Introduction To Statistics
Think & Do
`Version 4.1`
by Scott Stevens
Champlain College
Burlington, Vermont, USA

ISBN 978-0-9885572-2-2

- **Online Homework** WebAssign

  Online homework is available through WebAssign® (webassign.net). The problems are fully randomized which allows students to see detailed solutions with the option to *Practice Another Version*.
  A link for content and pricing details can be found at www.StevensStats.com

- **Online Video Lectures by Chapter**

  In these videos, the author summarizes the content, reviews the examples, and demonstrates step-by-step solutions to all of the *Your Turn* problems found in the text. A link to these pages can be found at www.StevensStats.com

- **Online Software Demonstrations and Videos**

  No software (aside from a calculator) is needed to complete the material in this text/workbook. It has not been written for use with any specific software in mind. However, much of the material is amenable to software applications and the results obtained from software are used throughout the book. Instructions and demonstration videos for various software packages (Excel, SPSS, TI-82, TI-83/84, & Minitab) can be found at www.StevensStats.com .

- **Instructor Version**

  Instructors are provided a full electronic pdf version of this text including detailed solutions to all examples, Your Turn problems, worksheets, discussions, and exercises within the text.

- **Early Correlation and Regression**

  Chapter 10, Correlation and Regression, has been written in such a way to be amenable for presentation directly after Chapter 3 for those instructors desiring an early introduction to this material. If this is done, Chapter 10.3 should be skipped.

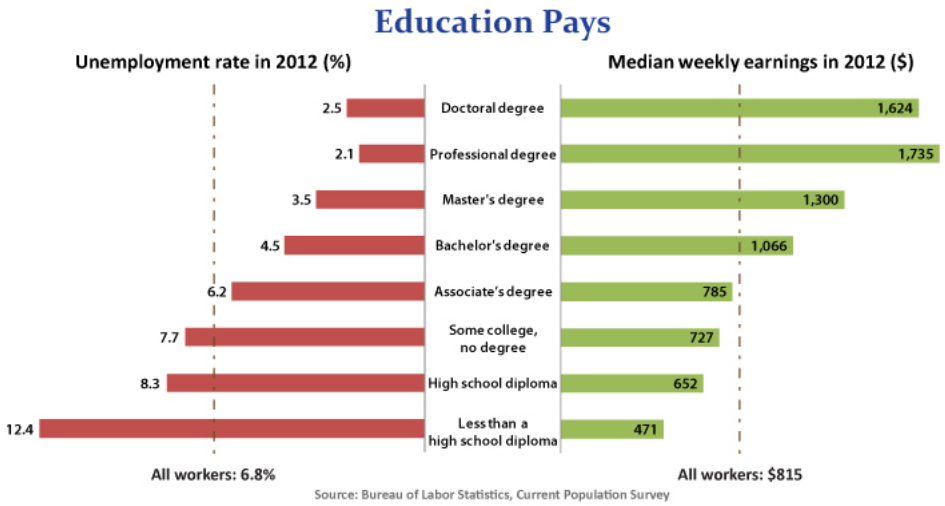
- **New to Version 4.1**

  This version contains many improvements including a revised chapter on two-sample hypothesis testing with a new section on proportions. A new chapter on Chi-Squared and ANOVA tests has been added to this version as well.

1 Introduction

Here are some *statistics* that students have found. I don't know for certain if any are true. Without formally fact-checking these numbers, consider what is required to make these claims. Are they believable?

1. The average pineapple plant is 3.47 feet tall.
2. Married men live, on average, ten years longer than unmarried men.
3. Right-handed people live, on average, nine years longer than left-handed people.
4. 1 in 112,000,000 people will die from a vending machine accident in a year.
5. 1 in 289,200,000 people will die from a shark attack per year.
6. Girls have more taste-buds than boys.
7. The proportion of people who can roll their tongue is between 65 and 81 percent.
8. The average person spends 2 weeks of life waiting for traffic lights to change.
9. Americans throw out 27% of the 350,000,000 pounds of food they buy each year.
10. 12% of lightning strikes occur at golf courses.
11. There are about 45 million tattooed Americans. About 17% of them will come to regret it.
12. 100,000 dogs are killed each year by falling out of pick-up trucks.
13. Based on a 2012 public opinion poll, about one in two people believe that colleges are not affordable.
14. There is a strong correlation between education and income & unemployment. *



*Published by the U.S. Bureau of Labor Statistics. http://www.bls.gov/emp/ep_chart_001.htm.

