Statistics for Scientists – CSC261

Introduction

Dr DEBAJYOTI PAL

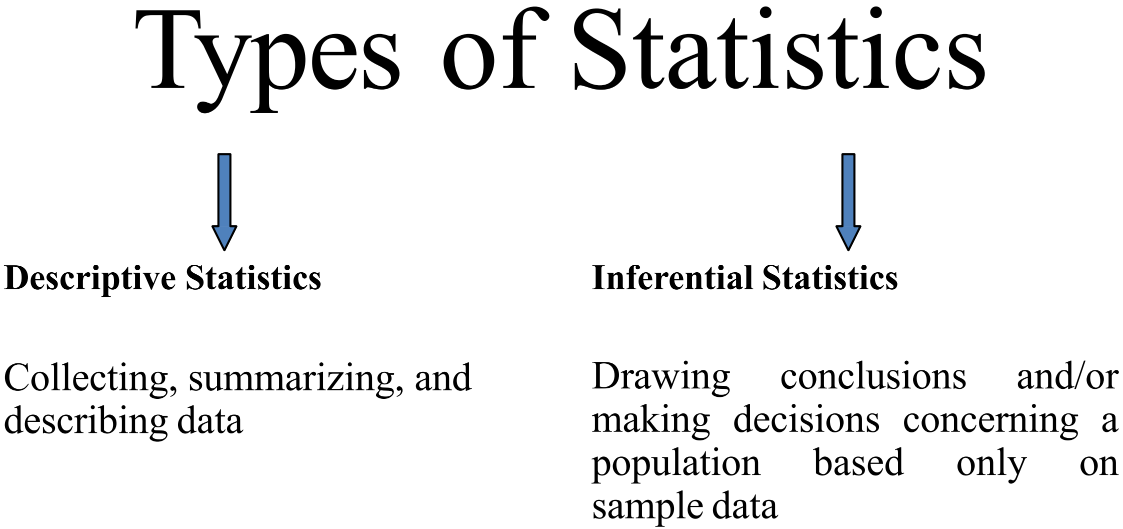
SCHOOL OF INFORMATION TECHNOLOGY,
KMUTT

What is Statistics?

- A branch of mathematics taking and transforming numbers into useful information for decision makers

- Methods for processing & analyzing numbers

- Methods for helping reduce the uncertainty inherent in decision making



Types of Statistics

Descriptive Statistics

Collecting, summarizing, and describing data

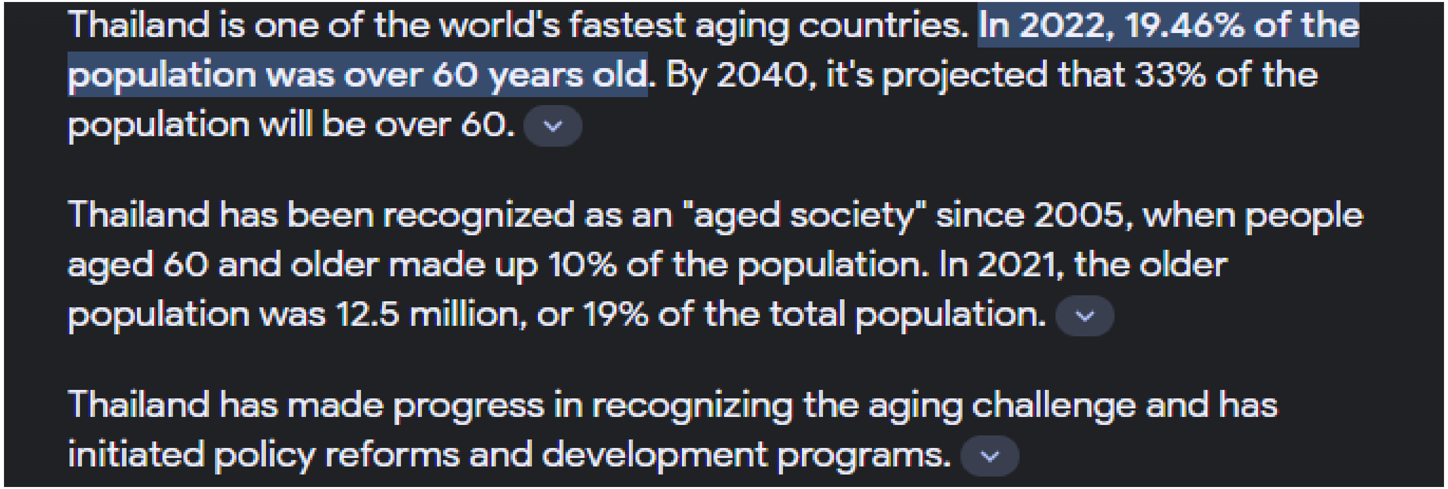
Inferential Statistics

Drawing conclusions and/or making decisions concerning a population based only on sample data

Why Study Statistics?

Decision Makers Use Statistics To:

- Present and describe business data and information properly
- Draw conclusions about large groups of individuals or items, using information collected from subsets of the individuals or items.
- Make reliable <u>forecasts</u> about a business activity??
- Improve business <u>processes.</u>



Thailand is one of the world's fastest aging countries. In 2022, 19.46% of the population was over 60 years old. By 2040, it's projected that 33% of the population will be over 60.

Thailand has been recognized as an "aged society" since 2005, when people aged 60 and older made up 10% of the population. In 2021, the older population was 12.5 million, or 19% of the total population.

Thailand has made progress in recognizing the aging challenge and has initiated policy reforms and development programs.

- Source – Wikipedia (accessed 22nd Jan 2024)

Introduction and Data Collection

Data: Are collection of any number of related observations

Data set: A collection of data is data set

Data point: A single observation

Raw data: Information before it arranged and analyzed

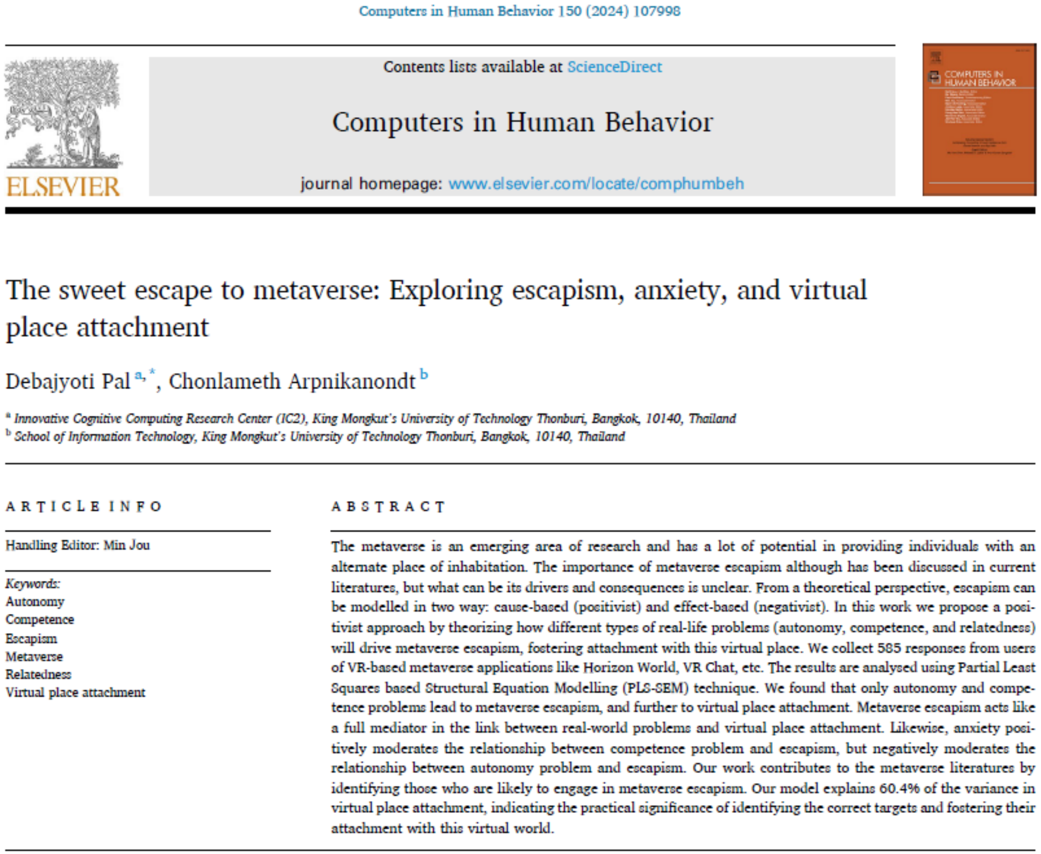
Data: Observation + Noise

Example of raw data:

High school and college CGPA	HS	College
	3.6	2.5
	2.6	2.7
	2.7	2.2
	3.7	3.2
	4.0	3.8

Criteria/Tests for Evaluating Data

Criteria	Issues	Remarks
Specifications & Methodology	**Data collection method, response rate, quality & analysis of data, sampling technique & size, questionnaire design, fieldwork.**	**Data should be reliable, valid, & generalizable to the problem.**
Error & Accuracy	**Examine errors in approach, research design, sampling, data collection & analysis, & reporting.**	**Assess accuracy by comparing data from different sources.**
Currency	**Time lag between collection & publication, frequency of updates.**	**Census data are updated by syndicated firms.**
Objective	**Why were the data collected?**	**The objective determines the relevance of data.**
Nature	**Definition of key variables, units of measurement, categories used, relationships examined.**	**Reconfigure the data to increase their usefulness.**
Dependability	**Expertise, credibility, reputation, and trustworthiness of the source.**	**Data should be obtained from an original source.**



Computers in Human Behavior 150 (2024) 107998

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

The sweet escape to metaverse: Exploring escapism, anxiety, and virtual place attachment

Debajyoti Pal[a], *, Chonlameth Arpnikanondt[b]

[a] Innovative Cognitive Computing Research Center (IC2), King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand
[b] School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand

