Chapter 6. Descriptive Statistics

FPT Education
FPT UNIVERSITY

February 16, 2022

Take a sample of n observations (data) $x_1, x_2, \cdots, x_n$ from a population.

Definition

(1) Sample mean $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

(2) Sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}.$$

Take a sample of n observations (data) $x_1, x_2, \cdots, x_n$ from a population.

Definition

(1) Sample mean $\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

(2) Sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}.$$

(3) Sample standard deviation $s = \sqrt{s^2}$.

FPT Education FPT UNIVERSITY

Take a sample of n observations (data) $x_1, x_2, \cdots, x_n$ from a population.

Definition

(1) Sample mean $\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

(2) Sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}.$$

(3) Sample standard deviation $s = \sqrt{s^2}$.

(4) The sample range is $r = \max(x_i) - \min(x_i)$.

FPT Education FPT UNIVERSITY

Take a sample of n observations (data) $x_1, x_2, \cdots, x_n$ from a population.

Definition

(1) Sample mean $\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

(2) Sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}.$$

(3) Sample standard deviation $s = \sqrt{s^2}$.

(4) The sample range is $r = \max(x_i) - \min(x_i)$.

These values are good for the estimation of mean, variance (standard deviation) of the investigating population.

Take a sample of n observations (data) $x_1, x_2, \cdots, x_n$ from a population.

Definition

(1) Sample mean $\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

(2) Sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}.$$

(3) Sample standard deviation $s = \sqrt{s^2}$.

(4) The sample range is $r = \max(x_i) - \min(x_i)$.

These values are good for the estimation of mean, variance (standard deviation) of the investigating population.

Example

The test scores of 16 students are listed below. Find sample mean and sample standard deviation.

32	41	47	50	56	58	58	61
62	68	75	82	85	90	92	98

Example

The test scores of 16 students are listed below. Find sample mean and sample standard deviation.

32	41	47	50	56	58	58	61
62	68	75	82	85	90	92	98

Answer: $\bar{x} = 65.9375, s = \sqrt{376.329} \approx 19.4.$

Example

The test scores of 16 students are listed below. Find sample mean and sample standard deviation.

32	41	47	50	56	58	58	61
62	68	75	82	85	90	92	98

Answer: $\bar{x} = 65.9375, s = \sqrt{376.329} \approx 19.4.$

$$\begin{aligned}\bar{x} =& \frac{1}{16}(32 + 41 + 47 + 50 + 56 + 58 + 58 + 61 \\ & + 62 + 68 + 75 + 82 + 85 + 90 + 92 + 98) = 65.9375 \\ s^2 =& \frac{1}{15}(\sum_{i=1}^{16} x_i^2 - 16\bar{x}^2) \\ =& \frac{1}{15}(32^2 + 41^2 + 47^2 + 50^2 + 56^2 + 58^2 + 58^2 + 61^2 \\ & + 62^2 + 68^2 + 75^2 + 82^2 + 85^2 + 90^2 + 92^2 + 98^2 - 16 * 65.9375^2) \\ =& 376.329\end{aligned}$$

Example

The test scores of 16 students are listed below. Find sample mean and sample standard deviation.

32	41	47	50	56	58	58	61
62	68	75	82	85	90	92	98

Answer: $\bar{x} = 65.9375, s = \sqrt{376.329} \approx 19.4.$

$$
\begin{aligned}
\bar{x} =& \frac{1}{16}(32 + 41 + 47 + 50 + 56 + 58 + 58 + 61 \\
&+ 62 + 68 + 75 + 82 + 85 + 90 + 92 + 98) = 65.9375 \\
s^2 =& \frac{1}{15}(\sum_{i=1}^{16} x_i^2 - 16\bar{x}^2) \\
=& \frac{1}{15}(32^2 + 41^2 + 47^2 + 50^2 + 56^2 + 58^2 + 58^2 + 61^2 \\
&+ 62^2 + 68^2 + 75^2 + 82^2 + 85^2 + 90^2 + 92^2 + 98^2 - 16 * 65.9375^2) \\
=& 376.329
\end{aligned}
$$

A stem-and-leaf diagram is a good way to obtain an informative visual display of a data set where each number x_i consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps:

(1) Divide each number x_i into two parts: a stem, consisting of one or more of the leading digits, and a leaf, consisting of the remaining digit.

