Statistics

Chapter 1: Introduction

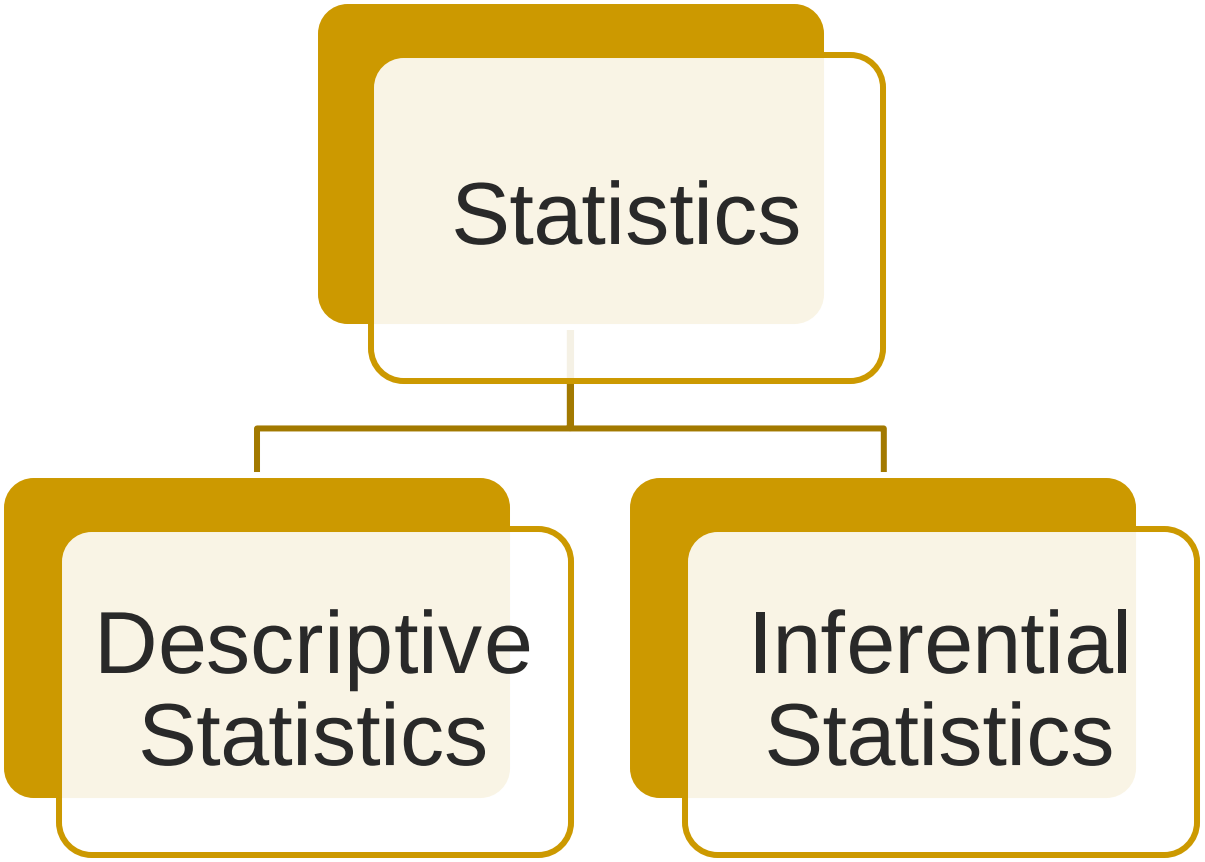
Where We're Going

- Introduction to the field of statistics
- How statistics applies to real-world problems
- Establish the link between statistics and data
- Differentiate between population and sample data
- Differentiate between descriptive and inferential statistics

1.1: The Science of Statistics

- **Statistics** is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing and interpreting numerical information.

