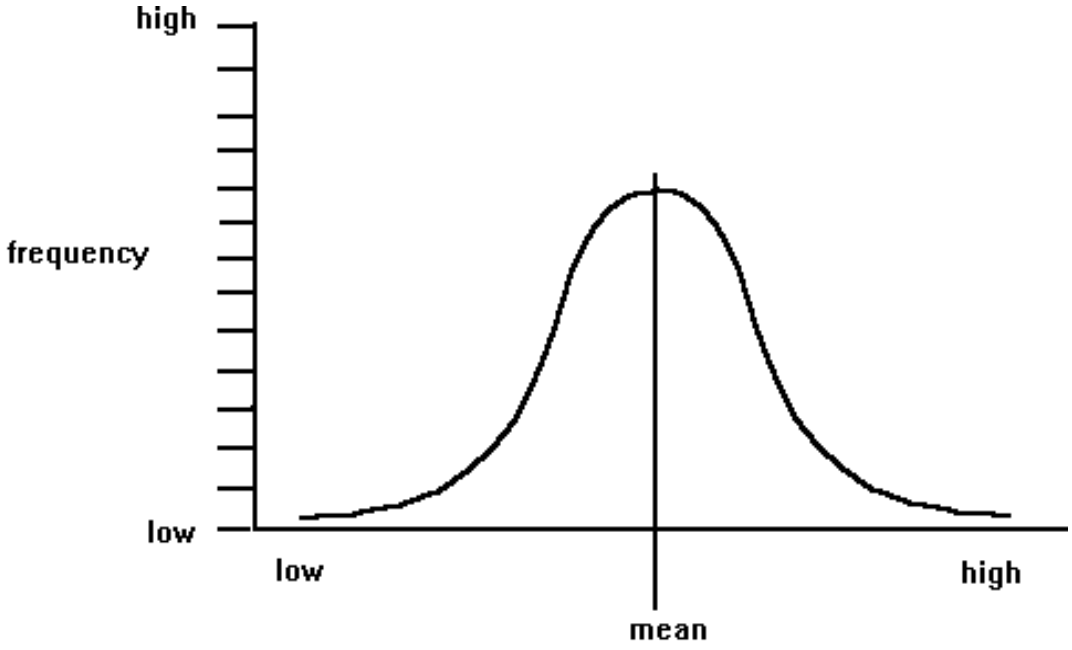
THE NORMAL CURVE AND "Z" SCORES:

The Normal Curve:
The "Normal" curve is a mathematical abstraction which conveniently describes many distributions of data in the natural and biological sciences. For example, the frequency distribution of adult heights can be approximated (or "modelled") by the normal curve; most people are around average height, but some people are taller and some are shorter. The more extreme the height (i.e., shorter or taller) the less commonly it occurs. Exceptionally tall or short individuals are very unusual. There are many other circumstances in nature where most people or "observations" are in the middle, with a few instances at either extreme of the distribution.

The normal curve is a theoretical ideal; the real-life frequency distribution of height, for example, is not perfectly normally distributed. However, the similarities are good enough for us to use the normal distribution as a description of height, instead of the real thing. The chief advantage of this is that it greatly simplifies matters: instead of having to describe the actual distributions of each and every particular characteristic that interests us, we can use the normal distribution as a convenient rough and ready description for all types of things.

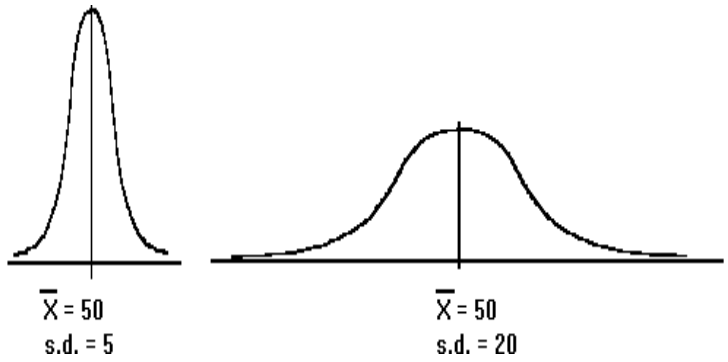


Properties of the Normal Distribution:
1. It is bell-shaped, and asymptotic at the extremes (that is, it approaches the horizontal axis at either extreme, but never actually touches it).
2. It is symmetrical around the mean of the distribution.
3. If a set of observations are normally distributed, their mean, median and mode all have the same value.
4. The normal curve can be specified completely, once its mean and standard deviation are known.
5. The area underneath the curve is directly proportional to the relative frequency of observations.

