Source: Juran's Quality Handbook: The Complete Guide to Performance Excellence, 7th Edition
ISBN: 9781259643613
Authors: Joseph A. De Feo

20.7. Statistical Tools for Improvement

This concept of a sampling distribution is fundamental to the two major areas of statistical inference, estimation and tests of hypotheses, which are discussed next.

20.7.1. Statistical Estimation: Point Estimation and Confidence Intervals

Estimation is the process of analyzing a sample result to predict the corresponding value of the population parameter. In other words, the process is to estimate a desired population parameter by an appropriate measure calculated from the sample values. For example, the sample of four batteries previously mentioned had a mean life of 31.0 hours. If this is a representative sample from the process, what estimate can be made of the true average life of the entire population of batteries? The estimation statement has two parts:

1. The point estimate is a single value used to estimate the population parameter. For example, 31.0 hours is the point estimate of the average life of the population.
2. The confidence interval is a range of values that include (with a preassigned probability called a confidence level[1]) the true value of a population parameter. Confidence limits are the upper and lower boundaries of the confidence interval. Confidence limits should not be confused with other limits (e.g., control limits, statistical tolerance limits).

Table 20.3 summarizes confidence limit formulas for common parameters. The following example illustrates one of these formulas:

Table 20.3 Summary of Confidence Limit Formulas (1 – α) (Confidence Level)

Source: Quality Planning and Analysis, Copyright 2007. Used by permission.	
Mean of a normal population (standard deviation known)	$\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ where $\overline{X}$ = sample average Z = normal distribution coefficient σ = standard deviation of population n = sample size
Mean of a normal population (standard deviation unknown)	$\overline{X} \pm t_{\alpha/2}\frac{S}{\sqrt{n}}$ where t = distribution coefficient (with n – 1 degrees of freedom) s = estimated σ (s is the sample standard deviation)
Standard deviation of a normal population	Upper confidence limit $= s\sqrt{\frac{n-1}{x^2_{\alpha/2}}}$ Lower confidence limit $= s\sqrt{\frac{n-1}{x^2_{1-\alpha/2}}}$ where x^2 = chi-square distribution coefficient with n – 1 degrees of freedom 1 – α = confidence level
Population fraction defective	See charts: *Ninety-five percent confidence belts for population proportion* and *Binomial Distribution* at the end of this chapter, pages 675–676.
Difference between the means of two normal populations (standard deviations σ_1 and σ_2 known)	$(\overline{X}_1 - \overline{X}_2) \pm Z_{\alpha/2}\sqrt{\frac{\alpha_1^2}{n_1} + \frac{\alpha_2^2}{n_2}}$
Difference between the means of two normal populations ($\sigma_1 = \sigma_2$ but unknown)	$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \times \sqrt{\frac{\Sigma(X-\overline{X}_1)^2 + \Sigma(X-\overline{X}_2)^2}{n_1+n_2-2}}$
Mean time between failures based on an exponential population of time between failures	Upper confidence limit $= \frac{2rm}{X^2_{\alpha/2}}$ Lower confidence limit $= \frac{2rm}{X^2_{1-\alpha/2}}$ where r = number of occurrences in the sample (i.e., number of failures) m = sample mean time between failures DF = 2r

Problem Twenty-five specimens of brass have a mean hardness of 54.62 and an estimated standard deviation of 5.34. Determine the 95 percent confidence limits on the mean. The standard deviation of the population is unknown.

Solution Note that when the standard deviation is unknown and is estimated from the sample, the$t$ distribution in Table 20.4 must be used. The$t$ value for 95 percent confidence is found by entering the table at 0.975 and 25 – 1, or 24, degrees of freedom[2] and reading a t value of 2.064.

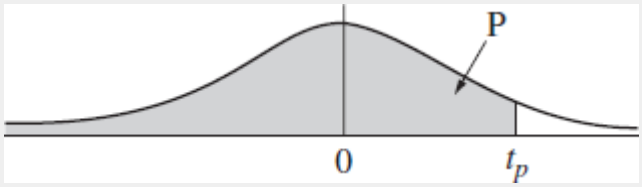
$$\text{Confidence limits} = \overline{X} = \pm t\frac{s}{\sqrt{n}}$$
$$= 54.62 \pm (2.064)\frac{5.34}{\sqrt{25}}$$
$$= 52.42 \text{ and } 56.82$$

Table 20.4 Distribution of *t*

Source: Introduction to Statistical Analysis, Copyright 1969. Used by permission.

Distribution of *t*

Value of *t* corresponding to certain selected probabilities (i.e., tail areas under the curve). To illustrate: the probability is .975 that a sample with 20 degrees of freedom would have *t* = +2.086 or smaller.



DF	$t_{.60}$	$t_{.70}$	$t_{.80}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977

15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660
120	0.254	0.526	0.845	1.289	1.658	1.980	2.358	2.617
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576

There is 95 percent confidence that the true mean hardness of the brass is between 52.42 and 56.82.

20.7.2. Determination of Sample Size

The only way to obtain the true value of a population parameter such as the mean is to measure (with a perfect measurement system) each and every individual within the population. This is not realistic (and is unnecessary when statistics are properly applied), so samples are taken instead. But how large a sample should be taken? The answer depends on (1) the sampling risks desired (alpha and beta risk, discussed further below and defined in Table 20.5), (2) the size of the smallest true difference that is desired to be detected, and (3) the variation in the characteristic being measured.

Table 20.5 Hypothesis Testing Definitions

Source: Quality Planning and Analysis, Copyright 2007. Used by permission.
Null hypothesis (H_0): Statement of no change or no difference. This statement is assumed true until sufficient evidence is presented to reject it.
Alternative hypothesis (H_a): Statement of change or difference. This statement is considered true if H_0 is rejected.
Type I error: The error in rejecting H_0 when it is true or in saying there is a difference when there is no difference.
Alpha risk: The maximum risk or maximum probability of making a type I error. This probability is preset, based on how much risk the researcher is willing to take in committing a type I error (rejecting H_0 wrongly), and it is usually established at 5% (or .05). If the *p*-value is less than alpha, reject H_0.
Significance level: The risk of committing a type I error.
Type II error: The error in failing to reject H_0 when it is false or in saying there is no difference when there really is a difference.
Beta risk: The risk or probability of making a type II error or overlooking an effective treatment or solution to the problem.
Significant difference: The term used to describe the results of a statistical hypothesis test where a difference is too large to be reasonably attributed to chance.
p-value: The probability of obtaining different samples when there is really no difference in the population(s)—that is, the actual probability of committing a type I error. The *p*-value is the actual probability of incorrectly rejecting the null hypothesis (H_0) (i.e., the chance of rejecting the null when it is true). When the *p*-value is less than alpha, reject H_0. If the *p*-value is greater than alpha, fail to reject H_0.
Power: The ability of a statistical test to detect a real difference when there really is one, or the probability of being correct in rejecting H_0. Commonly used to determine if sample sizes are sufficient to detect a difference in treatments if one exists. Power = $(1 - \beta)$, or 1 minus the probability of making a type II error.

