Work Sheets

A. Statistics

1. Bernoulli random variables take (only) the values 1 and 0.

Ans- True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

Ans- Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

Ans- Modeling bounded count data

4. Point out the correct statement.
a) The exponent of a normally distributed random variables follows what is called the log-normal distribution
b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
c) The square of a standard normal random variable follows what is called chi-squared distribution
d) All of the mentioned

Ans- d

5. ______ random variables are used to model rates.

Ans- Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.

Ans- False

7. Which of the following testing is concerned with making decisions using data?

Ans- Hypothesis

8. Normalized data are centered at______and have units equal to standard deviations of the original data.

Ans- 0

9. Which of the following statement is incorrect with respect to outliers?

Ans- c) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

- A normal distribution is the proper term for a probability bell curve.
- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.

- In a normal distribution, the mean, mode and median are all the same.

- Normal distribution curves are sometimes designed with a histogram inside the curve.

11. **How do you handle missing data? What imputation techniques do you recommend?**

Ans- There are many ways to handle missing data. Imputation simply means replacing the missing values with an estimate, then analysing the full data set as if the imputed values were actual observed values.
The following are common methods:

Mean imputation

Simply calculate the mean of the observed values for that variable for all individuals who are non-missing.

It has the advantage of keeping the same mean and the same sample size

Hot deck imputation

A randomly chosen value from an individual in the sample who has similar values on other variables.

In other words, find all the sample subjects who are similar on other variables, then randomly choose one of their values on the missing variable.

Cold deck imputation

A systematically chosen value from an individual who has similar values on other variables.

This is similar to Hot Deck in most ways, but removes the random variation.

Regression imputation

The predicted value obtained by regressing the missing variable on other variables.

So instead of just taking the mean, you're taking the predicted value, based on other variables. This preserves relationships among variables involved in the imputation model, but not variability around predicted values.

Stochastic regression imputation

The predicted value from a regression plus a random residual value.

This has all the advantages of regression imputation but adds in the advantages of the random component.

Most multiple imputation is based off of some form of stochastic regression imputation.

I would recommend mean imputation and regression imputation

12. What is A/B testing?

Ans - A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. AB testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

A/B testing allows individuals, teams, and companies to make careful changes to their user experiences while collecting data on the results. This allows them to construct hypotheses, and to learn better why certain elements of their experiences impact user behaviour.

13. Is mean imputation of missing data acceptable practice?

Ans- It's a popular solution to missing data, despite its drawbacks. Mainly because it's easy. It can be really painful to lose a large part of the sample you so carefully collected.
There are many reasons not to use mean imputation

a) imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. That’s a good thing.Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That’s good too. If all you are doing is estimating means (which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias your parameter estimate. It will still bias your standard error

b) We get the same mean from mean-imputed data that we would have gotten without the imputations. And yes, there are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small. Because the imputations are themselves estimates, there is some error associated with them. It treats it as real data. Ultimately, because your standard errors are too low, so are your p-values. which in turn giving Type I errors without realizing it.

14. What is linear regression in statistics?

Ans - Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0). The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. After a regression line has been computed for a group of data, a point which lies far from the line (and thus has a large residual value) is known as an outlier. Such points may represent erroneous data, or may indicate a poorly fitting regression line. Once a regression model has been fit to a group of data, examination of the residuals (the deviations from the fitted line to the observed values) allows to check the validity of assumption that a linear relationship exists.

15. What are the various branches of statistics?

Ans - Two branches, descriptive statistics and inferential statistics, comprise the field of statistics.

Descriptive Statistics

This branch of statistics focuses on collecting, summarizing, and presenting a set of data. Descriptive statistics has two parts

- Central tendency measures
- Variability measures

Measures of Central Tendency

Central tendency measures specifically help statisticians evaluate the distribution center of values. These tendency measures are:

Mean

Mean is a conventional method used to describe the central tendency. Typically, to calculate the average of values, count all values, and then divide them with the number of available values.

Median

It is the result that is in the middle of a set of values. An easy way to calculate the median is to edit the results in numerical journals and locate the result that is in the center of the distributed sample.

Mode

The mode is the frequently occurring value in the given data set.

Measures of Variability

The variability measure helps statisticians to analyse the distribution that is spreading from a specific data set. Some of the variables of variability include quartiles, ranges, variances, and standard deviation.

Inferential Statistics

This branch of statistics analyses sample data to draw conclusions about a population.

The inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population.

