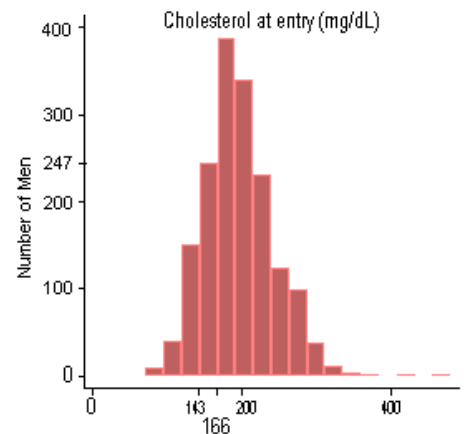


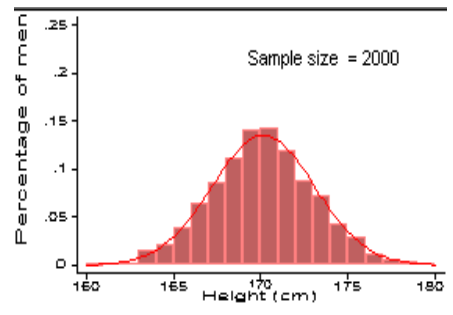
## The Normal Distribution

Quantitative variables are those that can take many values ranging from very small to very big values and they may also have negative or positive values. In other words their distribution can vary largely from small to large and from negative to positive values. Distribution of quantitative variables is displayed in a histogram.

For example the histogram opposite shows the distribution of the cholesterol levels of a sample of 1677 middle-aged British men obtained from a large study on the risk factors for mortality in London. The X axis displays the number of men and the Y axis displays the cholesterol level. The height of each bar shows the number of men whose cholesterol level was between the values at the base of the bar on Y axis. Therefore, if we summed the heights of all the bars in the histogram, we should get the total sample size which is 1677.

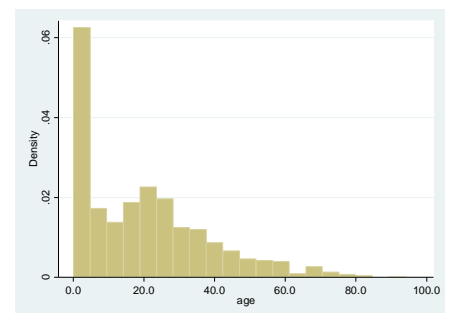


This distribution is not quite symmetrical or bell-shaped. If we take a bigger sample this may become more symmetrical. The graph show height of 2000 men. The histogram is symmetrical around its mean and like a bell. Usually the distribution becomes more symmetrical when the sample is larger. If we draw a line over this symmetrical histogram, we get a display of the **Normal distribution**.



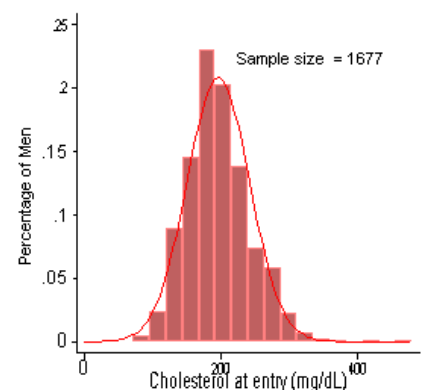
When the distribution of a quantitative variable is symmetrical around the mean of the variable i.e. about half of the values are on each side of the mean, we say the variable has normal distribution or it is normally distributed. The distributions of many quantitative variables are Normal especially when the sample size is large.

But some variables are not normally distributed even with large samples. For example the histogram opposite shows age distribution of 2975 burn patients in Sulaimani 2008. We can see from the histogram that the distribution is not symmetrical, we say the distribution is skewed.



The **Normal distribution** is the most important distribution of variables in statistics. It is also called *Gaussian distribution*, named after the German mathematician Carl Friedrich Gauss (1777-1855) who first described it. A Normal distributions has a distinctive bell shape. The mean and the standard deviation define the shape of the bell.

The graph below shows the histogram of the cholesterol levels in the sample of 1677 middle-age men with the Normal plot drawn around it. The histogram represents the sample but the normal plot represents the whole population represented by



this sample. This Normal distribution has a mean of 196 mg/dL and a standard deviation (SD) of 46 mg/dL

So let us see what will happen if change the mean and standard deviation. If we increase the mean, say to 220, the whole distribution (the bell) will move to right. If we decrease the mean, say to 150, the whole distribution will move to left. This is why mean is called a measure of location or central tendency of the distribution.

What happens if we change the standard deviation? If we increase the SD, the distribution (the bell) will be wider i.e. the values of the variables will be spread over a wider area. If we decrease the SD, the distribution will be narrower i.e. the distribution will be spread over a narrower area. This is why SD is called a measure of spread of the distribution around the mean. If the SD increases, the spread increases and the height of the distribution decrease. If the SD decreases, the spread decreases and the height of the distribution increases

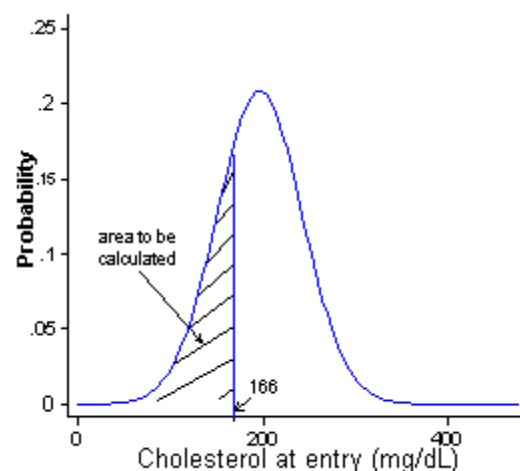
***Difference between histogram and normal plot***

The histogram represents the actual valued observed from one sample. The Normal plot is used to represent the distribution of everybody in the population based on the sample we have studied. The normal plot tells the probability of different values of the variable in the population i.e. on the long run. This is why the Y axis in the plot of a Normal distribution is called "probability".

