

Definitions:

- **data** = observations, collected or measured
- **Statistics** = study of data (well, duh!). Planning experiments, collecting, organizing, summarizing, analyzing, interpreting, presenting, and drawing conclusions from data.
- **Population** = complete set or collection we want information about. It includes **all** the objects you want info about.
- **census** = getting data from **every** member of a population
- **sample** = a subcollection of members from the population.

Example: The deer in Hixon Forest seem to be malnourished, possibly because there are too many of them. To study the health of the deer we need data such as weight, size, age, gender?, general health, etc. How do the terms above apply?

Sample data must be collected in an appropriate and suitably random way to be sure that it represents the entire population.

1.2 - Types of Data

Data comes from observations. From data we derive either statistics or parameters.

Definitions:

- **parameter** = a number that describes a population
- **statistic** = a number that describes a sample

Example: In healthy adult males the sciatic nerve conduction velocity has average 65 cm/msec and standard deviation 5 cm/msec. The conduction velocities of 16 subjects admitted to the poison control center of a metropolitan hospital with a diagnosis of methylmercury poisoning had a mean conduction velocity of 55 cm/msec and a variance of $49(\text{cm/msec})^2$.

What is the population and what is the sample? (There are really **two** populations here.)

Which numbers are parameters and which are statistics?

Data - numbers or labels?

- **qualitative data** = values are labels or non-numerical attributes, e.g. gender, political affiliation, eye color, etc. Also called categorical data. Numbers are sometimes used as convenient labels, but don't really have a numerical meaning.
- **quantitative data** = values are numerical, e.g. number of eggs in a clutch, mass of a frog. One way to further classify quantitative data is to distinguish between:
 - **discrete data** = the possible values for this data have gaps between them, e.g. number of eggs in a clutch, winnings from scratch off lottery ticket
 - **continuous data** = the possible values for this data have no gaps and form an interval or range, e.g. mass of a frog, methylmercury concentration

The scheme above is not the only way to classify data. Another useful classification scheme or hierarchy is as follows:

- **nominal level of measurement** - corresponds to categorical data, labels only and there is no order, e.g. eye color
- **ordinal level of measurement** - there is an ordering to data values and they can be compared, but differences do not make sense, e.g. letter grade, class at UW-L (freshman, sophomore, junior, senior)
- **interval level of measurement** - differences make sense, but there is not a natural zero (so ratios do not make sense), e.g. temperature ($80^{\circ}F$ is $40^{\circ}F$ hotter than $40^{\circ}F$, but it is not twice as hot); can you think of other examples?
- **ratio level of measurement** - differences and ratios make sense, e.g. mass, concentration, cholesterol level, height, volume, etc.

You should know both classification schemes. Being able to sort out what kind of data you have is often a good first step in any statistical analysis. Especially important is the being able to distinguish between qualitative (nominal and ordinal (usually)) and quantitative (interval and ratio). (The table on page 9 is a good summary of the second classification scheme.)

Homework: Section 1-2: 1-19 (odd).

1.3 - Design of Experiments

To succeed we must collect the data in a meaningful way.

A bad idea -

Definition: Voluntary Response Sample - the respondents themselves decide whether or not to be included.

Think about typical phone surveys. Do you hang up or refuse to answer? Do phone surveys reach an *unbiased* sample of the population?

The above is but one example of a bad way to collect data. If data is collected inappropriately, then it is useless.

Two primary sources of data:

- **Observational study** - make observations and measurements, but don't intervene or modify the subjects or objects being studied
- **Experiment** - apply a treatment and see how it affects the objects or subjects

We'll talk just a little about experimental and study design. Statisticians take entire courses about it, but the important thing to think about it is that your data really represents the entire population equally.

Definition: **Confounding** occurs when effects of variables are mixed and cannot be separated and identified.

Example: People drink lots of fluids, take zinc, and take vitamin C to try to get over the common cold. If they get over the cold faster, then what caused them to do so, if anything?

When there are many variables we want to control the effects of the variables that we are not of interest.

Definition: **Blinding.** Used to study effect of placebo versus treatment. **Single-blind:** subjects don't know. **Double-blind:** subjects and doctors don't know.

Definition: **Blocking.** Grouping together subjects, into blocks, with similar characteristics that might affect outcome of experiment.

Definition: **Randomization.** Randomly assigning subjects to treatment groups or within blocks to avoid biasing groups or blocks.

A guiding principle is that we want our results to be reproducible, that is, if we collect data again we'll get more or less the same results. This means that size of our samples needs to be large enough to allow for **replication**.

Definition: **Random Sample.** Every individual in the population has the same chance of being selected.

Definition: **Simple Random Sample.** Every sample of size n has the same chance of being selected.

A random sample and a simple random sample are not the same thing.

Important Note: Many of our statistical procedures will require that that our data is gathered from a *simple random sample*.

Some other sampling techniques:

- **systematic sampling** - select every k^{th} element
- **convenience sampling** - selecting the easiest to obtain data
- **stratified sampling** - divide the population into groups or strata and then select from each group.
- **cluster sampling** - divide the population area into clusters, then select some clusters and choose all members in those clusters.

We're not going to give a lot of examples here. See page 16 and 17.

Homework: 1,3,9-31 (odd) (you don't have to do every one, but do enough that you understand)

Suppose we have a data set:

6.2 16.5 19.1 23.8 26.4 26.7 29.1 31.2
34.0 35.6 38.1 38.4 38.4 40.2 41.0 47.9

To summarize this data we can make a frequency distribution:

Histogram.

Piegraphs and Bar Graphs.

Example:

Grade	Frequency
A	75
AB	61
B	129
BC	88
C	146
D	76
F	36
	611

Suppose you have a basketball team and you need to communicate, *using one number*, how tall the team is - what number do you use?

We need to choose a number that represents the typical, or average, or center of a data set. There are several ways to do this, and the best way to do this depends mostly on how the data is laid out.

1. **Mean** - average of the data. For a sample of size n with data values x_1, x_2, \dots, x_n , the mean is

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x}{n}.$$

For a population of size N , the mean is

$$\mu = \frac{\sum x}{N}.$$

The mean is the balancing point of the data set. If you add up the total distances to the mean, including the sign, you'll get zero.

2. **Median** - the central value of an ordered data set, half the data falls below the median and half falls above the median.

- for an odd number of ordered data values, the median is the middle data point
- for an even number of ordered data values, the median is the average of the two middle data points.
- there is not standard notation for the median, but to give it some notation we'll use \tilde{x} to denote the sample median

3. **Mode** - the most frequent (common) value.

- If two values occur more often than most, the data set is called **bimodal** (unlike the book, we don't require that they occur equally often)
- If there are many frequent values, then the data set is called **multimodal**.
- There doesn't have to be a mode.

Example: Consider the sample:

2, 7, 7, 8, 11, 13, 15, 33

Find the sample mean, \bar{x} . Also, make a dotplot to illustrate how the sample mean is the balancing point.

For this data set, does the sample mean give a good summary of the values?

Find the sample median, \tilde{x} . Add it to the dotplot. Is the median a good way to describe the center of this data set? How is different than the mean?

What is the mode? Is it useful?

Mean versus Median:

Skewness: a distribution of data is skewed if it is not symmetric and extends more to one side than the other.

Outliers: observations that are far from the others (how far?)

Generally if the data is not skewed (symmetric) and doesn't have outliers then the mean and the median will give similar results and the mean is preferable because it uses all of the data. When the data is skewed or has outliers, the median is sometimes preferable because the mean is affected by the skewness/outliers.

Example: Housing Prices / Starting Salaries.

Homework: Section 2-4: 3, 5, 9, 11, 17

2.5 - Measures of Variation

Consider two simple data sets:

9, 10, 11

5, 1015

What is the sample mean for each data set?

How are the data sets different?

We want to develop some ways to measure the variability or spread of the data:

1. **Range** = max - min. Measures total distance across data, but is sensitive to extremes.

2. **Standard deviation and variance** - measure the “average” distance from the mean.

Important formulas for standard deviation and variance:

definition formula for sample variance:

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

This formula isn't a particularly good one for hand calculation, but it is exactly the right formula for understanding variance. *Sample variance is (almost) the average squared deviation from the mean.*

hand calculation formula for variance:

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

formula for sample standard deviation:

However you calculated variance, just take the square root to get standard deviation.

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

population variance

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N} = \frac{1}{N} \left(\sum x^2 - \frac{(\sum x)^2}{N} \right)$$

population standard deviation

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

Understanding Standard Deviation

The value of the standard deviation tells us how spread out the data is, in particular it yields the "average" distance to the mean. The larger the standard deviation, the more spread out the data is.

Range and Standard Deviation:

$$\text{standard deviation} \approx \frac{\text{range}}{4}$$

Example: Test scores on a stats test range from about 98 to 50, estimate the standard deviation.

Empirical Rule: Also called the 68-95-99.7 rule. If the distribution of data is bell-shaped (symmetric and mound shaped) then,

- about 68% of the data falls within one standard deviation of the mean.
- about 95% of the data falls within two standard deviations of the mean.
- about 99.7% of the data falls within three standard deviations of the mean.

Example: The length of natural human pregnancies follow a bell-shaped distribution where the mean is 266 days and the standard deviation is 16 days.

Chebyshev's Theorem: The proportion, or fraction, of any set of data lying within K standard deviations of the mean is always *at least* $1 - 1/K^2$ where $K > 1$. If $K = 2$ or $K = 3$ we get:

- At least $3/4$ (75%) of all values lie within 2 standard deviations of the mean.
- At least $8/9$ (about 89%) of all values lie within 3 standard deviations of the mean.

Example: Scores on a stats test have mean 75 points and standard deviation 8 points. Using Chebyshev's Theorem what sorts of things can we say about the distribution of scores on this test?

Homework: Section 2-5: 3, 9, 19, 21, 24, 25

2.6 - Measures of Relative Standing

z-scores: Given an observation x , the z -score measures how many standard deviations away from the mean x is.

$$z = \frac{x - \mu}{\sigma} \text{ or } z = \frac{x - \bar{x}}{s}$$

A negative z -score corresponds to a value of x that is below the mean, while a positive z -score corresponds to a value of x that is above the mean.

Example: Billy scores 27 on the ACT test for which the population mean is $\mu = 20$ and the standard deviation is $\sigma = 3$ points. Sally scores 900 on the SAT which has mean $\mu = 800$ and standard deviation $\sigma = 40$. Who did better on their test?

Quartiles: Markers to divide data set into 4 equal quarters.

- Q_1 = median of lower half of data. 25% of the data is lower than Q_1 .
- Q_2 = overall median (* find this first *). 50% of the data is lower than Q_2
- Q_3 = median of upper half of data. 75% of the data is lower than Q_3 .

To find the quartiles, first sort the data from lowest to highest, then find Q_2 . Separate the data set into lower and upper halves, if n is odd, then discard the middle number. Find Q_1 and Q_3 by finding the median of the lower and upper halves of the data respectively.

Example: 2,7,7,8,11,13,15,33

Percentiles: P_1, P_2, \dots, P_{99} separate the data into 100 groups each containing about 1% of the data. If, for instance, you score in the 80th percentile on a test that means you've done better than 80% of people taking the test.

Given a score it is pretty easy to figure out what percentile it is in:

$$\text{percentile of value } x = \frac{\text{number of values less than } x}{\text{total number of values}} \cdot 100$$

Example: 2, 7, 7, 8, 11, 13, 15, 33. Find the percentile corresponding to the value 13.

It is a little trickier to go from the percentile to the score or value.

- n total number of values in the data set
- k percentile being used (e.g. for the 80th percentile, $k = 80$)
- L is the locator or index that gives the position of the value in data set. e.g. $L = 16$ means the 16th value in the sorted list.
- P_k is the value at the k th percentile
- 1. sort the data, low to high.
- 2. $L = \left(\frac{k}{100}\right) n$
- 3. if L is a whole number then P_k is the average of the L and $L + 1$ values in the list.
- 4. if L is not a whole number, then round it *up* to a whole number and use that value in the list for P_k .

Example: Find the 75th and 90th percentiles for this sample of 16 orangutan weights (in kg): 75.3, 75.6, 77.8, 81.3, 82.5, 82.6, 84.3, 85.0, 85.4, 87.5, 88.0, 88.3, 89.5, 90.1, 90.6, 92.3.

2.7 - Exploratory Data Analysis

Using tools to summarize, explore, and understand data.

Outliers: observations that are far from the others (how far?)

Inter-Quartile Range: The span of the middle 50% of the data:

$$\text{IQR} = Q_3 - Q_1$$

1.5 IQR Rule for Outliers: An observation that is more than $1.5 \times \text{IQR}$ from the nearest quartile is an outlier.

- Observations below: $Q_1 - 1.5 * \text{IQR}$
- Observations above: $Q_3 + 1.5 * \text{IQR}$

are called outliers.

Example: 2,7,7,8,11,13,15,33

Mild versus Extreme Outliers: An outlier is an observation that is more than $1.5 \times \text{IQR}$ away from the nearest quartile. If it is not more than $3 \times \text{IQR}$ away, then we call it a mild outlier. If it is more than $3 \times \text{IQR}$ away from the nearest quartile, then we call it an extreme outlier.

