Collaborative Statistics:

Text

by
Barbara Illowsky
and
Susan Dean

Collaborative Statistics: Text

By Barbara Illowsky and Susan Dean
De Anza College
Cupertino, CA

TABLE OF CONTENTS

Preface

This book is intended for introductory statistics courses being taken by students at two– and four–year colleges who are majoring in fields other than math or engineering. Intermediate algebra is the only prerequisite. The book focuses on applications of statistical knowledge rather than the theory behind it. The text is named *Collaborative Statistics* because students learn best by *doing*. In fact, they learn best by working in small groups. The old saying "two heads are better than one" truly applies here.

Our emphasis in this text is on four main concepts:

- thinking statistically
- incorporating technology
- working collaboratively
- writing thoughtfully

These concepts are integral to our course. Students learn the best by actively participating, not by just watching and listening. Teaching should be highly interactive. Students need to be thoroughly engaged in the learning process in order to make sense of statistical concepts. *Collaborative Statistics* provides techniques for students to write across the curriculum, to collaborate with their peers, to think statistically, and to incorporate technology.

This book takes students step by step. The text is interactive. Therefore, students can immediately apply what they read. Once students have completed the process of problem solving, they can tackle interesting and challenging problems relevant to today's world. The problems require the students to apply their newly found skills. In addition, technology (TI-86 and TI-83 graphing calculators are highlighted) is incorporated throughout the text and the problems, as well as in the special group activities and projects. A workbook accompanies the text. It contains labs that use real data and practices that lead students step by step through the problem solving process.

At De Anza, along with hundreds of other colleges across the country, the college audience involves a large number of ESL students as well as students from many disciplines. The ESL students, as well as the non-ESL students, have been especially appreciative of this text. They find it extremely readable and understandable. *Collaborative Statistics* has been used in classes that range from 20 to 120 students, and in regular, honor, and distance learning classes.

Susan Dean
Barbara Illowsky

Acknowledgments

We wish to acknowledge the many people who have helped us and have encouraged us in this project. At De Anza, Donald Rossi and Rupinder Sekhon and their contagious enthusiasm started us on our path to this book. Inna Grushko and Diane Mathios painstakingly checked every practice and homework problem. Inna also wrote the glossary and offered invaluable suggestions.

Kathy Plum co-taught with us the first term we introduced the TI-85. Lenore Desilets, Charles Klein, Kathy Plum, Janice Hector, Vernon Paige, Carol Olmstead, and Donald Rossi of De Anza College, Ann Flanigan of Kapiolani Community College, Birgit Aquilonius of West Valley College, and Terri Teegarden of San Diego Mesa College, graciously volunteered to teach out of our previous editions. Janice Hector and Lenore Desilets also contributed valuable problems. Diane Mathios and Carol Olmstead contributed labs to the workbook. In addition, Diane and Kathy have been our "sounding boards" for new ideas.

Jim Lucas and Valerie Hauber of De Anza's Office of Institutional Research, along with Mary Jo Kane of Health Services, provided us with a wealth of data.

We would also like to thank the thousands of students who have used this text. So many of them gave us permission to include their outstanding word problems as homework. They encouraged us to turn our note packet into this book, have offered suggestions and criticisms, and keep us going.

Finally, we owe much to Frank, Jeffrey, and Jessica Dean and to Dan, Rachel, Matthew, and Rebecca Illowsky, who encouraged us to continue with our work and who had to hear more than their share of "I'm sorry, I can't" and "Just a minute, I'm working."

Dear Student:

Have you heard others say, "You're taking statistics? That's the hardest course I ever took!" They say that, because they probably spent the entire course confused and struggling. They were probably lectured to and never had the chance to experience the subject. You will not have that problem. Let's find out why.

There is a Chinese Proverb that describes our feelings about the field of statistics:

I HEAR, AND I FORGET
I SEE, AND I REMEMBER
I DO, AND I UNDERSTAND

Statistics is a "do" field. In order to learn it, you must "do" it. We have structured this book so that you will have hands-on experiences. They will enable you to truly understand the concepts instead of merely going through the requirements for the course.

What makes this book and the accompanying workbook different from other texts? First, we have eliminated the drudgery of tedious calculations. You might be using computers or graphing calculators so that you do not need to struggle with algebraic manipulations. Second, this course is taught as a collaborative activity. With others in your class, you will work toward the common goal of learning this material.

Here are some hints for success in your class:

- Work hard and work every night.
- Form a study group and learn together.
- Don't get discouraged - you can do it!
- As you solve problems, ask yourself, "Does this answer make sense?"
- Many statistics words have the same meaning as in everyday English.
- Go to your teacher for help as soon as you need it.
- Don't get behind.
- Read the newspaper and ask yourself, "Does this article make sense?"
- Draw pictures - they truly help!

Good luck and don't give up!

