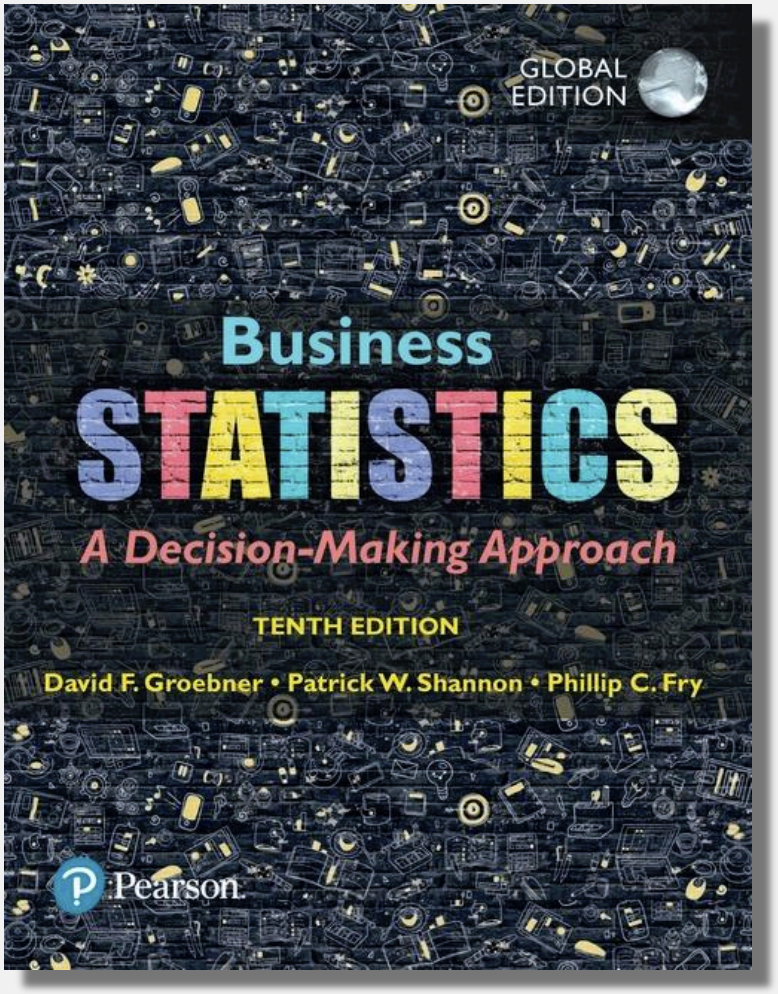
Chapter 1

The Where, Why, and How of Data Collection



Learning Outcomes

Outcome 1. Know the key data collection methods.

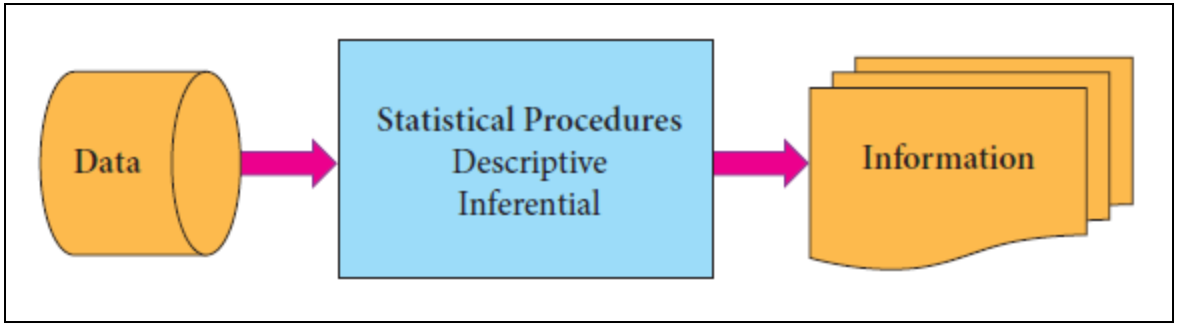
Outcome 2. Know the difference between a population and a sample.

Outcome 3. Understand the similarities and differences between different sampling methods.

Outcome 4. Understand how to categorize data by type and level of measurement.

1.1 What is Business Statistics?

- A collection of procedures and techniques used to convert data into meaningful information in a business environment



Data

Variables

Observations

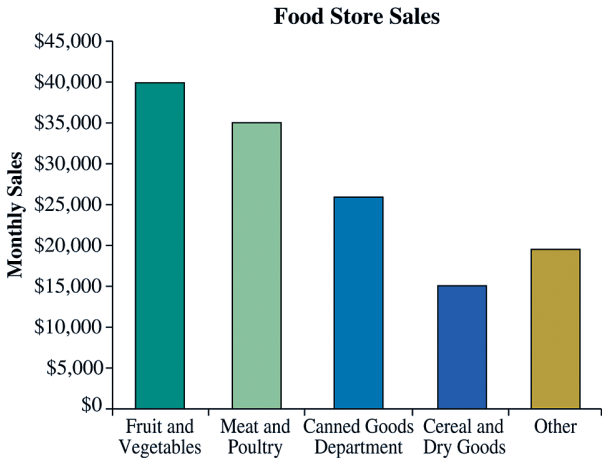
	A	B	C	D	E	F	G
1	Book	Units Sold	Pages	Competing Books	Advertising Budget	Age of Author	Market Classification
2	1	15000	176	5	25000	49	Social Sciences
3	2	140000	296	10	83000	57	Business
4	3	75000	483	7	40000	29	Business
5	4	100000	811	14	29000	37	Social Sciences
6	5	26000	302	9	52000	35	Business
7	6	33000	411	15	33000	43	Business
8	7	59000	333	7	19000	51	Social Sciences
9	8	103000	602	4	37000	62	Business
10	9	88000	504	12	51000	33	Social Sciences
11	10	10000	204	3	30000	50	Business
12	11	9000	376	4	19000	26	Business
13	12	77000	600	7	41000	40	Business
14	13	59000	400	3	26000	44	Social Sciences
15	14	183000	597	8	51000	59	Business
16	15	16000	126	1	27000	38	Social Sciences

Statistical Procedures

- Descriptive Statistics
  - Procedures and techniques designed to describe data
- Inferential Statistics
  - Tools and techniques that help decision makers to draw inferences from a set of data

Descriptive Procedures

- Charts, graphs, and tables



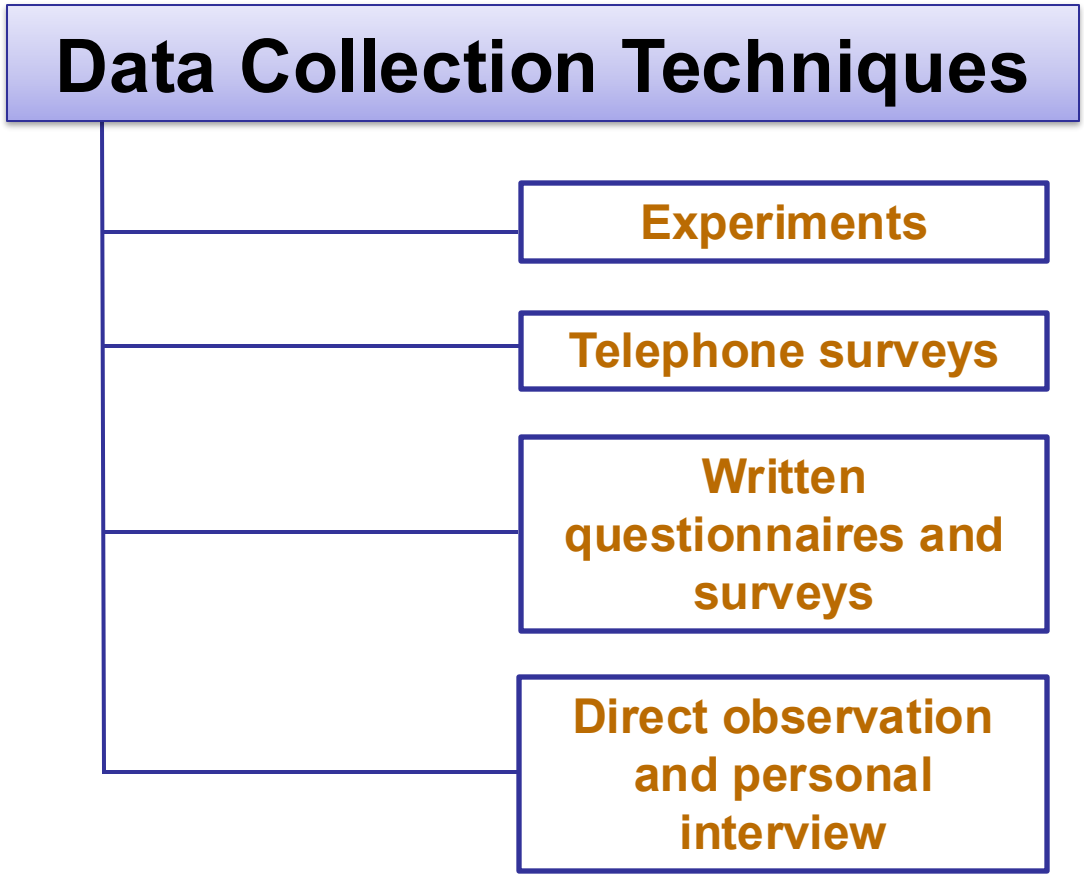
- Numerical measures

$$\text{Average} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

Inferential Procedures

- Estimation
  - e.g., Estimate the average family income of all families in a city based on the average income of a sample of families in that city.
- Hypothesis Testing
  - e.g., Use sample evidence to test the claim that the average family income exceeds $45,000 per year.