Example: 2,7,7,8,11,13,15,33

Simple Boxplot:

Modified Boxplot:

Homework: Section 2-7: 1, 2, 11,

Probability is the foundation for inferential statistics - generalizing from data to make statements. Suppose a large family has 10 children and all of them are boys. If we assume that this couple has normal fertility and can produce 50% boys and 50% girls, then the chance of having 10 boys is 1 in 1024 or less than 0.001. Because this seems so unlikely, we are forced to question the assumption that the couple has normal fertility. This is exactly what statisticians do - they reject explanations that are based on very low probabilities.

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

3-2 Probability Fundamentals

Definitions:

- **probability experiment** - an experiment, trial, or study in which the outcomes are random
- **sample space** - the collection or set of all possible outcomes of a probability experiment
- **event** - a subset consisting of some of the outcomes in the sample space
- **simple event** - an event consisting of a single outcome

Example: Roll a die once ...

Notation: P denotes probability. $A, B, \text{ and } C$ denote specific events. $P(A)$ is probability that event A occurs.

Determining probability.

Estimating Probability by Relative Frequency Approach: Conduct or observe an experiment/trial and count the number of times that A occurs. $P(A)$ is estimated by

$$P(A) \approx \frac{\text{number of times } A \text{ occurred}}{\text{number of times experiment/trial was repeated}}$$

A sample of 99 women who take a particular pregnancy test gives the following results:

	Positive Test Result (Pregnancy is indicated)	Negative Test Result (Pregnancy is not indicated)	
Subject is pregnant	80 (True positive)	5 (False negative)	
Subject is not pregnant	3 (False positive)	11 (True negative)	

We can use this sample to estimate various proportions (probabilities) of interest. For instance, if a woman is not pregnant, how likely is it that she will test positive? (False positive)

When all of the simple events has the same probability (like rolling a die or tossing a coin), then it is relatively simple to use a theoretical or classical approach to determine probability:

If a woman in this sample tests positive for pregnancy, what is the probability that she really is pregnant?

Classical/Theoretical Approach to Probability: If all the outcomes (simple events) in the sample space are equally likely, then

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

Example: Roll a die ...

Law of Large Numbers: As a probability experiment is repeated over and over, the relative frequency probability gets closer to the true probability.

Example: Toss a coin 100 times and you might get 54 heads.

Toss the coin 1000 times and we get 488 heads (I used the computer to do this).

Toss the coin 10000 times and we get 5051 heads.

Probability Bounds: Consider an event A

- If A is impossible, then $P(A) = 0$
- if A always happens, then $P(A) = 1$
- For any event A the probability of A must be between 0 and 1, that is, $0 \leq P(A) \leq 1$.

Sometimes we have to figure out the probability that an event does not occur. For instance if there is a 30% chance of rain today, then the probability that it will not rain today is?

Definition: The complement of an event A , denoted by \bar{A} , consists of all outcomes in which event A does NOT occur. [Note: in other texts the complement is denoted A' or A^c .]

Example: From the CIA Factbook. In the U.S. there are about 1.05 boys born for each girl. So for 105 boys, there are about 100 girls. The probability of a boy is then

$$P(\text{boy}) = \frac{105}{205} \approx 0.512.$$

What is the probability of a girl being born?

3-3 The Addition Rule

Example: Consider rolling a single die. Let A be the event that we roll an even number, let C be the event that we roll a low number. What is the probability that we roll an even number or a low number? (Note: in this class “or” is always inclusive in the sense that it means one or the other or both)

Venn Diagram for this example:

Formal Addition Rule:

Union of two events:

Intersection of two events:

Formal Addition Rule (again):

Definition: Events A and B are **disjoint** or **mutually exclusive** if they cannot occur simultaneously (they have no common outcomes, they do not intersect).

For instance “roll an even” and “roll an odd” are disjoint events.

Look at Problem 2 on Page 107.

Addition Rule for Disjoint Events: If A and B are disjoint, then

$$P(A \text{ or } B) = P(A) + P(B)$$

or,

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

Example: Probability of rolling an odd or an even?

Probability Rule for Complements:

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

Examples: Works problems 4, 8, 10, 14, 18

Homework: Section 3-3: 1, 3, 7, 9, 11, 13, 15, 17, 19, 23

3-4 Multiplication Rule

The last section was about finding $P(A \text{ or } B) = P(A \cup B)$. In this section we'll be dealing with finding $P(A \text{ and } B) = P(A \cap B)$.

Example: A = roll an even, C = roll a low. What is $P(A \cap C)$?

To understand how the multiplication rule works lets consider an idealized company or business with 100 employees. Twenty of the employees are managers and 80 are classified as clerical. Fifty of the employees are female and 50 are male. Furthermore, assume that job and gender have nothing to do with each other in this company.

	Managers	Clerical	Total
Male			50
Female			50
Total	80	20	100

If job and gender really have nothing to do with each other, then how many female managers should there be? How many male clerical workers? etc.

We can rewrite this problem in terms of probabilities.

To get the probability of each joint event (“and”) in this case we multiplied the probabilities of the events together. This works when the events are *independent*.

Definition: Two events A and B are independent if the occurrence of one does not affect the probability of the occurrence of the other. If A and B are not independent, then they are said to be dependent.

Example: Draw two cards from a shuffled deck of cards without putting a card back. Let A be the event that the first card is an ace. Let B be the event that second card is an ace. Are A and B independent?

In this example, the probability of B occurring depends on whether or not A has occurred. In this circumstance it makes sense to talk about the conditional probability of B .

Notation for conditional probability: $P(B|A)$ is the probability of event B occurring after event A has already occurred. (read $B|A$ as “ B given A ” or as “ B occurring after A has already occurred.”)

Example: Use conditional probability notation to describe the probabilities in the cards and aces example above.

Formal Multiplication Rule:

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

Example: Find the probability that the first card is an ace and the second card is an ace.

Intuitive Multiplication Rule: When the probability that A occurs (in one trial) and B occurs (in the next trial), multiply the probability of event A by the probability of event B , but be sure the probability of event B takes into account the previous occurrence of event A .

Example: See problem 10 on page 116.

Mathematical checks for independence: If any of the following equations is true, then A and B are independent events (if one equation is true, they will all be true - it is not necessary to check all three).

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \cap B) = P(A) \cdot P(B)$

*** The book does not have problems that ask you to determine if two events are independent, but I am likely to do so. ***

Example: High biological oxygen demand (BOD) and low pH are both indicators of stream pollutions. Of the many streams flowing into a large lake, 30% have high BOD, 20% have low pH, and 10% exhibit both indicators. Is the presence of high BOD independent of the presence of low pH?

Example: Roll two dice and consider their sum. Let A be the event that the sum of the dice is 6. Let B be the event that first die is a 2. Are A and B independent? Start by filling in all the possible sums:

	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

Now find $P(A)$, $P(B)$, $P(A \cup B)$ and use this information to determine if A and B are independent. (Another way to approach this is to think it through, if B occurs, does that alter the probability of A ?)

Example: Following the example above, let A be the event that the sum of the dice is 7 and let B be the event that first die is a 2. Are A and B independent?

Example: Draw a card from a well-shuffled deck. Let A be the event that the card is an ace. Let B be the event that the card is a heart. Are A and B independent?

Multiplication Rule for independent events: If A and B are known to be independent, then

$$P(A \cap B) = P(A) \cdot P(B)$$

IMPORTANT: A and B must be independent to apply this simplified multiplication rule, this is **not** always true.

Example: If the probability that a randomly selected baby in the U.S. will be a boy is 0.512, then what is the probability that four randomly selected babies will be boys?

3-5 Multiplication Rule: Beyond the Basics

Example: For simplicity, let $P(\text{boy}) = P(\text{girl}) = \frac{1}{2}$. What is the probability a couple with three children has three boys?

Example: What is the probability a couple with three children has at least one girl?

Probability of at least one:

$$P(\text{at least one}) = 1 - P(\text{none})$$

Computing conditional probability - definition:

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

Example: High biological oxygen demand (BOD) and low pH are both indicators of stream pollutions. Of the many streams flowing into a large lake, 30% have high BOD, 20% have low pH, and 10% exhibit both indicators. If a stream is known to have low pH, then what is the probability that it has high BOD?

Sometimes an intuitive approach to conditional probability is sufficient.

Intuitive approach to computing $P(B|A)$: To compute the conditional probability of B given A , assume that A has occurred and from there work out the probability that the event B will occur. (This is easiest in problems where the relative frequency approach is used so that you have the numerical data to work with instead of probabilities.)

Example: Work out Problem 18 on page 126.

- Chapter 1

- parameter vs. statistic
- population vs. sample
- qualitative vs. quantitative and discrete vs. continuous
- levels of measurement: nominal, ordinal, interval, ratio
- random sample vs. simple random sample (how are these different?)
- other kinds of sampling designs: systematic, convenience, stratified, cluster
- there are some other aspects of experimental design like blinding, replication, and randomization which we also talked about a little, but we'll have a most a question or two about these topics (worth only a small number of points)

- Chapter 2

- be able to determine classes (uses class boundaries as described in class without gaps between classes ... do not use the class limits that have gaps between them that are described in the book, the class width should be the distance between class boundaries without gaps)
- frequency distribution, relative frequency distribution (percents), cumulative frequency distribution
- histogram, bar graph, pie chart
- Center: mean, median, mode. What do they mean? Units and decimal places. Significant figures versus decimal places.
- Variation/Spread: range, standard deviation, variance. What do they mean? Units and decimal places. Significant figures versus decimal places.
- We are only worried about calculating sample mean, sample standard deviation, sample variance. Should be able to say/explain what these quantities represent.
- statistics should have units. statistics should be presented to one more significant figure than data except for sample variance which should be two more significant figures. intermediate figures should use extra decimal places.
- Empirical Rule - 68/95/99.7 only applies to bell-shaped data distributions.
- Chebyshev Rule applies to *any* distribution (at least $(1 - 1/k^2) * 100\%$ of data falls within k standard deviations of the mean)
- standard deviation \approx range / 4
- relative standing: z -scores, quartiles, percentiles.
- outliers, 1.5 IQR rule, boxplots (simple and modified - shows outliers)

- Chapter 3

- probability basics: event, outcome, $P(A)$, relative frequency and classical approach
- addition rule, A or B , union, addition rule, disjoint, complements, Venn diagram
- conditional probability, multiplication rule, A and B , intersection

In everyday the words risks and odds are used similarly, but in statistics they have particular meanings and are calculated differently.

Risk is familiar to those in health fields. Risk is the probability of an adverse outcome. Usually it is expressed as a decimal, but sometimes as a percentage. Often times it is expressed as a count. A risk of 0.3 means that 30 out every 100 people will have the adverse outcome.

Odds are more familiar to gamblers. Odds are not probabilities, but instead are ratios of probabilities. The odds of an event or the ratio of the probability the event will occur to the probability the event will not occur. (odds of $A = P(A)/P(\bar{A})$)

Example of odds: Consider rolling a single die. What are the odds that a two is rolled?

$$\text{odds of rolling two} = \frac{\text{prob. of rolling two}}{\text{prob. of not rolling two}} = \frac{\frac{1}{6}}{\frac{5}{6}} = \frac{1}{5} = 0.2$$

Odds of $\frac{1}{5}$ are often written as 1 : 5 This means the event happens one time for every five times it doesn't happen.

What are the odds a two is *not* rolled?

$$\text{odds of rolling two} = \frac{\text{prob. of not rolling two}}{\text{prob. of rolling two}} = \frac{\frac{5}{6}}{\frac{1}{6}} = \frac{5}{1} = 5$$

These odds would also be written as 5 : 1.

Note, the risk of rolling a two is $\frac{1}{6}$ and the risk of not rolling a two is $\frac{5}{6}$.

Formulas connecting odds and risk:

$$\text{risk} = \frac{\text{odds}}{1 + \text{odds}}, \quad \text{odds} = \frac{\text{risk}}{1 - \text{risk}}$$

Example: The risk of rolling a two is $\frac{1}{6}$, the odds of rolling a two are 1 : 5 or 1/5. Check out the formulas:

To understand risk and odds better, we'll use some made up data. Let's imagine that for a certain disease we have followed 10,000 smokers and 10,000 nonsmokers. 120 of the smokers get Disease X while 40 of the nonsmokers get Disease X. Thus there appears to be a higher chance of contracting Disease X among the smokers. The data is summarized in the table below:

	Disease X	No Disease X	Total
Nonsmoker	40	9,960	10,000
Smoker	120	9,880	10,000

The risk of Disease X for smokers is $120/10000 = .0012$ while the risk of Disease X for nonsmoker is $40/10000 = .0004$. Clearly there is an increased risk to smokers, how do we communicate this risk? We might regard not smoking as the "treatment" in this case, but how effective is the treatment?

Absolute Risk Reduction: measures the difference between the risks for the treatment and control groups.

$$\text{absolute risk reduction} = |P(\text{occurrence in treatment group}) - P(\text{occurrence in control group})|$$

here “occurrence” means occurrence of the disease or occurrence of some event.

Example: Nonsmokers = treatment group, Smokers = control group.

There is a total reduction in risk of _____ by being in the treatment group (the nonsmoking group). Since the reduction in risk is so small, it is hard to tell what it means. Its reciprocal is often more meaningful:

Number needed to treat: The number of subjects that must be treated in order to prevent one event:

$$\text{number to treat} = \frac{1}{\text{absolute risk reduction}}$$

(Round UP if needed)

Example: The absolute risk reduction for the nonsmoking treatment is _____. So

So for every additional _____ nonsmokers, one case of Disease X is prevented. If this were a vaccination or a medicine, we would have to treat _____ patients in order to prevent the disease once.

