7

Inference for Unstructured Multisample Studies

Chapter 6 introduced the basics of formal statistical inference in one- and two-sample studies. This chapter begins to consider formal inference for multisample studies, with a look at methods that make no explicit use of structure relating the samples (beyond time order of data collection). That is, the study of inference methods specifically crafted for use in factorial and fractional factorial studies and in curve- and surface-fitting analyses will be delayed until subsequent chapters.

The chapter opens with a discussion of the standard one-way model typically used in the analysis of measurement data from multisample studies and of the role of residuals in judging its appropriateness. The making of confidence intervals in multisample contexts is then considered, including both individual and simultaneous confidence interval methods. The one-way analysis of variance (ANOVA) test for the hypothesis of equality of several means and a related method of estimating variance components are introduced next. The chapter then covers the basics of Shewhart control (or process monitoring) charts. The $\bar{x}$, R, and s control charts for measurement data are studied. The chapter then closes with a section on p charts and u charts for attributes data.

7.1 The One-Way Normal Model

Statistical engineering studies often produce samples taken under not one or two, but rather many different sets of conditions. So although the inference methods of Chapter 6 are a start, they are not a complete statistical toolkit for engineering problem solving. Methods of formal inference appropriate to multisample studies are also needed.

This section begins to provide such methods. First the reader is reminded of the usefulness of some of the simple graphical tools of Chapter 3 for making informal comparisons in multisample studies. Next the "equal variances, normal distributions" model is introduced. The role of residuals in evaluating the reasonableness of that model in an application is explained and emphasized. The section then proceeds to introduce the notion of combining several sample variances to produce a single pooled estimate of baseline variation. Finally, there is a discussion of how standardized residuals can be helpful when sample sizes vary considerably.

7.1.1 Graphical Comparison of Several Samples of Measurement Data

Any thoughtful analysis of several samples of engineering measurement data should begin with the making of graphical representations of those data. Where samples are small, side-by-side dot diagrams are the most natural graphical tool. Where sample sizes are moderate to large (say, at least six or so data points per sample), side-by-side boxplots are effective.

Example 1 Comparing Compressive Strengths for Eight Different Concrete Formulas

Armstrong, Babb, and Campen did compressive strength testing on 16 different concrete formulas. Part of their data are given in Table 7.1, where eight different

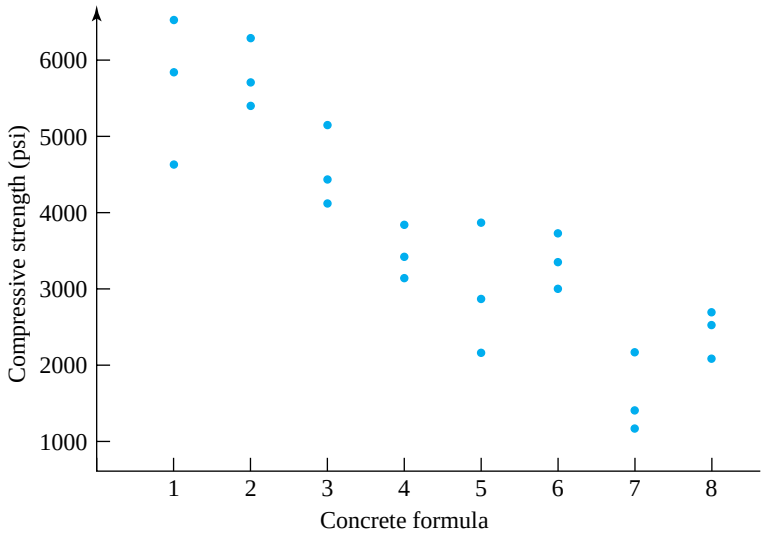

Figure 7.1 Side-by-side dot diagrams for eight samples of compressive strengths

Table 7.1
Compressive Strengths for 24 Concrete Specimens

Specimen	Concrete Formula	28-Day Compressive Strength (psi)
1	1	5,800
2	1	4,598
3	1	6,508
4	2	5,659
5	2	6,225
6	2	5,376
7	3	5,093
8	3	4,386
9	3	4,103
10	4	3,395
11	4	3,820
12	4	3,112
13	5	3,820
14	5	2,829
15	5	2,122
16	6	2,971
17	6	3,678
18	6	3,325
19	7	2,122
20	7	1,372
21	7	1,160
22	8	2,051
23	8	2,631
24	8	2,490

formulas are represented. (The only differences between formulas 1 through 8 are their water/cement ratios. Formula 1 had the lowest water/cement ratio, and the ratio increased with formula number in the progression .40, .44, .49, .53, .58, .62, .66, .71. Of course, knowing these water/cement ratios suggests that a curve-fitting analysis might be useful with these data, but for the time being this possibility will be ignored.)

Making side-by-side dot diagrams for these eight samples of sizes $n_1 = n_2 = n_3 = n_4 = n_5 = n_6 = n_7 = n_8 = 3$ amounts to making a scatterplot of compressive strength versus formula number. Such a plot is shown in Figure 7.1. The general message conveyed by Figure 7.1 is that there are clear differences in mean compressive strengths between the formulas but that the variabilities in compressive strengths are roughly comparable for the eight different formulas.