1.2: Types of Statistical Applications



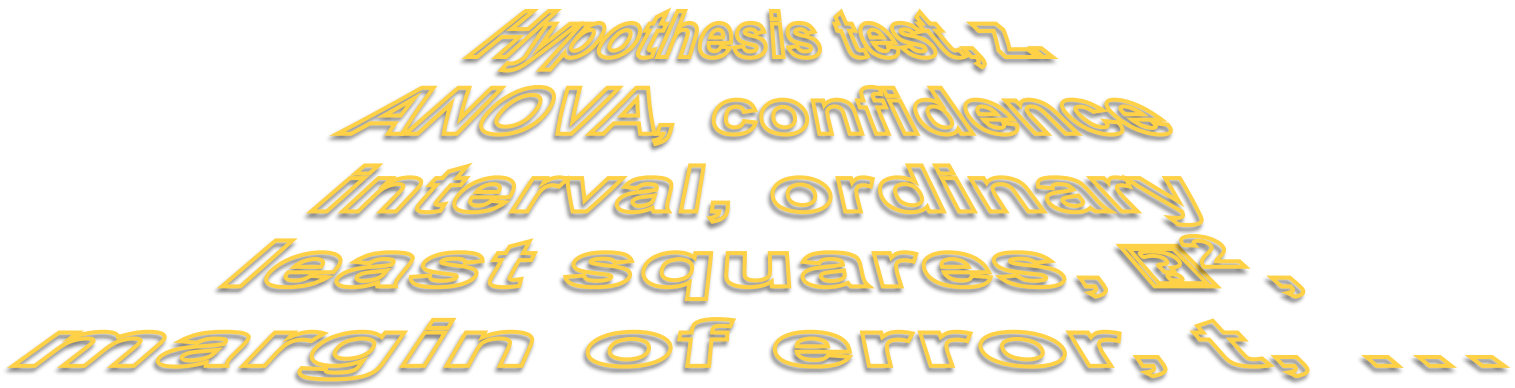
1.2: Types of Statistical Applications

- **Descriptive statistics** utilizes numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in a data set and to present that information in a convenient form.

Average, spread, range, frequency, histogram, median, scatter plot, mode, interquartile range,...

1.2: Types of Statistical Applications

- **Inferential statistics** utilizes sample data to make estimates, decisions, predictions or other generalizations about a larger set of data.



1.3: Randomness and Variability

- Experiments & Processes Are Not Deterministic

- Statistical techniques are useful for describing and understanding variability.

