The Normal Distribution

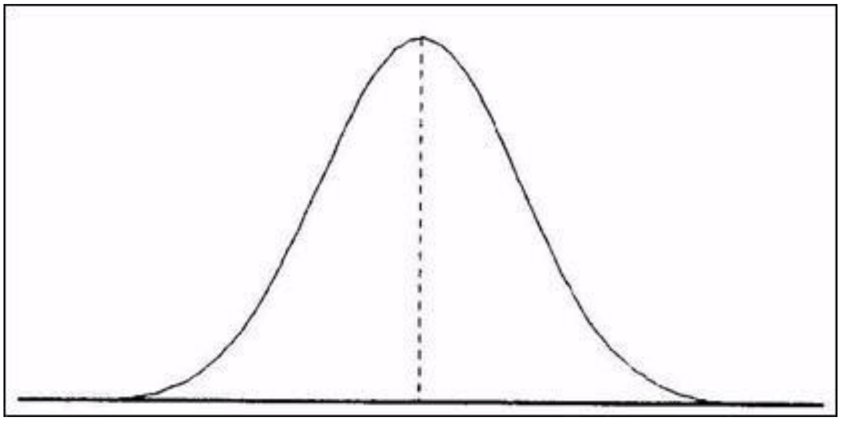
It might not always be obvious but we have been blessed with one fantastic stroke of luck as far as the study of statistics is concerned: the normal distribution. For centuries now mathematicians, statisticians and business people have realised that life isn't always as random and as difficult as it might seem. What we are about to look at first of all, then, is what the normal distribution is, what it means, how it can help us and how we can use it in our every day lives.

Many biological characteristics conform to a Normal distribution closely enough for it to be commonly used - for example, heights of adult men and women, blood pressures in a healthy population

http://bmj.com/collections/statsbk/2.shtml

What the normal distribution looks like

Here is a basic normal distribution curve:



What this graph shows us is that the X value (for example, heights of people, the blood pressure of healthy people, the sizes of shoes in stock in a shoe shop, the amount of water in bottles of mineral water …) are distributed in a symmetrical way along the Y axis, the frequency.

If a data set really is based on the normal distribution then we are lucky: we are lucky because everything about that data set is predictable; and because everything's predictable, it makes analysing and using them so much easier.

Check it out: do you want to prove that the heights of people are normally distributed? Then measure everyone in your year group … let me stress that you measure as much of the year group as possible otherwise you won't see the normal distribution so clearly: a minimum of 30 measurements anyway.

It also helps to turn your figures into a picture ... plot your data on a graph: this is what I call the **golden rule of data analysis**, plotting your raw data on a graph to see what it looks like.

This really works!

Honestly, the normal distribution really does work. The more things you look at, the more you'll find it. Go to a shoe shop and take a look at their stocks, if they'll let you, and you will find that there are a lot more shoes of adult size 4 - 8 than there are of adult size 10 - 12: the reason is all down to the normal distribution curve. The big bulk of people have medium size feet (eg 4 - 8) and only a few people have big feet: they are exceptional!

Similarly, if you went to a mineral water bottler and were able to measure the contents of their bottles you would find that they didn’t all contain exactly 1 litre or 330 ml … some would have 983 millilitres, others would have 1001 millilitres and a lot would have 1000 millilitres, or 1 litre. I’d be willing to bet that if our sample size were big enough, the frequency of the amounts of water in the bottles would be normally distributed.

As a matter of interest, if you have ever visited an old town or village in the UK or elsewhere in Europe did you notice that the doorways in medieval buildings are a lot smaller than the doors in modern buildings? The height of the average new door in the average British house is around 2 metres but in a medieval building it may be around 1½ metres. Medieval people were smaller than modern people; but builders have always taken the normal distribution of the heights of people into account … is this obvious? How do I know that builders have always taken the normal distribution into account even though the normal distribution hadn’t been written down and analysed in full until the last couple of centuries or so?

