Part I	Exploring and Understanding Data
Chapter 1	**Stats Starts Here**
Statistics is	a way of reasoning, along with a collection of tools and methods, designed to help us understand the world.
Statistics are	particular calculations made from data.
A statistic is	A numerical summary of data
Statistics is about	variation
Chapter 2	**Data**
Data are	values along with their context
The context for data values is provided by _______________	The "W's" **Why** do we care about the data? **Who** are the individuals described by the data? **What** variables do the data contain? When Where How **(Necessary)**
Three steps to doing Statistics right:	Think –were you're headed and why (the "W's"). Show – the mechanics of calculating statistics and making displays. Tell – what you've learned remembering the "4 Cs."
4 Cs: conclusions are ______	Clear, concise, complete, and in context.
Data table	An arrangement of data in which each row represents a case and each column represents a variable.
Case	An individual about whom we have data (row of data table)
Individual	Object described by a set of data (person, animal, thing, identifier variable)
Variable	Holds information about the same characteristic for many cases. (column of data table)
Variables can usually be identified as either ____ or ___:	Categorical or quantitative
Categorical variable	Places an individual into one of several groups or categories
Quantitative variable	Has numerical values (with units) that measure some characteristic of each individual.
Ordinal variable _____________ You must look at the ________ of your study to decide whether to treat it as ______ or ______	Reports order with out natural units. Why Categorical or quantitative
Identifier variable	ID number or other convention often used to protect confidentiality (Categorical variable with exactly one individual in each category)
Chapter 3	**Displaying and Describing Categorical Data**
Three things you should always do first with data:	1. Make a picture – a display will help you *think* clearly about patterns and relationships that may be hiding in your data. 2. Make a picture – *show* important features and patterns in your data 3. Make a picture – best way to *tell* others about your data.
To analyze categorical data, we often use ________ or _______	counts (frequencies) or percents (relative frequencies)

of individuals that fall into various categories.	
(Relative) Frequency table [Distribution of a categorical variable]	Lists the categories in a categorical variable and the (percentage) count of observations for each category.
Area principle	In a statistical display, each data value should be represented by the same amount of area.
(Relative Frequency) Bar chart	Shows a bar representing the (percentage) count of each category in a categorical variable.
Pie chart	Shows how a "whole" divides into categories by showing a wedge of a circle whose area corresponds to the proportion in each category.
Contingency table	Displays counts (percentages) of individuals falling into named categories on two (or more) variables, columns vs. rows. The table categorizes the individuals on all variables at once, to reveal possible patterns in one variable that may be contingent on the category of the other.
Marginal distribution	The distribution of one of the variables <u>alone</u> is seen in the totals found in the last row/column of a contingency table. (see frequency table)
Conditional distribution	The distribution of a variable restricting the *Who* to consider only a smaller group of individuals. [A single row (column) of the contingency table.]
Relationships among categorical variables are described by calculating ______ from the _____ given. This avoids ______________________	percents counts count variation between them.
Segmented Bar Chart	A stacked relative frequency bar chart (100% total). Often better than a pie chart for comparing distributions. [a pie chart within a bar chart]
Independent variables	The conditional distribution of one variable is the same for each category of the other. [if rows (columns) of contingency table have = distributions]
Simpson's paradox	When averages are taken across different groups, they can appear to contradict the overall averages
Chapter 4	**Displaying Quantitative Data**
Distribution of a quantitative variable	Tells us what values a variable takes and how often it takes them. Shows the pattern of variation of a (quantitative) variable.
Stem-and-leaf plot	A sideways histogram that shows the individual values. Bins/intervals might be the tens places with the ones places strung out sequentially to the right.
Back-to-back stem-and-leaf plot	Useful for comparing two related distributions with a moderate number of observations.
Dotplot	Graphs a dot for each case against a single axis.
(Relative Frequency) Histogram	Uses adjacent, equal-width bars to show the distribution of values in a quantitative variable. Each bar represents the (percentage) count falling in a particular interval of values. (% are useful for comparing

	several distributions with different numbers of observations.)
A good estimate for how many bars will give a decent histogram =	$\frac{\text{Number of observations}}{5}$
Once we make a picture, we describe a distribution by telling about its	Shape, center, spread, and any unusual features.
Shape	Uniform, single, multiple modes Symmetry vs. skewed
Uniform	A distribution that is roughly flat.
Mode	A hump or local high point in the shape of the distribution of a variable (unimodal, bimodal, multimodal).
Symmetric	A distribution where the two halves on either side of the center look approximately like mirror images of each other.
Skewed (left/right)	A non-symmetrical distribution where one tail stretches out further (to the left/right) than the other.
Center	A "typical" value that attempts the impossible, summarizing the entire distribution with a single number. {midpoint}
Spread	A numerical summary of how tightly the values are clustered around the "center." {range}
Outliers	Extreme values that don't appear to belong with the rest of the data.
Timeplot	Displays quantitative data collected over time (x-axis). Can reveal trends overlooked by histograms and stem-and-leaf plots that ignore time order. Often, successive values are connected with lines to show trends more clearly.
Time series	Measurements of a variable taken at regular time intervals.
Seasonal variation	A pattern in a time series that repeats itself at know regular intervals of time.
Chapter 5	**Describing Distributions Numerically**
Median	Middle value (balances data by counts) (equal-areas point)
Range	Max – min data values
*p*th percentile	Value such that *p* percent of the observations fall at or below it.
Lower quartile (Q1) Upper quartile (Q3)	Median of the lower half. (25^{th} percentile) Median of the upper half. (75^{th} percentile)
Interquartile range (IQR)	Q3 – Q1, the middle half of the data.
5-number summary	Max Q3 Median Q1 Min
Suspected outlier	If observation > Q3 + (1.5)(IQR) Or observation < Q1 – (1.5)(IQR)
Boxplot	Displays the 5-number summary as a central box with whiskers that extend to the non-outlying data values. Particularly effective for comparing groups. However, a histogram or stem-and-leaf plot is a clearer display of the shape of a distribution.
Mean	[Average]

