Statistics

Suppose the length x (in cm) of 10 screws is measured. The nominal length of the screws is 2.0cm but, because of small random errors in the production process, there is a random variation in the measured length. The set of measurements is called the *population*:

#	1	2	3	4	5	6	7	8	9	10
x cm	2.09	1.95	2.0	1.91	2.05	2.11	1.99	2.01	1.89	2.0

There are two quantities associated with the data: ***mean*** and ***standard deviation***

Mean

The mean of the data is the average:

Mean $\mu = \frac{\sum_1^n x_i}{N}$ = sum of measurements/number of measurements

N = number of measurements = 10

For the population μ =

(2.09 + 1.95 + 2.0 + 1.91 + 2.05 + 2.11 + 1.99 + 2.01 + 1.89 + 2.0)/10 = 2.0

Residual

The quantity $x_i - \mu$ is called a ***residual*** (or error). It is the difference of the i^{th} data value from the mean.

We can now add the residual row to the table:

measurement #	1	2	3	4	5	6	7	8	9	10
X cm	2.09	1.95	2.0	1.91	2.05	2.11	1.99	2.01	1.89	2.0
residual	0.09	-0.05	0.0	-0.09	-0.05	0.11	-0.01	0.01	-0.11	0.0

Standard Deviation

The standard deviation of the population is a measure of the spread of the data around the mean. Both the sum and the mean of the residuals is zero because, by definition, they are equally spread positive and negative. A new quantity is needed that is not zero and uses the squares of the residuals that are all positive. It is the root mean square residual, otherwise known as the ***standard deviation***:

Standard deviation of population = σ = $\sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$ = ***root mean square residual***

A small standard deviation means screw lengths are close to the mean value.

For the screw data (note that the squares of the residuals are all positive)

$\sigma = ((0.09^2 + 0.05^2 + 0.0 + 0.09^2 + 0.05^2 + 0.11^2 + 0.01^2 + 0.01^2 + 0.11^2 + 0.0)/10)^{0.5}$ =

$= ((0.0081 + 0.0025 + 0.0 + 0.0081 + 0.0025 + 0.0121 + 0.0001 + 0.0001 + 0.0121 + 0.0)/10)^{0.5}$

= 0.0675 (to 3 significant figures).

Sample standard deviation

There is another expression called the *sample standard deviation* which is slightly different and is a better approximation when N is a small sample taken from a much larger population. An example is the heights of college students measured at Algonquin College, which is a sample when compared to the much larger nation-wide population of students that are not measured. For the sample standard deviation the formula is:

$$\sigma_s = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N-1}}$$

It becomes the same as the population standard deviation when N is large but, because N-1 is smaller than N, always gives a slightly larger result. This expresses the fact that the mean calculated from a sample is never as accurate as that calculated from the entire population.

Standard Deviation of the Sample Mean

Suppose we repeatedly select samples of size n from a population. For each sample we will not quite get the same value of the sample mean. In fact the sample means will themselves have a mean value that is the same as the population mean but a standard deviation that is larger. As the size of the population increases this standard deviation of the sample means gets smaller until for very large samples it becomes zero.

The formula for the standard deviation of the sample mean = $\sigma_{sm} = \frac{\sigma}{\sqrt{n}}$

This is a famous formula because it shows that, for example, to reduce the error in the measurement of a sample mean by a factor of 10, you have to increase the sample size by a factor of 100. To put this into numbers, taking samples of size 9 and 900 respectively:

$\sigma_{sm(9)} = \frac{\sigma}{\sqrt{9}}$

$\sigma_{sm(900)} = \frac{\sigma}{\sqrt{900}}$

so $\sigma_{sm(900)} / \sigma_{sm(9)} = \frac{\sigma}{\sqrt{900}} / \frac{\sigma}{\sqrt{9}} = \frac{\sqrt{9}}{\sqrt{900}} = \sqrt{\frac{1}{100}} = 1/10$

Example

As part of a statistics teaching exercise, a teacher measures the arrival times of buses on a particular route over a very long period time so as to establish the population results. He finds the time interval between buses varies a lot due to the unpredictability of traffic conditions but has a mean of 20.0 minutes with a standard deviation of 5.0 minutes.

He wants his students to repeat the mean time interval measurement for N consecutive buses, where N can't be too big because the students can't stand there all day. What should N be so that the standard deviation on their mean interval measurements is 1.0 minute?

Answer

$\sigma_{sm} = \frac{\sigma}{\sqrt{n}}$, so we have $1.0 = 5.0/\sqrt{N}$ which gives $N = (5.0/1.0)^2 = 25$ buses

Samples of 25 buses will give a mean interval that has a standard deviation of 1.0 minute.

