5.1 BASIC CONCEPTS OF STATISTICS

Consider the following numerical data[1]:

10, 20, 20, 20, 30, 30, 40, 50, 70, 70, 80

The total number of entries is **n=11**.

In order to describe these data we use

- 3 measures of central tendency
- 3 measures of spread

The first three measures indicate a representative central value which best describes the data, while the second three measures indicate if our data are very close or dispersed to each other.

♦ MEASURES OF CENTRAL TENDENCY (The 3 M's)

A) MEAN = The sum of all values divided by n.

Here

$$\textbf{mean} = \frac{10+20+20+20+30+30+40+50+70+70+80}{11} = 40$$

B) MODE = the most frequent value

Here

mode = 20

C) MEDIAN = The value in the middle

(provided they have been placed in ascending order).

Here, it is the sixth number in the list

median = 30

---

[1] This set of values is either a **population** or a **sample.**
By a **population** we mean a complete set of values in some measurement. If the population is very large we usually consider a small **sample** of the population.

NOTICE

- For the data 10, 20, 30

  $$\text{Median} = 20$$

  For the data 10, 20, 30, 40

  $$\text{Median} = 25$$

  That is, for an **even** number of data,

  median = the mean of the two middle values

- The median is not the $\frac{n}{2}$-th entry as one would possibly expect.

  the median is the $\frac{n+1}{2}$-th entry.

  For example,

  if n=11, $\frac{n+1}{2}=6$, thus the median is the 6^{th} entry. See the example above;

  if n=10, $\frac{n+1}{2}=5.5$, thus the median is the mean of the 5^{th} and 6^{th} entries; for the 10 entries

  $$10, 20, 30, 40, 50, 60, 70, 80, 90, 100$$

  the median is the mean of 50 and 60. Hence **median = 55**

  The median is also denoted by $\mathbf{Q_2}$ (the index 2 will be clarified soon)

- The **mean** is denoted by μ (or by $\bar{x}$). In fact, we use

  the Greek letter μ for the whole population.
  the Latin letter $\bar{x}$ for a sample of the population.

  If our data are denoted by $x_1, x_2, \ldots, x_n$, the mean is given by

  $$\mu = \frac{x_1 + x_2 + x_3 + \cdots}{n}$$

  or otherwise

  $$\mu = \frac{\sum x_i}{n}$$

EXAMPLE 1

Find

a) the integers $a \le b \le c$, given that mean=4, mode=5, median=5.

The median implies that b=5. The mode implies that also c=5.

Then $\frac{a+5+5}{3}=4 \Leftrightarrow a+10=12 \Leftrightarrow a=2$

Therefore, the numbers are 2,5,5.

b) the integers $a \le b \le c \le d$, given that mean=5, mode=7, median=6.

The median implies that either b=c=6 or (b=5 and c=7)

Since the mode is 7 we obtain b=5 and c=d=7.

Then $\frac{a+5+7+7}{4}=5 \Leftrightarrow a+19=20 \Leftrightarrow a=1$

Therefore, the numbers are 1,5,7,7.

♦ MEASURES OF SPREAD

We use the same set of data

10, 20, 20, 20, 30, 30, 40, 50, 70, 70, 80

A) STANDARD DEVIATION

The **standard deviation** is perhaps the most "reliable" measure for spread, as it takes all data into consideration. It measures how far the entries from the mean are. It can be found by using the GDC (directions will be given later on).

The **standard deviation** is denoted[2] either by σ or by s_n.

For our example the GDC gives $\sigma = 22.96$.

---

[2] In fact,

the Greek letter σ is used for the whole population;

the Latin letter s_n is used for a sample of the population

B) RANGE = (maximum value) – (minimum value)

Here

$$\textbf{range} = 80-10 = \textbf{70}$$

C) INTERQUARTILE RANGE = IQR = $Q_3 - Q_1$

where

Q_1 = **LOWER QUARTILE** = *the median of the values before* Q_2

Q_3 = **UPPER QUARTILE** = *the median of the values after* Q_2

Here, before the median Q_2=**30**, *we have 5 numbers, hence*

Q_1=**20** (*this is the* 3rd *entry*)

Also,

Q_3=**70** (*it is the* 3rd *entry from the end*)

Therefore,

$$\textbf{IQR} = 70-20 = \textbf{50}$$

As the estimation of the values Q_1, Q_2, Q_3 *is quite tricky, let us see some extra cases in the following example.*

EXAMPLE 2 *Remember that*

- *for the value of the median* Q_2 *we consider the* $\frac{n+1}{2}$*th entry.*
- *for the values of Q1 and Q3 we consider only the entries* **before** *and the entries* **after** *the median respectively.*

a) *For n=7 entries:* **10, 20, 30, 40, 50, 60, 70**

The median is Q_2=40 (*the* 4th *entry*). *Hence* Q_1=20, Q_3=60.

b) *For n=8 entries:* **10, 20, 30, 40, 50, 60, 70, 80**

The median is Q_2=45 (*the* 4.5th *entry*). *Hence* Q_1=25, Q_3=65.

c) *For n=9 entries:* **10, 20, 30, 40, 50, 60, 70, 80, 90**

The median is Q_2=50 (*the* 5th *entry*). *Hence* Q_1=25, Q_3=75.

d) *For n=10 entries:* **10, 20, 30, 40, 50, 60, 70, 80, 90, 100**

Then Q_2=55 (*the* 5.5th *entry*). *Hence* Q_1=30, Q_3=80.