Your Turn

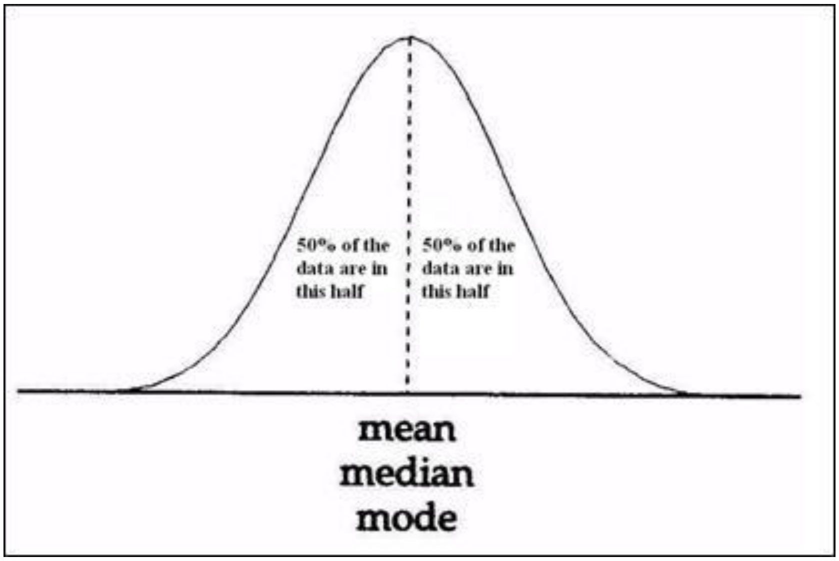
Find as many examples of where the normal distribution fits before you move on to the next section.

Why is the normal distribution so good then?

Because the normal distribution curve is so symmetrical there are two batches of things that are always true:

a in normally distributed data the mean = median = mode

Graphically, that means:

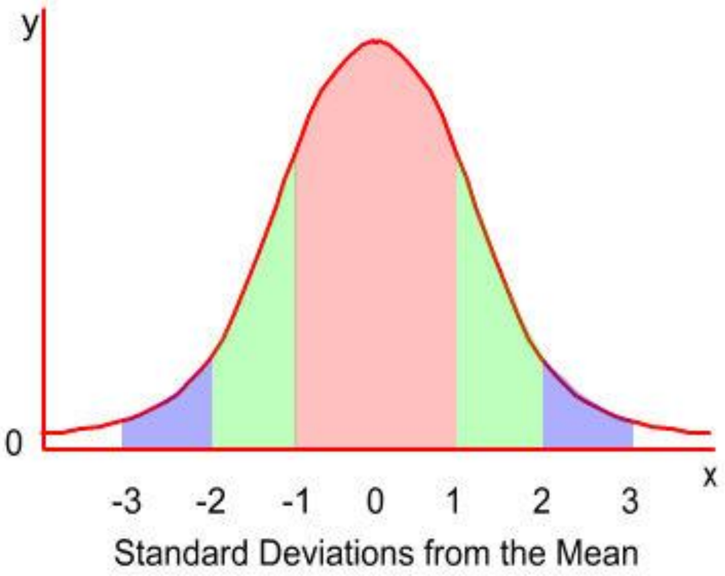

50% of the data are in this half
50% of the data are in this half
mean
median
mode

b we’ve already talked about standard deviations (SD) so we can now start to use them. Here are some fascinating facts about the link between the standard deviation and the normal distribution. In normally distributed data:

- 68.26% of the data will be found within one SD either side of the mean (±1SD)
- 95.44% of the data will be found within two SD either side of the mean(±2SD)

- 99.74% of the data will be found within three SD either side of the mean (±3SD)

This really is brilliant news! It's brilliant news because for *all* normal distributions about two thirds of all the results we get when a data set is normally distributed will be found within the ±1SD range, just over 95% of the data set will be within ±2SD of the mean and virtually 100% of the data will be within ±3SD of the mean. That's *always*, with *all* normally distributed data. The following graph summarises this situation quite well: where red = ±1SD, green = ±2SD and blue =±3SD



