

CHAPTER

1

Introduction



- ▼ 1.1 Overview of Course (Basic Concepts)
- ▼ Population
- ▼ Sample
- Random Sample
- Internal and External Validity

- 1.2 Why We Sample
- Sampling to Determine μ
- Sampling to Determine p
- Summary
- Exercises

Statistical research touches us all: the clothes we wear, the TV we watch, the cars we drive, the education we receive, the medicines we take, the movies we admire, the warnings against cigarette smoking and other products, and perhaps even the salary we earn. Statistical research touches almost every aspect of our life.

To be more specific, the techniques discussed in this text are used extensively in the fields of psychology, education, health, marketing, agriculture, criminology, factory production, biology, medicine, advertising, economics, and a host of other disciplines in the soft sciences, hard sciences, and nonsciences. In other words, the techniques are used in almost any discipline that requires the analysis of data. Examples will be drawn arbitrarily from all fields since the basic tenets of statistical analysis are much the same whatever your field of interest. ▼

We shall define **statistics** as follows:

Sta-tist'ics

The collection, organization, and interpretation of numerical data for the purposes of

- a. summarization, and
- b. drawing conclusions about populations based on taking samples from that population.

When statistical techniques are used primarily for the purpose of summarizing data, this is called **descriptive statistics**. Descriptive techniques are discussed mostly in chapters 2 through 4. However when statistical techniques are used for the purpose of drawing conclusions about populations based on samples drawn from that population, this is called **inferential statistics**. Inferential techniques are discussed in chapters 4 through 11. As you can see, the primary focus of this text is on inference.

In'fer-ence

Decision making based on samples drawn from populations.

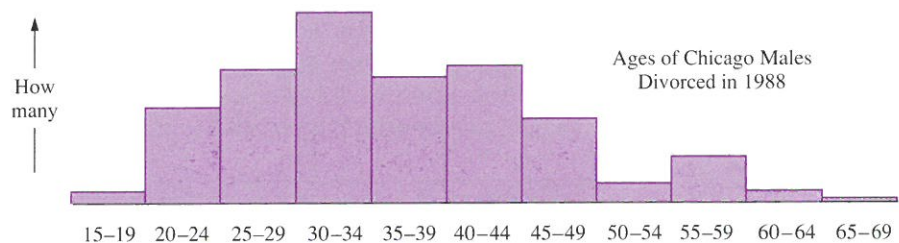
1.1 Overview of Course (Basic Concepts)

Although many of the following concepts are discussed in depth in subsequent chapters, certain terms are so crucial for a full understanding of the material that a brief overview is presented here.

Population

The term **population** has a unique meaning in statistics. It not only refers to a precisely defined entity or group of entities we wish to study (people, land, insects, or whatever) but also to the specific attribute we wish to measure.

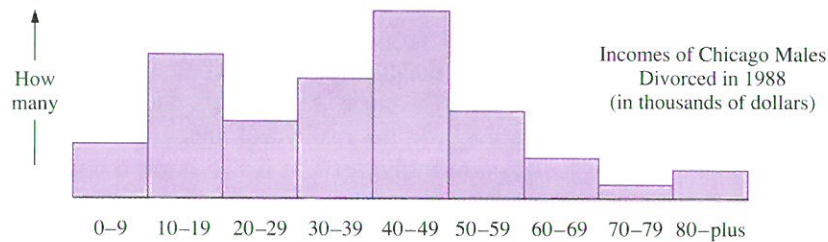
For instance, suppose we wish to study Chicago males divorced in 1988. Although in the real world, this might be considered a precisely defined population of individuals, it is not as yet a population in a statistical sense, because we have no attribute to measure. However, if we had said, we wish to study the *ages* of Chicago males divorced in 1988, then this would be a population in a statistical sense. Notice that a population in statistics must not only consist of the precisely defined entity or group of entities, in this case, Chicago males divorced in 1988, but also to a specific attribute we wish to measure, in this case, *age*. Suppose this population was presented as follows:



This is called a **histogram**.^{*} The height of the bars represents *how many* in each category. For instance, the category 30–34 years old is represented by the highest bar. This means more men were divorced in this 30–34-year-old age category than in any other age category. Histograms are of enormous value in statistics. A mere glance gives you substantial amounts of information. Notice very few men were divorced in the 50–54-year-old age category. Now, let's look at a different population.

