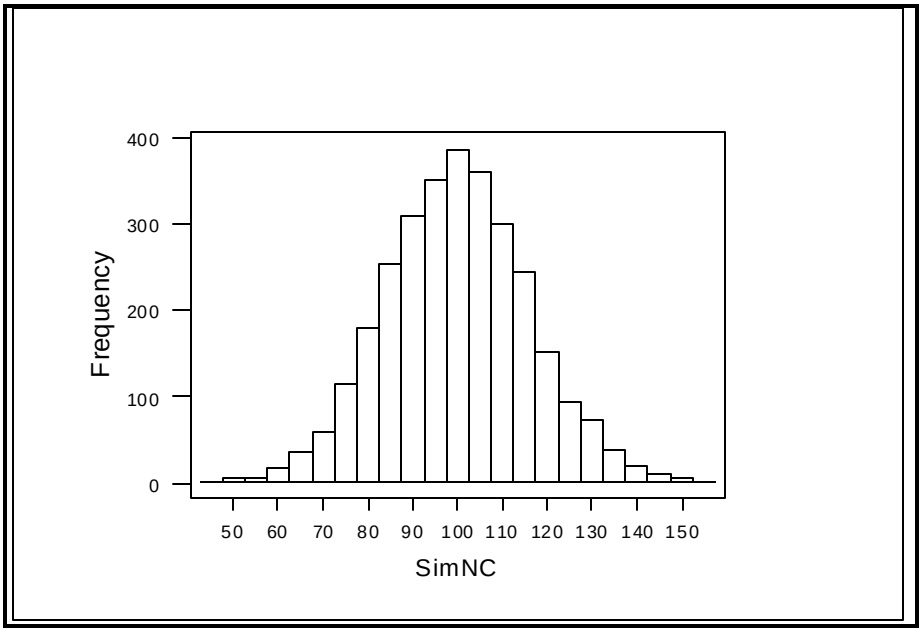
CHAPTER 6

THE NORMAL DISTRIBUTION

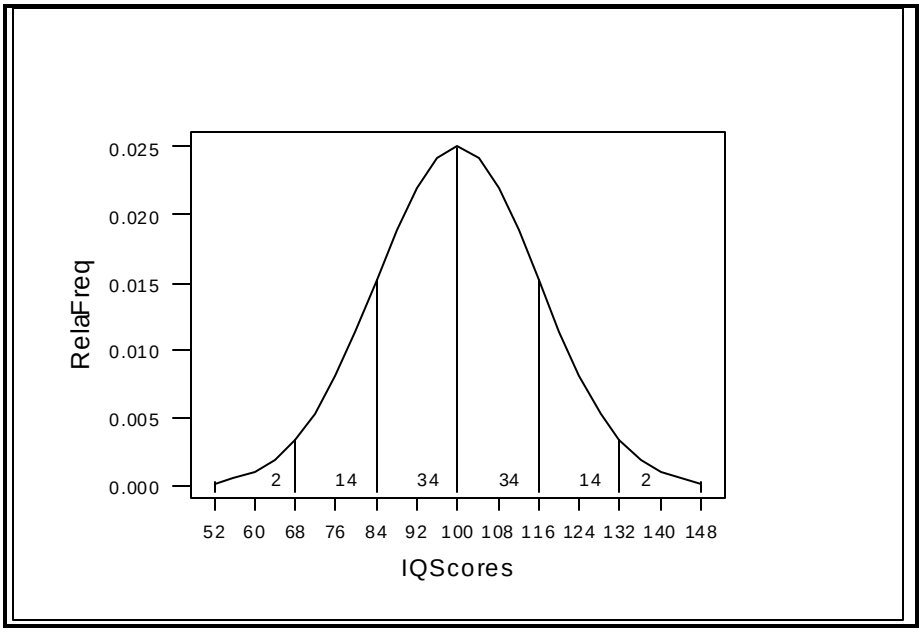
In the first Chapter, different shapes or forms of distributions (symmetrical or skewed, rectangular, U shaped, etc.) were discussed. One of the most popular distributions is that of a **UNIMODAL**, **SYMMETRICAL** distribution which is more typically called the **NORMAL DISTRIBUTION**. The normal distribution is important in statistics for many reasons but the most important one for our present purposes is the fact that many real life large sets of data resemble the pattern of a normal distribution. Since the normal distribution has certain fixed properties, it can be used to solve problems when the data set approximates the normal distribution. That is precisely what this section is about.

