CHAPTER 1

DESCRIPTION OF DATA

Objectives

After the completion of this chapter, the student should be able to calculate and interpret the following from a given sample of data:

(1) Measures of centrality: Arithmetic Mean, Mode, and Median. Relation between these three.
(2) Summary Measures: Quartiles, Deciles and Percentiles.
(3) Measures of variability: Range, Interquartile Range, Variance and Standard Deviation.
(4) Composite Summary Measures: Coefficient of Variation and the Coefficient of Skewness.
(5) Graphical representation of data: Box Plot, Histogram, Polygon and Ogive.

1.1 Introduction

Statistics is one of the most important branches of applied mathematics. It is used in fields like agriculture, physical sciences and human sciences. Descriptive statistics is mainly concerned with summary calculations and graphical displays of data.

In this module we will study the description of data obtained from samples.

1.2 Definitions

1.2.1 Population / Sample

A population is the total set of elements of interest for a given problem. A sample is a subset of a population. In later studies we use statistics obtained from samples to make predictions for the population the sample came from. This is called **Statistical Inference.**

1.2.2 The statistic

A statistic is a function of the sample data. Examples of a statistic are the arithmetic mean, mode, median and standard deviation. These will be defined in the following sections.

1.2.3 Arithmetic mean

The arithmetic mean is the sum of all the data divided by the number of observations. Thus if X_1, X_2, ..., X_n are n observations, then we write the arithmetic mean as follows:

$$\bar{X} = \frac{X_1+X_2+\ldots+X_n}{n}$$

$$\bar{X}=\frac{\sum X_i}{n}$$

The Greek capital letter Σ (Sigma) is used to indicate the sum of a number of elements. The subscript i in formula (1.1) is a discrete variable which assumes the values 1, then 2, 3,... up to the last observation n. We will often use this notation.

1.2.4 The mode

The mode is the most frequently occurring value. If it is important to know which observation occurred most often, the mode is determined.

1.2.5 The median

The median is the middle value when the data are arranged in numerical sequence. Fifty percent of the observations are greater than the median and 50% are less.

1.3 Measures of deviation

All the previous statistics, the mean, mode and median are measures of central tendency, i.e. a number which tries to indicate a "central" value in a set of data. Statistics which describe the spread or variation of data are called measures of deviation or dispersion.

1.3.1 The range

The range is the simplest statistic which gives a indication of the spread of a set of numbers. It is simply equal to the difference between the lowest and highest value in a sample. The range has the serious disadvantage in that it only uses the extreme two values of the sample and ignores the information contained in the rest of the sample. The following statistics overcome this problem.

1.3.2 The variance

If we add all the squared differences between each sample value and the sample mean and divide this quantity by (n -1), we obtain the variance:

$$S^2=\frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\ldots+(X_n-\bar{X})^2}{n-1}$$

And if we use the Σ notation:

$$S^2=\frac{\sum(X_i-\bar{X})^2}{n-1}$$

The **standard deviation** S is the square root of the variance:

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

1.3.3 **Example 1**

Consider the following data set:

X_i : 24, 13, 18, 35, 5, 28, 31, 24, 29

The mean:

$$\bar{X} = \frac{1}{n}\sum X_i$$
$$= \frac{24+13+...+29}{9}$$
$$= 23$$

The mode: the number 24 occurs twice, hence this is the mode.

The median: first arrange the data set in ascending order:

X_i : 5, 13, 18, 24, 24, 28, 29, 31, 35

↑

Hence the median is 24.

Note: If the set of data has an even number of elements we obtain the median by computing the mean of the two middle numbers.

The standard deviation:

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$
$$= \frac{(5-23)^2+(13-23)^2+...+(35-23)^2}{9-1}$$
$$= 9{,}49$$

Remarks:

(1) The mode and median are equal in this case, but this is in general not true.

(2) Another data set Y_i might have a equal mean but a smaller or greater standard deviation, depending on the variation in the data set.

1.4 Grouped data

When summarizing large masses of raw data, it is often useful to distribute the data into classes or categories. The number of individuals belonging to each class is determined, and this is called the class frequency. A tabular arrangement of data by classes together with the corresponding class frequency is called a frequency distribution or frequency table. Table 1.1 is a sample of 60 determinations of the copper content of a standard solution.

Table 1.1

COPPER DETERMINATIONS (p.p.m.)

61,0	65,4	60,0	59,2	57,0	62,5	57,7	56,2	62,9
56,5	60,2	58,2	56,5	64,7	54,5	60,5	59,5	61,6
58,7	54,4	62,2	59,0	60,3	60,8	59,5	60,0	61,8
64,5	66,3	61,1	59,7	57,4	61,2	60,9	58,2	63,0
56,0	59,4	60,0	62,9	60,5	60,8	61,5	58,5	58,9
61,2	57,8	63,4	58,9	61,5	62,3	59,8	61,7	64,0
62,5	60,8	63,8	59,9	60,5	62,7			

These observations were made to an accuracy of 3 significant numbers and to 1 decimal place. To obtain a good overall impression of such a large data set, it is necessary to group the data in a frequency distribution.

