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]

	$\bar{x} = \frac{\sum x}{n}$ 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	$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$ Sum of the squared deviations from the mean, divided by n – 1.
Standard deviation	$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$ 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 $z = \frac{(x - \bar{x})}{s}$
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.
Part II	**Exploring Relationships Between Variables**
Chapter 7	**Scatterplots, Association, and Correlation**
Scatterplot ______ is plotted on the x-axis. ______ is plotted on the y-axis.	Shows the relationship between two quantitative variables on the same cases (individuals). Explanatory (*independent*/input) variable Response (*dependent*/output) variable
Once we make a scatterplot, we describe association by telling about:	**1. Form**: straight, curved, no pattern, other? **2. Direction**: + or – slope? **3. Strength**: how much scatter {how closely points follow the form} **4. Unusual Features**: outliers, clusters, subgroups?
_____ is a deliberately vague term describing the relationship between two variables. If positive then ______________	Association increases in one variable generally correspond to increases in the other.

Correlation describes the ____ and ____ of the ____ relationship between two ______ variables, without significant ______	strength direction, linear quantitative outliers.
3 conditions needed for Correlation:	1. Quantitative Variables 2. Straight Enough 3. Outlier
The correlation coefficient is found by _______________. It's value ranges from ______ , it has no ______, and is immune to changes of _______	finding the average product of the z-scores (standardized values). $r = \frac{\sum z_x z_y}{n-1}$ -1 to +1 units. scale or order.
Perfect correlation r = _____, occurs only when __________.	± 1 the points lie exactly on a straight line. (you can perfectly predict one variable knowing the other)
No correlation r = ______, means that knowing one variable gives you __________	0 no information about the other variable.
You should give the ______ and ______ of x and y along with the correlation because …	Mean Standard deviation Correlation is not a complete description of two-variable data and the its formula uses means and standard deviations in the z-scores.
Scatterplots and correlation coefficients never prove ______	 causation.
Lurking variable	A variable other than x and y that simultaneously affects both variables, accounting for the correlation between the two.
To add a categorical variable to an existing scatterplot _____	 use a different plot color or symbol for each category.
Chapter 8	**Linear Regression**
Regression to the mean	Because the correlation is always less than 1.0 in magnitude, each predicted $\hat{y}$ tends to be fewer standard deviations from its mean than its corresponding x was from its mean. ($\hat{z}_y = rz_x$)
Residual If positive If negative	Observed value – predicted value $y - \hat{y}$ Then the model makes an underestimate. Then the model makes an overestimate.
Regression line Line of best fit For standardized values For actual *x* and *y* values	The unique line that minimizes the variance of the residuals (sum of the squared residuals). $\hat{z}_y = rz_x$ $\hat{y} = b_0 + b_1 x$
To calculate the regression line in real units (actual *x* and *y* values)	1. Find slope, $b_1 = \frac{rs_y}{s_x}$ 2. Find y-intercept, plug b_1 and point (x, y) [usually ($\bar{x}$, $\bar{y}$)] into $\hat{y} = b_0 + b_1 x$ and solve for b_0 3. Plug in slope, b_1, and y-intercept, b_0, into $\hat{y} = b_0 + b_1 x$

