**Chapter 6**

**Comparison of Means**

with Munir Ahmed

## [H1] Learning Objectives

## [INSERT NL]

1. Describe normal distribution
2. Detect outliers in the data
3. Test hypotheses and generalize beyond the sample
4. Conduct one-sample test of population means
5. Compare two-sample means
6. Create an X-bar control chart
7. Set upper and lower control limits
8. Examine in- and out-of-control observations
9. Create a risk-adjusted control chart

**[END NL]**

## [H1] Key Concepts

## [INSERT BL]

* Standard normal distribution
* Outlier
* Hypothesis tests
* Confidence intervals
* X-bar control chart
* Control limits
* Risk-adjusted control limits

## [END BL]

## [H1] Chapter at a Glance

In a sample of data, values for variables fluctuate. These shifts could be purely random. Managers and improvement teams need to separate out random changes from real changes. The magnitude of the fluctuations tells us a lot about whether changes are real or random—small ones are considered to be random, and large ones to be real changes in the value of the variable. Change in the magnitude of variables can be traced through changes in the mean of the variable.

When we want to compare two means, we have to compare the distributions of the observed values. A distribution shows the fluctuations in the estimate of the mean, mapping the value of a variable to the probability of observing the value. Equipped with these probabilities, managers can infer whether observed changes in a variable are small and possibly random or large and more likely to be real. This chapter shows how distributions are used to calculate that the probability of observing the difference in two means is so low as to be considered random chance. The chapter ends with ways to examine differences in mean over time, a procedure useful for quality control. The procedures at end of this chapter can help improvement teams test whether the changes they have introduced have led to real improvements. Chapter 6 can help teams distinguish between fake claims and true improvements.

## [H1] Normal Distribution

The distribution of a continuous random variable is called a *continuous probability density function*, *continuous probability distribution*, or just *continuous distribution*. *Normal distribution* is a particular type of continuous distribution that is very common. Many natural phenomena have normal distribution, including average cost, average satisfaction level, average blood pressure, or for that matter, as we will see shortly, the average of almost anything. The *normal probability distribution* is also important because it can approximate the distributions of various discrete random variables. The normal probability density curve for a continuous random variable *x* can be given by the mathematical expression

**[INSERT EQUATION]**

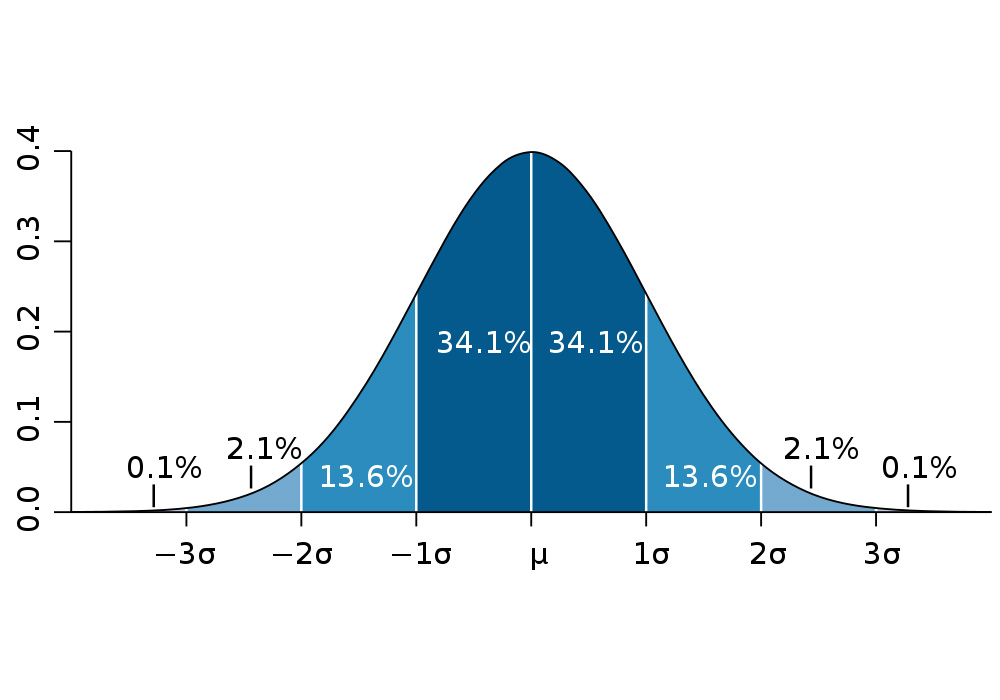
,

**[END EQUATION]**

where  is the population mean, and  is the population standard deviation of *x*, and the values of  and *e* are approximately 3.14159 and 2.71828, respectively. Because  and *e* are constants, the probabilities of random variable *x* are completely determined once the values of  and  are known. These two latter values are thus the parameters of the normal distribution. The normal distribution is displayed in exhibit 6.1.

**[INSERT EXHIBIT; please render in gray scale; convert light blue to dark blue, medium blue to medium gray, and dark blue to black]**

**Exhibit 6.1** Normal Distribution



**[END EXHIBIT]**

A normal distribution has the following properties:

**[INSERT BL]**

* It is a bell-shaped distribution.
* It is symmetric. Values around the average, in exhibit 6.1 shown as, are mirror images of each other.
* Three measures of central tendency—the mean, the median, and the mode—are all equal for this distribution. In exhibit 6.1, the mean is shown.
* Approximately 68 percent, 95 percent, and 99 percent of the data under the normal curve are contained within the 1, 2, and 3 standard deviations, respectively.
* The range for this distribution is.
* Total area under the normal curve is exactly 1, or 100 percent.

**[END BL]**

### [H2] Example 1: Normal Distribution

Suppose you are a regional clinic manager, and you want to understand what your competitors are charging for influenza vaccines. You discover that five clinics charge $30.00, $15.00, $20.00, $25.00, and $20.00, respectively. You can now calculate the average cost of influenza vaccine in your market like this:

**[INSERT EQUATION]**

.

**[END EQUATION]**

If you wanted to see the range into which the majority (68 percent) of the prices fell, you would first calculate the standard deviation:

**[INSERT EQUATION]**

**[INSERT EQUATION]**

**[END EQUATION]**

Therefore, 68 percent of the flu shot costs in the region are within one standard deviation ($5.70) of the mean ($22.00), so they range from $22.00 + $5.70 = $27.70 or $22.00 − $5.70 = $16.30.

## [H2] Standard Normal Distribution

The probabilities associated with normal distribution of *x* depend on the mean of *x* and its variance. The mean and standard deviation, of course, depend on the observed values of *x*. To easily discover probabilities of observing a particular value in the normal distribution, statisticians have created the standard normal distribution. The standard normal distribution has a mean of 0 and a standard deviation of 1. The standard normal distribution can be estimated by subtracting the mean from each observation of *x* and dividing by the standard deviation of *x*. The formula

**[INSERT EQUATION]**

**[END EQUATION]**

can be used. In this equation, *z* denotes the new standard normal variable, is the observed value of the variable, is the average of the observed values, and *s* is the standard deviation of the observed values. This formula ensures that *z* has a mean of 0 and a standard deviation of 1. The probability density function of the standard normal variable is

**[INSERT EQUATION]**

.

**[END EQUATION]**

Note that the standard normal distribution does not depend on the average or standard deviation of the variable; therefore, these values can be calculated beforehand and used only when needed. The probability of *z* falling between any two values can be calculated by examining the area under the earlier formula and reported in a table for reuse when needed. Because the normal distribution of a variable is continuous, the probability for any one value of the random variable is always 0. In other words, the chance of observing any one value is always 0. To properly use the normal distribution, and any continuous probability distribution, we need to calculate the area under the curve using two values of the random variable. We can also report a nonzero probability if we talk about *z* exceeding or being less than a single value, as again we can calculate the area under the curve that exceeds or is below *z*. Statisticians often refer to *one-sided tests* in which they examine the probability of a value exceeding a constant. Similarly, one can do a one-sided test of a value being less than a constant. In contrast, the *two-sided test* examines whether the probability of the observed values is above or below cutoff values.

### [H2] Example 2: Standard Normal Distribution

Assume that the average length of stay (LOS) for individuals having cardiac bypass surgery is normally distributed with a mean of 9 days (*µ* = 9.0) and a standard deviation of 1.25 days  
(*σ* = 1.25). You want to find out the probability that a random bypass patient will have a LOS of less than 8.0 days. Schematically, this is represented by P (*x* <8.0), and this is considered a one-sided test. We are only looking for the probability that a number is *less* than a certain number (vs. both less than and more than a certain number).

