BE601 - DATA ANALYTICS I

Chapter 1 - Describing data: Graphical

Uyen Vu ([uyen.vu@phdstudent.hhs.se](mailto:uyen.vu@phdstudent.hhs.se))

August 2nd, 2022

# Important terms and concepts

## Population vs sample

* population: the complete set of items that interest the investigator (N)
* sample: the observed subset (portion) of a population (n)

*Why do we need a sample? Why don’t we just study the whole population?* - Because of 2 factors: size and time constraints

## Random sampling vs systematic sampling

* random sampling: the procedure in which n objects are chosen from a population N totally by chance, the selection of one object does not influence the selection of any other object.
* systematic sampling (including stratified sampling and cluster sampling): the procedure in which the selection of objects is done with jth item from a population, where j = N/n. Randomly pick a number from 1 to j to obtain the first item to be included in the systematic sampling.

## Sampling errors vs Nonsampling errors

* parameter: numerical measure that describe the characteristic of a population N
* statistic: numerical measure that describe the characteristic of a sample n
* sampling errors are caused by sampling procedure:
  + the population and the sample are not relevant
  + survey subjects give inaccurate answers or are dishonest
  + no response to the questions
* nonsampling errors are caused not by the procedure of selection of objects from a population but by other procedures like designing questionnaires, analyzing data.

## Descriptive vs Inferential statistics

* descriptive statistics: focus on graphical and numerical procedures that summarize and process data
* inferential statistics: focus on using data to make predictions, forecasts, and estimates to make better decisions

\*Running a regression is a descriptive or inferential statistics?" -

# Classification of variables

* 2 common ways of classifying variables:
  + categorical vs numerical
  + qualitative vs quantitative

## Categorical vs Numerical

* Categorical variables: responses to categorical variables belong to groups or categories (yes/no question, gender, marital status, Likert scale)
* Numerical variables:
  + discrete variables: produced (usually) by a counting process
  + continuous variables: take value within a given range and usually produced by a measurement process

## Qualitative vs Quantitative

* Qualitative variables: nominal vs ordinal levels of measurement (no measurable meaning to the difference between responses)
* Quantitative variables: interval vs ratio levels of measurement (difference between response has measurable meaning)

# Graphs to describe categorical variables

* Graph to describe a single categorical variable
  + bar chart: cluster bar chart vs component bar chart
  + pie chart
  + pareto diagram:
* Graph to describe the relationship between 2 categorical variables:
  + component bar chart
  + cluster bar chart
* frequency distribution vs relative frequency distribution:
  + FD: comes in absolution value, described by *bar chart*
  + RFD: comes in percentage, described by *pie chart*

Figure. Example of frequency distribution and relative frequency distribution

Figure. Example of frequency distribution and relative frequency distribution

Figure. Example of bar chart

Figure. Example of bar chart

Figure. Example of pie chart

Figure. Example of pie chart

* Cross tables (crosstab, contingency table): is to list the observations for every combination of values for 2 different categorical/ ordinal variable.We can use *component bar chart* or *cluster bar chart* to visualize cross tables

Figure. Example of cross table

Figure. Example of cross table

Figure. Example of component bar chart

Figure. Example of component bar chart

Figure. Example of cluster bar chart

Figure. Example of cluster bar chart

* Pareto diagram: is a *bar chart* that displays the *frequency of defect causes* and orders them in a descending manner. This helps to separate the “vital view” from the “trivial many”.

Figure. Example of table ranking frequencies in decending order

Figure. Example of table ranking frequencies in decending order

Figure. Example of Pareto Diagram

Figure. Example of Pareto Diagram

# Graphs to describe time-series data

* We use line chart

# Graphs to describe numerical variables

* Graphs to describe single numerical variable
  + histogram and ogive (cumulative line graph)
  + stem-and-leaf
* Graphs to describe the relationship between 2 numerical variables: scatter plot
* Shape of distribution:
  + symmetry: obs are balanced or evenly distributed about its center
  + skewness: obs are unevenly distributed about its center
    - skewed-right (positively skewed)
    - skewed-left (negatively skewed)
* Stem-and leaf

# Misleading histogram and misleading time-series plot

# Using R to describe data by graphical