

# STATISTICS IS THE STUDY OF DATA

## Things we do with data

Gather data  
Organize  
Summarize  
Analyze  
Interpret

## Population vs. Sample

**Parameter:** A number that is a characteristic of a population

**Statistic:** A number that is a characteristic of a sample

Some Parameters and Statistics that we often use in Math 10 are AVERAGES or PROPORTIONS.  
(Some other parameters are median, standard deviation, variance; we will learn about these in chapter 2)

### Examples of Parameters:

*PROPORTION:* In a recent quarter, 39% of all De Anza College students were over age 25  
*AVERAGE:* In a recent quarter, the average age of all De Anza College students was 27.1 years  
*MEDIAN:* In a recent quarter, the median age of all De Anza College students was 22 years

### Examples of Statistics:

*PROPORTION:* 41% of a random sample of 200 students were over age 25.  
*AVERAGE:* The average age of a random sample of 200 De Anza College students was 28.1 years.

**Variable:** The variable is the characteristic of interest

**Examples of Variable:** "the age of a (one, individual) De Anza College student"  
"the distance that a (one, individual) student commutes to De Anza College "  
"the number of quarters that a (one, individual) student has attended De Anza College"

When asked to identify the variable, you are being asked to DESCRIBE the characteristic of interest. Your answer for a variable will be a sentence, not a number.

*Think of the variable as the question you are asking in order to obtain information.*

**Data:** The data are the information collected about the variable for individuals in the population or sample.

**Examples of Data:** 2.5 miles, 8.4 miles, 0.25 miles, 52 miles, . . . (commute distances)

**Examples of Data:** 2, 2, 5, 8, 3, 1, 7, 5, 5, 4, 6, 3, 3, 1, 2, . . . (number of quarters at De Anza)

*If you think of the variable as the question you are asking in order to obtain information, the data are the answers to the question.*

## Types of Variables and Data:

### Quantitative

Discrete

Continuous

**Qualitative** (also called Categorical):

**Statistical Methods:**

•**Descriptive statistics:**

•**Inferential statistics:**

**STATISTICS: INTERPRETING VOCABULARY**

**Question 1:** Suppose 1500 randomly selected registered voters in a large city are asked the following questions:

What is your age?      What is your annual income?      Do you intend to vote?

If you intend to vote, do you intend to vote in person or by absentee ballot?

A newspaper article discussing expected voter turnout reported that for this sample, the average annual income is \$61,000, the average age is 44, 63% of the registered voters sampled intend to vote, 37% intend to vote by absentee ballot while 63% intend to vote in person at the polling place.

- a. Describe the sample.
- b. What would be the appropriate target population for this sample?
- c. What are the variables? For each variable, is it quantitative or qualitative?
- d. Make up two examples of data that might have been given by somebody in the sample.
- e. Are the values reported in the newspaper statistics or parameters? Explain.

**Question 2:** A weight loss clinic is studying the exercise habits of its clients.

A random survey of 100 clients shows that 36% of clients exercised regularly before enrolling in the clinic's weight loss program and 71% of clients exercise regularly after enrolling in the clinic's weight loss program. Of those clients in the sample who exercise regularly, the average amount of time that clients exercise each week is 2.5 hours, and clients exercise on average 4 times a week.

List two qualitative variables:

List a quantitative discrete variable:

List a quantitative continuous variable:

## TYPES OF STATISTICAL STUDIES:

**Census:**

**Sampling:**

**Observational study:**

**Experiment:**

Vocabulary for Experiments:

Treatment

Response

Control Group

Placebo

Blinding

Double Blind

## CRITICAL EVALUATION OF STATISTICAL STUDIES AND RESULTS

### Common Problems in Statistics to beware of

- Problems with Samples: A sample should be representative of the population.  
A sample that is not representative of the population is biased.
- Self-Selected Samples
- Sample Size Issues
- Collecting data or asking questions in a way that influences the response
- Non-response or refusal of subject to participate
- Causality: a relationship between two variables does not necessarily imply that one **causes** the other to occur; they may both be related to some other variable.
- Self-Funded or Self-Interest Studies
- Misleading Use of Data: improperly displayed graphs, incomplete data, lack of context
- Confounding: when the effects of multiple factors on a response can not be separated.  
it becomes difficult or impossible to draw valid conclusions about the effect of each factor.

