Paper : DSE-A-1

(Biostatistics)

Full Marks : 50

The figures in the margin indicate full marks. Candidates are required to give their answers in their own words as far as practicable.

(1) (a) A **frequency distribution** is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval. The frequency is how often a value occurs in an interval while the distribution is the pattern of frequency of the variable[1].

(b) Two limitations of statistics are:

1. **Study of Numerical Facts only**: Statistics studies only such facts as can be expressed in numerical terms. It does not study qualitative phenomena like honesty, friendship, wisdom, health, patriotism, justice, etc[3].
2. **Study of Aggregates only**: Statistics studies only the aggregates of quantitative facts. It does not study statistical facts relating to any particular unit[3].

© A **discrete variable** is any variable that can only take on a certain number of distinct values, typically represented by whole numbers. These variables can be subdivided into dichotomous types, which only take two values, and polytomous types, taking three or more[4].

(d) **Primary data** refers to data that is collected firsthand by a researcher or a team of researchers for a specific research project or purpose. It is original information that has not been previously published or analyzed, and it is gathered directly from the source or through the use of data collection methods such as surveys, interviews, observations, and experiments[5]. For example, a restaurant might collect primary data by asking customers to rate their dining experience[5].

e) **Null Hypothesis:**

- **Definition:** The null hypothesis is a statement in statistical hypothesis testing that suggests there is no significant difference, effect, or relationship between groups or variables.
- **Example:** In a drug trial, the null hypothesis might state that there is no difference in the average recovery time between patients who receive the drug and those who receive a placebo.

[6].

(f) The **statistical error** is the difference between the collected data's obtained value and the actual value of the collected data[7]. A **mistake** arises because of miscalculations, use of wrong methods of calculations and wrong interpretation of the result[8]. So, while a statistical error is a deviation from accuracy or correctness, a mistake is an action or judgment that is misguided or wrong[9].

g) **Population and Sample:**

- **Population:** The population is the entire set of individuals or instances about whom statistical inferences are being made.
- **Sample:** A sample is a subset of the population selected for a particular study. It is used to make inferences about the entire population.

[10].

(h) A **cumulative frequency distribution** is the sum of the class and all classes below it in a frequency distribution[11]. It is used to identify the number of observations that lie above or below the particular frequency in the provided data set[12]. For example, if you're interested in studying a population to find out a "more" or "less" question, you might want to know how many people in a particular geographic area spend up to a certain amount per year in groceries[11].

(2) (a) What are the advantages of 'Arithmatic mean' and 'mode value'?

(a) The advantages of the **Arithmetic Mean** are[1234]:

1. **Easy to Understand**: The arithmetic mean is very easy to understand. It's simply the sum of all the numbers in a set divided by the number of numbers in that set.
2. **Useful for Large Data Sets**: If you have a large data set, the arithmetic mean can be a very useful statistic. It gives you a single number that summarizes the data and makes it easier to compare different data sets.
3. **Provides a Good Estimate of the "Typical" Value**: The arithmetic mean provides a good estimate of the "typical" value in a data set. This means that if you were to pick a random number from the data set, the arithmetic mean would be a good guess for what that number might be.
4. **Useful for Calculating Other Statistics**: The arithmetic mean is a useful statistic for calculating other statistics like the variance and standard deviation. These statistics give us information about how spread out the data is and how much it varies from the mean.
5. **Works with Different Types of Data**: The arithmetic mean can be used with different types of data, including interval, ratio, and some nominal data. This means that it's a versatile statistic that can be used in a wide range of situations.

The advantages of the **Mode** are[5][6][7][8]:

1. **Easy to Understand and Calculate**: Mode is one of the central tendencies that is easy to understand and calculate.
2. **Not Affected by Extreme Values**: Mode is not affected by extreme values.
3. **Easy to Identify in a Dataset**: Mode is easy to identify in a dataset as it depends on the frequency of the observations.
4. **Useful for Qualitative Data**: Mode is useful for qualitative data.
5. **Can be Computed in an Open-Ended Frequency Table**: Mode can be computed in an open-ended frequency table.
6. **Can be Located Graphically**: Mode can be located graphically.

(b) How does the standard deviation help for analysing the data in case of normal distribution? What is bimodal distribution?