Sincerely,
Susan Dean and Barbara Illowsky
De Anza College
21250 Stevens Creek Blvd.
Cupertino, California 95014

Chapter One:

SAMPLING AND DATA

SAMPLING AND DATA

You are probably asking yourself the question, "When and where will I use statistics?". If you read any newspaper or watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a news program on television, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques to analyze the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

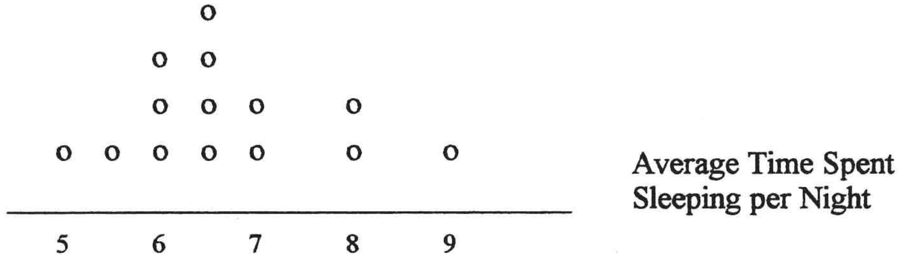
Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data are.

STATISTICS

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data.** We see and use data in our everyday lives. To be able to use data correctly is essential to many professions and is in your own best self-interest.

Collaborative Classroom Exercise

In your classroom, try this exercise. Have class members write down the average time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, if the data are the numbers 5, 5.5, 6, 6, 6, 6.5, 6.5, 6.5, 6.5, 7, 7, 8, 8, 9, the dot plot would be as follows:



Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?

Where do your data appear to cluster? How could you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called *descriptive statistics.* Two ways to summarize data are by graphing and by numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called *inferential statistics.* Statistical inference uses probability to determine if conclusions drawn are reliable or not.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

PROBABILITY

Probability is the mathematical tool used to study randomness. It deals with the chance of an event occurring. For example, if you toss a **fair** coin 4 times, the outcomes may not be 2 heads and 2 tails. However, if you toss the same coin 4,000 times, the outcomes will be close to 2,000 heads and 2,000 tails. The expected theoretical probability of heads in any one toss is 1/2 or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl Pearson who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction 996/2000 is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Today, probability is used to predict the likelihood of an earthquake, of rain, or whether you will get a A in this course. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is suppose to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

SOME KEY TERMS

In statistics, we generally want to study a **population**. You can think of a population as an entire collection of persons, things, or objects under study. To study the larger population, we select a **sample.** The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000 to 2,000 people are taken. The opinion poll is suppose to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that is a property of the sample. The average number of points earned in a math class at the end of a term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a number that is a property of the population. If we consider all math classes to be a population, then the average number of points earned per student in the population is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, notated by capital letters like X and Y, is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical. Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. Datum is a single value.

Example 1-1: Define the key terms from the following study: We want to know the average amount of money first year college students spend at ABC College on school supplies that do not include books. Three students spent $150, $200, and $225, respectively.

The **population** is all first year students attending ABC College this term.
The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).
The **parameter** is the average amount of money spent (excluding books) by first year college students at ABC College this term.
The **statistic** is the average amount of money spent (excluding books) by first year college students in the sample.
The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.
The **data** are the dollar amounts spent by the first year students. Examples of the data are $150, $200, and $225.

Collaborative Classroom Exercise

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

DATA

Data may come from a population or from a sample. Small letters like x or y generally are used to represent data values. Most data can be put into the following categories:

1. **Qualitative**
2. **Quantitative**: Quantitative data may be either **discrete** or **continuous.**

Qualitative data are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, and red. Blood type might be AB+, O-, or B+. Qualitative data are not as widely used as quantitative data because many numerical techniques do not apply to the qualitative data. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers and are usually the data of choice because there are many methods available for analyzing the data. Quantitative data are the result of *counting or measuring* attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and the number of students who take statistics are examples of quantitative data.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get 0, 1, 2, 3, etc.

All data that are the result of measuring are **quantitative continuous data** assuming that we can measure accurately. Measuring angles in radians might result in the numbers $\pi/6$, $\pi/3$, $\pi/2$, π, $3\pi/4$, etc. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

Example 1-2: Data sample of quantitative discrete data. The data are the number of books students carry in their backpacks. You sample five students. Two students carry 3 books, one student carries 4 books, one student carries 2 books, and one student carries 1 book. The numbers of books (3, 4, 2, and 1) are the quantitative discrete data.

Example 1-3: Data sample of quantitative continuous data. The data are the weights of the backpacks with the books in it. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are quantitative continuous data because weights are measured.

Example 1-4: Data sample of qualitative data. The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative data.

Note that you may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

Example 1-5: Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of." (Answers to the chapter examples are in the text. Look at the Table of Contents under "Answers to Chapter Examples.")