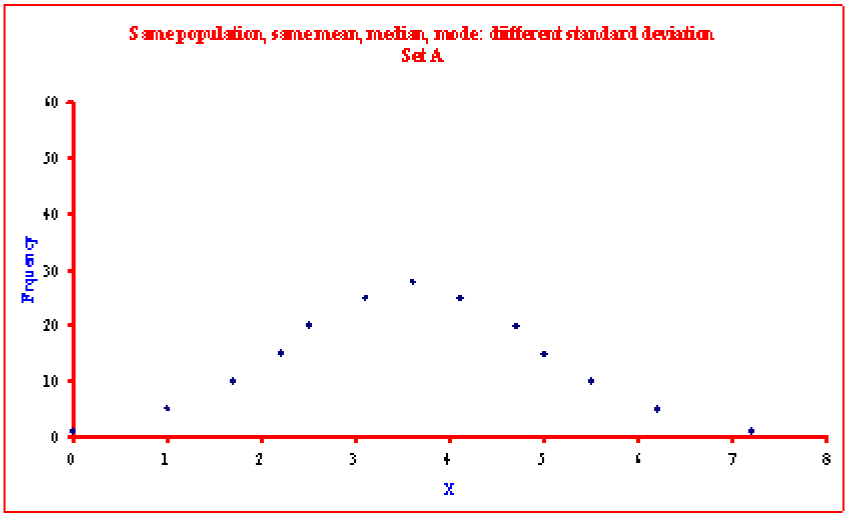
The Statistician's Approach to the Normal Distribution

Of course, there are data sets that are not normally distributed but generally, statisticians, and that means you and me, almost always begin by assuming that they are unless and until we discover otherwise. So, when we've cracked the normal distribution code, we've cracked a lot of statistical analysis. Honest, it's true!

Let's put Standard Deviation and the Normal Distribution Together

OK, let's put our money where our mouth is and prove everything we've said so far: that SD and the normal distribution do go together as we have said. Here is a data set with variables X and Ya: we've presented the data on a graph, too. We are **required** to **calculate** the mean and standard deviation of the X values; then we are required to calculate the values of the mean, $\bar{x}$, ±1SD, $\bar{x}$ ±2SD and $\bar{x}$ ±3SD

X	Ya
0	1
1	5
1.7	10
2.2	15
2.5	20
3.1	25
3.6	**28**
4.1	25
4.7	20



5	15
5.5	10
6.2	5
7.2	1

Here are the Mean and SD:

Mean	3.6000
SD	1.263

The values of $\bar{x}$ ±SD, with their workings, are:

	From	to
$\bar{x}$ **±1SD**	3.60 - 1*1.263 = 2.337	3.60 + 1*1.263 = 4.863
$\bar{x}$ **±2SD**	3.60 - 2*1.263 = -1.074	3.60 + 2*1.263 = 6.126
$\bar{x}$ **±3SD**	3.60 - 3*1.263 = -.187	3.60 + 3*1.263 = 7.389

Let’s do some exercises now.

Your Turn

Here are the results of some research that has been carried out, your job is to calculate the ranges covered by the ±SD we’ve just been talking about.

1 The heights of 92 young men are believed to be normally distributed with a mean of 1.75 metres and a standard deviation of 0.25 metres. Fill in the table below based on the information just given:

	From	to
$\bar{x}$ **±1SD**		
$\bar{x}$ **±2SD**		
$\bar{x}$ **±3SD**		

2 A baker has recorded the weights of loaves, taken at random, coming out of one of his ovens and she has found the following:

weight of loaf **(grammes)**	**frequency**
772.5	3
777.5	17
782.5	44
787.5	100
792.5	141

797.5	192
802.5	191
807.5	150
812.5	90
817.5	42
822.5	14
827.5	9

a plot these data on a graph

b calculate the mean and standard deviation of the weights of the loaves

c complete the following table:

	From	to
$\bar{x}$ **±1SD**		
$\bar{x}$ **±2SD**		
$\bar{x}$ **±3SD**		

d identify the $\bar{x}$ ±SD ranges on your graph.

e if we assume that the distribution of the weight of loaves is normal

i what is the probability that a loaf will weigh less than 800 grammes?

ii what is the probability of a loaf weighing more than 815 grammes?

(Question adapted from Oakshott pages 76 - 77)

Have you Spotted Something Really Interesting?

Take a look, for example, at your answers to question 1 that you've just done … then look at your answers to part 'c' of question 2 … compare that with the worked example for Ya … have you spotted something really interesting?

What's really interesting is that for the $\bar{x}$ ±SD work, all we need to know are the values of 'x', we don't need any 'y' values at all. In other words, when a data set is normally distributed, we work only from the 'x' values to find the **±**SD values: we sort out the 'y' values and interpretation completely separately, like we did with question 2 e parts i and ii.

