Q28
p. 27

Q28, Page 27

This problem presents data on a group of 33 elementary school students. The researchers counted the number of times the students talked to themselves. The number of times kids talk to themselves may be an indicator of future of future academic progress. The data they collected is as follows;

The Data

82	96	99	102	103	106	107	108	108	108
109	110	110	111	113	113	113	113	115	115
119	121	122	122	127	132	136	140	146	
108	108								
118	118								

Frequency Table

Let us construct a frequency table for this data, using the procedure we discussed in class.

Step 1. Range of the data.

Range

Range = high value − low value

= 146 − 82

= 64.

64

Step 2. Number of intervals (classes) for our frequency table.

As a rule of thumb any value between 5 and 20 is OK.

Less than 5 means you have over-simplified your data. More than 20 means you have over-elaborated.

So lets pick, hmm, 8 ?

So we need to determine the width of each interval, lets call it class width.

# of classes

$$\text{Class width} = \frac{\text{Range}}{\text{Num. of classes}}$$

$$= \frac{64}{8}$$

$$= 8.$$

Now. It is MUCH easier to work with class widths that are multiples of say 5, 10, 20 etc

class width 10

So lets pick, say, hmm, 10

# of classes 7

class width

Note that by picking 10 we will end up with $\frac{64}{10} \approx 7$ classes.

which is OK.

Step 3. Construct the frequency table

For our intervals we shall use the notation $[a, b)$ where

a is the lower limit,

b is the upper limit, but it is not included in the interval

Class notation

So e.g. $[80, 90)$ would mean a value of 80 will fall in this interval but a value of 90 will not, and will be tallied in the interval $[90, 100)$ $[a, b)$

If you prefer you can do $(a, b]$ and the opposite of what we just did will be right.

There is no hard rule on this. Different authors use different notation. But whatever you choose, stick with it and be consistent. DO not mix and match.

First interval

Now our first class must start at a value that encompasses our low value.

Last interval

Our last interval's upper limit ~~sho~~ must enclose the high value of the data.

So we shall have

Interval Structure

$[80, 90)$

$[90, 100)$

$\vdots$

$[140, 150)$

So for each class (interval) tally the number of times a data value falls in that class.

Then right out the frequency.

Frequency Table also called Frequency Distribution

Class	Tally	Frequency (f)
[80, 90)	/	1
[90, 100)	//	2
[100, 110)	///// ///	10
[110, 120)	///// ///// //	12
[120, 130)	////	4
[130, 140)	//	2
[140, 150)	//	2
		$\sum f = 33$

We shall now expand the frequency table.
We shall add the class mid point (also called the class mark).

Another way of expressing frequency is the Cumulative Frequency.
The cumulative frequency of a class is the sum of the frequencies up to that class.

Another form of frequency is the relative frequency. Relative frequency is the frequency of the interval divided by the total frequency.

The Cumulative Relative Frequency ~~is the~~ of an interval is the sum of the relative frequencies up to that interval.