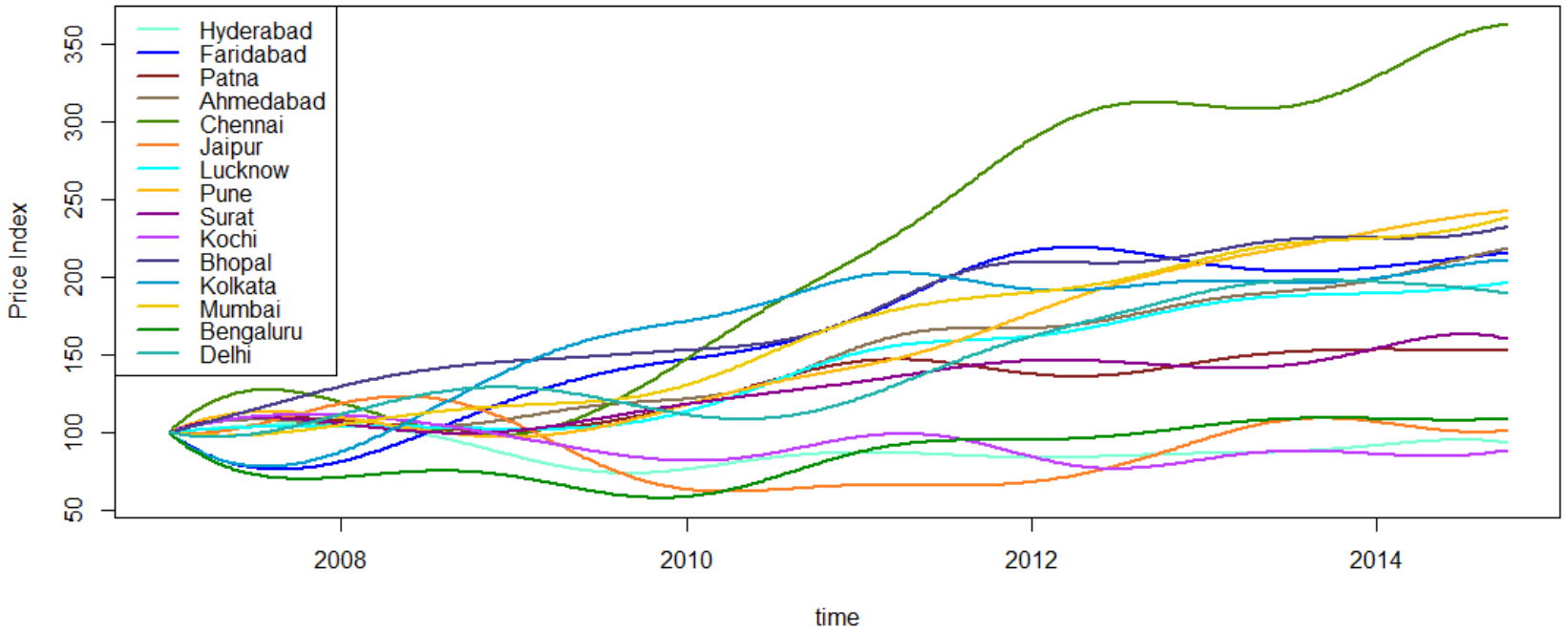
1.3: Randomness and Variability

- By variability, we mean successive observations of a system or phenomenon do ***not*** produce exactly the same result.

- Statistics gives us a framework for describing this variability and for learning about potential sources of variability.

1.3: Randomness and Variability

An Example of Variability



Price Index (Residex) of different cities with 2007 as base.

1.3: Randomness and Variability

- **Systematic errors,** also called fixed errors, are errors associated with using an inaccurate instrument. These errors can be detected and avoided by properly calibrating instruments

1.3: Randomness and Variability

- **Random errors** are generated by a number of unpredictable variations in a given measurement situation.

  Randomness often displays an underlying order that can be quantified, and thus used to advantage.

1.3: Randomness and Variability

Air traffic Example

- Based on joint project with BOBASMA, AAI & ISI (Antar and Deepayan, ISI Delhi)

1.3: Randomness and Variability

Air traffic example

- An airspace is safe, if under the nominal flying condition, the risk of occurrence of a fatal collision between two aircrafts flying in the airspace is smaller that the International standard, known as the Target Level of Safety (TLS), which is 0.000000005 (8 zeros) number of fatal accidents per flight hour.

1.3: Randomness and Variability

Air traffic example

- This means, an airspace is safe, when out of 2 crores flights flying nominally in that airspace in an hour time, on an average we will observe at most 1 fatal accident involving two such flights.

- Indian airspace consists of three large water bodies, namely, Bay of Bengal, Arabian Sea and Indian Ocean and the vast Indian Sub-Continent. It is typically known as the BOBASIO region.

1.3: Randomness and Variability

Air traffic example

- Till 2011, the safety monitoring of this airspace was done by agencies outside of India.
- Then the separation standards were
- 50 NM lateral separation between all the parallel routes;
- 10 minutes/80 NM longitudinal separation between the front and behind aircrafts.

1.3: Randomness and Variability

Air traffic example

- Since 2011, following this work, a reduced horizontal separation has been introduced in sixteen routes in the region:
- 50 NM lateral separation between all the parallel routes;
- 50 NM longitudinal separation between the front and behind aircrafts.

1.3: Randomness and Variability

Computer Engineering

ACM India Doctoral Dissertation Award 2014: Rijurekha Sen (IIT Bombay)
Thesis Title: "Different Sensing Modalities for Traffic Monitoring in Developing Regions"

