

## MODULE 3-Descriptive Statistics

**Descriptive statistics** are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population.

Descriptive statistics are broken down into

- **Measures of central tendency**
- **Measures of variability**

Measures of central tendency include the **mean, median, mode, percentile and quartile**

Measures of variability include **Range, Inter quartile range, Quartile deviation, Mean deviation, Variance, Standard deviation & Coefficient of variation**

### Central Tendency

Measures of central tendency focus on the average or middle values of data sets, whereas measures of variability focus on the dispersion of data. These two measures use graphs, tables and general discussions to help people understand the meaning of the analyzed data.

**Measures of central tendency** describe the center position of a distribution for a data set. A person analyzes the frequency of each data point in the distribution and describes it using the mean, median, or mode, which measures the most common patterns of the analyzed data set.

**Measures of variability** (or the measures of spread) aid in analyzing how dispersed the distribution is for a set of data. For example, while the measures of central tendency may give a person the average of a data set, it does not describe how the data is distributed within the set.

So, while the average of the data maybe 65 out of 100, there can still be data points at both 1 and 100.


**Measures of variability** help communicate this by describing the shape and spread of the data set. Range, quartiles, absolute deviation, and variance are all examples of measures of variability.

Consider the following data set: 5, 19, 24, 62, 91, 100. The range of that data set is 95, which is calculated by subtracting the lowest number (5) in the data set from the highest (100).



Arithmetic Mean

Arithmetic mean is the sum of all observations divided by number of observations.

formula = Sum of Observation ÷ Total numbers of Observations



$$\text{Arithmetic Mean Formula} = \frac{X_1 + X_2 + X_3 + \dots + X_n}{n}$$



### Arithmetic mean formula for frequency distribution

The formula for the arithmetic mean for frequency distribution can be derived by summing up the products of each variable and its frequency and then the result is divided by the sum of the frequencies.



#### Arithmetic Mean Formula

$$\text{Arithmetic Mean} = \sum \frac{X_i}{n}$$



$$\text{Arithmetic Mean} = \sum \frac{(f_i * X_i)}{f_i}$$



### Median

The median is simply the middle value of a data set. In order to calculate the median, all values in the data set need to be ordered, from either highest to lowest, or vice versa. If there are an odd number of values in a data set, then the median is easy to calculate.

If there is an even number of values in a data set, then the calculation becomes more difficult. Statisticians still debate how to properly calculate a median when there is an even number of values, but for most purposes, it is appropriate to simply take the mean of the two middle values.

Median in case of Ungrouped Data	
In this case we first arrange the observations in <b>increasing or decreasing</b> order then we use the following formulae for Median:	
If “n” is odd	$Median = \text{size of } \left(\frac{n+1}{2}\right) \text{th observation}$
If “n” is even	$Median = \frac{\text{size of } \left\{\left(\frac{n}{2}\right) \text{th} + \left(\frac{n}{2} + 1\right) \text{th}\right\} \text{ observation}}{2}$

### Median of Grouped Data

$$\text{Median of Grouped Data} = L + [(N/2 - C) / F]W$$

- **L**: Lower limit of median class
- **W**: Width of median class
- **N**: Total Frequency
- **C**: Cumulative frequency up to median class
- **F**: Frequency of median class

Weekly expenditure (\$)	Number of families (f)	Cumulative frequency (cf)
0 - 1000	28	28
1000 - 2000	46	74
2000 - 3000	54	128
3000 - 4000	42	170
4000 - 5000	30	200
	N = 200	

Median class =  $(N/2)^{\text{th}}$  value

=  $(200/2)^{\text{th}}$  value

= 100<sup>th</sup> value

Median class = 2000 - 3000

$l = 2000, N/2 = 100, C = 74, F = 54$  and  $W = 1000$

Median =  $2000 + ([100 - 74]/54) \times 1000$

=  $2000 + (26/54) \times 1000$

=  $2000 + 481.5$

= 2481.5

## THE MODE

The mode is the most commonly occurring number in the data set. The mode is best used when you want to indicate the most common response or item in a data set.

