Example: Engineers fabricating a new transmission-type electron multiplier created an array of silicon nanopillars on a flat silicon membrane. The precise structure can influence the electrical properties, so the heights of 50 nanopillars were measured in nanometers and presented below;

245	333	296	304	276	336	289	234	253	392
366	323	309	284	310	338	297	314	305	330
266	391	315	305	290	300	292	311	272	312
315	355	346	337	303	265	278	276	373	271
308	276	364	390	298	290	308	221	274	343

You have been asked to construct a frequency table for this data. Use your frequency table to demonstrate the class width (also called class interval), class mark (also called class mid-point).

We shall follow the procedure presented in class.

1. Range = max value − min value

= 392 − 221

= 171 nanometers

2. Let us divide our range into say 20 classes (or intervals)

$$\text{Class width} = \frac{\text{range}}{\text{No of classes}} = \frac{171}{20} = 8.55$$

To things easier to work with let us round this value up to the nearest multiple of 5 or 10.

So class size $\longrightarrow$ 10.

The class limits will now be chosen so they encompass the min value and the max value accordingly. Note that because we rounded the class with from $8.55 \longrightarrow 10$, we will not have exactly 20 classes but less.

3. So now we can construct the frequency table, also called the frequency distribution.

Interval notation. We shall use $[a, b)$ where a is the lower limit which is included, and b is the upper limit but is not included in the interval.

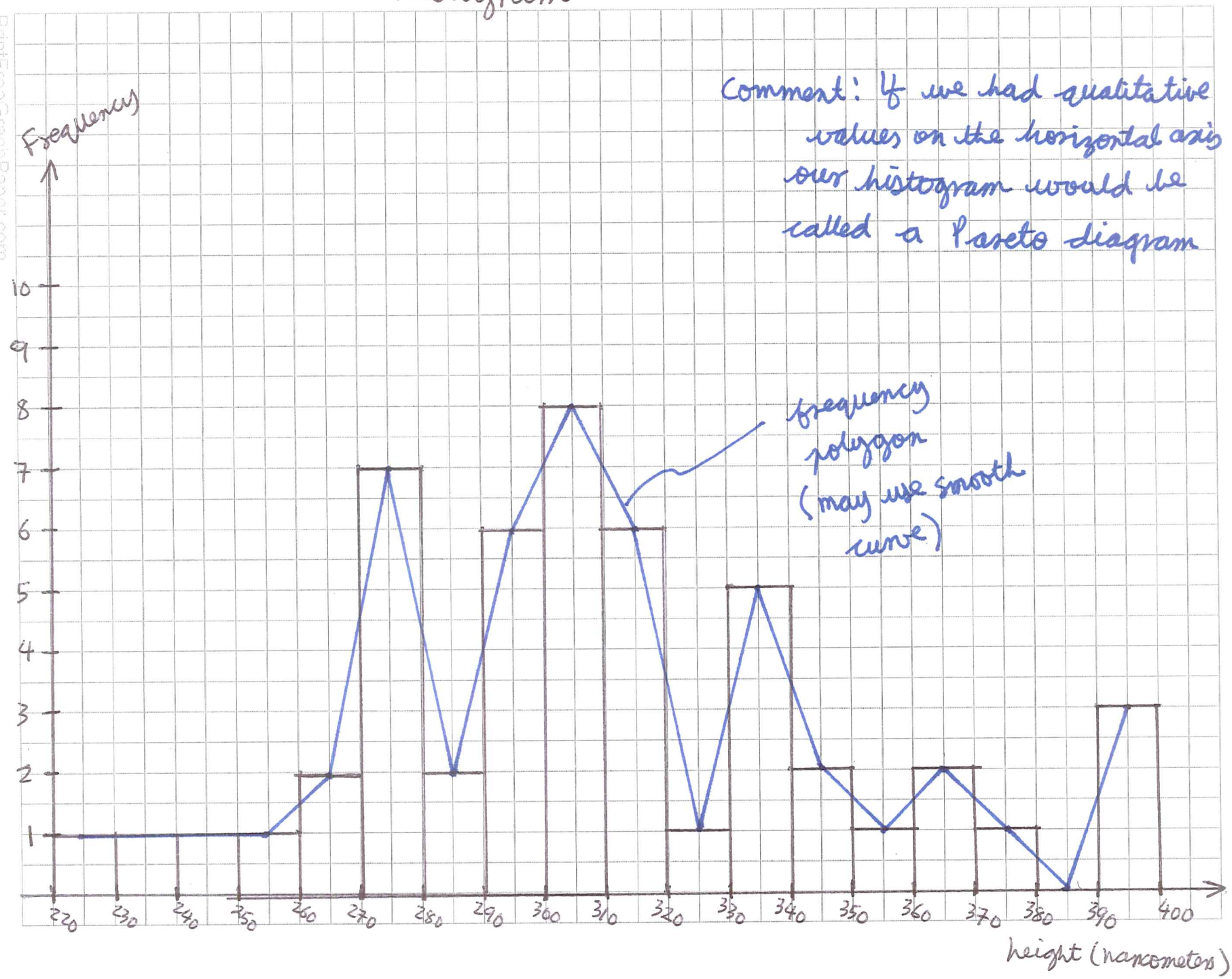
So lets say we have an interval $[220, 230)$. The data values 220, 225, 229, would count towards the frequency of this interval. A value of 230 is not in this interval and counts to frequency of $[230, 240)$.

