Introduction to the Practice of Basic Statistics

1.1

Data are **numbers with a context**.

Variables

- **Individuals**- objects described by a set of data.
- **Variable**- characteristic of an individual.
- **Questions to pursue** when studying Statistics:
    - *Why? What purpose do the data have?*
    - *Who? What individuals do the data describe? How many individ.?*
    - *What? How many variables? What is the unit of measurement?*

Categorical and Quantitative Variables

A ***categorical variable*** places an indi. into one of several groups.

A ***quantitative variable*** takes numerical values for which numerical operations make sense.

The ***distribution*** of a variable tells us what values it takes and how often it takes these variables.

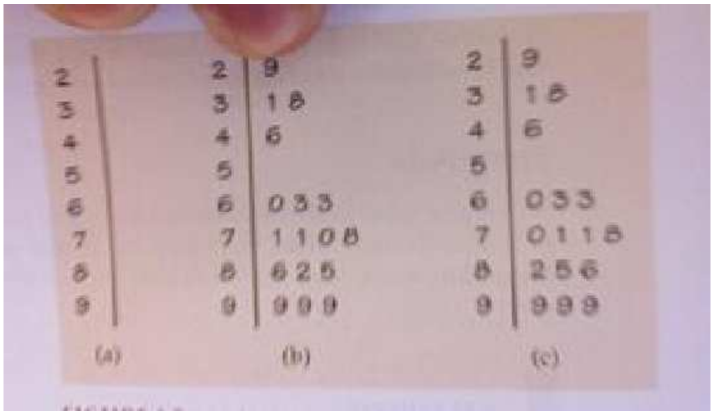
(7)

1.1 Displaying Distributions with Graphs

- **Exploratory Data Analysis**- examination of data to describe their main features,
    - Steps:
        - Examine variables, then study relationships.
        - Begin with a graph, then summarize numerical aspects.
- Charts, and especially bar graphs and pie charts, are excellent ways of representing data.

(11)

- **Stemplots**- are excellent ways to represent data, as are **back-to-back** stemplots. They do not work well for large sets of data



2	
3	
4	
5	
6	
7	
8	
9	

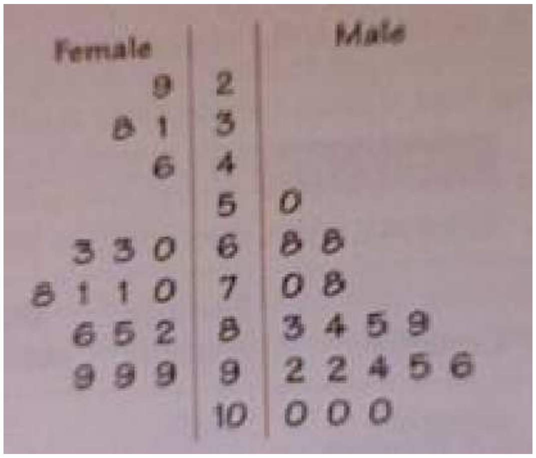
(a)

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

(b)

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

(c)



Female		Male
9	2	
8 1	3	
6	4	
	5	0
3 3 0	6	8 8
8 1 1 0	7	0 8
6 5 2	8	3 4 5 9
9 9 9	9	2 2 4 5 6
	10	0 0 0

(13)

- One can **trim** or **split** the data in order to deal with large amounts of data in a stemplot.
- **Histograms** are good for large amounts of data. One sets up "classes" of data, or simple ranges with a defined width. Individuals that fall into each class, or **frequencies**, are grouped together and then graphed on a bar graph.
- When analyzing graphs, one must look at the overall pattern for **deviations, shape, center, and spread**. Always pay attention to **outliers**.
- One must identify outliers and explain them.

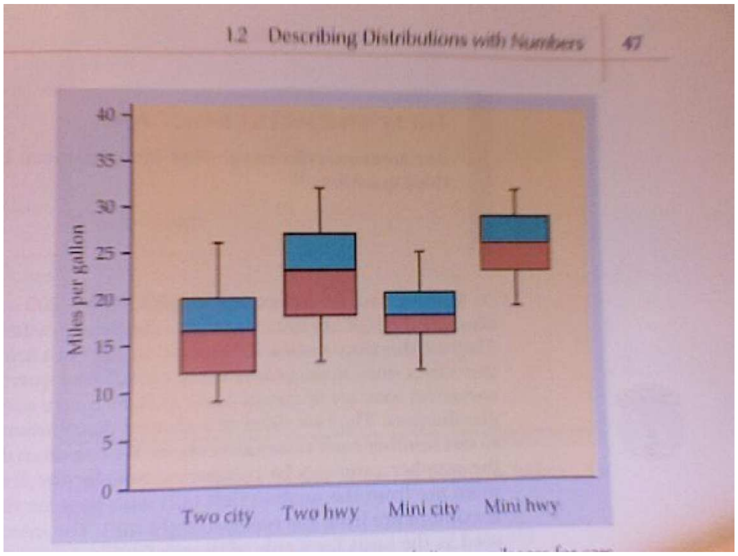
(19)

- **Time plots** are used when there is a change over time from which data is being collected.
- **Time series** are collections of measurements taken regularly over time.
- A **trend** is a persistent, long-term rise or fall.
- A patterin in a time series that repeats itself at known regular intervals of time is called **seasonal variation**. When a variation like this is pinpointed, data may be able to be **seasonally adjusted** to account for an expected difference.