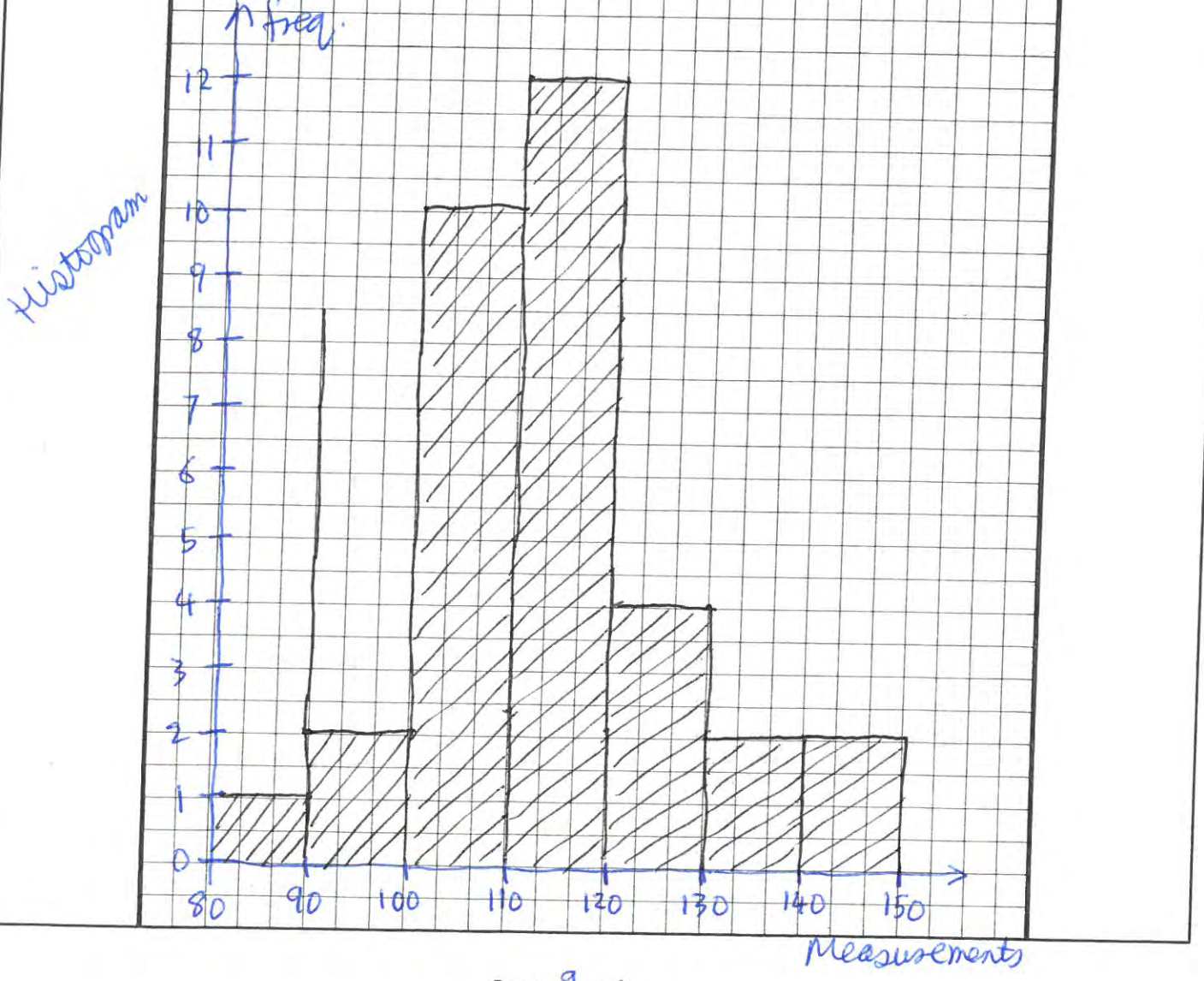
Class	Tally	Freq	Cf	Rel. Freq	Cum Rel Freq	Class Mark
[80, 90)	/	1	1	0.03	0.03	85
[90, 100)	//	2	3	0.06	0.09	95
[100, 110)	~~////~~ ~~////~~	10	13	0.30	0.39	105
[110, 120)	~~////~~ ~~////~~ //	12	25	0.36	0.75	115
[120, 130)	////	4	29	0.12	0.87	125
[130, 140)	//	2	31	0.06	0.93	135
[140, 150)	//	2	33	0.06	1.00	145
		$\Sigma f = 33$		1.00		

As we shall see, plotting the various frequencies can reveal important properties and information regarding the data.

Graphs of Frequency Distributions

Graphs of frequency distributions are graphical representations of the frequency table.

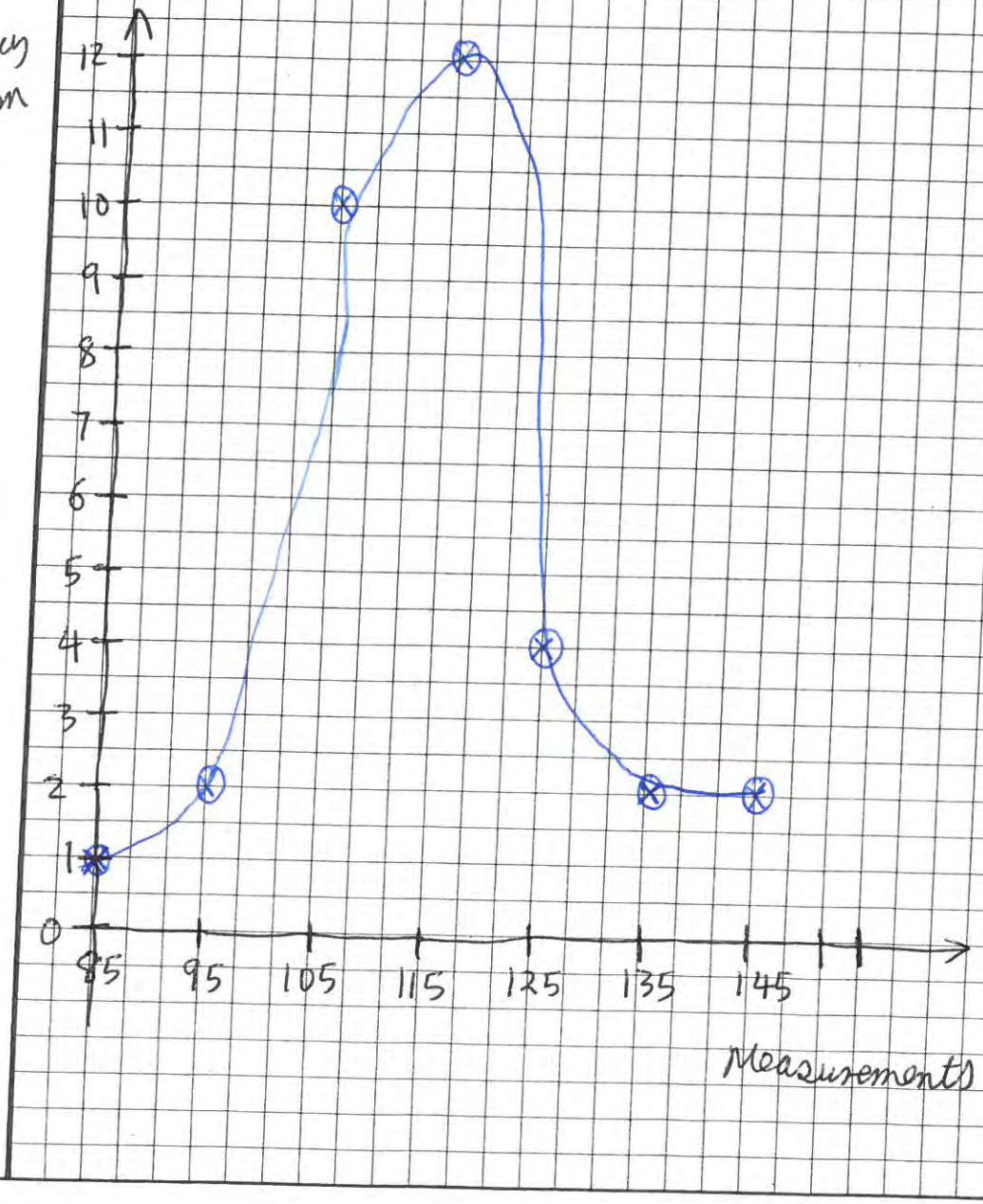
The histogram is a plot of frequency versus class intervals as a bar graph.



