

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

**HTTP://NOTES.BOOKHOLDERS.COM**

Statistics First Semester

### Statistics

- Statistics is the science of collecting, organizing, interpreting, and learning from data
- **Descriptive statistics** (coping with lots of numbers)
  - Draw a picture (graph, charts etc)
  - Calculate a few numbers which summarize the data (mean, median, percentile)
- **Inferential statistics**
  - How can one make decisions and predictions about a population even if we have data for relatively few subjects from that population?
  - We need to generalize the facts we learn from a sample ( i.e. a part of the population) to the entire population

**“Five W’s”** → Who (leftmost column of a table), What, When, Where, Why, and How

**Respondents** → Individuals who respond to a survey

**Subjects or participants** → People on whom we experiment

**Experimental units** → animals, plants, and other inanimate subjects

**Records** → rows

**Relational database** → two or more separate data tables are linked together so that information can be merged across them

**Data** → the value of the variables

### Variables

- What has been measured (columns of a data table)
- The aspect/characteristic that differs from subject to subject, individual to individual.
- **Categorical**
  - When a variable names categories and answers questions about how cases fall into those categories
  - **Ordinal variables**
    - Categories that have a natural ordering
    - Numbers could be assigned to categories
  - **Nominal variables**
    - Categories that have no natural ordering
- **Quantitative**
  - When a variable has measure numerical values with units and the variable tells us about the quantity of what is measured
  - Units tell how each value has been measured
    - Tell us the scale of measurement
    - Tell us how much of something we have or how far apart the two values are
  - **Continuous Variables**
    - May take on any value in some interval (numbers are close together)
    - Summarized in a grouped data frequency table (daily high temperatures)

**HTTP://NOTES.BOOKHOLDERS.COM**

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.

Accept books no matter where purchased.

Up to 3x more than Cash Now.

[www.BookHolders.com](http://www.BookHolders.com)

## HTTP://NOTES.BOOKHOLDERS.COM

- Discrete Variables
  - There is a natural gap between the values
  - May take on any value in some interval (numbers are close together)
  - Summarized in a grouped data frequency table (daily high temperatures)
- Interval data
  - No meaningful zero point; can't multiply or divide but the difference between two values is meaningful (temp.)
- Ratio data
  - Meaningful zero point; can multiply and divide (weight)
- Time series data
  - Ordered data values over time
- Cross sectional data
  - Data values observed at a single point in time
- Bias → **BAD!!** Sample doesn't represent the population
  - Selection Bias
    - Problem in sampling scheme; systematic tendency to exclude one kind of individual from the survey
    - Difference between population of interest and effective population
  - Non-response Bias
    - Subjects don't answer
    - Skip questions
  - Response Bias
    - Subjects lie
    - Interviewer effect
- Randomize
  - Randomization can protect you against factors that you know are in the data.
  - It can also help protect against factors you are not even aware of
  - Randomization gets rid of biases
  - Randomizing makes sure that *on the average* the sample looks like the rest of the population
  - Sample-to-sample differences are referred to as *sampling error*
- Sample Size
  - It is the size of the sample, not the size of the population, that makes the difference in sampling.
- Population
  - The entire group of individuals in which we are interested but can't usually assess directly.
- Sample
  - The part of the population we actually examine and for which we do have data.
- Parameter
  - Number describing a characteristic of the population.
- Statistic
  - Number describing a characteristic of a sample.

HTTP://NOTES.BOOKHOLDERS.COM

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold →

**Cash Now**

Instant Cash

Accept all books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

## HTTP://NOTES.BOOKHOLDERS.COM

- **Sampling Techniques**
  - **Nonstatistical Sampling**
    - **Convenience**
      - Collected in the most convenient manner for the researcher (ask whoever is around)
      - **Bias:** Opinions limited to individuals present
    - **Voluntary**
      - Individuals choose to be involved. These samples are very susceptible to being biased because different people are motivated to respond or not. Often called “public opinion polls,” these are not considered valid or scientific
      - **Bias:** Sample design systematically favors a particular outcome
  - **Statistical Sampling**
    - **Simple Random Sampling**
      - Every possible sample of a given size has an equal chance of being selected
      - The simplest way to obtain a sample is to draw names out of a hat
      - The sample can be obtained using a table of random numbers or computer random number generator
    - **Stratified**
      - Divide population into subgroups (called *strata*) according to some common characteristic
        - e.g., gender, income level
      - Select a simple random sample from each subgroup
      - Combine samples from subgroups into one
    - **Systematic**
      - Decide on sample size:  $n$
      - Divide ordered (e.g., alphabetical) frame of  $N$  individuals into groups of  $k$  individuals:  $k=N/n$
      - Randomly select one individual from the 1<sup>st</sup> group
      - Select every  $k^{\text{th}}$  individual thereafter
    - **Cluster**
      - Divide population into several “clusters,” each representative of the population (e.g., county)
      - Select a simple random sample of clusters
        - All items in the selected clusters can be used, or items can be chosen from a cluster using another probability sampling technique

### Survey Design

- Define the issue
- Define the population of interest
- Develop survey questions
- Pre-test the survey
- Determine the sample size and sampling method

HTTP://NOTES.BOOKHOLDERS.COM

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.

Accept books no matter where purchased.

Up to 3x more than Cash Now.

**SELLS**

[www.BookHolders.com](http://www.BookHolders.com)

**[HTTP://NOTES.BOOKHOLDERS.COM](http://NOTES.BOOKHOLDERS.COM)**

- Select sample and administer the survey

**Explanatory variable**

- Predictor, “cause”, available variable

**Response variable**

- Predicted, “effect”, interesting variable

**Bar graphs and pie charts**

- Many varieties, actual form of the graph depends on the use
- Height of bar or size of pie slice shows the frequency or percentage for each category (area principle)
- What is the graph communicating?

**Marginal Distributions**

- We can look at each categorical variable separately in a two-way table by studying the row totals and the column totals. They represent the marginal distributions, expressed in counts or percentages (They are written as if in a margin.)

**Simpson’s Paradox**

- An association or comparison that holds for all of several groups can reverse direction when the data are combined (aggregated) to form a single group. This reversal is called **Simpson’s paradox**.

**Symmetric or Skewed**

- A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.
- A distribution is **skewed to the right** if the right side of the histogram (side with larger values) extends much farther out than the left side
- It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.
- Mean moves in the direction of a skewed distribution

**Bimodal or Uniform Distribution**

**Outliers**

- Data points that don’t seem to fit in the distribution.
- Far to the left or right in the graph.