^{*}The histogram is discussed at length in chapter 2.

Suppose we wish to study the *income* of Chicago males divorced in 1988. Although in the real world, the population is still Chicago males divorced in 1988, when we talk about a different attribute, in this case, income, in statistics this constitutes a *different* population. Here the population is incomes of Chicago males divorced in 1988. Suppose we were to actually measure this population and represent the results as follows:



Again, a histogram is used to represent this data. Notice, more men were divorced in the 40-49 (\$40,000-\$49,000) income category than in any other income category, although the 10-19 income category was a close second. In statistics, the histogram is the workhorse of data presentation. We will use it extensively in this text.

To sum up, when we talk about a population, we not only refer to a precisely defined entity or group of entities (in the above case, Chicago males divorced in 1988) but we also refer to a specific attribute we wish to measure, either age, income, emotional stability, number of offspring, educational level, or whatever.

Population

Any precisely defined collection of values we wish to study

Sample

Measuring an entire population can be exceedingly time-consuming, costly, and in many cases physically impossible. **Samples**, on the other hand, can often be taken with relative ease.

Sample

Part of a population

In this case, since we sampled 5 from the basket, we call this a random sample of size $n = 5$, yielding ages 52, 37, 20, 32, and 29.

In actuality, writing thousands of numbers (sometimes millions) on slips of paper and rotating large baskets can be tedious, so in practice, we very often use a master list, say of all the Chicago males divorced in 1988, assign each a number, then use something called a **random digit table**.

Let's see how it works. Say we have a total population of 5000 Chicago males divorced in 1988. We would proceed as follows:

Random Digit Method


Random Digit Method																																																																	
<p>Sam Benton 0001</p> <p>Mel Sims 0002</p> <p>.</p> <p>Ferris Callo 0299</p> <p>Al Gee 0300</p> <p>John Deitski 0301</p> <p>Juan Eckstin 0302</p> <p>.</p> <p>Chang Hoo 4997</p> <p>Joe Doet 4998</p> <p>Harry Fudim 4999</p> <p>Bill Leeder 5000</p>	<p>Random Digit Table</p> <table border="1"> <tr><td>08141</td><td>31926</td><td>30566</td><td>63607</td></tr> <tr><td>71798</td><td>99058</td><td>06766</td><td>20575</td></tr> <tr><td>81679</td><td>51276</td><td>84580</td><td>15001</td></tr> <tr><td>72925</td><td>57040</td><td>57012</td><td>99058</td></tr> <tr><td>72844</td><td>80198</td><td>08598</td><td>21410</td></tr> <tr><td colspan="4"> </td></tr> <tr><td>45929</td><td>94739</td><td>65371</td><td>48423</td></tr> <tr><td>36663</td><td>64130</td><td>17730</td><td>75755</td></tr> <tr><td>10610</td><td>05794</td><td>04717</td><td>95862</td></tr> <tr><td>16688</td><td>23757</td><td>25018</td><td>50541</td></tr> <tr><td>33967</td><td>24160</td><td>99725</td><td>85113</td></tr> <tr><td colspan="4"> </td></tr> <tr><td>86980</td><td>09066</td><td>19347</td><td>60203</td></tr> <tr><td>.</td><td>.</td><td>.</td><td>.</td></tr> <tr><td>.</td><td>.</td><td>.</td><td>.</td></tr> <tr><td>.</td><td>.</td><td>.</td><td>.</td></tr> </table>	08141	31926	30566	63607	71798	99058	06766	20575	81679	51276	84580	15001	72925	57040	57012	99058	72844	80198	08598	21410					45929	94739	65371	48423	36663	64130	17730	75755	10610	05794	04717	95862	16688	23757	25018	50541	33967	24160	99725	85113					86980	09066	19347	60203
08141	31926	30566	63607																																																														
71798	99058	06766	20575																																																														
81679	51276	84580	15001																																																														
72925	57040	57012	99058																																																														
72844	80198	08598	21410																																																														
45929	94739	65371	48423																																																														
36663	64130	17730	75755																																																														
10610	05794	04717	95862																																																														
16688	23757	25018	50541																																																														
33967	24160	99725	85113																																																														
86980	09066	19347	60203																																																														
.	.	.	.																																																														
.	.	.	.																																																														
.	.	.	.																																																														

First, we write up a master list of the 5000 Chicago males divorced in 1988, assigning each a four-digit identifying number, similar to what we did in the Rotating Basket Method.