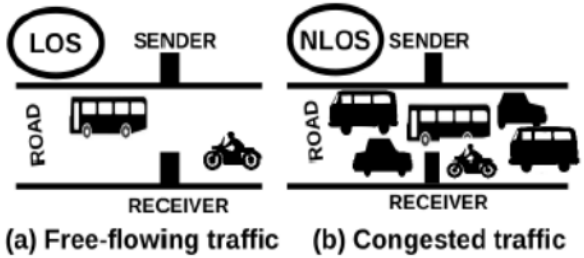


Figure 4: Wireless communication across road

- Kyun Queue: A Sensor Network System To Monitor Road Traffic Queues

1.3: Randomness and Variability

Computer Engineering

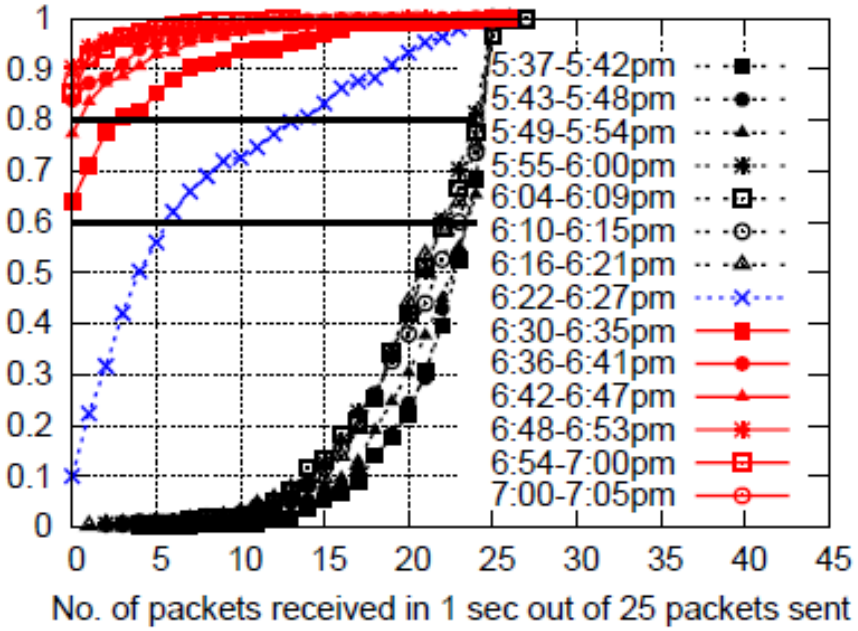


Figure 6: CDF of PRR

1.3: Randomness and Variability

Data Mining

ASA Sections on Statistical Computing: Student paper competition winner 2001: Rituparna Sen, University of Chicago
Jointly with Mark Hansen, Bell Labs

Predicting Web Users' Next Access Based on Log Data

1.3: Randomness and Variability

Data Mining

Table 1. Log Entries for Single Request to cm.bell-labs.com/stat/

Client IP	*Date*	*Time*	*Item*	*Bytes*	*Code*	*Referral*
cs.washington.edu	8/10/2000	15:23:04	/stat/index.html	9269	200	http://search.yahoo.com/search?p=network+stats
cs.washington.edu	8/10/2000	15:23:04	/stat/style/win.css	486	200	http://cm.bell-labs.com/stat/index.html
cs.washington.edu	8/10/2000	15:23:04	/stat/images/logo.gif	6562	200	http://cm.bell-labs.com/stat/index.html
cs.washington.edu	8/10/2000	15:23:05	/stat/images/tukey.gif	4441	200	http://cm.bell-labs.com/stat/index.html
cs.washington.edu	8/10/2000	15:23:05	/stat/images/pbar.gif	145	200	http://cm.bell-labs.com/stat/index.html

Given a visitor's previous activities on the site, we propose models that predict their next page request.

1.3: Randomness and Variability

Data Mining

- If the prediction is reasonably accurate, we might consider "prefetching" the page before the visitor requests it.
- A more conservative use for such predictions would be to simply update the freshness records in a proxy or network cache, eliminating unnecessary Get-If-Modified-Since requests.

1.4: Fundamental Elements of Statistics

- An **experimental unit** is an object about which we collect data.
  - Person
  - Place
  - Thing
  - Event

1.4: Fundamental Elements of Statistics

- An **population** is a set of units in which we are interested.
- Typically, there are too many experimental units in a population to consider *every* one.
- If we can examine every single one, we conduct a **census**.

1.4: Fundamental Elements of Statistics

- A **variable** is a characteristic or property of an individual unit.
- The values of these characteristics will, not surprisingly, vary.
- **Examples:** Highest daily temperature of Bangalore, white blood cell count of a person, time to crashing of a laptop.

1.4: Fundamental Elements of Statistics

How many variables have you measured?

- **Univariate data:** One variable is measured on a single experimental unit.
- **Bivariate data:** Two variables are measured on a single experimental unit.
- **Multivariate data:** More than two variables are measured on a single experimental unit.