**5 Number Summary**

- Min, Q1, Med, Q3, Max

**[HTTP://NOTES.BOOKHOLDERS.COM](http://NOTES.BOOKHOLDERS.COM)**

## How to Buy Textbooks

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

## How to Sell Textbooks @ BookHolders

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

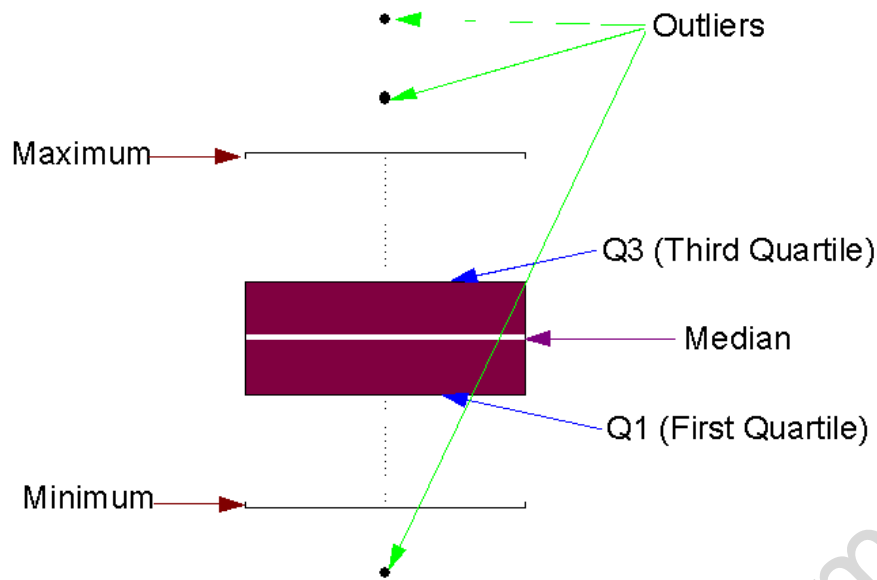
Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

[HTTP://NOTES.BOOKHOLDERS.COM](http://NOTES.BOOKHOLDERS.COM)



### Measuring the Spread

- 1. Range
  - Maximum – Minimum
- 2. InterQuartile Range
  - Q3 – Q1
- 3. Standard Deviation
  - Average squared distance from the mean

### Variance

- Average of the squared deviations
- Shows variation about the mean
- Sample variance:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

### Standard Deviation

- Square root of the variance
- Has the same units as the original data

[HTTP://NOTES.BOOKHOLDERS.COM](http://NOTES.BOOKHOLDERS.COM)

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept all books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

**HTTP://NOTES.BOOKHOLDERS.COM**

- Sample standard deviation:

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

**Z-score**

$$Z = \frac{\text{value} - \text{mean}}{\text{st.dev.}}$$

### Explanatory and Response Variable

- A **response variable** measures or records an outcome of a study. An **explanatory variable** explains changes in the response variable.
- Typically, the *explanatory* or *independent variable* is plotted on the *x* axis, and the *response* or *dependent variable* is plotted on the *y* axis.

### Interpreting Scatterplots

- After plotting two variables on a scatterplot, we describe the relationship by examining the form, direction, and strength of the association. We look for an overall pattern ...
  - Form: linear, curved, clusters, no pattern
  - Direction: positive, negative, no direction
  - Strength: how closely the points fit the “form”
    - With a strong relationship, you can get a pretty good estimate of *y* if you know *x*.
  - and deviations from that pattern (outliers) → In a scatterplot, outliers are points that fall outside of the overall pattern of the relationship.
- Linear, non-linear, and no-relationship (*X* and *Y* vary independently. Knowing *X* tells you nothing about *Y*.)

### Sum of Cross Products

- The cross-products of the z-scores showing a positive relationship will be positive and those showing a negative relationship will be negative
- The more positive cross-products we have, the greater their sum will be
- Ratio is correlation coefficient

$$r = \frac{\sum z_x z_y}{n-1}$$

- The correlation coefficient “*r*”

**HTTP://NOTES.BOOKHOLDERS.COM**

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

## HTTP://NOTES.BOOKHOLDERS.COM

- $r$  does not distinguish between  $x$  and  $y$
- $r$  has no units of measurement
- $r$  ranges from  $-1$  to  $+1$
- Correlation of zero means no linear relationship
- Correlation is not affected by changes in the center or scale of either variable
- Correlation is sensitive to unusual observations
- " $r$ " quantifies the **strength** and **direction** of a linear relationship between 2 quantitative variables.
  - **Strength:** how closely the points follow a straight line.
  - **Direction:** is positive when individuals with higher  $X$  values tend to have higher values of  $Y$ .
- No matter how strong the association,  $r$  does not describe curved relationships
  - Note: You can sometimes transform a non-linear association to a linear form, for instance by taking the logarithm. You can then calculate a correlation using the transformed data.
- **Lurking Variable**
  - A **lurking variable** is a variable not included in the study design that does have an effect on the variables studied.
  - Lurking variables can *falsely suggest* a relationship.
- Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other. The confounded variables may be either explanatory variables or lurking variables.
- **Association is not causation.** Even if an association is very strong, this is not by itself good evidence that a change in  $x$  will cause a change in  $y$ .
- **Regression Line**
  - 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  $x$ .
  - In regression, the distinction between explanatory and response variables is important.
  - The least-squares regression line is the unique line such that the sum of the squared vertical ( $y$ ) distances between the data points and the line is as small as possible.

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

- $y(\hat{\ })$  is the predicted  $y$  variable
- $b(0)$  is the  $y$ -intercept →  $b_0$  is the estimated average value of  $Y$  when the value of  $X$  is zero (if  $x = 0$  is in the range of observed  $x$  values)
- $b(1)$  is the slope →  $b_1$  measures the estimated change in the average value of  $Y$  as a result of a one-unit change in  $X$
- **Extrapolation** is the use of a regression line for predictions *outside the range of  $x$  values* used to obtain the line.
- $r^2$ , the **coefficient of determination**, is the square of the correlation coefficient.

HTTP://NOTES.BOOKHOLDERS.COM

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

**HTTP://NOTES.BOOKHOLDERS.COM**

- $r^2$  represents **the percentage of the variance in y** (vertical scatter from the regression line) **that can be explained by changes in x**.