Relative Risk: Quantifies the reduction in risk (or increase) by being in the treatment group. It is a ratio:

$$\text{relative risk} = \frac{\text{probability of occurrence in treatment group}}{\text{probability of occurrence in control group}}$$

Example: The relative risk ratio for the nonsmoking treatment is

So the nonsmokers get Disease X at _____ the rate of the smokers. Taking the reciprocal we see that smokers get Disease X at _____ times the rate of nonsmokers. However, this can be misleading since the overall risk is still fairly small.

Odds: The odds of an event A are

$$\text{odds of } A = \frac{P(A)}{P(\bar{A})}$$

(odds against A are just the reciprocal). To express the odds reduce the fraction complete to m/n and then write $m : n$.

Example: Roll two dice, the probability of getting two ones (snake eyes) is 1/36. What are the odds of getting snake eyes?

Odds of not getting snake eyes?

Example: In the casino game of Craps, the simplest bet is called a Pass bet and it is made on the first roll of the game. It's an even money bet so that you win or lose the amount that you bet. Two dice are rolled and if the total is 7 or 11 on the first roll you win, if the total is 2,3, or 12 on the first roll you lose. If you roll 4,5,6,8,9, or 10 on the first roll, then you keep rolling until you repeat this total for a win or roll a 7 for a loss. The probability of winning the pass bet is 244/495. What are the odds of a win?

How about the odds of losing?

An alternative way to compare the risk for a treatment group to a control group is to compute the: Odds Ratio:

$$\text{odds ratio} = \frac{\text{odds in favor of event for treatment group}}{\text{odds in favor of event for control group}}$$

Example: Use the odds ratio to see how increased the odds are for smokers to get Disease X.

	Disease X	No Disease X	Total
Nonsmoker	40	9,960	10,000
Smoker	120	9,880	10,000

A **rate** is a different way to express a relative frequency or probability. If the probability of catching a cold in the winter time is 0.103, another way to express this is to say that for every 1,000 people, 103 will catch colds. Expressed another way, the incidence or rate of colds is 103 colds per 1000 people.

If relative frequencies are being used, then the rate is

$$\left(\frac{a}{b}\right)k$$

where a is the frequency or count of people for whom the event occurred, b is the total number exposed or at risk for the event, and k is a convenient multiplier such as 1000, 10000, or even 100000.

Your book lists many important rates such as birth rates, fertility rates, infant mortality rate, incidence rate (for disease), etc. on page 137 of your text. For instance:

Example: The crude birthrate for the U.S.

Number of live births: 4,026,000

Population: 285,318,000

Example: Suppose you're going to make a sandwich, there are 4 choices of bread, 3 choices of meat, 5 choices of cheese, and 6 choices of sauce. You get to choose one of each thing. How many different sandwiches can be made?

Fundamental Counting Rule: If there are m ways to do the first thing, and n ways to the second thing, then there are $m \cdot n$ different ways to do them together.

In this section we're interested in the number of possibilities when grouping or scrambling objects.

Factorial: $n! = n(n-1)(n-2) \dots (3)(2)(1)$

$$0! = 1$$

$$1! = 1$$

$$2! = 2 \cdot 1 = 2$$

$$3! = 6 \quad 4! = 4 \cdot 3! = 24, \quad 5! = 5 \cdot 4! = 120, \text{ etc.}$$

Example: The simplest protein molecule in biology is called *vasopressin* and is composed of 8 amino acids that chemically bound together in a particular order. The order in which these amino acids occur is of vital importance to the proper functioning of vasopressin. If these 8 amino acids were placed in a hat and drawn out randomly one by one, how many different arrangements of these 8 amino acids are possible?

- Let A,B,C,D,E,F,G,H symbolize the 8 amino acids
- They must fill 8 slots:
 - There are 8 possible amino acids for the first slot, 7 for the second, ...
 - $8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 8! = 40320$.
 - Of all these 40320 possible orderings, humans can use just one.
 - Probability of correct order if order is random:

$$\frac{1}{40320}$$

- For more complicated biological molecules, the correct order probability is extremely small.

Consider the following two problems:

1. Consider the set $\{ p, e, n \}$. How many two-letter "words" (including nonsense words) can be formed?
2. Consider a set consisting of three males: $\{ \text{Paul, Ed, Nick} \}$ or for short $\{ p, e, n \}$. How many two man crews can be selected from this set?

What is the difference between these problems?

In the first problem, the *order* of the selection matters. These are called permutations.

In the second problem, the order of the selection does not matter. These are called combinations.

Definition: How many permutations of length k from n distinct objects. In a permutation, **order is important**.

Example: Find ${}_5P_3$ the number of arrangements of 5 distinct objects taken 3 at a time.

For instance, how many ways can 5 people sit on a bench if the bench can only seat 3 people?

Example: Find ${}_5P_5$ – the number of arrangements of 5 distinct objects taken 5 at a time.

e.g. A book shelf has space for 5 books. How many ways can 5 books be arranged on the shelf?

Definition: Combinations - how many ways to choose/select k objects from n distinct objects when the order of the items in the selection does not matter.

Example: Find ${}_8C_5$

Example: Find ${}_8C_8$

Example: How many ways can you select 5 out of 10 friends to a dinner party?

Example:

1. How many ways can a three person subcommittee be selected from a committee of seven people?

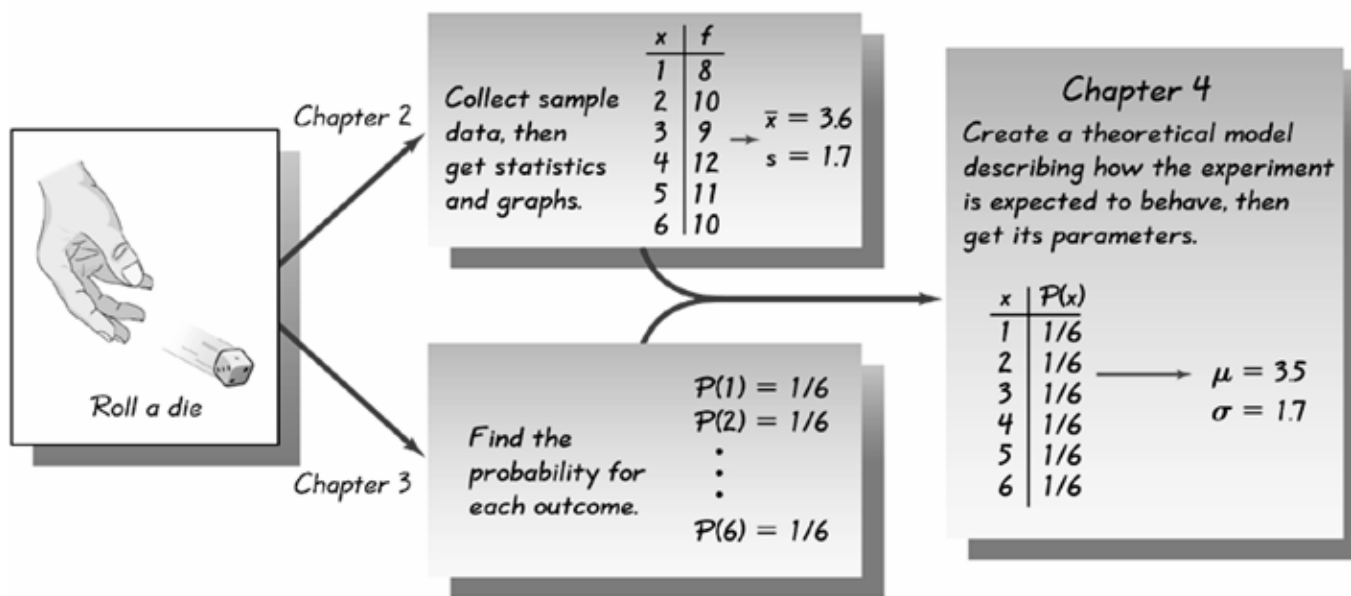
2. How many ways can a president, vice president, and secretary be selected from a committee of 7 people?

Example: From a standard 52 card deck, how many 7 card hands have exactly 4 clubs and 3 diamonds?

Example: WI Megabucks Lottery - to win you must select the correct set of 6 numbers from 49 numbers. How many possible lottery tickets are there?

When you buy a ticket, you actually purchase 2 sequences of 6 numbers, so what is the probability of winning?

Homework: Section 3-4: 17, 19, 21; Section 3-5: 1, 3, 7, 9, 11, 13, 17, 19; Section 3-8: 1-21 (odd)



4-2 Random Variables

Definitions: A **random variable** (often x) takes a value determined by a probability experiment, e.g. rolling a die, or number of girls among 10 random babies.

A **probability distribution** is a graph, table, or function that gives the probability for each value of the random variable.

A **discrete random variable** has values with gaps (Chapter 4), while a **continuous random variable** has infinitely many values in a range or interval without gaps (later Chapters).

Example: An experiment involves groups of four seedlings grown under controlled conditions. The random variable x is the number of seedlings in a group that are classified as “diseased”.

x	$P(x)$
0	0.805
1	0.113
2	0.057
3	0.009
4	0.002

The table is one way of displaying the probability distribution. Another way is with a probability histogram

The probability distribution gives us insight into the shape of the distribution (skewed, symmetric, bell-shaped, etc.), to locate the center and describe the variation the mean, variance, and standard deviation are useful.

Mean, μ , (also called expected value E):

$$\mu = E = \sum xP(x)$$

Variance, σ^2 :

$$\sigma^2 = \sum (x - \mu)^2 P(x) = \left(\sum x^2 P(x) \right) - \mu^2$$

Standard Deviation, σ :

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

Decimal Places - one more than used for values of x .

Example:

Example: An experiment involves groups of four seedlings grown under controlled conditions. The random variable x is the number of seedlings in a group that are classified as “diseased”.

<u>x</u>	<u>P(x)</u>
0	0.805
1	0.113
2	0.057
3	0.009
4	0.002

Example: (Problem 10 page 168) Assume that in a test of a gender-selection technique, a clinical trial results in 12 girls in 14 births. Refer to Table 4-1 and find the indicated probabilities.

- Find the probability of exactly 12 girls in 14 births.
- Find the probability of 12 or more girls in 14 births.
- Which probability is relevant for determining whether 12 girls in 14 births is unusually high?
- Does 12 girls in 14 births suggest that the gender-selection technique is effective?

Binomial Probability Distributions

When we repeat a bunch of independent yes/no trials and count the number of yesses, the result is often a binomial random variable. For instance, the count of the number of girls in 14 births in Table 4-1 is a binomial random variable or binomial probability distribution.

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

1. There are a fixed number of trials, n .
2. The trials are independent.
3. Each trial has only two outcomes.
4. The probabilities are the same for every trial.

Examples: Number of sixes rolled in 5 die tosses. Number of heads in 10 coin tosses. Number of girls among 14 random babies. Number of survivors among 3 snakebite victims.

Notation:

S - success, F - failure (these are arbitrary)

probability of success: $P(S) = p$

probability of failure: $P(F) = 1 - p = q$

number of trials: n

number of successes among the n trials: x (can be any number between 0 and n)

probability of success in one trial: p

probability of failure in one trial: q (note: $q = 1 - p$)

probability of x success in n trials: $P(x)$

Example: Suppose 70% of a certain kind of snakebite victims survive. If 3 people are bitten find the probability distribution for the number of survivors.

Binomial Probability Formula:

$$P(x) = {}_n C_x \cdot p^x \cdot q^{n-x} = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} \cdot p^x q^{n-x}$$

Example: If 50% of babies are girls, what is the probability of 12 girls in 14 births?

Use Formula:

Use Table A-1:

Example: If 50% of babies are girls, what is the probability of 12 or more girls in 14 births?

Use Formula:

Use Table A-1:

Example: #6 (page 176). (Introduce TI-83)

4-4: Mean, Variance, and Standard Deviation for Binomials

$$\mu = np, \sigma^2 = np(1 - p), \sigma = \sqrt{\sigma^2} = \sqrt{np(1 - p)}$$

Example: Toss a coin 100 times and count the number of heads. What is mean, variance, and standard deviation of the number of heads?

Example: If you tossed 63 heads in 100 tosses, would this be unusual or is the sort of typical result you would expect to observe due to chance?

4-5: Poisson Distribution (Discrete)

For modeling the number of occurrences of something per unit of time or space: x is the number of diseased trees per acre of forest, of x is the number of radioactive particles emitted by a sample every second.

Requirements:

- the random variable x is the number of occurrences of an event over some interval or per unit
- occurrences are random
- occurrences are independent of each other
- occurrences are uniformly distributed over the interval or unit (equally likely throughout)

The probability distribution function is

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

where μ is the mean number of occurrences per unit or interval (μ) must be specified. The standard deviation of x is $\sigma = \sqrt{\mu}$.

IMPORTANT:

unlike the binomial random variable where we specified a fixed number of trials n , there is no fixed number of trials here, the possible values for x have no upper limit.

Example: Suppose a radioactive sample emits an average of 6 particles every minute. Find the probability that exactly 3 particles are emitted in a minute.