For example, suppose it was important to detect that the mean life of the battery cited previously was 35.0 hours (recall that the intended value is 30.0 hours). Specifically, we want to be 80 percent certain of detecting this difference (this is the "power" of the test, and has a corresponding risk of $\beta = 0.2$; this means we are willing to take a 20 percent chance of failing to detect the five-hour difference when, in fact, it exists). Further, if the true mean was 30.0 hours, we want to have only a 5 percent risk of wrongly concluding it is not 30.0 hours (a risk of $\alpha = 0.05$). Then, using the following formula:

$$n = \left[\frac{\left(Z_{\alpha/2} + Z_{\beta}\right)_{\sigma}}{\mu - \mu_o}\right]^2$$

we plug in our values to obtain

$$n = \left[\frac{(1.96 + 0.84)\,10}{35 - 30}\right]^2 = 31.4$$

The required sample size is 32 (Gryna et al. 2007, p. 605).

Note that sample size sometimes is constrained by cost or time limitations; in addition, rules of thumb exist to estimate sample size. However, these potentially lead to gross under- or oversampling, with wasted time and effort. The recommended approach is to use power and sample size calculators (available online and in statistical software; these readily apply formulas appropriate for different sampling situations) in order to enter data collection and hypothesis testing with full knowledge of the statistically appropriate sample size.

20.7.3. Hypothesis Testing

A hypothesis, as used here, is an assertion about a population. Typically, the hypothesis is stated as a pair of hypotheses as follows: the null hypothesis (H_0) and an alternative hypothesis, H_a. The null hypothesis, H_0, is a statement of no change or no difference—hence, the term "null." The alternative hypothesis is the statement of change or difference—that is, if we reject the null hypothesis, the alternative is true by default.

For example, to test the hypothesis that the mean life of a population of batteries equals 30 hours, we state:

$$H_o : \mu = 30.0\,\text{hours}$$
$$H_a : \mu \neq 30.0\,\text{hours}$$

A hypothesis test is a test of the validity of the assertion, and is carried out by analyzing a sample of data. Sample results must be carefully evaluated for two reasons. First, there are many other samples that, by chance alone, could be drawn from the population. Second, the numerical results in the sample actually selected can easily be compatible with several different hypotheses. These points are handled by recognizing the two types of sampling errors, already alluded to above.

20.7.3.1. The Two Types of Sampling Errors

In evaluating a hypothesis, two errors can be made

1. Reject the null hypothesis when it is true. This is called a type I error, or the level of significance. The maximum probability of a type I error is denoted by α.
2. Fail to reject the null hypothesis when it is false. This is called type II error, and the probability is denoted by β.

These errors are defined in terms of probability numbers and can be controlled to desired values. The results possible in testing a hypothesis are summarized in Table 20.6. Definitions are found in Table 20.5. For additional detail on sampling errors in the context of quality, see Gryna et al. (2007).

Table 20.6 Type I (α) Error and Type II (β) Error

Suppose Decision of Analysis Is	Suppose the H_0 Is True	Suppose the H_0 Is False
Source: Quality Planning and Analysis, Copyright 2007. Used by permission.		
Fail to reject H_0	Correct decision $p = 1 - \alpha$	Wrong decision $p = \beta$
Reject H_0	Wrong decision $p = \alpha$	Correct decision $p = 1 - \beta$

20.7.3.2. Steps to Hypothesis Testing

As emphasized earlier, it is important to plan for data collection and analysis; an investigator ideally should arrive at the point of actual hypothesis testing with elements such as sample size already defined. Hypothesis testing often is an iterative process, however, and as mentioned above in the opening discussion of data collection, further data may be needed after initial collection, for example, to bolster sample sizes to obtain the desired power so that both type I and type II errors are defined in advance.

Generally, then, the steps to test a hypothesis are as follows:

1. State the practical problem.
2. State the null hypothesis and alternative hypothesis.
3. Choose a value for α (alpha). Common values are 0.01, 0.05, and 0.10.
4. Choose the test statistic for testing the hypothesis.
5. Determine the rejection region for the test (i.e., the range of values of the test statistic that results in a decision to reject the null hypothesis).
6. Obtain a sample of observations, compute the test statistic, and compare the value to the rejection region to decide whether to reject or fail to reject the hypothesis.
7. Draw the practical conclusion.

20.7.3.3. Common Tests of Hypotheses

No single means of organizing hypothesis tests can convey all the information that may be of interest to an investigator. Table 20.7 summarizes some common tests of hypotheses in terms of the formulas. Table 20.8 categorizes tests according to the question being asked and type of data. Figure 20.13 provides similar information but in the form of a roadmap to assist in deciding what hypothesis test(s) are appropriate. Readers may find that the combination of these presentations will provide the best understanding of what is a multifaceted topic.

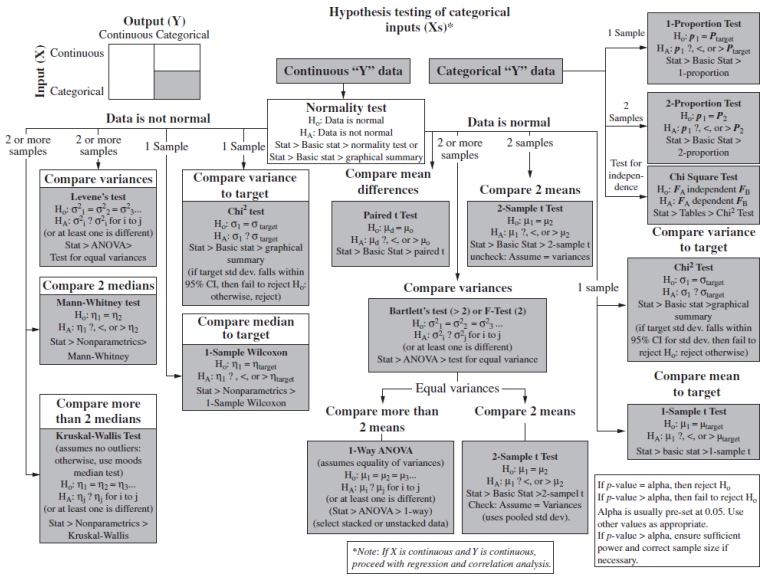
Table 20.7 Summary of Formulas on Tests of Hypotheses