### Residuals

- The distances from each point to the least-squares regression line give us potentially useful information about the contribution of individual data points to the overall pattern of scatter. These distances are called **“residuals.”**
- Standard deviation of the residual →

$$s_e = \sqrt{\frac{\sum e^2}{n-2}}$$

- Residuals Plot
  - Residuals are the distances between  $y$ -observed and  $y$ -predicted. We plot them in a **residual plot**.
  - If residuals are scattered randomly around 0, chances are your data fit a linear model, was normally distributed, and you didn't have outliers
  - Curved pattern—means the relationship you are looking at is not linear.

### How to:

First we calculate the **slope of the line,  $b_1$** ; from statistics we already know:

$$b_1 = r \frac{s_y}{s_x}$$

$r$  is the correlation.

$s_y$  is the standard deviation of the response variable  $y$ .

$s_x$  is the the standard deviation of the explanatory variable  $x$ .

Once we know  $b_1$ , the slope, we can calculate  $b_0$ , the **y-intercept**:

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

where  $\bar{x}$  and  $\bar{y}$  are the sample means of the  $x$  and  $y$  variables

### Randomness and Probability

- A phenomenon is **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 can be defined as the proportion of times the outcome would occur in a very long series of repetitions.

**HTTP://NOTES.BOOKHOLDERS.COM**

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

**HTTP://NOTES.BOOKHOLDERS.COM**  
Probability

$$P(A) = \frac{\text{Size of the Event A}}{\text{Size of the Sample Space S}}$$

•

### Simple Case

- If **outcomes** are **equally likely**

$$P(A) = \frac{\# \text{ outcomes in A}}{\text{Total \# outcomes}}$$

•

- Because some outcome must occur on every trial, the sum of the probabilities for all possible outcomes (the sample space) must be exactly 1.

•

### The Addition (OR) Rule

- **Mutually Exclusive**
  - Events contain no common outcomes
  - Intersection is empty
  - They can't both happen
- For **mutually exclusive** events A,B  
 $P(A \text{ or } B) = P(A) + P(B)$
- Simple events are mutually exclusive
  - Equally likely
  - $P(\text{Simple Event}) = 1/(\text{total \# of outcomes})$
- Complement of A
  - All outcomes **not** in A
  - $A^c$
  - $P(A^c) = 1 - P(A)$

### General Addition (OR) Rule

- For any events A, B  
 $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Independence

[HTTP://NOTES.BOOKHOLDERS.COM](http://NOTES.BOOKHOLDERS.COM)

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

**SALES**

[www.BookHolders.com](http://www.BookHolders.com)

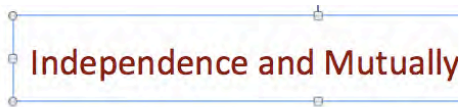
**HTTP://NOTES.BOOKHOLDERS.COM**

- A and B independent
- Are not related
- Knowing A does not give information about B
- A does not affect B

**Multiplication (AND) Rule**

- $P(A \text{ and } B) = P(A) P(B)$   
– If and only if A and B are independent.

**General Multiplication (AND) Rule**



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

- **Mutually Exclusive** : they can't both
  - $P(A \text{ and } B) = 0$
  - They aren't independent
- **OR rule for Independent Events**  
–  $P(A \text{ or } B) = P(A) + P(B) - P(A) P(B)$
- Also  
–  $P(A \text{ and } B) = P(B) P(A|B)$

**Conditional probability**

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

If  $P(A) > 0$  then

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

**Dependence**

- A says something about B
- $P(B|A)$  = conditional probability of B given A
- If A and B are independent,

$P(B|A) = P(B)$

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

**HTTP://NOTES.BOOKHOLDERS.COM**

### Tree Diagram

### Probability Distribution Function

- Pdf specifies the possible values a random variable can take on and their probabilities
  - 1. Prob.'s add up to 1.
  - $0 \leq P(X=k) \leq 1$  for every k.
  - 3. Other #'s have prob. = 0

### Expected Value of a Discrete Random Variable

- Mean of X over the long run
- The expected value of X is found by multiplying each possible value of X by its probability, and then adding the products i.e.
- $E(X) = \sum k P(X=k)$
- Sum over all possible values of k

### Variance of a Random Variable

- The variance and the standard deviation are the measures of spread that accompany the choice of the mean to measure center.
- The variance  $\sigma^2$  of a random variable is a weighted average of the squared deviations  $(X - \mu)^2$  of the variable X from its mean  $\mu$ . Each outcome is weighted by its probability in order to take into account outcomes that are not equally likely.
- The larger the variance of X, the more scattered the values of X on average. The positive square root of the variance gives the standard deviation  $\sigma$  of X.

### Variance of a discrete random variable

For a discrete random variable X  
with probability distribution →

Value of X	$x_1$	$x_2$	$x_3$	...	$x_k$
Probability	$p_1$	$p_2$	$p_3$	...	$p_k$

and mean  $\mu_X$  the variance  $\sigma^2$  of X is found by multiplying each squared deviation of X by its probability and then adding all the products.

$$\begin{aligned} \sigma_X^2 &= (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + \dots + (x_k - \mu_X)^2 p_k \\ &= \sum (x_i - \mu_X)^2 p_i \end{aligned}$$

## How to Buy Textbooks

need class notes?  
ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

## How to Sell Textbooks @ BookHolders

**SALES**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold →

**Cash Now**

Instant Cash

Accept all books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

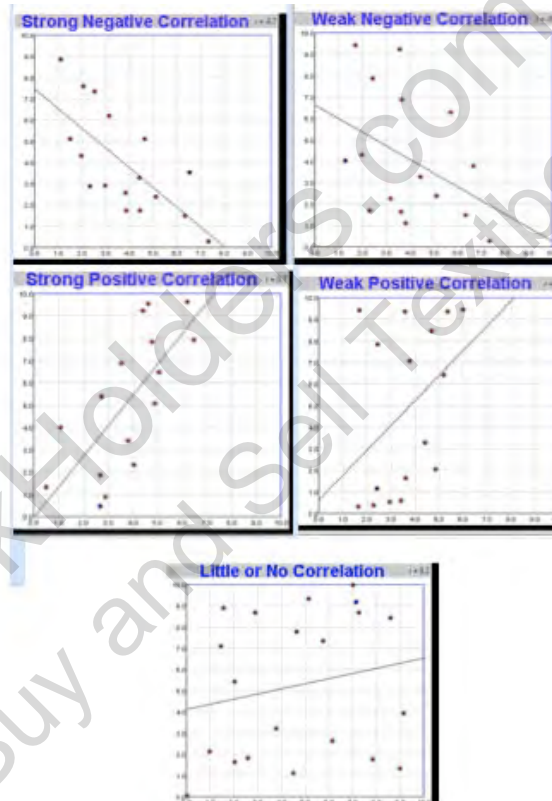
[www.BookHolders.com](http://www.BookHolders.com)