Histogram

The Frequency Polygon is a smooth curve of frequency versus class mark (class mid point)

Frequency Polygon



Cumulative Frequency Curve.

Also called Ogive (oh-jive)

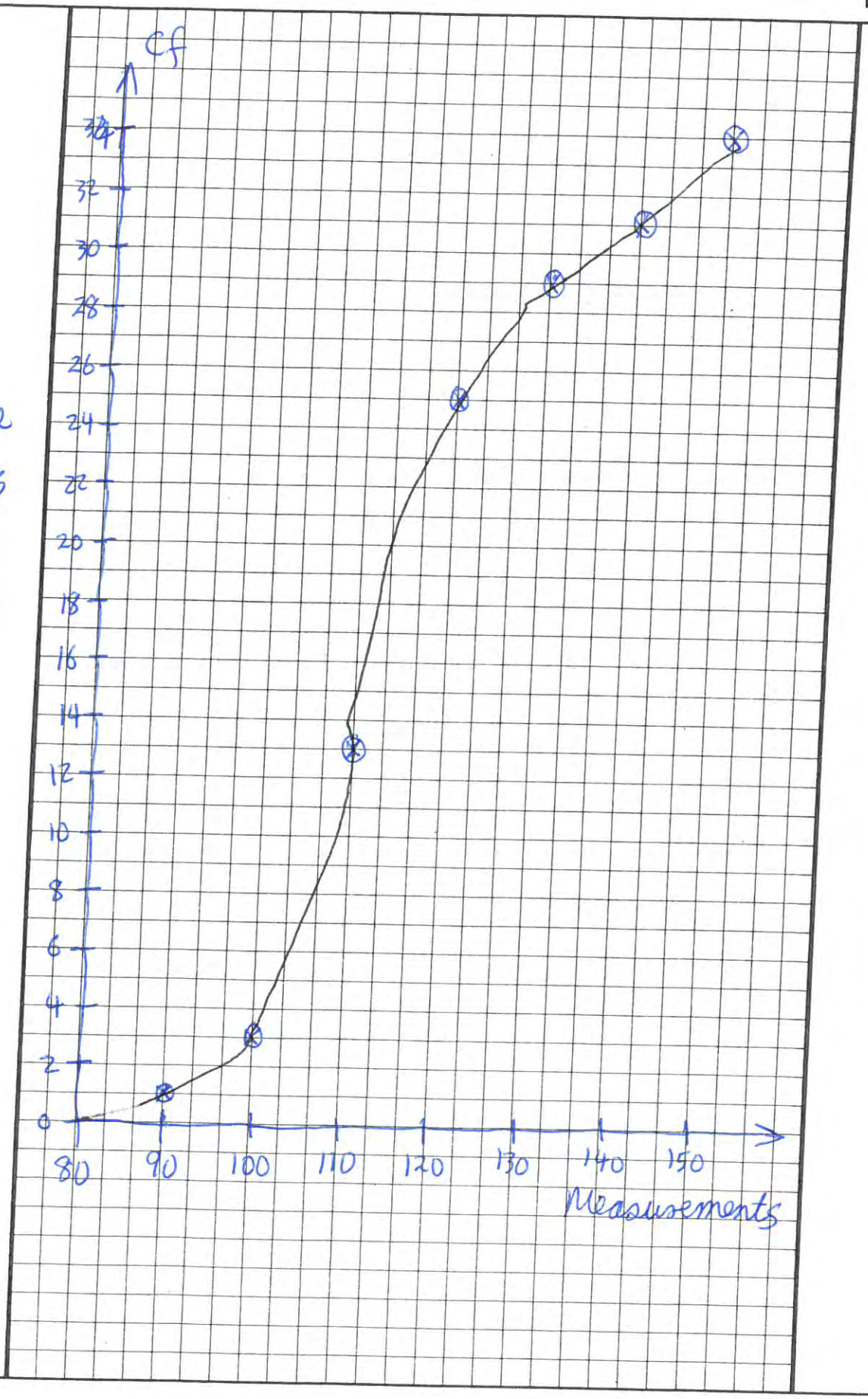
Also called S-curve due to its shape.