1.4: Fundamental Elements of Statistics

Siva Athreya
B.V. Rajarama Bhat
Jishnu Gupta Biswas
Mohan Delampady
Shreedhar
Manish Kumar
Aniruddha C. Naolekar
Anita Naolekar
Suresh Nayak
V.R. Padmawar
C.R.E. Raja
B. Rajeev
Jaydeb Sarkar
Ramesh Sreekantan
B. Sury
Yogeshwaran D
Parthanil Roy
Mathew Joseph
Maneesh Thakur
Soumyashant Nayak

- A **sample** is a subset of the population.
- From all professors in StatMath Bangalore, take a sample of size 3.
- One such sample can be:

  Mohan, Suresh, Jaydeb.

How to draw a random sample in R?

1.4: Fundamental Elements of Statistics

Descriptive Statistics

- The population or sample of interest
- One or more variables to be investigated
- Tables, graphs or numerical summary tools
- Identification of patterns in the data

Inferential Statistics

- Population of interest
- One or more variables to be investigated
- The sample of population units
- The inference about the population based on the sample data
- A measure of reliability of the inference

1.4: Fundamental Elements of Statistics

Exercise

- Suppose1000 cola consumers are given a blind taste test (i.e., a taste test in which the two brand names are disguised). Each consumer is asked to state a preference for brand A or brand B.

a. Describe the population.

b. Describe the variable of interest.

c. Describe the sample.

d. Describe the inference.

1.4: Fundamental Elements of Statistics

a. Because we are interested in the responses of cola consumers in a taste test, a cola consumer is the experimental unit. Thus, the population of interest is the collection or set of all cola consumers.

b. The characteristic that the experimenter wants to measure is the consumer's cola preference as revealed under the conditions of a blind taste test, so cola preference is the variable of interest.

1.4: Fundamental Elements of Statistics

c. The sample is the 1,000 cola consumers selected from the population of all cola consumers.

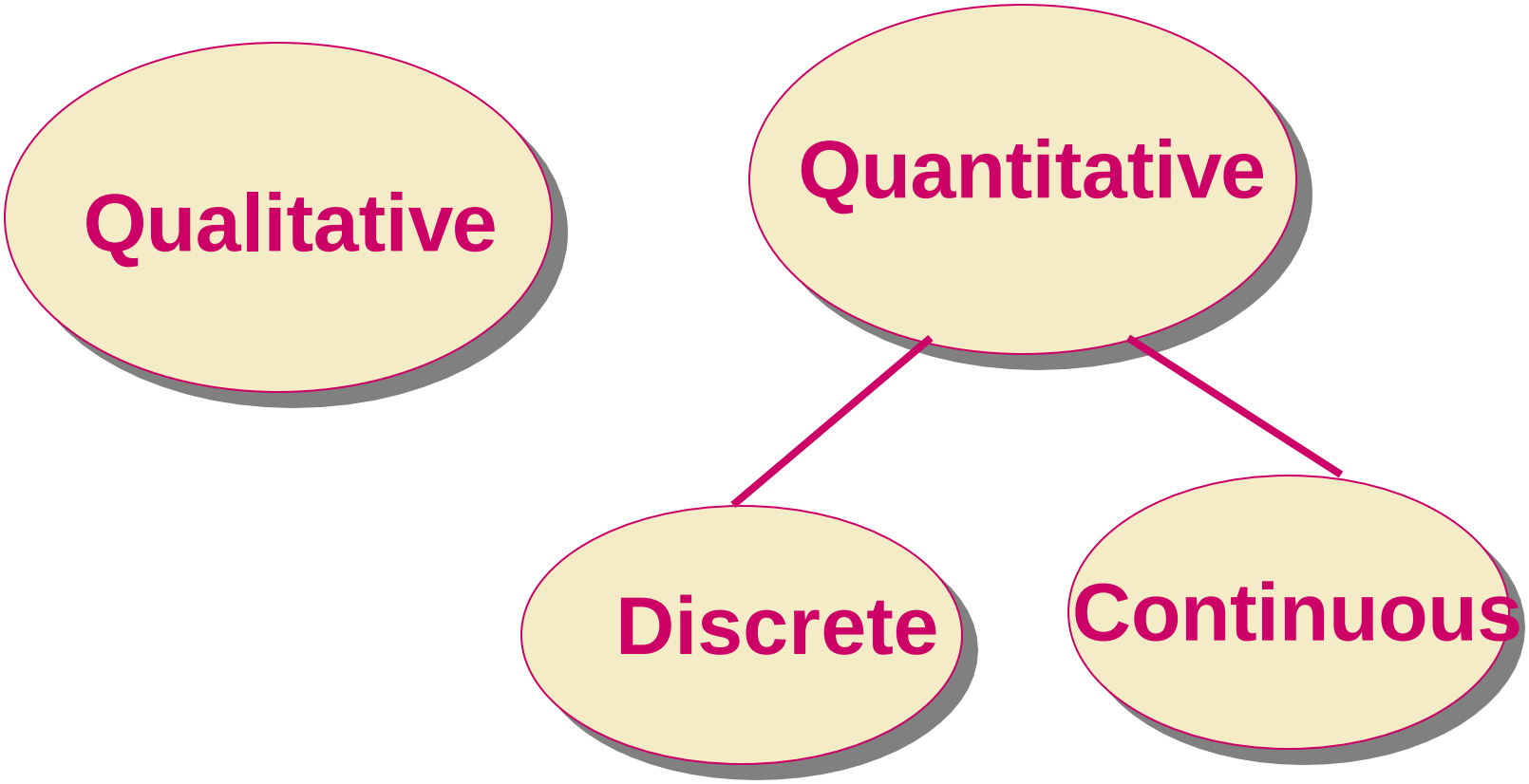
d. The inference of interest is the *generalization* of the cola preferences of the 1,000 sampled consumers to the population of all cola consumers. In particular, the preferences of the consumers in the sample can be used to *estimate* the percentage of all cola consumers who prefer each brand.

