Book: Chapter 1: Introduction and Framework.

Unit-I: Descriptive statistics: – The representation of collected data set/matrix allow the applications of various Statistical Methods

- Statistics is a collection of methods which help us to describe, summarize, interprete and analyse data. Drawing conclusions from data is vital in research, administration & business.

- Observations: The unit on which we measure data such as persons, cars, animals, or plants are called observations. The units/observations are represented by the Greek symbol ω.

- Population: The collection of all units is called population and is denoted by Ω.

- $\omega \in \Omega$, we mean a single unit out of all units.

- Sample: If we consider a selection of observations $\omega_1, \omega_2, \cdots, \omega_n$, then these observations are called Sample: Note that, a sample is always a subset of the populations $\{\omega_1, \omega_2, \cdots, \omega_n\} \subseteq \Omega$.

  e.g: We may be interseted in collecting information about those participating in a statistics course. All participants in the course constitute the population Ω and each participant refers to a unit or observation ω.

Conditional Probability:

2: Frequency Measures and Graphical Representation of

- Variables:

If we have specified the population of interest for a specific research question, we can think about what is of interest about our observations. A particular feature of these observations can be collected in a statistical variable X. Any information we are interested in may be captured in such a variable.

e.g:- Student Name, subject S_1, ...

- The formal definition of variable is $X : \Omega \to S$, $\omega \mapsto x$.

This definition state that a variable X takes a value x for each observation $\omega \in \Omega$, whereby the number of possible values is contained in the set S.

e.g: ① A variable X which refers to age may take any value between 1 and 125. Each person ω is assigned a value x which represents the age of this person.

② If X refers to gender, possible x-values are contained in $S = \{\text{male}, \text{female}\}$. Each observations ω is either male or female and this information summarized in X.

requency measures and Graphical Representation of Data using diffrent

...itative and Quantitative Variables:

Ve

- Qualitative: Qualitative variable are the variable which takes values that cannot be ordered in a logical or natural way.

  e.g: (i) the colour of the eye. (iii) Blood group.
  (ii) the name of a political party.

  Quantitative: Quantitative variables represent measurable quantities. The value which these variable can take can be ordered in a logical and natural way.

  eg: (i) Size of shoes
  (ii) price of houses
  (iii) number of semesters studied.

Discrete and continuous variables:

Discrete: Discrete variables which can only take a finite number of values.

- All qualititative variables are discrete, such as gender of person, the colour of the eye, or the region of a country.

But also quantitative variables can be disc
the size of shoes, or the number of semesters
studied would be discrete because the numbe
of values these variables can take is limited.

Continuous: Variables which can take an infinite number of values are called continuous variables.

e.g: ① the time it takes to travel to university
② the distance between planets.

- Sometimes, it is said that continuous variables are variables which are "measured rather than counted!".

* Scales:

The thoughts and considerations from above indicate that diffrent variables contain diffrent amount of information. A so useful classification of these considerations is given by the concept of the scale of a variable.

Nominal scale: The values of a nominal variable cannot be ordered.

eg: ① the gender of a person
② the status of an application (pending / not pending)

Ordinal scale: The values of an ordinal variable can be ordered. However, the diffrence between these values can not be interpreted in a meaningful way.

e.g: (i) the possible values of education level (none-primary education, secondary education, University degree). can be ordered meaningfully, but the diff-erences between these values cannot be interpreted.

(ii) final ranking at a sports championship.

Continuous scale: The values of a continuous variables can be ordered. Furthermore, the diffrence between these values can be interpreted in a meaningful way.

e.g: (i) the height of persons ... ordered (170 cm, 171, ...) & the diffrences between these values can be compared. Sometimes, the continuous scale is divided further into subscales.

Interval scale: Only diffrences between values, but not ratios, can be interpreted.

e.g: temperature (measured in °c): the diffrence between $-3°c$ and $6°c$ is $9°c$ but the ratio of $\frac{6}{-3} = -2$ does not mean that $-6°c$ is twice as cold as $3°c$.

Ratio scale: Both differences and ratio can be inter

e.g: ① An example is speed : 60 km/h is 40 km/h more than
Moreover 60 km/h is three times faster than 20 km/
② The production time of a car ③ Price of a chocolate bar

Absolute scale: The absolute scale is the same as the ratio scale, with the exception that the values are measured in "natural" units.

e.g: number of semester studied where no artificial unit such as km/h or °C is needed: the values are simply 1, 2, 3, ...

Grouped Data: Sometimes, data may be available in summarized form: instead of the original value, one may only know the category or group the value belong to.

e.g: ① It is often convenient in a survey to ask for the income per year by means of groups
(Rs: 0 - 20,000, 20,000 - 1,00,000, > 1,00,000)

② If there are many political parties in an election, those with a low number of voters are often summarized in a new category "other parties".

③ Instead of capturing the number of claims made by an insurance company customer, the variable "claimed" may denote whether or not the customer claimed at all (yes/no)

ata is available in grouped form, we call the respective variable capturing this information a grouped variable. Sometimes, these variables are also known as categorical variables. [which takes a finite, possibly small, number of values.]

Thus, any discrete [and/or] nominal / ordinal / qualitative variable may be regarded as a categorical variable.

- Binary variable: Any grouped or categorical variable which can only take two values is called a binary variable.

Qualitative data → discrete

Quantitative data → discrete (size of shoes or a grouped variable)
→ continuous (temperature)

Nominal variable → quantitative & discrete (color of eye)

continuous variable always quantitative (temperature)

categorical variables → qualitative (colour of the eye)
→ quantitative (satisfaction level on a scale 1 to 5)

Categorical variables are never continuous.

*Data collection: When collecting data, we may ask ourselves how to faciliate this in detail and much data needs to be collected.

Survey: Collect data by asking questions.

Experiments: data is obtained in controlled setting.

Observational data: data which is collected routinely, without survey / experiments.

Primary data & Secondary data. ← collect data by someone else.

data use ↓ Collect ^data ourselves via a survey or experiment

* Creating data set:

Key Points:

① The scale of variables is not only a formalism but an essential framework for choosing the correct analysis methods. This is particularly relevant for associating analysis, (Ch. 4), Statistical tests (Ch. 10) & linear regression (Ch-11).

② Even if variables are measured on a nominal scale (i.e. they are categorical / qualitative), we may choose to assign a number to each category of this variable. This eases implementation of some analysis methods introduced later in this book/course.

③ Data is usually stored in a data matrix, where the rows represent the observations and the column are variables. It can be analysed with Statistica software.