DEFINITION

Data are ___________________________ of ________________________ (such as measurements, genders, survey responses).

Statistics is the _______________________ of planning ______________________

and _____________________________, obtaining ______________________, and

then _____________________________, ______________________________,

____________________________, _______________________________,

____________________________, and drawing _____________________________

based on the _______________________.

A **population** is the complete collection of all _________________________ (scores, people, measurements, and so on) to be studied.

A **census** is the collection of ___________________ from ___________________ member of the population.

A **sample** is a ________________________ of members selected from a ____________________________.

Remember—garbage in, garbage out! Sample data must be collected through a

process of ____________________________ selection. If sample data are not

collected in an appropriate way, the data may be completely __________________!

1.2 STATISTICAL THINKING

Key Concept…

When conducting a statistical analysis of data we have collected or analyzing a statistical analysis done by someone else, we should not rely on blind acceptance of mathematical calculations. We should consider these factors:

- Context of the data
- Source of the data
- Sampling method
- Conclusions
- Practical implications

650	24249	0
1050	20666	0
967	19413	0
500	21992	0
1700	21399	0
2000	22022	0
1100	25859	0
1300	20390	0
1400	23738	0
2250	23294	0
800	19063	0
3500	30131	0
1200	18698	0
1250	25348	0

2250	25642	1
3000	23074	1
1750	28349	1
1525	24644	1
1500	23245	1
1500	24378	1
1250	23246	1
1200	23695	1
1600	23258	1
425	19325	1
1450	20397	1
900	17256	1
675	19545	1
1450	20780	1

Description: These data for the 1991 season of the National Football League were reported by the Associated Press.

Number of cases: 28

Variable Names:

1. TEAM: Name of team
2. QB: Salary ($thousands) of regular quarterback
3. TOTAL: Total team salaries ($thousands)
4. NFC: National Football Conference (1) or American Football Conference (0)

The Data:

TEAM	QB	TOTAL	NFC
BILLS	650	24249	0
BENGALS	1050	20666	0
BROWNS	967	19413	0
BRONCOS	500	21992	0
OILERS	1700	21399	0
COLTS	2000	22022	0
CHIEFS	1100	25859	0
RAIDERS	1300	20390	0
DOLPHINS	1400	23738	0
PATRIOTS	2250	23294	0
JETS	800	19063	0
STEELERS	3500	30131	0
CHARGERS	1200	18698	0
SEAHAWKS	1250	25348	0
FALCONS	2250	25642	1
BEARS	3000	23074	1
COWBOYS	1750	28349	1
LIONS	1525	24644	1
PACKERS	1500	23245	1
RAMS	1500	24378	1
VIKINGS	1250	23246	1
SAINTS	1200	23695	1
GIANTS	1600	23258	1
EAGLES	425	19325	1
CARDINALS	1450	20397	1
49ERS	900	17256	1
BUCCANEERS	675	19545	1
REDSKINS	1450	20780	1

Example 1: Refer to the data in the table below. The x-values are weights (in pounds) of cars; the y-values are the corresponding highway fuel consumption amounts (in mi/gal).

Car Weights and Highway Fuel Consumption Amounts

WEIGHT	4035	3315	4115	3650	3565
FUEL CONSUMPTION	26	31	29	29	30

a. Context of the data.
   i. Are the x-values matched with the corresponding y-values? That is, is each x-value somehow associated with the corresponding y-value in some meaningful way?

   ii. If the x and y values are matched, does it make sense to use the difference between each x-value and the y-value that is in the same column? Why or why not?

b. Conclusion. Given the context of the car measurement data, what issue can be addressed by conducting a statistical analysis of the values?

c. Source of the data. Comment on the source of the data if you are told the car manufacturers supplied the values. Is there an incentive for car manufacturers to report values that are not accurate?

d. Conclusion. If we use statistical methods to conclude that there is a correlation between the weights of cars and the amounts of fuel consumption, can we conclude that adding weight to a car causes it to consume more fuel?

Example 2: Form a conclusion about statistical significance. Do not make any formal calculations. Either use results provided or make subjective judgements about the results.

One of Gregor Mendel's famous hybridization experiments with peas yielded 580 offspring with 152 of those peas (or 26%) having yellow pods. According to Mendel's theory, 25% of the offspring should have yellow pods. Do the results of the experiment differ from Mendel's claimed rate of 25% by an amount that is statistically significant?

1.3 TYPES OF DATA

DEFINITION

A **parameter** is a _________________________ measurement describing some

characteristic of a __________________________.

A **statistic** is a _________________________ measurement describing some

characteristic of a _______________________.

Example 3: Determine whether the given value is a statistic or a parameter.

a. 45% of the students in a calculus class failed the first exam.

b. 25 calculus students were randomly selected from all the sections of calculus I. 38% of these student failed the first exam.

DEFINITION

Quantitative (aka numerical) data consist of ___________________________

representing ____________________ or __________________________________.

Categorical (aka qualitative or attribute) data consist of ___________________

or ______________________ that are not numbers representing counts or measurements.

Give 2 examples of

a. Quantitative data

b. Categorical data

DEFINITION

Discrete data result when the number of possible values is either a ____________________ number or a

___________________________ number.

Continuous (aka numerical) data result from ___________________________ many possible values that

correspond to some ____________________________ scale that covers a _________________________ of values without gaps, interruptions or jumps.

Give 2 examples of

a. Discrete data

b. Continuous data

DEFINITION

The **nominal level of measurement** is characterized by data that consists of

____________________, ______________________, or _____________________ only. The data cannot

be arranged in an _________________________ scheme (such as low to high).

Give 2 examples of the nominal level of measurement.

DEFINITION

Data are at the **ordinal level of measurement** if they can be ______________________ in some

_______________________, but differences (obtained by subtraction) between data values either cannot be determined or are meaningless.

Give 2 examples of the ordinal level of measurement.

DEFINITION

The **interval level of measurement** is like the ______________________ level, with the additional

property that the _________________________ between any two data values is meaningful. However,

data at this level do not have a natural _____________ starting point (where none of the quantity is present).

Give 2 examples of the interval level of measurement.

DEFINITION

The **ratio level of measurement** is like the _______________________ level, with the additional property that there is a natural _______________________ starting place (where zero indicates that none of the quantity is present). For values at this level, _______________________ and ____________________ are both meaningful.

Give 2 examples of the ratio level of measurement.

1.4 CRITICAL THINKING

*"**Lies, damned lies, and statistics**" is a phrase describing the persuasive power of numbers, particularly the use of statistics to bolster weak arguments, and the tendency of people to disparage statistics that do not support their positions. It is also sometimes colloquially used to doubt statistics used to prove an opponent's point.*

DEFINITION

A **voluntary response sample (aka self-select sample)** is one in which ________________________ themselves _______________________ whether to be included.

Give two examples of voluntary response samples.

CORRELATION AND CAUSALITY

Another way to __________________________________ statistical data is to find a statistical association between two variables and to conclude that one of the variables _______________________ (or directly affects) the other variable.

____________________________ DOES NOT IMPLY CAUSALITY!

REPORTED RESULTS

When collecting data from people, it is better to take the measurements _________________________ instead of asking subjects to _____________________ results.

Give two situations in which people might falsely report results.

SMALL SAMPLES

Conclusions should ______________ be based on samples that are far too small.

PERCENTAGES

Some studies will cite _________________________ or ______________________ percentages. Keep in mind that 100% of a quantity is __________ of the quantity. If there are references made to percentages which exceed 100%, such references are often not justified.

PERCENTAGE REVIEW

"of" means ___________________

Percent means per ____________________ so $n\% = \frac{n}{100}$

Percentage of: Change the % to $\frac{1}{100}$ then multiply.

Fraction to percentage: Change the fraction to a decimal by dividing the

______________________________ by the _____________________________, then multiply by 100 and put in the percent symbol.

Decimal to percentage: Multiply the decimal by _______ and put in the percent symbol.

Percentage to decimal: Remove the _____________________ symbol and divide by _______.

Example 4: Perform the indicated operation.

a. 12% of 1200

b. Write 5/8 as a percentage.

c. Write 8.5% as a decimal

d. Write 15% as a simplified fraction

LOADED QUESTIONS

If survey questions are not worded carefully, the results of a study can be misleading. Survey questions can be ________________________ or intentionally ______________________ to elicit a desired response.

ORDER OF QUESTIONS

Sometimes survey questions are unintentionally loaded by such factors as the _______________ of items being considered.

NONRESPONSE

A __________________________ occurs when someone either refuses to respond or is unavailable. Why do you think that more and more people are refusing to participate in polls?

MISSING DATA

Results can sometimes be dramatically affected by missing data. This can be due to a ______________ occurrence such as a subject __________________ ________ of a study for reasons unrelated to the study. Some data are missing due to special factors such as the tendency of people with low incomes to be less likely to report their income.

SELF-INTEREST STUDY

Some parties with interests to __________________ will sponsor studies. We should be wary of surveys in which the sponsor can enjoy monetary gains from the results.

PRECISE NUMBERS

Numbers which are __________________ should be rounded. 2,234,786 should be rounded to 2 million.

DELIBERATE DISTORTIONS

http://jezebel.com/5730719/the-depressing-realities-of-rape-statistics

1.5 COLLECTING SAMPLE DATA

If sample data are not collected in the appropriate way, the data may be so completely useless that no amount of statistical torturing can salvage them.

DEFINITION

In an **observational study**, we ______________________ and measure specific characteristics, but we don't attempt to ______________________ the subjects being studied.

In an **experiment**, we apply some ______________________ and then proceed to ______________________ its ______________________ on the subjects. Subjects in experiments are called experimental units.

Give one example of an

a. Observational study

b. Experiment

DEFINITION

In a **random sample**, members from the _______________________ are selected in such a way that each _______________________ member in the population has an _________________________ chance of being selected.

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

DEFINITION

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

Random sample versus simple random sample

Example: Consider a box with 100 marbles.

Random Sample: Reach in and select ______________ marble. Each marble has the __________ chance of being selected.

Simple Random Sample: Reach in and select marbles in ______________ of 6 (n = 6). No matter how many _____________ you do this, every possible group of six marbles has the ____________ chance of being selected. If you then try selecting groups of 17 (n = 17) marbles, you will also find that every possible group of 17 marbles has an equal chance of being selected.

Random, but not Simple Random: For the Presidential Election, let's say you select a random sample of all voting precincts in your state, then interview all the voters as they leave the polling place. The sample is ________________ because all ______________________ have an equal chance of being selected. The sample is not simple random, because those ______________________ from precincts that were not selected have no chance of being interviewed. This is also known as a Cluster Sample.

There is no such thing as a sample that is "Simple Random, but not Random" because n can also equal a

sample of size ____.

Read more:

http://wiki.answers.com/Q/What_is_the_difference_between_a_random_sample_and_a_simple_random_sample#ixzz21Z1axK9m

DEFINITION

In **systematic sampling**, we select some ______________________ point and then select every *k*th (such as every 20th) element in the population.

With **convenience sampling**, we simply use results that are very ______________ to get.

With **stratified sampling**, we ______________________ the population into at least two different subgroups (aka strata) so that subjects within the same subgroup share the same characteristics, such

as _____________ or _________ bracket, then we draw a sample from each _________________.

In **cluster sampling**, we first ______________________ the population area into sections or

______________________, then _________________________ select some of those clusters, and

then choose ___________________ the members from those selected clusters.

Example 5: Identify which type of sampling is used: random, systematic, convenience, stratified, or cluster.

a. Every 8th driver is stopped and interviewed at a sobriety checkpoint.

b. In a neighborhood, specific streets are randomly selected and all residents on the selected streets are polled.

c. At Mira Costa College, 500 male students and 500 female students are randomly selected to participate in a study.

d. Ms. Gracey surveyed the students in her class.

e. Telephone numbers are randomly generated. Those people are selected to be interviewed.

DEFINITION

In a **cross-sectional study**, data are ________________________, ________________________, and

__________________________ at one point in time.

In a **retrospective (aka case-control) study**, data are collected from the ________________________

by going back through time (through examination of records, interviews, etc).

In a **prospective (aka longitudinal or cohort) study**, data are collected in the ______________________

from groups sharing common factors (called cohorts).

Give one example of a

a. Cross-sectional study

b. Retrospective study

c. Prospective study

DESIGN OF EXPERIMENTS

RANDOMIZATION

Subjects are assigned to different ____________ through a process of __________________ selection.

REPLICATION

Replication is the ______________________ of an experiment on more than _________ subject. Use a sample size that is __________________ enough to let us see the true nature of any _______________, and obtain the sample using an appropriate method, such as one based on _________________________.

BLINDING

Blinding is a technique in which the _____________________ doesn't know whether he or she is receiving the _______________________ or the ____________________. In a double-blind experiment, both the subject and the _____________________________ do not know whether the subject received the treatment or the placebo.

DEFINITION

Confounding occurs in an experiment when you are not able to distinguish among the

_________________________ of different _________________________.

COMPLETELY RANDOMIZED EXPERIMENTAL DESIGN

Assign subjects to different treatment groups through a process of ______________________ selection.

RANDOMIZED BLOCK DESIGN

A **block** is a group of subjects that are ____________________________, but blocks differ in ways that might affect the __________________________ of an experiment. If testing one or more treatments within different blocks, use this experimental design.

1. ________________ blocks (or groups) of subjects with similar characteristics.

2. _____________________ assign treatments to the subjects within each block.

RIGOROUSLY CONTROLLED DESIGN

Carefully assign subjects to different treatment groups, so that those given each treatment are _________________________ in ways that are important to the ____________________________.

MATCHED PAIRS DESIGN

Compare exactly ____________ treatment groups (such as treatment and placebo) by using subjects matched in __________________ that are somehow related or have similar _____________________________.

SUMMARY

1. Use ______________________________ to assign subjects to different groups.

2. Use ______________________________ by repeating the experiment on enough subjects so that effects of treatments or other factors can be clearly seen.

3. __________________________ the effects of _______________________ by using such techniques as blinding and a completely randomized experimental design.

DEFINITION

A **sampling error** is the difference between a _______________________ result and the true ___________________________ result; such an error results from chance sample fluctuation.

A **nonsampling error** occurs when the sample data are incorrectly ________________________, recorded, or __________________________ (such as by selecting a biased sample, using a defective measurement instrument, or copying the data incorrectly).

Example 6: Identify the type of observational study (cross-sectional, retrospective, or prospective)

a. Physicians at the Mount Sinai Medical Center plan to study emergency personnel who worked at the site of the terrorist attacks in New York City on September 11, 2001. They plan to study these workers from now until several years into the future.

b. University of Toronto researchers studied 669 traffic crashes involving drivers with cell phones. They found that cell phone use quadruples the risk of a collision.

2.1 REVIEW AND PREVIEW

CHARACTERISTICS OF DATA

1. **Center**: A representative or average value that indicates where the
2. ________________________ of the data set is located.
3. **Variation**: A measure of the amount that data values ________________.
4. **Distribution**: The nature or shape of the ______________________ of the

   data over the ________________________ of values (such as bell-shaped, uniform, or skewed).
5. **Outliers**: Sample values that lie very far away from the vast _______________________ of the other sample values.
6. **Time**: Changing characteristics of the data over ______________.

2.2 FREQUENCY DISTRIBUTIONS

DEFINITION

A **frequency distribution (aka frequency table)** shows how a data set is _________________________ among all of several categories (or classes) by listing all of the _________________________ along with the number of data _______________________ in each of the categories.

Height (cm)	Frequency
170	7
172	2
174	3
176	1
178	4

Weekly wages in $ of 25 workers	Tally marks	Frequency
220 - 234	\\	2
235 - 249	\\\	3
250 - 264	~~\\\\~~ \\	7
265 - 279	\\\	3
280 - 294	~~\\\\~~ \\\	8
295 - 309	\	1
310 - 324	\	1
Total		25

DEFINITION

Lower class limits are the ___________________________ numbers that can belong to the different

_____________________________.

Upper class limits are the ___________________________ numbers that can belong to the different

_____________________________.

Class boundaries are the numbers used to ____________________________ the classes, but without the gaps created by class limits.

Class midpoints are the values in the ___________________________ of the classes. Each class midpoint is found by adding the lower class limit to the upper class limit and dividing the sum by 2.

Class width is the ________________________ between two consecutive lower class limits or two consecutive lower class boundaries.

PROCEDURE FOR CONSTRUCTING A FREQUENCY DISTRIBUTION

1. Determine the number of _____________________________. The number of classes should be between 5 and 20, and the number you select might be affected by the convenience of using large numbers.
2. Calculate the class width.

$$\text{class width} \approx \frac{(\qquad\qquad\qquad\qquad)-(\qquad\qquad\qquad\qquad)}{\text{number of classes}}$$

3. Choose either the minimum data value or a convenient value ______________ the minimum data value as the first lower class limits.
4. Using the first lower class limit and the class width, list the other lower class limits. (Add the

   class width to the ______________________ lower class limit to get the second lower class limit. Add the class width to the

   _________________________ lower class limit to get the third lower class limit, and so on).
5. List the lower class limits in a _______________________ column and then enter the upper class limits.
6. Take each individual data value and put a tally mark in the appropriate class.

   ______________________ the tally marks to find the total frequency for each class.

Example 1: Let's construct our own frequency distribution which summarizes the height distribution in our class.

Height of Students in Ms. Gracey's Class

RELATIVE FREQUENCY DISTRIBUTION

In a relative frequency distribution, the frequency of a class is replaced with a relative frequency (aka a proportion) or a percentage frequency. The sum of the

relative frequencies in a relative frequency distribution must be close to ______ or ______________.

$$\text{relative frequency} = \frac{}{}$$

$$\text{percentage frequency} = \frac{}{} \times 100$$

CUMULATIVE FREQUENCY DISTRIBUTION

The cumulative frequency for a class is the ______________________ of the _________________________

for that class and all _____________________________________ classes.