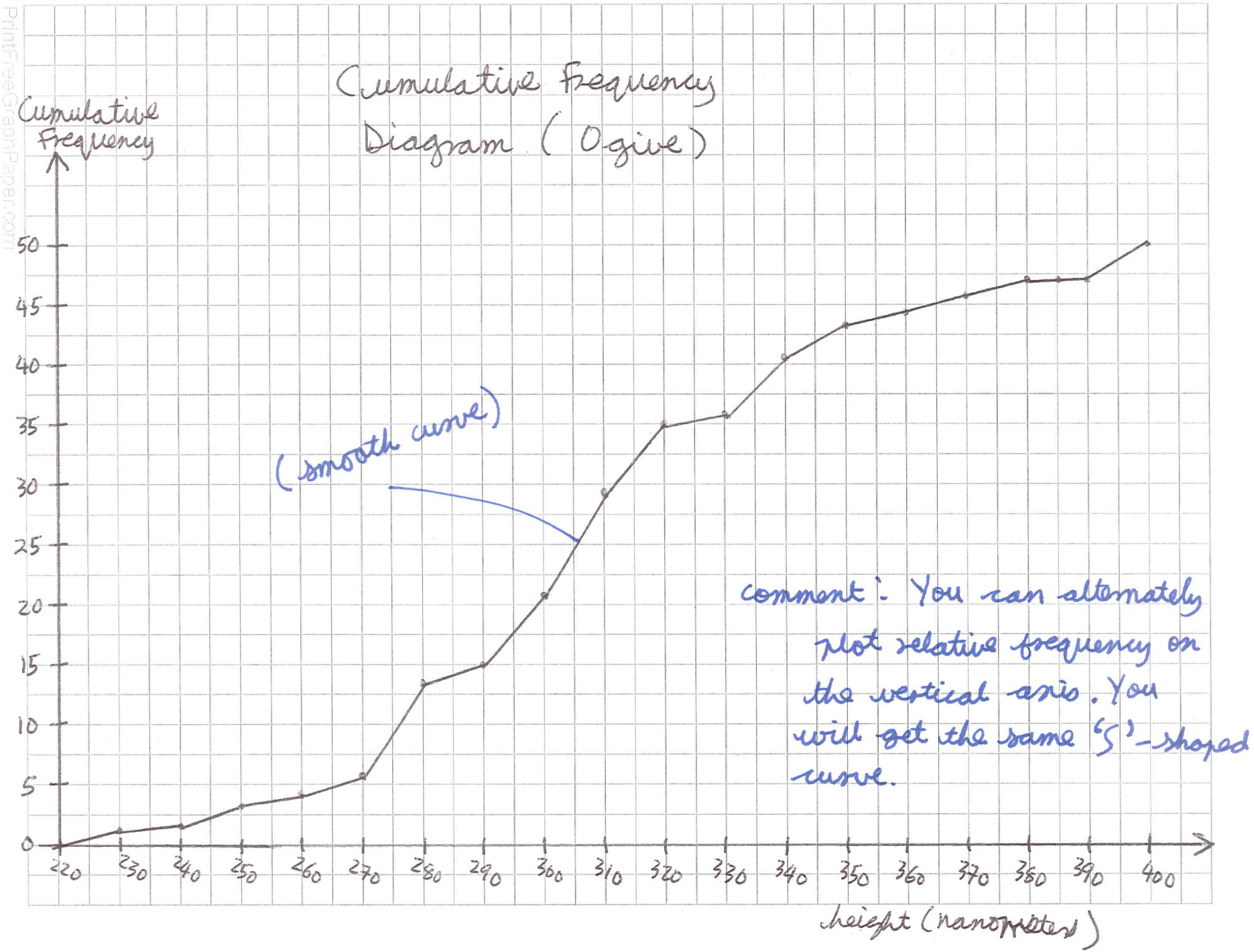
Class	Tally	Frequency	Class Mark	Cumulative Frequency[1]	Relative Frequency[2]
[220 - 230)	/	1	225	1	0.02
[230 - 240)	/	1	235	2	0.04
[240 - 250)	/	1	245	3	0.06
[250 - 260)	/	1	255	4	0.08
[260 - 270)	//	2	265	6	0.12
[270 - 280)	~~////~~ //	7	275	13	0.26
[280 - 290)	//	2	285	15	0.30
[290 - 300)	~~////~~ /	6	295	21	0.42
[300 - 310)	~~////~~ ///	8	305	29	0.58
[310 - 320)	~~////~~ /	6	315	35	0.70
[320 - 330)	/	1	325	36	0.72
[330 - 340)	~~////~~	5	335	41	0.82
[340 - 350)	//	2	345	43	0.86
[350 - 360)	/	1	355	44	0.88
[360 - 370)	//	2	365	46	0.92
[370 - 380)	/	1	375	47	0.94
[380 - 390)		0	385	47	0.94
[390 - 400)	///	3	395	50	1.00
		$\Sigma f = 50$			

