

Descriptive Statistics

Part -1

Sourav Karmakar

souravkarmakar29@gmail.com

Data and Variables

- **Data** are pieces of information about individuals organized into variables.
- By an **individual** (also called **record**), we mean a particular person or object.
- By a **variable**, we mean a particular characteristic of the individual.

The following dataset shows medical records from a particular survey:

Individuals	Variables						
	Gender (M/F)	Age	Weight (lbs.)	Height (in.)	Smoking (1=No, 2=Yes)	Race	
	Patient #1	M	59	175	69	1	White
	Patient #2	F	67	140	62	2	Black
	Patient #3	F	73	155	59	1	Asian

Patient #75	M	48	90	72	1	White	

Usually variables are arranged across columns while the individuals (records) are arranged across rows

Data and Variables

Variables can be classified into one of two types: categorical or quantitative.

- **Categorical variables** take category or label values and place an individual into one of several groups. Each observation can be placed in *only* one category, and the categories are mutually exclusive.
- **Quantitative variables** take numerical values and represent measurement.

Individuals	Variables						
	Gender (M/F)	Age	Weight (lbs.)	Height (in.)	Smoking (1=No, 2=Yes)	Race	
	Patient #1	M	59	175	69	1	White
	Patient #2	F	67	140	62	2	Black
	Patient #3	F	73	155	59	1	Asian

Patient #75	M	48	90	72	1	White	

In our example:

- *Gender* and *Smoking* are categorical variables.
- *Age*, *Weight* and *Height* are quantitative variables.

Data and Variables

We took a random sample from the 2000 U.S. Census. Here is part of the dataset:

Census is an official count or survey, especially of a population.

US Census 2000

	State	zipcode	Family_Size	Annual_income
1	Florida	32716	8	200
2	Alabama	35236	5	800
3	Florida	32116	6	13500
4	Florida	33679	5	21000
5	Alabama	36374	4	21000
6	California	94565	1	23000

Q.1. Who are the individuals described by this data?

- States
- ✓ People living in the United States in the year 2000
- People with families in the year 2000

The U.S. Census is completed by people living in the United States.

Q.2. What type of variable is Zipcode?

- ✓ Categorical
- Quantitative

Zipcode is a categorical variable because it categorizes individuals by geographic location

Q.3. What type of variable is Annual Income?

- Categorical
- ✓ Quantitative

Annual Income is a Quantitative variable because it assumes continuous values over a range and has arithmetic significance.

Categorical Variable

MOVIE

ID	NAME	DIRECTOR	TIME	RATING
1				4*
2				3*
3				5*
4				2*

STUDENTS

ID	NAME	GENDER (F/M)	SMOKING (Y/N)	FAV SUB
			3 ← Y → 0 2 ← N → 1	HISTORY ARTS MATH

What type of variable is rating?

✓ RATING is a categorical variable

What type of variable is Smoking?

✓ SMOKING is a categorical variable.

Rating



1* 3* 5*
—————> order.

ORDINAL VARIABLE

10	6 S
	4 NS

S → 1
NS → 0

(0.6) X

NOMINAL VARIABLE

Quantitative (continuous):-

Temp ($^{\circ}\text{C}$)

Relative Humidity (%)

Ratio .

Daily Income
✓ 0

Temp
0

Distribution of Variables

- We will begin our journey of Descriptive Statistics by exploring (or looking at) one variable at a time.
- As we saw in the previous slides, usually a variable in a data consist of long list of values (whether numerical or not) and are not very informative in that form.
- In order to convert these raw data into useful information we need to summarize and then examine the **distribution** of the variable. By **distribution** of a variable, we mean:
 - what values the variable takes, and
 - how often the variable takes those values.
- We will first learn how to summarize and examine the distribution of a single categorical variable, and then do the same for a quantitative variable.

Distribution of Variables

Examining Distribution of Categorical Variable

- Suppose we have taken a survey among the office-goers on the mode of daily commutes.
- Few office-going people were randomly chosen and they were asked how they travel from home to office on a regular basis. The dataset we obtain looks something like below.

Persons	Types of Daily Commute
1	Public Transport
2	Public Transport
3	Two - Wheeler
.	.
.	.
.	.
1198	Four - Wheeler
1199	Public Transport
1200	Two - Wheeler

Following are the findings from the given dataset

- There are three mode of transports that people avail.
 - Public Transport
 - Own Two-Wheeler vehicle
 - Own Four-Wheeler vehicle
- There are total 1200 rows or individuals in the dataset

