Statistical Data

Tushar B. Kute,
http://tusharkute.com

Statistical Data Analysis

- Statistical data analysis is a procedure of performing various statistical operations.
- It is a kind of quantitative research, which seeks to quantify the data, and typically, applies some form of statistical analysis.
- Quantitative data basically involves descriptive data, such as survey data and observational data.

Qualitative Data Type

- Qualitative or Categorical Data describes the object under consideration using a finite set of discrete classes.
- It means that this type of data can't be counted or measured easily using numbers and therefore divided into categories.
- The gender of a person (male, female, or others) is a good example of this data type.

Qualitative Data Type

- These are usually extracted from audio, images, or text medium.
- Another example can be of a smartphone brand that provides information about the current rating, the color of the phone, category of the phone, and so on.
- All this information can be categorized as Qualitative data. There are two subcategories under this:
  - Nominal
  - Ordinal

Nominal

- These are the set of values that don’t possess a natural ordering.
- Example: The color of a smartphone can be considered as a nominal data type as we can’t compare one color with others.
- It is not possible to state that ‘Red’ is greater than ‘Blue’.
- The gender of a person is another one where we can’t differentiate between male, female, or others.
- Mobile phone categories whether it is midrange, budget segment, or premium smartphone is also nominal data type.

Ordinal

- These types of values have a natural ordering while maintaining their class of values.
- If we consider the size of a clothing brand then we can easily sort them according to their name tag in the order of small < medium < large.
- The grading system while marking candidates in a test can also be considered as an ordinal data type where A+ is definitely better than B grade.

Ordinal

- These categories help us deciding which encoding strategy can be applied to which type of data.
- Data encoding for Qualitative data is important because machine learning models can’t handle these values directly and needed to be converted to numerical types as the models are mathematical in nature.
- For nominal data type where there is no comparison among the categories, one-hot encoding can be applied which is similar to binary coding considering there are in less number and for the ordinal data type, label encoding can be applied which is a form of integer encoding.

Quantitative Data Type

- This data type tries to quantify things and it does by considering numerical values that make it countable in nature.
- The price of a smartphone, discount offered, number of ratings on a product, the frequency of processor of a smartphone, or ram of that particular phone, all these things fall under the category of Quantitative data types.

Quantitative Data Type

- The key thing is that there can be an infinite number of values a feature can take.
- For instance, the price of a smartphone can vary from x amount to any value and it can be further broken down based on fractional values.
- The two subcategories which describe them clearly are:
  - Discrete
  - Continuous