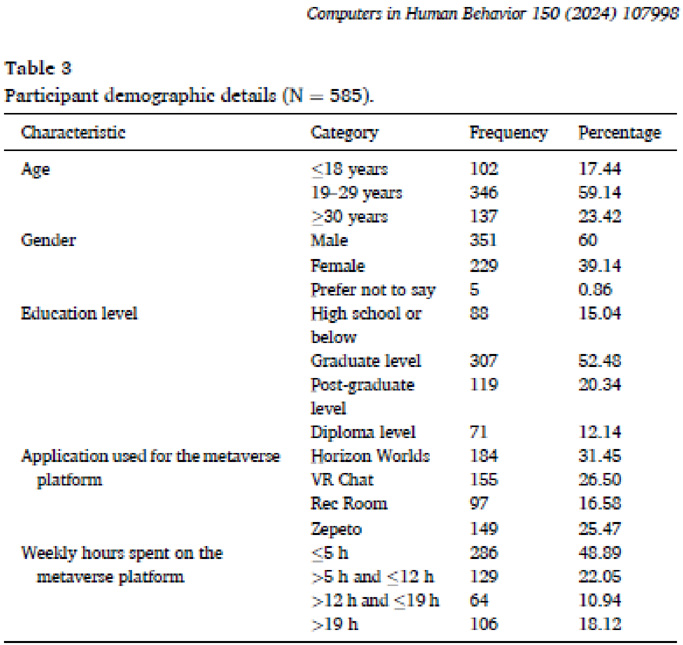
ARTICLE INFO

Handling Editor: Min Jou

Keywords:
Autonomy
Competence
Escapism
Metaverse
Relatedness
Virtual place attachment

ABSTRACT

The metaverse is an emerging area of research and has a lot of potential in providing individuals with an alternate place of inhabitation. The importance of metaverse escapism although has been discussed in current literatures, but what can be its drivers and consequences is unclear. From a theoretical perspective, escapism can be modelled in two way: cause-based (positivist) and effect-based (negativist). In this work we propose a positivist approach by theorizing how different types of real-life problems (autonomy, competence, and relatedness) will drive metaverse escapism, fostering attachment with this virtual place. We collect 585 responses from users of VR-based metaverse applications like Horizon World, VR Chat, etc. The results are analysed using Partial Least Squares based Structural Equation Modelling (PLS-SEM) technique. We found that only autonomy and competence problems lead to metaverse escapism, and further to virtual place attachment. Metaverse escapism acts like a full mediator in the link between real-world problems and virtual place attachment. Likewise, anxiety positively moderates the relationship between competence problem and escapism, but negatively moderates the relationship between autonomy problem and escapism. Our work contributes to the metaverse literatures by identifying those who are likely to engage in metaverse escapism. Our model explains 60.4% of the variance in virtual place attachment, indicating the practical significance of identifying the correct targets and fostering their attachment with this virtual world.



Computers in Human Behavior 150 (2024) 107998

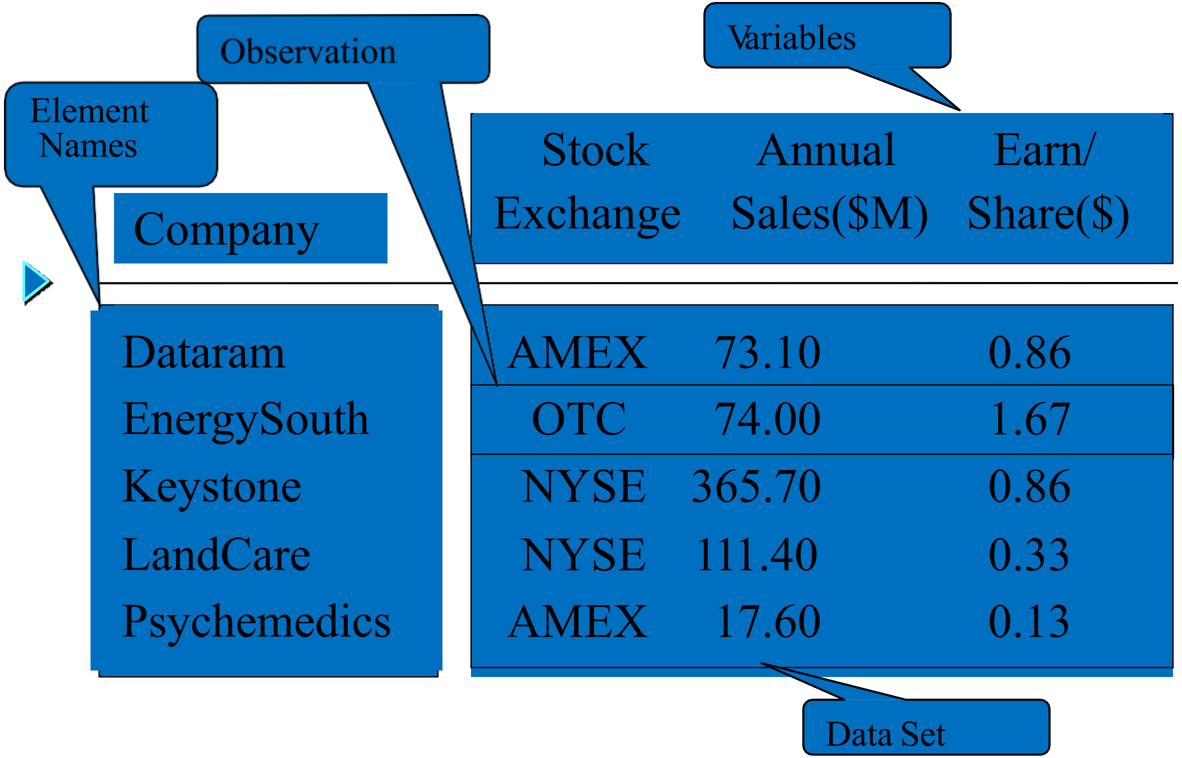
Table 3
Participant demographic details (N = 585).

Characteristic	Category	Frequency	Percentage
Age	≤18 years	102	17.44
	19–29 years	346	59.14
	≥30 years	137	23.42
Gender	Male	351	60
	Female	229	39.14
	Prefer not to say	5	0.86
Education level	High school or below	88	15.04
	Graduate level	307	52.48
	Post-graduate level	119	20.34
	Diploma level	71	12.14
Application used for the metaverse platform	Horizon Worlds	184	31.45
	VR Chat	155	26.50
	Rec Room	97	16.58
	Zepeto	149	25.47
Weekly hours spent on the metaverse platform	≤5 h	286	48.89
	>5 h and ≤12 h	129	22.05
	>12 h and ≤19 h	64	10.94
	>19 h	106	18.12

Elements, Variables, and Observations

- The elements are the entities on which data are collected.
- A variable is a characteristic of interest for the elements.
- The set of measurements collected for a particular element is called an observation.
- The total number of data values in a data set is the number of elements multiplied by the number of variables.

Data, Data Sets, Elements, Variables, and Observations



Observation

Variables

Element Names

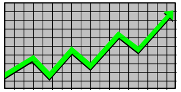
Data Set

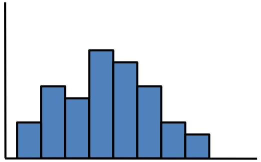
Company	Stock Exchange	Annual Sales($M)	Earn/ Share($)
Dataram	AMEX	73.10	0.86
EnergySouth	OTC	74.00	1.67
Keystone	NYSE	365.70	0.86
LandCare	NYSE	111.40	0.33
Psychemedics	AMEX	17.60	0.13

Descriptive Statistics

- Collect data
  - e.g., Survey

- Present data
  - e.g., Tables and graphs





- Characterize data
  - e.g., Sample mean = $\frac{\sum X_i}{n}$

Inferential Statistics

- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 120 pounds

Drawing conclusions about a large group of individuals based on a subset of the large group.

Basic Vocabulary of Statistics

POPULATION
A **population** consists of all the items or individuals about which you want to draw a conclusion.
Ex: People who live within 25 kms of radius from center of the city.

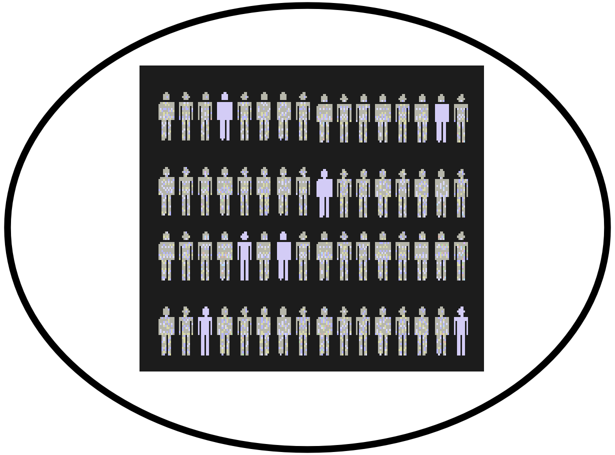
SAMPLE
A **sample** is the portion of a population selected for analysis. It has to be **representative**.

PARAMETER
A **parameter** is a numerical measure that describes a **characteristic** of a **population**.

STATISTIC
A **statistic** is a numerical measure that describes a **characteristic** of a **sample**.

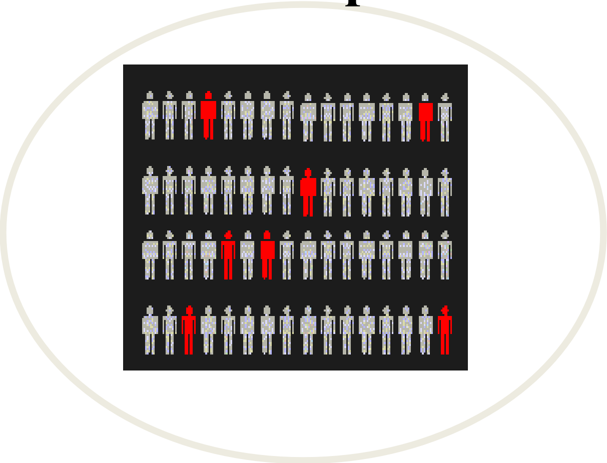
Population vs. Sample

Population

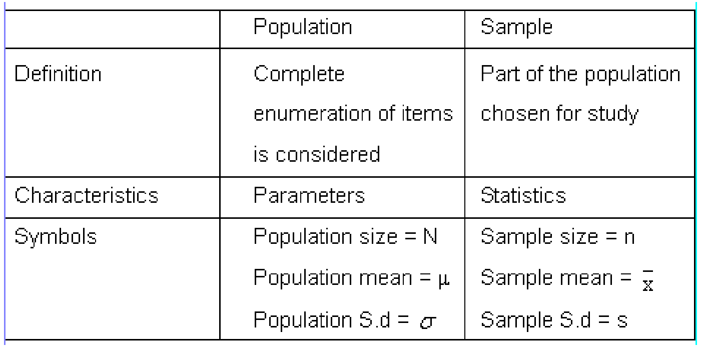


Measures used to describe the population are called **parameters**

Sample



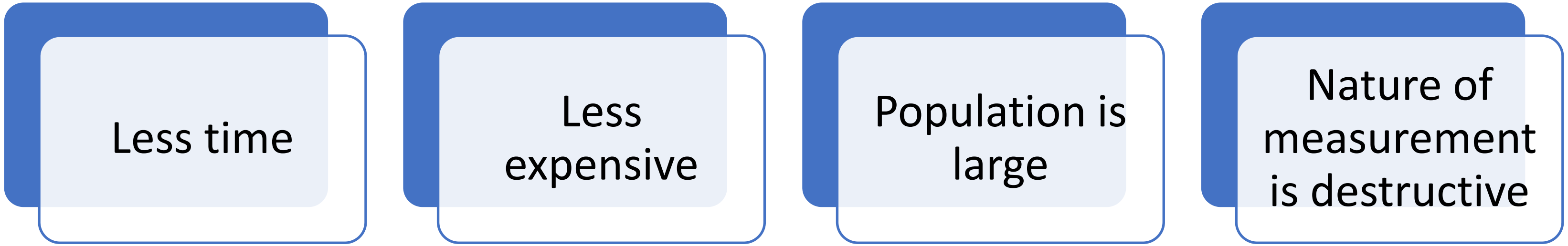
Measures computed from sample data are called **statistics**



