Calculations for data: Distributions

1. **Center**
   - **Mean** = add up all the numbers and divide by how many values you have
   - **Median** = middle value of ordered data
   - **Mode** = The value that has the highest frequency

2. **Spread**
   - **Range** = Maximum - Minimum
   - **Standard deviation** = the average distance the values are from the mean.
     - **Variance** = \( SD^2 \)

3. **Shape**
   - **Symmetric**
   - **Skewed right**
   - **Skewed left**

4. **Outliers** (sometimes)
   - Extreme values

---

**Example about standard deviation (SD):**

3, 3, 9, 15, 15  
Symmetric data

\[
\text{mean} = \frac{3+3+9+15+15}{5} = 9 = \text{median}
\]

\[SD = 6\]

---

1. 20, 20, 20, 20, 20

\[
\text{mean} = 20
\]

\[SD = 0\]

If all of the values in the list are the same, \( SD = 0 \)

---

3, 3, 9, 15, 15

Add 4: 7, 7, 13, 19, 19

\[
\text{mean} = 13 (\frac{9+4}{2})
\]

\[SD = 6\] remains the same

---

Multiply by 2: 6, 6, 18, 30, 30

\[
\text{mean} = 18 \ (6 \times 2)
\]

\[SD = 12 \ (6 \times 2)\]
### Quiz 1

#### Question 1

True or False:
The amount of rainfall in your state last month is an example of discrete data.

- **a)** True
- **b)** False

#### Question 2

True or False:
The standard deviation is the square of the variance.

- **a)** True
- **b)** False

#### Question 3

Which of the following is not affected by an extreme value in the data set?

- **a)** standard deviation
- **b)** range
- **c)** median
- **d)** mean

#### Question 4

Given a data set of all positive values, if the smallest value of a data set is divided by two, which of the following is true?

- **a)** The interquartile range increases.
- **b)** The standard deviation decreases.
- **c)** The range decreases.
d) The mean decreases.

**Question 5**

If the test scores of a class of 35 students have a mean of 71.1 and the test scores of another class of 28 students have a mean of 67.4, then the mean of the combined group is

\[
\text{mean} = \frac{\text{sum of the total}}{\text{n}}
\]

a) 67.750

\[
\text{class 1 = sum = mean(n)} = 71.1 (35) = 2488.5
\]

b) 69.250

\[
\text{class 2 = sum = mean(n)} = 67.4 (28) = 1887.2
\]

c) 69.456

d) 66.956

\[
\text{mean} = \frac{2488.5 + 1887.2}{35 + 28} = \frac{4375.7}{63} = 69.456
\]

**Question 6**

Given the first type of plot indicated in each pair, which of the second plots could not always be generated from it?

a) dot plot, histogram  
**yes**  
**histogram or box plot for first plot.**

b) stem and leaf, dot plot  
**yes**

c) histogram, stem and leaf  
**no**

d) dot plot, box plot  
**yes**

**Question 7**

A survey was conducted to gather ratings of the quality of service at local restaurants. Respondents rated on a scale of 0 (terrible) to 100 (excellent). The data are represented by the following stem plot.

```
3 | 2 4
4 | 0 3 4 7 8 9 9 9
5 | 0 1 2 3 4 5
6 | 1 2 5 6 6
7 | 0 1
8 |
9 | 2
```

The median response was

```
32, 34, 40, 43, 44, 47, 48, 49, 49, 49, 50, 51, 51
52, 53, 54, 55, 61, 62, 65, 66, 68, 70, 71, 92
```

a) 50

b) **51**

c) 51.5

d) 52
mean = \frac{-5 + (-1) + (-1) + 3 + 9}{5} = \frac{0}{5} = 1

median = \text{middle value of ordered data} = -1

mode = \text{value with the highest frequency} = -1

range = \text{highest value} - \text{lowest value} = 9 - (-5) = 14

Standard deviation:

<table>
<thead>
<tr>
<th>X</th>
<th>X - mean</th>
<th>(X - mean)^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5</td>
<td>-5 - 1 = -6</td>
<td>(-4)^2 = 36</td>
</tr>
<tr>
<td>-1</td>
<td>-1 - 1 = -2</td>
<td>(-2)^2 = 4</td>
</tr>
<tr>
<td>-1</td>
<td>-1 - 1 = -2</td>
<td>(0)^2 = 0</td>
</tr>
<tr>
<td>3</td>
<td>3 - 1 = 2</td>
<td>(2)^2 = 4</td>
</tr>
<tr>
<td>9</td>
<td>9 - 1 = 8</td>
<td>(8)^2 = 64</td>
</tr>
</tbody>
</table>

\[ \text{sum} = 112 \]

\[ \text{variance} = \frac{\text{sum of} \ (X - \text{mean})^2}{n - 1} = \frac{112}{4} = 28 = \sigma^2 \]

\[ \sigma = \sqrt{\text{variance}} = \sqrt{28} = 5.29 \]

using R-studio

> x = c(-5,-1,-1,3,9)
> mean(x)
[1] 1
> median(x)
[1] -1
> sd(x)
[1] 5.291503
Question 8

Calculate the mean, median, mode, range and standard deviation of the data: -5, -1, -1, 3, 9

a) mean = 1.0, median = -1, mode = -1, range = 14, standard deviation = 5.3

b) mean = 1.8, median = -5, mode = -1, range = 13, standard deviation = 5.2

c) mean = 1.0, median = 3, mode = -5, range = 15, standard deviation = 5.3

d) mean = 1.8, median = -1, mode = -1, range = 14, standard deviation = 5.2

e) None of the above

Question 9

Calculate the mean, median, mode, range and standard deviation of the data: -72, -40, -40, 56, 88

a) mean = -1.6, median = 56, mode = -72, range = 161, standard deviation = 69.4

b) mean = -1.6, median = -40, mode = -40, range = 160, standard deviation = 69.4

c) mean = 17.6, median = -40, mode = -40, range = 160, standard deviation = 69.4

d) mean = 17.6, median = -72, mode = -40, range = 159, standard deviation = 69.4

e) None of the above

Question 10

The boxplots shown below summarize two data sets, I and II. Based on the boxplots, which of the following statements about these two data sets CANNOT be justified?

a) The interquartile range of data set I is equal to the interquartile range of data set II. Yes

b) Data set I and data set II have the same number of data points.

c) The range of data set I is greater than the range of data set II. Yes

d) The median of data set I is equal to the median of data set II. Yes
Question 11

The distribution that has the box plot shown could be described as

- a) skewed left
- b) skewed right
- c) symmetrical
- d) inconclusive

[Box plot image]

Question 12

The figure below shows a cumulative relative frequency plot of 40 scores on a test given in a Statistics class. Which of the following conclusions can be made from the graph?

- a) The horizontal nature of the graph for test scores of 60 and below indicates that those scores occurred most frequently.
- b) 10% of the class scored 50 or less.
- c) 25% of the class scored 70 or less.
- d) Q1 = 70.
- e) 50% of the class scored 85 or less.
- f) Median = 85.
- g) 60 is at 15%.

[Graph image]

60 is at 15% False
b) There is greater variability in the lower 20 test scores than in the higher 20 test scores.

c) The median test score is less than 70.  

\[ \text{median} \approx 85 \]

\[ 80 \text{ is at the } 40\text{th percentile} \Rightarrow 40\% \text{ scored } 80 \text{ or less} \]

\[ 60\% \text{ scored above } 80 \]

**Question 13**

The weights of male and female students in a class are summarized in the following boxplots:

Which of the following is NOT correct?

a) About 50% of the male students have weights between 150 and 185 lbs.  

b) About 75% of the female students have weights more than 128 lbs.  

\[ \text{median} = \text{mean} \]

c) The mean weight of the female students is about 120 because of symmetry.

d) The median weight of the male students is about 166 lbs.

**Question 14**

Given a data set consisting of 33 unique whole number observations, its five-number summary is:

\[ [14, 25, 36, 47, 59] \]

What is the IQR?

a) 45  

\[ \text{IQR} = Q_3 - Q_1 = 47 - 25 = 22 \]

b) 22  

How many values are strictly less than 25?

\[ Q_1 = 25 \Rightarrow 25\% \text{ are less than } 25. \]

\[ 0.25(33) = 8.25 \text{ are less than } 25. \]

\[ Q_3 = 47 \Rightarrow 75\% \text{ of } 33 \text{ are less than } 47 \]

\[ 0.75(33) = 24.75 \text{ are less than } 47. \]