The formula that statisticians and analysts use in the calculation of a data set in statistics:

$$\text{Mode or } Z = L_1 + \frac{D_1}{D_1 + D_2} \times i$$

Where the modal class = the one with the highest frequency data interval.

$L_1$  = lower limit of the said modal class

$i$  = size of the class interval

$D_1 = f_m - f_1$   $D_2 = f_m - f_2$

$f_m$  = modal class frequency

$f_1$  = frequency of the class which precedes the modal class; and

$f_2$  = frequency of the class which succeeds the modal class

**Example 1:** 1, 2, 3, 4, 10, 10, 10

Here the mode is 10.

**Example 2:**

Class Interval	0–10	10–20	20–30	30–40	40–50
Frequency	8	5	10	4	7

Modal class = 20-30 as it has the data with the highest frequency

Size of the class interval,  $h = 10$

The lower limit of the above modal class,  $L = 20$

Frequency of the modal class,  $f_m = 10$

Frequency of the class which precedes the modal class,  $f_1 = 5$ ; and

Frequency of the class which succeeds the modal class  $f_2 = 4$

Therefore, putting all the values in the formula of mode,  $M = 20 + 10 \frac{(10-5)}{(10-5) + (10-4)}$

Mode = 24.54

## Percentile

**Percentiles** should not be confused with percentages. The percentages is used to express fractions of a whole, while percentiles are the values below which a certain percentage of the data in a data set is found.

In other words, the percentage score reflects how well the student did on the exam itself; the percentile score reflects how well he did in comparison to other students.

$$\text{Percentile} = \frac{\text{Number of Values Below "x"}}{\text{Total Number of Values}} \times 100$$

**Example 1:** The scores obtained by 10 students are 38, 47, 49, 58, 60, 65, 70, 79, 80, 92. Using the percentile formula, calculate the percentile for score 70?

**Solution:**

**Given:**

Scores obtained by students are 38, 47, 49, 58, 60, 65, 70, 79, 80, 92

Number of scores below 70 = 6

Using percentile formula,

Percentile = (Number of Values Below "x" / Total Number of Values)  $\times 100$

Percentile of 70

=  $(6/10) \times 100$

$$= 0.6 \times 100 = 60$$

**Therefore, the percentile for score 70 = 60**

**Example 2:** The weights of 10 people were recorded in kg as 35, 41, 42, 56, 58, 62, 70, 71, 90, 77. Find the percentile for the weight 58 kg.

**Solution:**

Given:

Weight of the people are 35, 41, 42, 56, 58, 62, 70, 71, 77, 90

Number of people with weight below 58 kg = 4

Using percentile formula,

Percentile = (Number of Values Below "x" / Total Number of Values)  $\times$  100

Percentile for weight 58 kg

$$= (4/10) \times 100$$

$$= 0.4 \times 100 = 40$$

Therefore, the percentile for weight 58 kg = 40

### CALCULATING THE $p$ TH PERCENTILE

**Step 1.** Arrange the data in ascending order (smallest value to largest value).

**Step 2.** Compute an index  $i$

$$i = \left( \frac{p}{100} \right) n$$

where  $p$  is the percentile of interest and  $n$  is the number of observations.

**Step 3.** (a) If  $i$  is not an integer, round up. The next integer greater than  $i$  denotes the position of the  $p$ th percentile.

(b) If  $i$  is an integer, the  $p$ th percentile is the average of the values in positions  $i$  and  $i + 1$ .

### Calculate 85<sup>th</sup> Percentile

**Step 1.** Arrange the data in ascending order.

**3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925**

**Step 2.**  $i = (p / 100)n = (85 / 100)12 = 10.2$

**Step 3.** Because  $i$  is not an integer, round up. The position of the 85th percentile is the next integer greater than 10.2, the 11th position.