Example 2

Comparing Empirical Spring Constants for Three Different Types of Springs

Hunwardsen, Springer, and Wattonville did some testing of three different types of steel springs. They made experimental determinations of spring constants for $n_1 = 7$ springs of type 1 (a 4 in. design with a theoretical spring constant of 1.86), $n_2 = 6$ springs of type 2 (a 6 in. design with a theoretical spring constant of 2.63), and $n_3 = 6$ springs of type 3 (a 4 in. design with a theoretical spring constant of 2.12), using an 8.8 lb load. The students' experimental values are given in Table 7.2.

These samples are just barely large enough to produce meaningful boxplots. Figure 7.2 gives a side-by-side boxplot representation of these data. The primary qualitative message carried by Figure 7.2 is that there is a substantial difference in empirical spring constants between the 6 in. spring type and the two 4 in. spring types but that no such difference between the two 4 in. spring types is obvious. Of course, the information in Table 7.2 could also be presented in side-by-side dot diagram form, as in Figure 7.3.

Table 7.2
Empirical Spring Constants

Type 1 Springs	Type 2 Springs	Type 3 Springs
1.99, 2.06, 1.99	2.85, 2.74, 2.74	2.10, 2.01, 1.93
1.94, 2.05, 1.88	2.63, 2.74, 2.80	2.02, 2.10, 2.05
2.30		

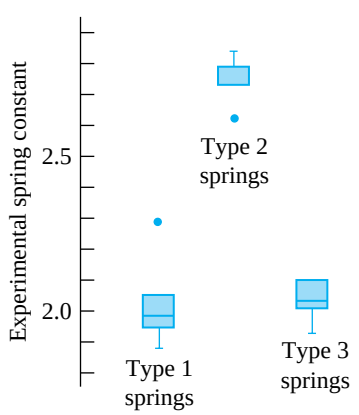

Figure 7.2 Side-by-side boxplots of empirical spring constants for springs of three types

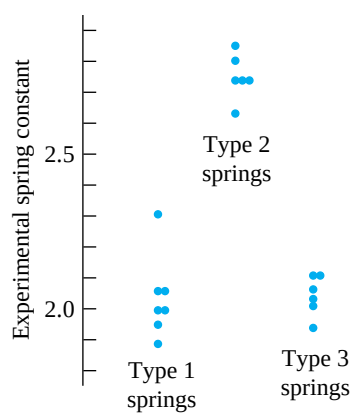

Figure 7.3 Side-by-side dot diagrams for three samples of empirical spring constants

Methods of formal statistical inference are meant to sharpen and quantify the impressions that one gets when making a descriptive analysis of data. But an intelligent graphical look at data and a correct application of formal inference methods rarely tell completely different stories. Indeed, the methods of formal inference offered here for simple, unstructured multisample studies are *confirmatory*—in cases like Examples 1 and 2, they should confirm what is clear from a descriptive or *exploratory* look at the data.

7.1.2 The One-Way (Normal) Multisample Model, Fitted Values, and Residuals

Chapter 6 emphasized repeatedly that to make one- and two-sample inferences, one must adopt a model for data generation that is both manageable and plausible. The present situation is no different, and standard inference methods for unstructured multisample studies are based on a natural extension of the model used in Section 6.3 to support small-sample comparison of two means. The present discussion will be carried out under the assumption that r samples of respective sizes $n_1, n_2, \ldots, n_r$ are independent samples from normal underlying distributions with a common variance—say, σ^2. Just as in Section 6.3 the $r = 2$ version of this **one-way** (as opposed, for example, to several-way factorial) **model** led to useful inference methods for $\mu_1 - \mu_2$, this general version will support a variety of useful inference methods for r-sample studies. Figure 7.4 shows a number of different normal distributions with a common standard deviation. It represents essentially what must be generating measured responses if the methods of this chapter are to be applied.

One-way normal model assumptions

In addition to a description of the one-way model in words and the pictorial representation given in Figure 7.4, it is helpful to have a description of the model in symbols. This and the next three sections will employ the notation

$$y_{ij} = \text{the } j\text{th observation in sample } i$$

The model equation used to specify the one-way model is then

One-way model statement in symbols

$$y_{ij} = \mu_i + \epsilon_{ij} \tag{7.1}$$

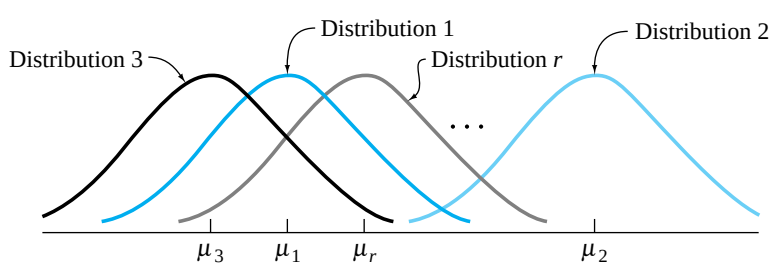

Figure 7.4 r normal distributions with a common standard deviation

where μ_i is the ith underlying mean and the quantities $\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{1n_1}, \epsilon_{21}, \epsilon_{22}, \ldots, \epsilon_{2n_2}, \ldots, \epsilon_{r1}, \epsilon_{r2}, \ldots, \epsilon_{rn_r}$ are independent normal random variables with mean 0 and variance σ^2. (In this statement, the means $\mu_1, \mu_2, \ldots, \mu_r$ and the variance σ^2 are typically unknown parameters.)