CRITICAL THINKING: INTERPRETING FREQUENCY DISTRIBUTION

In statistics, we are interested in the ______________________________ of the data, and in particular,

whether the data have a ____________________________ distribution.

NORMAL DISTRIBUTION

1. The ______________________________ start low, then increase to one or two high frequencies, then decrease to a low frequency.

2. The distribution is approximately __________________________________, with frequencies preceding the maximum being roughly a mirror image of those that follow the maximum.

GAPS

The presence of gaps can show that we have data from two or more different

_______________________________________. BE CAREFUL—the converse is not necessarily true!

Example 2: Consider the frequency distribution below.

Tar (mg) in filtered cigarettes	Frequency
2-5	2
6-9	2
10-13	6
14-17	15

a. Identify the class width

b. Identify the class midpoints

c. Identify the class boundaries

d. If the criteria are interpreted very loosely, does the frequency distribution appear to have a normal distribution?

Example 3: Listed below are amounts of strontium-90 (in millibecquerels) in a simple random sample of baby teeth obtained from Pennsylvania residents born after 1979. Construct a frequency distribution with eight classes. Begin with a lower class limit of 110, and use a class width of ten.

155 142 149 130 151 163 151 142 156 133 138 161 128 144 172 137 151 166 147 163 145 116 136 158 114 165 169 145 150 150 158 151 145 152 140 170 129 188 156

2.3 HISTOGRAMS

Key Concept…
In this section we discuss a visual tool called a histogram, and its significance in representing and analyzing data.

DEFINITION

A **histogram** is a graph consisting of bars of ____________________________ width drawn adjacent to each other without ______________________. The horizontal scale represents ____________________ of quantitative data values and the vertical scale represents _________________________________.

HORIZONTAL SCALE: Use class ___________________________________ or class ____________________________________

VERTICAL SCALE: Use class _____________________________

A **relative frequency histogram** has the same shape and horizontal scale as a histogram, but the vertical scale is marked with ____________________________ frequencies (as ____________________________ or ______________________________) instead of actual frequencies.

CRITICAL THINKING: INTERPRETING HISTOGRAMS

We ____________________________________the histogram to see what we can learn about

C____

V____

D____

O____

T____

Example 4: Use the frequency distribution from example 3 to construct a histogram.

2.4 STATISTICAL GRAPHICS

Key Concept…
In this section we discuss types of statistical graphs other than ___________________________.

Our objective is to identify a _____________________________________ graph for representing

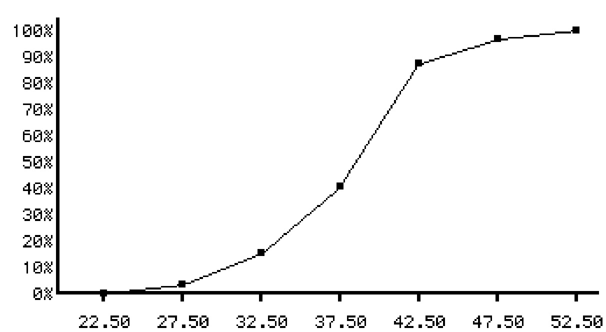
a __________________________________ set.

FREQUENCY POLYGON

A **frequency polygon** uses line segments connected to points directly above class

______________________________________ values.

A **relative frequency polygon** uses relative frequencies for the _________________________________ scale.



OGIVE

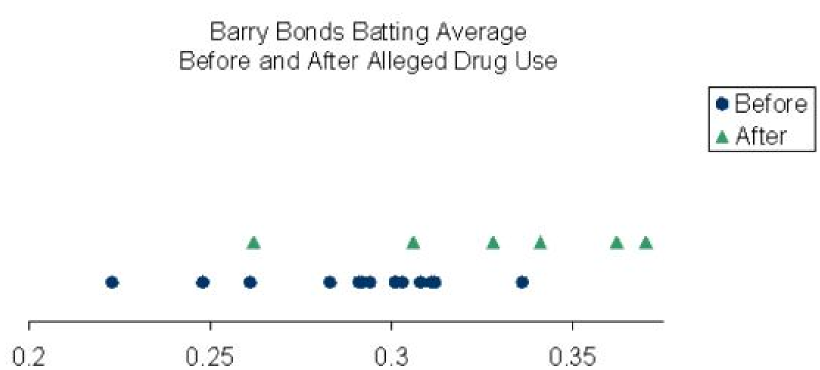
An **ogive (pronounced "oh-jive")** involves ________________________________ frequencies. Ogives are useful for determining the number of values below some particular value. An ogive is a ___________________ graph that depicts cumulative frequencies. An ogive uses class boundaries along the horizontal scale, and ____________________________ frequencies along the vertical scale.

For example, if you saved $300 in both January and April and $100 in each of February, March, May, and June, an ogive would look like Figure 1 .

Figure 1 Ogive of accumulated savings for one year.

DOTPLOTS

A **dotplot** consists of a graph in which each data value is plotted as a __________________________ or __________________ along a scale of values. Dots representing equal values are ____________________________.



STEMPLOTS

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

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

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

Example 1: Listed below are amounts of strontium-90 (in millibecquerels) in a simple random sample of baby teeth obtained from Pennsylvania residents born after 1979.

155 142 149 130 151 163 151 142 156 133 138 161 128 144 172 137 151 166 147 163 145 116 136 158 114 165 169 145 150 150 158 151 145 152 140 170 129 188 156

a. Construct a stemplot of the amounts of Strontium-90

i. What does the stemplot suggest about the distribution?

BAR GRAPH

A **bar graph** uses bars of ____________________ width to show frequencies of categories of

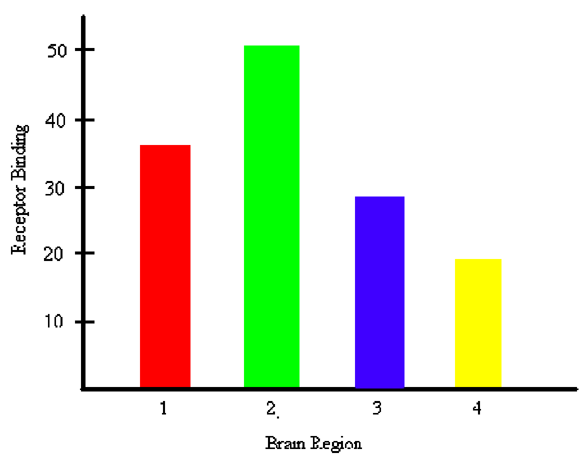
________________________________ data. The vertical scale represents __________________________

or _________________________ frequencies. The horizontal scale identifies the different

______________________ of qualitative data. The bars may or may not be separated by small gaps.

A **multiple bar graph** has two or more sets of bars, and is used to compare two or more

_________________________ sets.



PARETO CHARTS

A **Pareto chart** is a bar graph for _______________________________ data, with added stipulation that

the bars are arranged in descending order according to __________________________________. The

vertical scale represents __________________________________ or ______________________________

frequencies. The horizontal scale identifies the different categories of ___________________________ data.

PIE CHARTS

A **pie chart** is a graph that depicts ______________________________ data as slices of a

___________________________, in which the size of each slice is proportional to the frequency count for each category.

Example 2: Chief financial officers of U.S. companies were surveyed about areas in which job applicants make mistakes. Here are the areas and the frequency of responses: interview (452); résumé (297); cover letter (141); reference checks (143); interview follow-up (113); screening call (85).

a. Construct a pie chart representing the given data.

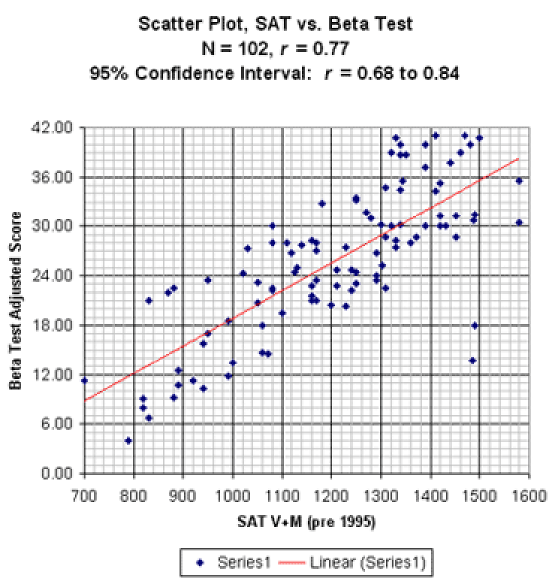
b. Construct a Pareto chart of the data.

c. Which graph is more effective in showing the importance of the mistakes made by job applicants?

SCATTERPLOTS

A **scatterplot (aka scatter diagram)** is a plot of ordered pair ________________________________

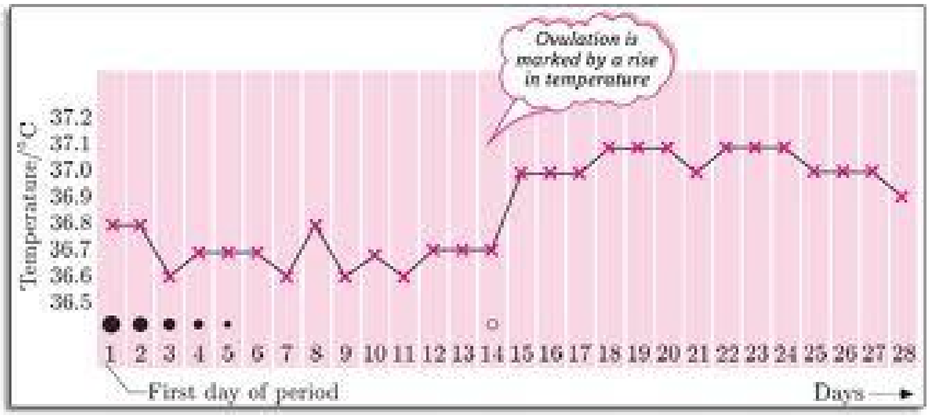
data with a horizontal *x*-axis and a vertical *y*-axis. The horizontal axis is used for the first (*x*) variable, and the vertical axis is used for the second variable. The pattern of the plotted points is often helpful in determining whether there is a ________________________________ between the two variables.



TIME-SERIES GRAPH

A **time-series graph** is a graph of *time-series data,* which are ___________________________________

data that have been collected at different points in _________________________.



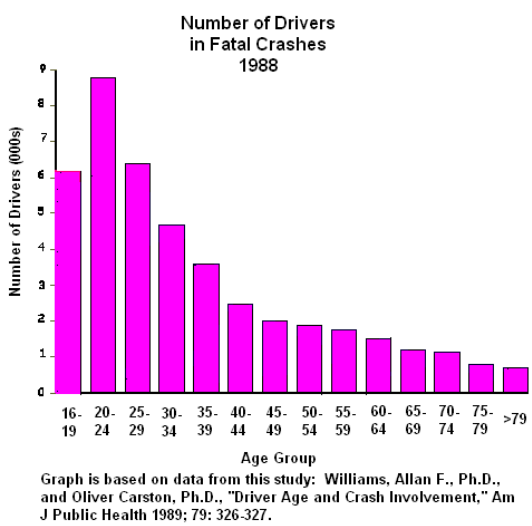
2.5 CRITICAL THINKING: BAD GRAPHS

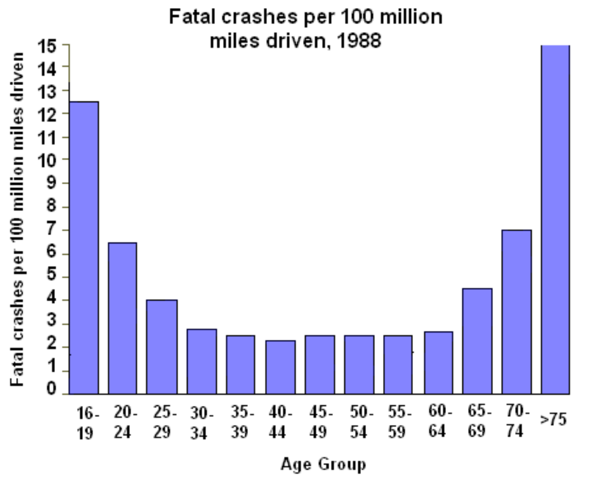
Nonzero axis

Some graphs are misleading because one or both of the ______________________ begin at some value

other than ________________________, so the differences are _____________________________________.

The following statistics suggest that 16-year-olds are safer drivers than people in their twenties, and that octogenarians are very safe. Is this true?





Graph is based on data from this study: Williams, Allan F., Ph.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.

Solution: No. As the following graph shows, the reason 16-year-old and octogenarians appear to be safe drivers is that they don't drive nearly as much as people in other age groups.

Pictographs

Drawings of objects, often called pictographs, are often misleading.

3.2 MEASURES OF CENTER

DEFINITION

A **measure of center** is a value at the ____________________________ or ________________________ of a data set.

DEFINITION

The **arithmetic mean (aka mean)** of a set of data is the _____________________ of ____________________________ found by _______________________ the ________________ values and ____________________________ the total by the ______________________________ of data values.

$$\text{mean} = \frac{\sum x}{n} = \underline{\hspace{10cm}}$$

**One advantage of the mean is that it is relatively ______________________, so that when samples are selected from the same population, sample means tend to be more ___________________ than other measures of center. Another advantage of the mean is that it takes every ____________ value into account. However, because the mean is _______________________________ to every value, just one ________________ value can affect it dramatically. Because of this fact, we say the mean is not a ________________________ measure of center.

NOTATION

Example 1: Find the mean of the following numbers:

17 23 17 22 21 34 27

DEFINITION

The **median** of a data set is the measure of center that is the ___________________________ value when the original data values are arranged in ____________________________ of increasing (or decreasing) magnitude. The median is often denoted _____________ (pronounced "x-tilde"). To find the median, first ____________________ the values, then follow one of these two procedures:

1. If the number of data values is ____________________, the median is the number located in the exact _____________________________ of the list.
2. If the number of data values is _____________________, the median is the __________________ of the ___________________________ two numbers.

**The median is a _______________________________ measure of center, because it does not change by _______________________ amounts due to the presence of just a few ________________________ values.

Example 2:

a. Find the median of the following numbers:

   17 23 17 22 21 34 27

b. Find the median of the following numbers

17 23 17 22 34 27

DEFINITION

The **mode** of a data set is the value that occurs with the greatest ________________________________.

A data set can have more than one mode, or no mode.

- When two data values occur with the same greatest frequency, each one is a ________________ and the data set is ________________________.
- When more than two data values occur with the same greatest frequency, each is a __________________ and the data set is said to be ___________________________________.
- When no data value is repeated, we say there is no _________________.

**The mode is the only measure of center that can be used with data at the ________________________ level of measurement.

Example 3:

a. Find the mode of the following numbers:

17 23 17 22 21 34 27

b. Find the mode of the following numbers

17 23 17 22 21 34 27 22

DEFINITION

The **midrange** of a data set is the measure of center that is the value ___________________________ between the ___________________________ and __________________________________ values in the original data set. It is found by adding the maximum data value to the minimum data value and then dividing the sum by two.

$$\text{midrange} = \underline{\hspace{10cm}}$$

**The midrange is rarely used because it is too _______________________ to extremes since it uses only the minimum and maximum data values.

Example 4: Find the midrange of the following numbers:

17 23 17 22 21 34 27

ROUND-OFF RULE FOR THE MEAN, MEDIAN, AND MIDRANGE

Carry _____________________ more decimal place than is present in the original data set. Because values of the mode are the same as some of the original data values, they can be left without any rounding.

MEAN FROM A FREQUENCY DISTRIBUTION

When working with data summarized in a frequency distribution, we don't know the __________________ values falling in a particular ________________. To make calculations possible, we assume that all sample values in each class are equal to the class ____________________________. We can then add the ____________________________ from each ________________________ to find the total of all sample values, which we can the ____________________________ by the sum of the frequencies, $\sum f$

$$\bar{x} = \frac{\sum (f \cdot x)}{\sum f}$$

Example 5: Find the mean of the data summarized in the given frequency distribution.

Tar (mg) in nonfiltered cigarettes	Frequency
10-13	1
14-17	0
18-21	15
22-25	7
26-29	2

WEIGHTED MEAN

When data values are assigned different weights, we can compute a weighted mean.

$$\bar{x} = \frac{\sum (w \cdot x)}{\sum w}$$

Example 6: A student earned grades of 92, 83, 77, 84, and 82 on her regular tests. She earned grades of 88 on the final and 95 on her class project. Her combined homework grade was 77. The five regular tests count for 60% of the final grade, the final exam counts for 10%, the project counts for 15%, and homework counts for 15%. What is her weighted mean grade? What letter grade did she earn?

SKEWNESS

A comparison of the ____, __________________________, and

_______________________ can reveal information about the characteristic of **skewness**. A distribution

of data is said to be ____________________________ if it is not ____________________________ and

extends more to one side than the other.

A Comparison of the Mean, Median, and Mode

The mean, median, and mode are affected by what is called skewness (i.e., lack of symmetry) in the data.

