Introduction to Statistics

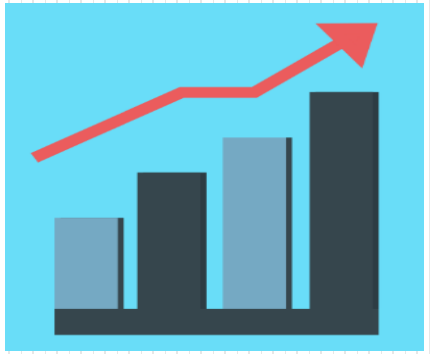
Dr. Farman Ali

Assistant Professor

DEPARTMENT OF SOFTWARE

SEJONG UNIVERSITY

Lecture-2



Course Syllabus

- Introduction to the course
- Sampling and data presentation
- Basic of probability
- Distributions
- Confidence intervals
- Hypothesis testing
- Correlation and simple linear regression
- Multiple regression

Outline

- Basic concepts
- Sampling
- Data Collection
- Data Representation

Some Basic Concepts

Variable

- A variable is a characteristic or attribute that can assume different values.
- A characteristic that changes or varies over time for different individuals or objects under consideration.

Examples

- Hair color
- White blood cell count
- Time to failure of a computer component.

Data

- An **experimental unit** is the individual or object on which a variable is measured.
- A **measurement** results when a variable is actually measured on an experimental unit
- A set of measurements, called **data**, can be **sample** or **population**.

Some Basic Concepts

Examples 1

Variable:

- Hair color

Experimental unit:

- Person

Typical Measurements

- Brown, black, blonde, etc.

Examples 2

Variable:

- Time until a light bulb burns out

Experimental unit:

- Light bulb

Typical Measurements

- 1500 hours, etc.

How many variables we are going to measured?

- Univariate data: One variable is measured on a single experimental unit (individual or object).
- Bivariate data: Two variables are measured on a single experimental unit (individual or object).
- Multivariate data: More than two variables are measured on a single experimental unit (individual or object).

Types of Variables

- Two main types of variables: Qualitative variables and Quantitative variables

Qualitative Variables: Whose range consists of qualities or attributes of objects.

Examples

- Hair color (black, brown, white)
- Make of car (Suzuki, Honda, etc.)
- Gender (male, female)
- Grades: (A, B, C, D, F)
- Level of satisfaction: (Very satisfied, satisfied, somewhat satisfied)
- Model of transportation: (Car, University Bus, Bike, Cycle etc.)

Types of Variables

- Two main types of variables: Qualitative variables and Quantitative variables

Quantitative Variables: Whose range consists of a numerical measurement characteristics of objects under study.

Examples

- Marks of students of statistics class in quiz 1
- Ages of students
- Salaries of faculty members
- Number of cars owned by faculty of Software dept

Types of Qualitative Variables

- **Nominal variable:** A qualitative variable that describes name an element of a population.

Examples

- Hair color (black, brown, white)
- Gender (male, female)
- Make of car (Suzuki, Honda, etc.)

Note: order of variables doesn't matter.

- **Ordinal variable:** A qualitative variable that incorporates an order position or ranking.

Examples

- Grades: (A, B, C, D, F)
- Level of Satisfaction: (Very satisfied, satisfied, somewhat satisfied)

Types of Quantitative Variables

- **Discrete variable:** A quantitative variable that can assume a countable number of values.