	Population	Sample
Definition	Complete enumeration of items is considered	Part of the population chosen for study
Characteristics	Parameters	Statistics
Symbols	Population size = N Population mean = μ Population S.d = σ	Sample size = n Sample mean = $\bar{x}$ Sample S.d = s

Which is better : Samplings or complete enumeration?

Benefits of sample



- Less time
- Less expensive
- Population is large
- Nature of measurement is destructive

Why Collect Data?

A marketing research analyst needs to assess the **effectiveness of a new television advertisement.**

A pharmaceutical manufacturer needs to determine whether a new **drug is more effective** than those currently in use.

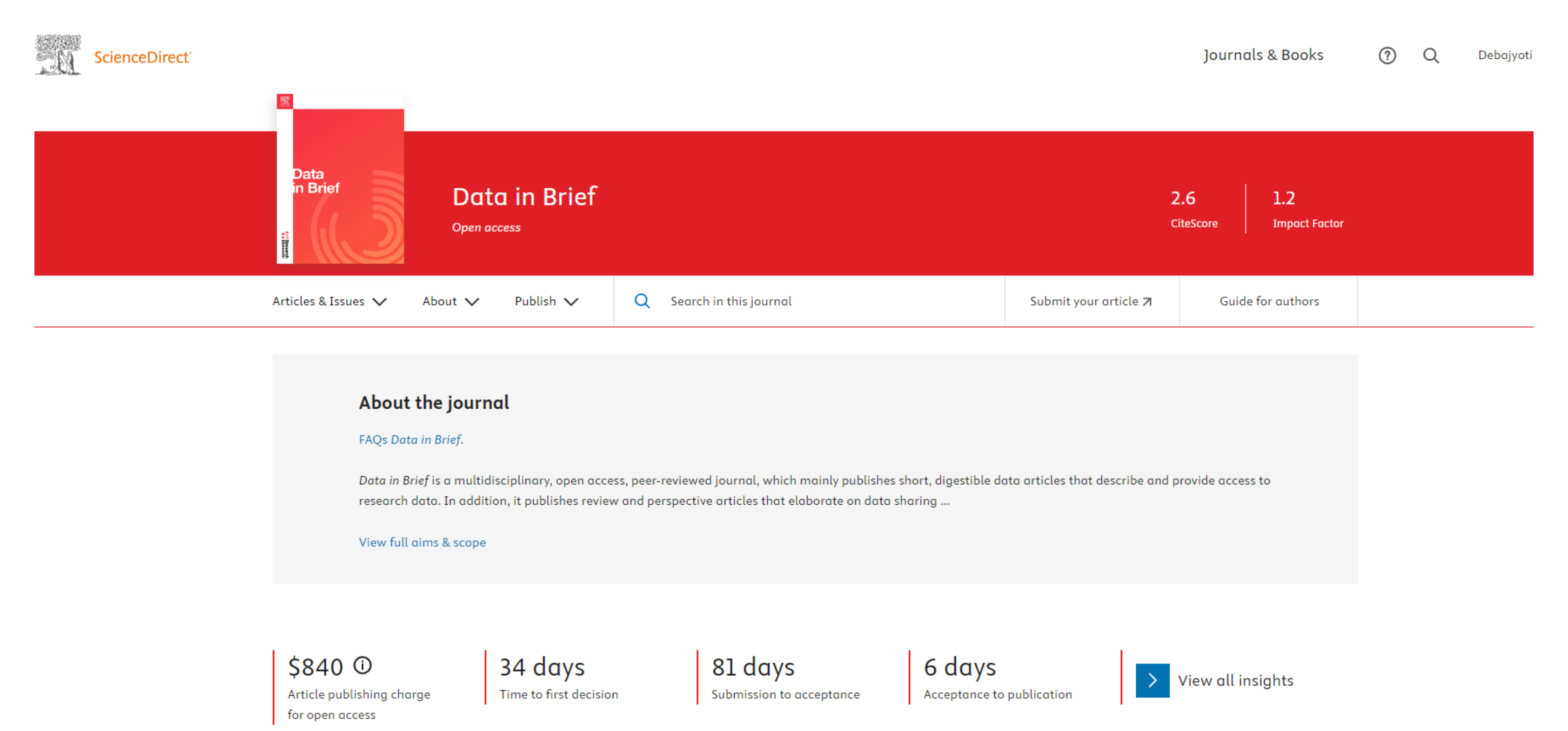
An operations manager wants to monitor a manufacturing process to find out whether the **quality**

of the product being manufactured is conforming to company **standards.**

An auditor wants to review the financial transactions of a company in order to determine whether the company is in **compliance** with generally accepted **accounting principles**.

Sources of Data

- Primary Sources: The data **collector is the one using** the data for analysis
  - Data from a political survey
  - Data collected from an experiment
  - Observed data
- Secondary Sources: The **person performing data analysis is not the data collector**
  - Analyzing census data
  - Examining data from print journals or data published on the internet.



Data in Brief

Open access

2.6 CiteScore | 1.2 Impact Factor

About the journal

FAQs *Data in Brief*.

Data in Brief is a multidisciplinary, open access, peer-reviewed journal, which mainly publishes short, digestible data articles that describe and provide access to research data. In addition, it publishes review and perspective articles that elaborate on data sharing ...

View full aims & scope

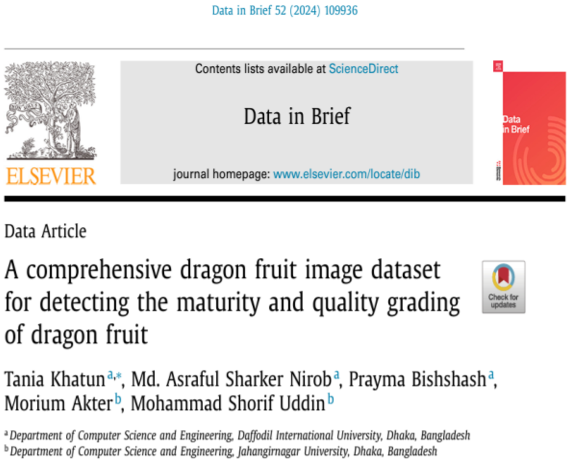
$840 ⓘ Article publishing charge for open access

34 days Time to first decision

81 days Submission to acceptance

6 days Acceptance to publication

View all insights



Data in Brief 52 (2024) 109936

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

ELSEVIER

Data Article

A comprehensive dragon fruit image dataset for detecting the maturity and quality grading of dragon fruit

Tania Khatun[a,*], Md. Asraful Sharker Nirob[a], Prayma Bishshash[a], Morium Akter[b], Mohammad Shorif Uddin[b]

[a] Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh
[b] Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh

Fig. 1. The real dragon fruit field from where we collected the dataset images.

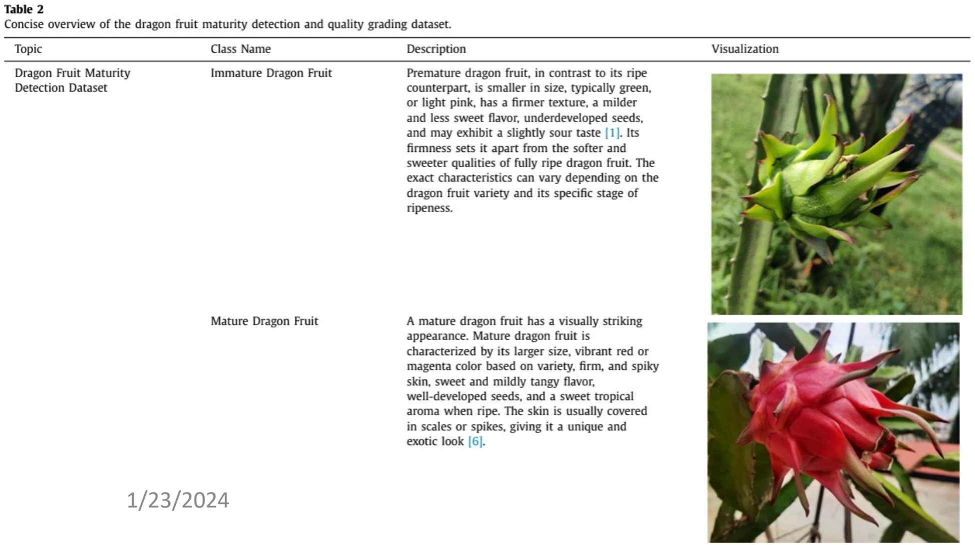


Table 2
Concise overview of the dragon fruit maturity detection and quality grading dataset.

