Chapter 1:
Defining and Collecting Data

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Motivation

- BlackSea Corp is intended to offer dental care benefit to its employees
- A group at the HR is assigned to conduct a survey to estimate the dental care cost of the employees
- **DCOVA:**
  - **DEFINE** the variables that are useful in identifying the above value
  - **COLLECT** the data by surveying a randomly selected sample
  - **ORGANIZE** the collected data
  - **VISUALIZE** them to understand the underlying patterns better
  - **ANALIZE** the organized data to identify the above value

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Variables

- **REMINDER:** Variables represent a characteristic; e.g. ZIPCODE

- Broadly speaking, there are two types of characters:
  - **Categorical** – ***qualitative*** description of variations

    Example: MALE and FEMALE
  - **Numerical** – ***quantitative*** description of variations

    Example: 1, 2, 3, 4, and 5

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Variables

- **Categorical** variables have values that can only be placed into categories.

Example:

NAME	CITY	High School Education	College Education
Helen Maxwell	OKC	YES	YES
Pete Gregory	NORMAN	YES	NO
Louise Medina	NORMAN	NO	NO
Phillip Cox	MOORE	YES	NO
Marie Henry	OKC	YES	YES
Taylor Reeves	EDMUND	NO	NO

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

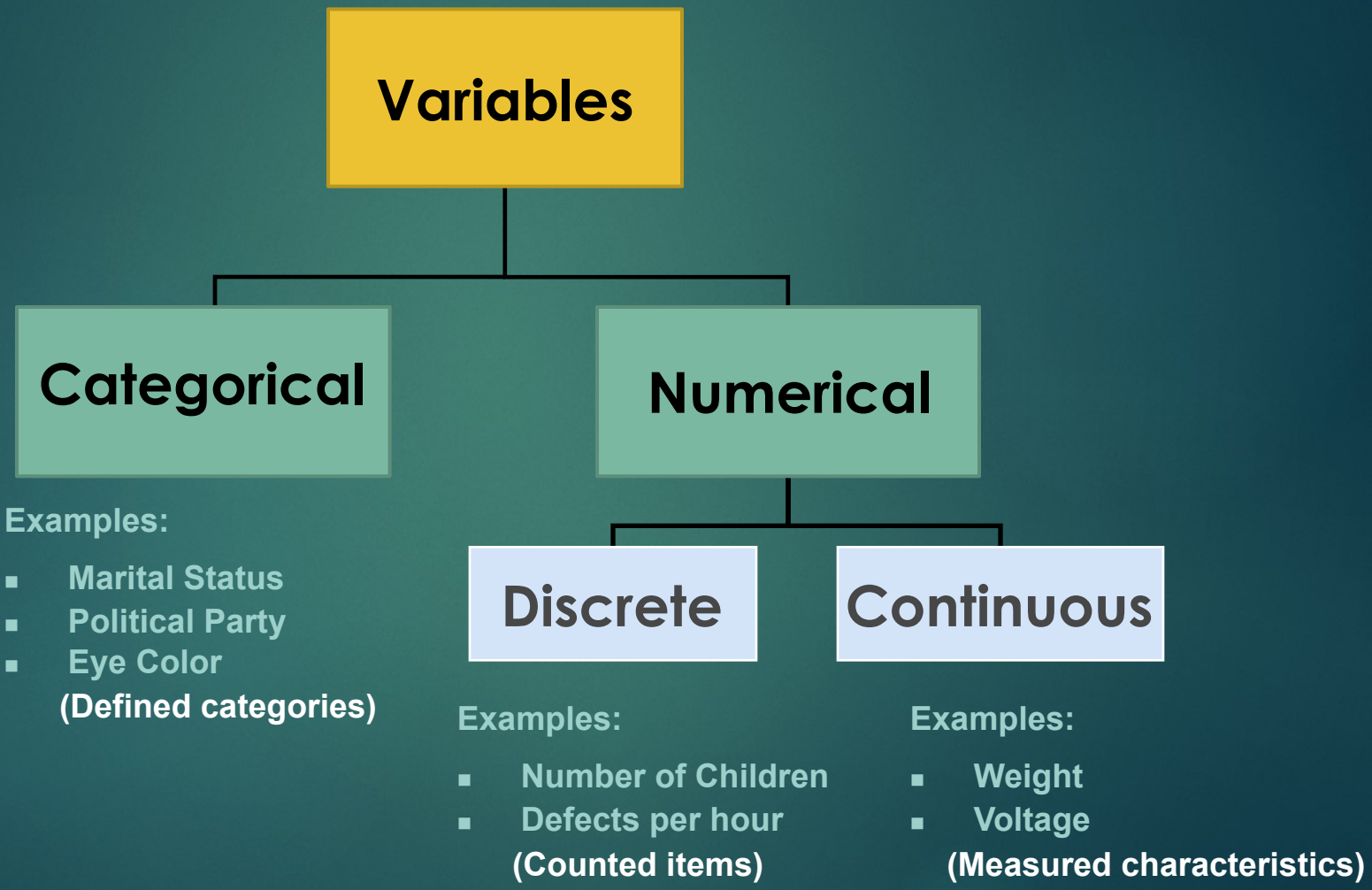
Variables