1.2

- A brief description of a distribution should include **shape, center, & spread**.
- **The mean $\bar{x}$ of a set of observations is the sum of their values divided by the number of observations.**

    - $\bar{x}$ **=(1/n)Σxsubi**
- Because the mean cannot resist the influence of extreme observations like outliers, we refer to it as **not being a *resistant measure.***
- **Median**-formal version of midpoint.
    - Steps to reaching the median.
        - **Arrange** all numbers from smallest to largest.
            - If number of observations *n* is odd, the median is the midpoint, found by using (n+1)/2.
            - If the number of observations *n* is even, the median is the average of the two midpoints, found using (n+1)/2.
- The median is a **resistant measure** and is used when outliers are present as a more accurate measurement.
- If mean and median are the same, the distribution is **exactly symmetric**. If they are not, the distribution is **skewed**.
- The simplest useful numerical description of a distribution consists of both a measure of center and **a measure of spread**.
- Percentiles, such as the ***pth* percentile***,* of a distribution to describe how many observations fall below said percentile.
- Two other percentiles often used, besides the median, are the **third** and **first quartile**, 75% and 25% respectively.
    - **Q1-** median of observations before median.
    - **Q3-** median of observations after the median.
- The **five number summary**:
    - **Minimum Q1 M Q3 Maximum**
    - Gives good idea of center, spread, etc.



  - A **boxplot** is a graph of the five number summary.
    - Central box spans quartiles Q1 and Q3.
    - Line in the box marks the median *M*.
    - Lines extend from the box out to the smallest and largest observations.
    - *Modified boxplots* plot suspected outliers individually.
- The **Interquartile range** or **IQR:**
  - **IQR= Q1-Q3**
- The **1.5 x IQR rule for suspected outliers**
  - Call an observation a suspected outlier if it falls more than 1.5 x IQR above the third quartile or below the first quartile.
- **Spread** is measured through ***Stanard Deviation***

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

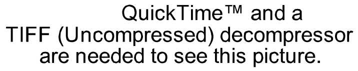
-

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

- **Finding standard deviation on a TI-83**
    - Press STAT and choose Edit from the EDIT menu
    - Enter your data in L1
    - Press 2nd QUIT
    - Press STAT and choose 1-Var Stats from the CALC menu
    - Press 2nd 1 (for L1)
    - Press Enter
- ***s,*** or **Standard Deviation**, is not resistant, and *s=0* only when there is *no spread*.
- When choosing which method of measuring spread is the most accurate, use the five number summary when outliers are present.

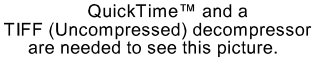
1.3

- Clear strategy for describing distributions:
    - **Always plot your data**: make a graph, usually a stemplot or histogram.
    - **Look for the overall pattern and for striking deviations** such as outliers.
    - **Calculate an appropriate numerical summary** to briefly describe center and spread.
- Fitting a smooth curve to a histogram gives us a **density curve.**
    - Density curves are done through software.
    - Smooth approximation to the irregular curvature of a histogram.
    - A density curve does not take outliers into consideration.
    - **The median of the curve** is the point with half the total area on each side. The **mean** is the balance point.


QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

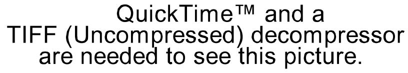
  -

- A **normal curve is bell-shaped**.
- ***μ*** is the symbol for the mean of an idealized distribution. The **standard deviation** symbol of such a curve is sigma, or (tail swing to the right).
- The **65-95-99.7 rule** shows how data should theoretically fall.


QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

-
  - Approximately **68%** of the obs fall within the mean of *μ.*
  - Approx **95%** of the obs. Fall within 2(sigma) of *μ.*
  - Approx **99.7**% of the obs fall within 3(sigma) of *μ*

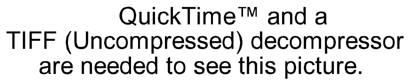
- Since normal distributions share many properties, we like to **standardize** their units to and μ:


QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

  -
  - The resulting number, the **z-score**, tells how many standard deviations the original observation falls from the mean, and in which direction.

(74)

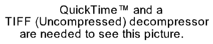
- Standarized observations from symmetric distributions are used to express things in a common scale. We might, for example, compare the heights of two children of different ages by calculating their z-score, and these heights could tell us where each child stands in the distribution of his or her age group.
- **Cumulative proportions** are areas under a distribution that show the proportion of observations in a distribution that lie at or below a given value.
- One can use a **standard normal table** from the book to find cumulative proportions
- EXMP:


QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

-

(80)

- The most useful tool for assessing normality is the **normal quantile plot**.
  - Used when the stemplot or histogram appears roughly symmetric and unimodal.
  - Process:
    - Arrange data from smallest → largest.
    - Do normal distribution calculations to find the z-scores at these same percentiles.
    - Plot each data point against the corresponding *z*.
    - Voila!


QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

  - If the points on a normal quantile plot lie close to a straight line, the plot indicates that the data are normal. Systematic deviations from a straight line indicate non-normal distribution. Outliers appear as points that are far away from the overall pattern of the plot.