Equation (7.1) says exactly what is conveyed by Figure 7.4 and the statement of the one-way assumptions in words. But it says it in a way that is suggestive of another useful pattern of thinking, reminiscent of the "residual" notion that was used extensively in Sections 4.1, 4.2, and 4.3. That is, equation (7.1) says that an observation in sample i is made up of the corresponding underlying mean plus some random noise, namely

$$\epsilon_{ij} = y_{ij} - \mu_i$$

This is a theoretical counterpart of an empirical notion met in Chapter 4. There, it was useful to decompose data into fitted values and the corresponding residuals.

In the present situation, since any structure relating the r different samples is specifically being ignored, it may not be obvious how to apply the notions of fitted values and residuals. But a plausible meaning for

$$\hat{y}_{ij} = \text{the fitted value corresponding to } y_{ij}$$

in the present context is the ith sample mean

ith sample mean

$$\boxed{\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}}$$

That is,

Fitted values for the one-way model

$$\boxed{\hat{y}_{ij} = \bar{y}_i} \qquad \textbf{(7.2)}$$

(This is not only intuitively plausible but also consistent with what was done in Sections 4.1 and 4.2. If one fits the approximate relationship $y_{ij} \approx \mu_i$ to the data via least squares—i.e., by minimizing $\sum_{ij}(y_{ij} - \mu_i)^2$ over choices of $\mu_1, \mu_2, \ldots, \mu_r$—each minimizing value of μ_i is $\bar{y}_i$.)

Taking equation (7.2) to specify fitted values for an r-sample study, the pattern established in Chapter 4 (specifically, Definition 4, page 132) then says that residuals are differences between observed values and sample means. That is, with

$$e_{ij} = \text{the residual corresponding to } y_{ij}$$

one has

Residuals for the one-way model

$$\boxed{e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_i} \tag{7.3}$$

Rearranging display (7.3) gives the relationship

$$y_{ij} = \hat{y}_{ij} + e_{ij} = \bar{y}_i + e_{ij} \tag{7.4}$$

which is an empirical counterpart of the theoretical statement (7.1). In fact, combining equations (7.1) and (7.4) into a single statement gives

$$y_{ij} = \mu_i + \epsilon_{ij} = \bar{y}_i + e_{ij} \tag{7.5}$$

This is a specific instance of a pattern of thinking that runs through all of the common normal-distribution-based methods of analysis for multisample studies. In words, equation (7.5) says

$$\text{Observation} = \text{deterministic response} + \text{noise} = \text{fitted value} + \text{residual} \tag{7.6}$$

and display (7.6) is a paradigm that provides a unified way of approaching the majority of the analysis methods presented in the rest of this book.

The decompositions (7.5) and (7.6) suggest that

1. the fitted values ($\hat{y}_{ij} = \bar{y}_i$) are meant to approximate the deterministic part of a system response (μ_i), and
2. the residuals (e_{ij}) are therefore meant to approximate the corresponding noise in the response (ϵ_{ij}).

The fact that the ϵ_{ij} in equation (7.1) are assumed to be iid normal $(0, \sigma^2)$ random variables then suggests that the e_{ij} ought to look at least approximately like a random sample from a normal distribution.

So the normal-plotting of an entire set of residuals (as in Chapter 4) is a way of checking on the reasonableness of the one-way model. The plotting of residuals against (1) fitted values, (2) time order of observation, or (3) any other potentially relevant variable—hoping (as in Chapter 4) to see only random scatter—are other ways of investigating the appropriateness of the model assumptions.

These kinds of plotting, which combine residuals from all r samples, are often especially useful in practice. When r is large at all, budget constraints on total data collection costs often force the individual sample sizes $n_1, n_2, \ldots, n_r$ to be fairly small. This makes it fruitless to investigate "single variance, normal distributions" model assumptions using (for example) sample-by-sample normal plots. (Of course, where all of $n_1, n_2, \ldots, n_r$ are of a decent size, a sample-by-sample approach can be effective.)

Example 1 *(continued)*

Returning again to the concrete strength study, consider investigating the reasonableness of model (7.1) for this case. Figure 7.1 is a first step in this investigation. As remarked earlier, it conveys the visual impression that at least the "equal variances" part of the one-way model assumptions is plausible. Next, it makes sense to compute some summary statistics and examine them, particularly the sample standard deviations. Table 7.3 gives sample sizes, sample means, and sample standard deviations for the data in Table 7.1.