3 conditions needed for Linear Regression Models: /* same as correlation */	1. Quantitative Variables 2. Straight Enough – check original scatterplot & residual scatterplot 3. Outlier (clusters) –points with large residuals and/or high leverage
R^2	The square of the correlation, *r*, between *x* and *y* The success of the regression model in terms of the fraction of the variation of *y* accounted for by the model. (XX% of the variability in *y* is accounted for by variation in *x*) (differences in *x* explain XX% of the variability in *y*)
A high R^2	Does not demonstrate the appropriateness of the regression.
Looking at a ________________ is a good way to check the Straight Enough Condition. It should be _________________	a scatterplot of the residuals vs. the *x*-values. (appropriateness) boring: uniform scatter with no direction, shape, or outliers..
The ____ is the key to assessing how well the model fits (extracts the form).	variation in the residuals
Standard deviation of the residuals, s_e	Gives a measure of how much the points spread around the regression line.
$1 - R^2$	The fraction of the original variation left in the residuals. (The percentage of variability not explained by the regression line.)
Extrapolations	Dubious predictions of *y*-values based on *x*-values outside the range of the original data.
Chapter 9	**Regression Wisdom**
What can go wrong with regression:	1. Inferring Causation 2.Extrapolation 3.Outliers and Influential Points 4.Change in Scatterplot Pattern 5.Means (or other summaries) rather than actual data.
High leverage points With enough leverage the ____ can appear deceptively small.	Have *x*-values far from $\bar{x}$ (($\bar{x}$,$\bar{y}$) is the fulcrum) and pull more strongly on the regression line. residuals
Leverage and residual produce three flavors of outliers:	1) Extreme Conformers: don't influence model but do inflate R2 2) Large Residuals: might not influence model much but aren't consistent with the overall form. 3) Influential Points: those that distort the model
Influential point [most menacing]	Omitting it from the data results in a very different regression model
Influential points are often difficult to detect because	They distort the model which causes their residual to be small.
The surest way to verify an outlier and its affects is to	Calculate the regression line with and without the suspect point.
A histogram of the residuals	Compliments a scatterplot of the residuals in the search for conditions, such as subsets, that may compromise the effectiveness of the regression model.
Consider comparing two or more regressions if you find	1) Points with large residuals and/or high leverage. 2) Change in Scatterplot Pattern as a result of changes over time or subsets that behave differently.

Regressions based on summaries of the data _______ Because _________	Tend to look stronger than the regression on the original data. Summary statistics are less variable than the underlying data.
Chapter 10	**Re-expressing Data: Get It Straight!**
Re-expression	A means of altering the data to achieve the conditions/structure necessary to utilize particular summaries or models.
Several reasons to consider a re-expression:	1. Make the form of a scatterplot straighter. 2. Make the scatter in a scatterplot more consistent (not fan shaped). 3. Make the distribution of a variable (histogram) more symmetric. 4. Make the spread across different groups (boxplots) more similar.
Ladder of Powers A good starting point is ______ If all else fails _______	Orders the effects that the re-expressions have on the data 2 1 ½ 0 -½ -1 y^2 y √y log y -1/√y -1/y taking logs. try whacking the data with two logs (log x and log y).
Base 10 logs are roughly	One less than the number of digits needed to write the number.
Re-expression limitations:	1. Can't straighten scatterplots that turn around. 2. Can't re-express "-" data values with √ (+constant to shift > 0) 3. Minimal affect on data values far from 1-100. (-constant to shift) 4. Can't unify multiple modes.
When discussing the accuracy or confidence of the linear regression model be sure to comment on both the __ & __	Appropriateness of the model as indicated by the residual plot Success of the model as indicated by R^2
Part III	**Gathering Data**
Chapter 11	**Understanding Randomness**
What is it about random selection that makes it seem fair?	1. Nobody can guess the outcome in advance. 2. Outcomes are equally likely.
Random event/phenomenon	We know what outcomes could happen, but not which particular values will happen. Outcomes that we cannot predict but that nonetheless have a regular distribution in very many repetitions.
If our goal as statisticians is to uncover the truth about the world around us, then randomness is both . . .	our greatest enemy and our most important tool.
Simulation	A sequence of random outcomes that model a situation, often difficult to collect data on and with a mathematical answer hard to calculate. Models random events by using random numbers to specify event outcomes with relative frequencies that correspond to the true real-world relative frequencies we are trying to model. An artificial representation of a random process used to study its long-term properties.
Component	The most basic situation in a simulation in which something happens at random. [random happening]
Outcome	An individual result of a component [result of random happening]

