BSc 1 (Game/Web Dev)
Maths 1
Statistics Example

The table below shows the weekly wage for a small firm of 100 workers. For the data contained in this table below, do the following:

(a) Set up the table necessary using the coding method necessary for the mean and standard deviation.

(b) Calculate the mean of the data.

(c) Calculate the standard deviation of the data.

(d) Calculate the semi-inter quartile range.

(e) Calculate the 10-90 Percentile range.

(f) What % of the data falls within ±1 standard deviation of the mean.

(g) Comment on this figure and how it relates to a normal distribution.

Weekly Wage (£)	***Frequency***
0-50	1
50-100	2
100-150	5
150-200	9
200-250	12
250-300	25
300-350	18
350-400	14
400-450	5
450-500	5
500-550	3
550-600	1

(a) Table is set up as follows:
Columns 1 and 7 correspond to the class definitions and corresponding frequencies respectively as given in the frequency distribution.
Column 2 contains the class mark of each class
Column 3 contains the coding index measured from the Assumed Mean Class Mark, A.
Column 4 holds the product of u with the corresponding frequency f from column 7.
Column 5 contains the coding index squared.
Column 4 holds the product of u^2 with the corresponding frequency f from column 7.

Wages (€)	Class Mark (€)	Index u	fu	u^2	fu^2	Freq (f)
0-50	25	-5	-5	25	25	1
50-100	75	-4	-8	16	32	2
100-150	125	-3	-15	9	45	5
150-200	175	-2	-18	4	36	9
200-250	225	-1	-12	1	12	12
250-300	275	0	0	0	0	25
300-350	325	1	18	1	18	18
350-400	375	2	28	4	56	14
400-450	425	3	15	9	45	5
450-500	475	4	20	16	80	5
500-550	525	5	15	25	75	3
550-600	575	6	6	36	36	1
		$\Sigma fu =$	44	$\Sigma fu^2 =$	460	

The sum of all the elements in column 4 yields $\sum f u$. This is required to determine the mean of the frequency distribution from

$$\bar{X} = A + \left(\frac{\sum f u}{N}\right)c$$

The sum of all the elements in column 6 yields $\sum f u^2$. This is required for the standard deviation from the equation

$$s = c\sqrt{\frac{\sum f u^2}{N} - \left(\frac{\sum f u}{N}\right)^2}$$

(b) Mean, $\bar{X}$
Using

$$\bar{X} = A + \left(\frac{\sum f u}{N}\right)c$$

we have, from the above table,

$$A = €275 \text{ and } \sum f u = 44$$

Then

$$\bar{X} = A + \left(\frac{\sum f u}{N}\right)c = 275 + \left(\frac{44}{100}\right)50 = €297$$

(c) Standard Deviation, s

Using

$$s = c\sqrt{\frac{\sum f u^2}{N} - \left(\frac{\sum f u}{N}\right)^2}$$

and the above table we find

$$\sum f u = 44 \text{ and } \sum f u^2 = 460$$

Then

$$s = c\sqrt{\frac{\sum f u^2}{N} - \left(\frac{\sum f u}{N}\right)^2} = 50\sqrt{\frac{460}{100} - \left(\frac{44}{100}\right)^2} = 50\sqrt{4.6 - 0.1936} \simeq €105$$

(d) The Semi-InterQuartile Range (SIQR)

This is defined to be

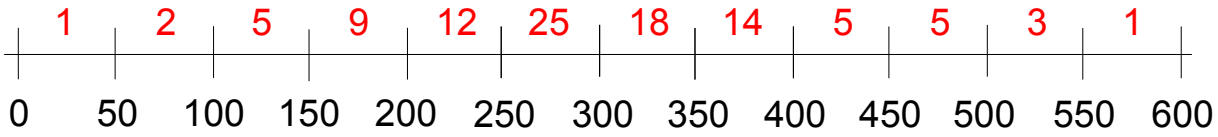
$$SIQR = \frac{Q_3 - Q_1}{2}$$

Therefore we need to determine what Q_1 and Q_3 are from the table above.