Example cont. What is the probability 2 or fewer particles or emitted in a minute?

Example cont. What is the probability 4 or more particles or emitted in a minute?

In the last chapter we examined discrete random variables - the possible values had gaps. Here we examine continuous random variables, where there are infinitely many possible values, without gaps, in some range.

There are some basic differences between our probability models for discrete and continuous probability models. To see this consider the following example:

Imagine a random number machine that can generate ANY real number between 0 and 1, actually in the interval $[0,1)$.

Suppose all the numbers in the interval are equally likely. What is the probability that we will observe 0.7231?

How about 0.3032485?

There is zero probability associated with individual values for a continuous random variable.

Spinner:

Probability Density Function: given by a **density curve** - a curve on or above the x -axis that has area 1 beneath it. Probability of a range of values is given by area under the curve:

A special kind of continuous random variable

- models many things found in nature, often related to natural selection
- symmetric, bell-shaped, uni-modal density curves
- completely determined by the mean, μ and the standard deviation, σ , we write $x \sim N(\mu, \sigma)$ for shorthand to say the random variable x is distributed as a normal random variable with mean μ and standard deviation σ

Example: $x = \text{length of human pregnancy (days)} \sim N(266, 16)$

$\mu = 266$ days , $\sigma = 16$ days

Picture:

Flowchart for normal probability calculations:

For all of the problems that follow in this section we will work with the same normal probability distribution $x = \text{pregnancy length in days} \sim N(266, 16)$

Example: What is the probability a randomly selected pregnancy lasts less than 240 days?

Example: What *percentage* of pregnancies last more than 301 days?

Example: What is the probability a pregnancy lasts between 251 and 301 days?

Example: How long are the longest 10% of pregnancies?

Example: How long are the “middle” 95% of pregnancies?

Using your calculator:

- Forward Probability Calculations:
 - normalcdf(a,b) finds the probability that z is between a and b in the standard normal
 - * $P(z < -2) = \text{normalcdf}(-1E5, -2)$ *the “EE” key*
 - * $P(-1.3 < z < 0.8) = \text{normalcdf}(-1.3, .8)$
 - * $P(z > 1.62) = \text{normalcdf}(1.62, 1E5)$
 - normalcdf(a,b, μ , σ) finds the probability that $x \sim N(\mu, \sigma)$ is between a and b
- Backward Probability Calculations:
 - invNorm(p) finds the z score so that the cumulative probability p is to the left of that z score
 - invNorm(p, μ , σ) finds x in $N(\mu, \sigma)$ so that there is probability p to the left of x

The value of a statistic, like the sample mean or the sample standard deviation, varies from one sample to the next even when the sample size is the same. If you take two different samples of 30 frogs and compute the average weight for each sample you'll very likely get two different sample means. Repeated samples of size 30 will all yield different results, but the sample means, collectively, will behave in a predictable way.

Definition: The **sampling distribution of the sample mean** shows all the possible values of the sample mean along with the probability of each possible value, that is, it is the probability distribution of the sample means for all samples having the same size.

Defintion: The value of any statistic, such as \bar{x} , varies from one sample to the next. This is called **sampling variability**.

Example: Consider the average value of two rolled dice. What are all the possible values of the sample mean and how likely is each value?

Understanding the sampling distribution of \bar{x} is useful for making inferences about the value of the population mean μ .

A sample proportion is the proportion of “successes” divided by the sample size. For example, if 6 out of 30 frogs have defects, then the sample proportion of frogs with defects is $6/30$ or 0.2 .

Definition: The **sampling distribution of the proportion** is the probability distribution of sample proportions, with all samples having the same size n .

We'll use sample proportions to make inferences about population proportions. For instance if in a random sample of 600 students, 450 students support caps on tuition raises, then the sample proportion of students supporting caps on tuition raises is 75% . It would be reasonable to say that around 75% of all students support caps on tuition raises.

If we consider many different samples of size 600 and look at all the different sample proportions, we would observe that the sample proportions follow, approximately, a normal distribution. See pictures at top of page 217.

Estimation: We will use data from samples to compute statistics that give us estimates of population parameters. Some statistics make better estimators than others. The ones that target population parameters: mean, variance, proportion. Statistics that do not target population parameters: median, range, and standard deviation.

Homework: 5-4: 1,3,5

5-5 Central Limit Theorem

Mean and standard deviation of the sample mean distribution:

Sample Means:

For samples of size n , the sample mean is a random variable with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

If the population we are sampling from is normal (exact) or the sample size is larger than 30 (approximate), then the sample means will be normally distributed.

Summary: population normal or $n \geq 30$, then

$$\bar{x} \tilde{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Since sample means are normally distributed (in many instances), we know that

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

will be from a standard normal distribution. This gives us the ability to decide if a sample mean is unusually far from a population mean, for instance.

Sample proportions:

A sample proportion is just the relative frequency of successes in a fixed number of trials or fixed sample size:

$$\hat{p} = \frac{X}{n},$$

where \hat{p} is the sample proportion and X is the count of successes.

If p is the population proportion of successes and $np \geq 5, n(1-p) \geq 5$, then the distribution of \hat{p} is approximately normal with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

Summary: if $np \geq 5$ and $n(1-p) \geq 5$, then

$$\hat{p} \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Gregor Mendel, a pioneer of modern genetics, was crossing pea plants with green pods with pea plants with yellow pods. His theory predicted that 25% of the offspring plants should have yellow pods. In his experiment, of 580 offspring plants, 152 had yellow pods while the rest had green pods. Thus, the sample proportion of plants in Mendel's sample with yellow pods was:

$$\hat{p} = \frac{152}{580} \times 100\% = 26.2\%.$$

Was Mendel wrong? How do we explain this discrepancy?

When the sample size is large and when the sample contains sufficiently many "successes" and "failures," then a normal distribution can be used approximate a binomial distribution. This can be useful for approximating binomial probabilities when n is large and we don't have a table or computer/calculator. We won't cover that as a stand-alone topic (Section 5-6), but it does form the basis for the math in this section.

Requirements for using a normal distribution to approximate a binomial distribution (or for using a normal distribution to approximate a distribution of sample proportions)

- the sample is a simple random sample
 - the conditions for a binomial distribution are satisfied (fixed n , independent trials, S or F, fixed p)
 - the sample is large enough and contains enough successes and failures:
 - number of successes ≥ 5
 - number of failures ≥ 5
 - Notice the smallest sample that can ever be used $n = 10$, but $n = 100$ or more may be inadequate if successes or failures are really dominant.
-

Sampling Distribution for sample proportions:

Consider a simple experiment where we "toss a lopsided" coin and assign the value of 1 for a success, with probability p , and the value of 0 for a failure, with probability $q = 1 - p$. For this random variable we have

x	$P(x)$
0	$1 - p$
1	p

This is just a numerical way of thinking of single binomial trial. Since this is a discrete random variable we can find the mean μ and standard deviation σ fairly easily:

Now imagine taking a sample of size n of categorical data and counting the number of successes, this is the same as finding the number of 1's for this random variable. Since we are taking a random sample of size n from this distribution we can apply the central limit theorem. The sample mean in this case is just the number of successes divided by the sample size, in other words, the sample mean will be the sample proportion. The central limit theorem tells us what the sampling distribution of the sample proportion will be:

If we know the population proportion and the sample size, then we can predict the distribution of sample proportions. We know how much sampling variability to expect.

Example: Suppose the population proportion is $p = 0.25$ and we are considering samples of size $n = 480$.

- What is the sampling distribution of sample proportions? (What sample proportions do we expect to observe?)

- Find the middle 95% of sample proportions.

Here is the main estimation idea - even if we do not know the population proportion we can estimate the amount of expected sampling variability and then use this estimate to construct a plausible range of values for the population proportion.

Definition: A **point estimate** is a single value (or point) used to approximate a population parameter.

The sample proportion \hat{p} is the best point estimate of the population proportion p .

Definition: A **confidence interval** (CI) is a range of values, or an interval, used to estimate the true value of a population parameter.

A CI has an associated confidence level that tells us about the quality of the estimate. It is the success rate of the procedure used to construct the CI.

Definition: The **confidence level** is the probability $1 - \alpha$ (often expressed as a percentage, like 95%) that is the proportion of times that the confidence interval actually contains the population parameter, assuming that the estimation process is repeated a large number of times.

Another way to say this is the confidence level gives the percentage of samples which give “good” intervals.

To construct a confidence interval we’ll need to find a **critical value**, this is a cut-off value in a probability distribution (in this case, the standard normal - z -scores) that separates the likely values from the unlikely values.

Common Critical Values from the Standard Normal:

Confidence Level	α	Critical Value $z_{\alpha/2}$
90%	.1	1.645
95%	.05	1.96
99%	.01	2.576

Definition: When data from a simple random sample is used to estimate a population parameter, the **margin of error**, E is the maximum likely difference between the point estimate and the true value of the population parameter. The margin of error is usually found by multiplying the critical value and the standard error of the estimate which is the (estimated) standard deviation of the sampling distribution of the point estimate. Simply stated, the margin of error is the maximum likely difference due to sampling variability.

The margin of error for proportions

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

This margin of error is large enough so that $100 \times (1 - \alpha)\%$ of CI’s constructed using this margin of error will contain the true value of the population proportion.

Level $1 - \alpha$ z -interval for a population proportion

- **What?** A plausible range of values that estimates the true value of the population proportion p along with a confidence level $1 - \alpha$ that expresses our belief in the estimate.
- **When?**
 1. The sample is a simple random sample.
 2. The sample contains at least 5 “successes” and 5 “failures” (if it does not, the underlying mathematical normal approximation to a binomial distribution will likely not be very accurate)
- **How?**
 - n is the sample size
 - $\hat{p} = \frac{\text{number of successes}}{n}$, $\hat{q} = 1 - \hat{p}$ (at the end you’ll round to three significant decimal places, you should keep at least four significant decimal places for intermediate calculations)
 - Find the critical value corresponding to the level of confidence $1 - \alpha$. If it isn’t in the table above, draw the picture and look up $z_{\alpha/2}$.
 - Margin of error $E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$
 - The confidence interval $\hat{p} - E < p < \hat{p} + E$ or equivalently

$$\hat{p} \pm E \text{ or } (\hat{p} - E, \hat{p} + E)$$

Example: In a study of perception, 80 men are tested and 7 are found to have red/green color blindness. Construct and interpret a 90% confidence interval estimate of the population proportion of men that have red/green color blindness.

If we want a more precise estimate we need to decrease the margin of error. The simplest way to do this is to increase the sample size.

Finding sample size to achieve a target margin of error:

- When an estimate \hat{p} is known (from a pilot study)

$$n = \left(\frac{z_{\alpha/2}}{E} \right)^2 \hat{p}\hat{q}$$

- When no estimate for \hat{p} is available

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

Example: Consider the previous example, how large of sample would be required to estimate the population proportion of men with red/green color blindness with margin of error 0.03 and confidence level 96%. Do it with and without the prior knowledge from the previous study.

Estimating a population mean when σ is known

If the population is normally distributed or if $n \geq 30$ we know that sample means are normally distributed as $\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

If we standardize in this distribution we have

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

and we can make a statement such as ... for 95% of samples the value of z will be:

$$-1.96 < z < 1.96.$$

Replacing z by the standardized formula above we have

$$-1.96 < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96.$$

Now if σ , the population standard deviation is known, and we have an actual simple random sample of size n that gives a sample mean \bar{x} , then we can solve for μ to get

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

This is exactly a 95% confidence interval for the population mean when σ is known.

Level $1 - \alpha$ z -interval for a population mean when σ is known

- **What?** A plausible range of values that estimates the true value of the population mean μ , when σ is known, along with a confidence level $1 - \alpha$ that expresses our belief in the estimate.
- **When?**
 1. The population standard deviation σ must be known.
 2. The sample is a simple random sample.
 3. The sample size $n \geq 30$ or the population is normally distributed (not OK if $n < 30$ and the population is not normally distributed)
- **How?**
 - n is the sample size
 - \bar{x} is the sample mean
 - σ is the population standard deviation
 - Find the critical value corresponding to the level of confidence $1 - \alpha$. If it isn't in the table above, draw the picture and look up $z_{\alpha/2}$.
 - Margin of error $E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
 - The confidence interval $\bar{x} - E < \mu < \bar{x} + E$ or equivalently

$$\bar{x} \pm E \text{ or } (\bar{x} - E, \bar{x} + E)$$

Example: The health of the bear population in Yellowstone National Park is monitored by periodic measurements taken from anesthetized bears. A sample of 54 bears has a mean weight of 182.9 lb. Assuming that σ is known to be 121.8 lb. Find a 99% confidence interval estimate of the mean of the population of all such bear weights. What aspect of this problem is not realistic.

To estimate the population mean as precisely as possible it is desirable to make the margin of error,

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

as small as possible.

There are three ways to do this corresponding to the three terms in the margin of error:

1. Make $z_{\alpha/2}$ smaller by decreasing the confidence level.
2. Make σ smaller (this would be unusual)
3. Make n larger

Of the three of these, increasing the sample size is often the easiest or best choice (though collecting more data is not always cheap in terms of time or money). Just how large should the sample be?