Hypothesis	Test Statistic and Distribution
Source: Quality Planning and Analysis, Copyright 2007. Used by permission.	
H_o: $\mu = \mu_0$ (the mean of a normal population is equal to a specified value μ_0; s is known)	$Z = \frac{\bar{X} - \mu_o}{\sigma/\sqrt{n}}$ Standard normal distribution
H_o: $\mu = \mu_0$ (the mean of a normal population is equal to a specified value μ_0; σ is estimated by s)	$t = \frac{\bar{X} - \mu_o}{s/\sqrt{n}}$ t distribution with $n - 1$ degrees of freedom (DF)
H_o: $\mu_1 = \mu_2$ (the mean of population 1 is equal to the mean of population 2; assume that $\sigma_1 = \sigma_2$ and that both populations are normal)	$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{1/n_1 + 1/n_2}\sqrt{[(n_1-1)s_1^2 + (n_2-1)s_2^2]/(n_1+n_2-2)}}$ t distribution with DF = $n_1 + n_2 - 2$
H_o: $\sigma = \sigma_0$ (the standard deviation of a normal population is equal to a specified value σ_0)	$X^2 = \frac{(n-1)s^2}{\sigma_o^2}$ Chi-square distribution with DF = $n - 1$
H_o: $\sigma_1 = \sigma_2$ (the standard deviation of population 1 is equal to the standard deviation of population 2; assume that both populations are normal)	$F = \frac{s_1^2}{s_2^2}$ F distribution with $DF_1 = n_1 - 1$ and $DF_2 = n_2 - 1$
H_o: $\hat{p} = p_0$ (the fraction defective in a population is equal to a specified value p_0; assume that = sample proportion	$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$ Standard normal distribution
H_o: $p_1 = p_2$ (the fraction defective in population 1 is equal to the fraction defective in population 2; assume that n_1p_1 and n_2p_2 are each ≥5)	$Z = \frac{X_1/n_1 - X_2/n_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$ $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$ Standard normal distribution
To test for independence in a J × K contingency table that cross-classifies the variable A and B H_o: A is independent of B H_a: A is dependent on B	$X^2 = \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{(f_{jk} - e_{jk})^2}{e_{jk}}$ Chi-square distribution with DF = $(J - 1)(K - 1)$ Where f_{jk} = the observed frequency of data for category j of variable A and to category k of variable B e_{jk} = the expected frequency = $f_{j0}f_{0k}/f_{00}$ f_{j0} = frequency total for category j for variable A f_{0k} = frequency total for category k of variable B f_{00} = frequency total for $J \times K$ table

Table 20.8 Hypothesis Testing Table

Tests of hypotheses organized by the question being asked. All tests assume a categorical X in the $Y = f(X)$ format. For example, X might be manufacturing plant, and there could be 1, 2, or more than two plants of interest in terms of output, Y. A continuous Y might be mean or standard deviation of daily units produced, a categorical Y might be proportion defective units produced in a single day.					
Question: Is There a Difference in the Parameter	**Number of Sample Groups**	**Continuous Y (Normal)**		**Categorical Y**	
		Parameter of Interest	**Test**	**Parameter of Interest**	**Test**
*ANOVA assumes both equal variances and normality.					
Source: Juran Institute, Inc. Used by permission.					
Compared to a target?	1	μ σ	1-sample *t* Chi-square	Proportion	1-proportion test
Between two groups?	2	μ σ	2-sample *t* F-test	Proportion	2-proportion test
Among all groups?	≥2	μ σ	ANOVA* Bartlett's	Proportion	Chi-square test of Independence

Figure 20.13 ***Hypothesis testing.***



The hypothesis testing procedure is illustrated through the following example:

1. State the practical problem. To investigate a problem with warping wood panels, it was proposed that warping was caused by differing moisture content in the layers of the laminated product before drying. The sample data shown in Table 20.9 were taken between layers 1-2 and 2-3. Is there a significant difference in the moisture content?
2. State the null hypothesis and alternative hypothesis:

$$H_o : \mu 1 - 2 = \mu 2 - 3$$
$$H_a : \mu 1 - 2 \neq \mu 2 - 3$$

3. Choose a value for α. In this example, a type I error (α) of 0.05 will be assumed.
4. Choose the test statistic for testing the hypothesis.

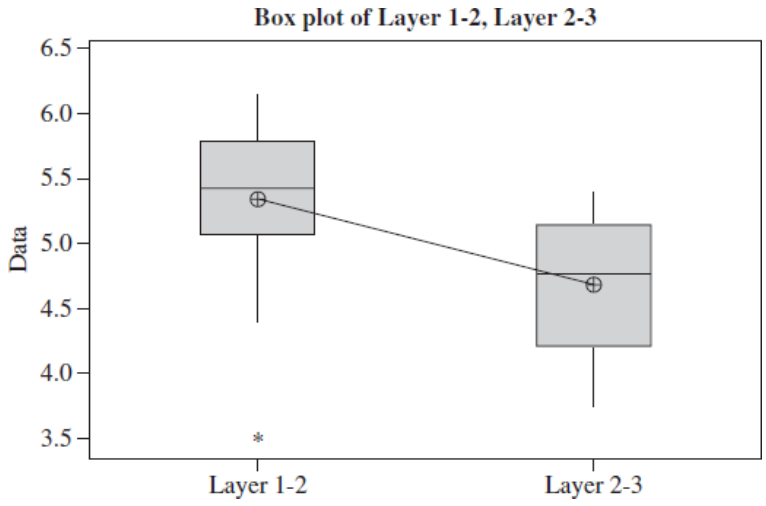
   Because we have two samples and desire to test for a difference in the means, a two-sample t-test is appropriate. (Note: A probability plot or test for normality will confirm the assumption of normality in the data. Also, an equal variance test concludes variances are approximately equal.)
5. Determine the rejection region for the test.

   The critical value defining the rejection region is approximately 2.0 (see Table 20.4); if the absolute value of the calculated t is larger than the critical value, then we reject the null hypothesis.
6. Obtain a sample of observations, compute the test statistic, and compare the value to the rejection region to decide whether to reject or fail to reject the hypothesis. A box plot (remember to plot the data!) suggests that the moisture content in Layer 1-2 tends to be higher than in Layer 2-3. Minitab output (see Fig. 20.14) shows that the calculated t is 4.18, which is in the rejection region.

Table 20.9 Moisture Content

Layer 1-2		Layer 2-3	
Source: Quality Planning and Analysis, Copyright 2007. Used by permission.			
4.43	4.40	3.74	5.14
6.01	5.99	4.30	5.19
5.87	5.72	5.27	4.16
4.64	5.25	4.94	5.18
3.50	5.83	4.89	4.78
5.24	5.44	4.34	5.42
5.34	6.15	5.30	4.05
5.99	5.14	4.55	3.92
5.75	5.72	5.17	4.07
5.48	5.00	5.09	4.54
5.64	5.01	4.74	4.23
5.15	5.42	4.96	5.07
5.64		4.21	

Figure 20.14 Box plot of Layer 1-2, Layer 2-3. (Quality Planning and Analysis, Copyright 2007. Used by permission.)