- **Numerical** variables have values that represent a counted or measured quantity.
  - **Discrete:** arise from a ***counting*** **process**
  - **Continuous:** arise from a ***measuring*** **process**

NAME	Age	Weight (lbs.)	Heart Rate (bpm)
Helen Maxwell	23	325.5	89
Pete Gregory	20	210	97
Louise Medina	25	187.2	90
Phillip Cox	20	157	101
Marie Henry	21	135.8	94
Taylor Reeves	21	140	90

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Types of Variables (summary)



Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Measurement Scales

CATEGORIAL VARIABLES

- **Nominal Scales**:

  The name says it all; e.g.: male and female

- **Ordinal Scales**:

  A ranking is involved; e.g.: very likely, likely, unlikely, very unlikely

  Example: In a survey, MLB fans were asked: how likely it is that your team can make it to the World Series this year? Their response is summarized below:

ID Number	Gender	Team	Response
12768	Male	NYM	Very Likely
13478	Female	SD	Unlikely
18435	Male	MIL	Unlikely
19654	Male	CIN	Very Unlikely
15675	Female	LAD	Likely

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Measurement Scales
Numerical VARIABLES

- **Ratio Scales:**

  When there is a true *zero point*; e.g.: weight

- **Interval Scales:**

  When there is no zero point; e.g.: GPA

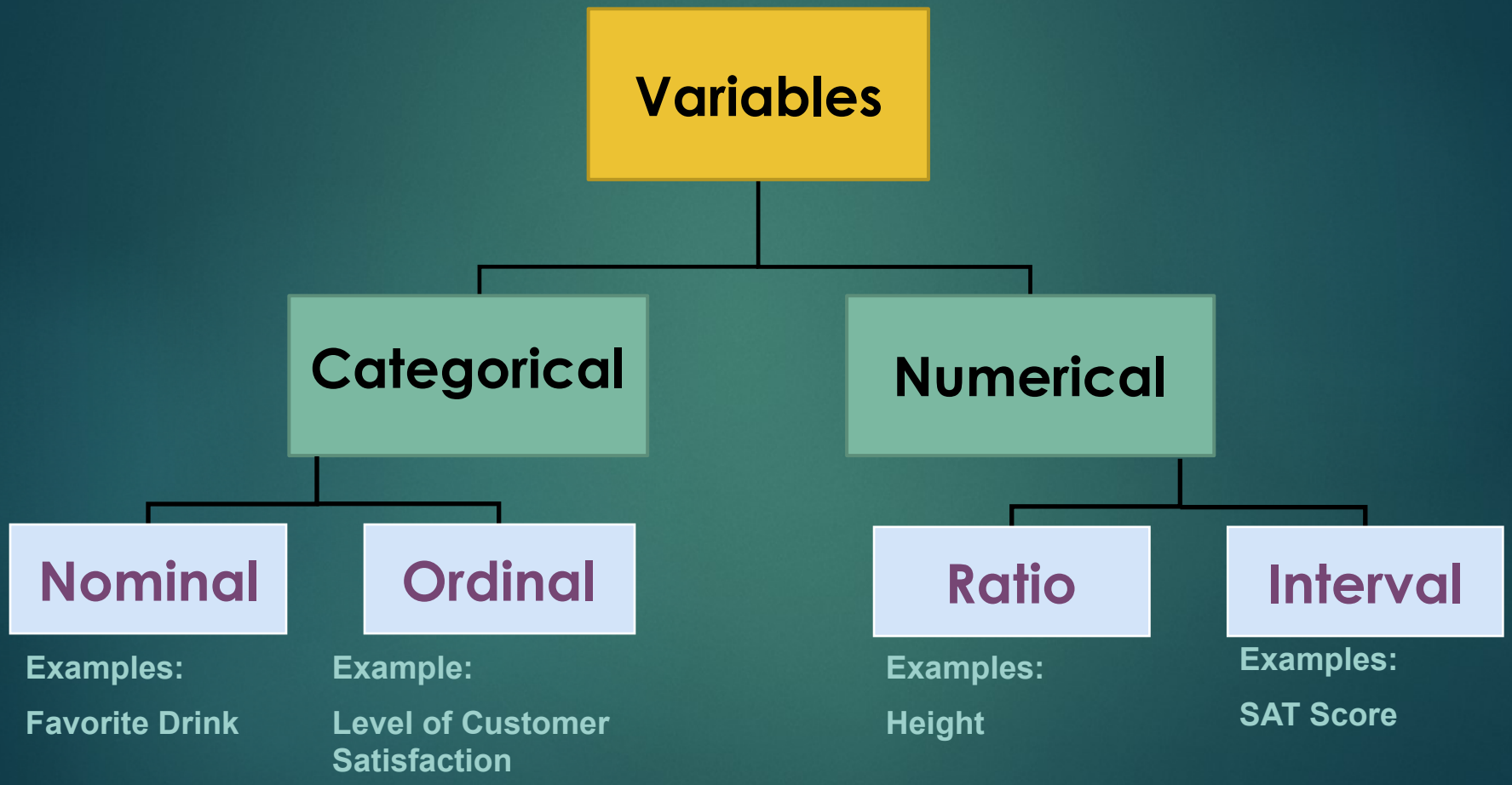
Name	Weight	GPA
Carlos Leonard	187.5	4.0
Trevor Cortez	250	3.8
Pamela Colon	203.6	4.0
Julio Greer	200.4	3.2
Katherine Ford	154.5	3.5