<b>Histogram</b>	<b>Normal plot</b>
Shows the distribution of the values observed in a sample	Shows the distribution of values as we think they are in the population from which the sample was derived
The Y axis represents the frequency or the proportion (percentage) of the values in the sample	The Y axis represents the probability of the values in population
The sum of all the bars of an histogram is equal to 100% because all the observed values are included in the plot;	The area under the Normal curve is also equal to 100% because the curve covers all possible values
We interpret a histogram by reading the frequency or the percentage corresponding to a selected bar	we interpret a Normal plot by calculating the area spanning over a selected interval

***Why do we need a normal plot?***

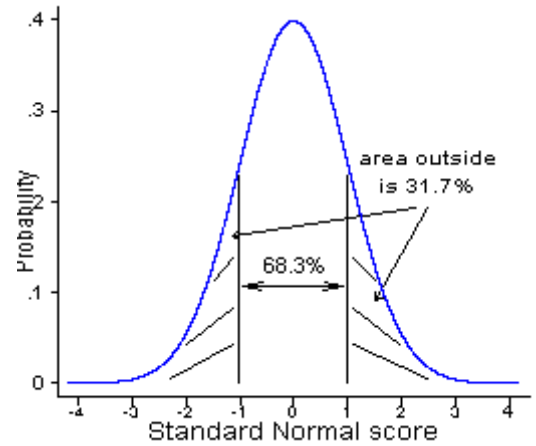
We need a normal plot in order to answer probabilities. For example, if we wanted to know the probability of a middle-aged man having a cholesterol level less that 166mg? If we talk about the sample, this is easy, we juts calculate the percentage of men who have cholesterol level below 166mg. but if we want to talk about all middle-age men, then we have to calculate this from a normal plot not from the histogram of the sample. This probability is shown on the normal plot opposite. The probability of having cholesterol below 166 is shown in the plot indicated by the marked area below the curve to the left of 166 mg. But how can



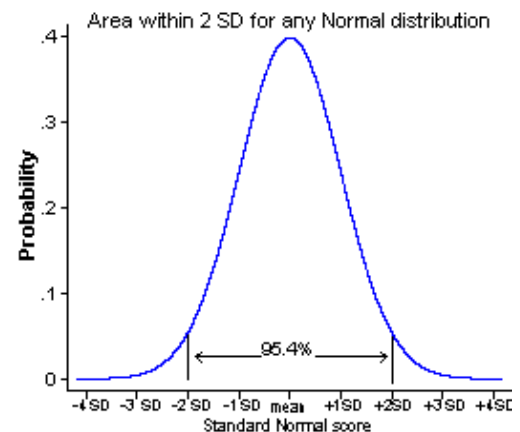
we compute this area? This is done by ready-made tables prepared from a special normal plot called standard normal distribution.

**The Standard Normal Distribution**

It is very difficult to calculate the area under the curve of a normal plot. Fortunately scientists have designed a special normal plot called standard normal distribution, and prepared tables from this plot which we can use. The **Standard Normal distribution** is a normal plot which has a mean equal to zero and standard deviation equal to one. The figure opposite shows a standard normal plot with a mean of zero and a SD of 1. Tables which are prepared from this plot are used to calculate probability of different values of variables which are normally distributed.



As we can see from this standard normal plot, 68% of the population falls within one SD from the mean i.e. 68% of the values are between 1 SD less than the mean to 1 SD more than the mean. The marked area under the plot indicates proportion of values which are outside this area. The probability of these values is calculated by referring to the standard tables such as shown below.



We can also see from the plot opposite that 95% of the values of any normally distributed variable falls within 2 SD from the mean.

The points on the X axis of this graph are called **standard normal scores** and they represent z values in the table. The values under P-lower represent the proportion of values below a corresponding z value. The values under P-upper represent the proportion of values above a corresponding z value.

Standard normal scores are calculated by subtracting the mean from the observed value and dividing the result by SD

$$z = (x - \mu) / \sigma$$

For example what is the z score for cholesterol level of 166 if the mean is 196 and SD is 46?

$$Z = (166 - 196) / 46 = -0.65$$

What does this mean?

If we look at the table, under z we search for -0.65, opposite this under p-lower we see 0.2578 which is the probability; it means that 25.8% of the population have a cholesterol level below 166mg. in other words the probability of having cholesterol level lower than 166 in

z	P-lower	P-upper
-3.00	0.0013	0.9987
-2.00	0.0228	0.9772
-1.00	0.1587	0.8413
-0.65	0.2578	0.7422
-0.50	0.1587	0.8413
0.00	0.5000	0.5000
0.50	0.6915	0.3085
0.65	0.7422	0.2578
1.00	0.8413	0.1587
1.50	0.9332	0.0668
1.64	0.9495	0.0505
1.96	0.9750	0.0250
2.00	0.9772	0.0228
3.00	0.9987	0.0013
4.00	0.9997	0.0003

this population is 26%.

Another example from the table and graph above: what proportion of the sample is below 2 SD ( $z=-2$ ) from the mean? Read P-lower opposite  $z= -2$  which is 0.0228. This means 2.3% of the sample is less than the mean by more than 2 SD. In other words the probability of a value being more than 2 SD less than the mean is 2.3%.