Examples

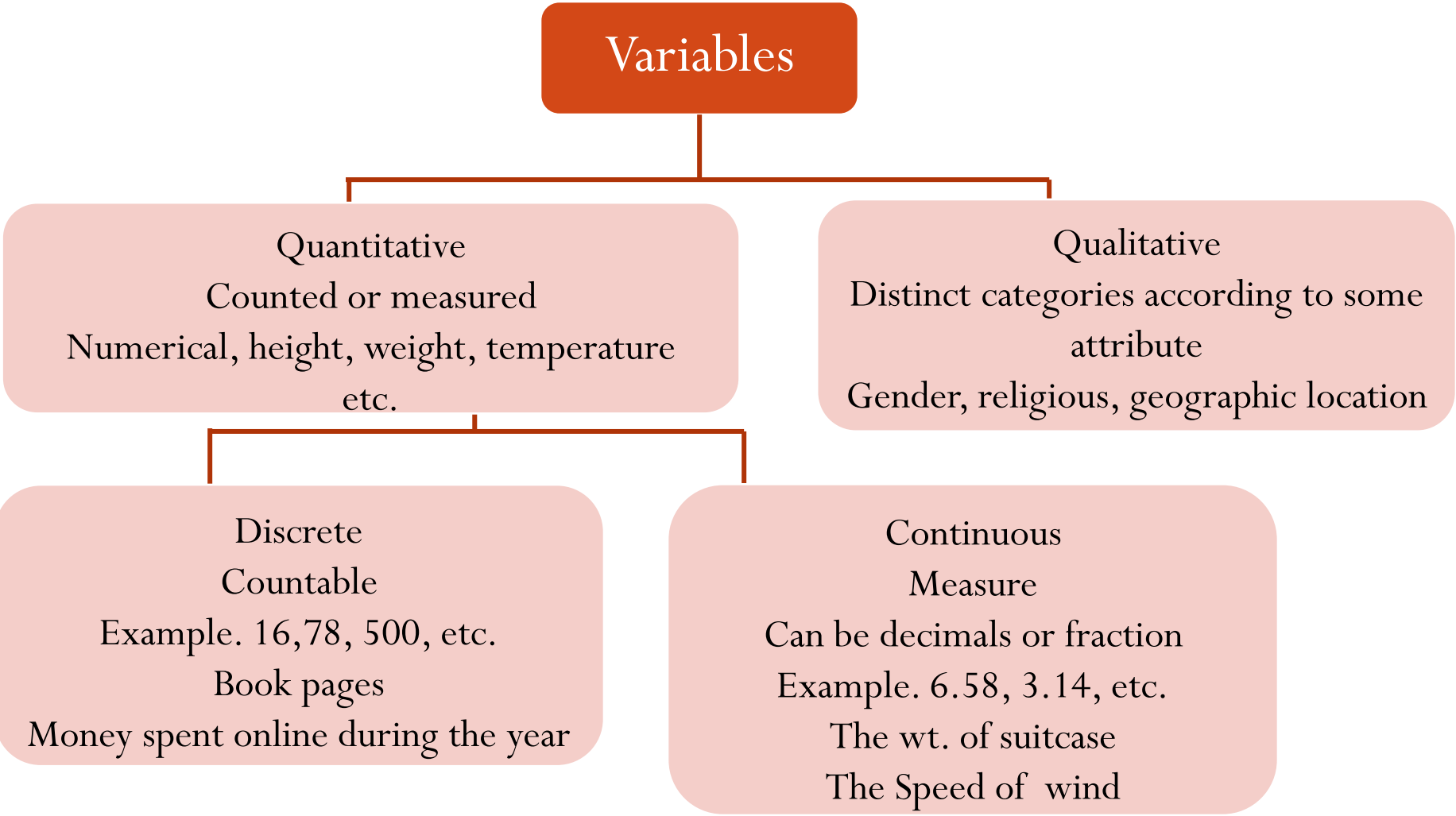
- Number of course for which you are currently registered.
- Total number of students in a class.
- Number of TV sets sold by a company

- **Continuous variable:** A quantitative variable that can assume a uncountable number of values.

Examples

- Weight of books
- Height of the students
- Amount of rain fall

Variables and types of data



Random Variables

➢ Variables whose values are determined by chance are called random variables.

Examples

- Automobile insurance, claim suppose 5% every year.
- Population: Census example
- Sample: may be biased.

Recorded Values and boundaries

Variable	Recorded Value	Boundaries
Length	18 centimeters (cm)	17.5-18.5 cm
Temperature	76^o Fahrenheit (^{o}F)	75.5-76.5 ^{o}F
Time	0.24 second (sec)	0.235-0.245 sec
Mass	3.8 grams (g)	3.75-3.85 g

How variables are categorized, counted or measured.

