CALIFORNIA STATE UNIVERSITY, SACRAMENTO
College of Business Administration

NOTES FOR DATA ANALYSIS

[Ninth Edition]

Manfred W. Hopfe, Ph.D.
Stanley A. Taylor, Ph.D.

NOTES FOR DATA ANALYSIS

[Ninth Edition]

As stated in previous editions, the topics presented in this publication, which we have produced to assist our students, have been heavily influenced by the *Making Statistics More Effective in Schools of Business* Conferences held throughout the United States. The first conference was held at the University of Chicago in 1986. The School of Business Administration at California State University, Sacramento, hosted the tenth annual conference June 15-17, 1995. Most recent conferences were held at Babson College, (June 1999) and Syracuse University (June 2000).

As with any publication in its developmental stages, there will be errors. If you find any errors, we ask for your feedback since this is a dynamic publication we continually revise. Throughout the semester you will be provided additional handouts to supplement the material in this book.

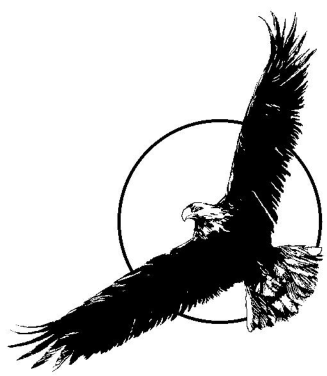
StatGraphics Plus for Windows (ver 4.0), the statistical software used in MIS 101 and MIS 206, will work only on a Pentium chip computer. For the chapter discussions, the term StatGraphics is generic for StatGraphics® Plus for Windows (ver 4.0)

Manfred W. Hopfe, Ph.D.
Stanley A. Taylor, Ph.D.

Carmichael, California
August 2000

TABLE OF CONTENTS

[intentionally left blank]



INTRODUCTION

The objective of this section is to ensure that you have the necessary foundation in statistics so that you can maximize your learning in data analysis. Hopefully, much of this material will be review. Instead of repeating Statistics 1, *the pre-requisite for this course*, we discuss some major topics with the intention that you will focus on concepts and not be overly concerned with details. In other words, as we "review" try to think of the overall picture!

Statistic vs. Parameter

In order for managers to make good decisions, they frequently need a fair amount of data that they obtain via a sample(s). Since the data is hard to interpret, in its original form, it is necessary to summarize the data. This is where statistics come into play -- a statistic is nothing more than a quantitative value calculated from a sample.

Read the last sentence in the preceding paragraph again. **A statistic is *nothing more than* a quantitative value *calculated from a sample*.** Hence, for a given sample there are many different statistics that can be calculated from a sample. Since we are interested in using statistics to make decisions there usually are only a few statistics we are interested in using. These useful statistics estimate characteristics of the population, which when quantified are called *parameters*.[1]

The key point here is that managers must make decisions based upon their perceived values of parameters. Usually the values of the parameters are unknown. Thus, managers must rely on data from the population (sample), which is summarized (statistics), in order to estimate the parameters.

Mean and Variance

Two very important parameters which managers focus on frequently are the ***mean*** and ***variance***[2]. The mean, which is frequently referred to as "the average," provides a measure of the central

[1] Greek letters usually denotes parameters.
[2] The square root of the variance is called a standard deviation.

tendency while the variance describes the amount of dispersion within the population. For example, consider a portfolio of stocks. When discussing the rate of return from such a portfolio, and knowing that the rate of return will vary from time period to time period[3] one may wish to know the average rate of return (mean) and how much variation there is in the returns [explain why they might be interested in the mean and variance].

Sampling Distribution

In order to understand statistics and not just "plug" numbers into formulas, one needs to understand the concept of a sampling distribution. In particular, one needs to know that ***every statistic has a sampling distribution, which shows every possible value the statistic can take on and the corresponding probability of occurrence***.

What does this mean in simple terms? Consider a situation where you wish to calculate the mean age of all students at CSUS. If you take a random sample of size 25, you will get one value for the sample mean (average)[4] which may or may not be the same as the sample mean from the first sample. Suppose you get another random sample of size 25, will you get the same sample mean? What if you take many samples, each of size 25, and you graph the distribution of sample means. What would such a graph show? The answer is that it will show the distribution of sample means, from which probabilistic statements about the **population** mean can be made.

Normal Distribution

For the situation described above, the distribution of the sample mean will follow a normal distribution. What is a **normal distribution**? The normal distribution has the following attributes:

- It depends on two parameters - the **mean** and **variance**
- It is bell-shaped
- It is symmetrical about the mean

[3] What is the random variable?
[4] The sum of all 25 values divided by 25.

[You are encouraged to use StatGraphics Plus and plot different combinations of means and variances for normal distributions.]

From a manager's perspective it is very important to know that with normal distributions approximately:

- 95% of all observations fall within 2 standard deviations of the mean
- 99% of all observations fall within 3 standard deviations of the mean.

Confidence Intervals

Suppose you wish to make an inference about the average income for a group of people. From a sample, one can come up with a **point estimate**, such as $24,000. But what does this mean? In order to provide additional information, one needs to provide a confidence interval. What is the difference between the following 95% confidence intervals for the population mean?

[23000 , 24500] and [12000 , 36000]

Hypothesis Testing

When thinking about hypothesis testing, you are probably used to going through the formal steps in a very mechanical process without thinking very much about what you are doing. Yet you go through the same steps every day.

Consider the following scenario:

> I invite you to play a game where I pull a coin out and toss it. If it comes up heads you pay me $1. Would you be willing to play? To decide whether to play or not, many people would like to know if the coin is fair. To determine if you think the coin is fair (a hypothesis) or not (alternative hypothesis) you might take the coin and toss it a number of times, recording the outcomes (data collection). Suppose you observe the following sequence of outcomes, here H represents a head and T represents a tail -
>
> **H H H H H H H H T H H H H H H T H H H H H H**
>
> What would be your conclusion? Why?

Most people look at the observations and notice the large number of heads (statistic) and conclude that they think the coin is not fair because the probability of getting 20 heads out of 22 tosses is very small, if the coin is fair (sampling distribution). It did happen; hence one rejects the idea of a fair coin and consequently does not wish to participate in the game.

