15 problems based on Descriptive Stats: 30 mins



Introduction to Statistics

1. Problem:

A researcher collects the following data on the heights (in cm) of a sample of five plants:

120, 125, 130, 135, 140 .

Classify the type of data as:

a) Structured b) unstructured

c) Numerical d) categorical

2. Problem:

A survey records the following data for 10 individuals: their age, favorite color, and hours

spent on social media per day.

Identify the types of data for:

a) Age

b) Favorite color

c) Hours spent on social media



Measures of Central Tendency

3. Problem:

Calculate the mean, median, and mode for the dataset:

3, 7, 7, 10, 15, 20 .

4. Problem:

The weights (in kg) of five parcels are: 12, 15, 18, 21, 25 .

Add an outlier weight of 50 . How does this affect the mean and median?



Measures of Dispersion

5. Problem:

Find the range and interquartile range (IQR) for the dataset:

5, 10, 15, 20, 25, 30, 35 .

6. Problem:

A dataset has a standard deviation of . If all values in the dataset are doubled, what is the 5

new standard deviation?

7. Problem:

Calculate the coefficient of variation for a dataset with a mean of 50 and a standard

deviation of .



Correlation and Skewness

8. Problem:

|  |  |  |  |
| --- | --- | --- | --- |
| Two variables, X | and Y | , have a correlation coefficient of 0.85 | . Interpret this value. |

9. Problem:   
A dataset has a positive skew. Which measure of central tendency (mean, median, or mode) is likely the largest?

When a dataset has a **positive skew** (also called **right skew**), the **mean** is usually the **largest** measure of central tendency.

* In a **positively skewed** distribution, there are a few **large outliers** (high values) pulling the tail to the right.
* These high values **increase the mean** more than they affect the **median** or **mode**.

### Order of Central Tendency in Positive Skew:

Mode<Median<Mean\{Mode} < \text{Median} < \{Mean}Mode<Median<Mean

Mean is the largest

10. Problem:   
Calculate the Pearson correlation coefficient for the following paired data: X : 1, 2, 3, 4   
Y : 2, 4, 6, 8

The Pearson correlation coefficient (**r**) is calculated using the formula:

r=n∑xy−(∑x)(∑y)[n∑x2−(∑x)2][n∑y2−(∑y)2]r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}r=[n∑x2−(∑x)2][n∑y2−(∑y)2]​n∑xy−(∑x)(∑y)​

### ****Step 1: Organize the data****

| **X** | **Y** | **XY** | **X2X^2X2** | **Y2Y^2Y2** |
| --- | --- | --- | --- | --- |
| 1 | 2 | 2 | 1 | 4 |
| 2 | 4 | 8 | 4 | 16 |
| 3 | 6 | 18 | 9 | 36 |
| 4 | 8 | 32 | 16 | 64 |
|  |  |  |  |  |
| **Σ** |  | **60** | **30** | **120** |
|  | **20** |  |  |  |
| **ΣX = 10**, **ΣY = 20**, **ΣXY = 60**, **ΣX² = 30**, **ΣY² = 120** |  |  |  |  |

### ****Step 2: Plug the values into the formula****

r=4(60)−(10)(20)[4(30)−(10)2][4(120)−(20)2]=240−200(120−100)(480−400)r = \frac{4(60) - (10)(20)}{\sqrt{[4(30) - (10)^2][4(120) - (20)^2]}} = \frac{240 - 200}{\sqrt{(120 - 100)(480 - 400)}}r=[4(30)−(10)2][4(120)−(20)2]​4(60)−(10)(20)​=(120−100)(480−400)​240−200​ r=4020×80=401600=4040=1r = \frac{40}{\sqrt{20 \times 80}} = \frac{40}{\sqrt{1600}} = \frac{40}{40} = 1r=20×80​40​=1600​40​=4040​=1

### ****Final Answer:****

r=1​

This indicates a **perfect positive linear correlation** between **X** and **Y**.



Five Point Summary and Visualization

11. Problem:   
Determine the five-point summary for the dataset: 5, 8, 12, 14, 18, 20, 24 .

To determine the **five-point summary** for the dataset [5,8,12,14,18,20,24][5, 8, 12, 14, 18, 20, 24][5,8,12,14,18,20,24], we need to calculate the following:

1. **Minimum (Min)**: The smallest value in the dataset.
2. **First Quartile (Q1)**: The median of the lower half of the dataset (25th percentile).
3. **Median (Q2)**: The middle value of the dataset (50th percentile).
4. **Third Quartile (Q3)**: The median of the upper half of the dataset (75th percentile).
5. **Maximum (Max)**: The largest value in the dataset.

