Statistics for Data Science - 1

Practice assignment solutions

Week 1

1. **A quality engineer wants to check the quality of steel rods produced in a steel factory. For this, 40 pieces of steel rods are randomly selected from the steel factory to assess their quality. Based on this, choose the correct option.**

   A. The population is all steel rods produced in all factories; the sample is the 40 steel rods selected from the given steel factory's production.

   B. The population is all steel rods produced in all factories; the sample is all the steel rods produced in the given steel factory.

   C. The population is all steel rods produced in the given steel factory; the sample is the 40 steel rods selected from the given steel factory's production.

   D. All the steel rods in the given steel factory are population; the sample is all steel rods in the given steel factory.

   Answer: C
   By definition, population is the entire collection of elements we are interested in. Here, the engineer wants to check the quality of steel rods produced in the given steel factory. Hence, population will be all steel rods produced in the given factory. Sample is a subset of the population which is being studied. Since the engineer is studying only a set of 40 rods collected from that factory, the sample is the set of 40 selected steel rods.

   Options A and B are wrong as the engineer is interested only in the given factory's rods and not of all factories'.

2. **Which of the following statements is(are) true?**

   A. All basic mathematical operations can be performed on some structured data.

   B. All basic mathematical operations can be performed on unstructured data.

   C. Email contents, text messages, and audio files are usually unstructured data.

   D. Height(cm), Weight(Kg), Age(years) are structured data.

   Answer: A, C, D
   Option (A) Consider the income of employees of a certain company. It is certainly a structured data and we can perform mathematical operations such as mean and/or sum on the income of employees.
   Option (B) Unstructured data may not necessarily have numeric properties; therefore,

in general, mathematical operations can not be performed on unstructured data.
Option (C) Email contents, text messages or audio files can not be organised in any specific way as these do not have any inherent order; therefore, these are unstructured data.
Option (D) Height, weight, and age have definite order and can always be arranged in an order. Hence, they are structured data.

3. **Values of temperature and humidity of a room are measured for 24 hours at a regular time interval of 30 minutes. Based on this, choose the correct option.**

    A. It is a cross-sectional data.

    B. It is time series data.

    C. None of the above.

Answer: B
Since the temperature and humidity are recorded over a period of time at regular intervals, the data collected is time series data.

4. **Which of the following is(are) numerical variable(s)?**

    A. Height(cm)

    B. Day of the week

    C. Jersey number of sports player

    D. Mobile number

    E. Email address

    F. Age in years

Answer: A, F
Option (A) Since height has numeric properties and can have arithmetic operations performed on it, it follows that height is a numerical variable.
Option (B) The days of a week belong to a certain category in the set {sunday, monday, . . . , saturday}. Hence, it is a categorical variable and not a numerical variable.
Option (C) Jersey numbers are just labels assigned to players for identification. As we can see, Dhoni's jersey number 7 is in no way greater than or lesser than Gambhir's jersey number 5. Similarly, a player with jersey number 12 is not the sum of two players whose jersey numbers add up to 12 (say jersey numbers 5 and 7). That is, it is meaningless to perform mathematical operations on any two jersey numbers. Therefore, it is not a numerical variable.
Option (D) Mobile numbers neither have any order nor can we perform any standard arithmetic operations on them. Hence, mobile number is also not a numerical variable.
Option (E) Since the email address also does not have any numeric property, it is also not a numerical variable.
option(F) Age has numeric property and we can perform arithmetic operations on age.

For instance, we can calculate average age of a group of people. Hence, it is a numerical variable.

5. **Which of the following variable(s) has(have) ratio scale of measurement?**

   A. Temperature in Kelvin

   B. Temperature in Centigrade

   C. Year

   D. Angle measured in degrees

   Answer: A, D
   Option (A) Temperature in Kelvin has ratio scale of measurement because temperature has meaningful intervals and it also has absolute zero.
   Option (B) Temperature in Centigrade has no absolute zero, however, we can perform addition and subtraction operations on it. Therefore, it comes under interval scale of measurement.
   Option (C) Year has ordinal scale of measurement as we can not perform addition and subtraction operations on year but we can arrange it in increasing or decreasing order.
   Option (D) Angles in degree has ratio scale of measurement as we can compare the intervals or differences between different angles and it also has absolute zero.