Topic	Class Name	Description	Visualization
Dragon Fruit Maturity Detection Dataset	Immature Dragon Fruit	Premature dragon fruit, in contrast to its ripe counterpart, is smaller in size, typically green, or light pink, has a firmer texture, a milder and less sweet flavor, underdeveloped seeds, and may exhibit a slightly sour taste [1]. Its firmness sets it apart from the softer and sweeter qualities of fully ripe dragon fruit. The exact characteristics can vary depending on the dragon fruit variety and its specific stage of ripeness.	
	Mature Dragon Fruit	A mature dragon fruit has a visually striking appearance. Mature dragon fruit is characterized by its larger size, vibrant red or magenta color based on variety, firm, and spiky skin, sweet and mildly tangy flavor, well-developed seeds, and a sweet tropical aroma when ripe. The skin is usually covered in scales or spikes, giving it a unique and exotic look [6].	

1/23/2024

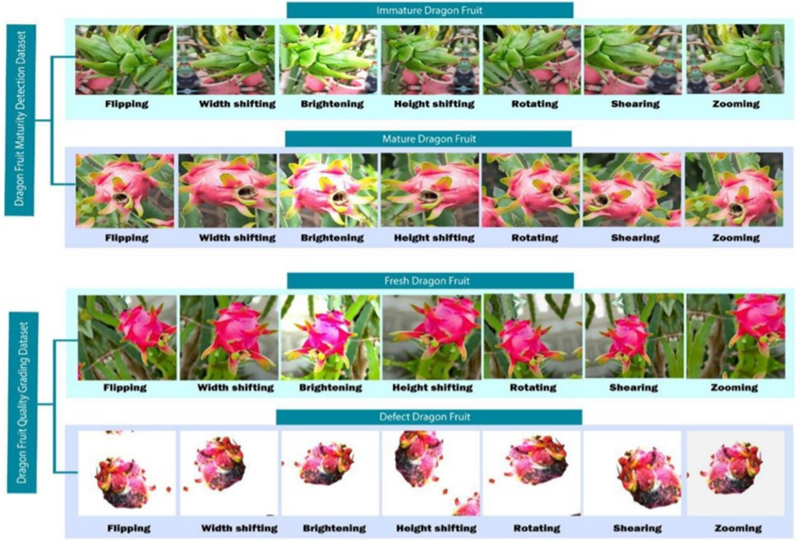


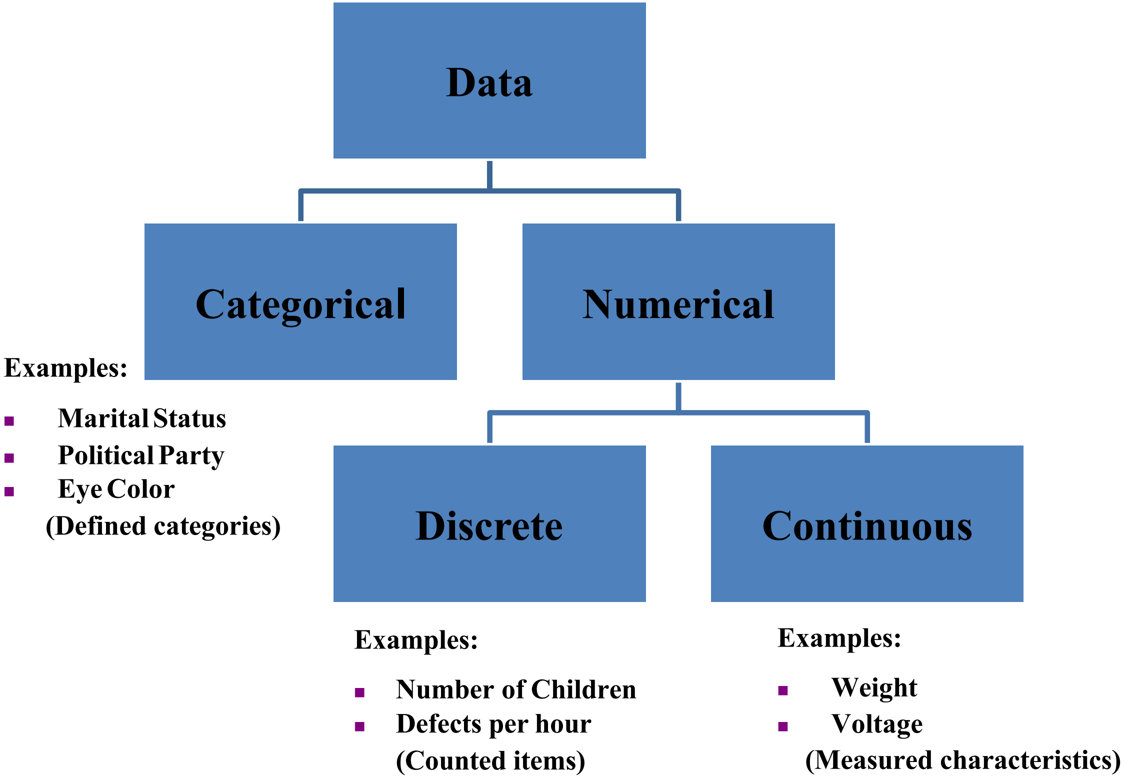
Fig. 3. Augmented images of dragon fruit dataset.

Types of Variables

- **Categorical** (qualitative) variables have values that can only be placed into categories, such as “yes” and “no.”

- **Numerical** (quantitative) variables have values that represent quantities.

Types of Data



Data

Categorical

Numerical

Discrete

Continuous

Examples:

- Marital Status
- Political Party
- Eye Color

(Defined categories)

Examples:

- Number of Children
- Defects per hour

(Counted items)

Examples:

- Weight
- Voltage

(Measured characteristics)

Scales of Measurement

Scales of measurement include:

Nominal	Interval
Ordinal	Ratio

The scale determines the amount of **information** contained in the data.

The scale indicates the data **summarization** and **statistical** analyses that are most appropriate.

Scales of Measurement

- Nominal

▷ Data are labels or names used to identify an attribute of the element.

▷ A nonnumeric label or numeric code may be used.

Scales of Measurement

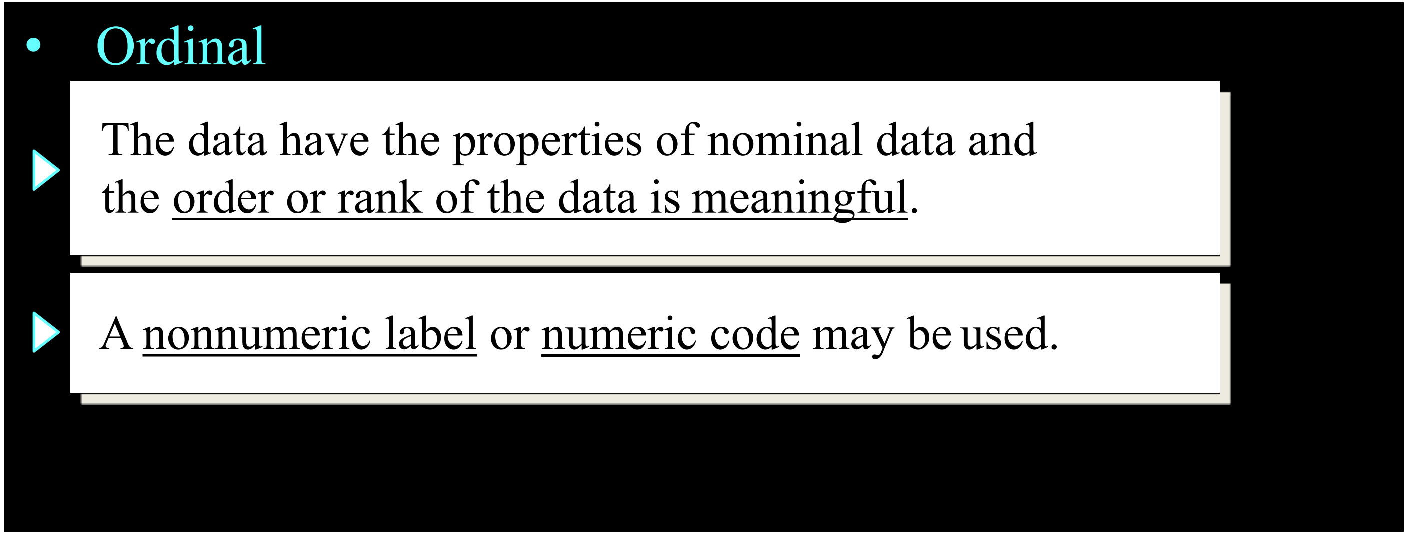
- Nominal

Example:

Students of a university are classified by the school in which they are enrolled using a nonnumeric label such as Business, Humanities, Education, and so on.

Alternatively, a numeric code could be used for the school variable (e.g. 1 denotes Business,2 denotes Humanities, 3 denotes Education, and so on).

Scales of Measurement



- Ordinal

▷ The data have the properties of nominal data and the order or rank of the data is meaningful.

▷ A nonnumeric label or numeric code may be used.

Scales of Measurement

- Ordinal

Example:

Students of a university are classified by their class standing using a nonnumeric label such as Freshman, Sophomore, Junior, or Senior.

Alternatively, a numeric code could be used for the class standing variable (e.g. 1 denotes Freshman, 2 denotes Sophomore, 3 denotes Junior, 4 denotes Senior).