Examples

- Can the data values be ranked, 1st place, 2nd place and so on.
- Can the data be organized into specific categories (rural, urban, etc.)
- Can the data be measured (height, temperature, time, length, IQ's etc.)

These types of classification needs measurement scale.

Measurement Scales

- These are simply ways to categorize different types of variables.
- The values of variable can be classified by measurement scale.

Four Scales of Measurement:

- Nominal Scale
- Ordinal Scale

For Qualitative Data

- Interval Scale
- Ratio Scale

For Quantitative Data

Nominal scale

- Nominal- categorical (names), No ranking or no order.
- Nominal scales are used for labeling variables, without any quantitative value.
- Classifies data into distinct categories where no ranking is implied. All we can say is that one is different from the other.

Examples

- Note that all of these scales are mutually exclusive (no overlap) and none of them have any numerical significance. No ranking or no order.

What is you gender

- M- Male
- F- Female

What is your hair color?

- 1- brown
- 2- black
- 3- Gray

Ordinal scale

- Ordinal-nominal, plus can be ranked (order)
- Data measured at this level can be placed into categories, and these categories can be ordered or ranked.

Examples

- Letter Grades A, B, C, D…,
- Guest speaker speech, Excellent, average, poor
- Restaurant Services, '1' for poor, '2' for average, '3' for very good and '4' for excellent.
- Student ranking 1^{st} , 2^{nd} , 3^{rd},…
- Faculty Ranks: Professor, Associate Professor, Assistant Professor, Lecturer

interval scale

- Interval- Ordinal,, plus intervals are consistent

Examples

- IQ score: meaningful difference between 107 and 108 IQ score.
- Temperature: meaningful difference between 15C and 16C.
- Distance: meaningful difference between 30km-40km

- Note: there is no true ZERO. IQ tests do not measure people who have no intelligence.
- Temperature 0C does not mean no heat at all. Etc.
- We can't say that temperature of 30^0 C is twice as hot as a temperature of 15^0C.

Ratio scale

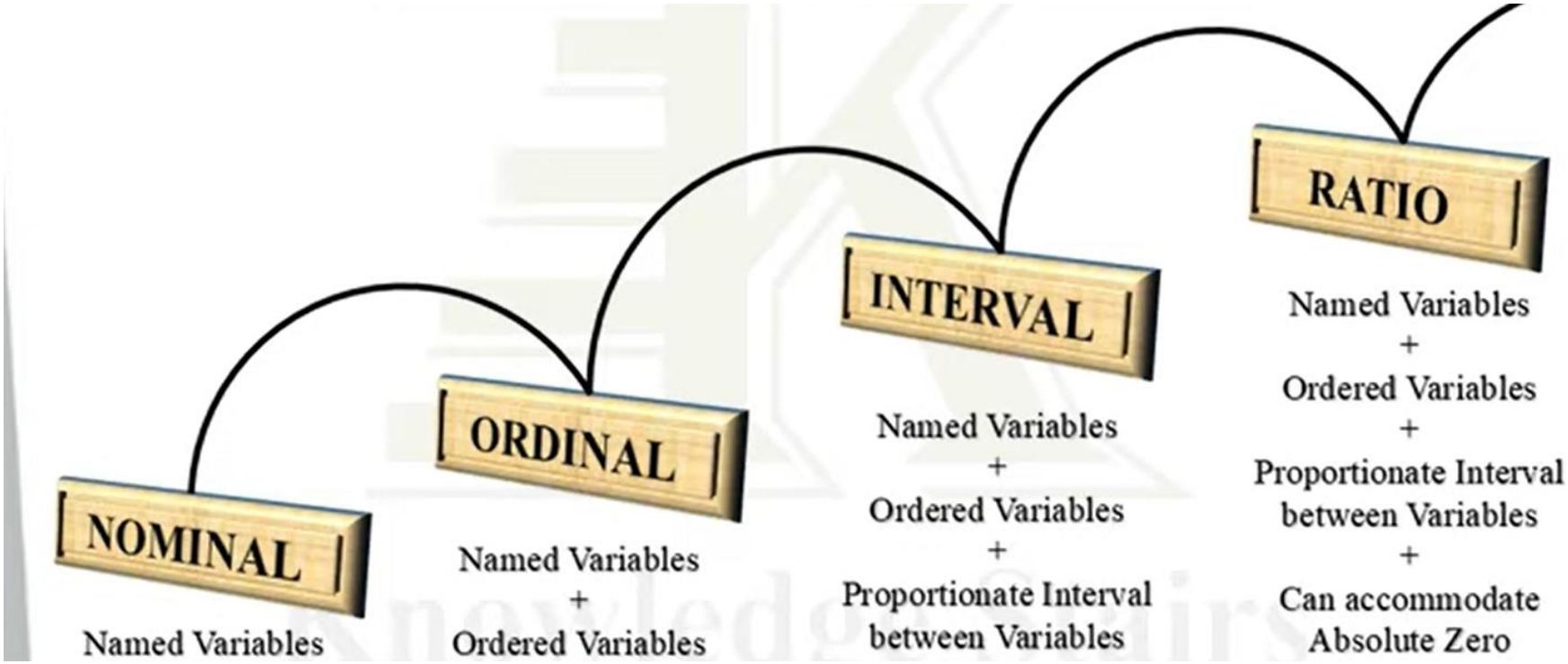
- Ratio- interval, plus ratios are consistent, true zero
- The zero in the scale makes this type of measurement unlike the other types of measurement, although the properties are similar to that of the interval level of measurement.