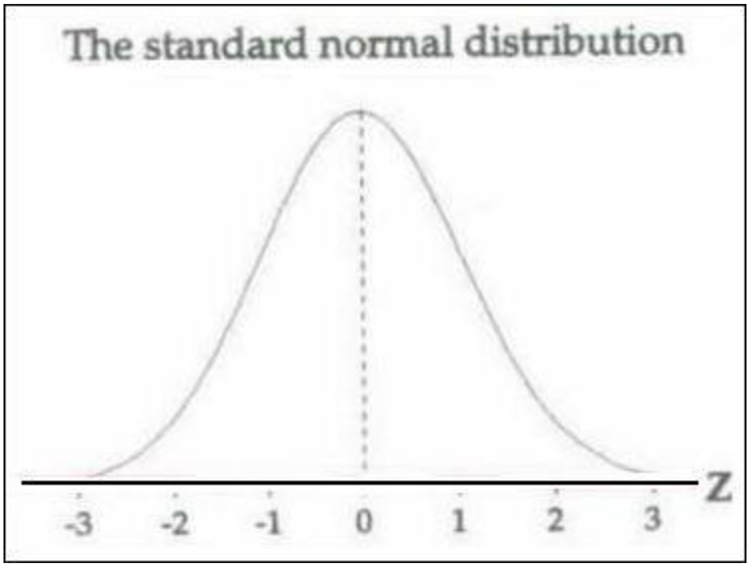
Now that takes us onto something really, really interesting: the standard ***normal*** distribution.

The Standard Normal Distribution

The normal distribution behaves exactly as we have seen: perfectly symmetrical about the mean, 50% of the data set to be found in each half of the distribution, $\bar{x}$ ±SD ranges that can help us with some of our analysis and so on. Hear this, though, there is such a thing as a standard normal distribution that helps us even more than the ordinary normal and here's how.

The standard normal distribution has a mean of 0 and a standard deviation of 1: ***always***. Exactly half of a standard normal distribution set will be bigger than the mean value and exactly half of the standard normal data set will be smaller than the mean.

When we plot the standard normal distribution on a graph, we find that it looks as follows, with the values on the 'X' axis being, usually -3, -2, -1, 0, 1, 2, 3. We normally don't go beyond ±3 because these numbers represent the number of SD away from the mean and we already know that almost 100% of a normally distributed data set is found within ±3SD of the mean.



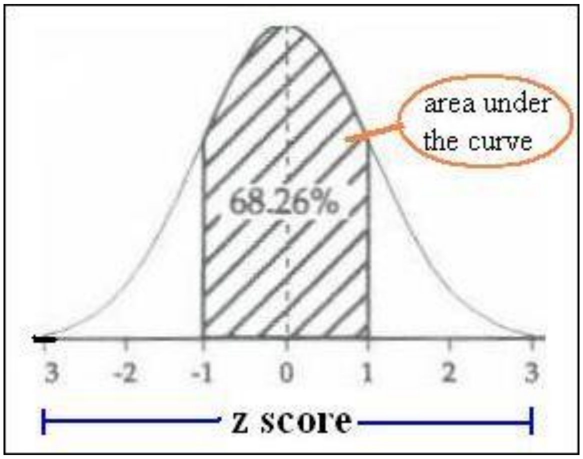
The best way to understand the standard normal distribution is to use it; and there are two parts to using this distribution; and we'll look at each in turn:

- normal distribution curve areas

- z score or the standard normal random variable

Normal Distribution Curve Areas

There is a table called the normal distribution curve areas: we need them and this is why and what we do with them.



The area under the normal curve at $\bar{x}$ ±1SD is 68.26%: that is, in a normally distributed data set, 68.26% of all 'X' values will lie within ±1SD of the mean.

Z Score or the Standard Normal Random Variable

We are concerned here with standardising the normal distribution: that is, using the idea of the standard normal distribution to solve problems with any other normally distributed data. We use the standard normal distribution because not many data sets, even though they might be normally distributed, have a mean of 0 and an SD of 1. Still, all we need to do is to carry out a quick and easy calculation and we can use the standard normal distribution … to calculate the z score in any situation, use this formula:

$$\text{z score} = z = \frac{x - \mu}{\sigma}$$

where:

x is a value from a data set

μ is the average of the data set (strictly speaking it's the population average and we should use $\bar{x}$

σ is the standard deviation of the data set

The value of z tells us how many SDs our data point is away from the mean:

- a positive value of z tells us that the data point is to the right of the mean
- a negative value of z tells us that the data point is to the left of the mean

Example: imagine that we have found that the value of 'x' is 10, the mean of the data set from which the value came is 2 and the SD of that data set is 6.4: **calculate** the z score from this information

$$z = \frac{10 - 2}{6.4} = \frac{8}{6.4} = 1.25$$

Using the normal distribution curve areas table, we find that a z score of 1.25 means? Read table

Your Turn

Get to understand and learn the z score formula by using it: try these.

