**Faculty of Computing and Information Technology/Department of Computer Science and Mathematics**

# BACS3013 Data Science

## Tutorial 3 (Data Collection, Sampling and Pre-processing - part 1)

**Note:**

Level of measurement: nominal, ordinal, interval, binary or ratio. (also known as scale of measurement)
Type of variable: continuous or discrete, qualitative or quantitative.

Q1. A group of women tried five brands of fingernail polish and ranked them according to preference. What level of measurement is this?

**The rankings are ordinal. While the rankings indicate which brand is preferred over another, they do not measure how much more they are preferred.**

Q2. What level of measurement is a person's "favorite sport"?

**Nominal. The variable, person's "favorite sport," is a label with no natural order and cannot be ranked or ordered.**

Q3. The Nielsen Ratings break down the number of people watching a particular television show by age. What level of measurement is age?

**Age is a ratio variable because it has a zero point, and the ratio between two values is meaningful.**

Q4. What type of variable is "pounds of popcorn" served at a movie theater?

**Continuous. "Pounds of popcorn" can assume any value within a range. There are no "gaps" in the scale.**

Q5. What type of variable is the number of auto accidents reported in a given month?

**Discrete. The number of auto accidents is counted and must be a whole number, such as 0, 500, or 3,125,874.**

Q6. The performance of personal and business investments is measured as a percentage, "return on investment." What type of variable is "return on investment"?

**Continuous. "Return on investment" can assume any value within a range. There are no "gaps" in the scale.**

Q7. Describe the difference between a population and a sample.

## BACS3013 Data Science

**A population is the entire set of individuals or objects that could be observed or measured. A sample is a subset or portion of a population.**

Q8. A New York newspaper reported the average gasoline prices in four metropolitan areas and used a bar chart to illustrate the differences. What type of statistics was shown? What activities did the newspaper use to make the report?

**The newspaper used descriptive statistics. The statistical techniques used to make the report were collecting data, summarizing the data, and presenting the data.**

Q9. A company was studying the demographics of their customers. As part of the study, they collected the following variables: gender, marital status, credit rating (low, medium, high), annual income, and age. Label each variable as qualitative or quantitative, discrete or continuous, and nominal, ordinal, interval, or ratio.

**Gender: qualitative, discrete, nominal;**
**Marital status: qualitative, discrete, nominal;**
**Credit rating: qualitative, discrete, ordinal;**
**Annual income: quantitative, continuous, ratio;**
**Age: quantitative, continuous, ratio.**

Q10. After the graduation ceremonies at a university, six graduates were asked whether they were in favor of (identified by 1) or against (identified by 0) abortion. Some information about these graduates is shown below.

| Graduate | Sex | Age | Abortion | Issue | Class Rank |
|---|---|---|---|---|---|
| Nancy | F | 22 | 1 | 3 | |
| Michael | M | 21 | 1 | 2 | |
| Tammy | F | 33 | 0 | 1 | |
| John | M | 38 | 0 | 20 | |
| Marlene | F | 25 | 1 | 4 | |
| David | M | 19 | 0 | 8 | |

a. How many variables are in the data set?
b. How many observations are in the data set?
c. Name the level of measurement for each of the above (Sex, Age, Abortion Issue, Class rank).
d. Which of the above (Sex, Age, Abortion Issue, Class rank) are categorical and which are continuous variables?
e. Are arithmetic operations appropriate for the variable "abortion issue"?

**ANS:**

**a.** **4**
**b.** **6**
**c.** **Sex: nominal**
**Age: ratio**
**Abortion Issue: nominal**
**Class rank: ordinal**
**d.** **Sex: categorical**
**Age: continuous**
**Abortion Issue: categorical**

**Class Rank: categorical**

**e. No**

Q11. An issue of Fortune Magazine reported that the following companies had the lowest sales per employee among the Fortune 500 companies.

| Company | Sales per Employee ($1000s) | Sales Rank |
|---|---|---|
| Seagate Technology | $42.20 | 285 |
| SSMC | 42.19 | 414 |
| Russell | 41.99 | 480 |
| Maxxam | 40.88 | 485 |
| Dibrell Brothers | 22.56 | 470 |

a. How many variables are in the above data set?
b. How many observations are in the above data set?
c. Name the scale of measurement for each of the variables.
d. Name the variables and indicate whether they are categorical or continuous.