Introduction to the Practice of Basic Statistics

2.1

- When examining relationships, one must keep in mind:
  - What *individuals* do the data describe?
  - What *variables* are present? How are they measured?
  - Which variables are *quantitative* and which are *categorical*?
- A **response variable** or **dependent variable** measures the outcome of a study. A **explanatory variable or independent variable** explains or causes changes in the response variable.
- The goal of a study is to determine if an explanatory variable actually ***causes*** a response variable to change.
- The most useful graph for displaying the relationship between two quantitative variables is a **scatterplot**.
  - Two variables, one set of individuals.
  - Explanatory variable goes on the X.
- Examining a scatterplot
  - Search for the overall pattern and deviations.
  - You can describe the overall pattern of a scatterplot by the **form, direction, and strength** of the relationship
- Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other and below-average values also tend to occur together.
- Two variables are **negatively associated** when above-average values of one accompany below-average values of the other, and vice-versa.
- If the relationship has a clear **direction**, we speak of either positive or negative association.

2.2

- **Correlation *r*** is the measure we use to determine the strength of a linear relationship.
  - The correlation measures the direction and strength of the linear relationship between two quantitative variables.
  - Suppose we have data on variables *x* and *y* for *n* individuals. The means and standard deviations of the two variables are xbar and s_x for the x-values and ybar and s_y for the y-values.

  - $$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$
- It makes no difference what values you make x & y.
- Positive r = positive association, negative r = negative association.
- Correlation **always between -1 and 1**.
- Correlation is **strongly affected by outliers** and is not resistant.

2.3

- A **regression line** is a straight line that describes how a response variable *y* changes as an explanatory variable *x* changes. We often use a regression line to **predict** the value of *y* for a given value of x. Regression, unlike correlation, requires that we have an explanatory variable and a response variable.
- The equation is
  - Y = a + bx
    - B is slope
    - A is the intercept
- We **extrapolate** when we use a regression line for prediction far outside the range of values of the explanatory variable *x* used to obtain the line. Such predictions are often not accurate.
- Since prediction errors most often lie in the response variables, we must find **error.**
  - ***Error = observed gain – predicted gain***
  - Errors are positive if they are above the line, and vice versa.
- The **least-squares regression line** is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.
  - Yhat= a + bx
  - b = r * $^{sy}/_{sx}$
  - a = meany – b * meanx
  - ALWAYS passes through the mean point (xbar, ybar)
- **r^2** is the square of the correlation, and it is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x.
  - Gives the amount of variation accounted for in the linear relationship.

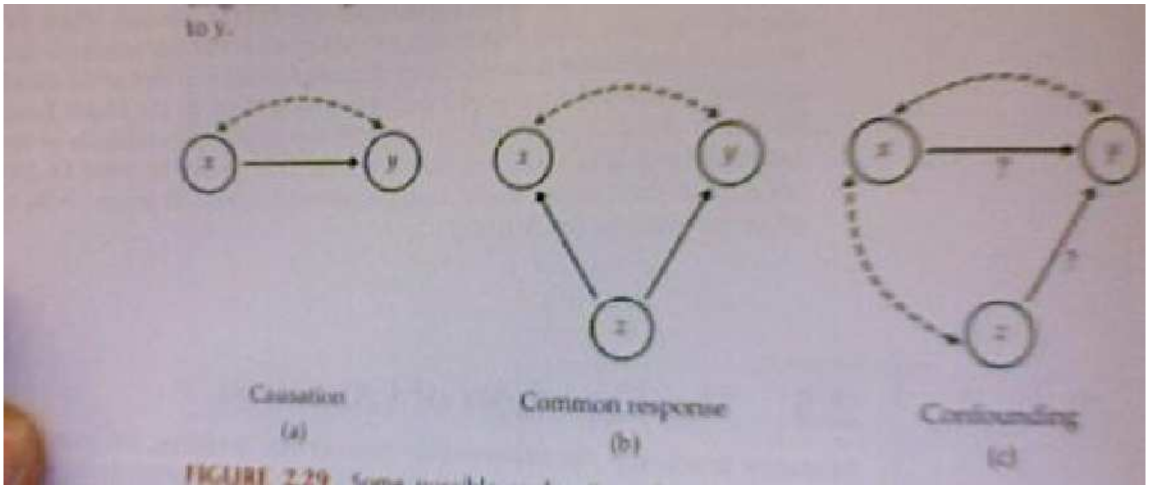
- We often **transform relationships**, especially one involving sizes. We usually use **logarithms** to rid the data of extreme outliers

2.4

- A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,
  - Residual = observed y -   predicted y
  - = y – yhat
  - **The mean of the least-squares residuals is always zero**.
- A **residual plot** is a scatterplot of the regression residuals can be made.
  - Since the mean of the residuals is always zero, a horizontal line at zero is centered in the scatter plot.
- A **lurking variable** is a variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.
- An association between an explanatory variable *x* and a response variable *y*, even if it is very strong, is not by itself good evidence that changes in x actually cause changes in *y*.
  - Those that aren't are called "nonsense variables."
- A correlation based on **averages** over many individuals is usually higher than the correlation between the same variables based on data for individuals.

2.5

- We often seek to understand if changes in the explanatory *cause* changes in the response variable. What ties between two variables, (or lurking ones), cause observed association?



