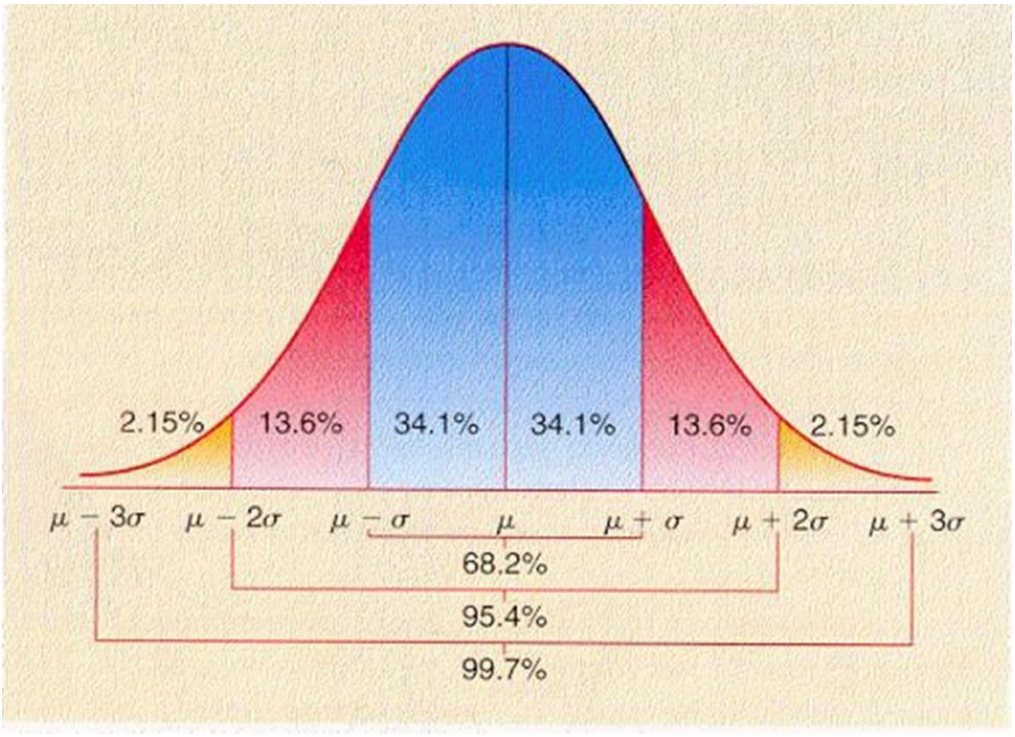
Empirical Rule of Standard Deviation

The Measures

- For symmetrical, bell shaped frequency distribution (also called **Normal Curve**), the range within which a given percentage of values of the distribution are likely to fall within a specified number of standard deviation of the mean is determined as follows:
    - μ ± $\boldsymbol{\sigma}$ covers approximately 68.27% of values in the data set
    - μ ± $\mathbf{2}\boldsymbol{\sigma}$ covers approximately 95.45% of values in the data set
    - μ ± $\mathbf{3}\boldsymbol{\sigma}$ covers approximately 99.73% of values in the data set

The Figure



Test Yourself 1

The following data give the number of passengers travelling by airplane from one city to another in one week.

115 112 129 113 119 124 132 120 110 116

Calculate the mean and standard deviation and determine the percentage of class that lie between **(i) μ ± σ; (ii)μ ± 2σ; and (iii) μ ± 3σ.**

What percentage of cases lie outside these limits?

Answer

x_i	x_i^2
115	13225
122	14884
129	16641
113	12769
119	14161
124	15376
132	17424
120	14400
110	12100
116	13456
$\sum x_i = 1200$	$\sum x_i^2 = 144436$

Variance, $\boldsymbol{\sigma}^2 = \frac{1}{N}\left[\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}\right] = \frac{1}{10}\left[144436 - \frac{(1200)^2}{10}\right] = 43.6$

$\text{SD} = \boldsymbol{\sigma} = 6.60$

The percentage of cases that lie between a given limit are as follows:

Interval	Values within Interval	Percentage of Population	Percentage falling outside
$\mu \pm \sigma$ $= 120 \pm 6.60$ $= 113.4 \; and \; 126.6$	115, 116, 119, 120,122, 124	60%	40%
$\mu \pm 2\sigma$ $= 120 \pm 2(6.60)$ $= 106.8 \; and \; 133.2$	110, 113, 115, 116,119, 120, 122,124,129, 132	100%	Nil

Shape characteristics of a distribution

Shape Characteristics

- The study of shape characteristics of a distribution is of crucial importance in comparing a distribution with other distributions.
- By shape characteristic of a distribution we refer to the extent of its **asymmetry and peakedness** relative to an agreed upon standard.
- The study of these two characteristics (i.e. asymmetry and peakedness) is accomplished through what is known as the measures of skewness and kurtosis, respectively.