1. Class mark = ~~mean~~ $\frac{\text{lower bound} + \text{upper bound}}{2}$

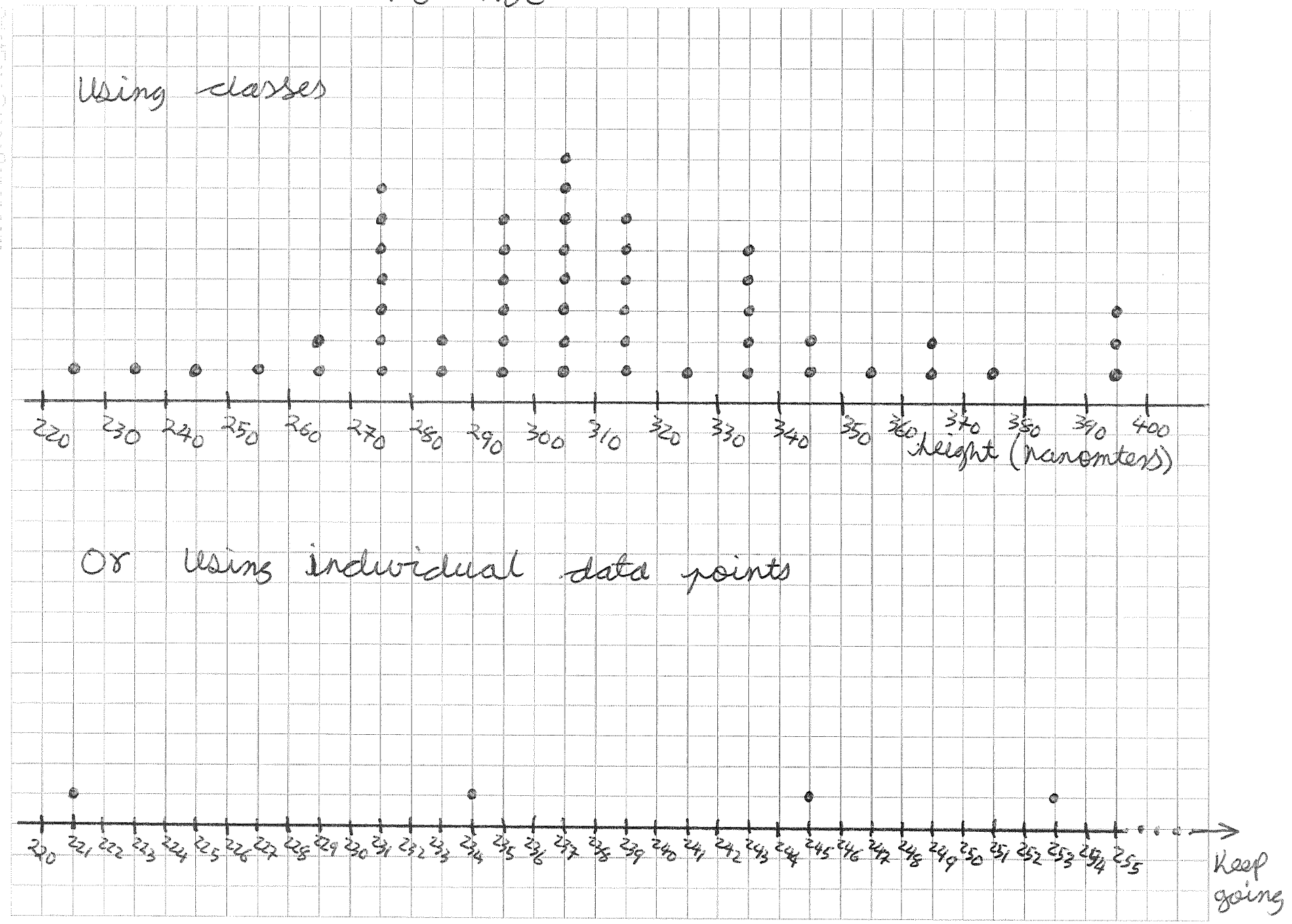
2. Rel. Freq. = $\frac{\text{Cum. Freq}}{\Sigma f}$

