**ELEMENTARY STATISTICS**

**IN AN UPSIDE DOWN CLASSROOM**

**(Math 190)**

Fall 2013

Skyline College

Research and Text by  
Terrie Nichols

Researched, Compiled, Edited and Written by

Terrie Nichols

Cuyamaca College

Math Department

Contents

Introduction 5

First Homework Assignment and Group Quiz 6

Preparing for Class 6

**MODULE 1** 7

**Types of Statistical Studies** 7

**and Producing Data** 7

1.1 Types of Statistical Studies 9

1.2 Collecting Data - Sampling 13

1.3 Collecting Data – Conducting an Experiment 17

**MODULE 2** 23

**Summarizing Data Graphically** 23

**and Numerically** 23

2.1.1 Distributions for Quantitative Data (Dotplots) 25

2.1.2 Distributions for Quantitative Data (Histograms) 29

2.2.1 Measures of Center (Getting a Feel for "Typical Values") 33

2.2.2 Measures of Center (Balancing on the Mean) 37

2.2.3 Measures of Center (The Median) 41

2.3 Quantifying Variability Relative to the Median 45

2.4 Quantifying Variability Relative to the Mean 49

**MODULE 3** 53

**Examining Relationships:** 53

**Quantitative Data** 53

3.1.1 Scatterplots 55

3.1.2 Linear Relationships and Correlation Coefficients (Introduction) 63

3.1.3 Causation and Lurking Variables 65

3.2 Fitting a Line 67

3.3.1 Residuals and the Best Fit Line 75

3.3.2 Assessing the Fit of a Line 85

**MODULE 4** 91

**Non-linear Models** 91

4.1.1 Non-linear Models (Exponential Relationships) 93

4.1.2 Exponential Equations – Part 1 99

5.1.2 Exponential Equations – Part 2 103

**MODULE 5** 109

**Relationships in Categorical Data** 109

**with an Intro to Probability** 109

5.1 Relationships in Categorical Data with an Intro to Probability 111

**MODULE 6** 119

**Probability and** 119

**Probability Distributions** 119

6.1.1 Introduction to Probability 119

6.1.2 Probability Rules 124

6.1.2 Practice Problems 132

Mod 6 Probability Distributions Warm Up – Area Under the Normal Curve 136

6.1.4 & 6.2 Probability Distributions for Random Variables 138

6.1.4 Practice Problems 148

**MODULE 7** 150

**Linking Probability to** 150

**Statistical Inference** 150

7.1.1 Distribution of Sample Proportions (Exploring with Skittles) 152

7.1.2 Distributions of Sample Proportions (Making Connections to Probability Models) 157

7.2.1 Introduction to Statistical Inference: Confidence Intervals 163

7.2.2 Introduction to Statistical Inference: Exploring Hypothesis Testing 171

7.2 Practice Problems 173

# Introduction

Forget what you know about the traditional math classroom where teachers lecture and students diligently take notes while struggling to understand the hieroglyphics materializing before them on the board. Learning math this way may work for some, but for most the traditional math classroom does not allow the student to engage with the course material until the student attempts to work through the practice problems a few days later (and even then … when faced with math homework, the student may suddenly prefer to do the dishes that have been sitting in the sink for days). To improve learning and increase student success, we'll study statistics in an upside down classroom – no more typical lectures, robotic note-taking, textbooks, skill and drill problems, or traditional high-stakes exams.

So what is an upside down classroom and how does it differ from the traditional math classroom? In the upside down classroom students are expected to review materials that would normally be presented in class (including the lecture) outside of class time.  This frees class time up to focus on discussion, collaborative work, and engagement with other activities that are traditionally done outside of class. More specifically, during class we'll work in groups on activities and problems that introduce and motivate the concepts in each of the course topics and/or review course topics and skills. This will allow me to work with students individually or in small groups during class to facilitate student mastery of concepts and acquisition of skills needed for the student to work through the "lecture" content and answer a few short questions online at home. After each course topic is introduced through our in-class group-work and BEFORE we introduce the next topic, the student is expected to work through the online materials (lecture content, questions, activities, checkpoints, and/or labs) outside of class. This online work replaces the typical homework assigned in traditional math classes. Please note – it is very important that each student completes the online work in a timely manner so that we can identify gaps in understanding and use class time to remediate those gaps before we move on to the next topic.

In summary, the student should expect to engage in the following activities during class.

* Group-work to motivate and introduce key concepts for each of the course topics
* Group-work to review skills and work through skill-building problems from topics previously discussed in class
* Class discussions of course topics
* Mini lectures as necessary to close gaps in concept mastery or skill attainment
* Mini on-the-spot poster presentations (with speed dating)

Furthermore, the student should expect to do the following outside of class.

* Read the online lecture content on the OLI (Open Learning Initiative) website (you'll find more information about OLI in the next section)
* Answer the embedded questions, complete the embedded activities, and StatTutor labs while working through the OLI content
* Complete the topic and module checkpoints (think of these as homework problems)

## First Homework Assignment and Group Quiz

This assignment will prepare you for the group quiz on the first day of class.

Read through the following *Introduction and Learning Strategies* sections of our class on OLI.

* Introduction
* Learning Strategies
* The Big Picture

On the first day of class, you will work with your group to reproduce (draw) the *Big Picture* and briefly describe its various parts without notes. This will be your first group quiz.

## Preparing for Class

Since we will work in groups every day and since your learning (and your grade) and your classmates' learning (and their grades) depend upon this group work, it is very important that you

* complete the assigned OLI work before class,
* attend class every day, and
* arrive on time and are ready to learn when class starts.

Being ready to learn means you

* have this workbook with you (but recall you should NOT work through the activities outside of class unless instructed to do so),
* are well-rested and ready to engage with the material and your classmates,
* have a writing utensil out and your cell phone put away, and
* have your graphing calculator with you (beginning with Module 3).

# **MODULE 1**

# **Types of Statistical Studies**

# **and Producing Data**

## 1.1 Types of Statistical Studies

Learning Objectives

* Distinguish between questions about a population and questions about cause-and-effect.
* Determine if a study is an experiment or an observational study.
* Explain how the study design impacts the types of conclusions that can be drawn.

1. Suppose a History instructor allows students to select one of the two topics below for a term paper they are required to write. Further suppose about half the class selects the first topic and half selects the second topic. Which group of students do you think would be more successful when writing their term paper? Explain.

|  |  |
| --- | --- |
| 1. Write a term paper on the history of the world. | 1. Write a term paper about the problems Martin Luther King had to overcome to lead the civil rights movement in the United States. |

1. Suppose we are researchers tasked with investigating the relationship between smoking and lung capacity. We begin by writing research questions that will guide our investigation. For each of the following research questions, decide whether the question would make it easier or harder for us to collect and analyze data to answer the question. Then rank all four questions from best (rank 1) to worst (rank 4). Explain your rankings, i.e. explain why the better ranked question(s) would make it easier to collect and analyze data, and why the worst ranked question(s) would make it more difficult to collect and analyze data.
2. How does smoking affect lung capacity?
3. On average, do women who smoke have less lung capacity than women who do not smoke?
4. Does smoking cause reduced lung capacity in women?
5. Are people who smoke more likely to ruin their lung capacity?
6. Copy your top two research questions below, and then label each one as a question about a ***population*** or a question about ***cause and effect***.
7. For each of your top two questions what are the groups being compared?
8. For each of your top two questions what is the ***variable*** we are trying to measure?
9. Two statistical studies are described below (we'll improve on each of the studies in subsequent activities, but for now, let's explore what we have).

|  |  |
| --- | --- |
| a) | b) |
| * Find 100 women age 20 who do not smoke. * Randomly assign 50 of the 100 women to the smoking treatment and the other 50 to the no-smoking treatment. * Those in the smoking group smoke a pack a day for 10 years, and those in the control group remain smoke-free for 10 years. * Measure and record lung capacity for each of the 100 women. * Analyze, interpret, and draw conclusions from the data. | * Find 100 women age 30 such that 50 of them have been smoking a pack a day for 10 years and 50 of them have been smoke-free for 10 years. * Measure lung capacity for each of the 100 women. * Analyze, interpret, and draw conclusions form the data. |

1. At the top of each column, label the statistical study as an ***observational study*** or an ***experiment*** (I know … we haven’t defined these terms yet … but give it your best shot). What features of the study made you label it an observational study or an experiment?
2. In each column of the column headers on the previous page indicate whether or not the study would provide convincing evidence about cause and effect. In the space below, explain why you labeled each study as ***provides convincing evidence about cause and effect*** or ***does NOT provide convincing evidence about cause and effect***.
3. Match your top two research questions to the corresponding study, and then write the research question in the appropriate column header on the previous page.
4. In each of the statistical studies listed on the previous page, what are the variables (note – the variables are the same in both studies)? Which is the ***explanatory variable*** and which is the ***response variable***? Explain.
5. Fisher’s Hypothesis entertains the following suppositions and draws the following conclusions. Suppose there is a gene that makes smoking appear to be very pleasurable. Further suppose the same gene causes emphysema, lung cancer, and throat cancer. Then people who have the gene would be more likely to smoke than people who do not have the gene. And people who have the gene would be more likely to have emphysema, lung cancer, or throat cancer.

Let’s make a Stats Academy hypothesis. Suppose there is a gene that makes smoking appear to be very pleasurable, and suppose that the presence of the gene makes people more likely to have lower than average lung capacity regardless of whether they smoke or not. Further suppose that the test for the gene is astronomically expensive (drain-the-entire-research-budget-to-test-just-one-person kind of expensive).

1. How would the Stats Academy gene affect either or both of our statistical studies?
2. Considering each of our statistical studies independently, could we reasonably control for the Stats Academy gene? If so, explain how. If not, explain why not. Be prepared to share your conclusions with the entire class.

## 1.2 Collecting Data - Sampling

Learning Objective

* Identify the sampling plan for a study. Recognize implications and limitations of the plan.

1. In the 1.1 activity, we examined two statistical studies which might be used to answer the question, "Does smoking affect lung capacity?" Each of the studies is copied below along with the research question the study seeks to answer.

|  |  |
| --- | --- |
| **Experiment**  Research Question: Does smoking cause reduced lung capacity in women? | **Observational Study**  Research Question: On average, do women who smoke have less lung capacity than women who do not smoke? |
| * Find 100 women age 20 who do not smoke. * Randomly assign 50 of the 100 women to the smoking treatment and the other 50 to the no-smoking treatment. * Those in the smoking group smoke a pack a day for 10 years, and those in the control group remain smoke-free for 10 years. * Measure and record lung capacity for each of the 100 women. * Analyze, interpret, and draw conclusions from the data. | * Find 100 women age 30 such that 50 of them have been smoking a pack a day for 10 years and 50 of them have been smoke-free for 10 years. * Measure and record lung capacity for each of the 100 women. * Analyze, interpret, and draw conclusions from the data. |

1. Review: What are the explanatory and response variables for these studies (they both have the same explanatory and response variables)?
2. Suppose you have been hired as the lead researcher tasked with answering the question, "Does smoking affect lung capacity?" You will be responsible for conducting either one of these statistical studies. As the lead researcher and responsible party, it is your choice which study to use.
3. Which will you choose? Explain.
4. Are there additional steps you would include to improve the study? Explain.
5. Since each study seeks to answer questions about all women, and since you do not have the time, personnel, and resources to study all women, your team will need to select a sample that is representative of the population – all women. Develop a sampling plan that describes exactly how your team will choose the sample so that it is representative of the population and avoids systematically favoring certain outcomes. Make your plan as detailed as possible. As you develop the plan list any outcomes you discover which may be favored inadvertently over others when collecting the sample and describe the steps you would take to avoid the subsequent bias.
6. In order to deliver results that have a profound impact on the health care of their patients (as well as the hospital's bottom line) Tri-City Memorial Hospital seeks to go above and beyond patients' basic expectations. Additionally, they would like to develop a plan to emotionally engage patients so their patients feel attached and loyal to Tri-City Memorial. They decide they need to first identify the customer service areas of patient care in which they are failing to exceed patients' basic expectations and/or failing to emotionally engage patients. They hire a firm to help them design a survey. Then over a period of three months, they give each patient the survey at checkout and ask the patient to complete it and mail it back in the self-addressed stamped envelope when the patient is feeling up to it. The sample population will be the patients who respond to the survey over a period of three months.
7. Identify the outcomes which could be favored by this sampling plan.
8. Could they design a sampling plan that would avoid the biases you discovered in part a?
9. As part of the Social Justice unit in your Statistics class, the class decides to research the question, "Are upper class people in the U.S. healthier than working class people in the U.S.?" The students develop a survey, and a sampling plan for administering the survey. Each student in the class agrees to visit a mall close to his or her home to find 10 people who appear to be lower class to take the survey and 10 people who appear to be upper class to take the survey. After each of the 40 students in the class collect their surveys, the data from 800 respondents will be analyzed.
10. Identify the outcomes which could be favored by this sampling plan and explain why these biases may appear.
11. Could the class design a sampling plan that would avoid the biases you discovered in part a?

## 1.3 Collecting Data – Conducting an Experiment

Learning Objectives

* Identify features of experiment design that control the effects of confounding.
* Explain how the study design impacts the types of conclusions that can be drawn.

1. The Hawthorne Works, in Cicero, Illinois, was a large factory complex built by Western Electric starting in 1905 and operating until 1983. It had 45,000 employees at the height of its operations. Besides telephone equipment, the factory produced a wide variety of consumer products, including refrigerators and electric fans. Hawthorne Works was named for Hawthorne, Illinois, a small town that was later incorporated into Cicero. The facility was so expansive; it contained a private railroad to move shipments through the plant to the nearby Burlington Northern Railroad freight depot, and workers regularly used bicycles for transit within the plant. It was purchased in the mid-1980s by the late Donald L. Shoemaker and replaced with a shopping center. (Wikipedia: <http://en.wikipedia.org/wiki/Hawthorne_Works>)

In addition to its enormous output of telephone equipment, Hawthorne Works was the site of some well-known industrial studies. In one of the studies, experimenters chose two women as test subjects and asked them to choose four other workers to join their respective test group. Together the women in each test group worked in a separate room over the course of five years (1927–1932) assembling telephone relays.

Output was measured mechanically by counting how many finished relays each worker dropped down a chute. This measuring began in secret two weeks before moving the women to an experiment room and continued throughout the study. In the experiment room, they had a supervisor who discussed changes with them and at times used their suggestions. Then the researchers spent five years measuring how different variables impacted the group's and individuals' productivity. (Wikipedia: <http://en.wikipedia.org/wiki/Hawthorne_effect>)

1. The researchers decided to study the effect of breaks on productivity. First they provided the workers with two 5 minute breaks (after a discussion with the workers on the best length of time), and then they changed to two 10-minute breaks (not the workers' preference). Productivity increased, but when the workers received six 5-minute breaks, the workers disliked it and reduced output.

Is this an experiment or an observational study? If it's an experiment, what is the treatment? If it's an observational study, what population is being studied?

What are the explanatory and response variables?

Is the explanatory variable the only variable impacting the response variable? If not, what are the confounding variables which impact the response variable?

1. Then the researchers decided to study the effect of the length of the workday on productivity. They shortened the work day by 30 minutes and output increased. But when they shortened it more, overall output decreased even though the output per hour increased. When they returned to the first condition (prior to shortening the work day by 30 minutes), output peaked.

Is this an experiment or an observational study? If it's an experiment, what is the treatment? If it's an observational study, what population is being studied?

What are the explanatory and response variables?

Is the explanatory variable the only variable impacting the response variable? If not, what are the confounding variables which impact the response variable?

1. The following description of a statistical study is excerpted from *Understandable Statistics Concepts and Methods, 10 Ed.*, by Brase and Brase (example 4, page 22).

*In 1778, Captain James Cook landed in what we now call the Hawaiian Islands. He gave the islanders a present of several goats, and over the years these animals multiplied into wild herds totaling several thousands. They eat almost anything, including the famous silver sword plant, which was once unique to Hawaii. At one time, the silver sword grew abundantly on the island of Maui … but each year there seemed to be fewer and fewer plants. Biologists suspected that the goats were partially responsible for the decline in the number of plants and conducted a statistical study that verified their theory.*

|  |  |
| --- | --- |
| *To test the theory, park biologists set up stations in remote areas of Haleakala. At each station two plots of land similar in soil conditions, climate, and plant count were selected. One plot was fenced to keep out the goats, while the other was not. At regular intervals a plant count was made in each plot.* | *SilverSwordPlantHaleakala.jpg*  Haleakalā silversword (Wikipedia) |

1. Is this an experiment or an observational study? If it's an experiment, what is the treatment? If it's an observational study, what population is being studied?
2. What are the explanatory and response variables?
3. Is there a control group? Explain.
4. What are the confounding variables this study attempts to control and how does the study attempt to control each?
5. The following description of a statistical study is excerpted from *Understandable Statistics Concepts and Methods, 10 Ed.*, by Brase and Brase (example 5, page 23).

*Can chest pain be relieved by drilling holes in the heart? For more than a decade, surgeons have been using a laser procedure to drill holes in the heart. Many patients report a lasting dramatic decrease in angina (chest pain) symptoms. Is the relief due to the procedure, or is it a placebo effect? A recent research project at Lenox Hill Hospital in New York City provided some information about this issue by using a completely randomized experiment. The laser treatment was applied through a less invasive (catheter laser) process. A group of 298 volunteers with severe, untreatable chest pain were randomly assigned to get the laser or not. The patients were sedated but awake. They could hear the doctors discuss the laser process. Each patient though he or she was receiving the treatment.*

1. Draw a diagram of the experiment.
2. What are the explanatory and response variables?
3. The laser patients did well, but the placebo group showed more improvement in pain relief. How could this be?
4. In its January 25th, 2012, issue, the Journal of the American Medical Association (JAMA) reported on the effects of overconsumption of low, normal, and high protein diets on weight gain, energy expenditure, and body composition.

**Context** The role of diet composition in response to overeating and energy dissipation in humans is unclear.

**Objective** To evaluate the effects of overconsumption of low, normal, and high protein diets on weight gain, energy expenditure, and body composition.

**Design, Setting, and Participants** A single-blind, randomized controlled trial of 25 US healthy, weight-stable male and female volunteers, aged 18 to 35 years with a body mass index between 19 and 30. The first participant was admitted to the inpatient metabolic unit in June 2005 and the last in October 2007.

**Intervention** After consuming a weight-stabilizing diet for 13 to 25 days, participants were randomized to diets containing 5% of energy from protein (low protein), 15% (normal protein), or 25% (high protein), which they were overfed during the last 8 weeks of their 10- to 12-week stay in the inpatient metabolic unit. Compared with energy intake during the weight stabilization period, the protein diets provided approximately 40% more energy intake.

**Main Outcome Measures** Body composition was measured by dual-energy x-ray absorptiometry biweekly, resting energy expenditure was measured weekly by ventilated hood, and total energy expenditure by doubly labeled water prior to the overeating and weight stabilization periods and at weeks 7 to 8.

**Results** Overeating produced significantly less weight gain in the low protein diet group compared with the normal protein diet group or the high protein diet group. Body fat increased similarly in all 3 protein diet groups and represented 50% to more than 90% of the excess stored calories. Resting energy expenditure, total energy expenditure, and body protein did not increase during overfeeding with the low protein diet. In contrast, resting energy expenditure and body protein (lean body mass) increased significantly with the normal and high protein diets.

**Conclusions** Among persons living in a controlled setting, calories alone account for the increase in fat; protein affected energy expenditure and storage of lean body mass, but not body fat storage.

1. Is this an experiment or an observational study?
2. What are the explanatory and response variables?
3. Is there a control group in this study?
4. What are the confounding variables this study attempts to control and how does the study attempt to control each?

# **MODULE 2**

# **Summarizing Data Graphically**

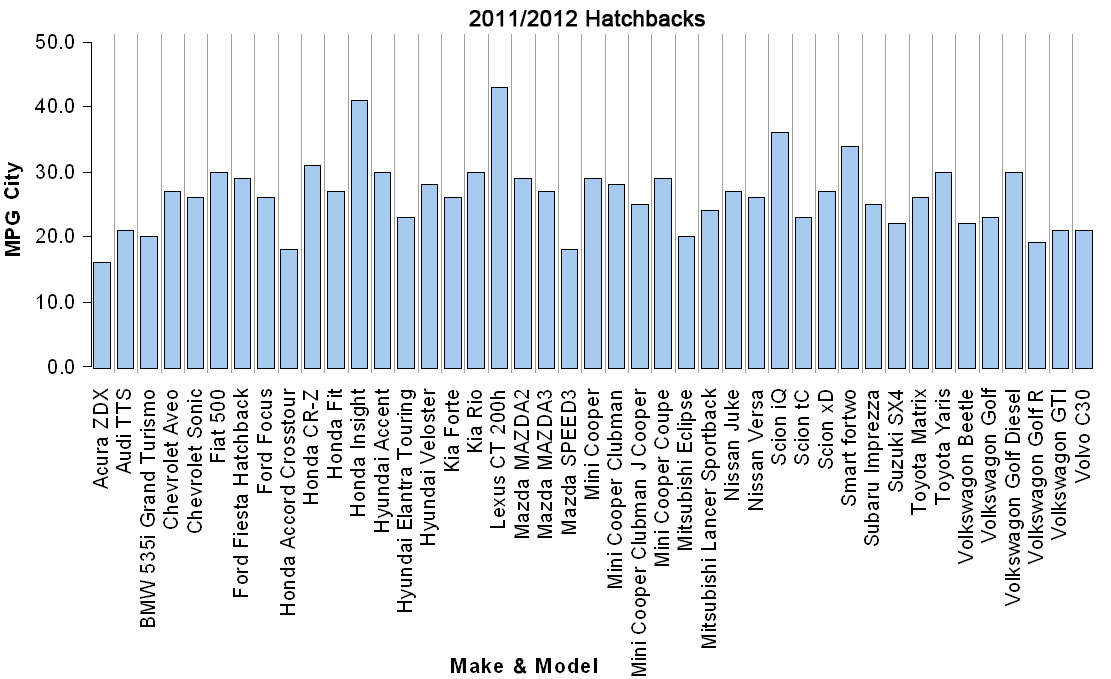
# **and Numerically**

## 2.1.1 Distributions for Quantitative Data (Dotplots)

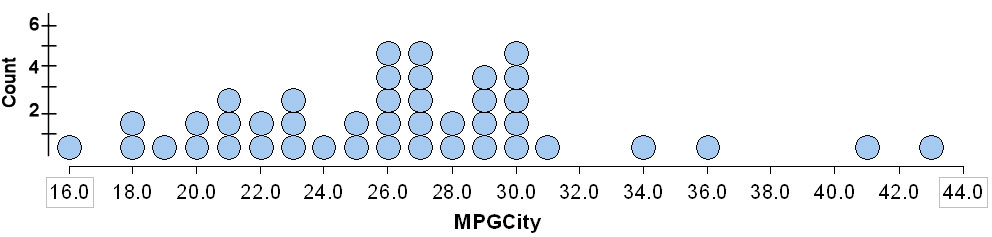
Learning Objective

* For the distribution of a quantitative variable, describe the overall pattern (shape, center, and spread) and striking deviations from the pattern.