[HTTP://NOTES.BOOKHOLDERS.COM](http://NOTES.BOOKHOLDERS.COM)

### 6.1

#### Scatterplot

- Plots one quantitative variable against another
- Questions relate two quantitative variables and ask whether there is an association between them
- What to look for:
  - **Direction**
    - Negative
    - Positive
  - **Form**
    - Straight line relationship
    - Curved
    - Something exotic
    - No pattern
    - **If relationship isn't straight, but curves gently, while still increasing or decreasing steadily, we can find ways to straighten it out**
    - **If it curves sharply, then you'll need more advanced methods**
  - **Strength (Between 0 and 1)**
    - How much scatter?
      - Tight cluster or single stream
  - **Unusual Features**
    - Check for outliers



### 6.2 (Variables in Scatterplots)

- Explanatory or predictor variable
  - x-axis
- Response variable

[HTTP://NOTES.BOOKHOLDERS.COM](http://NOTES.BOOKHOLDERS.COM)

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold →

**Cash Now**

Instant Cash

Accept all books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

**HTTP://NOTES.BOOKHOLDERS.COM**

- y-axis
- y-variable depends on x-variable and the x-variable acts independently to make y-respond

### 6.3 (Understanding Correlation)

Standardizing variables

- $Z_x = \frac{X - \bar{X}}{S_x}$
- $Z_y = \frac{y - \bar{y}}{S_y}$

Correlation Coefficient

- $r = \frac{\sum Z_x Z_y}{n-1}$
- $r = \frac{\sum (X - \bar{X})(y - \bar{y})}{(n-1)S_x S_y}$ 
  - Multiply each  $(X - \bar{X})$  by it's paired  $(y - \bar{y})$  and then take the sum

Correlation Conditions

- Correlation measures the strength of the linear association between two quantitative variables
  1. Quantitative variables condition
    - a. Correlation applies only to quantitative variables
  2. Linearity Condition
    - a. Correlation measures the strength only of the linear association and will be misleading if the relationship is not straight enough
  3. Outlier Condition

Correlation Properties

1. The sign of a correlation coefficient gives the direction of the association
2. Correlation is always between -1 and 1
3. Correlation treats x and y symmetrically
  - a. The correlation of x with y is the same as the correlation of y with x
4. Correlation has no units
5. Correlation is not affected by changes in the center or scale of either variable
  - a. Changing the units or baseline of either variable has no effect on the correlation coefficient because the correlation depends only on the z-scores

**HTTP://NOTES.BOOKHOLDERS.COM**

**How to Buy Textbooks**

need class notes?  
ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

## HTTP://NOTES.BOOKHOLDERS.COM

6. Correlation measures the strength of the linear association between the two variables
  - a. Variables can be strongly associated but still have a small correlation if the association is not linear
7. Correlation is sensitive to unusual observations
  - a. A single outlier can make a small correlation large or large correlation small

### Correlation Tables

- Efficient way to start to look at a large data set
- The diagonal cells of a correlation table always show correlations of exactly 1.000

### 6.4 (Lurking Variables and Causation)

- No matter how strong the association, no matter how large the r-value, no matter how straight the form, there is no way to conclude from a high correlation alone that one variable causes the other
  - Always the possibility that some third variable – a lurking variable – is affecting both of the variables you have observed

### 6.5 (The Linear Model)

#### Residuals

- We want to find the line that somehow comes closer to all points than any other line
- A linear model is just an equation of a straight line through the data
- A linear model can be written as  $\hat{y} = b_0 + b_1 X$  where  $b_0$  and  $b_1$  are numbers estimated from the data and  $\hat{y}$  is the predicted value
- The difference between these two is called the residual
- $e = y - \hat{y}$
- The residual value tells us how far the model's prediction is from the observed value at that point
  - A negative residual means the predicted value is too big – an overestimate
  - A positive residual shows the model makes an underestimate

#### The Line of Best Fit

- The line for which the sum of the squared residuals is smallest – often called the least squares line

### 6.6 (Correlation and the Line)

- If the model is a good one, the data will scatter closely around it
- Example
  - For the Lowe's sales data, the line is:
  - $\hat{y}(\text{sales}) = -19,679 + .346(\text{Improvements})$

HTTP://NOTES.BOOKHOLDERS.COM

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold →

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Accept books no matter where purchased.

Accept all books no matter where purchased.  
Up to 3x more than Cash Now.

[www.BookHolders.com](http://www.BookHolders.com)

## HTTP://NOTES.BOOKHOLDERS.COM

- What does this mean?
  - The slope .346 says that we can expect a year in which residential improvement spending is 1 million dollars higher to be one in which Lowe's sales will be about .346 \$M (\$364,000) higher.
    - The slope is .346 million dollars of sales per million dollars of improvements
      - Slopes are always expressed in y-units per x-units
  - The intercept, -19,679, is the value of the line when the x-variable is zero
- Slope of the line:  $b_1 = r \frac{(S_y)}{S_x}$  where  $r$ =correlation and  $\frac{(S_y)}{S_x}$  are the standard deviations
  - If the slope is positive, the correlation is positive as well
  - If the slope is negative, the correlation is negative as well
- Intercept:  $b_0 = \hat{y} - b_1 X$
- Least squares lines are commonly called regression lines
  - 3 conditions need to be checked for regression
    - 1. Quantitative Variables Condition
    - 2. Linearity Condition
    - 3. Outlier Condition

### Understanding Regression from Correlation

- For every standard deviation we deviate from the mean in x, we predict that y will be r standard deviations away from the mean in y

### 6.7 (Regression to the Mean)

- Each predicted y tends to be closer to its mean (in standard deviations) than its corresponding x was
  - This property of the linear model is called regression to the mean
    - This is why the line is called the regression line

### 6.8 (Checking the Model)

- Conditions for linear regression
  1. Quantitative Data Condition
  2. Linearity Assumption (if scatterplot looks reasonably linear)
  3. Outlier Condition
  4. Independence Assumption (the cases are a random sample from the population)
- A scatterplot of residuals versus the x-values should be a plot without patterns. It shouldn't have any interesting features – no direction, no shape. It should stretch horizontally, showing no bends, and it should have no outliers

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold →

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Accept books no matter where purchased.