## TYPES OF STATISTICAL STUDIES / CRITICAL EVALUATION

### Question 3:

**Study I:** Two algebra classes on similar schedules taught by the same instructor were given different types of tutoring resources during the quarter to determine whether the tutoring resources affected student outcome.

**Study II:** Researchers are studying whether retirement age affects the rate of memory problems in senior citizens. A survey of retired senior citizens showed that those who retired earlier tended to have a higher incidence of memory problems after retirement than those who retired at an older age.

**Study III:** 300 randomly selected individuals are asked if they had been on a diet in the last 8 weeks and how much their weight has changed over the last 8 weeks. Weight change for dieters and non-dieters are compared.

**Study IV:** 100 individuals are put on a low fat diet, 100 on a low carb diet and 100 eat their normal diet. Their weight change over an 8 week period is recorded.

a. For each of the above, what type of study is it?

I: \_\_\_\_\_

II: \_\_\_\_\_

III: \_\_\_\_\_

IV: \_\_\_\_\_

b. What problem can you see in Study II?

c. Which weight loss study (III or IV) do you think would give the best information about the effect of diet on weight loss? Why?

### Question 4:

A large city is proposing a parcel tax to support education. Each property owner would be assessed a tax of \$100 per property per year. The parcel tax will be voted on by voters in the next election. It will pass if  $\frac{2}{3}$  of the voters vote in favor of the tax.

- I. A group of parents and teachers supporting the parcel tax randomly select and call residents in the city. They identify themselves as members of the Parent Teachers Association for the school system and ask the person who answers the telephone call if they support the parcel tax.
- II. A TV news station in the city conducts a call-in survey. Viewers are asked whether they favor or oppose the tax. Viewers are asked to dial a toll free 800 number to record their votes. The poll is publicized and responses are solicited by announcements on the TV station's evening news programs.
- III. A professional polling organization conducts a survey by randomly calling selected residents in the city. If the resident is a registered voter, he or she is asked his/her their opinion about the proposed parcel tax. They are asked whether they favor the tax, oppose the tax, or have no opinion. These three choices are presented to the individual in random order, so that not all respondents hear the choices in the same order.

a. Which survey do you think would produce the most accurate prediction of the election results?

b. For each of the other two surveys, what problems do you think there might be with the information obtained? Explain your reasoning for your answers to these questions.

## TYPES OF SAMPLES:

A sample is a part of or a subset of a population.

A sample should be representative of the population

A sample that is not representative of the population is biased.

### Vocabulary and Concepts:

**Sampling Error:** Random error obtained by using part of the population to represent the whole population

**Non-Sampling Error:** Non-random error: improper data collection recording or sampling techniques, bias,

**Random Sample:** Every person has equal chance of being included in the sample

#### **Sampling Methods:**

- **simple random sample**
  
- **systematic sample**
  
- **stratified sample**
  
- **cluster sample**
  
- **convenience sample**

**Example:** (1 – 6 based on example 1.6 in Collaborative Statistics, chapter 1, by B. Illowsky & S. Dean [www.cnx.org](http://www.cnx.org))

Determine the type of sampling method used:

1. To form a recreational soccer team, a soccer coach randomly selects 6 players from a group of boys ages 8 to 10, 7 players from a group of boys ages 11 to 12, and 3 players from a group of boys age 13 to 14. \_\_\_\_\_
2. For a survey of human resource (HR) personnel at high tech companies, a pollster interviews all HR personnel in each of 5 randomly selected high tech companies. \_\_\_\_\_
3. In a survey of engineering salaries, a researcher selects engineers to interview by randomly selecting 50 women engineers and randomly selecting 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 selects 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 average. \_\_\_\_\_
7. In a study to learn what types of after school child care are used in their district, a school district administrator randomly selects 6 classes at each school and surveys all parents with children in the selected classes. \_\_\_\_\_