Trial	The sequence of several components representing events that we are pretending will take place.
Response variable	The result of each trial with respect to what we were interested in.
Chapter 12	**Sample Surveys**
Population	The entire group of individuals or instances about whom we hope to learn, but examining all of them is usually impractical, if not impossible.
Sample	A (representative) subset of a population, examined in hope of learning about the population.
Sample survey	A study that asks questions of a sample drawn from some population in the hope of learning something about the entire population.(Polls)
Statistic They are written in __________	Any summary calculated from the (sampled) data. Latin ($\bar{x}$, *s*, *r*, *b*, $\hat{p}$)
Parameters They are written in __________	Key numbers in mathematical models used to represent reality. Greek (μ, σ, ρ, β, p)
Population parameter	A numerically valued attribute of a model for a population, often unknowable and estimated from sampled data.
Sample statistic	Correspond to, and thus estimate, a population parameter.
Representative Sample	Statistics computed from it accurately reflect the corresponding population parameters.
Bias	Any systematic failure of a sampling method to represent its population. It is almost impossible to recover from.
5 common bias errors:	1. Voluntary response – individuals can choose on their own whether to participate in the sample. Always yields invalid samples. 2. Convenience – when the sample is comprised of individuals readily available. Always yields a non-representative sample. 3. Undercoverage – when individuals from a subgroup of the population are selected less often than they should be. 4. Nonresponse – when a large fraction of those sampled will not or cannot respond. 5. Response – when respondents' answers might be affected by suvey design, such as question wording or interviewer behavior.
____ is often the best use of time and resources when sampling or surveying.	Reducing biases
Randomization	The best defense against bias. (stirring to make sure that on average the sample looks like the rest of the population)
Simple random sample (SRS)	A sample in which each set of n elements in the population has an equal chance of selection. The standard method of utilizing radomization to make the sample representative of the population of interest.
Sampling variability Sampling error	The natural tendency of randomly drawn samples to differ from each other.
The precision of the statistics of	

a sample depend on __________ not _______________	the sample size (soup spoon) its fraction of the larger population.
Census	A sample that consists of the entire population.
Sampling frame	A list of individuals, which clearly defines but may not be representative of the entire population, from which the sample is drawn.
Stratified samples	These samples can reduce sampling variability by identifying homogeneous subgroups and then randomly sampling within each.
Cluster samples	These samples randomly select among heterogeneous subgroups that each resemble the population at large, making our sampling tasks more manageable.
Systematic samples	These samples can work, when there is no relationship between the order of the sampling frame and the variables of interest, and are often the least expensive method of sampling. But we still want to start them randomly.
Multistage sample	A sampling scheme that combines several sampling methods.
Identify the W's: Why What Who When, Where, and How /* previously Who < What */	 Population and associated sampling frame. Parameter of interest and variables measured. Sample actually drawn. Given by the sampling plan.
Chapter 13	**Experiments and Observational Studies**
Observational study Retrospective Prospective	A study based on data in which no manipulation of factors has been employed (researchers don't assign choices). Usually focuses on estimating differences between groups but is not possible to demonstrate a causal relationship. Often used when an experiment is impractical. Subjects are selected and then their previous conditions or behaviors are determined. Subjects are followed to observe future outcomes. No treatments are deliberately applied.
To prove a cause-and-effect relationship we need to perform ___________	a valid experiment.
An experiment ______________ to create treatments, __________ to these treatment levels, and then ________________________ across treatment levels.	manipulates factor levels randomly assigns subjects compares the responses of the subject groups (boxplots are often a good choice for displaying results of groups)
Factor	A variable whose levels are controlled by the experimenter.
Level	The specific values that the experimenter chooses for a factor.
Treatment	The process, intervention, or other controlled circumstance applied to randomly assigned experimental units. Treatments are the different levels of a single factor or are made up of combinations of levels of two or more factors.
_____ are individuals on whom an experiment is performed.	Experimental units