First, convert 8.0 to a standard *z* score using the equation

**[INSERT EQUATION]**

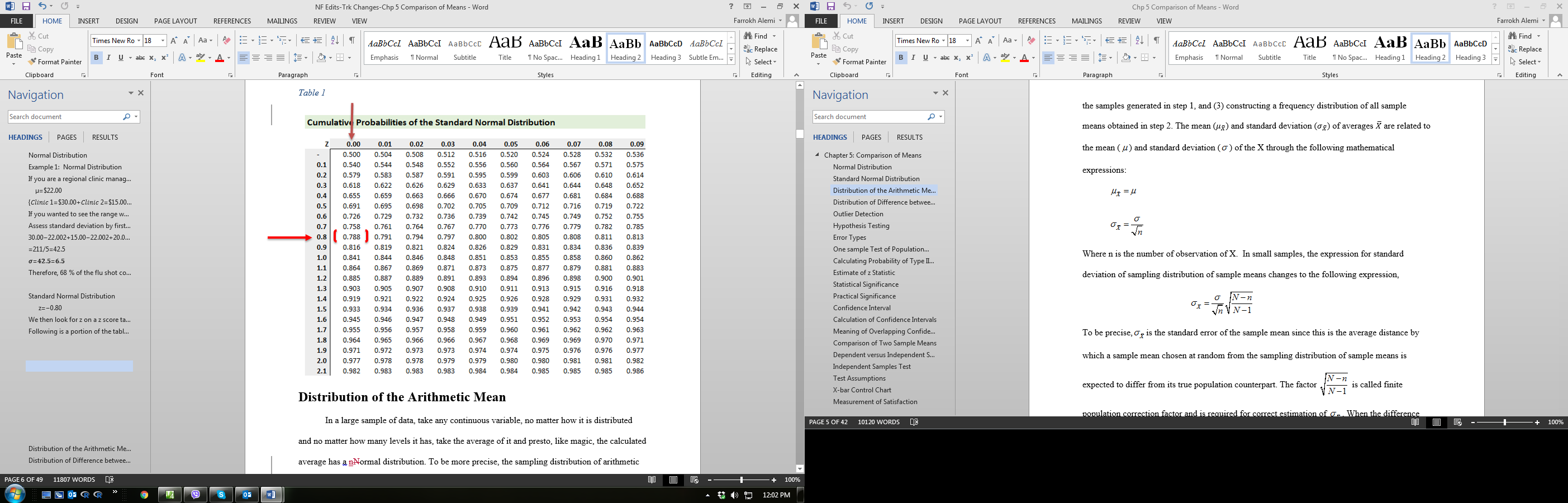
.

**[END EQUATION]**

Then the analyst looks for the probability of observing the *z* value of .8 in a standard normal table. P (*z* <0.80) is found by looking at the rows and columns of the table. In exhibit 6.2, we can see that the row *z* = 0.8 and the column 0.00 show the probability of observing *z* as .800. The cell corresponding to row 0.8 and column 0.00 shows the number 0.788. Therefore, the probability of observing *z* <0.8 is 78.8 percent. The probability of observing a *z* value greater than or equal to 0.8 is 1 − 0.788. If we are doing a two-sided test, we need to estimate the probability that *z* <−0.8, which we can do through the same table after a bit of rearrangement. Because standard normal tables are symmetric, we know that at mean 0, the probability is 0.5 (i.e., 50 percent of the data fall below the mean). The probability from −0.8 to 0 is the same as 0 to 0.8, so the change from the mean is 0.788 − 0.5 = 0.288. Now we can estimate that the probability of *z* <−0.8 is 0.5 − 0.288 = 0.212. In this way, we can determine that 21.2 percent of patients will have a stay of fewer than eight days.

**[INSERT EXHIBIT]**

**Exhibit 6.2** Cumulative Probability Distribution



## [END EXHIBIT]

## [H2] Everything Becomes Normal

In a large sample of data, take the average of any continuous variable—no matter how it is distributed and no matter how many levels it has—and presto, like magic, the calculated average has a normal distribution. To be more precise, the sampling distribution of the arithmetic mean can be obtained by the following steps:

**[INSERT NL]**

1. Draw all possible samples of a fixed size *n* from a population of size *N*, where *N* is very large relative to *n*.

2. Calculate a sample arithmetic mean from each of the samples generated in step 1.

3. Construct a frequency distribution of all sample means.

**[ND NL]**

The mean () and standard deviation () of averages are related to the mean () and standard deviation () of the *x* through the mathematical expressions

**[INSERT EQUATION]**

and

**[END EQUATION]**

In this equation, *n* is the number of observations of *x*. In small samples, the expression for the standard deviation of the sampling distribution of sample means changes to the expression

**[INSERT EQUATION]**

.

**[END EQUATION]**

In fact,is the standard error of the sample mean, because this is the average distance by which a sample mean chosen at random from the distribution is expected to differ from its true population counterpart. The factor

**[INSERT EQUATION]**



**[END EQUATION]**

is called the *finite population correction factor*,and it is required for the correct estimation of . When the difference between *N* and *n* is large, the value of the finite population correction factor is approximately 1.

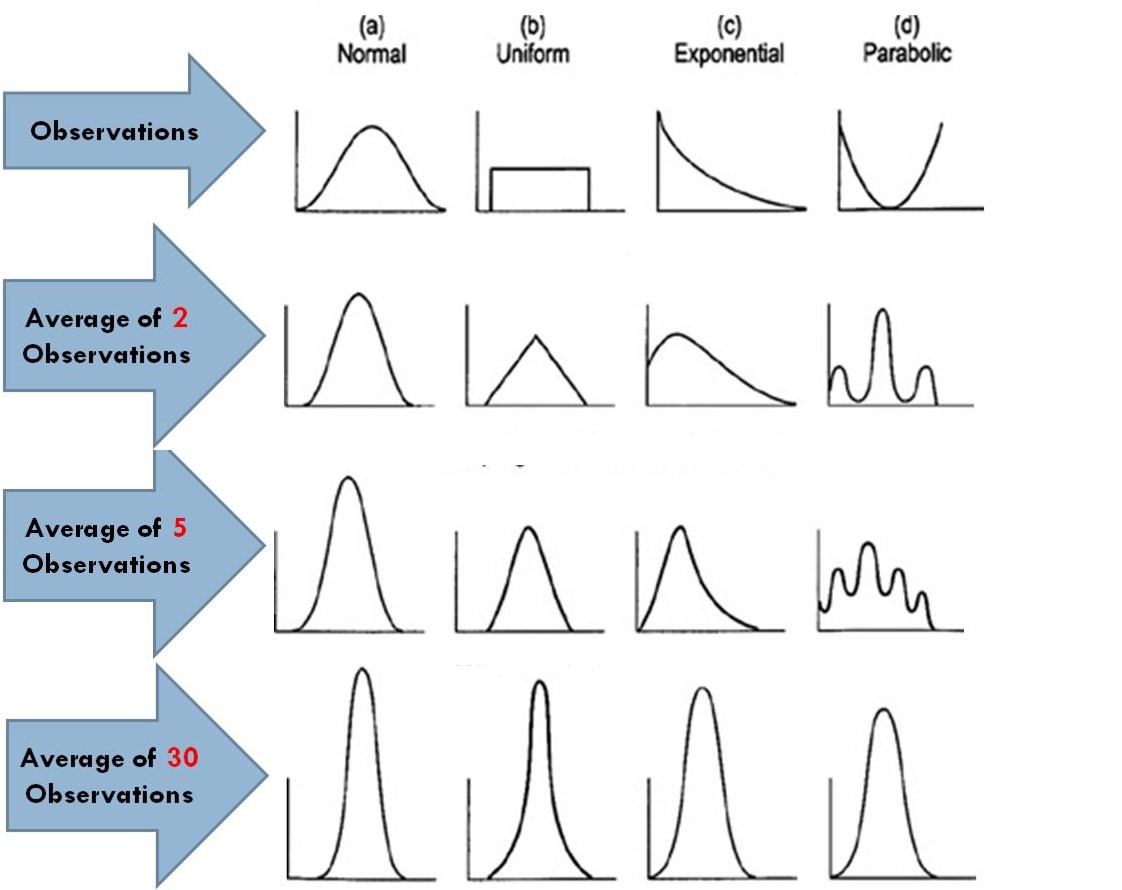
The *central limit theorem* is one of the most important theorems underlying classical inferential statistics. In basic terms, the theorem states that as sample size *n* becomes larger, the mean becomes more normally distributed. When a sample is sufficiently large (usually taken to mean 30 or more cases), we can assume that the mean of the sample has a normal distribution. The central limit theorem works for any distribution of the observations. In other words, given a sufficiently large sample size, the normality of distribution of the mean is guaranteed for all observations regardless of their distributions. For observations that are initially normal, the mean has a normal distribution with a sample size as small as two cases. For symmetrical distributions, a sample smaller than 30 is sufficient. For asymmetrical distributions, a minimum of 30 may be required.

The central limit theorem works only when a very large number of samples are drawn from a target population. For small populations, accuracy often requires sampling with replacement. This need for a large number of samples is often referred to as the *law of large numbers*. The law of large numbers and the central limit theorem usually work hand in hand, and their mechanics can be easily demonstrated by constructing a simple simulation in Excel.

Exhibit 6.3 shows the distribution of both the original observation and the average of the observations as the number of cases averaged increases. In the exhibit, we start with different types of distributions. The initial row shows *uniform, exponential*,and *parabolic* distributions. These distributions are not symmetric and unimodal. Clearly, these distributions are not normal. In the second row, we repeatedly take two observations from the distribution and plot the average of these two observations. Already, the distributions are becoming more like normal distributions. In the third row, we have plotted the distribution average of five observations. The symmetric unimodal shape of normal distribution can be seen clearly for uniform and exponential observations. The average of five observations from exponential distributions does not look completely symmetric, but it is unimodal. Even the average of five observations from parabolic distribution seems to wiggle its way to a near-normal, unimodal, and symmetric distribution. In the last row, we see the distribution of the average of 30 observations; all non-normal distributions are nearly normal. As the number of observations increases, the average increasingly has a normal distribution. This makes normal distribution relevant for the analysis of average of observations, even when these observations are not normal. Like the phrase “all roads lead to Rome,” as the number of cases in the sample increases, the average ends up having a normal distribution.

**[INSERT EXHIBIT; convert to gray scale; turn blue into medium gray, turn text in arrows white, and make red text black]**

**Exhibit 6.3** Effect of Averaging on Progression Toward Normal Distribution

****

**[END EXHIBIT]**

For any two variables *x* and *y* that are random and normally distributed, their difference is also a randomly distributed normal variable. Thus, whenever we are interested in comparing the difference between two arithmetic means whose sampling distributions are approximately normal, the central limit theorem and the law of large numbers both apply. In this situation, it is useful to remember that for two random variables *x* and *y*, the mean of the difference of these variables is equal to the difference of their means (i.e., ).