- Here is Figure 15.6, which showed a normal curve, a negatively skewed curve, and a positively skewed curve:

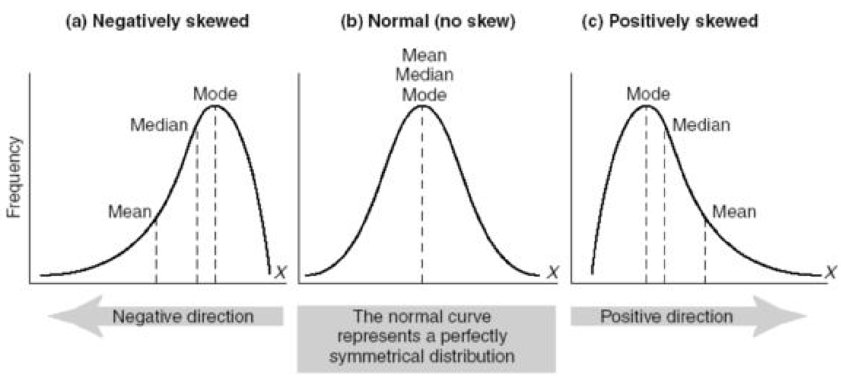


FIGURE 15.6 Examples of normal and skewed distributions

- Look at the above figure and note that when a variable is normally distributed, the mean, median, and mode are the same number.
- When the variable is skewed to the left (i.e., negatively skewed), the mean shifts to the left the most, the median shifts to the left the second most, and the mode the least affected by the presence of skew in the data.
- Therefore, when the data are negatively skewed, this happens:
  mean < median < mode.
- When the variable is skewed to the right (i.e., positively skewed), the mean is shifted to the right the most, the median is shifted to the right the second most, and the mode the least affected.
- Therefore, when the data are positively skewed, this happens:
  mean > median > mode.
- If you go to the end of the curve, to where it is pulled out the most, you will see that the order goes mean, median, and mode as you "walk up the curve" for negatively and positively skewed curves.

3.3 MEASURES OF VARIATION

DEFINITION

The **range** of a set of data values is the ____________________________ between the

_____________________________ and the ___________________________ data value.

DEFINITION

The **standard deviation** of a set of sample values, denoted by s, is a measure of ____________________ of values about the _________________. It is a type of ____________________________ deviation of values from the mean that is calculated by using either of the following formulas:

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

or

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

- The standard deviation is a measure of ___________________________ of all values from the ________________________.
- The value of the standard deviation is usually __________________________.
  - It is zero only when all of the data values are the same ________________________.
  - It is never _________________________.
- Larger values of the standard deviation indicate ______________________ amounts of ____________________________.
- The value of the standard deviation can increase dramatically with the inclusion of one or more _____________________________.
- The units of the standard deviation are the same units as the original ________________ values.

General Procedure for Finding Standard Deviation (1st formula)	Specific Example Using the Following Numbers: 2, 4, 5, 16
Step 1: Compute the mean $\overline{x}$	
Step 2: Subtract the mean from each individual sample value	
Step 3: Square each of the deviations obtained from Step 2.	
Step 4: Add all of the squares obtained from Step 3.	
Step 5: Divide the total from Step 4 by the number $n-1$, which is one less than the total number of sample values present.	
Step 6: Find the square root of the result from Step 5. The result is the standard deviation.	

STANDARD DEVIATION OF A POPULATION

The definition of standard deviation and the previous formulas apply to the standard deviation of

____________________________ data. A slightly different formula is used to calculate the standard

deviation σ of a _____________________________________: instead of dividing by $n-1$, we divide by the population size N.

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

DEFINITION

The **<u>variance (aka dispersion aka spread)</u>** of a set of values is a measure of ______________________

equal to the _________________________ of the ______________________________

_______________________________.

Sample variance: s^2

Population variance: σ^2

**The sample variance is an unbiased estimator of the _____________________ variance, which means

that values of s^2 tend to target the value σ^2 of instead of systematically tending to

_______________________________ or underestimate σ^2.

USING AND UNDERSTANDING STANDARD DEVIATION

One simple tool for understanding standard deviation is the ________________ ________________ of ____________________, which is based on the principle that for many data sets, the vast majority (such as 95%) lie within ____ standard deviations of the ______________________.

RANGE RULE OF THUMB

Interpreting a known value of the standard deviation: We informally defined __________________ values in a data set to be those that are typical and not too _________________________. If the standard deviation of a collection of data is ______________________, use it to find rough estimates of the _______________________ and __________________________ values as follows:

minimum "usual " value = (mean) – 2 x (standard deviation)

maximum "usual " value = (mean) + 2 x (standard deviation)

Estimating a value of the standard deviation *s*: To roughly estimate the standard deviation from a collection of ________________________ sample data, use

$$s \approx \frac{\text{range}}{4}$$

Example 1: Use the range rule of thumb to estimate the ages of all instructors at MiraCosta if the ages of instructors are between 24 and 60.

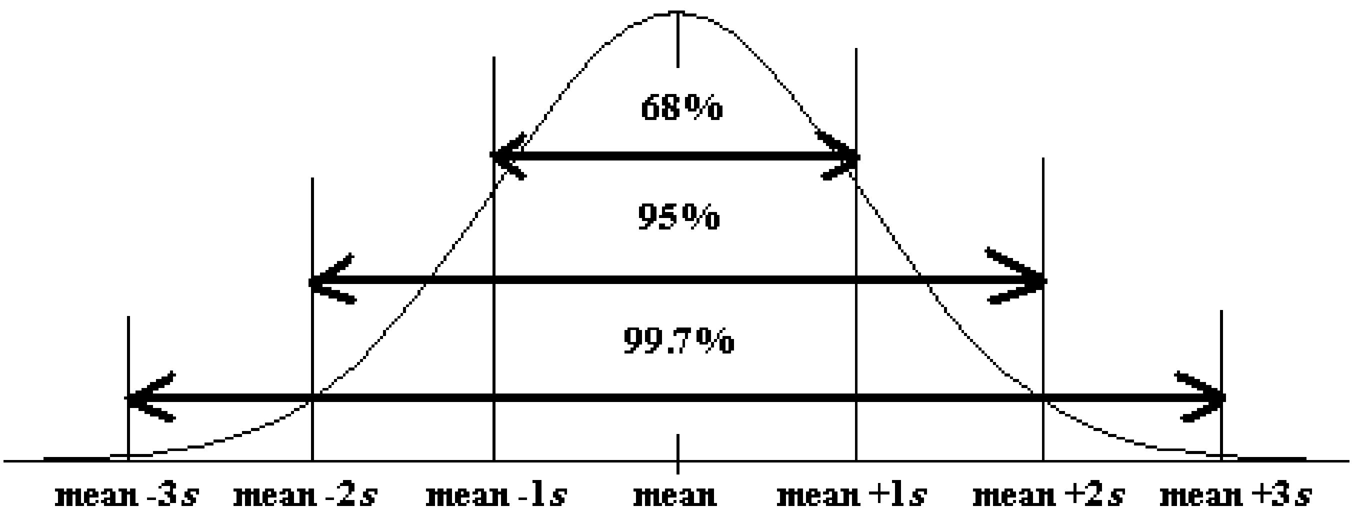
EMPIRICAL (OR 68-95-99.7) RULE FOR DATA WITH A BELL-SHAPED DISTRIBUTION

Another concept that is helpful in interpreting the value of a standard deviation is the

________________________ rule. This rule states that for data sets having a _______________________

that is approximately _____________________________, the following properties apply:

- About 68% of all values fall within 1 standard deviation of the mean
- About 95% of all values fall within 2 standard deviations of the mean
- About 99.7% of all values fall within 3 standard deviations of the mean



Example 2: The author's Generac generator produces voltage amounts with a mean of 125.0 volts and a standard deviation of 0.3 volt, and the voltages have a bell-shaped distribution. Use the empirical to find the approximate percentage of voltage amounts between

a. 124.4 volts and 125.6 volts

b. 124.1 volts and 125.9 volts

CHEBYSHEV'S THEOREM

The ______________________________(or fraction) of any data set lying within *K* standard deviations of the mean is always ________ _____________________ $1-\frac{1}{K^2},\ K\geq 1$. For *K* = 2 or *K* = 3, we get the following statements:

- At least ¾ or 75% of all values lie within 2 standard deviations of the mean.
- At least 8/9 or 89% of all values lie within 3 standard deviations of the mean.

COMPARING VARIATION IN DIFFERENT POPULATIONS

When comparing ________________________ in _____________ different sets of ________________, the ________________________ deviations should be compared only if the two sets of data use the same _________________ and ____________________ and they have approximately the same ______________.

DEFINITION

The **coefficient of variation (aka CV)** for a set of nonnegative sample or population data, expressed as a percent, describes the standard deviation ______________________________ to the ____________________________, and is given by the following:

Sample : $CV=\frac{s}{\bar{x}}\cdot 100\%$

Population: $CV=\frac{\sigma}{\mu}\cdot 100\%$

Example 3: Find the coefficient of variation for each of the two sets of data, then compare the variation.

The trend of thinner Miss America winners has generated charges that the contest encourages unhealthy diet habits among young women. Listed below are body mass indexes (BMI) for Miss America winnersfrom two different time periods.

BMI (from the 1920s and 1930s): 20.4 21.9 22.1 22.3 20.3 18.8 18.9 19.4 18.4 19.1

BMI (from recent winners): 19.5 20.3 19.6 20.2 17.8 17.9 19.1 18.8 17.6 16.8

3.4 MEASURES OF RELATIVE STANDING AND BOXPLOTS

BASICS OF *Z*-SCORES, PERCENTILES, QUARTILES, AND BOXPLOTS

A ______________________ (aka standard value) is found by converting a value to a

_____________________________ scale.

DEFINITION

The ***z* score (aka standard value)** is the number of _________________________ deviations a given value x is above or below the _________________________. The z score is calculated by using one of the following:

$$\text{Sample:} \quad z = \frac{x - \bar{x}}{s} \qquad \text{Population:} \quad z = \frac{x - \mu}{\sigma}$$

ROUND-OFF RULE FOR *Z* SCORES

Round *z* scores to ______________________ decimal places. This rule is due to the fact that the standard table of *z* scores (Table A-2 in Appendix A) has *z* scores with two decimal places.

Z SCORES, UNUSUAL VALUES, AND OUTLIERS

In Section 3.3 we used the _______________________ ____________________ of ____________________ to conclude that a value is _____________________ if it is more than 2 standard deviations away from the ______________________. It follows that unusual values have *z* scores less than ____________ or greater than _____________.

Example 1: The U.S. Army requires women's heights to be between 58 inches and 80 inches. Women have heights with a mean of 63.6 inches and a standard deviation of 2.5 inches. Find the *z* score corresponding to the minimum height requirement and find the *z* score corresponding to the maximum height requirement. Determine whether the minimum and maximum heights are unusual.

DEFINITION

Percentiles are measures of _____________________________, denoted _____________________________, which divide a set of data into ______________________ groups with about _______________ of the values in each group. The process of finding the percentile that corresponds to a particular data value *x* is given by the following:

Percentile of *x* = --

NOTATION

n

k

L

P_k

Example 2: Use the given sorted values, which are the number of points scored in the Super Bowl for a recent period of 24 years.

36 37 37 39 39 41 43 44 44 47 50 53 54 55 56 56 57 59 61 61 65 69 69 75

a. Find the percentile corresponding to the given number of points.
   i. 65
   ii. 41

b. Find the indicated percentile or quartile.
   i. Q_1
   ii. P_{80}
   iii. P_{95}

DEFINITION

Quartiles are measures of ____________________________, denoted ____________________________, which divide a set of data into _____________________ groups with about ______________ of the values in each group.

FIRST QUARTILE:

SECOND QUARTILE:

THIRD QUARTILE:

DEFINITION

For a set of data, the **5-number summary** consists of the ____________________ value, the ______________ _______________________, the ________________ (aka ______________________ _________________________), the ___________________ ________________________, and the ________________ value.

A **boxplot (aka box-and-whisker diagram)** is a graph of a data set that consists of a ____________________ extending from the _________________________ value to the ____________________________ value, and a ___________________ with lines drawn at the ________________________ _______________________, the _______________________, and the ____________________.

OUTLIERS

When ___________________________ data, it is important to ____________________ and

_______________________ outliers because they can strongly affect values of some important

statistics, such as the __________ and ___________________________ _______________________.

In __________________________ ___________________________, a data value is an

________________________________ if it is…

above quartile 3 by an amount greater than 1.5 x inner quartile range or **below quartile 1 by an amount greater than 1.5 x inner quartile range**

_______________________ _______________________ are called __________________________ or

__________________________ boxplots, which represent ________________________________ as

special points. A **modified boxplot** is a boxplot constructed with these modifications: (1) A special

symbol, such as an ________________________ or point is used to identify ______________________

and (2) the solid horizontal line extends only as far as the minimum and maximum values which are not outliers.

Example 3: Use the given sorted values, which are the number of points scored in the Super Bowl for a recent period of 24 years to construct a boxplot. Are there any outliers?

36 37 37 39 39 41 43 44 44 47 50 53 54 55 56 56 57 59 61 61 65 69 69 75

Outlier check:

PUTTING IT ALL TOGETHER

We have discussed several basic tools commonly used in statistics. When designing an

___________________________, __________________________ data, reading an article in a professional

journal, or doing anything else with data, it is important to consider certain key factors, such as:

- ________________________________ of the data
- ______________________________ of the data
- _______________________________ method
- Measures of ___________________________
- Measures of ___________________________
- _______________________________
- _______________________________
- Changing __________________________ over ___________________
- __________________________________ implications

4.1 REVIEW AND PREVIEW

RARE EVENT RULE FOR INFERENTIAL STATISTICS

If, under a given assumption, the ___________________________ of a particular observed is extremely

_________________________, we conclude that the ______________________________ is probably not

________________________.

4.2 BASIC CONCEPTS OF PROBABILITY

PART 1: BASICS OF PROBABILITY

In considering _____________________________, we deal with procedures that produce

_____________________________.

DEFINITION

An **event** is any ____________________________ of _____________________ or __________________

of a _____________________________.

A **simple event** is an ____________________________ or ____________________ that cannot be

further broken down into simpler ___________________________.

The **sample space** for a ____________________________ consists of all possible ___________________

_________________________.

NOTATION

P

$A,\ B,$ and C

$P(A)$

1. **Relative Frequency Approximation of Probability**

   Conduct (or ____________________) a _____________________________, and count the number

   of times that event *A* _________________________ occurs. Based on these actual results, *P(A)*

   is approximated as follows:

   P(A) = --

2. **Classical Approach to Probability (Requires _____________________ _____________________ Outcomes)**

   Assume that a given procedure has *n* different _______________________ events and that each

   of these simple events has an ____________________ chance of ___________________________.

   If an event *A* can occur in *s* of these *n* ways, then

   P(A) = -- = -------------

3. **Subjective Probabilities**

   P(A) is ___________________________ by using knowledge of the _________________________

   circumstances.

Example 1: Identifying Probability Values

a. What is the probability of an event that is certain to occur?

b. What is the probability of an impossible event?

c. A sample space consists of 10 separate events that are equally likely. What is the probability of each?

d. On a true/false test, what is the probability of answering a question correctly if you make a random guess?

e. On a multiple-choice test with five possible answers for each question, what is the probability of answering correctly if you make a random guess?

Example 2: Adverse Effects of Viagra
When the drug Viagra was clinically tested, 117 patients reported headaches, and 617 did not (based on data from Pfizer, Inc.).

a. Use this sample to estimate the probability that a Viagra user will experience a headache.

b. Is it unusual for a Viagra user to experience headaches?

c. Is the probability high enough to be of concern to Viagra users?

LAW OF LARGE NUMBERS

As a procedure is ____________________________ again and again, the ___________________________

_____________________________ probability of an event tends to approach the _____________________

probability. The ________ _______ _________________ __________________________ tells us that

relative frequency approximations tend to get better with more _____________________.

PROBABILITY AND OUTCOMES THAT ARE NOT EQUALLY LIKELY

One common ___________________________ is to __________________________ assume that outcomes are _____________________________ likely just because we know nothing about the likelihood of each outcome.

Example 3: Flip a coin 50 times and record your results.

a. What is the sample space?

b. What is the probability of getting a result of heads?

SIMULATIONS

Many procedures are so___________________________ that the classical approach is impractical. In such cases, we can more easily get good estimates by using the ____________________________ frequency approach. A ____________________________ of a procedure is a process that behaves in the same way as the ___________________________ itself, so that ____________________________ results are produced.

COMPLEMENTARY EVENTS

Sometimes we need to find the probability that an event *A* ________________ ____________________ occur.

DEFINITION

The **complement** of event *A*, denoted by $\bar{A}$, consists of all outcomes in which event *A* does _______ occur.

Example 4: Find the probability that you will select the incorrect answer on a multiple-choice item if you randomly select an answer.

ROUNDING OFF PROBABILITIES

When expressing the value of a probability, either give the _________________ fraction or decimal or round off final results to _________________ significant digits. All digits in a number are __________________________ except for the ________________ that are included for proper placement of the decimal point.

PART 2: BEYOND THE BASICS OF PROBABILITY: ODDS

Expressions of likelihood are often given as _________________, such as 50:1 (or 50 to 1). Because the use of odds makes many ___________________________ difficult, statisticians, mathematicians, and scientists prefer to use __________________________________. The advantage of odds is that they make it easier to deal with money transfers associated with ________________________, so they tend to be used in ____________________, ___________________________, and ______________________.

DEFINITION

The **actual odds against** of event *A* occurring are the ratio ________________, usually expressed in the form of _________________ or ________________, where *a* and *b* are integers having no common factors.

The **actual odds in favor** of event *A* occurring are the ratio _______________, which is the ______________________ of the actual odds against that event.

The **payoff odds** against event *A* occurring are the ratio of ________________ ________________ (if you win) to the amount _______________.