Returning to the data, we see that the 85th percentile is the data value in the 11th position, or 3730

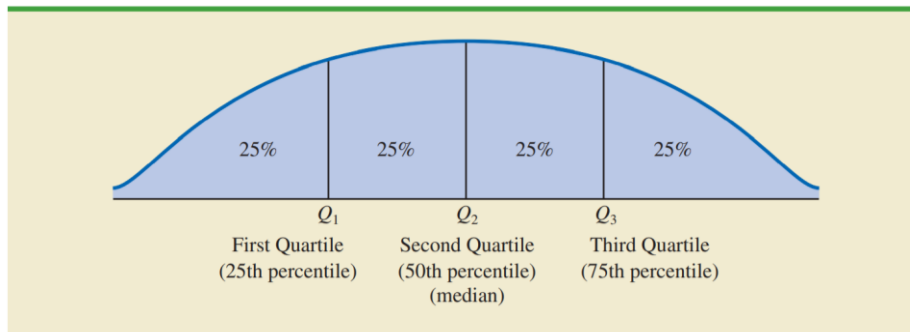
## Quartile

It is often desirable to divide data into four parts, with each part containing approximately one-fourth, or 25% of the observations. Figure 3.1 shows a data distribution divided into four parts. The division points are referred to as the quartiles and are defined as

Q1 first quartile, or 25th percentile

Q2 second quartile, or 50th percentile (also the median)

Q3 third quartile, or 75th percentile.



**Example: 3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925**

The starting salary data are again arranged in ascending order. We already identified  $Q_2$ , the second quartile (median), as 3505.

3310 3355 3450 3480 3480 3490 3520 | 3540 3550 3650 3730 3925

The computations of quartiles  $Q_1$  and  $Q_3$  require the use of the rule for finding the 25th and 75th percentiles. These calculations follow.

For  $Q_1$ ,

$$i = \left( \frac{p}{100} \right) n = \left( \frac{25}{100} \right) 12 = 3$$

Because  $i$  is an integer, step 3(b) indicates that the first quartile, or 25th percentile, is the average of the third and fourth data values; thus,  $Q_1 = (3450 + 3480)/2 = 3465$ .

For  $Q_3$ ,

$$i = \left( \frac{p}{100} \right) n = \left( \frac{75}{100} \right) 12 = 9$$

Again, because  $i$  is an integer, step 3(b) indicates that the third quartile, or 75th percentile, is the average of the ninth and tenth data values; thus,  $Q_3 = (3550 + 3650)/2 = 3600$ .

The quartiles divide the starting salary data into four parts, with each part containing 25% of the observations.

3310	3355	3450		3480	3480	3490		3520	3540	3550		3650	3730	3925
				$Q_1 = 3465$				$Q_2 = 3505$				$Q_3 = 3600$		
								(Median)						

### Quartiles for Grouped Data

Quartiles are values that split up a dataset into four equal parts.

You can use the following formula to calculate quartiles for grouped data:

$$Q_i = L + (C/F) * (iN/4 - M)$$

where:

- L: The lower bound of the interval that contains the  $i^{\text{th}}$  quartile
- C: The class width
- F: The frequency of the interval that contains the  $i^{\text{th}}$  quartile
- N: The total frequency
- M: The cumulative frequency leading up to the interval that contains the  $i^{\text{th}}$  quartile

Class Interval	Frequency	Cumulative Frequency
1-5	6	6
6-10	19	25
11-15	13	38
16-20	20	58
21-25	12	70
26-30	11	81
31-35	6	87
36-40	5	92

Third quartile (Q3) of this distribution.

The value at the third quartile will be located at position  $(iN/4)$  in the distribution.

Thus,  $(iN/4) = (3 \cdot 92/4) = 69$ .