To get things started, look at the following histogram of a distribution that was simulated by Minitab randomly sampling 3000 values from a normal distribution where the mean is 100 and the standard deviation is 16. This is similar to IQ data.



Note that the pattern is one where the greatest number of frequencies are in the middle of the scale, say from 90 to 110, while there are fewer and fewer frequencies as you move to the right or to the left of the middle, say near 50 or 150. This is the typical unimodal symmetrical pattern that is called the normal distribution. Of course, since these data were randomly simulated, the pattern is not exactly smooth. In addition, if I did a second or third random simulation, the resulting distributions would be somewhat different from the one above.

Mathematically, normal distributions are continuous distributions where the curve at the top is smoothed. Using the same characteristics as above, with mean equal to 100 and standard deviation equal to 16, I have drawn a smoothed normal distribution below.



First note that the baseline X axis goes from about 52 to 148. In normal distributions, the data will extend on either side of the mean approximately 3 standard deviations. So, in this case, 3 standard deviations would be 3 times 16 or about 48 points. If you add 48 to 100 and subtract 48 from 100, you get the low and high points as shown on the graph. Technically, the distribution keeps going on and on but, practically speaking, after 3 standard deviations, you more or less run out of data. Secondly, notice that the Y or vertical axis (in this graph) is relative frequency. All that means is that the middle of the graph has more of the frequencies in the distribution whereas as you move to the left or right of center, the frequencies at each data location become fewer and fewer. Finally, notice that I have placed numbers <u>within</u> certain sections of the graph. For example, from 52 to 68, you see the number '2'. The number within each segment of the graph above indicates the percentage of the total distribution (out of 100%) that falls within that segment. Thus, 2 percent of the total normal distribution in the graph above falls between approximately 52 and 68. Between 68 and 84, there is approximately 14 percent of the total distribution. And, between 84 and 100, there is approximately 34 percent. Since a normal distribution is symmetrical around the middle, you will find the same percents to the right of 100: from 100 to 116 there is 34%, from 116 to 132 there is 14%, and from 132 to the approximate maximum value of 148 there is the final 2%. Even though different normal distributions will have different values along the baseline, <u>all normal distributions will have the same percentages of the total area within these standard deviation segments</u>.

Recall from the section on position measures, we could also indicate position in

terms of z scores; ie, how many standard deviation units a score is from the mean. In the distribution above, the standard deviation is 16. Since 100 is the mean value, it would have a z score of 0. A score of 84 would be 16 points or 1 standard deviation below the mean and would have a z value of -1. The z scores for the values of 68 and 52 would be -2 and -3 respectively. Above the mean, the scores the same IQ score point distances away from the mean would have the same z score values except that the sign would be positive, indicating 'above the mean''. Thus, the scores of 116, 132, and 148 would have z score values of +1, +2, and +3 respectively. Again, all normal distributions will have these same percentage areas within the indicated z score segments. To help you keep these in mind, look at the line graph below with the IQ scale and z score equivalents.

IQ Score	52	68	84	100	116	132	148
z Score	-3	-2	-1	0	+1	+2	+3
Area Between		2%	14%	34%	34%	14%	2%

It is a good idea to try to memorize the values of 34, 14, and 2 as the approximate percentage values for the 1 standard deviation segments going left or right on either side of the mean, for working problems that fit the normal distributionm model.

In addition to the z score values that could be placed along the baseline and the percentage values between certain raw and z score segments, we could also place along the baseline the approximate percentile rank values. Remember, the percentile rank is the percentage of the distribution that falls below some specific value. For example, as the most obvious case, the percentile rank for an IQ score of 100, that also has a z value of 0, is 50 since this is the point at which 50% of the distribution falls below. Using the areas between z score segments, we can list out several other percentile rank values. See the expanded line graph below.