Sample size for a desired margin of error: Given a margin of error, a confidence level (and corresponding critical value), and a known (or estimated) population standard deviation, σ , then the sample size required to achieve the margin of error E is:

$$z = \lceil \left(\frac{z_{\alpha/2} \sigma}{E} \right)^2 \rceil.$$

Example: An economist wants to estimate the mean income for the first year of college students who majored in biology. How many such incomes must be found if we want to be 95% confident that the sample mean is within \$500 of the true population mean. Assume that a previous study has estimated $\sigma = \$6250$.

6-4 Estimating a population mean when σ is unknown

Proceeding as in 6-3 if the population is normally distributed or if $n \geq 30$, then the sample means are still normally distributed, however we can't standardize unless we know σ . Instead we estimate σ using the *sample* standard deviation s . This gives a t -score instead of z -score:

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

Similar to the standard normal distribution we can find critical values for the t -distribution, $t_{\alpha/2}$, that carve out the middle $1 - \alpha$. So for with probability $1 - \alpha$ a sample will have

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

We can solve this for μ to get

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

t-distributions

Level $1 - \alpha$ t -interval for a population mean when σ is unknown

- **What?** A plausible range of values that estimates the true value of the population mean μ , when σ is unknown, along with a confidence level $1 - \alpha$ that expresses our belief in the estimate.
- **When?**
 1. The population standard deviation σ must be unknown.
 2. The sample is a simple random sample.
 3. The sample size $n \geq 30$ or the population is normally distributed (not OK if $n < 30$ and the population is not normally distributed)
- **How?**
 - n is the sample size
 - \bar{x} is the sample mean
 - s is the sample standard deviation
 - Find the t critical value corresponding to the level of confidence $1 - \alpha$. Use Table A-3 to find $t_{\alpha/2}$ in the t -distribution with degrees of freedom $df = n - 1$.
 - Margin of error $E = t_{\alpha/2} \frac{s}{\sqrt{n}}$
 - The confidence interval $\bar{x} - E < \mu < \bar{x} + E$ or equivalently

$$\bar{x} \pm E \text{ or } (\bar{x} - E, \bar{x} + E)$$

Example: Do problem 18a on page 296.

Homework: 6-4: 1, 2, 3, 4, 5, 6, 7, 8, 9, 13, 17

- Chapter 3
 - 3-4 Understanding and mathematically proving independence, multiplication rule for independent events
 - 3-5 probability of at least one, conditional probability formula
 - 3-8 factorial, permutations, combinations, probability of getting certain combinations
- Chapter 4
 - discrete random variables: mean (expected value), variance, standard deviation
 - binomial random variable (special discrete): computations of cumulative and individual probabilities, table, mean, variance, standard deviation
- Chapter 5
 - conditions for a binomial random variable
 - continuous random variable and density curve (what is a density curve?)
 - normal distributions: forward and backward normal probability calculations
 - sampling distributions (what are they in general?)
 - sampling distribution of sample means: mean and standard deviation in all cases
 - sampling distribution of sample means: shape of distribution (when is it normal?) - Central Limit Theorem
 - sample mean normal probability calculations
- Chapter 6
 - what is a confidence interval?
 - what does 95% or 99% confidence mean? what doesn't it mean?
 - confidence interval for a population proportion (conditions!)
 - confidence interval for a population mean when σ is known.

1. In a study of the effects of acid rain on fish populations in Adirondack mountain lakes, samples of yellow perch, *Perca flavescens*, were collected. Forty percent of the fish had gill filament deformities and 70% were stunted. Twenty percent exhibited both abnormalities.
 - (a) Find the probability that a randomly sampled fish will be free of both symptoms.
 - (b) If a fish has a gill filament deformity, what is the probability it will be stunted?
 - (c) Are the two symptoms independent of each other? Explain.

2. A large fruit-eating bat called the black flying fox, *Pteropus alceto*, occupies a large mangrove swamp on Indooroopilly Island. Assume that about 80% of these bats are infected with an ectoparasitic mite and 30% have larger tick parasites. Twenty percent are infected with both.
 - (a) Find the probability that a randomly chosen bath will have some parasites?
 - (b) If a randomly chosen bat has mites, what is the probability that it will not have ticks?
 - (c) Are the precense of the two types of ectoparasites independent of each other? Justify.

3. Suppose a physical education class is made up of 25 students, 10 of whom are classified as cigarette smokers. A random sample of 6 students is to be chosen for an exercise physiology experiment. What is the probability that exactly half the sample will be smokers?

4. Suppose that you have read that the leopard frog, *Rana pipiens*, has a sex ratio in most populations of 60% females and 40% males. If this is true, what is the probability that

(a) in a random sample of 15 individuals 5 or fewer will be male?

(b) in a random sample of 13 individuals exactly 8 will be female?

5. The serum cholesterol levels of a certain population of boys follow a normal distribution with a mean of 170 mg/dl and a standard deviation of 30 mg/dl.
- (a) Find the probability that a randomly chosen boy has a serum cholesterol level of 155 mg/dl or less.
- (b) Find the percentage of boys with values between 125 mg/dl and 215 mg/dl.
- (c) If a doctor wants to examine the 3% of boys with the highest cholesterol levels for possible health problems, what values of the cholesterol level should be included?
- (d) Find the probability that the **mean** serum cholesterol level of a random sample of 25 boys is below 182 mg/dl.
- (e) Determine the probability that the mean serum cholesterol level of a random sample of 100 boys is below 164 mg/dl.

Testing claims about data. If we are trying to demonstrate a particular theory or effect, then we must rule out randomness or chance as a reasonable explanation for the data or effect we are observing.

Mendel revisited: According to Mendel's theory, 25% of the hybrid pea plants should have had yellow pea pods. If he produced 580 of these hybrid pea plants, then we showed that the sampling distribution of the sample proportion of yellow pea pods was

$$\hat{p} \sim N\left(.25, \sqrt{\frac{(.25)(.75)}{580}}\right) = N(.25, .0180)$$

So a typical sample proportion, out of 580 plants, is around 0.25, with a typical deviation of 0.0180.

Mendel Example 1: In Mendel's actual experiment he found 152 out of 580 hybrid pea plants had yellow pea pods. Does this provide evidence that the true proportion of yellow pea pods is actually greater than 0.25?

Mendel Example 2: Suppose Mendel had found 167 out of 580 hybrid pea plants had yellow pea pods. Does this provide evidence that true proportion of yellow pea pods in this particular genetic hybrid is greater than 0.25?

z -test for a population proportion p

- **What?** Compare an unknown population proportion, p , to a hypothesized value, p_0
- **When?**
 - Simple Random Sample
 - The expected number of successes, $np_0 \geq 5$ and expected number of failures, $nq_0 = n(1 - p_0) \geq 5$ (this is not the number of successes and failures in your sample)
- **How?**
 1. Identify claim to be tested. Put it in symbols. $p = p_0, p > p_0, p < p_0, p \neq p_0$
 2. Give the symbolic form that must be true when the original claim is false.
 3. The alternative hypothesis, H_1 is the one that uses $<, >$, or \neq . The null is the one that uses $=$.

 4. Choose your significance level α . (0.05 and 0.01 are common - more about this soon)
 5. For the proportion test, we'll use the (approximate) sampling distribution of sample proportions which will follow a normal distribution:

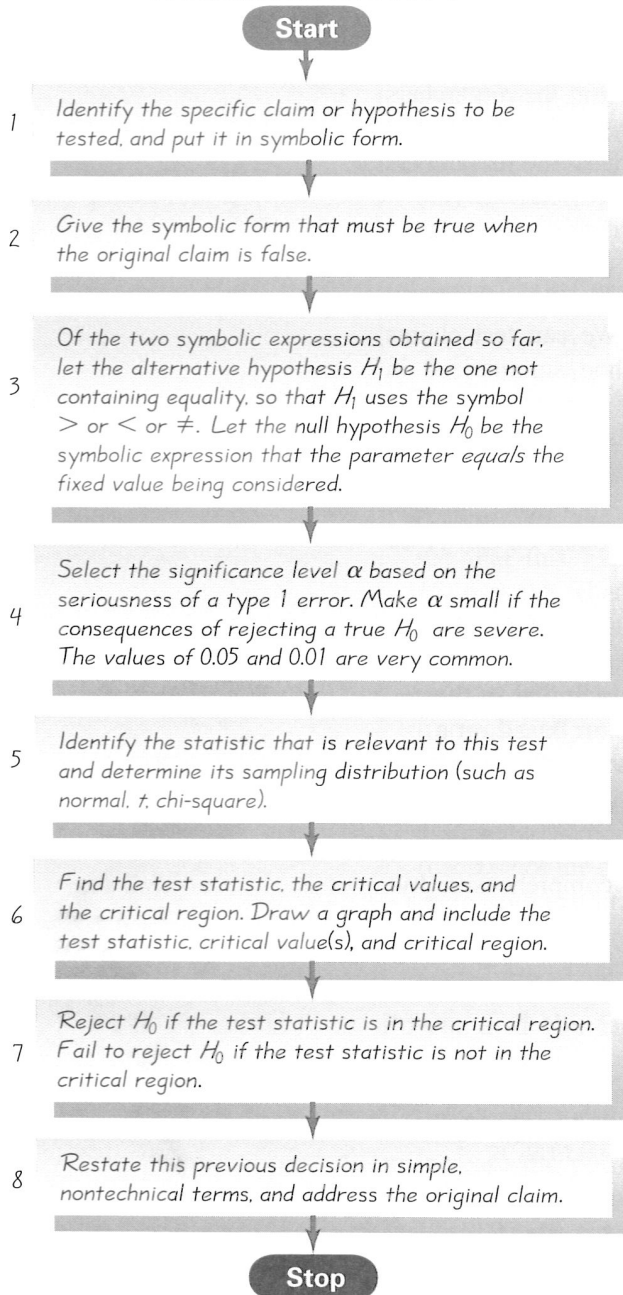
$$\hat{p} \sim N\left(p_0, \sqrt{\frac{p_0 q_0}{n}}\right)$$

6. Find the test statistic and P -value (Draw a picture).

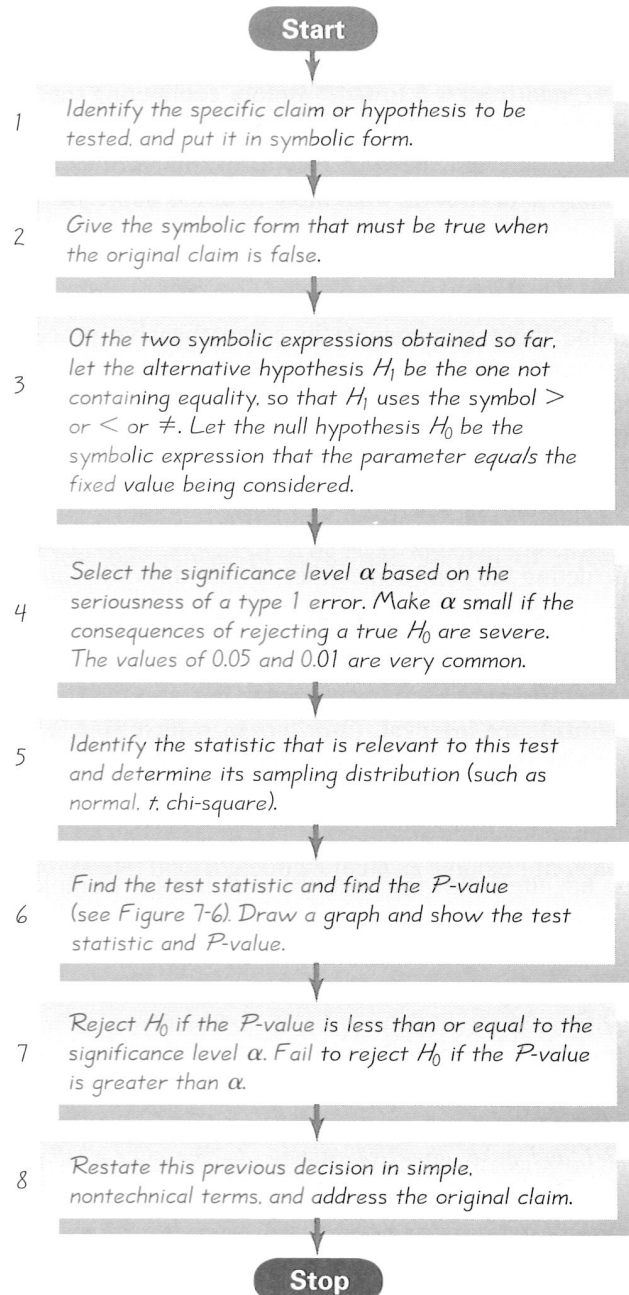
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

7. Reject H_0 if the P -value is less than or equal to the significance level α . Fail to reject H_0 if the P -value is greater than α .
8. Restate the previous decision in simple nontechnical terms and address the original claim.

Traditional Method



P-Value Method



Confidence Interval Method

Construct a confidence interval with a confidence level selected as in Table 7-2.

Because a confidence interval estimate of a population parameter contains the likely values of that parameter, reject a claim that the population parameter has a value that is not included in the confidence interval.

Table 7-2 Confidence Level for Confidence Interval

Significance Level for Hypothesis Test		Two-Tailed Test	One-Tailed Test
		0.01	99%
0.05		95%	90%
0.10		90%	80%

Mendel Example: 152 out of 580 hybrid pea plants had yellow pea pods. Does this provide evidence that true proportion of yellow pea pods in this particular genetic hybrid is greater than 0.25?