Table 1.2 COPPER DETERMINATIONS

Classes copper conc. (ppm)	Class Mark X_i	Frequency f_i	Cumulative Frequency
54,0 - 55,9	54,95	2	2
56,0 - 57,9	56,95	8	10
58,0 - 59,9	58,95	14	24
60,0 - 61,9	60,95	21	45
62,0 - 63,9	62,95	10	55
64,0 - 65,9	64,95	4	59
66,0 - 67,9	66,95	1	60

1.4.1 **Classes**

The left column contains the class intervals. In setting up the class intervals we must consider:

(1) **The number of classes**

As a guide Sturge's Rule can be used:

$$K = 1 + (3{,}3\,Log\ n)$$

Where K=the number of classes
n=the sample size

In this example the number of classes is given by:

$$K = 1 + (3{,}3\,Log\,n)$$
$$= 1 + 3{,}3\,Log\,60)$$
$$= 6{,}87 \approx 7$$

(2) **Class size**

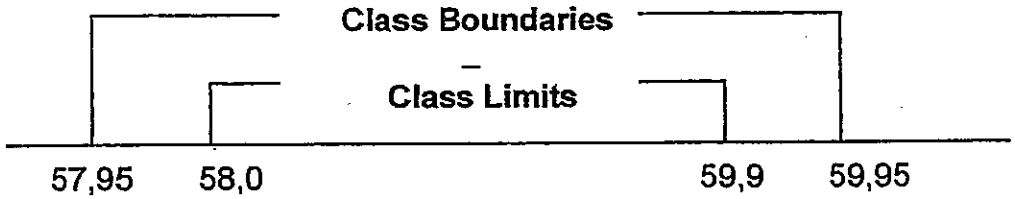
The class size is obtained by dividing the Range (highest - lowest number in data) by the number of classes obtained above. In this example the range is equal to 66,3 - 54,4 =11,9 we find the class size to be 11,9/7 =1,7. For practical reasons this is rounded to the nearest whole number. In this example we will have a class width of 2 ppm.

(3) **Class limits**

Since 54,4 was the lowest determination, we take our first class within the limits 54,0 and 55,9 (see table 1.2). The next class starts with a lower limit of 56,0 and ends with a higher limit of 57,9. This is continued until the highest determination falls in the last class.

(4) **The class boundaries**

The lower class boundary and upper class boundary of the first class will be 53,95 and 55,95 respectively. This gives us a class ppm size of 2 ppm. Since the copper determinations were accurate to one decimal place we calculate the class boundaries to two decimal places. This will prevent that an observation will fall on a class boundary. Schematically, the class limits and boundaries of the third class can be shown as follows:



(5) **The class mark**

All the determinations in each class are represented by a single number; the Class Mark. It is found by dividing the sum of the limits of a class by 2. Thus for the first class in table 1.2; (54,0 + 55,9)/2 = 54,95.

(6) **The class frequency**

The number of observations in each class is found in the third column of table 1.2.

1.4.2 **Relative frequency distributions**

The relative frequency of a class is the frequency of the class divided by the total frequency of all classes. It is generally expressed as a percentage. The relative frequency of the class 58,0 - 59,5 in table 1.2 is $(14/60)\times100 = 23,3\%$.

If the frequencies in a frequency table are replaced by the relative frequency, the resulting table is a relative frequency table.

1.4.3 Cumulative frequency distributions

The total frequency of all values less than the upper class boundary of a given class interval, is called the cumulative frequency up to and including the class interval. For example: the cumulative frequency up to and including the class interval 58 - 59,5 is 2 + 8 + 14 = 24, signifying that 24 determinations showed a copper concentration of 59,95 ppm or less.

1.4.4 The relative cumulative frequency distribution

If the cumulative frequencies are divided by the total frequency and expressed as a percentage, we obtain the relative cumulative frequency distribution.

1.5 Measures of central tendency

1.5.1 The arithmetic mean

The same definition as in 1.2.3 is used. The X_i are (in the case of a frequency table) taken as the class midpoints. If the class midpoints $X_1, X_2, ..., X_k$ (and thus the classes themselves) occur $f_1, f_2, ..., f_k$ times respectively, the arithmetic mean is:

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + ... + f_k X_k}{f_1 + f_2 + ... + f_k}$$

$$\bar{X} = \frac{\sum f_i X_i}{n}$$

Remark:
Since we let a single class midpoint represent all the determinations which are grouped in that class, we lose information. The arithmetic mean is thus not the exact mean but an approximation.