## REPRESENTATIONS OF CATEGORICAL (QUALITATIVE) DATA

### Qualitative (categorical) data can be organized and summarized using tables and graphs

**Tables:** showing counts (frequencies) and percentages or proportions (relative frequencies)

**Pie Charts:** categories of data are represented by wedges in the circle proportional in size to the percent of individuals in each category

**Bar Graphs:** The length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal

**Pareto Charts:** Bars are sorted into order by category size (largest to smallest)

*It is helpful to look at a variety of charts to decide which best displays the data.*

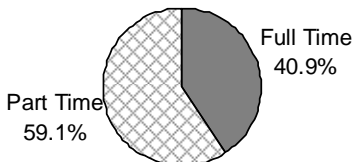
De Anza College		
	Number	Percent
Full-time	9,200	40.9%
Part-time	13,296	59.1%
Total	22,496	100%

Spring 2010

Foothill College		
	Number	Percent
Full-time	4,059	28.6%
Part-time	10,124	71.4%
Total	14,183	100%

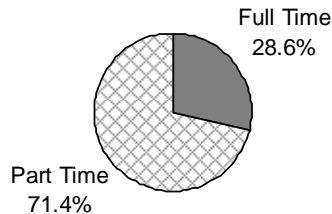
De Anza College

■ Full Time □ Part Time



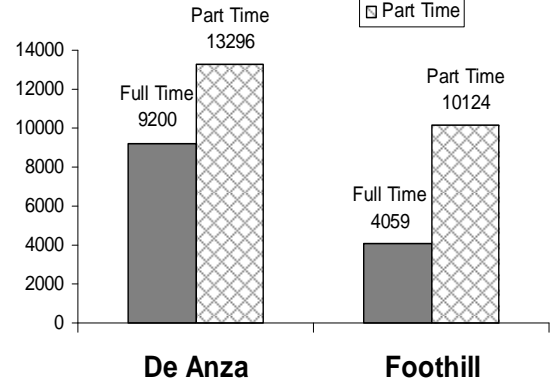
Foothill College

■ Full Time □ Part Time



Student Status

■ Full Time □ Part Time



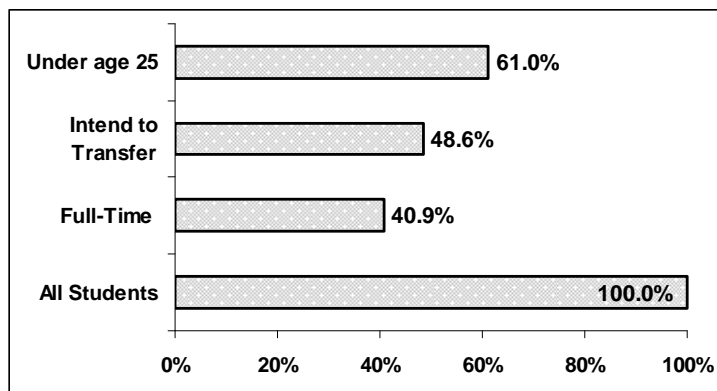
### Percentages that add to more than 100%:

*Bar chart is appropriate to compare relative size of categories*

*Pie chart can not be used.*

De Anza College Spring 2010

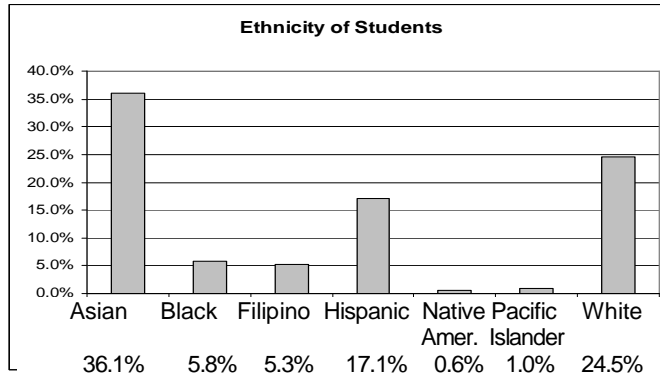
Characteristic/Category	Percent
Full-Time Students	40.9%
Students who intend to transfer to a 4 year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%



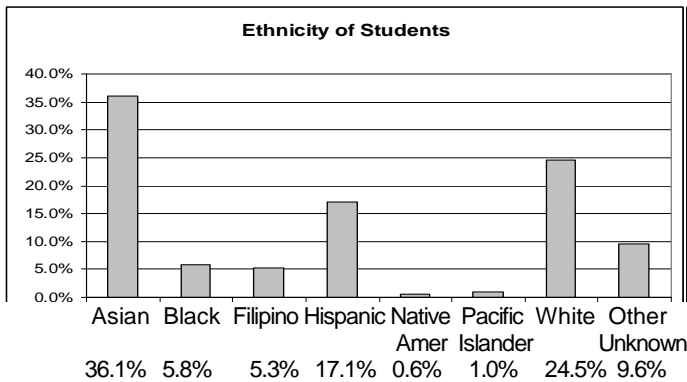
**Missing data: Ethnicity of Students De Anza College Fall Term 2007 (Census Day)**

	Frequency	Percent
Asian	8,794	36.1%
Black	1412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

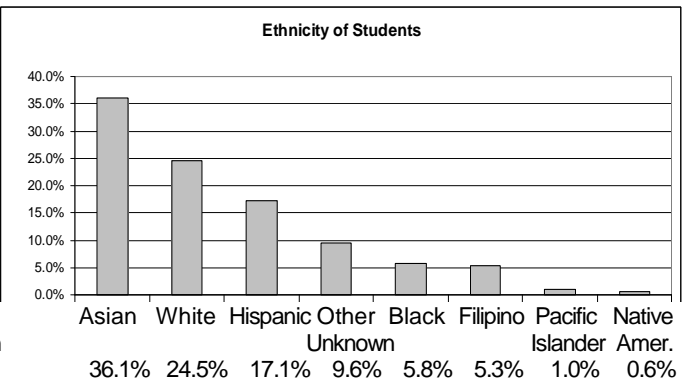
**Bar Chart with Missing Data**



**Bar chart with Other/Unknown category**

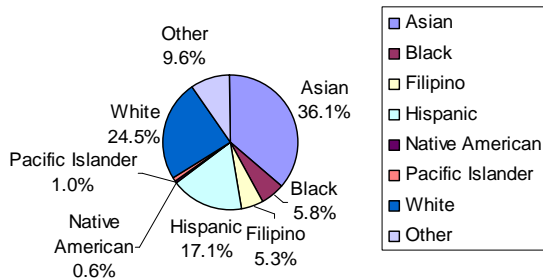


**Pareto Chart: Bars sorted by size**

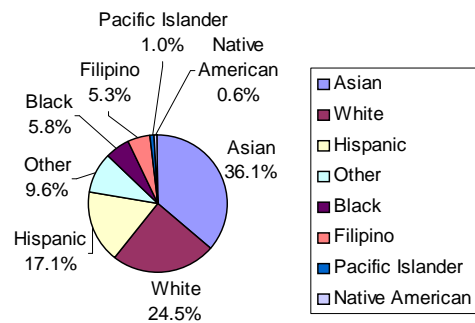


**Pie Charts:** can use only if there is not missing data; both pie charts include the Other/Unknown category

**Ethnicity of Students**



**Ethnicity of Students**



Pie Chart is easier to understand visually when categories are sorted into order by size.