Example 4: Finding Odds in Roulette
A roulette wheel has 38 slots. One slot is 0, another is 00, and the others are numbered 1 through 36, respectively. You place a bet that the outcome is an odd number.

a. What is your probability of winning?

b. What are the actual odds against winning?

c. When you bet that the outcome is an odd number, the payoff odds are 1:1. How much profit do you make if you bet $18 and win?

4.3 ADDITION RULE

DEFINITION

A **compound event** is any event combining ________________ or more ______________________ events.

NOTATION

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

FORMAL ADDITION RULE

The **formal addition rule**:

$P(A \text{ or } B) =$ ____ where $P(A \text{ and } B)$

denotes the probability that _________ and ___________ both occur at the ___________________ time

as an ________________________ in a ___________________ or ______________________________.

INTUITIVE ADDITION RULE

The **intuitive addition rule**: To find $P(A \text{ or } B)$, find the ___________ of the ___________________

of ways that event __________ can occur and the number of ways that event ___________ can occur,

adding in such a way that every ____________________ is counted only ________________.

$P(A \text{ or } B)$ is equal to that ______________, ___________________________ by the total number of

__________________________ in the ________________________ space.

DEFINITION

Events *A* and *B* are **disjoint (aka mutually exclusive)** if they cannot ___________________ at the same

____________________.

COMPLEMENTARY EVENTS

Recall that the complement of event *A* is denoted _____________, and consists of all the

________________________ in which event *A* _____________ ________ occur. An event and its

complement must be ________________________, because it is _______________________________ for

an event and its complement to occur at the same time. Also, we can be sure that *A* either does or does

not occur, which implies that either ___________ or _____________ must occur.

Example 1: Sobriety Checkpoint
When the author observed a sobriety checkpoint conducted by the Dutchess County Sheriff Department, he saw that 676 drivers were screened and 6 were arrested for driving while intoxicated. Based on those results, we can estimate the $P(I) = 0.00888$, where I denotes the event of screening a driver and getting someone who is intoxicated. What does $P(\overline{I})$ denote and what is its value?

RULES OF COMPLEMENTARY EVENTS

$$P(A)+P(\overline{A})=1$$
$$P(\overline{A})=1-P(A)$$
$$P(A)=1-P(\overline{A})$$

Example 2: Use the data in the table below, which summarizes challenges by tennis players (based on the data reported in USA Today). The results are from the first U.S. Open that used the Hawk-Eye electronic system for displaying an instant replay used to determine whether the ball is in bounds or out of bounds. In each case, assume that one of the challenges is randomly selected.

	Was the challenge to the call successful?	
	Yes	**No**
Men	201	288
Women	126	224

a. If *S* denotes the event of selecting a successful challenge, find $P\left(\overline{S}\right)$.

b. If *M* denotes the event of selecting a challenge made by a man, find $P\left(\overline{M}\right)$.

c. Find the probability that the selected challenge was made by a man or was successful.

d. Find the probability that the selected challenge was made by a woman or was successful.

e. Find $P(\text{challenge was made by a man or was not successful})$.

f. Find $P(\text{challenge was made by a woman or was not successful})$.

4.4 MULTIPLICATION RULE: BASICS

NOTATION

$$P\left(A \text{ and } B\right) =$$

$$P\left(B \mid A\right) =$$

DEFINITION

Two events A and B are **independent** if the occurrence of one does not ____________________ the

_________________________ of the occurrence of the other. If A and B are not

____________________________, they are said to be **dependent**.

Example 1: Give an example of

a. Two independent events

b. Two dependent events

FORMAL MULTIPLICATION RULE

The **formal multiplication rule**:

$P(A \text{ and } B) =$ ____.

If A and B are ____________________________ events, $P(B|A)$ is the same as

__________________________.

INTUITIVE ADDITION RULE

When finding the probability that event A occurs in one trial and event B occurs in the next trial,

__________________________ the probability of event A by the probability of event B, but be sure that

the ___________________________ of event B takes into account the previous ____________________

of event A.

Example 2: Use the data in the table below, which summarizes blood groups and Rh types for 100 subjects.

	O	A	B	AB
Rh^+	39	35	8	4
Rh^-	6	5	2	1

a. If 2 of the 100 subjects are randomly selected, find the probability that they are both group O and type Rh^+.

i. Assume that the selections are made with replacement.

ii. Assume that the selections are made without replacement.

b. People with blood that is group O and type Rh^- are considered to be universal donors, because they can give blood to anyone. If 4 of the 100 subjects are randomly selected, find the probability that they are all universal recipients.
   i. Assume that the selections are made with replacement.

   ii. Assume that the selections are made without replacement.

Example 3: Suppose that you are married and want to have 3 children. Assume that the probability for you to give birth to a girl is equal to the probability for you to give birth to a boy, and that you only give birth to one child at a time.

a. Make a tree diagram and list the sample space.

b. What is the probability that you have all girls?

c. What is the probability that you have 2 boys?

d. What is the probability that you have at least one girl?

TREATING DEPENDENT EVENTS AS INDEPENDENT: THE 5% GUIDELINE FOR CUMBERSOME CALCULATIONS

If calculations are very cumbersome and if a ____________________ size is no more than

____________ of the size of the population, treat the selections as being _________________________

(even if the selections are made without __________________________, so they are technically

__________________________).

Example 4: A quality control analyst randomly selects three different car ignition systems from a manufacturing process that has just produced 200 systems, including 5 that are defective.

a. Does this selection process involve independent events?

b. What is the probability that all three ignition systems are good? (Do not treat the events as independent).

c. Use the 5% guideline for treating the events as independent, and find the probability that all three ignition systems are good.

d. Which answer is better: The answer from part (b) or the answer from part (c)? Why?

4.5 MULTIPLICATION RULE: COMPLEMENTS AND CONDITIONAL PROBABILITY

COMPLEMENTS: THE PROBABILITY OF "AT LEAST ONE"

- At least one is equivalent to _______________ or _______________.
- The _______________________of getting at least one item of a particular type is that you get ____________ items of that type.

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

Example 1: Provide a written description of the complement of the following event:
When Brutus asks five different women for a date, at least one of them accepts.

Example 2: If a couple plans to have 8 children what is the probability that there will be at least one girl?

CONDITIONAL PROBABILITY

DEFINITION

A **conditional probability** of an event is a ________________________ obtained with the additional

___________________________ that some other event has already _____________________________.

$P(B|A)$ denotes the _____________ probability of an event *B* occurring, given that event *A* has

already ____________.

$$P(B|A) = \rule{8cm}{0.4pt}$$

INTUITIVE APPROACH TO CONDITIONAL PROBABILITY

The _______________________________ probability of *B* __________________*A* can be found by

__________________________ that event *A* has occurred, and then calculating the probability that

event *B* will ___________________________.

Example 3: Use the table below to find the following probabilities.

	Did the Subject Actually Lie?	
	No (Did Not Lie)	**Yes (Lied)**
Positive test result (Polygraph test indicated that the subject lied)	15 (false positive)	42 (true positive)
Negative test result (Polygraph test indicated that the subject did not lie)	32 (true negative)	9 (false negative)

a. Find the probability of selecting a subject with a positive test result, given that the subject did not lie.

b. Find the probability of selecting a subject with a negative test result, given that the subject lied.

c. Find $P(\text{negative test result} \mid \text{subject did not lie})$.

d. Find $P(\text{subject did not lie} \mid \text{negative test result})$.

e. Are the results from (c) and (d) equal?

Example 4: The Orange County Department of Public Health tests water for contamination due to the presence of *E. coli* bacteria. To reduce the laboratory costs, water samples from six public swimming areas are combined for one test, and further testing is done only if the combined sample fails. Based on past results, there is a 2% chance of finding *E. coli* bacteria in a public swimming area. Find the probability that a combined sample from six public swimming areas will reveal the presence of *E. coli* bacteria.

4.6 COUNTING

FUNDAMENTAL COUNTING RULE

For a ________________________ of two ___________________ in which the first event can occur

________ ways and the second event can occur _________ ways, the events together can occur a total

of ________________ ways.

Example 1: How many different California vehicle license plates (not specialized plates) are possible if the first, fifth, sixth, and seventh digits consist of a number from 1-9, and the second, third, and fourth digits have letters?

NOTATION

The **factorial symbol(!)** denotes the product of decreasing positive whole numbers.

Example 2: Evaluate 5!

FACTORIAL RULE

A collection of ______ different items can be _____________________ in order _____________ in different ways.

Example 3: Find the number of ways that 8 people can be seated at a round table.

PERMUTATIONS RULE (WHEN ITEMS ARE ALL DIFFERENT)

Requirements:

1. There are ________ ___________________________ items available.
2. We select ________ of the ________ items (without replacement).
3. We consider ______________________________ of the same items to be _____________________ sequences. This would mean that ABC is different from CBA and is counted separately.

If the preceding requirements are satisfied, the number of _____________________________ (aka _________________________) of ______ items selected from ________ different available items (without replacement) is

$${}_nP_r = \underline{\hspace{6cm}}$$

Example 4: A political strategist must visit state capitols, but she has time to visit only three of them. Find the number of different possible routes.

PERMUTATIONS RULE (WHEN SOME ITEMS ARE IDENTICAL TO OTHERS)

Requirements:

1. There are ________ items available, and some items are_________________ to others.

2. We select __________ of the _____ items (without replacement).

3. We consider ______________________________ of distinct items to be ______________________ sequences.

If the preceding requirements are satisfied, and if there are _______ alike, _________ alike, ...,

_________ alike, the number of _________________________ or ____________________________ of all

items selected without replacement is

____

Example 5: In a preliminary test of the MicroSort gender-selection method, 14 babies were born and 13 of them were girls.

a. Find the number of different possible sequences of genders that are possible when 14 babies are born.

b. How many ways can 13 girls and 1 boy be arranged in a sequence?

c. If 14 babies are randomly selected, what is the probability that they consist of 13 girls and 1 boy?

d. Does the gender-selection method appear to yield a result that is significantly different from a result that might be expected from random chance?

COMBINATIONS RULE

Requirements:

1. There are ________ __________________________ items available.
2. We select ________ of the ________ items (without replacement).
3. We consider ______________________________ of the same items to be the _________________. This would mean that ABC is the same as CBA.

If the preceding requirements are satisfied, the number of _____________________________ of _____ items selected from ________ different items is

$$_nC_r = \rule{5cm}{0.4pt}$$

Example 6: Find the number of different possible five-card poker hands.

Example 7: The Mega Millions lottery is run in 12 states. Winning the jackpot requires that you select the correct five numbers between 1 and 56, and, in a separate drawing, you must also select the correct single number between 1 and 46. Find the probability of winning the jackpot.

5.2 RANDOM VARIABLES

DEFINITION

A **random variable** is a ___________________________ (typically represented by _____) that has a ______________________ ________________________ value, determined by ____________________, for each ________________________ of a _____________________________.

DEFINITION

A **probability distribution** is a ___________________________ that gives the ____________________ for each value of the ___________________ _____________________________. It is often expressed in the format of a _________________________, ___________________________, or __________________________.

NOTE

If a probability value is very small, such as 0.000000123, we can represent it as 0+ in a table, where 0+ indicates that the probability value is a very small positive number. Why not represent this as 0?

Recall the tree diagram we made for a couple having 3 children:

DEFINITION

A **discrete random variable** has either a ____________________________ number of

_________________________ or a ___________________________ number of values, where

____________________________ refers to the fact that there might be _________________________

many values, but they can be _________________________________ with a _________________________

process, so that the number of values is 0 or 1 or 2 or 3, etc.

A **continuous random variable** has __________________________ many values, and those values can be

associated with ______________________________ on a ____________________________ scale without

_______________ or ______________________________.

Example 1: Give two examples of

a. Discrete random variables

b. Continuous random variables

GRAPHS

There are various ways to graph a __________________________ distribution, but we will consider only

the _____________________________ _____________________________. A probability histogram is

similar to a relative frequency histogram, but the vertical scale shows __________________________

instead of __________________________ frequencies based on actual sample events.

REQUIREMENTS FOR A PROBABILITY DISTRIBUTION

1. $\sum P(x) = 1$ where *x* assumes all possible values. The sum of all probabilities must be __________, but values such as 0.999 or 1.001 are acceptable because they result from ______________________ errors.
2. $0 \le P(x) \le 1$ for every individual value of *x*.

MEAN, VARIANCE, AND STANDARD DEVIATION

1. $\mu = \sum\left[x \cdot P(x)\right]$
2. $\sigma^2 = \sum\left[(x-\mu)^2 \cdot P(x)\right]$
3. $\sigma^2 = \sum\left[x^2 \cdot P(x)\right] - \mu^2$
4. $\sigma = \sqrt{\sum\left[x^2 \cdot P(x)\right] - \mu^2}$

ROUND-OFF RULE FOR μ, σ, and σ^2

Round results by carrying one more ________________________ place than the number of decimal places used for the ________________________ variable ________. If the values of _______ are _________________________, round to one decimal place.

IDENTIFYING UNUSUAL RESULTS WITH THE RANGE RULE OF THUMB

The range rule of thumb may be helpful in ___________________________ the value of a __________________________ _________________________. According to the _________________ ____________________ of __________________, most values should lie within __________ standard deviations of the ____________________; it is ________________________ for a value to differ from

the mean by ______________ than ________ standard deviations.

Maximum usual value = _____ + ______

Minimum usual value = _____ - ______

IDENTIFYING UNUSUAL RESULTS WITH PROBABILITIES

x successes among *n* trials is an unusually high number of successes if the _________________________

of _____ or more ______________________ is unlikely with a probability of _____________ or

_______________.

x successes among *n* trials is an unusually low number of successes if the _________________________

of _____ or fewer ______________________ is unlikely with a probability of _____________ or

_______________.

RARE EVENT RULE FOR INFERENTIAL STATISTICS

If, under a given ___________________________, the probability of a particular ___________________

event is extremely small, we conclude that the ___________________________ is probably not

__________________________.

Example 2: Based on information from MRINetwork, some job applicants are required to have several interviews before a decision is made. The number of required interviews and the corresponding probabilities are: 1 (0.09); 2 (0.31); 3 (0.37); 4 (0.12); 5 (0.05); 6 (0.05).

a. Does the given information describe a probability distribution?

b. Assuming that a probability distribution is described, find its mean and standard deviation.

c. Use the range rule of thumb to identify the range of values for usual numbers of interviews.

d. Is it unusual to have a decision after just one interview. Explain.

DEFINITION

The **expected value** of a ______________________ random variable is denoted by _______, and it represents the ____________________ ________________ of the ____________________. It is obtained by finding the value of $\sum\left[x\cdot P(x)\right]$.

$$E=\sum\left[x\cdot P(x)\right]$$

Example 3: There is a 0.9968 probability that a randomly selected 50-year old female lives through the year (based on data from the U.S. Department of Health and Human Services). A Fidelity life insurance company charges $226 for insuring that the female will live through the year. If she does not survive the year, the policy pays out $50,000 as a death benefit.

a. From the perspective of the 50-year-old female, what are the values corresponding to the two events of surviving the year and not surviving?

b. If a 50-year-old female purchases the policy, what is her expected value?

c. Can the insurance company expect to make a profit from many such policies? Why?

5.3 BINOMIAL PROBABILITY DISTRIBUTIONS

DEFINITION

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

1. The procedure has a ________________ __________________ of trials.
2. The trials must be ___________________________.
3. Each trial must have all _____________________ classified into _________ __________________ (commonly referred to as ___________________ and ________________________).
4. The probability of a ______________________ remains the _____________ in all trials.

NOTATION FOR BINOMIAL PROBABILITY DISTRIBUTIONS

S and F (success and failure) denote the two possible categories of outcomes

$P(S) = p$

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

n

x

p

q

$P(x)$

Example 1: A psychology test consists of multiple-choice questions, each having four possible answers (a, b, c, and d), one of which is correct. Assume that you guess the answers to six questions.

a. Use the multiplication rule to find the probability that the first two guesses are wrong and the last four guesses are correct.

b. Beginning with WWCCCC, make a complete list of the different possible arrangements of 2 wrong answers and 4 correct answers, then find the probability for each entry in the list.

c. Based on the preceding results, what is the probability of getting exactly 4 correct answers when 6 guesses are made?

d. Now use the Binomial Probability Formula to find probability of getting exactly 4 correct answers when 6 guesses are made.

BINOMIAL PROBABILITY FORMULA

$$P(x)=\frac{n!}{(n-x)!x!}\cdot p^x \cdot q^{n-x} \qquad \text{for } x=0,1,2,\ldots,n$$

Example 2: Assuming the probability of a pea having a green pod is 0.75, use the binomial probability formula to find the probability of getting exactly 2 peas with green pods when 5 offspring peas are generated.

5.4 MEAN, VARIANCE, AND STANDARD DEVIATION FOR THE BINOMIAL DISTRIBUTION

Any Discrete pdf	**Binomial Distributions**
1. $\mu = \sum\left[x \cdot P(x)\right]$	1. $\mu = np$
2. $\sigma^2 = \sum\left[(x-\mu)^2 \cdot P(x)\right]$	
3. $\sigma^2 = \sum\left[x^2 \cdot P(x)\right] - \mu^2$	2. $\sigma^2 = npq$
4. $\sigma = \sqrt{\sum\left[x^2 \cdot P(x)\right] - \mu^2}$	3. $\sigma = \sqrt{npq}$

RANGE RULE OF THUMB

Maximum usual value:

Minimum usual value:

Example 1: Mars, Inc. claims that 24% of its M&M plain candies are blue. A sample of 100 M&Ms is randomly selected.

a. Find the mean and standard deviation for the numbers of blue M&Ms in such groups of 100.

b. Data Set 18 in Appendix B consists of 100 M&Ms in which 27 are blue. Is this result unusual? Does it seem that the claimed rate of 24% is wrong?