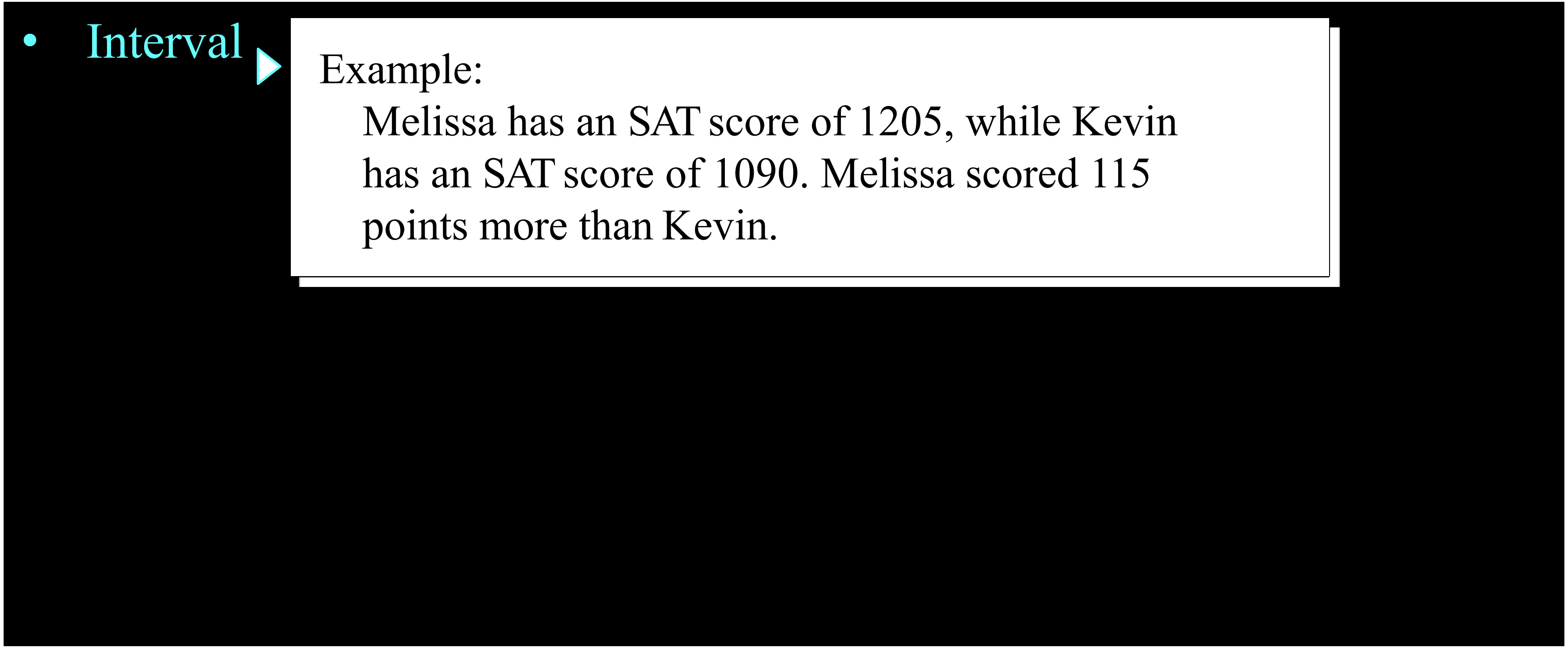
Scales of Measurement

- Interval

- The data have the properties of ordinal data, and the interval between observations is expressed in terms of a **fixed unit of** measure.

- Interval data are always numeric.

Scales of Measurement

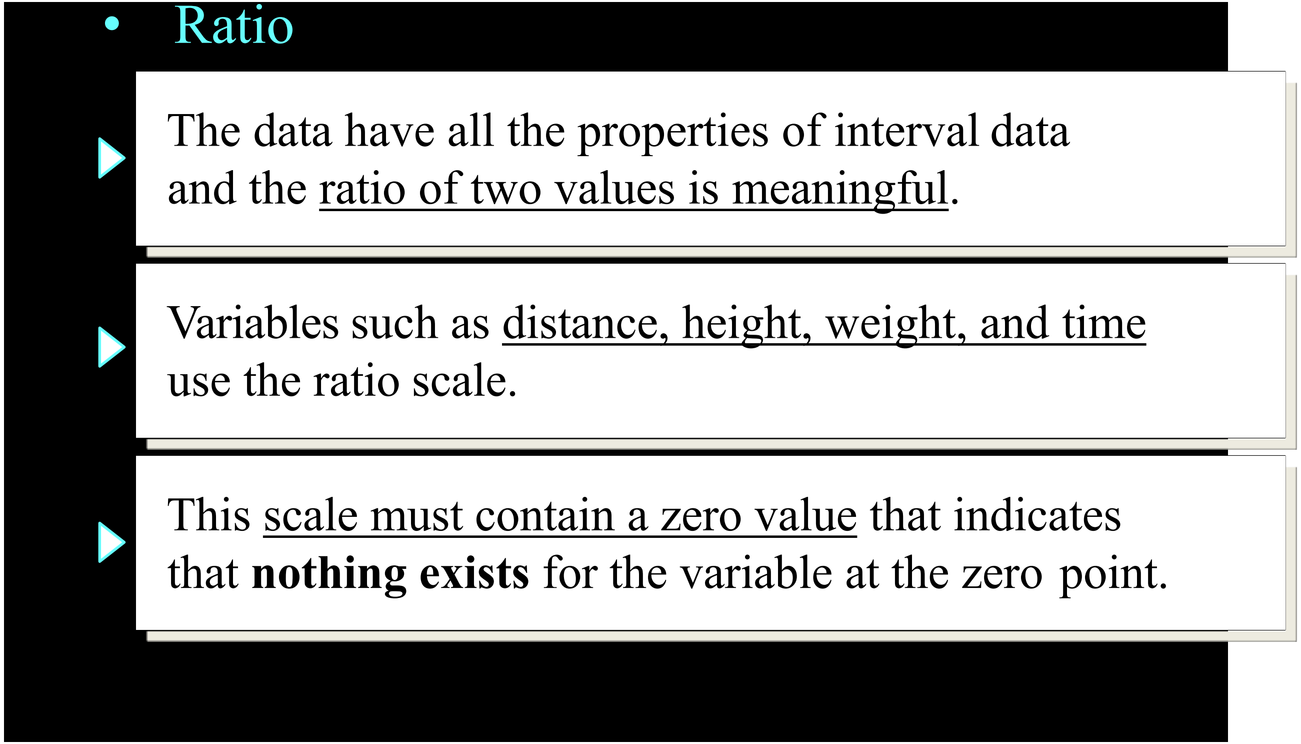


- Interval

Example:

Melissa has an SAT score of 1205, while Kevin has an SAT score of 1090. Melissa scored 115 points more than Kevin.

Scales of Measurement



- Ratio

The data have all the properties of interval data and the ratio of two values is meaningful.

Variables such as distance, height, weight, and time use the ratio scale.

This scale must contain a zero value that indicates that **nothing exists** for the variable at the zero point.

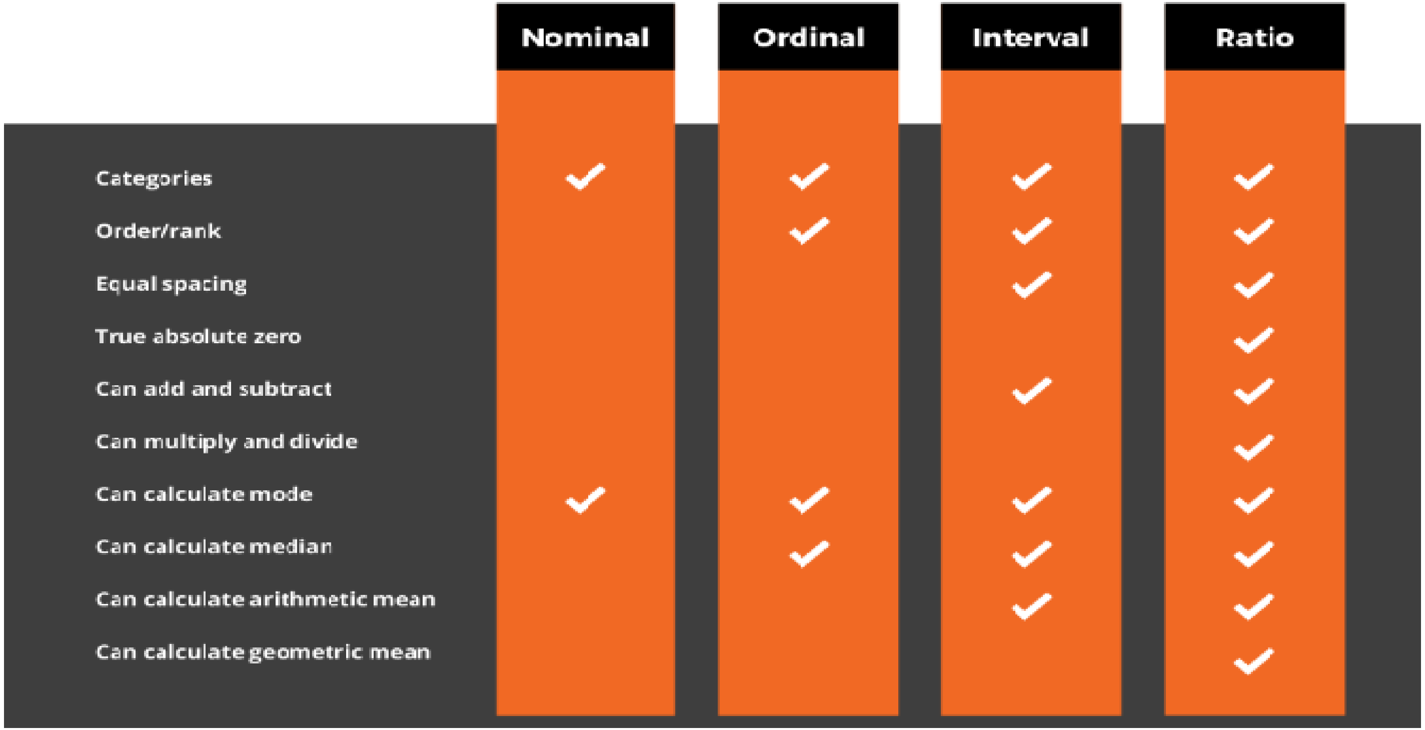
Scales of Measurement

- Ratio

Example:

Melissa’s college record shows 36 credit hours earned, while Kevin’s record shows 72 credit hours earned. Kevin has twice as many credit hours earned as Melissa.