Because the calculated t is larger than the critical value, the associated p-value is $< \alpha$, and we reject the null hypothesis, H_0.

N	Mean	StDev	SE Mean	
Layer 1-2	25	5.350	0.613	0.12
Layer 2-3	25	4.689	0.499	0.10

$$\begin{gathered}\text{Difference} = \text{m}\mu(\text{Layer}\,1-2)\bar{\text{m}}\mu(\text{Layer}\,2-3)\\ \text{Estimate for difference}: 0.660901\\ 95\text{percent CI for difference}: (0.343158, 0.978644)\\ \text{T}-\text{test of difference} = 0(\text{vs.not} =): t-\text{value} = 4.18 p-\text{value} = 0.000\,\text{DF} = 48\end{gathered}$$

Both use pooled StDev = 0.5587.

1. Draw the practical conclusion. We conclude that the moisture content in Layer 1-2 is higher than the moisture content of Layer 2-3.

20.7.4. Nonparametric Hypothesis Tests, Data Transformation, and Bootstrapping

The preceding discussion has focused on "parametric" hypothesis tests (so-called because they rely on parameter estimation). Often, it is the case that one or more of the assumptions underlying the parametric tests are violated. In particular, practitioners frequently face skewed or otherwise nonnormal data, and application of parametric tests that assume bell-shaped data distribution may lead to erroneous conclusions and inappropriate action. Fortunately, options are available; these include nonparametric tests, data transformation, and bootstrapping.

Nonparametric hypothesis tests avoid violating key assumptions by virtue of being "distribution-free"; that is, they are not strictly dependent on particular distributions (such as a normal distribution); however, nonparametric tests have their own set of assumptions of which investigators should be aware. In effect, these methods typically transform the original data into ranks, and hypothesis tests then are carried out on the ranked data. Although nonparametric methods are not nearly as well developed and frequently are statistically less powerful compared to parametric tests, they are available for basic one-, two-, and two or more sample tests (see the bottom of Table 20.7 and the left side of the roadmap in Fig. 20.13). See Sprent and Smeeton (2001) for more on traditional nonparametric methods. New methods continue to emerge, for example, wavelets and nonparametric Bayesian techniques; see Kvam and Vidakovic (2007).

Data transformation allows one to take data that violate some assumption of a parametric test and change them so that the assumption no longer is violated. For example, nonnormal data, or sample data with unequal variances can be changed to new numbers that are normal or have equal variances. Three common methods are

20.7.4.1. Power Functions

Traditionally, standard functions such as taking the square (x^2), square root ($x^{1/2}$), log (log10(x)), natural log (ln(x)), or inverse (x^{-1}) were used because they could easily be done with a calculator. Trial and error often is needed to find a function that appropriately transforms the data to meet the test assumptions.

20.7.4.2. Box-Cox Transformation

This method provides simultaneous testing of power functions to find an optimum valueλ that minimizes the variance. Typically, one selects a power (value of λ) that is understandable and within a 95 percent confidence interval of the estimated λ (e.g., square: $\lambda = 2$; square root: $\lambda = 0.5$; natural log: $\lambda = 0$; inverse: $\lambda = -1$). The Box-Cox transformation does not work with negative numbers.

20.7.4.3. Johnson Transformation

This method selects an optimal function among three families of distributions (bounded, unbounded, lognormal). While effective in situations where Box-Cox does not work, the resulting transformation is not intuitive.

These methods are easy to apply (with software), and allow use of the more powerful parametric tests. However, the transformed data do not necessarily have intuitive meaning.

Bootstrapping is one of a broader class of computation-intensive resampling methods. Rather than assuming any particular distribution of a test statistic (such as normal), the distribution is determined empirically. More specifically, a statistic of interest (such as the mean) is repeatedly calculated from different samples drawn themselves, with replacements, from a sample. The distribution of these calculated statistics then is used as the basis for determining the probability of obtaining any particular value by chance. Itself a nonparametric approach, bootstrapping is a flexible method that gradually is gaining acceptance. For more information on the method and applications, see Davison and Hinkley (2006).

20.7.5. Correlation and Regression Analysis

Correlation and regression analysis help us understand relationships. More specifically, regression analysis is the modeling of the relationships between independent and dependent variables, while correlation analysis is a study of the strength of the linear relationships among variables. From a practical perspective, simple linear regression examines the distribution of one variable (the response, or dependent variable) as a function of one or more independent variables (the predictor, or independent variable) held at each of several levels. Note that the cause-and-effect relationship is stated explicitly, and it is this relationship that is tested to determine its statistical significance. In addition, regression analysis is used in forecasting and prediction based on the important independent variables, and in locating optimum operating conditions. In contrast, correlation typically looks at the joint variation of two variables that have not been manipulated by the experimenter, and there is no explicit cause-and-effect hypothesis.

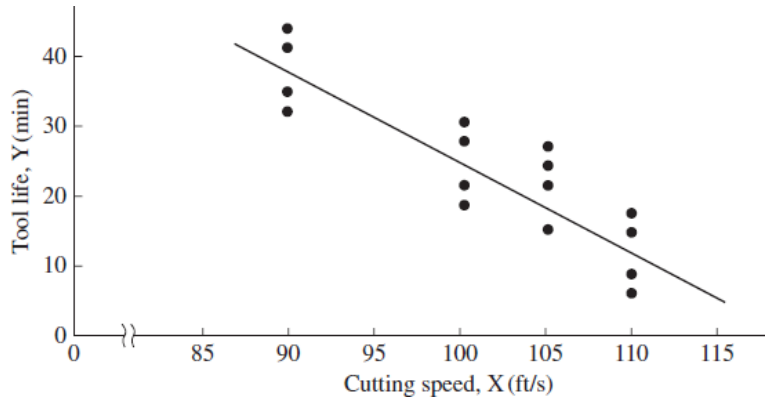
For example, suppose that the life of a tool varies with the cutting speed of the tool and we want to predict life based on cutting speed. Thus, life is the dependent variable (Y) and cutting speed is the independent variable (X). Data are collected at four cutting speeds (Table 20.10).

Table 20.10 Cutting Speed

X	Y	X	Y	X	Y	X	Y
(X, in feet per minute versus tool life; Y, in minutes)							
Source: Quality Planning and Analysis, Copyright 2007. Used by permission.							
90	41	100	22	105	21	110	15
90	43	100	35	105	13	110	11
90	35	100	29	105	18	110	6
90	32	100	18	105	20	110	10

Remembering to always plot the data, we note that a scatter plot (Fig. 20.15) suggests that life varies with cutting speed (specifically, life decreases with an increase in speed) and also varies in a linear manner (i.e., increases in speed result in a certain decrease in life that is the same over the range of the data). Note that the relationship is not perfect—the points scatter about the line.

Figure 20.15 Tool life (Y) versus cutting speed (X). (Quality Planning and Analysis, Copyright 2007.) Used by permission.