6.2 Stem-and-Leaf Diagrams

A stem-and-leaf diagram is a good way to obtain an informative visual display of a data set where each number x_i consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps:

(1) Divide each number x_i into two parts: a stem, consisting of one or more of the leading digits, and a leaf, consisting of the remaining digit.

(2) List the stem values in a vertical column.

A stem-and-leaf diagram is a good way to obtain an informative visual display of a data set where each number x_i consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps:

(1) Divide each number x_i into two parts: a stem, consisting of one or more of the leading digits, and a leaf, consisting of the remaining digit.

(2) List the stem values in a vertical column.

(3) Record the leaf for each observation beside its stem.

6.2 Stem-and-Leaf Diagrams

A stem-and-leaf diagram is a good way to obtain an informative visual display of a data set where each number x_i consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps:

(1) Divide each number x_i into two parts: a stem, consisting of one or more of the leading digits, and a leaf, consisting of the remaining digit.

(2) List the stem values in a vertical column.

(3) Record the leaf for each observation beside its stem.

(4) Write the units for stems and leaves on the display.

6.2 Stem-and-Leaf Diagrams

A stem-and-leaf diagram is a good way to obtain an informative visual display of a data set where each number x_i consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps:

(1) Divide each number x_i into two parts: a stem, consisting of one or more of the leading digits, and a leaf, consisting of the remaining digit.

(2) List the stem values in a vertical column.

(3) Record the leaf for each observation beside its stem.

(4) Write the units for stems and leaves on the display.

A stem-and-leaf diagram is a good way to obtain an informative visual display of a data set where each number x_i consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps:

(1) Divide each number x_i into two parts: a stem, consisting of one or more of the leading digits, and a leaf, consisting of the remaining digit.

(2) List the stem values in a vertical column.

(3) Record the leaf for each observation beside its stem.

(4) Write the units for stems and leaves on the display.

Example

The listing scores of 17 students in a TOEIC test are listed below

55	115	225	240	330	335
385	400	405	405	495	495

The stem-and-leaf diagram:

Stem	Leaves
5	5
11	5
22	5
24	0
33	0 5
38	5
40	0 5 5
49	5 5

Example

The listening scores of 17 students in a TOEIC test are listed below

55	115	225	240	330	335
385	400	405	405	495	495

The stem-and-leaf diagram:

Stem	Leaves
5	5
11	5
22	5
24	0
33	0 5
38	5
40	0 5 5
49	5 5

Definition

(1) The sample median is a measure of central tendency that divides the data into two equal parts, half below the median and half above. If the number of observations is even, the median is halfway between the two central values.

(2) The sample mode is the most frequently occurring data value.

Definition

(1) The sample median is a measure of central tendency that divides the data into two equal parts, half below the median and half above. If the number of observations is even, the median is halfway between the two central values.

(2) The sample mode is the most frequently occurring data value.

(3) An ordered set of data is divided into four equal parts, the division points are called quartiles:

Definition

(1) The sample median is a measure of central tendency that divides the data into two equal parts, half below the median and half above. If the number of observations is even, the median is halfway between the two central values.

(2) The sample mode is the most frequently occurring data value.

(3) An ordered set of data is divided into four equal parts, the division points are called quartiles:

- The first quartile, q_1 or Q_1: is a value that has approximately 25% of the observations below.

Definition

(1) The sample median is a measure of central tendency that divides the data into two equal parts, half below the median and half above. If the number of observations is even, the median is halfway between the two central values.

(2) The sample mode is the most frequently occurring data value.

(3) An ordered set of data is divided into four equal parts, the division points are called quartiles:

- The first quartile, q_1 or Q_1: is a value that has approximately 25% of the observations below.
- The sample median or second quartile, q_2 or Q_2, has approximately 50% of the observations below its value.

Definition

(1) The sample median is a measure of central tendency that divides the data into two equal parts, half below the median and half above. If the number of observations is even, the median is halfway between the two central values.

(2) The sample mode is the most frequently occurring data value.

(3) An ordered set of data is divided into four equal parts, the division points are called quartiles:

- The first quartile, q_1 or Q_1: is a value that has approximately 25% of the observations below.
- The sample median or second quartile, q_2 or Q_2, has approximately 50% of the observations below its value.
- The third quartile, q_3 or Q_3, has approximately 75% of the observations below its value.

