Unit I: Descriptive Statistics

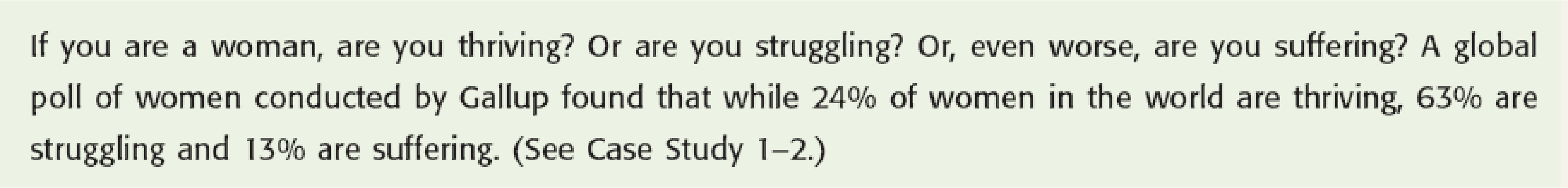
Essential Statistics for Data Science

STATISTICS

The science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions

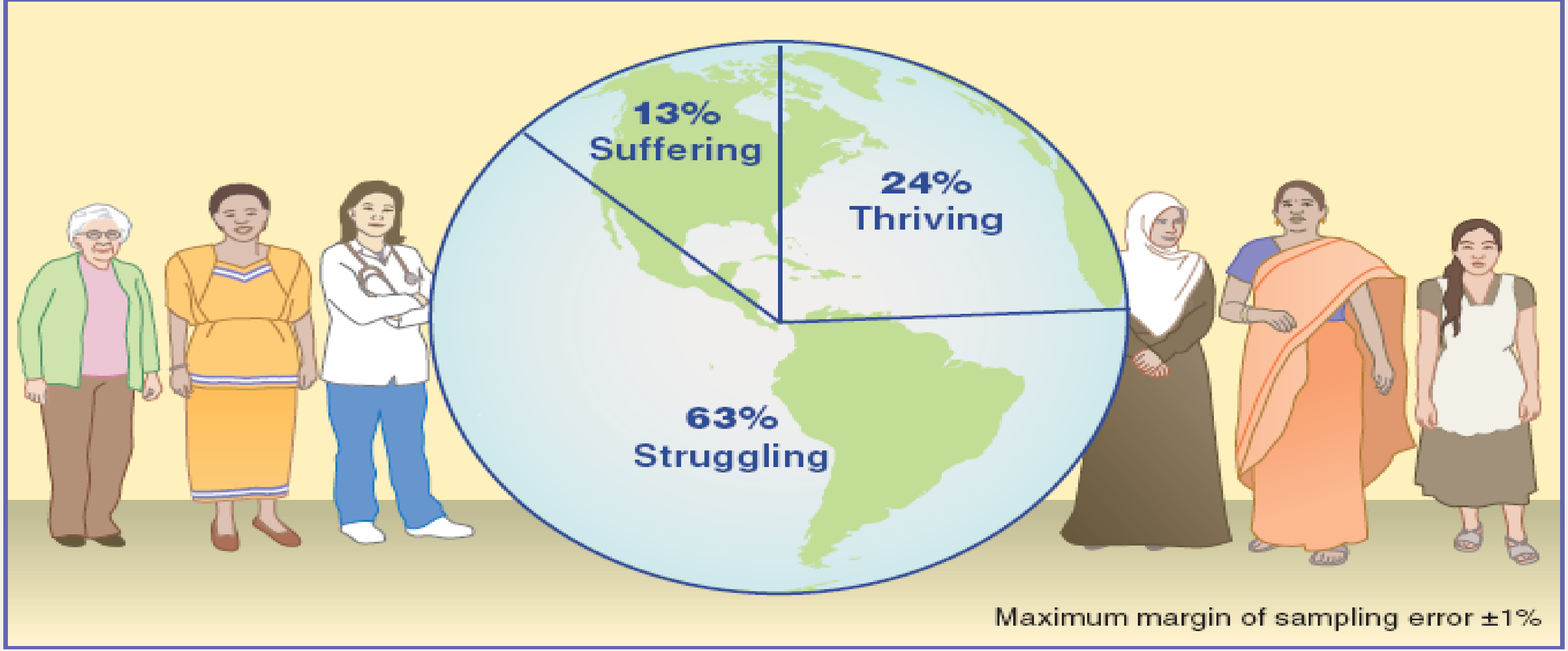
Statistical analysis – used to manipulate summarize, and investigate data, so that useful decision-making information results.