Relationship between the Normal Curve and the Standard Deviation:

Remember that a set of scores has a mean (a measure of "typical" performance) and a standard deviation (a measure of the extent to which the scores are spread out around the mean).

The "Normal Curve" is really a family of curves, since different values for the mean and s.d. will produce differently-shaped normal curves.



However, all normal curves share the same property: the standard deviation cuts off a constant proportion of the distribution of scores.

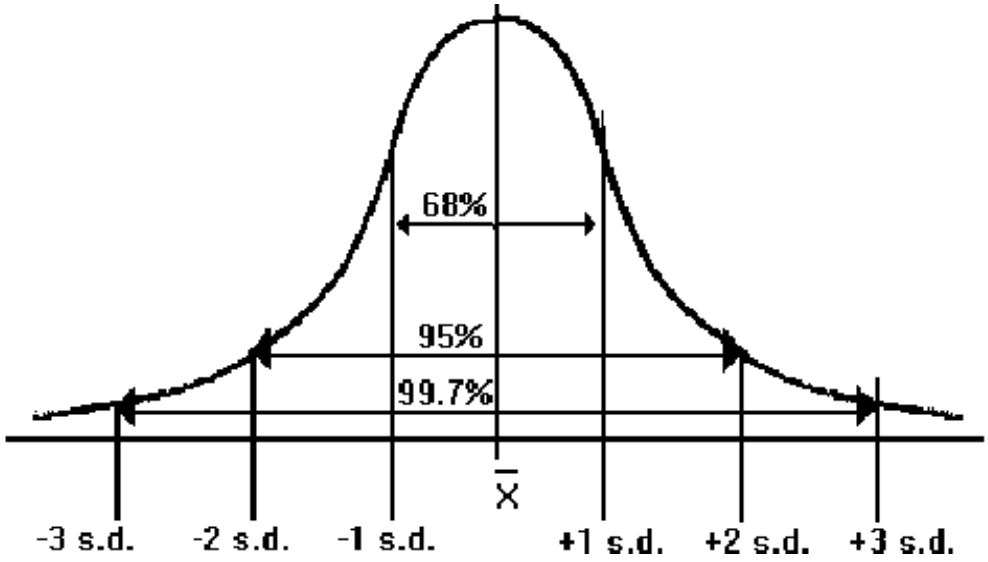
Therefore, if we know a set of scores has a normal distribution, and we know the mean and the standard deviation of those scores, then we can know how many scores fall within any particular limits that we are interested in.

It's worth knowing the following facts about the relationship between the normal distribution and standard deviations:

(a) if a set of scores are normally distributed, then roughly 68% of them fall within the limits of the mean plus or minus one standard deviation. For example, I.Q. is roughly normally distributed, and has a mean of 100 and a standard deviation of 15. This means that about 68% of the population would be expected to have an I.Q. between 85 (100 - 15) and 115 (100 + 15).

(b) Similarly, roughly 95% of a set of normally-distributed scores fall within the range of the mean plus or minus two standard deviations. Thus 95% of the population would be expected to have an I.Q. between 70 and 130.

(c) Roughly 99.7% of a set of scores fall within the range of the mean plus or minus 3 standard deviations. Thus, 99.7% of the population would be expected to have an I.Q. somewhere between the limits of 55 and 145.



These are only rough approximations: in real life, it might be that the true proportions are somewhat different from these theoretical ideals. However, these are useful facts to know; given just three pieces of information (the mean and s.d. of a set of scores, plus the fact that the scores are normally distributed) you can to a large extent reconstitute the original data and work out quite a lot about its characteristics! Thus, if we know that a person's I.Q. is 130, we have a fair idea of how he stands in relation to the rest of the population: we know that

such a high I.Q. is quite unusual. (130 is 2 standard deviations above the mean. 95% of the population have an I.Q. between 70 and 130, so 5% must have an I.Q. outside of these limits. Since the normal curve is symmetrical, this means that half of 5% must have an I.Q. of less than 70, and the remaining half of this 5% must have an I.Q. above 130. Therefore, we can say that our individual is in the top 2.5% of the population, as far as his I.Q. is concerned).

