**Statistics Fundamentals Assignment**

**Based on "Session 1: (Statistics Fundamentals)"**

**Part 1: True/False Questions (1 point each)**

*Indicate whether each statement is True (T) or False (F).*

1. **T/F:** Statistics are primarily used for summarizing data and have little relevance in AI and Machine Learning.False
2. **T/F:** A population refers to a subset of individuals selected from a larger group for study.False
3. **T/F:** Categorical variables can be either nominal or ordinal. True
4. **T/F:** The number of cars in a parking lot is an example of a continuous variable. False
5. **T/F:** Descriptive statistics help make generalizations about a larger population based on sample data. False
6. **T/F:** In a data matrix, each row represents a variable and each column represents a case. False
7. **T/F:** The mean is highly sensitive to outliers. True
8. **T/F:** The median is calculated by summing all values and dividing by the number of values. False
9. **T/F:** A dataset can have more than one mode. True
10. **T/F:** The Interquartile Range (IQR) is less sensitive to outliers than the range. True
11. **T/F:** Variance is expressed in the same units as the original data, making it easily interpretable. False
12. **T/F:** A small standard deviation indicates that data values are spread out over a wider range. False
13. **T/F:** Histograms are useful for visualizing the shape of data distribution. True
14. **T/F:** The probability of an impossible event is 1. False
15. **T/F:** If two events are independent, the occurrence of one affects the occurrence of the other. False
16. **T/F:** Bayes' Theorem allows us to update probability estimates based on new evidence. True
17. **T/F:** Inferential statistics are used to summarize data, not to draw conclusions. False
18. **T/F:** The Central Limit Theorem states that the sampling distribution of the sample mean will always be skewed, regardless of sample size. False
19. **T/F:** The null hypothesis (H0​) is a statement that there is no effect or no difference. True
20. **T/F:** If the p-value is greater than the significance level (α), we reject the null hypothesis. False

**Part 2: Short Answer Questions (3 points each)**

*Answer the following questions concisely.*

1. Explain two reasons why learning statistics is important for AI and Machine Learning.

 **understanding data behavior**.

**model evaluation and validation**.

1. Differentiate between a "Population" and a "Sample" in statistical terms, providing an example for each.

 **Population:** The **entire group of individuals, items, or observations** you’re interested in studying. It includes **all possible members** of a defined group.

**Example:** All university students in Egypt.

 **Sample:** A **subset selected from the population** for analysis, used when it's impractical to study the whole population.

**Example:** 5 randomly selected university students from Cairo, Alexandria, and Mansoura universities.

1. Describe the difference between "Nominal" and "Ordinal" categorical variables with an example for each.

 **Nominal Variables:**  
**Categorical variables with no inherent order or ranking** between the categories. They simply name or label different groups.

**Example:**  
Blood types: A, B, AB, O

**Ordinal Variables: Categorical variables with a meaningful order or ranking** among the categories, but the differences between the ranks aren’t necessarily equal.

**Example:**  
Customer satisfaction levels: Poor, Fair, Good, Excellent

1. What is the primary purpose of a "Frequency Table" and when would you use it instead of a "Data Matrix"?

**Quickly summarize the distribution of one variable**, especially for categorical or discrete numerical data.

**Identify the most or least common values** in a dataset.

1. Given the dataset [10,12,15,12,18,13,12], calculate the:

a. Mean

b. Median

c. Mode

**a. Mean:**

13.14

**b. Median:** 12

**Mode:12**

1. Explain why the median is often a better measure of central tendency than the mean for skewed distributions.

The **median is not affected by outliers**

The **mean is sensitive to outliers**

1. Define "Standard Deviation" and explain what a large standard deviation indicates about a dataset.

-The **standard deviation** is a measure of the **spread or dispersion of a dataset relative to its mean.** It indicates how much, on average, each data point differs from the mean ofthe dataset.

**-** A **large standard deviation** means that the data points are **widely spread out around the mean** showing high variability. This suggests that values in the dataset differ greatly from the average value.