It is a plot of Cumulative frequency versus upper limit of the class.

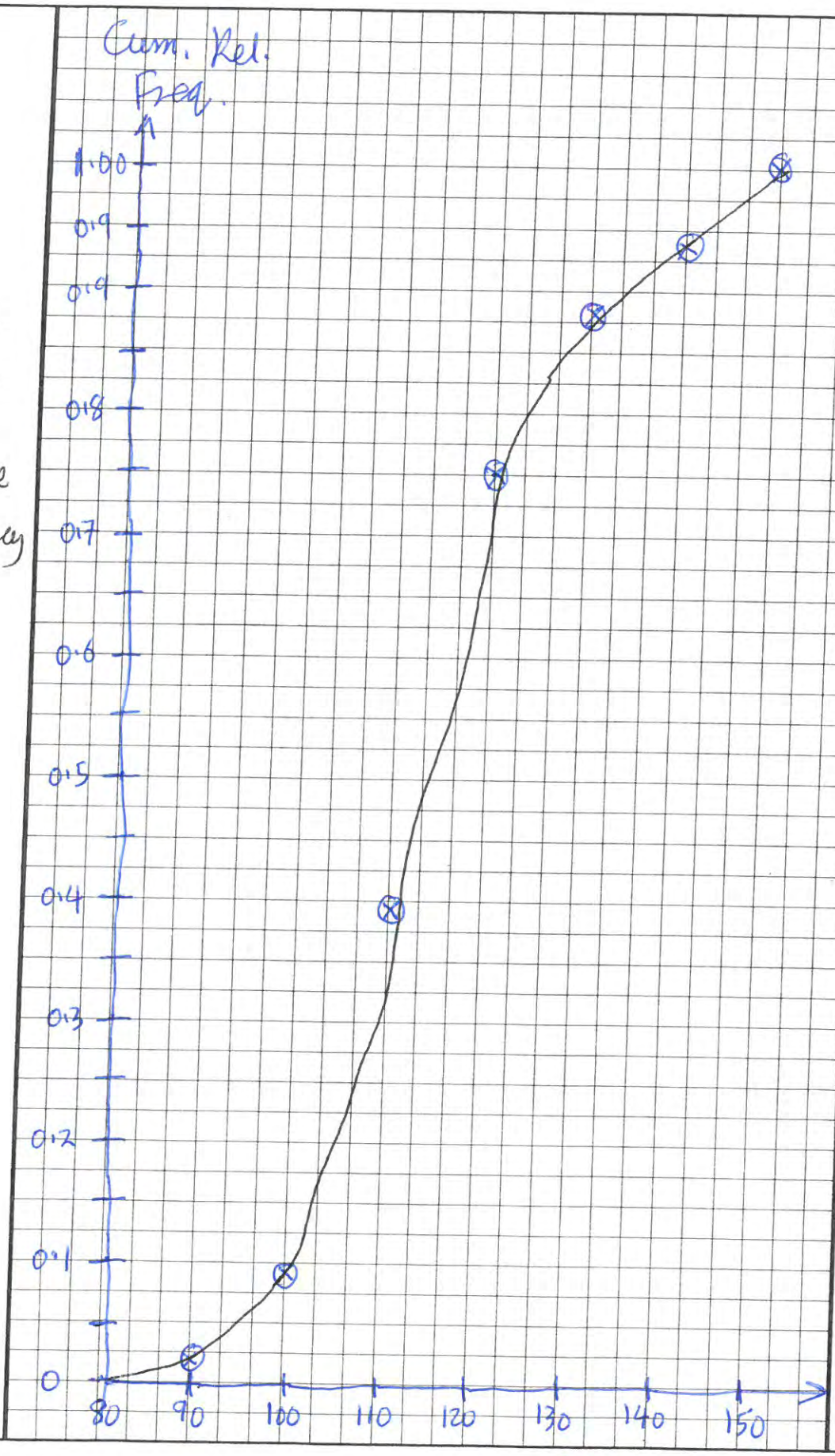
A variation is to plot the Cumulative Relative Frequency.

Some authors plot the cumulative frequency (or cumulative relative frequency) versus the class mark. The resulting curve will have the exact same shape. The difference is that it will be shifted to the left by the class mark value.

Ogive using Cum. Freq.