"Z"-scores:

It is useful to know that roughly 68% of cases fall within the limit of plus or minus one standard deviation either side of the mean. However, we often want to know how many scores fall within other limits, ones not defined in terms of whole standard deviations.

Also, we often want to compare subjects' performance on one test with their performance on another test, in circumstances where the two tests are not measured in the same way. For example, suppose that a subject obtained the same score of 130 on two tests, A and B, but that on test A subjects had a mean of 100 and a standard deviation of 10, whereas on test B, subjects had a mean of 70 and a standard deviation of 5. Our subject's score is clearly better than average on both tests and it looks as if he's done much better than average on the second test. However, it would be nice to be able to compare his performance on the two tests in a more precise way than this, and the use of z-scores enables us to do so.

Z-scores provide a way of standardising a person's score, with reference to the rest of the scores in that group (i.e., taking into account the mean and standard deviation of that group's set of scores).

A z-score expresses a particular score in terms of how many standard deviations it is away from the mean of the set of scores.

$$z = \frac{X - \overline{X}}{s}$$

where X is our raw score (the particular score we wish to convert to a z-score);
$\overline{X}$ is the mean of the scores;
and s is the standard deviation of the scores.

This formula has two components:

(a) first, we find the difference between our raw score and the mean; we do this by subtracting the mean from our raw score (X- $\overline{X}$).

(b) second, we divide this difference by the standard deviation. This tells us how many standard deviations (and bits of standard deviations) our raw score is away from the mean of the scores. The logic behind doing this is that the significance of a particular deviation of a score from the mean of the set of scores depends on how spread out the scores are in the first place. For example, imagine having a raw score of 55 and a mean of 50. We have a difference of 5. If the standard deviation is tiny (say 0.01), then this difference of 5 means a lot - it represents a lot of standard deviations and hence suggests that our score is very different from the mean. If the standard deviation is big (say 20) then a difference of 5 isn't very much at all - most of the scores will be quite a bit different from the mean.

By converting a raw score into a z-score, all we are doing is changing the scale: we have now converted our score so that it is on a scale measured in standard deviations. Our raw score was a measurement on a raw score scale, which had a particular mean and standard deviation. All we have done by turning our raw score into a z-score, is to change the scale on which the score is expressed: we have transformed our raw score so that it is now on a z-score scale which always has a mean of 0 and a standard deviation of 1. (This makes sense when you think about it: by turning a raw score into a z-score, we are re-defining it in terms of its difference from the mean of the set of scores to which it belongs. Therefore if a score is the same as the mean of its parent population, it must have a z-score of 0, since it does not differ from the mean! If it is one standard deviation away from the mean, t must have a z-score of 1).

The advantages of turning a raw score into a z-score are

(a) on a z-score scale, the relationship between a score and the distribution of scores to which it belongs is made much clearer. For example, if you told me that a person had a raw score of 70 on a test for which the mean was 60 and the standard deviation was 5, I would know that the person had obtained a relatively high score, but his precise standing on the test in relation

to his peers would not be immediately apparent to me. However, if you converted the same score into a z-score, and told me that this person's z-score was 2, then I would instantly know that this was exceptionally good performance. (Remember that z-scores are measured in terms of standard deviations. If someone gets a z-score of 2, they are 2 standard deviations above the mean of the population of scores. Since I know that 95% of the population would probably have scored between the mean plus or minus 2 standard deviations, this means that only 2.5% of the population would have been likely to have produced a score this high).

(b) we can convert *any* set of scores which are normally distributed into a standardised set of z-scores with a mean of 0 and a standard deviation of 1. This is very convenient when one remembers that the area under the normal curve is proportional to the number of scores. Once scores are turned into z-scores, the area under the normal curve always the same value: 1. (This is expressed as a proportion; you might find it easier to think of it as a percentage, i.e. as 100%).

