Homework 1 (Theoretical part)

Assignment's date: 5 October 2023
Student ID: 1933541

Question 1: What is Statistics and its relationship with other disciplines? Make examples of use in cybersecurity.

Statistics is a discipline or a science that collects and analyses data. There are many disciplines that use statistics:

In Biology, the use of statistics is known as biostatistics or biometrics. It consists in collecting and analysing the data received from experiments in order to decide results. Also, meteorology uses statistics in stochastic-dynamic prediction and weather forecasting.

Physics and Mathematics use probability theory and statistics dealing with the estimation of large populations. For example, the phenomenological results of thermodynamics were developed using the mechanics of statistics.

Information technology uses statistics to predict particular outcomes. In particular, statistics is used also in cybersecurity to identify intrusions and anomalous behaviour and therefore protect against cyber-attacks. Using Statistics, machine learning and Big Data analytics were developed tools to perform anomaly detection. Statistical techniques are used also in classification, data mining, streaming data analysis, graph analysis, and machine learning.

Question 1-bis: What is the Difference between Descriptive and Inferential Statistics.

The difference between Descriptive and Inferential Statistics is that: Descriptive Statistics consists in observe and analyse data from an entire known population. Inferential Statistics consists in observe a sample from a larger, unknown population. It involves making conclusions about the larger population based on the analysis of the sample.

Question 2: Describe the concepts of variable, attribute, population, sample and dataset

- Attribute: is a general abstract concept. In fact, it is a characteristic or a quality of an object.
- Variable: is a logical set of attributes. It can also vary across the variable's domain among individuals and from one observation to another.
- Population: is a set of similar items called statistical units by which is possible to collect data that are necessary for a specific survey.
- Sample: is a population's subset. It must be representative and not biased, this means that the shape of the sample's result has to be similar to the population one.
- Dataset: is a collection of data that is usually represented by a table. It is composed by some columns that are the attribute observed, and many rows which are the data collected from every statistical unit.
- Level of measurement: Level of measurement refers to the process of categorizing data based on their characteristics and properties. It represents the different set of values that a variable can assume.

Question 3: Briefly describe the main sampling techniques

The sampling techniques are used to choose a representative sample from a population. Reducing the number of individuals in a study reduces the cost and workload.

There are different sampling techniques and they can be subdivided into two groups: probability sampling and non-probability sampling. In probability (random) sampling, you start with a complete sampling frame of all individuals from which you select your sample. In this way, all individuals have a chance of being chosen for the sample, so it is possible to have a generalise results of the study. In non-probability sampling, you do not start with a complete sampling frame, so some individuals have no chance of being selected. This method is usually cheaper. Some examples are:

1. Simple random sampling (Probability): consists to give an identificatory number to each individual in a population, and then use a table of random numbers to decide which individuals have to be included. Whit this technique, each individual is chosen entirely by chance and each member of the population has an equal probability, of being selected.
2. Clustered sampling (Probability): the population is subdivided in subgroups, called cluster, that are used as the sampling unit and they are randomly selected to be included in the study. Cluster sampling can be efficient when a study takes place over a wide geographical region. But, if the chosen clusters are not representative of the population, the risk of bias will increase.
3. Convenience sampling (Non-Probability): all participants are selected based on availability and willingness to take part. There is a significant risk of bias and the sample may not be representative.
4. Snowball sampling (Non-Probability): Consists to ask to existing subjects nominate other individuals known to them, so the sample increases in size like a rolling snowball. Snowball sampling can be effective when a sampling frame is difficult to identify. However, there is a significant risk of selection bias.

Question 4: Briefly describe the main experiment designs.

The experiment design, is the design of any task that aims to describe and explain the variation of information under conditions that are hypothesized to reflect the variation.

Some experiment designs are:

1. Bayesian experimental design provides a general probability-theoretical framework from which other theories on experimental design can be derived. It is based on Bayesian inference to interpret the data collected during the experiment. This allows accounting for both any prior knowledge on the parameters to be determined as well as uncertainties in observations.
2. Optimal designs are a class of experimental designs that are optimal respect to some statistical criterion. This classification was created by Kirstine Smith and it allow parameters to be estimated without bias and with minimum variance. Also it can reduce the costs of experimentation, because non-optimal design requires a greater number of experimental runs to estimate the parameters with the same precision as an optimal design.

Bibliography:

1. Lecture notes from the lesson of the statistics course
2. Statistics.com: Link 2
3. Udemy.com: Link 3
4. Imperial college of London: Link 4
5. Wikipedia - Design of experiments: Link 5
6. Health knowledge: Link 6
7. Wikipedia- variables and attributes: link 7