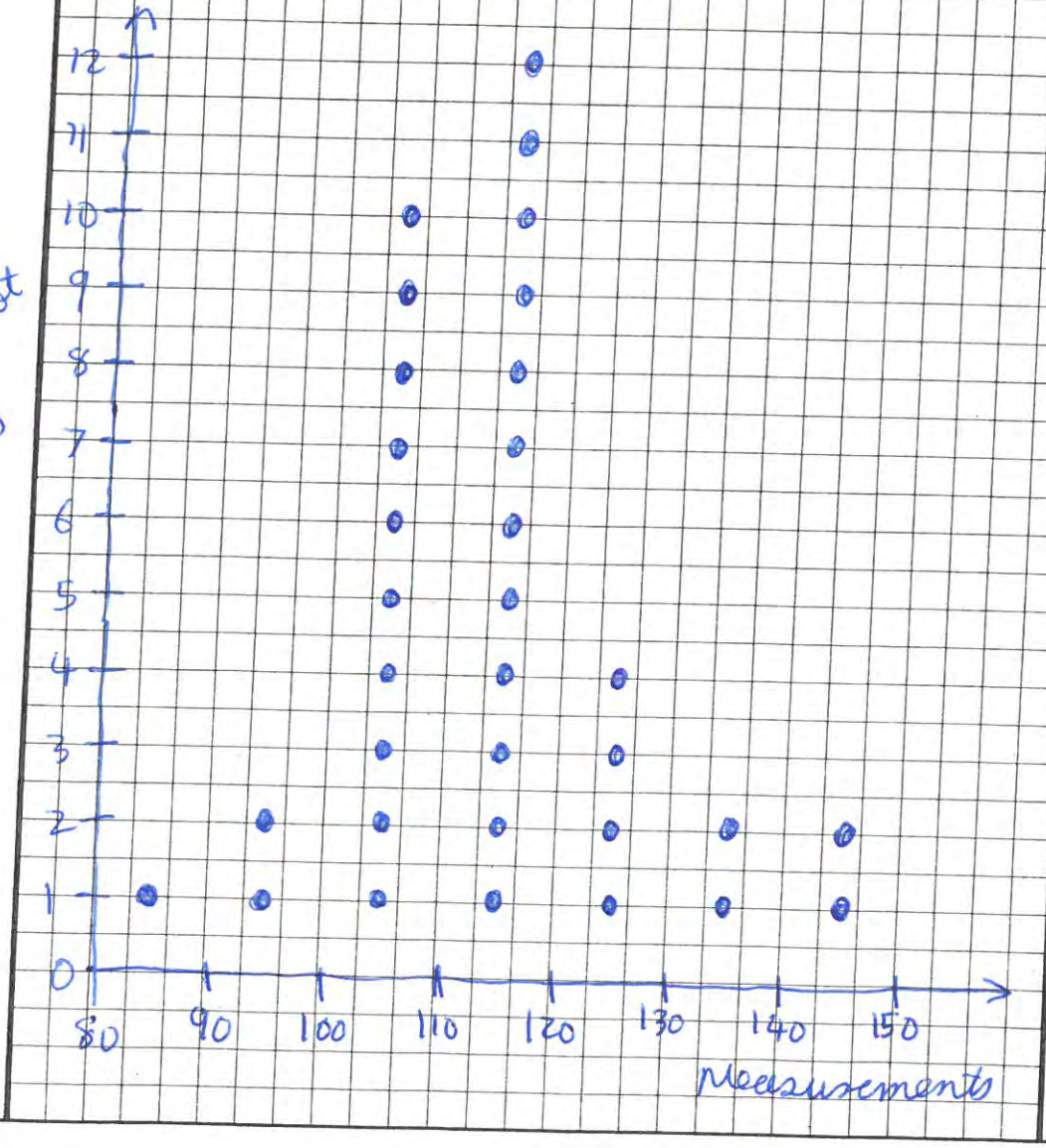
Ogive using Cum. Relative Frequency



The Dotplot

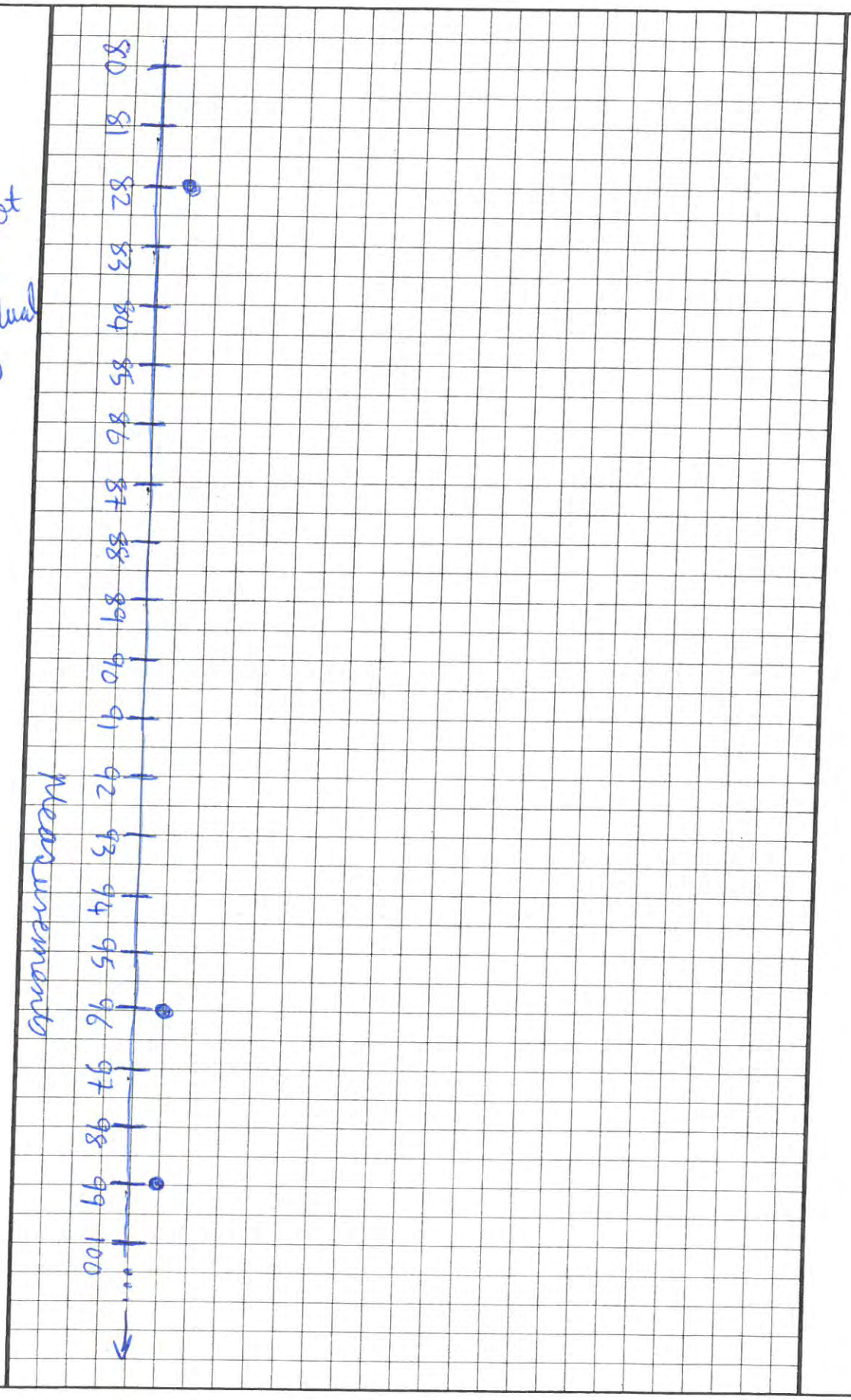
The dot plot uses dots to represent