Often, it is valuable to obtain a regression equation. In this case, we have a linear relationship in the general form provided by

$$Y = \beta_0 + \beta_1 X + \in$$

where β_0 and β_1 are the unknown population intercept and slope, and ε is a random-error term that may be due to measurement errors and/or the effects of other independent variables. This model is estimated from sample data by the form

$$\hat{Y} = b_0 + b_1 X$$

where $\hat{Y}$ is the predicted value of Y for a given value of X and b_0 and b_1 are the sample estimates of β_0 and β_1. Estimates usually are found by least-squares methods; formulas can be found in statistics books such as Kutner et al. (2004).

For this example, the resulting prediction equation is

$$\text{Tool life} = 106.90 - 1.3614\,(\text{cutting speed})$$

This equation can be used to predict tool life by plugging in values of cutting speed. Extreme caution should be used in making predictions outside the actual sample space (e.g., for cutting speeds above or below the tested maximum or minimum), however, as these are tenuous without confirmation by observation.

Although a prediction equation can be found mathematically, it should not be used without knowing how "good" it is. A number of criteria exist for judging the adequacy of the prediction equation. One common measure is R_2, the proportion of variation explained by the prediction equation. R_2, or the coefficient of determination, is the ratio of the variation due to the regression to the total variation. The higher R_2, the greater the probable utility of the prediction equation in estimating Y based on X.

Another measure of the degree of association between two variables is the simple linear correlation coefficient,*r*. This is the square root of the coefficient of determination, so that the values of *r* range from –1 to +1. A positive *r* is consistent with a positive relationship (an increase in one variable is associated with an increase in the other), whereas the opposite is true of a negative *r* (an increase in one variable is associated with a decrease in the other). Scatter plots are strongly recommended when interpreting correlations, especially as very different patterns can result in identical values of *r*. The significance level of *r* varies with sample size; statistical software is recommended to obtain exact significance levels.

The above discussion introduces simple linear correlation and regression—the direction and strength of a relationship between two variables, or prediction of a dependent variable, Y, from a single predictor variable, X. A natural extension of this is multiple regression that allows for two or more independent variables. For a discussion of how to estimate and examine a multiple regression prediction equation, see Kutner et al. (2004).

20.7.6. Analysis of Variance

Analysis of Variance (ANOVA) is an approach related to linear regression, falling into the class of what are called general linear models. However, unlike regression, the X is discrete rather than continuous (noting that general linear models actually can blend characteristics of both regression and ANOVA). In ANOVA, the total variation of all measurements around the overall mean is divided into sources of variation that are then analyzed for statistical significance. It is used in situations where the investigator is interested in comparing the means among two or more discrete groups. For example, an investigator may be interested in comparing performance among three different machine configurations. The ANOVA analysis detects a difference somewhere among the means (i.e., at least one mean is different from the others), and confidence intervals or follow-up tests such as pairwise comparisons can be applied to determine which mean (or means) is different. ANOVA is the basis for design of experiments, discussed next.

20.7.7. Design of Experiments

With origins in the pioneering work in agriculture of Sir Ronald A. Fisher, designed experiments have taken on an increasingly significant role in quality improvement in the business world. This section will first compare the classical and designed approaches to experimentation, thereby providing the reader with an understanding as to the limitations of traditional methods and the power of contemporary methods. Next, basic concepts and terminology will be introduced in the context of an example improvement problem, followed by an overview of different types of designs and the typical progression through a series of designed experiments. The section finishes with the related topic of Taguchi designs.

20.7.7.1. Contrast between the Classical and Contemporary Methods of Experimentation

The classical method of experimentation is to vary one factor at a time (sometimes called OFAT), holding everything else constant. By way of example, and to illustrate the need for designed experiments, consider the case of a certain fellow who decided he wanted to investigate the causes of intoxication. As the story goes, he drank some whiskey and water on Monday and became highly inebriated. The next day, he repeated the experiment holding all variables constant except one… he decided to replace the whiskey with vodka. As you may guess, the result was drunkenness. On the third day, he repeated the experiment for the last time. On this trial, he used bourbon in lieu of the whiskey and vodka. This time it took him two days just to be able to gather enough of his faculties to analyze the experimental results. After recovering, he concluded that water causes intoxication. Why? Because it was the common variable!

The contrast between this traditional method and the designed approach is striking. In particular, a designed approach permits the greatest information to be gained from the fewest data points (efficient experimentation), and allows the estimation of interaction effects among factors. Table 20.11 compares these two approaches in more detail for an experiment in which there are two factors (or variables) whose effects on a characteristic are being investigated (the same conclusions hold for an experiment with more than two factors).

Table 20.11 Comparison of Classical and Modern Methods of Experimentation

Criteria	Classical	Modern
Source: Quality Planning and Analysis, Copyright 2007. Used by permission.		
Basic procedure	Hold everything constant except the factor under investigation. Vary that factor and note the effect on the characteristic of concern. To investigate a second factor, conduct a separate experiment in the same manner.	Plan the experiment to evaluate both factors in one main experiment. Include in the design measurements to evaluate the effect of varying both factors simultaneously.
Experimental conditions	Care should be taken to have material, workers, and machine constant throughout the entire experiment.	Realizes difficulty of holding conditions reasonably constant throughout an entire experiment. Instead, experiment is divided into several groups or blocks of measurements. Within each block, conditions must be reasonably constant (except for deliberate variation to investigate a factor).
Experimental error	Recognized but not stated in quantitative terms.	Stated in quantitative terms.
Basis of evaluation	Effect due to a factor is evaluated with only a vague knowledge of the amount of experimental error.	Effect due to a factor is evaluated by comparing variation due to that factor with the quantitative measure of an experimental error.
Possible bias due to sequence of measurements	Often assumed that sequence has no effect.	Guarded against by randomization.
Effect of varying both factors simultaneously ("interaction")	Not adequately planned into experiment. Frequently assumed that the effect of varying factor 1 (when factor 2 is held constant at some value) would be the same for any value of factor 2.	Experiment can be planned to include an investigation for interaction between factors.
Validity of results	Misleading and erroneous if interaction exists and is not realized.	Even if interaction exists, a valid evaluation of the main factors can be made.
Number of measurements	For a given amount of useful and valid information, more measurements are needed than in the modern approach.	Fewer measurements needed for useful and valid information.
Definition of problem	Objective of experiment frequently not defined as necessary.	Designing the experiment requires defining the objective in detail (how large an effect do we want to determine, what numerical risks can be taken, etc.).
Application of conclusions	Sometimes disputed as applicable only to the controlled conditions under which the experiment was conducted.	Broad conditions can be planned in the experiment, thereby making conclusions applicable to a wider range of actual conditions.