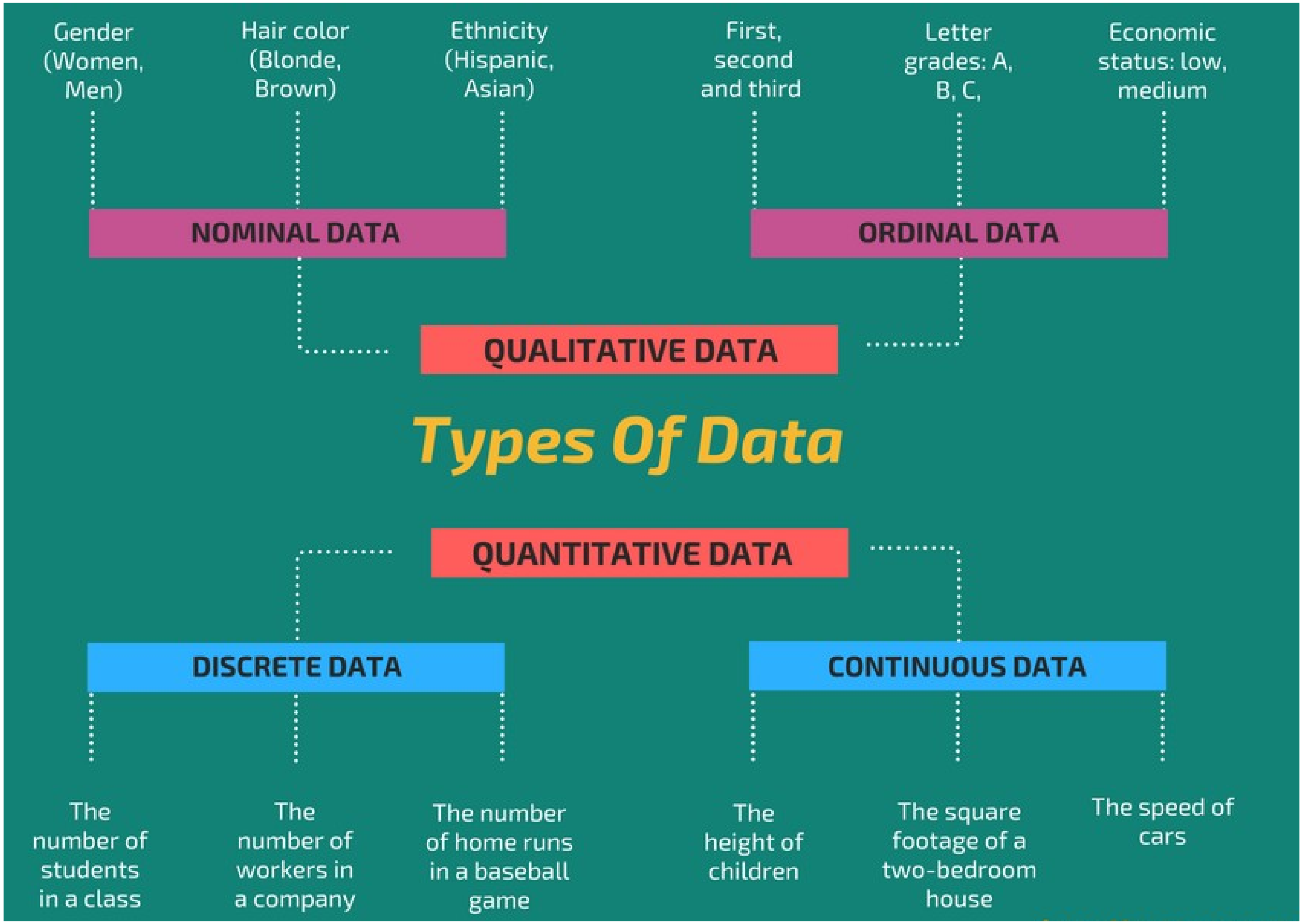
Discrete

- The numerical values which fall under are integers or whole numbers are placed under this category.
- The number of speakers in the phone, cameras, cores in the processor, the number of sims supported all these are some of the examples of the discrete data type.

Continous

- The fractional numbers are considered as continuous values.
- These can take the form of the operating frequency of the processors, the android version of the phone, wifi frequency, temperature of the cores, and so on.