1.5.2 The mode

The same definition as for ungrouped data applies. In the case of a frequency table the mode is approximated as follows:

$$Mo = L + \left(\frac{D_1}{D_1 + D_2} \right) C$$

Where L = lower boundary of modal class.
D_1 = Difference between modal frequency and frequency of class <u>before</u> modal class
D_2 = Difference between modal frequency and frequency of the class <u>after</u> modal class.
C = Class width

1.5.3 The median

For grouped data the median is given by

$$Me = L + \left(\frac{\frac{n}{2} - (\sum f)_1}{f_{Me}} \right) C$$

where L = lower class boundary of the median class (i.e. the class containing the median)

n = number of items in the data (i.e. total frequency)

$(\Sigma f)_1$ = the sum of the frequencies before the class which contains the median.

C = the class size.

1.5.4 Quartiles, deciles and percentiles

The median divides the set of data in two equal parts. in a similar way this set could be divided in four equal parts called **quartiles;** Q_1, Q_2, Q_3 respectively. Ten equal parts called **deciles** denoted by D_1, D_2,... D_9 or one hundred equal parts denoted by f_1, f_2, .., f_{99} (percentiles) are also possible.

Note: The median is thus equal to Q_2 , D_2 and P_{50}

To calculate P_k, k = 1, 2,..., 99 we generalize formula (1.6)

$$P_k = L + \left(\frac{\frac{kn}{100} - (\sum f)_1}{f_{p_k}} \right) C$$

1.5.5 The standard deviation

For grouped data the standard deviation is calculated in the same way as for ungrouped data. The square of each deviation of a class midpoint from the mean multiplied by the frequency of that specific class has to be added, divided by (n - 1) and the square root of the result taken.:

$$S = \sqrt{\frac{f_1(X_1 - \bar{X})^2 + f_2(X_2 - \bar{X})^2 + \ \ldots \ + f_k(X_k - \bar{X})^2}{n-1}}$$

And if we use the Σ notation:

$$S = \sqrt{\frac{\sum f_i(X_i - \bar{X})^2}{n-1}}$$

Example 2

Using the data in Table 1.2, we will determine the following:

The arithmetic mean:

$$\bar{X} = \frac{\sum f_i X_i}{n}$$
$$= \frac{2\times 54{,}95 + 8\times 56{,}95 + \ldots + 1\times 66{,}95}{60}$$
$$= 60{,}45 \ ppm$$

The mode:

$$Mode = L + \left(\frac{D_1}{D_1 + D_2} \right) c$$
$$= 59{,}95 + \left(\frac{21 - 14}{(21 - 14) + (21 - 10)} \right) \times 2$$
$$= 60{,}73 \ ppm$$

Interpretation: The copper concentration which occurs most often is 60,73 ppm.

The median:

$$Me = L + \left(\frac{\frac{n}{2} - (\sum f)_1}{f_{me}} \right) c$$
$$= 59{,}95 + \left(\frac{30 - 24}{21} \right) \times 2$$
$$= 60{,}52 \ ppm$$

Interpretation: 30 (50%) copper determinations were below 60,52 ppm.

The 25th percentile:

$$P_{25} = 57{,}95 + \left(\frac{\frac{25\times 60}{100} - 10}{14} \right) \times 2$$
$$= 58{,}66 \ ppm$$

Interpretation: 15 (25%) of the copper determinations were below 58,66 ppm.

The 75th percentile:

$$p_{75}=59{,}95+\left(\frac{\frac{75\times60}{100}-24}{21}\right)\times2$$
$$=61{,}95\ ppm$$

Interpretation: 45 (75%) of the determinations were below 61,95 ppm.

The range:

Range = Maximum - Minimum
= 66,3 - 54,4
= 11,9 ppm

Interpretation: The data stretches over an interval of 11,9 ppm.

The standard deviation:

$$S=\sqrt{\frac{\sum f_i(X_i-\overline{X})^2}{n-1}}$$
$$=\sqrt{\frac{2\times(54{,}95-60{,}45)^2+\ldots+1\times(66{,}95-60{,}45)^2}{60-1}}$$
$$=2{,}54\ ppm$$