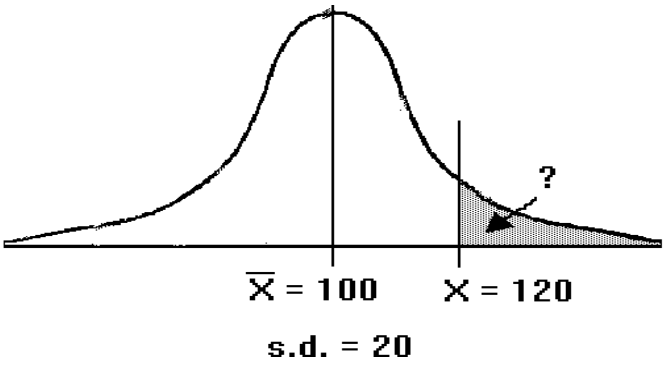
If you find this hard to grasp, try to understand the following examples. Firstly, what does the entire normal distribution represent? It represents all (i.e., 100%) of the scores (1, if expressed as a proportion). Secondly, what part of the curve represents 50% of the scores? half of it (i.e., 50%, or .5 if expressed as a proportion).

In the back of many statistics, you will find a table entitled "Area under the normal curve". For any particular z-score, the table will tell you two things. The column entitled "area between the mean and z" will tell you what proportion of the area under the normal curve lies between the mean of the set of scores and your z-score. The column entitled "area beyond z" will tell you what proportion of the are under the normal curve lies beyond your z-score. The area under the curve is proportional to the number of scores falling within those limits; and the z-score is directly equivalent to your original raw score. Thus, if have a raw score and you want to know what proportion of scores fall beyond it (for example, above it), or if you want to know what proportion of scores fall between it and the mean of the distribution, all you have to do is to (a) convert your raw score into a z-score, and then (b) look up the relevant area under the normal curve in a table, to find out what proportion of scores fall within the limits that you are interested in.

Examples of the use of z-scores:

(a) If a set of scores on a test has a mean of 100 and a s.d. of 20, what proportion of subjects are likely to have obtained scores of 120 or more?

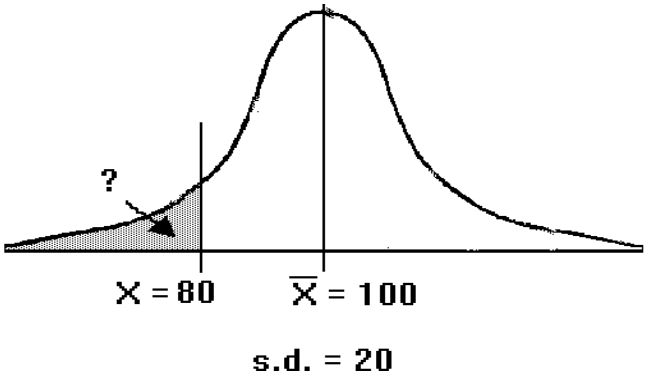
It always helps to draw a rough graph so that you can keep track of what you are trying to do.



(i) Convert the raw score of 120 into a z-score. (120-100)/20 = 20/20 = 1.00

(ii) Look up this z-score in the table. We want to know how many people scored 120 or more, and so the relevant column in the table is the one entitled "area beyond z". The area beyond a z-score of 1.00 is .16. In other words, .16 of the population would be expected to score 120 or above. To express this as a percentage, simply multiply the proportion by 100: thus 16% of the population would be expected to score 120 or higher. If you want to know how many subjects, multiply the proportion by the total number of subjects. Thus, if we had 200 subjects, .16 * 200 = 32 subjects would be expected to score 120 or higher.