Skewness

- The term skewness means the lack of symmetry; it may be either positive or negative.
- When the skewness is positive the associated distribution is called positively skewed.
- When the skewness is negative the associated distribution is negatively skewed.
- Simply, for a distribution,
    - Mean > Median > Mode: The distribution is **positively skewed**
    - Mean < Median < Mode: The distribution is **negatively skewed**
    - Mean = Median = Mode: The distribution is **symmetric**

Measuring Skewness

- Pearson's coefficient for Skewness,

$$\boldsymbol{Sk_p} = \frac{\boldsymbol{Mean - Mode}}{\boldsymbol{SD}} = \frac{\boldsymbol{3\,(Mean - Median)}}{\boldsymbol{SD}}$$

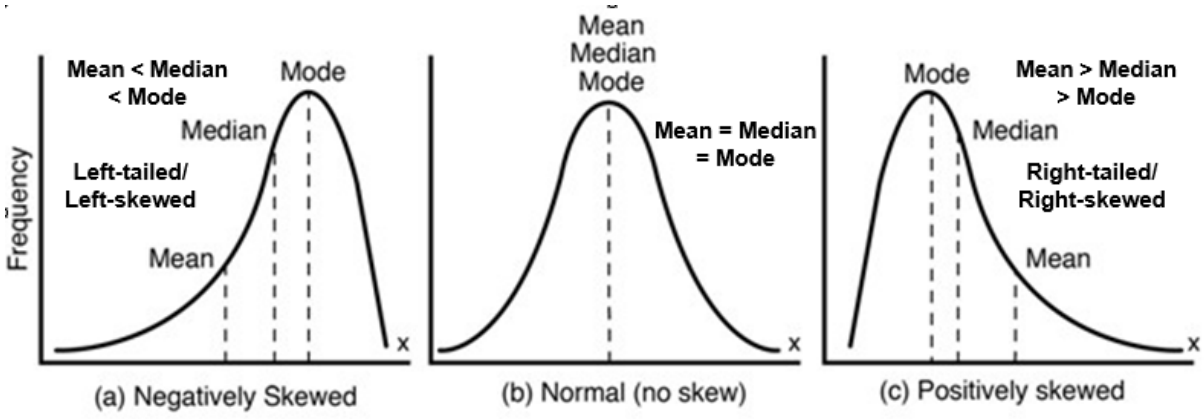
- Then if,
    - $Sk_p > 0$: The distribution is **positively skewed**
    - $Sk_p < 0$: The distribution is **negatively skewed**
    - $Sk_p = 0$: The distribution is **symmetric**

- Bowley's coefficient for Skewness i.e. quartile skewness coefficient,

$$\boldsymbol{Sk_b = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}}$$

- Then if,
    - $Sk_b > 0$: The distribution is **positively skewed**
    - $Sk_b < 0$: The distribution is **negatively skewed**
    - $Sk_b = 0$: The distribution is **symmetric**