Example 2: In a study of 420,095 cell phone users in Denmark, it was found that 135 developed cancer of the brain or nervous system. If we assume that the use of cell phones has no effect on developing such cancer, then the probability of a person having such a cancer is 0.000340.

a. Assuming that cell phones have no effect on developing cancer, find the mean and standard deviation for the numbers of people in groups of 420,095 that can be expected to have cancer of the brain or nervous system.

b. Based on the results from part (a), is it unusual to find that among 420,095 people, there are 135 cases of cancer of the brain or nervous system? Why or why not?

c. What do these results suggest about the publicized concern that cell phones are a health danger because they increase the risk of cancer of the brain or nervous system?

6.2 THE STANDARD NORMAL DISTRIBUTION

UNIFORM DISTRIBUTIONS

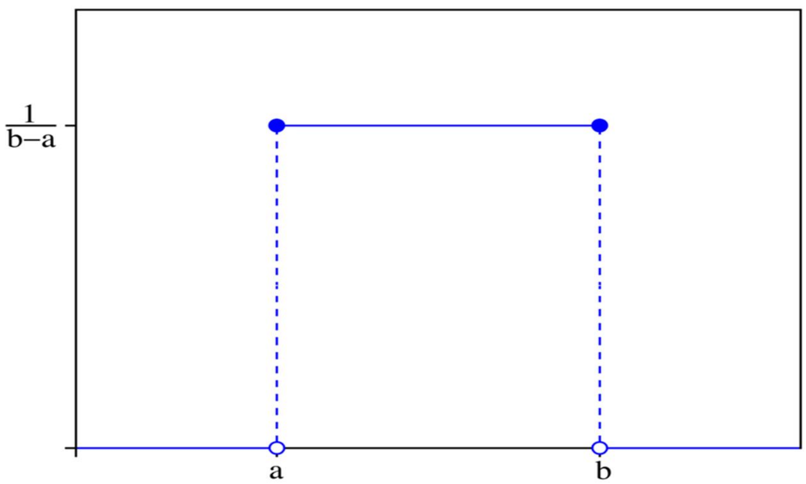
The ______________________ ________________________ allows us to see two very important properties:

1. The _______________ under the _______________ of a ______________ distribution is equal to ____________.

2. There is a _________________________ between _______________ and __________________________ (or _______________________ frequency), so some ________________________ can be found by _________________ the corresponding ________________.

DEFINITION

A ________________________ _________________ __________________ has a **uniform distribution** if its values are spread ______________________ over the ____________________ of __________________________. The graph of a uniform distribution results in a ______________________________ shape.



Example 1: The Newport Power and Light Company provides electricity with voltage levels that are uniformly distributed between 123.0 volts and 125.0 volts. That is, any voltage amount between 123.0 volts and 125.0 volts is possible, and all of the possibilities are equally likely. If we randomly select one of the voltage levels and represent its value by the random variable *x*, then *x* has a distribution that can be graphed.

a. Sketch a graph of the uniform distribution of voltage levels.

b. Find the probability that the voltage level is greater than 124.0 volts.

c. Find the probability that the voltage level is less than 123.5 volts.

d. Find the probability that the voltage level is between 123.2 volts and 124.7 volts.

e. Find the probability that the voltage level is between 124.1 volts and 124.5 volts.

The graph of a probability distribution, such as part (a) in the previous example is called a

______________________ ________________________. A density curve must satisfy the following two requirements.

1. The total ____________________ under the _________________ must equal __________.
2. Every point on the ________________ must have a vertical ________________ that is

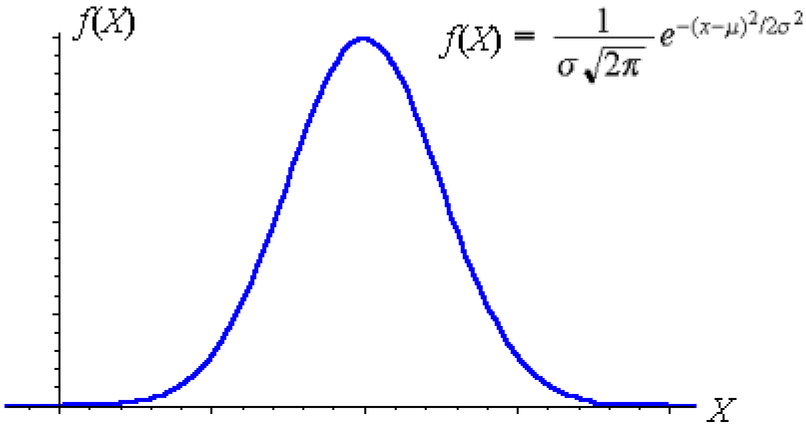
    ____________ or ____________________.

DEFINITION

The **standard normal distribution** is a ___________________________ ________________________

___________________________ with ______________ and ______________. The total

______________ under its ___________________ ________________ is equal to __________.



FINDING PROBABILITIES WHEN GIVEN *z* SCORES

Using table ________, we can find ______________ or ___________________ for many different __________________. Such areas can also be found using a _______________________ ________________________. When using Table A-2, it is essential to understand these points:

1. Table A-2 is designed only for the ___________________ _____________ distribution, which has a mean of ____________ and a standard deviation of ______.
2. Table A-2 is on ________ pages, with one page for _______________ ______________ and the other page for _______________ ____________.
3. Each value in the body of the table is a ____________________ _________ from the _______________ up to a ________________ _______________ above a specific __________________.
4. When working with a __________________, avoid confusion between _________________ and _________________.

   z score: ______________________ along the _____________________ scale of the standard normal distribution; refer to the _____________________ column and _______________ row of Table A-2.

Area: ___________________ under the ___________________; refer to the values in the

__________________ of Table A-2.

5. The part of the _______________ denoting ______________________ is found across the

___________ ___________ of Table A-2.

NEGATIVE *z* Scores

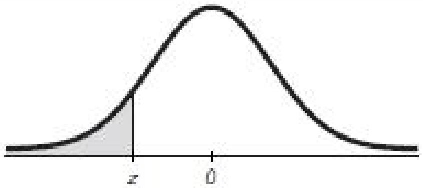


TABLE A-2 Standard Normal (*z*) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.50 and lower	.0001									
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	[illegible]	[illegible]	[illegible]	.0023	.0023	.0022	.0021	[illegible]	.0020	.0019

POSITIVE *z* Scores

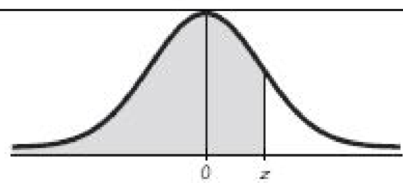


TABLE A-2 (continued) Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.61
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.754
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.813
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.862
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.91
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.93

NOTATION

$$P(a < z < b)$$

$$P(z > a)$$

$$P(z < a)$$

Example 2: Assume that thermometer readings are normally distributed with a mean of 0ºC and a standard deviation of 1.00 ºC. A thermometer is randomly selected and tested. In each case, draw a sketch and find the probability of each reading. The given values are in Celsius degrees.

a. Less than -2.75

b. Greater than 2.33

c. Between 1.00 and 3.00

e. Greater than 3.68

d. Between -2.87 and 1.34

USING THE TI-84

FINDING *z* SCORES WITH KNOWN AREAS

1. Draw a bell-shaped curve and _____________________ the _____________ under the _________________ that _______________________ to the ________________ probability. If that region is not a _________________ region from the ________________, work instead with a known region that is a cumulative region from the ______________.

2. Using the ______________________ __________________ from the ____________, locate the __________________ probability in the _______________ of Table A-2 and identify the ____________________ _______________.

NOTATION

The expression z_{α} denotes the *z* score with an area of _________ to its ____________.

Example 3: Find the value of $z_{.075}$.

Example 4: Assume that thermometer readings are normally distributed with a mean of 0ºC and a standard deviation of 1.00 ºC. A thermometer is randomly selected and tested. In each case, draw a sketch and find the probability of each reading. The given values are in Celsius degrees.

a. Find the 1st percentile.

b. If 0.5% of the thermometers are rejected because they have readings that are too low and another 0.5% are rejected because they have readings that are too high, find the two readings that are cutoff values separating the rejected thermometers from the others.

6.3 APPLICATIONS OF NORMAL DISTRIBUTIONS

TO STANDARDIZE VALUES USE THE FOLLOWING FORMULA:

STEPS FOR FINDING AREAS WITH A NONSTANDARD NORMAL DISTRIBUTION:

1. Sketch a ____________________ curve, label the _____________________ and the specific

   ________________, then __________________ the region representing the desired

   _________________________.

2. For each relevant value *x* that is a _______________________ for the shaded region, convert

   the relevant value to a standardized _____________.

3. Refer to table ____________ or use a _____________________ to find the _________________ of the shaded region.

Example 1: Assume that adults have IQ scores that are normally distributed with a mean of 100 and a standard deviation of 15.

a. Find the probability that a randomly selected adult has an IQ that is less than 115.

b. Find the probability that a randomly selected adult has an IQ greater than 131.5 (the requirement for the Mensa organization).

c. Find the probability that a randomly selected adult has an IQ between 90 and 110 (referred to as the normal range).

d. Find the probability that a randomly selected adult has an IQ between 110 and 120 (referred to as bright normal).

e. Find P_{30}, which is the IQ score separating the bottom 30% from the top 70%.

f. Find the first quartile Q_1, which is the IQ score separating the bottom 25% from the top 75%.

g. Find the third quartile Q_3, which is the IQ score separating the top 25% from the others.

h. Find the IQ score separating the top 37% from the others.

FINDING VALUES FROM KNOWN AREAS

1. Don't confuse ___________________ and ______________. Remember, ______________ are ___________________ along the ________________ scale, but ____________________ are ____________________ under the ______________.
2. Choose the correct ________________ of the ___________________. A value separating the top 10% from the others will be located on the ______________ side of the graph, but a value separating the bottom 10% will be located on the __________________ side of the graph.
3. A ________________ must be _____________________ whenever it is located in the ______________ half of the ___________________ distribution.

4. Areas (or ____________________) are ___________________ or ________ values, but they are never __________________.

Always use graphs to ___________________________!!!

STEPS FOR FINDING VALUES USING TABLE A-2:

1. Sketch a ___________________ distribution curve, enter the given ______________________ or _________________________ in the appropriate _________________ of the ____________________, and identify the _________________ being sought.
2. Use Table A-2 to find the ________________ corresponding to the ______________________ ___________________ area bounded by ____. Refer to the _______________ of Table A-2 to find the ______________ area, then identify the corresponding _______________.
3. Solve for _____ as follows:
4. Refer to the ________________ of the ________________ to make sure that the solution makes ______________________!

Example: Engineers want to design seats in commercial aircraft so that they are wide enough to fit 99% of all males. Men have hip breadths that are normally distributed with a mean of 14.4 inches and a standard deviation of 1.0 inch. Find the hip breadth for men that separates the smallest 99% from the largest 1 % (aka P_{99}).

6.5 THE CENTRAL LIMIT THEOREM

Key Concept…

In this section, we introduce and apply the ______________________ ______________________

________________________. The central limit theorem tells us that for a

__________________________ with ____________ distribution, the ______________________

of the _______________________ ____________________ approaches a __________________

_________________________ as the sample size __________________________. This means

that if the sample size is _____________ enough, the _________________________ of

________________ ____________________ can be approximated by a ____________________

_______________________, even if the original population is _____________ normally

distributed. If the original population has ___________________ _______ and _______________

_____________________________ _______, the ______________ of the ____________________

____________________ will also be _____, but the _______________________

_________________________ of the _____________________ _______________________ will

be _________, where ____ is the ______________________ size.

It is essential to know the following principles:

1. For a ____________________________ with any _____________________, if

   _______________, then the sample means have a ____________________________ that can

   be approximated by a ____________________________distribution, with mean _____ and

   standard deviation ___________.

2. If _____________ and the original population has a _______________ distribution, then the

   __________________ ____________________ have a ________________________

   distribution with mean ____ and standard deviation ___________.

3. If _____________ and the original population does not have a _____________________

distribution, then the methods of this section ___________ ____________ ______________.

NOTATION

If all possible _______________________ _____________________ of size ____ are selected from a population with mean ____ and standard deviation _____, the mean of the ____________________ ___________________ is denoted by______, so ______________ = ______. Also, the standard deviation of the sample means is denoted by ______, so ____________ = ____________. _________ is called the __________________ ________________ of the mean.

APPLYING THE CENTRAL LIMIT THEOREM

Example 1: Assume that SAT scores are normally distributed with mean $\mu = 1518$ and standard deviation $\sigma = 325$.

a. If 1 SAT score is randomly selected, find the probability that it is between 1440 and 1480.

b. If 16 SAT scores are randomly selected, find the probability that they have a mean between 1440 and 1480.

c. Why can the central limit theorem be used in part (b) even though the sample size does not exceed 30?

Example 2: Engineers must consider the breadths of male heads when designing motorcycle helmets. Men have head breadths that are normally distributed with a mean of 6.0 inches and a standard deviation of 1.0 inch.

a. If one male is randomly selected, find the probability that his head breadth is less than 6.2 inches.

b. The Safeguard Helmet company plans an initial production run of 100 helmets. Find the probability that 100 randomly selected men have a mean head breadth of less than 6.2 inches.

c. The production manager sees the result from part (b) and reasons that all helmets should be made for men with head breadths less than 6.2 inches, because they would fit all but a few men. What is wrong with that reasoning?

7.2 ESTIMATING A POPULATION PROPORTION

DEFINITION

A **point estimate** is a _______________________ value (or _______________) used to _________________________ a _______________________ parameter.

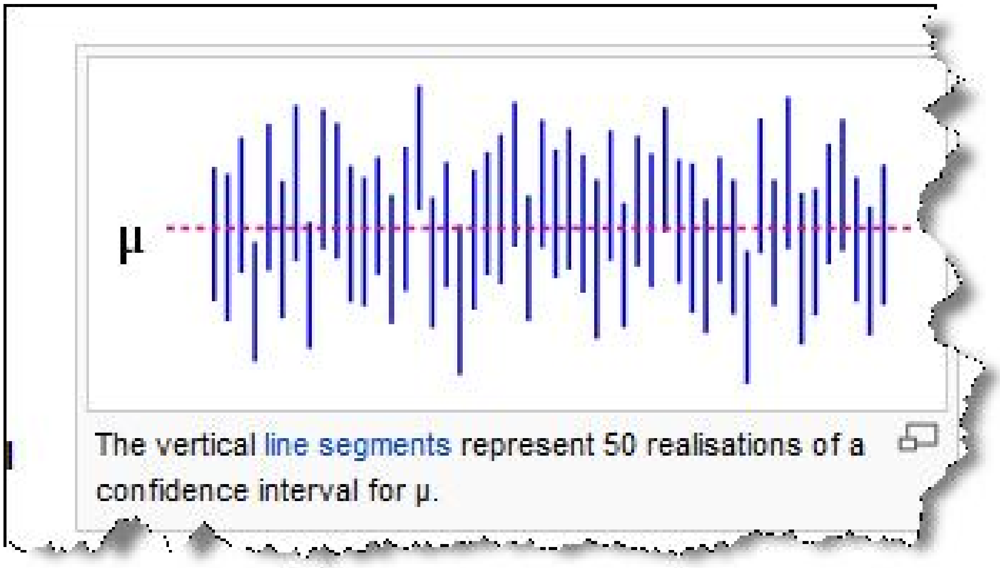
The _______________________ ____________________ ______ is the best ___________________ ____________________ of the __________________ _____________________ _____.

DEFINITION

A **confidence interval (aka _______________________ ___________________)** is a _____________________ (or an __________________________) of _______________ used to _______________________ the ____________ value of a _______________________ ___________________________. A ____________________ _______________________ is often abbreviated as CI.

DEFINITION

The **confidence level** is the ___________________________ ______________ (often expressed as the equivalent percentage value) that the _______________ _______________________ actually does ___________________ the _______________________ _____________________, assuming that the ___________________________ process is _______________________ a _________ number of times. (The _______________________ ____________________ is also called the _________________ of ___________________, or the _______________ _________________________).



The vertical line segments represent 50 realisations of a confidence interval for μ.

CRITICAL VALUES

The methods of this section (and many others) include a reference to a __________________

__________________ that can be used to ______________ between _________________

____________________ that are ______________ to ________________ and those that are

_______________ to _____________. Such a _________________ is called a _________________

_______________.

DEFINITION

A **critical value** is the ____________________ on the ______________________ separating

_______________ ______________________ that are likely to occur from those that are

___________________ to occur. The number ______ is a ___________________ __________________

that is a _______________ with the property that it ____________________ an _______________ of

___________ in the ________________ tail of the ______________________ ________________

distribution.

Example 1: An interesting and popular hypothesis is that individuals can temporarily postpone their death to survive a major holiday or important event such as a birthday. In a study of this phenomenon, it was found that in the week before and the week after Thanksgiving, there were 12,000 total deaths, and 6062 of them occurred in the week before Thanksgiving.

a. What is the best point estimate of the proportion of deaths in the week before Thanksgiving to the total deaths in the week before and the week after Thanksgiving?

b. Construct a 95% confidence interval estimate of the proportion of deaths in the week before Thanksgiving to the total deaths in the week before and the week after Thanksgiving.

c. Based on the result, does there appear to be any indication that people can temporarily postpone their death to survive the Thanksgiving holiday? Why or why not?