Definition

(1) The sample median is a measure of central tendency that divides the data into two equal parts, half below the median and half above. If the number of observations is even, the median is halfway between the two central values.

(2) The sample mode is the most frequently occurring data value.

(3) An ordered set of data is divided into four equal parts, the division points are called quartiles:

- The first quartile, q_1 or Q_1: is a value that has approximately 25% of the observations below.
- The sample median or second quartile, q_2 or Q_2, has approximately 50% of the observations below its value.
- The third quartile, q_3 or Q_3, has approximately 75% of the observations below its value.

(4) The interquartile range, $IQR = q_3 - q_1$.

Definition

(1) The sample median is a measure of central tendency that divides the data into two equal parts, half below the median and half above. If the number of observations is even, the median is halfway between the two central values.

(2) The sample mode is the most frequently occurring data value.

(3) An ordered set of data is divided into four equal parts, the division points are called quartiles:

- The first quartile, q_1 or Q_1: is a value that has approximately 25% of the observations below.
- The sample median or second quartile, q_2 or Q_2, has approximately 50% of the observations below its value.
- The third quartile, q_3 or Q_3, has approximately 75% of the observations below its value.

(4) The interquartile range, $IQR = q_3 - q_1$.

Example

Use the given sample data to find the sample quartiles, the sample mode and the IQR.

$$55, 52, 52, 52, 49, 74, 67, 55.$$

Example

Use the given sample data to find the sample quartiles, the sample mode and the *IQR*.

$$55, 52, 52, 52, 49, 74, 67, 55.$$

Answer: We rearrange the increasing data:

$$49, 52, 52, 52, 55, 55, 67, 74.$$

Example

Use the given sample data to find the sample quartiles, the sample mode and the IQR.

$$55, 52, 52, 52, 49, 74, 67, 55.$$

Answer: We rearrange the increasing data:

$$49, 52, 52, 52, 55, 55, 67, 74.$$

We the size of sample $n = 8$.

Example

Use the given sample data to find the sample quartiles, the sample mode and the IQR.

$$55, 52, 52, 52, 49, 74, 67, 55.$$

Answer: We rearrange the increasing data:

$$49, 52, 52, 52, 55, 55, 67, 74.$$

We the size of sample $n = 8$.

(1) $\frac{n+1}{2} = 4.5$, hence the sample median $Q_2 = \frac{x_4+x_5}{2} = \frac{52+55}{2} = 53.5$.

Example

Use the given sample data to find the sample quartiles, the sample mode and the IQR.

$$55, 52, 52, 52, 49, 74, 67, 55.$$

Answer: We rearrange the increasing data:

$$49, 52, 52, 52, 55, 55, 67, 74.$$

We the size of sample $n = 8$.

(1) $\frac{n+1}{2} = 4.5$, hence the sample median $Q_2 = \frac{x_4+x_5}{2} = \frac{52+55}{2} = 53.5$.

(2) $\frac{n+1}{4} = 2.25$, hence the first quartile $Q_1 = \frac{x_2+x_3}{2} = \frac{52+52}{2} = 52$.

Example

Use the given sample data to find the sample quartiles, the sample mode and the *IQR*.

$$55, 52, 52, 52, 49, 74, 67, 55.$$

Answer: We rearrange the increasing data:

$$49, 52, 52, 52, 55, 55, 67, 74.$$

We the size of sample $n = 8$.

(1) $\frac{n+1}{2} = 4.5$, hence the sample median $Q_2 = \frac{x_4+x_5}{2} = \frac{52+55}{2} = 53.5$.

(2) $\frac{n+1}{4} = 2.25$, hence the first quartile $Q_1 = \frac{x_2+x_3}{2} = \frac{52+52}{2} = 52$.

(3) $\frac{3(n+1)}{4} = 6.75$, hence the third quartile $Q_3 = \frac{x_6+x_7}{2} = \frac{55+67}{2} = 61$.

Example

Use the given sample data to find the sample quartiles, the sample mode and the IQR.

$$55, 52, 52, 52, 49, 74, 67, 55.$$

Answer: We rearrange the increasing data:

$$49, 52, 52, 52, 55, 55, 67, 74.$$

We the size of sample $n = 8$.