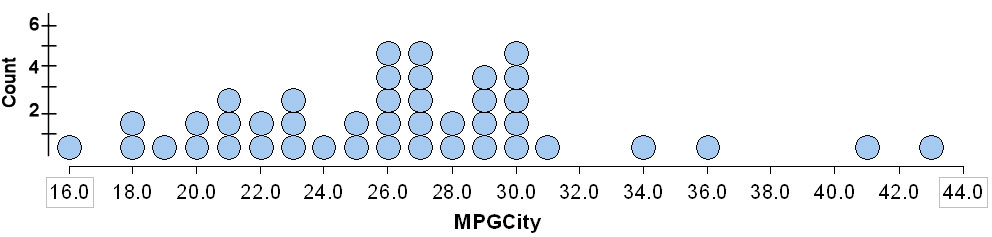
1. Use the case-value graph below to answer each of the following questions.



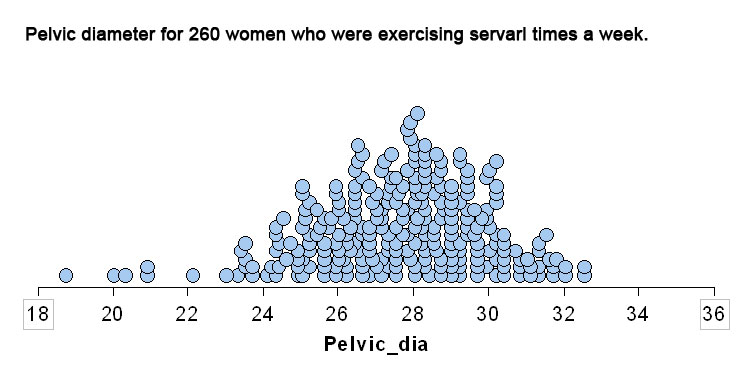
1. How many hatchbacks get 23 mpg in the city?
2. What are the mpg rates that occur most frequently?
3. Which mpg rate is the middle rate?
4. Take a look at the questions you answered on the previous page. Were you asked to examine the individuals cases and their values (i.e. the individual hatchbacks), or were you asked to describe patterns in data and/or to create summaries about the group (in this case the group would be 2011/2012 Hatchbacks)?
5. Was it awkward to work with the case-value graph on the previous page? Why or why not?
6. Let's answer the same questions using a dotplot to represent the same data for the variable MPGCity.

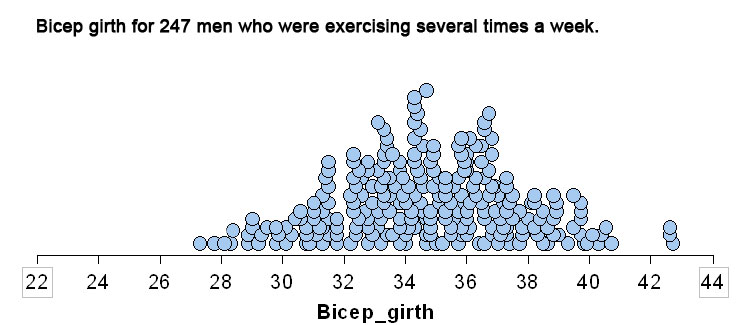


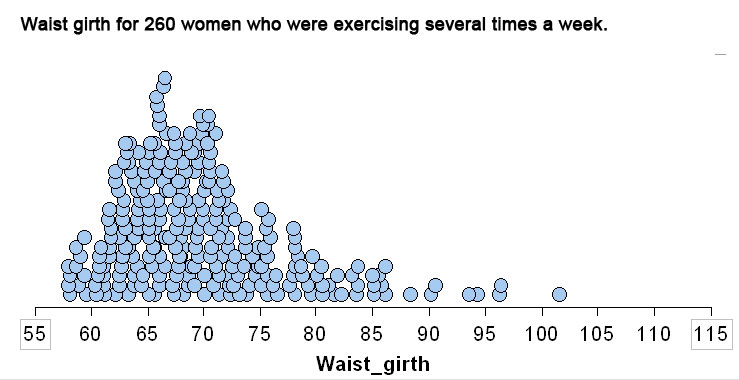
1. How many hatchbacks get 23 mpg in the city?
2. What are the mpg rates that occur most frequently?
3. Which mpg rate is the middle rate? What did you do to find this "middle"?
4. Which graph gives a more immediate summary of the distribution for the MPGCity variable (the case-value graph on the first page of this activity or the dotplot on the second page)? Explain.
5. Use the dotplot to summarize the data even further.



1. What is the overall range for the mpg values?
2. What is a typical range (the main clump of data) for the mpg values?
3. What is the average mpg?
4. What are the unusual mpg rates?
5. For each of the following dotplots, draw a smooth curve outlining the distribution, and then describe the shape of the distribution.





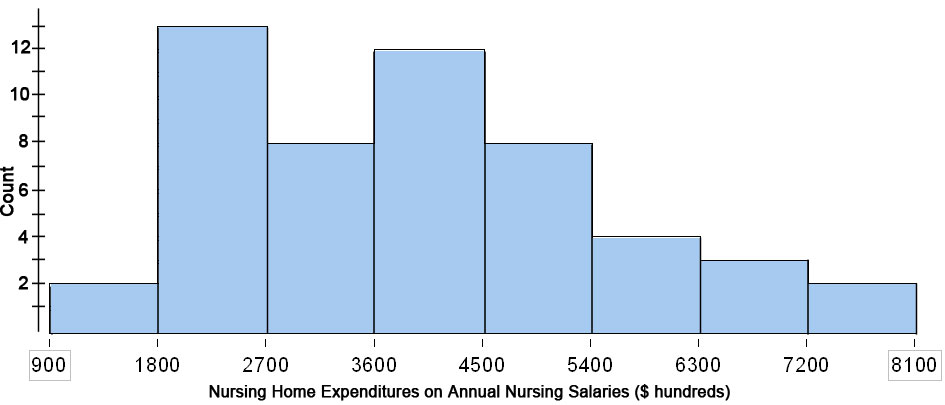


## 2.1.2 Distributions for Quantitative Data (Histograms)

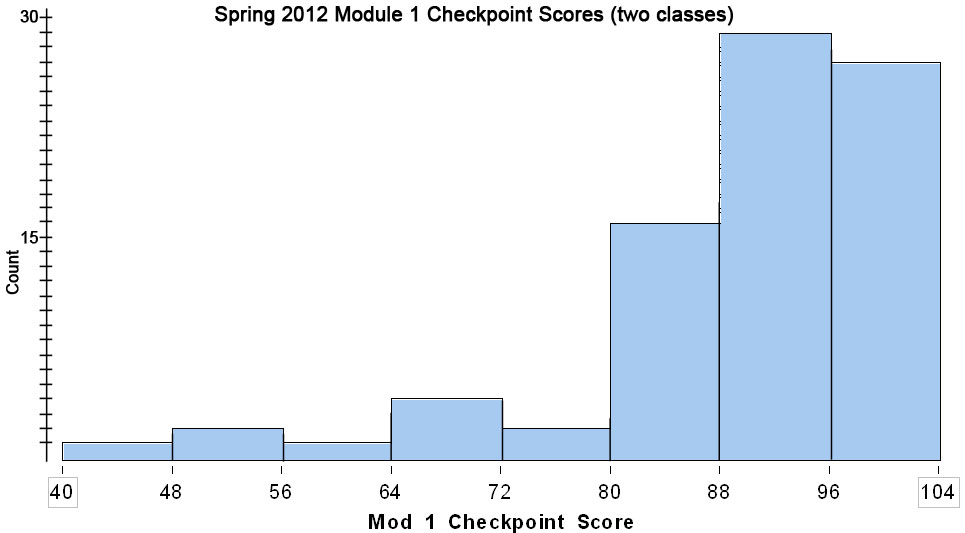
Learning Objective

* For the distribution of a quantitative variable, describe the overall pattern (shape, center, and spread) and striking deviations from the pattern.

1. The data in the histogram below were collected by the Department of Health and Social Services of the State of New Mexico and cover 52 of the 60 licensed nursing facilities in New Mexico in 1988. The quantitative variable *Expenditures on Nursing Salaries* is in hundreds of dollars.

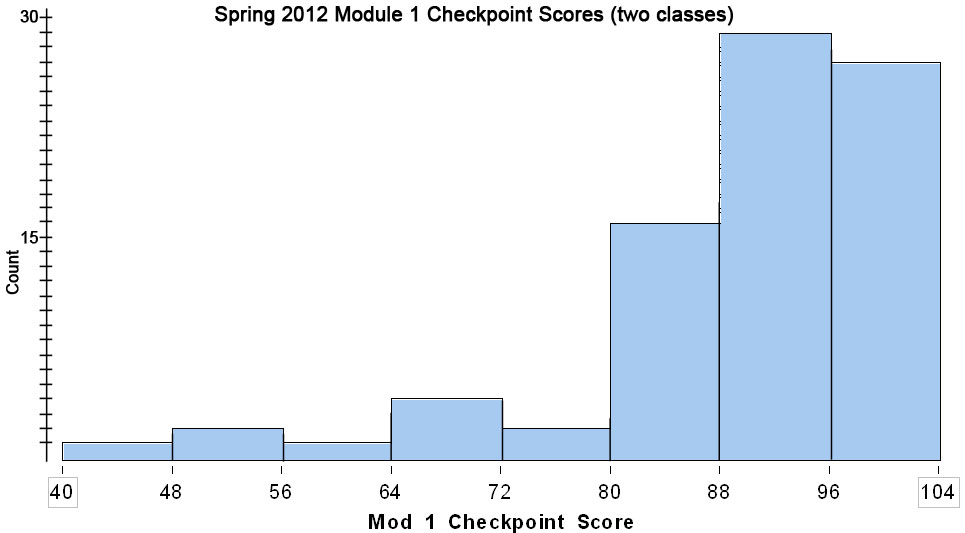


1. How many nursing homes spent between $450,000 and $540,000 annually on nursing salaries?
2. What was the frequency of nursing homes that spent between $180,000 and $270,000 annually on nursing salaries?
3. What percentage of nursing homes spent between $450,000 and $540,000 annually on nursing salaries (this percentage is called the **relative frequency**)? Explain how you found the relative frequency of nursing homes that spent between $450,000 and $540,000 annually on nursing salaries.
4. What percentage of nursing homes spent between $180,000 and $270,000 annually on nursing salaries (again … this percentage is called a **relative frequency**)?
5. The following is a histogram indicating the distribution of scores on the Spring 2012 Module 1 Checkpoint for my combined classes.

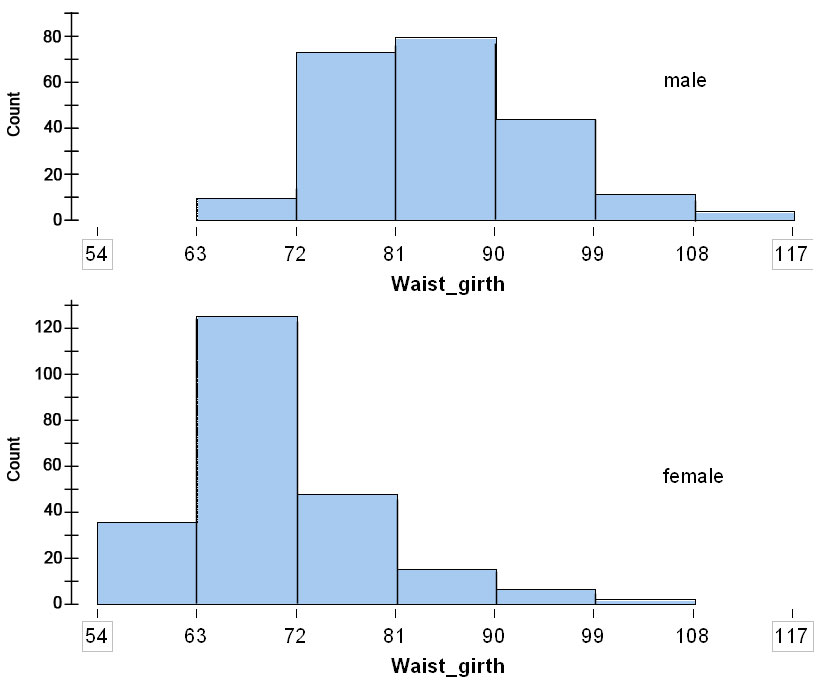


What percentage of students scored below 80% (assume left-hand endpoints are included in each bin)?

1. For each of the following questions, answer the question if the histogram in number 2 above provides enough information to answer it. If not, write "not enough information".
2. What percentage of the students who took the exam scored between 88 and 96 points?
3. What is the lowest grade on the Module 1 Checkpoint?
4. What percentage of students scored less than 90?
5. How many students did not pass the exam, if "not passing" is a score of 70 or less?
6. The following is a histogram indicating the distribution of scores on the Spring 2012 Module 1 Checkpoint for my combined classes.



1. Describe the shape of the distribution of the Module 1 Checkpoint Scores in the histogram above (try to write your response like a statistician).
2. Describe where the center appears to be in this distribution (again … try to write like a statistician).
3. Describe the range for this data (you should come up with a single number).
4. Describe any apparent outliers.
5. Here are data from adults (247 men and 260 women) who were exercising several hours a week. Indicate whether you think the following statements are **valid** or **invalid** (and try to explain why).



1. Typical females have a smaller waist girth than males.
2. The overall range in waist girth is smaller for females than for males.
3. A medium size pair of pants will fit a woman with a waist girth between 72 and 76 centimeters, so I medium size pair of pants with fit about 20% of the women in this sample.

## 2.2.1 Measures of Center (Getting a Feel for "Typical Values")

Learning Objective

* Summarize and describe the distribution of a quantitative variable in context. Describe the overall pattern (shape, center, and spread) and striking deviations from the pattern.

**Warm-up**

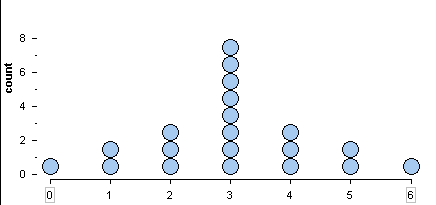
We'll calculate the mean height in cm of the students in this class (and we'll save your height data in an Excel file for a Module 7 activity). First, you'll need to calculate your height in centimeters using the following formula (round the result to one decimal place).

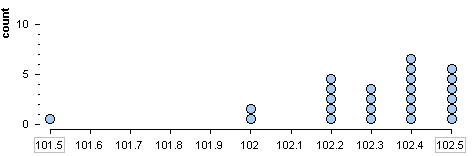
Height in Centimeters = (Height in Inches) 0.393700787 **Height in Cm =**

Next – go to front of the room and enter your height in centimeters in the Excel file.

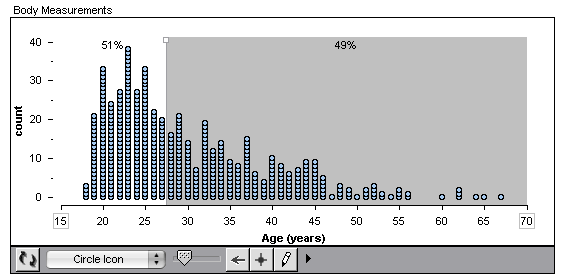
After all students have entered their heights in the Excel file, calculate the mean height.

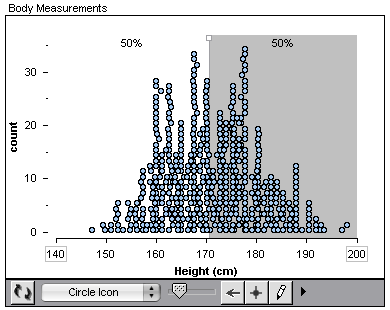
1. Use the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.
2. Find five digits that have a median of 7 and a mean of 7 (repeats allowed).
3. Pick five digits that have a median of 7 and a mean that is less than 7 (repeats allowed.) Give the mean and median of your 5 digits. Find a different set of 5 digits that work.
4. Pick five digits that have a median of 7 and a mean that is more than 7 (repeats allowed.) Give the mean and median of your 5 digits. Find a different set of 5 digits that work
5. Imagine that you have a bag filled with 8 numbers such that the mean and the median are both 6.
6. You draw a 4 out of the bag and replace it with a 1. Does the mean of the numbers in the bag get bigger, smaller, or stay the same? What about the median? Provide a detailed explanation of your reasoning.
7. You draw a number out of the bag. It is an 8. You replace it with 8 ones. Does the mean of the numbers in the bag get bigger, smaller, or stay the same? What about the median? Explain.
8. For each of the distributions pictured below, find the mean and the median.





1. For each distribution give an estimate for the median. Then say whether the mean is probably greater than, less than, or about equal to the median and then provide a detailed explanation of your reasoning.





1. Which of the following distributions is likely to have a mean that is smaller than the median? Explain why.
2. the salaries of NBA basketball players
3. repeated measurements of the volume of peanut M & Ms in a "one-pound" bag
4. the scores on our first exam, the Module 1 Checkpoint, where most of the grades were A’s and B’s, and only a few F’s

## 2.2.2 Measures of Center (Balancing on the Mean)

Learning Objective

* Summarize and describe the distribution of a quantitative variable in context. Describe the overall pattern (shape, center, and spread) and striking deviations from the pattern.

1. Borna Jabbapore is struggling in Math 160 (he failed to apply himself at first, but his effort is steadily improving). To date, his Checkpoint quiz scores are 0, 17, 36, 42, 48, 54, 73, and 90 (each Checkpoint is worth 100 points). What is his mean quiz score? Be sure to use the “x-bar” notation, , to indicate the mean . For example if his mean quiz score is 83, you would write .
2. For each of the following, create a dotplot of 8 quiz scores which meets the specified conditions **and has the same average (mean) quiz score as Borna’s**. Indicate the mean on the dotplot with a triangle. Then after you create the dotplot, describe its shape, center, and spread.
3. All the quiz scores are the same.

0 10 20 30 40 50 60 70 80 90 100

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

1. Most of the quiz scores are the same, but one quiz score is 20. Explain what you had to do to keep the mean quiz score the same as Mr. Jabbapore’s.

0 10 20 30 40 50 60 70 80 90 100

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

1. One quiz score is 100. Explain what you had to do to keep the mean quiz score the same as Borna’s.

0 10 20 30 40 50 60 70 80 90 100

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

1. Two quiz scores are 100. Explain what you had to do to keep the mean quiz score the same as Borna’s.

0 10 20 30 40 50 60 70 80 90 100

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

1. What is the maximum number of 100-point quiz scores possible with a mean quiz score the same as Borna’s? How do you know?
2. Draw two dotplots of 8 quiz scores such that: a) both have the same mean as Borna’s (indicate the mean with a triangle), b) one has very little spread, and c) one has a lot of spread. Be sure to use a triangle to indicate the mean on each graph.

0 10 20 30 40 50 60 70 80 90 100

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

0 10 20 30 40 50 60 70 80 90 100

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

1. Write Borna’s mean quiz score below, then record each of his individual quiz scores in the table below. Finally, for each quiz score, record its distance from the mean (use negative numbers to represent “below” the mean and positive numbers to represent “above” the mean).

Borna’s Mean Quiz Score:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Quiz Score** |  |  |  |  |  |  |  |  |
| **Distance from the Mean** |  |  |  |  |  |  |  |  |

Now add up the signed distances from the mean. What value did you get?

1. Make up a list of eight quiz scores between 0 and 100 points with a mean greater than or equal to 75 AND more quiz scores below the mean than at or above the mean. Record the mean quiz score, and then calculate the “signed” distances from the mean and include these in the table.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Quiz Score** |  |  |  |  |  |  |  |  |
| **Distance from the Mean** |  |  |  |  |  |  |  |  |

Now add up the signed distances from the mean. What value did you get?

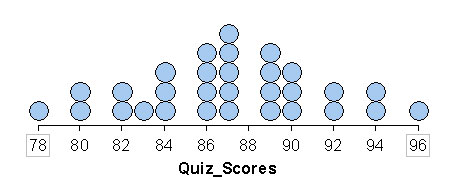
1. For any set of data, if we add up the signed distances from the mean, do you think we will always get the same number? Provide a detailed explanation of your reasoning.
2. Explain why folks call the mean the “balancing point”.

## 2.2.3 Measures of Center (The Median)

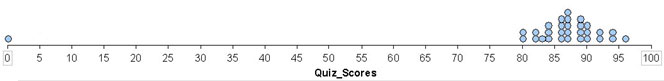
Learning Objectives

* Summarize and describe the distribution of a quantitative variable in context. Describe the overall pattern (shape, center, and spread) and striking deviations from the pattern.
* Relate measures of center and spread to the shape of the distribution. Choose the appropriate measure for different contexts.

1. Borna has taken another quiz. His quiz scores are now 0, 17, 36, 42, 48, 54, 73, 90, and 99. Find his median quiz score.
2. Below is a dotplot of Hilda Branstetter’s Math 160 quiz scores.



1. Find Hilda’s median quiz score.
2. Find Hilda’s mean quiz score.
3. Which would best describe Hilda’s overall performance on the quizzes, the mean or the median?
4. Describe the shape of this distribution.
5. Ooops, we recorded Hilda’s first quiz score wrong. Rather than a 78, she earned a 0 on the first quiz (she was sick and missed class that day). Here’s the revised dotplot of her quiz scores.