20.7.7.2. Concepts and Terminology—An Example Designed Experiment

Suppose that three detergents (A, B, C) are to be compared for their ability to clean clothes in an automatic washing machine. The "whiteness" readings obtained by a special measuring procedure are the dependent, or response, variable. The independent variable under investigation (detergent) is a factor, and each variation of the factor is called a level; in this case, there are three levels. A treatment is a single level assigned to a single factor, detergent A. A treatment combination is the set of levels for all factors in a given experimental run. A factor may be qualitative (different detergents) or quantitative (water temperature). Finally, some experiments have a fixed-effects model (i.e., the levels investigated represent all levels of concern to the investigator—for example, three specific washing machines or brands). Other experiments have a random effects model, that is, the levels chosen are just a sample from a larger population (e.g., three operators of washing machines). A mixed-effects model has both fixed and random factors.

Figure 20.16 outlines six possible designs of experiments, starting with the classical design in (a). Here, all factors except detergent are held constant. Thus, nine tests are run, three with each detergent with the washing time, make of machine, water temperature, and all other factors held constant. One drawback of this design is that the conclusions about detergent brands apply only to the specific conditions of the experiment.

Figure 20.16 Some experimental designs. (Quality Planning and Analysis, Copyright 2007. Used by permission.)

A	B	C
-	-	-
-	-	-
-	-	-

(a)

I	II	III
A	B	C
A	B	C
A	B	C

(b)

I	II	III
C	B	B
A	C	B
A	A	C

(c)

I	II	III
B	A	C
C	C	A
A	B	B

(d)

	I	II	III
1	C	A	B
2	B	C	A
3	A	B	C

(e)

I	II	III
ABC	ABC	ABC
1 ---	---	---
2 ---	---	---
3 ---	---	---

(f)

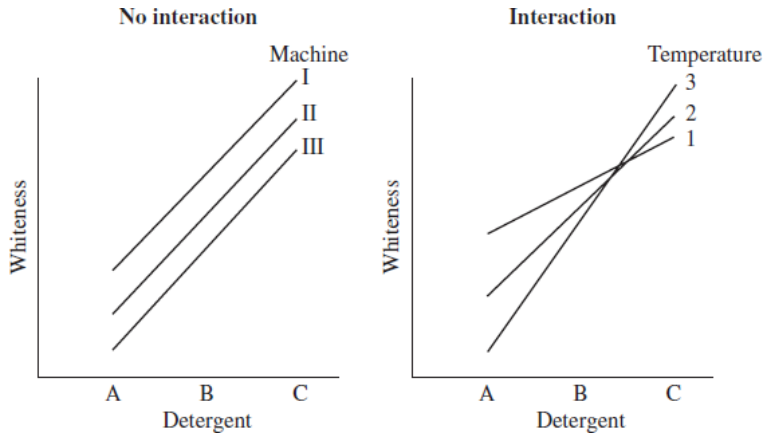
Design (b) recognizes a second factor at three levels (i.e., washing machines brands I, II, and III). However, in this design, it would not be known whether an observed difference was due to detergents or washing machine (they are said to be confounded).

In design (c), the nine tests are assigned completely at random, thus the name "completely randomized design." However, detergent A is not used with machine brand III, and detergent B is not used with machine brand I, thus complicating the conclusions.

Design (d) shows a randomized block design. Here each block is a machine brand, and the detergents are run in random order within each block. This design guards against any possible bias due to the order in which the detergents are used and has advantages in the subsequent data analysis and conclusions. First, a test of hypothesis can be run to compare detergents and a separate test of hypothesis run to compare machines; all nine observations are used in both tests. Second, the conclusions concerning detergents apply for the three machines and vice versa, thus providing conclusions over a wider range of conditions.

Now suppose that another factor such as water temperature is also to be studied, using the Latin square design shown in (e). Note that this design requires using each detergent only once with each machine and only once with each temperature. Thus, three factors can be evaluated (by three separate tests of hypothesis) with only nine observations. However, there is a danger. This design assumes no interaction among the factors. No interaction between detergent and machine means that the effect of changing from detergent A to B to C does not depend on which machine is used, and similarly for the other combinations of factors. The concept of interaction is shown in Fig. 20.17. There is no interaction among the detergents and the machines. But the detergents do interact with temperature. At high temperatures, C is the best performer. At low temperatures, A performs best.

Figure 20.17 Interaction. (Quality Planning and Analysis, Copyright 2007. Used by permission)



Finally, the main factors and possible interactions could be investigated by the factorial design in design (f) inFig. 20.16. Factorial means that at least one test is run for every combination of main factors, in this case 3 × 3 × 3 or 27 combinations. Separate tests of hypothesis can be run to evaluate the main factors and also possible interactions. Again, all the observations contribute to each comparison. When there are many factors, a portion of the complete factorial (i.e., a "fractional factorial") is useful when experimental resources are limited (see its application in a sequential testing approach, below).

Most problems can be handled with one of the standard experimental designs or a series of these. Designs can be classified by the number of factors to be investigated, the structure of the experimental design, and the kind of information the experiment is intended to provide (Table 20.12). For a description of both the design and analysis of various design structures, see Box et al. (2005). Another excellent general reference is Myers et al. (2009) for a detailed look at response surface designs.

Table 20.12 Classification of Designs

Design	Type of Application
Source: Adapted from JQH5, Table 47.3.	
Completely randomized	Appropriate when only one experimental factor is being investigated
Factorial	Appropriate when several factors are being investigated at two or more levels and interaction of factors may be significant
Blocked factorial	Appropriate when number of runs required for factorial is too large to be carried out under homogeneous conditions
Fractional factorial	Appropriate when many factors and levels exist and running all combinations is impractical
Randomized block	Appropriate when one factor is being investigated and experimental material or environment can be divided into blocks or homogeneous groups
Balanced incomplete block	Appropriate when all the treatments cannot be accommodated in a block
Partially balanced incomplete block	Appropriate if a balanced incomplete block requires a larger number of blocks than is practical
Latin square	Appropriate when one primary factor is under investigation and results may be affected by two other experimental variables or by two sources of nonhomogeneity. It is assumed that no interactions exist.
Youden square	Same as Latin square, but number of rows, columns, and treatments need not be the same
Nested	Appropriate when objective is to study relative variability instead of mean effect of sources of variation (e.g., variance of tests on the same sample and variance of different samples)
Response surface	Objective is to provide empirical maps (contour diagrams) illustrating how factors under the experimenter's control influence the response
Mixture designs	Use when constraints are inherent (e.g., the sum of components in a paint must add to 100%)

A sequential approach to experimentation often can be helpful. Briefly, a typical sequence of designed experiments will allow an experimenter to quickly and efficiently narrow down a large number of possible factors (or X's in the Y = f(X) terminology of Lean Six Sigma) to find out which are most important, and then refine the relationships to find optimal settings for each of the vital few factors. The steps might be as follows:

1. *Screening experiment*. In this stage, a fractional factorial design may be applied that does not allow interactions to be detected, but can ferret out which of many factors have the greatest main effect.
2. *Fractional factorial design*. The smaller number of factors identified in the screening experiment are tested to allow detection of interaction effects.
3. *Full factorial design*. A small number of factors (usually no more than five) are tested to allow all main effects and higher-order (e.g., three-way, four-way) interactions to be detected and accounted for. Such designs also can detect curvature that indicates a potential optimum.
4. *Response surface design*. By adding data points in particular ways (e.g., a composite design), an experimenter can build on earlier experiments to fully characterize nonlinear relationships and pinpoint optimal settings.