1.2 Procedures for Collecting Data



Experiments

- Experiment
  - A process that produces a single outcome whose result cannot be predicted with certainty.
- Experimental design
  - A plan for performing an experiment in which the variable of interest is defined.

Experiment – French Fry Example

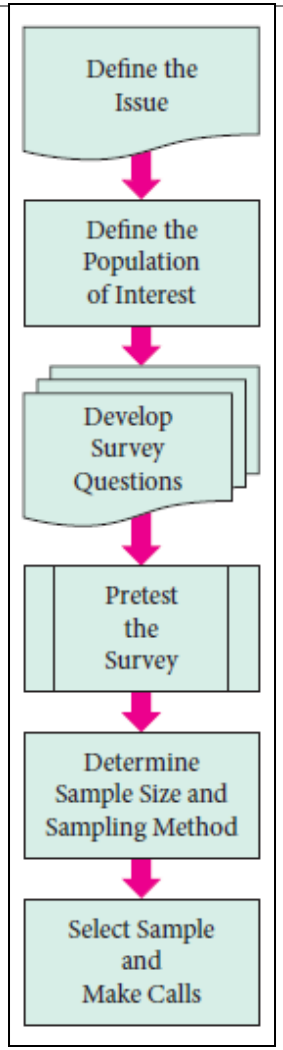
Blanch Time	Blanch Temperature	Potato Category 1	2	3	4
10 minutes	100				
	110				
	120				
15 minutes	100				
	110				
	120				
20 minutes	100				
	110				
	120				
25 minutes	100				
	110				
	120				

Telephone Surveys

- Typical Applications
  - Political Polls
  - Market Research
  - Customer Satisfaction Studies
  - Healthcare Studies
  - Other?

Telephone Surveys

Major Steps for a Telephone Survey



Telephone Surveys

- Issues and Challenges
  - Respondents Screen Calls – Won't Answer
  - Respondents Not Home
  - Cell Phone Numbers Must be Manually Dialed
  - Survey Must be Short – Respondents Hang Up
  - Requires Trained Callers
  - Other?

Telephone Surveys

- Closed-End Questions
  - Questions that require the respondent to select from a short list of defined choices
- Example:
  - "What is your preferred rental car company?" _____

Telephone Surveys

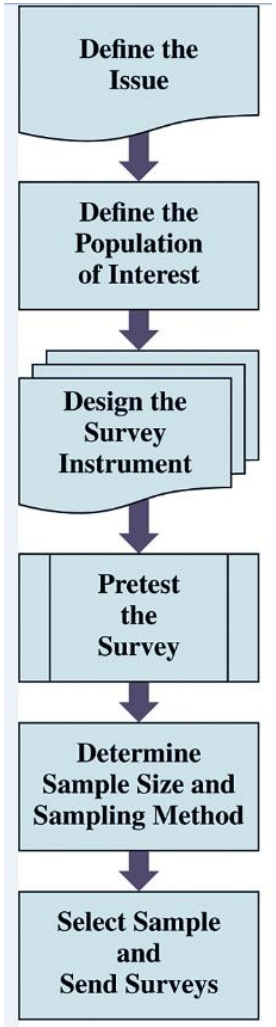
- Demographic Questions
  - Questions relating to the respondents' characteristics, backgrounds, and attributes
- Examples:
  - What is your gender? _____Male
    _____Female
  - List your age on your last birthday? _____

Written Questionnaires

- Differences from Telephone Surveys
  - Must be Carefully Formatted
  - Need Motivational Statement of Purpose
  - Can More Easily Allow for Open-Ended Questions
  - Less Expensive to Administer
  - Other?

Written Questionnaires

Major Steps for a
Written Questionnaire Survey



Written Questionnaires

- Open-End Questions
  - Questions that allow respondents the freedom to respond with any value, words, or statements of their own choosing
- Example
  - "Under what circumstances would you favor a reduction in the Federal long-term capital gains tax rate?" _____

Written Questionnaires

- Typical Applications
  - Employee Surveys
  - College Course/Professor Evaluations
  - Customer Satisfaction Studies
  - Academic Research
  - Other?

Written Questionnaires

- Advantages
  - Questions are Standardized for All Respondents
  - If On-Line, Data are Captured Electronically
  - Potentially Allows for More Questions
- Disadvantages
  - Responses May Not Be Timely
  - Potentially Low Response Rate (5-10%)
  - No Chance to Clarify Questions

Observations and Interviews

- Direct Observations
  - Data that are being collected is physically observed and the data recorded based on what takes place in the process.
  - Subjective and time-consuming
- Personal Interviews
  - Structured: questions are scripted
  - Unstructured: begin with one or more broadly stated questions, with further questions being based on the responses

Direct Observation

- Direct Observation – Typical Applications
  - Determine Customer Product Preferences
  - Determine Rate of Compliance With Mandatory Seatbelt Use
  - Assess Compliance With Company Safety Policies

Direct Observation

- Direct Observation - Challenges
  - Avoid “Hawthorne Affect”
  - Maintain Consistency Among Observers
  - Having Properly Trained Observers
  - Avoid Observer Bias
  - Costly
  - Time Consuming
  - Other?

Interview

- Interview – Typical Applications
  - Customer Satisfaction Measurement
  - Supplier Qualification Assessment
  - Health Status Measurement
  - Other?

Interview Types

- Structured Interviews
  - Verbally Administered
  - Specific Set of Predetermined Questions
  - Questions Asked in the Same Order to All Respondents
  - No Follow-Up or Elaboration of Answers

Interview Types

- Unstructured Interviews
  - No Predetermined Set of Questions
  - Begin With a General Question Reflecting the Overall Objective of the Interview
  - Allow the Discussion to Go in Any Direction
  - Probe Respondents for Clarification and Elaboration of Responses

Interview

- Interview – Challenges
  - Interviewer Bias
  - Capturing Responses Accurately
  - Time Requirements
  - Limits on Number of Interviewees
  - Interviewer Training
  - Consistency Among Interviewers
  - Other?

Data Collection Techniques

Data Collection Method	Advantages	Disadvantages
Experiments	Provide controls Preplanned objectives	Costly Time-consuming Requires planning
Telephone Surveys	Timely Relatively inexpensive	Poor reputation Limited scope and length
Mail Questionnaires Written Surveys	Inexpensive Can expand length Can use open-end questions	Low response rate Requires exceptional clarity
Direct Observation Personal Interview	Expands analysis opportunities No respondent bias	Potential observer bias Costly

Other Data Collection Techniques

- Focus Groups
  - A group interview using a facilitator
  - Obtains data about the group's perspectives and opinions
  - The group's responses are coded into categories for analysis
  - The members of the group have some common attributes

Other Data Collection Techniques

- Documents and Records
  - Studying existing data (secondary data) contained in databases, financial records, annual reports, etc.
  - Potentially less expensive than generating primary data but may not be complete.