Up to 3x more than Cash Now.

[www.BookHolders.com](http://www.BookHolders.com)

**HTTP://NOTES.BOOKHOLDERS.COM**

- Equal Spread Condition  $e^2$ 
  - This condition requires that the scatter is about equal for all x-values. It's often checked using a plot of residuals against predicted values
- Standard deviation of the residuals:  $se = \frac{\sqrt{\sum e^2}}{n-2}$

### 6.9 (Variation in the Model and $R^2$ )

- The squared correlation,  $r^2$ , gives the fraction of the data's variation accounted for by the model, and  $1 - r^2$  is the fraction of the original variation left in the residuals
  - For the Lowe's sale model,  $r^2 = .976^2 = .952$  and  $1 - r^2$  is .048, so only 4.8% of the variability in Sales has been left in the residuals
  - 95.2% of the variability in Lowe's Sales is accounted for by variation in residential Improvement expenditures
- A correlation of .80 gives an  $R^2$  four times as strong as a correlation of .40 and accounts for four times as much of the variability

How Big Should  $R^2$  Be?

- Data from scientific experiments often have  $R^2$  in the 80% to 90% range and even higher
- An  $R^2$  of 100% is a perfect fit, with no scatter around the line. The  $s_e$  would be 0

### 6.10 (Is the Regression Reasonable?)

- Always be skeptical and ask yourself if the answer is reasonable

## Chapter 12

### 12.1 (Comparing Two Means)

How can we tell if a difference we observe in the sample means indicates a real difference in the underlying population means?

- We'll need to know the sampling distribution model and standard deviation of the difference
  - Once we know those, we can build a confidence interval and a test hypothesis

If the sample means come from independent samples the variance of their sum or difference is the sum of their variances

As long as the two groups are independent, we find the standard deviation of the difference between the two sample means by adding their variances and then taking the square root

**HTTP://NOTES.BOOKHOLDERS.COM**

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept all books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

**HTTP://NOTES.BOOKHOLDERS.COM**

$$SD(\bar{y}_1 - \bar{y}_2) = \sqrt{Var(\bar{y}_1) + Var(\bar{y}_2)}$$

$$SD(\bar{y}_1 - \bar{y}_2) = \sqrt{\left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2}$$

$$SD(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Since usually we don't know the true standard deviations of the two groups,  $\sigma_1$  and  $\sigma_2$ , so we substitute the estimates  $s_1$  and  $s_2$ , and find a standard error

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2}$$

When the conditions are met, the standardized sample difference between the means of two independent groups,

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\bar{\mu}_1 - \bar{\mu}_2)}{SE(\bar{y}_1 - \bar{y}_2)}$$

can be modeled by a Student's t-model with a number of degrees of freedom found with a special formula. We estimate the standard error with

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

## 12.2 (The Two Sample T-Test)

We start by hypothesizing a value for the true difference of the means (the hypothesized difference -  $\Delta_0$ )

- We then take the ratio of the difference in the means from our samples to its standard error and compare that ratio to a critical value from a Student's t-model

When the appropriate assumptions and conditions are met, we test the hypothesis:

$H_0: \bar{\mu}_1 - \bar{\mu}_2 = \Delta_0$  where the hypothesized difference  $\Delta_0$  is almost always 0.

$H_0: \bar{\mu}_1 = \bar{\mu}_2$

$H_A: \bar{\mu}_1 > \bar{\mu}_2$

$H_A: \bar{\mu}_1 < \bar{\mu}_2$

$H_A: \bar{\mu}_1 \neq \bar{\mu}_2$

We use the statistic:

**HTTP://NOTES.BOOKHOLDERS.COM**

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

**HTTP://NOTES.BOOKHOLDERS.COM**

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE(\bar{y}_1 - \bar{y}_2)}$$

The standard error of  $(\bar{y}_1 - \bar{y}_2)$  is:  $SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s^2_1}{m} + \frac{s^2_2}{n}}$

When the null hypothesis is true, the statistic can be closely modeled by a Student's t-model with a number of degrees of freedom given by a special formula. We use that model to compare our t-ratio with a critical value for t or to obtain a p-value

Instructions: find the observed t-value by finding the standard error → then find the degrees of freedom (from technology) → then find the t-value for those degrees of freedom and alpha level → then if the observed t-value is larger, smaller, or not equal to (depending on what we are looking for from the alternative hypothesis) than the t-value we found using the degrees of freedom, we reject the null hypothesis (or use p-value and compare p-value to alpha level)

### 12.3 (Assumptions and Conditions)

#### Independence Assumption

- Randomization Condition
- 10% Condition

#### Normal Population Assumption

- Nearly Normal Condition
  - Must check this for both groups
    - $n < 15$  → don't use these methods
    - $15 < n < 40$  → do not work with severely skewed data
    - $40 < n$  → use these methods

#### Independent Groups Assumption

- Two groups must be independent of each other

### 12.4 (A Confidence Interval for the Difference Between Two Means)

A hypothesis test really says nothing about the size of the difference. All it says is that the observed difference is large enough that we can be confident it isn't zero

- Rejecting a null hypothesis simply says that the observed statistic is unlikely to have been observed if the null hypothesis were true

When the conditions are met, we are ready to find a **two-sample t-interval** for the difference between means of two independent groups,  $\bar{u}_1$  and  $\bar{u}_2$ . The confidence interval is:  $(\bar{y}_1 - \bar{y}_2) \pm t^*_{df} \times SE(\bar{y}_1 - \bar{y}_2)$  where the standard error of the difference of the means is:

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s^2_1}{m} + \frac{s^2_2}{n}}$$

$t^*_{df}$  depends on the particular confidence level, and on the number of degrees of freedom

### 12.5 (The Pooled t-Test)

Willing to assume that the variances of the groups are equal (at least when the null hypothesis is true) → save some degrees of freedom by pooling (a little more accurate)

**HTTP://NOTES.BOOKHOLDERS.COM**

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

**HTTP://NOTES.BOOKHOLDERS.COM**

$$s^2_{\text{pooled}} = \frac{(n_1 - 1)s^2_1 + (n_2 - 1)s^2_2}{(n_1 - 1) + (n_2 - 1)} \rightarrow \text{pooled variance}$$