## [H2] Outlier Detection

As we noted earlier, approximately 68.27 percent, 95.45 percent, and 99.7 percent of the data under the normal curve are contained within the 1, 2, and 3 standard deviations, respectively. In standardized form, these areas refer to *z* values of 1, 2, and 3, respectively. These cutoffs are often used to identify *outliers*, or extreme values. When normally distributed *x* values are transformed into *z* scores, we expect 95 percent of the variable values to lie within a *z* score of 2, and 99 percent of the variable values to lie within a *z* score of 3. Values outside the 95 percent range are typically classified as outliers that merit individual investigation.

## [H1] Hypothesis Testing

The term *hypothesis testing* refers to a series of steps that statisticians perform when generalizing information from a sample to its corresponding population. A test of hypothesis can be constructed for any population parameter of interest, including the population mean. A hypothesis test about the unknown population mean () refers to calculating the sample mean and assessing the likelihood of that sample mean being reasonably close to. Hypothesis testing involves a well-defined set of tasks, including the following six:

**[INSERT NL]**

1. *Null and alternative hypotheses*. These are two contradictory and mutually exclusive terms. The null hypothesis,, is representative of the status quo, while the alternative hypothesis,, reflects the research question being evaluated (i.e., the statement for which we are seeking evidence). In the absence of any evidence to the contrary, the null hypothesis is considered to be true. For example, a manager believes that the actual number of average hours worked by employees per week is less than the 40 currently understood to be true. In this example, because the burden of proof lies on the manager, the null hypothesis is that the mean number of hours is 40. The alternative statement could be either that the number of worked hours is different from 40 (a two-tailed test) or that the number of worked hours is less than 40 (a one-tailed test). The choice between these two alternatives really depends on the context of the analysis and the philosophy that the analyst adheres to.
2. *Level of significance*. Because the whole population is not visible and only a sample is used in order to make generalizations about unknown population parameters, error in hypothesis testing is inevitable. The error can take two forms. *Type I error*, denoted by  and also known as the *level of significance*, refers to a situation in which a particular null hypothesis is true but, based on sample evidence, the analyst rejects it. *Type II error*, denoted by, refers to the error of not rejecting a false null hypothesis. Type I error is examined in a test of hypothesis. In most studies, the probability of type I error is fixed at .05, or 5 percent. This means that the analyst considers it acceptable to be wrong 5 out of 100 times, if the same study is repeated a very large number of times. For situations in which a 5 percent chance of error is considered too high, a lower significance level, such as 1 percent or 0.1 percent, can be selected.
3. *Test statistic*. This is a mathematical expression that is based on the design of the study being conducted and the choice of the appropriate distribution. Examples include using a *z* distribution to conduct a one-sample (or two-sample) test about the population means, an *F* test comparing the means of two or more independent groups, or a  test for association between two categorical variables.
4. *Observed value of the test statistic*. This value is representative of sample results. The objective is to compare the observed value of the test statistic with a critical value (next step) in order to arrive at a decision about the null hypothesis. The observed value of the test statistic is obtained by resolving the expression (or formula) for the test statistic.
5. *Critical value of the test statistic*. This value represents the theoretical or expected value of the test statistic; it is representative of the distribution of the unknown population parameter that is considered true under the null hypothesis. Critical values can be obtained from published tables that are usually available at the end of most standard statistics textbooks or can be calculated by using simple functions in a spreadsheet program such as Excel.
6. *Conclusion*. In this step, sample (or observed) results are compared with the hypothesized (or theoretical) values to let the analyst arrive at a formal conclusion about the null hypothesis. Generally, the null hypothesis is rejected if, in absolute terms, the observed value of the test statistic exceeds the corresponding critical value. The analyst fails to reject the null hypothesis if the absolute value of the test statistic is smaller than the corresponding critical value.

**[END NL]**

The only two options at this stage are (1) to reject the null hypothesis or (2) to fail to reject the null hypothesis. The null hypothesis can never be completely accepted because doing so requires examining each element in the population—which, if it is even possible, defeats the main purpose of hypothesis testing. It should be noted that this approach of comparing the critical and observed values of the test statistic to arrive at a conclusion about the null hypothesis is known as the *critical value approach*.

## [H2] Error Types

In a test of a hypothesis, the analyst’s decision can lead to four possible situations, two correct and two incorrect. These are discussed in the following list and summarized in exhibit 6.4.

**[INSERT BL]**

* *Type I error.* As noted earlier, the probability of rejecting the null hypothesis when it is true is called a type I error, represented by. Other names for this error include *level of significance* and *false positive*. For example, imagine a healthcare manager who wants to determine whether mothers like the new maternity birthing rooms, and he queries all of the women who gave birth in the month of August (when the new birthing rooms had been open for the full calendar year). The null hypothesis is that the mothers do not like the new rooms; the alternate hypothesis is that the women do like the new rooms. The majority of the sample of women having babies in August liked the rooms, but during the rest of the year, the majority of the women did not. If the manager assumes that all the women liked the rooms, a type I error has occurred.
* *Confidence level*. When the null hypothesis is true and, based on sample evidence, it is not rejected, no error is involved. The probability of this outcome is known as the *level of confidence* or *true negative*. This probability is the complement of.
* *Type II error.* The probability of not rejecting the null hypothesis when it is false is a type II error, represented by. Alternately, this error is referred to as a *false negative*. For example, suppose a clinic administrator needs to understand how many urinalysis tests are done in a year. She samples all the tests done in a year by counting all the urinalysis tests done on Fridays. The null hypothesis is that there are fewer than 500 tests done per month; the alternative hypothesis is that there are more than 500 tests done per month. If a type II error occurs, the manager will incorrectly fail to reject the null hypothesis or, in this case, accept that fewer than 500 urinalysis tests are done per month—when in truth more than 500 tests are done per month.
* *Power of the test.* This is the probability of rejecting the null hypothesis when the null hypothesis is indeed false. This probability is also called a *true positive*, and it is the complement of. It should be noted that in any particular scenario, the null hypothesis is either true or false. If the null hypothesis is true, then based on sample evidence, it will be either rejected or not rejected. Thus, the level of significance and the level of confidence always add up to 1. Similarly, if the null hypothesis is false, then based on sample evidence, it will be either rejected (no error) or not rejected (type II error). Thus the probability of a type II error and a power of the test of a hypothesis complement each other and always add up to 1.

**[END BL]**

**[INSERT EXHIBIT]**

**Exhibit 6.4** Types of Errors

|  |  |  |  |
| --- | --- | --- | --- |
|  | | *True State of Hypothesis* | |
| *True* | *False* |
| Decision based on evidence from sample | Fail to reject hypothesis | Correct decision, true negative | Type II error, false negative |
| Reject hypothesis | Type I error, false positive | Correct decision, true positive |

**[END EXHIBIT]**

## [H2] One-Sample Test of Population Mean

One of the simplest tests of hypothesis for a variable in any population of interest concerns the *measure of center* of that variable. Given its analytically desirable mathematical properties, the arithmetic mean is usually the measure of choice in such situations. The simplest test of hypothesis about an unknown population mean involves comparing the value of the sample mean with a corresponding hypothesized value to determine whether the two values come from the same population. This test, known as a one-sample *z* test, can be performed by following the general steps in a test of hypothesis that were discussed earlier in this chapter.

For example, a particular health-related standardized test has a mean score of 500 and a standard deviation of 100 in the population. Test the claim that a sample of 30 students with a mean test score of 525 and standard deviation of 75 comes from this population.

**[INSERT EQUATION]**

Null hypothesis: .

Alternative hypothesis:.

Level of significance: = .05.

Test statistic: .

Observed value: .

Critical value: |*z*| = 1.96.

**[END EQUATION]**

Because the critical value exceeds the observed value in absolute terms, we fail to reject the null hypothesis.

Statisticians use the term *failure to reject*, a double negative, to emphasize that statistical tests cannot accept a hypothesis. While managers and policymakers do not have the luxury of speaking in double negatives, they would conclude from the statistician’s findings that the sample comes from a population with a mean score of 500 and a standard deviation of 100. Statisticians are careful about their statements, whereas managers and policymakers are being practical.

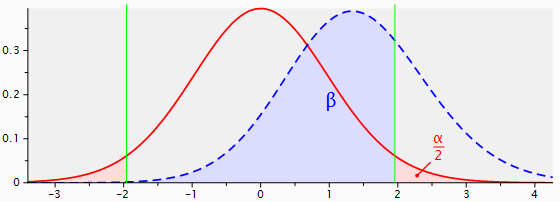
## [H2] Calculating Probability of a Type II Error

The level of significance or probability of a type I error is explicitly fixed in a test of hypothesis. This is not true for the probability of a type II error. In a one-sample *z* test about****, the probability of a type II error depends on a number of factors, such as the maximum expected difference between the  and its hypothesized value (effect size), the probability of a type I error, and sample size *n*. For the one-sample *z* test illustration presented earlier, the probability of a type II error can be calculated as outlined in the following sections.

Our null (solid red line) and alternative (dashed blue line) hypotheses are shown in standardized form in exhibit 6.5. Because the probability of a type II error is the probability of not rejecting the null hypothesis when, in fact, the null hypothesis is false, we are basically interested in calculating the shaded region in the graph.