IQ Score	52	68	84	100	116	132	148
z Score	-3	-2	-1	0	+1	+2	+3
Area Between		2%	14%	34%	34%	14%	2%
Percentile Rank	>1	2	16	50	84	98	>99

For the example above, we see that the percentile rank for an IQ score of 52 is less than 1. Recall that we don't usually list out percentile ranks as either 0 or 100. For a score value of 68, which includes about 2% from the bottom, the percentile rank is 2. The I Q value of 84, which is 1 standard deviation below the mean, has a percentile rank of approximately 16. On the right side of the mean, the percentile rank values for IQ's of 116, 132, and 148 are 84, 98, and 99+ respectively. Again, even though the raw score scale (IQ in this case) will differ from data set to data set, the percentile rank values for certain z score position along the baseline will be the same for all normal distributions. Again, it

would be helpful to try to remember these percentile rank values for the whole number z score positions along the baseline.

Using A Normal Distribution Table

If you are working a normal distribution problem and (for example) the percentile rank you want is for a score that falls exactly on one of the whole number z score positions (like -2 or +1), then the answer will easily be found by remembering the areas of 34, 14, and 2, or the percentile rank values listed above along the baseline. However, what if you had a normal distribution problem where you wanted to know the percentile rank for an IA value of 76 or 108? Since these two values do not fall exactly on the whole number z score positions, we cannot use the 34, 14, and/or 2 percent rules to help us find the answers. For a score of 76, we do know that the percentile rank value would be between 2 and 16, or for the value of 108, we know that the percentile would be between 50 and 84. But, exactly how far between 2 and 16, or 50 and 84? For answers to percentile rank problems that fall between various whole number z scores, we need the help of a table. Such a table is called the 'Area Under the Normal Curve' table, and such tables can be found in the backs of most books on statistics. See the table below.

AREA UNDER THE NORMAL DISITRUBITON

z values are at the side and top; areas are in body of table.
For areas, put decimal 2 places to RIGHT for percentage.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	000	004	008	012	016	019	023	027	031	035
0.1	039	043	047	051	055	059	063	067	071	075
0.2	079	083	087	091	094	098	102	106	110	114
0.3	117	121	125	129	133	136	140	144	148	151
0.4	155	159	162	166	170	173	177	180	184	187
0.5	191	195	198	201	205	208	212	215	219	222
0.6	225	229	232	235	238	242	245	248	251	254
0.7	258	261	264	267	270	273	276	279	282	285
0.8	288	291	293	296	299	302	305	307	310	313
0.9	315	318	321	323	326	328	331	334	336	338
1.0	341	343	346	348	350	353	355	357	359	362

1.1	364	366	368	370	372	374	377	379	381	383
1.2	384	386	388	390	392	394	396	398	399	401
1.3	403	404	406	408	409	411	413	414	416	417
1.4	419	420	422	423	425	426	427	429	430	431
1.5	433	434	435	437	438	439	440	441	442	444
1.6	445	446	447	448	449	450	451	452	453	454
1.7	455	456	457	458	459	459	460	461	462	463
1.8	464	464	465	466	467	467	468	469	469	470
1.9	471	471	472	473	473	474	475	475	476	476
2.0	477	477	478	478	479	479	480	480	481	481
2.1	482	482	483	483	483	484	484	485	485	485
2.2	486	486	486	487	487	487	488	488	488	489
2.3	489	489	489	490	490	490	490	491	491	491
2.4	491	492	492	492	492	492	493	493	493	493
2.5	493	494	494	494	494	494	494	494	495	495
2.6	495	495	495	495	495	496	496	496	496	496
2.7	496	496	496	496	496	497	497	497	497	497
2.8	497	497	497	497	497	497	497	497	498	498
2.9	498	498	498	498	498	498	498	498	498	498
3.0	498	498	498	498	498	498	498	498	499	499

Notice that the top and side columns are for z values with the side column being the tenths place of the z value, and the top being the hundredths place of the z value. For example, a z of .34 would be the 4th row at the side (.3) and the 5th column at the top (.04). Negative z values are handled in exactly the same manner. The body of the table represents the area under the normal curve from the mean or a z value of 0 out to that particular z score point. This table always gives areas from the mean out to some other z score value!