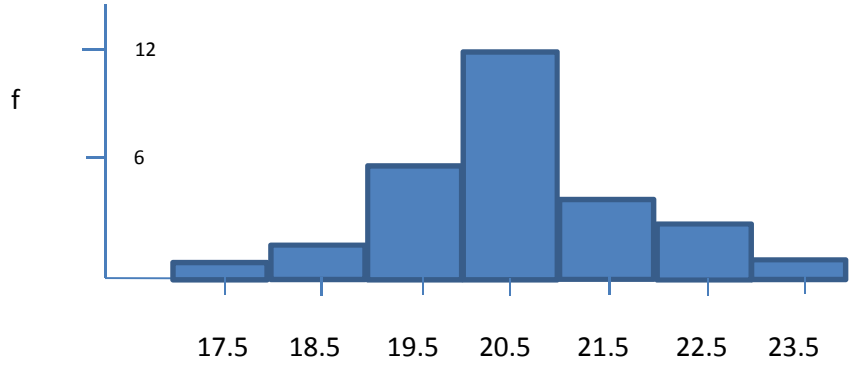
Class mark, Class width, Frequency and Relative Frequency

Suppose we measure the lengths of a larger number of screws and record the number of screws in each range of 1mm around the 20mm (2cm) nominal value. The number for a specific range is called the ***frequency***. When the frequency is divided by the population, or sample size, it is called the ***relative frequency***.
The results for a sample of 30 screws measured to the nearest 0.1mm might be:

range mm	16-6.9	17-17.9	18–18.9	19–19.9	20-20.9	21-21.9	22-22.9	23-23.9	24-4.9	
frequency f	0	1	2	6	12	5	3	1	0	
relative frequency fr	0 = 0	1/30 = 0.0333	2/30 = 0.0666	6/30 = 0.2	12/30 = 0.4	5/30 = 0.1667	3/30 = 0.1	1/30 = 0.0333	0 = 0	
class mark x	16.5	17.5	18.5	19.5	20.5	21.5	22.5	23.5	24.5	

An important property of the relative frequency is that the sum of all the relative frequencies equals 1.0: $\sum fr = 0.0333 + 0.0667 + 0.2 + 0.4 + 0.1667 + 0.1 + 0.0333 = 1.0$
(We will see this result again later where it translates to the area under the Standard Normal Distribution being equal to 1.0)

These data can be plotted on a **histogram** where each entry is a rectangle. The height of the rectangle is the frequency (or relative frequency) and the base of the rectangle is the range or **class width** (0.1 for the screw data):



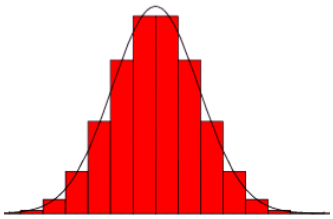
- the **class** for each rectangle determines the actual range (17-17.9, 18-18.9, …) of values included in the frequency
- the **class mark** is the value at the centre of each rectangle (17.5, 18.5, 19.5,…)

When presented this way we have lost information about the exact measured lengths of the screws and only know what class a given screw falls in. To calculate the mean the best we can do is use the class mark and either the frequency or the relative frequency:

a. Using the frequency.
Mean $\mu = (1x17.5 + 2x18.5 + 6x19.5 + 12x20.5 + 5x21.5 + 3x22.5 + 1x23.5)/30 = 20.53$

b. Using the relative frequency
Mean $\mu = 0.0333x17.5 + 0.0666x18.5 + 0.2x19.5 + 0.4x20.5 + 0.1667x21.5 + 0.1x22.5 + 0.0333x23.5 = 20.53$

When the data set is very large and the quantity being measured is well defined with a small random error and measurements are made to high precision the histogram becomes a smooth curve called a ***Normal Distribution*** (or ***Gaussian Distribution*** or ***Bell curve***):