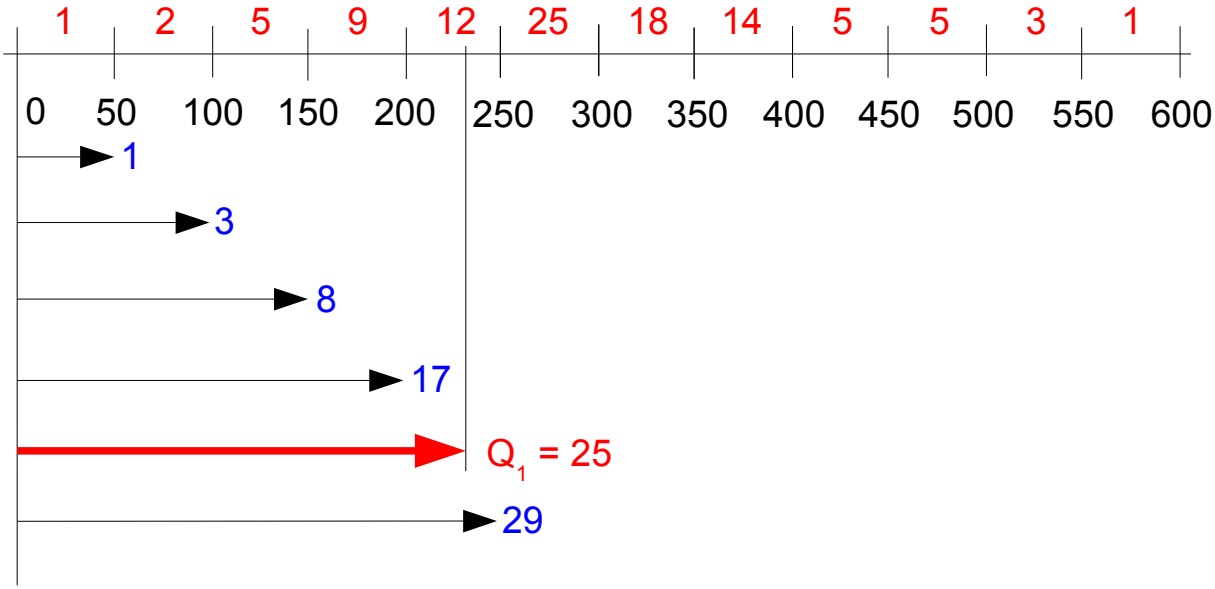
Q_1 is the X value below which 25% (25 in our example because N=100) of the observations (people here) fall.

Q_3 is the X value below which 75% of the observations fall.

To facilitate the calculation to follow (and that of the 10-90 Percentile range, etc.) it helps to visualise the frequency distribution as a number line; i.e.



The frequencies of the classes are given in red and the class boundaries in black. We can then use this representation to determine Q_1 and Q_3. This is shown below for Q_1.