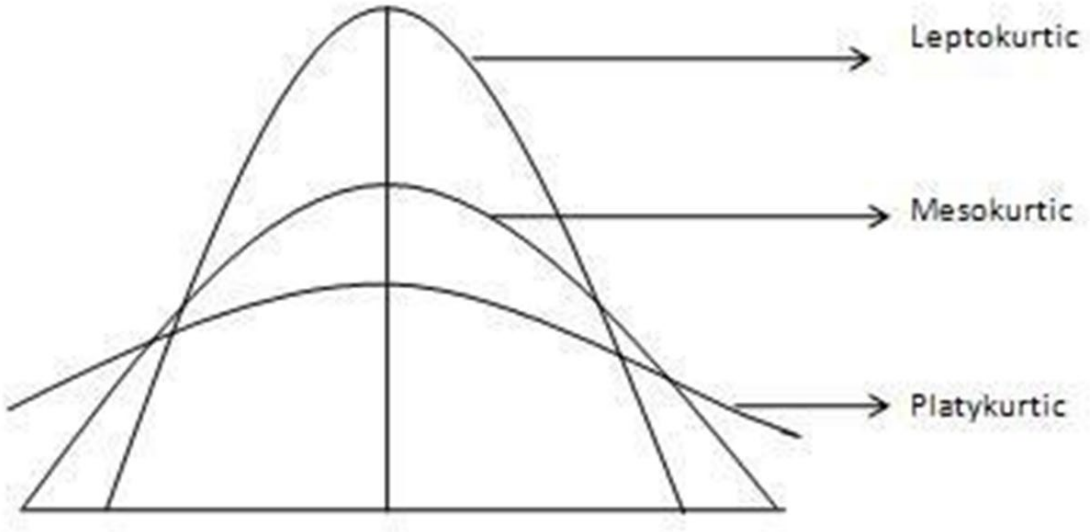
Skewness of Distribution



Kurtosis

- There is considerable variation among symmetrical distributions.
- For instance, they can differ markedly in terms of **peakedness**. This is what we call **kurtosis**.
- **Kurtosis**, as defined by Spiegel, is the degree of **peakedness** of a distribution, usually taken in relation to a normal distribution.
    - A curve having relatively higher peak than the normal curve, is known as **leptokurtic**.
    - A curve, which is neither too peaked nor too flat topped, is known as **mesokurtic**.
    - A curve that is more flat topped than the normal curve is called **platykurtic**.

Kurtosis of Distribution



Summary

- Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.
- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. It helps to determine if the curve is more or less high as compared to the normal curve.

Test Yourself 2

1. If for a distribution Mean =18, Median= 32, and Mode= 36, the distribution is negatively skewed. [Mean < Median < Mode]

2. If for a distribution Mean = 20, Median= 26.4, and SD= 3.3, the distribution is negatively skewed.

$Sk_p = \frac{3(Mean-Median)}{SD}$ = -5.818<0 [Negatively skewed]

3. If for a distribution, Mean = 35.6, Mode = 24, and SD= 5.2, what is the skewness coefficient of the distribution?

$Sk_p = \frac{Mean-Mode}{SD}$ = 2.2307>0. The distribution is positively skewed.

Box Plot

- A box plot is a graphic display that shows the general shape of a variable's distribution.
- It is based on five descriptive statistics:
    - The minimum value,
    - The first quartile (Q_1),
    - Median,
    - Third quartile (Q_3), and
    - The maximum value

Example

- Pizza Hut offers free delivery of its pizza within 15 miles. Mr. Rahman, the owner, wants some information on the time it takes for delivery e.g. how long does a typical delivery take, or within what range of times will most deliveries be completed. For a sample of 20 deliveries, he determined the following information:

Minimum value = 13 minutes

Q_1 = 15 minutes

Median = 18 minutes

Q_3 = 22 minutes

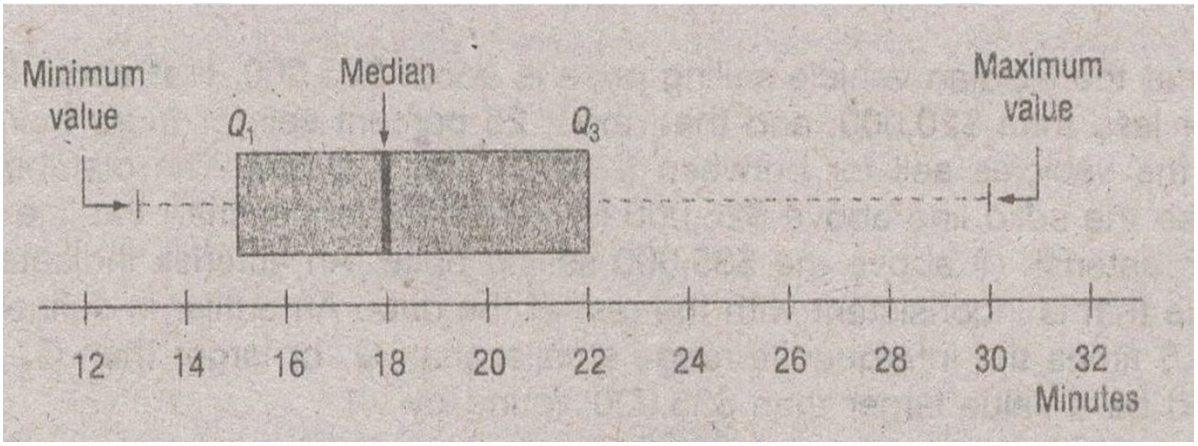
Maximum value = 30 minutes

Develop a box-plot for the delivery times. What conclusions can you make about the delivery times?

Solution

- In order to draw box plot follow the steps mentioned below:
- Step 1: Create an appropriate scale along the horizontal axis.
- Step 2: Draw a box that starts at $Q1$ (15 minutes) and ends at $Q3$ (22 minutes)
- Step 3: Place a vertical line to represent the median (18 minutes)
- Step 4: Extend the horizontal lines from the box out to the minimum value (13 minutes) and the maximum value (30 minutes). These horizontal lines are sometimes called "whiskers" due to the resemblance with cat's whiskers.

Box Plot for Pizza Delivery



Interpretation of the Box Plot

- The box plot shows that the middle 50 percent of the deliveries take between 15 minutes and 22 minutes. The distance between the ends of the box, 7 minutes, is the inter quartile range i.e. the distance between the first and third quartile. That shows the spread or dispersion of the majority of deliveries.
- The box plot also reveals that the distribution of the delivery times is positively skewed. The guiding principle for such conclusion are
    - The dashed line to the right of the box from 22 minutes (Q_3) to the maximum time of 30 minutes is longer than the dashed line from the left of 15 minutes (Q_1) to the minimum value of 13 minutes.
    - The median is not in the middle in the center of the box. The distance from the first quartile to the median is smaller than the distances from the median to the third quartile.

Test Yourself 3

Construct a box plot for the data given below and hence comment on the skewness of the distribution:

99	75	84	33	45	66	97	69	55	61
72	91	74	93	54	76	62	91	77	68

Minimum value = 33

Maximum value = 99

Median = 73

First Quartile Q_1:

$\frac{in}{4} = \frac{1*20}{4} = 5$, an integer

$Q_1 = \frac{1}{2}[61 + 62] = 61.5$

Second Quartile Q_2:

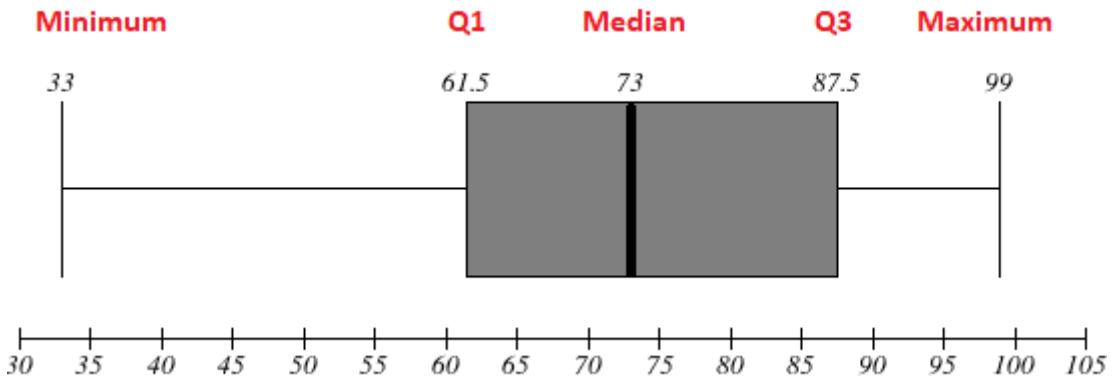
$\frac{in}{4} = \frac{2*20}{4} = 10$, an integer

$Q_2 = \frac{1}{2}[72 + 74] = 73$

Third Quartile Q_3:

$\frac{in}{4} = \frac{3*20}{4} = 15$, an integer

$Q_3 = \frac{1}{2}[84 + 91] = 87.5$



Using Bowley's coefficient for Skewness we get

$Sk_b = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = 0.1154 > 0$, the distribution is positively skewed.