Measurement Levels : Comparison



	Nominal	Ordinal	Interval	Ratio
Categories	✓	✓	✓	✓
Order/rank		✓	✓	✓
Equal spacing			✓	✓
True absolute zero				✓
Can add and subtract			✓	✓
Can multiply and divide				✓
Can calculate mode	✓	✓	✓	✓
Can calculate median		✓	✓	✓
Can calculate arithmetic mean			✓	✓
Can calculate geometric mean				✓

Activity 1

Colleges and universities are requiring an increasing amount of information about applicants before making acceptance and financial aid decisions. Classify each of the following types of data required on a college application as quantitative or qualitative:

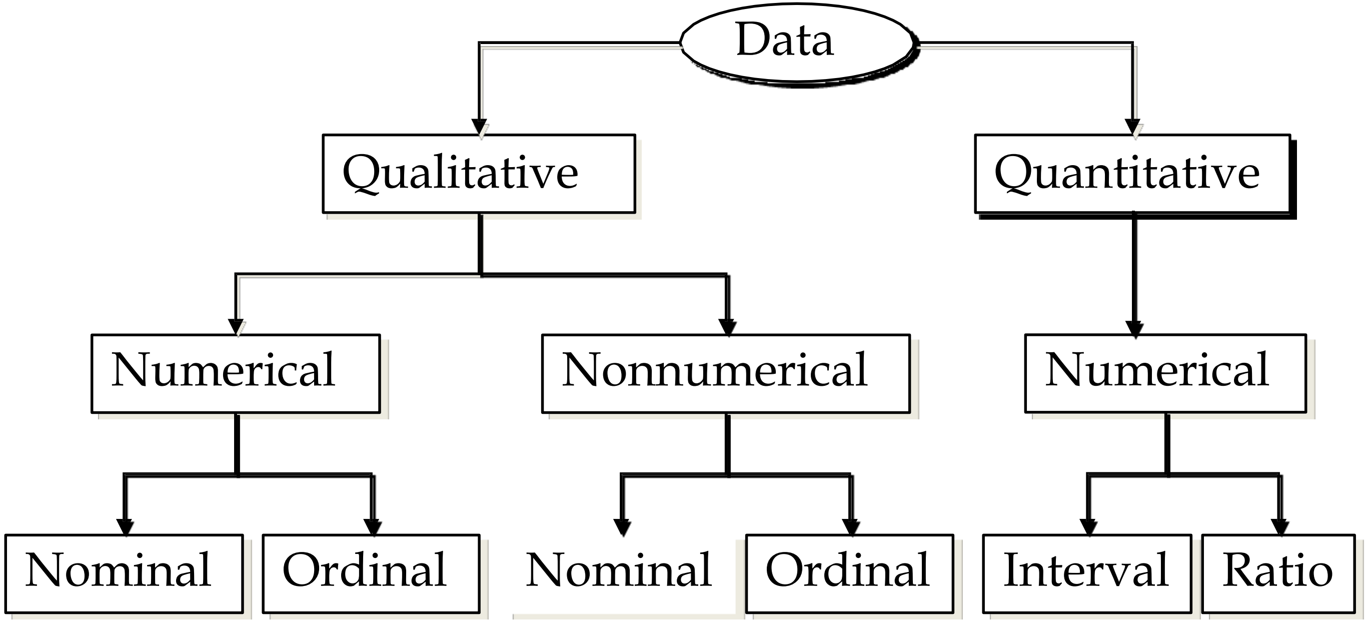
- High school GPA
- High school class rank
- Applicants score on the SAT or ACT
- Gender of applicant
- Parent’s income
- Age of applicant

Activity 2

- *The Sport Journal* (Winter 2004) reported on a study of a speed-training program for high school football players. Each participant was timed in a 40-yard sprint both before and after training. The researchers measured two variables: (1) the difference between the before and after sprint times (in seconds), and (2) the category of improvement (“improved”, “no change”, and “worse”) for each player.

- (a) Identify the type (qualitative or quantitative) of each variable measured
- (b) A total of 14 high school football players participated in the speed-training program. Does the data set collected represent a population or a sample? Explain

Scales of Measurement



Data

Qualitative

Quantitative

Numerical

Nonnumerical

Numerical

Nominal

Ordinal

Nominal

Ordinal

Interval

Ratio

Cross-Sectional Data

- Cross-sectional data are collected at the same or approximately the **same point in time**.

- Example: data detailing the number of building permits issued in June 2017 in each of the provinces of Thailand

Time Series Data

- Time series data are collected over several time periods.

- Example: data detailing the number of building permits issued in a city in the last 36 months

Data Acquisition Considerations

Time Requirement

- Searching for information can be time consuming.
- Information may no longer be useful by the time it is available.

Cost of Acquisition

- Organizations often charge for information even when it is not their
- primary business activity.

Data Errors

- Using any data that happens to be available or that were acquired with
- little care can lead to poor and misleading information.

Descriptive Statistics

- Descriptive statistics are the tabular, graphical, and numerical methods used to summarize data.

Example: Hudson Auto Repair

- The manager of Hudson Auto would like to have a better understanding of **the cost** of parts used in the engine tune-ups performed in the shop. He examines 50 customer invoices for tune-ups. The costs of parts, rounded to the nearest THB, are listed on the next slide.

Example: Hudson Auto Repair

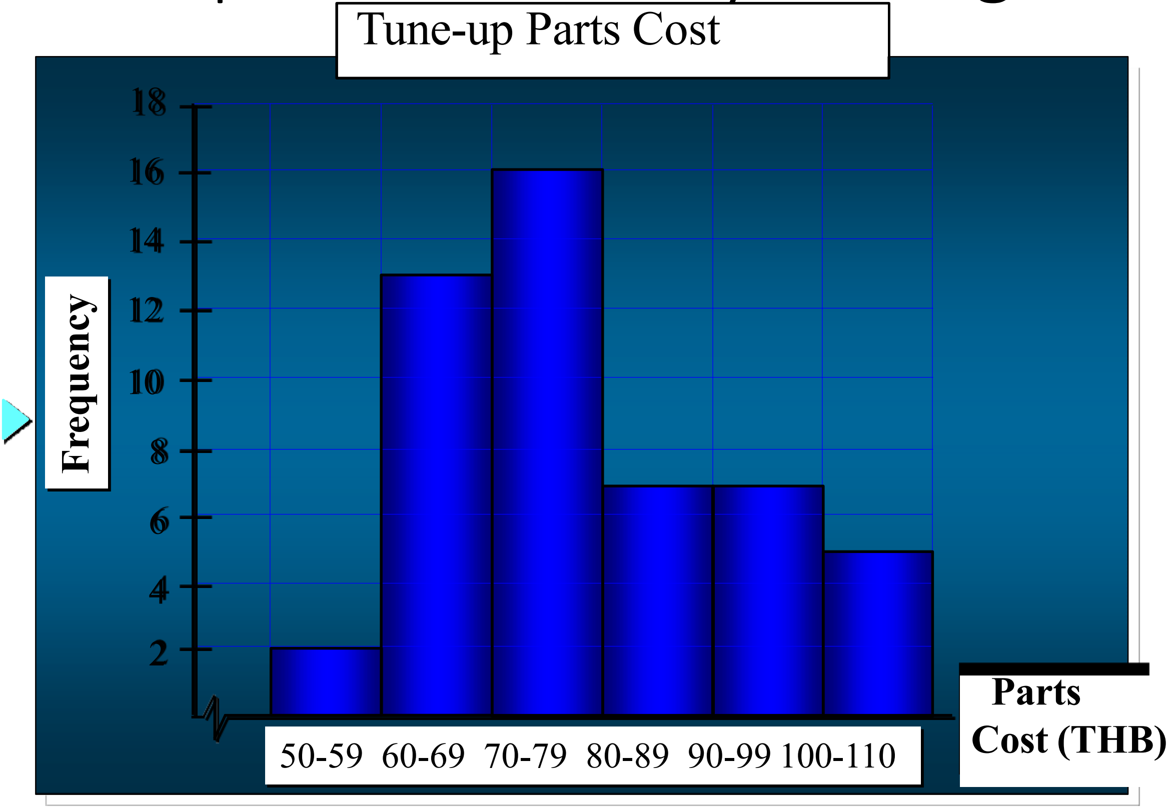
91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

Tabular Summary:
Frequency and Percent Frequency

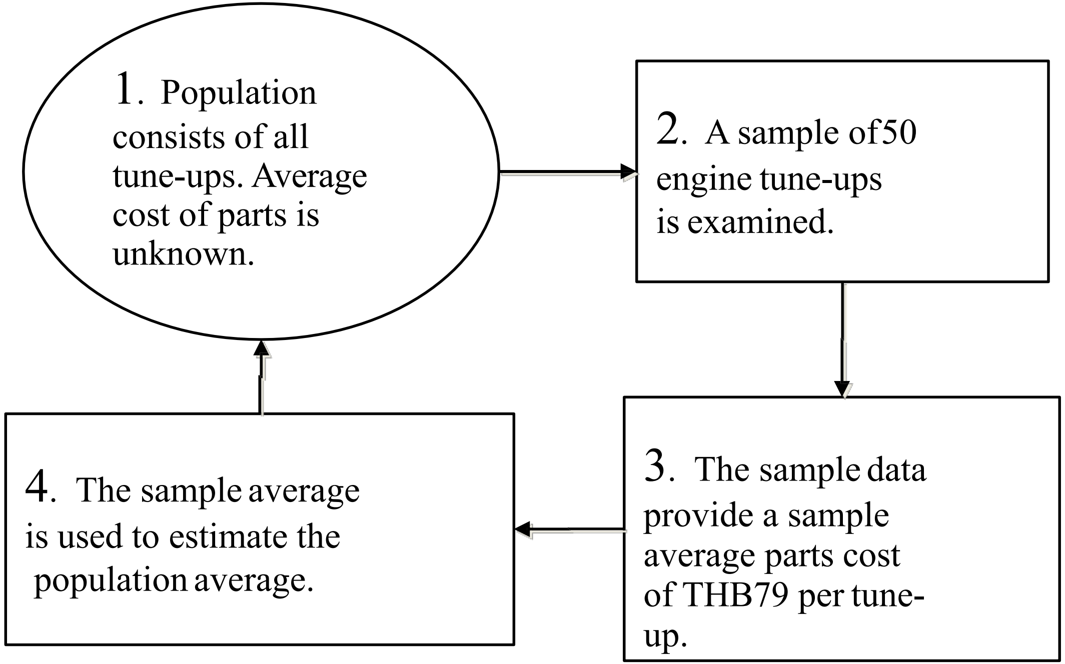
Parts Cost (THB)	Parts Frequency	Percent Frequency
	2	4
50-59	13	26
60-69	16	32
70-79	7	14
80-89	7	14
90-99	5	10
100-109	50	100

??????

Graphical Summary: Histogram



Process of Statistical Inference



1. Population consists of all tune-ups. Average cost of parts is unknown.

2. A sample of50 engine tune-ups is examined.

3. The sample data provide a sample average parts cost of THB79 per tune-up.

4. The sample average is used to estimate the population average.

Sampling Methods

Defining the Target Population

- It is critical to the success of the research project to clearly define the target population.
- Rely on logic and judgment.
- The population should be defined in connection with the objectives of the study.

Technical Terminology

- An element is an object on which a measurement is taken.
- A population is a collection of elements about which we wish to make an inference.
- Sampling units are nonoverlapping collections of elements from the population that cover the entire population.