(1) $\frac{n+1}{2} = 4.5$, hence the sample median $Q_2 = \frac{x_4+x_5}{2} = \frac{52+55}{2} = 53.5$.

(2) $\frac{n+1}{4} = 2.25$, hence the first quartile $Q_1 = \frac{x_2+x_3}{2} = \frac{52+52}{2} = 52$.

(3) $\frac{3(n+1)}{4} = 6.75$, hence the third quartile $Q_3 = \frac{x_6+x_7}{2} = \frac{55+67}{2} = 61$.

(4) The sample modes: 52.

Example

Use the given sample data to find the sample quartiles, the sample mode and the IQR.

$$55, 52, 52, 52, 49, 74, 67, 55.$$

Answer: We rearrange the increasing data:

$$49, 52, 52, 52, 55, 55, 67, 74.$$

We the size of sample $n = 8$.

(1) $\frac{n+1}{2} = 4.5$, hence the sample median $Q_2 = \frac{x_4 + x_5}{2} = \frac{52+55}{2} = 53.5$.

(2) $\frac{n+1}{4} = 2.25$, hence the first quartile $Q_1 = \frac{x_2 + x_3}{2} = \frac{52+52}{2} = 52$.

(3) $\frac{3(n+1)}{4} = 6.75$, hence the third quartile $Q_3 = \frac{x_6 + x_7}{2} = \frac{55+67}{2} = 61$.

(4) The sample modes: 52.

(5) $IQR = Q_3 - Q_1 = 61 - 52 = 9$.

Example

Use the given sample data to find the sample quartiles, the sample mode and the IQR.

$$55, 52, 52, 52, 49, 74, 67, 55.$$

Answer: We rearrange the increasing data:

$$49, 52, 52, 52, 55, 55, 67, 74.$$

We the size of sample $n = 8$.

(1) $\frac{n+1}{2} = 4.5$, hence the sample median $Q_2 = \frac{x_4 + x_5}{2} = \frac{52+55}{2} = 53.5$.

(2) $\frac{n+1}{4} = 2.25$, hence the first quartile $Q_1 = \frac{x_2 + x_3}{2} = \frac{52+52}{2} = 52$.

(3) $\frac{3(n+1)}{4} = 6.75$, hence the third quartile $Q_3 = \frac{x_6 + x_7}{2} = \frac{55+67}{2} = 61$.

(4) The sample modes: 52.

(5) $IQR = Q_3 - Q_1 = 61 - 52 = 9$.

Example

Use the given sample data to find the sample quartiles, the sample mode and the IQR.

$$55, 52, 52, 52, 49, 74, 67, 55.$$

Answer: We rearrange the increasing data:

$$49, 52, 52, 52, 55, 55, 67, 74.$$

We the size of sample $n = 8$.

(1) $\frac{n+1}{2} = 4.5$, hence the sample median $Q_2 = \frac{x_4 + x_5}{2} = \frac{52+55}{2} = 53.5$.

(2) $\frac{n+1}{4} = 2.25$, hence the first quartile $Q_1 = \frac{x_2 + x_3}{2} = \frac{52+52}{2} = 52$.

(3) $\frac{3(n+1)}{4} = 6.75$, hence the third quartile $Q_3 = \frac{x_6 + x_7}{2} = \frac{55+67}{2} = 61$.

(4) The sample modes: 52.

(5) $IQR = Q_3 - Q_1 = 61 - 52 = 9$.

Construction of frequency distribution: divide the range of the data into intervals (called class intervals, cells, or bins). The bins should be of equal width.

Example

We have frequency distribution of MAS291 final scores of the class IS1402.

scores	frequency	cumulative frequency distribution	relative frequency	cumulative relative frequency
0-2	2	2	$2/30$	$2/30$
2-4	5	7	$5/30$	$7/30$
4-6	8	15	$8/30$	$15/30 = 0.5$
6-8	11	26	$11/30$	$26/30$
8-10	4	30	$4/30$	1

Construction of frequency distribution: divide the range of the data into intervals (called class intervals, cells, or bins). The bins should be of equal width.