Usually called ___________ or ___________ when human.	Subjects Participants
Response	A variable whose values are compared across different treatments.
The 4 principals of experimental design:	1. **Control** sources of variation other than the factors we are testing by making conditions as similar as possible for all treatment groups. 2. **Randomize** subjects to treatments to even out effects that we cannot control. 3. **Replicate** over as many subjects as possible. Would like to get results from a representative sample of the population of interest. 4. **Block** and then randomize within to reduce the effects of identifiable attributes of the subjects that cannot be controlled.
Control group	The experimental units assigned to a baseline treatment level, typically either the default treatment, which is well understood, or a null, placebo treatment. Their responses provide a basis for comparison.
Statistically significant	When an observed difference is too large for us to believe that it is likely to have occurred naturally (only by chance).
Placebo	A (fake) treatment known to have no effect, administered so that all groups experience the same conditions.
Placebo effect	The tendency of many human subjects (often 20% or more of experimental subjects) to show a response even when administered a placebo.
Blinding	Individuals associated with an experiment are not aware of how subjects have been allocated to treatment groups.
2 main classes of individuals who can affect the outcome of an experiment: Single-blind Double-blind	1. those who could influence the results (subjects, treatment administrators, or technicians) 2. those who evaluate the results (judges, treating physicians, etc.) When every individual in ***either*** of these classes is blinded. When everyone in ***both*** classes is blinded.
Block	Same idea for experiments as stratifying is for sampling. Group together subjects that are similar and randomize within those groups as a way to remove unwanted variation (of the differences between the groups so that we can see the differences caused by the treatments more clearly) (Doing parallel experiments on different groups.)
Matching	In a retrospective or prospective study, subjects who are similar in ways not under study may be paired and then compared with each other on the variables of interest as a way to reduce unwanted variation in much the same way as blocking.
Designs: Randomized block design Completely random design	 The randomization occurs only within blocks. All experimental units have an equal chance of receiving any particular treatment.
The best experiments are usually:	Randomized, comparative, double-blind, placebo-controlled.
Lurking (Confounding) variables are outside influences	

that make it ______________ we are modeling with ___________	harder to understand the relationship regression and observational studies (a designed experiment).
Lurking variable	Creates an association between two other variables that tempts us to think that one may cause the other. [regression analysis or observational study]
Confounding	Some other variable associated with a factor has an effect on the response variable. [experiments] Arises when the response we see in an experiment is at least partially attributable to uncontrolled variables.
Part IV	**Randomness and Probability**
Chapter 14	**From Randomness to Probability**
Probability	The proportion of times the event occurs in many repeated trials of a random phenomenon. (the long-term relative frequency of an event)
The rules and concepts of probability give us a language to talk and think about ________	Random phenomena.
Trial	A single attempt or realization of a random phenomenon.
Outcome	The value measured, observed, or reported for each trial.
Event	A combination of outcomes usually for the purpose of attaching a probability to them. Denoted with bold capital letters, **A**, **B**, or **C**.
Independent	The outcome of one trial doesn't influence or change the outcome of another.
The Law of Large Numbers (LLN)	The long-run relative frequency of repeated independent events settles down to the true probability as the number of trials increases.
Law of Averages (The lesson of LLN)	Assumes that the more something hasn't happened, the more likely it becomes. Random processes don't need to compensate in the short run to get back to the right long-run probabilities. (Streaks happen - the coin can't remember what happened and make things come out right.)
________ is just a casual term for probability.	Relative frequencies [at the beginning of the year it was a code-word for percent]
Probability of the event **A**	The likelihood of the event's occurrence $P(\mathbf{A}) = \dfrac{\text{number of outcomes in } \mathbf{A}}{\text{total number of outcomes}}$ if outcomes are equally likely $0 \leq P(\mathbf{A}) \leq 1$
Sample Space, ***S***	The collection of all possible outcome values.
Something Has to Happen Rule	The sum of the probabilities of all possible outcomes must be 1. $P(\boldsymbol{S}) = 1$
Complement Rule	The probability of an event occurring is 1 minus the probability that it doesn't occur. $P(\mathbf{A}) = 1 - P(\mathbf{A}^c)$
Disjoint	Two events that share no outcomes in common, mutually exclusive. (outcomes cannot happen **at the same time,** one prevents the other)
The _____________ says: If **A** and **B** are disjoint events, Then the probability of **A** *or* **B**	Addition Rule $P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B})$; ∪ = Union
An assignment of probabilities to outcomes is legitimate if	Each probability is between 0 and 1(inclusive) The sum of the probabilities is 1