Other Data Collection Techniques

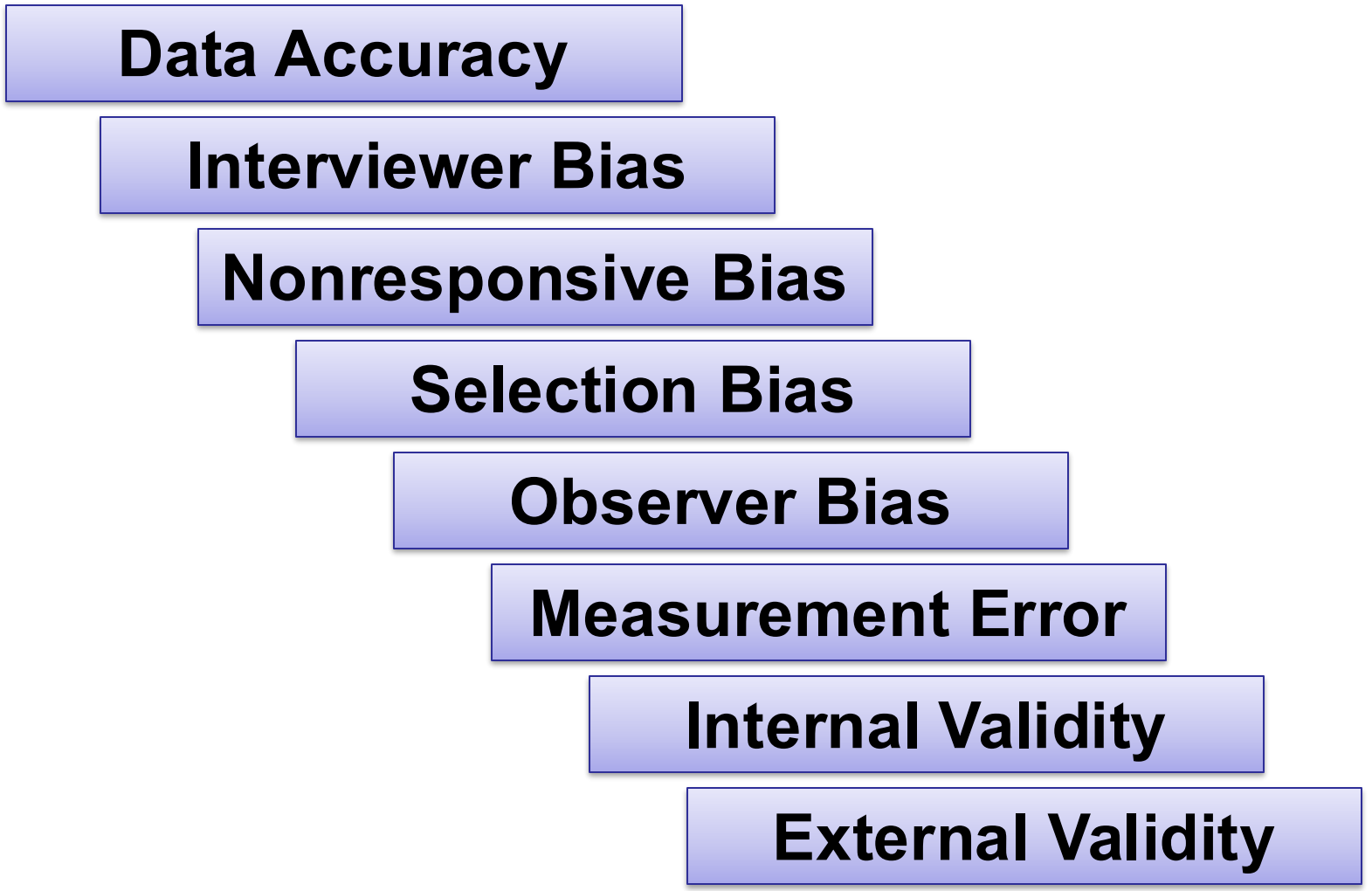
- Bar Codes

  - Bar Code Provides Information About a Product (Price, Supplier, etc.)
  - A Bar Code Scanner Automatically Captures the Data

Other Data Collection Techniques

- RFID (Radio Frequency Identification Device)
  - Transmitting Devices Attached to a Product
  - RFID Receiver Reads Data Signal
  - Holds More Data than a Bar Code
  - Does Not Require Line of Sight
  - Works Better than Bar Codes in Harsh Environments
  - Signals Can Be Compromised

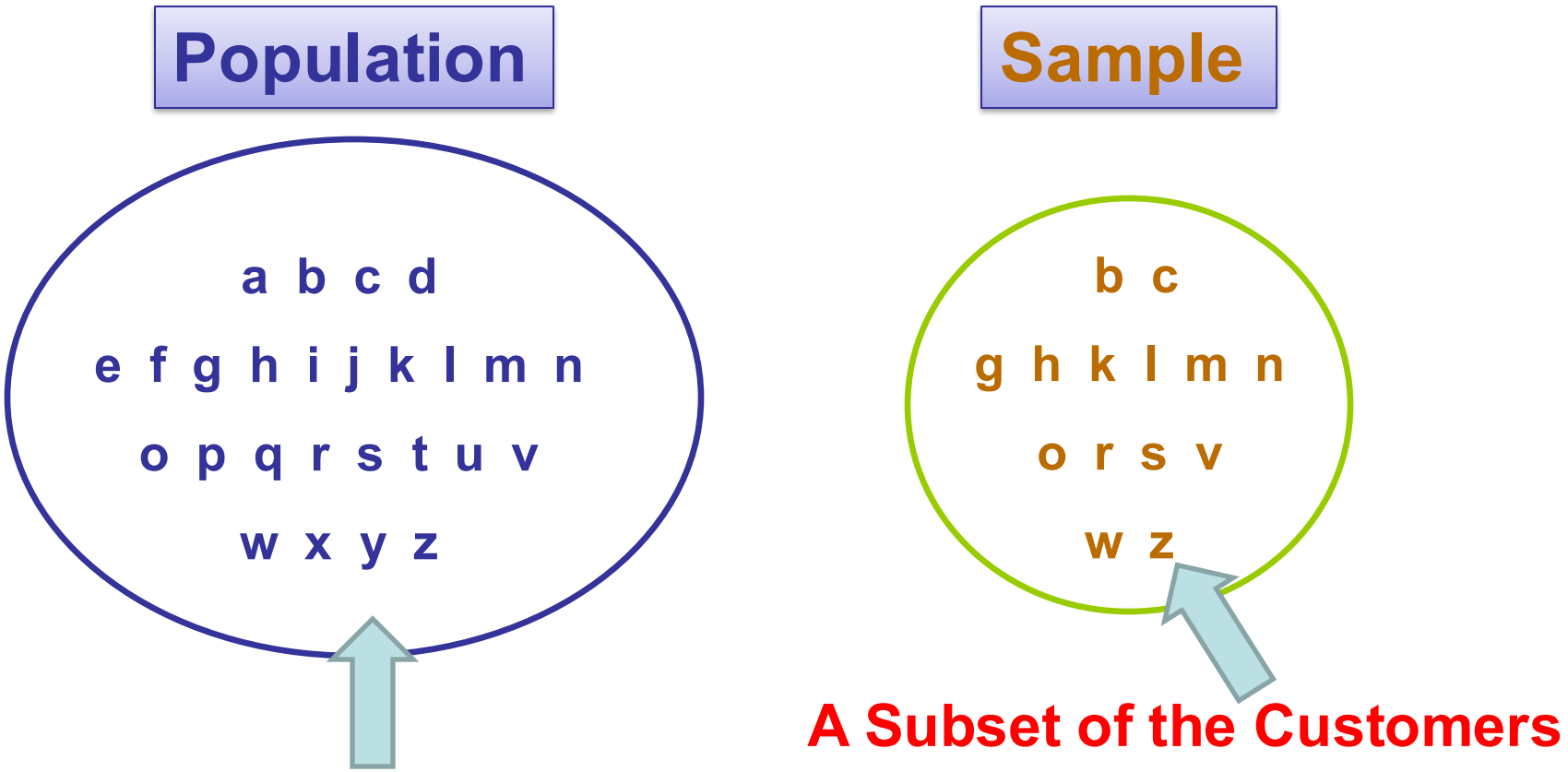
Data Collection Issues



1.3 Populations, Samples, and Sampling Techniques

- Population
  - The set of all objects or individuals of interest or the measurements obtained from all objects or individuals of interest
- Sample
  - A subset of the population
- Census
  - An enumeration of the entire set of measurements taken from the whole population

Population vs. Sample



Parameters

- Parameters
  - Descriptive numerical measures, such as an average or a proportion, that are computed from an entire population
- Examples:

The average yards gained per play by all NFL teams in the 2016 season

The proportion of all university students in California who have more than $40,000 in student loans

Statistics

- Statistics
  - Descriptive numerical measures, such as an average or a proportion, that are computed from a sample selected from a population
- Examples:

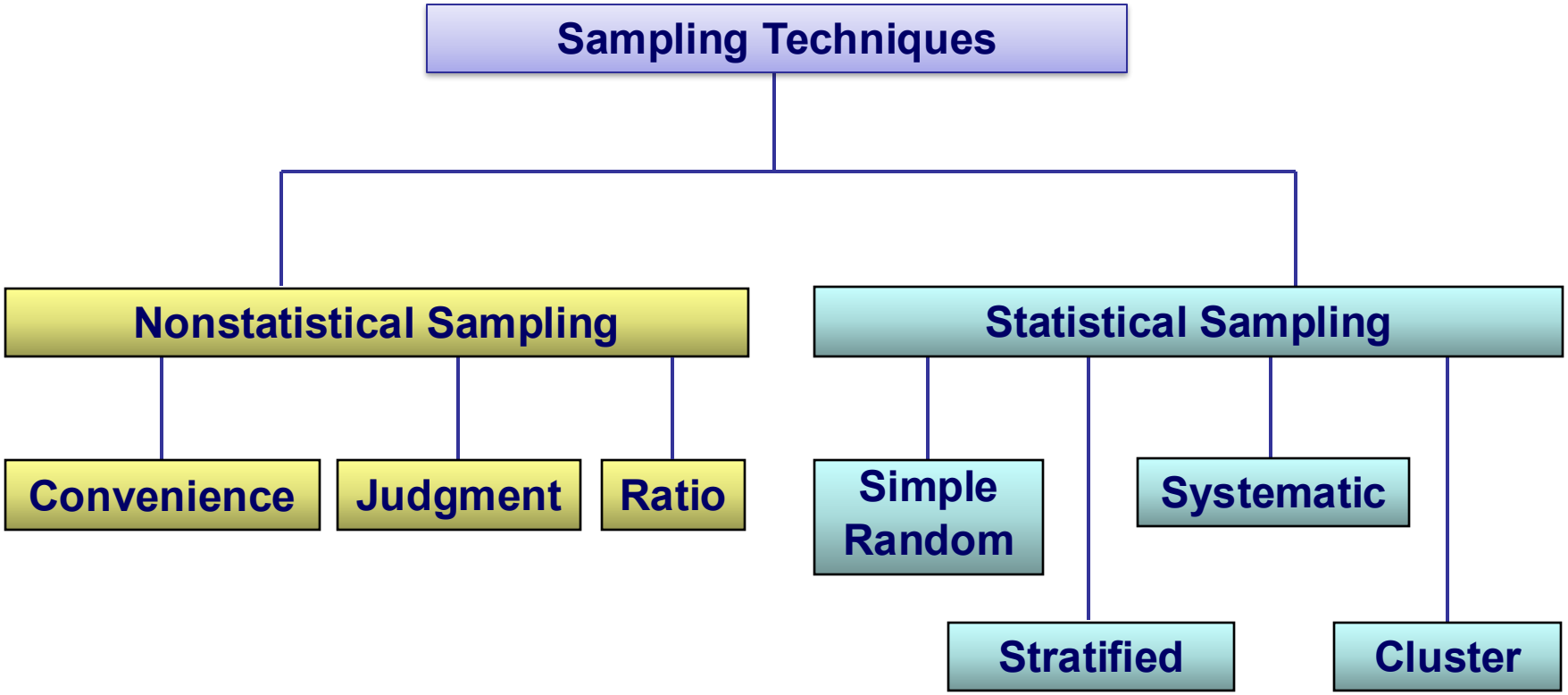
  The average credits taken by a sample of students at a university

  The proportion of defective parts in a sample of parts selected from the parts made by a automotive supply company

Sampling Techniques

- Statistical
  - Sampling methods that use selection techniques based on chance selection
- Nonstatistical
  - Methods of selecting samples that use convenience, judgment, or other non-chance processes

Sampling Techniques



Nonstatistical Sampling

- Convenience
  - Collected in the most convenient manner for the researcher

Example:

A University wishes to gather data from students to determine their interest in changing from a quarter system to a semester system. The sample is selected by stopping some students who enter the library.

Nonstatistical Sampling

- Judgment
  - Based on judgments about who in the population would be most likely to provide the needed information

Example:

A manager wishes to interview her company's suppliers to determine their satisfaction with the ordering process. She selects her ten largest suppliers to interview.

Nonstatistical Sampling

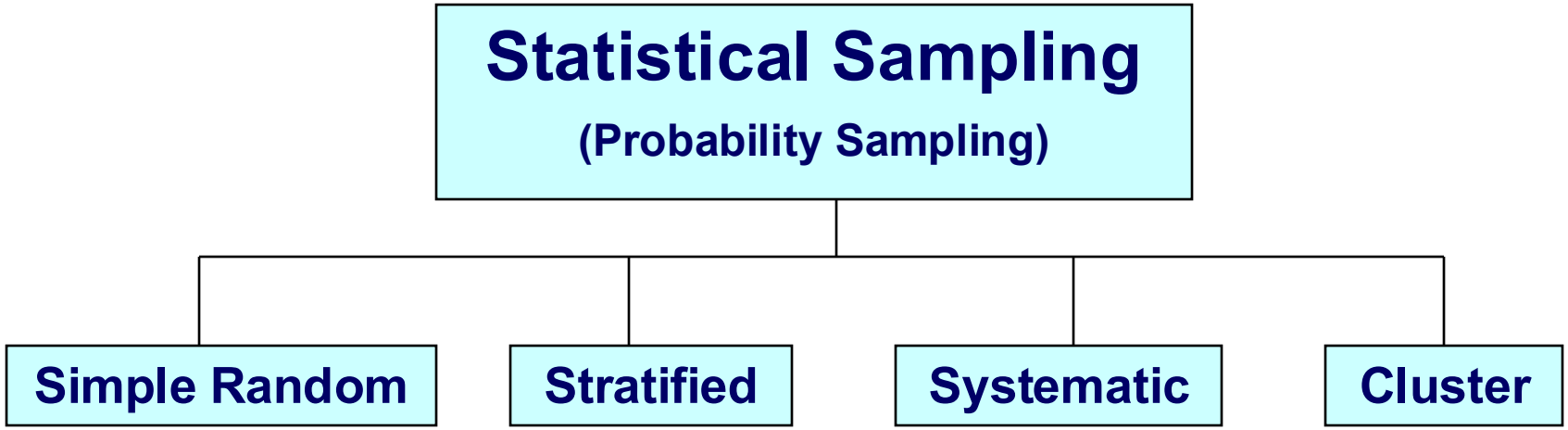
- Ratio
  - A sample in which the sample size selected from a given segment of the population is proportional to the number of items in the population belonging to that segment.

Example:

A marketing manager is interested sampling his target population of customers. The target population contains 60% females and 30% college graduates. He will make sure that the sample contains 60% females and 30% college graduates.

Statistical Sampling

- Items of the sample are chosen based on known or calculable probabilities.



Statistical Sampling

- Also called probability (or random) sampling
- Allows every item in the population to have a known or calculable chance of being included in the sample

Main Types of Statistical Sampling

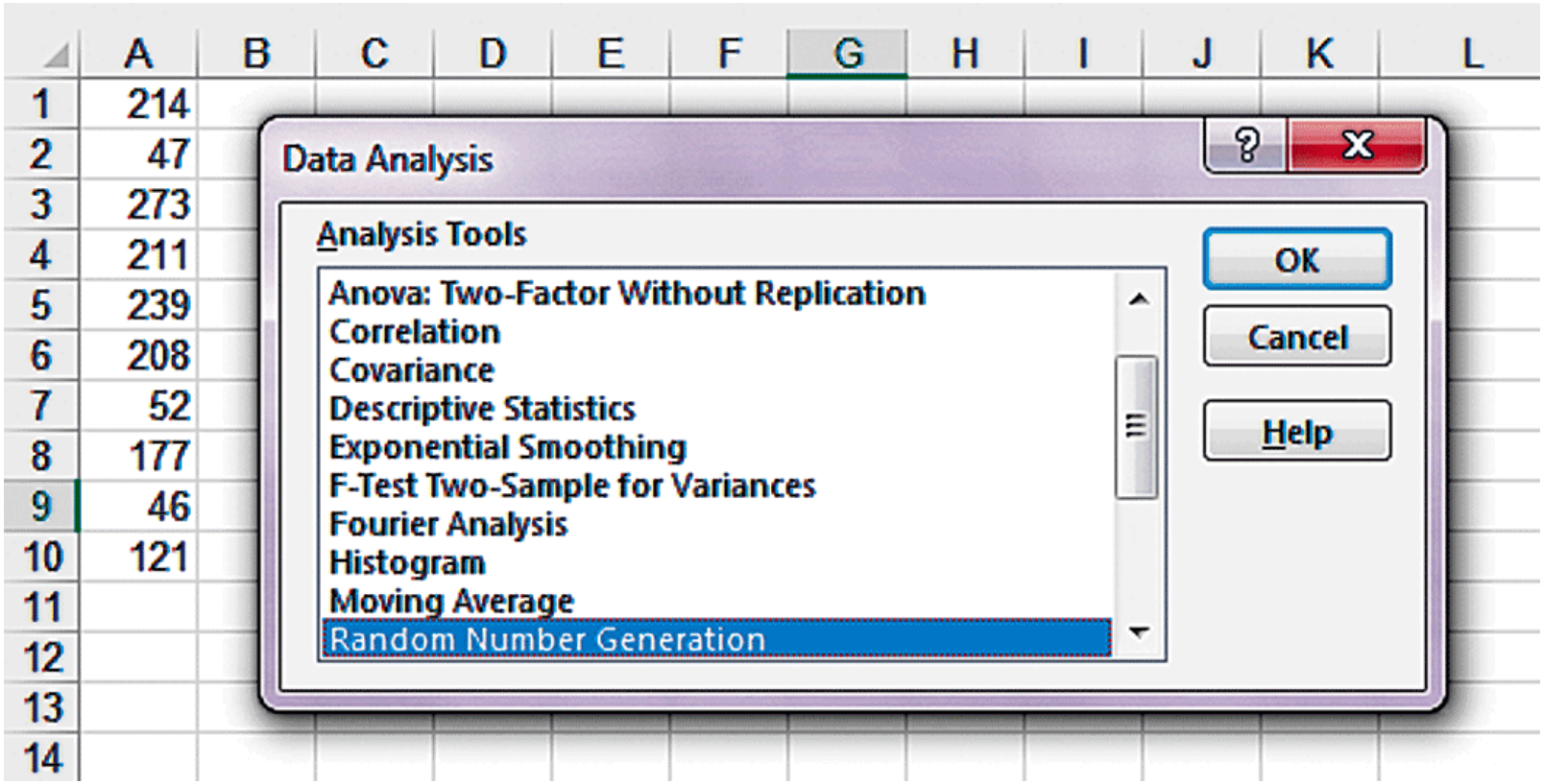
- Simple Random Sampling
- Stratified Random Sampling
- Systematic Random Sampling
- Cluster Random Sampling