1.1 Statistics and Data

- **Definitions**
  - **Data** is a collection of observations about the members of a group - human or otherwise.
  - A **population** is the complete collection of all members in a group.
  - A **sample** is a sub-group of the population.
  - A **parameter** is a numerical description of some characteristic of a **population**.
  - A **statistic** is a numerical description of some characteristic of a **sample**.
  - **Statistics** (The Practice) comes in two different flavors
    - **Descriptive Statistics** is the collection, organization, and presentation of data.
    - **Inferential Statistics** is the art/science of making inferences (estimates, predictions) about population parameters from sample statistics.

    The link between these types of statistics is **Probability**.

- **Example:** Suppose I am in charge of lunches at Scooter's Summer Super-Fun Camp. I am looking to save a buck by serving a cheaper brand of macaroni and cheese than I currently serve because I suspect most of the kids won't be able to tell the difference. Before acting on this plan I decide to do a survey on a subgroup of kids at the camp. In a sample of 40 kids, it turns out that 26 were not able to tell the difference between the cheap stuff and the good stuff. I then, through the process described in Chapter 8, conclude that most kids at camp can not tell the difference but there is a 3% chance that I am wrong.
  - The **data** here consists of the results from the 40 kids in the survey.
  - The **population** I am concerned with is all of the kids at the camp.
  - The **sample** is the subgroup of 40 kids I select for my survey.
  - The **parameter** I seek is the percentage of all kids at camp who can tell the difference.
  - The **statistic** is that 65% (26/40) of kids in the sample can't tell the difference. This is a **descriptive statistic**.
  - The process of estimating the population parameter from the sample statistic is a form of **inferential statistics**. The 3% chance that I have reached the wrong conclusion is a **probability**.

- **Your Turn:** In a sample of 230 college students, the average number of hours slept per night is 6.2 hours. After analyzing the data, I am 90% confident that the average number of hours of sleep for all college students is between 5.9 and 6.5 hours per night. Determine the following:
  - What is the population I am studying?
  - What is the sample?
  - What is the statistic?
  - Describe the parameter we are seeking to find?
  - What is the estimate on this parameter?
  - What is the probability that my conclusion about the parameter is correct?

- **Qualitative and Quantitative Data**
  - **Qualitative (categorical) data** consist of non-numerical categories such as name, eye color, gender, college attended. Some appear to be numerical, such as ID numbers.
  - **Quantitative data** consist of numerical counts or measurements such as age, test-scores, rainfall, weight or the *number* of red cars. Quantitative Data can be subdivided as follows:
    - **Discrete data** can only take certain values within a given range - usually counts. There are gaps between possible data values. Examples include the number of cars sold by color, the number of children in a family, or calendar years (2011, 2012, ...).
    - **Continuous data** can take on any value in a given range - usually measurements such as time, length, volume, or weight. In between any two measurements exists another. Continuous data often appear to be discrete because of the measuring device.

- **Levels of measurements** (given here lowest to highest)
  1. **Nominal measurements** consist of categories, names, labels, etc., which can not be ordered, added, or averaged. Examples generally come from qualitative data but might be disguised numerically by an identification number of some sort.
  2. **Ordinal measurements** can be ordered (ranked) but the difference between measurements are not well defined. Examples: grades (A,B,C,D,F), hotel ratings, levels of pain.
  3. **Interval measurements** are like ordinal but better because differences are meaningful. However, zero is arbitrary and ratios do not make sense. Examples: dates, non-Kelvin temperatures.
  4. **Ratio measurements** are the best for numerical comparisons. Differences are meaningful, zero is not arbitrary, and ratios make sense. Examples: time, temperature in degrees Kelvin, counts, prices, weights, lengths, etc.

- **Examples:** Determine whether the given data is qualitative or quantitative. If it is quantitative, categorize it as discrete or continuous. Then, determine the level of measurement of the data collected.

  (a) The finishing times of the top 100 runners in the Boston Marathon.
  Answer: Quantitative, Continuous, Ratio

  (b) The ISBN numbers for all of your textbooks.
  Answer: While these are numerical, the numbers are acting as names so the data is qualitative and the level of measurement is nominal.