At first glance, it might seem worrisome that in this table s_1 is more than three times the size of s_8. But the sample sizes here are so small that a largest ratio of

Table 7.3
Summary Statistics for the Concrete Strength Study

i, Concrete Formula	n_i, Sample Size	$\bar{y}_i$, Sample Mean (psi)	s_i, Sample Standard Deviation (psi)
1	$n_1 = 3$	$\bar{y}_1 = 5{,}635.3$	$s_1 = 965.6$
2	$n_2 = 3$	$\bar{y}_2 = 5{,}753.3$	$s_2 = 432.3$
3	$n_3 = 3$	$\bar{y}_3 = 4{,}527.3$	$s_3 = 509.9$
4	$n_4 = 3$	$\bar{y}_4 = 3{,}442.3$	$s_4 = 356.4$
5	$n_5 = 3$	$\bar{y}_5 = 2{,}923.7$	$s_5 = 852.9$
6	$n_6 = 3$	$\bar{y}_6 = 3{,}324.7$	$s_6 = 353.5$
7	$n_7 = 3$	$\bar{y}_7 = 1{,}551.3$	$s_7 = 505.5$
8	$n_8 = 3$	$\bar{y}_8 = 2{,}390.7$	$s_8 = 302.5$

Table 7.4
Example Computations of Residuals for the Concrete Strength Study

Specimen	i, Concrete Formula	y_{ij}, Compressive Strength (psi)	$\hat{y}_{ij} = \bar{y}_i$, Fitted Value	e_{ij}, Residual
1	1	5,800	5,635.3	164.7
2	1	4,598	5,635.3	−1,037.3
3	1	6,508	5,635.3	872.7
4	2	5,659	5,753.3	−94.3
5	2	6,225	5,753.3	471.7
⋮	⋮	⋮	⋮	⋮
22	8	2,051	2,390.7	−339.7
23	8	2,631	2,390.7	240.3
24	8	2,490	2,390.7	99.3

sample standard deviations on the order of 3.2 is hardly unusual (for $r = 8$ samples of size 3 from a normal distribution). Note from the F tables (Tables B.6) that for samples of size 3, even if only 2 (rather than 8) sample standard deviations were involved, a ratio of sample variances of $(965.6/302.5)^2 \approx 10.2$ would yield a p-value between .10 and .20 for testing the null hypothesis of equal variances with a two-sided alternative. The sample standard deviations in Table 7.3 really carry no strong indication that the one-way model is inappropriate.

Since the individual sample sizes are so small, trying to see anything useful in eight separate normal plots of the samples is hopeless. But some insight can be gained by calculating and plotting all $8 \times 3 = 24$ residuals. Some of the calculations necessary to compute residuals for the data in Table 7.1 (using the fitted values appearing as sample means in Table 7.3) are shown in Table 7.4. Figures 7.5 and 7.6 are, respectively, a plot of residuals versus fitted y (e_{ij} versus $\bar{y}_{ij}$) and a normal plot of all 24 residuals.

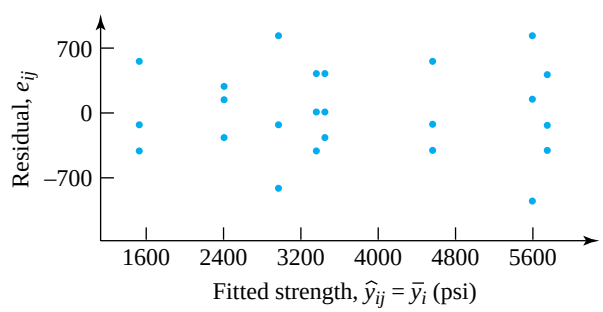

Figure 7.5 Plot of residuals versus fitted responses for the compressive strengths

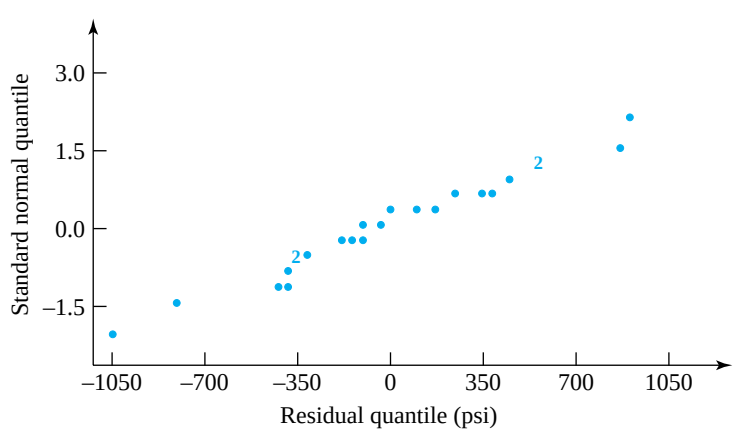

Figure 7.6 Normal plot of the compressive strength residuals

Example 1 *(continued)*

Figure 7.5 gives no indication of any kind of strong dependence of σ on μ (which would violate the "constant variance" restriction). And the plot in Figure 7.6 is reasonably linear, thus identifying no obvious difficulty with the assumption of normal distributions. In all, it seems from examination of both the raw data and the residuals that analysis of the data in Table 7.1 on the basis of model (7.1) is perfectly sensible.