INTRODUCTION



If you are a woman, are you thriving? Or are you struggling? Or, even worse, are you suffering? A global poll of women conducted by Gallup found that while 24% of women in the world are thriving, 63% are struggling and 13% are suffering. (See Case Study 1–2.)

HOW WOMEN RATE THEIR LIVES



Maximum margin of sampling error ±1%

Data source: Gallup poll of adult women aged 15 and older conducted during 2011 in 147 countries and areas.

WHY STATISTICS MATTERS

Statistical procedures are used to advance our knowledge of ourselves and the world around us.

groups of people, gathering scores, and analyzing

Statistics helps to strengthen your critical thinking skills and reasoning abilities

The Sun Will Explode in Less Than 6 Years (9/19/2002) - to evaluate the validity of such assertions to see if they stand up to scientific scrutiny

Statistics enables you to understand research results in the professional journals of your ar ea of specialization

knowledgeable and competent professional in your area of study.

STATISTICS AND OTHER STATISTICAL TERMS

Research, a systematic inquiry in search of knowledge, involves the use of statistics.

Set of scores, referred to as data.

Before the scores have undergone any type of statistical transformation or analysis, they are called raw scores.

A population consists of all elements – individuals, items, or objects – whose characteristics are being studied. The population that is being studied is also called the target population.

A portion of the population selected for study is referred to as a sample.

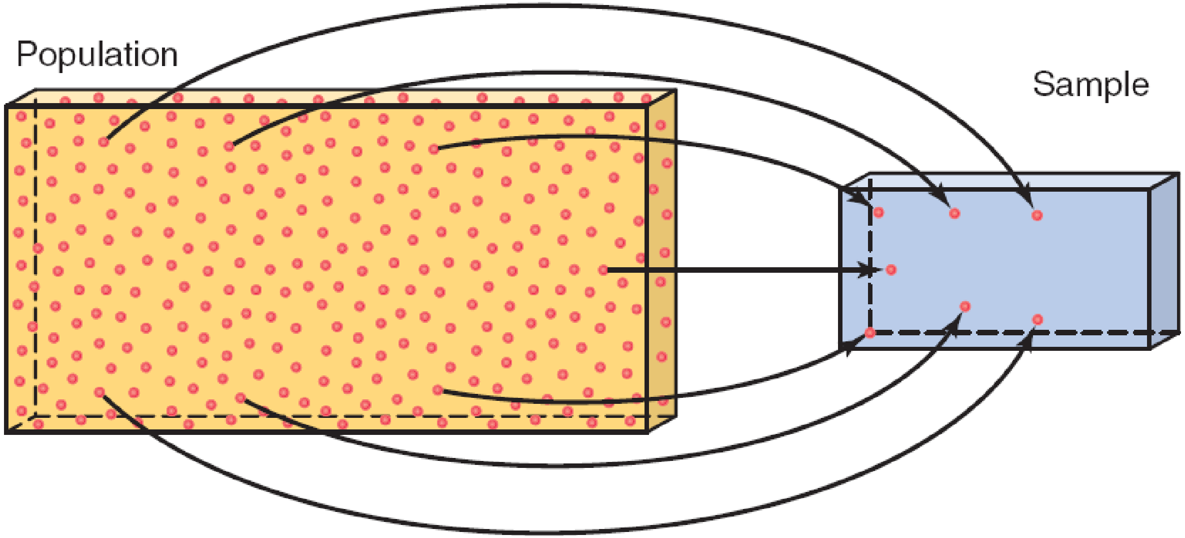
Usually populations are so large that a researcher cannot examine the entire group.

Therefore, a **sample** is selected to represent the population in a research study.

The goal is to use the results obtained from the sample to help answer questions about the population.

A sample drawn in such a way that each element of the population has a chance of being selected is called a **random sample.**