Notice the steps in the above scenario

1. State hypothesis
2. Collect data
3. Calculate statistic
4. Determine likelihood of outcome, if null hypothesis is true
5. If the likelihood is small, then reject the null hypothesis
   If the likelihood is not small, then do not reject the null hypothesis

The one question that needs to be answered is "what is small?" To quantify what *small* is one needs to understand the concept of a Type I error. (We will discuss this more in class.)

P-Values

In order to simplify the decision-making process for hypothesis testing, ***p-values*** are frequently reported when the analysis is performed on the computer. In particular a p-value[5] refers to where in the sampling distribution the test statistic resides. Hence the decision rules managers can use are:

- If the p-value is $\leq$ alpha, then reject Ho
- If the p-value is $>$ alpha, then do not reject Ho.

The p-value may be defined as *the probability of obtaining a test statistic equal to or more extreme than the result obtained from the sample data, given the null hypothesis H_0 is really* true.

[5] Referred to frequently in statistical software as a Prob. Level or Sig. Value.

QUALITY -- COMMON VS SPECIFIC VARIATION

During the past decade, the business community of the United States has been placing a great deal of emphasis on quality improvement. One of the key players in this quality movement was the late W. Edwards Deming, a statistician, whose philosophy has been credited with helping the Japanese turn their economy around.

One of Deming's major contributions was to direct attention away from inspection of the final product or service towards monitoring the process that produces the final product or service with emphasis of statistical quality control techniques. In particular, Deming stressed that in order to improve a process one needs to reduce the variation in the process.

Common Causes and Specific Causes

In order to reduce the variation of a process, one needs to recognize that the total variation is comprised of **common causes** and **specific causes**. At any time there are numerous factors which individually and in interaction with each other cause detectable variability in a process and its output. Those factors that are not readily identifiable and occur randomly are referred to as the **common causes**, while those that have large impact and can be associated with special circumstances or factors are referred to as **specific causes**.

To illustrate *common causes versus specific causes*, consider a manufacturing situation where a hole needs to be drilled into a piece of steel. We are concerned with the size of the hole, in particular the diameter, since the performance of the final product is a function of the precision of the hole. As we measure consecutively drilled holes, with very fine instruments, we will notice that there is variation from one hole to the next. Some of the possible common sources can be associated with the density of the steel, air temperature, and machine operator. As long as these sources do not produce significant swings in the variation they can be considered common sources. On the other hand, the changing of a drill bit could be a specific source provided it produces a significant change in the variation, especially if a wrong sized bit is used!

In the above example what the authors choose to list as examples of common and specific causes is not critical, since what is a common source in one situation may be a specific source in another and vice versa. What is important is that one gets a feeling of a specific source, something that can produce a significant change and that there can be numerous common sources that individually have insignificant impact on the process variation.

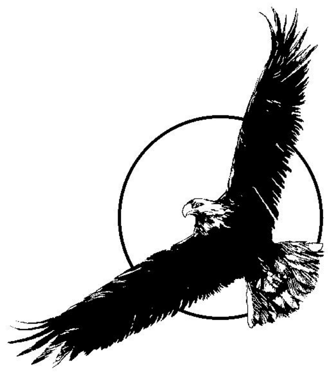
Stable and Unstable Processes

When a process has variation made up of only common causes then the process is said to be a stable process, which means that the process is in statistical control and remains relatively the same over time. This implies that the process is predictable, but does not necessarily suggest that the process is producing outputs that are acceptable as the amount of common variation may exceed the amount of acceptable variation. If a process has variation that is comprised of both common causes and specific causes then it is said to be an unstable process, which means that the process is not in statistical control. An unstable process does not necessarily mean that the process is producing unacceptable products since the total variation (common variation + specific variation) may still be less than the acceptable level of variation.

In practice one wants to produce a quality product. Since quality and total variation have an inverse relation (i.e. less {more} variation means greater {less} quality), one can see that a goal towards achieving a quality product is to identify the specific causes and eliminate the specific sources.[1] What is left then is the common sources or in other words a stable process. Tampering with a stable process will usually result in an increase in the variation that will decrease the quality. Improving the quality of a stable process (i.e. decreasing common variation) is usually only accomplished by a structural change, which will identify some of the common causes, and eliminate them from the process.

For a complete discussion of identification tools, such as time series plots to determine whether a process is stable (is the mean constant?, is the variance constant?, and is the series random -- i.e. no pattern?) see the Stat Graphics Tutorial. The runs test is an identification tool that is used to identify nonrandom data.

[intentionally left blank]



CONTROL CHARTS

In this section we first provide a general discussion of control charts, then follow up with a description of specific control charts used in practice. Although there are many different types of control charts, our objective is to provide the reader with a solid background with regards to the fundamentals of a few control charts that can be easily extended to other control charts.

Control charts are statistical tools used to distinguish common and specific sources of variation. The format of the control chart, as shown in Figure 1 below, is a group made up of three lines where the center line = process average, upper control limit = process average + 3 standard deviations and lower control limit = process average - 3 standard deviations.