Histogram

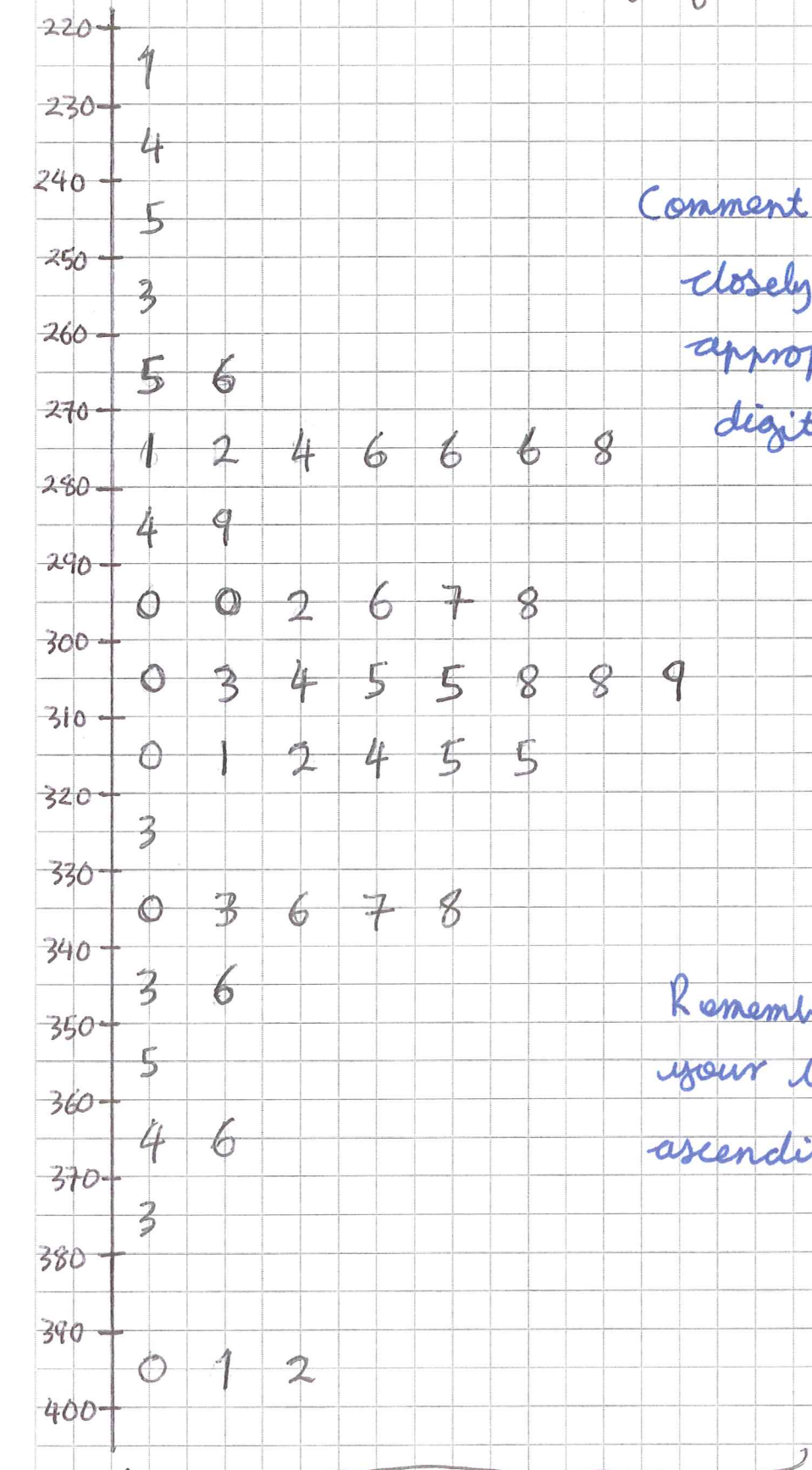



Cumulative Frequency Diagram (Ogive)
Cumulative Frequency
50
45
40
35
30
25
20
15
10
5
0
220
230
240
250
260
270
280
290
300
310
320
330
340
350
360
370
380
390
400
Height (nanometers)
(smooth curve)
comment: You can alternately plot relative frequency on the vertical axis. You will get the same 'S'-shaped curve.

Dot Plot


Using classes
220
230
240
250
260
270
280
290
300
310
320
330
340
350
360
370
380
390
400
Height (nanometers)
OR using individual data points
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
Keep going

Stem-and-leaf diagram



Comment: Study your data closely to choose stem appropriately so that one digit goes into the leaf.

Remember to always arrange your leaf elements in ascending order.

Stem-and-leaf

(Alternate format)

22	1
23	4
24	5
25	3
26	5 6
27	1 2 4 6 6 6 8
28	4 9
29	0 0 2 6 7 8
30	0 3 4 5 5 8 8 9
31	0 1 2 4 5 5
32	3
33	0 3 6 7 8
34	3 6