Second, we start anywhere in the Random Digit Table, penciling off a section of four-digit numbers. Say for instance, we arbitrarily choose column 3, fifth number down to start. Then, we start circling numbers of 5000 or less in this section.* If we wish 5 in our sample, then we circle only 5 numbers.

Third, we locate the specific 5 Chicago males who were picked and determine their ages.


I'm one of the five selected and I'm 26.




I'm 41.




Me? 39.



43.



I'm 32.



So, this is another random sample of $n = 5$. This time the sample yielded ages 26, 41, 39, 43, and 32.

The random digit table, in effect, takes the place of the rotating basket. Numbers are written down for us, mixed thoroughly, and presented in table

*Since 5000 is a 4-digit number, we circle only groups of 4 digits. Note that in a random digit table, any digit sequence is possible. For instance, the first four-digit sequence we had chosen, 0859, was just as likely to occur as any other four-digit sequence, say 1234 or 0000 or 4444.

form. Our job is simply to reach into the random digit table and pull out any desired number of candidates.

The point is whether we use the rotating basket or the random digit table, everyone has an equal or near equal chance of being selected on each and every pick, and every sample of the same size has an equal chance of being selected. Experience has shown that samples selected in this way generally give reliable information about a population (provided certain assumptions are met, which will be discussed as we proceed).

Let me add: in many industry and research applications, obtaining a random sample is often one of the most difficult aspects of performing a statistical investigation. As a result, a number of other methods have been developed in an attempt to obtain samples representative of a population (stratified, cluster, systematic sampling, random intact groups, random assignment, haphazard sampling, and others), however many of these methods come with risks that are difficult to assess. Therefore these methods will not be used or discussed in this textbook. All sampling in this text assumes use of the rotating basket or random digit table methods.

Internal and External Validity

The ability to achieve a true representation of the population from a sample is often referred to as the ability of the experiment to achieve **external validity**. Although a random sample is an important factor in achieving external validity, a random sample by no means guarantees it. Other factors must be considered before we can safely use our random sample as truly representative of the population, and generally these factors are categorized under threats to either internal or external validity.

Internal Validity

The certainty that the observations in our sample group are *accurate* measures of the characteristics we set out to measure.

In other words, from our sample group, did we extract honest, accurate, and reliable information? If yes, then we have achieved internal validity.

In the case of Chicago males divorced in 1988, internal validity simply means: were the five ages we obtained from our sample *accurate* reflections of the true ages of the five males? In other words, did we obtain the real ages of these five males? Did we ask them in a telephone conversation and possibly risk

So, where we can, we usually take a sample in the hope the sample will give us accurate information about a population without having to actually measure the population in its entirety.

Samples, if properly taken, can tell us much about a population. However, samples, if improperly taken can mislead. Let us look at some examples of how samples can mislead.

For instance, 5 young men sampled walking out of the student cafeteria at Washington State University yielded an average height of $6'3\frac{1}{2}"$. Can we conclude the average height of *all* Washington State men is $6'3\frac{1}{2}"$? I doubt it. A sample of French women at a local convention showed that 80% were severely underweight. Can we conclude 80% of *all* French women are severely underweight? A poll of 25 joggers in Virginia Beach showed an average annual salary of \$183,000. Would this describe the *true* average annual salary of Virginia Beach joggers—or did we, perhaps, run into a pack of Wall Street tycoons on vacation?

What's wrong with the above samples? Chances are, they are not truly representative of the population from which they were drawn.