Example 2 *(continued)*

The spring testing data can also be examined with the potential use of the one-way normal model (7.1) in mind. Figures 7.2 and 7.3 indicate reasonably comparable variabilities of experimental spring constants for the $r = 3$ different spring types. The single very large value (for spring type 1) causes some doubt both in terms of this judgment and also (by virtue of its position on its boxplot as an outlying value) regarding a "normal distribution" description of type 1 experimental constants. Summary statistics for these samples are given in Table 7.5.

Table 7.5
Summary Statistics for the Empirical Spring Constants

i, Spring Type	n_i	$\bar{y}_i$	s_i
1	7	2.030	.134
2	6	2.750	.074
3	6	2.035	.064

Without the single extreme value of 2.30, the first sample standard deviation would be .068, completely in line with those of the second and third samples. But even the observed ratio of largest to smallest sample variance (namely $(.134/.064)^2 = 4.38$) is not a compelling reason to abandon a one-way model description of the spring constants. (A look at the F tables with $\nu_1 = 6$ and $\nu_2 = 5$ shows that 4.38 is between the $F_{6,5}$ distribution .9 and .95 quantiles. So even if there were only two rather than three samples involved, a variance ratio of 4.38 would yield a p-value between .1 and .2 for (two-sided) testing of equality of variances.) Before letting the single type 1 empirical spring constant of 2.30 force abandonment of the highly tractable model (7.1) some additional investigation is warranted.

Sample sizes $n_1 = 7$ and $n_2 = n_3 = 6$ are large enough that it makes sense to look at sample-by-sample normal plots of the spring constant data. Such plots, drawn on the same set of axes, are shown in Figure 7.7. Further, use of the fitted values ($\bar{y}_i$) listed in Table 7.5 with the original data given in Table 7.2 produces

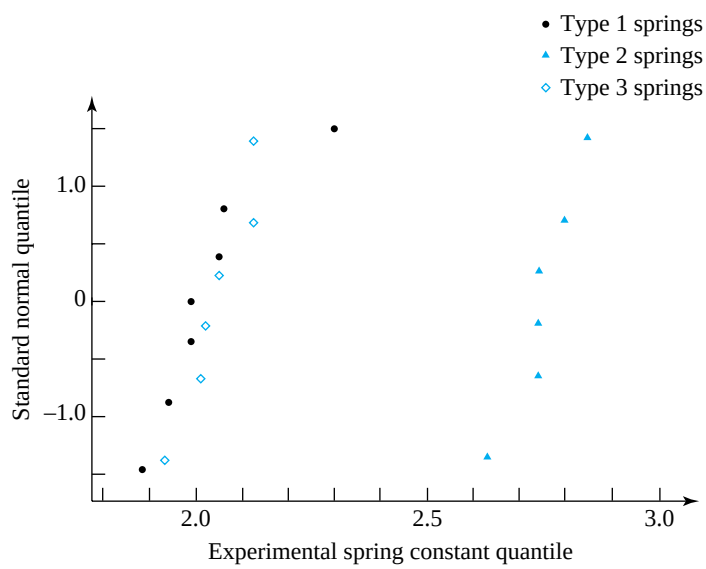

Figure 7.7 Normal plots of empirical spring constants for springs of three types

Table 7.6
Example Computations of Residuals for the Spring Constant Study

i, Spring Type	j, Observation Number	y_{ij}, Spring Constant	$\hat{y}_{ij} = \bar{y}_i$, Sample Mean	e_{ij}, Residual
1	1	1.99	2.030	−.040
⋮	⋮	⋮	⋮	⋮
1	7	2.30	2.030	.270
2	1	2.85	2.750	.100
⋮	⋮	⋮	⋮	⋮
2	6	2.80	2.750	.050
3	1	2.10	2.035	.065
⋮	⋮	⋮	⋮	⋮
3	6	2.05	2.035	.015

19 residuals, as partially illustrated in Table 7.6. Then Figures 7.8 and 7.9, respectively, show a plot of residuals versus fitted responses and a normal plot of all 19 residuals.

Example 2
(*continued*)

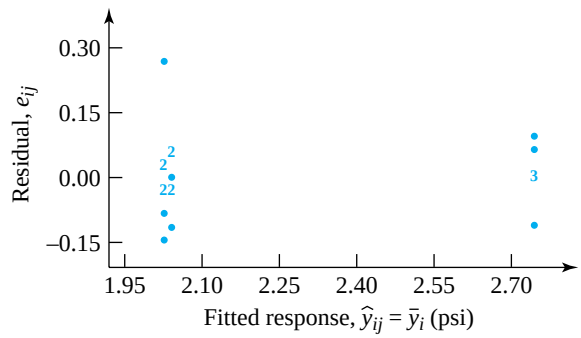

Figure 7.8 Plot of residuals versus fitted responses for the empirical spring constants