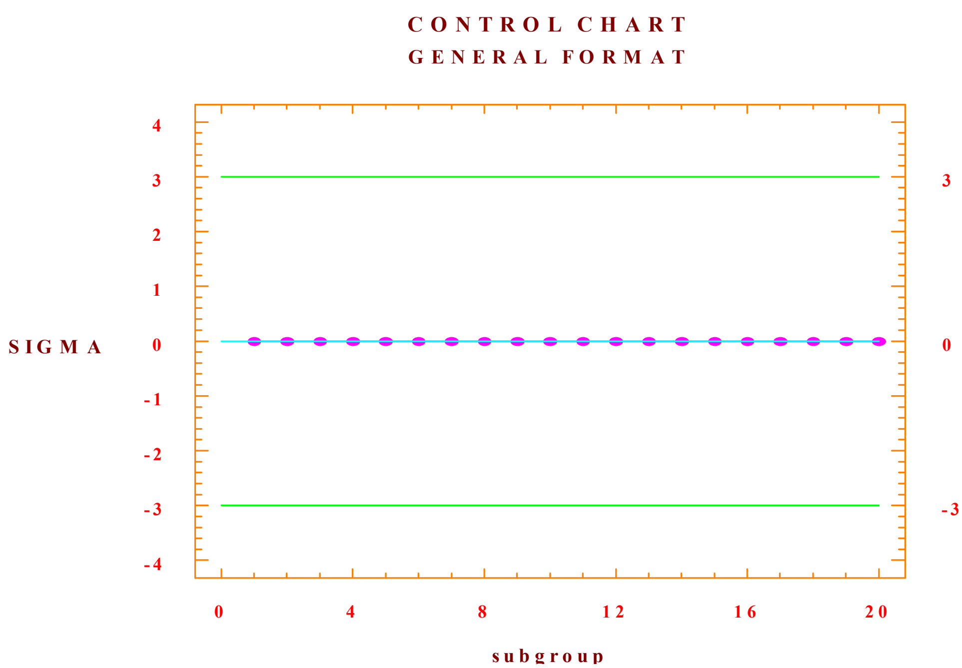


Figure 1. Control Chart (General Format)

The control charts are completed by graphing the descriptive statistic of concern, which is calculated for each subgroup. There are usually 20 to 30 subgroups used per each graph. The concept of how to form subgroups is very important and will be discussed later. For now it is important to state that *the horizontal axis is time*, so that we can view the graphed points from earliest to latest as we read the graph.

Recall that our goal in constructing control charts is to detect sources of specific variation, which, if they exist can be eliminated, thereby decreasing the variation of the process and hence increasing quality. Furthermore, recall that the existence of specific variation is the difference between an unstable process and a stable process. Therefore the detection of specific variation will be equivalent to being able to differentiate between unstable and stable processes.

Since stable processes are made up of only common causes of variation, the control charts of stable processes will exhibit no pattern in the time series plot of the observations. Departures, i.e. a pattern in the time series plot, indicate an unstable process that means that specific sources of variation exist, which need to be exposed of and eliminated in order to reduce variation and hence improve quality. As we consider each control chart, we will focus on whether there is any information in the series of observations that would be evident by the existence of a pattern in the time series plot of the observations.

Rather than showing what the control chart of a stable process looks like, it is helpful to first consider charts of unstable processes that occur frequently on practice.

We present seven graphs on the following pages for consideration. The following will summarize the seven examples displayed:

> Note that in Figure 2. Chart A the process appears to be fairly stable with the exception of an outliner (see subgroup 7). If this were the case then one would want to determine what caused that specific observation to be outside the control limits and based upon that source take appropriate action.
>
> In Figure 3, Chart B, note that there are two observations, close to each other that are outside the control limits. When this occurs there is much stronger evidence that the process is out of control than in Figure 2. Chart A. Again one would need to investigate the reason for these outliers and take appropriate action.
>
> Illustrated in Figure 3. Charts C and D is the concept of a trend. Notice in Chart C there is a subset of observations that constitute a downward trend, while in Figure 3. Chart D there is a subset that constitutes an upward trend.

In Figure 3, Chart E, a cyclical pattern is depicted. These types of patterns occur frequently when the process is subject to a seasonal influence. If this is the case, then one needs to account for the seasonality and make the necessary adjustments.

Presented in Figure 3. Chart F, is a situation where there is a change in the level of a process. Notice how the level slides upward, thereby indicating a change in the level. In this situation, one would need to ascertain why the slide took place and then take appropriate action.

The final case illustrated, Figure 3. Chart G, is one where there is a change in the variance (dispersion). Notice that the first part of the sequence has a much smaller variance than the latter part. Clearly an event occurred which altered the variance and needs to be dealt with appropriately.

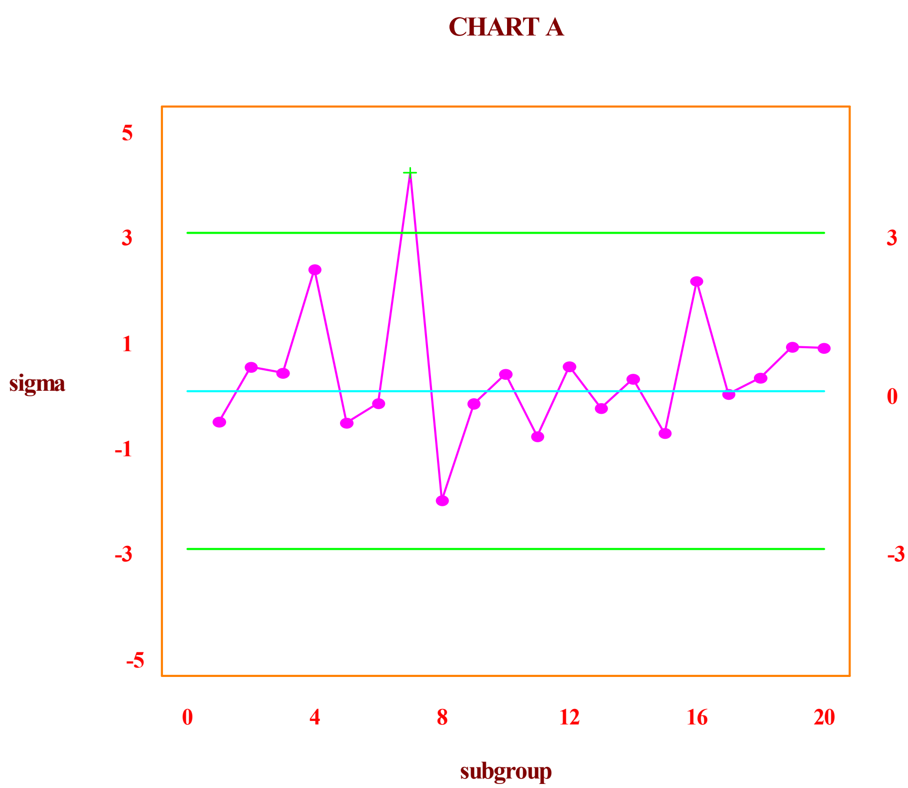


Figure 2. Chart A

Charts B through F appear in Figure 3 on the next page.