Example 2: In a study of 420,095 cell phone users in Denmark, it was found that 135 developed cancer of the brain or nervous system. Prior to this study of cell phone use, the rate of such cancer was found to be 0.0340% for those not using cell phones.

a. Use the sample data to construct a 95% confidence interval estimate of the percentage of cell phone users who develop cancer of the brain or nervous system.

b. Do cell phone users appear to have a rate of cancer of the brain or nervous system that is different from the rate of such cancer among those using cell phones? Why or why not?

DEFINITION

When the data from a ____________________ ___________________ sample are used to

______________________ a ________________ ______________ ___, the **margin of error**, denoted

by ____, is the ____________________ likely _____________________ (with probability __________)

between the ___________________ _____________________ ____________________ _____

and the _____________ _______________ of the ______________________ _____________________

_____. The _____________________ of ___________ ____ is also called the _____________________

__________________ of the _____________________ and can be found by _______________________

the ____________________ value and the ______________________ ___________________ of

_____________________ ______________________ as shown in the formula below:

ROUND-OFF RULE FOR CONFIDENCE INTERVAL ESTIMATES OF *p*

Round the confidence interval __________________ for _____ to ___________ _____________________

_____________________.

DETERMINING SAMPLE SIZE

Suppose we want to __________________ data in order to _______________ some _______________

_______________________. How do we know how many sample items must be obtained? If we solve the

________________ for _________________ of _________________ _____ for ____, we get the

first formula below. Note that this formula requires ______. If no such estimate is known, we replace

_________ by __________ and replace _________ by ______, which is shown in the second formula.

When an estimate ____ is known:

When no estimate ____ is known:

ROUND-OFF RULE FOR DETERMINING SAMPLE SIZE

If the computed sample size _____ is not a _______________ _____________, round the value of

____ _____ to the next _____________ ______________ number.

Example 3: As your text was being written, former NYC mayor Rudolph Giuliani announced that he was a candidate for the presidency of the United States. If you were a campaign worker and needed to determine the percentage of people that recognized his name, how many people should you have surveyed to estimate that percentage? Assume that you wanted to be 95% confident that the sample percentage was in error by no more than 2 percentage points, and also assume that a recent survey indicated that Giuliani's name is recognized by 10% of all adults (based on data from a Gallup poll).

7.3 ESTIMATING A POPULATION MEAN: SIGMA KNOWN

Key Concept…

In this section we present methods for ________________________ a _________________

____________________. In addition to knowing the values of the ____________________ data or

_____________________, we must also know the value of the ____________________

________________ _______________________, _____. Here are three concepts that should be

learned in this section.

1. We should know that the ________________ ________________ ______ is the best ______________________ ______________________ of the _________________________ ______________________ _____.

2. We should learn how to use __________________ ________________ to construct a ___________________ ______________________ for _________________________ the value of a _______________________ _______________, and we should know how to ___________________ such _________________________ _______________________.

3. We should develop the ability to _______________________ the ____________ ___________ necessary to _______________________ a _________________________ ___________________.

POINT ESTIMATE

The ________________ _______________ ______ is an _________________ estimator of the ______________________ ________________ _____, and for many populations, ____________________ ________________ tend to _______ less than other measures of ___________________, so the _______________ __________________ ______, is usually the best ______________________ _________________ of the __________________ ______________ _____.

KNOWLEDGE OF SIGMA

The methods of this section require that we know _____, but in 7.4 we will learn methods to _____________________ a _____________________ __________ without knowledge of the value of ______.

NORMALITY REQUIREMENT

The population must either be _______________________ _________________ or _____________. If _____________, the population does not need to have a ____________________ that is __________________ _________________ as long as it is ______________________ _________________. As long as there are no _____________________ and if a _____________________ of the ________________ ________________ is not __________________ different from being ____________________________, the ________________ requirement is satisfied.

SAMPLE SIZE REQUIREMENT

The ______________________ sample size actually depends on how much the ____________________ ______________________ departs from a ________________ ______________________. Sample sizes of __________ to __________ are sufficient if the population has a ______________________ that is not far from __________________, but some other populations have _____________________ that are extremely far from ________________ and ____________________ _______________ greater than __________ might be necessary.

CONFIDENCE LEVEL

The ______________________ ______________________ is associated with a ___________________ ___________________, such as ____________ or ________________. The __________________ _______________ gives us the ___________________ _____________ of the ___________________ used to construct the confidence interval. Remember the _____ is the ______________________ of the _________________ __________________.

Example 1: Find the indicated critical value $z_{\alpha/2}$.

a. Find the critical value that corresponds to a 98% confidence level.

b. $\alpha = .04$

PROCEDURE FOR CONSTRUCTING A CONFIDENCE INTERVAL FOR μ WITH KNOWN σ.

1. Verify that the _________________________ are ______________________.

2. Refer to table ____________ or use ______________________ to find the __________________ ____________________ __________ that corresponds to the desired _______________________ ____________________.

3. Evaluate the ____________________ of ________________ _________________________.

4. Using the value of the ________________________ __________________ of __________________ ____ and the value of the __________________ ____________________ _____, find the values of the __________________ ________________________ ___________________: _________________ and ___________________. Substitute those values in the _______________ _________________ for the _________________ ________________: ________________________ or ___________________________ or ___________________________.

5. Round the resulting values by using the following round-off rule.

ROUND-OFF RULE FOR CONFIDENCE INTERVALS USED TO ESTIMATE μ

1. When using the ________________________ set of _______________ to ______________________ a confidence __________________, round the ________________________ _____________________ _______________ to ____________ ______________ __________________ place than is used for the ______________________ set of data.
2. When the ____________________ set of data is ____________________ and only the ____________________ ______________________ (________) are used, round the _______________________ ______________________ limits to the same number of digits as the _________________ mean.

Example 2: A simple random sample of 40 salaries of NCAA football coaches has a mean of $415,953. Assume that $\sigma = \$463,364$.

a. Find the best point estimate of the mean salary of all NCAA football coaches.

b. Construct a 95% confidence interval estimate of the mean salary of an NCAA football coach.

c. Does the confidence interval contain the actual population mean of $474,477?

Example 3: Polling organizations typically generate the last digits of telephone numbers so that people with unlisted numbers are included. Listed below are digits randomly generated by STATDISK. Such generated digits are from a population with a standard deviation of 2.87.

1 1 7 0 7 4 5 1 7 6

a. Use the methods of this section to construct a 95% confidence interval estimate of the mean of all such generated digits.

b. Are the requirements for the methods of this section all satisfied? Does the confidence interval from part (a) serve as a good estimate for the population mean? Explain.

FINDING THE SAMPLE SIZE REQUIRED TO ESTIMATE A POPULATION MEAN

Objective:

Notation:

Requirements:

ROUND-OFF RULE FOR SAMPLE SIZE *n*

If the ______________________ sample size ___ is _________ a ________ _________________, round

the value of ____ _____ to the next _____________ _______________ ________________.

Example 4: A researcher wants to estimate the mean grade point average of all current college students in the United States. She has developed a procedure to standardize scores from colleges using something other than a scale from 0 and 4. How many grade point averages must be obtained so that the sample mean is within 0.1 of the population mean. Assume that a 90% confidence level is desired. Also assume that a pilot study showed that the population standard deviation is estimated to be 0.88.

7.4 ESTIMATING A POPULATION MEAN: SIGMA NOT KNOWN

Key Concept…

In this section, we present methods for _________________________ a ___________________

________________________ when the population ____________________ __________________

_____ is not known. With _________ unknown, we use the ___________________ ___

_________________________ instead of a ______________________ _____________________,

assuming the relevant __________________________ are satisfied. The ____________________

______ _________________________ was developed by William Gosset (1876-1937). William

Gosset was a Guinness Brewery employee. He needed a distribution that could be used with small

samples. The brewery where he worked did not the publication of research results so he

published under the pseudonym "_____________________________". In real circumstances,

_________ is typically ______________, which makes the methods of this section

_______________________ and ________________________.

POINT ESTIMATE

The ________________ _______________ _____ is an _________________ estimator of the

_____________________ ________________ ____.

STUDENT *t* DISTRIBUTION

If a population has a ____________________________ distribution, then the distribution

is a ________________________ __________ _________________________ for

all samples of size ____. A _________________ _____ ____________________ is referred to as a

____ ______________________________. Because we _______ __________ know the value of the

________________________ __________________ _______________________ _____, we

_______________ it with the value of the ___________________ ______________________

__________________________ _____, but this introduces another source of __________________,

especially with _________________ ________________. In order to maintain a desired

_________________________ _______________, we compensate for this additional unreliability by

making the ________________ _________________________ _________________: we use

______________ _______________ ______ that are ______________________ than the

_________________________ ____________________ of _______ from the ____________________

________________________. A _________________ _______________ of _______ can be found

using ______________________ or ____________________ ______.

DEFINITION

The number of **degrees of freedom** for a collection of __________________ __________________ is the ____________________ of ________________ __________________ that can _______________ after certain restrictions have been _______________________ on all data values. The number of ______________________ of _____________________ is often abbreviated as _______.

For example: If 10 students have quiz scores with a mean of 80, we can freely assign values to the first ____________ scores, but the _____________ score is then ____________________________. The ______________ of the 10 scores must be _______________so the _______________ score must be _________ _______________ the ______________ of the ______________ ___________ scores. Because the first 9 scores can be _________________________ selected to any values, we say there are _____________ ___________________ of _______________________ ___________________________. For the applications of this section, the number of degrees of freedom is simply the __________________ ____________ ______________ _________.

Example 1: A sample size of 21 is a simple random sample selected from a normally distributed population. Find the critical value $t_{\alpha/2}$ corresponding to a 95% confidence level.

PROCEDURE FOR CONSTRUCTING A CONFIDENCE INTERVAL FOR μ WITH UNKNOWN σ.

1. Verify that the ________________________ are _____________________.
2. Using ______________ _______________ of __________________, refer to table ____________ or use _____________________ to find the __________________ ____________________ __________ that corresponds to the desired ______________________ ____________________. For the _______________________ _________________________, refer to the "________________ in ____________ _________________".
3. Evaluate the ____________________ of _______________ _________________________.
4. Using the value of the _______________________ _________________ of ________________ ____ and the value of the _________________ ____________________ _____, find the values of the _________________ ________________________ __________________: ________________ and __________________. Substitute those values in the ______________ ___________________ for the __________________ _________________.
5. Round the resulting values by using the following round-off rule.

ROUND-OFF RULE FOR CONFIDENCE INTERVALS USED TO ESTIMATE μ

1. When using the _______________________ set of ______________ to ____________________ a confidence _________________, round the _______________________ _________________ ______________ to ___________ _____________ _________________ place than is used for the ____________________ set of data.
2. When the ____________________ set of data is ____________________ and only the _______________ _______________ (_______) are used, round the ____________________ ______________________ limits to the same number of digits as the ________________ mean.

Example 2: In a study designed to test the effectiveness of acupuncture for treating migraine, 142 subjects were treated with acupuncture and 80 subjects were given a sham treatment. The numbers of migraine attacks for the acupuncture treatment group had a mean of 1.8 and a standard deviation of 1.4. The numbers of migraine attacks for the sham treatment group had a mean of 1.6 and a standard deviation of 1.2.

a. Construct a 95% confidence interval estimate of the mean number of migraine attacks for those treated with acupuncture.

b. Construct a 95% confidence interval estimate of the mean number of migraine attacks for those given a sham treatment.

c. Compare the two confidence intervals. What do the results suggest about the effectiveness of acupuncture?

IMPORTANT PROPERTIES OF THE STUDENT *t* DISTRIBUTION

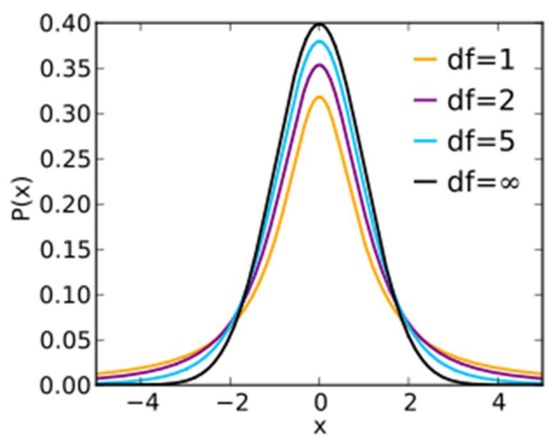
1. The Student *t* distribution is _____________________ for different __________________ ________________.

2. The Student *t* distribution has the ___________ general ________________ _________________ as the ____________________ ___________________ distribution, but it reflects the greater _____________________ (with ___________________ distributions) that is expected of ______________ ____________________.

3. The Student *t* distribution has a mean of _________ (just as the _________________ ___________________ distribution has a mean of _________).

4. The standard _____________________ of the Student *t* distribution ___________ with the ________________ size, but is ________________ than ______ (unlike the ________________________ _________________ distribution, which has _____________).

5. As the _________________ _______________ _____________________, the Student *t* distribution gets _____________________ to the ________________ ____________________ ________________________.



CHOOSING THE APPROPRIATE DISTRIBUTION

It is sometimes difficult to decide whether to use the __________________ __________________

____ ______________________ or the __________________________ _________

______________________.

METHOD	CONDITIONS
Use normal (z) distribution	σ ____________ and ______________ distributed population or σ known and ________________
Use *t* distribution	σ ___________ _______________ and __________________ distributed population or σ _____________ ________________ and _______________
Use a nonparametric method or bootstrapping	Population is ______________ __________________ distributed and _____________

Example 3: Choosing distributions. You plan to construct a confidence interval for the population mean μ. Use the given data to determine whether the margin of error *E* should be calculated using a critical value of $z_{\sigma/2}$ from the normal distribution, $t_{\sigma/2}$ from a *t* distribution, or neither (methods of this chapter cannot be used).

a. $n=7,\ \bar{x}=80,\ s=8$, and the population has a very skewed distribution

b. $n=150,\ \bar{x}=23.5,\ \sigma=0.2$, and the population has a skewed distribution

c. $n=10,\ \bar{x}=65,\ s=12$, and the population has a normal distribution

d. $n=13,\ \bar{x}=5,\ \sigma=3$, and the population has a normal distribution

e. $n=92,\ \bar{x}=20.7,\ s=2.5$, and the population has a skewed distribution

FINDING A POINT ESTIMATE AND *E* FROM A CONFIDENCE INTERVAL

The ________________ ____________________ _______ is the value _________________________ between the __________________________ _______________________ ________________________.

The ______________ of _____________________ ________ is _____________ ________________ the ________________________ between those ______________________.

Point estimate of μ :

Margin of error:

USING CONFIDENCE INTERVALS TO DESCRIBE, EXPLORE, OR COMPARE DATA

In some cases, we might use a _____________________ _____________________ to achieve an ultimate goal of ______________________ the ________________ of a ____________________________ ___________________________. In other cases, _______________________ ______________________ might be among the different _____________________ used to _____________________, ____________________, or ________________________ data sets. When two or more data sets have ______________________________ confidence intervals, one could _____________________________ conclude that there does not appear to be a significant difference between the estimated ______________________.

TI-83/84 PLUS

Example 4: In a sample of seven cars, each car was tested for nitrogen-oxide emissions (in grams per mile) and the following results were obtained: 0.06, 0.11, 0.16, 0.15, 0.14, 0.08, 0.15 (based on data from the EPA).

a. Assuming that this sample is representative of the cars in use, construct a 98% confidence interval estimate of the mean amount of nitrogen-oxide emissions for all cars.

b. If the EPA requires that nitrogen-oxide emissions be less than 0.165 g/mi, can we safely conclude that this requirement is being met?

Example 5: Listed below are 12 lengths (in minutes) of randomly selected movies from Data Set 9 in Appendix B.

110 96 125 94 132 120 136 154 149 94 119 132

a. Construct a 99% confidence interval estimate of the mean length of all movies.

b. Assuming that it takes 30 minutes to empty a theater after a movie, clean it, allow time for the next audience to enter, and show previews, what is the minimum time that a theater manager should plan between start times of movies, assuming that this time will be sufficient for typical movies?

8.1 REVIEW AND PREVIEW

DEFINITION

In statistics, a **hypothesis** is a _______________ or _______________________ about a

_______________________ of the _______________________________.

A **hypothesis test (aka test of significance)** is a ________________________ for testing a

_______________________ about a _________________________ of a __________________________.

8.2 BASICS OF HYPOTHESIS TESTING

PART 1: BASICS CONCEPTS OF HYPOTHESIS TESTING

The methods presented in this chapter are based on the _________ ___________ __________ for

_____________________ _____________________.

RARE EVENT RULE FOR INFERENTIAL STATISTICS

If, under a given assumption, the ___________________________ of a particular observed is extremely

_________________________, we conclude that the ______________________________ is probably not

_______________________.

WORKING WITH THE STATED CLAIM: NULL AND ALTERNATIVE HYPOTHESES

The **null hypothesis** denoted by ________ is a _____________________ that the value of a

________________________ ______________________________ is __________________ to some

__________________ value. The term ________ is used to _______________________ ________

_________________ or _______ ____________________ or _______ ____________________________.

The **alternative hypothesis** denoted by _________ or ________ or ________ is the __________________

that the ______________________ has a value that somehow _________________ from the

___________ ______________________.

For the methods of this chapter, the ______________________ form of the ____________________

_____________________ must use one of these symbols: _____, _____, _____.

IDENTIFYING ______ AND ______

START

1
- Identify the specific ____________ or__________________ to be tested
  - Express it in __________________ form

2
- Give the symbolic form that must be __________ when the __________________ ______________ is __________________

3
- Using the two ________________________ expressions obtained so far, identify the ______________ ____________________ ______ and the _____________________ _________________________ _________
  - _______ is the symbolic expression that _________ __________ contain ______________________
  - _______ is the symbolic expression that the _________________________ ___________________the ______________ value being __________________________

Example 1: Examine the given statement, then express the null hypothesis and the alternative hypothesis in symbolic form.

a. The majority of college students have credit cards.

b. The mean weight of plastic discarded by households in one week is less than 1 kg.