Different types of inferential statistics include:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis

B. SQL

1. Which of the following is/are DDL commands in SQL?

   Ans- A) Create B) Delete C) Alter

2. Which of the following is/are DML commands in SQL?
   Ans- A) Update C) Select B) Delete

3. Full form of SQL is:
   Ans - Structured Query Language

4. Full form of DDL is:
   Ans- Data Definition Language

5. DML is:
   Ans- Data Manipulation Language

6. Which of the following statements can be used to create a table with column B int type and C float type?

Ans- Create Table A (B int,C float)

7. Which of the following statements can be used to add a column D (float type) to the table A created above?

   Ans- B) Alter Table A ADD COLUMN (D float)

8. Which of the following statements can be used to drop the column added in the above question?

   Ans- B) Alter Table A Drop Column D

9. Which of the following statements can be used to change the data type (from float to int) of the column D of table A created in above questions?

   Ans. B) Alter Table A Alter Column D int

10. Suppose we want to make Column B of Table A as primary key of the table. By which of the following statements we can do it?

    Ans- C) Alter Table A Add Primary key B

11. What is data-warehouse?

A **Data warehouse** is typically used to connect and analyse business data from heterogeneous sources. The data warehouse is the core of the BI system which is built for data analysis and reporting. It is a process of transforming data into information and making it available to users in a timely manner to make a difference. A Data Warehouse works as a central repository where information arrives from one or more data sources.

Data may be:

1. Structured
2. Semi-structured
3. Unstructured data

The data is processed, transformed, and ingested so that users can access the processed data in the Data Warehouse through Business Intelligence tools, SQL clients, and spreadsheets.

Three main types of Data Warehouses (DWH) are:

1. **Enterprise Data Warehouse (EDW):**

Enterprise Data Warehouse (EDW) is a centralized warehouse. It provides decision support service across the enterprise. It also provide the ability to classify data according to the subject and give access according to those divisions.

2. **Operational Data Store**:

Operational Data Store, which is also called ODS, are nothing but data store required when neither Data warehouse nor OLTP systems support organizations reporting needs. In ODS, Data warehouse is refreshed in real time.

3. **Data Mart:**

A data mart is a subset of the data warehouse. It specially designed for a particular line of business, such as sales, finance, sales or finance. In an independent data mart, data can collect directly from sources.

12. What is the difference between OLTP VS OLAP?

Online Analytical Processing (OLAP) – Online Analytical Processing consists of a type of software tools that are used for data analysis for business decisions. OLAP provides an environment to get insights from the database retrieved from multiple database systems at one time.

Examples – Any type of Data warehouse system is an OLAP system. Uses of OLAP are as follows:

- Spotify analysed songs by users to come up with the personalized homepage of their songs and playlist.
- Netflix movie recommendation system.

Online transaction processing (OLTP) – Online transaction processing provides transaction-oriented applications in a 3-tier architecture. OLTP administers day to day transaction of an organization.

Examples – Uses of OLTP are as follows:

- ATM center is an OLTP application.
- OLTP handles the ACID properties during data transaction via the application.
- It's also used for Online banking, Online airline ticket booking, sending a text message, add a book to the shopping cart.

13. What are the various characteristics of data-warehouse?

The key characteristics of a data warehouse are as follows:

- Some data is denormalized for simplification and to improve performance

- Large amounts of historical data are used
- Queries often retrieve large amounts of data
- Both planned and ad hoc queries are common
- The data load is controlled

In general, fast query performance with high data throughput is the key to a successful data warehouse.

14. What is Star-Schema?

Star schema is the fundamental schema among the data mart schema and it is simplest. This schema is widely used to develop or build a data warehouse and dimensional data marts. It includes one or more fact tables indexing any number of dimensional tables. The star schema is a necessary case of the snowflake schema. It is also efficient for handling basic queries.
It is said to be star as its physical model resembles to the star shape having a fact table at its center and the dimension tables at its peripheral representing the star's points.

In Star Schema, Business process data, that holds the quantitative data about a business is distributed in fact tables, and dimensions which are descriptive characteristics related to fact data.

15. What do you mean by SETL?

Semantic Extract-Transform-Load

Semantic ETL (SETL), a unified framework for processing and integrating data semantically by bridging Semantic Web and Data Warehouse technologies.

SETL allows including semantically annotated data (Resource Description Framework data) in the analytical process. To process a Non Semantic Data Source, it builds a semantic layer on top of the source.