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Numerical Variable	*Level of Measurement*
Temperature (in degrees Celsius or Fahrenheit)	Interval
Standardized exam score (e.g., ACT or SAT)	Interval
Height (in inches or centimeters)	Ratio
Weight (in pounds or kilograms)	Ratio
Age (in years or days)	Ratio
Salary (in American dollars or Japanese yen)	Ratio

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Types of Measures (summary)



Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Class Activity

- A group of risk management analysts have gathered data on:
  - **Car Loan Payments**, which varies btw. $250 and $1000
  - **Age Group**, which is either Young, Middle age, or Senior
  - **Family Size**, which varies btw. 1 to 12
  - **Vehicle Type**, which is either Compact, Sedan, SUV, or Truck

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Class Activity (cont' d)

- What type of variable is **Car Loan Payments** (which varies btw. $250 and $1000)?
  - Categorical and Ordinal
  - Categorical and Nominal
  - Numerical and Discrete with Interval Value
  - Numerical and Discrete with Ratio Value
  - Numerical and Continuous with Interval Value
  - Numerical and Continuous with Ratio Value

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Class Activity (cont' d)

- What type of variable is **Car Loan Payments** (which varies btw. $250 and $1000)?
  - Categorical and Ordinal
  - Categorical and Nominal
  - Numerical and Discrete with Interval Value
  - Numerical and Discrete with Ratio Value
  - Numerical and Continuous with Interval Value
  - **Numerical and Continuous with Ratio Value**

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Class Activity (cont' d)

- What type of variable is **Age Group** (which varies btw. Young, Middle age, and Senior)?
  - Categorical and Ordinal
  - Categorical and Nominal
  - Numerical and Discrete with Interval Value
  - Numerical and Discrete with Ratio Value
  - Numerical and Continuous with Interval Value
  - Numerical and Continuous with Ratio Value

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Class Activity (cont' d)

- What type of variable is **Age Group** (which varies btw. Young, Middle age, and Senior)?
  - **Categorical and Ordinal**
  - Categorical and Nominal
  - Numerical and Discrete with Interval Value
  - Numerical and Discrete with Ratio Value
  - Numerical and Continuous with Interval Value
  - Numerical and Continuous with Ratio Value

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Class Activity (cont' d)