1.6 Graphs

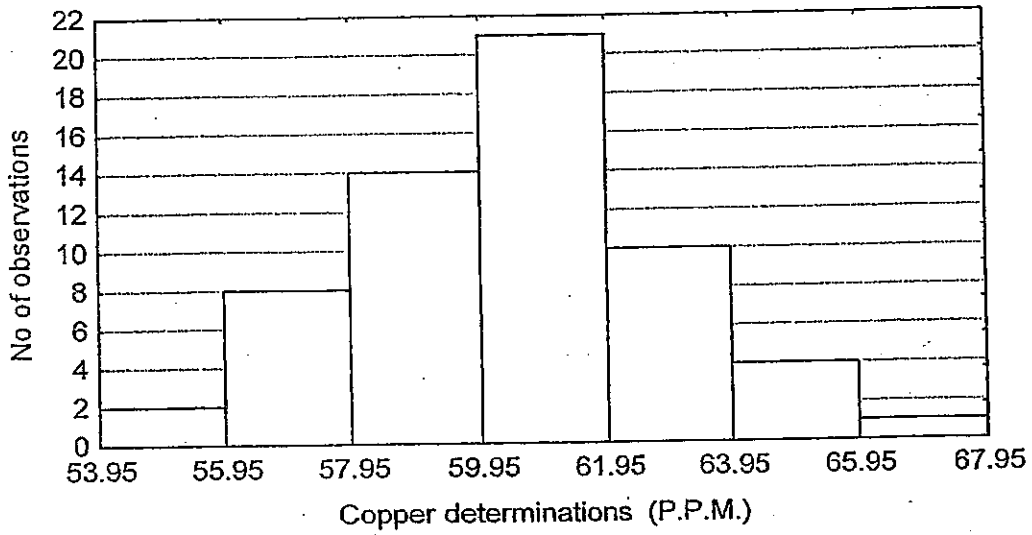
1.6.1 The histogram

A histogram consists of a set of touching bars, each having its:

(a) basis on a horizontal axis were the first bar starts at the lowest *class boundary* in the frequency table and ends at the upper boundary of the first class. The subsequent bars all end at the upper boundaries of their classes.

(b) Area proportional to its class frequency. If the class intervals all have equal width, the heights of the rectangles are proportional to the class frequencies and it is then customary to take the height numerically equal to the class frequencies. If class intervals do not have equal width, the heights must be adjusted.

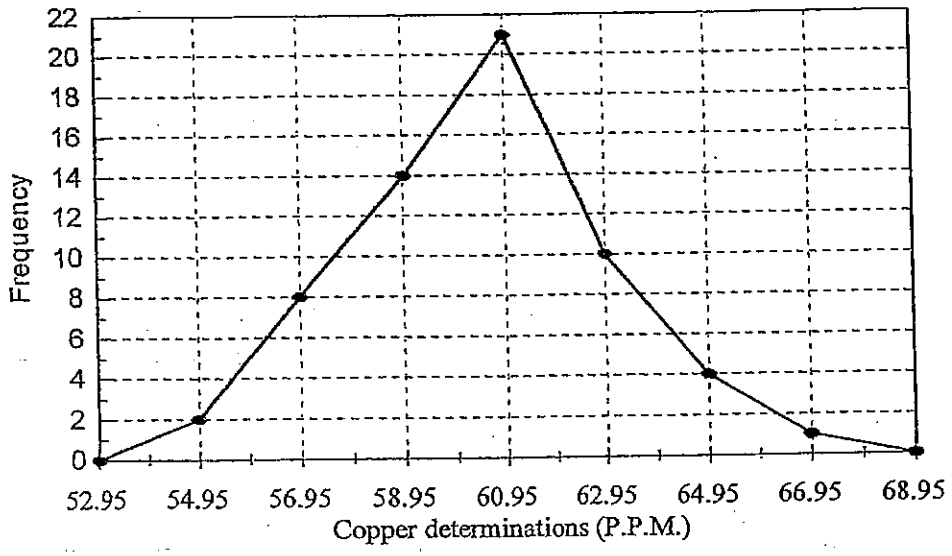
HISTOGRAM



1.6.2 The frequency polygon

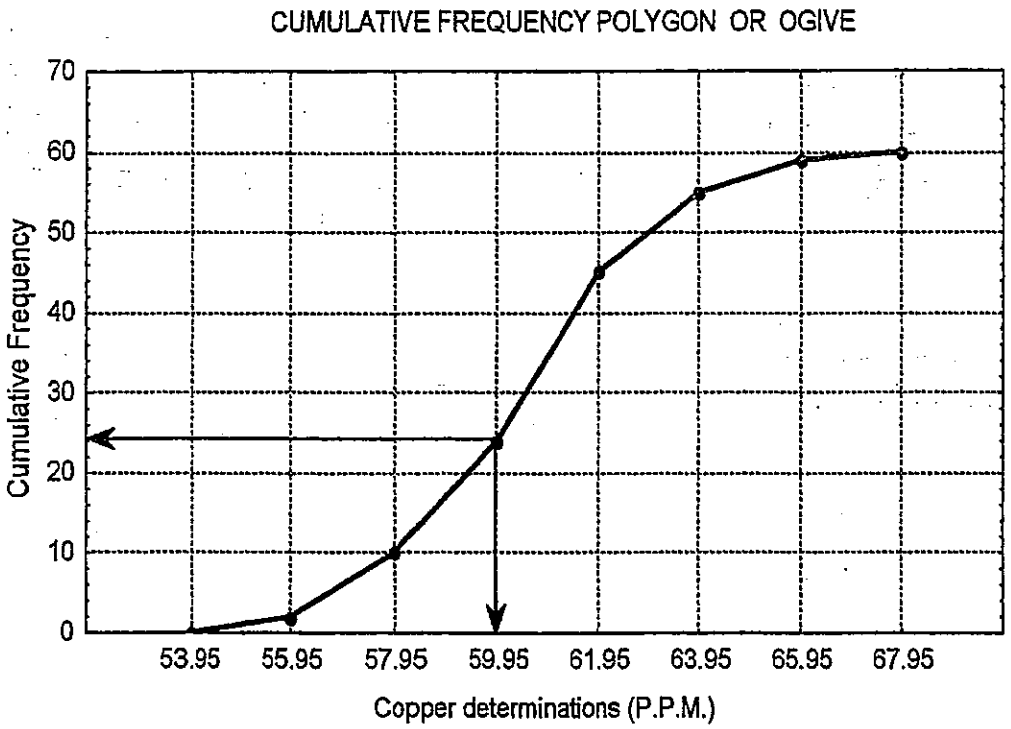
A frequency polygon is a line graph of class frequency plotted against class midpoint. It can be obtained by connecting the tops of the rectangles in the histogram. This graph is used when the underlying data is continuous (like copper determinations) instead of categorical, the line can than be used for interpolation. Normally, a frequency polygon curve is closed. To achieve this, we subtract a class width from the lowest class mark, add a class width to the highest class mark and plot these points on the X-axis so that we can close the curve.