SETL follows a demand-driven approach to design a (MD(Multidimensional Model)) SDW. The first step of this approach is to identify and analyse the information requirements of business users and decision makers. Then, based on the gathered requirements, the next two steps of the Data Warehouse are to build the Data Warehouse schema and to build the ETL. As the data in a SDW should be semantically connected, we assume the SDW to be a Knowledge base and it allows to produce an MD schema for the SDW to benefit from OLAP. The semantic ETL process populates the SDW from the data sources according to the semantics captured in the schema.

C. Machine Learning

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:

   Ans – 4

2. In which of the following cases will K-Means clustering fail to give good results?

   a. Data points with outliers

b. Data points with different densities

c. Data points with round shapes

d. Data points with non-convex shapes

Ans - d) 1,2and4

3. The most important part of _____is selecting the variables on which clustering is based

Ans - d) formulating the clustering problem

4. The most commonly used measure of similarity is the ____ or its square.

Ans- a) Euclidean distance

5. ________is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.

Ans- b) Divisive clustering

6. Which of the following is required by K-means clustering?

a) Defined distance metric

b) Number of clusters

c) Initial guess as to cluster centroids

d) All answers are correct

Ans- d) All answers are correct

7. The goal of clustering is to-

   Ans- a) Divide the data points into groups

8. Clustering is a-

   Ans- b) Unsupervised learning

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

   a) K- Means clustering

   b) Hierarchical clustering

   c) Diverse clustering

   d) All of the above

   Ans- d) All of the above

10. Which version of the clustering algorithm is most sensitive to outliers?

    Ans- a) K-means clustering algorithm

11. Which of the following is a bad characteristic of a dataset for clustering analysis-

    a) Data points with outliers

    b) Data points with different densities

    c) Data points with non-convex shapes

    d) All of the above

Ans- d) All of the above

12. For clustering, we do not require-

Ans- a) Labeled data

13. **How is cluster analysis calculated?**

Clusters can be calculated using various grouping methods. These can be divided into

- graph-theoretical
- hierarchically
- partitioning
- optimizing

The hierarchical cluster analysis follows three basic steps: 1) calculate the distances, 2) link the clusters, and 3) choose a solution by selecting the right number of clusters.

The k-Means method is one of the most widely used non-hierarchical methods. It is a partitioning method, which is particularly suitable for large amounts of data.

- First, an initial partition with k clusters (given number of clusters) is created.
- Then, starting with the first object in the first cluster, Euclidean distances of all objects to all cluster foci are calculated.
- If an object is detected whose distance to the center of gravity of the own cluster is greater than the

distance to the center of gravity (centroid) of another cluster, this object is shifted to the other cluster.

- Finally, the centroids of the two changed clusters are calculated again, since the compositions have changed here.
- These steps are repeated until each object is located in a cluster with the smallest distance to its centroid (center of the cluster) (optimal solution).

14. How is cluster quality measured?

"external clustering quality metrics" like Purity, Homogeneity, Completeness, V-Measure, Precision, Recall, Normalized Mutual Information and CMM (for streaming environments).

"internal clustering quality metrics" includes Silhouette coefficient and Sum of Squared Distances (SSQ).

To measure a cluster's fitness within a clustering, we can compute the average silhouette coefficient value of all objects in the cluster. To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set. The silhouette coefficient and other intrinsic measures can also be used in the elbow method to heuristically derive the number of clusters in a data set by replacing the sum of within-cluster variances.

15. What is cluster analysis and its types?

Cluster analysis is an exploratory analysis that tries to identify structures within the data. Cluster

analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known. The different cluster analysis methods that SPSS offers can handle binary, nominal, ordinal, and scale (interval or ratio) data.
Cluster analysis is often used in conjunction with other analyses (such as discriminant analysis). three methods for the cluster analysis are: **K-Means Cluster, Hierarchical Cluster, and Two-Step Cluster.**
K-means cluster is a method to quickly cluster large data sets. The researcher define the number of clusters in advance. This is useful to test different models with a different assumed number of clusters.

Hierarchical cluster is the most common method. It generates a series of models with cluster solutions from 1 (all cases in one cluster) to n (each case is an individual cluster).

Two-step cluster analysis identifies groupings by running pre-clustering first and then by running hierarchical methods. Because it uses a quick cluster algorithm upfront, it can handle large data sets that would take a long time to compute with hierarchical cluster methods. In this respect, it is a combination of the previous two approaches. Two-step clustering can handle scale and ordinal data in the same model, and it automatically selects the number of clusters.

END of Document