- What type of variable is **Family Size** (which varies btw. 1 and 12)?
  - Categorical and Ordinal
  - Categorical and Nominal
  - Numerical and Discrete with Interval Value
  - Numerical and Discrete with Ratio Value
  - Numerical and Continuous with Interval Value
  - Numerical and Continuous with Ratio Value

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Class Activity (cont' d)

- What type of variable is **Family Size** (which varies btw. 1 and 12)?
  - Categorical and Ordinal
  - Categorical and Nominal
  - Numerical and Discrete with Interval Value
  - **Numerical and Discrete with Ratio Value**
  - Numerical and Continuous with Interval Value
  - Numerical and Continuous with Ratio Value

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Class Activity (cont' d)

- What type of variable is **Vehicle Type**(which is either Compact, Sedan, SUV, or Truck)?
  - Categorical and Ordinal
  - Categorical and Nominal
  - Numerical and Discrete with Interval Value
  - Numerical and Discrete with Ratio Value
  - Numerical and Continuous with Interval Value
  - Numerical and Continuous with Ratio Value

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Class Activity (cont' d)

- What type of variable is **Vehicle Type**(which is either Compact, Sedan, SUV, or Truck)?
  - Categorical and Ordinal
  - **Categorical and Nominal**
  - Numerical and Discrete with Interval Value
  - Numerical and Discrete with Ratio Value
  - Numerical and Continuous with Interval Value
  - Numerical and Continuous with Ratio Value

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Assignment:

MyStatLab: Homework, Section 1.1.-2.

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

DCOVA

Defining,

Collecting,

Organizing, **V**isualizing, and **A**nalyzing

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Data Collection

- 1st Step: identify the data source
- 2nd Step: decide whether you want to collect the data from the entire population or just a sample.
- 3rd Step: clean the collected data
- 4th Step: recode the variables

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Data Sources

- **Primary Sources:**

  The data collector is the one using the data for analysis

  - Data from a political survey
  - Data collected from an experiment

- **Secondary Sources:**

  The person performing data analysis is not the data collector

  - Analyzing census data
  - Examining data published on the internet.

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Data Sources (cont'd)

- Data distributed by an organization or an individual (Regional Economic Data Collected by Dallas Fed)
- The outcomes of a designed experiment

  (Reports on Selected Field Experiments in Economics)
- The responses from a survey

  (The BLS Job Openings and Labor Turnover Survey)
- The results of conducting an observational study

  (Marketing Departments often use focus groups to elicit unstructured responses)
- Data collected by ongoing business activities

  (CompuStat Dataset)

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Data Cleaning

- Often find "irregularities" in the data
  - Typographical or data entry errors
  - Values that are impossible or undefined
  - Missing values
  - Outliers
- When found these irregularities should be reviewed / addressed
- Both Excel & Minitab can be used to address irregularities

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Data Recoding

- Recoding a variable can either supplement or replace the original variable.
- Recoding a categorical variable involves redefining categories.
- Recoding a quantitative variable involves changing this variable into a categorical variable.
- When recoding be sure that the new categories are mutually exclusive (categories do not overlap) and collectively exhaustive (categories cover all possible values).

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Data Recoding

- Recoding a variable can either supplement or replace the original variable.

Student ID	Class Designation
2347890	Freshman
2457914	Freshman
2347689	Sophomore
2596801	Freshman
2890756	Junior
2334909	Sophomore
2889123	Senior
2009748	Freshman
2764310	Junior

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Data Recoding (cont'd)

Freshman & Sophomore → Lower

Senior & Junior → Upper

Student ID	Class Designation	Designation Recoded
2347890	Freshman	**Lower**
2457914	Freshman	**Lower**
2347689	Sophomore	**Lower**
2596801	Freshman	**Lower**
2890756	Junior	**Upper**
2334909	Sophomore	**Lower**
2889123	Senior	**Upper**
2009748	Freshman	**Lower**
2764310	Junior	**Upper**

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Data Recoding (cont'd)

- When recoding be sure that the new categories are:
  - **mutually exclusive** (categories do not overlap)
  - **collectively exhaustive** (categories cover all possible values).

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Population

Sample

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Population vs. Sample

POPULATION consists of all the items or individuals about which you want to draw a conclusion. The population is the "large group"

SAMPLE is the portion of a population selected for analysis. The sample is the "small group"

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Population vs. Sample

We are often interested to learn about the population, but it is not possible for us to collect data as population of interest grows in size.