Example: Environmental concerns often conflict with modern technology, as is the case with birds that pose a hazard to aircraft during takeoff. An environmental group states that incidents of bird strikes are too rare to justify killing the birds. A pilot's group claims that among aborted takeoffs leading to aircraft going off the end of the runway, 10% are due to bird strikes. Use a 0.05 significance level to test that claim. Sample data consist of 74 aborted takeoffs in which the aircraft overran the runway. Among the 74 cases, 5 were due to bird strikes.

In a hypothesis test we have to choose between two competing theories. For instance, Mendel observed that 152 out of 580 or 26.2% of the hybrid pea pods were yellow so there are two possible explanations. One explanation is that the true proportion of yellow pea pods is actually greater than the original 25% that Mendel proposed. This is called the alternative hypothesis. Another explanation is that the true proportion of yellow pea pods is 25% and the fact that we observed 26.2% is explained by simple random variation. This is called the null hypothesis.

H_1 = there is a statistically significant effect

H_0 = there is not a statistically significant effect, only random variation

We choose between the two hypotheses by assuming the null hypothesis, H_0 is true, and asking if our sample data is unusual when H_0 is true. If the sample data is unusual when H_0 is true, then H_0 probably isn't true and we'll assume H_1 is true, otherwise we'll say we don't have enough evidence for H_1 to be true.

In the P -value approach we find the probability of sample data like ours (or more in favor of H_1) if H_0 is true. Another approach is to simply compute the test statistic and determine if it is far enough in the extreme ... this is called the *critical value approach*.

In the critical value approach we use the significance level and the direction of the hypothesis to determine cutoff values for the test statistic ... these define a rejection region.

Example: Critical Value Approach: 152 out of 580 hybrid pea plants had yellow pea pods. Does this provide evidence that true proportion of yellow pea pods in this particular genetic hybrid is greater than 0.25?

Example: Critical Value Approach: Environmental concerns often conflict with modern technology, as is the case with birds that pose a hazard to aircraft during takeoff. An environmental group states that incidents of bird strikes are too rare to justify killing the birds. A pilot's group claims that among aborted takeoffs leading to aircraft going off the end of the runway, 10% are due to bird strikes. Use a 0.05 significance level to test that claim. Sample data consist of 74 aborted takeoffs in which the aircraft overran the runway. Among the 74 cases, 5 were due to bird strikes.

Is there enough evidence to prove that a new medication is effective in reducing the average blood serum cholesterol level below 140 mg/dl?

Is the average height of a division one college basketball player greater than 6'4"?

If we can collect sample data about a quantitative random variable, then we can answer questions about the population mean value using statistical methods.

***z*-test for a population mean μ**

- **What?** Compare an unknown population mean, μ , to a hypothesized value, μ_0
- **When?**
 - Simple Random Sample (this is important - data that is carelessly collected may be useless no matter what statistical procedure you apply!)
 - The population standard deviation σ must be known
 - The population is normally distributed or $n \geq 30$ (or both)
- **How?**
 1. Identify claim to be tested. Put it in symbols. $\mu = \mu_0, \mu > \mu_0, \mu < \mu_0, \mu \neq \mu_0$
 2. Give the symbolic form that must be true when the original claim is false.
 3. The alternative hypothesis, H_1 is the one that uses $<, >, \text{ or } \neq$. The null is the one that uses $=$.

4. Choose your significance level α . (0.05 and 0.01 are common - more about this soon)
5. For the mean test, we'll use the sampling distribution of sample means which follows (approximately) a normal distribution:

$$\bar{x} \sim N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$$

6. Find the test statistic and P -value (Draw a picture).

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

7. Reject H_0 if the P -value is less than or equal to the significance level α . Fail to reject H_0 if the P -value is greater than α .
8. Restate the previous decision in simple nontechnical terms and address the original claim.

Example: The mean body temperature of healthy adults is less than 98.6°F . The sample data is: $n = 106$, $\bar{x} = 98.20^{\circ}\text{F}$. Assume that $\sigma = 0.62$ and the significance level is $\alpha = 0.01$. Use a P -value approach.

Example: Repeat with a critical value approach.

Example: The mean starting salary for college graduates who have taken a statistics course is equal to \$46,000. Sample data: $n = 65$, $\bar{x} = \$45,678$. Assume that $\sigma = 9900$ and that the significance level is $\alpha = 0.05$. Use a P -value approach.

Example: Repeat with a critical value approach.

Sometimes unusual events do happen. It really could happen that in a sample of 12 randomly selected babies there will be 11 girls! It shouldn't happen very often, but it could happen. Suppose a doctor makes a claim that they have a gender selection method and they demonstrate it by producing 11 girl babies in the next 12 births. On the basis of the data we would likely conclude that the gender selection method is effective, but what if it isn't? What if the doctor is a fraud (knowingly or unknowingly) and just happens to have beat the odds (the probability of 11 or more girls out of 12 babies is roughly $13/4096 \approx .00317$ and Mother Nature simply provided 11 girls out of 12 babies - then we have made an error (a Type I error as we'll see below).

These kind of errors are not computational errors, they are simply the results of occasionally collecting extreme data that "beats the odds." Just as people win prizes in lotteries, poker players occasionally draw straight flushes, sometimes random variation gives us extreme data that leads to the wrong conclusion.

Types of Errors:

		True State of Nature	
		The null hypothesis is true.	The null hypothesis is false.
Decision	We decide to reject the null hypothesis ($P \leq \alpha$)	Type I Error (rejecting a true null hypothesis) α	Correct decision
	We fail to reject the null hypothesis ($P > \alpha$)	Correct decision	Type II Error (failing to reject a false null hypothesis) β

See bottom of page 329 for a mnemonic for remembering the types of errors.

Analogy: Jury Trial.

Definition: The **power** of a hypothesis test is the probability $(1 - \beta)$ of rejecting a false null hypothesis when using a particular α , sample size n , the assumed value of the population parameter, and a particular assumed value of the population parameter that is an alternative to the value assumed in the null hypothesis. Essentially, the power is the probability of correctly detecting a particular true alternative hypothesis.

HW: 37-40, 42.

***t*-test for a population mean μ**

- **What?** Compare an unknown population mean, μ , to a hypothesized value, μ_0
- **When?**
 - Simple Random Sample (this is important - data that is carelessly collected may be useless no matter what statistical procedure you apply!)
 - The population standard deviation σ must be **unknown**
 - The population is normally distributed or $n \geq 30$ (or both)
- **How?**
 1. Identify claim to be tested. Put it in symbols. $\mu = \mu_0, \mu > \mu_0, \mu < \mu_0, \mu \neq \mu_0$
 2. Give the symbolic form that must be true when the original claim is false.
 3. The alternative hypothesis, H_1 is the one that uses $<, >$, or \neq . The null is the one that uses $=$.

 4. Choose your significance level α . (0.05 and 0.01 are common - more about this soon)
 5. For the mean test, we'll use the t test statistic from Student's t -distribution with degrees of freedom $n - 1$.
 6. Find the test statistic and the range of P -values from Table A3.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Recall that the degrees of freedom is $n - 1$ (round down if necessary). We'll have to approximate the P -value with a range of values from Table A3.

7. Reject H_0 if the P -value is less than or equal to the significance level α . Fail to reject H_0 if the P -value is greater than α .
8. Restate the previous decision in simple nontechnical terms and address the original claim.

Example: Find a range of values for P - value

- Left-tailed test with $n = 12$ and test stat. $t = -0.855$

- Two-tailed test with $n = 9$ and test stat. $t = 1.577$

- Right-tailed test with $n = 15$ and test stat $t = 2.100$

Example: Claim: The mean body temperature of healthy adults is less than 98.6°F . The sample data is: $n = 35$, $\bar{x} = 98.20^{\circ}\text{F}$, $s = 0.62$. The significance level is $\alpha = 0.01$.

What would the critical value be for t in the previous example?
What type of error might you have just made?

Example: The mean starting salary for college graduates who have taken a statistics course is equal to \$46,000. Sample data: $n = 27$, $\bar{x} = \$45,678$, $s = \$9900$ The significance level is $\alpha = 0.05$. Use P-value approach.

Critical value?

Type of error possible?

HW: 1-8, 9, 11, 17

Do more boys play video games than girls? Do a higher proportion of women who enter UW-L graduate than men? Do more people with advanced lung cancer survive with chemotherapy or with radiation?

To answer these questions we need to collect data from each of the *two* populations.

Conditions:

- Collect two *independent* samples - one from each population. The sample values from population must have nothing to do with the sample values collected from the other population.
- Each sample must have at least 5 success and 5 failures.

Hypothesis Tests: testing $p_1 = p_2$ versus $p_1 <$ or $>$ or $\neq p_2$

***z*-test for comparing population proportions p_1 and p_2**

- **What?** Compare two unknown population proportions p_1 and p_2 using an independently selected sample from each population.
- **When?**
 - Independent Simple Random Samples from the two populations
 - The number of success and failures in each of the two samples must be at least 5 each.

- **How?**

1. Identify claim to be tested. Put it in symbols. $p_1 = p_2, p_1 > p_2, p_1 < p_2, p_1 \neq p_2$
2. Give the symbolic form that must be true when the original claim is false.
3. The alternative hypothesis, H_1 is the one that uses $<, >, \text{ or } \neq$. The null is the one that uses $=$.

4. Choose your significance level α - this is the probability of a Type I error, but the smaller that α is (all other things being equal) then the more difficult it becomes to demonstrate a true alternative hypothesis (β gets larger, that is, the test loses power)
5. Verify the conditions required for the test.
6. Test statistic:

First find the **pooled estimate of the equal proportions** p_1 and p_2 :

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}, \text{ and } \bar{q} = 1 - \bar{p}$$

Now compute the z test statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$\hat{p}_1 = \frac{x_1}{n_1}, \quad \hat{p}_2 = \frac{x_2}{n_2}.$$

Now find the P -value and/or the critical values using the standard normal distribution as we have done with other tests based on z test statistics.

7. Reject H_0 if the P -value is less than or equal to the significance level α . Fail to reject H_0 if the P -value is greater than α .
8. Restate the previous decision in simple nontechnical terms and address the original claim.

Confidence Interval for the difference of Two Population Proportions:

- **What?** Estimate the difference between two unknown population proportions $p_1 - p_2$.
- **When?**
 - Independent Simple Random Samples from the two populations
 - The number of success and failures in each of the two samples must be at least 5 each.

- **How?**

For confidence level $1 - \alpha$, determine the critical value $z_{\alpha/2}$ from the standard normal distribution.

The point estimate is $\hat{p}_1 - \hat{p}_2$.

The margin of error is

$$z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Putting it all together we get the CI for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}.$$

Example: #18 on page 387.

In the previous example we worked a 95% CI for $p_{\text{XSORT}} - p_{\text{YSORT}}$ to be (0.0224,0.264).

Notice we don't have estimates of either proportion by itself only the difference between them. From this interval estimate we know that $p_{\text{XSORT}} - p_{\text{YSORT}}$ is a positive number which tells us p_{XSORT} has to be bigger than p_{YSORT} . Furthermore, we know p_{XSORT} is anywhere from 0.0224 to 0.264 larger than p_{YSORT} . A good explanation of this result is:

We are 95% confident that the XSORT methods produces girls at higher rate than the YSORT method produces boys. The difference appears to be between 0.0224 and 0.264.

What if the interval for $p_{\text{XSORT}} - p_{\text{YSORT}}$ had been (-0.173,-.052)?

What if the interval for $p_{\text{XSORT}} - p_{\text{YSORT}}$ had been (-0.087,.062)?

In this section we compare the averages of two populations using data *independently* collected from each of the populations:

Definitions:

Two samples are **independent** if the sample values selected from one population are not related to or somehow matched or paired with the sample values selected from the other population.

Two samples are **dependent** if the members of one sample can be used to determine the members of the other sample (pairs for instance).

t-test for comparing population means μ_1 and μ_2

- **What?** Compare two unknown population means μ_1 and μ_2 using an independently selected sample from each population.
- **When?**
 - Independent Simple Random Samples from the two populations
 - For each sample we require that the sample size be at least 30 or that the sample comes from a normally distributed population. *t*-tests are robust against departures from normality, for modestly sized samples, $15 \leq n < 30$, as long as the data does not have outliers and is not extremely far from being normal as verified with a histogram and/or boxplot, then a *t*-procedure.
- **How?**
 1. Identify claim to be tested. Put it in symbols. $\mu_1 = \mu_2, \mu_1 > \mu_2, \mu_1 < \mu_2, \mu_1 \neq \mu_2$
 2. Give the symbolic form that must be true when the original claim is false.
 3. The alternative hypothesis, H_1 is the one that uses $<, >, \text{ or } \neq$. The null is the one that uses $=$.