1. What is a "Box Plot" and what five key numbers does it display?

A box plot (or box-and-whisker plot) is a graphical representation of a dataset’s distribution that shows its spread, center, and potential outliers in a compact, visual form. It’s especially useful for comparing distributions across groups.

**The Five Key Numbers a Box Plot Displays (the Five-Number Summary):**

1. **Minimum** — the smallest data value (excluding outliers)
2. **First Quartile (Q1)** — the 25th percentile
3. **Median (Q2)** — the 50th percentile (middle value)
4. **Third Quartile (Q3)** — the 75th percentile
5. **Maximum** — the largest data value (excluding outliers)
6. State the "Addition Rule" for mutually exclusive events in probability and provide an example.

If two events cannot happen at the same time (i.e., they are mutually exclusive), the probability of either event occurring is the sum of their individual probabilities.

P(A or B)=P(A)+P(B)P(A \text{ or } B) = P(A) + P(B)P(A or B)=P(A)+P(B)

**Example:**  
When rolling a standard 6-sided die:

* Let A = rolling a 2 → P(A)=16P(A) = \frac{1}{6}P(A)=61​
* Let B = rolling a 5 → P(B)=16P(B) = \frac{1}{6}P(B)=61​

Since you can't roll both a 2 and a 5 at the same time, these are mutually exclusive.

P(2 or 5)=16+16=26=13P(2 \text{ or } 5) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}P(2 or 5)=61​+61​=62​=31​

1. Briefly explain the concept of "Conditional Probability" and provide its formula.

**Concept:**  
Conditional probability is the probability of an event A occurring given that another event Bhas already occurred. It reflects how the likelihood of an event changes when we know additional information.

**Formula:**

P(A∣B)=P(A∩B)P(B)P(A|B) = \frac{P(A \cap B)}{P(B)}P(A∣B)=P(B)P(A∩B)​

Where:

* P(A∣B)P(A|B)P(A∣B) = Probability of A given B
* P(A∩B)P(A \cap B)P(A∩B) = Probability of both A and B happening
* P(B)P(B)P(B) = Probability of event B (must be greater than 0)

1. In the context of hypothesis testing, what is the significance of the "P-value" and the "Significance Level (α)"?

The p-value tells you how compatible your data is with H₀.

The significance level (α) sets the standard for how unlikely the result must be before we reject H₀.

1. Describe what a "Confidence Interval" represents and how it is interpreted.

A **confidence interval (CI)** is a **range of values, derived from sample data, that is likely to contain the true population parameter (like a mean or proportion) with a certain level of confidence.**

**How It’s Interpreted:**  
If you construct a **95% confidence interval,** it means that if you were to take many random samples and build a confidence interval from each**, about 95% of those intervals would contain the true population parameter.**

**Part 3: Problem Solving (5 points each)**

*Solve the following problems, showing your work where applicable.*

1. Probability: You have a bag containing 5 red balls, 3 blue balls, and 2 green balls.

a. What is the probability of drawing a red ball?

b. What is the probability of drawing a blue or a green ball?

c. If you draw a red ball and do not replace it, what is the probability of drawing another red ball?

a) 0.5

b) 0.5

c) 4/9 ≈ 0.444

1. Measures of Dispersion: For the dataset [2,5,7,8,10,12,15,18], calculate the:

a. Range

b. Interquartile Range (IQR) - Show steps for finding Q1 and Q3.

a) Range = 16

b) IQR = 8.75 (with Q1 = 5.5 and Q3 = 14.25)

1. Hypothesis Testing Scenario: A new teaching method is introduced, and the school claims that students' average test scores will increase from the historical average of 75. A sample of 40 students who used the new method achieved an average score of 78 with a sample standard deviation of 10.

***Use hypotheses testing codes from attachments to answer the following questions:***

a. State the null and alternative hypotheses for this scenario.

b. If the p-value for this test is 0.03, and the significance level (α) is set at 0.05, what decision would you make regarding the null hypothesis? Justify your answer.