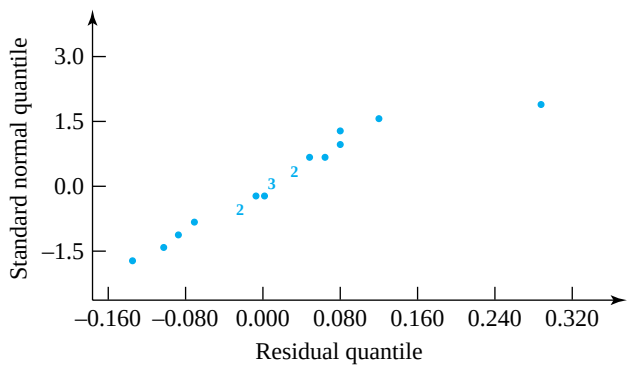

Figure 7.9 Normal plot of the spring constant residuals

But Figures 7.8 and 7.9 again draw attention to the largest type 1 empirical spring constant. Compared to the other measured values, 2.30 is simply too large (and thus produces a residual that is too large compared to all the rest) to permit serious use of model (7.1) with the spring constant data. Barring the possibility that checking of original data sheets would show the 2.30 value to be an arithmetic blunder or gross error of measurement (which could be corrected or legitimately force elimination of the 2.30 value from consideration), it appears that the use of model (7.1) with the $r = 3$ spring types could produce inferences with true (and unknown) properties quite different from their nominal properties.

One might, of course, limit attention to spring types 2 and 3. There is nothing in the second or third samples to render the "equal variances, normal distributions" model untenable for those two spring types. But the pattern of variation for springs of type 1 appears to be detectably different from that for springs of types 2 and 3, and the one-way model is not appropriate when all three types are considered.

7.1.3 A Pooled Estimate of Variance for Multisample Studies

The "equal variances, normal distributions" model (7.1) has as a fundamental parameter, σ, the standard deviation associated with responses from any of conditions $1, 2, 3, \ldots, r$. Similar to what was done in the $r = 2$ situation of Section 6.3, it is typical in multisample studies to **pool** the r sample variances to arrive at a single estimate of σ derived from all r samples.

Definition 1

If r numerical samples of respective sizes $n_1, n_2, \ldots, n_r$ produce sample variances $s_1^2, s_2^2, \ldots, s_r^2$, the **pooled sample variance**, s_{p}^2, is the weighted average of the sample variances, where the weights are the sample sizes minus 1. That is,

$$s_{\mathrm{p}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_r - 1)s_r^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_r - 1)} \tag{7.7}$$

The **pooled sample standard deviation**, s_{p}, is the square root of s_{p}^2.

Definition 1 is just Definition 14 in Chapter 6 restated for the case of more than two samples. As was the case for s_{p} based on two samples, s_{p} is guaranteed to lie between the largest and smallest of the s_i and is a mathematically convenient form of compromise value.

Equation (7.7) can be rewritten in a number of equivalent forms. For one thing, letting

The total number of observations in an r-sample study

$$n = \textstyle\sum_{i=1}^{r} n_i = \text{the total number of observations in all } r \text{ samples}$$

it is common to rewrite the denominator on the right of equation (7.7) as

$$\sum_{i=1}^{r}(n_i - 1) = \sum_{i=1}^{r} n_i - \sum_{i=1}^{r} 1 = n - r$$

And noting that the ith sample variance is

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

the numerator on the right of equation (7.7) is

$$\sum_{i=1}^{r}(n_i - 1)\left(\frac{1}{(n_i - 1)}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2\right) = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \qquad \textbf{(7.8)}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{n_i} e_{ij}^2 \qquad \textbf{(7.9)}$$

Alternative formulas for s_P^2

So one can define s_P^2 in terms of the right-hand side of equation (7.8) or (7.9) divided by $n - r$.

Example 1 *(continued)*

For the compressive strength data, each of $n_1, n_2, \ldots, n_8$ are 3, and s_1 through s_8 are given in Table 7.3. So using equation (7.7),

$$\begin{aligned} s_P^2 &= \frac{(3-1)(965.6)^2 + (3-1)(432.3)^2 + \cdots + (3-1)(302.5)^2}{(3-1) + (3-1) + \cdots + (3-1)} \\ &= \frac{2[(965.6)^2 + (432.3)^2 + \cdots + (302.5)^2]}{16} \\ &= \frac{2{,}705{,}705}{8} \\ &= 338{,}213 \text{ (psi)}^2 \end{aligned}$$

and thus

$$s_P = \sqrt{338{,}213} = 581.6 \text{ psi}$$

One estimates that if a large number of specimens of any one of formulas 1 through 8 were tested, a standard deviation of compressive strengths on the order of 582 psi would be obtained.

The meaning of s_P

s_P is an estimate of the intrinsic or baseline variation present in a response variable at a fixed set of conditions, calculated supposing that the baseline variation is constant across the conditions under which the samples were collected. When that supposition is reasonable, the pooling idea allows a number of individually unreliable small-sample estimates to be combined into a single, relatively more reliable combined estimate. It is a fundamental measure that figures prominently in a variety of useful methods of formal inference.