Distribution of Categorical Variable

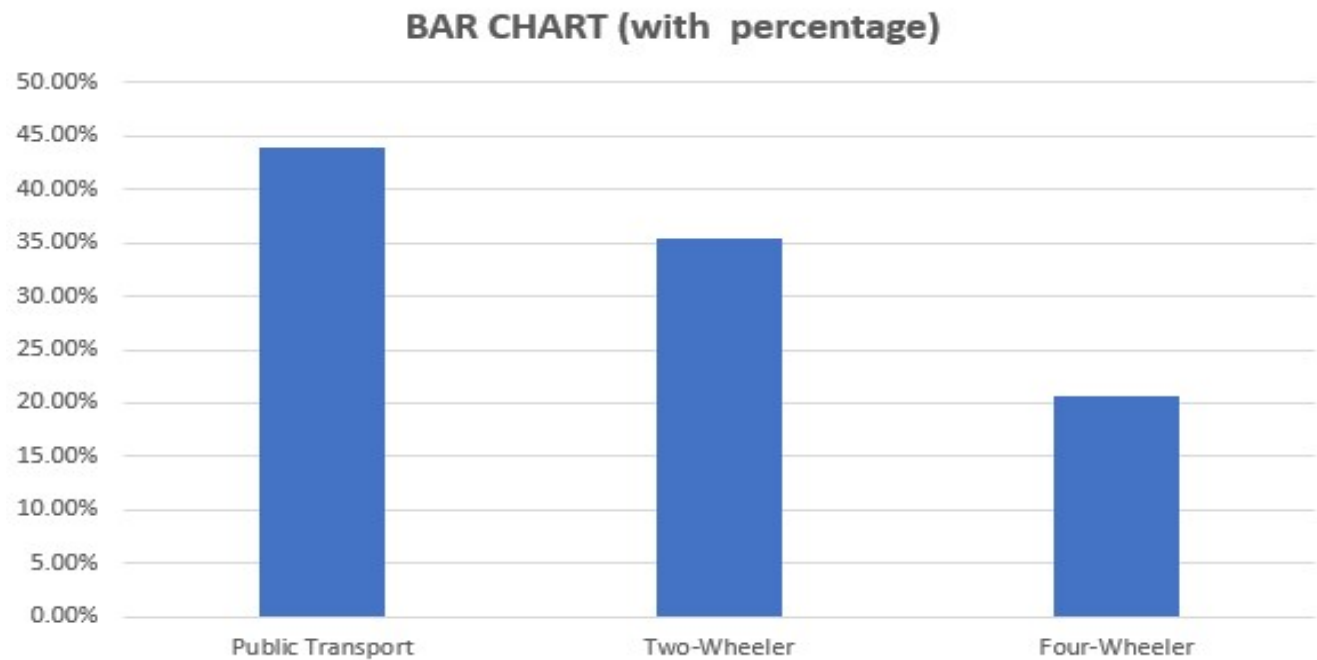
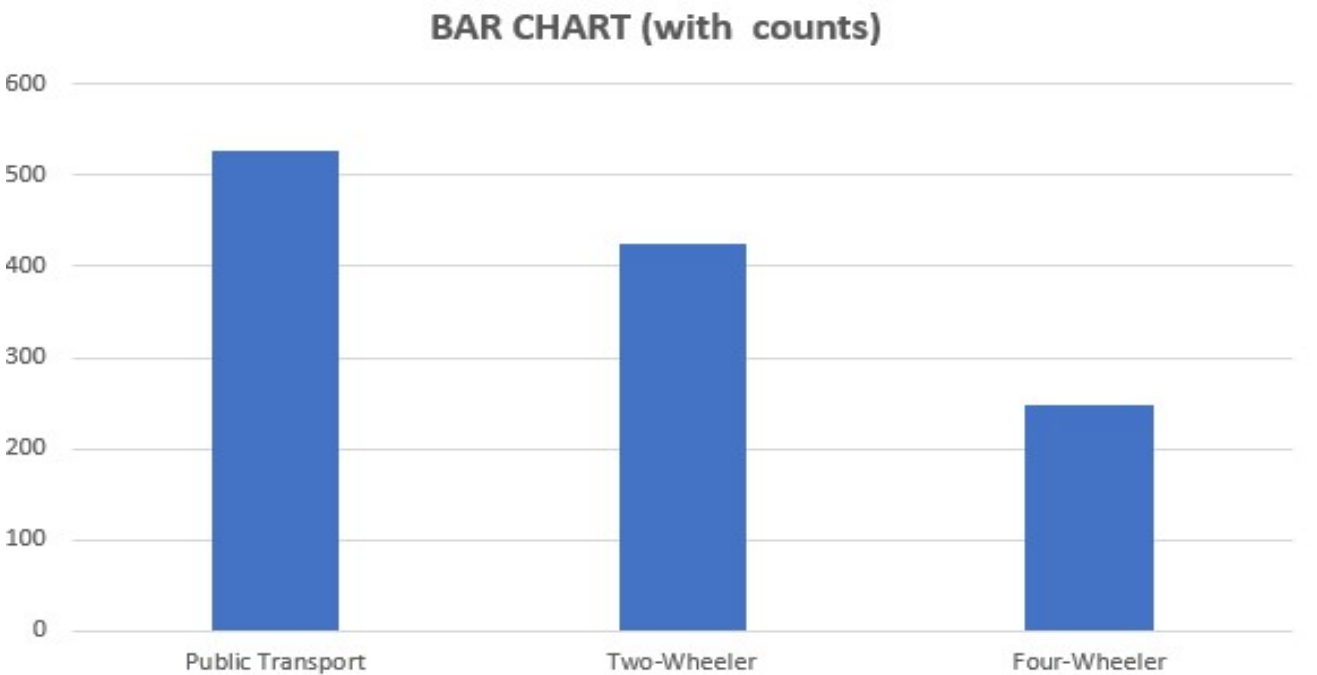
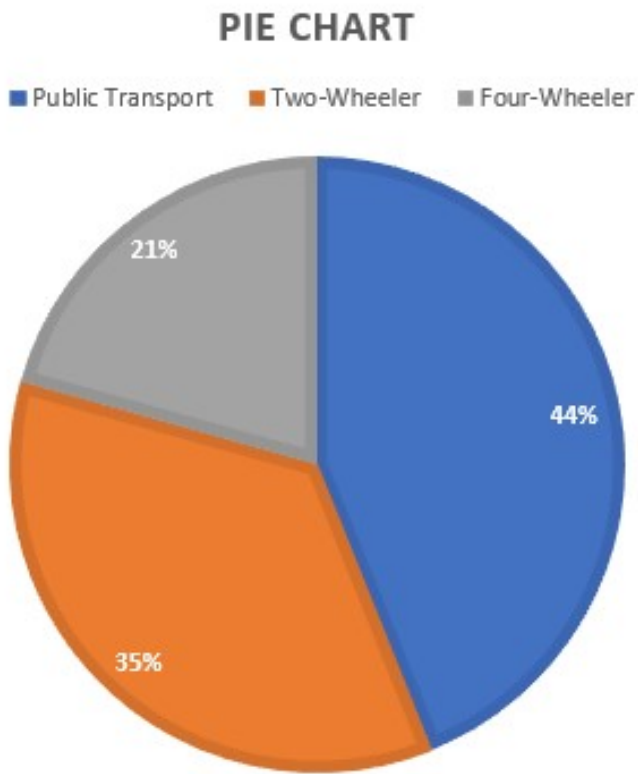
- Suppose we ask: *What percentage of sampled office-goers fall into each category?*
- This question will be easily answered once we summarize and look at the **distribution** of the variable *“Types of Daily Commute”*.
- In order to summarize the distribution of a categorical variable, we first create a table of the different values (**categories**) the variable takes, how many times each value occurs (**count**) and, more importantly, how often each value occurs (by converting the counts to **percentages**); this table is called a frequency distribution. Here is the frequency distribution for our example:

Categories	Count	Percentage
Public Transport	527	43.92%
Two-Wheeler	425	35.42%
Four-Wheeler	248	20.67%
Total	1200	100.00%

Distribution of Categorical Variable

We can express the result pictorially with the help of Bar charts or Pie Charts

Categories	Count	Percentage
Public Transport	527	43.92%
Two-Wheeler	425	35.42%
Four-Wheeler	248	20.67%
Total	1200	100.00%



Distribution of Variables

Examining Distributions of Quantitative Variable

Histogram: Intervals

Break the range of values into intervals and count how many observations fall into each interval.

Examples: Here are the exam grades of 15 students-

88, 48, 60, 51, 57, 85, 69, 75, 97, 72, 71, 79, 65, 63, 73

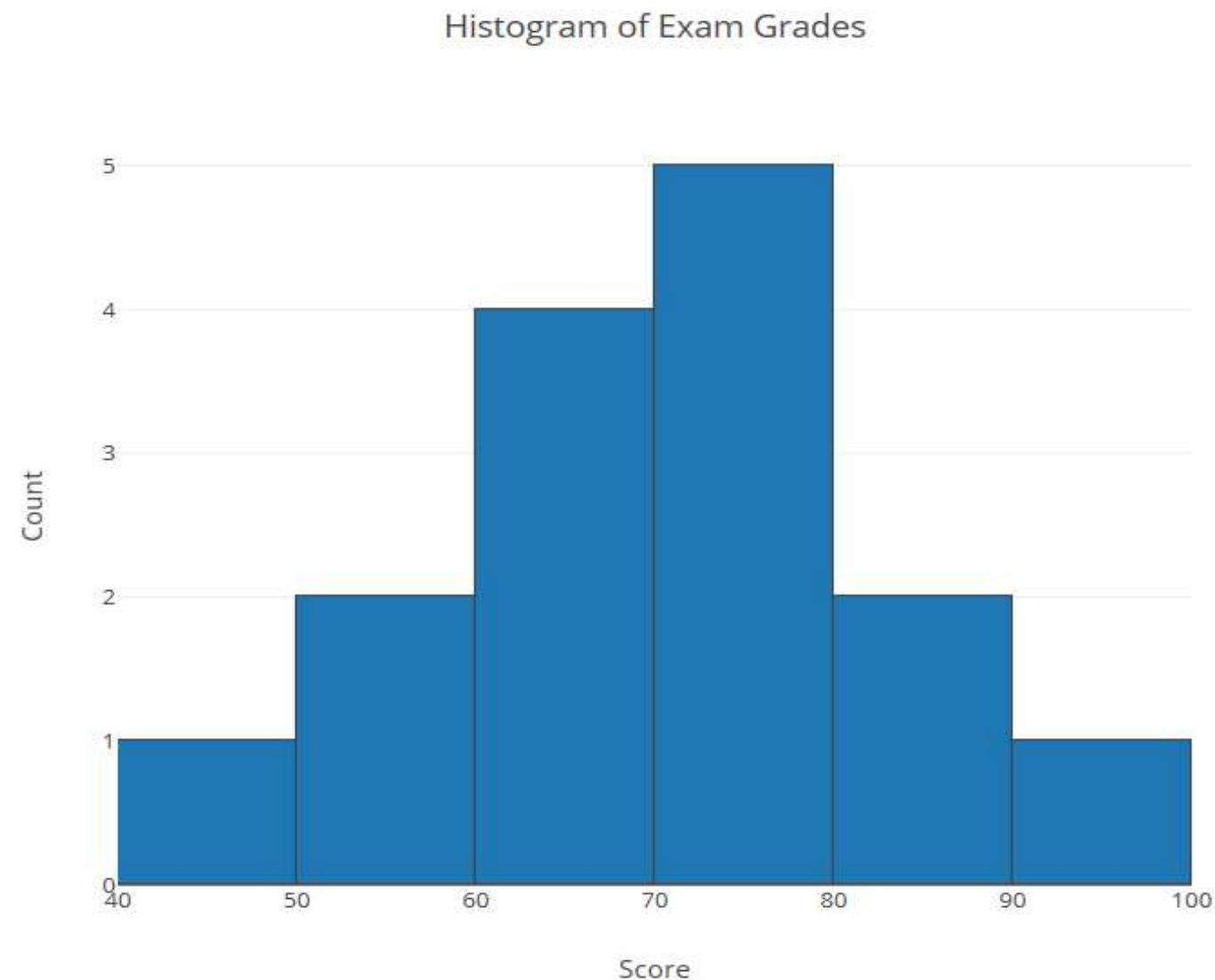
We first need to break the range of values into intervals (also called "bins" or "classes"). By counting how many of the 15 observations fall in each of the intervals, we get the table as shown:

Score Intervals	Counts
[40-50)	1
[50-60)	2
[60-70)	4
[70-80)	5
[80-90)	2
[90-100]	1

Distribution of Quantitative Variable

Score Intervals	Counts
[40-50)	1
[50-60)	2
[60-70)	4
[70-80)	5
[80-90)	2
[90-100]	1

To construct the histogram from this table we plot the intervals on the X-axis and show the number of observations in each interval (frequency of the interval) on the Y-axis, which is represented by the height of a rectangle located above the interval.



Distribution of Quantitative Variable

Measure of Central Tendency:

The three main numerical measures for the center of a distribution are the **mode**, the **mean** and the **median**.

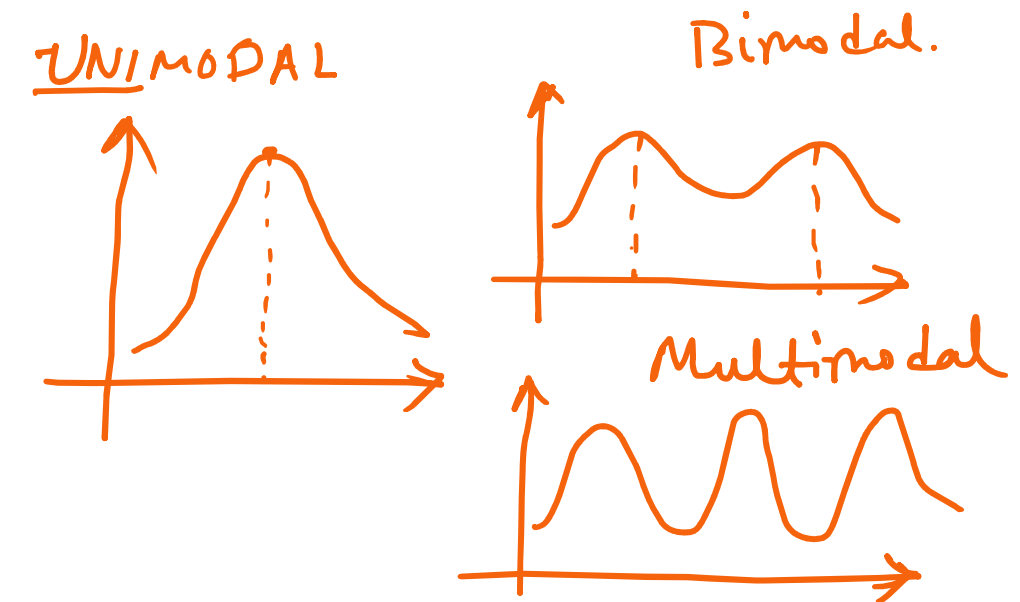
Mode:

The mode is the most commonly occurring value in a distribution.

Mean:

The mean is the average of a set of observations (i.e., the sum of the observations divided by the number of observations). If the n observations are $x_1, x_2, x_3, \dots, x_n$ then their mean or average which we call \bar{x} is calculated as:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$



Distribution of Quantitative Variable

Median:

The median is the midpoint of the distribution. It is the number such that half of the observations fall above, and half fall below.

Calculation of Median:

- Order the data from smallest to largest.
- Consider whether n , the number of observations, is even or odd.
 - If n is **odd**, the median M is the center observation in the ordered list. This observation is the one "sitting" in the $\frac{(n+1)}{2}$ **spot** in the ordered list.
 - If n is **even**, the median M is the **mean** of the **two center observations** in the ordered list. These two observations are the ones "sitting" in the $\frac{n}{2}$ and $(\frac{n}{2} + 1)$ spots in the ordered list.

Distribution of Quantitative Variable

Calculation of Median (Example):

Consider the following numbers:

12, 18, 11, 4, 14, 18, 11, 5, 12, 7, 3, 12, 23, 9, 16, 1, 5, 17, 3, 13, 21

- The ordered numbers from small to large is:

1, 3, 3, 4, 5, 5, 7, 9, 11, 11, 12, 12, 13, 14, 16, 17, 18, 18, 21, 23

- Here $n = 21$ (odd). Therefore, the $\frac{(n+1)}{2}$ *th* data i.e. 11th data here will be the median.
Hence, median = 12.