35	5
36	4 6
37	3
38	
39	0 1 2
40	

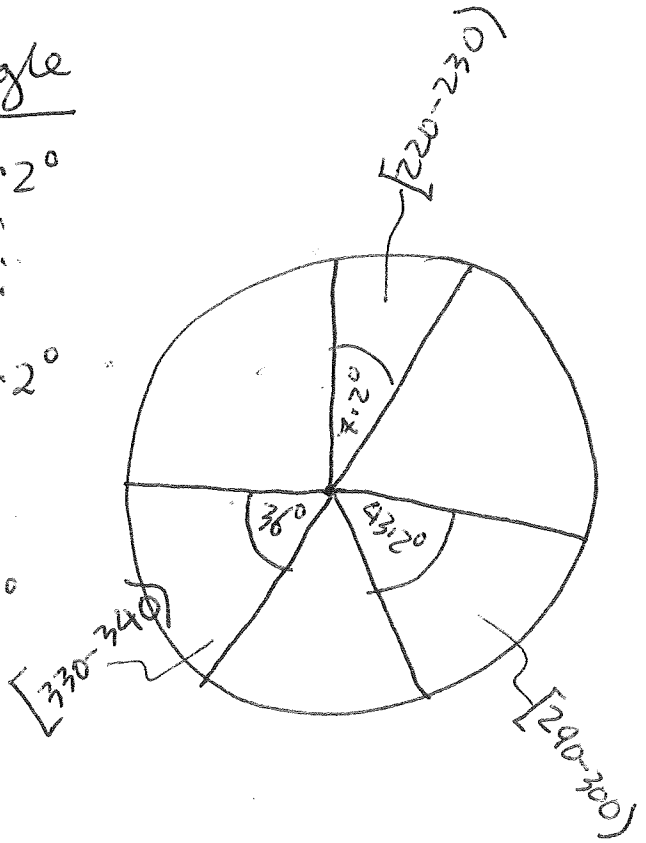
Pie chart

We shall slice up a circular pie based on frequency.

$$\text{Each interval's slice} = \frac{\text{Frequency}}{\text{Total Frequency}} * 360^\circ$$

So

Class	Frequency	Angle
[220-230)	1	7.2°
⋮	⋮	⋮
[290-300)	6	43.2°
⋮	⋮	⋮
[330-340)	5	36°
⋮	⋮	⋮



Students,

Please see page 3 of the class webpages to see how all of the above methods can be done in Excel.

Measures of Center

Median:

$n = 50$ (even number)

$\frac{n}{2} = 25$, $\frac{n+2}{2} = 26$

we find the average of data points at 25th and 26th positions when data arranged in increasing order.

If you have already drawn a stem-and-leaf diagram, you can get the information directly from it.

Median $\tilde{x} = \frac{305 + 305}{2} = 305$ nanometers

Mode: This is the data value with the highest frequency. If there is a two-way tie then the data is said to be bimodal and both values are modes.

