Lecture 25

Types of Data, Sampling

STAT 225 Introduction to Probability Models
April 16, 2104

Whitney Huang
Purdue University

Agenda

PURDUE UNIVERSITY

1. Definitions
2. Types of Data
3. Sampling

Probability vs. Statistics

PURDUE UNIVERSITY

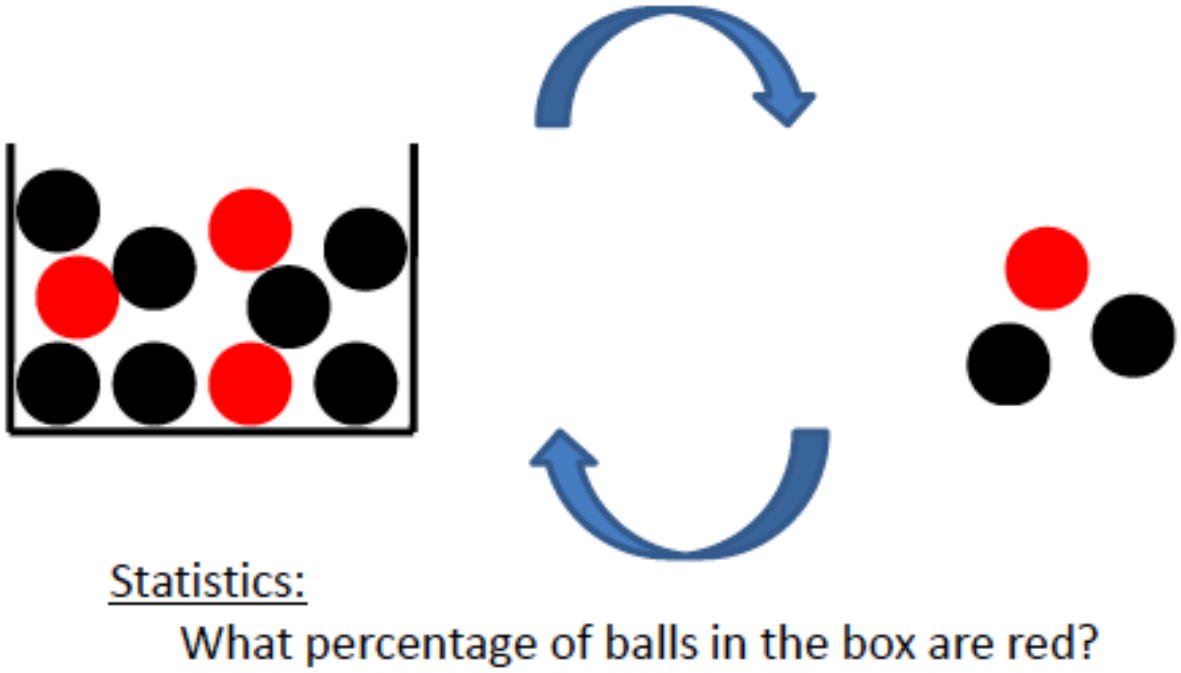

Figure : Taken from JHU Statistical Computing by Hongkai Ji

Definitions

- Statistics is the science of collecting, analyzing, presenting, and interpreting data

PURDUE UNIVERSITY

Definitions

- Statistics is the science of collecting, analyzing, presenting, and interpreting data
- Data set: all the data collected in a particular study

PURDUE UNIVERSITY

Definitions

- Statistics is the science of collecting, analyzing, presenting, and interpreting data
- Data set: all the data collected in a particular study
- Elements are the individual entities of a data set

PURDUE UNIVERSITY

Definitions

- Statistics is the science of collecting, analyzing, presenting, and interpreting data
- Data set: all the data collected in a particular study
- Elements are the individual entities of a data set
- A variable is a characteristic of interest for the elements

PURDUE UNIVERSITY

PURDUE UNIVERSITY

Definitions

- Statistics is the science of collecting, analyzing, presenting, and interpreting data
- Data set: all the data collected in a particular study
- Elements are the individual entities of a data set
- A variable is a characteristic of interest for the elements
- An observation is the set of measurements obtained for a particular element

Types of variables

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement

Types of variables

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement
  - Nominal: order does not matter e.g Gender

Types of variables

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement
  - Nominal: order does not matter e.g Gender
  - Ordinal: order does matter e.g. Education levels

Types of variables

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement
  - Nominal: order does not matter e.g Gender
  - Ordinal: order does matter e.g. Education levels
- Quantitative variable: has numeric values that indicate how much or how many of something. Quantitative data uses either the **interval** or **ratio** scale