Technical Terms

- A sampling frame is a list of sampling units.
- A sample is a collection of sampling units drawn from a sampling frame.
- Parameter: numerical characteristic of a population
- Statistic: numerical characteristic of a sample

Errors of nonobservation

- The deviation between an estimate from an ideal sample and the true population value is the sampling error.

- Almost always, the sampling frame does not match up perfectly with the target population, leading to errors of coverage.

Errors of nonobservation

- Nonresponse is probably the most serious of these errors.
  - Arises in three ways:
    - Inability of the person responding to come up with the answer
    - Refusal to answer
    - Inability to contact the sampled elements

Errors of observation

- These errors can be classified as due to the interviewer, respondent, instrument, or method of data collection.

Interviewers

- Interviewers have a direct and dramatic effect on the way a person responds to a question.
    - Most people tend to side with the view apparently favored by the interviewer, especially if they are neutral.
    - Friendly interviewers are more successful.
    - In general, interviewers of the same gender, racial, and ethnic groups as those being interviewed are slightly more successful.

Respondents

- Respondents differ greatly in motivation to answer correctly and in ability to do so.
- Obtaining an honest response to sensitive questions is difficult.
- Basic errors
  - Recall bias: simply does not remember
  - Prestige bias: exaggerates to ‘look’ better
  - Intentional deception: lying
  - Incorrect measurement: does not understand the units or definition

Census Sample

- A census study occurs if the entire population is very small, or it is reasonable to include the entire population (for other reasons).

- It is called a census sample because data is gathered on every member of the population.

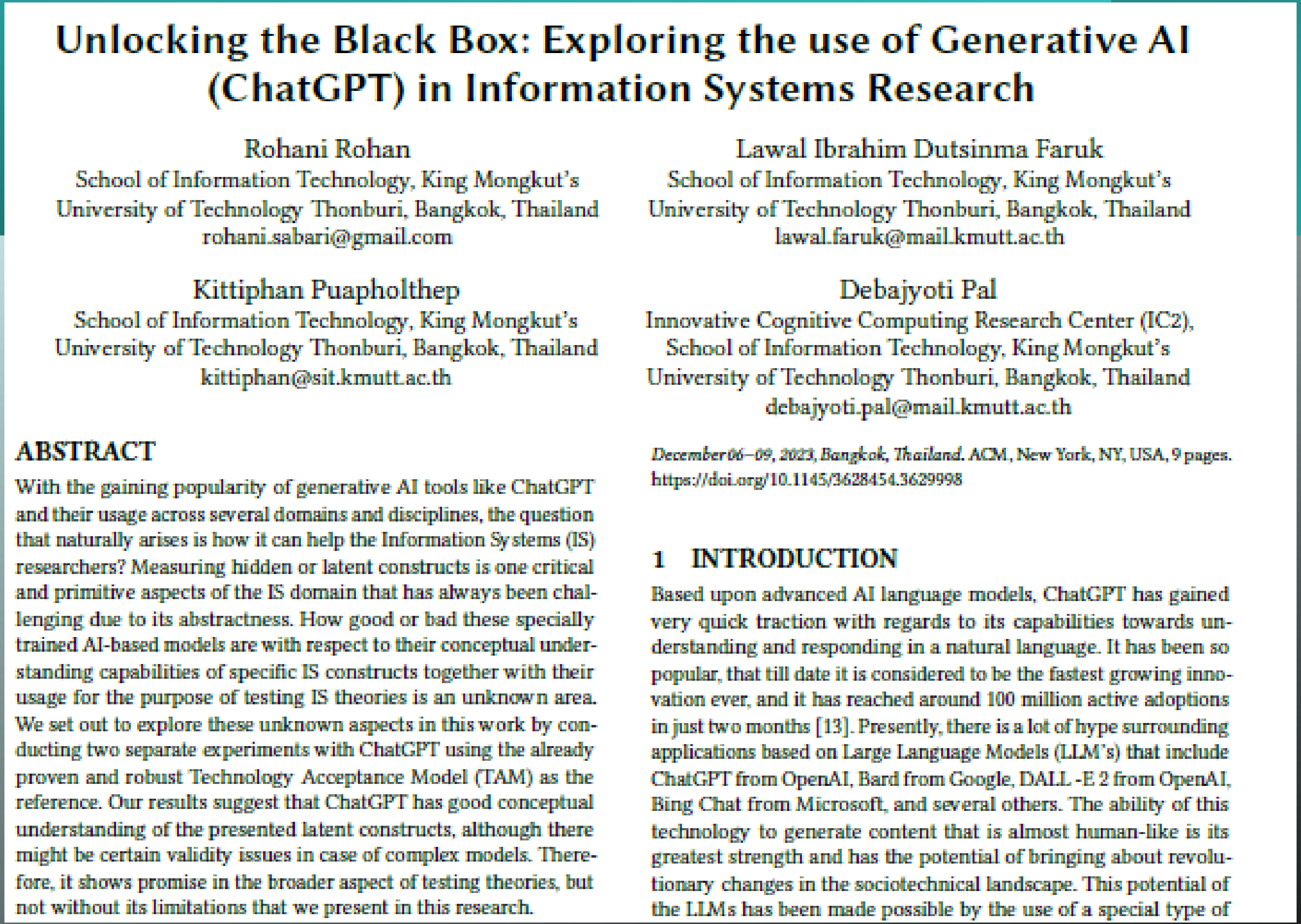
Why sample?

- The population of interest is usually too large to attempt to survey all of its members.
- A carefully chosen sample can be used to represent the population.
  - The sample reflects the characteristics of the population from which it is drawn.

Probability versus Nonprobability

- **Probability Samples:** each member of the population has a known non-zero probability of being selected
  - Methods include random sampling, systematic sampling, and stratified sampling.
- **Nonprobability Samples:** members are selected from the population in some nonrandom manner
  - Methods include convenience sampling, judgment sampling, quota sampling, and snowball sampling

Any Alternatives for Sampling yet?
What can be a possibility?



Unlocking the Black Box: Exploring the use of Generative AI (ChatGPT) in Information Systems Research

Rohani Rohan
School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
rohani.sabari@gmail.com

Lawal Ibrahim Dutsinma Faruk
School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
lawal.faruk@mail.kmutt.ac.th

Kittiphan Puapholthep
School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
kittiphan@sit.kmutt.ac.th

Debajyoti Pal
Innovative Cognitive Computing Research Center (IC2), School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
debajyoti.pal@mail.kmutt.ac.th

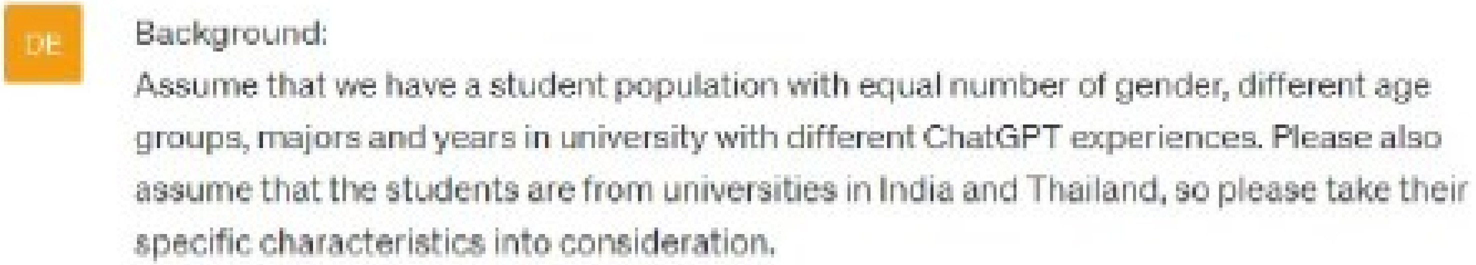
ABSTRACT

With the gaining popularity of generative AI tools like ChatGPT and their usage across several domains and disciplines, the question that naturally arises is how it can help the Information Systems (IS) researchers? Measuring hidden or latent constructs is one critical and primitive aspects of the IS domain that has always been challenging due to its abstractness. How good or bad these specially trained AI-based models are with respect to their conceptual understanding capabilities of specific IS constructs together with their usage for the purpose of testing IS theories is an unknown area. We set out to explore these unknown aspects in this work by conducting two separate experiments with ChatGPT using the already proven and robust Technology Acceptance Model (TAM) as the reference. Our results suggest that ChatGPT has good conceptual understanding of the presented latent constructs, although there might be certain validity issues in case of complex models. Therefore, it shows promise in the broader aspect of testing theories, but not without its limitations that we present in this research.

December 06–09, 2023, Bangkok, Thailand. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3628454.3629998

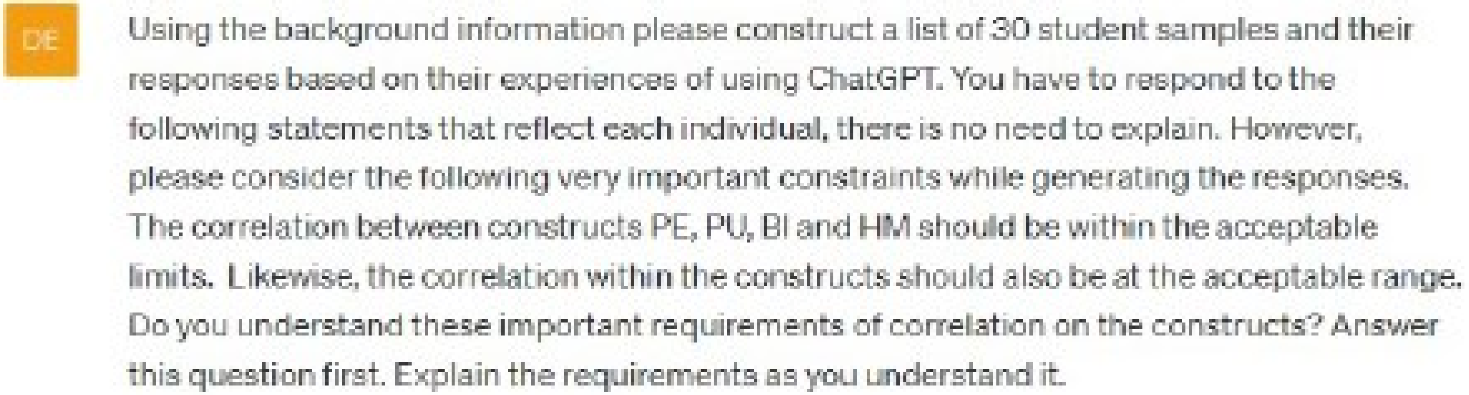
1 INTRODUCTION

Based upon advanced AI language models, ChatGPT has gained very quick traction with regards to its capabilities towards understanding and responding in a natural language. It has been so popular, that till date it is considered to be the fastest growing innovation ever, and it has reached around 100 million active adoptions in just two months [13]. Presently, there is a lot of hype surrounding applications based on Large Language Models (LLM's) that include ChatGPT from OpenAI, Bard from Google, DALL -E 2 from OpenAI, Bing Chat from Microsoft, and several others. The ability of this technology to generate content that is almost human-like is its greatest strength and has the potential of bringing about revolutionary changes in the sociotechnical landscape. This potential of the LLMs has been made possible by the use of a special type of



