**AW2c Chebyshev’s Theorem**

Not necessarily every data set is distributed to follow normal distribution. In such cases, how do we estimate what percentage of data is covered by the standard deviation? Well, we use **Chebyshev’s theorem**. This states that for any set of data (either population or sample) and for any constant k greater than 1, the proportion of the data that must lie within k standard deviations on either side of the mean is at least

1 – 1/k2 (1)

Assumption: k > 1.

If we wanted to calculate the percentage of the distribution that will lie within 2 standard deviations of the mean value, then the value of k = 2. Percentage of data points within the interval mean ± 2 standard deviations = 1 – 1/22 = 1 – ¼ = ¾ = 0.75. Therefore, 75% of the data values from this distribution (either population or sample distribution) will lie within the interval: mean ± 2 standard deviations. You can contrast this with the normal distribution that has 95% of all the data within $\overbar{x}$=±2standard deviations. Data generally come distributed in many shapes and forms and Chebyshev’s theorem just generalizes them all.

Equation (1) also provides us with a shortcut to calculate the appropriate range, as we demonstrate below in Example.

**Example**

A football club keep up to date records on all club members who play for one of the club teams. Information from the club accountant show that the average age for players is 28 years with a corresponding standard deviation of 3. We will apply Chebyshev’s theorem to determine the interval age range to contain 90% of these footballers.

Given that mean = 28, standard deviation = 3, and percentage within interval mean ± k standard deviations = 90%, we solve equation (5.20) with proportion = 0.9 for k.

1 – 1/k2 =0.9

Solving this equation gives:

$k= \sqrt{10}$ = 3.16

Age interval = 28 ± 3.16 \* 3 = 18.52 to 37.48

For mean = 28 and standard deviation = 3, at least 90% of the footballer age values will lie between 18.52 and 37.48 years.

You can create a quick simulation showing for every k what percentage of the sample/population is included in ±k number of standard deviations, as well as what is the value of k for different percentages of sample/population to be included in the range.

To calculate the proportion (percentage) of the sample/population that lie within k standard deviations on either side of the mean, we use equation (1):

% of population $=1-\frac{1}{k^{2}}$ (2)

If, on the other hand, we want to calculate the value of k (i.e. the number of standard deviations) on either side of the mean, providing we know proportion (percentage) of the sample/population, then we solve the same equation for %:

$k=\sqrt{\frac{1}{1 - \% of population}}$ (3)

The two simulations are performed below in Figure 1.



Figure W3.1

**Excel solution**

k Cells A4:A24 Values

% of population Cell B4 Formula: =1-(1/A4^2)

Copy down from B5:B24

% of population Cells D4:D12 Values

k Cell E4 Formula: =SQRT(1/(1-D4))

Copy down from E5:E12

We can see that as we manually change k in column A, the % of population included in the range of the mean ± k number of standard deviations changes in column B. For k=2 we get 75%, as we already know, for k=3 we get 89%, etc. The curve that these two variables form is inverse exponential, and it looks as per Figure 2. In our spreadsheet we went to k=100, which gives us 99.99%. Clearly this is an asymptotic function that tends towards 100, but it never reaches it.



Figure 2

The other simulation in columns D:E does the opposite, i.e. for any given % of population that we would like to include in the range of the mean ± k number of standard deviations, find what is the number of k. We picked some round numbers for the % of population, such as 90%, 95% and 99%, and got the values of k respectively as 3.16, 4.47 and 10. This inverse relationship is depicted in Figure 3.



Figure 3

This curve is a pure exponential curve that asymptotically approaches 1 as the number of k grows.

#### Check your understanding

1 Amazing Bookstore has sold the following number of books during a particular week: 103, 106, 114, 177, 111, 162, 148, 119, 120, and 144. Calculate: (a) sample mean and standard deviation, (b) use Chebyshev’s theorem to find an interval centred about the mean in which you would expect 75% of the weekly sales to fall, and (c) use Chebyshev’s theorem to find an interval centred about the mean in which you would expect 96% of the weekly sales to fall.