1. Find her revised median quiz score. Is it less than, more than, or about the same as her previous median quiz score? Explain why.
2. Find her revised mean quiz score. Is it less than, more than, or about the same as her previous mean quiz score? Explain what happened.
3. Which would best represent her performance on the quizzes, the mean or the median? Explain.
4. Describe the shape of this distribution.
5. You’re being recruited by a firm to apply for a job, but the application process involves a lot of work and is very time consuming, and you know that new hires typically put in very long hours for the first couple of years. So you decide to look at the firm’s typical salaries to determine if it will be worth your effort to go through the application process (not to mention those long hours for a couple of years if you get the job).
6. While investigating the salary information, you learn that the mean salary is $175,000 per year. Will you apply for the job (you know the work will be interesting, and you suspect that you will enjoy the folks you will be working with, but those long hours for the first couple of years … hmmm)?
7. You’re just about to start the application process when you learn that the median salary is $19,500 per year (and remember … those first couple of years you’ll be putting in some really long hours each and every week). What happened? How could the mean salary be $175,000 and the median salary only be $19,500 per year?
8. If you could only learn about one measure of center before deciding to invest your time and energy in the application process, which would you like to know, the mean or the median? Explain your choice.

## 2.3 Quantifying Variability Relative to the Median

(This is a modified version of a Los Medanos College activity developed for their Path2Stats program.)

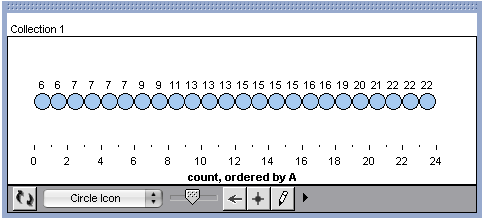
Learning Objectives

* Create and interpret different graphs of the distribution of quantitative data.
* Summarize and describe the distribution of quantitative data in context. Describe the overall pattern (shape, center, and spread) and striking deviations from the pattern.
* Compare distributions from two or more groups.

**Variability, Quartiles, and the Interquartile Range**

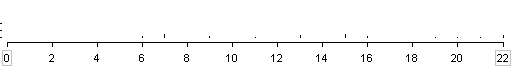
A statistician will describe a distribution by describing the shape, giving a measure of central tendency (mean or median), and also giving a measure of variability.

1. Find the median for this set of data.

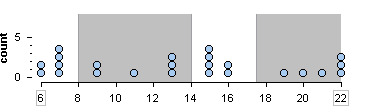


Statisticians measure spread relative to the median by marking the quartiles. Quartiles divide the data into four groups, with 25% of the data in each group. The median is the second quartile.)

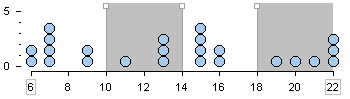
Mark the quartiles on the number line. Let's create a boxplot for these data. Use the quartile marks to make a box. Then mark the lowest data value and the highest data value. Connect these to the box with a horizontal line.



1. When we have a distribution, we can set dividers with “equal widths” or “equal counts”.



* 1. For Graph A are the four areas created with “equal widths” or “equal counts” dividers?

Which kind of dividers are used in Graph B (equal widths or equal counts)?

* 1. Write percents representing the percent of data in each of the four shaded sections for both graphs.
  2. Which kind of divider is related to a boxplot? Why does this make sense?

Draw the boxplot for this data. Show the percentages above your boxplot.

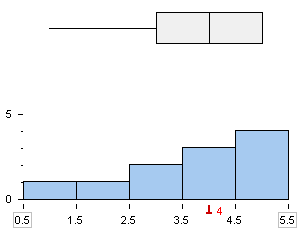
* 1. Which kind of divider is related to a histogram? Why does this make sense?

Draw a histogram with four bars for this data. Show the percentages above your histogram.

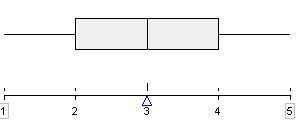
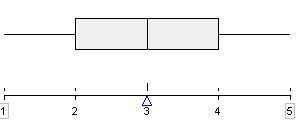
1. Match the histograms to the boxplots.

|  |  |
| --- | --- |
| Picture 25 | Picture 19 |
| Picture 17 | Picture 23 |
| Picture 15 | Picture 16 |

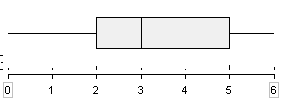
1. Make up a data set (n=11) that fits this pair of graphs. Then find a different data set (n = 11) that also fits these two graphs.



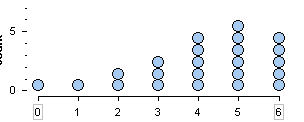
1. Make up a data set (n=10) with the most amount of spread possible that fits this boxplot. Then make up a data set (n=10) with the least amount of spread possible that fits this boxplot.

1. Make up a data set with 10 numbers that matches this boxplot. Make a histogram of your data.



1. Draw a boxplot for the data shown in the dotplot.



## 2.4 Quantifying Variability Relative to the Mean

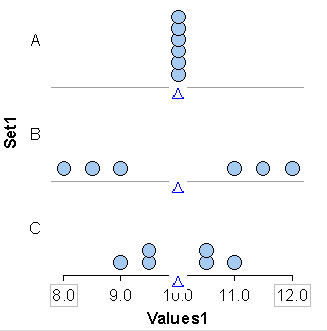
(This is a modified version of a Los Medanos College activity created for their Path2Stats project).

Learning Objectives

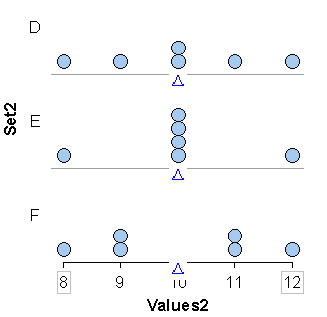
* Summarize and describe the distribution of quantitative data in context. Describe the overall pattern (shape, center, and spread) and striking deviations from the pattern.
* Relate measures of center and spread to the shape of the distribution. Choose appropriate measures for different context.

**Exploring the notion of variability (we have not defined it yet)**

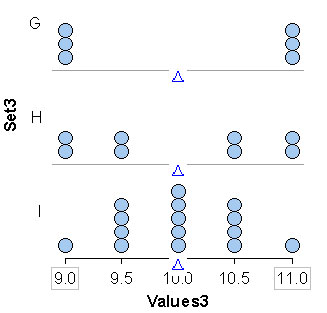
1. Below is a set of three distributions labeled A, B, and C. Assume the scales to be the same on each horizontal axis.
2. Use your own visual sense to order this set of distributions from least amount of variability to most amount of variability.
3. Explain what you observed or did to order this set of distributions from least amount of variability to most amount of variability.



1. Below is another set of three distributions labeled D, E, and F. Assume the scales to be the same on each horizontal axis.
2. Use your own visual sense to order this set of distributions from lest amount of variability to most amount of variability.
3. Explain what you observed or did to order this set of distributions from least amount of variability to most amount of variability.



1. Below is another set of three distributions labeled G, H, and I. Assume the scales to be the same on each horizontal axis.
2. Use your own visual sense to order this set of distributions from lest amount of variability to most amount of variability.
3. Explain what you observed or did to order this set of distributions from least amount of variability to most amount of variability.



1. Get ready to speed date. Work with your group members to prepare a poster presentation of your work from 1, 2, and 3 above. Choose one member of the group to be the presenter (the presenter will explain what y’all did to order each set of distributions). The other members of your group will be the “daters”.
2. (Wait for the mini-lecture to do this one) Explain how to find ADM and then find the ADM for each distribution in Set 3.
3. (Again … wait for the mini-lecture to do this one). Explain how to find the standard deviation (or just write down the formula) and then find the standard deviation for each distribution in Set 3.

# **MODULE 3**

# **Examining Relationships:**

# **Quantitative Data**

## 3.1.1 Scatterplots

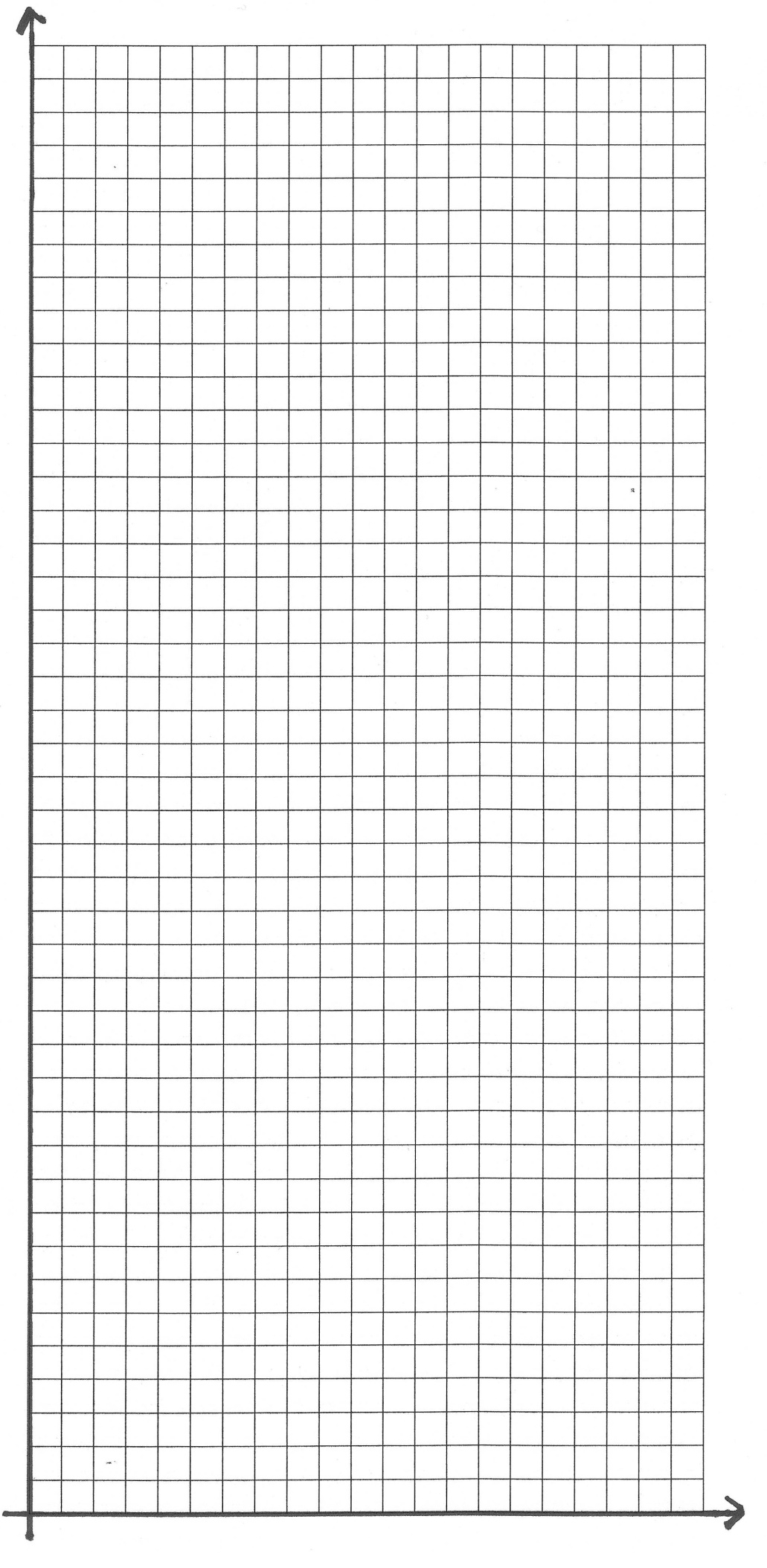
Learning Objectives

* Use a scatterplot to display the relationship between two quantitative variables.
* Describe the overall pattern and striking deviations from the pattern.

The following table describes the number of AIDS related deaths in New York City. Let *n* represent the number of AIDS related deaths in hundreds at time *t* (in years) since 1980 (in other words *t* = 6 would represent 1986 and *n* = 2 would be 200 deaths).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **AIDS and HIV Deaths in  New York City (1981 – 1994)** | | | Year | Number of Deaths (in hundreds) | | 1981 | 1 | | 1982 | 2 | | 1983 | 6 | | 1984 | 11 | | 1985 | 18 | | 1986 | 27 | | 1987 | 34 | | 1988 | 43 | | 1989 | 54 | | 1990 | 57 | | 1991 | 65 | | 1992 | 70 | | 1993 | 74 | | 1994 | 84 | | 1. Why would we say that each of these variables is a quantitative variable as opposed to a categorical variable? 2. Which is the explanatory variable and which is the response variable? |

1. Use the grid on the back of this page to make a scatterplot of the data (i.e. plot the data points). Be sure to place the explanatory variable on the horizontal axis (a.k.a. the x-axis) and the response variable on the vertical axis (a.k.a. the y-axis), and label your axes. Also, select the scales on each axis so that you use as much of the grid as possible. Which basic shape do these data make in the scatterplot?



1. What does each dot represent in your scatterplot?
2. Now we'll describe the overall pattern of the relationship between these two variables (using **direction**, **form**, and **strength**) and striking deviations (if any) from the overall pattern.
3. What is the **direction** of your scatterplot – positive (increasing), negative (decreasing), or neither?
4. What overall **form** does the data in your scatterplot have (linear or curvilinear – if you don't know what these are – ask the teacher).
5. What is the **strength** of the relationship of the data in your scatterplot (**you may want to work on the next numbered problem before answering this**).
6. Are there any **striking deviations** from the overall pattern (a.k.a. **outliners**)?
7. These scatterplots show various body measurements for 34 adults who exercise several times each week. Describe the overall pattern with regard to direction, form, and strength. Then describe any striking deviations from the pattern.

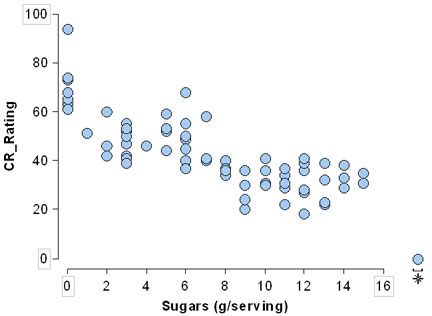
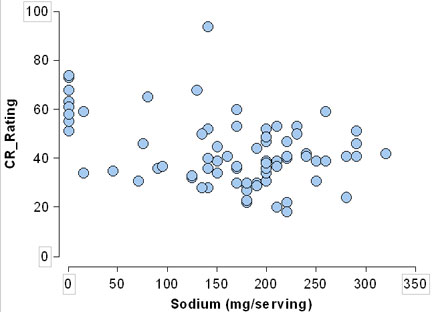
|  |  |  |
| --- | --- | --- |
| ***Scatterplot 1*** | ***Scatterplot 2*** | ***Scatterplot 3*** |
| Picture 6 | Picture 4 | Picture 5 |

1. For each scatterplot, describe the overall pattern of the relationship between the two variables (using direction, form, and strength) and striking deviations (outliers) from the overall pattern.

|  |  |  |
| --- | --- | --- |
| ***Scatterplot 4*** | ***Scatterplot 5*** | ***Scatterplot 6*** |
| Picture 8 | Picture 6 | Picture 7 |

|  |  |  |
| --- | --- | --- |
| ***Scatterplot 7*** | ***Scatterplot 8*** | ***Scatterplot 9*** |
| Picture 8 | Picture 5 | Picture 2 |

1. Each scatterplot below relates one ingredient to the Consumer Report rating for 77 breakfast cereals. For each scatterplot, describe the overall pattern of the relationship between the two variables (using direction, form, and strength) and striking deviations (outliers) from the overall pattern. Also, for each scatterplot, indicate the explanatory and response variables and then describe what each dot represents

Overall pattern: Overall pattern:

Explanatory variable: Explanatory variable:

Response variable: Response variable

Each dot represents: Each dot represents:

1. Match each pair of variables to a scatterplot, and label the variables on the appropriate axes. Briefly explain your reasoning.
   * 1. x = city miles per gallons and y = highway miles per gallon for 10 cars
     2. x = sodium (mg/serving) and y = Consumer Report quality rating for 10 salted peanut butters
     3. x = price ($) and y = Consumer Report quality rating for 10 bicycle helmets

|  |  |  |
| --- | --- | --- |
| ***Scatterplot 10*** | ***Scatterplot 11*** | ***Scatterplot 12*** |
| Picture 3 | Picture 12 | Picture 1 |

1. For each of the scatterplots in the previous problem, describe what a dot represents.
2. Oops … we did not quite have all the AIDS related death data for New York City.

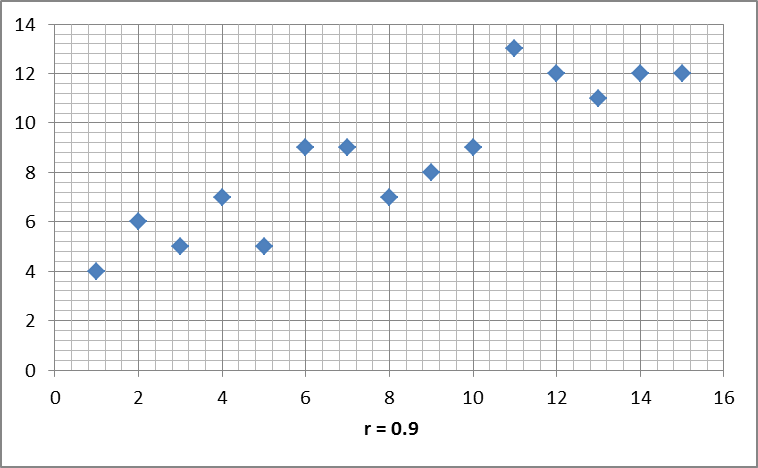
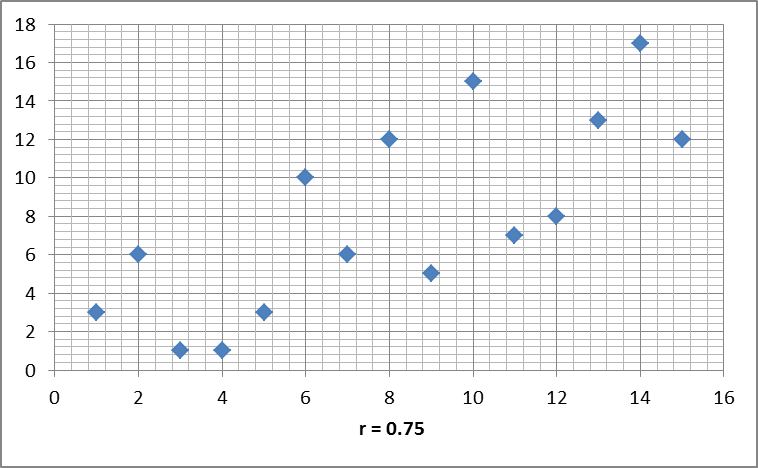
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **AIDS and HIV Deaths in  New York City (1981 – 2010)** | | | Year | Number of Deaths | | 1981 | 59 | | 1982 | 201 | | 1983 | 593 | | 1984 | 1107 | | 1985 | 1828 | | 1986 | 2720 | | 1987 | 3350 | | 1988 | 4300 | | 1989 | 5358 | | 1990 | 5724 | | 1991 | 6475 | | 1992 | 6985 | | 1993 | 7429 | | 1994 | 8355 | | 1995 | 8322 | | 1996 | 6078 | | 1997 | 3428 | | 1998 | 2795 | | 1999 | 2805 | | 2000 | 2710 | | 2001 | 2577 | | 2002 | 2554 | | 2003 | 2520 | | 2004 | 2387 | | 2005 | 2318 | | 2006 | 2090 | | 2007 | 1988 | | 2008 | 1940 | | 2009 | 1589 | | 2010 | 1413 | | Here is the scatterplot for this data.  Grid-AIDS-Deaths-to-2010   1. Could we say that the form of this data is linear? Curvilinear? Explain. 2. Is it possible to force this data set to have a linear form?   **WARNING: Soon - you will need to bring your graphing calculator to class every day!** |

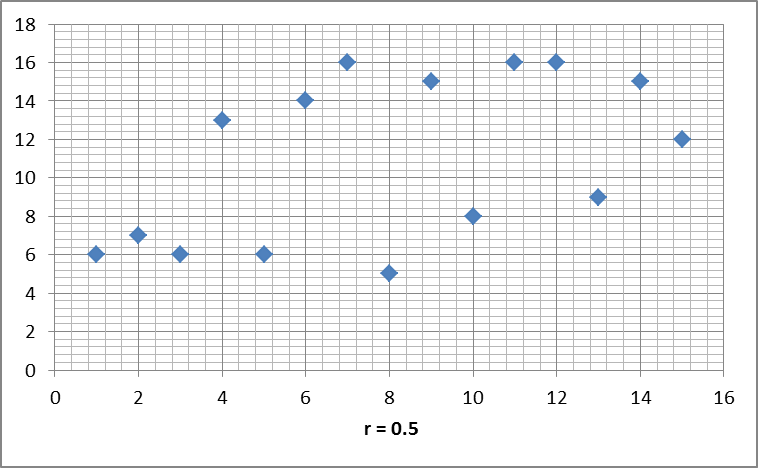
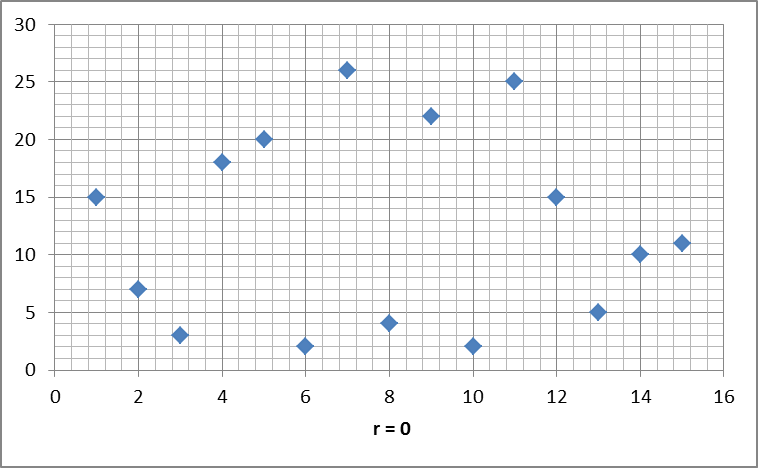
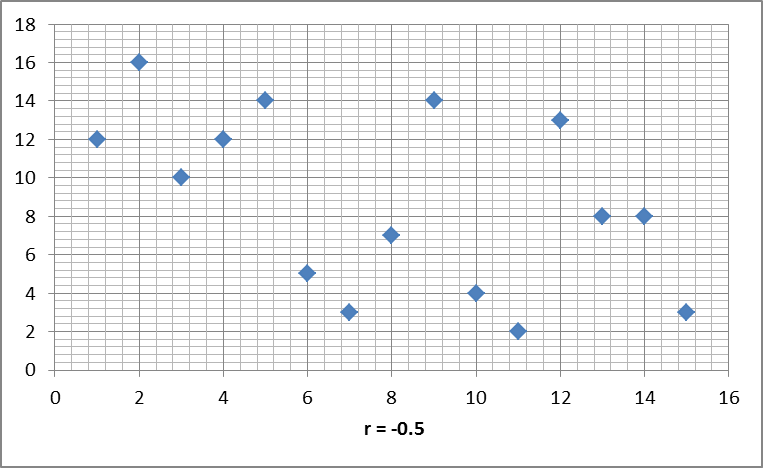
## 3.1.2 Linear Relationships and Correlation Coefficients (Introduction)

Learning Objectives

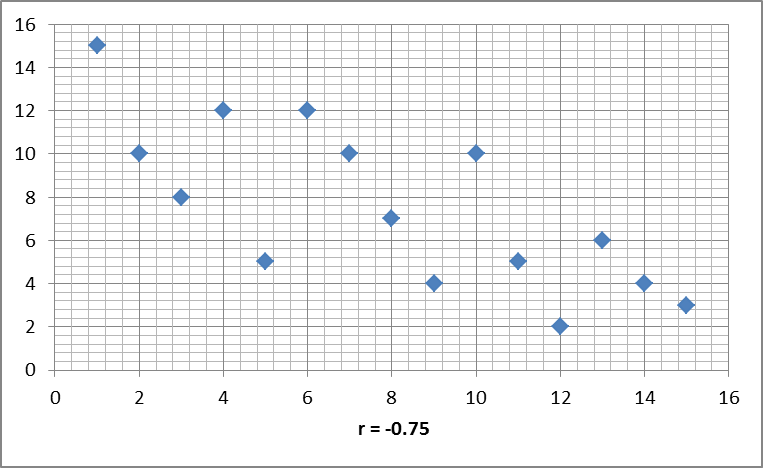
* Interpret the value of the correlation coefficient.
* Recognize the limitations of the correlation coefficient as a measure of the relationship between two quantitative variables.