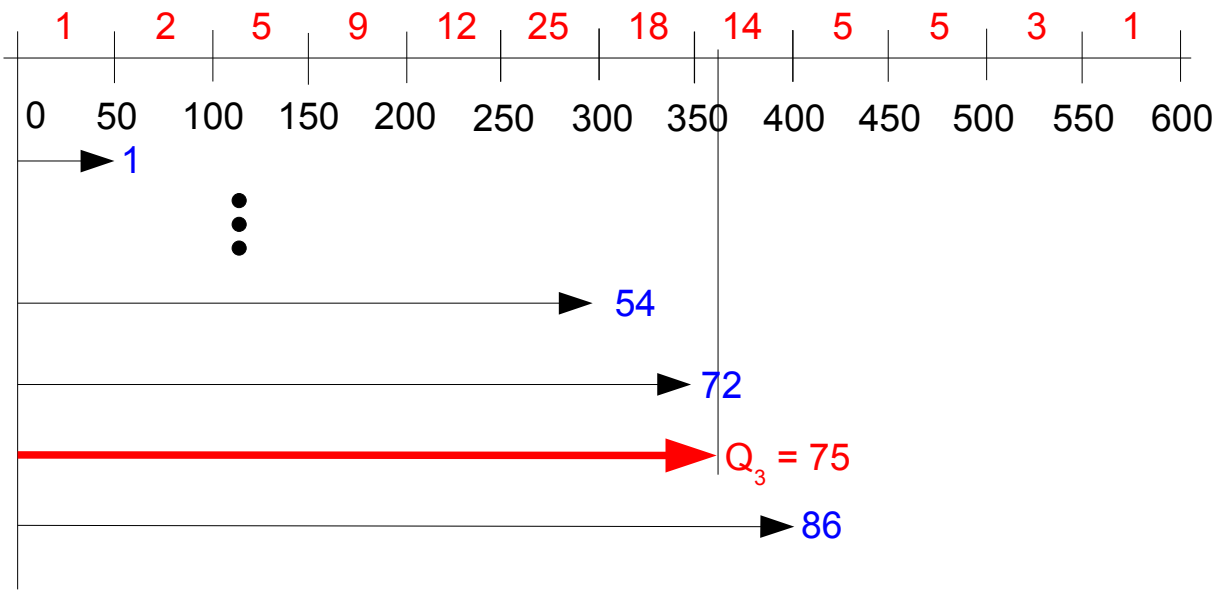


Then Q_1 is somewhere between 200 and 250. At 200, the cumulative frequency is 17 and

so we need 8 of the 12 observations in the 200-250 class to bring the cumulative frequency up to 25; i.e. to Q_1.
Thus

$$Q_1 = \underbrace{200}_{17 \text{ people}} + \underbrace{\frac{8}{12} \times 50}_{\text{Extra 8}} = 200 + 33.33 \simeq €\,233.33$$

For Q_3 we have



Then Q_3 is somewhere between 350 and 400. At 350, the cumulative frequency is 72 and so we need 3 of the 14 observations in the 350-400 class to bring the cumulative frequency up to 75; i.e. to Q_3.
Thus

$$Q_1 = \underbrace{350}_{72 \text{ people}} + \underbrace{\frac{3}{14} \times 50}_{\text{Extra 3}} = 350 + 10.71 \simeq €\,360.71$$

The SIQR is then determined:

$$SIQR = \frac{Q_3 - Q_1}{2} = \frac{360.71 - 233.33}{2} = \frac{126.38}{2} = €\,63.19$$

(e) The 10-90 Percentile Range
This is defined to be

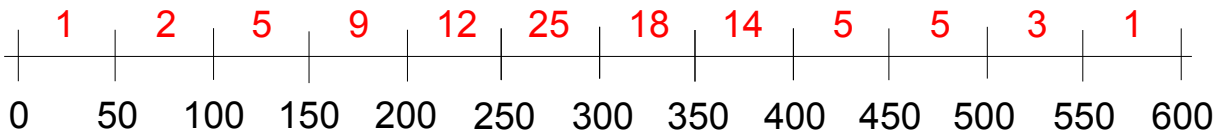
$$\text{10-90 Percentile Range} = P_{90} - P_{10}$$

Therefore we need to determine what P_{10} and P_{90} are from the table above.

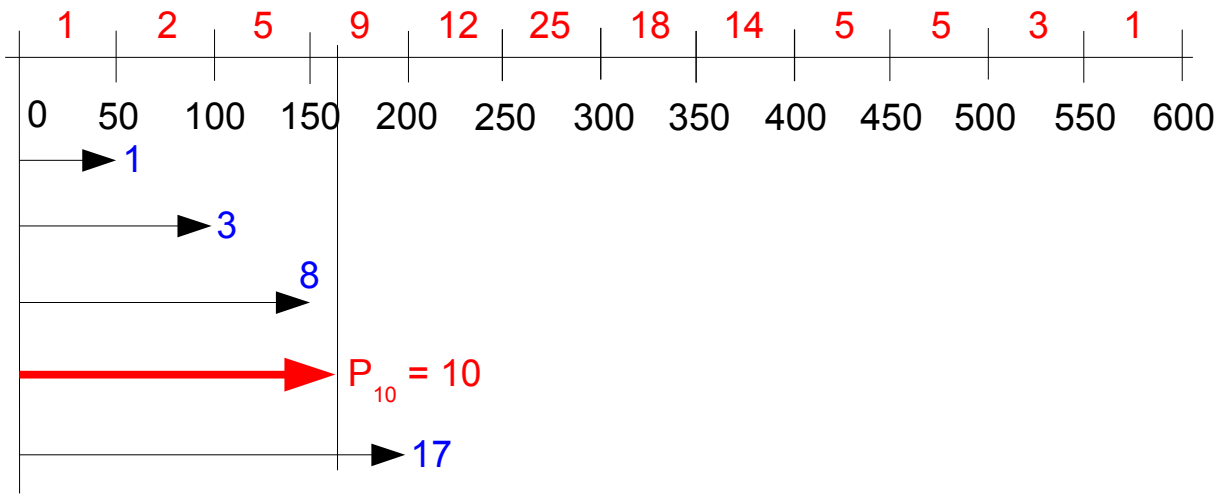
P_{10} is the X value below which 10% (10 in our example because N=100) of the observations (people here) fall.