Distribution of Quantitative Variable

Measure of Dispersion / Spread:

These measures provide different ways to quantify the variability of the distribution. We will discuss the following three most commonly used measures of spread.

- Range
- Variance & Standard Deviation and
- Inter-Quartile Range (IQR)

Range:

Range is exactly the distance between the smallest data point (min) and the largest one (Max). i.e.,
 $\text{Range} = \text{Max} - \text{Min}$

Example-1: 12, 18, 11, 4, 14, 18, 11, 5, 12, 7, 3, 12, 23, 9, 16, 1, 5, 17, 3,

Here Max = 23 and Min = 1.

Hence, Range = Max – Min = 23 - 1 = 22

Distribution of Quantitative Variable

Variance & Standard Deviation:

If the n observations are x_1, x_2, \dots, x_n their mean or average is $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. Then the variance and the standard deviation is calculated as:

$$\text{Variance (Var)} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{and Standard Deviation } (\sigma) = \sqrt{\text{Var}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Distribution of Quantitative Variable

Example of calculation of Variance and Standard Deviation:

Consider the following example Dataset:

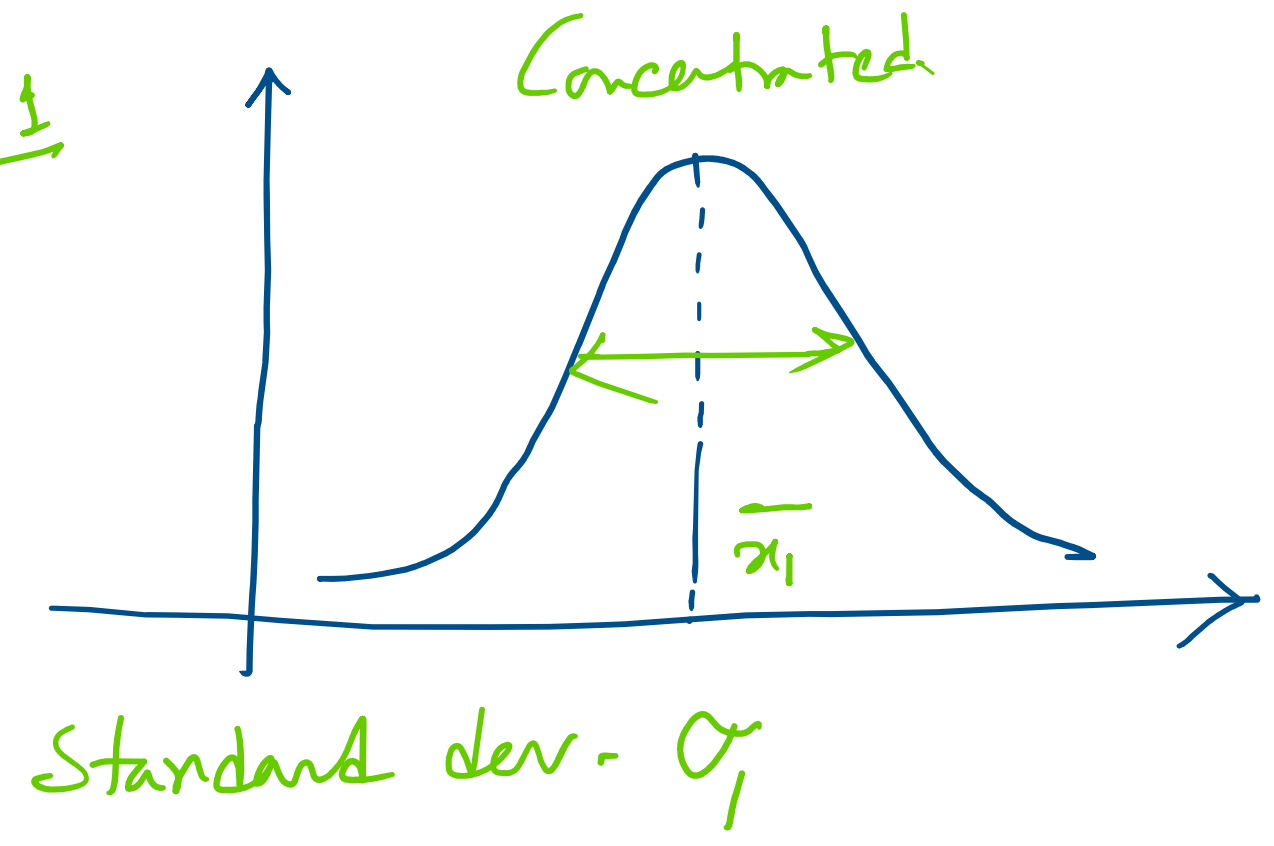
Sl. No.	Values
1	12
2	5
3	10
4	8
5	5
6	11
7	7
8	9
9	15
10	6

- From the dataset we can observe that number of observations = 10
- Average / Mean = $\frac{12+5+10+8+5+11+7+9+15+6}{10} = 8.8$
- Build the table as shown next.