1. For each of the scatterplots rank the strength of relationship (how well it fits the form – which is linear for each of these scatterplots). Use only rankings from 0 to 1 where 0 indicates that the data does not fit the form and a 1 indicates that the data is perfectly linear. You may use rankings such as 0.7 etc. If the direction is negative, then indicate this by attaching a negative sign to your ranking.

… continued on the back

1. Let r be each the rankings you guesstimated for each of the scatterplots in the previous problem. Then you guesstimated the correlation coefficient for each scatterplot. Here are the actual correlation coefficients for the scatterplots. Go back and label each with the correlation coefficient that is the best indicator of the strength of the relationship for the data (be sure to use the notation “r = ” with the number). r = 0, r = 0.5, r = -0.5, r = 0.75, r = -0.75, r = 0.9, r = -0.9, r = 1, r = -1
2. Create two scatterplots each with 10 data points (one with a correlation coefficient of r = -0.8 and one with a correlation coefficient of 0.8).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **WARNING: Soon - you will need to bring your graphing calculator to class every day!** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

## 3.1.3 Causation and Lurking Variables

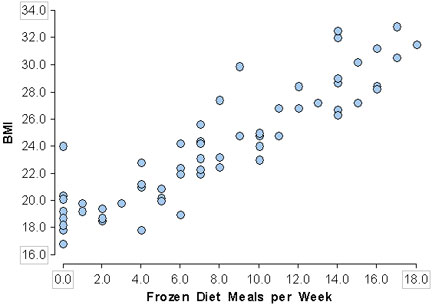
Learning Objectives

* Distinguish between association and causation.
* Identify lurking variables that may explain an observed relationship.

1. For each of the following scatterplots:
2. identify the explanatory and response variables,
3. guesstimate the correlation coefficient r (where the correlation coefficient indicates both the strength of the association and the direction of the relationship between the variables),
4. if the correlation coefficient indicates a strong association, write a sentence indicating that the explanatory variable “causes” the response variable (even if it doesn't), and
5. determine whether or not the sentence you wrote in part c) makes sense (explain why or why not).

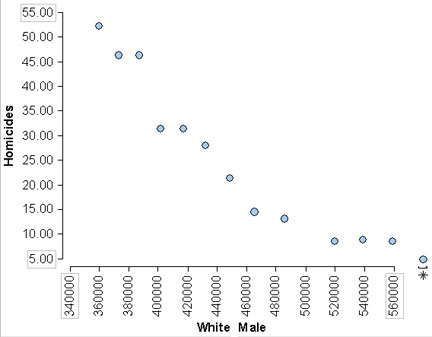
This scatterplot represents 60 female Cuyamaca College students between the ages of 18 and 35 who were selected at random. They were asked how many frozen diet meals (on average) did they eat each week. Then their body mass index (BMI) was measured. A BMI under 18.5 is considered “underweight”. A BMI between 18.5 and 24.9 is considered “normal weight”. A BMI between 25 and 29.9 is considered overweight. And a BMI of 30 or more is considered “obese”.

1. Explanatory & response variables:
2. Guesstimate correlation coef:
3. Causation sentence:
4. Sentence makes sense (explain):



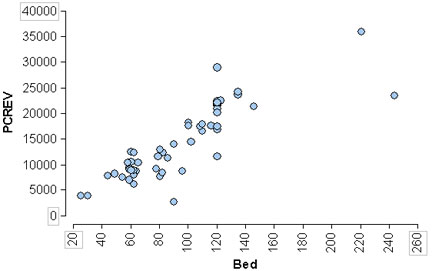
These data were collected by J.C. Fisher and used in his paper: "Homicide in Detroit: The Role of Firearms", Criminology, vol.14, 387-400 (1976). The data are on the homicide rate in Detroit for the years 1961-1973. The number of “Homicides” are given per 100,000 people, and “White Males” indicates the white male population in Detroit.

1. Explanatory & response variables:
2. Guesstimate correlation coef:
3. Causation sentence:
4. Sentence makes sense (explain):



These data were collected by the Department of Health and Social Services of the State of New Mexico and cover 52 of the 60 licensed nursing facilities in New Mexico in 1988. “PCREV” is the annual total patient care revenue for the nursing home (in hundreds of dollars). “Bed” is the number of beds in the nursing home.

1. Explanatory & response variables:
2. Guesstimate correlation coef:
3. Causation sentence:
4. Sentence makes sense (explain):



**WARNING: Soon - you will need to bring your graphing calculator to class every day!**

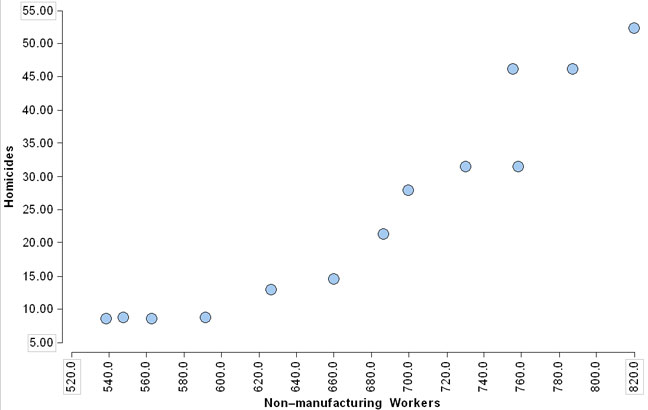
## 3.2 Fitting a Line

**WARNING: You need your graphing calculator to complete this activity!**

Learning Objectives

* For a linear relationship, use the least squares regression line to summarize the overall pattern and to make predictions.

1. These data were collected by J.C. Fisher and used in his paper: "Homicide in Detroit: The Role of Firearms", Criminology, vol.14, 387-400 (1976). The data are on the homicide rate in Detroit for the years 1961-1973. The number of “Homicides” are given per 100,000 people, and “Non-Manufacturing Workers” indicates the number of non-manufacturing workers in thousands.



1. Use a straight edge to draw a straight line that "best fits" the data (note: your best-fit line need not pass through any of the data points).
2. Label the point on your line that corresponds to 610,000 non-manufacturing workers in Detroit. Now use your line (and the point you just labeled) to predict the number of homicides when there are 610,000 non-manufacturing workers in Detroit.
3. Use the graph of your best-fit line to predict the number of homicides when there are 790,000 non-manufacturing workers in Detroit.
4. Use the graph of your best-fit line to estimate the number of non-manufacturing workers in Detroit when there are 40 homicides.
5. Now label any two points on your line (again these points need not actually be data points), and you may use the point(s) you are already labeled if you want to.

***O.K. we're going to take a little sidetrack and come back to this problem later …***

1. Mathematicians often use the slope-intercept formula, , to represent the equation of a line that best fits a scatterplot. Which letter in this formula represents the slope of a line and which letter represents the y-intercept?
2. However, statisticians often write the slope-intercept formula for the equation of a line as . Is the formula essentially the same as the formula? If not, how is it different? If so, which letter represents the slope and which letter represents the y-intercept in our new linear formula, ?

***O.K. let's get back to the scatterplot and our best-fit line.***

1. Write the two points you labeled on the scatterplot (from the previous page) in the space below. Then use the points to find the equation of your best-fit line in our "new" slope-intercept form (don't worry if you're rusty with finding the slope and y-intercept … StatCrunch and/or your graphing calculator will be able to do this for you … but not every Econ or Biology teacher etc. will let you use your graphing calculator, so we need to practice this skill).
2. Now rewrite the **equation** of your best-fit line in the space below and then use it (not the actual line you drew on the scatterplot) to answer the following questions. Be sure to indicate what each variable represents.

Equation of the best-fit line:

1. How many homicides should we expect in Detroit when there are 633,000 non-manufacturing workers? Does your result make sense, why or why not?
2. How many homicides should we expect in Detroit when there are zero non-manufacturing workers? Does your answer make sense, why or why not?
3. In part a) we used **interpolation** to make a prediction, and in part b) we used **extrapolation** to make a prediction (perhaps we should define these terms together in class right now).

Interpolation:

Extrapolation:

1. Which do you think is a more reliable predictor, interpolation or extrapolation? Why or why not?
2. The formulas below are used by statisticians to find the slope and y-intercept of the one-and-only true best fit line.

For the best-fit line the slope is and the y-intercept is .

Now we're going to find the slope *b* and y-intercept *a* of the best-fit line. Since *b* is part of the formula for the y-intercept, *a*, we need to calculate the slope *b* first.

1. Identify each part of the best-fit slope (the first one is done for you).

is the correlation coefficient.

is

is

1. Use the Starbucks data listed below to calculate each part of the best-fit slope .

Step 1: Find the mean values:

|  |  |
| --- | --- |
| **Years (since 1990)** | **# of SB Stores** |
| 0 | 84 |
| 1 | 116 |
| 2 | 165 |
| 3 | 272 |
| 4 | 425 |
| 5 | 676 |
| 6 | 1015 |
| 7 | 1412 |
| 8 | 1886 |
| 9 | 2498 |
| 10 | 3501 |
| 11 | 4709 |
| 12 | 5886 |
| 13 | 6294 |

Step 2: Find the standard deviations.

Recall and .

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  | Fill out the table (round to 4 decimal places), and then use the values in the table to find each of the following.  =  =  =  = |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Step 3: Now let's find the correlation coefficient .

Rewrite the values found on the previous page here.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

Fill out the table below and round to 4 decimal places (note - you can copy your calculations for and from the table you filled out on the previous page).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Calculate and round to 4 decimal places.

1. Now we're finally ready to calculate the best-fit slope . Be careful … you need to use in the top of the fraction and in the bottom of the fraction.

1. Now we just need to calculate the y-intercept . Oh … but there's good news. We've already found all the parts. We just need to gather them up.

1. Now put it all together … write the equation of the one-and-only true best-fit line, .

Best-fit line:

1. Let's compare the one-and-only true best-fit line you found in part e) above and the equation of the line you graphed with the Starbucks scatterplot on the first page (you found the equation of your sketched line in number 4 on page 2). Graph both equations with the Starbucks data in the same graphing window. How does each line compare the data in scatterplot? How do the lines compare? Is each a good fit with the data?

|  |  |
| --- | --- |
| **Years (since 1990)** | **# of Starbucks Stores** |
| 0 | 84 |
| 1 | 116 |
| 2 | 165 |
| 3 | 272 |
| 4 | 425 |
| 5 | 676 |
| 6 | 1015 |
| 7 | 1412 |
| 8 | 1886 |
| 9 | 2498 |
| 10 | 3501 |
| 11 | 4709 |
| 12 | 5886 |
| 13 | 6294 |

1. O.K. now we'll use our graphing calculators to find the equation of the one-and-only true best-fit line (a.k.a **the least squares regression line** or **regression equation** or **regression line** etc.) for the Starbucks data. No need to take notes (we'll do this together several times during the next week, and then you can take notes … just walk through it with me this time).

**WARNING**! When you use your graphing calculator to find the equation of the best-fit line, be sure that the correlation coefficient *r* is displayed with the regression equation. After we find the regression line, if you don't see *r*, ask me about it.

**Please continue to bring your graphing calculator to class every day.**

## 3.3.1 Residuals and the Best Fit Line

**Activity 1**

1. Vitruvius connected the proportions of the male figure to the proportions used in classical architecture and wrote that a man’s arm span is equal to his height. Below is a scatter plot and best-fit line (also called a **least squares regression line** or just a **regression line**) for the arm span and height measurements (in cm) for 11 men.
2. Use the data points to find the observed heights (heights of actual men) for the indicated arm spans.

Arm span = 173cm Observed Height =

Arm span = 196cm Observed Height =

1. Use the graph of the least squares regression line (a.k.a. best-fit line) to predict the height of a man with the same arm spans.

Arm span = 173cm Predicted Height =

Arm span = 196cm Predicted Height =

Obviously the graph of the least squares regression line does not predict the height of every man accurately.

1. Consider the man with arm span of 173cm. Is the predicted height from the graph of the least squares regression line (the best-fit line) an over estimate or an underestimate?
2. What is the error (a.k.a. the residual) in the predicted height for a man with an arm span of 173cm (the error or residual is the observed height minus the predicted height)?
3. Consider the man with arm span of 196cm. Is the predicted height from the graph of the least squares regression line (the best-fit line) an over estimate or an underestimate?
4. What is the error (a.k.a. the residual) in the predicted height for a man with an arm span of 196cm (the error or residual is the observed height minus the predicted height)?
5. The table below gives the height (in feet) and speed (in mph) for twenty roller coasters. The equation of the best-fit line (i.e. the least squares regression line) for these data is given below.

**predicted speed** = 40.31 + 0.18(**height**)

**Roller Coasters**

1. Use the data from the table to find the observed speeds (i.e. the actual speeds) for roller coasters with the following heights.

Height = 72ft Observed Speed =

Height = 415ft Observed Speed =

1. Use the equation to find the predicted speed for roller coasters with the following heights.

Height = 72ft Predicted Speed =

Height = 415ft Predicted Speed =

1. Consider the roller coaster with a height of 72 ft. Is the predicted speed an over estimate or an underestimate?
2. Consider the roller coaster with a height of 415 ft. Is the predicted speed an over estimate or an underestimate?
3. What is the error (a.k.a. the residual) in the predicted speed for a roller coaster that is 415 ft?

|  |  |
| --- | --- |
| **Height** | **Speed** |
| 57 | 50 |
| 72 | 44 |
| 84 | 50 |
| 105 | 70 |
| 118 | 56 |
| 150 | 72 |
| 160 | 85 |
| 161 | 65 |
| 170 | 74 |
| 181 | 70 |
| 185 | 70 |
| 205 | 82 |
| 208 | 77 |
| 209 | 80 |
| 226 | 70 |
| 249 | 78 |
| 318 | 95 |
| 415 | 100 |
| 420 | 120 |
| 456 | 128 |

**Activity 2**

In the previous lesson, you predicted the value of the response variable knowing the value of the *explanatory variable* (also known as the *predictor variable*) using a best-fit line. So, how do you identify the line that is the best fit? In this class we’ll use technology to find the equation of the best-fit line, but what does it mean to say that a particular line is the best fit? In this lesson, you will investigate this question with the goal of developing a method for determining which line is indeed the best-fit line.

1. Students at a school collected information from classmates in grades 1, 3, 5, and 7, weighing both the students and their backpacks. The scatterplot gives the student’s body weight in pounds and the **backpack weight as a percentage of body weight**. Which line appears to be the best fit?

|  |  |
| --- | --- |
| A)  **Predicted BP % Weight** = -17.7143 + 0.4143**(Body Wt)** | B) **Predicted BP % Weight** = 7.7872 + 0.1219**(Body Wt)** |
|  |  |
|  |  |
| C) **Predicted BP % Weight** = 8.0588 + 0.1412**(Body Wt)** | D) **Predicted BP % Weight** = 19.5 |
|  |  |
|  |  |

1. We’ll do some math to determine if we picked the best-fit line. Begin by using the equation of each line to complete the table below. When necessary, round to four decimal places.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Body Weight** | **Backpack % Weight** | **Line A Predicted BP % Weight** | **Line B Predicted BP % Weight** | **Line C Predicted BP % Weight** | **Line D Predicted BP % Weight** |
| 34 | 9 | -3.6281 | 11.9318 | 12.8596 | 19.5 |
| 44 | 16 | 0.5149 | 13.1508 | 14.2716 | 19.5 |
| 48 | 10 | 2.1721 | 13.6384 | 14.8364 | 19.5 |
| 51 | 18 |  |  |  |  |
| 54 | 15 | 4.6579 | 14.3698 | 15.6836 | 19.5 |
| 59 | 7 | 6.7294 | 14.9793 | 16.3896 | 19.5 |
| 62 | 18 | 7.9723 | 15.3450 | 16.8132 | 19.5 |
| 68 | 32 | 10.4581 | 16.0764 | 17.6604 | 19.5 |
| 72 | 7 | 12.1153 | 16.5640 | 18.2252 | 19.5 |
| 75 | 12 | 13.3582 | 16.9297 | 18.6488 | 19.5 |
| 76 | 21 | 13.7725 | 17.0516 | 18.7900 | 19.5 |
| 82 | 15 |  |  |  |  |
| 95 | 20 | 21.6442 | 19.3677 | 21.4728 | 19.5 |
| 104 | 26 | 25.3729 | 20.4648 | 22.7436 | 19.5 |
| 119 | 18 | 31.5874 | 22.2933 | 24.8616 | 19.5 |

1. Now which of the four lines do you think results in the best overall prediction of backpack weight as a percentage of body weight? Why? How are you selecting the best-fit line?
2. For each linear model, complete the table and then follow the directions under the scatterplot.

|  |  |  |  |
| --- | --- | --- | --- |
| **Body Weight** | **Backpack % Weight** | **Line A Predicted % Weight** | **Error (a.k.a Residual)**  A) **Predicted BP % Weight** = -17.7143 + 0.4143**(Body Wt)**   * Circle the student for which the line comes closest to predicting the backpack weight as a percent of body weight (there is only one). * Circle the student for which the actual backpack weight as a percent of body weight is furthest from the prediction. |
| 34 | 9 | -3.6281 | 12.628 |
| 44 | 16 | 0.5149 | 15.485 |
| 48 | 10 | 2.1721 | 7.8279 |
| 51 | 18 | 3.415 | 14.5850 |
| 54 | 15 | 4.6579 | 10.3421 |
| 59 | 7 | 6.7294 | 0.2706 |
| 62 | 18 | 7.9723 |  |
| 68 | 32 | 10.4581 | 21.5419 |
| 72 | 7 | 12.1153 | -5.1153 |
| 75 | 12 | 13.3582 | -1.3582 |
| 76 | 21 | 13.7725 | 7.2275 |
| 82 | 15 | 16.2583 | -1.2583 |
| 95 | 20 | 21.6442 | -1.6442 |
| 104 | 26 | 25.3729 | 0.6271 |
| 119 | 18 | 31.5874 |  |

B) **Predicted BP % Weight** = 7.7872 + 0.1219**(Body Wt)**

* Circle the student for which the line comes closest to predicting the backpack weight as a percent of body weight (there is only one).
* Circle the student for which the actual backpack weight as a percent of body weight is furthest from the prediction.

|  |  |  |  |
| --- | --- | --- | --- |
| **Body Weight** | **Backpack % Weight** | **Line B Predicted % Weight** | **Error (a.k.a Residual)** |
| 34 | 9 | 11.9318 | -2.9318 |
| 44 | 16 | 13.1508 | 2.8492 |
| 48 | 10 | 13.6384 |  |
| 51 | 18 | 14.0041 | 3.9959 |
| 54 | 15 | 14.3698 | 0.6302 |
| 59 | 7 | 14.9793 | -7.9793 |
| 62 | 18 | 15.3450 | 2.6550 |
| 68 | 32 | 16.0764 | 15.9236 |
| 72 | 7 | 16.5640 |  |
| 75 | 12 | 16.9297 | -4.9297 |
| 76 | 21 | 17.0516 | 3.9484 |
| 82 | 15 | 17.783 | -2.7830 |
| 95 | 20 | 19.3677 | 0.6323 |
| 104 | 26 | 20.4648 | 5.5352 |
| 119 | 18 | 22.2933 | -4.2933 |

C) **Predicted BP % Weight** = 8.0588 + 0.1412**(Body Wt)**

* Circle the student for which the line comes closest to predicting the backpack weight as a percent of body weight (there is only one).
* Circle the student for which the actual backpack weight as a percent of body weight is furthest from the prediction.

|  |  |  |  |
| --- | --- | --- | --- |
| **Body Weight** | **Backpack % Weight** | **Line C Predicted % Weight** | **Error** |
| 34 | 9 | 12.8596 | -3.8596 |
| 44 | 16 | 14.2716 | 1.7284 |
| 48 | 10 | 14.8364 | -4.8364 |
| 51 | 18 | 15.26 | 2.74 |
| 54 | 15 | 15.6836 | -0.6836 |
| 59 | 7 | 16.3896 |  |
| 62 | 18 | 16.8132 | 1.1868 |
| 68 | 32 | 17.6604 | 14.3396 |
| 72 | 7 | 18.2252 | -11.2252 |
| 75 | 12 | 18.6488 | -6.6488 |
| 76 | 21 | 18.7900 |  |
| 82 | 15 | 19.6372 | -4.6372 |
| 95 | 20 | 21.4728 | -1.4728 |
| 104 | 26 | 22.7436 | 3.2564 |
| 119 | 18 | 24.8616 | -6.8616 |

D) **Predicted BP % Weight** = 19.5

* Circle the student for which the line comes closest to predicting the backpack weight as a percent of body weight (there is only one).
* Circle the student for which the actual backpack weight as a percent of body weight is furthest from the prediction.

|  |  |  |  |
| --- | --- | --- | --- |
| **Body Weight** | **Backpack % Weight** | **Line D Predicted % Weight** | **Error** |
| 34 | 9 | 19.5 | -10.5 |
| 44 | 16 | 19.5 | -3.5 |
| 48 | 10 | 19.5 | -9.5 |
| 51 | 18 | 19.5 | -1.5 |
| 54 | 15 | 19.5 | -4.5 |
| 59 | 7 | 19.5 | -12.5 |
| 62 | 18 | 19.5 |  |
| 68 | 32 | 19.5 | 12.5 |
| 72 | 7 | 19.5 | -12.5 |
| 75 | 12 | 19.5 | -7.5 |
| 76 | 21 | 19.5 | 1.5 |
| 82 | 15 | 19.5 | -4.5 |
| 95 | 20 | 19.5 | 0.5 |
| 104 | 26 | 19.5 |  |
| 119 | 18 | 19.5 | -1.5 |

1. Now use the “errors” column in each of the previous four tables to complete the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Which measures of the total error help you determine how well a line fits the data?** | | | |
| *Line* | *Sum of Errors* | *Sum of Absolute*  *Value of Errors (SAE)* | *Sum of Squares of Errors (SSE)* |
| A |  |  |  |
| B |  |  |  |
| C |  |  |  |
| D |  |  |  |

Which total error do you think would indicate the best fit line? Why?