(If the two sample sizes are equal, this is just the average of the two variances)  
Substitute this pooled variance in place of each of the variances in the standard error formula

$$SE_{\text{pooled}}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s^2_{\text{pooled}}}{n_1} + \frac{s^2_{\text{pooled}}}{n_2}} = s_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

df=(n<sub>1</sub>-1)+(n<sub>2</sub>-1) → MUCH simpler than with the two-sample t test

Need to add the **Equal Variance Assumption** to the pooled t-test

- The variances of the two populations from which the samples have been drawn are equal

The conditions for the pooled t-test for the difference between two means of two independent groups are the same as for the two-sample t test with the additional assumption that the variances of the two groups are the same

The corresponding pooled-t confidence interval is:  $(\bar{y}_1 - \bar{y}_2) \pm t^*_{df} \times SE_{\text{pooled}}(\bar{y}_1 - \bar{y}_2)$ ,

where the critical value  $t^*$  depends on the confidence interval and is found with (n<sub>1</sub>-1)+(n<sub>2</sub>-1) degrees of freedom

Because the advantages of pooling are small, and you are allowed to pool only rarely (when the equal variances assumption is met), don't do it unless you are POSITIVE

**How to find 2-sample t-test confidence interval on calculator:**

Hit 2nd[F7](Ints)→4:2- SampTInt. In the first window that comes up, choose Stats. Then either say yes or no to pooled variance (say yes to pooled variance if it is known that the population variances are equal)

**How to find 2-sample t-test on calculator:**

**Hypothesis Testing for the Difference between Two Population Means.**

Use as a null hypothesis  $H_0: \mu_1 = \mu_2$

Hit 2[ndF6](Tests)→4:2-SampTTest.

Choose Stats in the first window ...put in alternate hypothesis; yes or no for pooled depending on if variances are the same, and calculate for results.

**12.6 (Tukey's Quick Test)**

7(.05)-10(.01)-13(.001)

- Count how many values in the high group are higher than all values of the low group
  - If the total of these exceedences is 7 or more, we can reject the null hypothesis (at alpha=.05)

**12.7 (Paired T-Test)**

You must decide whether the data are paired from understanding how they were collected and what they mean (most commonly when we have data on the same cases in two different circumstances)

**HTTP://NOTES.BOOKHOLDERS.COM**

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept all books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

## HTTP://NOTES.BOOKHOLDERS.COM

A paired t-test is just a one-sample t test for the mean of the pairwise differences (sample size is the number of pairs)

Conditions:

- Paired Data Assumption (data must actually be paired)
- Independence Assumption
  - Randomization Condition
  - 10% Condition
- Normal Population Assumption
  - Nearly Normal Condition

We test the hypothesis:  $H_0: \bar{u}_d = \Delta_0$

$t = \frac{\bar{d} - \Delta_0}{SE(\bar{d})}$  where  $\bar{d}$  is the mean of the pairwise differences,  $n$  is the number of pairs, and

$SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$  where  $s_d$  is the standard deviation of the pairwise differences

When the conditions are met and the null hypothesis is true, the sampling distribution of this statistic is a Student's t-model with  $n-1$  degrees of freedom and we use that model to obtain the P-value

The confidence interval is:  $\bar{d} \pm t^*_{n-1} \times SE(\bar{d})$

**On calculator: find the values of  $SE(\bar{d})$  and  $t$  and  $p$  values by plugging in  $n$ ,  $(\bar{d})$ , and  $s_d$  on a normal t-test model**

### Chapter 13

#### 13.1 (Goodness of Fit Tests)

To test if there is a recognizable pattern, we test the table's goodness-of-fit, where fit refers to the null model proposed. Example: Here the null model is that there is no pattern, that the distribution of up days should be the same as the distribution of trading days overall

Assumptions and Conditions:

- Counted Data Condition
  - The data must be counts for the categories of a categorical variable
- Independence Assumption
- Randomization Condition
- Sample Size Assumption
  - Check the expected cell frequency condition
- Expected Cell Frequency Condition
  - We should expect to see at least 5 individuals in each cell

Chi-Square Model

- Found by adding up the sum of the squares of the deviations between the observed and expected counts divided by the expected counts

HTTP://NOTES.BOOKHOLDERS.COM

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

## HTTP://NOTES.BOOKHOLDERS.COM

- $\chi^2 = \sum_{\text{all cells}} \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$
- The number of degrees of freedom for a goodness-of-fit test is  $k-1$ , where  $k$  is the number of cells
- A small chi-square statistic means that our model fits the data well, so a small value gives us no reason to doubt the null hypothesis
- Example of hypothesis:
  - $H_0$ : The days of the work week are distributed among the up days as they are among all trading days
  - $H_A$ : The trading days model does not fit the up days distribution

### The Chi Square Calculation

1. Find the expected values
  - a. These come from the null hypothesis model. Every null model gives a hypothesized proportion for each cell. The expected value is the product of the total number of observations times this proportion
2. Compute the residuals (Obs. - Exp)
3. Square the residuals (Obs. - Exp)<sup>2</sup>
4. Find  $\frac{(\text{Obs.} - \text{Exp.})^2}{(\text{Obs.} - \text{Exp.})}$  for each cell
5. Find the sum of the components  $\chi^2 = \sum_{\text{all cells}} \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$
6. Find the degrees of freedom
  - a. It's equal to the number of cells minus one
7. Test the hypothesis
  - a. Large chi-square values mean lots of deviation from the hypothesized model, so they give small P-values. Look up the critical value from a table of chi-square values or use calculator to find the P-value directly

On TI-89 → 2<sup>nd</sup> → F1 → Chi2GOF → enter in data to lists and enter in degrees of freedom

### 13.2 (Interpreting Chi-Square Values)

- Goodness-of-fit tests are often performed by people who have a theory of what the proportions should be in each category and who believe their theory to be true
- The only null hypothesis available is that the proposed theory is true
  - We can never confirm the null hypothesis...we can point out that the data are consistent with the proposed theory

### 13.3 (Examining the Residuals)

- When we reject a null hypothesis in a goodness-of-fit test, we can examine the residuals in each cell to learn more
- Whenever we reject a null hypothesis, it's a good idea to examine the residuals. (We don't need to do that when we fail to reject because when the  $\chi^2$  value is small, all of its components must have been small.)