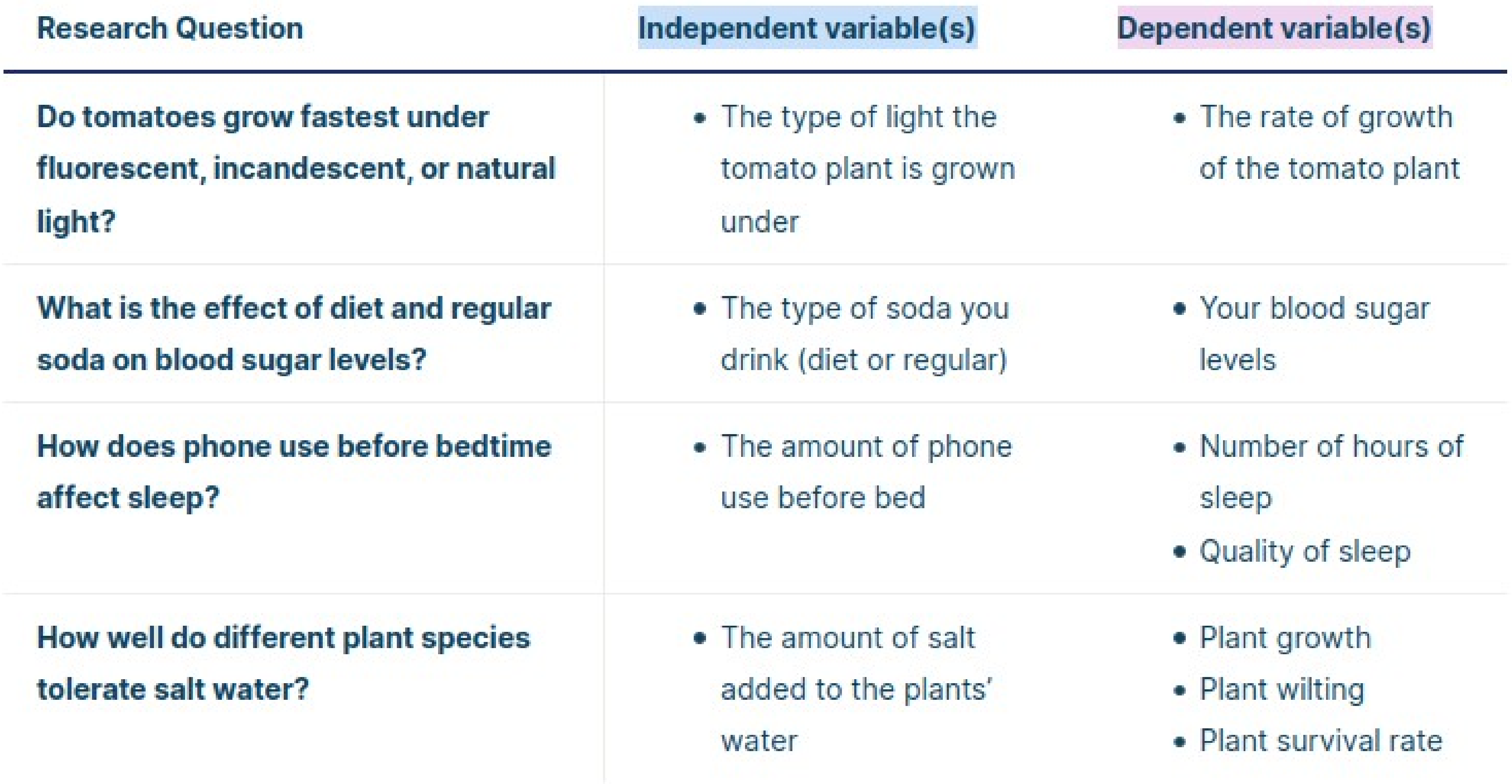
Summary


Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed of cars

Types of variables

- In research, variables are any characteristics that can take on different values, such as height, age, species, or exam score.
- In scientific research, we often want to study the effect of one variable on another one. For example, you might want to test whether students who spend more time studying get better exam scores.
- The variables in a study of a cause-and-effect relationship are called the independent and dependent variables.
  - The independent variable is the cause. Its value is independent of other variables in your study.
  - The dependent variable is the effect. Its value depends on changes in the independent variable.