 4. Choose your significance level α - this is the probability of a Type I error, but the smaller that α is (all other things being equal) then the more difficult it becomes to demonstrate a true alternative hypothesis (β gets larger, that is, the test loses power)
 5. Verify the conditions required for the test.
 6. Test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

This (approximate) *t* test statistic comes from a *t*-distribution. To determine the degrees of freedom we use one of the two following techniques which, while they give different results, seldom result in different conclusions:

- For hand calculations (as in this book) we use a conservative (less likely to lead to type I errors) estimate:

$$\text{df} = \text{smaller of } n_1 - 1 \text{ and } n_2 - 1$$

- For computer calculations (including TI-83 and other software)

$$\text{df} = \frac{(A + B)^2}{\frac{A^2}{n_1 - 1} + \frac{B^2}{n_2 - 1}}$$

where

$$A = \frac{s_1^2}{n_1} \text{ and } B = \frac{s_2^2}{n_2}.$$

Now find a range of P -values using table A-3 and/or the critical values using the standard normal distribution as we have done with other tests based on t -distributions.

7. Reject H_0 if the P -value is less than or equal to the significance level α . Fail to reject H_0 if the P -value is greater than α .
8. Restate the previous decision in simple nontechnical terms and address the original claim.

t -interval for estimating difference of population means $\mu_1 - \mu_2$

- **What?** Estimate the difference between two unknown population means $\mu_1 - \mu_2$ using an independently selected sample from each population.
- **When?** - exactly the same conditions as for the t -test above.
- **How?** The level $1 - \alpha$ CI for estimating the difference $\mu_1 - \mu_2$:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where the degrees of freedom, df , used to determine the critical value from the t -distribution are determined exactly as in the t -test above.

Example:

Copper sulfate is routinely used to control algal blooms in ponds and lakes. An ichthyologist believes that copper sulfate has an adverse effect on the gill filaments of several species of fish, including largemouth bass, reducing the number of mucus cells in these species. To test her belief, she recorded the number of mucus cells per square micron in the gill filaments of untreated fish and in fish exposed to copper sulfate at 1 mg/l:

	n	\bar{x}	s
Untreated	11	15.0	2.4
Exposed	14	10.0	2.1

Assume normality for the data.

- Does the ichthyologist have support for her contention? Conduct an appropriate hypothesis test.
- Develop a 99% CI for $\mu_U - \mu_E$. Explain in words what this calculation represents.

Main idea: want to control four sources of variation that are not relevant

Example: Does color make a difference in the number of flies caught on fly paper? Compare red versus yellow fly paper.

<u>Experiment 1</u>	<u>Experiment 2</u>
30 red strips, 30 yellow strips	Hang 30 pairs of strips in 30 random locations - each pair has one yellow and one red strip.
Independent samples.	Not independent samples - dependent samples.
Hang in 60 random locations.	Look at difference within each pair, say Diff = flies on red - flies on yellow
Compare mean fly counts.	Is the average Diff significantly $\neq, <, >, 0$?
Much of the differences in means might be explained by differences in locations where the strips are hung.	Differences are not explained by location, only color of the fly paper or

For the matched pairs analysis, Experiment 2, our sample data might look like this:

Location	Red Fly Paper	Yellow Fly Paper	Diff = # flies red - # flies yellow
1	112	88	$118 - 88 = 24$
2	100	94	$100 - 94 = 6$
3	60	50	$60 - 50 = 10$
4	80	92	$80 - 92 = -12$
\vdots	\vdots	\vdots	\vdots

Generate a single sample of differences and do a one sample *t*-procedure on the differences ...

Example: A potential side effect of using oral contraceptives is an increase in blood pressure. Given below are the systolic blood pressures of 10 women measured before beginning and after having taken an oral contraceptive for a 6-month period. Assume systolic blood pressures are normally distributed.

Woman	Before	While Taking
1	113	118
2	117	123
3	111	114
4	107	115
5	115	122
6	134	140
7	121	120
8	108	105
9	106	111
10	125	129

- Do the data suggest that using an oral contraceptive increases systolic blood pressure? What type of error (I or II) could you have made here? Explain.
- Give a complete estimate of the average increase in systolic blood pressure.

Correlation means two variables are related in some way.

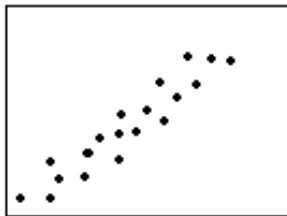
Example: Age and height of children aged 3 to 9 years:

Age (years)	Height (inches)
3	38
3	31
4	35
6	42
6	49
7	45
9	51

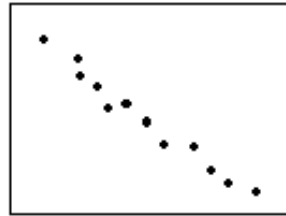
Scatterplot: A graph in which the paired (x,y) sample data are plotted with a horizontal x -axis and a vertical y -axis.

Correlation from inspection of a scatterplot:

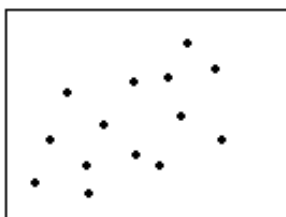
Degree of Correlation



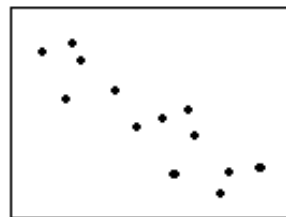
Strong Positive



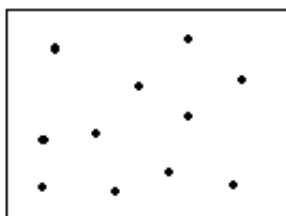
Strong Negative



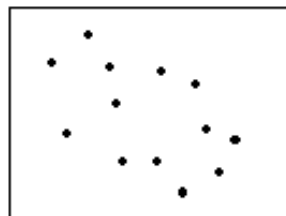
Weak Positive



Moderate Negative



None



Weak Negative

Linear Correlation Coefficient, r :

- Measures the strength and direction of a linear relationship
- r has no units
- $-1 \leq r \leq 1$
- $r = 1$ means a perfect, positive, linear relationship
- $r = -1$ means a perfect, negative, linear relationship
- r close to 0 is a sign of a weak linear relationship or no linear relationship.
- Weak, moderate, strong:

- The correlation coefficient can always be computed from (x, y) data, but in order to make meaningful inferences about correlation the paired (x, y) data must be from a *random* sample. The data must appear to follow a line (see the scatterplot). Outliers must be removed if they are errors, because outliers are especially problematic here.

$$r = \frac{\sum \left[\frac{(x - \bar{x})(y - \bar{y})}{s_x s_y} \right]}{n - 1}$$

Understanding Formula

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

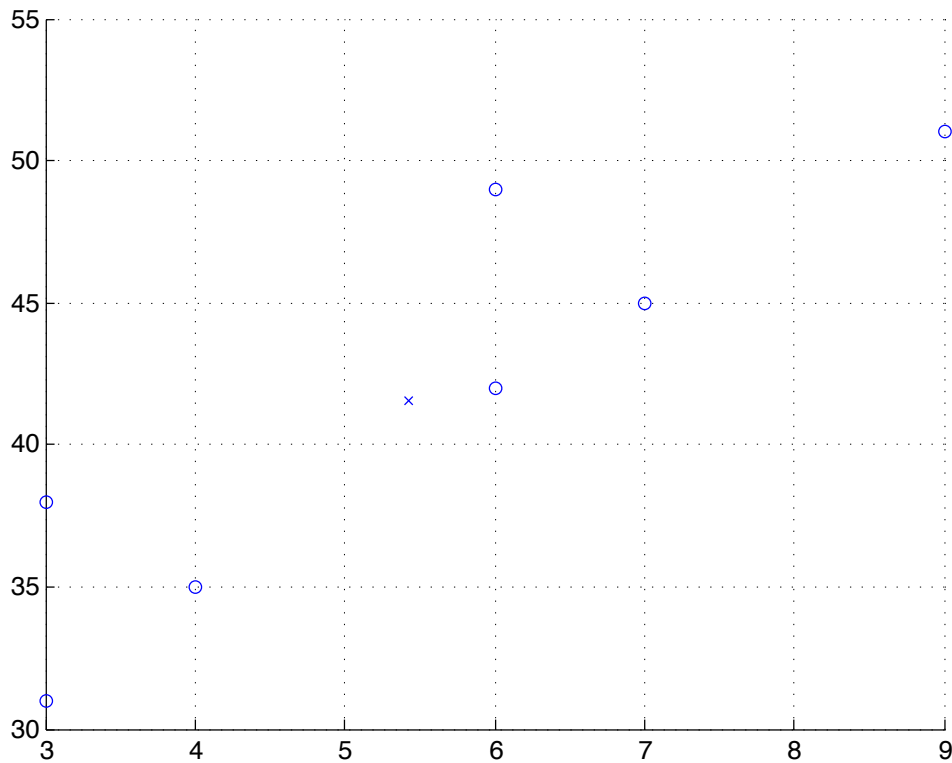
for hand calculation

Example: Compute the linear correlation coefficient:

Age (years)	Height (inches)
3	38
3	31
4	35
6	42
6	49
7	45
9	51

Example: How does it work? The point (\bar{x}, \bar{y}) is called the centroid. For the age-height data the centroid is $(\bar{x}, \bar{y}) = (5.43, 41.57)$. Essentially, linear correlation works by locating the data relative to the centroid. This happens by standardizing each of the x scores into a z score, and doing the same for y scores. For this data, $s_x = 2.22, s_y = 7.35$.

x	y	$z_x = \frac{x-5.43}{2.22}$	$z_y = \frac{y-41.57}{7.35}$	$z_x \cdot z_y$
3	38	-1.095	-.4857	.53166
3	31	-1.095	-1.438	1.5741
4	35	-.6441	-.8939	.57579
p 6	42	.25676	.0585	.01502
6	49	.25676	1.0109	.25955
7	45	.70721	.46667	.33003
9	51	1.6081	1.283	2.0632
				5.3494



TI-83 for correlation (and coefficient of determination):

Put data into two lists, say x in L_1 and y in L_2 . Then $\text{STAT} \rightarrow \text{CALC} \rightarrow 8:\text{LinReg}(a+bx) L_1, L_2$. If you don't see r , then press CATALOG and scroll down to DiagnosticOn and press ENTER a couple of times.

Hypothesis Test for Correlation.

Does the (simple) random sample provide evidence of a linear correlation between these variables for the population? The linear correlation coefficient for the population is called ρ (“rho”).

For the formal hypothesis test we use a t -test. The hypotheses are

$$H_0 : \rho = 0, \quad H_1 : \rho \neq 0$$

The t test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

and has $n - 2$ degrees of freedom.

Example: Assume the sample of age-height data is a simple random sample of children ages 3-10. Does this data provide evidence of a positive linear correlation between age and height for the entire population of such children?

Age (years)	Height (inches)
3	38
3	31
4	35
6	42
6	49
7	45
9	51

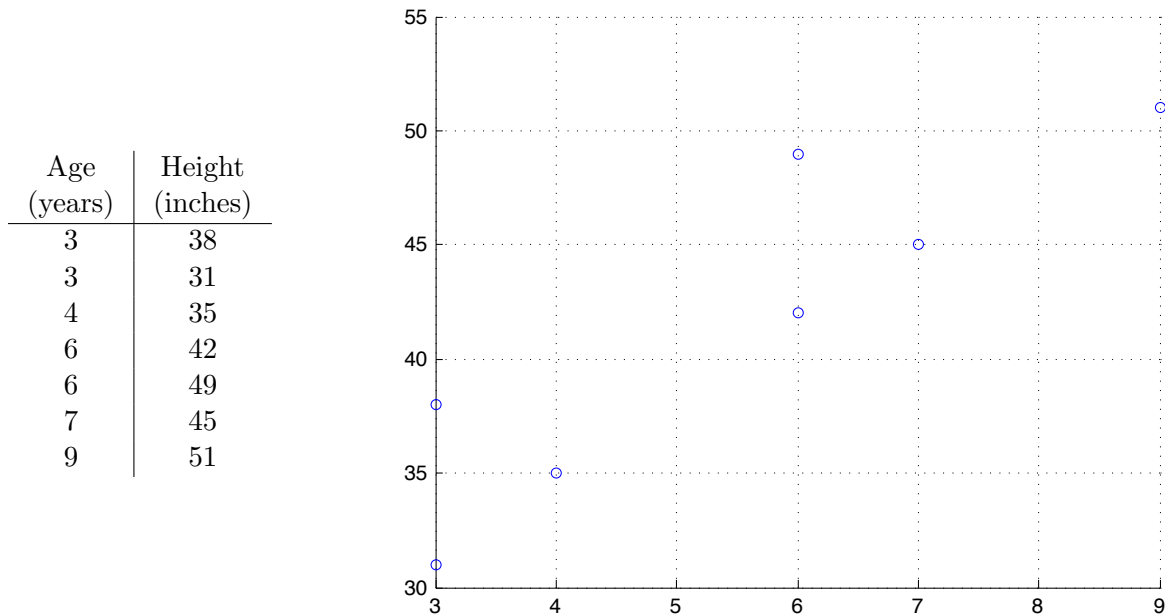
For correlation analysis the choice of x and y variable does not matter - the linear correlation coefficient is independent of which variable is called x and which is called y . In this section we are trying to build a model to predict one variable from another. In statistics this is called *regression*.

Example: For our sample data of children's ages and heights we might like to design a model which uses the age of children to predict their height. Since the scatterplot suggested a strong linear relationship and the linear correlation coefficient suggested a strong positive linear correlation, we might seek a linear model of the form

$$\hat{y} = a + bx = b_0 + b_1x$$

where x represents the age of the children and \hat{y} the predicted height of the children.