FREQUENCY POLYGON



1.6.3 Ogives

A graph showing the cumulative frequency which is less than any upper class boundary plotted against this upper class boundary, is called a cumulative frequency polygon or ogive.



By using this polygon, we can see that aproximately 24 copper determination readings were below 60 ppm.

1.6.4 Relative histogram, frequency polygon and ogive

The same definitions as in 1.6.1, 1.6.2 and 1.6.3 apply, except that the frequencies are expressed as a percentage of the total.

Remark:

It is important to note that the total area of the rectangles of the histogram and the area under the frequency polygon equal 100% in this case.

1.7 Relation between mean, median and mode

If the frequency polygon is perfectly symmetrical, the mean, median and mode coincide. The following figures show the relative positions of these statistics when the frequency curves have positive and negative skewness.

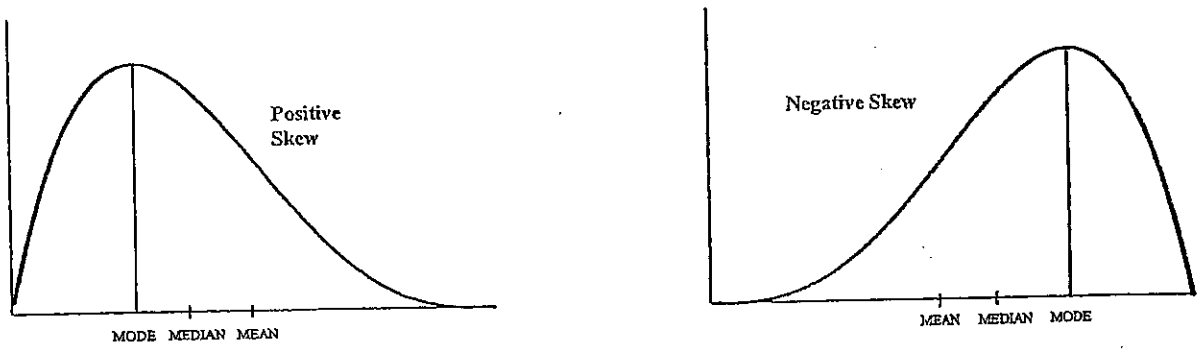


Figure 5

Summary Measures

1.8.1 Interquartile range

A useful measure of variability is the interquartile range. It represents the middle 50 % of the observations. It the difference between the third quartile and the first quartile.

It will be noted that $P_{25} = Q_1$ and $P_{75} = Q_3$. The quantity $Q_3 - Q_1$ is called the Interquartile Range and gives a measure of the middle 50% of the data. In this case:

$Q_3 - Q_1 = 66,95 - 58,66$
$= 3,29$ ppm

1.8.2 Coefficient of variation

The standard deviation on it self does not indicate a **relative** measure of variability. For this the coefficient of variation can be used which is defined by $V = \frac{S}{\bar{X}}$.

For this example:

$$V = \frac{S}{\bar{X}}$$
$$= \frac{2,54}{60,45} = 0,042$$

Interpretation: The variation relative to the mean is 0,042 (4,2%).

1.8.3 Coefficient of skewness

Two useful measures of skewness are given by Pearson's first and second coefficients of skewness, respectively: $S_K = \frac{\bar{X} - Mo}{S}$ and $S_K = \frac{3(\bar{X} - Me)}{S}$.