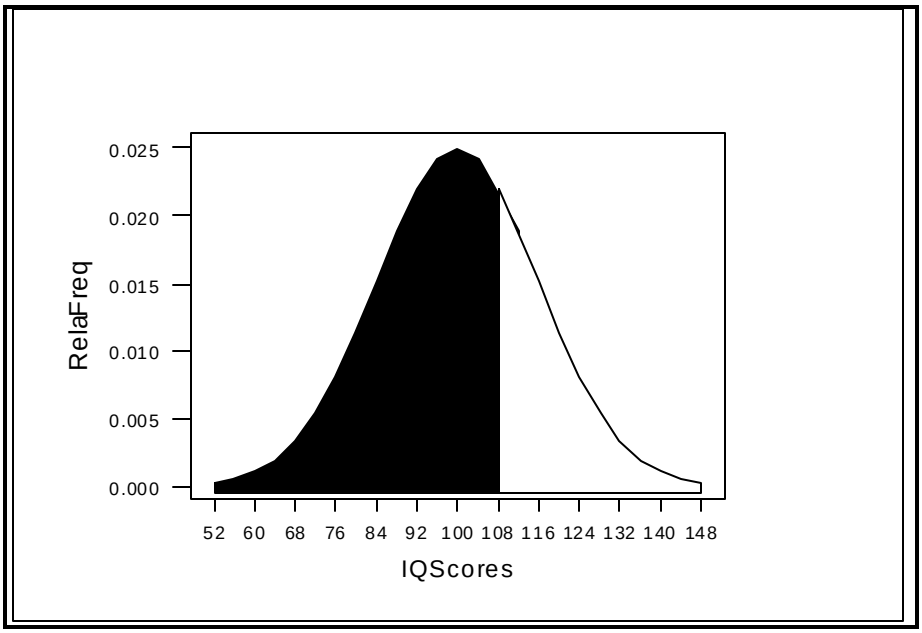
What if you wanted the percentile rank for a score of 116? We know of course, from the diagram, that the percentile rank is about 84. But, let's see how we would obtain that

value using the area table. The first thing you have to do is to convert the score value into a z score. For a score of 116, we would have the following: (116 - 100) / 16 = 1. Since we know that 50 percent of the distribution is below a score of 100, what we need to add to that is the area between 100 and 116. Since the area table always works from the mean (100 in this case) out to some other point (116 or a z of 1 in this case), we can find that additional area by looking up the area value for a z score of 1 from the area table. For a z of 1, we need the 1.0 row and the .00 (or first) column. The intersection of the 1.0 row and the .00 column gives us an area value of 341. The values in the table are given in proportions and we would need to move the decimal place two places to the right to make it into a percentage. In this case, 34.1 percent is the figure. So, with 50 percent below the mean and 34.1 percent from the mean out to a z of 1 or the score of 116, the percentile rank will be a little more than 84. We knew that anyway, right?

