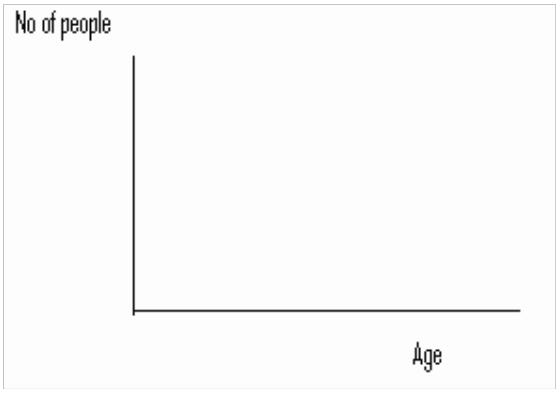
Chebyshev's Theorem and The Empirical Rule

Suppose we ask 1000 people what their age is. If this is a representative sample then there will be very few people of 1-2 years old just as there will not be many 95 year olds. Most will have an age somewhere in their 30's or 40's. A list of the number of people of a certain age may look like this:

Age	Number of people
0	1
1	2
2	3
3	8
..	..
..	..
30	45
31	48
..	..
..	..
60	32
61	30
..	..
..	..
80	6
81	3

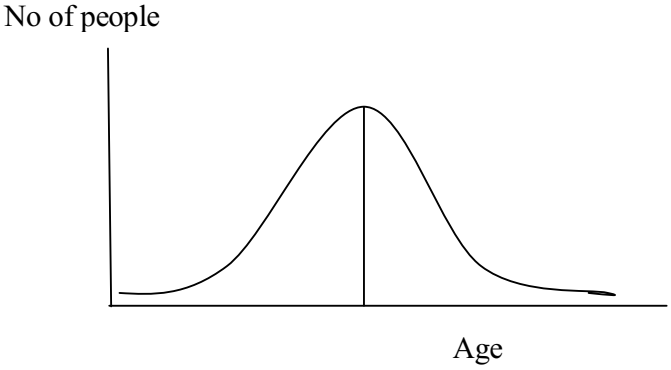
Next, we can turn this list into a scatter diagram with age on the horizontal axis and the number of people of a certain age on the vertical axis.



From the statistical point of view a scatter diagram may have two shapes.

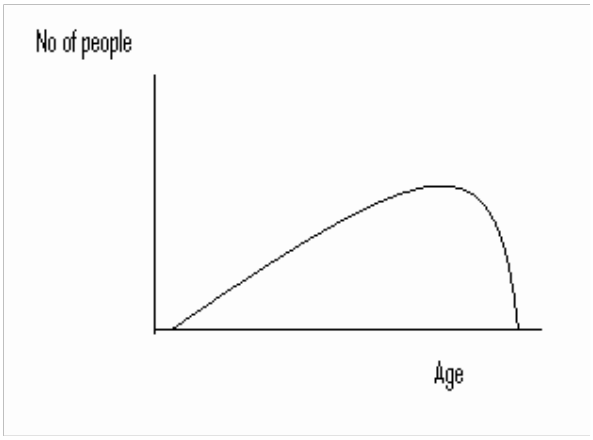
Ron Souverein and Nada Wray

It may be shaped or at least looks approximately like a 'bell curve', which looks like this:



A 'bell curve' is perfectly symmetrical with respect to a vertical line through its peak and is sometimes called a "Gauss curve" or a "normal curve".

The second shape a scatter diagram may have is anything but a normal curve as in the next drawing:



We can do a lot of good statistics with the normal curve, but virtually none with any other curve.

Let us assume that we have recorded the 1000 ages and computed the mean and standard deviation of these ages. Assuming the mean age came out as 40 years and the standard deviation as 6 years we can do the following predictions.

Chebyshev's Theorem

In the case of a scatter diagram that seems to be anything but a normal curve, all we can go by is Chebyshev's theorem. This very important but rarely used theorem states that in those cases where we have a non-normal distribution, the following can be said abut the individual data, which in this case are the ages:

- At least 75% of all the ages will lie in the range of $\overline{X} \pm 2 \cdot s$.
  In our case this means that at least 75% of the people will have an age in the range of $40 \pm 2 \cdot 6 = 40 \pm 12$ years which simplifies to a range of 28 to 52 years.
- At least 88.9% of all the ages will lie in the range of $\overline{X} \pm 3 \cdot s$.
  In our case this means that at least 88.9% of the people will have an age in the range of $40 \pm 3 \cdot 6 = 40 \pm 18$ years which simplifies to a range of 22 to 58 years.
- At least 93.75% of all the ages will lie in the range of $\overline{X} \pm 4 \cdot s$.
  In our case this means that at least 93.75% of the people will have an age in the range of $40 \pm 4 \cdot 6 = 40 \pm 24$ years which simplifies to a range of 16 to 64 years.
- At least 96% of all the ages will lie in the range of $\overline{X} \pm 5 \cdot s$.
  In our case this means that at least 96% of the people will have an age in the range of $40 \pm 5 \cdot 6 = 40 \pm 30$ years which simplifies to a range of 10 to 70 years.
- At least 97.2% of all the ages will lie in the range of $\overline{X} \pm 6 \cdot s$.
  In our case this means that at least 97.2% of the people will have an age in the range of $40 \pm 6 \cdot 6 = 40 \pm 36$ years which simplifies to a range of 4 to 76 years.

How can we calculate these percentages? To calculate the 75%, the 88.9%, the 93.75%, etc, we look at the number of standard deviations in the respective intervals. The 75% goes together with 'mean ± 1 standard deviation', the 88.9% with 'mean ± 2 standard deviations', the 93.75% with 'mean ± 3 standard deviations', and the 96% with 'mean ± 4 standard deviations'. In general you can say that the percentage of people with an age in the range of "mean ± k standard deviations" can be found by calculating the value of the quantity $1 - \frac{1}{k^2}$ and then converting that into a percentage. Summarizing the above we get the following table:

Interval	k	$1-\frac{1}{k^2}$	%
$\bar{X} \pm 2 \cdot s$	2	$1-\frac{1}{2^2}=0.75$	75
$\bar{X} \pm 3 \cdot s$	3	$1-\frac{1}{3^2}=0.889$	88.9
$\bar{X} \pm 4 \cdot s$	4	$1-\frac{1}{4^2}=0.9375$	93.75
$\bar{X} \pm 5 \cdot s$	5	$1-\frac{1}{5^2}=0.96$	96
$\bar{X} \pm 6 \cdot s$	6	$1-\frac{1}{6^2}=0.972$	97.2

Do we have to restrict ourselves to whole numbers as values for k? No, we may take any value for k as long as it larger than 1. For instance, for k = 2.5 we get the result that $1-\frac{1}{2.5^2}=0.84$ *or* 84 % in the interval $40 \pm 2.5 \cdot 6 = 40 \pm 15$ years

Example 1:

Students Who Care is a student volunteer program in which college students donate work time in community centers for homeless people. Professor Gill is the faculty sponsor for this student volunteer program. For several years Dr. Gill has kept a record of the total number of work hours volunteered by s student in the program each semester. For students in the program, for each semester the mean number of hours was 29.1 hours with a standard deviation of 1.7 hours. Find an interval for the number of hours volunteered in which at least 88.9% of the students in this program would fit.

Solution:
From the table above we see that a percentage of 88.9 will coincide with an interval of $29.1 \pm 3 \cdot 1.7 = 29.1 \pm 5.1$ hours. This can be rewritten as an interval from 24 to 34.2 hours volunteered each semester.

Example 2:

The East Coast Independent News periodically runs ads in its own classified section offering a month's free subscription to those who respond. This way management can get a sense about the number of subscribers who read the classified section each day. Careful records have been kept over a period of 2 years. The mean number of responses was 525 with a standard deviation of 30. What is the smallest percentage of responses in the interval between 375 and 675?

Solution:
The difference between the mean of 525 and the upper limit of this interval is 150. This is 5 standard deviations since $150 / 30 = 5$. The same is true for the difference between the mean and the lower limit of this interval. According to the table above this coincides with 96%.

The Empirical Rule

When the data values seem to have a normal distribution, or approximately so, we can use a much easier theorem than Chebyshev's.

The "empirical rule" states that in cases where the distribution is normal, the following statements are true:

- Approximately 68% of the data values will fall within 1 standard deviation of the mean.
- Approximately 95% of the data values will fall within 2 standard deviations of the mean.
- Approximately 99.7% of the data values will fall within 3 standard deviations of the mean.

Example 3:
The average salary for graduates entering the actuarial field is $60,000. If the salaries are normally distributed with a standard deviation of $5000, then what percentage of the graduates will have a salary between $50,000 and $70,000?

Solution:
Both $50,000 and $70,000 are $10,000 away from the mean of $60,000. This is two standard deviations away from the mean, so 95% of the graduates will have a salary in this interval.