5. *Evolutionary Operation (EVOP).* Once an improved process is in production mode, evolutionary operation techniques can be used to conduct many small experiments on production units over time. Although individual changes are small, the cumulative effect over time can be quite large, and exemplifies the power of continuous improvement. See Box and Draper (1969) for a classic text on this subject.

For a series of four papers on sequential experimentation, see Carter (1996). Emanuel and Palanisamy (2000) discuss sequential experimentation at two levels and a maximum of seven factors.

20.7.8. Taguchi Approach to Experimental Design

Professor Genichi Taguchi uses an approach to experimental design that has three purposes:

1. Design products and processes that perform consistently on target and are relatively insensitive ("robust") to factors that are difficult to control.
2. Design products that are relatively insensitive (robust) to component variation.
3. Minimize variation around a target value.

Thus, although cited in this "improvement tools" section because of its association with DOE, the approach is meant to provide valuable information for product design and development (see the section "Statistical Tools for Designing for Quality" in this chapter). Taguchi divides quality control into online control (e.g., diagnosing and adjusting a process during production) and offline control that encompasses the engineering design process and its three phases: systems design, parameter design, and tolerance design. For an extensive bibliography and a summary of some controversial aspects of the Taguchi approach, see Box and Draper (1969, pp. 47.58 and 47.59).

Many books are available that cover DOE for engineering and manufacturing applications. For readers in nonmanufacturing environments, Ledolter and Swersey (2007) may be of interest. Another text readers may find useful for not only classical but more contemporary techniques (e.g., Bayesian inference, kriging) is del Castillo (2007).

20.7.9. Discrete Event and Monte Carlo Simulation

Advances in user-friendly software make computer simulations increasingly accessible to quality practitioners that do not have a strong background in mathematics, programming, or modeling. Numerous types of simulation models exist, but two that may be of most interest to readers are discrete event and Monte Carlo simulations. These can be powerful methods for making process improvements; in particular, modeling provides a means of asking "what if?" questions and rapidly testing the effects of process changes and potential solutions in a safe, low-risk environment.

20.7.9.1. Discrete Event Simulation

Discrete event simulation (DES) attempts to mimic situations in which there are distinct, recognizable events and transactions. In a hospital, for example, arrival of patients at an emergency department and subsequent steps in patient care represent specific events that combine into a flow of transactions: arrival, registration, triage, nursing assessment, physician assessment, etc., through inpatient admission, discharge, or transfer. Discrete event simulation enables system components to be changed and tracks the resulting process flow over time to help understand the relationships among inputs, outputs, and process variables.

Typically, a process flow diagram (or process "map") that graphically displays the sequence and flow of activities forms the basis for a discrete event simulation. A discrete event simulation takes this basic flow diagram and adds inputs and process variables that govern the flow of transactions. Following on the hospital example, these include inputs (such as patient arrivals), human resources (e.g., number of nurses, physician schedules, overtime availability, skill levels, pay rates, etc.), equipment resources (e.g., types and number of beds, imaging equipment, etc.), rules for flow (the required sequence of steps, batching of inputs or outputs, priority rules, exceptions, decisions), resource acquisition (what resources are needed to complete an activity (e.g., one RN or one physician's assistant; two RNs; one RN and one physician, etc.), activity cycle times (work time, wait time), and similar details.

Once these details are built into the model, it "runs" by tracing the path of units (patients, in the hospital example) from arrival through to exit from the process. Patients are processed in accordance with the activities, rules, and constraints, and any relevant attributes (patient-specific characteristics) that may be assigned to them (e.g., acuity level, age, gender). The output consists of a multitude of descriptive statistics and measures that portray the collective behavior of the process as the various players interact and move through time.

Although every model is different and details vary, there are basic steps that should be a part of every simulation study. These steps and related questions are (adapted from Law and Kelton 2000):

1. State the problem and question(s) being asked. What is the business need for the simulation? What problem is to be fixed? What answers are being sought?
2. Prepare a plan for the simulation study. Who needs to be involved? What data are needed and how will data be collected? What alternative scenarios are to be tested? What are the milestones and timeline for completion?
3. Collect data. What is my current state? What are the data for alternative scenarios? Are there gaps in the data, and how will they be handled?
4. Build and validate a conceptual model. Given available data, what is the general structure of the model? What will be the inputs, process variables, and outputs? What statistical accumulators are needed, and where? If the model is built, will it provide the answers to the questions?
5. Build and validate an operational model. Are the model components necessary and sufficient? Does the model produce results consistent with the current state?
6. Design scenarios or experiments needed to answer the questions. What model parameters will be changed? Which are fixed? What combinations of factors need to be tested?
7. Run the scenarios or experiments to obtain the needed outputs. Are the results reproducible? Are additional scenarios or experiments suggested?
8. Analyze and interpret the data. What are the statistical results? Do the descriptive statistics and/or statistical tests indicate meaningful effects? What are the answers to the original questions? Are additional questions raised?

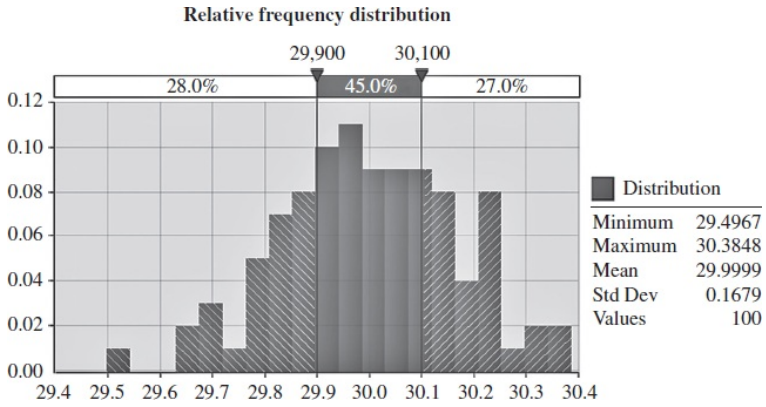
As emphasized at the beginning of this chapter, formulation of the question(s) being asked is a critical first step to the successful application of simulation modeling. Failure to have a clear understanding of what the model is being asked to do leads to poorly constructed models, models with insufficient inputs or process detail, or overly complicated models that take unnecessary time and effort to build and run. In addition, a clearly communicated business need will garner the stakeholder support needed to collect data, evaluate the model, and implement suggested changes.