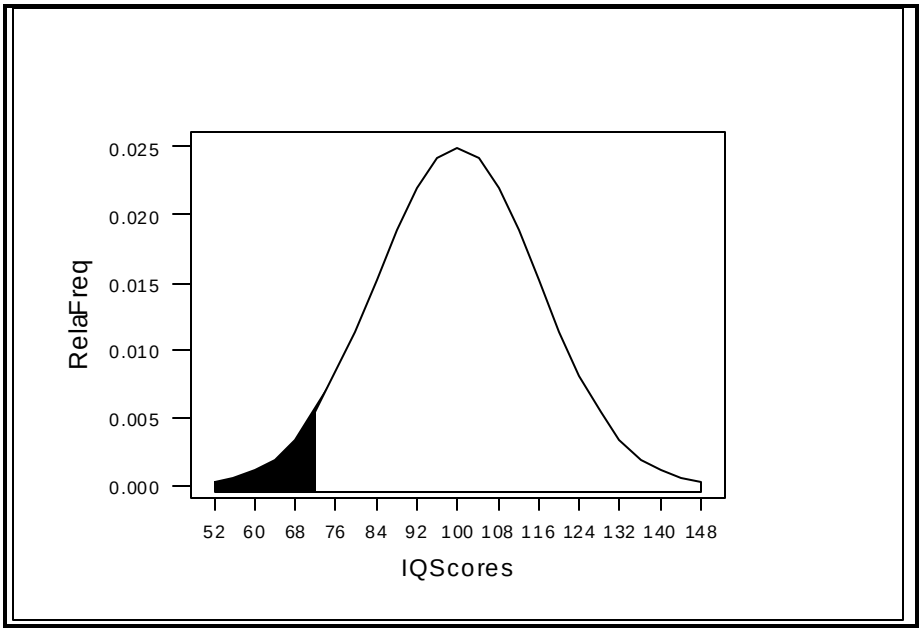
Finding the Percentile Rank for a Given Score

Perhaps the most common problem involving the use of a normal distribution is the case where one wants to find the percentile rank for a specific score value. As I said above, if that score value falls on some whole number z score position (-2 or +1 for example), then the problem is easily solved by remembering the 34, 14, and 2 values. But, if the score falls between these whole number points, the problem is not quite so easy. What if we work a problem that is not so easily obtained from the normal curve diagram. Perhaps we would like to find the percentile rank for a score of 108. Look at the graph at the top of the next page; the shaded part is the area we are interested in finding.

First we need to convert 108 to a z score. In this case, the z score would be: (108 - 100) / 16 = .5. A score of 108 is one half of a standard deviation above the mean. We now use the area table and look for the .5 row and (again) the .00 column. The intersection gives us an area of 191 which, if you move the decimal 2 places to the right, is about 19 percent. The 19 percent is the area between the mean, 100 and 108. Therefore, a score of 108 has a percentile rank of about 50 + 19 = 69.



What about a value that is below the mean? What is the percentile rank for a score of 72? See the figure below. Again, we need to find the z score. For 72 we have: (72 - 100) / 16 = - 1.75. We then look up 1.75 in the area table at the intersection of 1.7 and .05. This



gives us an area value from the mean of 459 or, if you move the decimal 2 places to the right, about 46 percent. However, this is the area from the mean or middle down to the z value of -1.75! What we actually need is the area below a z of -1.75. Since there is 50

percent from the middle down (totally) and 46 percent from the middle down to a score of 72, the area below 72 will be the difference between 50 and 46 = 4. The score of 72 in this normal distribution has an approximate percentile rank of 4. For scores above the mean, find the area from the table and add 50%. For scores below the mean, find the area from the table and subtract from 50%.

So far, I have not emphasized specific Minitab procedures for accomplishing certain statistical tasks, though I have mentioned (primarily at the ends of Chapters) some commands in Minitab that could be useful. However, here I break a bit from that pattern. It happens that Minitab has two commands that are very helpful when solving problems related to the normal distribution. At this point, I introduce the first of those two commands. It is called the **CDF** command, and cdf stands for 'cumulative distribution function'. In short, what cdf does is to accumulate the area up to some point that you specify in some particular normal distribution. For example, we looked at problems involving IQ scores of 108 and 72. Using the area table is rather cumbersome and, you must remember to either add 50% to the tabled value or subtract the tabled value from 50%. However, with the cdf command, Minitab takes care of all that work for you. Look at the commands below for finding the percentile ranks for scores of 108 and 72. Note: **MTB >** is the main command prompt where you indicate the command and the score value, while the **SUBC>** is called a subcommand where you indicate the distributional type, in this case a normal distribution with a mean of 100 and a standard deviation of 16.