1 you are told that the value of 'x' is 25.75 and that the average of the data set that this value comes from is 12 and its SD is 5.5: **calculate** the z score from these data.

Answer:

$$z = \frac{25.75 - 12}{5.5} = \frac{13.75}{5.5} = 2.5$$

2 In a previous question we were given some bread baking data, repeated below. **Calculate** the z scores for loaf weights of

- 784 grammes
- 800 grammes
- 825 grammes

and **say whether** the z score values you obtain seem reasonable from our discussion up to this point.

A baker has recorded the weights of loaves, taken at random, coming out of one of his ovens and she has found the following:

weight of loaf	***frequency***
(grammes)	
772.5	*3*
777.5	*17*
782.5	*44*
787.5	*100*
792.5	*141*
797.5	*192*
802.5	*191*
807.5	*150*
812.5	*90*
817.5	*42*
822.5	*14*
827.5	*9*

average	800.003
standard deviation	10.018

Answers:

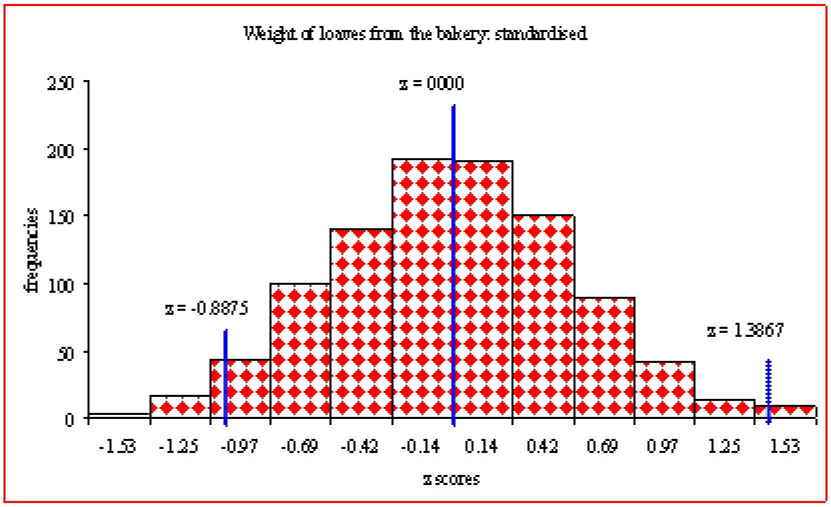
$$Z = (784 - 800)/10.018 = -1.597$$

confirm that you appreciate that the z score for a weight of 784 grams is **negative**.

Z = (800—800)/10.018 = 0

Z = (825—800)/10.018 = 2.496

We should be able to see that these values fit very well with what we did when we looked at the baker's data before. You should already have a graph of these data and can readily see how these z scores really do fit perfectly the picture we have created. Here's our graph of the bakery data; look at these z scores there. Notice the scale on the 'x' axis, the weights, we have standardised those values by converting all of the weights given in the table to its z score and using that as the scale. This makes applying the results to this question more understandable.



Using the Standard Normal Distribution

Since the standard normal distribution is so symmetrical and predictable, we can use it for all sorts of things now. We can use the standard distribution to help us to answer such questions as:

what proportion of my loaves are heavier than, say, 810 grammes or less than, say, 775 grammes or even exactly 800 grammes?

To do this we don't need to do much more than we've done already. We have our raw data, we have the mean and the standard deviations and we have our z score all we need now is the table of Normal Distribution Curve Areas and apply what we find to our z scores. We'll come back to the baker shortly but first let's go back to another question we've already done.

Example: earlier we were given an 'x' value of 10, a mean of 2 and an SD of 6.4. From this information we found the z score of 1.25.