  (c) Each of 12 hotels are ranked by the number of stars.
  Answer: Quantitative, Discrete, Ordinal

  Your Turn

  (a) Each of 30 people in a stats class must categorize themselves as conservative, moderate, or liberal.

  (b) The number of donuts made by a baker on Sunday morning.

  (c) The high temperatures for each day this week in degrees Celsius.

1.2 Sampling

- **Census and Parameters -vs- Samples and Statistics**
  - When you gather data from every member of a population it is called a **census** and the results are called **parameters**.
  - When you gather data from a subgroup of a population it is called a **sample** and the results are called **statistics**.

- **Random and Simple Random Samples**
  - In a **random sample** every member of the population has an equal chance of being selected.
  - A **simple random sample** is a random sample where every sample of the same size has the same chance of being selected. There can be no sub-grouping of the population.
  - A simple random sample is always random but the reverse is not necessarily true.

- **Examples:**
  Classify each sampling method as simple random, random but not simple random, or neither.

  1. In a class with 200 males and 300 females, I want to select 50 students for a survey.
     (a) I randomly select 20 males and 30 females for the survey.
     Answer: This is random because all students have a 1/10 chance of being picked, but it is not simple random because I can't have a sample of say 25 males and 25 females.
     (b) I put all 500 students in a list and randomly select 50 students.
     Answer: This is simple random sample because everyone has an equal (1/10) chance of being picked **and** no sample of 50 has a better chance of being picked.
     (c) I randomly select 25 males and 25 females.
     Answer: Neither. Males have a 25/200 chance of being picked and females have a 25/300 chance of being picked. Not all students have the same probability of being picked so it is not even a random sample.

  2. **Your Turn:** In my class I have 30 students, there are 5 rows of 6 students each. I want to select 12 students for a survey. Classify each sampling method as simple random, random but not simple random, or neither.
     (a) I put all 30 students names in a basket and randomly select 12 students.

     (b) I randomly select two of the five rows and choose all students in each of these rows.

     (c) I randomly select one odd row and one even row and take all students in the chosen rows.

- **Sampling Strategies**
  - In a **systematic sample**, every n^{th} member of the population is selected.
  - In a **convenience sample**, the most convenient subgroup is selected.
  - A **stratified sample** is one in which the population is divided into two or more sub-groups, called strata, that share similar characteristics. We then draw a random sample from each strata. This is good if you want to make sure that some members from all strata are present in the sample. Note: A stratified sample may be random but it won't be simple random.
  - In **cluster sampling**, we divide the population into groups (or clusters), then randomly select some of those clusters. Once a cluster is selected all the members of that cluster are included in the sample. This method is often used for convenience purposes. Note: Again, a cluster sample may be random but it won't be simple random.
- **Examples:** Classify each sampling method as **systematic**, **convenience**, **stratified**, **cluster**, or **none of these**. Does the method constitute a random sample? If it is random, is it simple random?

1. You're considering a lunch-delivery business and want to gather lunch-break data on a sample of employees from your target population of 20 local businesses.
   (a) You randomly select 3 of the businesses and interview all the employees from those businesses.
   Answer: This is a cluster sample (the employees are clustered by the business employing them). It is random because all members of your target population have a 3/20 chance of being selected. It is not simple random because you have grouped your subjects prior to sampling.
   (b) You get a sample of 100 by randomly selecting 50 employees from labor and 50 employees from management.
   Answer: This is a stratified sample (the strata are labor and management). It is unlikely to be random unless there are an equal number of labor and management employees in your target population. Either way, it is not simple random because not all samples of 100 have the same probability of being selected.
2. **Your Turn:** Suppose you want to gather household income information from a sample of 10 houses on North Street. The house numbers start at 1 and end at 100 without any missing house numbers.
   (a) You randomly select 5 even numbered houses and 5 odd numbered houses.
   (b) You take every 10 $^{\text{th}}$ house starting at number 7.
   (c) You group the houses as #'s 1 - 10, 11 - 20, ..., 91 - 100. You then randomly select one of these groups to be in the sample.
   (d) You randomly select 10 numbers between 1 and 100 and select those 10 houses.
   (e) You include the first 10 houses where someone answers the door.