Causation
(a)

Common response
(b)

Confounding
(c)

FIGURE 2.29 Some possibl

- Dashed lines mean observed association.
- Solid lines explain direct cause and effect association.
- **As we see above, it can seem apparent that in (b), x and y are related. They are, however, not. They just seem to be.**
- Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other. The confounded variables may either be explanatory variables or lurking variables.
- **Strong association between two variables is not by itself good evidence that there is a cause-effect link between the variables**.
- **Confounding** of two variables, (either explanatory or lurking), means that we cannot distinguish their effects on the response variable.
- One can establish a ***direct, causal link*** between x and y only through experiment.
- Criteria for establishing causation when we cannot do an experiment:
  - *The association is strong.*
  - *The association is consistent.*
  - *Higher does are associated with stronger responses.*
  - *The alleged cause precedes the effect in time.*
  - *The alleged clause is plausible*.

Intro to the Practice of Basic Statistics

3.1

- **Designs** focus on patterns for collecting data.
    - How many individuals will be studied?
    - How shall one select the individuals?
    - Etc.
- To rely on **anecdotal evidence** is to base evidence on haphazardly selected individual cases, which often come to our attention because they are striking in some way. These cases need not be representative of any larger group of cases.
- Many nations have federally run data collecting agencies that collect all kinds of data.
    - www.fedstats.gov
- **Available data** are data that were produced in the past for some other purpose but that may help answer a present question.
- **Sampling** for data and **experimenting** are two very essential methods of data collection.
- An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses.
- An **experiment** deliberately imposes some treatment on individuals in order to observe their responses.

3.2

- A **study** is an experiment when we actually do something to people, animals, or objects in order to observe the response.
    - **Experimental Units**- Individuals under experimentation
        - **Subjects** – Humans under expermentation
    - **Treatment** – condition applied to the units
- In principle experiments can give good evidence for causations.
- Experiments are sometimes simple, with a **Treatment → Observed Response** format.
- **Control Groups** are often used as groups that get a faux experiment to see if situations like the *placebo effect* are in action.
- The design of a study is **biased** if it systematically favors certain outcomes.

- The **design of an experiment** first describes the response variable or variables, the factors, and the layout of the treatments with comparison as the leading principle.
- To avoid bias, rely on chance. **Randomization** is the way to achieve this.
- When relying on chance, one must make the field of individuals large enough to combat the luck of the draw. (One group out of two gets a majority of the better-abled individuals.
- **Principles of Experimental Design:**
    - *Control* the effects of lurking variable son the response, most simply by comparing two or more treatments.
    - *Randomize* using impersonal chance to assign experimental units to treatments.
    - *Repeat* each treatment on many units to reduce chance variation.
- An observed effect so large that it would rarely occur by chance is called **statistically significant.**
- Often, we run **double-blind** experiments, especially in medicine, in which neither the subjects nor the personnel who worked with them knew which treatment any subject had received.
- A serious weakness in experimentation is a **lack of realism**, in which the the subjects or treatments or setting of an experiment may not realistically duplicate the conditions we really want to study.
    - One cannot always generalize conclusions to some setting wider than that of the actual experiment.
- For even more accuracy than a completely randomized experiment, one may use **matched pair designs** which compares just two treatments.
    - Often matches individuals that are similar in many categories.
    - Because the two are so similar, the comparison is more efficient.
- **Block designs** are experiments in which a *block* of experimental units that are known to be similar in some way before the experiment are grouped together because they are expected to affect the response to the treatments

3.3

- The entire group of individuals that we want information about is called the **population**.
- A **sample** is part of the population that we actually examine in order to gather information.

- The design of a sample survey refers to the method used to choose the sample from the population.
- A **voluntary response sample** consists of people who choose themselves by responding to a general appeal. Voluntary response samples are biased because people with strong opinions, especially negative opinions, are likely to respond.
- A **simple random sample (SRS)** of size *n* consists of *n* individuals from the population chosen in such a way that every set on *n* individuals has an equal chance to be the sample actually selected.
- A **probability sample** is a sample chosen by chance. We must know what samples are possible and what chance, or probability, each possible sample has.
- To select a **stratified random sample**, first divide the population into groups of similar individuals, called **strata**. Then, choose a separate SRS in each stratum and combine these SRSs to form the full sample.
- A common means of restricting random selection is to choose the sample in stages.
    - *Used for households or people.*
- A **multistage sampling design** selects successively smaller groups within the population in stages, resulting in a sample consisting of clusters of individuals. Each stage may employ an SRS, a stratified sample, or another type of sample.
- **Undercoverage** occurs when some groups in the population are left out of the process of choosing the sample.
- **Nonresponse** occurs when an individual chosen for the sample can’t be contacted or does not cooperate.
- **Response bias** is often the fault of behavior by either the respondent or the interviewer.
    - Illegal or unpopular behavior often incites lies.
- The **wording of questions** is the most important influence on the answers given to a sample survey.
    - Confusing or leading questions can introduce small bias.

3.4

- **Statistical inference** is when assumptions about a larger population are made through the results of the smaller, sampled population.

- A **parameter** is a number that describes the population. A parameter is a fixed number, but in practice we do not know it's value.
- A **statistic** is a number that describes a sample. We often use a statistic to estimate an unknown parameter.
- Often, we use **sample variability** to make many samples. This helps us get a better picture of a possible trend and to account for changes form sample to sample. A **histogram** is then made.
- Process:
    - Take a large number of samples
    - Calculate the sample proportion phat for each sample.
    - Make a histogram of the values of phat.
    - Describe the histogram.
- Using random digits from a table or computer software to initiate chance behavior is called **simulation**.
- The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.
- **Bias** concerns the center of the sampling distribution. A statistic used to estimate a parameter is **unbiased** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.
- The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined by the sampling design and the sample size *n*. Statistics from larger probability samples have smaller

spreads.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

- **To reduce bias**, use random sampling. When we start with a list of the entire population, simple random sampling produces unbiased estimates- the values of a statistic computer from an SRS neither consistently overestimate nor consistently underestimate the value of the population parameter.
- **To reduce the variability** of a statistic from an SRS, use a larger sample. You can make the variability as small as you want by taking a large enough sample.
- **Large SRSs** are often a close estimate of the truth.
- **A margin of error** comes from the results and sets bounds of the size of the likely error.