Sl. No.	Values	Values - Mean	(Values - Mean) ²
1	12	3.2	10.24
2	5	-3.8	14.44
3	10	1.2	1.44
4	8	-0.8	0.64
5	5	-3.8	14.44
6	11	2.2	4.84
7	7	-1.8	3.24
8	9	0.2	0.04
9	15	6.2	38.44
10	6	-2.8	7.84

- From the last table, $Variance (Var) = \frac{1}{10} \sum (Values - Mean)^2 = \frac{95.6}{10} = 9.56$
- Hence, $Standard\ Deviation (\sigma) = \sqrt{9.56} = 3.092$

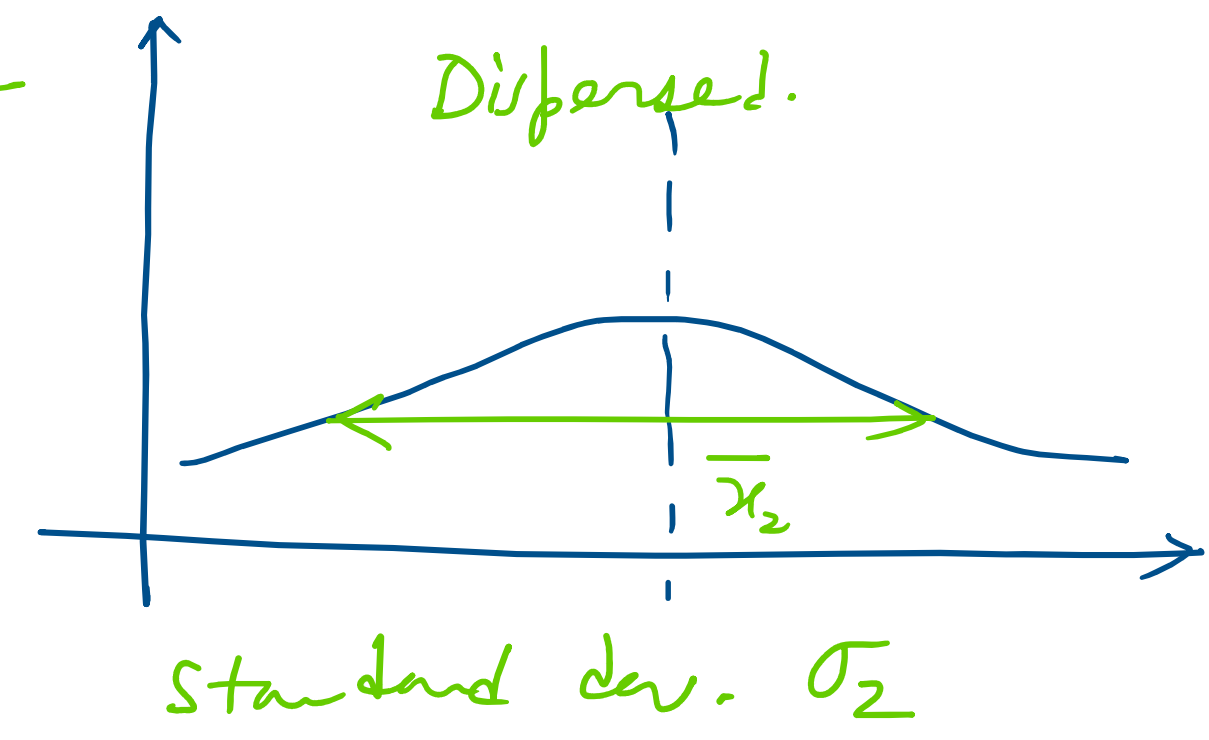
Dist-1



COV₁ $\left(\frac{\sigma_1}{\bar{x}_1} \right)$

COV₁ < COV₂

Dist-2



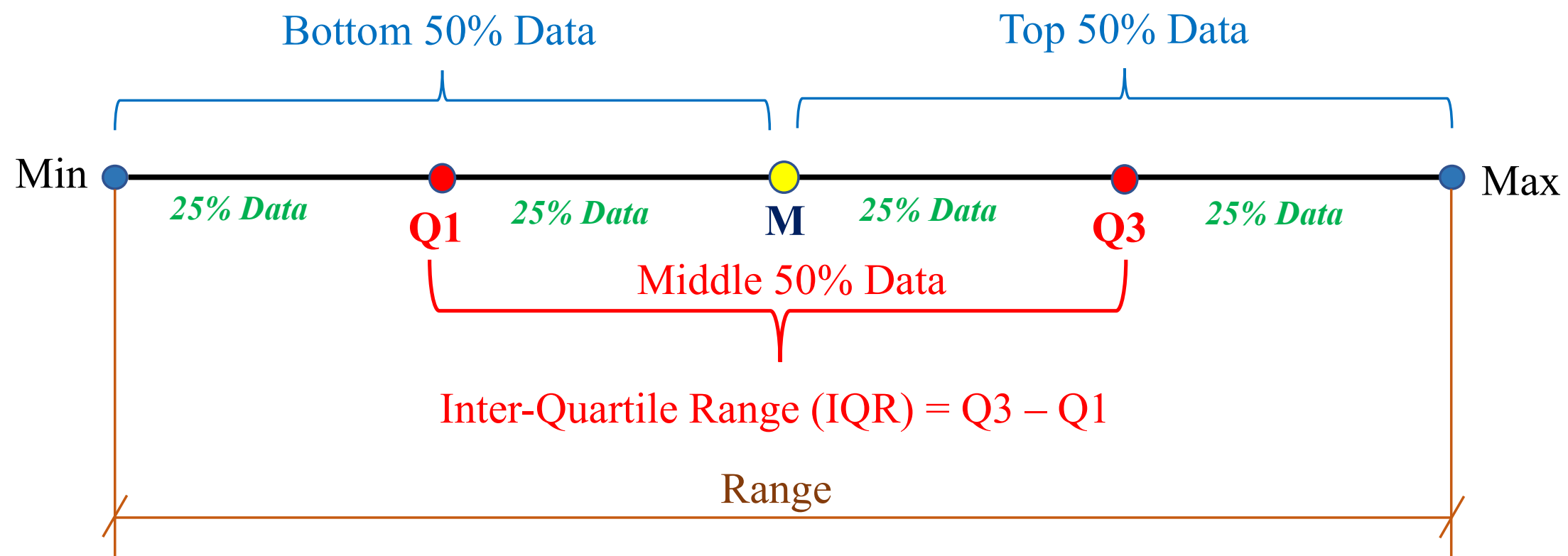
COV₂ $\left(\frac{\sigma_2}{\bar{x}_2} \right)$