the frequency in the vertical plane.

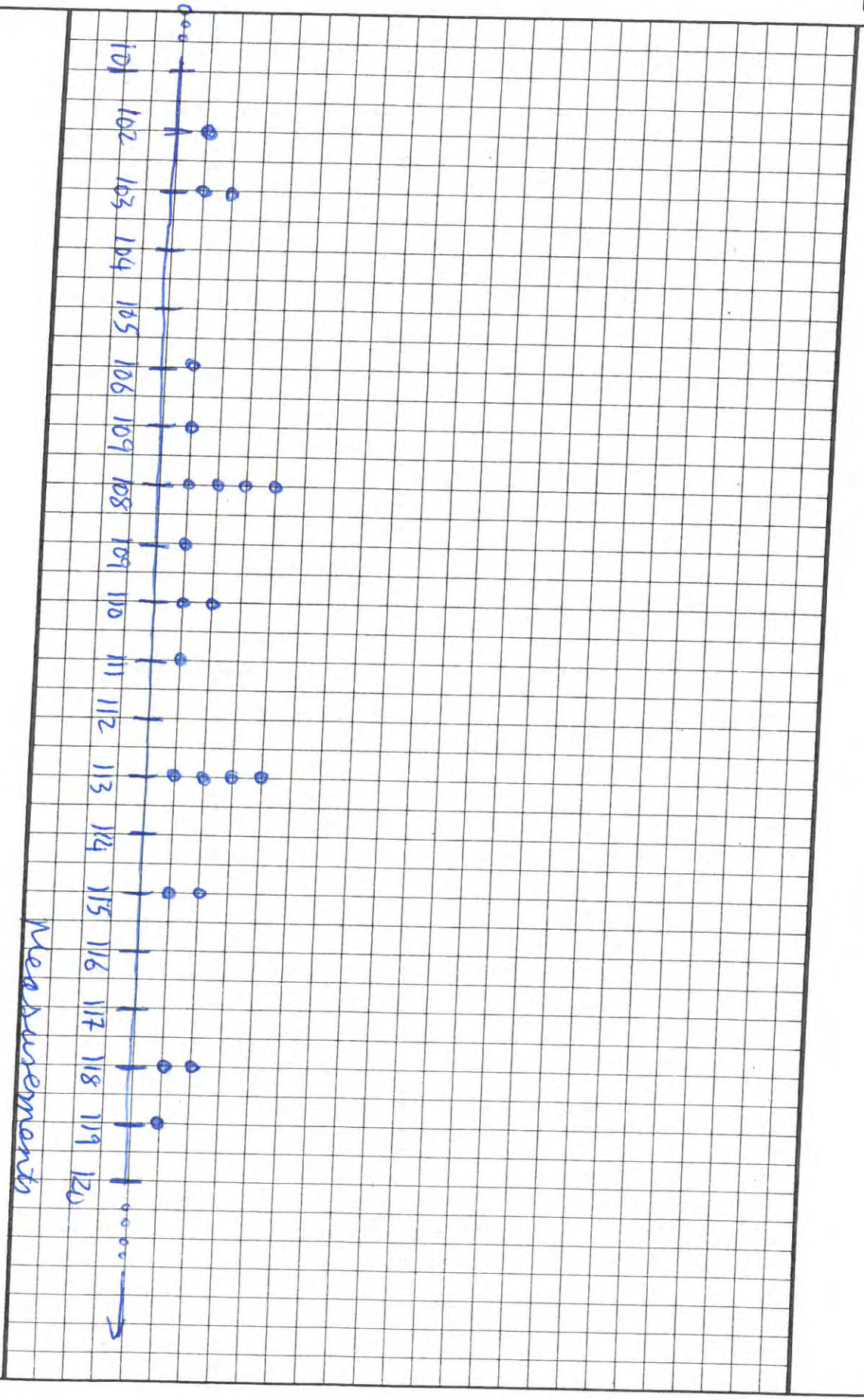
It may be based on individual data points (which is tedious) or the class intervals.