If all values have the same frequency then there is NO mode.

For data, looking at the stem-and-leaf diagram, our mode is 276.

Arithmetic mean:

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

$$= \frac{\sum_{i=1}^{50} x_i}{50}$$

$$= \frac{15379}{50}$$

$$= 307.58 \text{ nanometers}$$

Measures of Dispersion

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

here I was reading values for stem-and-leaf plot

$$= \frac{(221-307.58)^2 + (234-307.58)^2 + (245-307.58)^2 + \ldots}{(50-1)}$$

$$= 1511.432 \text{ nanometers}^2$$

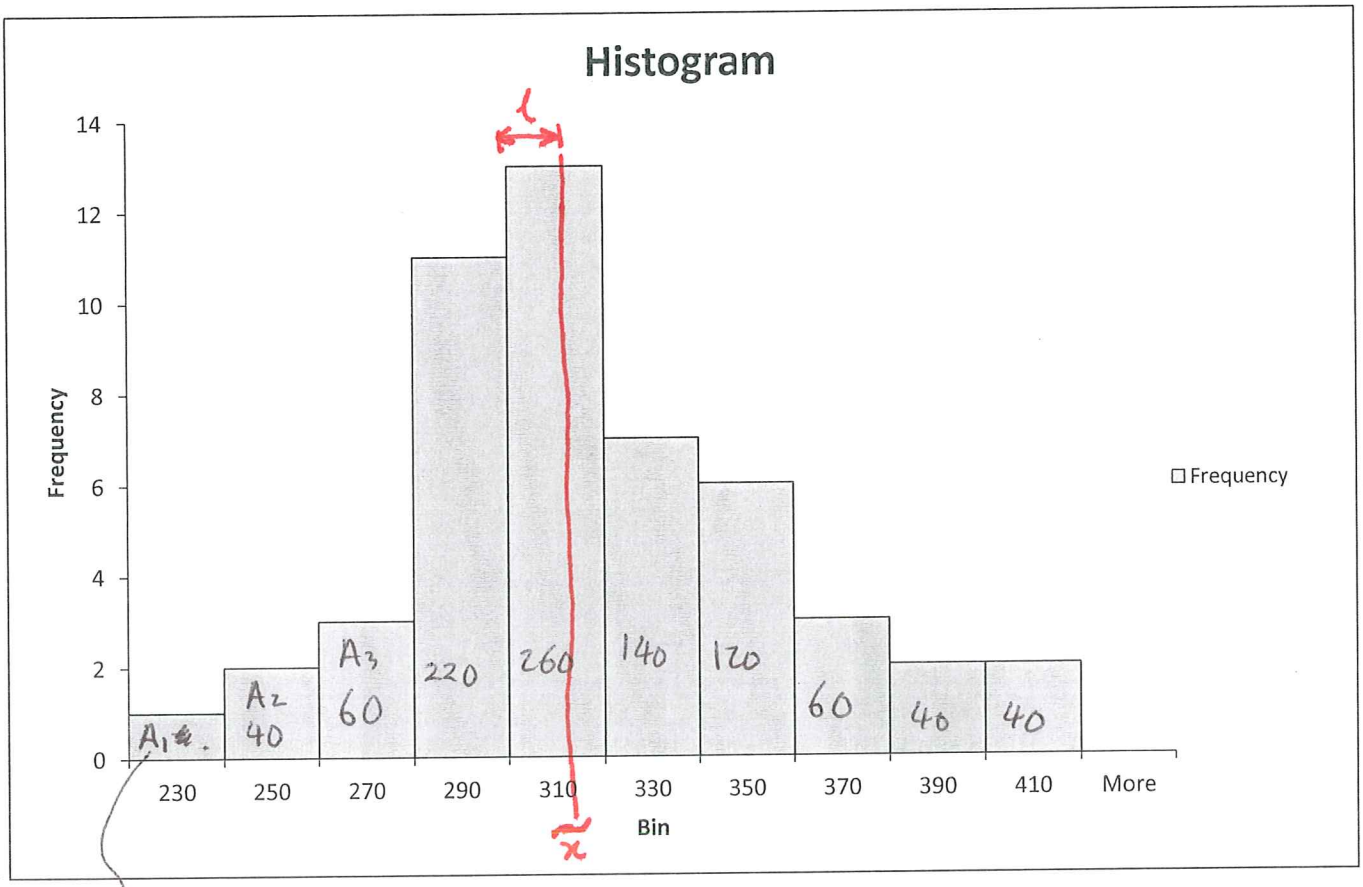
Standard deviation, $s = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