P_{90} is the X value below which 90% of the observations fall.

As for the SIQR above, we return to our number line representation of the frequency distribution; i.e.



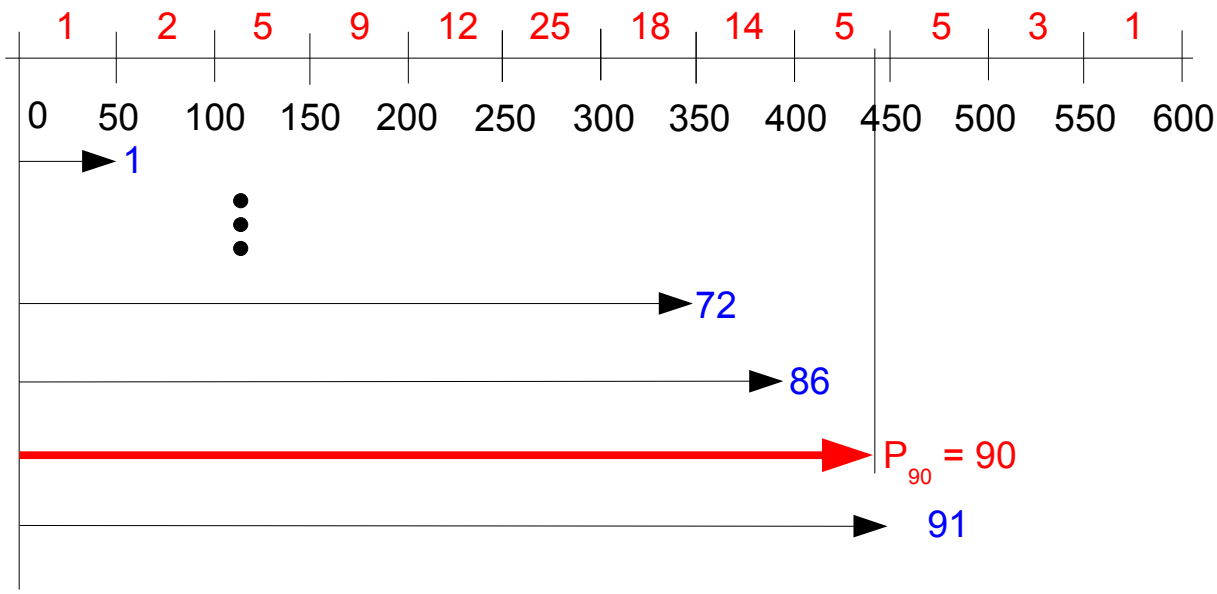
For P_{10} we have



Then P_{10} is somewhere between 150 and 200. At 150, the cumulative frequency is 8 and so we need 2 of the 9 observations in the 150-200 class to bring the cumulative frequency up to 10; i.e. to P_{10}.
Thus

$$P_{10} = \underbrace{150}_{\text{8 people}} + \underbrace{\frac{2}{9}\times 50}_{\text{Extra 2}} = 150 + 11.11 \simeq €161.11$$

For P_{90} we have



Then P_{90} is somewhere between 400 and 450. At 400, the cumulative frequency is 86 and

so we need 4 of the 5 observations in the 400-450 class to bring the cumulative frequency up to 90; i.e. to P_{90}.
Thus

$$P_{90} = \underbrace{400}_{86 \text{ people}} + \underbrace{\frac{4}{5} \times 50}_{\text{Extra 4}} = 400 + 40 = €\,440.00$$

The 10-90 Percentile Range is then determined:

$$P_{90} - P_{10} = €\,440.00 - €\,161.11 = €\,278.89$$

(f) % of the data falling within ±1 standard deviation of the mean
Here we are required to estimate how many people (observations) fall in the interval from $\bar{X} - s$ to $\bar{X} + s$.