The ____________ says: If **A** and **B** are independent events, Then the probability of **A** *and* **B**	Multiplication Rule $P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B})$; ∩ = Intersection
Chapter 15	**Probability Rules!**
________ and ________ should be used to display the sample space and aid probability calculations.	Venn diagrams, two-way contingency tables
When working with probabilities: "or" is the ______ of the two events and translates into _____ "and" is the ______ of the two events and translates into _____ "not" & "at least …" often indicate __________	Union + Intersection X Complement
The ____________ says: For any two events, **A** and **B**, The probability of **A** *or* **B**	General Addition Rule (avoids double counting when not disjoint) $P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \cap \mathbf{B})$
Conditional Probability [restricts the "Who"]	$P(\mathbf{B} \mid \mathbf{A}) = \dfrac{P(\mathbf{A} \cap \mathbf{B})}{P(\mathbf{A})}$ $P(\mathbf{B} \mid \mathbf{A})$ is read "the probability of **B** given **A**."
The ____________ says: For any two events, **A** and **B**, The probability of **A** *and* **B**	General Multiplication Rule (adjusts for non-independence) $P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B}\mid\mathbf{A})$
Events **A** and **B** are disjoint	When $P(\mathbf{B}\mid\mathbf{A}) = 0$ (From conditional probability formula because the intersection as shown in a Venn diagram is 0.) /*see IL notes*/
Events **A** and **B** are independent	When $P(\mathbf{B}\mid\mathbf{A}) = P(\mathbf{B})$
Tree diagram	Useful for showing sequences of (conditional) events and when utilizing the General Multiplication Rule. The probabilities of each set of branches as well as disjoint final outcomes sum to 1.
Reverse Conditioning We have $P(\mathbf{A}\mid\mathbf{B})$ but want ____ We need to find _____ and ___ With the help of ____________	 $P(\mathbf{B}\mid\mathbf{A})$ $P(\mathbf{A} \cap \mathbf{B})$, $P(\mathbf{A})$ a tree diagram
Chapter 16	**Random Variables**
Random variable Discrete random variable Continuous random variable	A variable, denoted by a capital letter (X, Y, Z etc.), whose value is a numerical outcome of a random event. The theoretical data (possible outcomes) of a probability model. Has a finite number of possible outcomes. Takes all values in an interval of numbers (infinite or bounded).
Probability model	A function that associates a probability *P* with each value of a discrete random variable *X*, denoted $P(X = x)$,

	or with any interval of values of a continuous random variable.
Probability histogram	Pictures the probability distribution of a discrete random variable. (a relative frequency histogram for a very large number of trials)
Density curve	Pictures the probability distribution of a continuous random variable (normal distributions are 1 type)
Expected value of a random variable.	The mean over the long run of a random variable. If the random variable is discrete, multiply each possible value by the probability that it occurs, and find the sum: $\mu_x = E(X) = \Sigma x_i p_i$
Variance of a random variable.	The expected value of the squared deviation from the mean $\sigma^2_x = \mathrm{Var}(X) = \Sigma(x_i - \mu_x)^2 p_i$
Standard deviation of a random variable	Describes the spread of the model $\sigma_x = \mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$
μ_{a+bX} = ______ σ_{a+bX} = ______	$a + b\mu_X$ (a and b are constants) $b\sigma_X$
μ_{X+Y} = ______ μ_{X-Y} = ______ $\sigma_{X\pm Y}$ = ______	$\mu_X + \mu_Y$ $\mu_X - \mu_Y$ $\sqrt{\sigma^2_X + \sigma^2_Y}$, if X and Y are independent. (Pythagorean Theorem of Statistics)
$X_1 + X_2 \neq$ _____ μ_{X1+X2} = ______ μ_{X1-X2} = ______ $\sigma_{X1\pm X2}$ = ______	2X , (X_1 & X_2 are distinct random variables with the same μ and σ. They aren't like terms) $\mu_{X1} + \mu_{X2} = 2\mu_X$ $\mu_{X1} - \mu_{X2} = 0$ $\sqrt{\sigma^2_{X1} + \sigma^2_{X2}} = \sqrt{2\sigma^2_X} = \sigma_X\sqrt{2}$
If two independent continuous random variables have Normal models,	So does their sum or difference.