6. **Which of the following mathematical operation(s) can be performed on interval variables?**

   A. Addition

   B. Subtraction

   C. Multiplication

   D. Division

   Answer: A, B
   Addition and subtraction can be performed on variables with interval scale of measurement as they have a definite difference between them but multiplication and subtraction upon these variables is not possible because difference between them is not comparable; moreover, they do not have absolute zero.

7. **Pin code is a numerical variable.**

   A. True

   B. False

   Answer: B
   Pin code is not a numerical variable as there is no ordering possible among various pin codes. For example, there is no order among pin codes 100002, 500001, 500002 i.e., pin code 100002 is neither greater nor lesser than 500001 or 500002.

8. **Which of the following is(are) expected while selecting a sample for a population?**

    A. Sample should be a subset of the population.

    B. Sample can contain data that is not from the population.

    C. Sample should be representative of the characteristics of different elements in the population.

    D. Sample need not be representative of the characteristics of different elements in the population.

    Answer: A, C
    By definition, a sample must be a subset of the population and must be representative of the characteristics of different elements of population. The purpose of a sample is to get information about the population.

9. **In the 2011 Cricket ODI World Cup quarter-final match between India and Australia, a media organization estimated that Australia would beat India by 50 runs if Australia bats first, based on the information of matches played between the two teams previously. Which branch of statistics does the above analysis belong to?**

    A. Descriptive Statistics

    B. Inferential Statistics

    Answer: B
    Making predictions from the data comes under inferential statistics. Here, media makes prediction based on the information it has about the matches played between two teams previously; therefore, given analysis belongs to inferential statistics.

10. **A class teacher wants to collect feedback from students of the class. The teacher hands out a blank sheet to each student to obtain descriptive input and suggestions on the class. The data collected by the class teacher is:**

    A. Structured Data

    B. Unstructured Data

    Answer: B
    Students are going to add feedback and suggestions in non organised or non tabular format. Hence, generated data will be unstructured.

11. **Variables with an interval scale of measurement may be converted into a ratio scale of measurement by performing?**

    A. Addition operation

    B. Subtraction operation

    C. Multiplication operation

D. Cannot be converted to ratio variables.

Answer: B
Variables with interval scale of measurement can be converted into other variables with ratio scale of measurement by performing subtraction.
If we make a new variable by subtracting a variable with an interval scale, the new variable will have absolute zero as one of its values. We obtain absolute zero when we subtract one value from itself. So, the new variable has a ratio scale of measurement. For example, in restaurant 1 and restaurant 2, the rating given by users is noted. The rating given by the user should be an integer from 1 to 5. So, the rating given by user to restaurant 1 and 2 is an interval scale since it has fixed measure and no absolute zero. But if we are interested in the difference (absolute value) between ratings given by the same user, then the new variable can take values from 0 to 4. This variable has absolute zero. Hence, it has a ratio scale of measurement.

12. **Which of the following operations can be valid for categorical variables?**

    A. Addition

    B. Subtraction

    C. Comparison ($>$, $<$, $=$)

    D. Multiplication

    E. Division

Answer: C
Arithmetic operations can not be performed on categorical variables because they do not have numeric properties. From the given operations, the only operation applicable on categorical variables is comparison.

13. **What is the scale of measurement for the amount of money you have?**

    A. Nominal

    B. Ordinal

    C. Ratio

    D. Interval

Answer: C
Amount of money can have a meaningful interval. It also has an absolute zero. Hence, it comes under the ratio scale of measurement.

14. **What is the scale of measurement for the military titles - Major, Captain, Colonel?**

    A. Nominal

    B. Ordinal

    C. Ratio

D. Interval

Answer: B
Military titles have a definite rank but they do not have numeric properties; therefore, they have ordinal scale of measurement.