1. The number of pairs of shoes you own.
2. The type of car you drive.
3. Where you go on vacation.
4. The distance it is from your home to the nearest grocery store.
5. The number of classes you take per school year.
6. The tuition for your classes
7. The type of calculator you use.
8. Movie ratings.
9. Political party preferences.

10. Weight of sumo wrestlers.
11. Amount of money (in dollars) won playing poker.
12. Number of correct answers on a quiz.
13. Peoples' attitudes toward the government.
14. IQ scores. (This may cause some discussion.)

SAMPLING

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.**

Two common methods of sampling are **with replacement** and **without replacement.** If each member of a population may be chosen more than once then the sampling is with replacement. If each member may be chosen only once, then the sampling is without replacement.

One of the most important methods of obtaining samples is called **random sampling.** If each member of a population has an equal chance of being selected for the sample, the sample is called a **simple random sample.** Two simple random samples would contain members equally representative of the entire population. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 32 members including Lisa. To choose a simple random sample of size 3 from the other members of her class, Lisa first lists the last names of the members of her class together with a two-digit number as shown below.

00 Anselmo	**09 Jiao**	**18 Price**	**27 Tallai**
01 Bautista	**10 Khan**	**19 Quizon**	**28 Tran**
02 Bayani	**11 King**	**20 Reyes**	**29 Wai**
03 Cheng	**12 Legeny**	**21 Roquero**	**30 Wood**
04 Cuarismo	**13 Lundquist**	**22 Roth**	
05 Cunningham	**14 Macierz**	**23 Rowell**	
06 Fontecha	**15 Motogawa**	**24 Salangsang**	
07 Hong	**16 Okimoto**	**25 Slade**	
08 Hoobler	**17 Patel**	**26 Stracher**	

Lisa can either use a table of random numbers (found in many statistics books as well as mathematical handbooks) or a calculator or computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are:

.94360 .99832 .14669 .51470 .40581 .73381 .04399.

Lisa reads two-digit groups until she has chosen three class members (that is, she reads .94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers .94360 and .99832 do not contain appropriate two digit numbers. However the third random number, .14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cunningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, and Cunningham, and Cuarismo.

Sometimes, it is difficult or impossible to obtain a simple random sample because populations are too large. Then we choose other forms of sampling methods that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a sample from each stratum. For example, you could stratify (group) your college population by department and then choose a simple random sample from each stratum to get a stratified random sample.

To choose a **cluster sample**, divide the population into sections and then randomly select some of the sections. All the members from these sections are in the cluster sample. For example, if you randomly sample four departments from your stratified college population (randomly choose four departments from all of the departments), the four departments make up the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every nth piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. You start by randomly picking one of the first 50 names and then choose every 50th name thereafter. Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is nonrandom is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favors certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (for example, they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough or representative of the population. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

Example 1-6: Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience) .

1. A soccer coach selects 6 players from a group of boys aged 8 to 10, 7 players from a group of boys aged 11 to 12, and 3 players from a group of boys aged 13 to 14 to form a recreational soccer team.
2. A pollster interviews all human resource personnel in five different high tech companies.
3. An engineering researcher interviews 50 women engineers and 50 men engineers.
4. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
5. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
6. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

MORE ON SAMPLING

If we were to examine two samples representing the same population, they would, more than likely, not be the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will seem natural.

Example 1-7: Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task.

Suppose we take two different samples.

First, we use convenience sampling and survey 10 students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class . The amount of money they spend is as follows:

$128 $87 $173 $116 $130 $204 $147 $189 $93 $153

The second sample is taken by using a list from the P.E. department of senior citizens who take P.E. classes and taking every 5th senior citizen on the list, for a total of 10 senior citizens. They spend:

$50 $40 $36 $15 $50 $100 $40 $53 $22 $22

Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average part-time student. Both samples are biased. Also, in both cases, not all students have a chance to be be in either sample.

Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

No. Never use a sample that is not representative or does not have the characteristics of the population.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he/she has a corresponding number. The students spend:

$180 $50 $150 $85 $260 $75 $180 $200 $200 $150

Do you think this sample is representative of the population?

Yes. It is chosen from different disciplines across the population.

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Collaborative Classroom Exercise

As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

1. To find the average GPA of all students in a university, use all honor students at the university as the sample.

2. To find out the most popular cereal among young people under the age of 10, stand outside a large supermarket for three hours and speak to every 20th child under age 10 who enters the supermarket.

3. To find the average annual income of all adults in the United States, sample U. S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.

4. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.

5. To determine the average cost of a two day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

VARIATION IN DATA

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8 16.1 15.2 14.8 15.8 15.9 16.0 15.5

Measurements of the amount of beverage in a 16 ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

VARIATION IN SAMPLES

It was mentioned previously that two or more samples from the same population and having the same characteristics as the population may be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students sleep each night and use all students at their college as the population. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample even though both samples have the characteristics of the population. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) would be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This *variability in samples* cannot be stressed enough.