**[INSERT EXHIBIT; convert to gray scale. Graph lines and green lines should be black; background should be transparent; lavender section under graph line should be medium gray]**

**Exhibit 6.5** Error Types



**[END EXHIBIT]**

The first step is to use information from the distribution of under the null hypothesis to calculate critical values of *z*. These critical values () are marked with green vertical lines in exhibit 6.5. Because our sample mean (which is also the expected mean) of 525 is greater than the hypothesized mean of 500, we simply need to calculate the area under the curve, which represents the alternative hypothesis below the *z* value of 1.96. This probability can be obtained from a standard normal probability table or from functions available in standard spreadsheet programs (e.g., NORMDIST in Excel). For our illustration, this probability is 0.7226. Thus, the probability of a type II error for our test of hypothesis is 72.26 percent.

## [H2] Estimate of *Z* Statistic

In most practical situations, information about population parameters such as mean and standard deviation is not readily available. The unknown population mean is typically not a problem because the test of hypothesis is usually designed around that parameter. However, the unavailability of standard deviation does pose a problem, because without this parameter, the standard error of the sample mean cannot be calculated. To solve this issue, we can substitute the value of the unbiased estimate of the sample standard deviation in lieu of the population standard deviation, . The resulting distribution, known as *student’s t-distribution*, is then near normal. As *n* approaches infinity, the distribution of *t* approaches that of *z*.Mathematically, the *t*-statistic can be defined as

**[INSERT EQUATION]**

.

**[END EQUATION]**

The careful reader will notice that the only difference between the formulas for the *z* score and the *t* score is the way the standard error is estimated. For the *t* score, we use sample standard deviation, *s*, as an estimate of the unknown population standard deviation, . It should be noted that unlike the *z* distribution, the *t*-distribution has only one parameter—degrees of freedom. For a one-sample *t*-test, degrees of freedom are equal to *n* – 1, where *n* is sample size.

## [H2] Significance

*Statistical significance* is the term used when the null hypothesis is rejected in a test of hypothesis. In such a situation, we say, “The test of hypothesis is significant,” or simply, “the test is significant.” Statistical significance is often reported as a *p*-value, which is the area under the curve of the null hypothesis that is not enclosed in the critical values of the test statistic (i.e., the rejection region). Because the *p*-value is a probability, it can be directly compared with the level of significance, . The rule of thumb is to reject the null hypothesis if the *p*-value exceeds. This method of arriving at a conclusion about the null hypothesis is called the *p‑value approach*, and it is mathematically equivalent to the critical value approach.

The concept of statistical significance is useful because it provides an objective rule for rejecting (or not rejecting) the null hypothesis. However, it suffers from a serious drawback. Because the value of the test statistic inversely depends on the standard error, which is in turn inversely related to the square root of the sample size, this value can be manipulated by varying sample size. In other words, there exists some sample size for each test of hypothesis at which the null hypothesis can be rejected regardless of the magnitude of difference specified. Thus, we need guidelines that let us evaluate the results of a test of hypothesis beyond merely its statistical significance. Statisticians have developed several measures that can be used to evaluate such practical significance. Such measures are collectively called measures of *effect size*.

For the one-sample *t*-test case, a commonly used measure of effect size is Cohen’s *d*. This statistic is typically defined as

**[INSERT EQUATION]**

.

**[END EQUATION]**

When defined in this way, this measure of effect size is independent of sample size, *n*. For our one-sample *z* test illustration, the value of Cohen’s is

**[INSERT EQUATION]**

.

**[END EQUATION]**

Although the exact interpretation of the value of *d* is contextual, in most situations, *d* values of .2, .5, and .8 are used to identify small, medium, and large effect sizes, respectively (Cohen 1992).

In electronic health records, where thousands and sometimes millions of cases are examined, every difference may end up being statistically significant. In these situations, it is important to rely on both effect size and statistical significance to determine the significance of the findings.

## [H2] Confidence Interval

In a previous section, we considered two equivalent approaches when making a decision about the null hypothesis: the critical value approach and the *p*-value approach. A third approach, called the *confidence value approach*, is also available and is analytically equivalent to the other two approaches. Its advantage is that, instead of providing a point estimate (a single guess) of the unknown population parameter, it provides a range of values, called a *confidence interval*, that are attributable to the parameter being investigated in repeated replications of the test of hypothesis. A confidence interval can be calculated from the mathematical expression for the test statistic. In our sample *z* test about, the test statistic had the following expressions:

**[INSERT EQUATION]**

**[END EQUATION]**

Algebraic recalculation of the in the formulas provides us with a formula for their range of values. This is defined as

**[INSERT EQUATION]**

.

**[END EQUATION]**

If is calculated at type I error (alpha) levels of 0.05, it has a 95 percent confidence interval. The 95 percent confidence interval is a range of values of that, 95 percent of the time, contains the true population mean. Note that the confidence interval is for the population mean and not the sample mean. The probability of observing the population mean outside the confidence interval is relatively small—in this case 0.05. The 95 percent confidence interval (critical value of 1.96) for a sample of 30 patients with an average sample mean of 525 and population standard deviation of 100 is calculated as

**[INSERT EQUATION]**

.

**[END EQUATION]**

The population mean falls within the range of 489.22 and 560.78, and it does so 95 percent of the time. The population mean of 500 falls within this range. Therefore, we cannot reject the hypothesis that this sample of 30 patients came from a population with a mean of 500 and a variance of 100.

A convenient way to test whether samples are from the same population is to examine the confidence intervals of their distributions. When two confidence intervals overlap, then the two samples could have come from the same population; if not, we can reject the hypothesis.

## [H1] Comparison of Two-Sample Means

The general method of hypothesis testing described in the previous section and applied to the case of a single sample can be extended to two samples. The exact form of the test statistic depends on whether the samples are dependent or independent. In *dependent samples*, respondents (or subjects) are matched with each other on some criterion of interest. This is the case when the same group of respondents provides more than one measurement on the same variable. This is also the case when two groups of respondents share common values on a set of variables. In contrast, two samples are considered *independent* when they comprise different individuals who are not matched with each other.

A *dependent samples test*, also known as a *repeated measures test* or a *paired samples test* depending on the exact scenario, is a special case of a one-sample test performed on the difference between two vectors of measurements. Comparing the mean of a set of observations *X*1 with the mean of a second set of observations *X*2 is equivalent to performing a one-sample test on the difference *D* between the two variables where *D* = *X*1 – *X*2. All remaining steps are the same as those discussed in the *z* test and *t*-tests sections corresponding to a one-sample test about in the earlier sections of this chapter.

When means on a variable of interest between two independent samples are compared, the null hypothesis is that both means come from the same population against the alternative that the means come from different populations. The general steps involved in the test of hypothesis remain the same; only the expression for null alternative hypotheses and the test statistic change. For an independent samples test about two population means, the null and alternative hypotheses are

**[INSERT EQUATION]**

 and

**[END EQUATION]**

The test statistic takes the following expression if population standard deviations are known:

**[INSERT EQUATION]**

,

**[END EQUATION]**

where subscripts denote independent samples (or groups). If population standard deviations are unknown, the test statistic used is

**[INSERT EQUATION]**

.

**[END EQUATION]**

The degrees of freedom for an independent samples *t*-test are .

Several mathematical assumptions need to be satisfied in order to use parametric tests, such as *z* and *t*-tests, of the population mean. For one-sample tests of the mean, including the dependent samples test (which is essentially a special case of the one-sample test), the assumptions of normality and independence need to be satisfied. For the independent samples test, an additional assumption of homogeneity of variance needs to be satisfied, as well as the other two assumptions.

The *normality* *assumption* requires that the distribution of sample means be normal and is typically automatically satisfied in cases for which group sizes exceed 30. For smaller sample sizes, this assumption can be satisfied by showing evidence that the population values of the variable under investigation are normally distributed.

The *homogeneity of variance assumption* is required for the independent-samples *t*-test and requires that the variance of observations be equal to each other in the two groups being compared. Although several formal tests, such as the Levene's test and Hartley's Fmax test, are available to test this assumption, it is usually considered satisfied if the ratio of the larger-to-smaller standard deviation is no larger than 3.

The *independence assumption* requires that the respondents (or subjects) in the two groups being compared are unrelated to each other; it can be easily satisfied by either following a random scheme when selecting the sample from the population or by randomly allocating participants to either of the two groups (e.g., in the case of experiments).

## [H1] Control Chart with Normal Distribution

In the rest of this chapter, we examine how to create two types of control charts: XmR and X‑bar. Control charts plot data over time. A manager may want to construct a control chart as opposed to simply testing a hypothesis or reporting a confidence interval for two reasons:

**[INSERT NL]**

1. *To discipline intuitions*. Data on human judgment shows that we—meaning all of us, including you—have a tendency to attribute system improvements to our own efforts and skill and system failure to chance events. In essence, we tend to fool ourselves. Control charts help see through these mistakes, delineating whether the improvement is real or we have just been lucky.

In a field such as medicine, if a poor outcome occurs over time, the natural tendency is to think of it as poor care. But such judgments are misleading—from time to time there will be unusual outcomes. If administrators focus on these occasional outcomes, they risk punishing good clinicians whose efforts have failed by chance. Instead, they should focus on patterns of good or bad outcomes. Control charts help you see whether there is a pattern in the data and move you away from making judgments about quality through case-by-case review.

1. *To tell a story*. P-charts display the change over time. These charts tell how the system was performing before and after change. They are testimonials to the success of improvement efforts. Telling these stories helps the organization to

**[INSERT BL]**

* celebrate small-scale successes (an important step in keeping the momentum for continuous improvement) and
* communicate to people who are not part of the cross-functional team; such communications help pave the way for eventual organization-wide implementation of change.