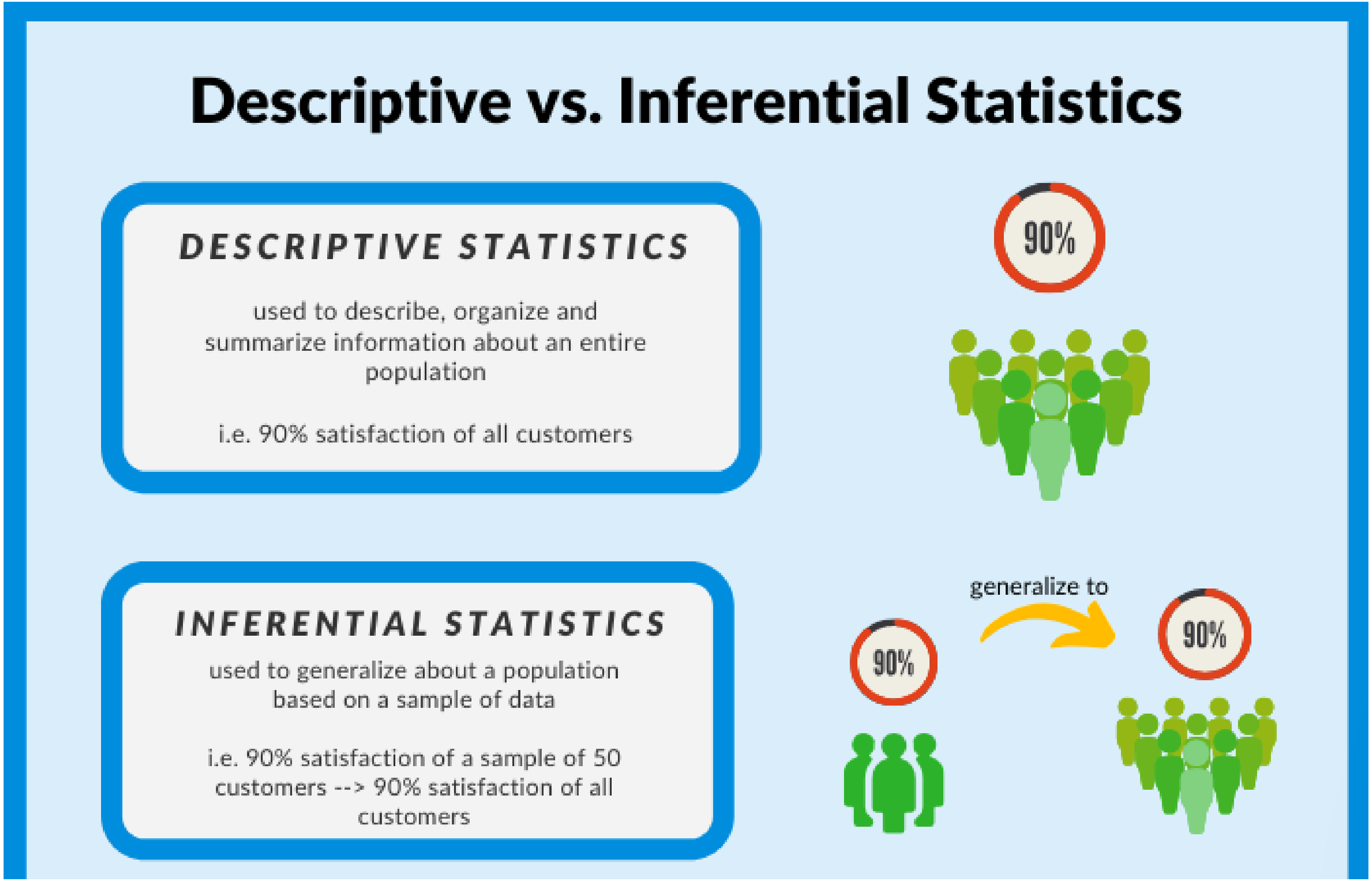
Samples are normally drawn from only the portion of the population that is accessible, referred to as the **sampling frame**


Population
Sample

Descriptive statistics sum up and condense a set of raw scores so that overall trends in the data become apparent.

Percentages and averages are examples of descriptive statistics.

Inferential statistics involve predicting characteristics of a population based on data obtained from a sample.