On occasion, it is helpful to have not only a single number as a data-based best guess at σ^2 but a confidence interval as well. Under model restrictions (7.1), the variable

$$\frac{(n-r)s_{\mathrm{P}}^2}{\sigma^2}$$

has a χ^2_{n-r} distribution. Thus, in a manner exactly parallel to the derivation in Section 6.4, a two-sided confidence interval for σ^2 has endpoints

Confidence limits for the one-way model variance

$$\boxed{\frac{(n-r)s_{\mathrm{P}}^2}{U} \quad \text{and} \quad \frac{(n-r)s_{\mathrm{P}}^2}{L}} \tag{7.10}$$

where L and U are such that the χ^2_{n-r} probability assigned to the interval (L, U) is the desired confidence level. And, of course, a one-sided interval is available by using only one of the endpoints (7.10) and choosing U or L such that the χ^2_{n-r} probability assigned to the interval $(0, U)$ or (L, ∞) is the desired confidence.

Example 1 *(continued)*

In the concrete compressive strength case, consider the use of display (7.10) in making a two-sided 90% confidence interval for σ. Since $n - r = 16$ degrees of freedom are associated with s_{P}^2, one consults Table B.5 for the .05 and .95 quantiles of the χ^2_{16} distribution. These are 7.962 and 26.296, respectively. Thus, from display (7.10), a confidence interval for σ^2 has endpoints

$$\frac{16(581.6)^2}{26.296} \quad \text{and} \quad \frac{16(581.6)^2}{7.962}$$

So a two-sided 90% confidence interval for σ has endpoints

$$\sqrt{\frac{16(581.6)^2}{26.296}} \quad \text{and} \quad \sqrt{\frac{16(581.6)^2}{7.962}}$$

that is,

$$453.7\text{psi} \quad \text{and} \quad 824.5\text{psi}$$

7.1.4 Standardized Residuals

In discussing the use of residuals, the reasoning has been that the e_{ij} are meant to approximate the corresponding random errors ϵ_{ij}. Since the model assumptions are

that the ϵ_{ij} are iid normal variables, the e_{ij} ought to look approximately like iid normal variables. This is sensible rough-and-ready reasoning, adequate for many circumstances. But strictly speaking, the e_{ij} are neither independent nor identically distributed, and it can be important to recognize this.

As an extreme example of the dependence of the residuals for a given sample i, consider a case where $n_i = 2$. Since

$$e_{ij} = y_{ij} - \bar{y}_i$$

one immediately knows that $e_{i1} = -e_{i2}$. So e_{i1} and e_{i2} are clearly dependent.

One can further apply Proposition 1 of Chapter 5 to show that if the sample sizes n_i are varied, the residuals don't have the same variance (and therefore can't be identically distributed). That is, since

$$e_{ij} = y_{ij} - \bar{y}_i = \left(\frac{n_i - 1}{n_i}\right) y_{ij} - \frac{1}{n_i} \sum_{j' \neq j} y_{ij'}$$

it is the case that

$$\operatorname{Var} e_{ij} = \left(\frac{n_i - 1}{n_i}\right)^2 \sigma^2 + \left(-\frac{1}{n_i}\right)^2 (n_i - 1)\sigma^2 = \frac{n_i - 1}{n_i}\sigma^2 \qquad \textbf{(7.11)}$$

So, for example, residuals from a sample of size $n_i = 2$ have variance $\sigma^2/2$, while those from a sample of size $n_i = 100$ have variance $99\sigma^2/100$, and *one ought to expect residuals from larger samples to be somewhat bigger in magnitude than those from small samples*.

A way of addressing at least the issue that residuals need not have a common variance is through the use of **standardized residuals**.

Definition 2

If a residual e has variance $a \cdot \sigma^2$ for some positive constant a, and s is some estimate of σ, the **standardized residual** corresponding to e is

$$\boxed{e^* = \frac{e}{s\sqrt{a}}} \qquad \textbf{(7.12)}$$

The division by $s\sqrt{a}$ in equation (7.12) is a division by an estimated standard deviation of e. It serves, so to speak, to put all of the residuals on the same scale.

Plotting with standardized residuals

Standardized residuals for the one-way model

$$e_{ij}^{*} = \frac{e_{ij}}{s_{\mathrm{P}}\sqrt{\dfrac{n_i - 1}{n_i}}} \tag{7.13}$$

is a somewhat more refined way of judging the adequacy of the one-way model than the plotting of raw residuals e_{ij} illustrated in Examples 1 and 2. When all n_i are the same, as in Example 1, the plotting of the standardized residuals in equation (7.13) is completely equivalent to plotting with the raw residuals. And as a practical matter, unless some n_i are very small and others are very large, the standardization used in equation (7.13) typically doesn't have much effect on the appearance of residual plots.