Of course, this leads us to the question: how do we obtain samples that *are* truly representative of the population from which they are drawn? To start off, we should have a **random sample**.

Random Sample

We have seen in the above examples that it is not enough to take just any sample from a population. Samples can easily give false impressions. Even if we try to be fair when taking a sample, often the information we obtain will give us misleading information about a population.

To better guarantee a true representation of a population, the sample should be a random sample, which leads to the following definition.

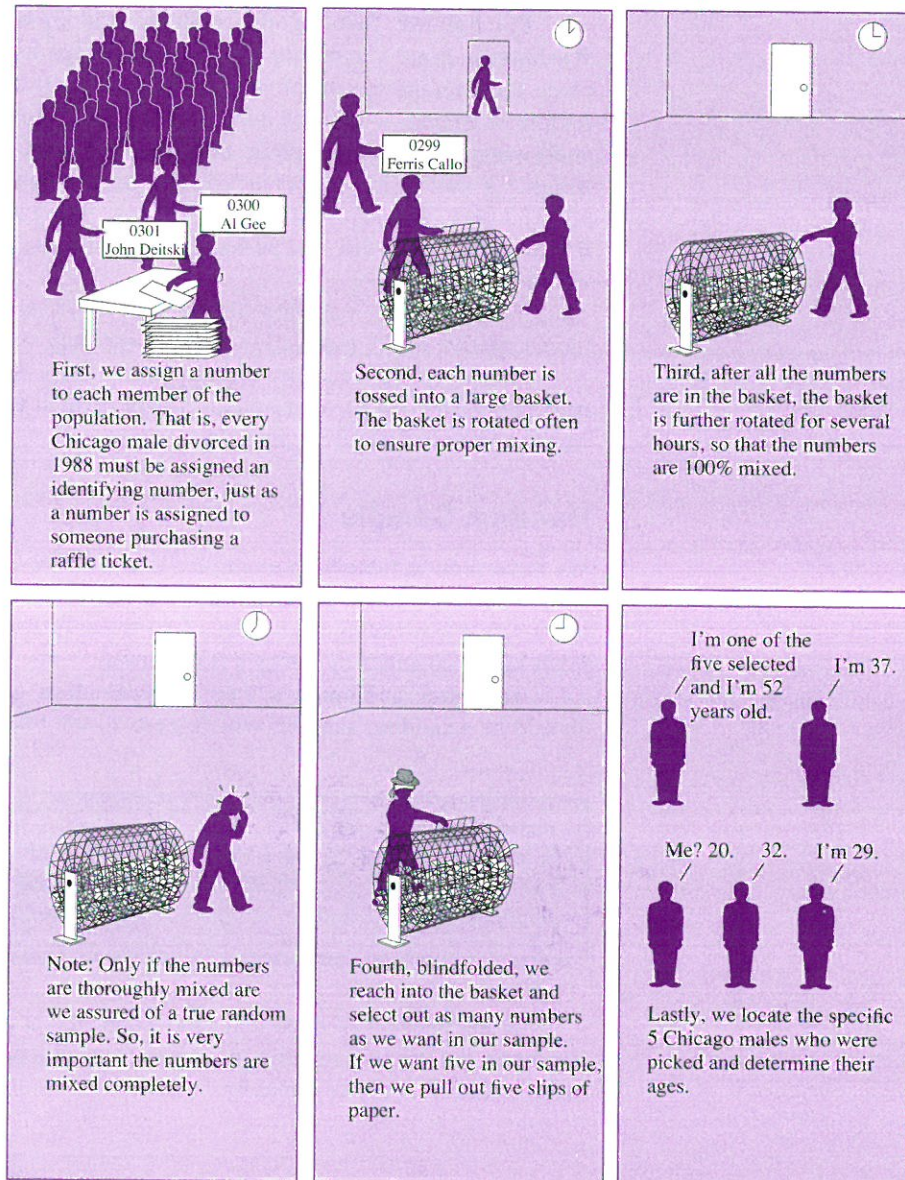
Random Sample

A sample drawn from a population such that each and every member of the population has an equal chance of being included in the sample. In addition, every sample of the same size has an equal chance of being selected.