The variability of a statistic from a random sample does not depend on the size of the population, as long as the population is at least 100 times larger than the sample.

Intro to the Practice of Basic Statistics

4.1

- **Random** in statistics does not mean haphazard, but rather a kind of order that emerges in the long run. We call a phenomenon random if individual outcomes are uncertain but there is nonetheless a regular distribution of outcomes in a large number of repetitions.
- The **probability** of any outcome of a random phenomenon is the proportion of times the outcome would occur in a very long series of repetitions.
- A long series of **independent trials** that don't affect each other must be enacted before we can study probability of an outcome effectively.

4.2

- **Probability models** are descriptions of random phenomenon in the language of math.
  - List of all outcomes and the probability of each outcome.
- The **sample space *S*** of a random phenomenon is the set of all possible outcomes.
  - ***S* = {………………………………}**
- An **event** is an outcome or a set of outcomes of a random phenomenon. That is, an event is a subset of the sample space.
- **Rules of probabilities:**
  - ***A probability is a number between 0 & 1.***
  - ***All possible outcomes together must have probability 1.***
  - ***If two events have no outcomes in common, the probability that one or the other occurs is the sum of their individual probabilities.***
  - ***The probability that an event does not occur is 1 minus the probability that the event does occur.***
  - ***Multiplication Rule***

    Two events A and B are independent if knowing that one occurs does not change the probability that the other occurs.
    - *P(A and B) = P (A)P(B)*
  - ***Compliment Rule***

    The complement of any event A is the event that A does not occur, written as A^c= 1- P(A)

- If a random phenomenon has *k* possible outcomes, all equally likely, thene ach individual outcome has probability 1/*k*. The probability of any event *A* is:
    - *P(A)* = Count of outcomes in A / Count of Outcomes in S
        - = *Count of outcomes in A / k*

4.3

- A **random variable** is a variable whose value is a numerical outcome of a random phenomenon.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

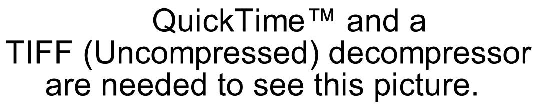
-
- **Probability Histograms** compare the probability model for random digits.
- A **continuous random variable *X*** takes all values in an interval of numbers. The **probability distribution of *X*** is described by a density curve. The probability of any event is the area under the density curve and above the values of X that make up the event.
- **All continuous probability distributions assign probability 0 to every individual outcome**.

4.3

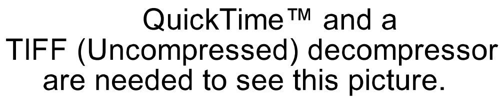
- The mean of a probability distribution **describes the long-run average outcome**, **not the direct center of the data.**
- We cannot refer to the mean of a probability distribution as "xbar," for the mean of a probability distribution does not describe data that has an equal likelihood of happening. Rather, **the mean of a probability distribution is referred to as μ, and sometimes μ_X**

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

-
- The mean of a random variable *X* is often called **the expected value of *X*.**
- μ is a **parameter** and xbar is a **statistic**.
- **Law of Large Numbers:**
  - *Draw independent observations at random from any population with finite mean* μ. *As the number of observed values eventually approaches the mean of the draw increases, the mean*xbar *of the observed values eventually approaches the mean* μ *of the population as closely as you specified and then* ***stays that close**.*
- The more variable the outcomes, **the more trials are needed** to ensure that the mean outcome *xbar* is close to the distribution mean *μ*.


QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

-
- We write the variance as .


QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

-
- If random variables are independent, this kind of association between their values is ruled out and their variances do add. Two random variables *X and Y* are **independent** if knowing that any event involving X alone did or did not occur tells us nothing about the occurrence of any event involving *Y* alone.
- When random variables are not independent, the variance of their sum depends on the **correlation** between them as well as on their individual variables.

- **The correlation between two independent random variables is zero**.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Intro to the Practice of Basic Statistics

Introduction

- The link between probability and data is formed by the ***sampling distributions*** of statistics. It is an idealized mathematical description of the results of an indefinitely large number of samples, rather than the results of a particular 1000 samples.

The Distribution of a Statistic

A statistic from a random sample or randomized experiment is a random variable. The probability distribution of the statistic is its **sampling distribution**.

Population Distribution

The **population distribution** of a variable is the distribution of its values for all members of the population. The population distribution is also the probability distribution of the variable when we choose one individual at random from the population.

5.1

- The **count** is the number of occurrences of some outcome in a fixed number of observations. It is often a random variable, like *X*.
- If the number of observations is *n*, then the **sample proportion** is *phat=X/n*.
  - For example, if three hundred people were asked if they agree with a statement or not, and 200 say they agree, then the **count** is 200 and the **sample proportion** is phat = (200)/(300) = 0.66.
- The distribution of a count *X* depends on how the data are produced.

