relationship between the two variables x and y. The graph of the regression equation is called the **regression line**.

The regression line is the line that "best" fits the data in the least-squares sense.

The values of $\hat{y}$ predicted by the regression equation and corresponding to the values of x probably do not equal the actual y data values.

For a pair of sample x- and y-values, the **residual** is the difference between the observed sample value of y and the y-value predicted by the regression equation.

$$\text{residual} = \text{sample } y \text{ - predicted } y = y - \hat{y}$$

The regression line is the line that best fits the data in the sense that the sum of the squares of the residuals is a minimum, i.e. $\sum(y - \hat{y})^2$ is a minimum.