CONVERTING SAMPLE DATA TO A TEST STATISTIC

Test statistic for proportion:

Test statistic for mean:

Example 2: Find the value of the test statistic. The claim is that less than ½ of adults in the United States have carbon monoxide detectors. A KRC Research survey of 1005 adults resulted in 462 who have carbon monoxide detectors.

TOOLS FOR ASSESSING THE TEST STATISTIC: CRITICAL REGION, SIGNIFICANCE LEVEL, CRITICAL VALUE, AND *P*-VALUE

The _______________ ________________________ alone usually _________ _______ give us enough information to make a decision about the ___________ being _________________. The following tools can be used to ______________ and ________________________ the __________ ________________________.

- π The **critical region (aka rejection region)** is the __________ of all ______________ of the _____________ _______________ that cause us to ________________ the __________________ ____________________ _____
- π The **significance level (denoted by ___)** is the ______________ that the ________________ __________________ will fall in the _____________________ __________________ when the ____________ _______________________ is actually ______. If the ________________ __________________ falls in the _________________ ______________________, we _________ the _______________ ____________________, so ______ is the ____________________ of making the _________________ of _______________ the ___________ _________________ when it is ________________.
- π A **critical value** is any value that _________________ the ____________________ ______________________ from the _______________ of the _____________ _______________ that _______ ___________ lead to _______________________ of the ____________ _______________________. The _______________ _____________________ depend on the nature of the _______________

______________________, the __________________ ____________________ that applies, and the __________________ _____________ of _____. The procedure can be summarized as follows:

Critical region in the left tail:

Critical region in the right tail:

Critical region in two tails:

π The ***P*-value (aka *p*-value or probability value)** is the ____________________ of getting a ________________ of the ________________ ___________________ that is _____ ____________ _____ _________________ as the one representing the ________________ __________, assuming that the ___________ ______________ is ______________. *P*-values can be found __________________ finding the _____ __________________ the ______________ _________________.

DECISIONS AND CONCLUSIONS

***P*-value method:** Using the ____________________ ___________ _____:

If *P*-value ______ _____, ___________ ______

If *P*-value ______ _____, _________to __________ _____

Traditional method: If the ________________ ________________ falls _______________ the ________________ __________________, _______________ _____. If the ____________ ________________ ________ ____________ fall ______________ the _________________ ___________, ______________ to _______________ ______.

Confidence intervals: A ______________ _________________ ______________ of a _________________ _________________ contains the _____________ values of that ______________________. If a _______________ ____________________ does __________ ________________ a _____________ value of a _______________ ___________________, ________________ that _____________.

Example 3: Use the given information to find *P*-value.

a. The test statistic in a right-tailed test is $z = 2.50$

b. The test statistic in a two-tailed test is z = -0.55

c. With $H_1: p \neq \frac{3}{4}$, the test statistic is z = 0.35

d. With $H_1: p < 0.777$, the test statistic is z = -2.95

Example 4: State the final conclusion in simple non-technical terms. Be sure to address the original claim. Original claim: The percentage of on-time U.S. airline flights is less than 75%. Initial conclusion: Reject the null hypothesis.

ERRORS IN HYPOTHESIS TESTS

		TRUE STATE OF NATURE	
		THE NULL HYPOTHESIS IS TRUE	THE NULL HYPOTHESIS IS FALSE
DECISION	We decide to reject H_0	TYPE I ERROR	CORRECT DECISION
	We fail to reject H_0	CORRECT DECISION	TYPE II ERROR

Example 5: Identify the type I error and the type II error that correspond to the given hypothesis. The percentage of Americans who believe that life exists only on earth is equal to 20%.

COMPREHENSIVE HYPOTHESIS TEST

CONFIDENCE INTERVAL METHOD

For _______________ hypothesis tests _____________ a ___________________ interval with a

___________________ _______________ of _____________; but for a ____________________

hypothesis test with ______________________ _______________ ____, construct a

___________________ ______________ of ___________.

A ___________________ ____________________ ________________ of a __________________

_____________________ contains the ______________ values of that parameter. We should therefore _______________ a __________ that the population parameter has a _____________

that is ________ included in the ________________ ____________________.

8.3 TESTING A CLAIM ABOUT A PROPORTION

PART 1: BASIC METHODS OF TESTING CLAIMS ABOUT A POPULATION PROPORTION *p*

OBJECTIVE

NOTATION

$n =$ $p =$

$\hat{p} = \frac{}{}$ $q =$

REQUIREMENTS

1. The ________________ observations are a ____________ ___________ sample.
2. The ___________________ for a _______________ _________________ are satisfied.

3. The conditions ___________ and ____________ are _________ satisfied so the _______________ _____________________ of _______________ proportions can be ____________________ by a _____________ _____________ with ______________ and ________________. Note that ____ is the ________________ ___________________ used in the ____________________.

TEST STATISTIC FOR TESTING A CLAIM ABOUT A PROPORTION

$z = \underline{\hspace{4cm}}$

$P -$values:

Critical values:

FINDING THE NUMBER OF SUCCESSES *x*

Computer software and ________________ designed for _______________ tests of _______________ usually require ______________ consisting of the __________________ ___________ ____ and the number of ____________ ____, but the ____________________ ____________________ is often given instead of ____.

Example 1: Identify the indicated values. Use the normal distribution as an approximation to the binomial distribution. In a survey, 1864 out of 2246 randomly selected adults in the United States said that texting while driving should be illegal (based on data from Zogby International). Consider a hypothesis test that uses a 0.05 significance level to test the claim that more than 80% of adults believe that texting while driving should be illegal.

a. What is the test statistic?

b. What is the critical value?

c. What is the *P*-value?

d. What is the conclusion?

Example 2: The company Drug Test Success provides a "1-Panel-THC" test for marijuana usage. Among 300 tested subjects, results from 27 subjects were wrong (either a false positive or a false negative). Use a 0.05 significance level to test the claim that less than 10% of the test results are wrong. Does the test appear to be good for most purposes?

a. Identify the null hypothesis

b. Identify the alternative hypothesis

c. Identify the test statistic

d. Identify the *P*-value or critical value(s)

e. What is your final conclusion?

Example 3: In recent years, the town of Newport experienced an arrest rate of 25% for robberies (based on FBI data). The new sheriff compiles records showing that among 30 recent robberies, the arrest rate is 30%, so she claims that her arrest rate is greater than the 25% rate in the past. Is there sufficient evidence to support her claim that the arrest rate is greater than 25%?

a. Identify the null hypothesis

b. Identify the alternative hypothesis

c. Identify the test statistic

d. Identify the *P*-value or critical value(s)

e. What is your final conclusion?

8.4 TESTING A CLAIM ABOUT A MEAN: SIGMA KNOWN

TESTING CLAIMS ABOUT A POPULATION MEAN (WITH σ KNOWN)
OBJECTIVE
NOTATION

$n =$	$\mu_{\overline{x}} =$
$\overline{x} =$	$\sigma =$

REQUIREMENTS

1. The ________________ is a ______________ ____________ ___________ (_____).

2. The ___________________ of the _______________ _________________ ____________________ ________ is ______________.

3. The _____________________ is ___________________ _________________ and/or _________.

TEST STATISTIC FOR TESTING A CLAIM ABOUT A MEAN (WITH σ KNOWN)

$z = \rule{4cm}{0.4pt}$

$P-$values:

Critical values:

Example 1: When a fair die is rolled many times, the outcomes of 1, 2, 3, 4, 5, and 6 are equally likely, so the mean of the outcomes should be 3.5. The author drilled holes into a die and loaded it by inserting lead weights, then rolled it 40 times to obtain a mean of 2.9375. Assume that the standard deviation of the outcomes is 1.7078, which is the standard deviation for a fair die. Use a 0.05 significance level to test the claim that outcomes from the loaded die have a mean different from the value of 3.5 expected with a fair die.

a. Identify the null hypothesis

b. Identify the alternative hypothesis

c. Identify the test statistic

d. Identify the *P*-value or critical value(s)

e. What is your final conclusion?

Example 2: Listed below are recorded speeds (in mi/h) of randomly selected cars traveling on a section of Highway 405 in Los Angeles (based on data from Sigalert). That part of the highway has a posted speed limit of 65 mi/h. Assume that the standard deviation of speeds is 5.7 mi/h and use a 0.01 significance level to test the claim that the sample data is from a population with a mean greater than 65 mi/h.

68 68 72 73 65 74 73 72 68 65 65 73 66 71 68 74 66 71 65 73
59 75 70 56 66 75 68 75 62 72 60 73 61 75 58 74 60 73 58 75

a. Identify the null hypothesis

b. Identify the alternative hypothesis

c. Identify the test statistic

d. Identify the *P*-value or critical value(s)

e. What is your final conclusion?

8.5 TESTING A CLAIM ABOUT A MEAN: SIGMA NOT KNOWN

TESTING CLAIMS ABOUT A POPULATION MEAN (WITH σ NOT KNOWN)

OBJECTIVE

NOTATION

$n =$

$\mu_{\bar{x}} =$

$\bar{x} =$

$s =$

REQUIREMENTS

1. The ________________ is a ______________ ____________ ___________ (_____).

2. The __________________ of the _______________ _________________ ___________________ ________ is ______ ____________.

3. The ____________________ is __________________ _________________ and/or ________.

TEST STATISTIC FOR TESTING A CLAIM ABOUT A MEAN (WITH σ KNOWN)

$t =$ ______________

$P-$values:

Critical values:

CHOOSING THE CORRECT METHOD

When _________________ a _____________ about a _________________ __________, first be sure that the sample data have been collected with an appropriate ______________________ method. If we have a _________________ ______________________ _________________, a _________________ test of a _____________ about ____ might use the __________________ _______ _____________________, the _______________________ distribution, or it might require _________________________ methods or ___________________ resampling techniques.

To test a _____________ about a ____________________ ___________, use the ___________________ _____ ____________________ when the sample is a _____________________ ______________________ _______________, ____ is _________ _______________, and ____________________ or ____________ of these conditions is satisfied:

The ________________ is ___________________ distributed or ____________.

Example 1: Determine whether the hypothesis test involves a sampling distribution of means that is a normal distribution, Student *t* distribution, or neither.

a. Claim about FICO credit scores of adults: $\mu = 678,\ n = 12,\ \overline{x} = 719,\ s = 92.$ The sample data appear to come from a population with a distribution that is not normal and σ is not known.

b. Claim about daily rainfall amounts in Boston: $\mu < 0.20 \text{ in.},\ n = 52,\ \overline{x} = 0.10 \text{ in.},\ s = 0.26 \text{ in.}$ The sample data appear to come from a population with a distribution that is very far from normal, and σ is known.

FINDING *P*-VALUES WITH THE STUDENT *t* DISTRIBUTION

1. Use software or a ____________________ _____________________.

2. If ______________________ is not available, use Table A-3 to identify a _____________ of

   ______________ as follows: Use the number of _____________ of _______________ to

   ______________ the ______________ row of Table A-3, then determine where the __________

   _________________ lies ________________ to the _________________ in that ___________.

   Based on a comparison of the ____ ______________ _____________________ and the ______

   _____________ in the row of Table A-3, ________________ a ______________ of

   _______________ by referring to the _____________ ________________ given at the

   _______ of Table A-3.

Example 2: Either use technology to find the *P*-value or use Table A-3 to find a range of values for the *P*-value.

a. Movie Viewer Ratings: Two-tailed test with $n = 15$, and test statistic $t = 1.495$.

b. Body Temperatures: Test a claim about the mean body temperature of healthy adults. Left-tailed test with $n = 11$ and test statistic $t = -3.518$.

Example 3: Assume that a SRS has been selected from a normally distributed population and test the given claim. A SRS of 40 recorded speeds (in mi/h) is observed from cars traveling on a section of Highway 405 in Los Angeles. The sample has a mean of 68.4 mi/h and a standard deviation of 5.7 mi/h (based on data from Sigalert). Use a 0.05 significance level to test the claim that the mean speed of all cars is greater that the posted speed limit of 65 mi/h.

a. Identify the null hypothesis

b. Identify the alternative hypothesis

c. Identify the test statistic

d. Identify the *P*-value or critical value(s)

e. What is your final conclusion?

Example 2: Assume that a SRS has been selected from a normally distributed population and test the given claim. The trend of thinner Miss America winners has generated charges that the contest encourages unhealthy diet habits among young women. Listed below are body mass indexes (BMI) of recent Miss America winners. Use a 0.01 significance level to test the claim that recent Miss America winners are from a population with a mean BMI less than 20.16, which was the BMI for winners from the 1920s and 1930s.

19.5 20.3 19.6 20.2 17.8 17.9 19.1 18.8 17.6 16.8

a. Identify the null hypothesis

b. Identify the alternative hypothesis

c. Identify the test statistic

d. Identify the *P*-value or critical value(s)

e. What is your final conclusion?

9.2 INFERENCES ABOUT TWO PROPORTIONS

OBJECTIVES

NOTATION FOR TWO PROPORTIONS

$p_1 =$

$\hat{p}_1 = \frac{}{}$

$n_1 =$

$\hat{q}_1 =$

$x_1 =$

The corresponding notations $p_2, n_2, x_2, \hat{p}_2,$ and $\hat{q}_2$ apply to population 2.

POOLED SAMPLE PROPORTION

The ___________________ ____________________ ______________________ is

denoted by _____ and is given by:

REQUIREMENTS

1. The __________________ _________________________ are from __________ _____________

___________ samples that are ____________________________.

2. For each of the ______________ samples, the number of _________________ is _____ _____________

___________ and the number of ____________________ is at ____________________ ______. That is,

____________________ and _____________________ for each of the two samples.

TEST STATISTIC FOR TWO PROPORTIONS (WITH $H_0 : p_1 = p_2$)

$z = \rule{6cm}{0.4pt}$

$P -$ value:

Critical values:

CONFIDENCE INTERVAL ESTIMATE OF $p_1 = p_2$

The confidence interval estimate of the ____________________ _______________ is:

where the ___________________ of _____________________ ______ is given by

Rounding: Round the confidence interval limits to ____________ significant digits.

CAUTION!!! When testing a claim about ____________ population proportions, the ________________ method and the _________________________ method are equivalent, but they _________ ______ equivalent to the ____________________________ __________________________ method!!! If you want to _____________ a claim about ______________ _____________________ _________________,

use the ___________ method or the __________________________ method; if you want to

_______________________ the _________________________ between ______ ________________

_______________________, use a ___________________________ __________________________.

Example 1: In a 1993 survey of 560 college students, 171 said they used illegal drugs during the previous year. In a recent survey of 720 college students, 263 said that they used illegal drugs during the previous year (based on data from the National Center for Addiction and Substance Abuse at Colombia University). Use a 0.05 significance level to test the claim that the proportion of college students using illegal drugs in 1993 was less than it is now.

Example 2: Among 2739 female atom bomb survivors, 1397 developed thyroid diseases. Among 1352 male atom bomb survivors, 436 developed thyroid diseases (based on data from "Radiation Dose-Response Relationships for Thyroid Nodules and Autoimmune Thyroid Diseases in Hiroshima and Nagasaki Atomic Bomb Survivors 55-58 Years After Radiation Exposure," by Imaizumi, et al., *Journal of the American Medical Association,* Vol. 295, No. 9).

a. Use a 0.01 significance level to test the claim that the female survivors and male survivors have different rates of thyroid diseases.

b. Construct the confidence interval corresponding to the hypothesis test conducted with a 0.01 significance level.

c. What conclusion does the confidence interval suggest?

9.3 INFERENCES ABOUT TWO MEANS: INDEPENDENT SAMPLES

INDEPENDENT SAMPLES WITH σ_1 AND σ_2 UNKNOWN AND NOT ASSUMED EQUAL

DEFINITION

Two ________________________ are **independent** if the ________________ ____________________

from one population _________ __________ ____________________ ____ or somehow

___________________ ____________ or _____________________ with the ____________________

______________ from the other population.

Two ____________________ are **dependent** if the sample values are ___________.

Inferences about Means of Two Independent Populations, With σ_1 and σ_2 Unknown and Not Assumed to be Equal

NOTATION

Population 1:

$\mu_1 =$ $s_1 =$

σ_1

$\bar{x}_1 =$ $n_1 =$

The corresponding notations for ______, ______, ______, ______,and ______ apply to population _______.

REQUIREMENTS

1. _____ and _____ are _______________ and it is not _____________ that _____ and ______ are __________.

2. The _______ samples are ________________.

3. Both samples are ____________________ __________________ _________________.

4. Either or both of these conditions are satisfied: The two _____________ ___________ are both _________ (with _________ and __________) or both samples come from populations having _________________ _______________.

HYPOTHESIS TEST STATISTIC FOR TWO MEANS: INDEPENDENT SAMPLES

$t =$ ________________________________

Degrees of Freedom: When finding ________________ ________________ or _______________, use the following for determining the number of degrees of freedom.