Examples

- It is used to measure height, phone calls received, area, weight etc. if one-person weight is 100 pounds, while other is 50 pounds, then the ratio is 2:1

- The researcher should note that among these levels of measurement, the nominal level is simply used to classify data, whereas the levels of measurement described by the interval level and the ratio level are much more exact.
- Remember -> Nominal, ordinal, and interval data are discrete and Ratio data are continuous.

Some Basic Concepts



Variables and types of Data

Determine the measurement level.

Variable	Nominal	Ordinal	Interval	Ratio	Level
Hair Color					
Zip Code					
Letter Grade					
IQ Score					
Height					
Age					
Temperature (F)					

Variables and types of Data

Determine the measurement level.

Variable	Nominal	Ordinal	Interval	Ratio	Level
Hair Color	Yes	No			Nominal
Zip Code	Yes	No			Nominal
Letter Grade	Yes	Yes	No		Ordinal
IQ Score	Yes	Yes	Yes	No	Interval
Height	Yes	Yes	Yes	Yes	Ratio
Age	Yes	Yes	Yes	Yes	Ratio
Temperature (F)	Yes	Yes	Yes	No	Interval

Data collection

- **Why Statisticians collect data?**
- Data can be used to describe situations or events.
- Manufacturer can make a smart marketing strategy if he knows the purchasing power of the consumers.
- With the help of data buyers can make an intelligent decisions, what stock to buy etc.
- For these purposes statisticians need data. There are different way to collect data. The most common way is to collect data through surveys.

- Telephone Survey
- Mailed questionnaire
- Personal interviews

Telephone Surveys

- **Advantages**
- Telephone surveys are less expensive
- People are more frank, to express their judgement, as no face to face communication.

- **Disadvantages**
- No all people have chanced to survey, as sometimes they don't receive the phone, or they are at work, when the call was made.
- Some people have unlisted numbers, or don't have phone.

Mailed Questionnaire

- **Advantages**
- Mail Survey are less costly than personal survey, and can cover the wide area than telephone survey.
- People may remain anonymous if they want.

- **Disadvantages**
- Low number of replies.
- Incorrect responses.
- May trouble to read or understand question. etc.