MTB > cdf 108;
SUBC> norm 100 16.

```
      x          P( X <= x)
  108.0000        0.6915     <--------- Minitab prints out about 69 percent
```

MTB > cdf 72;
SUBC> norm 100 16.

```
      x          P( X <= x)
   72.0000        0.0401     <--------  Minitab prints out about 4 percent
```

Finding the Score that has a Given Percentile Rank

Another problem you may want to work is to find a score that has a particular percentile rank. This is exactly the opposite process from what is discussed above. Again, from the normal curve diagram, some problems like this would be easy to solve. For example, what is the score value for a percentile rank of 16? For our data, the answer is 84. What is the score for a percentile rank of 84? Again, the answer is easy and is 116. But, what if the score we are looking for falls between the nice percentile rank values listed in the diagram? For example, what is the score for a percentile rank of 20 or 75? We could guesstimate from the diagram but, in most instances, that would not be quite

accurate enough. Again, we can use the area table to work these types of problems.

What is the score that has a percentile rank of 20? The first thing we need to solve this problem is to find the z score that is associated with a percentile rank of 20. From looking at the diagram at the bottom of page 55, we know that the z score for a value that has a percentile rank of 20 would be somewhere between -1 and 0. But exactly where between -1 and 0 would it be? Remember, before, we found the z score, went to the table, and then found our percentile rank. In this case, we have the percentile rank. If we knew the area from the score to the middle, we could back track and then find the z score that had that much area from the mean out to that point. So, if a score has a percentile rank of 20, how much area is there from that point to the mean? The answer is 30 percent. Now we enter the area table in the body and look for the area value as close to the number 300 (which is 30% if you move the decimal 2 places to the right) as possible. In the table we find 299 (29.9%) and the number 302 (30.2%). Of the two, 299 is closer to 300. What is the z score for this value of 299? Going to the side and to the top, we have a z of .84. But, keep in mind that a percentile rank of 20 is below the mean and therefore the z must be negative or -.84. With the z value of -.84, we can substitute into the z score formula to find the X or score value. We would have:

$$-.84 = (X - 100) / 16$$
$$-13.44 = X - 100$$
$$X = 86.56$$

Thus, the score that has a percentile rank of 20 is 86.56 in a normal distribution where the mean is 100 and the standard deviation is 16.

What about the problem of finding the score value that has a percentile rank of 75? Again, from the diagram on page 55, we would see that a value with a percentile rank of 75 would have a z value between 0 and 1. Since a percentile rank of 75 would have 25% between that point and the mean, we would need to look up 25% (or as close to it as possible) in the body of the area table and see what z score is associted with that percentage. From the area table (top of page 57) we see the values of 248 and 251. Of the two, the value of 251 or 25.1% is the closest to 25%. If you then go to the left and to the top of the table, you will find that 251 is associated with a z score of .67 and it is positive since the 75th percentile rank is above the mean. To find the IQ score value:

$$.67 = (X - 100) / 16$$
$$10.72 = X - 100$$
$$X = 110.72$$

The IQ for a z of .67 is approximately 111.

The process of finding a score given a percentile rank is: find the area from the percentile rank to the mean, find that area in the body of the area table, find the z score, and then, as a last step, find the X score by using the z score formula.

Again, Minitab has a command that can easily find these values for us. If you are given the score and want the percentile rank, we saw above that the **CDF** command was helpful. But, in the case of finding the score given the percentile rank, we need to go through the process in the opposite order. Actually, what we need is a command like cdf that will do the opposite calculation. That command in Minitab is called **INVCDF** or the inverse of what cdf will do. In this case, at the **MTB>** prompt, we will indicate which percentile rank we are using, and using the **SUBC>** line to again indicate what type of a normal distribution we are working with. See below. It is important to input the percentile rank as a decimal value, not a whole number.

MTB > invcdf .25;
SUBC> norm 100 16.

```
P( X <= x)        x
   0.2500      89.2082     <--------- Minitab prints out a value of 89.2
```

MTB > invcdf .75;
SUBC> norm 100 16.