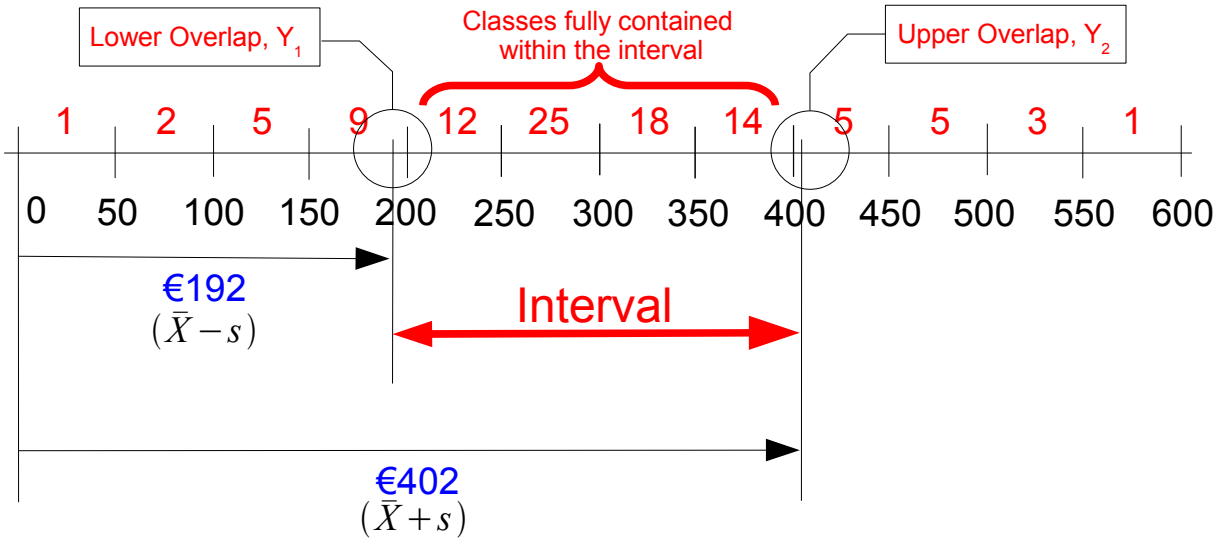
We have calculated the mean and standard deviation already in parts (b) and (c):

Mean, $\bar{X} = €\,297$ and Standard Deviation, $s = €\,105$

Then, the interval we're interested in is

$$\bar{X} - s = €\,192 \leq X \leq €\,402 = \bar{X} + s$$

Using our number line representation of the frequency distribution, we have the following construction:



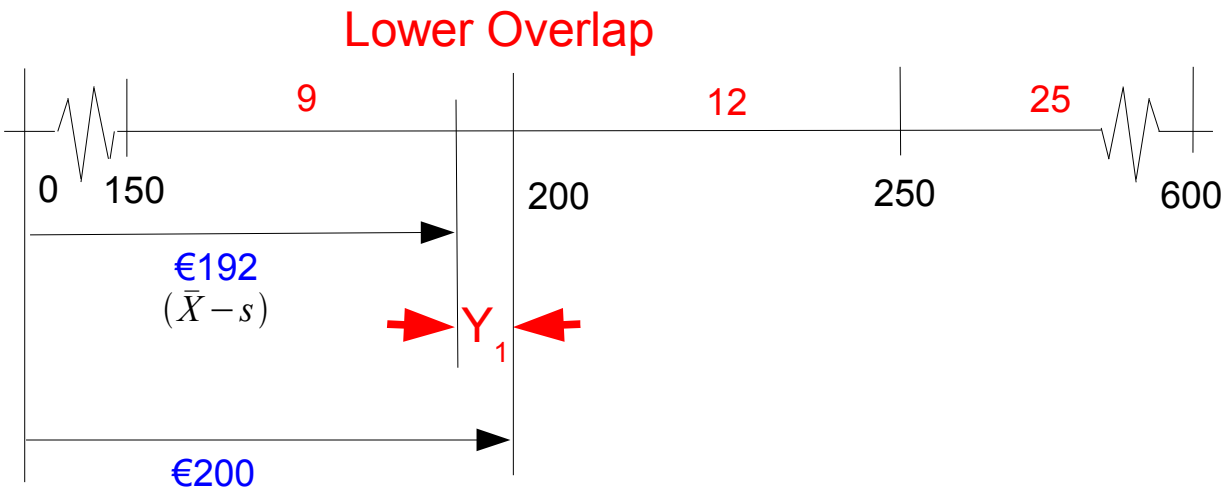
We can see that four classes are contained in their entirety within the interval we interested in. These are the 200-250, 250-300, 300-350, and 350-400 classes. Therefore the people (observations) within these classes are within our interval of interest. So our interval, at the very least, contains 12+25+18+14 = 69 People.