Personal Survey / Interview

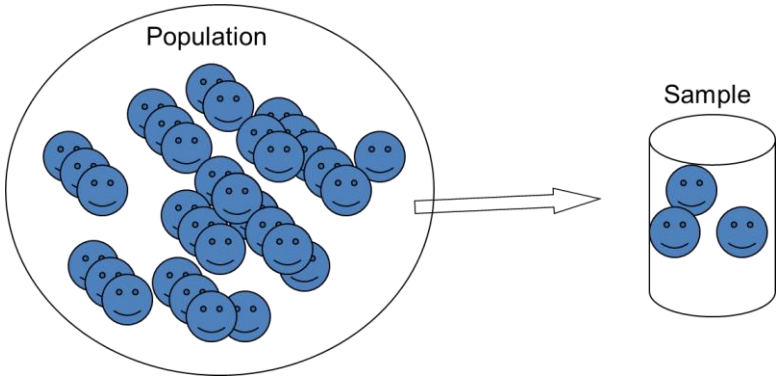
- **Advantages**
- Detail answer from the respondents

- **Disadvantages**
- More costly.
- Unfair choice of respondents.

- Researchers use sample to collect data, which saves time and money, but sample may be unfair or biased.
- To get sample that are unbiased, that give each subject in the population an alike chance of being selected, statisticians practice four methods of sampling.

Algebraic Form of Data

- **Population**: entire collection of objects or outcomes
  - Heights of all students at a university
- **Sample**: subset of a population
  - Students in Statistic course
  - Students on the basketball team
  - 100 randomly chosen students



Sample of size n: $X_1, X_2, X_3, \ldots, X_n$

Four types of sampling

➢ Simple random sampling

- A simple random sample of size n is a sample chosen by a method in which each collection of *n* population items is equally likely to comprise the sample.

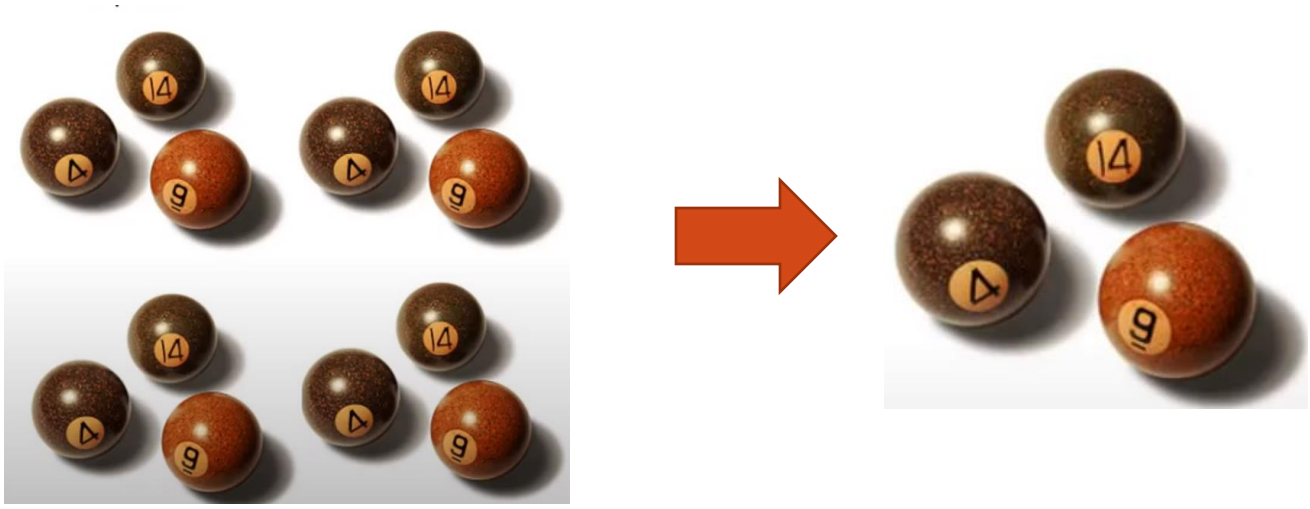
- Assign each member of the population a unique ID, generate random numbers to choose which ones are sampled.