That's when SAMPLING METHODS become important.

AND

That's when INFERENTIAL STATISTICS becomes handy!

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Sampling
Sampling Frame

- The listing of items that make up the population
- Frames are data sources; e.g.; directories such as Employers Identification Number (EIN) at the IRS
- If a frame excludes certain portions of the population, then the results will be inaccurate

  (e.g., considering only the large firms and ignoring the medium or small firms)
- Using different frames to generate data can lead to dissimilar conclusions

  (e.g., considering only the large firms in one study and only the small firms in another study)

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Sampling
Non-probability Sampling

- Cannot be used for inferential statistics since it is not a good representation of population
- It is of two general types:
  - Convenience Sample; e.g., Social Media Polls
  - Judgment Sample; e.g., Expert Polls
- Once you understand the limitations of Non-probability Sampling, you may be able to use them properly (e.g., pilot studies)

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Sampling

Probability Sampling

Probability Sampling

- **Simple Random**
- **Systematic**
- **Stratified**
- **Cluster**

The *characteristics* of the items in the population are NOT considered in these sampling methods.

The *characteristics* of the items in the population are considered in these sampling methods.

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Probability sampling with no particular attention to items characteristics:

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Sampling

Simple Random Sampling

- **N=800** (Number of Employees at BlackSea Corp.)
- **n*=32** (Desired sample size for a survey on dental care costs)
- Expected response rate=80%
- **n=40** (Number of employees that the survey is sent to)
- Use the Table of Random Numbers (PP. 742-743) to choose the employees in the sample:
  - 1st Step: Assign codes to the employees: 001, 002,…, 800
  - 2nd Step: Choose a random number from the table
  - 3rd Step: Choose those codes that are greater than 000 and less than 800, and send the survey to the employees who have been assigned with those codes
- For more, see Using a Table of Random Numbers in Simple Random Sampling (PDF) on the D2L

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Sampling
Simple Random Sampling

- Initially, and I put emphasis on "initially," the probability of choosing the items from the sample is the same for all of them
- Initially, and put emphasis on "initially" again, any given item has the same chance of being chosen in Simple Random Sampling
- Why do I put an emphasis on "initially"?
  - Choosing items with replacement

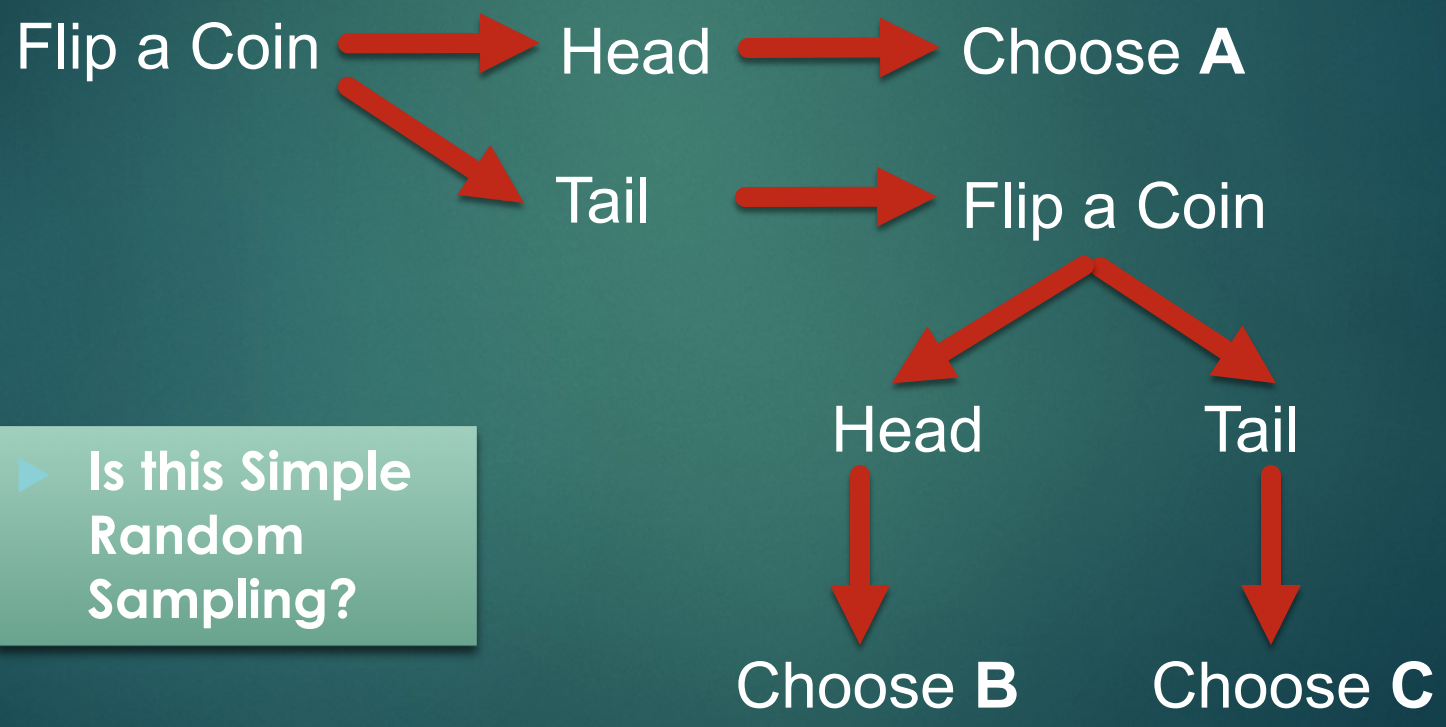
  vs.