(b) In the case of a **normal distribution**, the **standard deviation** is a particularly valuable measure of spread[1][2]. It represents the typical distance between each data point and the mean[3]. Here's how it helps in analyzing data:

1. **Identifying Variability**: The standard deviation summarizes the variability in a dataset. Smaller values indicate that the data points cluster closer to the mean, signifying relatively consistent values. Conversely, higher values signify that the values spread out further from the mean, indicating more dissimilar data values and a higher likelihood of extreme values[3].
2. **Understanding Probability**: In a normal distribution, a specific percentage of data falls within a certain number of standard deviations from the mean. For example, about 68% of the data falls within one standard deviation from the mean, about 95% falls within two standard deviations, and about 99.7% falls within three standard deviations[1]. This property is known as the Empirical Rule or the 68-95-99.7 rule[4].
3. **Standardizing Distributions**: The standard deviation is used to standardize a normal distribution, transforming it into a standard normal distribution with a mean of 0 and a standard deviation of 1[5]. This standardization allows for easy calculation of the probability of certain values occurring and facilitates the comparison of different datasets[5].

A **bimodal distribution** is a probability distribution with two modes or peaks[6][7]. This type of distribution often arises due to some underlying phenomena or when two different groups are lumped together[6][7]. For example, if you measure the heights of two different species of plants in a field without realizing it, the resulting distribution of heights would be bimodal[7]. Similarly, if you were to analyze the distribution of customers at a restaurant by hour, you'd likely find a bimodal distribution with peaks during lunch and dinner hours[7]. Recognizing a bimodal

distribution can provide valuable insights into the underlying processes or populations that are affecting the data[6].

© Five persons A, B, C, D, E occupy seats in a row at random. What is the probability that A and B sit next to each other?

The probability that two specific people (in this case, A and B) will sit next to each other in a row can be calculated as follows:

1. **Total number of arrangements**: The total number of ways in which 5 people can be seated in a row is given by 5!, which is 120.
2. **Favorable arrangements**: Consider A and B as a single entity. Then, we have 4 entities (AB, C, D, E). The number of ways these can be arranged is 4!, which is 24. However, within the AB entity, A and B can switch places, so we have 2! = 2 ways of arranging A and B. Therefore, the total number of favorable cases is 4! * 2!, which is 48.

So, the probability that A and B sit next to each other is given by the number of favorable arrangements divided by the total number of arrangements, which is

48/120=0.4

So, there is a 40% chance that A and B will sit next to each other.

Superiority of Standard Deviation:

The standard deviation is often regarded as superior to other measures of dispersion for several reasons:

1. **Incorporation of All Data Points:**

- The standard deviation takes into account the deviation of each data point from the mean, considering the entire dataset. This means that it considers the variability across all data points rather than just focusing on extremes.

2. **Sensitivity to Variability:**

- The standard deviation is sensitive to changes in individual data points, providing a more nuanced understanding of the variability within the dataset. It reflects the average deviation from the mean and is not heavily influenced by outliers.

3. **Mathematical Properties:**

- The standard deviation is mathematically well-defined and allows for the application of statistical inference techniques. It plays a crucial role in parametric statistical tests and is a fundamental component in the calculation of confidence intervals.

4. **Comparison between Distributions:**

- Standard deviation allows for the comparison of the spread of different datasets, enabling researchers to assess and compare the variability within and between groups.

Chief Defect of Standard Deviation:

While the standard deviation is a powerful measure of dispersion, its chief defect lies in the fact that it is influenced by outliers or extreme values. Outliers can significantly impact the standard deviation, leading to potentially misleading interpretations of variability. In cases where the dataset contains extreme values, the standard deviation may not accurately represent the typical spread of the majority of the data.

Now, let's calculate the mean and standard error of grain length based on the given data:

```plaintext
Grain Length (mm) | No. of Grains
------------------|--------------
9-11              | 3
12-14             | 5
15-17             | 9
18-20             | 3
```

Calculations:

1. Calculate the Mean (μ):
   $\mu = \frac{\sum (x \cdot f)}{N}$
   where x is the midpoint of the class interval, f is the frequency, and N is the total number of observations.
   $\mu = \frac{(10 \cdot 3)+(13 \cdot 5)+(16 \cdot 9)+(19 \cdot 3)}{20} = \frac{30+65+144+57}{20} = \frac{296}{20} = 14.8 \text{ mm}$
2. Calculate the Standard Error (SE):
   $SE = \frac{\sigma}{\sqrt{N}}$
   where σ is the standard deviation, and N is the total number of observations.

Since the standard deviation is not provided, we'll assume it. Let's assume $\sigma = 2$:

$$SE = \frac{2}{\sqrt{20}} \approx \frac{2}{4.47} \approx 0.45 \text{ mm}$$

Therefore, the mean grain length is 14.8 mm, and the standard error is approximately 0.45 mm (assuming a standard deviation of 2).