### Step-by-Step Calculation:

1. **Minimum (Min)**: The smallest value in the dataset is **5**.
2. **Maximum (Max)**: The largest value in the dataset is **24**.
3. **Median (Q2)**:
   * Since the dataset has 7 elements, the median is the middle element.
   * The middle element is the 4th value, which is **14**.
4. **First Quartile (Q1)**:
   * The lower half of the data (values less than the median) is [5,8,12][5, 8, 12][5,8,12].
   * The median of this subset is **8** (middle value).
5. **Third Quartile (Q3)**:
   * The upper half of the data (values greater than the median) is [18,20,24][18, 20, 24][18,20,24].
   * The median of this subset is **20** (middle value).

### Five-Point Summary:

* **Min** = 5
* **Q1** = 8
* **Median (Q2)** = 14
* **Q3** = 20
* **Max** = 24

Thus, the five-point summary is:

5, 8, 14, 20, 24\{5, 8, 14, 20, 24}5, 8, 14, 20, 24

12. Problem:   
A box plot shows the median closer to Q1, with a long tail extending to the right. What does this indicate about the dataset's skewness?

If a box plot shows the median closer to **Q1** (the first quartile), with a **long tail extending to the right**, this indicates that the dataset is **positively skewed** or **right-skewed**.

### Explanation:

* **Median closer to Q1**: The median being closer to the first quartile suggests that the lower half of the data points (from the minimum to the median) are more tightly packed compared to the upper half (from the median to the maximum).
* **Long tail to the right**: The presence of a long tail extending to the right indicates that there are some higher-value outliers or a small number of extreme values that pull the distribution to the right.

In summary, this type of box plot suggests that the **majority of the data** is concentrated on the **lower end**, with a few larger values causing the **right-skewness** in the distribution.

13. Problem:   
Construct a histogram for the following dataset: 2, 2, 3, 3, 3, 4, 5, 6, 6, 7 .

Suggest appropriate bin sizes.

To construct a histogram for the dataset [2,2,3,3,3,4,5,6,6,7][2, 2, 3, 3, 3, 4, 5, 6, 6, 7][2,2,3,3,3,4,5,6,6,7], we need to decide on appropriate bin sizes. A histogram is a way to represent the frequency distribution of a dataset, and the bins (or intervals) determine how the data is grouped.

### Step 1: Calculate the range of the data

The range is the difference between the maximum and minimum values in the dataset.

Range=Maximum value−Minimum value=7−2=5\{Range} = \{Maximum value} - \{Minimum value} = 7 - 2 = 5Range=Maximum value−Minimum value=7−2=5

### Step 2: Choose an appropriate number of bins

A commonly used rule of thumb for choosing the number of bins is Sturges' formula:

Number of bins=1+log⁡2(n)\{Number of bins} = 1 + \log\_2(n)Number of bins=1+log2​(n)

where nnn is the number of data points. For this dataset, n=10n = 10n=10.

Number of bins=1+log⁡2(10)≈1+3.32=4.32≈4 bins\{Number of bins} = 1 + \log\_2(10) \approx 1 + 3.32 = 4.32 \approx 4 \{ bins}Number of bins=1+log2​(10)≈1+3.32=4.32≈4 bins

So, we will use 4 bins.

### Step 3: Determine the bin width

The bin width can be calculated by dividing the range by the number of bins:

Bin width=RangeNumber of bins=54=1.25\{Bin width} = \frac{\{Range}}{\{Number of bins}} = \frac{5}{4} = 1.25Bin width=Number of binsRange​=45​=1.25

Since we want whole numbers for the bin edges, we can round this to 1.5 or 2.

### Step 4: Create bins and classify data

If we use bin width 2, we would have the following bins:

* Bin 1: [2, 4)
* Bin 2: [4, 6)
* Bin 3: [6, 8)

### Step 5: Frequency count in each bin

* Bin 1: [2, 4) contains 2, 2, 3, 3, 3, 4 (6 data points).
* Bin 2: [4, 6) contains 5 (1 data point).
* Bin 3: [6, 8) contains 6, 6, 7 (3 data points).