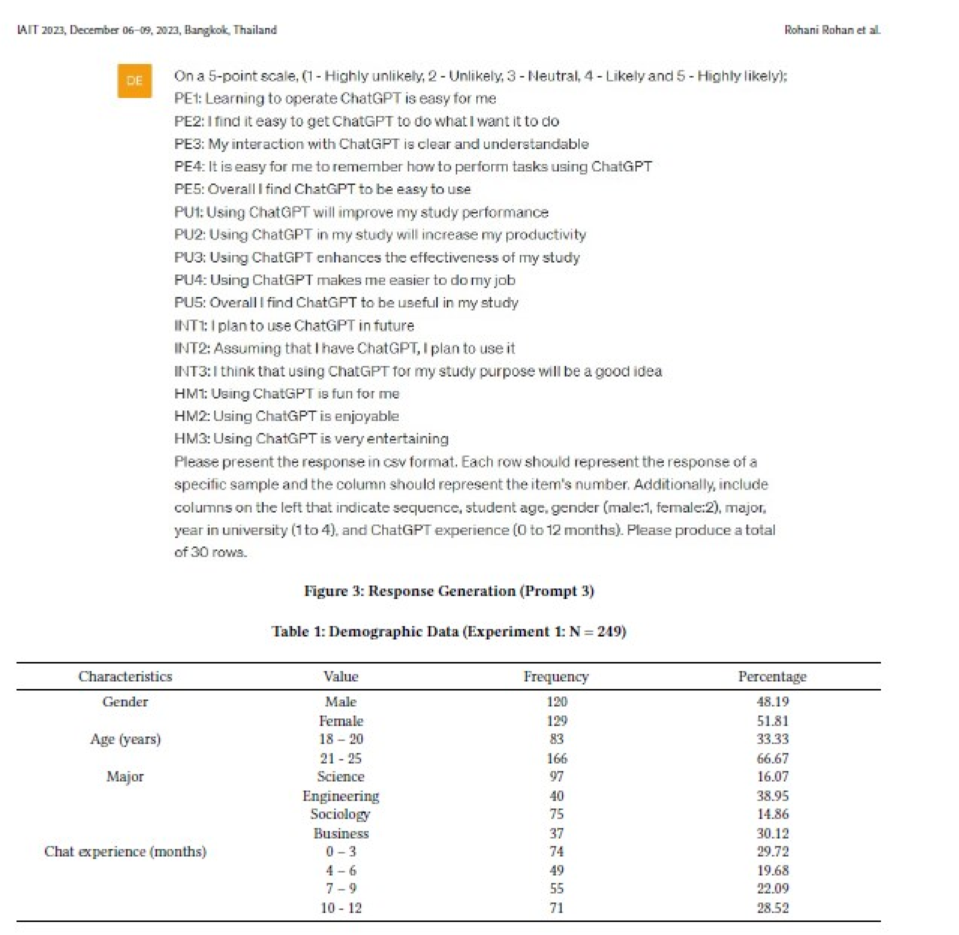
Background:
Assume that we have a student population with equal number of gender, different age groups, majors and years in university with different ChatGPT experiences. Please also assume that the students are from universities in India and Thailand, so please take their specific characteristics into consideration.

Figure 1: Background Prompt (Prompt 1)



Using the background information please construct a list of 30 student samples and their responses based on their experiences of using ChatGPT. You have to respond to the following statements that reflect each individual, there is no need to explain. However, please consider the following very important constraints while generating the responses. The correlation between constructs PE, PU, BI and HM should be within the acceptable limits. Likewise, the correlation within the constructs should also be at the acceptable range. Do you understand these important requirements of correlation on the constructs? Answer this question first. Explain the requirements as you understand it.

Figure 2: Sample Setup and Constraints (Prompt 2)



On a 5-point scale, (1 - Highly unlikely, 2 - Unlikely, 3 - Neutral, 4 - Likely and 5 - Highly likely);
PE1: Learning to operate ChatGPT is easy for me
PE2: I find it easy to get ChatGPT to do what I want it to do
PE3: My interaction with ChatGPT is clear and understandable
PE4: It is easy for me to remember how to perform tasks using ChatGPT
PE5: Overall I find ChatGPT to be easy to use
PU1: Using ChatGPT will improve my study performance
PU2: Using ChatGPT in my study will increase my productivity
PU3: Using ChatGPT enhances the effectiveness of my study
PU4: Using ChatGPT makes me easier to do my job
PU5: Overall I find ChatGPT to be useful in my study
INT1: I plan to use ChatGPT in future
INT2: Assuming that I have ChatGPT, I plan to use it
INT3: I think that using ChatGPT for my study purpose will be a good idea
HM1: Using ChatGPT is fun for me
HM2: Using ChatGPT is enjoyable
HM3: Using ChatGPT is very entertaining
Please present the response in csv format. Each row should represent the response of a specific sample and the column should represent the item's number. Additionally, include columns on the left that indicate sequence, student age, gender (male:1, female:2), major, year in university (1 to 4), and ChatGPT experience (0 to 12 months). Please produce a total of 30 rows.

Figure 3: Response Generation (Prompt 3)

Table 1: Demographic Data (Experiment 1: N = 249)

Characteristics	Value	Frequency	Percentage
Gender	Male	120	48.19
	Female	129	51.81
Age (years)	18 – 20	83	33.33
	21 - 25	166	66.67
Major	Science	97	16.07
	Engineering	40	38.95
	Sociology	75	14.86
	Business	37	30.12
Chat experience (months)	0 – 3	74	29.72
	4 – 6	49	19.68
	7 – 9	55	22.09
	10 - 12	71	28.52

Random Sampling

Random sampling is the purest form of probability sampling.

- Each member of the population has an equal and known chance of being selected.
- When there are very large populations, it is often ‘difficult’ to identify every member of the population, so the pool of available subjects becomes biased.
  - You can use software, such as SPSS to generate random numbers or to draw directly from the columns

Systematic Sampling

- **Systematic sampling** is often used instead of random sampling. It is also called an Nth name selection technique.

- After the required sample size has been calculated, every Nth record is selected from a list of population members.

- As long as the list does not contain any hidden order, this sampling method is as good as the random sampling method.

- Its only advantage over the random sampling technique is simplicity (and possibly cost effectiveness).

Stratified Sampling

- **Stratified sampling** is commonly used probability method that is superior to random sampling because it reduces sampling error.

- A stratum is a subset of the population that share at least one common characteristic; such as males and females.

    - Identify relevant stratums and their actual representation in the population.

    - Random sampling is then used to select a *sufficient* number of subjects from each stratum.

    - Stratified sampling is often used when one or more of the stratums in the population have a low incidence relative to the other stratums.

Cluster Sampling

- Cluster Sample: a probability sample in which each sampling unit is a collection of elements.
- Effective under the following conditions:
  - A good sampling frame is not available or costly, while a frame listing clusters is easily obtained
- Examples of clusters:
  - City blocks – political or geographical
  - Housing units – college students
  - Hospitals – illnesses

Convenience Sampling

- **Convenience sampling** is used in exploratory research where the researcher is interested in getting an inexpensive approximation.
- The sample is selected because they are convenient.
- It is a nonprobability method.
  - Often used during preliminary research efforts to get an estimate without incurring the cost or time required to select a random sample

Judgment Sampling

- **Judgment sampling** is a common nonprobability method.
- The sample is selected based upon judgment.
  - an extension of convenience sampling
- When using this method, the researcher must be confident that the chosen sample is truly representative of the entire population.

Snowball Sampling

- **Snowball sampling** is a special nonprobability method used when the desired sample characteristic is rare.

- It may be extremely difficult or cost prohibitive to locate respondents in these situations.

- This technique relies on referrals from initial subjects to generate additional subjects.

- It lowers search costs; however, it introduces bias because the technique itself reduces the likelihood that the sample will represent a good cross section from the population.

Sample Size?

- The more heterogeneous a population is, the larger the sample needs to be.
- Depends on topic – frequently it occurs?
- For probability sampling, the larger the sample size, the better.
- With nonprobability samples, not generalizable regardless – still consider stability of results

Response Rates

- About 20 – 30% usually return a questionnaire
- Follow up techniques could bring it up to about 50%
- Still, response rates under 60 – 70% challenge the integrity of the random sample
- How the survey is distributed can affect the quality of sampling

Activity

- Poll on alien spacecraft: “have you ever seen anything that you believe was a spacecraft from another planet? This was the question put to 1,500 American adults in a national poll conducted by ABC News and The Washington Post. The pollsters used random-digit telephone dialing to contact adult Americans until 1,500 responded. Ten percent (i.e., 150) of the respondents answered that they had, in fact, seen an alien spacecraft. No information was provided on how many adults were called and, for one reason or another, did not answer the question.

- A) Identify the data collection method
- B) Identify the target population
- C) Comment on the validity of the survey results

Activity

- The PROMISE Software Engineering Repository, hosted by the University of Ottawa, is a collection of publicly available datasets to serve researchers in building prediction software models. A PROMISE dataset on software reuse, saved in in the **SWREUSE** file, provides information on the success or failure of reusing previously developed software for each in a sample of 24 new software development projects (*Data source: IEEE Transactions on Software Engineering, vol. 28, 2002*). Of the 24 projects, 9 were judged failures and 15 were successfully implemented.

- A) Identify the experimental units for this study.
- B) Describe the population from which the sample is selected.
- C) What is the variable of interest in the study? Is it quantitative or qualitative?
- D) Critically evaluate the statement “Since 15/24 = 0.625, it follows that 62.5% of all new software development projects will be successfully implemented”.