**[END BL]**

**[END NL]**

You can present the data without plotting it and showing it. But without charts and plots, people will not be able to understand it well. Numbers are not enough. Plots and charts, especially those showing change over time, make people connect a story with the data; they end up feeling and understanding the data better. For many people, seeing the data is believing it. When posted publicly, control charts prepare the organization for change. They are also useful for reporting and explaining the experience of one unit of the organization to others.

In quality improvement, the purpose of data collection and analysis is not to lay blame but to assist improvement efforts. The purpose is not to prove, but to improve. Often data sets are small and conclusions refer only to the process at hand; findings are not intended to be generalized to other environments.

**[H2] Elements of Control Charts**

Control charts have a set of common elements:

**[INSERT BL]**

* *X*-axis shows periods of time
* *Y*-axis shows the observed values
* UCL line shows the upper control limit
* LCL line shows the lower control limit
* 95 percent or 99 percent of data fall within the UCL and LCL
* Values outside the control limits mark statistically significant changes and suggest a change in the underlying process

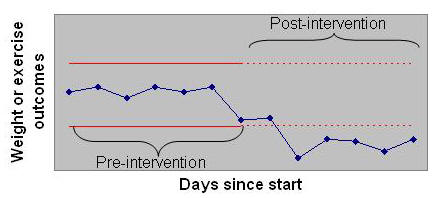
**[END BL]**

A control chart visually displays the possible impact of a change. Control charts allow managers and improvement teams to see whether the process changes they have introduced have led to the desired change. To do so, they contrast the outcomes before an intervention with outcomes after it. Typically, the control limits are drawn so that they reflect the distribution of outcomes before the intervention. If limits were calculated from data in the pre-intervention period, the limits are extrapolated to the post-intervention period; then, post-intervention observations are visually compared to the pre-intervention limits. If any points fall outside the limits, then the analyst concludes that the post-intervention outcomes come from a distribution that is different from pre-intervention outcomes.

See exhibit 6.6 for an example of the display of control limits. Two lines show the control limits, both in the same red color. The higher line is called the *upper control limit* (UCL) and the lower line is called the *lower control limit* (LCL). The solid portion of these two lines shows the periods used to set the limit; the dashed line is an extrapolation of the limit to other periods. Typically, the control limits are displayed without markers. The observations are shown as a line with a marker to highlight analysis for each period separately. If an observation falls outside the control limits, then the chance of observing this situation is small, and we can reject the hypothesis that the observation belongs to the distribution of pre-intervention data.

**[INSERT EXHIBIT; please convert to gray scale; make background transparent and everything else black, please. Make upper line dashes, and to the right, make it smaller dashes. Make lower line dots, and on right, make them smaller dots. Capitalize “Outcomes.” Replace label on uppermost line with “Post-intervention UCL” and lowermost line with “Pre-intervention LCL”]**

**Exhibit 6.6** Example Control Chart



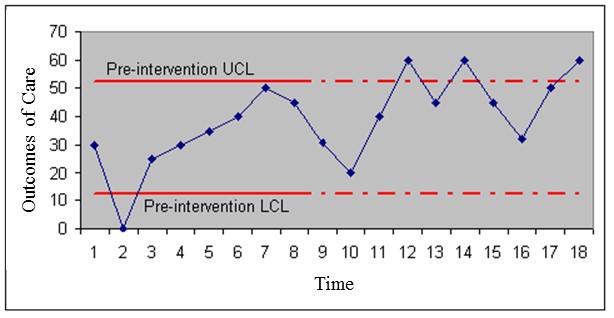
**[END EXHIBIT]**

Control limits can also be calculated from the post-intervention period and projected back to the pre-intervention period (although this is seldom done). The interpretation is still the same. We are comparing the two periods. Because the results will radically differ, it is important to judiciously select the periods from which the control limits are calculated. The selection depends on the inherent variability in the pre- or post-intervention periods. Control limits should be calculated from the period with least variability. Typically, this is done by visually looking at the variability in the pre- and post-intervention data. Because some control chart techniques ignore outliers (e.g., the Tukey chart, discussed in chapter 9), a visual review of variability could be misleading. A more reasonable approach might be to calculate the control limits two different ways and choose the limits that have a smaller spread. This will produce control limits that are tighter and more likely to detect changes in underlying processes.

Exhibit 6.7 shows the control limits derived from pre-intervention data. The control limits are shown as a solid red line. They are extended to the post-intervention period, shown as a dashed red line. When points fall outside the limit, we can conclude that, in these periods, the process had outcomes that were different from the pre-intervention pattern.

**[INSERT EXHIBIT; please convert to gray scale; make background transparent and everything else black, please. Make upper line dashes, and to the right, make it smaller dashes. Make lower line dots, and on right, make them smaller dots. Make middle (blue) line solid black. Capitalize remove “of care” on y axis.]**

**Exhibit 6.7** Control Limits Derived from the Pre-intervention Period

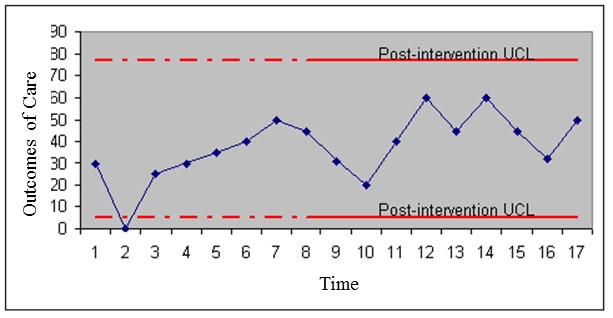


**[END EXHIBIT]**

Exhibit 6.8 shows the control limits for the same data drawn from post-intervention data. Note that this time around, the control limits are calculated from the post-intervention period, shown as a solid red line. They are extended to the pre-intervention period, shown as a dashed red line. The control chart compares the pre-intervention observations to the control limits calculated from the post-intervention data. Points that fall outside the limit indicate periods in which the data were unlikely to come from the post-intervention period.

**[INSERT EXHIBIT please convert to gray scale; make background transparent and everything else black, please. Make upper red line dashes, and to the left, make it smaller dashes. Make lower line dots, and on left, make them smaller dots. Make middle (blue) line solid black. Capitalize remove “of care” on y axis.]]**

**Exhibit 6.8** Control Limits Derived from the Post-intervention Period



**[END EXHIBIT]**

Both analyses are based on the same pre- and post-intervention data, contrasting the observation in one period to control limits derived from the other period. Note that in these data the limits are tighter when they are calculated from the post-intervention period, as shown in Exhibit 6.8. This is the correct way to analyze the data because it is based on the tighter control limits.

## [H2] XmR Control Chart

Two widely used control charts are based on standard normal distributions. These are XmR charts (*X* stands for “observation,” and *mR* stands for “moving range”) and X-bar (the bar indicates the average of the observations in the sample) charts. XmR is described in this section and X-bar in the next.

XmR charts are used widely. They were first developed by Walter A. Shewart (1931) and sometimes are referred to as *Shewart charts*. An XmR chart assumes:

**[INSERT NL]**

1. There is one observation per period.
2. Patients’ case mix or risk factors do not change over the period. If the chart monitors the same patient over time, there is little need to measure the severity of the patient, as this is unlikely to change in short periods. If data come from different patients at different times, it is important to verify that these patients have a similar prognosis or severity of illness on admission.
3. Observations are measured in an interval scale (i.e., the observation values can be meaningfully added or divided).
4. Observations are independent of each other, meaning that knowledge of one observation gives insight into what the next value will be. Outcomes of infectious diseases are usually not considered independent, as knowing that a patient has an infectious disease at period *t* increases the probability of infection for period *t* + 1.

**[END NL]**

All of these assumptions must be met before the use of XmR charts makes sense. An analyst using an XmR chart does not explicitly assume that the observed values have a normal distribution. The XmR chart relies on differences in consecutive values, which can be assumed to be normally distributed or to be more normal than the observations themselves. We know that the average (and by extension the difference) of any two observations is more likely to have a normal distribution than the original observations. Therefore, the assumption of normal distribution may be approximately met in XmR charts. Take a look again at the second row of exhibit 6.3to see how the assumptions of normal distribution may be reasonable, even when the original data are not normal. The second row shows the distribution of the average of two points, which, except for division by a constant, is the same as distribution of a difference.

If there is an intervention, we need to decide whether the control limits should be calculated from the pre-intervention or post-intervention period. Select the period with the least variability, which will produce the tightest control limit. The variability in pre- and post-intervention periods can be examined visually or by calculating the difference between the maximum and minimum value in each period. Calculate the control limit from the pre-intervention period if it has the smaller difference. Otherwise, calculate the control limits from the post-intervention period. Control limits are calculated from one period and extended to the other so that we can judge whether the pre- and post-intervention periods differ.

Control limits in XmR charts are calculated from the moving range (mR). A range is based on the absolute value of consecutive differences in observations.

**[INSERT BL]**

* Estimate the average of the moving range.
* Count the number of periods, *n*.
* Calculate the absolute value of the difference of every consecutive value, the moving range.
* Add the moving ranges and divide by *n* − 1 to get the average moving range.

**[END BL]**

The UCL is the average of the observations, plus a constant E, times the average moving range. The constant E depends on how many consecutive observations are included in the moving range. When two consecutive observations are examined, the constant is 2.66 for control limits that include 99 percent of the data. If the moving range is calculated from two consecutive periods, the correction factor E for control limits that include 95 percent of the data is 1.88. Then the UCL can be calculated as

**[INSERT EQUATION]**

UCL = Average of observations + 1.88 × Average of moving range.

**[END EQUATION]**

Similarly, the LCL is calculated as

**[INSERT EQUATION]**

LCL = Average of observations – 1.88 × Average of moving range.