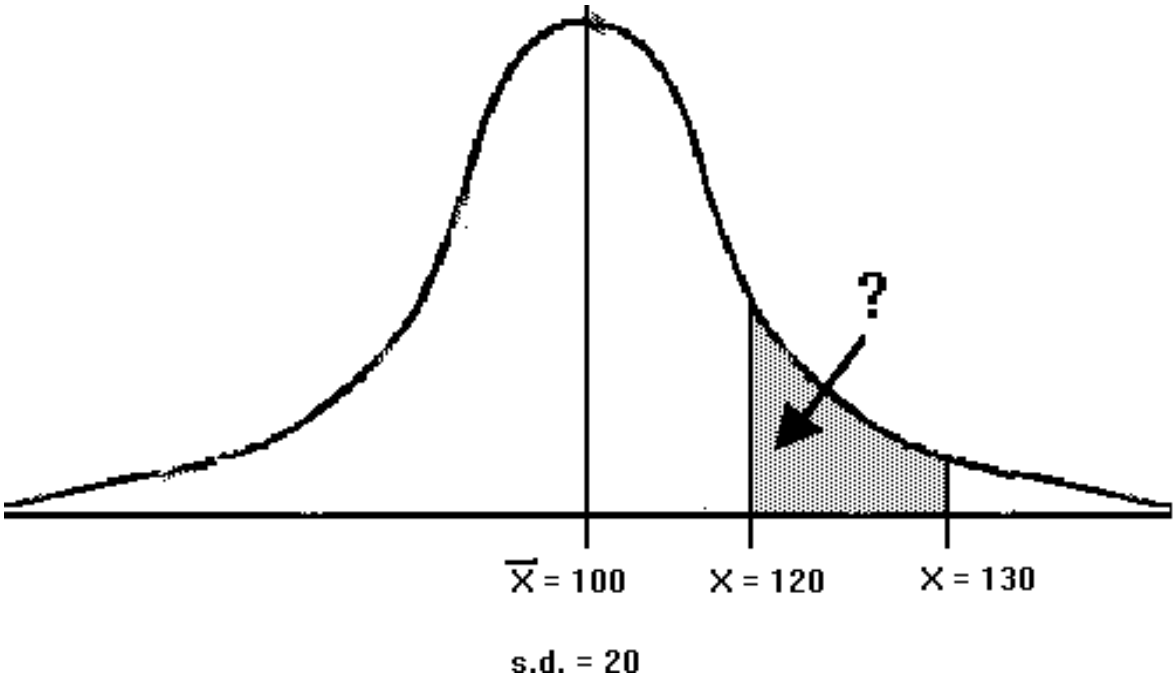
(b) If a set of scores have a mean of 100 and a standard deviation of 20, what proportion of subjects would be expected to score 80 or less?



(80-100)/20 = -20/20 = -1.00. This is exactly the same type of problem as before, despite the minus sign! Since the normal curve is symmetrical, the table of areas only needs to show the areas corresponding to positive z-scores. The area beyond -1.00 is effectively the same as the area beyond +1.00, so just ignore the negative sign when you use the table. 0.16 of scores lie beyond a z-score of 1.00, and hence .16 of scores lie beyond a z-score of -1.00. We can conclude that 16% of subjects scored 80 or less.

(c) What proportion of scores fall between two scores?

For example, if a set of scores have a mean of 100 and a standard deviation of 20, what proportion of scores would be expected to fall between 120 and 130? This is slightly trickier - again, it helps if you graph the problem you are trying to solve.



The graph makes it apparent that what we are trying to find is the area between the z-scores corresponding to 120 and 130.

(i) First, find the z-scores for 120 and 130. The z-score for 120 is (120-100)/20 = 1.00. The z-score for 130 is (130-100)/20 = 1.50.

(ii) Look up the "area beyond z" for the z-score of 1. This is .1587.

(iii) Look up the "area beyond z" for the z-score of 1.5. This is .0668

(iv) The area that we want is the area beyond 1.00, minus the area beyond 1.50. This corresponds to the proportion of scores falling between these two limits. .1587 - .0668 = .0919. Thus 9.19% of subjects would be expected to score between 120 and 130.

(d) What raw score cuts off the top 20% of scores in our population of scores?

This is the inverse type of problem to the ones mentioned above. Now, we are given the proportion and have to work backwards to find the raw score that separates this proportion from the rest of the scores. To do this, we need a different formula - one that converts a z-score into a raw score:

$$X = \overline{X} + (z * s)$$

This says "the raw score equals the mean of the raw scores plus the z-score times the standard deviation of the raw scores". (Strictly speaking, the brackets are superfluous, but I've included them to emphasise that you work out the bit within the brackets first).

Suppose our set of scores has a mean of 100 and a standard deviation of 20. We now need to do the following:

(i) We want to find the z-score beyond which 20% of the scores fall. 20% as a proportion is .20. Use the "area beyond z" column in the table, and find .20. Find the z-score which corresponds to this value. It is 0.84

(ii) Insert this z-score into the equation above.

X = 100 + (0.84 * 20)

X = 116.8. In other words, 20% of the scores would be expected to be 116.8 or higher.