Descriptive vs. Inferential Statistics

DESCRIPTIVE STATISTICS

used to describe, organize and summarize information about an entire population

i.e. 90% satisfaction of all customers

INFERENTIAL STATISTICS

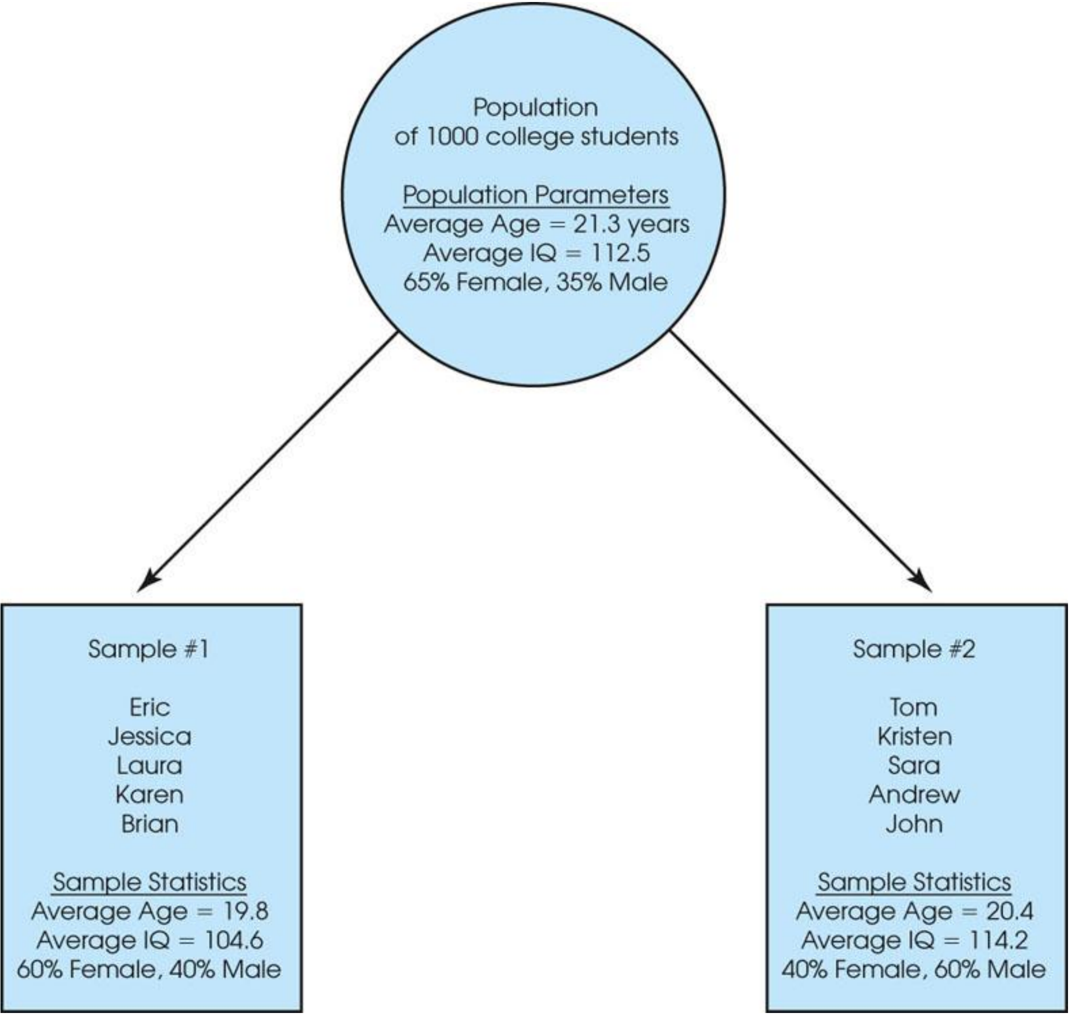
used to generalize about a population based on a sample of data

i.e. 90% satisfaction of a sample of 50 customers --> 90% satisfaction of all customers

90%
90%
generalize to
90%

A **statistic is a numerical value that originates from a sample, whereas a parameter is a numerical value that represents an entire population.**

The discrepancy between a sample statistic and its population parameter is called **sampling error**.



Population
of 1000 college students

Population Parameters
Average Age = 21.3 years
Average IQ = 112.5
65% Female, 35% Male