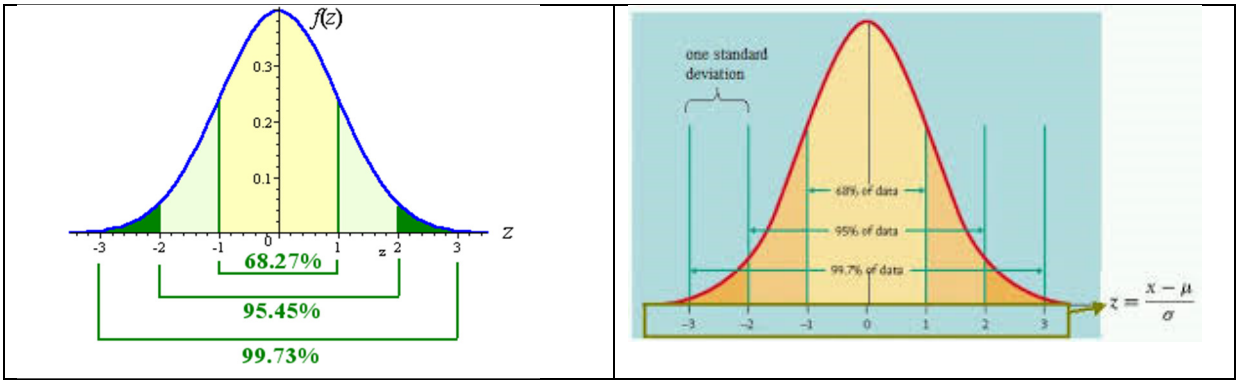


Standard Normal Distribution and z-value

The data from any Normal Distribution can be reduced to lie on a universal curve called the **Standard Normal Distribution** when the following changes are made:

1. The data is presented in terms of the number of standard deviations from the mean z:
   z = number of standard deviations from mean = $(x- \mu)/ \sigma$
2. The relative frequency is divided by the class width to become the probability density f(z)

The Standard Normal Distribution is centred at z = 0, has a standard deviation of 1.0 and a height of 0.3989. Since all Normal Distributions can be reduced to lie on the Standard curve, the results from it can be used to make specific predictions about any particular Normal Distribution.



From page 2 we saw that the sum of the relative frequencies fr = 1.0. In this plot it translates to the area under the curve = 1.0. This means that **the area under the curve between 0 and a particular value of z corresponds to the relative frequency of data lying in that range**.
From measuring the areas it is found that:

- the area between z = 0 and z = 1.0 is 0.3413 which means 68.26% of data lies within +/- 1 standard deviation from the mean,
- the area between z = 0 and z = 2.0 is 0.4772 which means 95.44% of data lies within +/- 2 standard deviation from the mean,
- the area between z = 0 and z = 3.0 is 0.4987 which means 99.74% of data lies within +/- 3 standard deviation from the mean,

Because this useful result concerning the area can be used in practical calculations, the area under the curve has been calculated for a range of positive values of z.

Area Under the Standard Normal Distribution

z	*Area*	*z*	*Area*	*z*	*Area*
0.0	0.0000	1.0	0.3413	2.0	0.4772
0.1	0.0398	1.1	0.3643	2.1	0.4821
0.2	0.0793	1.2	0.3849	2.2	0.4861
0.3	0.1179	1.3	0.4032	2.3	0.4893
0.4	0.1554	1.4	0.4192	2.4	0.4918
0.5	0.1915	1.5	0.4332	2.5	0.4938
0.6	0.2257	1.6	0.4452	2.6	0.4953
0.7	0.2580	1.7	0.4554	2.7	0.4965
0.8	0.2881	1.8	0.4641	2.8	0.4974
0.9	0.3159	1.9	0.4713	2.9	0.4981
1.0	0.3413	2.0	0.4772	3.0	0.4987

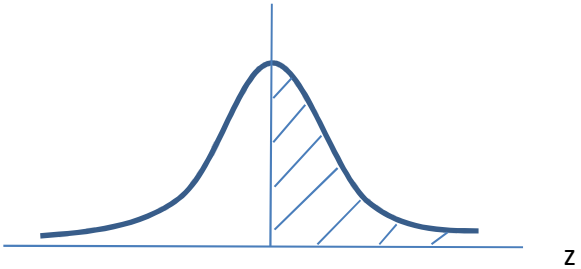
Since the Standard Normal curve is symmetric about the mean, this table also applies to negative z values (data below the mean).

Examples using the Standard Normal Distribution

Because all normal distributions are described by this standard curve, we can now predict the results we expect to get from any statistical data that is normally distributed.

Example 1. The lengths of a nominal 20mm screws follow a Normal distribution. From measurements on a sample it is found that the mean length is 20.00mm and the standard deviation is 1.00mm. A store sells 2000 of this type of screw to customers.

a. How many screws are longer than 20.00mm?
For 20.00mm the z value = (20.00 – 20.00)/1.00 = 0
So the question is how many lie above the mean? The answer is 0.5, or the hatched area shown under the right-hand half of the curve. The number of screws = 0.5 x 2000 = 1000 screws

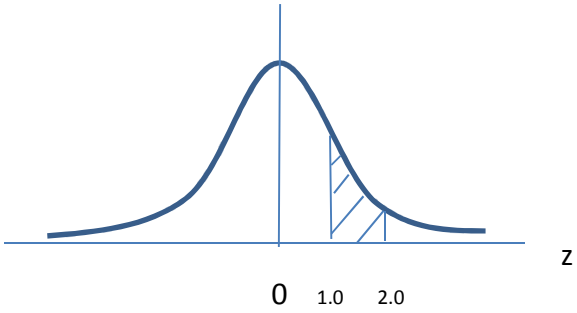


b. How many screws will have lengths between 21.00mm and 22.00mm?
For 21.00mm the z value = (21.00 – 20.00)/1.00 = 1.00
For 22.00mm the z value = (22.00 – 20.00)/1.00 = 2.00

So the question is how many screws lie between 1 and 2 standard deviations above the mean?