However, the interval boundaries do not coincide with class boundaries and so we have to measure the overlap at each end. For clarity we have labeled these Y_1 for the lower overlap and Y_2 for the upper.

How do we calculate how many people (observations) fall into these overlaps?

Let us consider each overlap in turn.

Y_1 The lower overlap.



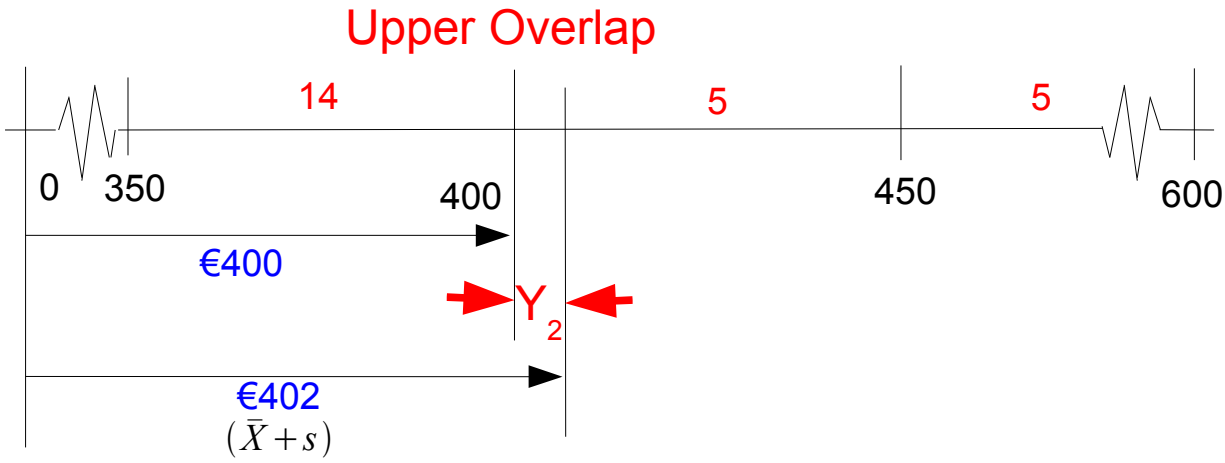
From the construction above, we can see that the lower boundary on our interval extends into the 150-200 class by €8. The fractional overlap is then

$$\text{Fractional Overlap} = \frac{8}{50} = 0.16$$

with the number of people falling into this overlap being this fraction times the class frequency:

Number of people in the lower Overlap = Class Frequency $\times$ Fractional Overlap
= $9 \times 0.16 \simeq 2$ persons

Y_2 The upper overlap.



From the construction above, we can see that the upper boundary on our interval extends into the 400-450 class by €2. The fractional overlap is then

$$\text{Fractional Overlap} = \frac{2}{50} = 0.04$$

with the number of people falling into this overlap being this fraction times the class frequency:

Number of people in the upper Overlap = Class Frequency $\times$ Fractional Overlap
= $5 \times 0.04 \simeq 1$ person

Then the total number of people falling within 1 standard deviation of the mean is 72 people.

N.B.
It could be argued logically that the 1 person in the upper overlap is so small as to be ignored. While it is inadvisable to do this, you could take the total number of people from both overlaps before rounding them up , sum them, and then round up. If you do this then the number of people in the overlap regions would be

$$\underbrace{9\times 0.16}_{Y_1}+\underbrace{5\times 0.04}_{Y_2}=1.44+0.2=1.64\simeq 2$$

Then our answer is reduced by one observation and we get 71 in the interval.

(g) A normal distribution has 68.36% falling within 1 standard deviation of the mean. Using our method above, we get 71-72 falling within this interval and so we would conclude that the data is fairly normally distributed.
In general, if this percentage falls between 63 to 73% then we would conclude that it is fairly normally distributed. Of course, the closer this percentage is to 68%, the more normally distributed the data is.