NOTICE

The square of the standard deviation is called **variance**. That is

$$\textbf{variance} = \sigma^2 \quad \textbf{or} \quad s_n^{\,2}$$

For our example, $\sigma^2 = 22.96^2 = 527.27$

♦ USE OF GDC

We can use the GDC to easily obtain all these measures.

For Casio CFX we select

- MENU
- STAT
- Complete List 1 with values of x (our data)
- CALC
- (1VAR): We obtain all the statistics.

Notice that

The standard deviation in the GDC is denoted by σ_x

The variance is not given; it is simply the square of σ_x

♦ BOX AND WHISKER PLOT

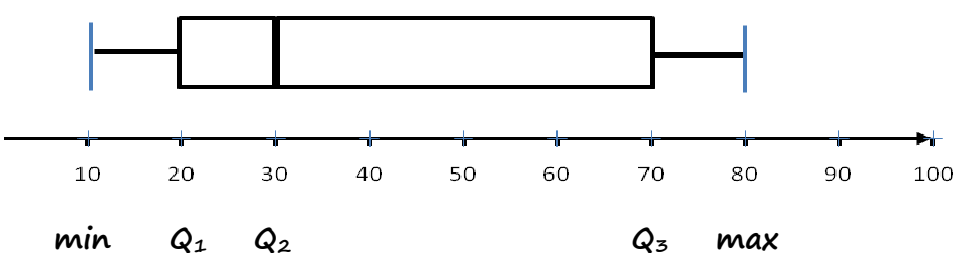
Consider again the initial example

10, 20, 20, 20, 30, 30, 40, 50, 70, 70, 80

In an appropriate horizontal scale we mark 5 figures:

$$\textbf{min},\ Q_1,\ Q_2,\ Q_3,\ \textbf{max}$$

in the following way:



This diagram is helpful, particularly when we have a large number of entries. It shows the "density" of data within the whole range. In fact, the box plot splits the whole range of data in 4 intervals. Generally speaking, each interval contains 25% of the entries. Thus the following conclusions can be drawn:

The lowest 25% is below Q_1 The upper 25% is above Q_3
The lowest 50% is below Q_2 The upper 50% is above Q_2
The middle 50% is between Q_1 and Q_2

♦ MORE DETAILS

1) Percentiles

The values Q_1, Q_2, Q_3 are also called

Q_1 : 25^{th}-percentile
Q_2 : 50^{th}-percentile
Q_3 : 75^{th}-percentile

Other percentiles may also be defined in a similar way; we will give further examples in the next paragraph.

2) **Outliers**

Very extreme values in a set of data (that is very small or very large) may give a false impression for out data. They are known as outliers. We agree that

an **outlier** is any value

$$\text{below } Q_1 - 1.5\times IQR$$
$$\text{or above } Q_3 + 1.5\times IQR,$$

Such a value is viewed as being too far from the central values to be reasonable. In our example,

$$Q_1 - 1.5\times IQR = 20 - 1.5\times 50 = -55$$

$$Q_3 + 1.5\times IQR = 70 + 1.5\times 50 = 145$$

i.e. there are no outliers.

♦ MORE ON VARIANCE – STANDARD DEVIATION **(only for HL)**

If our data are $x_1, x_2, \ldots, x_n$

the **variance** is given by $\sigma^2 = \dfrac{\sum (x_i - \mu)^2}{n}$

the **standard deviation** is given by $\sigma = \sqrt{\dfrac{\sum (x_i - \mu)^2}{n}}$

For our example,

$$\textbf{variance} = \frac{(10-40)^2 + (20-40)^2 + (20-40)^2 + \cdots + (80-40)^2}{11} = \textbf{527.27}$$

$$\textbf{standard deviation} = \sqrt{527.27} = \textbf{22.96}$$

The variance measures the spread of the data as in fact we find

- the distance of each entry from the mean
- the squares of these distances
- the average of all these square distances

An alternative and more practical formula for the **variance** is given by

$$\sigma^2 = \frac{\sum x_i^{\,2}}{n} - \bar{x}^2$$

For our example, we have $\bar{x}=40$ and

$$\frac{\sum x_i^{\,2}}{n} = \frac{10^2 + 20^2 + 20^2 + \cdots + 80^2}{11} = \frac{23400}{11} = 2127.27$$

Hence

$$\sigma^2 = 2127.27 - 40^2 = \textbf{527.27}$$

Proof of the alternative formula

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum (x_i^{\,2} - 2x_i\bar{x} + \bar{x}^2)}{n}$$

$$= \frac{\sum x_i^{\,2}}{n} - 2\bar{x}\frac{\sum x_i}{n} + \frac{\sum \bar{x}^2}{n} = \frac{\sum x_i^{\,2}}{n} - 2\bar{x}\bar{x} + \frac{n\bar{x}^2}{n}$$

$$= \frac{\sum x_i^{\,2}}{n} - 2\bar{x}^2 + \bar{x}^2 = \frac{\sum x_i^{\,2}}{n} - \bar{x}^2$$

5.2 FREQUENCY TABLES – GROUPED DATA

Consider again the numerical data:

10, 20, 20, 20, 30, 30, 40, 50, 70, 70, 80

The total number of entries is **n=11**.