Statisticians square the errors and then find the line that minimizes the sum of the squared errors. The line that has the smallest sum of the squared errors is called the *least squares regression line*. This line minimizes the sum of the squares of the errors, when compared to **all** other possible lines.

1. Recall the formula for the correlation coefficient r which we studied in a previous handout. Why would the correlation coefficient be a good indicator for how well the graph of the least squares regression line fits the data?



## 3.3.2 Assessing the Fit of a Line

Learning Objective

* Use residuals and to assess the fit of a line.

**Warning: We'll work through portions of this activity together as a class, so even if you know how to use your graphing calculator to find the regression equation, please do not work ahead.**

1. Source for the following data: Hosmer, D.W. and Royston, P.R. (2003) "Using Fractional Polynomial to Model Continuous Covariates in Regression Analysis". The data in the scatterplot provide the gestational age in weeks (GAWKS) and the abdominal circumference (AC) in millimeters (mm).

|  |  |
| --- | --- |
| **Abdominal Circumference (AC)** | **Gestational Age in Weeks (GAWK)**   1. Enter these data in your graphing calculator. 2. Create a scatterplot for these data. 3. Looking at your scatterplot, do you think a linear model would be a good fit for this data? If so what do you think the correlation coefficient *r* will be? 4. Use your graphing calculator to find a linear regression equation for these data. Record the value for the linear regression equation here along with the linear regression equation.   Regression equation:    ***O.K. you can work in groups now.*** |
| 62 | 12.85714 |
| 101 | 15.71429 |
| 104 | 14.28571 |
| 129 | 17.85714 |
| 141 | 19.14286 |
| 166 | 20 |
| 176 | 22.28572 |
| 221 | 24.42857 |
| 221 | 26.71428 |
| 266 | 28.85714 |
| 274 | 31.42857 |
| 269 | 33.57143 |
| 315 | 36 |
| 341 | 38.42857 |
| 331 | 39.85714 |
| 383 | 41.57143 |

1. Use your linear regression model to predict the gestational age in weeks when the abdominal circumference is 331 mm. Compare this gestational age with the original data in the scatterplot. Did your linear model overestimate or underestimate the gestational age for an abdominal circumference of 331 mm? To calculate the “signed” error, simply take the actual gestational age in weeks when the abdominal circumference is 331 mm (from the data) and subtract the gestational age in weeks from your model when you input x = 331. So what is the signed error?
2. Use your linear regression model to predict the gestational age in weeks when the abdominal circumference is 104 mm. Compare this gestational age with the original data in the scatterplot. Did your linear model overestimate or underestimate the gestational age for 104 mm? We also call the signed error, the residual. What is the residual for 104 mm?
3. Complete the table below to find the gestational age in weeks residual for each abdominal circumference (coming attraction – in problem number 4 we'll learn how to use our graphing calculators to complete a table like this).

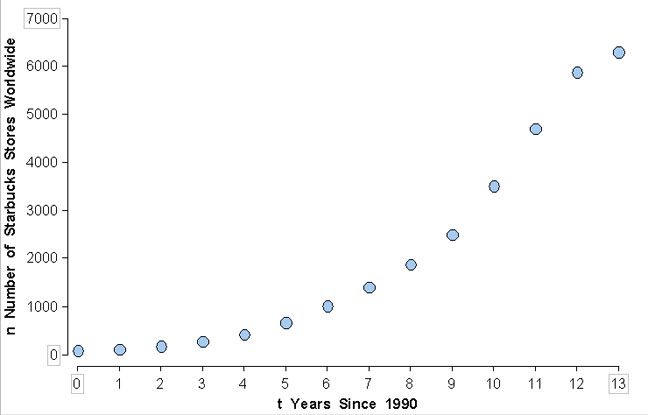
|  |  |  |  |
| --- | --- | --- | --- |
| **AC** | **GAWK** | **Prediction** | **Residual** |
| 62 | 12.85714 |  |  |
| 101 | 15.71429 |  |  |
| 104 | 14.28571 |  |  |
| 129 | 17.85714 |  |  |
| 141 | 19.14286 |  |  |
| 166 | 20 |  |  |
| 176 | 22.28572 |  |  |
| 221 | 24.42857 |  |  |
| 221 | 26.71428 |  |  |
| 266 | 28.85714 |  |  |
| 274 | 31.42857 |  |  |
| 269 | 33.57143 |  |  |
| 315 | 36 |  |  |
| 341 | 38.42857 |  |  |
| 331 | 39.85714 |  |  |

1. Now plot the residuals with their corresponding abdominal circumferences (AC on the horizontal or x-axis and the residual or “signed” error for the GAWK on the vertical or y-axis).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Is there a readily discernible pattern to the residuals scatterplot? If so, what is it?

1. Now let’s find a new regression equation for some Starbucks data. The number of Starbucks stores worldwide has increased substantially since 1991. Let t represent the number of years since 1990 and n represent the number of Starbucks stores worldwide t years after 1990.



1. Which do you think would be the best fit for the Starbucks scatterplot - a linear or curvilinear regression model?
2. The Starbucks data are listed in the table in part d on the next page. Use your graphing calculator to make a scatterplot for these data (it should look like the scatterplot above).
3. Now use your graphing utility to find the **LINEAR** regression equation for this data (even if you decided that a curvilinear regression model would be a better fit). Record the value for the linear regression equation here along with the linear regression equation.

*Continued next page …*

1. Use the linear regression model you found in part c) to find the residuals for the number of Starbuck’s stores corresponding to the year after 1990 and fill in the following table. Hmmm … would like to learn how to use your graphing calculator to complete the table?

|  |  |  |  |
| --- | --- | --- | --- |
| **Year Since 1990** | **Number of Startbucks Stores Worldwide** | **Prediction** | **Residual** |
| 0 | 84 |  |  |
| 1 | 116 |  |  |
| 2 | 165 |  |  |
| 3 | 272 |  |  |
| 4 | 425 |  |  |
| 5 | 676 |  |  |
| 6 | 1015 |  |  |
| 7 | 1412 |  |  |
| 8 | 1886 |  |  |
| 9 | 2498 |  |  |
| 10 | 3501 |  |  |
| 11 | 4709 |  |  |
| 12 | 5886 |  |  |
| 13 | 6294 |  |  |

1. Now use your graphing calculator to make a scatterplot of the residuals with the year since 1990 on the x-axis and the residual or “signed” error in predicting the number of Starbucks stores on the y-axis. Draw the scatterplot in the space below and label the axes appropriately.
2. Is there a readily discernible pattern to the Starbucks residuals scatterplot? If so, describe it.
3. Recall we created a residuals scatterplot by hand in number 3 for the GAWK data and a residuals scatterplot using our graphing calculators in number 5 for the Starbucks data.
4. Which of the two data sets (the GAWK data in number 3 or the Starbucks data in number 5) appeared to be more linear and which appeared to be more curvilinear?
5. Compare and contrast the pattern or lack of pattern in the residual plots from number 3 and number 5 above.
6. So if there is a clear pattern in the residuals plot, would you say that a linear model would be a good fit for the data, i.e. would you say that a linear model would best capture patterns in the data? Explain.

# **MODULE 4**

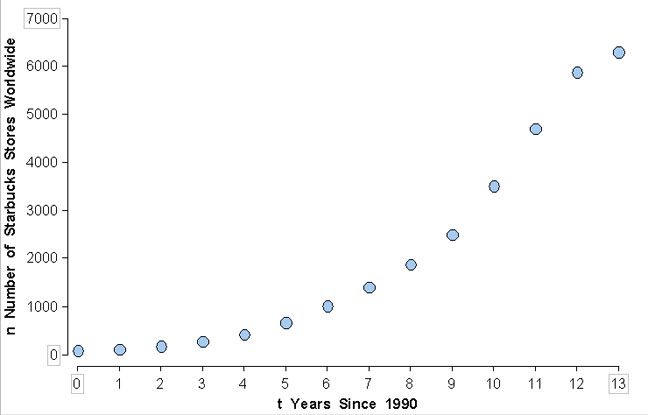
## **Non-linear Models**

## 4.1.1 Non-linear Models (Exponential Relationships)

Learning Objective

* Use an exponential model (when appropriate) to describe the relationship between two quantitative variables. Interpret the model in context.

1. Starbucks Revisited: The number of Starbucks stores worldwide has increased substantially since 1991. Let t represent the number of years since 1990 and n represent the number of Starbucks stores worldwide t years after 1990.



Which do you think would be the best fit for the Starbucks scatterplot - a linear or curvilinear regression model?

1. Recall the Starbucks data.

Which of the following do you think is the more reasonable statement? Explain your choice and give examples to support your conclusion.

1. The number of Starbucks stores is approximately doubling each year.
2. The number of Starbucks stores is increasing by roughly a factor of 1.5 each year (i.e. 1.5 times the number of Starbucks stores from the previous year).
3. The number of Starbucks stores is increasing by roughly a factor of 1.4 each year (i.e. 1.4 times the number of Starbucks stores from the previous year).

|  |  |
| --- | --- |
| **Year Since 1990** | **Number of Starbucks Stores Worldwide** |
| 0 | 84 |
| 1 | 116 |
| 2 | 165 |
| 3 | 272 |
| 4 | 425 |
| 5 | 676 |
| 6 | 1015 |
| 7 | 1412 |
| 8 | 1886 |
| 9 | 2498 |
| 10 | 3501 |
| 11 | 4709 |
| 12 | 5886 |
| 13 | 6294 |

1. Now use your graphing calculator to find the exponential regression model for the Starbucks data and note the value. Record the value below along with your exponential regression equation.

Exponential regression equation:

1. How is the base in your exponential model related to your choice of a, b, or c in number 2 above? Does it make sense to call the base the **growth factor** (explain why or why not)?
2. Use your model to find the number of Starbucks stores in 1990, (i.e. year 0). How is this result related to your model? Does it make sense to call this the **initial value** (explain why or why not)?
3. Quick sidetrack … given the exponential equation , identify the base multiplier and the initial value and explain what each represents.

Growth Factor =

Initial Value =

1. Rewrite your exponential model for the Starbucks data in the space below, and then use your exponential model to answer the following questions (no need to show your work).

Exponential Model:

1. How many Starbucks stores were there in 2004? Does this seem reasonable? Why or why not?
2. How many Starbucks stores were there in 2008? Does this seem reasonable? Why or why not?
3. How many Starbucks stores will there be in 2015? Does this seem reasonable? Why or why not?
4. The $10,000 question … and some of us may not be able to answer it yet (we may have to wait until 4.1.2 or 4.1.3 activity to answer this). When will there be a Starbucks store for every man, woman, and child in the United States (assume a US population of 300,000,000)?
5. Re-write your linear and exponential models for the number of Starbucks stores (and their respective values) below and then use the models to fill in the following table. Just looking at the table, which appears to be the better predictor?

Linear Model: Exponential Model:

|  |  |  |  |
| --- | --- | --- | --- |
| **Year Since 1990** | **Number of Starbucks Stores Worldwide** | **Linear Prediction** | **Exponential Prediction** |
| 0 | 84 |  |  |
| 1 | 116 |  |  |
| 2 | 165 |  |  |
| 3 | 272 |  |  |
| 4 | 425 |  |  |
| 5 | 676 |  |  |
| 6 | 1015 |  |  |
| 7 | 1412 |  |  |
| 8 | 1886 |  |  |
| 9 | 2498 |  |  |
| 10 | 3501 |  |  |
| 11 | 4709 |  |  |
| 12 | 5886 |  |  |
| 13 | 6294 |  |  |

1. Now graph your linear regression model, your exponential regression model, and your scatterplot for the Starbucks data in the same graphing window (if you need help choosing a graphing window, be sure to ask). Draw a sketch of your graphing window in the space below. Which appears to be a better fit, the linear or the exponential model? How is this reflected in the residual plot you sketched in number 5 of 3.3.2 activity?
2. Which had the higher value for the Starbucks data, the linear or the exponential model? How is this reflected in the residual plot you sketched in number 5 of 3.3.2 activity?
3. To determine some of the basic exponential shapes we might find in scatterplots, draw some basic “exponential” curves below (keep in mind we can flip the basic exponential graph about the x and y axes, and we can shift it up, down, left, or right). You should draw at least four.

## 4.1.2 Exponential Equations – Part 1

Learning Objectives

* Know the meaning of an exponential equation.
* Know the graphical significance of  and  for an equation of the form .
* Recognize the graph of an exponential equation.

**Exponential Equation**

Definition: The equation, where , , and  is called an exponential equation.

Hmmm … why do you suppose ?

Determine which of the following are exponential equations. If the equation is an exponential equation, identify  and .

1. 
2. 
3. 

**Graphs of Exponential Equations**

1. Graph  in your graphing calculator (be sure to select an appropriate graphing window so the important features of the graph are displayed and the graph is not scrunched up in one little area on your screen).

|  |  |
| --- | --- |
|  | 1. Use the space to the left to draw a quick sketch of your graphing window (be sure to indicate x-min, x-max, y-min, and y-max for your graphing window). 2. Find the *y*-intercept and label it on your sketch above. |

1. How is the y-intercept in the graph related to the equation, ?
2. Fill in the table and then answer the following question: What pattern do you notice in the output values (i.e. the y values)? Hint: after you fill in the table, try exploring the ratio of successive outputs (if you have no idea what I'm talking about, be sure to ask me). How is the "pattern" related to the equation that you graphed?

|  |  |
| --- | --- |
| ***x*** |  |
|
| – 1 |  |
| 0 |  |
| 1 |  |
| 2 |  |
| 3 |  |

1. For , use your graphing calculator to make a table of values AND to make a graph of the equation. Provide a quick sketch of your graphing window and label the *y*-intercept.
2. Identify *a* and *b*.
3. Identify the y-intercept.
4. Identify the ratio of successive outputs.

|  |  |
| --- | --- |
| ***x*** | ***y*** |
| – 2 |  |
| – 1 |  |
| 0 |  |
| 1 |  |
| 2 |  |
| 3 |  |

1. For, use your graphing calculator to make a table of values AND to make a graph of the equation. Provide a quick sketch of your graphing window and label the *y*-intercept.
2. Identify *a* and *b*.
3. Identify the y-intercept.
4. Identify the ratio of successive outputs.

|  |  |
| --- | --- |
| ***x*** | ***y*** |
| – 2 |  |
| – 1 |  |
| 0 |  |
| 1 |  |
| 2 |  |
| 3 |  |

1. To explore each of the following properties, you'll need to refer to your work on the previous pages.

**Growth Factor**

Look back at your work on the previous page, how is the ratio of successive outputs related to the base ? In your own words state what the growth factor of an exponential equation is and how it works.

**The *y*-Intercept of an Exponential Equation**

Look back at your work on the previous page, how is the *y*-intercept related to ?

Complete the following statement: For an exponential equation of the form , the *y*-intercept is \_\_\_\_\_ , and the growth factor is \_\_\_\_\_\_.

1. Set up an equation to solve, and then use your graphing calculator to solve it (if you don't know how, be sure to ask). To show your work, write down the equation you are solving, and provide a quick sketch of your graphing window. If necessary, round your results to six decimal places. Be sure to "check" your work. If you would prefer to solve the equation by hand, you can do so, but be sure to show each step.
2. If find x when y = 2 . b) If find x when y = 10.3923.

## 5.1.2 Exponential Equations – Part 2

Learning Objectives

* Use the growth factor and the y-intercept to find an exponential equation.
* Given an exponential equation identify the annual percent increase.
* Given the annual percent increase find the base of an exponential equation.

**Using the Growth Factor and y-Intercept to Find an Exponential Equation**

Let  represent time in the following exponential equation: .

What is the -intercept and what does it mean?

What is the base  and what does it mean?

Now make an input output table for, and actually calculate the ratio of successive outputs. Is it reasonable to claim that each output is obtained by multiplying the previous output by the base ?

|  |  |
| --- | --- |
|  |  |
| –2 |  |
| –1 |  |
| 0 |  |
| 1 |  |
| 2 |  |

So … inquiring minds want to know … if I told you that we have a NEW exponential equation and the **initial value** (the value of the output when the input is zero) of our new exponential equation is 4 and that each output is obtained by multiplying the previous output by , could you quickly find an exponential equation to model this situation? Do it.

1. The y-intercept of an exponential equation is 0.75 and each output is obtained by multiplying the previous output by 4.2. Find the exponential equation.
2. In a laboratory experiment a colony of 100 bacteria is established and the growth rate of the colony is monitored. The experimenters discover that the colony’s population triples every day. Let  represent the number of bacteria  days after the colony was established. Can this situation be modeled by an exponential equation? If so, do it. If not, explain why?
3. Suppose a flu epidemic has broken out at your school. Assume that on February 10th (day zero), a total of 20 people come down with the flu and that each day thereafter the total number of people who have gotten the flu doubles. Let  represent the total number of people who have gotten the flu at  days after February 10th.
4. Find an exponential equation to model the flu epidemic (be sure to identify your variables).
5. How many people will have the flu after 5 days?
6. How long will it take for 2000 people to get sick?
7. How many people will have the flu after 12 days? Does this make sense? Why or why not?
8. Each output of an exponential equation is obtained by multiplying the previous output by 1.42, and the graph of the exponential equation passes through the point (0, 97.56). Find the exponential equation.

**Exponential Equations and Percent Increase**

1. The exponential equation models the number of Starbucks stores worldwide years after 1990. Fill in the input/output table and then answer the following questions.

|  |  |  |
| --- | --- | --- |
| Years since 1990 | # Starbucks stores | 1. Explain how to find the percent increase in the number of Starbucks stores from year to year. 2. What is the percent increase in the number of Starbucks stores from 1990 to 1991 (calculate by hand and show your work). |
| 0 |  |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |
| 8 |  |
| 9 |  |
| 10 |  |

1. What is the percent increase in the number of Starbucks stores from 1999 to 2000 (calculate by hand and show your work)?
2. Hmmm … what is the percent increase in the number of Starbucks stores for any two consecutive years?
3. What is the base of the exponential model?
4. How is the percent increase in the number of Starbucks stores related to the base of the exponential model? Explain why this is true.
5. Each year after 2008, a company's annual sales were 1.082 times the previous year's sales. The initial sales in 2008 were $47,000.
6. Find an exponential model for the company's sales (be sure to identify your variables).
7. What is the annual percent increase in sales for the company? How do you know?
8. The following table gives the number of Kohl's stores for various years.

|  |  |  |
| --- | --- | --- |
| **Year** | **# Kohl's Stores** | 1. Let represent the number of Kohl's stores years since 1990. Make a scatterplot of these data (be sure to choose an appropriate graphing window). Would it be best to model the data using linear regression or exponential regression? How do you know? |
| 1992 | 79 |
| 1996 | 150 |
| 1999 | 259 |
| 2002 | 457 |
| 2005 | 675 |
| 2008  **Source*:*** *Intermediate Algebra: Functions & Authentic Applications, 4th Ed,* by Jay Lehman (#38 p. 217). | 1004 |

1. Based on your response to part a, find a regression model for these data.
2. What is the annual percent increase in the number of Kohl's stores? How do you know?
3. The percentage of households that have cell phones but no landlines is modeled by the exponential regression equation where represents the number of years since 2000 and represents the percentage of cell-phone only households. **Source*:*** *Intermediate Algebra: Functions & Authentic Applications, 4th Ed,* by Jay Lehman (#39 p. 217).
4. What percent of households were cell-phone only in 2009?
5. When will the percentage of cell-phone only households exceed 50%? (You may need your graphing calculator to answer this one).
6. What is the annual percent increase in the percentage of cell-phone only households? How do you know?
7. From 1960 to 2008, the federal debt increased by approximately 8.21% each year. The federal debt was $232.21 billion in 1960. Let represent the federal debt in billions of dollars years since 1960. Find an exponential equation for the annual federal debt since 1960.
8. A person invested $7000 in the stock of a start-up bio-medical company in San Diego. In its first year after going public, the FDA approved human trials for the company’s cloning technique for skin grafting of burn victims (including cloning techniques for ear and nose replacement etc). The annual percent increase of the stock’s value was 257%. Let represent the value of the stock years after the initial investment. If the company’s stock value were to continue to increase at this rate, find an exponential equation for the value of the stock.

# **MODULE 5**

# **Relationships in Categorical Data**

# **with an Intro to Probability**

## 5.1 Relationships in Categorical Data with an Intro to Probability

Learning Objectives

* Use a two-way table to analyze the association between two categorical variables.
* Given a two-way table, interpret (in context) joint, marginal, and conditional probability.
* Create a hypothetical two-way table to answer complex probability questions.