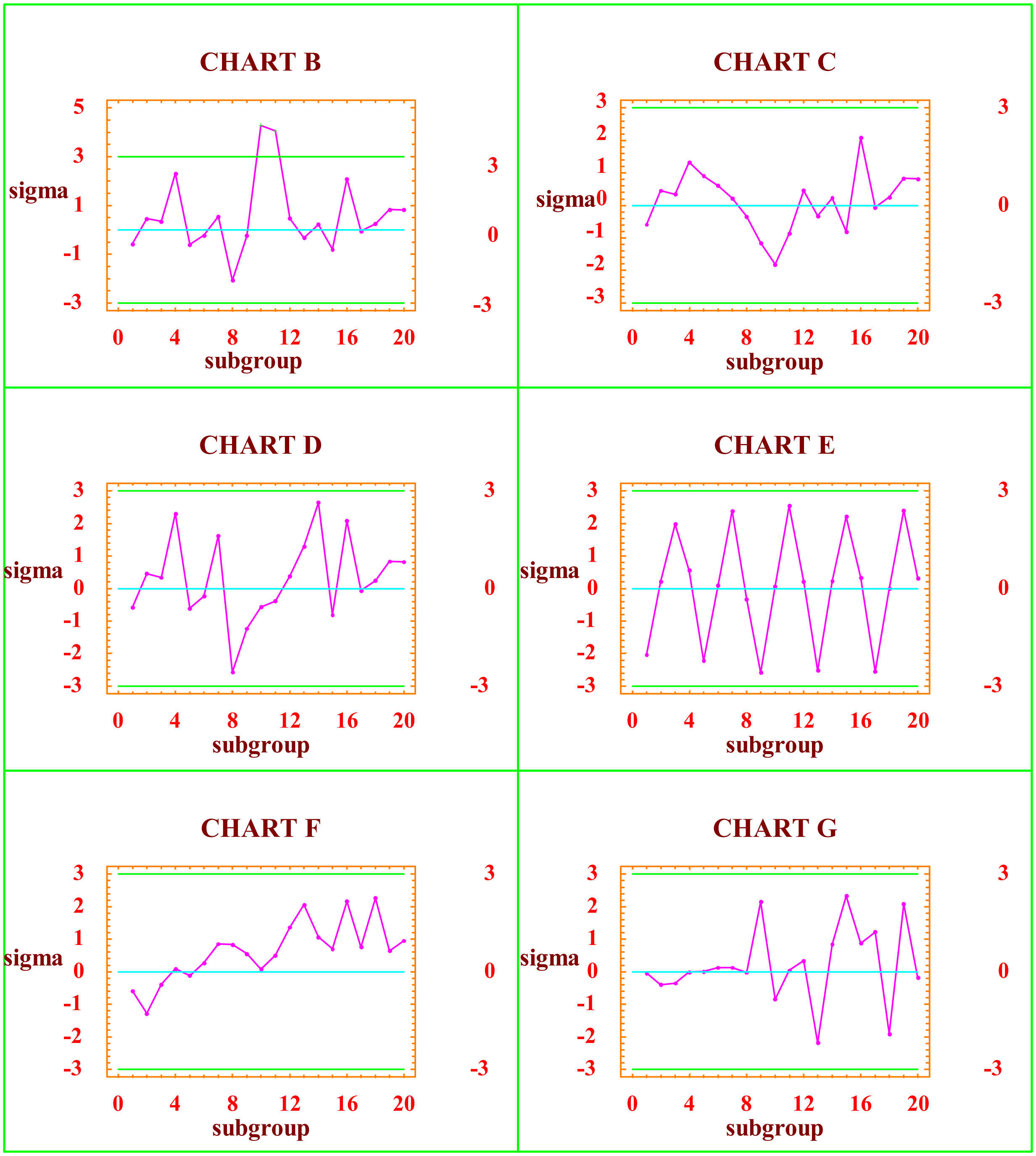


Figure 3. Charts B through G

Types Of Control Charts

As we mentioned previously, there are a large number of different control charts that are used in practice but for our purposes we will consider just a few. For a given application the type of control chart that should be employed depends upon the type of data being collected. There are three general classes of data:

- continuous data
- classification data
- count data

Continuous data is measurable data such as thickness, height, cost, sales units, revenues, etc. The latter two classes (classification and count) are examples of attribute data. For classification, data is bi-polar, for example, success/failure, good/bad, yes/no or conforming/non-conforming. Count data is rather straightforward -- number of customers served during the lunch hour, number of blemishes per sheet (8' by 4') of particleboard, number of failed parts per case, and so forth.

For many applications the data to be collected can be either ***continuous*** or ***attribute***. For example, when considering the size of holes discussed earlier one can record the diameter in millimeters (continuous) or as simply acceptable or unacceptable (attribute). Whenever possible, one should elect to record continuous data since fewer measurements are required per subgroup for continuous charts, 1 to 10, than for attribute charts which typically require 30 to 1000. The fewer the number of observations needed, the quicker the possible response time when problems surface.

We now consider examples for each of the control charts stated previously. First we will consider continuous data, in particular the X-bar and R charts. Then we will consider the P chart (classification data). Lastly we present the C chart (count data).

Continuous Data

X-bar and R Charts

To demonstrate the X-bar and R charts, we utilize data generated over a twenty-week period of time from the SR Mattress Co. The daily output of usable mattress frames for both shifts are shown below:

SR Mattress Company					
Week	**Mon**	**Tue**	**Wed**	**Thur**	**Fri**
1	53	56	44	57	51
2	46	58	53	59	46
3	47	56	55	44	57
4	58	53	46	44	51
5	50	55	55	46	46
6	54	55	44	51	53
7	54	54	54	49	55
8	46	58	52	51	58
9	46	49	46	45	52
10	54	47	55	45	47
11	48	51	46	54	49
12	58	45	55	44	45
13	56	44	54	56	52
14	49	48	55	53	57
15	59	45	54	58	50
16	53	50	44	55	53
17	54	50	59	45	52
18	58	51	55	47	55
19	56	44	46	52	53
20	54	47	51	54	59

Table 1. SR Mattress Company Data

The first question one needs to answer before analyzing the data is "How will the subgroups be formed?" We will address this issue later, but to keep things simple, we will define the subgroups as being made up of the 5 daily outputs for each shift ***per week***. In their respective time series plots, x-bar equals 51.42 with the lower and upper control limits of 44.931 and 59.909, respectively. [When using StatGraphics, grid lines appear in the graphs and the control limits are not initially shown. One can insert the control limits by left clicking on the graph (pane) and then right clicking in order to "pull up" the options selection. We eliminated the background grids in order to highlight the other features.