Pearson's second coefficient of skewness for this example equals:

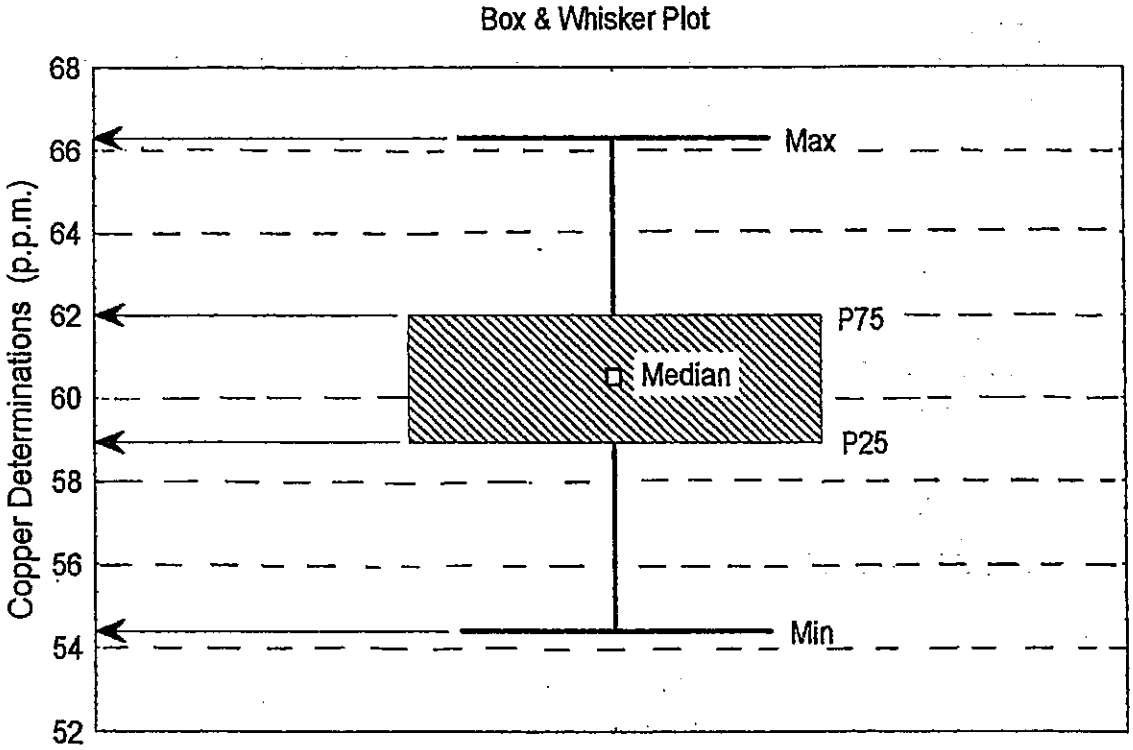
$$S_K = \frac{3(\bar{X} - Me)}{S}$$

$$= \frac{3 \times (60{,}45 - 60{,}52)}{2{,}54}$$

$$= -0{,}083$$

1.8.4 The box and whisker plot

A graphical representation of the range, the quartiles and their relative positions with regard to each other can be provided by a *box and whisker plot*. The plot begins with a line segment starting at the minimum observed level and ends with a line segment ending at the maximum value. The box begins at the first quartile and ends at the third, and is divided at the median. The overall length of the plot gives the range. The length of the box provides the interquartile range. For the example above, the box and whisker plot looks as follows:



Interpretation: The distribution shows a slight negative skew. (The histogram and polygon confirm this finding). Also, the "box" part of the graph shows that the interquartile range is quite small compared to the range of the data, this means that 50% of the data is concentrated on a relatively small interval.

NB: When the raw data is no longer available the lower boundary of the first class and the upper boundary of the last class are taken as the box plot's minimum and maximum respectively.

Example 2

The following sample data applies to the average yield of a final product (in grams) from each litre of chemical feedstock:

Yield (grams)	Class midpoint	Number of batches
25,5 to under 26,5	26	2
26,5 to under 27,5	27	3
27,5 to under 28,5	28	4
28,5 to under 29,5	29	5
29,5 to under 30,5	30	7
30,5 to under 31,5	31	11
31,5 to under 32,5	32	14
32,5 to under 33,5	33	3
33,5 to under 34,5	34	1

Remarks:

(1) In this frequency table, the class limits and class borders coincide: for all calculations we treat class limits and borders as the same.

(2) Instead of writing a class like "27,5 to under 28,5", a more mathematically elegant way is to designate this class as "[27,5 - 28,5)", where the "[" bracket indicates that the 27,5 is **included**, and the ")" bracket indicates that the 28,5 is **excluded** in the class.

a) **The mean:**

$$\bar{X} = \frac{\sum f_i X_i}{n}$$

$$= \frac{2\times 26 + 3\times 27 + \ldots + 34\times 1}{50}$$

$$= 30{,}44 \ g$$

b) **The median:**

$$Me = L + \left[\frac{\frac{n}{2} - (\sum f)_1}{f_{Me}}\right] C$$

$$= 30{,}5 + \left[\frac{25 - 21}{11}\right] \times 1$$

$$= 30{,}86 \ g$$

c) **The mode:**

$$Mo = L + \left[\frac{D_1}{D_1 + D_2}\right] C$$

$$= 31{,}5 + \left[\frac{3}{3+9}\right] \times 1$$

$$= 31{,}75 \ g$$

d) **The variance:**

$$S^2 = \frac{\sum f_i (X_i - \bar{X})^2}{n-1}$$

$$= \frac{2 \times (26 - 30{,}44)^2 + 3 \times (27 - 30{,}44)^2 + \ldots + 1 \times (34 - 30{,}44)^2}{50 - 1}$$