1.4: Fundamental Elements of Statistics

- A **measure of reliability** is a statement about the degree of uncertainty associated with a statistical inference.

Based on our analysis, we think 56% of soda drinkers prefer Pepsi to Coke, ± 5%.

1.5: Types of Data



1.5: Types of Data

- **Qualitative Data** are measurements that cannot be recorded on a natural numerical scale, but are recorded in categories.
  - Live on/off campus
  - Choice of elective course
  - Gender
  - Favourite IPL team

1.5: Types of Data

- **Quantitative Data** are measurements that are recorded on a naturally occurring numerical scale.
  - Age
  - Marks in exam
  - Salary
  - Cost of books this semester

1.5: Types of Data

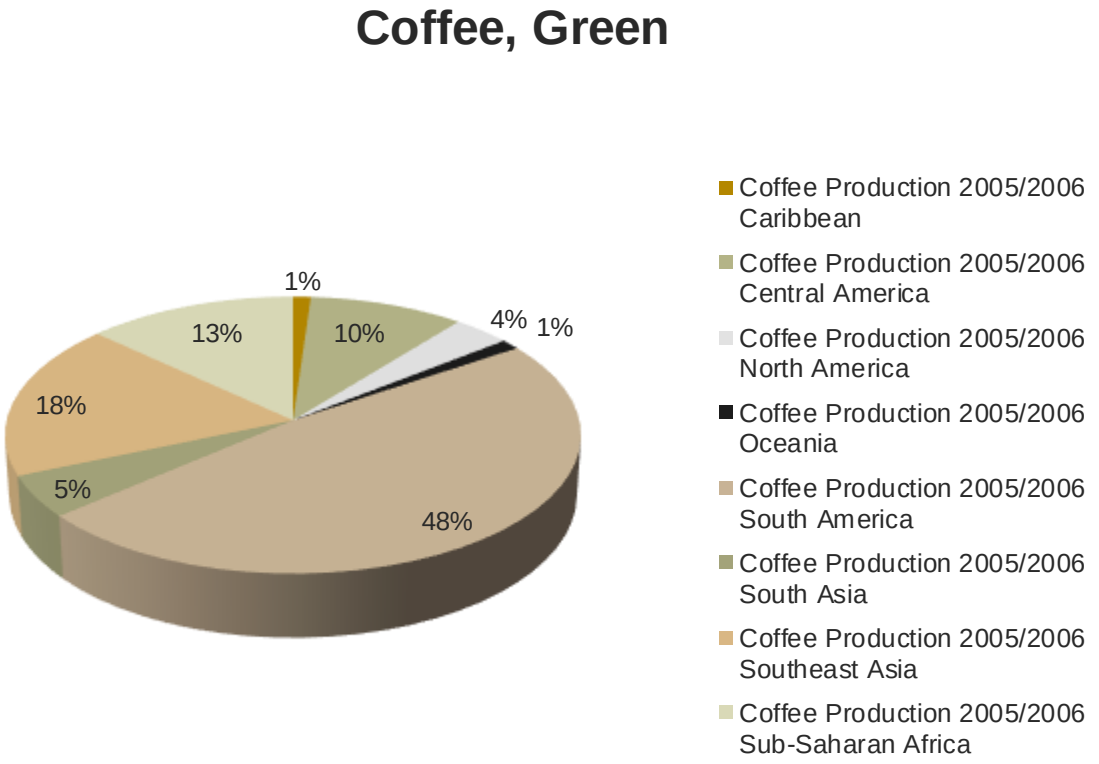
- For each orange tree in a grove, the number of oranges is measured.
  - **Quantitative discrete**