Lottery

Four types of sampling

➢ Simple random Sampling

- Random sample is a subset of the population in which each member of the subset has an equal probability of being selected. An example of a random sample would be the names of 30 employees being chosen out of a hat from a company of 300 employees.



Four types of sampling

➢ Simple of Convenience- mall Surveys

- Convenience sampling(also known as availability sampling) is a particular kind of non-probability sampling technique that depend on data collection from population members who are conveniently available to contribute in study. Facebook polls or questions can be stated as a common example for convenience sampling.
- Can be biased toward certain members of the population
- Used when it is not feasible to draw a random sample.

street questionnaires

Four types of sampling

Random and Convenience Sampling

- **Population:** 1000 steel rods manufactured by a machine in the last hour
- **Simple random sampling:** 50 rods chosen by a random number generator.
- **Sample of convenience:** 50 most recently produced rods

Four types of sampling

➢ Systematic Sampling

- Sample members from a larger population are selected according to a random starting point and a fixed periodic interval. This interval, called the sampling interval, is calculated by dividing the population size by the desired sample size.

1	**2**	**3**	**4**
5	6	7	8
9	10	11	12
13	14	15	16

➢ Suppose you have a number of students lined up in a row:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

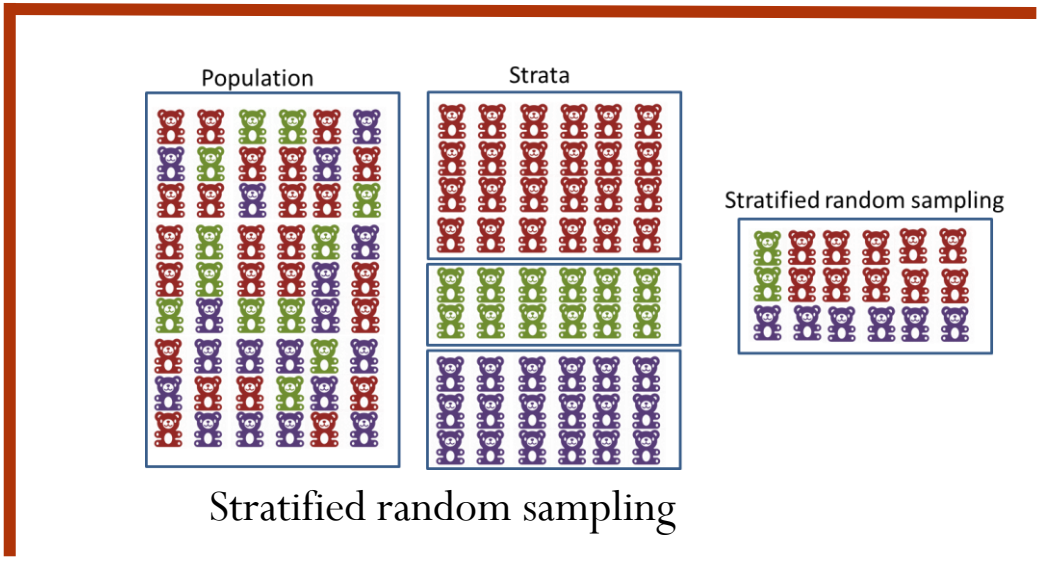
16/4=4

➢ Here we might take a sample every 4 elements, or 1 in 4 elements from the population. (1, 5, 9, 13) or (2, 6, 10, 14), etc.

Four types of sampling

➢ **Stratified Sampling-** divide population into 'layers'

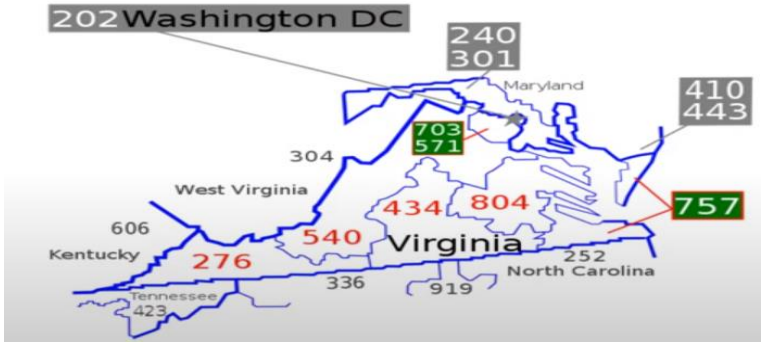
- Stratified random sampling is a method of sampling that involves the division of a population into smaller groups known as strata.
- For example, one might divide a sample of adults into subgroups by age, like 17-28, 29-38, 39-48, 49-58, etc.