Types of variables

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement
  - Nominal: order does not matter e.g Gender
  - Ordinal: order does matter e.g. Education levels
- Quantitative variable: has numeric values that indicate how much or how many of something. Quantitative data uses either the **interval** or **ratio** scale
  - Interval: difference of quantities that are meaningful but ratios of quantities that cannot be compared e.g. temperature with the Celsius scale

Types of variables

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement
  - Nominal: order does not matter e.g Gender
  - Ordinal: order does matter e.g. Education levels
- Quantitative variable: has numeric values that indicate how much or how many of something. Quantitative data uses either the **interval** or **ratio** scale
  - Interval: difference of quantities that are meaningful but ratios of quantities that cannot be compared e.g. temperature with the Celsius scale
  - Ratio: ratios of quantities that are meaningful e.g. Height

Cross-sectional vs. Time series data

We have two types of data set based on how the data were collecting

- Cross-sectional: data collected at the same or approximately the same point in time
- Time series: data collected over several time periods

Example 61

PURDUE UNIVERSITY

Grade	Major	GPA	Credit hours
Sophomore	Psychology	3.14	30
Senior	Spanish	2.89	105
Senior	Religion	3.01	99
Freshman	Philosophy	2.45	12

1. How many elements are in the data set?
2. How many variables are in the data set?
3. What type of variable is each variable in the data set (be sure to answer both qualitative or quantitative as well as nominal, ordinal, interval, or ratio).

Example 61 cont'd

Solution.

1. 4 elements in total
2. 4 variables in this data set. They are Grade, Major, Credit hours, and GPA
3. Grade: qualitative (ordinal); Major: qualitative (nominal); GPA: quantitative (interval); Credit hours: quantitative (ratio)

Example 62

For this example, answer what type of variable each of the following are

1. Smoking status
2. Income
3. Level of satisfaction
4. clothing size (s, m, l, xl)
5. time taken to run a mile

PURDUE UNIVERSITY

Example 62 cont'd

Solution.

1. qualitative (nominal)
2. quantitative (ratio) or qualitative (ordinal)
3. qualitative (ordinal)
4. qualitative (ordinal)
5. quantitative (ratio)

Example 63

For this problem, state whether the variables included are cross-sectional or time series

1. Current GPAs of Purdue Statistics Graduate Students
2. GPA of Sanvesh during his time at Purdue
3. Value of Gordan Gecko's portfolio over the previous 3 years
4. Value of all portfolio's at Charles Schwaab in January 2008
5. Total salary of the LA Lakers throughout the 1990s
6. Salaries of all NBA teams in 1994.

Example 63 cont'd

Solution.

1. cross-sectional
2. time series
3. time series
4. cross-sectional
5. time series
6. cross-sectional

Statistical Sampling

In Statistics, sampling is procedure to select a subset from a statistical population that is representative of the population. There are several types of sampling as follows:

- Simple random sampling (SRS): a sample selected such that each element in the population has the same probability of being selected

Statistical Sampling

In Statistics, sampling is procedure to select a subset from a statistical population that is representative of the population. There are several types of sampling as follows:

- Simple random sampling (SRS): a sample selected such that each element in the population has the same probability of being selected
- Stratified random sample: elements in the population are first divided into groups and a simple random sample is then taken from each group

Sampling cont'd

- Probability sampling: elements in the population are selected with a known probability of being included in a sample

PURDUE UNIVERSITY

Sampling cont'd

- Probability sampling: elements in the population are selected with a known probability of being included in a sample
- Cluster sampling: the elements in the population are first divided into separate groups called clusters and then a simple random sample of the clusters is taken that all elements in a selected cluster are part of a sample

PURDUE UNIVERSITY

Sampling cont'd

- Probability sampling: elements in the population are selected with a known probability of being included in a sample
- Cluster sampling: the elements in the population are first divided into separate groups called clusters and then a simple random sample of the clusters is taken that all elements in a selected cluster are part of a sample
- Systematic sampling: randomly select one of the first k elements from the population and then every k_{th} element thereafter is picked

Sampling cont'd

- Probability sampling: elements in the population are selected with a known probability of being included in a sample
- Cluster sampling: the elements in the population are first divided into separate groups called clusters and then a simple random sample of the clusters is taken that all elements in a selected cluster are part of a sample
- Systematic sampling: randomly select one of the first k elements from the population and then every k_{th} element thereafter is picked
- Convenience sampling: elements selected from the population on the basis of convenience

Sampling cont'd

- Probability sampling: elements in the population are selected with a known probability of being included in a sample
- Cluster sampling: the elements in the population are first divided into separate groups called clusters and then a simple random sample of the clusters is taken that all elements in a selected cluster are part of a sample
- Systematic sampling: randomly select one of the first k elements from the population and then every k_{th} element thereafter is picked
- Convenience sampling: elements selected from the population on the basis of convenience
- Judgment sampling: elements are selected from the population based on the judgment of the person doing the study.