Using the table of Normal Distribution Curve Areas, we know that z = 1.25 has a value of 0.3944. We might want to know, using this z score, how many things in the data set will be

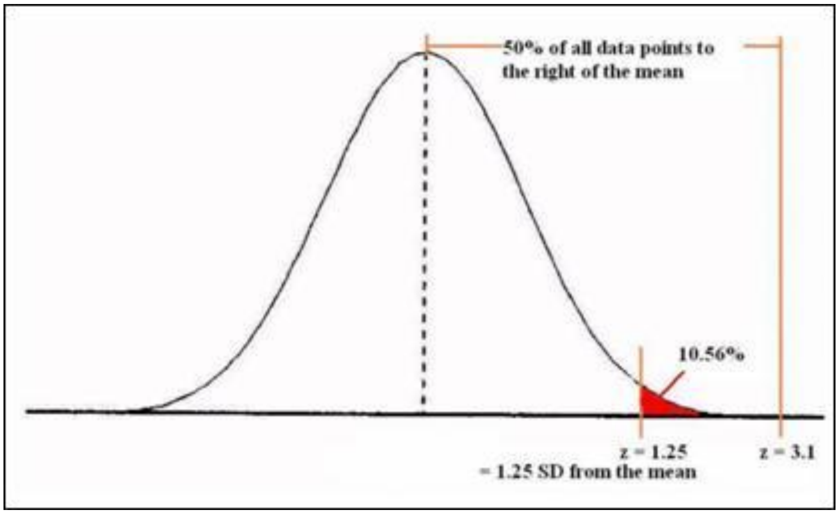
- bigger than 10
- smaller than 10

Let's put this onto a graph to appreciate the power of the standard normal distribution and its ability to help us here.

It’s important to keep reminding ourselves that half of a normal distribution contains 50% of all of the data points in a data set. Hence, 50% of all data points will be greater than the mean and 50% of all data points will be less than the mean: that’s so simple, but worth marks in an exam, honest!

How many data points are bigger than 10, to answer the question, though? Well, the graph shows that 10.56% of the data points are bigger than 10

How many data points are smaller than 10? Since we know that 10.56% of the data points are bigger than 10, it MUST be true that 100% - 10.56% of all data points are smaller than 10 = 89.5%



Where did we get our answers from?

Please note, most importantly, the areas given in our standard normal distribution table are for only HALF of the distribution; but since the distribution is symmetrical, that doesn’t matter except that we add on an extra bit to our calculations to make sure we take this into account.

To discover the proportion of the data set that will have a value of greater than 10, the calculation is:

0.5 - 0.3944 = 0.1056 = 10.56% of the data set will have a value of more than 10

Notice, we subtract the 0.3944 from 0.5000 because half of a standard normally distributed data set is bigger than the mean and the answer of z = 1.25 puts that data point at the mean plus 1.25SD. So, to find the proportion of ‘x’ values greater than 10 in this case, we need to subtract 0.3944 from 0.5000. Study the graph carefully and that should help if you have any doubts:

See Question 1 ‘a’ for help on what to do with the 0.500 when z is negative if you can’t work it out yet.

The best thing to do now is to work, work, work with these ideas. Here are four questions: we give you fully worked answers to two of them and skeleton answers to the other two. Aw, aren’t we kind!!

Your Turn

1 Back to the bakery:

what percentage of loaves will be

a larger than 775 grammes

b smaller than 810 grammes

c between 805 and 820 grammes

average	800.0000
standard deviation	10.018

2 The average amount of milk in a bottle marked for sale as containing one litre is 1.025 litres. If the distribution of milk contents is normal and the standard deviation of amounts is 0.05 litres, what is the percentage of bottles likely to contain

a more than 1 litre

b less than 1 litre

3 A certain type and make of car tyre has a mean life of 60,000 kilometres and a standard deviation of 8,300 km:

a what proportion of tyres will need to be placed under guarantee if the manufacturer guarantees all tyres for 45,000 km?

b what proportion of tyres are likely to need replacing between 50,000 and 55,000 km?

4 Good news! So said an advertisement from the MatVest Bank … use your credit card on purchases over £x in November and you will receive a free gift … however, it needs to restrict the number of gifts to only 7.5% of all credit card customers. If the average expenditure is £135 with a standard deviation of £55, how much should £x be?