**ANS:**

**a. 2**
**b. 5**
**c. Sales per employee: ratio; Sales rank: ordinal**
**d. Sales per employee: continuous ; Sales rank: categorical**

Q12. A pharmaceutical company is performing clinical trials on a new drug that is intended to relieve symptoms for allergy sufferers. Twelve percent of the 300 clinical trial participants experienced the side effect dry mouth.

a. What is the population being studied?
b. What is the sample being studied?
c. Based on the sample, what percentage of the population do you think would suffer from dry mouth?

**ANS:**

**a. All allergy sufferers**
**b. The 300 participants**
**c. 12%**

Q13. A pharmaceutical company is performing clinical trials on a new drug that is intended to relieve symptoms for allergy sufferers. Twelve percent of the 300 clinical trial participants experienced the side effect dry mouth.

a. What is the population being studied?
b. What is the sample being studied?

c. Based on the sample, what percentage of the population do you think would suffer from dry mouth?

**ANS:**

**a. All allergy sufferers**

**b. The 300 participants**

**c. 12%**

Q14. Suppose the current weather report for your area contains the following information. Specify the measurement scale for each of the variables.

| | |
|---|---|
| Temperature | 84o |
| Wind Speed | 10 mph |
| Wind Direction | (from the) South |
| Sky Description | Sunny |
| Molds Level | High |

**ANS:**

**Temperature - interval**<br>
**Wind Speed - ratio**<br>
**Wind Direction - nominal**<br>
**Sky Description - nominal**<br>
**Molds Level - ordinal**

Q15. Molly Porter owns and operates two convenience stores, one on the East side of the city and the other on the South side. She has workforce-planning decisions to make and has collected some recent sales data that are relevant to her decisions. Listed below are the monthly sales ($000) at her two stores for the past six months.

| Store | March | April | May | June | July | August |
|---|---|---|---|---|---|---|
| East | 102 | 100 | 103 | 105 | 109 | 106 |
| South | 72 | 74 | 81 | 86 | 92 | 93 |

a. Is the data set cross-sectional or time series data? Explain.

b. Comment on any apparent patterns you see in the data.

**ANS:**

**a. Time series data for two variables: monthly sales for East store and monthly sales for South store.**

**b. Both stores have been experiencing an overall rise in sales during the past six months. The South store's increase in sales (as a percentage of sales) has been greater than the East store's increase. The increases might be temporary, due to the seasonal nature of demand. It is also possible that the increases will continue.**

Q16. The Bureau of Transportation Statistics Omnibus Household Survey is conducted annually and serves as an information source for the U.S. Department of Transportation. In one part of the survey the person being interviewed was asked to respond to the following statement: "Drivers of motor vehicles should be allowed to talk on a hand-held cell phone while driving." Possible response were strongly agree, somewhat agree, somewhat disagree, and strongly disagree. Forty-four respondents said that they strongly agree with this statement, 130 said that they somewhat agree, 165 said they somewhat disagree, and 741 said they strongly disagree with this statement (Bureau of Transportation website, August 2010).

a. Do the responses for this statement provide categorical or continuous data?

b. Would it make more sense to use averages or percentage as a summary of the responses for this statement?

c. What percentage of respondents strongly agrees with allowing drivers of motor vehicles to talk on a hand-held cell phone while driving?

d. Do the results indicate general support for or against allowing drivers of motor vehicles to talk on a hand-held cell phone while driving?

**Ans:**

**a. Categorical**

**b. Percentages**

**c. 44 of 1080 respondents or approximately 4% strongly agree with allowing drivers of motor vehicles to talk on a hand-held cell phone while driving.**

**d. 165 of the 1080 respondents or 15% of said they somewhat disagree and 741 or 69% said they strongly disagree. Thus, there does not appear to be general support for allowing drivers of motor vehicles to talk on a hand-held cell phone while driving.**

Q17. A BusinessWeek North America subscriber study collected data from a sample of 2861 subscribers. Fifty-nine percent of the respondents indicated an annual income of $75,000 or more and 50% reported having an American Express credit card.

a. What is the population of interest in this study?

b. Is annual income a categorical or continuous variable?

c. Is ownership of an American Express card a categorical or continuous variable?

d. Does this study involve cross-sectional or time series data?

e. Describe any statistical inferences BusinessWeek might make on the basis of the survey.

**a. All subscribers of Business Week in North America at the time the survey was conducted.**

**b. continuous**

**c. Categorical (yes or no)**

**d. Cross-sectional - all the data relate to the same time.**

**e. Using the sample results, we could infer or estimate 59% of the population of subscribers have an annual income of $75,000 or more and 50% of the population of subscribers have an American Express credit card.**