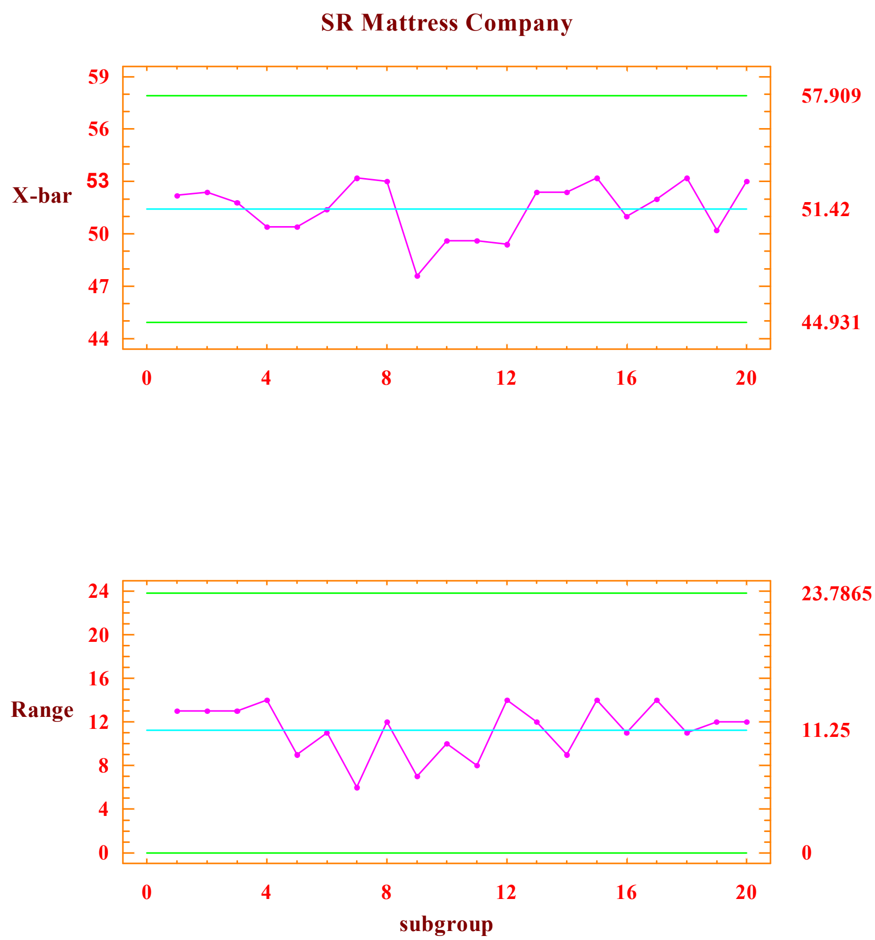


Figure 4. SR Mattress Company X-Bar and Range Chart

From these charts, the X-bar chart and range chart, we can see that none of the values are outside the control limits, thereby suggesting a possible stable process. On closer examination one may see some possible patterns that should be investigated for possible sources of specific variation. Do you see any such patterns? If so, what might be a possible scenario to describe the pattern and what type of action might management take if your scenario is true.

Given the previous example, hopefully the reader has an intuitive feel for what X-bar charts and R charts represent. We will leave it to the computer to calculate the upper and lower control limits.

Before moving on, we need to take another look at the question about how the subgroups were defined. The division described above will highlight differences between different weeks. However, what if there was a difference between the days of the workweek? For example, what if a piece of required machinery is serviced after closing every Wednesday, resulting in higher outputs every Thursday, would our sub-grouping detect such an impact? In this case one might choose to subgroup by day of the week. Hopefully, one can see how the successful implementation of control charts may depend upon the design of the control chart itself that is a function of knowing as much as possible about possible sources of specific variation.

Two final points about continuous variable control charts. The first is that when the subgroups are of size one, the X-bar chart is the same as a chart for the original series. In this case the R chart may be replaced by a moving average chart based upon past observations. The second point is that in our scenario we required each subgroup to be of the same size (equal number of observations). For example, what if there were holidays in our sample? In this case an R chart, where the statistic of concern is the range, could be replaced by an S chart, which relies on the sample standard deviation as the statistic of concern. In practice, the R charts are used more frequently with exceptions such as the holiday situation just noted.

P Charts

The P chart is very similar to the X-bar chart except that the statistic being plotted is the sample proportion rather than the sample mean. Since the proportion deals with the percentage of successes[6], clearly the appropriate data for P charts needs to be attribute data where the outcomes for each trial can be classified as either a success or a failure (conform or non-conform, yes or no, etc.). The subgroup size must be equal so the proportion can be determined by dividing the outcome by the subgroup size.

To illustrate the P chart, a situation is considered where we are concerned about the accuracy of our data entry departments work. In auditing their work over the last 30 days, we randomly selected a sample of 100 entries for each day and classify each entry as correct or incorrect. The results of this audit are as follows:

Day	# Incorrect	Day	# Incorrect
1	2	16	1
2	7	17	5
3	6	18	9
4	2	19	6
5	4	20	4
6	3	21	3
7	2	22	3
8	6	23	5
9	6	24	3
10	2	25	6
11	4	26	6
12	3	27	5
13	6	28	2
14	2	29	3
15	4	30	4

Table 2. Number Incorrect Entries in Sample Size of 100

Given the data above, one can easily calculate the proportions of incorrect entries per day by taking the number of incorrect entries and dividing by the total number of entries for that day, which in our example were 100 each day. This may seem to be an unnecessary task at this time, since we are

[6] Recall the binomial distribution where one of the parameters is the probability of success.

essentially just scaling the data. This scaling, however, does allow us to work with the P statistic, rather than the total number of occurrences that would produce another type of chart called the NP chart. We have chosen not to discuss the NP chart since it provides the same information as the P chart for subgroups of the same size, while the P chart allows us more flexibility, so that we can consider cases when the subgroups are not all of the same sample size.[7] The P chart for the data entry example is shown below.