We call x the independent variable and y the dependent variable. y represents the observed values of the dependent variable (height) while \hat{y} represents predicted values (by the model).



A possible model might be $\hat{y} = 25 + 2.5x$

y -intercept:

slope (related to marginal change):

Plot the actual line on the scatterplot (don't just guess). To do this, predict the value of y for two values of x at the extremes, plot, and connect.

workspace:

Find the best possible model. To do this we want to fit the predict the observed data as closely as possible:

<http://hadm.sph.sc.edu/courses/J716/demos/leastquares/leastquaresdemo.html>

Least Squares Fit Line (Least Squares Regression Line:

IDEA: minimize the vertical differences (residuals) between the observed values of y and predicted values of y (the \hat{y} 's). That is, find a and b in $\hat{y} = a + bx$ so that the sum of the squared residuals $\sum(y - \hat{y})^2$ is as small as possible (minimized). (See the demo above.)

NOTE: On your calculator (TI-83'ish) you'll use "Option 8" to compute $\hat{y} = a + bx$, while in your book the notation is $\hat{y} = b_0 + b_1x$. So $a = b_0$ is the y -intercept and $b = b_1$ is the slope.

Computational Formulas:

$$b = b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = r \frac{s_y}{s_x}$$

$$a = b_0 = \bar{y} - b\bar{x} = \bar{y} - b_1\bar{x}$$

Example:

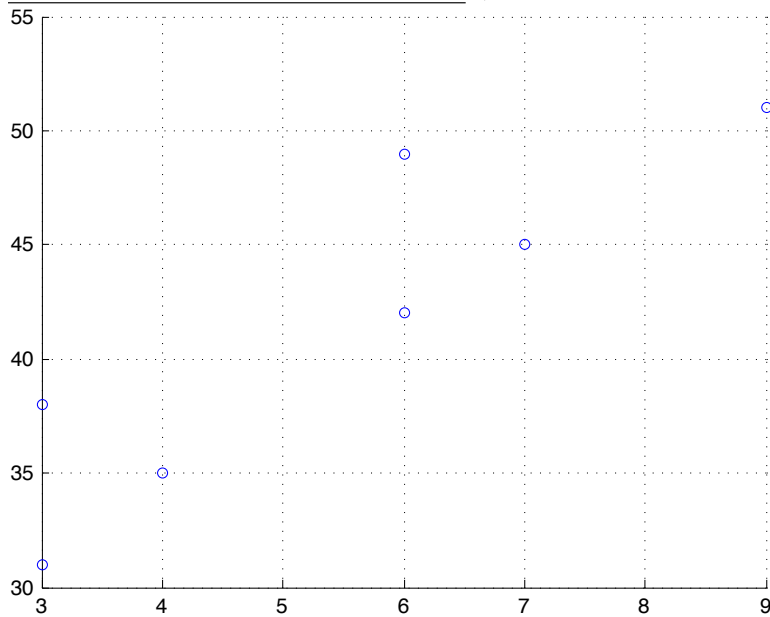
Age - x (years)	Height - y (inches)	xy	x ²	y ²
3	38	114	9	1444
3	31	93	9	961
4	35	140	16	1225
6	42	252	36	1764
6	49	294	36	2401
7	45	315	49	2025
9	51	459	81	2601
38	291	1667	236	12421

Using your calculator: Put your x and y data into two lists and then press **STATS** → **CALC** → **8 LinReg** xlist, ylist. The calculator will tell you a and b which are the value of the intercept and slope, respectively.

The model for predicting height from age:

Rounding Issues:

Making predictions from the model: (only if there is a significant linear correlation)



The right way to think to this model is that for each value of x (the age of a child) it predicts the average height of children of that age. Pretend for a moment that we set $x = 5$, so that we are predicting the height of 5 year old children. According to the model if $x = 5$ we have $\hat{y} \approx 40.3$ inches.

From here we can make two sorts of statistical inference:

1. We can try to predict the **average** height of all 5 year olds.
2. We try to predict the height of an **individual** 5 year old (a much larger range ...).

There are a few special kinds of statistical inference for correlation and linear regression.

- test to see if there is truly a linear correlation $H_1 : \rho >< \neq 0$ (t -test is 9-2)
- the true or population slope is β_1 , a hypothesis test for testing a hypothesized value of this slope, $\beta_{1,0}$ is based on a t test statistic.
- estimate the true (population) value of the slope β_1
- there is a confidence interval for predicting the population mean value of y , μ_y for a fixed value of x , e.g. what is the population mean height of a 5 year old?
- we can also predict a range of values for the next data point, y , at a fixed value of x , this is called a *prediction interval*, e.g. what is the height of an individual 5 year old? or give an interval that contains the heights of 95% of 5 year olds (covered in (9-4))

Not all of these topics are covered in this text. This doesn't mean that they aren't valuable topics. It simply means that these authors have made the choice about which topics to include.

Required Conditions:

- We have a simple random sample
- There is evidence of a linear relationship (first check the scatter plot, then run the hypothesis test for linear correlation)
- The residuals, $y - \hat{y}$, are normally distributed with mean 0 and standard deviation σ . (Basically the spread of the deviations from the line is bell-shaped and the same for every value of x .)

To estimate the average amount of deviation of the data about the least-squares line we have

$$\sigma \approx s = \sqrt{\frac{\sum(\hat{y} - y)^2}{n - 2}} = s_y \sqrt{\frac{(n - 1)(1 - r^2)}{n - 2}}$$

To do inference for the population slope, β_1 , the population mean response (for a fixed x) μ_y , and the prediction of y , we will use the statistics b_1 , $\hat{\mu}_y = \hat{y}$, and \hat{y} . For each of these statistics we need its standard error:

$$SE_{b_1} = \frac{s}{s_x \sqrt{n - 1}} = \frac{b_1 \sqrt{1 - r^2}}{r \sqrt{n - 2}}$$

$$SE_{\hat{\mu}} = s \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(n - 1)s_x^2}}$$

$$SE_{\hat{y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n - 1)s_x^2}}$$

Inference for β_1 :

- CI for $\beta_1 : b_1 \pm t_{\alpha/2} SE_{b_1}$, $df = n - 2$
- HT for $H_1 : \beta_1 >, <, \neq \beta_{1,0}$, $t = \frac{b_1 - \beta_{1,0}}{SE_{b_1}}$, $df = n - 2$. If we test against a hypothesized value of zero for the population slope, then we are testing to see if there is a linear relationship or not (the same as the linear correlation test - this hypothesis test is equivalent in that case).

CI for the mean response: (at a specific or fixed value of x)

Use the regression model to predict the mean response at $x : \hat{\mu} = b_0 + b_1x$, then the CI for the population mean at x is

$$\hat{\mu} \pm SE_{\hat{\mu}}$$

Prediction interval for individual observation: (at a specific or fixed value of x)

Use the regression model to predict the mean response at $x : \hat{y} = b_0 + b_1x$, then the CI for the population mean at x is

$$\hat{y} \pm SE_{\hat{y}}$$

To determine whether the least squares line of regression ($y_i = b_0 + b_1x$) obtained from a SRS of two quantitative variables provides a valid estimate of the population regression line $\mu_y = \beta_0 + \beta_1x$. The statistical model for simple linear regression is $y_i = \beta_0 + \beta_1x_i + \varepsilon_i$.

Example: Consider our (simple random) sample of age and height data for children:

Age (years)	Height (inches)
3	38
3	31
4	35
6	42
6	49
7	45
9	51

So far we have worked out that

$$r = .890, \quad r^2 = .792, \quad \hat{y} = b_0 + b_1x = 25.625 + 2.9375x, \quad \bar{x} = 5.43$$

Some other useful information for doing statistical inference for regression:

$$s_y = 7.345, \sigma \approx s = s_y \sqrt{\frac{(n-1)(1-r^2)}{n-2}} = 3.670$$

- Is there evidence that the population slope is greater than 2 inches per year?

- Give and interpret a 95% CI for the population slope.

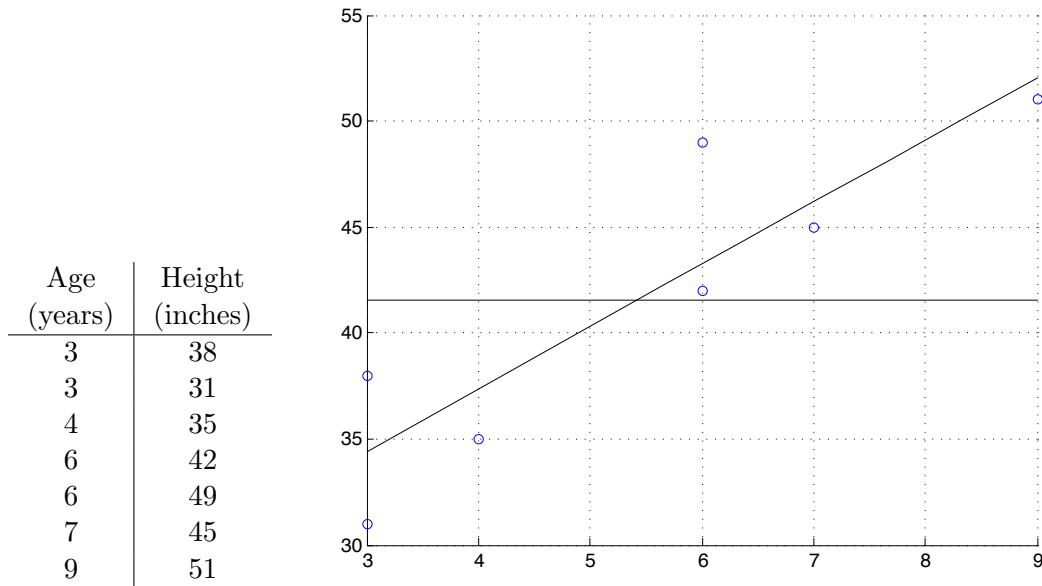
- Give and interpret a 95% CI for the population mean height of a 5 year old.

- Give and interpret a 95% prediction interval for the range of heights of individual 5 year olds?

Consider the data in problem 11 on page 471.

1. Make a scatterplot of the circumference (x) versus height (y) for the trees. Does it suggest the relationship is linear?
2. Is there evidence of a significant linear correlation between the circumference and height of the trees? Conduct the appropriate hypothesis test.
3. What is the linear regression model for predicting height of the tree (in feet) from the circumference (in feet)?
4. Interpret the slope in the context of this problem (what does it mean)?
5. Is there evidence to show that, on average, these trees increase more 3 feet in height for each additional foot of circumference?
6. With 95% confidence, estimate the slope of the population regression line.
7. Estimate, with 95% confidence, the population mean height of trees that have circumference of 4 feet. Interpret your result.
8. Give an interval that predicts a range of values containing 95% of the heights of individual trees having circumference of 4 feet. Interpret your result.
9. (Will cover on Thursday) What percentage of the variation in heights of the trees is explained by the linear relationship with circumference?

Example: Consider, for the last time, our sample of age and height data for children:



If there were not a significant linear correlation then the best prediction of height would be $\hat{y} = \bar{y} = 41.57$ inches. So the 4 year old kid that is 35 inches tall has a **total deviation** from that baseline of $y - \bar{y} = 35 - 41.57 = -6.57$.

Since a linear model is appropriate here, the linear model predicts that a 4 year old will be $\hat{y} = 25.625 + 2.9375(4) = 37.375$ inches tall or reasonably close to the actual value of 35 inches. We say that **explained deviation** by the model is $\hat{y} - \bar{y} = 37.375 - 41.57 = -4.195$.

The **unexplained deviation** by the model is $y - \hat{y} = 35 - 37.375 = -2.375$. (This distance is also called the residual and the goal of the finding the least-squares line is to make the residuals small.)

$$\text{total deviation} = \text{explained deviation} + \text{unexplained deviation.}$$

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

Squaring and summing the deviations gives a measure of the variation:

$$\text{total variation} = \text{explained variation} + \text{unexplained variation}$$

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2$$

Definition:

The **coefficient of variation** is the amount of variation in y that is explained by the regression line. It is computed as

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

It is easiest to compute by simply squaring the linear correlation coefficient!

Example: For the age and height data we have $r = 0.890$ and $r^2 = .792$. This means that 79.2% of the total variation in the children's heights can be explained by the linear association between age and height. It follows that 20.8% of the total variation remains unexplained.

Weak, strong, moderate, revisted:

Additional HW Problems: 9-4: 1-4.

Is the proposed discrete probability distribution correct? This test examines the difference between the observed data and the data that would be expected if the proposed probabilities/proportions held exactly. If the deviation is too great, then the proposed probabilities are likely wrong.

Example: Coal mining recors from the period 1851 to 1962 revealed 182 explosions that killed 10 or more workers. If the distribution of accidents by day of the week were uniform, the approximately one-sixth of the accidents could be expected to have occurred on any work day. Do the data support the uniform hypothesis?

Day	Mon	Tue	Wed	Thu	Fri	Sat
Frequency	19	34	33	36	35	29

One hundred random mud samples were taken from a lake bottom in order to determine whether two species of the genus *Stylaria* are associated, that is, tend to be occur together. *Stylaria* are oligochaete worms related to earthworms. Analyze the results below to determine if a significant association between these two species exists.

		Species B		Total
		Present	Absent	
Species A	Present	50	10	60
	Absent	25	15	40
Total		75	25	