	$\overline{x} = \frac{\sum x}{n}$ Add up all the numbers and divide by n (balance point, by size) (balances deviations)
Deviation	How far each data value is from the mean.
Variance	$s^2 = \frac{\sum (x-\overline{x})^2}{n-1}$ Sum of the squared deviations from the mean, divided by n – 1.
Standard deviation	$s = \sqrt{\frac{\sum (x-\overline{x})^2}{n-1}}$ The square root of the variance (gets us back to the original units)
Report summary statistics to ____________ decimal places ____________________________	1 or 2 more than the original data.
When describing the distribution of a quantitative variable, if the shape is skewed then report _________________ If the shape is symmetric then report _________________ and repeat calculations without ___________ if present.	median and IQR (they are based on position) mean and standard deviation (they are based on size/value) outliers
A complete analysis of data almost always includes:	Verbal, visual, and numerical summaries.
Answers are _____, not ______	sentences, numbers
Chapter 6	**The Standard Deviation as a Ruler and the Normal Model**
Adding (subtracting) a constant to every data value ___________ the same constant to measures of position/center and __________ measures of spread.	adds (subtracts) does not change
Multiplying (dividing) every data value by a constant ___________ the same constant to measures of position/center and ___________ measures of spread.	multiplies (divides) multiplies (divides)
Changing the center and spread of a variable is equivalent to __	changing its units.
Standardizing	Uses the standard deviation as a ruler to measure distance from the mean creating z-scores $z = \frac{(x-\overline{x})}{s}$
z-scores tell us ______________ important uses are:	the number of standard deviations a value is from the mean. 1. Comparing values from different distributions (decathlon events) or values based on different units.

	2. Identifying unusual or surprising values among data. 3.
Units can be eliminated by ____ ___________ have no units.	standardizing the data. z-scores
When we standardize data to get _____ we do two things. First we _____ the data by subtracting the mean. Then we _____ the data by dividing by their standard deviation.	z-scores shift rescale
Standardizing has the following affect on the distribution of a variable:	Shape – is not changed. Center – the mean is shifted to 0 Spread – the standard deviation is rescaled to 1
If the distribution of a quantitative variable is ______ and ______ then the we can replace histograms by approximating the distribution with ________	unimodal roughly symmetric a normal model.
_________ are summaries of the data denoted with _______ mean, __ standard deviation, __	Statistics Latin letters $\bar{x}$, s
_________ are numerically valued attributes [statistics] of a model (they don't come from the data, they just specify the model) denoted with _______ mean, __ standard deviation, __	Parameters Greek letters μ, σ
A normal model is constructed from a rather complex equation only dependent on parameters for ______ and ______.	$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ mean, standard deviation $N(\mu, \sigma)$
The distribution of each normal model is ____, ____, and ____ as show by its density curve. We call it a density curve because the equation for the normal model adjusts the scale (of *y*, height) so that the area under the curve = ___ and gives the ______ for the distribution.	unimodal, symmetric, and bell-shaped 1 relative frequency
This scaling is extremely important in conceptualizing how unusual a value(z-score) is.	Specifically, it allows us to convert standard deviations into percents that are much easier to comprehend.
To avoid having to work with the complicated normal model	we convert our data to z-scores and use just one Standard Normal Model *N*(0,1) and its associated table.

equation or lug around a myriad of tables for every possible $N(\mu, \sigma)$	
Normal percentile	Read from a table of normal probabilities, it gives the percentage of values in a standard normal distribution found lying below a particular z-score.
The easiest conversion (from standard deviations to percents) is to remember the ___________ rule. About ____ of the data fall within 1 standard deviation of the mean, about ____ within 2 and about _____ within 3.	68,95,99.7 68% 95% 99.7%
Use this TI function if asked to find % or area	normalcdf(lower z-score, upper z-score)
Use this TI function If given % or area	invNorm(area to left) output is z-score that may have to be converted back
___________ is a more precise method than a histogram of checking the nearly normal condition, that the shape of the data's distribution is _________ and _________	A normal probability plot unimodal roughly symmetric
If the normal probability plot is roughly ___________ Then a normal model ________________________	a diagonal straight line will approximate the (actual) data well.
The _____________ of a normal curve identifies one standard deviation from the mean.	Inflection point
3 reasons normal distributions are important in statistics:	1. Good descriptions for some distributions of real data. 2. Good approximations to many kinds of chance outcomes. 3. Utilized in many statistical inference procedures.