1. What is the difference between categorical data and quantitative data? Give examples of each.
2. Use the data from our class survey to fill in the table.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Eye Color** | | | | | |
| **Hair Color** |  | | Blue | Green | Brown | Black | **Total** | |
| Blonde | |  |  |  |  |  | |
| Red | |  |  |  |  |  | |
| Brown | |  |  |  |  |  | |
| Black | |  |  |  |  |  | |
| **Total** | |  |  |  |  |  | |

1. How many students in our class have brown hair and green eyes?
2. How many have black hair and brown eyes?
3. How many students in our class have brown hair?
4. How many students in our class have brown eyes?

*Continued next page …*

1. What do the values in the bottom row represent (excluding the lower right-hand corner)?
2. What do the values in the right hand column represent (excluding the lower right-hand corner)?

*By the way … for any two-way table representing two categorical variables the totals in the bottom row and right-hand column (excluding the lower right-hand corner) are the* ***marginal distributions****.*

1. What does the value in the bottom right-hand corner represent? How did you get this number? Did you add the values in the right-hand column, or did you add the values along the bottom row? Would either method always work? Why or why not?
2. What fraction of students in our class have brown hair and brown eyes? What fraction of students in our class have green eyes?
3. Suppose the number of students in our class with blonde hair was the same, but the total number of students in the class was actually 500, would reporting the number of students with blonde hair still mean the same thing (in other words … would reporting the number of blondes out of 500 students mean the same thing as reporting the SAME number of blondes out of approximately 42 students)? Why or why not? How could you make the difference evident even though the number of blondes is the same in both situations?
4. A few years ago, Myra Snell (the inspiration behind Cuyamaca’s “Stats Academy”) gave her PreStats students a quiz containing a question from a national statistics exam. Twenty-nine students took the quiz. She compared their performance on a particular question to the performance of the 1470 students included in the national sample. 24 of her students got the item right compared to 1001 of the national sample. Did Myra’s PreStats students do as well as the national sample? Use math to support your conclusion.
5. This data comes from a study of the factors that impact birth weight. Here the variable *Visit Doctor* indicates whether a woman visited a physician during the first trimester of her pregnancy. The variable *Low Weight* indicates whether a baby was born weighing under 2500 grams.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | | **Low Weight** | |  | |
|  |  | **No** | | **Yes** | | **Row Totals** | |
| **Yes** | 66 | | 23 | |  | |
| **No** | 64 | | 36 | |  | |
| **Column Totals** |  | |  | |  | |

1. Complete the table to find the marginal distributions, and then fill in the bottom right-hand corner.
2. Does visiting a doctor during the early stages of pregnancy seem to be associated with a lower incidence of low weight births? Identify the explanatory and response variables, and be sure to use math to support your conclusion and show your work.
3. If a woman does not visit the doctor during the first trimester of her pregnancy, how many times more likely is it that she will have a low weight baby?
4. This table is based on records of accidents compiled by a State Highway Safety and Motor Vehicles Office (the marginal distributions and the lower right-hand corner have been filled in for you). Are people less likely to have a fatal accident if they are wearing a seatbelt? Be sure to clearly identify the explanatory and response variables, use math to support your conclusion, and show your work.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Injury** | |  |
|  |  | **Nonfatal Injury** | **Fatal Injury** | **Row Total** |
| **Seat belt** | 412,368 | 510 | 412,878 |
| **No seat belt** | 162,527 | 1,601 | 164,128 |
| **Column Total** | 574,895 | 2,111 | 577,006 |

1. A study in Sweden looked at the impact of playing soccer on the incidence of arthritis of the hip or knee. They gathered information on former elite soccer players, people who played soccer but not at the elite level, and those who never played soccer. Fill in the marginal distributions and the lower right-hand corner. Does this study suggest that playing soccer makes someone more likely to have arthritis of the hip or knee? Be sure to clearly identify the explanatory and response variables, use math to support your conclusion, and show your work.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **Soccer Player** | | |  |
|  |  | **Elite** | **Non-elite** | **Did not play** | **Column Totals** | |
| **Arthritis** | 10 | 9 | 24 |  | |
| **No arthritis** | 61 | 206 | 548 |  | |
| **Row Totals** |  |  |  |  | |

1. Four hundred seventy-eight students in grades 4 through 6 in selected schools in Michigan, were asked the following question.

*Which of the following would make you popular among your friends? Rank in order.*

* *Making good grades*
* *Having lots of money*
* *Being good at sports*
* *Being handsome or pretty*

The table below lists the number of students (by gender) who gave the indicated factor a ranking of 1 (most important factor in making them popular amongst their friends).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Most Important Popularity Factor** | | | | | |
| **Gender** |  | | Grades | Money | Sports | Looks | **Row Totals** | |
| Girl | | 55 | 17 | 38 | 141 |  | |
| Boy | | 39 | 17 | 127 | 44 |  | |
| **Column Totals** | |  |  |  |  |  | |

1. What proportion of the total number of students considers looks to be the most important factor in making them popular amongst their friends?
2. If we were to randomly select a student, what is the probability that the student would think Money is the most important factor in making them popular amongst their friends?

*To calculate the proportion in part a) we used two numbers in the margin that relate to just one of the categorical variables (Student Response). This is therefore called a****marginal proportion****. Also since we only used numbers from the margin, the probability we found in part b) is called a* ***marginal probability****.*

1. However, we’re not really interested in the marginal probabilities or marginal proportions. We’re interested in answering the question, “Are boys more likely to think that money is the most important factor in making them popular amongst their friends?” So we need to use **conditional proportions** (proportions that are “conditioned” on the explanatory variable) to answer this question … just like we did in numbers 4, 5, & 6 above (only we didn’t know we were calculating conditional proportions when we were working on those problems).
2. This data is from a 5-year experiment with physicians between the ages of 40 and 84, published in 1988 by the Steering Committee of the Physicians Health Study Research Group. The physicians participating in the study were randomly selected to receive an aspirin or a placebo. The pills looked the same and the physicians did not know which they were taking.   
     
   Does this data suggest that taking aspirin reduces the risk of heart attack? Use probabilities (rather than percents) to answer the question and indicate whether the probabilities are marginal or conditional probabilities. If they are conditional probabilities, indicate which variable the probabilities were conditioned on. **Before moving on, be sure to show your answers to this problem to the instructor**.

|  |  |  |
| --- | --- | --- |
|  | **Heart attack** | **No heart attack** |
| **Aspirin** | 104 | 10,933 |
| **Placebo** | 189 | 10,845 |

1. **Popularity Revisited:** Four hundred seventy-eight students in grades 4 through 6 in selected schools in Michigan, were asked the following question.

*Which of the following would make you popular among your friends? Rank in order.*

* *Making good grades*
* *Having lots of money*
* *Being good at sports*
* *Being handsome or pretty*

The table below lists the number of students (by gender) who gave the indicated factor a ranking of 1 (most important factor in making them popular amongst their friends).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Most Important Popularity Factor** | | | | | |
| **Gender** |  | | Grades | Money | Sports | Looks | **Row Totals** | |
| Girl | | 55 | 17 | 38 | 141 |  | |
| Boy | | 39 | 17 | 127 | 44 |  | |
| **Column Totals** | |  |  |  |  |  | |

1. Suppose we randomly select a student. What is the probability that the student is a boy? Is this a marginal or conditional probability? How do you know?
2. What is the probability that a student ranked grades as the most important factor in making him/her popular amongst his/her friends? Write your response in a sentence and then write it using the appropriate probability notation (if you don’t know how to do this, be sure to ask). Is this a marginal or conditional probability? How do you know? **Show the instructor your answers to part b**.
3. What is the probability that a boy ranked sports as the most important factor in making him popular amongst his friends? Is this a marginal or conditional probability? How do you know? Write your response in a sentence and then write it using the appropriate probability notation. **Show the instructor your answers to part c**.
4. What is the probability that a girl ranked looks as the most important factor in making her popular amongst her friends? Is this a marginal or conditional probability? How do you know? Write your response in a sentence and then write it using the appropriate probability notation. **Show the instructor your answers to part d**.
5. **Joint Probability**: Suppose we randomly select a student. What is the probability that the student is a boy **and** the student ranked grades as the most important factor in making him popular amongst his friends? Write your response in a sentence and then write it using the appropriate probability notation. **Show the instructor your answers to part e**.
6. Find and **interpret** the following probabilities and then indicate whether you have found a marginal probability, a conditional probability, or a joint probability.

i) P(Money|Girl)

ii) P(Sports)

1. P(Girl and Sports)

# **MODULE 6**

# **Probability and**

# **Probability Distributions**

## 6.1.1 Introduction to Probability

Learning Objectives

* Interpret (in context) a probability as a long-run relative frequency of an event.
* Interpret a probability distribution as a description of long-run behavior of a random variable.
* Recognize features of a probability distribution.

1. The coin-toss experiment.
2. If we toss a fair coin (i.e. heads and tails are equally likely outcomes), what is the theoretical probability that the outcome will be heads?

P(Heads) =

1. Each person in your group should conduct the coin-toss experiment alone by tossing a coin (fairly so that heads and tails are equally likely outcomes) ten times and recording the outcomes in the table below. With each trial (i.e. each coin toss) record an H if the outcome is heads, and record a T if the outcome is tails.

We're interested in the event that the outcome is heads when we toss the coin. What is the fraction of outcomes that are heads using **your** 10 trials?

Your 10 trials =

Now find the fraction of outcomes that are heads using your group's total number of trials.

Group's trials

Now find the fraction of outcomes that are heads using the total number of trials conducted by the entire class.

Class trials

|  |  |
| --- | --- |
| **Toss** | **Outcome** |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |
| 8 |  |
| 9 |  |
| 10 |  |

1. As the number of trials increases what happens to the fraction of outcomes that are heads?
2. The Monty Hall Problem (Let's Make a Deal): Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a piggy-bank containing $250,000; behind the other two doors, goats. You get to choose a door and take home the prize behind that door. You pick a door (but the door is not opened), and Monty Hall, the game show host who knows what is behind each door, opens one of the other doors, revealing a goat. So now there are two doors still closed; the original door you chose and just one of the two you did not choose. Monty says to you, "Do you want to switch your choice to the remaining closed door?"
3. Should you switch doors? Before we answer this question, let’s look at the probability of winning.

When you make the original door selection out of three possible doors, what is the probability that the door you selected conceals the piggy-bank containing $250,000 and is therefore the winner? What is the probability that the door you selected conceals a goat and is therefore a loser?

P(selected door is a winner) = P(selected door is a loser loser) =

Sidetrack: Hmmm … interesting … what is the sum of these probabilities, and will this always happen when you consider the probability of all possible outcomes?

Now, suppose Monty opens one of the doors concealing a goat, so now there are only two closed doors; the one you chose and one you did not choose. What is the probability that you will win the money if you do NOT switch doors? What is the probability that you will win the money if you do switch doors?

P(win without switching doors) = P(win with switching doors) =

Sidetrack: Did it happen again? Do these probabilities add up to the same number?

1. We'll work in pairs (or groups of three if there are an odd number in your group) to conduct an experiment. One member of your group will be the game show host, Monty Hall, and the other group member(s) will be the contestant. We’ll play Let’s Make a Deal using the chart on the next page (the teacher will demonstrate). Monty will fill out the chart completely, and then your group will switch roles so that the new Monty can fill out the chart.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Door1.jpg | Door2.jpg | Door3.jpg |  |  |
| **Trial** | **Piggy-bank OR Goat** | **Piggy-bank OR Goat** | **Piggy-bank OR Goat** | **Switch Doors Yes (Y) No (N)** | **Win (W) Lose (L)** |
| 1 |  |  |  |  |  |
| 2 |  |  |  |  |  |
| 3 |  |  |  |  |  |
| 4 |  |  |  |  |  |
| 5 |  |  |  |  |  |
| 6 |  |  |  |  |  |
| 7 |  |  |  |  |  |
| 8 |  |  |  |  |  |
| 9 |  |  |  |  |  |
| 10 |  |  |  |  |  |
| 11 |  |  |  |  |  |
| 12 |  |  |  |  |  |
| 13 |  |  |  |  |  |
| 14 |  |  |  |  |  |
| 15 |  |  |  |  |  |
| 16 |  |  |  |  |  |
| 17 |  |  |  |  |  |
| 18 |  |  |  |  |  |
| 19 |  |  |  |  |  |
| 20 |  |  |  |  |  |

1. O.K. let’s take a look at how well the contestant did.

Considering only the trials where your contestant did NOT switch doors, what fraction of the outcomes were wins?

Considering only the trials where your contestant did switch doors, what fraction of the outcomes were wins?

=

1. Now let’s compute the fraction of outcomes that were wins using the trials conducted by your **entire group**.

CLASS TRIALS WIHTOUT SWITCHING:

CLASS TRIALS WITH SWITCHING:

1. Now let’s compute the fraction of outcomes that were wins using the trials conducted by the entire class.

CLASS TRIALS WIHTOUT SWITCHING:

CLASS TRIALS WITH SWITCHING:

1. Based on the results in part e), what do you think the empirical probability of winning is if you **do not** switch doors and the empirical probability of winning is if you **do** switch doors (i.e. the long run probabilities if we could play *Let’s Make a Deal* millions of times)?

P(win without switching doors) = P(win with switching doors) =

Sidetrack: Did it happen again? Do these probabilities add up to 1?

1. Do the results in part f) match your predictions in part a)? Explain why or why not (do not just flippantly write down an explanation here … really think about it as I may grade your group based on this answer).

## 6.1.2 Probability Rules

Learning Objective

* Reason from Probability distributions or use probability rules to answer probability questions.

The **sample space S** of a random phenomenon such as flipping a coin or selecting a door or rolling dice is the set of all possible outcomes. For example, if we flip one coin twice, one possible outcome is we get a heads on the first flip and a tails on the second flip or HT. Another possible outcome is that we get a tails on the first flip and a heads on the second flip or TH. So the sample space of flipping one coin twice is

S = {HH, HT, TH, TT}.

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. For example, the event that we get exactly one heads when we flip two coins in order is

E = {HT, TH}.

And the probability of an event E is

.

1. Suppose we flip a fair coin three times and record the outcome of the three tosses.
2. What is the sample space S for this experiment? How many outcomes are in S?
3. We’re interested in the event, E, that we roll heads exactly once. How many outcomes are there in E? List them.
4. Find P(E).
5. Again, suppose we flip a coin three times and record the outcome of the three tosses.
6. Find each of the following probabilities.

P(No heads) =

P(exactly one heads) =

P(exactly two heads) =

P(exactly three heads) =

Have we found the probability of every possible outcome in S? If not, list the outcome(s) in S for which we have not found the probability.

1. Find the sum of the probabilities you found in part a).

P(No heads) + P(exactly one heads) + P(exactly two heads) + P(exactly three heads) =

Hmmm … have we seen this before? Do you think this will always happen?

Write a **rule about the sum of the probabilities** of all the events in a sample space S.

1. Do you think the sum of the probabilities of events from a sample space S could ever be greater than 1?
2. We’re still tossing a coin three times and recording the outcome of the three tosses.
3. Find each of the following.

P(two heads OR three tails) =

P(two heads) + P(three tails) =

1. Find each of the following.

P(no heads OR no tails) =

P(no heads) + P(no tails) =

1. Based on your results from a) and b) above, write a **general rule about the probability of an “OR” statement**.

P(A or B) =

1. Do you think P(no tails OR at least one heads) = P(no tails) + P(at least one heads)? Verify your response by using the sample space and counting the outcomes in the event to find each of the following.

P(no tails OR at least one heads) =

P(no tails) + P(at least one heads) =

Hmmm … does the rule you wrote in part c) hold up? Why or why not?

1. If necessary, go back to part c) above and revise the rule you wrote in part c) so that it will always hold true and then apply it in the space below to P(no tails OR at least one heads).
2. Use the following table of probabilities to calculate each of the following.

|  |  |  |  |
| --- | --- | --- | --- |
| **Fatal Injury by Sex and Cause in the United States Rates per 100,000 Population**  But first … complete the table by filling in marginal distributions and the lower right-hand corner. | | | |
|  | Male | Female | Row Totals |
| Motor Vehicle | 34.4 | 12.0 |  |
| Fall | 6.4 | 5.4 |  |
| Suicide | 18.6 | 5.4 |  |
| Homicide | 17.0 | 4.4 |  |
| Other Injury | 29.1 | 10.6 |  |
| Column Totals |  |  |  |
| *Source: American Journal of Public Health 1989; 79:1396-1400* | | | |

1. Find and interpret P(Male) and P(Female). Round to two decimal places. If you need help with the interpretation … let me know.
2. What should we get if we add P(Male) + P(Female)? Why?
3. Find each of the following and then interpret the first three (round to two decimal places).

P(Male) =

P(Suicide) =

P(Male and Suicide) =

P(Male) P(Suicide) =

1. Find each of the following and then interpret the first three (round to two decimal places).

P(Female) =

P(Motor Vehicle) =

P(Female and Motor Vehicle) =

P(Female) P(Motor Vehicle) =

1. Write a **general rule to calculate P(A and B)**.
2. Find each of the following and then interpret the first three.

P(Female) =

P(Fall) =

P(Female and Fall) =

P(Female) P(Fall) =

Hmmm … does the general rule you wrote in part e) hold up? If not, modify the general rule for P(A and B) in part e) above.

1. Find and interpret P(Motor Vehicle | Male). Recall this is a conditional probability and you need to find the probability that the fatal injury was a motor vehicle accident given the randomly selected dead person is a male (i.e. find the probability of the fatal injury being a motor vehicle accident using only the males). Do you recall how to find conditional probabilities? If not you can review the 5.1 handout or ask me.
2. Compare the following probability pairs.

|  |  |
| --- | --- |
| P(Homicide | Male) =  P(Homicide) = | P(Other Injury | Female) =  P(Other Injury) = |

What can you say, so far, about P(B | A)? This is actually going to be the **definition of independence**, but don't worry about that right now.

1. Find and interpret each of the following.

P(Homicide | Female) =

P(Homicide) =

Do your results contradict your conclusion in part h) above?

As it turns out two events A and B are **independent** when P(B | A) = P(B) or they are almost equal (taking into consideration that all the probabilities may have been rounded etc). If two events are not independent (a.k.a. **dependent**), then P(B | A) will not be equal or almost equal to P(B) . So you should modify your conclusion in part h) to say, "P(B | A) = P(B) means that A and B are independent events."

1. Hmmm … let's see if we can write a general rule for P(A and B) without having to worry about whether or not the events are independent. Find each of the following and then try to write the general rule for P(A and B).

|  |  |
| --- | --- |
| P(Homicide | Male) =  P(Male) =  P(Male) P(Homicide | Male) =  P(Homicide and Male) = | P(Other Injury | Female) =  P(Female) =  P(Female) P(Other Injury | Female) =  P(Other Injury and Female) = |

P(A and B) =

1. Complete the following to summarize **the probability rules**. Go back through this activity if you need to.
2. For any event A, .
3. If *S* is a sample space, then P(*S*) = .
4. The sum of the probabilities of all possible disjoint events in a sample space is .
5. If A and B are disjoint events (no outcomes in common), then P(A or B) = .
6. If A and B are NOT disjoint events (share at least one outcome), then   
     
   P(A or B) = .
7. For any event A, P(A does not occur) = .
8. If P(B | A) = , then A and B are independent events.
9. When A and B are independent events then P(A and B) = .
10. Regardless of whether A and B are independent events, P(A and B) = .
11. Use part i) and division to write the rule P(B | A) = .

### 6.1.2 Practice Problems

Do these problems on a separate piece of paper. Label the assignment "6.1.2 Practice Problems". Be sure to show your work and/or support your reasoning. The correct result without a sufficient amount of correct and appropriate work is worth zero points. I may collect this assignment at the beginning of our next class. (Some problems were taken from *Essential Statistics* by David S. Moore, 2010 W. H. Freeman and Company, pp 213 - 216)

|  |  |  |  |
| --- | --- | --- | --- |
| **Fatal Injury Probabilities by Sex and Cause in the United States** | | | |
|  | Male | Female | Row Totals |
| Motor Vehicle | 34.4 | 12.0 | **46.4** |
| Fall | 6.4 | 5.4 | **11.8** |
| Suicide | 18.6 | 5.4 | **24.0** |
| Homicide | 17.0 | 4.4 | **21.4** |
| Other Injury | 29.1 | 10.6 | **39.7** |
| Column Totals | **105.5** | **37.8** | **143.3** |
| *Source: American Journal of Public Health 1989; 79:1396-1400* | | | |

1. Let A be the event that a victim of a fatal injury was a woman and B the event that the death was a suicide. Using the table above, a student found P(???) = 0.04. The actual probability notation for P(???) is
2. P(A and B). **b)** P(A | B). **c)** P(B | A).
3. Again, let A be the event that a victim of a fatal injury was a woman and B the event that the death was a suicide. The probability of B conditioned on A is expressed in probability notation as
4. P(A and B). **b)** P(A | B). **c)** P(B | A).
5. Again, let A be the event that a victim of a fatal injury was a woman and B the event that the death was a suicide. Find the probability of B conditioned on A (and express the result using the appropriate probability notation).
6. Of people who died in the United States in recent years, 86% were white, 12% were black, and 2% were Asian (this ignores a small number of deaths that were other races). Diabetes caused 2.8% of deaths among whites, 4.4% among blacks, and 3.5% among Asians. Find the probability that a randomly chosen death is a white who died of diabetes.
7. New York State’s “Quick draw” lottery moves right along. Players choose between one and ten numbers from the range 1 to 80; 20 winning numbers are displayed on a screen every four minutes. If you choose just one number, your probability of winning is 20/80, or 0.25. Lester plays one number 8 times in a row as he sits in a bar. What is the probability that all 8 bets lose?
8. People with type O-negative blood are universal donors (any patient can receive a transfusion of O-negative blood). Only 7.2% of the American population has O-negative blood. If 10 people appear at random to give blood, what is the probability that at least 1 of them is a universal donor?
9. Before electronics took over, slot machines worked like this. You pull the lever to spin three wheels. Each wheel has 20 symbols, all equally likely to show when the wheel stops spinning, and the three wheels are independent of each other. Suppose that the middle wheel has 9 cherries among its 20 symbols, and the left and right wheels have 1 cherry each.
10. You win the jackpot if all three wheels show cherries. What is the probability of winning the jackpot (a tree diagram may help you calculate the probability … if you don’t know how to use one … ask).
11. There are three ways that the three wheels can show two cherries and one symbol other than a cherry. Find the probability of each of these ways.
12. What is the probability that the wheels stop with exactly two cherries showing among them?
13. A striking trend in higher education is that more women than men reach each level of attainment. Here are the counts (in thousands) of earned degrees in the United States in the 2010-2011 academic year (classified by level and gender).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Bachelor’s** | **Master’s** | **Professional** | **Doctorate** | **Total** |
| **Female** | 986 | 411 | 52 | 32 | 1481 |
| **Male** | 693 | 260 | 45 | 27 | 1025 |
| **Total** | 1679 | 671 | 97 | 59 | 2506 |

1. If you choose a degree recipient at random, what is the probability that the person you choose is a woman?
2. What is the conditional probability that you choose a woman, given that the person chosen received a doctorate?
3. Are the events “choose a woman” and “choose a doctoral degree recipient” independent? How do you know?
4. What is the probability that a randomly chosen degree recipient is a woman and the degree is a doctorate?
5. Use the College-Degree data from the previous problem to answer the following questions.
6. What is the probability that a randomly chosen degree recipient is a man?
7. What is the conditional probability that the person chosen received a bachelor’s degree, given that he is a man?
8. What is the probability that a randomly chosen degree recipient is a man and the degree conferred is a doctorate?