Sample #1

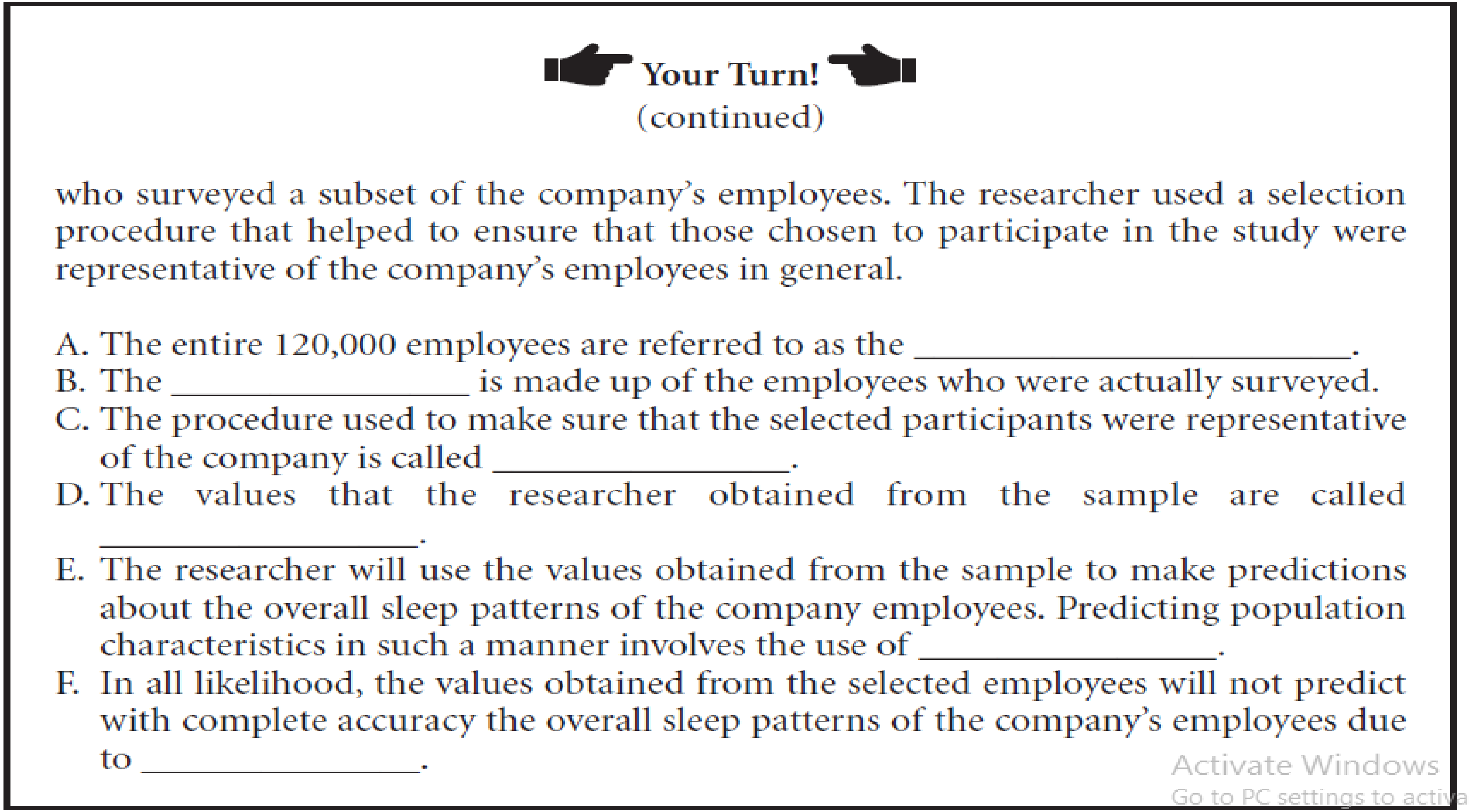
Eric
Jessica
Laura
Karen
Brian

Sample Statistics
Average Age = 19.8
Average IQ = 104.6
60% Female, 40% Male

Sample #2

Tom
Kristen
Sara
Andrew
John

Sample Statistics
Average Age = 20.4
Average IQ = 114.2
40% Female, 60% Male



Your Turn!

(continued)

who surveyed a subset of the company's employees. The researcher used a selection procedure that helped to ensure that those chosen to participate in the study were representative of the company's employees in general.

A. The entire 120,000 employees are referred to as the ______________________.

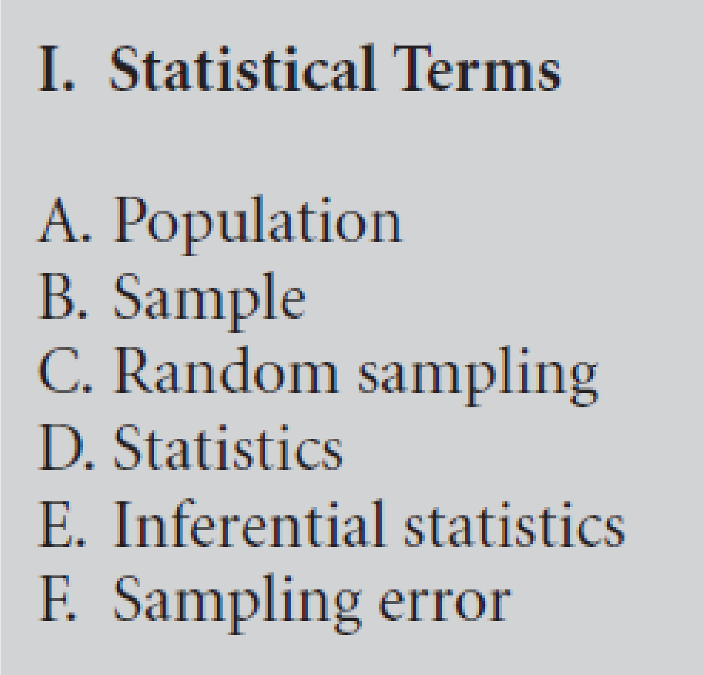
B. The ______________ is made up of the employees who were actually surveyed.