$$s = \sqrt{1511.432}$$

$$= 38.87 \text{ nanometers}$$

Finding Median from Histogram

The median corresponds to 1/2 of area ~~under~~ enclosed by histogram



$A_1 = 20(1) = 20$

Total area = 1,000 sq units

$\frac{1}{2}$ of total area = 500 sq units

Summing area from left cumulatively

@ 290 → Area = 340 sq units

@ 310 → Area = 600 sq units

so we know median will fall in 310 interval

so

$$340 + 13.l = 500$$

$$l = \frac{500 - 340}{13} = 12.31$$

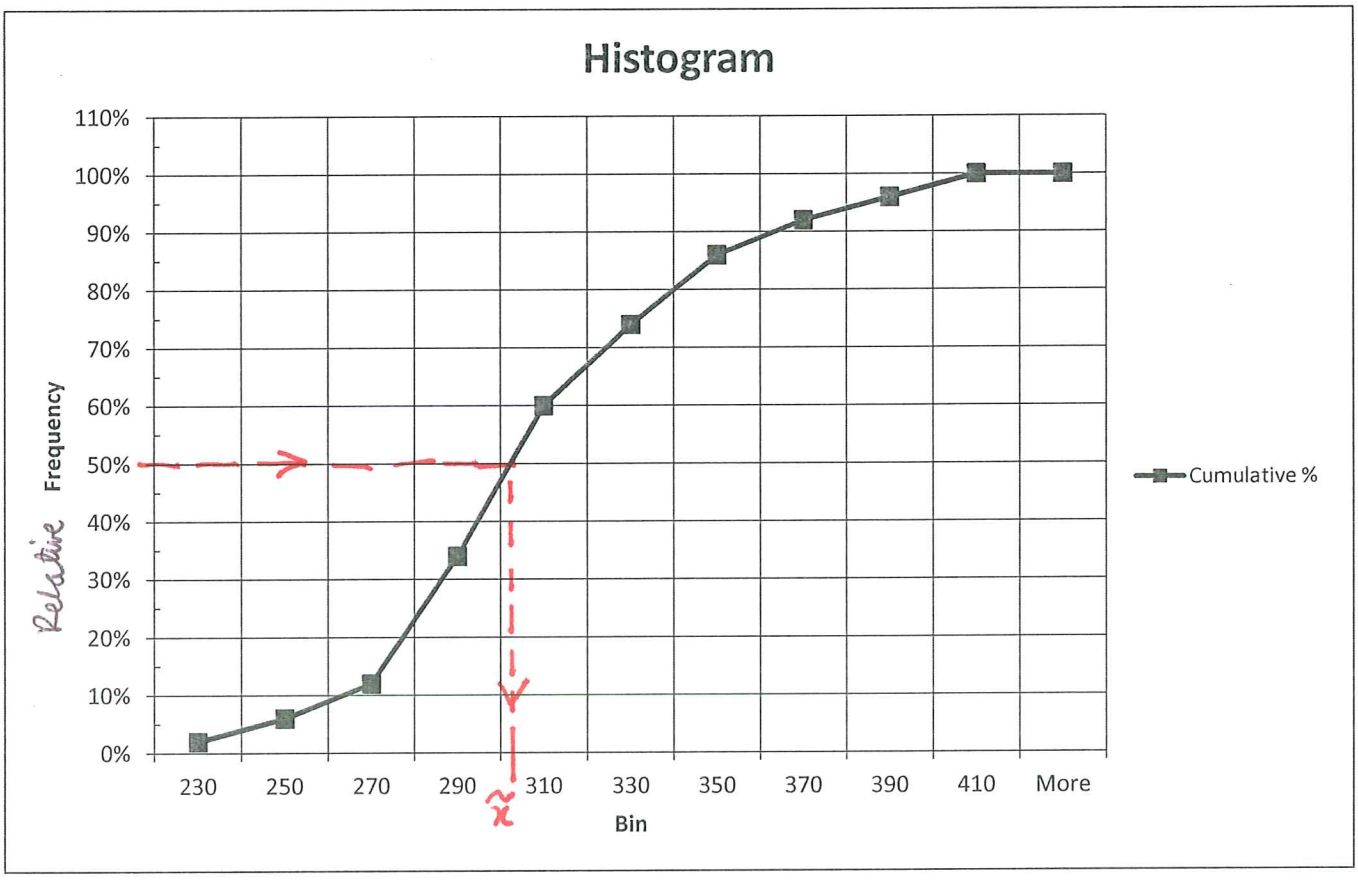
Therefore median

$$\tilde{x} = 300 + 12.31 = 312.31$$

You can follow this process to find your quartiles, percentiles etc.