From the table the areas corresponding to these z values are 0.3413 and 0.4772 respectively. The area between them = 0.4772 – 0.3413 = 0.1359 which is the fraction of the screws in this range.

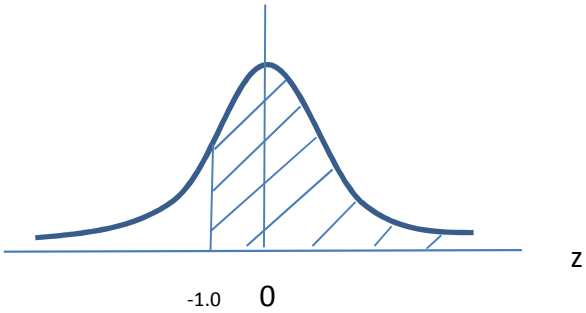
The actual number of screws = 0.1359 x 2000 = 272 screws.



c. How many screws are longer than 19.00mm?

For 19.00mm the z value = (19.00 – 20.00)/1.00 = -1.0.

A –ve z means it's on the left-hand side.



From the table z = 1.0 gives the area = 0.3413 that must be added to the 0.5 area on the right-hand side to get the total fraction longer than 19.00mm.

Total area = 0.3413 + 0.5 = 0.8413.

Total number of screws = 0.8413 x 2000 = 1683.

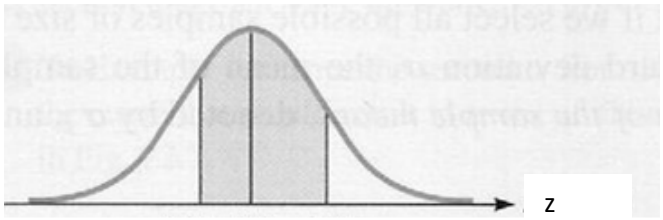
Example 2. The lifetimes of a type of 60W light bulb follow a Normal distribution. From measurements on a sample it is found that the mean lifetime is 300 days and the standard deviation is 60 days. A store sells 1000 of this type of light bulb to customers.

a. How many of the light bulbs will last between 270 and 420 days?

For 270 days z = (270 – 300)/60 = -30/60 = -1/2 = -0.5 (negative - on the left-hand side).

For 420 days z = (420 – 300)/60 = 120/60 = 2.0 (positive – on right-hand side)

On the Standard Normal Distribution this range of z corresponds to the following shaded area:

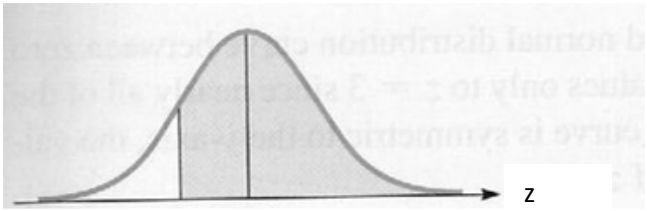


-0.5 0 2.0

From the table, the area corresponding to z = 0.5 is 0.1915 and the area corresponding to z = 2.0 is 0.4772.
Total area = 0.1915 + 0.4772 = 0.6687 =
Number lasting between 270 and 420 days = 0.6687 x 1000 = 669 light bulbs

b. How many will last more than 240 days?

240 days gives z = (240 – 300)/60 = -1 which lies below the mean.
The data we are interested in is the total shaded area above z = -1:

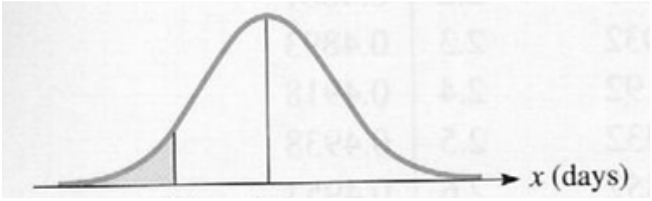


-1.0

From the table z = -1 corresponds to an area = 0.3413.
Total shaded area = 0.3413 + 0.5 = 0.8413.
Number of light bulbs that last more than 240 days = 0.8413 x 1000 = 841 light bulbs

c. How many will last less than 200 days?

200 days gives a z = (200 – 300)/60 = -1.667 which lies below the mean so the data we are interested in is the shaded area below z = -1.667:



-1.667

From the table this z corresponds to an area ~ 0.452
The shaded area below this is therefore = 0.5 – 0.452 = 0.048
Number of light bulbs lasting less than 200 days = 0.048 x 1000 = 48 light bulbs