C. The procedure used to make sure that the selected participants were representative of the company is called ______________.

D. The values that the researcher obtained from the sample are called ______________.

E. The researcher will use the values obtained from the sample to make predictions about the overall sleep patterns of the company employees. Predicting population characteristics in such a manner involves the use of ______________.

F. In all likelihood, the values obtained from the selected employees will not predict with complete accuracy the overall sleep patterns of the company's employees due to ______________.



I. Statistical Terms

A. Population
B. Sample
C. Random sampling
D. Statistics
E. Inferential statistics
F. Sampling error

MEASUREMENT

A variable is anything that varies or that can be present in more than one form or amount.
Variables describe differences.

Qualitative variables differ in kind rather than amount
– such as eye color, gender.

Quantitative variables differ in amount –
-such as scores on a test, annual incomes.

Discrete variables cannot be divided or split into intermediate values, but rather can be measured only in whole numbers.

21 students

Continuous variables , on the other hand, can be broken down into fractions or smaller units.

A newborn baby can weigh 7 pounds, 7.4 pounds

LIMITS

Intermediate values bounded by what is referred to as ***real limits .***

1. Identify the unit of measurement. If the value reported is a whole number, the unit of measurement is 1.
2. Using a calculator, divide the unit of measurement in half.
3. For lower limits (LL), subtract the value obtained in Step 2 from the reported value.
4. For upper limits (UL), add the value obtained in Step 2 to the reported value.

15 is a whole number, rather than a fraction or decimal.

So the unit of measurement is 1.

Half of 1 is .5.