20.7.9.2. Monte Carlo

Named after the famed gambling destination, this method seeks to account for uncertainty (variability) in inputs and carry this forward into probability distributions of outcomes. Essentially, instead of using single, fixed values in equations [such as Y = f(X)], distributions are used for the inputs (Xs), and samples repeatedly are drawn from the distributions, yielding a distribution of outputs (Y values) instead of a single value. For example, while the forecasted net return on a new product could simply be stated as an expected $10 million, it would be useful to know the probability of achieving this, or that the uncertainty in the forecast is such that there is a high probability of a negative return.

By way of illustration, assume we have three components, A, B, and C that are assembled end-to-end to create a final product. If the mean lengths are 5, 10, and 15 mm, then we can simply add these together to arrive at an expected mean combined total length of 5 mm + 10 mm + 15 mm = 30 mm. However, we know from the concept of statistical variation that there will be variation in the components. Assuming we sample populations of each component and find the respective distributions for each of A, B, and C, what can we expect the overall distribution of assembled product length to look like? By repeatedly taking a random sample from each distribution and adding the lengths, Monte Carlo simulation generates a distribution of the total length. Figure 20.18 shows the relative frequency distribution of the combined lengths of the three components from a Monte Carlo simulation with each of the three components having a standard deviation of 0.1 mm. The mean expected combined total length is almost exactly 30 mm, but the simulation shows the variation around this, with only 45% of assembled components expected to be within +/− 0.1 mm of the total mean value. This approach provides substantially more information than the single estimate of 30 mm.

Figure 20.18 ***Result of Monte Carlo simulation showing a relative frequency distribution of combined total length of three components A, B, and C that individually have normal distributions of 5, 10, and 15 mm, respectively, each with a standard deviation of 0.1 mm. The mean expected combined total length is approximately 30 mm, but the simulation shows the variation around this, e.g., that only 45% of assembled components are expected to be within +/− 0.1 mm of this mean value.***



20.7.9.3. Simulated DOE

As tools evolve, they are being combined in new ways. One example is the combination of Monte Carlo, discrete event simulation, and DOE. Briefly, this approach involves a discrete event simulation (DES) that uses probability distributions for the input and/or process variables (Monte Carlo), and the investigator changes these variables (as factors) following a structured, designed approach (DOE). While any results and conclusions should be treated as preliminary until verified by actual experimentation, this can be particularly useful in environments where real-life changes may be difficult or dangerous to make.

20.7.10. Additional Advanced Analysis Tools

For practitioners faced with more complex scenarios such as multiple variables (more than one y and/or x), nonlinear data, or categorical outputs, extensions of the general linear models and other alternatives are available. In particular are methods for multivariate analysis; this refers to statistical techniques that simultaneously analyze multiple measurements on subjects. Many techniques are extensions of the univariate (single-variable distributions) and bivariate (correlation, regression) methods dealt with above. Beyond the scope of this chapter, these include:

- *Multiple regression*. Applies when the investigator has a single, continuous dependent variable and multiple, continuous independent variables (Xs) of interest.
- *Nonlinear regression*. Useful when data cannot easily be treated by standard linear methods (note that curvilinear data do not necessarily require nonlinear methods).
- *Nonparametric linear regression*. Applies when the usual assumptions of regression are violated.
- *Multiple discriminant analysis*. Used in situations with a single, categorical (dichotomous or multichotomous) dependent variable (Y) and continuous independent variables (Xs).
- *Logistic regression*. Also known as logit analysis, this is a combination of multiple regression and multiple discriminant analysis in which one or more categorical or continuous independent variables (Xs) are used to predict a single, categorical dependent variable (Y). Odds ratios often are computed with this method.
- *Multivariate analysis of variance and covariance (MANOVA, MANCOVA)*. Dependence techniques that extend ANOVA to allow more than one continuous, dependent variable (Y) and several categorical independent variables (Xs).
- *Principal component analysis (PCA) and common factor analysis*. These methods analyze interrelationships among a large number of variables and seek to condense the information into a smaller set of factors without loss of information.
- *Cluster analysis*. An interdependence technique that allows mutually exclusive subgroups to be identified based on similarities among the individuals. Unlike discriminant analysis, the groups are not predefined.
- *Canonical correlation analysis*. An extension of multiple regression that correlates simultaneously several continuous dependent variables (Ys) and several continuous independent variables (Xs).
- *Conjoint analysis*. Often used in marketing analyses, this method helps assess the relative importance of both attributes and levels of complex entities (e.g., products). It is useful when trade-offs exist when making comparisons.
- *Multidimensional scaling*. An interdependence method (also called perceptual mapping), this seeks to transform preferences or judgments of similarity into a representation by distance in multidimensional space.
- *Correspondence analysis*. Another interdependence technique; this accommodates the perceptual mapping of objects (such as products) onto a set of categorical attributes. This method allows both categorical data and nonlinear relationships.

Readers are encouraged to research any techniques that appear to fit their need; although complex, these are powerful means of getting useful information from data. Some useful references include:

Multivariate techniques:

Hair, J. F., Jr., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (2006).

Multivariate Data Analysis. Pearson Prentice-Hall, Upper Saddle River, NJ.

Affifi, A., Clark, V. A., and May, S. (2004). *Computer-Aided Multivariate Analysis* (4th ed.).

Chapman and Hall/CRC Press, Boca Raton, FL.

Coleman, S, Greenfield, T., Stewardson, D., and Montgomery, D. C. (2008). *Statistical Practice in Business and Industry.* John Wiley & Sons, Hoboken, NJ. (see Chapter 13).

Hypothesis testing and DOE:

Box, G. E. P., Hunter, J. S., and Hunter, W. G. (2005). *Statistics for Experimenters: Design, Innovation and Discovery* (2nd ed.). Wiley-Interscience, Hoboken, NJ.

Logistic regression, Poisson regression, odds ratios:

Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* John Wiley & Sons, New York.

Nonparametric:

Sprent, P., and Smeeton, N. C. (2001). *Applied Nonparametric Statistical Methods* (3rd ed.). Chapman and Hall/CRC Press, Boca Raton, FL.

[1] *A confidence level is the probability that an assertion about the value of a population parameter is correct. Confidence levels of 90, 95, or 99 percent are usually used in practice.

[2] A mathematical derivation of degrees of freedom is beyond the scope of this book, but the underlying concept can be stated. Degrees of freedom (DF) is the parameter involved when, for example, a sample standard deviation is used to estimate the true standard deviation of a universe. DF equals the number of measurements in the sample minus some number of constraints estimated from the data to compute the standard deviation. In this example, it was necessary to estimate only one constant (the population mean) to compute the standard deviation. Therefore, DF = 25 – 1 = 24.