**[END EQUATION]**

Once the control limits have been calculated, we can construct the control chart. First plot the *x*-axis (time) and *y*-axis (observations). Plot the observed values for each period. Plot the control limits over the entire period (show the calculated control limit as a solid line and the extended portion as a dashed line). If your report can use color, show UCLs and LCLs in red, without markers.

Points within control limits are controlled variations. These points do not show real change, though data seem to rise and fall. These are merely random variations. Points outside the limits show real change. If a point falls outside the limits, we need to investigate what change in the process might have led to it. In other words, we need to search for a special cause. Once a control chart has been constructed, it is a useful way of telling an improvement story. Distribute the chart by electronic media, add it to the company newsletter, or display it as an element of a storyboard. When you show your staff a chart, demonstrate that you have verified assumptions, check that your chart is accurately labeled, and include your interpretation of the finding.

## [H2] X-bar Control Chart

This chapter introduced the normal distribution. The normal distribution assumes that observed values are collected in one sample. However, it is often difficult to collect all observed values in the same period. Furthermore, the analyst often wishes to look for improvement over time, which requires at least two samples—one before the intervention and one after. When samples are drawn from different periods, an X-bar chart can be used to display and analyze the data.

X-bar charts are often used to examine satisfaction ratings over time for a health service. Managers are often interested in how satisfaction is changing over time. We introduce the concepts behind an X-bar chart with an example from analyzing satisfaction ratings. The Agency for Healthcare Research and Quality developed the Consumer Assessment of Healthcare Providers and Systems (CAHPS) surveys to measure patient satisfaction. These surveys ask patients to rate their experience with healthcare encounters, including hospital and outpatient encounters. Separate surveys have been developed to report experiences with health plans, home health agencies, nursing homes, and a variety of other healthcare settings. A CAHPS survey typically has 40 questions. These are two examples:

**[INSERT BL]**

* In the last 12 months, how often did your personal doctor explain things in a way that was easy to understand?

**[INSERT SUBLIST]**

* + Never
  + Sometimes
  + Usually
  + Always

**[END SUBLIST]**

* In the last 12 months, how often did your personal doctor listen carefully to you?

**[INSERT BL]**

* + Never
  + Sometimes
  + Usually
  + Always

**[END SUBLIST]**

**[END BL]**

**[INSERT BOX]**

|  |
| --- |
| **Nominal, Ordinal, and Interval Scales**  In a nominal scale, numbers are assigned to objects in order to identify them, but the numbers themselves have no meaning. For example, the diagnosis related group code 240 (myocardial infarction) is a nominal scale. An ordinal scale is a scale in which the numbers preserve rank. In an ordinal scale, a score of 8 is more than 4 but not necessarily twice more than 4. Risk ratings are usually ordinal. An interval scale requires not only that numbers preserve the order but also that they be preserved in correct magnitudes. Thus, a score of 8 is twice 4. The difference between two interval scores is meaningful, while the difference between two ordinal scores is not. Number of patients is an example of an interval scale.  **[END BOX]** |

The response to each question is categorical—there are different categories of answers (such as “Never” in our example). These responses preserve the order of satisfaction. When a person responds “Never,” all we know is that the event happens less frequently than “Sometimes.” We do not know how many times less frequent. The overall satisfaction rating is calculated by scoring each of the individual questions and adding the scores. It is generally assumed that the overall satisfaction rating has an interval scale and a normal distribution. This assumption may be wrong, especially given that satisfaction ratings are made on an ordinal scale and it does not make sense to add ordinal scales. The Centers for Medicare & Medicaid Services publishes satisfaction scores for hospital care on its website. A quick review of these data shows that these ratings have a normal distribution. The data show a bell-shaped symmetric curve. Surprisingly, adding ordinal scores has produced a normal interval scale. Despite wrong assumptions, we have arrived at seemingly right conclusions.

In analyzing satisfaction ratings, X-bar charts are used to track changes over time. In analysis of the health status of a group of patients, X-bar charts are used to trace the group's health status. Exhibit 6.9 provides an example of data we might analyze in this manner. Over several periods, we sampled four patients and asked them to rate a clinic’s services.

**[INSERT EXHIBIT]**

**Exhibit 6.9** Satisfaction with Clinic Services

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Ratings* | | | |
| *Period* | *Patient 1* | *Patient 2* | *Patient 3* | *Patient 4* |
| 1 | 80 | 84 | 82 | 80 |
| 2 | 70 | 72 | 74 | 70 |
| 3 | 76 | 78 | 76 | 78 |
| 4 | 80 | 78 | 78 | 80 |

**[END EXHIBIT]**

Look at the data—there are wide variations. Could they be the result of chance? The first step is to calculate the average for each period. Do so by summing the ratings in each period and dividing by the number of cases in that period. Thus, for period 1, the average will be 80 + 84 + 82 + 80 divided by 4, which is 81.5. Calculate the averages for all periods. Then, plot the data. Create an *x*-*y* plot where the *x*-axis is time and the *y*-axis is the average satisfaction rating. Exhibit 6.10 shows the average satisfaction over the four periods.

**[INSERT EXHIBIT; please convert to gray scale; make line black]**

**Exhibit 6.10** Average Satisfaction Rating Over Time

**[END EXHIBIT]**

The plot tells us more about the data. If the analyst adds the UCL and LCL (between which one expects 95 percent of the data to fall) to the plot, she can see whether the changes she has observed are random or real. The UCL and LCLs in an X-bar chart are based on the assumption that the data are normally distributed. So before we calculate these limits, we need to check and see whether the assumptions are met. The assumptions of an X-bar chart are the following:

**[INSERT BL]**

* *Continuous interval scale.* The variable being averaged must be a continuous variable on an interval scale, where the differences between the scores are meaningful. An ordinal scale cannot be averaged. Yet, as we discussed earlier, one can arrive at correct conclusions despite wrong assumptions. Satisfaction rating and health status ratings behave like interval scales, though they are measured on ordinal responses to various questions.
* *Independent events*. The observations over each period are not affected by the previous observations. In our example, the satisfaction ratings in period 2 should not be affected by ratings in period 1. This assumption would be violated in an example in which the same patient is rating the unit in every period. It is likely that this patient's first impression affects subsequent evaluations. The assumption seems reasonable when different patients are rating in different periods.
* *Normal distribution*. If we were to stack all the ratings, most will fall on the average rating, some on each side. A normal distribution suggests that the stack will peak on the average, slowly decline on both sides of the average, and the shape of the curve will be symmetrical. The **law of large numbers** says that no matter what the distribution of a variable is, the average of the variable will tend to have a normal distribution. As the number of cases for calculation of the average increases, the average is more likely to be normal. A minimum of four cases is needed for applying the law of large numbers.
* *Constant variance*. Deviations from the average should not consistently increase or decrease over time. This assumption can be verified on a control chart.

**[END BL]**

When the assumptions of normal distribution are met, we can proceed to the next step of calculating UCLs and LCLs. To calculate the limits, we follow these steps:

**[INSERT NL]**

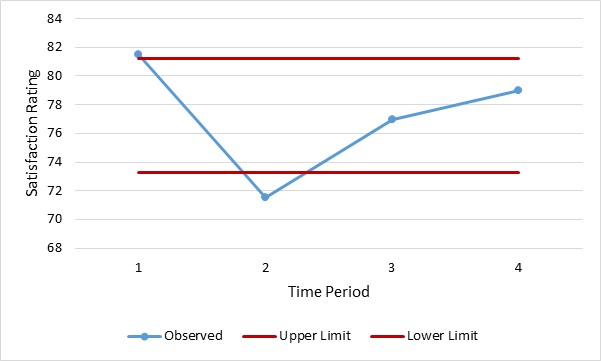
1. Calculate the grand average—the average of all ratings across all periods. Do not calculate the grand average by averaging the mean of each period. When these means are based on different numbers of ratings, the grand average will be incorrect. The correct way to calculate the grand average for all periods is to sum all the ratings for all the periods and divide this sum by the number of ratings across all periods.
2. Calculate the standard deviation for the observed ratings. Using the function StDev in Excel, the standard deviation of all observations is 4.06.
3. Estimate the standard deviation for observations in each period by the square root of the number of cases in the period. Thus, for period 1, this will be 4.06 ÷ square root of  
   4 = 2.03.
4. Calculate the UCL for each period as the grand average, plus 1.96,multiplied by the standard deviations of the time. In a normal distribution, the mean plus and minus 1.96, multiplied by the standard deviation of the distribution, contains 95 percent of the data. For our first period, this will be 77.25 + 1.96 × 2.03 = 81.23. The use of 1.96 makes sense only if the assumption of normal distribution of the data can be verified. Otherwise, when there are too few observations averaged, we suggest you use the *t*-student value corresponding to the sample size.
5. Calculate the lower limit for each period as the grand average minus 1.96 times the standard deviation of that period.

**[END NL]**

Exhibit 6.11 shows how the control chart will look. UCLs and LCLs are drawn with no markers, in red, to designate these lines as a cutoff point, beyond which the interpretation of the observations changes. Note the observed line is drawn with a marker so that attention is focused on the observed experience within individual periods.

**[INSERT EXHIBIT; convert to gray scale; make blue line black; make upper red line dashed; make lower red line dotted. Delete legend at bottom. Label uppermost line “UCL” and lowermost line “LCL”; label middle line “Observed”]**

**Exhibit 6.11** Satisfaction with Clinic Services



**[END EXHIBIT]**

The control limits are straight lines because, in every period, we sampled the same numbers of cases. If this were not the case, the control limits would not be a straight line. It would be tighter when the sample size was larger and looser when the sample size was smaller. Note that the first period’s average observation was higher than the UCL and the second period’s average was lower than the LCL. In the first period, patients were unusually satisfied. Patients rated the clinic services the lowest in the second period. The change in the ratings for these two periods was not the result of chance events. It marked a real shift in satisfaction with clinic services.