Simple Random Sampling

- Every possible sample of a given size has an equal chance of being selected
- Selection may be with replacement or without replacement
- The sample can be obtained using a table of random numbers or computer random number generator

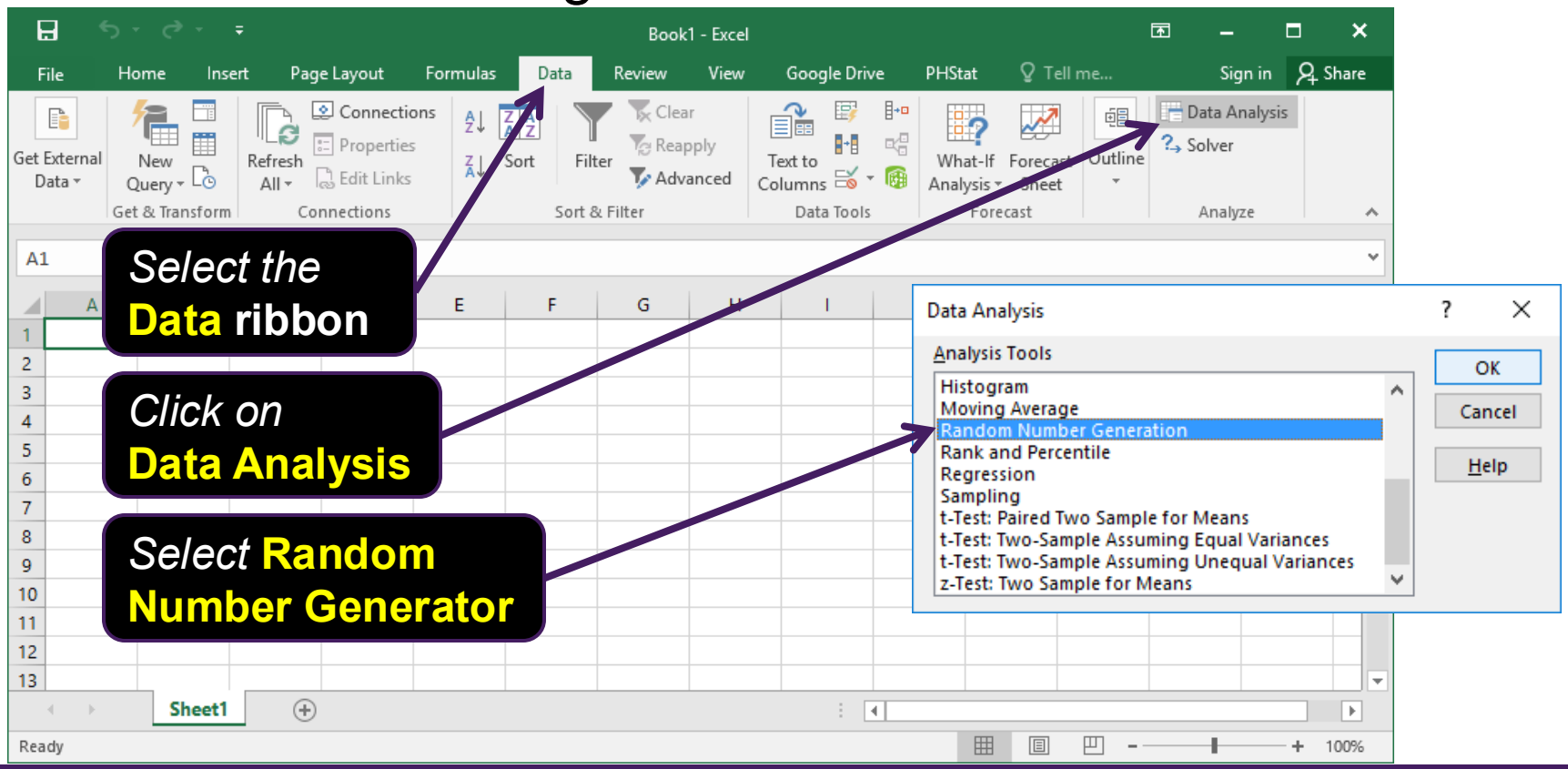
Simple Random Sampling

- Using Excel - Data Analysis – Random Number Generation

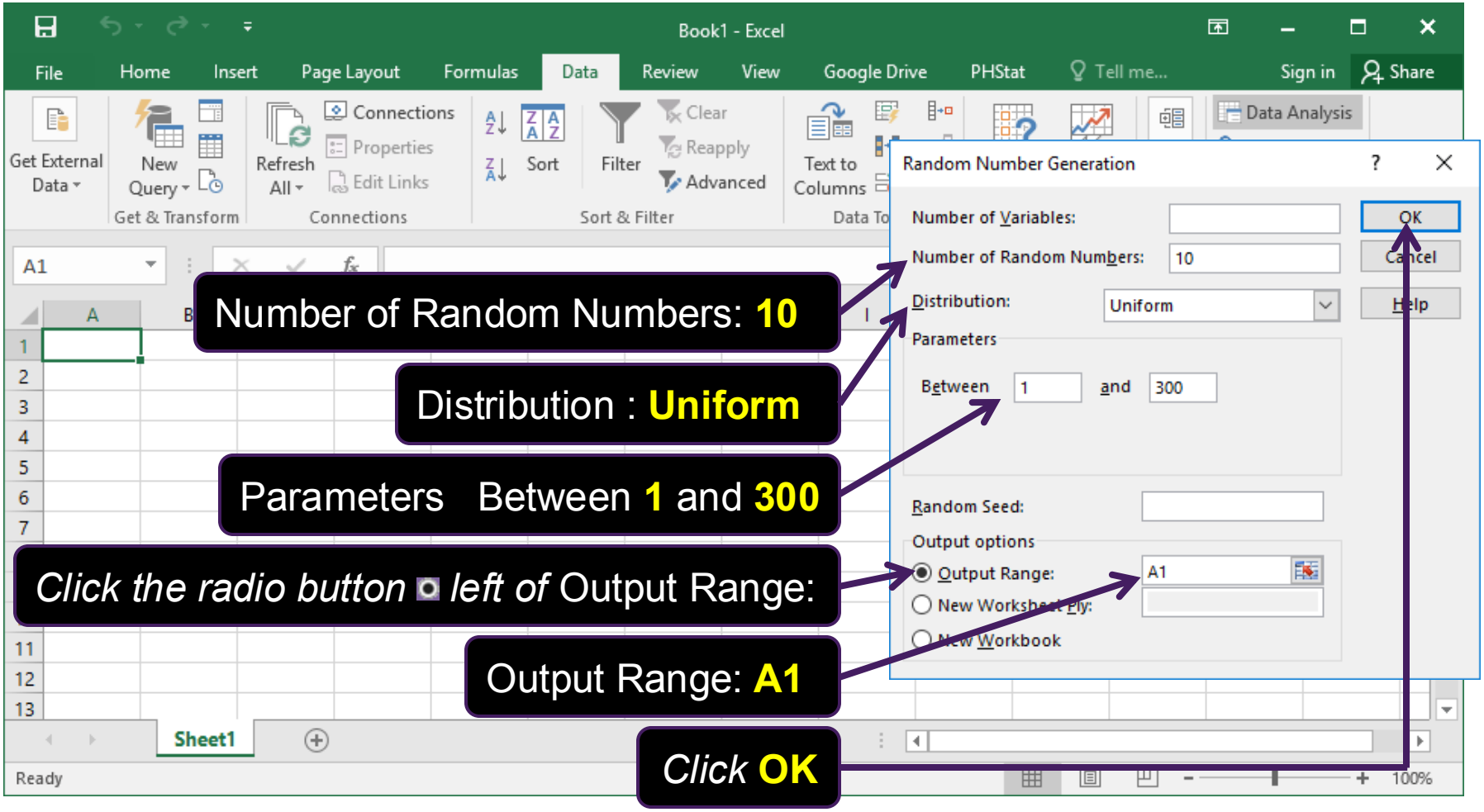


Simple Random Sampling

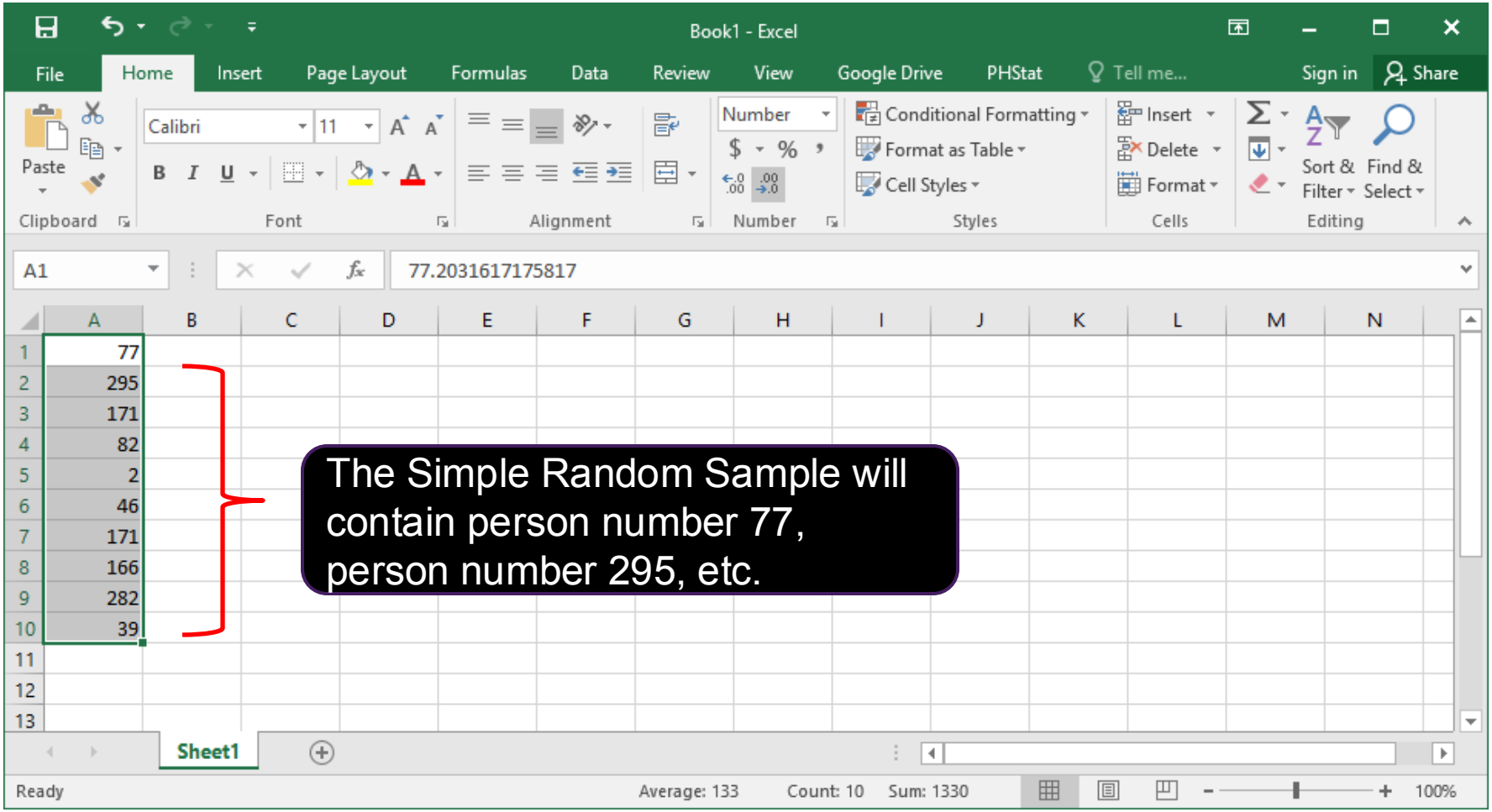
Objective: Population contains 300 people. Use Excel to generate 10 random numbers between 1 and 300. Values must be integer numbers.