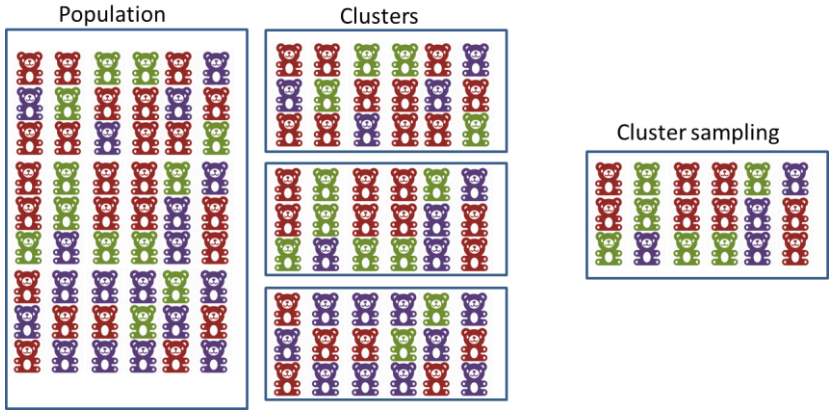


Four types of sampling

➢ **Cluster-** use intact groups

- With cluster sampling, the researcher divides the population into separate groups, by some means such as geographic area or school zone etc. called clusters. Then randomly choose some of these clusters as a whole as a subjects of the samples.





Cluster sampling

Raw Data

- Data collected in original form is called raw data OR data which is not organized is called raw data.
- Data in its original form is called un-grouped data.
- This means raw data is also called ungrouped data.
- Raw data refers to any data object that hasn't undergone through processing, either manually or through computer software.

Organizing Data

- To get an understanding of the data, it is organized and arranged into a meaningful form. This is done by the following methods:

- **Classification**

- **Tabulation** (e.g. simple tables, frequency tables, stem and leaf plots etc.)

- **Graphs** (Bar Graph, Pie chart, Histogram, etc)

Classification

- The process of arranging data into homogenous group or classes according to some common characteristics present in the data is called classification.

Example:

The process of sorting letters in a post office (classified according to the cities and further arranged according to streets)

Bases of Classification

- There are four important bases of classification

- Qualitative Base
- Quantitative Base
- Geographical Base
- Chronological or Temporal base

Bases of Classification

- **Qualitative Base:** When the data are classified according to some quality or attributes such as sex, religion, etc.

- **Quantitative Base:** when the data are classified by quantitative characteristics like heights, weights, ages, income etc.

- **Geographical Base:** When the data are classified by geographical regions or location, like states, provinces, cities, countries, etc.

- **Chronological or Temporal Base:** When the data are classified or arranged by their time of occurrence, such as years, months, weeks, days etc. (e.g. Time series data)

Types of Classification

- There are three main types of classifications:

- **One-way classification:** If we classify observed data keeping in view single characteristic, this type of classification is known as one-way classification.
  Examples: The population of world classified by religion as Muslim, Christian etc.

- **Two-way classification:** If we consider two characteristics at a time in order to classify the observed data then we are doing two-way classifications.
  Examples: The population of world classified by religion and sex.

- **Multi-way classification:** If we consider two characteristics at a time in order to classify the observed data then we are doing multi-way classifications.
  Examples: The population of world classified by religion, sex and Literacy.

Summary

- Basic concepts
- Sampling
- Data Collection
- Data Representation


Thank
You !