## [H2] Risk-Adjusted X-bar Control Chart

Risk adjustments are needed so that we can differentially attribute outcomes to the patient’s prognosis, as opposed to clinical or managerial interventions in processes of care. Alemi and Sullivan (2001) have proposed how to construct risk-adjusted X-bar charts, and here we replicate their recommendations.

For constructing a risk-adjusted X-bar chart, two data are needed: (1) a continuous observed outcome, collected over time, across a sample of patients; and (2) an expected outcome for each patient. The purpose of the risk-adjusted control chart is to determine whether outcomes have changed beyond what can be expected from the patients’ condition. If they have, the clinician has provided better- or worse-than-expected care. If not, changes in patients’ conditions explain the outcomes. The data needed are available in many circumstances. Expected outcomes can be based on a clinician’s review of patients or can be deduced from many commercial and noncommercial severity indices.

To help the reader understand risk-adjusted control charts, this chapter will present data from a recent analysis we conducted on diabetic patients of an outpatient clinic. Type 2 diabetes mellitus affects millions of Americans each year and, if not controlled, can result in considerable morbidity. The question of interest to the clinicians was whether they had improved over time in helping their patients control their disease. We thought if we look at the average experience of the patients of several providers, we would be able to speak to the skills of the provider in helping her patients control their diabetes. For our outcome variable, we decided to focus on hemoglobin A1C levels (HgbA1C) measured in patients with type 2 diabetes. Studies have shown that the microvascular complications of retinopathy, nephropathy, and neuropathy can be prevented with good control of the blood sugar levels. Measuring HgbA1C gives information on how well-controlled levels have been over the preceding eight weeks. We reviewed the data on 60 type 2 patients in a family practice clinic of five providers for 21 consecutive months and present the data for two of those providers. We will use this data set to demonstrate how to create a risk-adjusted X-bar control chart. Six steps are involved in constructing a risk-adjusted X‑bar chart.

### [H3] Step 1: Assumptions

The analyst must verify five assumptions before proceeding with the construction of an X-bar control chart. These include the following:

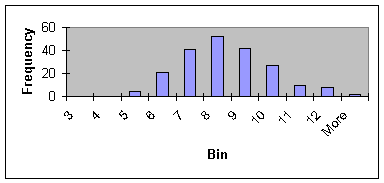
**[INSERT NL]**

1. Observations are on a continuous interval scale. In our example, HgbA1C measures preserve both order and difference among patients, and therefore occurs on an interval scale.
2. The second assumption is that each observation is independent. The measurement of HgbA1C in one patient does not influence the measurement of another patient or of the same patient in another period, which meets the assumption. Examining correlation among HgbA1C values for the same patient at different times can also test assumptions of independence—large positive correlations suggest a lack of independent observations.
3. Each period has more than four observations. When there are too few observations, the assumption of normal distribution may be wrong. In our example, there are more than five observations in each period, so this assumption is met.
4. The fourth assumption is that observations have a normal, bell-shaped curve. The chi‑square statistic can be used to test whether HgbA1Cs observed have a normal distribution. Instead of conducting a chi-square test, we prefer to examine the data visually by constructing a histogram (exhibit 6.12). The distribution is symmetric and peaks in the middle—it has the bell-shaped curve of a normal distribution. Therefore, the fourth assumption that observations have a normal distribution is not rejected.
5. The fifth assumption is related to the equality of variances of observations across periods. Analysis of variance (ANOVA) can be used to test the equality of variance of observations in different periods. We prefer to test the assumption quickly through calculating average ranges. When the ranges of observations in different periods are not two or three multiples of each other, then we accept the assumption of the equality of variance. In this case, average ranges differ from a low of 5.8 to a high of 11.6; all seem to be in the same ballpark. Therefore, we accept the assumption of equality of variances over time.

**[END NL]**

**[INSERT EXHIBIT; please render in gray scale; make the image larger, eliminate the gray background, and make the bars black]**

**Exhibit 6.12 Distribution of HgbA1C Data**



**[END EXHIBIT]**

### [H3] Step 2: Determine the Average in Each Period

Exhibit 6.13 shows the average of observations for each period. Each row represents an individual patient and each column is a separate period. The observations for each column are summed and then divided by the number of observations for that period:

**[INSERT EQUATION]**

*A*i = *∑*j = 1…n*i* *Aij* ÷ *ni* .

**[END EQUATION]**

In the formula, *Ai* is the average of observations for period *i*, *Aij* is the observation *j* in period *i*, and *ni* is the number of observations for period *i*.

**[INSERT EXHIBIT]**

**Exhibit 6.13 Data for Patients of Provider 1 over Seven Periods**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Patient* | *HgbA1C Levels* | | | | | | |
| *3 Months* | *6 Months* | *9 Months* | *12 Months* | *15 Months* | *18 Months* | *21 Months* |
| 1 | 8.8 | 9.9 | 10.2 | 10.5 |  | 11.0 |  |
| 2 | 6.7 | 9.0 | 9.0 | 9.2 | 8.8 | 9.2 | 8.5 |
| 3 | 8.2 | 7.6 | 10.8 |  | 9.2 |  | 8.9 |
| 4 |  | 6.1 |  | 7.7 | 7.1 |  | 9.0 |
| 5 | 9.1 | 9.6 |  | 9.6 |  | 8.8 |  |
| 6 |  | 11.1 | 11.5 |  |  |  |  |
| 7 | 9.1 |  |  |  | 8.4 | 7.8 | 8.5 |
| 8 |  |  |  | 9.7 | 9.0 | 11.7 |  |
| 9 | 5.9 | 7.6 | 7.8 | 6.1 | 7.0 | 6.3 |  |
| 10 | 7.4 | 7.0 |  | 5.7 |  | 7.0 |  |
| 11 |  | 7.2 | 7.4 | 7.0 |  | 7.3 | 8.1 |
| 12 | 8.9 | 8.1 | 9.1 | 9.7 | 7.9 | 8.1 |  |
| 13 | 5.9 | 7.1 | 7.8 |  |  | 7.1 | 7.8 |
| 14 |  |  |  | 9.9 | 13.5 |  |  |
| 15 | 5.1 |  | 5.7 | 5.5 | 5.0 |  |  |
| 16 | 7.4 |  |  |  |  | 7.6 |  |
| 17 | 7.1 | 6.2 |  | 7.0 |  | 7.2 | 7.1 |
| 18 | 8.3 |  | 9.2 | 7.9 |  |  | 9.0 |
| 19 | 7.0 | 8.6 | 8.9 | 8.6 | 8.2 | 10.1 | 10.0 |
| 20 | 6.4 |  | 6.9 | 6.2 | 7.0 | 8.4 |  |
| 21 |  | 7.0 |  | 7.4 |  |  |  |
| Average | 7.4 | 8.0 | 8.7 | 8.0 | 8.3 | 8.4 | 8.5 |

**[END EXHIBIT]**

### [H3] Step 3: Measure Risk or Expected Values

Extensive literature exists regarding factors that increase risk of complications from diabetes. Many of these are factors that clinicians can encourage patients to change (e.g., smoking). To measure risk, we decided to focus on variables that providers have little control over and that could make diabetes management more difficult. We looked at age of onset, as data show that patients will have less control over their disease over time. We looked at number of medications, as patients’ ability to control their diabetes may be hampered by their need to take medication. For the 60 patients of the five providers in our sample data, we regressed HgbA1C levels (averaged across periods) on two independent variables: number of medications and age of onset of diabetes. We used the regression equation to predict expected HgbA1C levels for each patient at each period. Our equation is

**[INSERT EQUATION]**

HgbA1C level = 8.58 + 0.76 × (number meds) − 0.03 × (age of onset).

**[END EQUATION]**

For each patient at each period, we calculated a predicted HgbA1C. For example, for patient 1 at three months (the first period), the number of medicines was three and the age was 65. Using this regression equation, we calculated the expected level of 8.2 for his A1c level. The calculated values are our expectation regarding the patients’ ability to control their diabetes. Exhibit 6.14 shows the observed and expected values for one provider in two periods.

**[INSERT EXHIBIT]**

**Exhibit 6.14 Expected and Observed Values for Two Periods**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Patient* | *3 Months* | | *6 Months* | |
| *Observed* | *Expected* | *Observed* | *Expected* |
| 1 | 8.8 | 8.2 | 9.9 | 8.2 |
| 2 | 6.7 | 9.1 | 9 | 9.1 |
| 3 | 8.2 | 9.3 | 7.6 | 9.3 |
| 4 |  |  | 6.1 | 9.1 |
| 5 | 9.1 | 8.4 | 9.6 | 8.4 |
| 6 |  |  | 11.1 | 8.8 |
| 7 | 9.1 | 8.7 |  |  |
| 8 |  |  |  |  |
| 9 | 5.9 | 8.3 | 7.6 | 8.3 |
| 10 | 7.4 | 7.2 | 7 | 7.2 |
| 11 |  |  | 7.2 | 7.9 |
| 12 | 8.9 | 8 | 8.1 | 8 |
| 13 | 5.9 | 7.8 | 7.1 | 7.8 |
| 14 |  |  |  |  |
| 15 | 5.1 | 6.8 |  |  |
| 16 | 7.4 | 7.4 |  |  |
| 17 | 7.1 | 7.3 | 6.2 | 7.3 |
| 18 | 8.3 | 8.6 |  |  |
| 19 | 7 | 9.5 | 8.6 | 9.5 |
| 20 | 6.4 | 7.8 |  |  |
| 21 |  |  | 7 | 6.8 |

**[END EXHIBIT]**

### [H3] Step 4: Expected Average for Each Period

### To calculate the average of the expected values for period *Ei*, add all expected values and divide by the number of observations. If *Ei,j* is the expected value for patient *j* in period *i*, then the average of these values, *Ei*, can be calculated as

**[INSERT EQUATION]**

.