Therefore, Lower Limit of 15 is → 15 − .5 = 14.5

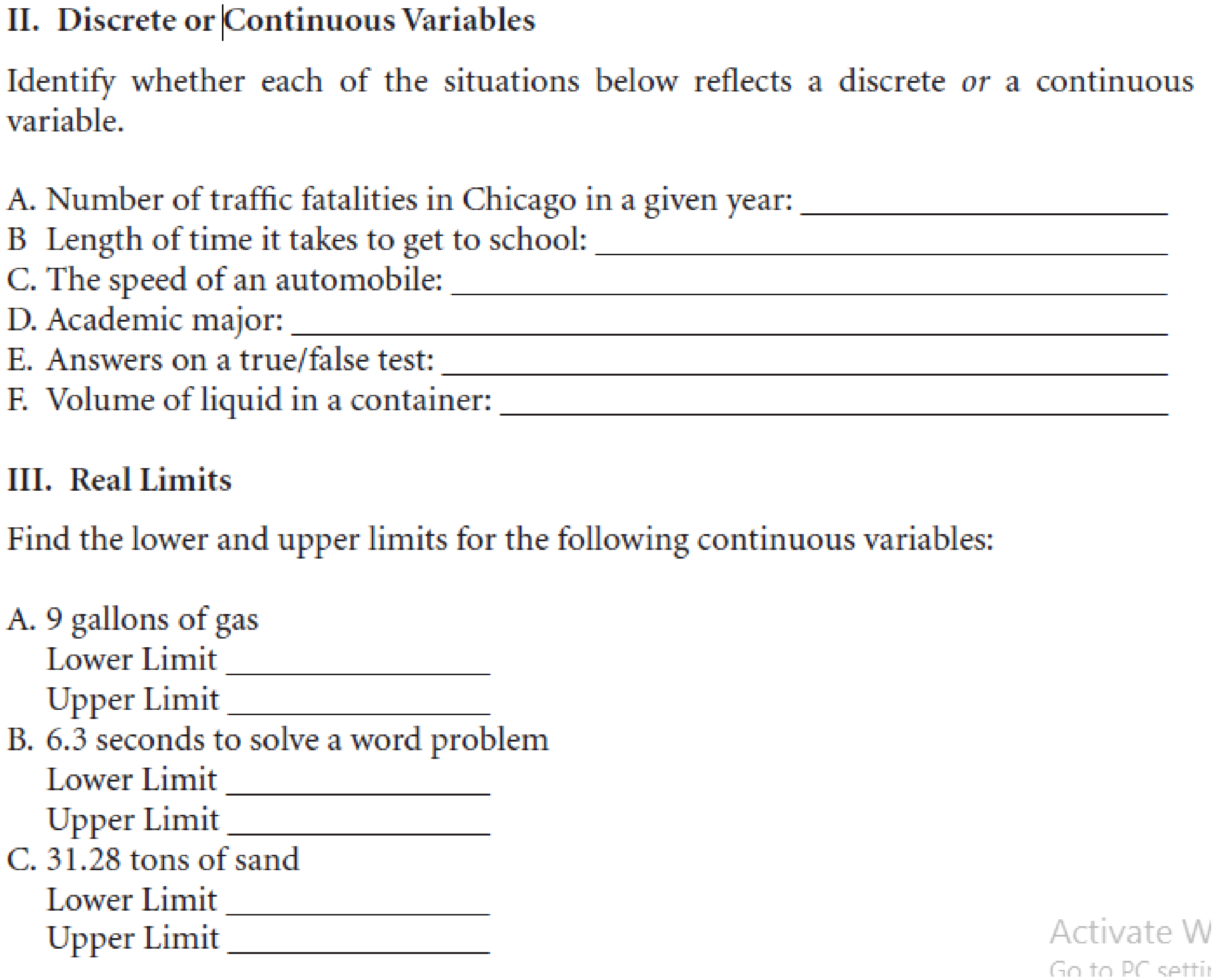
Upper Limit of 15 is → 15 + .5 = 15.5

6.8

1/10 = .1   Half of .1 = .05

Lower Limit of 6.8 is → 6.8 − .05 = 6.75

Upper Limit of 6.8 is → 6.8 + .05 = 6.85



II. Discrete or Continuous Variables

Identify whether each of the situations below reflects a discrete *or* a continuous variable.

A. Number of traffic fatalities in Chicago in a given year: ______________________

B Length of time it takes to get to school: ______________________

C. The speed of an automobile: ______________________

D. Academic major: ______________________

E. Answers on a true/false test: ______________________

F. Volume of liquid in a container: ______________________

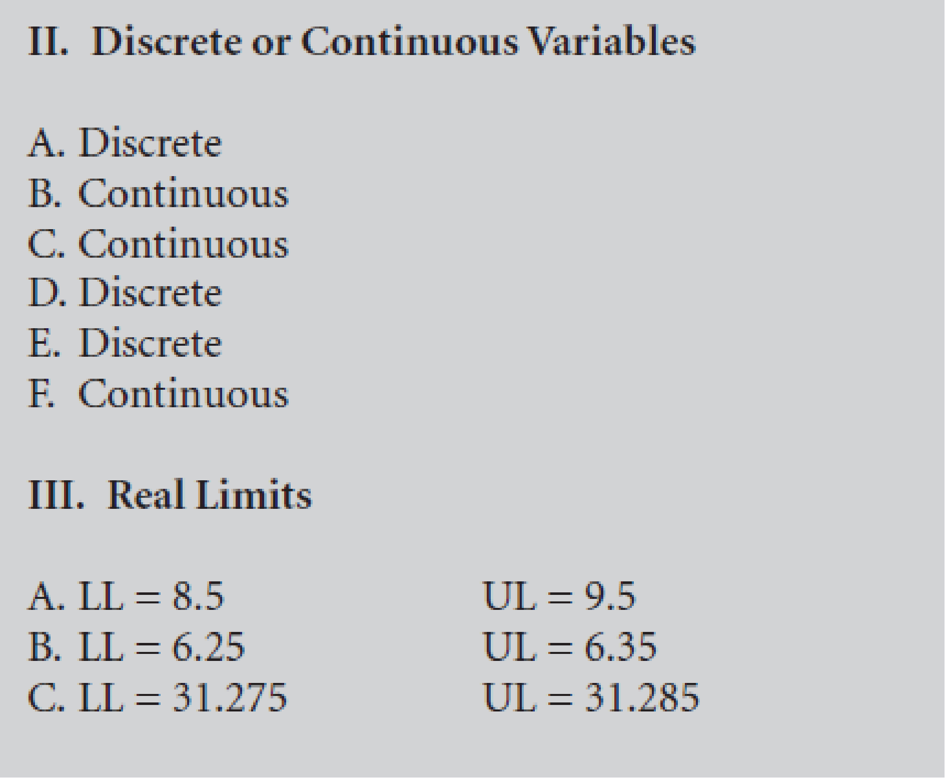
III. Real Limits

Find the lower and upper limits for the following continuous variables:

A. 9 gallons of gas
  Lower Limit ________________
  Upper Limit ________________

B. 6.3 seconds to solve a word problem
  Lower Limit ________________
  Upper Limit ________________

C. 31.28 tons of sand
  Lower Limit ________________
  Upper Limit ________________



II. Discrete or Continuous Variables

A. Discrete
B. Continuous
C. Continuous
D. Discrete
E. Discrete
F. Continuous

III. Real Limits

A. LL = 8.5	UL = 9.5
B. LL = 6.25	UL = 6.35
C. LL = 31.275	UL = 31.285

Scales of Measurement

Nominal Scale.