Answers

1 a) Z = (775–800)/10.018 = –2.496 therefore, 100–.64 = 99.36%

b) Z = (810−800)/10.018 = .998 therefor 84%

c) For this answer, we need to do two calculations, as you see here:

Z = (805−800)/10.018= .449 therefor = 69%

Z= (820−800)/10.018 = 1.996 therefor = 97%

So the percentage of loaves between 805 and 820 grammes is 97% — 69% = 28%

2 a) $z = \frac{1-1.025}{0.05} = \frac{-.025}{0.05} = -0.5$

the percentage of bottles containing more than 1 litre of milk is, therefore,

= 0.6915 = 69.15%

b) $z = \frac{1-1.025}{0.05} = \frac{-.025}{0.05} = -0.5$

the percentage of bottles containing less than 1 litre of milk is, therefore,

= 0.3085 = 30.85%

Make sure you agree with this … maybe a bit tricky for you: sketch your own graph to see what's going on if you're unsure.

3 a) $z = \frac{45{,}000-60{,}000}{8{,}300} = \frac{-15{,}000}{8{,}300} = -1.8072 \approx -1.81$

therefore = 3.51% of all tyres are likely to need replacing under warranty. We would advise the manufacturer to limit guarantee to 45,000 km!

b) $z = \frac{55{,}000-60{,}000}{8{,}300} = \frac{-5{,}000}{8{,}300} = -0.6024 \approx -0.60$ = 22.57%

$z = \frac{50{,}000-60{,}000}{8{,}300} = \frac{-10{,}000}{8{,}300} = -1.2048 \approx -1.20$ = 38.49%

Therefore 38.49% - 22.57% = 15.92% of tyres will need replacing between 50,000 and 55,000 km.

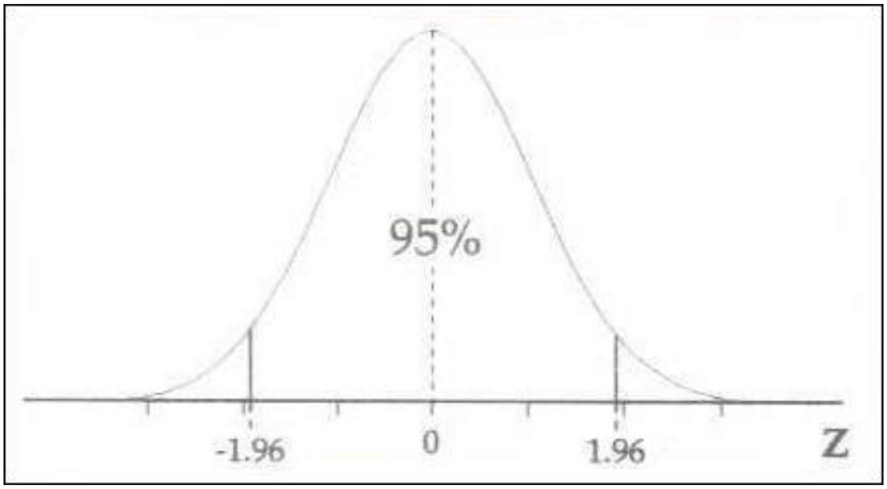
4 £x should be around £214: found by substituting for £x in the z score formula.

Confidence Intervals and Confidence Levels

Without even realising it, we have begun to use what are called **confidence intervals**; and many students think that they are hideous things to use. You're already experts with confidence levels and didn't even realise it.

When we are discussing the results of surveys and samples and interviews, we need to be sure that the statistics we generate are meaningful and accurate. To help us in this quest, statisticians have developed confidence levels and confidence intervals. We are going to use both of these ideas now.

The **confidence level** tells you how sure you can be; and it is expressed as a percentage, commonly 95% or 99%. Graphically, a 95% confidence interval looks like this:



A 95% confidence level tells us that 95% of an entire data set is found within ±1.96Sd of the mean. Why 1.96SD … look at the table and under 1.96 you will find the value 0.4750 which is half of 0.95 or 955. Hence, the whole range from -1.96SD to +1.96SD must contain 0.4750 + 0.4750 = 0.95 = 95% of the data set.

A 99% confidence level is found within ±2.58SD (well, the table shows that 2.57 = 0.4949 and 2.58 shows 0.4951 … usually you will see 2.58 rather than 2.58 or even 2.575): sketch a graph of that if you wish!

So, if we use a confidence level of 95%, we can say that we are 95% confident, or certain, that 95% of the data set, including the mean, is contained in that range. Similarly, if we use a confidence level of 99%, we are confident that 995 of the data set, including the mean, is contained in that range.