**[END EQUATION]**

For the first three months in exhibit 6.14, the average of expected values is 8.2; and for the second three-month period, the average of the expected values is 8.4. The following expected averages for subsequent periods for this one provider were 8.3, 8.3, 8.6, 8.2, and 8.6.

### [H3] Step 5: Standard Deviation of the Difference

Suppose shows the difference between observed and expected values for patient *j* in period *i*; that is,

**[INSERT EQUATION]**

.

**[END EQUATION]**

Furthermore, suppose is the average of the differences for the period *i*:

**[INSERT EQUATION]**

*Di* = *∑j* = 1…*ni**Di,j* ÷ *ni* .

**[END EQUATION]**

The standard deviation of the differences, S*i*, is calculated as

**[INSERT EQUATION]**

*Si* = [*∑j* = 1,…,*ni* (*Di*,*j* − *Di*)2 ÷ (*ni*-1)]0.5.

**[END EQUATION]**

Note that the standard deviation of each period depends on the number of observation in the period. As the number of observations increases, standard deviation decreases and, as we will see shortly, control limits are set tighter. The chance of observing points out of the control limits increases. The calculation of standard deviation using the first two periods for the 21 patients of one provider are shown in exhibit 6.15.

**[INSERT EXHIBIT]**

**Exhibit 6.15 Calculation of Standard Deviation of Differences** in Two Periods

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Patient* | *3 months* | | *6 months* | |
| *Difference = Observed − Expected* | *(Difference − Average of Difference)2* | *Difference = Observed − Expected* | *(Difference − Average of Difference)2* |
| 1 | 0.6 | 1.9 | 1.7 | 3.92 |
| 2 | −2.4 | 2.8 | −0.1 | 0.02 |
| 3 | −1.1 | 0.12 | −1.7 | 2.07 |
| 4 |  |  | −3 | 7.29 |
| 5 | 0.7 | 2.04 | 1.2 | 2.04 |
| 6 |  |  | 2.4 | 6.76 |
| 7 | 0.4 | 1.27 |  |  |
| 8 |  |  |  |  |
| 9 | −2.4 | 2.83 | −0.7 | 0.23 |
| 10 | 0.2 | 0.96 | −0.2 | 0.01 |
| 11 |  |  | −0.7 | 0.18 |
| 12 | 0.9 | 2.75 | 0.1 | 0.13 |
| 13 | −1.9 | 1.42 | −0.7 | 0.24 |
| 14 |  |  |  |  |
| 15 | −1.7 | 0.98 |  |  |
| 16 | 0.1 | 0.64 |  |  |
| 17 | −0.2 | 0.3 | −1.1 | 0.72 |
| 18 | −0.3 | 0.2 |  |  |
| 19 | −2.5 | 3.21 | −0.9 | 0.48 |
| 20 | −1.4 | 0.48 |  |  |
| 21 |  |  | 0.2 | 0.24 |
| Average of difference | −0.7 |  | −0.3 |  |
| Sum |  | 21.89 |  | 24.33 |
| Standard deviation of difference |  | 1.25 |  | 1.37 |

**[END EXHIBIT]**

### [H3] Step 6: Control Limits

Control limits are typically set two or three standard deviations away from the expected values. When the control limits are two standard deviations away from the expected values, 95 percent of the data fall within the limits. There is a 5 percent chance of erroneously concluding that the system is out of control. When the limits are set three standard deviations away from the expected values, 99.7 percent of the data are expected to fall within the limits, and the chance of making an erroneous conclusion drops to 0.3 percent. Tighter limits are chosen when the cost of making an erroneous conclusion is high. Wider limits are chosen when it is important to detect changes in the process, even if the analyst makes an erroneous conclusion 5 percent of the time. The UCL for period *i*, shown as UCL*i*, is calculated as

**[INSERT EQUATION]**

UCL*i* = *Ei* + *t* × *Si*.

**[END EQUATION]**

In this equation, *t* is a constant that depends on the number of cases used in the period and the confidence interval adopted. Exhibit 6.16 gives the *t*-values for various sample sizes and confidence intervals.

**[INSERT EXHIBIT]**

**Exhibit 6.16 T-values for Various Sample Sizes and Confidence Intervals**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Sample Size* | | | | |
| *5* | *10* | *15* | *20* | *25* |
| 95% confidence | 2.57 | 2.22 | 2.13 | 2.09 | 2.06 |
| 99% confidence | 4.03 | 3.17 | 2.95 | 2.84 | 2.79 |
| *Note*: For sample sizes not provided, estimate proportional to the nearest values or look on the internet for Student’s *t* calculators. | | | | | |

**[END EXHIBIT]**

Thus, for the 10 cases in period 1, the UCL1 is calculated with the equation

**[INSERT EQUATION]**

UCL1 = *E*1 + *t × S*1.

**[END EQUATION]**

So

**[INSERT EQUATION]**

UCL1 = 8.6 + 2.22 × 1.26.

**[END EQUATION]**

The LCL for period *i*, shown as LCL*i*, is calculated as

**[INSERT EQUATION]**

LCL*i*= *Ei* − *t × Si*.

**[END EQUATION]**

Thus, for the ten cases in the first period, the LCL1 is calculated with the equation

**[INSERT EQUATION]**

LCL1 = *E*1 − *t × S*1, so

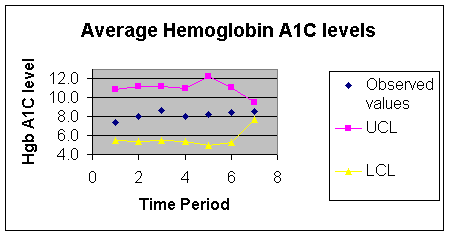
LCL1 = 8.6 − 2.22 × 1.26 = 5.81.

**[END EQUATION]**

Sometimes LCLs are negative numbers. If it is not possible to observe a negative number, as often is the case, the LCL is set to zero. In this case, the LCL is a positive number; therefore, we do not need to change it. When control limits have been calculated for all periods, these limits are plotted. Exhibit 6.17 shows the observed values plotted against seven periods for the cases of one provider. Lower and upper control lines are superimposed on these figures to enable quick interpretation of the findings. Similar figures can be constructed for other providers, thus making it possible to compare providers’ performance despite differences in their case mix.

**[INSERT EXHIBIT. Please convert to gray scale. Eliminate gray background. Make image larger. Make the squares in the pink lines into normal circles. Make the pink line dashes and label it “UCL”. Label the blue line “Observed values” And make it solid. Make the yellow triangles normal circles, and connect them with a dotted line, and label the line “LCL.” Delete legend.]**

**Exhibit 6.17** Control Chart for HgbA1C Levels Compared to Expected Levels

 **[END EXHIBIT]**

Points that fall within control limits indicate variations that can be expected by chance alone. Points outside the two limits indicate observations that are not within our expectations. For example, in the diabetes data, a point above the control limit indicates a period in which patients’ HgbA1C is worse than expected. Any point below the LCL indicates a period in which HgbA1C is better. In our data, no points were outside control limits. Over time, both clinicians had maintained the HgbA1C of their patients at the same levels. Interventions to encourage patients to lower their HgbA1C had not paid off beyond what could have been expected from the patients’ conditions.

### [H3] Distribute Findings

In the final step, the chart and the findings are distributed to the improvement team. Analysts should adhere to the following principles when distributing the findings:

**[INSERT NL]**

1. Always list assumptions and visually display the test of the assumptions. The normal distribution of the averages should be displayed.
2. Explain how expected values, risk scores, were calculated. Provide evidence that the expectations are reasonable. State clearly that if expectations change, the conclusions will change too.
3. Provide the control chart. Make sure that control limits are lines without markers and observed averages are lines with markers. Both control limits may be the same color, as the UCL and LCL are understood in the context. A legend may not be needed if labels are understood from the context.
4. Write a brief interpretation of the findings, including whether the process has changed. Annotate the control chart if special causes variation occurs (e.g., “new improvement started,” “Dr. A joined the practice,” “New medication was started”).
5. Distribute the analysis to both the improvement team and, with their approval, to other process owners who might be interested in the team’s deliberation.   
   **[END NL]**

Note that reports have a life of their own. Years later, people inside and outside of the organization may still refer to them. Keep in mind how the report may be of use after current improvement efforts have ended.

## [H1] Summary

This chapter introduced the distribution of means or averages. It showed how these distributions can be used in testing of hypotheses. The last part of the chapter focused on X-bar and XmR control charts. These control charts examine averages calculated over several periods. We showed how the control charts can be constructed. The chapter ended with an example of a risk-adjusted control chart.

## [H1] Supplemental Resources

Problem set, solutions to problems, multimedia presentations, SQL code, and other related materials are on the course website.

**[H1] References**

Alemi F., and T. Sullivan. 2001. “Tutorial on Risk Adjusted X-bar Charts: Applications to Measurement of Diabetes Control.” *Quality Management in Healthcare* 9 (3): 57–65.

Cohen, J. 1992. “[A Power Primer.](https://www.ncbi.nlm.nih.gov/pubmed/19565683)” *Psychological Bulletin* 112 (1): 155–59.

Shewart, W. A. 1931. Economic Control of Quality of Manufactured Product. New York: D. Van Nostrand.