Standard deviation
mean = COV (Coefficient of Variation)

Distribution of Quantitative Variable

Inter Quartile Range (IQR):

The IQR measures the variability of a distribution by giving us the range covered by the *MIDDLE* 50% of the data. The following figure illustrates the idea.



M: Median of the data

Q1: First Quartile of the data (one quarter of the data points fall below it)

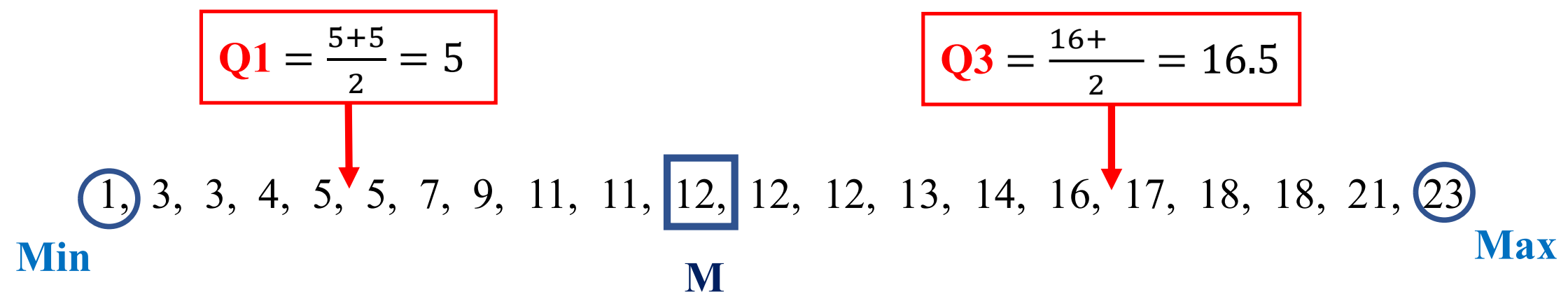
Q3: Third Quartile of the data (three quarters of the data points fall below it)

Distribution of Quantitative Variable

Example of Calculating Inter Quartile Range (IQR):

Consider the observations: 12, 18, 11, 4, 14, 18, 11, 5, 12, 7, 3, 12, 23, 9, 16, 1, 5, 17, 3, 13, 21

- Sort the observations in ascending order:



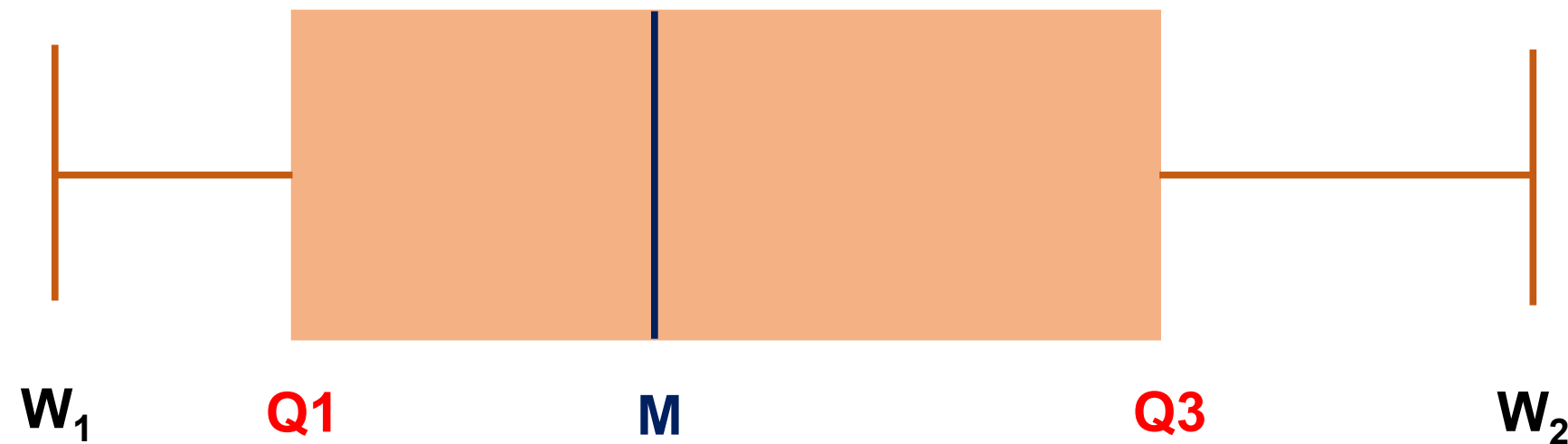
- Find the Median (M), followed by First Quartile (Q1) and Third Quartile (Q3)
- Then calculate Inter Quartile Range (IQR) = $Q3 - Q1 = 16.5 - 5 = 11.5$

The combination of all five numbers (min, Q1, M, Q3, Max) is called the **five number summary**, and provides a quick numerical description of both the center and spread of a distribution.

