As the basis for today's lab, we are interested in the following hypothetical question:   
Let's say that 1994 Chevrolet Cavaliers are recalled because of a slight defect in the suspension. Of those recalled, only 30% actually requre repairs. A small dealer with an overworked service department hopes that no more than 25 of his 100 recalled cars will require repairs. What is the chance of his being so lucky?

**Starting S-Plus**

To start S-Plus, click *Start*, then *Programs*, then *Statistics & Mathematics*, then *S-PLUS 2000*. We will not need to incorporate any data files for today's exercise, but we will need to create a new data set from scratch. To do this, click *File*, then *New*, then choose "Data Set".

**Binomial probability distribution**

Imagine that Y is the number of recalled cars (out of 100) that need repairs. We are interested in the probability distribution for Y; that is, we want to know what probabilities are associated with the different potential values for Y (0, 1, 2, ..., 100). The probability associated with each value can be calculated by hand or calculator according to the binomial formula, although this is a time-consuming process. Instead, we use S-Plus to calculate these probabilities for us, and then display them graphically.

First, we need to fill our data set's first column with the numbers 0 through 100, which are all the possibilities for Y. Then, we will use R to determine the probabilities (according to the binomial distribution) for each of these values, and put those in the second column.

To get the probabilities corresponding to each value, use *Data*, then *Distribution functions*. Choose your first column (where the possible Y values are) as the source column, and "Density" as the "Result type". Of course, "binomial" should be the distribution, with probability 0.30 and sample size 100. Now, the second column contains the probability associated with each value in the first column. Rename this column something like "Bin.prob", to denote that these are the exact probabilities obtained using the binomial distribution. Also, you may want to change the precision to allow for more than 2 decimal places to be displayed

**Graph the binomial probability distribution**

Go to *Graph*, then *Bar with base at Y min*. Enter the first column as the X column, and enter the second column as the Y column. This shows a histogram of the probabilities associated with 0,1,2,...,100 (out of 100) cars needing repairs. Notice how normal the graph of the distribution appears. We can superimpose the normal curve that approxmiates this distribution on top of the bar graph.

To do this, we note that the normal curve should have the same mean and variance as the binomial distribution. The binomial random variable Y has a mean equal to n\*p=100\*0.3=30 and a variance equal to n\*p\*(1-p)=100\*0.3\*0.7=21. So, we need to draw in a normal density function with mean 30 and variance 21 (standard deviation about 4.582576). We can evaluate the normal density at points 0,1,...,100, using the same procedure we used for the binomial. Go to *Data*, then *Distribution functions* as we did before; choosing normal as the distribution, and entering mean 30 and standard deviation 4.582576. Insert these values into column 3. To add the density curve to the bar graph, choose *Insert*, then *Plot*, then *Line plot*. Choose the first column as the X column, and the third column as the Y column. This should overlay a normal density line onto your barplot of binomial probabilities.

**Finding the probability that no more than 25 cars need repair**

To find the exact probability (using the binomial distribution) that 25 or fewer cars need repair, we just need to add the values in the first 26 rows of the density column. (This encompasses the probability that 0 cars need repair, 1 car needs repair, 2 cars need repair, ... , up to 25 cars need repair.) To do this, issue the command

sum(SDF1[1:26,2])

in the "Commands" window.

Now, we want to find an estimate for this answer using the normal approximation to the binomial. We know that the question the dealer is interested in can be stated in two ways "no more than 25 cars will need repair" and "less than 26 cars will need repair". This means that, given the continuity correction, we are interested in the probability under the normal curve (centered at 30 and with variance 100\*0.3\*0.7) to the left of 25.5. To find this probability, enter

pnorm(25.5, mean=30, sd=sqrt(100\*0.3\*0.7))

in the "Commands" window. How does this compare with the more exact answer obtained above? What if we had not used the continuity correction, and we had just found the probability to the left of 25, or to the left of 26 (depending on the original statement of the question)? To find what the approxmiations to the binomial might be if we had not used the continuity correction, issue the commands:

pnorm(25.0, mean=30, sd=sqrt(100\*0.3\*0.7)) #find prob. to left of 25

and

pnorm(26.0, mean=30, sd=sqrt(100\*0.3\*0.7)) #find prob. to left of 26

What does this exercise say about the efficacy of using the normal distribution to approximate the binomial (given that the sample size is large enough)? Also, what does this say about the use of the continuity correction for similar problems?