HTTP://NOTES.BOOKHOLDERS.COM

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Get Paid

**Cash Now**

Instant Cash

Accept all books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

**HTTP://NOTES.BOOKHOLDERS.COM**

- Formula to standardize a residual:  $\frac{(Obs. - Exp.)}{\sqrt{Exp.}}$ 
  - Basically making them a z-score

**13.4 (The Chi-Square Test of Homogeneity)**

- Example: Null hypothesis is that the proportions choosing each alternative are the same for each country
- Difference between this and goodness-of-fit tests is that the goodness-of-fit test compared our observed counts to the expected counts from a given model. The test of homogeneity, has a null hypothesis that the distributions are the same for all the groups
  - The test examines the differences between the observed counts and what we'd expect under that assumption of homogeneity
- Assumptions and Conditions
  - Counted Data Condition
  - Independence Assumption
    - Randomization Condition
  - Sample Size Assumption
  - Expected Cell Frequency Assumption
- Component =  $\frac{(Obs. - Exp.)^2}{(Obs. - Exp.)}$
- $\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp}$
- The degrees of freedom are different that they were for the goodness-of-fit test
  - $(R-1)(C-1)$ , where R is the number of rows and C is the number of columns
- How to find expected values

	A	B	C	Total
Group A	25	5	10	40
Group B	10	45	5	60
TOTAL	35	50	15	100

- If we want to find the expected value under homogeneity for Group A who prefer C we would do:  $\frac{40 \times 15}{100} = 6$
- Find all expected values using this method, and then plug into calculator using the new formula for degrees of freedom

**13.5 (Comparing Two Proportions)**

- A chi-square test on a 2 X 2 table, which has only 1 df, is equivalent to testing whether two proportions are equal (z-test)
  - z-test can also give confidence interval
- Confidence Interval for the Difference of Two Proportions
  - When the conditions are met, we can find the confidence interval for difference of two proportions,  $p_1 - p_2$ . The confidence interval is

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

## HTTP://NOTES.BOOKHOLDERS.COM

$(\hat{p}_1 - \hat{p}_2) \pm z^* SE(\hat{p}_1 - \hat{p}_2)$ , where we find the standard error of the difference as

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{(\hat{p}_1 \hat{q}_1)}{n_1} + \frac{(\hat{p}_2 \hat{q}_2)}{n_2}}$$
 from the observed proportions.

The critical  $z^*$  depends on the particular confidence level that you specify

### 13.6 (Chi-Square Test of Independence)

- For any two events, A and B, to be independent the probability of event A given that event B occurred must be the same as the probability of event A
- Instead of asking: "Are the two groups homogeneous?" we know ask "Are the two groups independent?"
- Example:  $H_0$ : personal appearance and age are independent  
 $H_A$ : personal appearance and age are not independent
- Assumptions and Conditions
  - Counted Data Condition
  - Randomization Condition
  - Expected Cell Frequency Condition

## Chapter 14

### 14.1 (The Population and the Sample)

- $u_y = \beta_0 + \beta_1 x$
- We write  $u_y$  instead of  $y$  because the regression line assumes that the means of the  $y$  values for each value of  $x$  fall exactly on the line
- If we want to account for each individual value of  $y$  in our model, we have to include these errors, which we denote by  $\varepsilon$ :  $y = \beta_0 + \beta_1 x + \varepsilon$
- We estimate the  $\beta$ 's by finding a regression line,  $\hat{y} = b_0 + b_1 x$ 
  - The residuals  $e = y - \hat{y}$ , are the sample-based versions of the errors,  $\varepsilon$

### 14.2 (Assumptions and Conditions)

- Linearity Condition
  - If a scatterplot looks straight
  - Quantitative Variable Condition
- Independence Assumption
  - Randomization Condition
- Equal Variance Assumption
  - The variability of  $y$  should be about the same for all values of  $x$
  - Equal Spread Condition
- Normal Population Assumption
  - Nearly Normal Condition
  - Normal Probability Plot
  - Outlier Condition

### 14.3 (Regression Inference)

HTTP://NOTES.BOOKHOLDERS.COM

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

## HTTP://NOTES.BOOKHOLDERS.COM

- What is the standard deviation of this distribution? What aspects of the data affect how much the slope vary from sample to sample
  - Spread around the line
    - Less scatter around the line means the slope will be more consistent from sample to sample
    - We measure the spread around the line with the residual standard deviation:

$$s_e = \frac{\sqrt{\sum (y - \hat{y})^2}}{n - 2}$$

- Spread of the x's
  - A large standard deviation of x,  $s_x$  provides a more stable regression
- Sample Size
  - A larger sample size gives more consistent estimates from sample to sample

- Standard error of the slope:  $SE(b_1) = \frac{s_e}{s_x \sqrt{n-1}}$ ,  $s_e$  = error standard deviation

- The sampling distribution for the regression slope
  - When the conditions are met, the standardized estimated regression slope,  $t = \frac{b_1 - \beta_1}{SE(b_1)}$ , follows a Student's t model with n-2 degrees of freedom. We

estimate the standard error with  $SE(b_1) = \frac{s_e}{s_x \sqrt{n-1}}$ , where  $s_e = \frac{\sqrt{\sum (y - \hat{y})^2}}{n - 2}$ ,

n is the number of data values, and  $s_x$  is the standard deviation of the x-values

- The usual null hypothesis about the slope is that it's equal to 0 because if the slope were zero, there wouldn't be much left of our regression equation
- The t-test for the regression slope
  - When the assumptions and conditions are met, we can test the hypothesis  $H_0: \beta_1 = 0$  vs.  $H_A: \beta_1 \neq 0$  (or a one-sided alternative hypothesis) using the standard estimated regression slope,  $t = \frac{b_1 - \beta_1}{SE(b_1)}$ , which follows a

Student's t model with n-2 degrees of freedom

- The confidence interval for the regression slope  $b_1 \pm [(t^*_{n-2})(SE(b_1))]$ , where the critical value  $t^*$  depends on the confidence level and has n-2 degrees of freedom

On TI-89 → 2<sup>nd</sup> → F6 → LinReg T Test

### 14.4 (Standard Errors for Predicted Values)

- The confidence interval for the predicted mean value

HTTP://NOTES.BOOKHOLDERS.COM

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold →

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

## HTTP://NOTES.BOOKHOLDERS.COM

- When the conditions are met, we find the confidence interval for the predicted mean value  $u_v$  at a value  $x_v$  as  $\hat{y}_v \pm [(t^*_{n-2})(SE)]$ , where the

$$\text{standard error is } SE(\hat{u}_v) = \sqrt{(SE^2(b_1))(x_v - \bar{x})^2 + \frac{s_e^2}{n}}$$