$$= 3{,}61 \ g^2$$

e) **Coefficient of variation:**

$$V = \frac{S}{\bar{X}}$$

$$= \frac{1{,}90}{30{,}44}$$

$$= 0{,}062 \ (6{,}2\%)$$

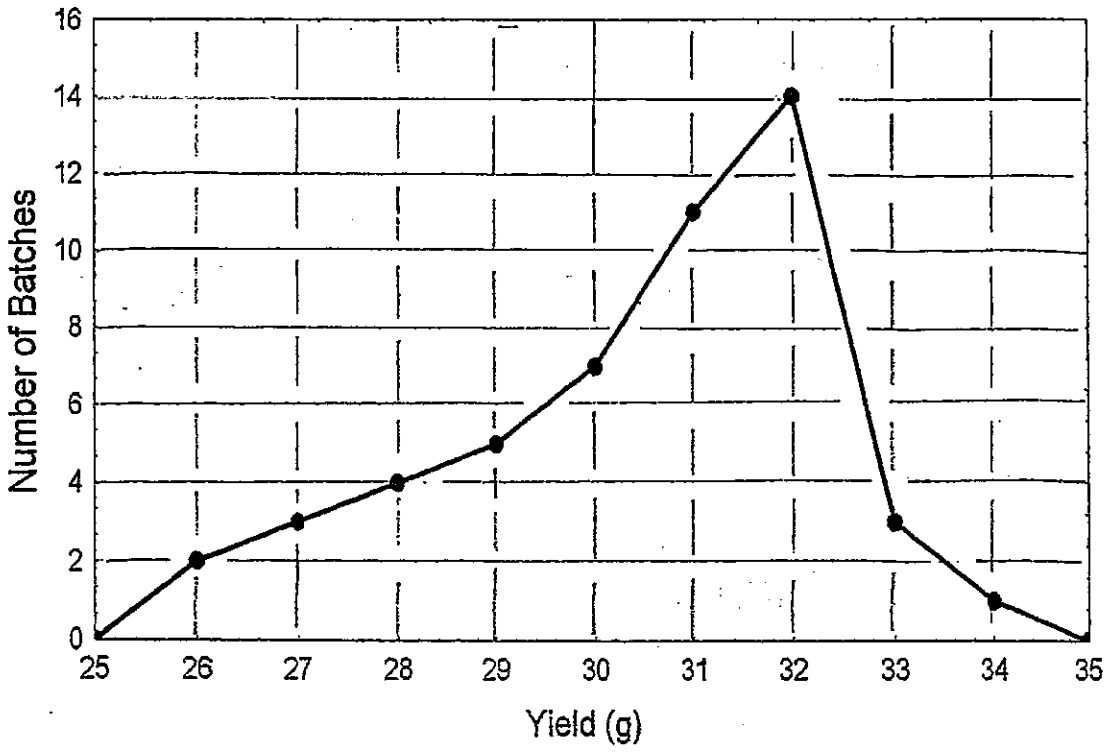
f) **Coefficient of skewness:**

$$S_K = \frac{3(\bar{X} - Me)}{S}$$

$$= \frac{3 \times (30{,}44 - 30{,}86)}{1{,}90}$$

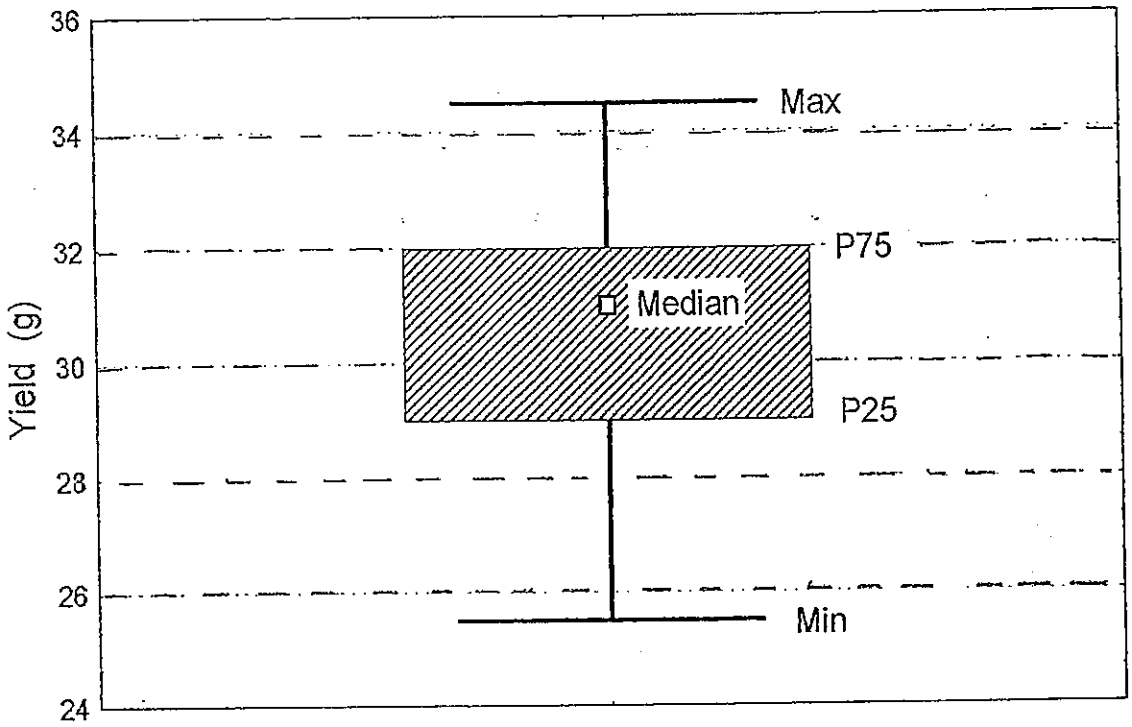
$$= -0{,}66$$

g) **The frequency polygon:**



It is left to the student as an exercise to find the interquartile range. [ans. 3,078].

h) **The box and whisker plot:**



Exercises

1. Use the ungrouped data in table 1.1 to calculate the
   (a) mean [ans. 60,37]
   (b) median [ans. 60,50]
   (c) standard deviation [ans. 2,54].

   Compare your results with the statistics obtained from the calculations based on the grouped data in table 1.2. Comment on the differences.

2. Produce a frequency table using the data in table 1.1 using a class width of 1 ppm starting with a lower class limit of 54 ppm. Use this table and calculate the
   (a) mean [ans. 60,37 ppm]
   (b) standard deviation [ans. 2,53 ppm].

   Compare your results with those obtained from table 1.6. Comment on any differences.

3. Use table 1.2 to draw a relative ogive. What percentage of the determinations are less than 60,95 ppm?

4. A series of unconfined compression tests was carried out on an undisturbed sample taken from a soil stratum. The values of the measured strengths are given in the table below:

Strength (kN/m^2)	Frequency
24,0 - 26,2	3
26,3 - 28,5	6
28,6 - 30,8	8
30,9 - 33,1	9
33,2 - 35,4	9
35,5 - 37,7	6
37,8 - 40,0	4

Compute the following sample statistics:

(a) The mean strength.
(b) The median strength.
(c) The interquartile range.
(d) The standard deviation.

Construct the following graphical representations of the above frequency table:

(e) A relative polygon.
(f) A box plot.
(g) Give the interpretations of the answers given in (b) and (c) above in the context of the problem.
(h) Compute the coefficient of skewness and explain the outcome in terms of your polygon.