The least-specific measurement scale is the nominal scale, which simply c lassifies observations into different categories.

Ex: Religion, types of trees, and colors.

These variables have no quantitative value

Ordinal Scale:

In addition to classifying observations into different categories, this scale also permits *ordering, or ranking, of the observations .*

Ex: horse race with the horses arriving at the finish line in different amounts of time so that there will be first-, second-, and third-place winners

Ordinal scales do not indicate how much difference exists between observations.

Interval Scales.

*With **interval scales, on the other hand, there is equal distance between units on the scale.***

Ex: Temperature in degrees Fahrenheit is measured on an interval scale.

The difference between 10°F and 30°F is the same as the difference between 40 °F and 60°F (20°F in each case).

Ratio Scale.

*On **ratio scales, the real, absolute-zero point is known.***

Experimentation

Researchers want to discover relationships between variables.

This is often accomplished through experimentation.

The manipulated variable is called the independent variable.

The variable that is measured to see if it has been affected by the independent variable is called the dependent variable.

NONEXPERIMENTAL RESEARCH

Non-experimental research in which variables that already exist in different values are passively observed and analyzed rather than being actively manipulated.

Correlation research involves using statistical procedures to analyze the degree o f relationship between two variables.

MATH REVIEW

Rounding : generally rounded to two decimal places (to the nearest hundredth)

9.34782 rounds to 9.35

123.39421 drops to 123.39

74.99603 rounds to 75 or 75.00

Proportions and Percentages

A proportion is a part of a whole number that can be expressed as a fraction or as a decimal.

In a class of 40 students, six earned As : fraction (6/40) or as a decimal (.15).

To change a decimal (proportion) to a percentage : .15 x 100 = 15%

Signed Numbers

Addition.

$(-6) + (-8) + (10) + (-7) + (5) + (-1) = 15 + (-22) = -7$

Subtraction

$(-14) - (-5) = -14 + 5 = -9$

Multiplication.

$-12 \times (-3) = 36$

Division.

$6 \div (-3) = -2$

Order of Operations

Parentheses, Exponents, Multiplication, Division, Addition, Subtraction

$$\begin{aligned} &7 \times [4 - (3 \times 6)] \\ &= 7 \times (4 - 18) \\ &= 7 \times (-14) \\ &= -98 \end{aligned}$$

$$\begin{aligned} &-3 + [(2 \times 4) - (7 \times 7)] + 8 \\ &= -3 + (8 - 49) + 8 \\ &= -3 + (-41) + 8 \\ &= -36 \end{aligned}$$

Suppose you have the following *X* and *Y* scores:

X	Y	X^2	XY	$(X - 3)$	$(Y + 2)$	$(Y + 2)^2$
6	4	36	24	3	6	36
5	7	25	35	2	9	81
9	2	81	18	6	4	16
8	1	64	8	5	3	9
28	14	206	85	16	22	142

$\Sigma X = 28$ Simply add all of the *X* values.

$(\Sigma X)^2 = 784$ Remember, parentheses first. Sum the *X* values. Then, square the sum of the *X* values.

$\Sigma X^2 = 206$ Here, exponents come first. Square each *X* value first, then find the sum of the X^2 column.

Suppose you have the following *X* and *Y* scores:

X	Y	X^2	XY	$(X - 3)$	$(Y + 2)$	$(Y + 2)^2$
6	4	36	24	3	6	36
5	7	25	35	2	9	81
9	2	81	18	6	4	16
8	1	64	8	5	3	9
28	14	206	85	16	22	142

$\Sigma XY = 85$ Multiplication comes before addition. Thus, multiply each *X* value by each *Y* value (XY), then add the column for the XY values.

$\Sigma(X - 3) = 16$ Subtract 3 from each *X* value, then add the $(X - 3)$ column.

$\Sigma(Y + 2)^2 = 142$ Parentheses first, then exponents, then addition. Thus, add 2 to each *Y* value, then square the $(Y + 2)$ values, and finally sum the $(Y + 2)^2$ column.