The Binomial Setting

1. There are a fixed number *n* of observations**.**

2. The *n* observations are all independent.

3. Each observation falls into one of just two categories, which for convenience we call “success” and “failure.”

4. The probability of a success, call it *p*, is the same for each observation.

Binomial Distributions

The distribution of the count X of success in the binomial setting is called the **binomial distribution** with parameters n and p. The parameter n is the number of observations, and p is the probability of a success on any one observation. The possible values of X are the whole numbers from 0 to n. As an abbreviation, we say that X is $B(n,p)$.

- ***The most important skill for using binomial distribution sis the ability to recognize the situations to which they do an don't apply.***
- Choosing an SRS from a population is not quite a binomial setting.

Sampling Distribution of a Count

A population contains proportion of p of successes. If the population is much larger than the sample, the count X of successes in a SRS of size n has approximately the binomial distribution $B(n,p)$.

The accuracy of this approximation improves as the size of the population increases relative to the size of the sample. As a rule of thumb, we will use the binomial sampling distribution for counts when the population is at least 20 times as large as the sample.

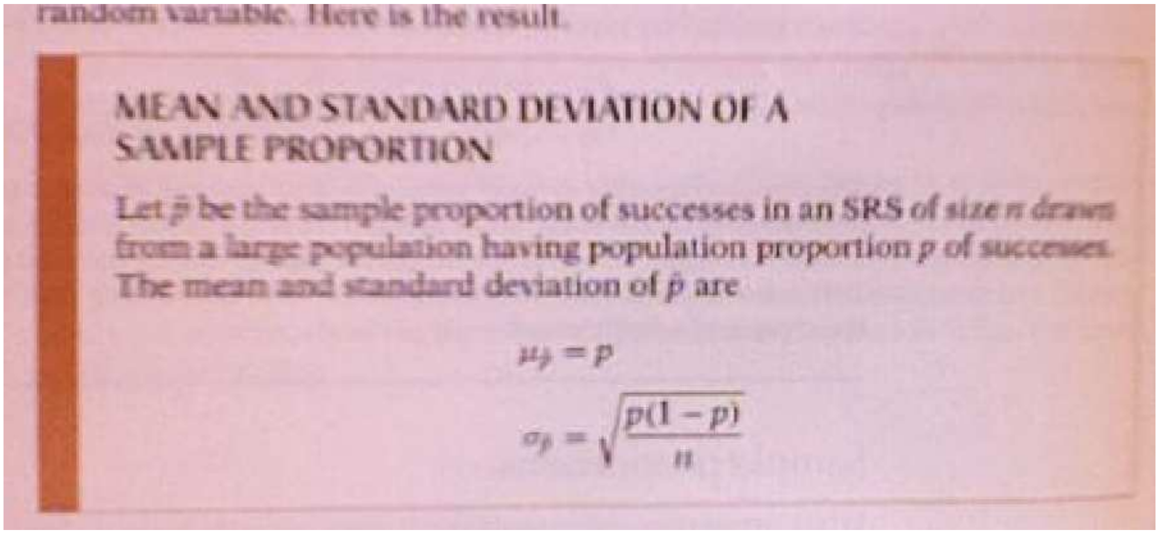
Binomial Mean and Standard Deviation

If a count X has the binomial distribution $B(n,p)$, *then*

$$\mu_x = np$$

$$\sigma_x = (np(1-p))^{1/2}$$

- **Proportion is ALWAYS a number between 0 and 1. Do NOT get it confused with the count.**



random variable. Here is the result.

MEAN AND STANDARD DEVIATION OF A SAMPLE PROPORTION

Let $\hat{p}$ be the sample proportion of successes in an SRS *of size n drawn* from a large population having population proportion *p of successes.* The mean and standard deviation of $\hat{p}$ are

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

5.2

- **Averages are less variable than individual observations.**
- **Averages are more normal than individual observations.**

Mean and Standard Deviation of a Sample Mean

Let xbar be the mean of an SRS of size *n* from a population having mean *μ* and standard deviation σ. The mean and standard deviation of xbar are:

$$\mu_{xbar} = \mu$$

$$\sigma_{xbar} = \sigma / (\sqrt{n})$$

Sampling Distribution of a Sample Mean

If a population has the *N(μ,* σ*),* then the sample mean xbar *of n* independent observations has the *N(μ, (* σ */ √n))*

Central Limit Theorem

Draw an SRS of size *n* from any population with mean μ and finite standard deviation σ. When *n* is large, the sampling distribution of the sample mean xbar is approximately normal:

Central Limit Theorem

Draw an SRS of size *n* from any population with mean μ and finite standard

deviation σ. When *n* is large, the sampling distribution of the sample mean xbar is approximately normal:

Xbar is approximately *n (μ, (* σ*/ (√n))*

- **The normal approximation for the sample proportions and counts is an example of the central limit theorem.**
- **More general versions of the central limit theorem say that the distribution of a sum or average of many small random quantities is close to normal.**
- **Any linear combination of independent normal random variables is also normally distributed**

Intro to the Practice of Basic Statistics

Introduction

- When you use statistical inference, you are acting **as if the data come from a random sample or a randomized experiment**.