  - Choosing items without replacement

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Class Activity

- We are interested in choosing a sample of size **n=1** from a population of size **N=3**, which is made of items **A**, **B**, and **C**.
- Here's how we choose the sample:

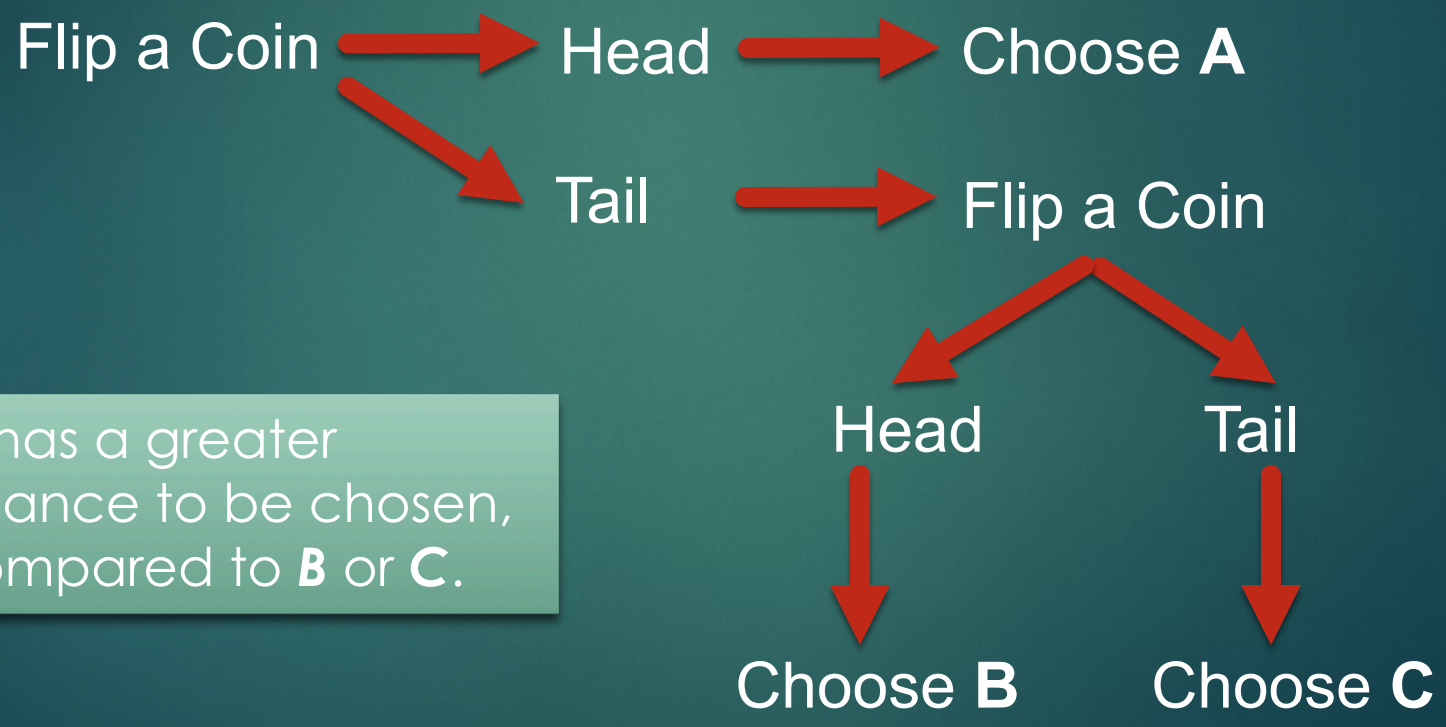


- **Is this Simple Random Sampling?**

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Class Activity (cont' d)