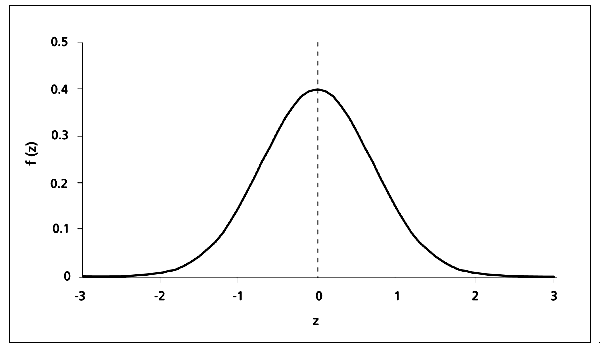
6.1.3 will cover *Introduction to Random Variables*, and then we’ll be done with Topic 6.1, but you should start working through the material in Topic 6.1 on OLI as soon as possible as it is a long section.

## Mod 6 Probability Distributions Warm Up – Area Under the Normal Curve

Learning Objectives

* Graph a normal curve and summarize its important properties.
* Apply the empirical rule (the 68-95-99.7 rule) to solve simple probability problems.

1. The scores on a statistics exam were approximately normal with mean 78 and standard   
   deviation 7. Below is a generic Normal Distribution. Fill in the numbers on the x-axis to make the distribution represent the statistics exam scores.



1. Now mark the middle 68%, 95% and 99.7% of the graph.
2. What percent of the students scored between a 71 and an 85?
3. What was the range of scores for the lowest 16% of the class?
4. What percent of the class scored higher than 85?
5. The length of human pregnancies from conception to birth varies according to a distribution that is approximately normal with a mean of 266 days and a standard deviation of 16 days.
6. Draw the density curve for the normal distribution of pregnancy lengths. Mark and label the location of the mean and one, two, and three standard deviations above and below the mean.
7. What percent of the pregnancies last between 250 and 282 days?
8. What are the upper and lower bounds for the middle 95% of all pregnancy lengths?
9. What percent of the pregnancies last between 218 and 314 days?
10. What percent of the pregnancies last 266 days or fewer?
11. What percent of the pregnancies last between 266 and 282 days?
12. What percent of the pregnancies last fewer than 250 days?
13. What percent of the pregnancies last between 234 and 314 days?
14. How long are the longest 2.5% of pregnancies?

## 6.1.4 & 6.2 Probability Distributions for Random Variables

Learning Objectives

* Distinguish between discrete random variables and continuous random variables.
* Use probability distributions for discrete and continuous random variables to estimate probabilities and identify usual events.

**Random Variables**

1. Suppose we conduct an experiment by flipping a fair coin three times.
2. Are the outcomes “heads” and “tails” values of a categorical variable or a quantitative variable? Explain.
3. Would you say that the outcome “the number of heads in three tosses” is a value of a categorical variable or a quantitative variable? Explain.
4. A **random variable** is a numerical measure of the outcome of an experiment, i.e. a quantitative variable. Random variables are typically denoted with capital letters near the end of the alphabet such as *X* or *Y*.
5. Suppose we toss a coin three times. Is the variable *X* = “the number of heads” a random variable? Explain.
6. Still tossing a coin three times … list all possible values for the random variable *X* = “the number of heads”. Would you say that the set of all possible values of *X* is finite or infinite?
7. Suppose we roll a pair of dice. Describe a random variable for this experiment.
8. Suppose we roll a pair of fair dice (one red die and one green die) and let *Y* represent the sum of the dots on the faces of the dice.
9. List all possible outcomes in the table below.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Red** | **Green** | **Y = Sum** |  | **Red** | **Green** | **Y = Sum** |  | **Red** | **Green** | **Y = Sum** |
|  |  |  |  |  |  |  |  |  |  |  |

1. Use the table below to list only the distinct values of the random variable *Y* and the probability distribution for *Y*, i.e. the probability of each value of *Y* (rounded to four decimal places). Then make a histogram for the distribution of the random variable *Y* and write the probability P(*Y*) at the top of each rectangle in your histogram.

1 2 3 4 5 6 7 8 9 10 11 12 13

P(*Y*)

*Y*

.17.16.15

.14

.13

.12

.11

.10

.09

.08

.07

.06

.05

.04

.03

.02

.01

Probability.jpg

|  |  |
| --- | --- |
| *Y* | P(*Y*) |
|  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Hmmm … would you say that the set of all possible values of the random variable *Y* is finite or infinite? |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

1. Now calculate the area of each rectangle in the histogram you created in part b). **Round to four decimal places and write the area inside of the rectangle**. Then add the areas of the rectangles to get the total area of the histogram. What is the total area of the histogram? Does this make sense? Why or why not?
2. Again, suppose that we roll a pair of fair dice. Use the areas in the histogram to find the following probabilities.

***Note:*** *In problem number 3, the height of our rectangles also represented the probability, but we were only able to set up the probability histogram this way because the width of each rectangle was one. So in this particular case the height was equal to the area of the rectangle. It must ALWAYS be the case that the AREA of each rectangle represents the probability, so it may not always be the case that the height (or the y-axis) will represent the probability.*

**Discrete vs Continuous Random Variables**

1. How many numbers do you think are in the interval 0 to 1 or the interval 5 to 8?
2. A random variable is called **discrete** if it can assume only a finite number of values OR if it can assume infinitely many values that increase in discrete steps (for example 4, 7, 10, 13, … goes on forever but makes “discrete” jumps of 3 each time). A random variable is called **continuous** if it can assume an infinite number of values over an interval (no discrete jumps). For each of the following, determine if it is a **discrete random variable** or a **continuous random variable**.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Experiment** |  | **Random Variable *X*** | **Discrete or Continuous?** |
| a) | Flip a coin three times | *X* = | total number of heads |  |
| b) | Randomly select a student who took a true/false test with 100 questions | *X* = | number of questions answered correctly |  |
| c) | Randomly select a mutual fund | *X* = | the number of companies in the portfolio |  |
| d) | Randomly select 50 students | *X* = | the exact average weight of the group in pounds |  |
| e) | Randomly select a woman | *X* = | the exact height of the woman in inches |  |

1. England's Department of Health tracks the wait times for the 1st hospital outpatient appointment following a general practitioner referral. For the month ending March 31st, 2010, data was collected from 974,683 patients across the country. (For our purposes we are counting a wait time from 0 up to 2 weeks as 2 weeks, and a wait time from 2 weeks up to 4 weeks as 4 weeks, and a wait time from 4 weeks up to 6 weeks as 6 weeks, and so on).

**Commissioner Based Hospital Waiting Times for 1st Outpatient Appointment**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Weeks** | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18+ |
| **# Patients** | 388,609 | 283,225 | 171,913 | 86,461 | 33,947 | 9,428 | 922 | 71 | 107 |

*Source: England's Department of Health*

<http://www.dh.gov.uk/en/Publicationsandstatistics/Statistics/Performancedataandstatistics/HospitalWaitingTimesandListStatistics/index.htm>

1. What is the probability that a patient will have to wait 2 weeks for a first outpatient appointment?
2. What is the probability that a patient will have to wait 8 weeks for a first outpatient appointment?
3. Let *X* represent the wait time (in weeks) after a general practitioner's referral that a patient must wait for the first outpatient appointment. Fill in the following table (round to 4 decimal places) and then use it to make a probability histogram for this data (center each rectangle on the given value of *X* and label each rectangle with its area). ***Warning***: the **area** of each rectangle in the histogram must be equal to the probability that a patient will wait that many weeks for an appointment.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***X*** | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
| **P(*X*)** |  |  |  |  |  |  |  |  |  |

Hmmm … given the way we have set up *X,* do   
 you think *X* is a   
 **discrete** or **continuous**   
 random variable?  
Circle one.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

1. Calculate the total area of your probability histogram, and show your work. Note: if the total area is not equal to one, should we be concerned? Explain.
2. Use the areas in your probability histogram to determine the probability that a patient will have to wait less than 8 weeks for first outpatient appointment.
3. O.K. let's find the mean wait time in two ways. First find the mean wait time using the table on the previous page. Then find the mean wait time using the probabilities from the table in part c). Show your work for each method. The two "means" should match … do they?
4. Based on your work in part e) above, write a general formula the ***mean of a discrete random variable*** using probabilities.

1. The formula for the ***standard deviation, , of a discrete random variable*** is given below (you'll explore why we use this formula on OLI). Use the formula to find the standard deviation for *X* (the number of weeks after a general practitioner's referral that a patient must wait for the first outpatient appointment).

1. The height *X* (in inches) of a randomly selected woman has the N(64, 2.7) distribution (hmmm … if you don’t know what this notation means … be sure to ask).
2. Draw the probability density curve for the normal distribution of heights of women (if you don't know what values to place on each axis or how to draw a Normal density curve, be sure to ask).

Do you think the   
 height of a randomly   
 selected woman is a   
 **discrete** or **continuous**   
 random variable?  
Circle one.

1. What do you think the area under the density curve you drew in part a) is? Explain.
2. What is the probability that a randomly selected woman will be 64 inches or taller? Be sure to write your result using probability notation.
3. What is the probability that a randomly selected woman will be between 61.3 and 66.7 inches tall? Be sure to write your result using probability notation.
4. What is the probability that a randomly selected woman will be between 58.6 and 72.1 inches tall? Be sure to write your result using probability notation.
5. What is the probability that a randomly selected woman will be between 68 and 72 inches (huh)?
6. **WE’LL DO THIS ONE AS A CLASS.** To determine the probability that a randomly selected woman will be between 68 and 72 inches, we need to learn about **the standard normal distribution** and z-scores. If *X* is an observation from a distribution that has mean and standard deviation , the **standardized value** of *X* is .
7. Use the - score formula to standardize the heights of women in the table below.

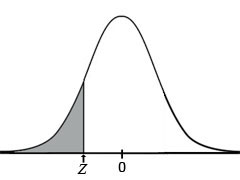
What is the standardized mean height of women, and what is the standardized standard deviation for the heights of women?

Draw the density curve for the **standard normal distribution** N(0, 1), i.e. in this case, the standardized density curve for the heights of women.

|  |  |
| --- | --- |
| ***X*:  Heights of Women (inches)** | ***Z*: Standardized Heights of Women** |
| 55.9 |  |
| 58.6 |  |
| 61.3 |  |
| 64 |  |
| 66.7 |  |
| 69.4 |  |
| 72.1 |  |

1. Now we can find the probability that a randomly selected woman is between 68 and 72 inches tall by using the Z-table on the next page to calculate the area under the standard normal curve corresponding to these heights.

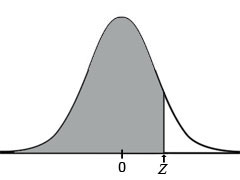
|  |  |
| --- | --- |
| i) | First standardize, i.e. find the z-scores corresponding to the heights 68 and 72. |
| ii) | Next draw the density curve for the **standard normal distribution** N(0, 1), locate the appropriate z-scores on the horizontal axis, and shade in the area corresponding to the probability you're trying to find (**for this problem you can use the curve you drew above**). |
| iii) | Now use the Z-table and any necessary arithmetic to find the area of the region you shaded. |
| iv) | So what is ? |



**Z-Table (area to the left of Z)**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **0** | **0.01** | **0.02** | **0.03** | **0.04** | **0.05** | **0.06** | **0.07** | **0.08** | **0.09** |
| **-3.4** | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0002 |
| **-3.3** | .0005 | .0005 | .0005 | .0004 | .0004 | .0004 | .0004 | .0004 | .0004 | .0003 |
| **-3.2** | .0007 | .0007 | .0006 | .0006 | .0006 | .0006 | .0006 | .0005 | .0005 | .0005 |
| **-3.1** | .0010 | .0009 | .0009 | .0009 | .0008 | .0008 | .0008 | .0008 | .0007 | .0007 |
| **-3.0** | .0013 | .0013 | .0013 | .0012 | .0012 | .0011 | .0011 | .0011 | .0010 | .0010 |
| **-2.9** | .0019 | .0018 | .0018 | .0017 | .0016 | .0016 | .0015 | .0015 | .0014 | .0014 |
| **-2.8** | .0026 | .0025 | .0024 | .0023 | .0023 | .0022 | .0021 | .0021 | .0020 | .0019 |
| **-2.7** | .0035 | .0034 | .0033 | .0032 | .0031 | .0030 | .0029 | .0028 | .0027 | .0026 |
| **-2.6** | .0047 | .0045 | .0044 | .0043 | .0041 | .0040 | .0039 | .0038 | .0037 | .0036 |
| **-2.5** | .0062 | .0060 | .0059 | .0057 | .0055 | .0054 | .0052 | .0051 | .0049 | .0048 |
| **-2.4** | .0082 | .0080 | .0078 | .0075 | .0073 | .0071 | .0069 | .0068 | .0066 | .0064 |
| **-2.3** | .0107 | .0104 | .0102 | .0099 | .0096 | .0094 | .0091 | .0089 | .0087 | .0084 |
| **-2.2** | .0139 | .0136 | .0132 | .0129 | .0125 | .0122 | .0119 | .0116 | .0113 | .0110 |
| **-2.1** | .0179 | .0174 | 0.170 | .0166 | .0162 | .0158 | .0154 | .0150 | .0146 | .0143 |
| **-2.0** | .0228 | .0222 | .0217 | .0212 | .0207 | .0202 | .0197 | .0192 | .0188 | .0183 |
| **-1.9** | .0287 | .0281 | .0274 | .0268 | .0262 | .0256 | .0250 | .0244 | .0239 | .0233 |
| **-1.8** | .0359 | .0351 | .0344 | .0336 | .0329 | .0322 | .0314 | .0307 | .0301 | .0294 |
| **-1.7** | .0446 | .0436 | .0427 | .0418 | .0409 | .0401 | .0392 | .0384 | .0375 | .0367 |
| **-1.6** | .0548 | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 | .0475 | .0465 | .0455 |
| **-1.5** | .0668 | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 | .0582 | .0571 | .0559 |
| **-1.4** | .0808 | .0793 | .0778 | .0764 | .0749 | .0735 | .0721 | .0708 | .0694 | .0681 |
| **-1.3** | .0968 | .0951 | .0934 | .0918 | .0901 | .0885 | .0869 | .0853 | .0838 | .0823 |
| **-1.2** | .1151 | .1131 | .1112 | .1093 | .1075 | .1056 | .1038 | .1020 | .1003 | .0985 |
| **-1.1** | .1357 | .1335 | .1314 | .1292 | .1271 | .1251 | .1230 | .1210 | .1190 | .1170 |
| **-1.0** | .1587 | .1562 | .1539 | .1515 | .1492 | .1469 | .1446 | .1423 | .1401 | .1379 |
| **-0.9** | .1841 | .1814 | .1788 | .1762 | .1736 | .1711 | .1685 | .1660 | .1635 | .1611 |
| **-0.8** | .2119 | .2090 | .2061 | .2033 | .2005 | .1977 | .1949 | .1922 | .1894 | .1867 |
| **-0.7** | .2420 | .2389 | .2358 | .2327 | .2296 | .2266 | .2236 | .2206 | .2177 | .2148 |
| **-0.6** | .2743 | .2709 | .2676 | .2643 | .2611 | .2578 | .2546 | .2514 | .2483 | .2451 |
| **-0.5** | .3085 | .3050 | .3015 | .2981 | .2946 | .2912 | .2877 | .2843 | .2810 | .2776 |
| **-0.4** | .3446 | .3409 | .3372 | .3336 | .3300 | .3264 | .3228 | .3192 | .3156 | .3121 |
| **-0.3** | .3821 | .3783 | .3745 | .3707 | .3669 | .3632 | .3594 | .3557 | .3520 | .3483 |
| **-0.2** | .4207 | .4168 | .4129 | .4090 | .4052 | .4013 | .3974 | .3936 | .3897 | .3859 |
| **-0.1** | .4602 | .4562 | .4522 | .4483 | .4443 | .4404 | .4364 | .4325 | .4286 | .4247 |

*Continued on the back …*



**Z-Table Continued (area to the left of Z)**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **0** | **0.01** | **0.02** | **0.03** | **0.04** | **0.05** | **0.06** | **0.07** | **0.08** | **0.09** |
| **0.0** | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 | .5239 | .5279 | .5319 | .5359 |
| **0.1** | .5398 | .5438 | .5478 | .5517 | .5557 | .5596 | .5636 | .5675 | .5714 | .5753 |
| **0.2** | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 | .6026 | .6064 | .6103 | .6141 |
| **0.3** | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 | .6406 | .6443 | .6480 | .6517 |
| **0.4** | .6554 | .6591 | .6628 | .6664 | .6700 | .6736 | .6772 | .6808 | .6844 | .6879 |
| **0.5** | .6915 | .6950 | .6985 | .7019 | .7054 | .7088 | .7123 | .7157 | .7190 | .7224 |
| **0.6** | .7257 | .7291 | .7324 | .7357 | .7389 | .7422 | .7454 | .7486 | .7517 | .7549 |
| **0.7** | .7580 | .7611 | .7642 | .7673 | .7704 | .7734 | .7764 | .7794 | .7823 | .7852 |
| **0.8** | .7881 | .7910 | .7939 | .7967 | .7995 | .8023 | .8051 | .8078 | .8106 | .8133 |
| **0.9** | .8159 | .8186 | .8212 | .8238 | .8264 | .8289 | .8315 | .8340 | .8365 | .8389 |
| **1.0** | .8413 | .8438 | .8461 | .8485 | .8508 | .8531 | .8554 | .8577 | .8599 | .8621 |
| **1.1** | .8643 | .8665 | .8686 | .8708 | .8729 | .8749 | .8770 | .8790 | .8810 | .8830 |
| **1.2** | .8849 | .8869 | .8888 | .8907 | .8925 | .8944 | .8962 | .8980 | .8997 | .9015 |
| **1.3** | .9032 | .9049 | .9066 | .9082 | .9099 | .9115 | .9131 | .9147 | .9162 | .9177 |
| **1.4** | .9192 | .9207 | .9222 | .9236 | .9251 | .9265 | .9279 | .9292 | .9306 | .9319 |
| **1.5** | .9332 | .9345 | .9357 | .9370 | .9382 | .9394 | .9406 | .9418 | .9429 | .9441 |
| **1.6** | .9452 | .9463 | .9474 | .9484 | .9495 | .9505 | .9515 | .9525 | .9535 | .9545 |
| **1.7** | .9554 | .9564 | .9573 | .9582 | .9591 | .9599 | .9608 | .9616 | .9625 | .9633 |
| **1.8** | .9641 | .9649 | .9656 | .9664 | .9671 | .9678 | .9686 | .9693 | .9699 | .9706 |
| **1.9** | .9713 | .9719 | .9726 | .9732 | .9738 | .9744 | .9750 | .9756 | .9761 | .9767 |
| **2.0** | .9772 | .9778 | .9783 | .9788 | .9793 | .9798 | .9803 | .9808 | .9812 | .9817 |
| **2.1** | .9821 | .9826 | .9830 | .9834 | .9838 | .9842 | .9846 | .9850 | .9854 | .9857 |
| **2.2** | .9861 | .9864 | .9868 | .9871 | .9875 | .9878 | .9881 | .9884 | .9887 | .9890 |
| **2.3** | .9893 | .9896 | .9898 | .9901 | .9904 | .9906 | .9909 | .9911 | .9913 | .9916 |
| **2.4** | .9918 | .9920 | .9922 | .9925 | .9927 | .9929 | .9931 | .9932 | .9934 | .9936 |
| **2.5** | .9938 | .9940 | .9941 | .9943 | .9945 | .9946 | .9948 | .9949 | .9951 | .9952 |
| **2.6** | .9953 | .9955 | .9956 | .9957 | .9959 | .9960 | .9961 | .9962 | .9963 | .9964 |
| **2.7** | .9965 | .9966 | .9967 | .9968 | .9969 | .9970 | .9971 | .9972 | .9973 | .9974 |
| **2.8** | .9974 | .9975 | .9976 | .9977 | .9977 | .9978 | .9979 | .9979 | .9980 | .9981 |
| **2.9** | .9981 | .9982 | .9982 | .9983 | .9984 | .9984 | .9985 | .9985 | .9986 | .9986 |
| **3.0** | .9987 | .9987 | .9987 | .9988 | .9988 | .9989 | .9989 | .9989 | .9990 | .9990 |
| **3.1** | .9990 | .9991 | .9991 | .9991 | .9992 | .9992 | .9992 | .9992 | .9993 | .9993 |
| **3.2** | .9993 | .9993 | .9994 | .9994 | .9994 | .9994 | .9994 | .9995 | .9995 | .9995 |
| **3.3** | .9995 | .9995 | .9995 | .9996 | .9996 | .9996 | .9996 | .9996 | .9996 | .9997 |
| **3.4** | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9998 |

### 6.1.4 Practice Problems

Do these problems on a separate piece of paper. Label the assignment "6.1.4 Practice Problems". Be sure to show your work and/or support your reasoning. The correct result without a sufficient amount of correct and appropriate work is worth zero points. I may collect this assignment at the beginning of our next class.

1. The heights of men have a normal distribution with mean 69.0 inches and standard deviation of 2.8 inches. Let *X* be the height of a randomly chosen man. Use z-scores, the standard normal distribution, and the z-table to find each of the following probabilities (be sure to draw and shade in the normal density curve for each probability BEFORE trying to calculate the z-scores and the probability). Show your work.
2. Now use the steps below for your TI-83 or TI-84 graphing calculator to find the same probabilities as in number 1 above.