Distribution of Quantitative Variable

Box and Whisker Plot:

This is the way to visualize the distribution of quantitative variable using **Five Number Summary**.



Five Number Summary

- **M**: Median
- **Q1**: First Quartile
- **Q3**: Third Quartile
- **W₁**: Lower Whisker
- **W₂**: Upper Whisker

There are several methods of drawing whiskers. We'll use the $1.5 \times IQR$ criterion, also known as the **Tukey method** for plotting whiskers. If *Max* and *Min* are the maximum and minimum values in the dataset respectively, then:

- **If:** $Min \geq Q1 - 1.5 \times IQR$, **Then:** $W_1 = Min$
Else: W_1 is at minimum value that is $\geq Q1 - 1.5 \times IQR$
- **If:** $Max \leq Q3 + 1.5 \times IQR$, **Then:** $W_2 = Max$
Else: W_2 is at maximum value that is $\leq Q3 + 1.5 \times IQR$

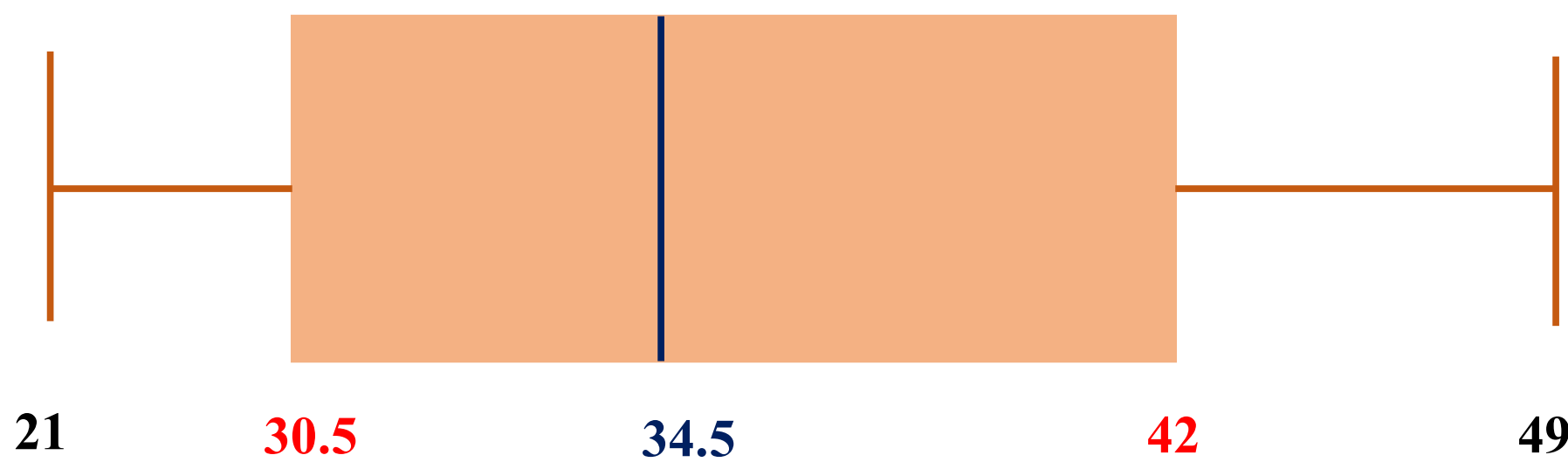
Distribution of Quantitative Variable

Example of Box and Whisker Plot:

Following are the ages of the actresses who won academy award (Oscar) from the year 1970 to 2013:

34, 34, 27, 37, 42, 41, 36, 32, 41, 33, 31, 74, 33, 49, 38, 61, 21, 41, 26, 80, 42, 29, 33, 36, 45, 49, 39, 34, 26, 25, 33, 35, 35, 28, 30, 29, 61, 32, 33, 45, 29, 62, 22, 44

On examining the data we find the following 5 number summary

- Min = 21
 - Q1 = 30.5
 - M = 34.5
 - Q3 = 42
 - Max = 80
- 
- A box and whisker plot representing the distribution of actress ages. The plot features a central orange box with a dark blue vertical line indicating the median at 34.5. The left whisker extends to the minimum value of 21, and the right whisker extends to the maximum value of 49. The box itself spans from the first quartile (Q1) at 30.5 to the third quartile (Q3) at 42. The values 21, 30.5, 34.5, 42, and 49 are labeled below the plot line in black, red, and blue colors respectively.
- $IQR = Q3 - Q1 = 42 - 30.5 = 11.5$
 - $Q1 - 1.5 \times IQR = 30.5 - 1.5 \times 11.5 = 13.25$ and $Min > 13.25$. Hence, W_1 is at *Min* (i.e. at 21)
 - $Q3 + 1.5 \times IQR = 42 + 1.5 \times 11.5 = 59.25$ and $Max > 59.25$. Hence, W_2 is at 49 (which the maximum value that is less than $Q3 + 1.5 \times IQR$)

Thank You