- If we want to capture an individual price, we need to use a wider interval, called a prediction interval
- The prediction interval for an individual value
  - When the conditions are met, we can find the prediction interval for all values of  $y$  at value  $x_v$  as  $\hat{y}_v \pm [(t^*_{n-2})(SE)]$ , where the standard error is

$$SE(\hat{y}_v) = \sqrt{(SE^2(b_1))(x_v - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$$

### 14.5 (Using Confidence and Prediction Intervals)

- $ME = [(t^*)(SE(\hat{y}_v))]$
- Prediction interval (a wider interval):  $\hat{y} \pm [(t^*)(SE(\hat{y}_v))]$

### 14.6 (Extrapolation and Prediction)

- Extrapolation
  - Dangerous because they require the additional assumption that nothing about the relationship between  $x$  and  $y$  changes, even at extreme values of  $x$  and beyond
    - If you extrapolate far into the future, be prepared for the actual values to be (possibly quite) different from your predictions

### 14.7 (Unusual and Extraordinary Observations)

- Outliers
  - Points with large residuals
- Leverage
  - If  $x$ -value is far from the mean of the  $x$ -values it has high leverage
    - If the point lines up with the pattern of the other points, it doesn't always change our estimate of the line
- Influence
  - A point is influential is omitting it from the analysis gives a very different model

### 14.8 (Working with Summary Values)

- Scatterplots of statistics summarized over groups tend to show less variability than we would see if we measured the same variables on individuals
  - Summary statistics themselves vary less than the data on individuals

### 14.9 (Linearity)

- Check the linearity condition by plotting the residuals versus either the  $x$ -variable or the predicted variables
  - Residual plots should have no pattern

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold → Instant Cash

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

**[HTTP://NOTES.BOOKHOLDERS.COM](http://NOTES.BOOKHOLDERS.COM)**

- Can straighten it out using reciprocals

Chapter 15 (pg.483)

### 15.1 (The Multiple Regression Model)

$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \dots + b_k X_k$  where  $b_0$  is still the intercept and each  $b_k$  is the estimated coefficient of its corresponding predictor  $X_k$

The residuals are  $e = y - \hat{y}$

The degrees of freedom is the number of observations (n) minus one for each coefficient estimated:  $df = n - k - 1$  where k is the number of predictor variables and n is the number of cases

Standard deviation of the residuals:  $S_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - k - 1}}$

The t ratio measures how many standard errors the coefficient is away from 0

- Using a student's t-model, we can use its P-value to test the null hypothesis that the true value of the coefficient is 0

### 15.2 (Interpreting Multiple Regression Coefficients)

- In a multiple regression, coefficients have a more subtle meaning. Each coefficient takes into account the other predictor(s) in the model

### 15.3 (Assumptions and Conditions for the Multiple Regression Model)

$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \dots + b_k X_k + \epsilon$  (with error in model)

- Linearity Assumption (need to make sure there is no bend or other nonlinearity)
- Independence Assumption
  - Randomization Condition
- Equal Variance Assumption
  - Equal Spread Condition
    - A scatterplot of residuals vs. predicted values should show no evidence of changing spread
- Normality Assumption
  - Nearly Normal Condition

### 15.4 (Testing the Multiple Regression Model)

Test the null hypothesis that all the slope coefficients are zero:

$$H_0: \beta_1 = \dots = \beta_k = 0$$

$$H_A: \text{at least one } \beta \neq 0$$

If we reject the null, we conclude that the multiple regression model for predicting "house prices" with these "five variables" is better than just using the mean

We can test this hypothesis with an F-test (It's the generalization of the t-test to more than one predictor). The sampling distribution of the statistic is labeled with the letter F. The F-distribution has two degrees of freedom, k, the number of predictors, and n-k-1

**[HTTP://NOTES.BOOKHOLDERS.COM](http://NOTES.BOOKHOLDERS.COM)**

**How to Buy Textbooks**

need class notes? ask for help notes

**BUY**

**Online** Search by class, author, title, or ISBN at [bookholders.com](http://bookholders.com). Purchase for pickup at local store during convenient hours.

**Self-Serve** Purchase at local store using simple and easy self-serve stations.

**Full-Serve** Purchase with the help of clerk that can answer your questions about a book or class.

4509A College Ave  
301-209-9313

[www.BookHolders.com](http://www.BookHolders.com)

**How to Sell Textbooks @ BookHolders**

**Two Options**

**Advantage**

Get Paid → Drop Off → Gets Sold →

**Cash Now**

Instant Cash

Accept books no matter where purchased.  
Up to 3x more than Cash Now.

Accept books no matter where purchased.

[www.BookHolders.com](http://www.BookHolders.com)

**HTTP://NOTES.BOOKHOLDERS.COM**

$$t_{n-k-1} = \frac{b_j - 0}{SE(b_j)}$$

A confidence interval for each slope  $b_j$  is:

$$b_j \pm t_{n-k-1}^* \times SE(b_j)$$

If we fail to reject the null hypothesis for a multiple regression coefficient, it does not mean that the corresponding predictor variable has no linear relationship to  $y$ . It means that the corresponding predictor contributes nothing to modeling  $y$  after allowing for all the other predictors

### 15.5 (Adjusted $R^2$ and the F-Statistic)

The  $R^2$  value tells us how much of the variation in  $y$  is accounted for by the model with all the predictor variables included

$SSE = \sum e^2$  is the sum of squared residuals ( $e$  is for error)

A larger SSE (And thus  $s_e$ ) means that the residuals are more variable and that our predictions will be correspondingly less precise

The total variation of the response variable,  $y$ , which is called the Total Sum of Squares and is denoted SST:  $SST = \sum (y - \bar{y})^2$

$SST = SSR + SSE$  where  $SSR = \sum (\hat{y} - \bar{y})^2$  (Regression sum of squares)  $\rightarrow$  tells us how much of the total variation in the response is due to the regression model

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$  ... When the SSE is nearly 0, the  $R^2$  value will be close to 1

$F = \frac{MSR}{MSE}$  where  $MSR = SSR/k$  and  $MSE$  is  $SSE/(n-k-1)$  and  $MST$  is  $SST/n-1$

Adjusted  $R^2$  imposes a "penalty" for each new term that's added to the model in an attempt to make models of different sizes (number of predictors) comparable

$$R^2_{adj} = 1 - \left( \frac{\frac{SSE}{n-k-1}}{\frac{SST}{n-1}} \right)$$

**ANOVA:**

**HTTP://NOTES.BOOKHOLDERS.COM**