Example

We have frequency distribution of MAS291 final scores of the class IS1402.

scores	frequency	cumulative frequency distribution	relative frequency	cumulative relative frequency
0-2	2	2	$2/30$	$2/30$
2-4	5	7	$5/30$	$7/30$
4-6	8	15	$8/30$	$15/30=0.5$
6-8	11	26	$11/30$	$26/30$
8-10	4	30	$4/30$	1

Histogram

The histogram is a visual display of the frequency distribution.

(1) Label the bin (class interval) boundaries on a horizontal scale.

Histogram

The histogram is a visual display of the frequency distribution.

(1) Label the bin (class interval) boundaries on a horizontal scale.

(2) Mark and label the vertical scale with the frequencies or the relative frequencies.

Histogram

FPT Education
FPT UNIVERSITY

The histogram is a visual display of the frequency distribution.

(1) Label the bin (class interval) boundaries on a horizontal scale.
(2) Mark and label the vertical scale with the frequencies or the relative frequencies.
(3) Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.

Histogram

The histogram is a visual display of the frequency distribution.

(1) Label the bin (class interval) boundaries on a horizontal scale.
(2) Mark and label the vertical scale with the frequencies or the relative frequencies.
(3) Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.

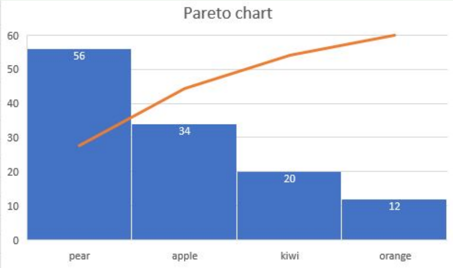
Histogram

The histogram is a visual display of the frequency distribution.

(1) Label the bin (class interval) boundaries on a horizontal scale.
(2) Mark and label the vertical scale with the frequencies or the relative frequencies.
(3) Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.

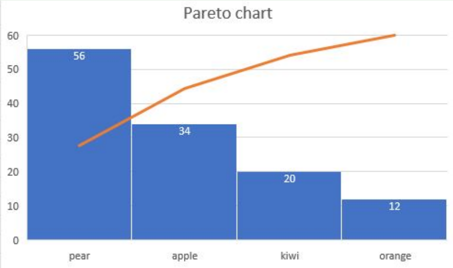
Example

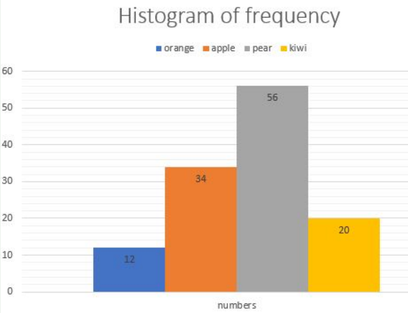
fruits	numbers
orange	12
apple	34
pear	56
kiwi	20



Example

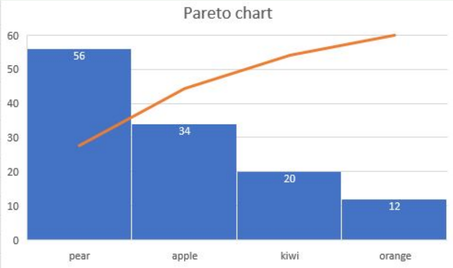
fruits	numbers
orange	12
apple	34
pear	56
kiwi	20