5(a). Table 1.3 shows some data on laboratory measurements on galvanized iron coat weights. Use 10 classes, starting with 3,90 - 4,04, 4,05 - 4,19 etc. Produce a frequency table which summarizes the data in Table 1.3. This table must contain the following columns:

(i) class midpoints (X_i)
(ii) frequency of occurrence of each class (f_i)
(iii) relative frequency column
(iv) cumulative frequency column.

Table 1.3 Weights of coating of 100 sheets of galvanized iron [units grams per square cm.(x0,01)]

4,47	4,35	4,58	3,95	4,22	4,77	4,52	4,95	5,10	4,67	4,51	4,48	4,80	4,39	5,39
4,37	4,72	4,24	4,19	4,62	4,28	4,07	4,56	4,89	5,00	4,66	4,13	4,91	4,73	4,01
4,03	3,98	4,53	4,92	4,57	5,12	4,89	4,58	4,36	4,59	4,44	4,59	4,69	4,35	4,46
4,39	4,78	4,61	4,52	3,99	4,23	4,37	4,48	4,93	5,20	4,49	4,62	4,67	4,67	4,76
4,54	4,00	4,58	4,93	4,55	4,21	4,36	4,50	4,97	4,55					

(b) Use your frequency distribution to calculate the following:
   (i) the mean [ans.4,555] g/cm^2
   (ii) the median [ans.4,553] g/cm^2
   (iii) the mode [ans.4,545] g/cm^2
   (iv) the standard deviation [ans.0,310] g/cm^2.

(c) Interpret the value of the median for this sample.

(d) Find the coefficient of skewness for the frequency distribution.

(e) Draw a frequency polygon and a box plot.

(f) If it can be assumed that the data is approximately normally distributed, find the two values between which approximately 68% of the data will be. [ans. 4,245 g/cm^2; 4,865] g/cm^2

(g) In the light of your answers on (d) and (e) above, do you think that the calculation in (f) can be justified?

6. By itself, the standard deviation does not convey the relative degree of variability. For example, consider the mean and standard deviation of the hand length and reach of men:

Hand length	**Reach**
μ = 18,90 cm	μ = 82,12 cm
σ = 0,05 cm	σ = 0,19 cm

The respective standard deviations suggest that *reach* has about 4 times the variability of *hand length*. Find the relative variability by calculating the coefficient of variation. Comment on your answers.