Median From Ogive

Median corresponds to 50% mark



based on the scale of your axis, read off the median value.

Quartiles, percentiles, ... can be found in the same way

Quartiles

For our nanopillar data:

$Q_1 \longrightarrow 0.25(50) = 12.5$ position in ascending order.

From stem-and-leaf plot

12 position ↓ 276

12.5 position ↓ $? = x$

13 position ↓ 278

by interpolation

$$\frac{x - 276}{278 - 276} = \frac{12.5 - 12}{13 - 12}$$

$x = 276 + 0.5(2) = 277$ nanometers is our Q_1

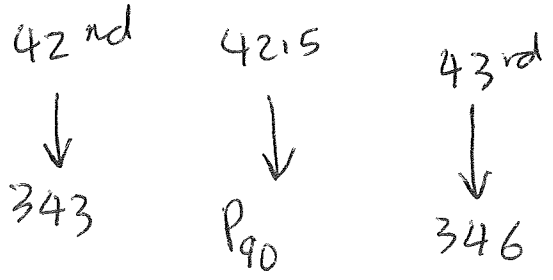
$Q_3 \longrightarrow 0.75(50) = 37.5$

likewise $Q_3 = 330 + 3(0.5) = 331.5$ nanometers

Percentiles

The 85th percentile is the data value at position $0.85(50) = 42.5$ when arranged in ascending order.

Again I will 'cheat' and use the stem-and-leaf diagram so I dont have to manually write out all the data in increasing order



$$P_{90} = 343 + 3(0.5) = 344.5 \text{ nanometers.}$$

This means 90% of the data values are ~~equal~~ less than or equal to 344.5 nanometers.

Outliers

Outliers are 'wierdly' high or low data values. They could be due to some anomaly, measurement error, experimental mis-procedure (if that's a word) etc.

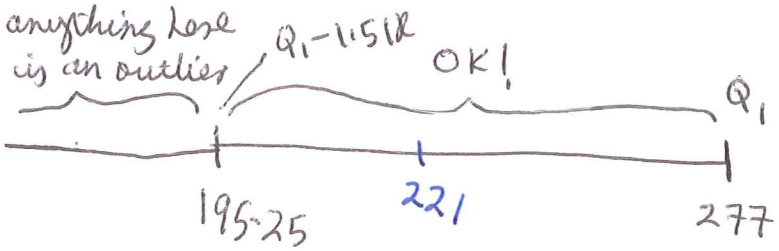
Let us test the low and high values in our data and see if they are outliers.

Low Value:

$$Q_1 - 1.5IR = Q_1 - 1.5(Q_3 - Q_1)$$

$$= 277 - 1.5(331.5 - 277)$$

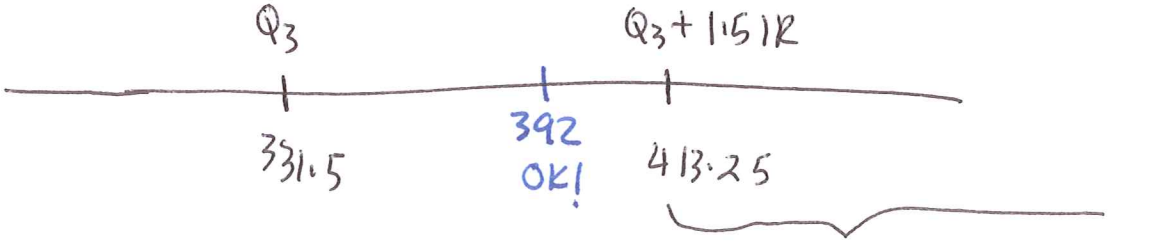
$$= 195.25$$



Our low value 221 is fine and not an outlier

We can check the high value.

$$Q_3 + 1.5\,IR = 331.5 + 1.5(Q_3 - Q_1)$$
$$= 331.5 + 1.5(331.5 - 277)$$
$$= 413.25$$



anything here is an outlier

The high value is not an outlier.

This nanopillar data does not contain any outliers.

In practice outliers may bias or skew your data. Typically, once identified, we eliminate them from the data.