1. In this book we use the conservative estimate: df = ___________ of _______ and __________.

2. Statistical software packages typically use the more accurate but more difficult estimate given below:

$$\text{df} = \frac{(A+B)^2}{\dfrac{A^2}{n_1-1}+\dfrac{B^2}{n_2-1}}, \quad A = \frac{s_1{}^2}{n_1}, \quad B = \frac{s_2{}^2}{n_2}$$

***P*-values and critical values:** Use Table A-3.

CONFIDENCE INTERVAL ESTIMATE OF $\mu_1 - \mu_2$: INDEPENDENT SAMPLES

The confidence interval estimate of the difference _____________ is

and the number of degrees of freedom df is as described above for hypothesis tests.

EQUIVALENCE OF METHODS

Example 1: Determine whether the samples are independent or dependent.

a. To test the effectiveness of Lipitor, cholesterol levels are measured in 250 subjects before and after Lipitor treatments.

b. On each of 40 different days, the author measured the voltage supplied to his home and he also measured the voltage produced by his gasoline powered generator.

Example 2: Assume that the two samples are independent simple random samples selected from normally distributed populations. Do not assume that the population standard deviations are equal. A simple random sample of 13 four-cylinder cars is obtained, and the braking distances are measured. The mean braking distance is 137.5 feet and the standard deviation is 5.8 feet. A SRS of 12 six-cylinder cars is obtained and the braking distances have a mean of 136.3 feet with a standard deviation of 9.7 feet (based on Data Set 16 in Appendix B).

a. Construct a 90% CI estimate of the difference between the mean braking distance of four-cylinder cars and six-cylinder cars.

b. Does there appear to be a difference between the two means?

c. Use a 0.05 significance level to test the claim that the mean braking distance of four-cylinder cars is greater than the mean braking distance of six-cylinder cars.

TI-83/84 PLUS

9.4 INFERENCES FROM DEPENDENT SAMPLES

Key Concept…

In this section we present methods for testing hypotheses and constructing confidence intervals

involving the ________________ of the ____________________ of the _______________ of

two _______________ _____________________. With ________________ samples, there is

some ____________________ whereby each value in one sample is _____________ with a

__________________ value in the other sample. Here are two typical

examples of dependent samples:

- Each pair of sample values consists of two measurements from the _____________ subject
- Each pair of sample values consists of a _____________ __________.

Because the hypothesis test and CI use the same __________________ and _________________

______________, they are _______________ in the sense that they result in the

______________ ____________________. Consequently, the ______________ hypothesis that

the _____________ _______________ _______________ _____ can be tested by determining

whether the __________________ __________________ includes ______. There are no exact

procedures for dealing with ___________________ samples, but the _____ __________________

serves as a reasonably good approximation, so the following methods are commonly used.

Inferences about Means of Two Dependent Populations

NOTATION

$d =$ $s_d =$

μ_d

$\bar{d} =$ $n =$

REQUIREMENTS

1. The ____________ data are ____________________.

2. The samples are ________________ __________________ _________________.

3. Either or both of these conditions are satisfied: The number of ___________ of ______________ is ______________ (__________) or the pairs of values have __________________ that are from a population that is approximately __________________.

HYPOTHESIS TEST FOR DEPENDENT SAMPLES

$t =$ ______________________________________

Degrees of Freedom: _____________
***P*-values and critical values:** Use Table A-3.

CONFIDENCE INTERVALS FOR DEPENDENT SAMPLES

where

and

Example 1: Assume that the paired sample data are SRSs and that the differences have a distribution that is approximately normal.

a. Listed below are BMIs of college students.

April BMI	20.15	19.24	20.77	23.85	21.32
September BMI	20.68	19.48	19.59	24.57	20.96

i. Use a 0.05 significance level to test the claim that the mean change in BMI for all students is equal to 0.

ii. Construct a 95% CI estimate of the change in BMI during freshman year.

iii. Does the CI include zero, and what does that suggest about BMI during freshman year?

b. Listed below are systolic blood pressure measurements (mm Hg) taken from the right and left arms of the same woman. Use a 0.05 significance level to test for a difference between the measurements from the two arms. What do you conclude?

Right arm	102	101	94	79	79
Left arm	175	169	182	146	144

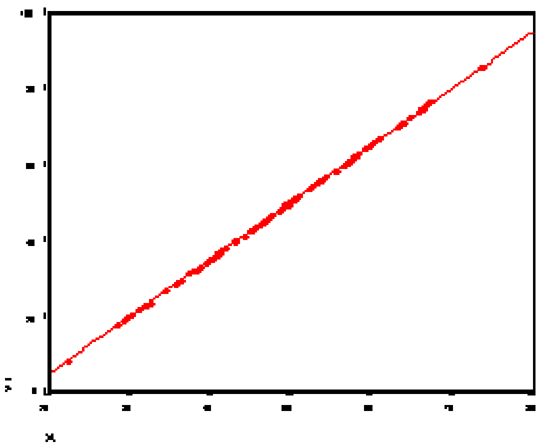
10.2 CORRELATION

DEFINITION

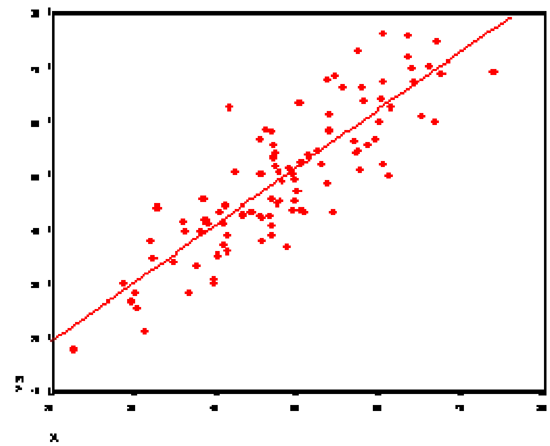
A **correlation** exists between two _______________ when the ______________ of one variable are somehow ____________________ with the values of the other variable.

EXPLORING THE DATA

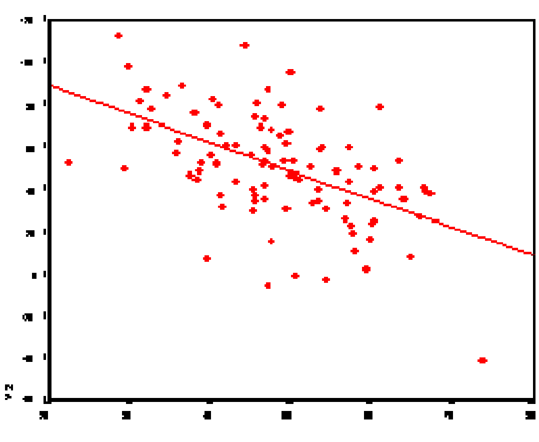
r = 1.00



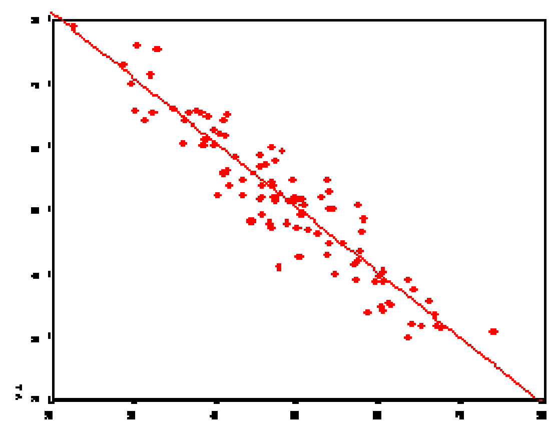
r = .85

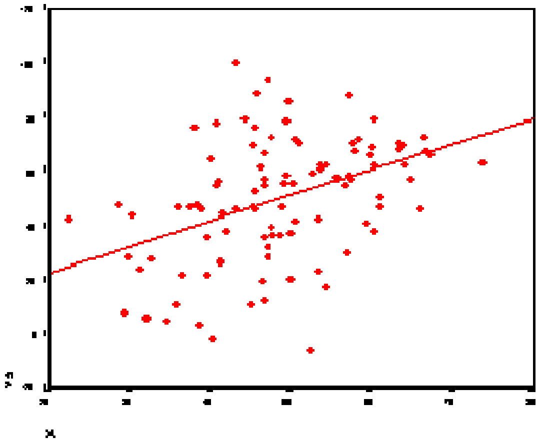


r = -.54

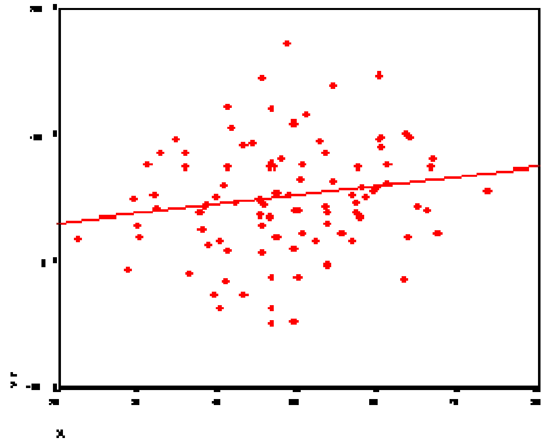


r = -.94

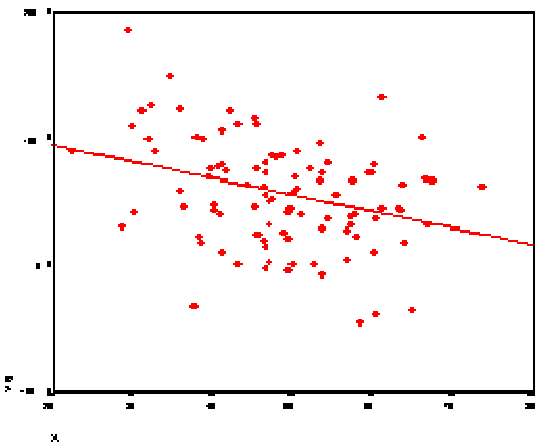




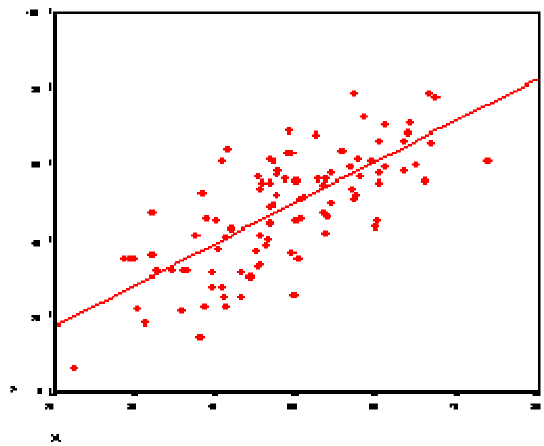
r = .33



r = .39



r =- .17



r = .42

DEFINITION

The **linear correlation coefficient *r*** measures the _______________ of the _______________ _________________ between the ___________________ _______________________ _____ and _______________ in a _______________. The linear correlation coefficient is sometimes referred to as the _____________________ __________________ _________________ ___________________ ____________________ in honor of Karl Pearson who originally developed it. Because the linear ___________________ coefficient _____ is calculated using _________________ data, it is a ___________________ _____________________. If we had every pair of ____________________ values, it would be represented by ____ (Greek letter rho).

OBJECTIVE

NOTATION FOR THE LINEAR CORRELATION COEFFICIENT

$n =$	$(\Sigma x)^2 =$
$\Sigma =$	$\Sigma xy =$
$\Sigma x =$	$r =$
$\Sigma x^2 =$	$\rho =$

REQUIREMENTS

1. The ________________ of ________________ __________data is a SRS of ____________________ data.

2. Visual examination of the ______________________ must confirm that the points _______________ a straight-line __________________.

3. Because results can be _________________ affected by the presence of ____________________,

any ________________________ must be _________________ if they are known to be ____________.

The effects of any other _______________ should be considered by calculating ____ with and without

the ________________ included.

FORMULAS FOR CALCULATING *r*

$$r = \frac{\qquad}{\qquad}$$

$$r = \frac{\qquad}{\qquad}$$

where _______ is the _____ ___________ for the sample value _____ and _____ is

the _____________ for the sample value ______.

INTERPRETING THE LINEAR CORRELATION COEFFICIENT *r*

Computer Software

If the _______________ computed from _____ is less than or equal to the ____________________

____________, conclude that there is a ______________ correlation. Otherwise, there is not

____________________ evidence to support the ____________________ of linear _________________.

Table A-5

If the ____________________ ____________ of _______, denoted _____, exceeds the value in Table A-5,

conclude that there is a _______________ correlation. Otherwise, there is not sufficient evidence to

________________ the conclusion of a linear correlation.

ROUNDING THE LINEAR CORRELATION COEFFICIENT *r*

Round the ___________________ _______________ ________________ ____ to _______ decimal places so that its value can be compared to critical values in Table A-5. Keep as many decimal places during the process and then ___________ at the end.

PROPERTIES OF THE LINEAR CORRELATION COEFFICIENT *r*

1. The value of ____ is always between ____ and ____ inclusive. That is ___________________.
2. If all values of _____________ variable are ________________ to a different ______________, the value of _____ _________ ________ change.
3. The value of ____ is ______ affected by the choice of ____ or ____.
4. _____ measures the _____________ of a _______________ relationship. It is not designed to measure the strength of a ___________________ that is _______ linear.
5. _____ is very sensitive to ____________ in the sense that a ____________ outlier can ______________________ affect its value.

COMMON ERRORS INVOLVING CORRELATION

1. A common ________________ is to _________________ that _________________ implies ___________________.
2. Another error arises with data based on ____________. Average ___________________ __________________ variation and may __________ the ________________ ___________________.
3. A third error involves the property of ________________. If there is no linear _____________, there might be some other ______________ that is not __________________.

Example 2: The paired values of the CPI and the cost of a slice of pizza are listed below.

CPI	30.2	48.3	112.3	162.2	191.9	197.8
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00

a. Construct a scatterplot

b. Find the value of the linear correlation coefficient *r*

c. Find the critical values of *r* from Table A-5 using a significance level of 0.05.

d. Determine whether there is sufficient evidence to support a claim of a linear correlation between the two variables.

10.3 INFERENCES ABOUT TWO MEANS: INDEPENDENT SAMPLES

PART 1: BASIC CONCEPTS OF REGRESSION

Two variables are sometimes related in a ______________________ way, meaning that given a value for one variable, the _____________ of the other variable is ___________________ determined without any _____________, as in the equation $y = 6x + 5$. Statistics courses focus on __________________________ models, which are equations with a variable that is not ___________________ completely by the other variable.

DEFIITION

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

algebraically describes the ____________________ between the two variables _____ and _____. The _______________ of the _______________ equation is called the **regression line (aka line of best it, or least-squares line)**. The regression equation expresses a relationship between the _______________ variable _____ (aka ________________ variable or __________________ variable) and ______ (called the _______________ variable, or ___________________ variable). The slope and y-intercept can be found using the following formulas:

The ___________________ line fits the __________________ points __________!

OBJECTIVE

NOTATION

	Population Parameter	Sample Statistic
y-intercept of regression equation		
Slope of regression equation		
Equation of the regression line		

REQUIREMENTS

1.The __________________ of __________________ ___________data is a SRS of ____________________ data.

2. Visual examination of the ________________________ must confirm that the points _______________ a straight-line ____________________.

3. Because results can be ___________________ affected by the presence of _____________________, any __________________________ must be ___________________ if they are known to be ____________________. The effects of any other ________________ should be considered by calculating _____ with and without the __________________ included.

FORMULAS FOR FINDING THE SLOPE _____ AND *y*-INTERCEPT _____ IN THE REGRESSION EQUATION ____________________

Slope:

where ____ is the _________ correlation coefficient, _____ is the ______________

_________________ of the ____ values, and _____ is the standard deviation of the _____ values

y-intercept:

ROUNDING THE SLOPE AND THE *y*-INTERCEPT

Round _____ and _____ to _________ ______________ ________________.

USING THE REGRESSION EQUATION FOR PREDICTIONS

1. Use the regression equation for ____________________ only if the ________ of the ______________ line on the ________________ confirms that the ______________ __________ ________ the points reasonably well.
2. Use the regression equation for ________________ only if the ________ correlation coefficient ____ indicates that there is a ___________ correlation between the two variables.
3. Use the regression line for predictions only if the _________ do not go much _____________ the __________ of the available __________ data.
4. If the regression equation does not appear to be ______________ for making ____________,

the best ________________ value is its ________________ estimate, which is its

_____________ _____________.

Example 1: The paired values of the CPI and the cost of a slice of pizza are listed below.

CPI	30.2	48.3	112.3	162.2	191.9	197.8
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00

a. Find the regression equation, letting the first variable be the predictor (x) variable.

b. Find the best predicted cost of a slice of pizza when the CPI is 182.5.

PART 2: BEYOND THE BASICS OF REGRESSION

DEFINITION

In working with two variables _________________ by a regression equation, the <u>**marginal change**</u> in a ________________ is the ________________ that it changes when the other variable changes by exactly ______ unit. The slope ___ in the regression equation represents the ______________ ___________ in ___ that occurs when ____ changes by _____ unit.

DEFINITION

In a ____________________, an **outlier** is a point lying ______ away from the other data points. Paired sample data may include one or more **influential points**, which are _____________ that ______________ affect the __________ of the _______________ ___________.

DEFINITION

For a pair of sample ____ and ____ values, the **residual** is the ______________ between the ______________ sample value of ____ and the _________ that is _________________ by using the ________________ equation. That is,

Residual = _________________ ___ - __________________ ___ = ___________

DEFINITION

A ______________ line satisfies the **least-squares property** if the ______ of their _____________ is the ________________ sum possible.

DEFINITION

A **residual plot** is a ____________________ of the __________ values after each of the _____________________ values has been ___________________ by the __________________ value ____________. That is, a residual plot is a graph of the points __________________.