Simple Random Sampling

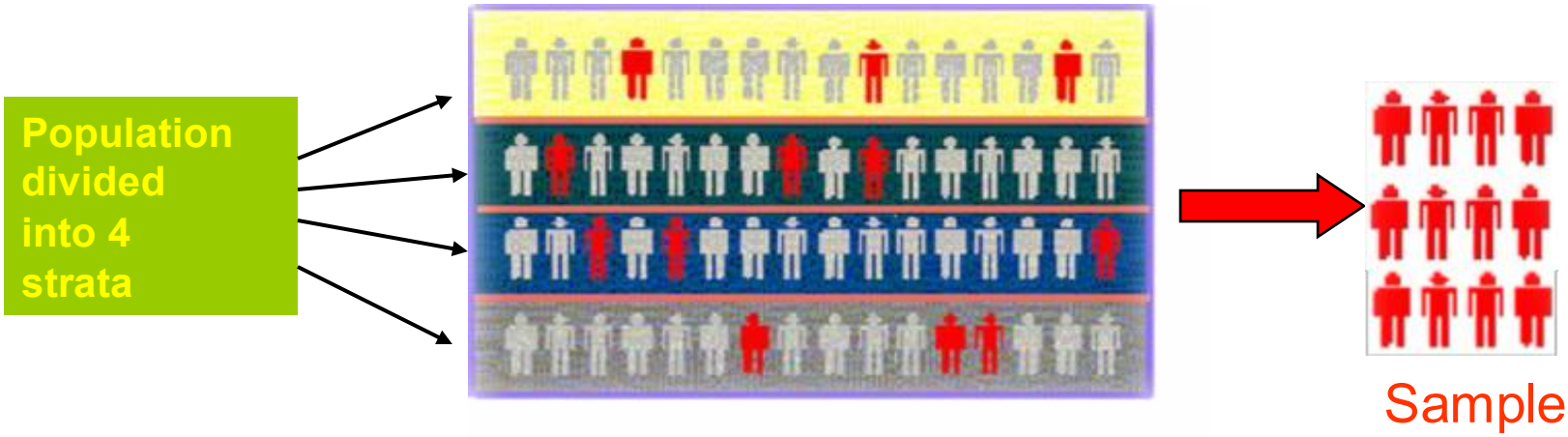


Simple Random Sampling

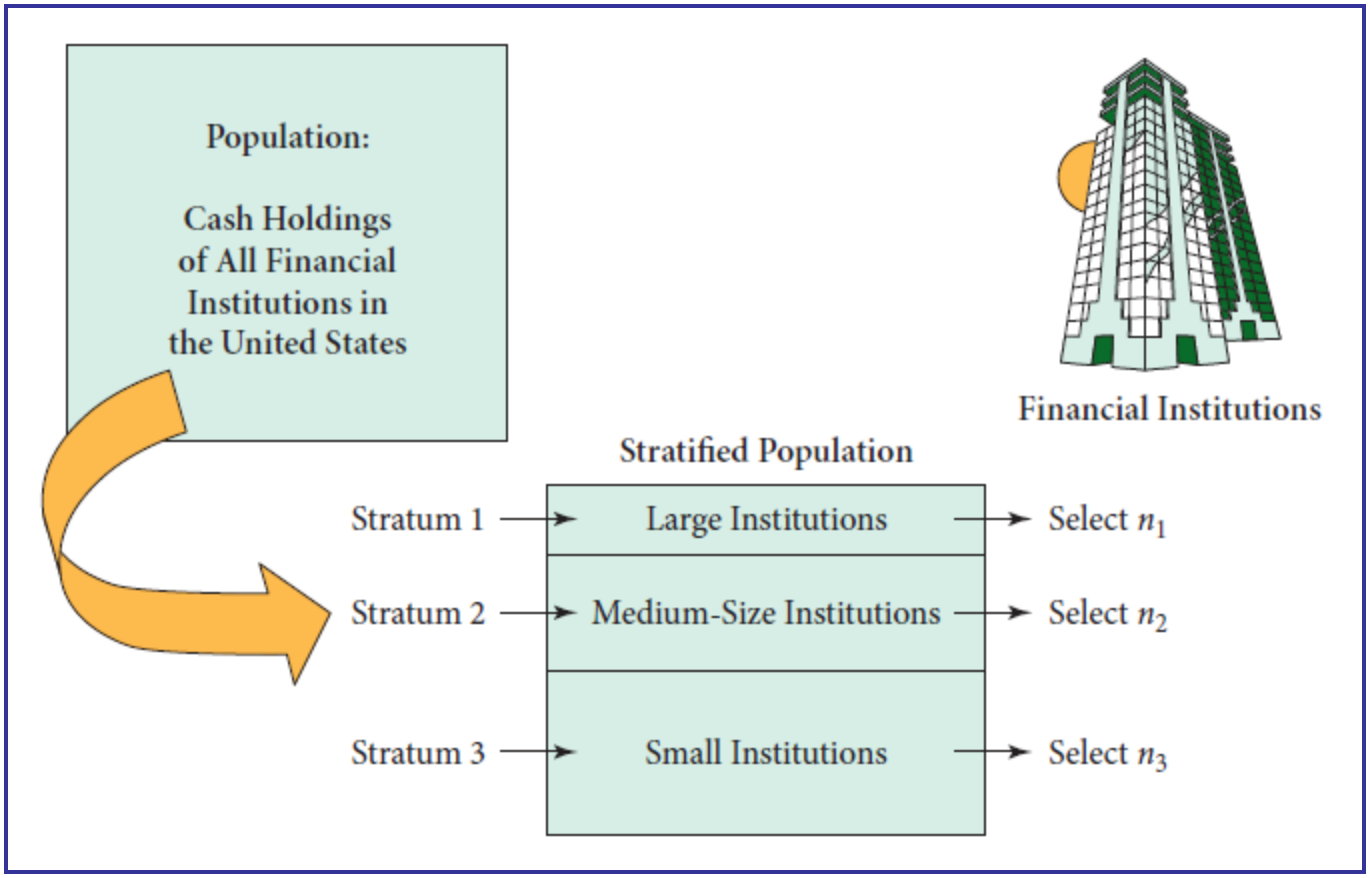


Stratified Random Sampling

- Divide population into subgroups (called *strata*) according to some common characteristic
  - e.g., gender, income level
- Select a simple random sample from each subgroup
- Combine samples from subgroups into one



Stratified Sampling Example



Stratified Sampling

- Advantage:

  If the strata are correctly determined, the data of interest will be fairly homogeneous within each stratum.