Many decades of experience have shown that results from *random samples* can be trusted to give reliable information about a population. However, random samples are not easy to come by.

One old-fashioned method used to obtain a random sample and, perhaps, still the best method, is similar to a way lottery or raffle winners are selected. We will call it the **rotating basket** method, and it proceeds as follows:

Rotating Basket Method



one or more of them not telling the truth for any number of reasons? Did we ask to see their driver's licenses or birth certificates, then verify this information with hospital birth records? Put another way: how well-grounded, sound, and supportable are the five measurements of age we obtained?

External Validity

The certainty that our *methods and presence* in no way jeopardize our ability to use the random sample as a true representative of the population.

For instance, with a caged animal certain feeding habits can be accurately measured and may be repeatable from experiment to experiment (thus achieving internal validity). However, if we measure the feeding habits of those same animals in nature, the results could be quite different. In other words, does the caging of the animals or our presence introduce an influence (and thus threaten external validity, our ability to use the sample results as representative of the feeding habits of these animals in nature)?

To achieve external validity, then, we must ask ourselves not only whether we randomly sampled and achieved internal validity but also whether our test methods or presence in any way influenced our results. This is often a difficult question to answer, but the question must be addressed. Poor methodology has been known to drastically influence results and is often the reason why statistical studies measuring the same phenomenon vary so widely.

Internal and external validity in the case of the Chicago males divorced in 1988 hardly seems a question. Accurate ages are relatively easy to obtain and verify in this particular instance. And we can control threats to external validity rather easily by merely keeping the study secret. However, in most experiments in psychology, education, and a number of other fields, the question of validity turns into a veritable nightmare and, along with the difficulties of obtaining a random sample, often becomes a second or third follow-up course in statistics dealing mostly with these issues.

So, to sum up, not only must we extract honest, *accurate*, and reliable information about that which we are measuring (internal validity), but we must make sure our *methods and presence* in no way influence our ability to use the results as truly representative of the population (external validity).

All samples used in this text are assumed to be *internally and externally valid random samples*.

1.2 Why We Sample

We sample to determine certain characteristics of a population without having to measure the entire population. What are these characteristics, you might ask. Actually, there are many characteristics we might look for, but in broad terms we most often wish to identify either μ or p , defined as:

μ ($m\bar{u}$)

The arithmetic mean or average* value of a population.

p

The proportion (or percentage) of a population that possesses a certain attribute.

Greek letters are often used in mathematics to represent specific quantities. μ (pronounced $m\bar{u}$) is the twelfth letter of the Greek alphabet and is used to represent the average value of the population.

Sampling to Determine μ

One of the most frequent uses of sampling is to gather information about μ , the average value of some population, which is discussed at length in chapters 5 through 8. Let's look at an example.

Suppose we wish to determine μ , the average age of all Chicago males divorced in 1988. To measure the thousands of males in this population would be time-consuming and tedious. Instead, we might simply take a random sample as follows.†

Example — A random sample of $n = 5$ was drawn from a population of Chicago males divorced in 1988 and yielded the following ages: 52, 37, 20, 32, and 29. Calculate the average age of these five men.

Solution To take an average, we add up the ages and divide by 5.

$$\bar{x} = \frac{52 + 37 + 20 + 32 + 29}{5} = 34 \text{ years old}$$

*Technically, the word *average* refers to a broad number of measures, however, we shall employ the word *average* in its common usage as the sum of a collection of numbers divided by the number of values in that collection.

†Keep in mind, all samples used in this text are assumed to be both internally and externally valid random samples.

Notice that the symbol \bar{x} (x bar) was used to represent the average age of this sample. This allows us to differentiate between a sample average age (\bar{x}) and a population average age (μ).

Because we took a *random* sample and obtained an average age of 34 years old, we can now state μ should be “approximately equal to” 34 years old. In other words,

$$\bar{x} \approx \mu$$

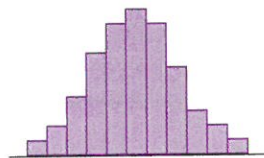
When we randomly sample, the sample average (\bar{x}) is approximately equal to the population average (provided certain restrictions are met regarding sample size, discussed below).