* Press **2nd** and then **Vars** (to get the **DISTR** menu)
* Select **normalcdf(**
* Complete the entry to obtain   
  **normalcdf(*lower bound, upper bound, Mean, Standard Deviation*)**   
  by substituting in the appropriate values and closing the parentheses.
* Press **ENTER**

1. (Use either the z-score method or your graphing calculator). An average light bulb manufactured by the Acme Corporation lasts 300 days with a standard deviation of 50 days. Assuming that bulb life is normally distributed, what is the probability that an Acme light bulb will last
2. at most 365 days?
3. at least 14 months (assume one month is 30 days)?
4. Between 6 and 18 months (assume a month is 30 days)?
5. Suppose scores on an IQ test are normally distributed. If the test has a mean of 100 and a standard deviation of 10, find each of the following.
6. The probability that a person who takes the test will score between 90 and 110.
7. The probability that a person who takes the test will score between 82 and 123.
8. England's Department of Health tracks the wait times for the 1st treatment after diagnosis for all types of cancers. For the 2nd quarter of 2012, data was collected from 4455 patients across the country

**# of Patients Receiving 1st Treatment for Cancer Within the Indicated Number of Days**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Days** | 31 | 38 | 48 | 62 | 76 | 90 | 104 |
| **# Patients** | 1380 | 808 | 973 | 1047 | 108 | 84 | 55 |

*(31 days indicates that a patient received the 1st treatment during the period from 0 days up to 31 days, and 38 days indicates that a patient received 1st treatment during the period from 31 days up to 38 days, etc.). Source: England's Department of Health* <http://www.dh.gov.uk/en/Publicationsandstatistics/Statistics/Performancedataandstatistics/HospitalWaitingTimesandListStatistics/index.htm>

1. Let *X* represent the number of days within which 1st treatment occurred. Fill in the probability table below. No need to show your work here.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***X*** | 31 | 38 | 48 | 62 | 76 | 90 | 104 |
| **P(*X*)** |  |  |  |  |  |  |  |

1. Now use the probabilities you found in part a) to calculate the mean, , and standard deviation, , for the discrete random variable *X*. Do show your work.

# **MODULE 7**

# **Linking Probability to**

# **Statistical Inference**

## 7.1.1 Distribution of Sample Proportions (Exploring with Skittles)

Learning Objectives

* Distinguish between a sample statistic and a population parameter.
* Explain the concept of sampling variability.
* Describe the sampling distribution of sample proportions.

**Activity 1**

For this activity we define the population to be all Skittles. Open your bag of Skittles (do NOT eat any of them yet), dump them out on your paper plate, and count and record the number of Skittles (then have one of your group members count your Skittles to verify). To make sure each student has the same number of Skittles, the instructor will randomly remove Skittles from students with more than 60 and randomly distribute additional Skittles to students with fewer than 60.

We're interested in "predicting" the proportion of purple Skittles in the population of Skittles, so let  
 *X* represent the color of a randomly selected Skittle.

1. Describe the population and sample(s), if any, for this activity.
2. Does *X* represent categorical or quantitative data? So, which sample statistic should we use, proportions, , or means, , to summarize the sample data (i.e. the data for each bag of skittles)?
3. Once every student has the same number of Skittles, record each of the following for your sample.

|  |  |
| --- | --- |
| n = |  |
| # red = | (proportion of red Skittles) |
| # green = | (proportion of green Skittles) |
| # purple = | (proportion of purple Skittles) |
| # orange = | (proportion of orange Skittles) |
| # yellow = | (proportion of yellow Skittles) |

**Side Note:** We use lower case letters to represent sample statistics and upper case letters to represent population parameters. For example the sample size is represented with a lower case ***n*** and the population size with an upper case ***N***, and the sample proportion is represented with lower case (pronuounced p-hat) and the population proportion with an upper case ***P***.

As the instructor enters your data into Tinkerplots, please verify that Tinkerplots calculates the correct proportions for your sample. Note: You are not expected to know how to use Tinkerplots as the instructor is merely using it to display and summarize the Skittles data.

1. Would you expect all of the sample proportions of purple Skittles to be approximately the same? Why or why not?
2. Would you expect your sample proportion of purple Skittles to be the same as the population proportion of purple Skittles ? Why or why not?
3. Typically statisticians cannot know the population proportion, but the Skittles folks claim that they produce an equal number of each of the five colors of Skittles, so according to the Skittles folks, what is the population proportion of purple Skittles?

1. Your instructor has displayed a dot-plot for the distribution of the sample proportions of purple Skittles from this class along with the samples from other classes. Sketch the histogram for this distribution. Indicate your sample proportion of purple Skittles and the population proportion of purple Skittles on the horizontal axis. Finally, mark the mean of the (i.e. the mean of the sample proportions of purple Skittles) on the horizontal axis (the instructor will use Tinkerplots to find the mean of for you). What approximate shape does the data have?

*Continued next page …*

1. What do you notice about the population proportion and the mean of all the ? Do you think this would happen if we only used the data from five randomly selected bags of 60 Skittles to calculate the mean of ? What do you think the mean of the sample proportions, , would be if we could use the data from all possible samples of 60 Skittles?
2. Look at the graph you sketched on the previous page. Is your sample proportion unusual (i.e. is your within the bulk of the data … maybe the middle 95%, or is it in one of the tails)?
3. If we did not know the population proportion , would it be appropriate to infer the population proportion from your sample proportion (i.e. can we use your sample proportion to estimate the population proportion )? How do you know?
4. Given a sample proportion that lies within the bulk of the sample proportion data (i.e. not in the tails), let's calculate the maximum error in using to estimate the population proportion, . Your instructor will calculate one standard error for the sample proportions of purple skittles (for now, you can think of standard error as standard deviation, i.e. variability, for categorical data). Mark three standard errors on each side of the population proportion on your plot in number 7 (just like you used to do for standard deviation).
5. Would you say that the bulk of the sample proportions of purple Skittles is within 1, 1.5, 2, or 2.5 standard errors of the population proportion ?

The bulk of our sample proportions is within standard error(s) of P.

1. Now we'll use your answer to part a) to calculate the average error due to chance for using any one of the sample proportions within the bulk of the data to estimate the population proportion.

Error = (convert the error to a percent)

So … if we use any one of the sample proportions in the bulk of the data, i.e.   
within standard error(s) of P (or the mean of the sample proportions if we don't know P) to estimate P, the maximum error in our estimation will be %. We also say that this is the expected error in estimating the population proportion.

1. Our sample size is 60 Skittles. Suppose we used a smaller sample size … perhaps only 10 Skittles. Do you think the expected error would be about the same, smaller, or larger? Explain - and you should use words like *variability* and *standard error* (i.e. standard deviation for categorical data) in your explanation.
2. Since we know the population proportion , you can calculate the error if we were to use your sample to estimate the sample proportion. Do it. Is your sample proportion within the expected error we found in part b?
3. Would you say that your sample of Skittles is usual or unusual? Explain.

Ahhh … in real life we don't know the population proportion , so if we use a sample proportion to estimate , we won't be able to determine whether the estimation is within the expected error (i.e. we won't be able to determine whether our sample is usual or unusual). But … we can use a confidence interval to say things like, we are 95% confident that the population proportion is within a particular error.

**Warning: You should still bring your graphing calculator to class every day.**

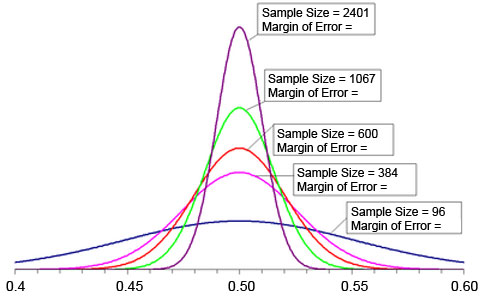
## 7.1.2 Distributions of Sample Proportions (Making Connections to Probability Models)

Learning Objective

* Draw conclusions about a population from a simulation and from a model of the sampling distribution.

**Warm-up**

For each of the following probability density curves, find the margin of error for the 95% confidence interval.



1. Assume that 20% of all Skittles are purple (i.e. suppose that the population proportion is ). What is the probability that a random sample of 60 Skittles estimates this population proportion within 6.5%? To answer this question, we'll develop a probability model for inference (i.e. we'll develop a probability model to draw conclusions about the population – such as estimating the population proportion).
2. Your instructor has displayed the distribution of sample proportions of purple Skittles on the board along with the mean of the sample proportions (i.e. the mean of ). Sketch the histogram for this distribution of sample proportions in the space below. Indicate the mean of the sample proportions as well as the population proportion on the horizontal axis.
3. So far we are only equipped to answer probability questions such as, "What is the probability that a random sample of 60 Skittles estimates this population proportion within 6.5%," if we are working with a Normal (or approximately Normal) distribution.

Is the distribution of sample proportions of purple Skittles above approximately Normal? How do you know?

Normal distributions have many different shapes. Some are tall and skinny. Some are short and fat. So how do you know that data is approximately normally distributed? Also, what does approximately normally distributed mean?

Hmmm … to answer the previous set of questions, we'll need to explore what the population proportion means in terms of successes and failures. So, for the population of all Skittles, what does the population proportion of purple Skittles mean in terms of successes and failures?

Since each of the our samples consists of 60 Skittles (i.e. the sample size is ), what is the expected number of successes when randomly selecting a Skittle from a sample and testing it for "purple" (i.e. how many of the Skittles would we expect to be purple)? What did you do to find the expected number of successes?

What is the expected number of failures when randomly selecting a Skittle for your sample and testing it for failure? What did you do to find the expected number of failures?

So if we have a sample of Skittles, and is the population proportion of, say green Skittles, what is the expected number of successes? What is the expected number of failures?

As it turns out a sampling distribution (the distribution of all possible samples of size ) is Normal if the expected number of successes is at least 10 **AND** the expected number of failures is at least 10. Write formulas to represent these facts. Then use the formulas to determine if the distribution of all sample proportions of purple Skittles is Normal.

1. Hmmm … what is the minimum sample size that would allow us to conclude that the distribution of sample proportions of purple Skittles is Normal?

O.K. we can continue to work on the question at hand, "What is the probability that a random sample of 60 Skittles estimates this population proportion within 6.5%?"

1. Convert the error to a decimal and use the population proportion to calculate the range of random sample proportions which can estimate this population proportion to within . Write the range in the space below, mark these values on the horizontal axis of the histogram you drew in part a). Be sure to shade in the appropriate region of the graph.
2. We'll use z-scores to calculate the probability that a random sample of 60 Skittles falls within the shaded region of your histogram (see why we needed the distribution to be Normal???). Hmmm … we need the *standard deviation* for our z-score formula from Module 6, but we're working with categorical data, so we can't use *standard deviation* like we do for quantitative data. With categorical data, we use *standard error* rather than *standard deviation* (although we also refer to standard error as standard deviation). Describe the similarities between each categorical data formula below and its comparable quantitative data formula.

|  |  |
| --- | --- |
| **CATEGORICAL DATA** | **QUANTITATIVE DATA** |
|  |  |
|  |  |

1. O.K. use z-scores to calculate the probability that a random sample of 60 Skittles falls within the shaded region of your histogram in part a). Show your work.
2. What is the probability that a random sample of 60 Skittles estimates this population proportion within 6.5%?
3. Assume that 20% of all Skittles are green (i.e. again). What is the probability that a random sample of 60 Skittles estimates this population proportion within 5% (We haven't mentioned it before, but this probability is called a **confidence level** and it's the same percentage for the associated **confidence interval**.) Your instructor should display the histogram for the sample proportions of green Skittles, and then you should follow the steps of the previous problem to answer this question. Be sure to start by first verifying that the distribution of the sample proportion of green Skittles is indeed Normal (otherwise you cannot use the z-scores to calculate the probability).
4. Find each of the following.
5. What is the margin of error for the 95% confidence level for the sample proportion of green Skittles?
6. What is the range for the 95% confidence interval?
7. What is the margin of error for a 68% confidence level for the sample proportions of green Skittles?
8. What is the range for the 68% confidence interval?

**WARNING: At the beginning of our next class, you may have another group quiz on the BIG PICTURE OF STATISTICS (you'll have to draw the big picture and label and briefly describe each component).**

## 7.2.1 Introduction to Statistical Inference: Confidence Intervals

Learning Objectives

* Use sample data and properties of the sampling distribution of a sample proportion to reason informally about estimating a population proportion.
* Explain how the margin of error and the confidence level for a confidence interval are related to the sampling distribution.

**Review**

**WARNING:** We may have a verbal group quiz on this review section and/or the subsequent activities.

To answer the next few questions, you may want to review The Big Picture in the introductory section of our class on OLI.

1. What is statistics all about (i.e. what is its purpose)?
2. If statistics is a process, what are the three major steps in the process and how do the four components of the Big Picture of Statistics fit into these three major steps?

Statistics as a process and the Big Picture of Statistics

Step 1:

Big Picture Component(s):

Step 2:

Big Picture Component(s):

Step 3:

Big Picture Component(s):

1. So what is *statistical inference*?
2. What are the two types of inference we'll use in this class and how is each used?

**Activity 1**

The following is an excerpt from the *Sleep Studies* section of the National Sleep Foundation's website (<http://www.sleepfoundation.org/article/sleep-topics/sleep-studies>).

According to the National Institutes of Health, 50 to 70 million Americans are affected by chronic sleep disorders and intermittent sleep problems that can significantly diminish health, alertness and safety. Untreated sleep disorders have been linked to hypertension, heart disease, stroke, depression, diabetes and other chronic diseases. Sleep problems can take many forms and can involve too little sleep, too much sleep or inadequate quality of sleep.

... Sleep problems can have serious consequences. According to the National Highway Traffic Safety Administration, drowsy driving claims more than 1,500 lives and causes at least 100,000 motor vehicle crashes each year.

According to the March 7, 2011, article, *Annual Sleep in America Poll Exploring Connections with Communications Technology Use and Sleep*, nearly 60% of Americans polled indicate they have a sleep problem every night or almost every night. The website further states, "The 2011 Sleep in America*®*annual poll was conducted for the National Sleep Foundation by WB&A Market Research, using a random sample of 1,508 adults between the ages of 13-64. The margin of error is 2.5 percentage points at the 95% confidence level.”

1. For statistical inference, what are the four most relevant pieces of information in the above paragraph?
2. What does the 60% represent?
3. Suppose we magically collect all possible samples of 1,508 adults between the ages of 13 to 64, and create a distribution of the sample proportions of those with a sleep problem every night (or almost every night). Would the distribution be Normal? Explain.
4. What is the range of the 95% confidence interval, and what can you infer from this about all American adults between the ages of 13 and 64?
5. What would the standard error (a.k.a. standard deviation for the distribution of sample proportions) be? Can we actually calculate the standard error? If so, do it. If not, explain why not?
6. What is the confidence level for a margin of error of ?

**Activity 2**

The following is an excerpt from the New York Times article, *How the Poll Was Conducted*, published October 23, 2009. It describes how a poll about texting while driving was conducted.

The latest New York Times/CBS News Poll is based on telephone interviews conducted Oct. 5 through 8 with 829 adults throughout the United States.

The sample of land line telephone exchanges called was randomly selected by a computer from a complete list of more than 69,000 active residential exchanges across the country. The exchanges were chosen so as to ensure that each region of the country was represented in proportion to its population.

Within each exchange, random digits were added to form a complete telephone number, thus permitting access to listed and unlisted numbers alike. Within each household, one adult was designated by a random procedure to be the respondent for the survey. To increase coverage, this land line sample was supplemented by respondents reached through random dialing of cell-phone numbers. The two samples were then combined.

Interviewers made multiple attempts to reach every phone number in the survey, calling back unanswered numbers on different days at different times of both day and evening. The combined results have been weighted to adjust for variation in the sample relating to geographic region, sex, race, marital status, age and education. In addition, the land line respondents were weighted to take account of household size and number of telephone lines into the residence, while the cell-phone respondents were weighted according to whether they were reachable only by cell-phone or also by land line.

In theory, in 19 cases out of 20, overall results based on such samples will differ by no more than 3 percentage points in either direction from what would have been obtained by seeking to interview all American adults. For smaller subgroups, the margin of sampling error is larger. Shifts in results between polls over time also have a larger sampling error.

1. Interpret the statement: "In theory, in 19 out of 20 cases, overall results based on such samples will differ by no more than 3 percentage points in either direction from what would have been obtained by seeking to interview ALL American adults."
2. If 80% of the respondents to the New York Times poll said that using a hand-held cell phone while driving should be illegal:
3. If we were able to repeat the sampling method above to obtain all possible samples of 829 adults in the United States, would the distribution of sample proportions be Normal? How do you know?
4. Calculate the standard error.
5. What is the range of the 95% confidence interval, and what can you infer about ALL American adults from this?
6. What is the confidence level for a margin of error of ?
7. If 27% of the respondents to the New York Times poll said that using a hands-free cell phone while driving should be illegal:
8. If we were able to repeat the sampling method above to obtain all possible samples of 829 adults in the United States, would the distribution of sample proportions be Normal? How do you know?
9. Calculate the standard error.
10. What is the range of the 95% confidence interval, and what can you infer about ALL American adults from this?
11. What is the confidence level for a margin of error of ?

**Summary**

1. Suppose we have categorical data with a population proportion of successes. If we could magically find all random samples of size from this population:
2. **(Center)** What is the mean for the distribution of all random samples?
3. **(Spread)** What do we use to talk about variability for the distribution of random samples?
4. **(Shape)** How do we know that the distribution of random samples is Normal?
5. Explain how the margin of error and the confidence level are related to the distribution of sample proportions?
6. Confidence Interval Formula: You've been inadvertently using the following formula. What does each piece of the formula represent?

Confidence Interval =

## 7.2.2 Introduction to Statistical Inference: Exploring Hypothesis Testing

Learning Objective

* Use sample data and properties of the sampling distribution of a sample proportion to reason informally about testing a claim about a population proportion.

**(Testing a Claim About a Population)** We need a volunteer who will stand by the statement, "I claim that I make 80% of my basketball free throws." We will of course give the volunteer a few practice shots before testing his or her claim.

1. To test our volunteer's claim, we need to take a sample of the volunteer's free throws. How large should the sample size be for us to assume that the sample distribution (the distribution of the percentage of shots made for all possible samples of this size) is Normal?
2. O.K. our volunteer will now attempt free throws. Record the number of successes and determine the percentage of baskets made?

Number of baskets made = Percentage of baskets made =

1. For each of the following statements, fill in the blanks and then assume the statement is true and explain what the statement would infer (if anything) about our volunteer's claim.
2. "Someone who makes 80% of his or her free throws is unlikely to make % out   
   of free throws."

If this statement were true, what would it infer (if anything) about the claim that our volunteer makes 80% of his or her free throws?

1. "Someone who makes 80% of his or her free throws is likely to make % out  
   of free throws."

If this statement were true, what would it infer (if anything) about the claim that our volunteer makes 80% of his or her free throws?

1. Since we selected a sample size that would ensure the sampling distribution is Normal, first explain how we might be able to determine which of the two previous statements is true then do it.
2. So, should we accept the claim that our volunteer makes 80% of his or her free throws? Why or why not?

### 7.2 Practice Problems

Read the summary below and then answer the questions on a separate piece of paper. Label the assignment "7.2 Practice Problems". Be sure to show your work and/or support your reasoning. The correct result without a sufficient amount of correct and appropriate work/explanation is worth zero points. I may collect this assignment at the beginning of our next class.

**Confidence in Institutions Summary**

Information in this summary was obtained from Gallup's *Confidence in Institutions* poll at <http://www.gallup.com/poll/1597/confidence-institutions.aspx>. Gallup's polling methodology is provided on the back of this page.

Americans' confidence in the United States Supreme Court is down nearly twenty percentage points from its all-time high since Gallup began tracking public confidence in American institutions in May 1973. In June 2012 only 37% of Americans reported a great deal or quite a lot of confidence in the United States Supreme Court compared to an all-time high of 56% in May 1985 and September 1988.

Americans' confidence in the U.S. military, the institution with the highest confidence rating in June 2012 (75% of adults surveyed reported a great deal or quite a lot of confidence in the US military), is down only seven percentage points from an all-time high of 85% in June 2003. The church or organized religion has lost twenty-four percentage points with 44% of those surveyed reporting a great deal or quite a lot of confidence in the church in June 2012 compared to 68% in May 1975. However Americans' confidence in the US Congress is down nearly thirty percentage points with only 13% of those surveyed in June 2012 reporting a great deal or quite a lot of confidence in Congress compared to 42% in May 1975. Banks have suffered an even greater loss in confidence, nearly forty percentage points. In June 2012, 21% of those surveyed reported a great deal or quite a lot of confidence in banks compared to 60% in April 1979.

1. What type of statistical measure is used throughout the summary?
2. What is the population from which the sample is drawn?
3. The description of Gallup's polling methodology (see below) reports that "one can say with 95% confidence that the margin of error is ±4 percentage points." What does this mean?
4. Overall, do you believe the statistics used in the summary are reasonably good estimates of the population? Explain.
5. In June 2012 only 44% of those surveyed reported a great deal or quite a lot of confidence in the church or organized religion. Find the 99% confidence interval for the sample distribution. Does the 99% confidence interval indicate that in June 2012 it was likely that only a minority of Americans have confidence in the church or organized religion?

**Gallup Polling Methodology:** Results are based on telephone interviews conducted June 7-10, 2012 with a random sample of 1004 adults aged 18+, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, one can say with 95% confidence that the margin of error is ±4 percentage points. Interviews are conducted with respondents on landline telephones and cellular phones, with interviews conducted in Spanish for respondents who are primarily Spanish-speaking. Each sample includes a minimum quoto of 400 cell phone respondents and 600 landline respondents, with additional minimum quotas among landline respondents by region. Landline numbers are chosen at random among listed telephone numbers. Cell phone numbers are selected using random digit dial methods. Landline respondents are chosen at random within each household on the basis of which member had the most recent birthday. Samples are weighted by gender, age, race, Hispanic ethnicity, education, region, adults in the household, and phone status (cell phone only, landline only, both, unlisted landline number, and cell phone mostly). Demographic weighting targets are based on the March 2011 Current Population Survey figures for the age 18+ non-institionalized population living in U.S. telephone households. All reported margins of sampling error include the computed design effects for weighting and sample design.