Dot Plot using classes

Dot Plot Using Individual Values




101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
Measurements

Stem and Leaf

Stem and Leaf Diagram

The data is arranged into a stem and each values last digit represents it in the leaf.

You must study your data carefully to select an appropriate stem. There are no hard rules, and each data set will be different.

Stem	Leaf
8	2
9	6 9
10	2 3 3 6 7 8 8 8 9
11	0 0 1 3 3 3 3 5 5 8 8 9
12	1 2 2 7
13	2 6
14	0 6

Upon completion, ALWAYS go back into each leaf and rearrange the values in increasing order. ALWAYS!!

You may use your classes as the stem.

Stem	Leaf
[80, 90)	2
[90, 100)	6 9
[100, 110)	2 3 3 6 7 8 8 8 8 9
[110, 120)	0 0 1 3 3 3 3 5 5 8 8 9
[120, 130)	1 2 2 7
[130, 140)	2 6
[140, 150)	0 6

Stem & leaf using class intervals

Upon completion of initial tally, rearrange each leaf in increasing order, otherwise your diagram will be incorrect.

Pie Chart

Pie Chart.

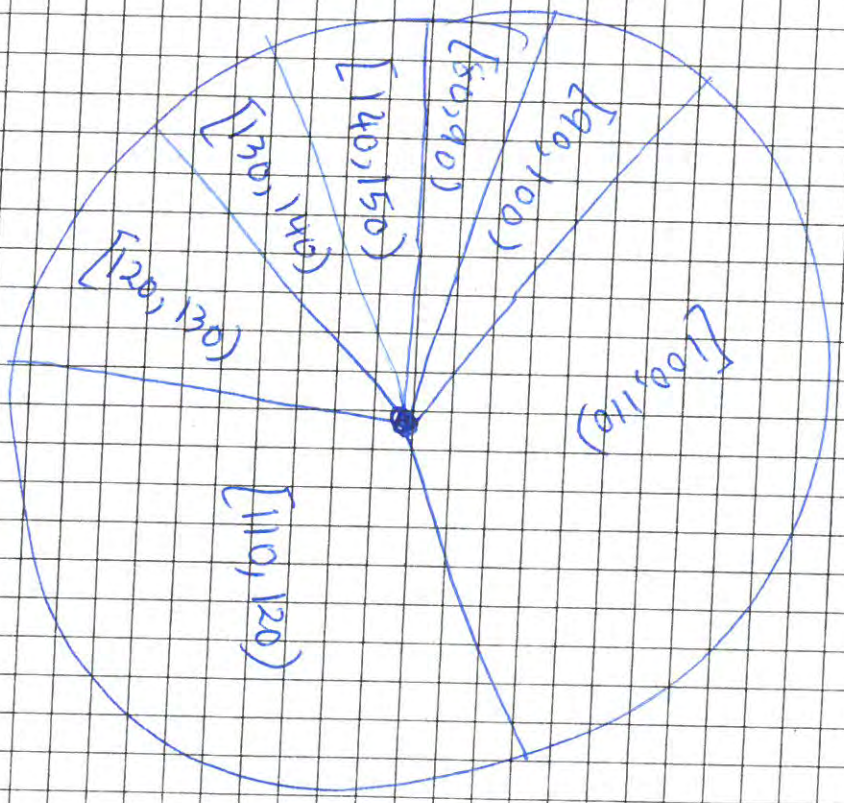
In the Pie chart we slice up the pie such that each class is a sector of the circle. The angle of each class is calculated as;

$$\text{Angle} = \frac{\text{Frequency}}{\text{Total Frequency}} * 360^\circ$$

So,

Class	Freq	Angle
[80, 90)	1	$1/33 * 360^\circ = 10.9^\circ$
[90, 100)	2	$2/33 * 360^\circ = 21.8^\circ$
[100, 110)	10	$10/33 * 360^\circ = 109.1^\circ$
[110, 120)	12	$12/33 * 360^\circ = 130.9^\circ$
[120, 130)	4	$4/33 * 360^\circ = 46.64^\circ$
[130, 140)	2	$2/33 * 360^\circ = 21.8^\circ$
[140, 150)	2	$2/33 * 360^\circ = 21.8^\circ$
	33	

Pie Chart



Descriptive Statistics

Measures of Center identify a 'central' value to represent the data. There are many measures of center

Median: This is the middle value when the data are listed in increasing order. If you have an even number of data, there will be two 'middle' numbers. The average of those will be your median.