Does this allow us to state: the average age of all Chicago males divorced in 1988 is exactly 34 years old? The answer is, no. A *random* sample merely gives an approximation. However, it does allow us to state: the average age of all Chicago males divorced in 1988 is *approximately* 34 years old. ■

One word about sample size: In the above example, we used as our sample size $n = 5$, that is, we used 5 males selected from our population. It is generally preferable to keep your sample size as large as possible. In fact, as a general rule, random samples should be kept above 30.

It is preferable to keep your sample size at 30 or more.

In the above case, it would have been preferable to have selected 30 or more males. First, larger samples tend to give closer approximations of μ and, second, certain restrictions apply when your sample size is smaller than 30. Perhaps the most formidable restriction when using samples under 30 is that your population must be at least somewhat bell-shaped in appearance, as follows.



For samples under 30, populations must be at least somewhat bell-shaped* to ensure $\bar{x} \approx \mu$.

If you keep your sample size at 30 or above, this restriction does *not* apply. Samples above 30 tend to give sample averages (\bar{x} 's) that are close approximations of μ for almost *any shaped* population.

*This restriction grows more critical the smaller the sample size.

Sampling to Determine p

Another frequent use of sampling is to gather information about p , the proportion (or percentage) of some population that possesses a certain attribute, which is discussed at length in chapters 3, 4, 10, and 11.

Let's look at an example.

Suppose we wish to determine p , the proportion (or percentage) of Chicago males divorced in 1988, say for instance, who are fifty or more years old. Again, to measure the thousands of males in this population would be time-consuming and tedious. Instead, we might take a random sample as follows.

Example

A random sample of $n = 100$ males were drawn from a population of Chicago males divorced in 1988 and yielded 12 males who were fifty or more years old. Calculate the proportion of males who were fifty or more years old. Express this proportion as a percentage.

Solution

The proportion of our sample who are fifty or more years old is simply 12 out of 100 or

$$p_s = \frac{12}{100}$$

To convert this to a percentage, we multiply by 100, as follows:

$$\text{Percentage} = \text{Fraction} \times 100 = \frac{12}{100} \times 100 = 12$$

$$p_s = 12\%$$

In other words, because 12 out of 100 males were fifty or more years old, this is $\frac{12}{100}$ or 12% of the sample. Notice that the symbol p_s (p sub s) was used to represent the proportion of males fifty or more. This allows us to differentiate between a sample proportion (p_s) and a population proportion (p).

Since we took a *random* sample and obtained $p_s = 12\%$, the true proportion of Chicago males fifty or more years old must be *approximately* equal to 12%. In other words,

$$p_s \approx p$$

When we randomly sample, the sample proportion (p_s) is approximately equal to the population proportion (p).

Does this allow us to state that exactly 12% of *all* Chicago males divorced in 1988 were fifty or more years old? The answer is, no. Remember, random samples merely give us approximations. However, it does allow us to state: *approximately* 12% of all Chicago males divorced in 1988 were fifty or more years old. ■

Again, it is best to keep your sample size as large as possible. Generally, larger samples give closer approximations of p and there are other benefits, which will be discussed later in the text.

Summary

Definitions:

Statistics: The collection, organization, and interpretation of numerical data for the purposes of summarization (descriptive statistics) and drawing conclusions about populations based on taking samples from that population (inferential statistics).

Inference: Decision making based on samples drawn from populations.

Population: Any precisely defined collection of values we wish to study.

Sample: Part of a population.

Random sample: A sample drawn from a population such that each and every member of the population has an equal chance of being included in the sample. In addition, every sample of the same size has an equal chance of being selected.

Internal validity: The certainty that the observations in our sample group are accurate measures of the characteristics we set out to measure.

External validity: The certainty that our methods and presence in no way jeopardize our ability to use the random sample as a true representative of the population.

Valid random sample: A sample that is randomly selected and possesses both internal and external validity. In this text, we use only valid random samples, unless specified.