An alternative way of presentation is the **frequency table**:

Data x	Frequency f
10	1
20	3
30	2
40	1
50	1
70	2
80	1
	n=11

Let us study again the basic measures for these data.

♦ MEASURES OF CENTRAL TENDENCY (The 3 M's)

A) MEAN = The sum of all values divided by n.

The MEAN is given by

$$\textbf{mean} = \frac{1\times10+3\times20+2\times30+1\times40+1\times50+2\times70+1\times80}{11} = 40$$

In general, given that f_i is the frequency of the entry x_i, the formula is

$$\mu=\frac{f_1x_1+f_2x_2+f_3x_3+\cdots}{n} \qquad \text{or otherwise} \qquad \mu=\frac{\sum f_ix_i}{n}$$

B) MODE = the most frequent value

It is very obvious now. The entry x of the highest frequency is

mode = 20

C) MEDIAN = The value in the middle

It is still the entry in position $\frac{n+1}{2}$, that is the 6^{th} entry.

We can easily see that this is 30.

It helps here to add an extra column in the table above with the so-called **cumulative frequencies:**

Data x	Frequency f	Cumulative frequency (c.f.)
10	1	1
20	3	4
30	2	6
40	1	7
50	1	8
70	2	10
80	1	11
	n=11	

It simply gives the total number of entries up to each row. For example, the total number of entries up to 20 is 1+3=4.
The MEDIAN, i.e. the 6^{th} entry, is 30.

♦ MEASURES OF SPREAD

A) STANDARD DEVIATION

Again, it can be directly obtained by the GDC.

For our example the GDC gives **σ = 22.96.**

Thus the **variance** is **σ^2 = 527.27**

B) RANGE = (maximum value of x) – (minimum value of x)

It is very obvious here

$$\textbf{range} = 80-10 = \textbf{70}$$

C) INTERQUARTILE RANGE = IQR = $Q_3 - Q_1$

The cumulative frequency table helps here as well.

The median Q_2=**30** is in the 6th position.

Thus, before the median we have 5 entries. Since $\frac{n+1}{2}=3$,

Q_1=20 (this is the 3rd entry)

and

Q_3=70 (this is the 3rd entry from the end)

Therefore,

$$\textbf{IQR} = 70-20 = \textbf{50}$$

♦ USE OF GDC

We can use the GDC to easily obtain all these measures.

For Casio CFX we select

- MENU
- STAT
- Complete List 1 with values of x (our data)
  List 2 with frequencies
- CALC
- SET: we check the first two lines
  The first line is OK. (1Var XList :List1)
  For the second line (1Var Freq :----), select between
  F1: enter 1, if there are no frequencies
  F2: enter List 2 to consider frequencies
- Go back (EXIT)
- 1VAR: We obtain all the statistics.

Check the value of **n** first (number of entries), to ensure that all data have been considered.

NOTICE (for the GDC)

- The variance is not given; it is simply the square of σ_x
- Since the GDC gives **minX**,**Q1**,**Med**,**Q3**,**maxX** remember that

  Range = **maxX** − **minx** Interquartile Range = **Q3** − **Q1**

  The box and whisker plot uses exactly those 5 measures
- Extra information given:

  Σx : the sum of all entries, i.e. $x_1+x_2+x_3+\ldots$

  Σx²: the sum of the squares, i.e. $x_1^2+x_2^2+x_3^2+\ldots$

  s_x : it is known as unbiased st. deviation (not in the syllabus!)

♦ GROUPED DATA

Suppose that 100 students took an exam and obtained scores from 1 to 60 (full marks), according to the following table:

Score (x)	Midpoint (for x)	No of students (frequency f)	Cumulative frequency (cf)
$0<x\leq10$	5	8	8
$10<x\leq20$	15	12	20
$20<x\leq30$	25	10	30
$30<x\leq40$	35	25	55
$40<x\leq50$	45	35	90
$50<x\leq60$	55	10	100
		n=100	

i.e. 8 students obtained scores from 1 up to 10, and so on.

- The mean and the standard deviation are still calculated as in a usual frequency table, but now $x_1,x_2,x_3,\ldots$ are the midpoints of the intervals.

  For example,

$$\mu=\frac{8\times5+12\times15+10\times25+25\times35+35\times45+10\times55}{100}=34.7$$