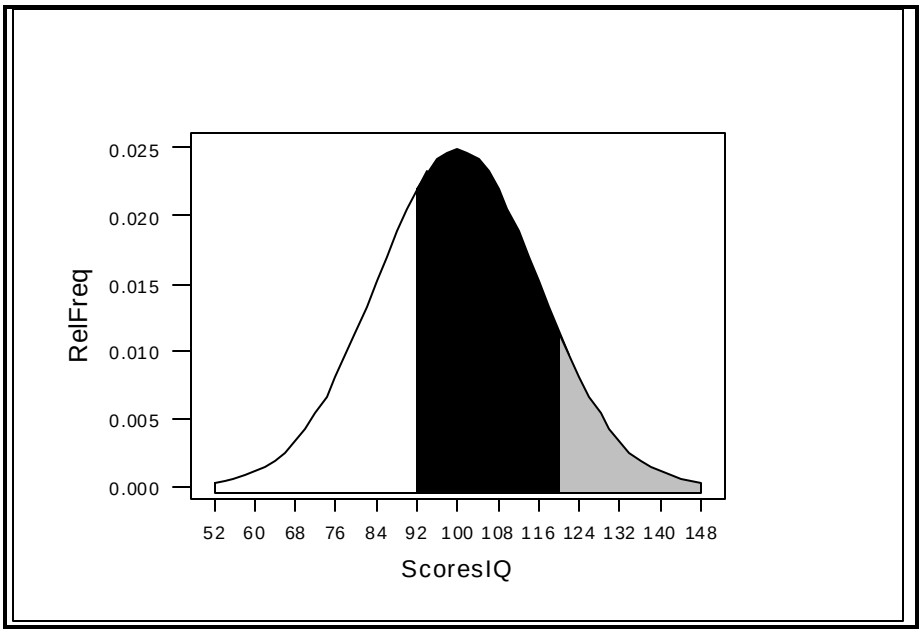
```
P( X <= x)        x
   0.7500     110.7918     <-------- Minitab prints out a value of 110.8
```

As you can see, the use of the cdf and invcdf commands makes the solution of normal distribution problems quite easy, if you are using Minitab.

Finding the Area between Two Points

While the most common normal distribution problems are those of either finding the percentile rank for a given score or finding the score for a given percentile rank, another problem that is sometimes encountered is of the nature: what is the area between two points? Again, if the problem involves data points that fall on whole number z score positions, then this problem is easily solved. For example, what is the area between scores of 84 and 132? Well, we can look at the normal distribution on page 54 and see that from 84 to 132 would encompass the areas of 34% (from 84 to 100), 34% (from 100 to 116), and 14% (from 116 to 132). Thus, the area between 84 and 132 would be 34% + 34% + 14% = 82%.

However, what if the area you want falls between IQ values like 92 and 120? Sure, we could estimate it from the normal distribution diagram on page 54 but, more than likely, we would need to be more accurate than that. The area we want is shown in the following diagram as dark shading.



To use the area table, we would first need to convert the scores of 92 and 120 into z values. For 92, it would be (92 - 100) = -8 divided by 16 or -.5 as a z value. For the score of 120, it would be (120 - 100) = 20 divided by 16 or a z value of 1.25. Looking up the z values of -.5 and 1.25 in the area table, we would have areas from the mean to those z values of 19.1% and 39.4% respectively. Thus, adding them together, the approximate area between the IQ scores of 92 and 120 would be 58.5 percent. If we let Minitab help us with this problem, we would need to use the cdf command twice, and then subtact the smaller value from the larger value. See the following.

MTB > cdf 92;
SUBC> norm 100 16.

```
      x     P( X <= x)
   92.0000       0.3085     <-------- Minitab gives this below 92
```

MTB > cdf 120;
SUBC> norm 100 16.

```
      x     P( X <= x)
  120.0000       0.8944   <--------- Minitab gives this below 120
```

Therefore, by subtracting 30.85% from 89.44%, we obtain 58.59%.

Useful Minitab Commands

CDF INVCDF

Practice Problems

Assume that you are working with a normal distribution where the mean is 50 and the standard deviation is 8. Answer the following.

1. What are the percentile ranks for scores of 30, 45 and 62?

2. What are the scores that have percentile ranks of 20, 35, 55 and 70?

3. How much area is there between the scores of 40 and 52?