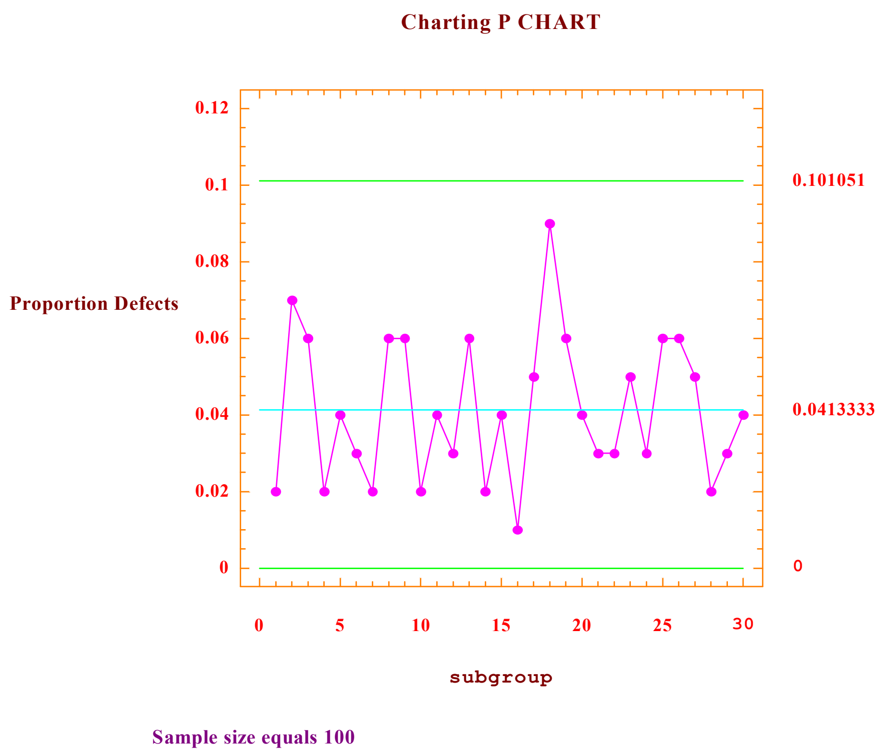


Figure 5. Proportion Control Chart

From the P chart displayed above, one can see that all of the observed values fall within the control limits and that there does not appear to be any significant pattern. One might be concerned with the

[7] When the sample sizes are different the calculations become more complicated. For our purposes we will just note this and leave the details for the software programmers.

value for the 18th observation that is .09 and look to see if a particular event triggered this *larger* value. Keep in mind, however, that common variation may very well cause this *larger variation*.

C Charts

The C chart is based upon the statistic that counts the number of occurrences in a unit, where the unit may be related to time or space. Whereas the P chart was related to the binomial distribution, the C chart is related to the Poisson distribution. To demonstrate the C chart we consider a situation where we are interested in the number of defective parts produced daily at the AKA Machine Shop. Over the past 25 days the number of defective parts per day are shown below:

Day	**# Defective Parts**	**Day**	**# Defective Parts**
1	5	14	7
2	10	15	3
3	7	16	4
4	5	17	8
5	8	18	5
6	8	19	3
7	8	20	6
8	5	21	10
9	7	22	1
10	7	23	6
11	10	24	5
12	6	25	4
13	6		

Table 3. Number of Defective Parts per Day

The C chart[8], which appears on the next page, shows that the process appears to be stable. In particular, there are no values outside the control limits, nor does there appear to be any systematic pattern in the data. (Note: no reference made to sample size.)

[8] The notation in the StatGraphics software may confuse you as it relates the C chart option with "count of defects" and the U chart option with "defects per unit". We are not discussing the U chart in class or this write up. The U chart allows for the "units" to change from subgroup to subgroup.

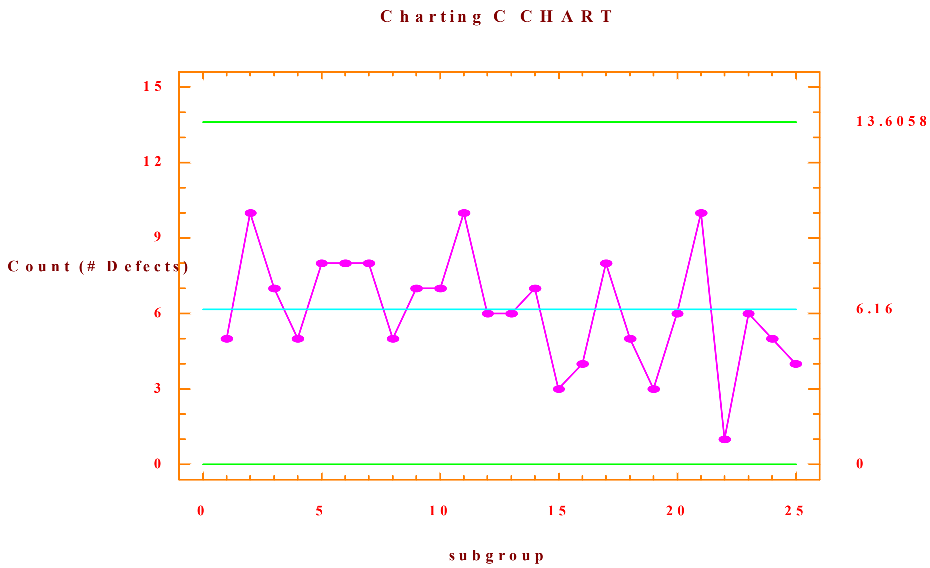


Figure 6. Count Control Chart

Conclusion

In our discussion of control charts we first discussed the common attributes of different control charts available (center line, upper control limit and lower control limit) and focused on what one looks for in trying to detect sources of specific variation (outliers, trends, oscillating, seasonality, etc.). We then looked at some of the most commonly used control charts in practice, namely the X-bar and R, P, and C control charts.

What differentiates the various control charts is the statistic that is being plotted. Since different types of data can produce different types of statistics it is clear that the type of data available will suggest the type of statistic that can be calculated and hence the appropriate control chart.