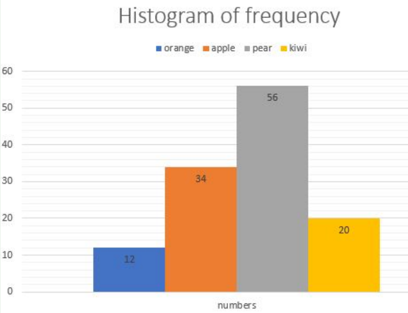




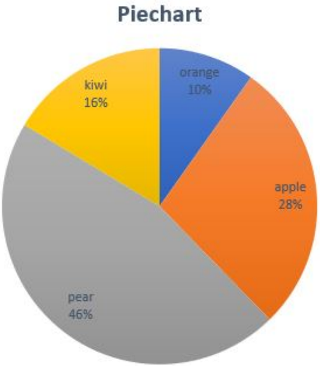
Example

fruits	numbers
orange	12
apple	34
pear	56
kiwi	20

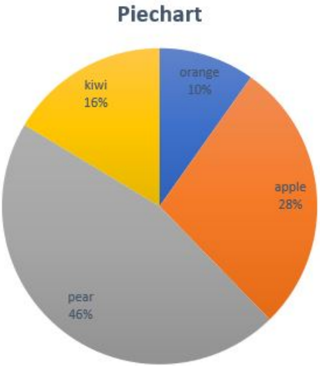




fruits	numbers
orange	12
apple	34
pear	56
kiwi	20



fruits	numbers
orange	12
apple	34
pear	56
kiwi	20

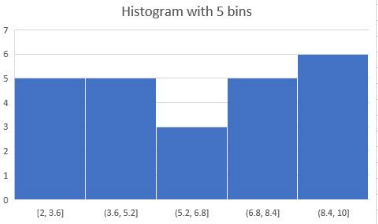


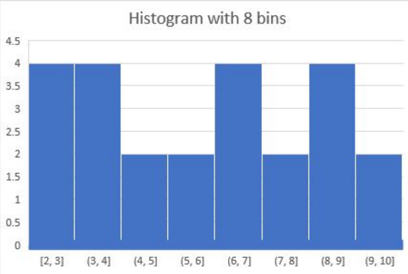
Example

Given marks of 24 students in a mathematics examination

2	2.5	3	3	3.5	4	4	4	4.5	4.5	6	6
6.5	7	7	7	7.5	8	8.5	8.5	9	9	9.5	10

Draw a histogram with 5 bins and a histogram with 8 bins.



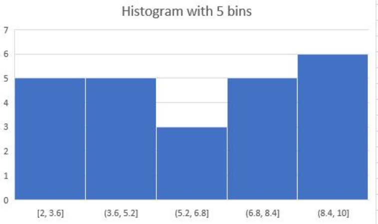


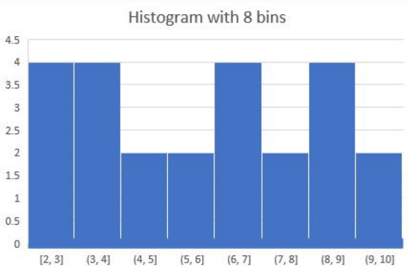
Example

Given marks of 24 students in a mathematics examination

2	2.5	3	3	3.5	4	4	4	4.5	4.5	6	6
6.5	7	7	7	7.5	8	8.5	8.5	9	9	9.5	10

Draw a histogram with 5 bins and a histogram with 8 bins.



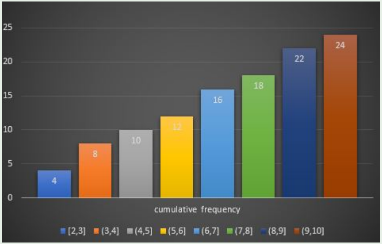


Example

Given marks of 24 students in a mathematics examination

2	2.5	3	3	3.5	4	4	4	4.5	4.5	6	6
6.5	7	7	7	7.5	8	8.5	8.5	9	9	9.5	10

Draw histograms of cumulative frequency and cumulative relative frequency with 8 bins.

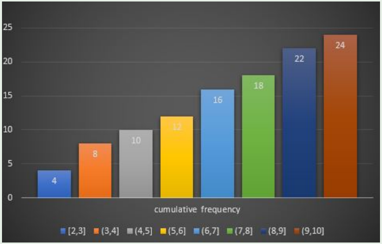


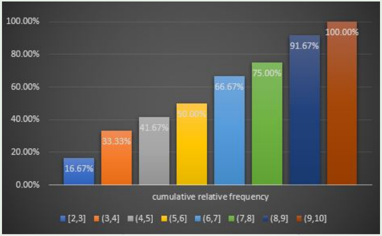
Example

Given marks of 24 students in a mathematics examination

2	2.5	3	3	3.5	4	4	4	4.5	4.5	6	6
6.5	7	7	7	7.5	8	8.5	8.5	9	9	9.5	10

Draw histograms of cumulative frequency and cumulative relative frequency with 8 bins.





Example

Given marks of 24 students in a mathematics examination

2	2.5	3	3	3.5	4	4	4	4.5	4.5	6	6
6.5	7	7	7	7.5	8	8.5	8.5	9	9	9.5	10

Draw histograms of cumulative frequency and cumulative relative frequency with 8 bins.