  If so, then the sum of the sample sizes needed from each stratum will be less than would have been needed had a simple random sample been selected. Thus sampling costs are reduced using stratified random sampling.

Systematic Random Sampling

- Decide on sample size: n
- Divide ordered (e.g., alphabetical) frame of N individuals into groups of k individuals: $k = N / n$
- Randomly select one individual from the 1st group
- Select every k^{th} individual thereafter

N = 64

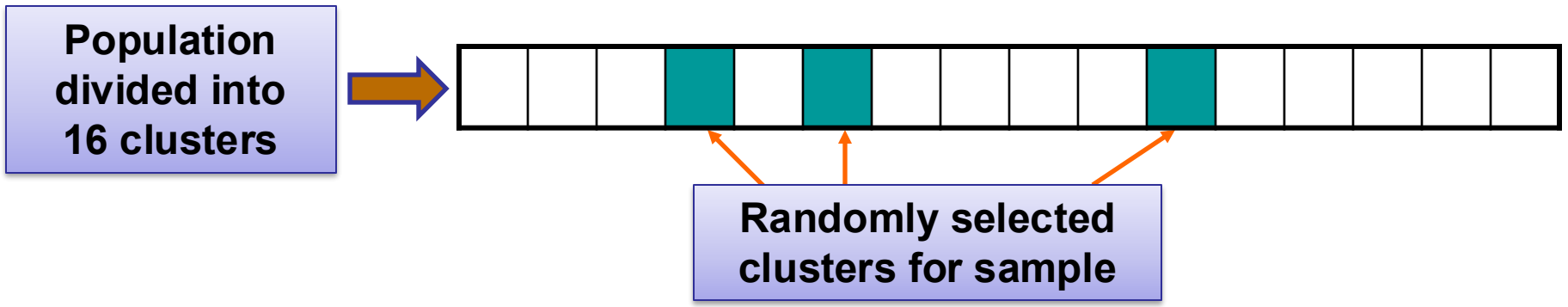
n = 8

k = 8

First Group

Cluster Sampling

- Divide population into several “clusters,” each representative of the population (e.g., county)
- Select a simple random sample of clusters
  - All items in the selected clusters can be used, or items can be chosen from a cluster using another probability sampling technique



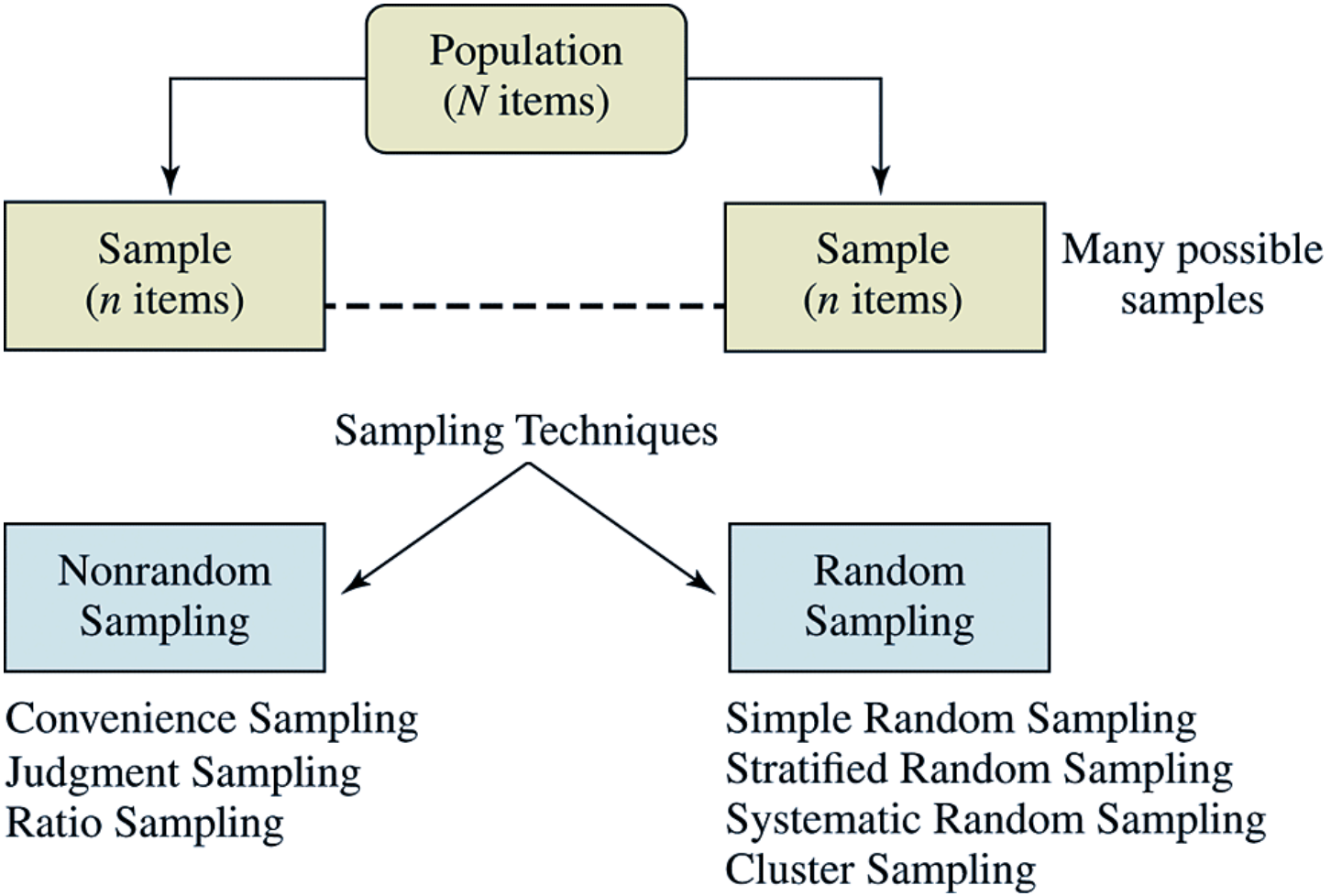
Cluster Sampling

- Example:

  A warehouse manager wishes to select a statistical sample of products that are located across the warehouse in 10 aisles, in bins stacked three high, in 20 stacks per aisle. A total of 10 x 3 x 20 = 600 bins are in the warehouse. Each bin is a cluster.

  Randomly select 5 clusters (bins) and sample each item in the 5 bins selected.

Sampling Methods - Summary



1.4 Data Types and Data Measurement Levels

The starting point in analyzing data is to know what kind of data you have collected.

Data Types

- Quantitative:
  - measurements whose values are inherently numerical
    - discrete (e.g. number of children)
    - continuous (e.g. weight, volume)
- Qualitative:
  - data whose measurement scale is inherently categorical (e.g. marital status, political affiliation, eye color)

Data Types

- Time-Series:
  - a set of consecutive data values observed at successive points in time (e.g. stock price on daily basis for a year)
- Cross-Sectional:
  - A set of data values observed at a fixed point in time (e.g. bank data about its loan customers)