We can sample for a number of reasons, but in broad terms we most often wish to identify either μ ($m\bar{u}$) or p , defined as follows:

μ : the arithmetic mean or average value of a population.

p : the proportion (or percentage) of a population that possesses a certain attribute or characteristic.

$\bar{x} \approx \mu$: when we randomly sample, the sample average (\bar{x}) is approximately equal to the population average (μ). In this case, it is preferable to keep your sample size at 30 or more.

$p_s \approx p$: when we randomly sample, the sample proportion (p_s) is approximately equal to the population proportion (p). As a general rule, it is best to keep your sample size as large as possible.

Sampling techniques, as we will learn, normally come with restrictions and conditions, however when small sample sizes are used, additional conditions and restrictions usually apply. All this is discussed in the following chapters.

Exercises

Note that full answers for exercises 1–5 and abbreviated answers for odd-numbered exercises thereafter are provided in the Answer Key.

1.1 If we define statistics as the collection, organization, and interpretation of numerical data for the purposes of (a) summarization and (b) drawing conclusions about populations based on taking samples from that population, find a newspaper or magazine article or TV advertisement that uses statistics (the *New York Times* and *Newsweek* magazine are good sources). Bring in the article to discuss in class and state whether the data represents a population or sample.

1.2 Suppose we wished to estimate the average age of Chicago males divorced in 1988 and took a random sample of $n = 100$ yielding $\bar{x} = 35.2$ years old.

- What is the population?
- What symbol do we use to represent the average age of the population?
- How many were in your sample?
- What symbol is used to represent the average age of the sample?
- Do we know the average age of the population?

1.3 Suppose we wish to take a random sample of $n = 7$ from a population of 5000 Chicago males divorced in 1988. First we assign each male a number from 0001 to 5000. Explain how we might use the following random digit table to select our sample:

```

94620 27963 96478 21559 19246 88097 44926
60947 60775 73181 43264 56895 04232 59604
27499 53523 63110 57106 20865 91683 80688
01603 23156 89223 43429 95353 44662 59433
00815 01552 06392 31437 70385 45863 75971

83844 90942 74857 52419 68723 47830 63010
06626 10042 93629 37609 57215 08409 81906
56760 63348 24949 11859 29793 37457 59377
64416 29934 00755 09418 14230 62887 92683
63569 17906 38076 32135 19096 96970 75917

```

1.4 Suppose we wished to estimate the average age of Chicago males divorced in 1988 and took a random sample of $n = 7$ yielding the ages: 26, 61, 39, 43, 31, 22, and 37.

- Calculate \bar{x} , the average age of the sample.
- What can we say about μ , the average age of all Chicago males divorced in 1988?

1.5 Suppose we wished to estimate the proportion of Chicago males divorced in 1988 that possess the attribute of *blue eyes* and took a random sample of $n = 100$, discovering that 18 males from this sample had blue eyes.

- Calculate p_s , the proportion of your sample that had blue eyes.
- Express p_s as a percentage.
- What can we say about p , the proportion of all Chicago males divorced in 1988 who have blue eyes?

1.6 Briefly state why we study statistics.

1.7 In a medical study, a researcher wished to estimate the average length of time needed for a particular nurse-in-training to draw a series of blood specimens. A sample of the nurse's work over several months yielded the following times: 10, 6, 5, 14, 6, and 13 (in minutes).

- What is the population?
- What is it about the population we wish to determine? And what symbol is used to represent this?
- What is the sample? Calculate \bar{x} .
- Is this sample representative of the population? Discuss:
 - randomness.
 - internal validity.
 - external validity.
 - sample size.

1.8 It is often said prior to an election that only a small number of people are needed to predict the outcome of a major election. The statement refers to use of a valid random sample. If we wish to predict the results of an upcoming election, discuss selecting a sample by:

- polling the class.
- having your classmates randomly poll friends.
- selecting 400 names from a telephone book and calling.

1.9 To determine the average height of Washington State University males, a campus newsletter sampled 50 males yielding an average height of $5'9\frac{1}{2}''$.

- What is the population?
- What is the sample?
- How might you obtain a valid random sample?