The interval that contains the third quartile will be the 21-25 interval since 69 is between the cumulative frequencies of 58 and 70.

- $Q_i = L + (C/F) * (iN/4 - M)$
- $Q_3 = 21 + (4/12) * ((3)(92)/4 - 58)$
- $Q_3 = 24.667$

## Measures of Variability

A measure of variability is a summary statistic that represents the amount of dispersion in a dataset.

While a measure of central tendency describes the typical value, measures of variability define how far away the data points tend to fall from the center. We talk about variability in the context of a distribution of values.

A low dispersion indicates that the data points tend to be clustered tightly around the center. High dispersion signifies that they tend to fall further away.

Some commonly used measures of variability are.

1. Range,
2. Inter quartile range
3. Quartile deviation
4. Mean deviation
5. Variance
6. Standard deviation
7. Coefficient of variation

### RANGE

The simplest measure of variability is the range.

Range= Largest value -Smallest value

### Interquartile range

A measure of variability that overcomes the dependency on extreme values is the interquartile range (IQR). This measure of variability is the difference between the third quartile, Q3, and the first quartile, Q1. In other words, the interquartile range is the range for the middle 50% of the data

### Quartile deviation

The Quartile Deviation can be defined mathematically as half of the difference between the upper and lower quartile. Here, quartile deviation can be represented as QD; Q3 denotes the upper quartile and Q1 indicates the lower quartile. Quartile Deviation is also known as the Semi Interquartile range.

$$\text{Quartile Deviation Formula} = \frac{Q_3 - Q_1}{2}$$

### Interquartile range

The quartiles divide the starting salary data into four parts, with each part containing 25% of the observations.

3310	3355	3450	3480	3480	3490	3520	3540	3550	3650	3730	3925
			$Q_1 = 3465$		$Q_2 = 3505$ (Median)		$Q_3 = 3600$				

$$IQR = Q_3 - Q_1$$



the quartiles are Q3 **3600** and Q1 **3465**. Thus, the interquartile range is **3600 -3465= 135**.

### Mean deviation

The difference between each  $x_i$  and the mean ( $\bar{x}$  for a sample,  $\mu$  for a population) is called a deviation about the mean.

### Variance

The **variance** is a measure of variability that utilizes all the data. The variance is based on the difference between the value of each observation ( $x_i$ ) and the mean.

The difference between each  $x_i$  and the mean ( $\bar{x}$  for a sample,  $\mu$  for a population) is called a deviation about the mean.

### Sample variance

For a sample, a deviation about the mean is written ( $x_i - \bar{x}$ ); In the computation of the sample variance, the deviations about the mean are squared and divided by ( $n-1$ )

### Population variance

For a population, a deviation about the mean is written ( $x_i - \mu$ ). If the data are for a population, the average of the squared deviations is called the population variance. The population variance is denoted by the Greek symbol  $\sigma^2$ .

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$
$\sigma^2$ = population variance $x_i$ = value of $i^{th}$ element $\mu$ = population mean $N$ = population size	$s^2$ = sample variance $x_i$ = value of $i^{th}$ element $\bar{x}$ = sample mean $n$ = sample size

### Standard deviation

The standard deviation is defined to be the positive square root of the variance. Following the notation, we adopted for a sample variance and a population variance, we use **s** to denote the sample standard deviation and  **$\sigma$**  to denote the population standard deviation. The standard deviation is derived from the variance in the following way

### Sample

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$$

### Population

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

Sample standard deviation =  $s = \sqrt{s^2}$

Population standard deviation =  $\sigma = \sqrt{\sigma^2}$

### **Coefficient of variation**

In some situations, we may be interested in a descriptive statistic that indicates how large the standard deviation is relative to the mean. This measure is called the coefficient of variation and is usually expressed as a percentage.

#### COEFFICIENT OF VARIATION

$$\left( \frac{\text{Standard deviation}}{\text{Mean}} \times 100 \right) \%$$