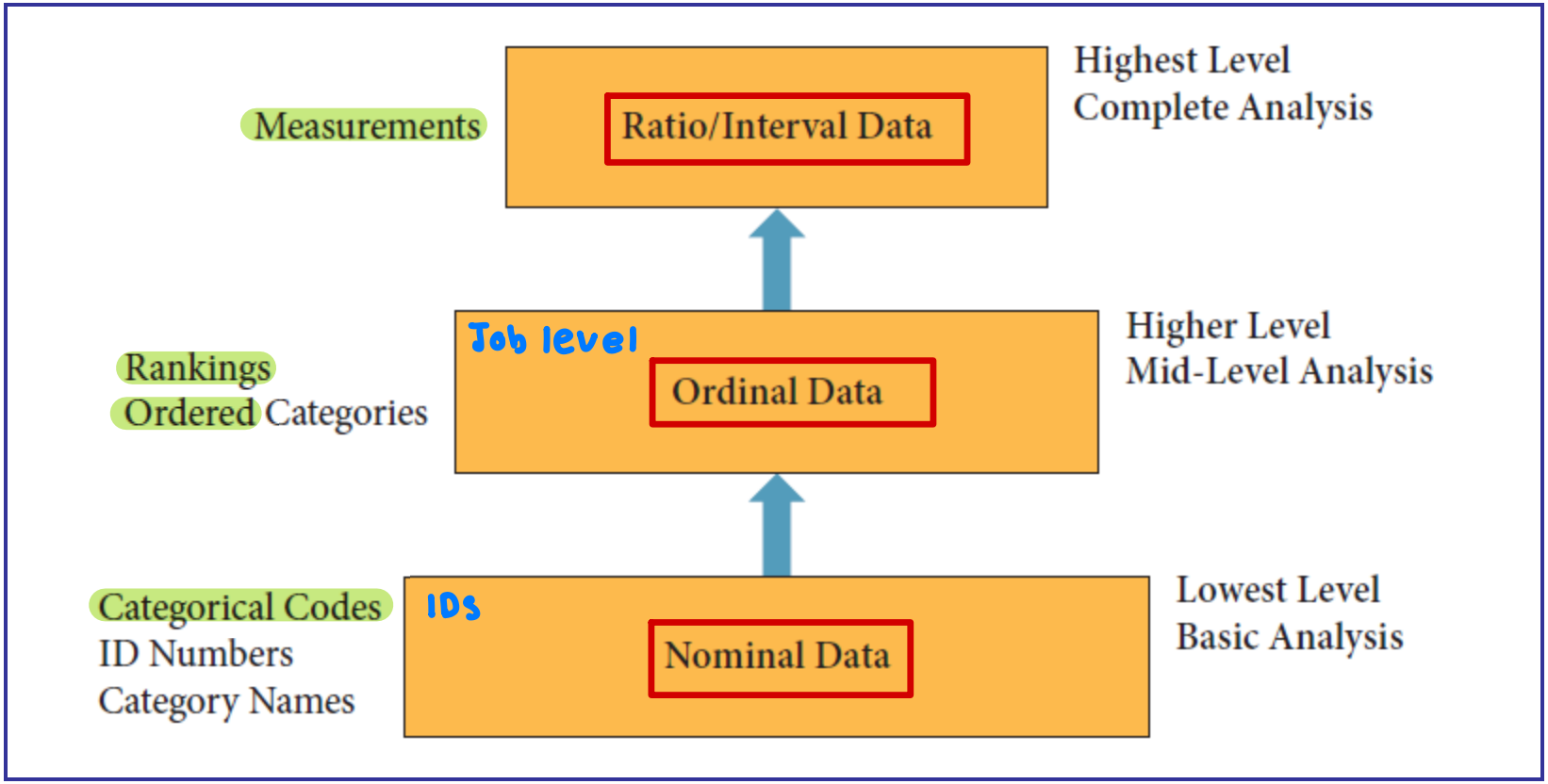
Data Timing Example

	Sales (in $1000s)			
	2009	2010	2011	2012
Atlanta	435	460	475	490
Boston	320	345	375	395
Cleveland	405	390	410	395
Denver	260	270	285	280

Time Series Data

Cross Sectional Data

Data Measurement Levels



Nominal Data Examples

- Student ID Number
- Favorite Color
- Gender
- Marital Status
- Other?

Ordinal Data Examples

- College Class Standing (UG =1, Grad = 2)
- Age Group
  - Under 18 =1
  - 18-30 = 2
  - 31-60 = 3
  - Over 60 =4
- Satisfaction Level
  - VS=1, S=2 Neutral=3, D=4, VD=5

If there's no true 0 it's an interval

Me

Interval/Ratio Data Examples

- Temperature (22^0, 87^0, 0^0, 32^0, 45.92^0,etc)

  Note: Temperature is interval scale because there is no true zero value - 0^0, is like any other temperature because 0^0 does not imply that there is no temperature.

The following are examples of Ratio level data.

What does it mean 0 when it comes to PH?

Me

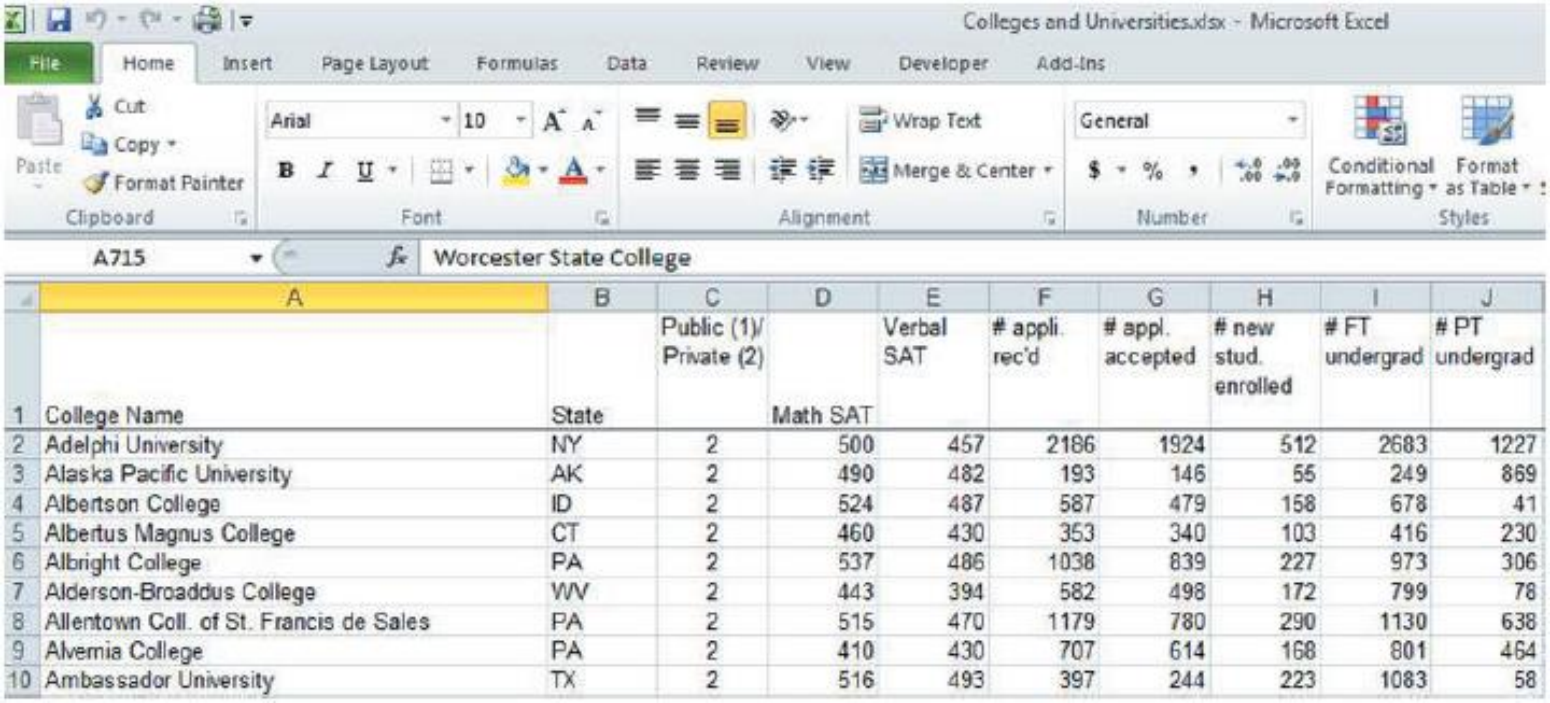
- Weight (1.35 oz., 14.5 lbs., 306 tons, etc.)
- Time (33.05 sec., 1.2 hrs., 7 weeks, etc.)
- Pay Rate per Hour ($33.50, $8.42, $15.45, etc.)
- Interest Rates (4.05%, 7.70%, 5.25%, etc.)

Categorizing Data

- Identify each factor in the data set.
- Determine whether the data are time-series or cross-sectional.
- Determine which factors are quantitative data and which are qualitative data.
- Determine the level of data measurement for each factor.

Data Categorization Example

Cross-sectional data



	A	B	C	D	E	F	G	H	I	J
1	College Name	State	Public (1)/ Private (2)	Math SAT	Verbal SAT	# appli. rec'd	# appl. accepted	# new stud. enrolled	# FT undergrad	# PT undergrad
2	Adelphi University	NY	2	500	457	2186	1924	512	2683	1227
3	Alaska Pacific University	AK	2	490	482	193	146	55	249	869
4	Albertson College	ID	2	524	487	587	479	158	678	41
5	Albertus Magnus College	CT	2	460	430	353	340	103	416	230
6	Albright College	PA	2	537	486	1038	839	227	973	306
7	Alderson-Broaddus College	WV	2	443	394	582	498	172	799	78
8	Allentown Coll. of St. Francis de Sales	PA	2	515	470	1179	780	290	1130	638
9	Alvernia College	PA	2	410	430	707	614	168	801	464
10	Ambassador University	TX	2	516	493	397	244	223	1083	58

Qualitative, nominal-level data

Quantitative, interval, ratio-level data

Data Types and Data Measurement Levels - Summary