Confidence levels enable us to find confidence intervals. A **confidence interval** is a range of data in which the mean is expected to lie. The confidence interval is based on a confidence level and is found as follows:

$$\text{Confidence Interval}_{\alpha} = \bar{x} \pm z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

where

α is the confidence level usually given as 0.05 for a 95% confidence level, 0.01 for a 99% confidence level and so on

$\bar{x}$ is the mean of a data set

z is the z score

σ is the standard deviation of a data set

We know, we know, it looks horrific; but you’ve trusted us so far so why stop now?

Here’s an **example** to start you off:

Find the confidence interval for an airline that has sampled 225 of its flights and found that an average of 11.6 seats per flight are unoccupied with a standard deviation of 4.1 seats: use the 95% confidence level.

$$\text{Confidence Interval}_{0.05} = 11.6 \pm 4.1_{0.05/2} * \frac{4.1}{\sqrt{225}} = 11.6 \pm 1.96 * \frac{4.1}{\sqrt{225}} = 11.6 \pm 0.54$$

So the limits of the confidence interval are

11.6 - 0.54 and 11.6 + 0.54

which are

11.06 to 12.14

What this means is that the airline is 95% confident that the mean of this distribution is within these limits, so that the average number of unoccupied seats on one of its flights is in the range 11.06 to 12.04.

Further example: a supermarket has asked 49 of its customers about their age, in order to target their customers as effectively as possible. The average age of those questioned was 40.1 years with a standard deviation of 8.6 years.

a Using a 95% confidence interval, advise the store on the age range that it should target for, say, its future advertising campaigns.

b Using a 99% confidence interval, advise the store on the age range that it should target for, say, its future advertising campaigns.

Answers:

a)

$$\text{Confidence Interval}_{0.05} = 40.1 \pm 1.96_{0.05/2} * \frac{8.6}{\sqrt{49}} = 40.1 \pm 1.96 * \frac{8.6}{\sqrt{49}} = 40.1 \pm 2.41$$

the confidence interval, and therefore the target age range, is 37.69 to 42.51 years

b)

$$\text{Confidence Interval}_{0.01} = 40.1 \pm 2.58_{0.01/2} * \frac{8.6}{\sqrt{49}} = 40.1 \pm 2.58 * \frac{8.6}{\sqrt{49}} = 40.1 \pm 3.17$$

the confidence interval, and therefore the target age range, is 36.93 to 43.27 years

Note: as we should expect, the 99% confidence interval is larger than the 95% confidence interval.

Your Turn

1 A survey of 200 people revealed a mean amount of cash in their wallets or purses of £40 with a standard deviation of £12.

a Estimate the 95% confidence interval from these data.

b Estimate the 90% confidence interval from these data.

a)

$$\text{Confidence Interval}_{0.05} = 40 \pm 1.96_{0.05/2} * \frac{12}{\sqrt{200}} = 40 \pm 2.58 * \frac{12}{\sqrt{200}} = 40 \pm 1.66$$

therefore the 95% confidence interval is from £38.34 to £41.66 … so we can be 95% confident that the group of people from which this sample came from are likely to have between £38.34 and £41.66 in their wallets or purses at any one time.

b) $$\text{Confidence Interval}_{0.1} = 40 \pm 1.645_{0.10/2} * \frac{12}{\sqrt{200}} = 40 \pm 1.645 * \frac{12}{\sqrt{200}} = 40 \pm 1.40$$

therefore the 90% confidence interval is from £38.60 to £41.40 … so we can be 90% confident that the group of people from which this sample came from are likely to have between £38.60 to £41.40 in their wallets or purses at any one time.

2 The weights of 10 Sparrows caught by a biology student for analysis were, in grammes

27.4, 28.3, 30.9, 26.9, 27.9, 28.2, 28.6, 29.1, 29.9, 30

The student believes that the weights of Sparrows are normally distributed:

a construct a 95% confidence interval

b construct a 90% confidence interval

that will contain the true mean of the weights of Sparrows.

Conclusions

There we are: standard deviations and the normal distribution all sewn up nicely, thank you!

These two aspects of statistics have worried generations of students … until now! If you have worked methodically and systematically through this work, you really will be ready to face much of what statistical life can throw at you.

If you have any doubts about what's presented here, read, reread and read again; work through all of the exercises; question and challenge everything.