- For a particular day, the number of cars entering a college campus is measured.
  - **Quantitative discrete**

- Time until a light bulb burns out
  - **Quantitative continuous**

1.6: Collecting Data

- Published Source
- Designed Experiment
- Observational Study



SOURCE: United States Department of Agriculture
Foreign Agricultural Service

1.6: Collecting Data

- **Published Source**
  - Journal
  - Book
  - Newspaper
  - Magazine
  - (Reliable) Web Site

1.6: Collecting Data

- **Designed Experiment**
  - Strict control over the experiment and the units in the experiment

1.6: Collecting Data

- **Designed Experiment**

A designed experiment in the medical field involving human subjects is referred to as a *clinical trial*. One recent clinical trial was designed to determine the potential of using aspirin in preventing heart attacks. Volunteers were randomly divided into two groups—the *treatment* group and the *control* group.

1.6: Collecting Data

- **Designed Experiment**

Each volunteer in the treatment group took one aspirin tablet a day for one year, while the volunteers in the control group took an aspirin-free placebo made to look identical to an aspirin tablet. Because the volunteers did not know which group, treatment or control, they were assigned to, such a clinical trial is called a *blind study*.

1.6: Collecting Data

- **Observational Study**
  - Observe units in natural settings
  - No control over behavior of units

1.6: Collecting Data

- **Example of Observational Study: Survey**

How do consumers feel about using the Internet for online shopping? To find out, United Parcel Service (UPS) commissioned a nationwide survey of 5,118 U.S. adults who had conducted at least two online transactions in 2015. One finding from the study is that 74% of online shoppers have used a smartphone to do their shopping.

1.6: Collecting Data

- A **representative sample** exhibits characteristics typical of those possessed by the target population.
- A **random sample** of *n* units is selected in such a way that every different sample of size *n* has the same chance of being selected.

1.7:Sources of bias and error

- **Statistical thinking** involves applying rational thought and the science of statistics to critically assess data and inferences.

1.7: Sources of bias and error

- **Selection bias** results when a subset of the experimental units in the population have been excluded so that these units have no chance of being selected in the sample.
- Eg, Online surveys will leave out the section of population who do not have access to the internet.

1.7: Sources of bias and error

- **Response bias:** Responses obtained are systematically biased
- This is common with sensitive issues: “Do you wear a mask when going out?”

1.7: Sources of bias and error

- **Nonresponse bias** results when the researchers conducting a survey or study are unable to obtain data on all experimental units selected for the sample.
- This becomes important if the non-respondents and respondents differ greatly on the issue at hand.
- For eg, for a survey of passengers regarding service of an airline, passengers who are dissatisfied are more like to respond than satisfied customers.

1.7: Sources of bias and error

- **Measurement error** refers to inaccuracies in the values of the data recorded. In surveys, this kind of error may be due to ambiguous or leading questions and the interviewer's effect on the respondent.
- This can happen due to leading questions as: "Do you prefer Candidate X, father of two and university professor, or Candidate Y, who got the nomination despite his shady past?"