One final but important point is that the control charts generated, including those in this write up, frequently use the data set being examined to construct centerline and control limits (upper and lower). The problem this may cause is that if the process is unstable then the data it generates may alter the components of the control chart (different centerlines and different control limits) and hence be unable to detect problems that may exist. For this reason, in practice, when a process is believed to be stable the resulting statistics are frequently used to establish the control limits (center, upper, and lower) for future windows. What we mean by *window* is that if we decide to monitor say 30 subgroups at a time, as time evolves subgroups are added and consequently the same number are dropped from the other end, hence a revolving window. Useful software, such as StatGraphics, will allow one to specify the limits as an option.

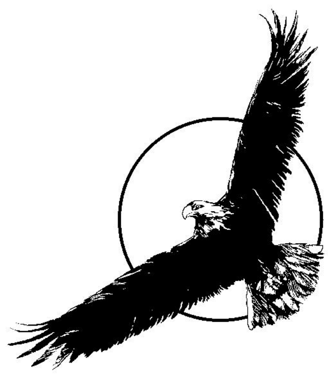
In summary:

- X-bar and Range charts are used when sample subgroups are of equal size, sample subgroups are taken at equal time intervals, and the subgroup means and range of highest and lowest values are of interest.

- Proportion charts are used when samples are of equal size and the defect proportions are of interest.

- Count charts are used when either the sample size is unknown or the sample sizes are not uniform.

[intentionally left blank]



TRANSFORMATIONS & RANDOM WALK

In the previous chapter we focused our attention on viewing variability as being comprised of two parts, common variation and specific variation. With the exception of manufacturing systems, most economic variables when viewed in their measured formats demonstrate sources of specific variation. In data analysis, whether we are trying to forecast or explain economic relationships, our goal is to model those sources of specific variation with the result being that only common variation is "left over." This can be depicted by the expression:

ACTUAL = FITTED + ERROR.

Where the FITTED values are generated from the model (specific variation), the ACTUAL values are the observed values and the ERROR values represent the differences and are a function of common sources of variation. If the common sources of variation of the model appear to be random, the model may better predict future outcomes as well as providing a more thorough understanding of how the process works.

Random Walk

One of the simplest, yet widely used models in the area of finance is the random walk model. A common and serious departure from random behavior is called a *random walk*. By definition, a series is said to follow a random walk if the first differences are random. What is meant by *first differences* is the difference from one observation to the next, which if you think about as the steps of a process and the sequence of steps as a walk, suggest the name random walk. (Do not be mislead by the term "random" in "random walk." A random walk is not random.) Relating this back to the equation we see that the ACTUAL values are the observed values for the current time period, while the FITTED values are the last periods observed values.

Hence we can write the equation as:

$$X_t = X_{t-1} + e_t$$

where: X_t is the value in time period t,
X_{t-1} is the value in time period t-1 (1 time period before)
e_t is the value of the error term in time period t.

Since the random walk was defined in terms of first differences, it may be easier to see the model expressed as:

$$X_t - X_{t-1} = e_t$$

Therefore, as one can see from the resulting equation, the series itself is not random. However, when we take the first differences the result is a transformed series $X_t - X_{t-1}$, which is random.

To illustrate the random walk model, we consider the series of stock prices for Nike as it was posted on the New York Stock Exchange at the end of each month, from May 1995 to May 2000. The time sequence plot of the series Nike (see data file) is shown in the figure below.

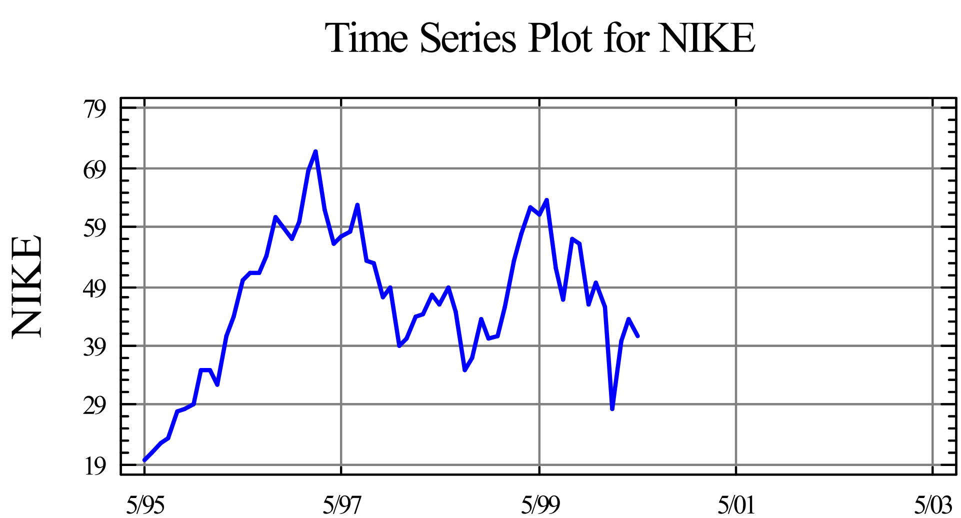


Figure 1. Times Sequence Plot of Nike

As one can see the original series for Nike does not appear to be random. In fact, when the nonparametric runs test is performed on the original series, the p-value is 0.000020, which indicates compelling evidence to reject the null hypothesis. Hence, the *original* series of Nike is not random.

H_0: The [original] series is random[9]
H_1: The [original] series is **NOT** random

Now consider the first differences of Nike with the time series plot shown below:

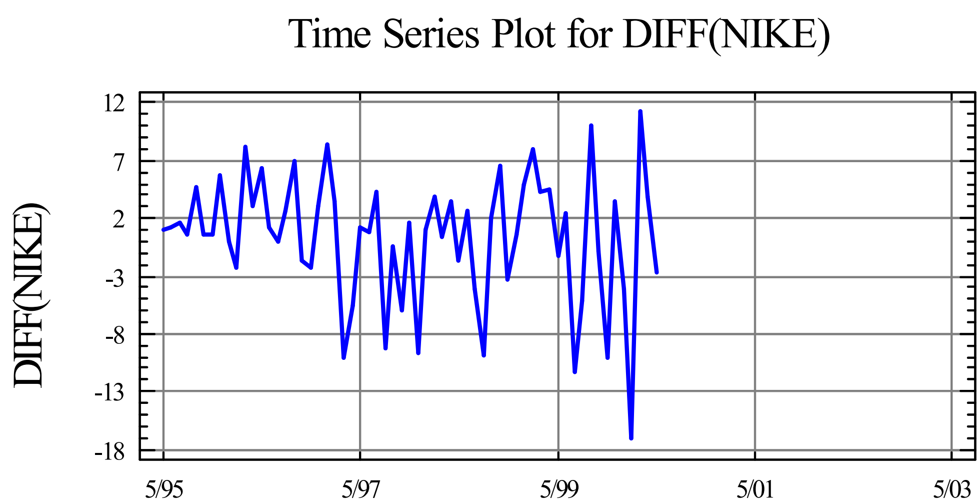


Figure 2. First Differences of Nike

As we can see from the time series plot, by taking first differences the transformed series appears to be random. (Note that we are only discussing whether the series is random, nothing is being said about it being stable since the variance increases with time.) To confirm our visual conclusion that the differenced series is random, we perform the runs test and find out that the p-value is 0.7191.

[9] The use of [original] is for emphasis only ... it is not normally used when stating the null hypothesis.

The p-value exceeds $\alpha = 0.05$ and thus provides supporting evidence to retain the null hypothesis, the differenced series is random, and thus the stock price of Nike tends to follow a random walk model.

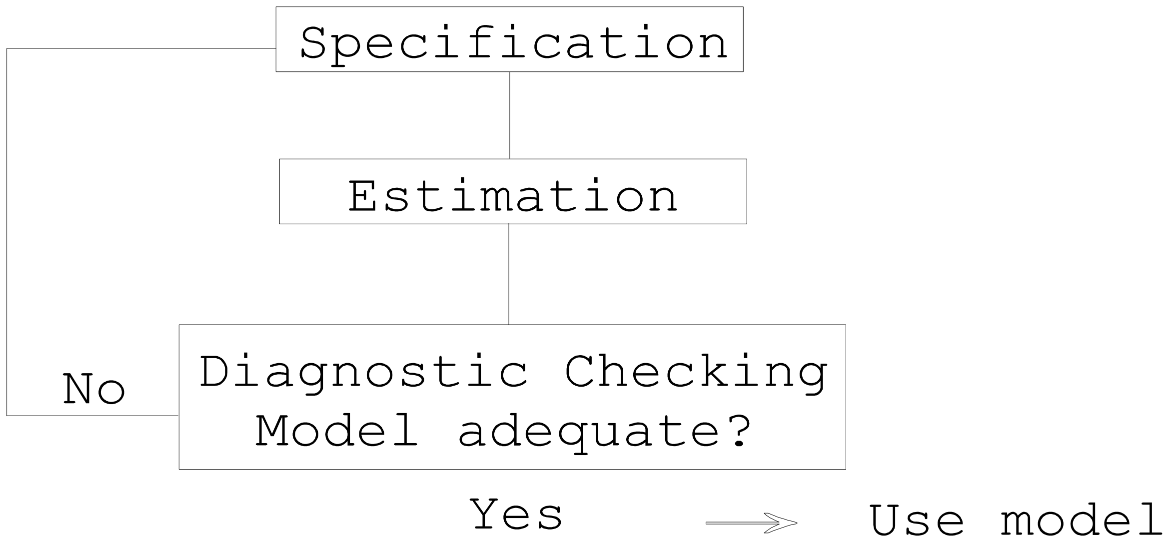
H_0: The (first differenced) series is random[10]
H_1: The (first differenced) series is **NOT** random

Information is not lost by differencing. In fact, use of differencing, or inspecting changes, is a very useful technique for examining the behavior of meandering time series. Stock market data generally follows a random walk and by differencing, we are able to get a simpler view of the process.

[10] Use of [first differenced] for emphasis only. (See footnote 10.)

MODEL BUILDING

Building a statistical model is an iterative process as depicted in the following flowchart:



As one can see, when constructing a statistical model for use there are three phases that must be followed. In fact most models used in practice require going through the three phases multiple times, as seldom is the model builder satisfied without refining the initial model at least once.

Each of these phases is discussed below in general terms, for all statistical models, and later will be described in detail for specific models (regression, time series, etc.)

Specification

The specification or identification phase involves answering two questions:

1. What variables are involved?

and

2. What is the mathematical relationship between variables?

When establishing a mathematical model there are parameters involved which are unknown to the practitioner. These parameters need to be estimated, hence, the need for the estimation phase which is discussed in the next section. When answering the questions above, it is essential that the model builder use economic theory to help establish a tentative model. A model that is based upon theory has a much better chance of being useful than one based upon guesswork.

Estimation

As mentioned previously, the models developed in the specification phase possess parameters that need to be estimated. To obtain these estimates, one gathers data and then determines the estimates that best fit the data. In order to obtain these estimates, one has to establish a criterion that can be used to ascertain whether one set of estimates is "better" than another set. The most commonly used criterion is referred to as the least squares criterion which, in simple English, means that the error terms which represent the differences between the actual and fitted values, when squared and added up will be minimized. The reason for using the squared terms is so that the positive and negative residuals do not cancel each other out. For our purposes, it will suffice to state that the computer will generate these values for us by using StatGraphics Plus.

Diagnostic Checking

The third phase is called the diagnostic checking phase and basically involves answering the question:

Is the model adequate?

If the answer to the above question is ***no***, then something about the model needs modification and the builder returns to the specification phase and goes through the entire three phase process again. If the answer to the above question is ***yes***, then the model is ready to use.

When in the process of discerning whether the model is adequate, a number of attributes about the model need to be considered:

1. How well does the model fit the data?
2. Do the residuals (actual - fitted) from the model contain any information that should be incorporated into the model? (i.e. is there information in the data that has been ignored in the creation of the model.)
3. Does the model contain variables that are useless and hence should be eliminated from the model?
4. Are the estimates derived from the estimation phase influenced disproportionately by certain observations (data)?
5. Does the model make economic sense?
6. Does the model produce valid results?

As stated previously, when the model builder is able to answer affirmatively to each of the above questions, *and only then*, are they able to use the model for their desired purpose.

[intentionally left blank]