Note that in a stem and leaf diagram you have already arranged the data from lowest to highest.

From your stem and leaf diagram read the median.

It is the value at position

$\frac{33}{2} = 16.5 \rightarrow 17^{th}$ position

Median ($\tilde{x}$)

Median ($\tilde{x}$) = 113 .

[Always append relevant units to your answer]

Mean
$\bar{x}$

The Mean ($\bar{x}$) also called the Arithmetic mean. ("average")

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where n is your sample size

For us,

$$\bar{x} = \frac{(82 + 96 + 99 + 102 + \ldots + 146)}{33}$$

$$= 113.72$$

Mode

The Mode is the value with the high frequency

In this data we have a two way tie — 108 and 113. So this data was two modes, or we say it is bimodal.

If there is a 3 or more ties then we have multi modal data.

If ALL the data have a frequency of 1 each, then there is no mode.

Measures of Dispersion

A measure of dispersion is a descriptive statistic that quantifies the variability in the data. In other words how much, on average, is each value different from the mean.

Variance s^2

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

The square root of the variance is called the standard deviation.

Std Deviation s

$$\text{Standard deviation } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

For our data,

$$s^2 = \frac{(82-113.72)^2 + (96-113.72)^2 + \ldots}{33-1}$$

$$= 162.39$$

Standard deviation

$$s = \sqrt{162.39}$$

$$= 12.74$$

As always, append the relevant units to each one.

Quartiles & Percentiles

Quartiles divide the data into quarters. The First quartile, also the called the Lower Quartile has one-fourth of the data below it and three-fourth of the data above it.

The second quartile has two-fourths of the data below it and two fourths of the data above it.

The second quartile is therefore the same as the median.

The third quartile has three fourths of the data below it and one quarter above it. It is also called the upper quartile.

Percentiles divide your data into hundredths.

The 10th percentile has has 10% of the data below it and 90% above it. The 50th percentile has 50% of the data above it and fifty percent below it. It is therefore identical to the median and also the second quartile.

The 75th percentile is identical to the third (upper) quartile.

The procedure to calculate quartiles and percentiles are the same as calculating the median.

Procedure :

First Quantile (Q_1)

lower quartile

$Q_1 \longrightarrow 0.25n = 0.25(33)$

$= 8.25$ position when data is in increasing order.

From stem-leaf diagram

8th position	8.25 position	9th position
↓	↓	↓
107	$? = Q_1$	108

by interpolation

$$\frac{Q_1 - 107}{8.25 - 8} = \frac{108 - 107}{9 - 8}$$

$$\Rightarrow Q_1 = \frac{1}{1} \cdot 0.25 + 107 = 107.25$$

85th percentile

Lets calculate the 85th percentile. (P_{85})

$P_{85} \rightarrow 0.85(33) = 28.05$ position

28th	28.05th	29th
↓	↓	↓
122	$? = P_{85}$	127

$$\frac{P_{85} - 122}{28.05 - 28} = \frac{127 - 122}{29 - 28}$$

$$P_{85} = \frac{5}{1} \cdot 0.05 + 122$$

$$= 122.25$$

As always, append applicable units.

Also, interpolation is optional, you may round off to the nearest position

So for example, for the P_{85}, you could have used the 29th position.

For the Q_1 calculation, you may have used the 9th position and so on.

In the real world, when working with large ~~data~~ the difference in results for interpolation and rounding the position, is negligible.

Outliers

Outliers are values in your data which you suspect are anomalous are due to some error in the measurement process.

Typically once identified we remove outliers from our data before any further analysis.

A value (x_i) on the high end is an outlier if

$$x_i > Q_3 + 1.5\,IR$$

where Q_3 = third quantile

IR = interquartile range, $Q_3 - Q_1$

On the low end of the data, a value is an outlier if

$x_i < Q_1 - 1.5IR$

where Q_1 is the first quartile.

So lets check our data.

$Q_1 \longrightarrow 0.25(33) = 8.25 \approx 9^{th}$ position

$Q_1 = 108$

$Q_3 = 0.75(33) = 24.75 \approx 25^{th}$ position

$Q_3 = 119$

$IR = Q_3 - Q_1 = 119 - 108 = 11$

$1.5IR = 1.5(11) = 16.5$

$Q_3 + 1.5IR = 119 + 16.5 = 135.5$

Any value greater than 135.5 is an outlier

So the values 136, 140 and 146 are outliers. They are unusually high.

Now we check on the low end.

$Q_1 - 1.5\,IR = 108 - 16.5 = 91.5$

Any value less than 91.5 is an outlier.

So 82 is an outlier. It is an anomalous measurement.

So in real life, before calculating our descriptive statistics we would check for outliers and remove them before any calculations or analysis

So for data of size 33 we have over 30 pages of calculations. In real life we use computers to do all this analysis, graphs etc. Please review the tutorial video to see how this is done.

Review the other solved problem on this topic to see how you can use other graphs to get some of the descriptive statistics.