### Final Histogram

* Bin 1: 6 data points
* Bin 2: 1 data point
* Bin 3: 3 data points



Application Problems

14. Problem:   
A factory measures daily production output (units): 200, 210, 190, 220, 230, 240, 205 .

Find the standard deviation.

To calculate the **standard deviation** of the given dataset, we will follow these steps:

### Given data:

**Daily production output**: 200, 210, 190, 220, 230, 240, 205

### Step 1: Find the **mean** (average)

Mean=200+210+190+220+230+240+2057\{Mean} = \frac{200 + 210 + 190 + 220 + 230 + 240 + 205}{7}Mean=7200+210+190+220+230+240+205​ Mean=15357=219.29 (rounded to 2 decimal places)\{Mean} = \frac{1535}{7} = 219.29 \, (\{rounded to 2 decimal places})Mean=71535​=219.29(rounded to 2 decimal places)

### Step 2: Find the squared differences from the mean for each data point

* (200−219.29)2=(−19.29)2=372.40(200 - 219.29)^2 = (-19.29)^2 = 372.40(200−219.29)2=(−19.29)2=372.40
* (210−219.29)2=(−9.29)2=86.42(210 - 219.29)^2 = (-9.29)^2 = 86.42(210−219.29)2=(−9.29)2=86.42
* (190−219.29)2=(−29.29)2=857.90(190 - 219.29)^2 = (-29.29)^2 = 857.90(190−219.29)2=(−29.29)2=857.90
* (220−219.29)2=(0.71)2=0.50(220 - 219.29)^2 = (0.71)^2 = 0.50(220−219.29)2=(0.71)2=0.50
* (230−219.29)2=(10.71)2=114.67(230 - 219.29)^2 = (10.71)^2 = 114.67(230−219.29)2=(10.71)2=114.67
* (240−219.29)2=(20.71)2=428.67(240 - 219.29)^2 = (20.71)^2 = 428.67(240−219.29)2=(20.71)2=428.67
* (205−219.29)2=(−14.29)2=204.26(205 - 219.29)^2 = (-14.29)^2 = 204.26(205−219.29)2=(−14.29)2=204.26

### Step 3: Find the **variance**

Variance=∑Squared differencesN=372.40+86.42+857.90+0.50+114.67+428.67+204.267\{Variance} = \frac{\sum \{Squared differences}}{N} = \frac{372.40 + 86.42 + 857.90 + 0.50 + 114.67 + 428.67 + 204.26}{7}Variance=N∑Squared differences​=7372.40+86.42+857.90+0.50+114.67+428.67+204.26​ Variance=2064.827=295.97\{Variance} = \frac{2064.82}{7} = 295.97Variance=72064.82​=295.97

### Step 4: Find the **standard deviation**

The standard deviation is the square root of the variance:

Standard deviation=295.97=17.21 (rounded to 2 decimal places)\{Standard deviation} = \sqrt{295.97} = 17.21 \, (\{rounded to 2 decimal places})Standard deviation=295.97​=17.21(rounded to 2 decimal places)

### Conclusion:

The **standard deviation** of the factory's daily production output is **17.21** units.

15. Problem:   
 You are analyzing sales data for two products.

Product A: Mean sales = 100 , Standard deviation = 20 , Standard deviation = 30 Product B: Mean sales = 150

### Given data:

* **Product A**: Mean = 100, Standard Deviation = 20
* **Product B**: Mean = 150, Standard Deviation = 30

### Now, we calculate the **Coefficient of Variation (CV)** for both products:

CV=Standard DeviationMean×100\{CV} = \frac{\{Standard Deviation}}{\{Mean}} \times 100CV=MeanStandard Deviation​×100

### For Product A:

CVA=20100×100=20%\{CV}\_A = \frac{20}{100} \times 100 = 20\%CVA​=10020​×100=20%

### For Product B:

CVB=30150×100=20%\{CV}\_B = \frac{30}{150} \times 100 = 20\%CVB​=15030​×100=20%

### Correct Answer:

Both **Product A** and **Product B** have the same **relative variability**, with a **Coefficient of Variation (CV) of 20%**.

Thus, **the correct answer is that both products have the same relative variability**.