6.1

- An estimate without an indication of its variability is of little value.
- The **68-95-99.7 rule** comes back into play here. When looking at a normal distribution, the probability is that 95 percent of your data will fall within 2 standard deviations of the mean.
- If we're looking at a sample mean "xbar," we know that that is a sample estimator of the population mean μ. We can safely say in a large enough sample that the true μ can be found within 95 percent of the data, or within two standard deviations. This is called a **95% confidence interval for μ.**
  - **Estimate ± margin of error**

Confidence Interval

A level *C* **confidence interval** for a parameter is an interval computed from a sample data by a method that has probability ***C*** of producing an interval containing the true value of the parameter.

1. It is an interval of the form (a,b), where a and b are numbers computed from the data..
2. It has a property called a confidence level that gives the probability that interval covers the parameter.
3. A confidence level **does not have to be 95%**

- Any normal curve has probability *C* between the point *z** standard deviations below the mean and the point *z** above the mean.
- ***z**** is simply the exact amount of standard deviations away from the mean could be. Refer to **table D in the back of the book**.
- The unknown population mean μ lies between

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

  -

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

- **Margin of Error too large?**
    - Use a lower level of confidence (smaller C)
    - Increase the sample size (larger n)
    - Reduce .
    - A wise user of statistics never plans data collection without at the same time planning the inference. You can arrange to have both high confidence and a small margin of error. **To obtain a desired margin of error *m*, just use the equation for margin of error above, substitute the value of *z** for your desired confidence level, and solve for the sample size n.**

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

- It is always safe to round **up** to the next higher whole number when finding n because this will give us a small margin of error.
- **CAUTION:**

- **THESE CONDITIONS MUST BE MET TO USE A FORMULA FOR INFERENCE.**
  - The data should be an **SRS from the population**.
  - The formula is **not correct for probability sampling designs more complex than SRS**.
  - There is **no correct method for inference for haphazardly collected data with bias of unknown size**.
  - Because "xbar" is not resistant, **outliers can have a large effect** on the confidence interval.
  - If the **sample size is small and the population is not normal**, the true **confidence level will be different** from the value C used in computing the interval.
  - **You must know the standard deviation of the population**.

6.2

- Confidence intervals are appropriate when our goal is to estimate a population parameter. The second common type of inference is directed at a quite different goal: **to assess the evidence provided by the data in favor of some claim about the population.**
- A **significance** test is a formal procedure for comparing observed data with a hypothesis whose truth we want to assess.
- The **hypothesis** is a statement about the parameters in a population.

Null Hypothesis

The statement being tested in a test of significance is called the **null hypothesis**. The test of significance is designed to assess the strength of the evidence against the null hypothesis. Usually the null hypothesis is a statement of "no effect" or "no difference."

We abbreviate this as H_0.

- H_A is the **alternative hypothesis** and it states that the statement we hope or suspect is true instead of the null hypothesis.
- Hypotheses always refer to some population or model, not to a particular outcome. **For this reason, we must state the null and the alternative hypotheses in terms of population parameters**.
- It is not always clear, in particular, whether the alternative hypothesis should be **one sided or two sided**, which refers to whether a parameter

differs from its null hypothesis value in a specific direction or in either direction.

- **Test Statistics**
  - The test is based on a statistic that estimates the parameter that appears in the hypotheses. Usually this is the same estimate we would use in a confidence interval for the parameter. When the null hypothesis is true, we expect the estimate to take a value near the parameter value specified by the null hypothesis.
  - Values of the estimate far from the parameter value specified by the null hypothesis give evidence against the null hypothesis. The alternative hypothesis determines which directions count against the null hypothesis.
  - To assess how far the estimate is from the parameter, standardize the estimate. In many common stat situations, the test statistic has the form:
    - ***z* = (estimate-hypothesized value)÷(standard deviation of the estimate)**

- A **test statistic** measures compatibility between the null hypothesis and the data. We use it for the probability calculation that we need for our test of significance. It is a random variable with a distribution that we know.
- A test of significance finds the probability **as extreme or more extreme than the actual observed outcome**.

P-Value
The probability, computed assuming that the null value is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the **P-value** of the test. The smaller the P-value, the stronger the evidence against the null value provided by the data.

- After one translates the test statistic into a P-value to quantify the evidence of the null hypothesis, we need to **compare the P-value we calculated with a fixed value that we regard as decisive.** This is called the **significance level**, or α.
- When we choose a significance level, we are saying **in advance how much evidence against the null hypothesis will be needed to reject it.**

- For instance, if we say that α = 0.05, then we are requiring that **the data give evidence against the null hypothesis so strong that it would happen no more than 5% of the time**.

Statistic Significange

If the P-value is as small or smaller than α, we say that the data are **statistically significant at level α.**

- **Four steps common to all tests of significance**
  - State the ***null hypothesis and the alternative hypothesis****.*
  - Calculate the value of the ***test statistic*** on which the test will be based. This statistic usually measures how far the data are from the null hypothesis.
  - **Find the P-value** for the observed data. This is the probability, calculated assuming that the null hypothesis is true, that the test statistic will weigh against the null hypothesis at least as strongly as it does for these data.
  - **State a conclusion!** You can choose a significance level α, how much evidence against the null hypothesis, you regard as decisive. **If the P-value is less than or equal to α, you conclude that the alternative hypothesis is true. If it is greater than α, you conclude that the data do not provide sufficient evidence to reject the null hypothesis**.
    - Sentence form.

Z Test for a Population Mean

To test the hypothesis $H_o : \mu = \mu_o$ based on an SRS of size n from a population with unknown mean μ and standard deviation , compute the test statistic

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

In terms of a standard normal random variable *Z*, the P-value for a test of the null hypothesis against

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

These P-values are exact if the population distribution is normal and are approximately correct for large *n* in other cases.