- **No it is not**. This is **Probability Sampling**, because the sample is chosen **randomly**, and the **probability of choosing each item is known**. But it is not **Simple Random Sampling**, because the **probability of choosing the items are not the same for all of them**.



A has a greater chance to be chosen, compared to ***B*** or ***C***.

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Sampling
Systematic Sampling

- **N=800** (Number of Employees at BlackSea Corp.)
- **n*=32** (Desired sample size for a survey on dental care costs)
- Expected response rate=80%
- **n=40** (Number of employees that the survey is sent to)
- Systematic Sampling:
  - 1st Step: Assign codes to the employees: 001, 002,…, 800
  - 2nd Step: Partition the employees into 40 groups. (Each group has 20 employees)
  - 3rd Step: Select a random number from the first group (using the table of random numbers, for instance)
  - 4th Step: Include every twentieth individual after the first selection

e.g., **015** → Sample includes ID No. 015, 035, 055, 075, 095,…, and 795

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Probability sampling with particular attention to items characteristics:

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Sampling
Stratification

- **Stratum:** a group of items in the population that share a common characteristics; e.g. personnel at the same level of organizational hierarchy at BlackSea Corp
- Stratification:
  - 1st Step: Define a characteristic: Managerial vs. Non-managerial staff members
  - 2nd Step: Divide the population into different strata based on the above characteristics
  - 3rd Step: Select a simple random sample from each stratum, and combine the results given the share of each stratum in the entire population.
- For more, see Selecting a Stratified Sample (PDF) on the D2L

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Sampling
Cluster Sampling

- **Cluster:** Naturally occurring groups; e.g. Different states in which BlackSea Corp has an office
- Cluster Sampling:
  - 1st Step: Identify the clusters: e.g., 12 states
  - 2nd Step: Divide the population into different clusters
  - 3rd Step: Select a simple random sample from each cluster, and combine the results.
- Cluster sampling often requires large sample sizes.

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Probability Sampling: Comparing Sampling Methods

- Simple random sample and Systematic sample
  - Simple to use
  - May not be a good representation of the population's underlying characteristics
- Stratified sample
  - Ensures representation of individuals across the entire population
- Cluster sample
  - More cost effective
  - Needs larger sample to acquire the same level of precision

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Assignment:

MyStatLab: Homework, Section 1.3.-4.

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

- Prior to the presidential election of 1936, Literary Digest had successfully predicted the results of five presidential elections.
- For 1936, they sent out 10 million ballots! 2.3 million ballots were received. The result was Alf Landon, with a large margin.
- Franklin D. Roosevelt was elected, and Alf Landon received fewest electoral votes ever.
- Literary Digest ended its publication in less than two years afterwards.

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Literary Digest

- What caused this?
  - Low response rate: less than 25% of those who received the ballot responded
  - Nonresponse error: FDR voters were less likely to send back the ballots
- What did we learn?
  - The size of the sample does not matter much
  - What matters is the sampling method.
- What are the implications for Social Media Polls, then?

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Excel Exercise

- File name: **Ch01_Recoding_and_Sampling.xlsx**

  Note: This file is uploaded under Lecture Presentations.
- Sheets: NOTES and DATASET1
- Tasks:
  - Recoding *Class* variable in a survey
  - Recoding *GPA* variable in a survey
  - Generating random numbers in Excel.

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Required Readings:

- Probability Sampling examples, uploaded under Lecture Presentation.
- Section 1.5: PP. 24-26
- 'THINK ABOUT THIS' box: PP. 26-27

Copyright © 2015, 2012, 2009 Pearson Education, Inc.

Assignment:

MyStatLab: Homework, Section 1.5.

Copyright © 2015, 2012, 2009 Pearson Education, Inc.