Types of variables

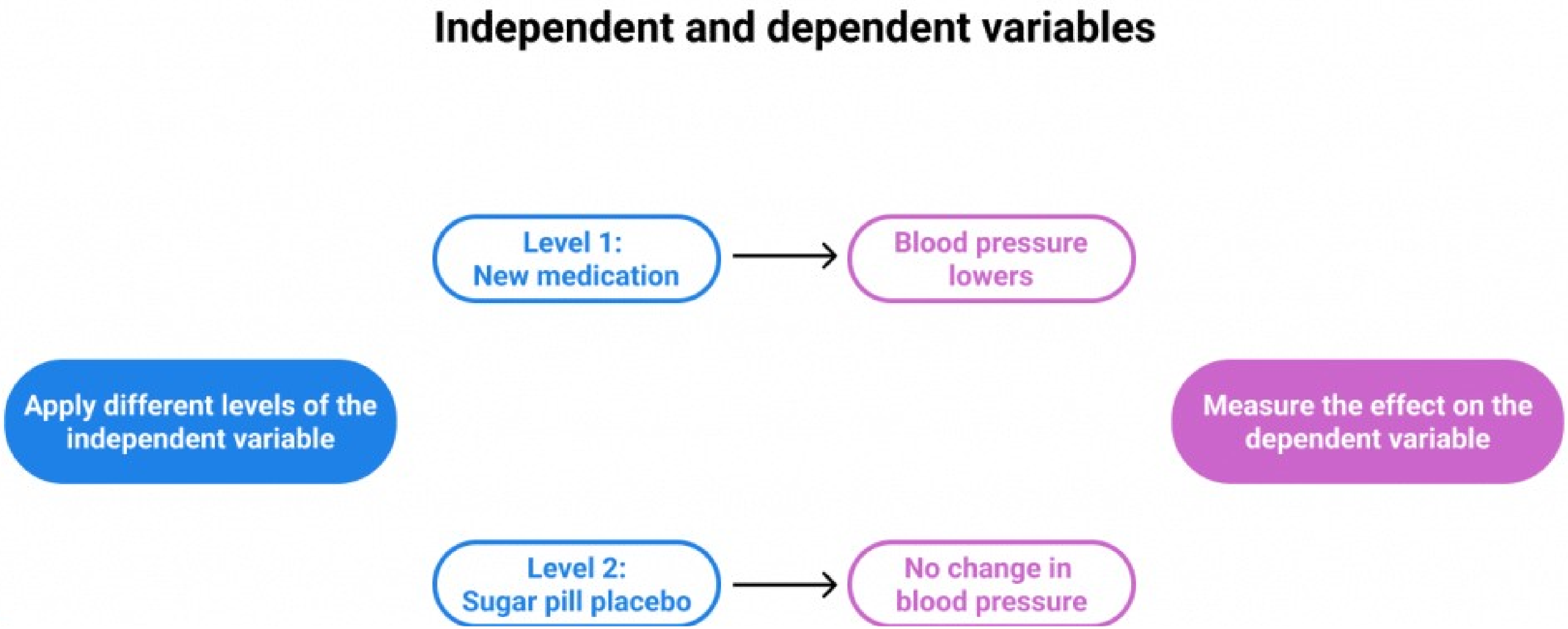


Research Question	Independent variable(s)	Dependent variable(s)
Do tomatoes grow fastest under fluorescent, incandescent, or natural light?	• The type of light the tomato plant is grown under	• The rate of growth of the tomato plant
What is the effect of diet and regular soda on blood sugar levels?	• The type of soda you drink (diet or regular)	• Your blood sugar levels
How does phone use before bedtime affect sleep?	• The amount of phone use before bed	• Number of hours of sleep • Quality of sleep
How well do different plant species tolerate salt water?	• The amount of salt added to the plants' water	• Plant growth • Plant wilting • Plant survival rate

Example:

- You are studying the impact of a new medication on the blood pressure of patients with hypertension.
- To test whether the medication is effective, you divide your patients into two groups. One group takes the medication, while the other group takes a sugar pill placebo.
  - Your independent variable is the treatment that you vary between groups: which type of pill the patient receives.
  - Your dependent variable is the outcome that you measure: the blood pressure of the patients.

Example:



Independent and dependent variables

Apply different levels of the independent variable

Level 1: New medication → Blood pressure lowers

Level 2: Sugar pill placebo → No change in blood pressure

Measure the effect on the dependent variable

Example:

- Imagine that a tutor asks 100 students to complete a maths test. The tutor wants to know why some students perform better than others. Whilst the tutor does not know the answer to this, she thinks that it might be because of two reasons:
  - (1) some students spend more time revising for their test; and (2) some students are naturally more intelligent than others. As such, the tutor decides to investigate the effect of revision time and intelligence on the test performance of the 100 students.
- Dependent Variable: Test Mark (measured from 0 to 100)
- Independent Variables: Revision time (measured in hours) Intelligence (measured using IQ score)

Other names for independent variables

- Sometimes, the variable you think is the cause might not be fully independent – it might be influenced by other variables. In this case, one of these terms is more appropriate:
  - Explanatory variables (they explain an event or outcome)
  - Predictor variables (they can be used to predict the value of a dependent variable)
  - Right-hand-side variables (they appear on the right-hand side of a regression equation).

Other names for dependent variables

- Dependent variables are also known by these terms:
  - Response variables (they respond to a change in another variable)
  - Outcome variables (they represent the outcome you want to measure)
  - Left-hand-side variables (they appear on the left-hand side of a regression equation)

Univatiate Data

- This type of data consists of only one variable.
- The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes.
- It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.
- The example of a univariate data can be height.

Univatiate Data

Heights (in cm)	164	167.3	170	174.2	178	180	186

Univatiate Data

- Suppose that the heights of seven students of a class is recorded, there is only one variable that is height and it is not dealing with any cause or relationship.
- The description of patterns found in this type of data can be made by drawing conclusions using central tendency measures (mean, median and mode), dispersion or spread of data (range, minimum, maximum, quartiles, variance and standard deviation) and by using frequency distribution tables, histograms, pie charts, frequency polygon and bar charts.

Bivatiate Data

- This type of data involves two different variables.
- The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.
- Example of bivariate data can be temperature and ice cream sales in summer season.

Bivatiate Data

TEMPERATURE(IN CELSIUS)	ICE CREAM SALES
20	2000
25	2500
35	5000
43	7800

Bivatiate Data

- Suppose the temperature and ice cream sales are the two variables of a bivariate data.
- Here, the relationship is visible from the table that temperature and sales are directly proportional to each other and thus related because as the temperature increases, the sales also increase.
- Thus bivariate data analysis involves comparisons, relationships, causes and explanations.
- These variables are often plotted on X and Y axis on the graph for better understanding of data and one of these variables is independent while the other is dependent.

Multivatiate Data

- When the data involves three or more variables, it is categorized under multivariate.
- Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

Multivatiate Data

- It is similar to bivariate but contains more than one dependent variable.
- The ways to perform analysis on this data depends on the goals to be achieved.Some of the techniques are regression analysis,path analysis,factor analysis and multivariate analysis of variance (MANOVA).

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com