Two-Sided Significance Tests and Confidence Intervals

A level α two-sided significance test **rejects a hypothesis $H_o : \mu = \mu_o$ exactly when the value μ_o falls outside a level 1- α confidence interval for μ.**

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

- **The P-value is the smallest level** α at which the data are significant.

6.3

- **There is no sharp border between “significant” and “not significant,”** only increasingly strong evidence as P-value decreases.
- We use the **5% standard** often.
- Formal statistical inference cannot correct basic flaws in the design.

We must often analyze data that do not arise from randomized samples or experiments. To apply statistical inference to such data, we must have confidence in a probability model for the data.

Intro to the Practice of Basic Statistics

7.1

Standard Error

When the standard deviation of a statistic is estimated from the data, the result is called the **standard error** of the statistic.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

- **When σ is not known**, we estimate it with the sample standard deviation *s*, and then we estimate the standard deviation of xbar by *s*/√*n*. This quantity is the standard error of the sample mean xbar seen above.
- When we substitute the standard error *s*/√*n* for the standard deviation σ / √*n* of xbar, the **statistic does *not* have a normal distribution**. It has a distribution called a ***t-distribution****.*

The t Distributions

Suppose that an SRS of size *n* is drawn from an *N* (*μ*, σ*)* population. Then, the **one-sample *t* statistic**

t = *(xbar-μ) / (s/√n)*

has the ***t* distribution** with *n*-1 **degrees of freedom**.

- **Degrees of freedom** are a measure of how much precision an estimate of variability has.
- **Table D** on page T-11 in the back of the book gives **critical values** for the *t* distribution.

The One-Sample t Confidence Interval

Suppose that an SRS of size *n* is drawn from a population having unknown mean *μ*. A level *C* **confidence interval** for *μ* is

$$Xbar \pm t^* (s/\sqrt{n})$$

where *t** is the value for the *t*(*n-1*) density curve with area *C* between *–t** & *t**. The quantity

$$t^* (s/\sqrt{n})$$

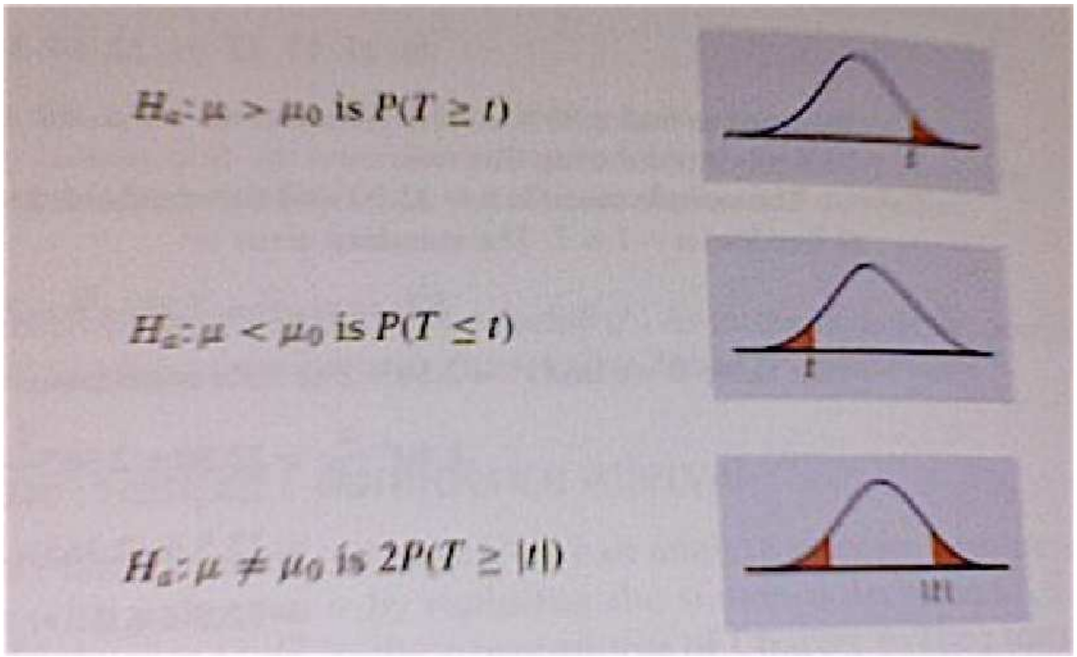
is the **margin of error**. This interval is exact when the population distribution is normal and is approximately correct for large *n* in other cases.

Z Test for a Population Mean

- Suppose that an SRS of size *n* is drawn from a population having unknown mean *μ*. To test the hypothesis $H_o : \mu = \mu_o$ based on an SRS of size *n*, compute the one-sample *t* statistic

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

-
- In terms of a standard normal random variable *T* having the *t(n – 1)* distribution, the P-value for a test of the null hypothesis against



$H_a: \mu > \mu_0$ is $P(T \geq t)$

$H_a: \mu < \mu_0$ is $P(T \leq t)$

$H_a: \mu \neq \mu_0$ is $2P(T \geq |t|)$

- These P-values are exact if the population distribution is normal and are approximately correct for large *n* in other cases.

- It is wrong to examine the data first and then decide to do a one-sided test in the direction indicated by the data.
- In a **matched pairs** study, subjects are matched in pairs and the outcomes are compare within each matched pair.