Example 2 *(continued)*

In the spring constant study, allowing for the fact that sample 1 is larger than the other two (and thus according to the model (7.1) should produce larger residuals) doesn't materially change the outcome of the residual analysis. To see this, note that using the summary statistics in Table 7.5,

$$s_{\mathrm{P}}^2 = \frac{(7-1)(.134)^2 + (6-1)(.074)^2 + (6-1)(.064)^2}{(7-1)+(6-1)+(6-1)} = .0097$$

so that

$$s_{\mathrm{P}} = \sqrt{.0097} = .099$$

Then using equation (7.13), each residual from sample 1 should be divided by

$$.099\sqrt{\frac{7-1}{7}} = .0913$$

to get standardized residuals, while each residual from the second and third samples should be divided by

$$.099\sqrt{\frac{6-1}{6}} = .0900$$

Clearly, .0913 and .0900 are not much different, and the division before plotting has little effect on the appearance of residual plots. By way of example, a normal plot of all 19 standardized residuals is given in Figure 7.10. Verify its similarity to the normal plot of all 19 raw residuals given in Figure 7.9 on page 454.

Example 2
(*continued*)

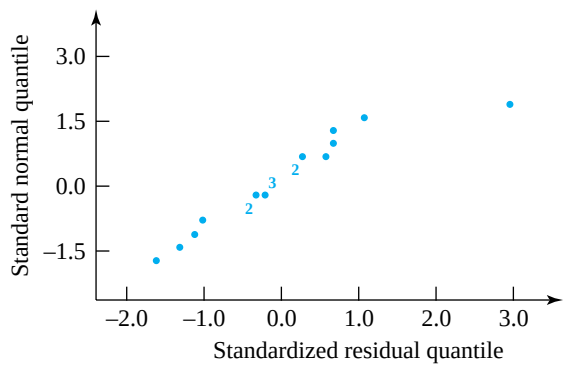

Figure 7.10 Normal plot of the spring constant standardized residuals

The notion of standardized residuals is often introduced only in the context of curve- and surface-fitting analyses, where the variances of residuals $e = (y - \hat{y})$ depend not only on the sizes of the samples involved but also on the associated values of the independent or predictor variables ($x_1, x_2, \ldots,$ etc.). The concept has been introduced here, not only because it can be of importance in the present situation if the sample sizes vary widely but also because it is particularly easy to motivate the idea in the present context.

Section 1 Exercises

1. Return again to the data of Example 1 in Chapter 4. These may be viewed as simply $r = 5$ samples of $m = 3$ densities. (For the time being, ignore the fact that the pressure variable is quantitative and that curve fitting seems a most natural method of analysis to apply to these data.)
   (a) Compute and make a normal plot of the residuals for the one-way model. What does the plot indicate about the appropriateness of the one-way model assumptions here?
   (b) Using the five samples, find s_P, the pooled estimate of σ. What does this value measure in this context? Give a two-sided 90% confidence interval for σ based on s_P.
2. In an ISU engineering research project, so called "tilttable tests" were done in order to determine the angles at which vehicles experience lift-off of the "high-side" wheels and begin to roll over. So called "tilttable ratios" (which are the tangents of angles at which lift-off occurs) were measured for four different vans with the following results:

Van #1	Van #2	Van #3	Van #4
1.096, 1.093	.962, .970	1.010, 1.024	1.002, 1.001
1.090, 1.093	.967, .966	1.021, 1.020	1.002, 1.004
		1.022	

(Notice that Van #3 was tested five times while the others were tested four times each.) Vans #1 and #2 were minivans, and Vans #3 and #4 were full-size vans.
   (a) Compute and normal-plot residuals as a crude means of investigating the appropriateness of the one-way model assumptions for tilttable ratios. Comment on the appearance of your plot.
   (b) Redo part (a) using standardized residuals.
   (c) Compute a pooled estimate of the standard deviation based on these four samples. What is s_P supposed to be measuring in this example? Give a two-sided 95% confidence interval for σ based on s_P.

7.2 Simple Confidence Intervals in Multisample Studies

Section 6.3 illustrates how useful confidence intervals for means and differences in means can be in one- and two-sample studies. Estimating an individual mean and comparing a pair of means are every bit as important when there are r samples as they are when there are only one or two. The methods of Chapter 6 can be applied in r-sample studies by simply limiting attention to one or two of the samples at a time. But since individual sample sizes in multisample studies are often small, such a strategy of inference often turns out to be relatively uninformative. Under the one-way model assumptions discussed in the previous section, it is possible to base inference methods on the pooled standard deviation, s_{p}. Those tend to be relatively more informative than the direct application of the formulas from Section 6.3 in the present context.

This section first considers the confidence interval estimation of a single mean and of the difference between two means under the "equal variances, normal distributions" model. There follows a discussion of confidence intervals for any linear combination of underlying means. Finally, the section closes with some comments concerning the notions of individual and simultaneous confidence levels.

7.2.1 Intervals for Means and for Comparing Means

The primary drawback to applying the formulas from Section 6.3 in a multisample context is that typical small sample sizes lead to small degrees of freedom, large t multipliers in the plus-or-minus parts of the interval formulas, and thus long intervals. But based on the one-way model assumptions, confidence interval formulas can be developed that tend to produce shorter intervals.

That is, in a development parallel to that in Section 6.3, under the one-way normal model,

$$T = \frac{\bar{y}_i - \mu_i}{\dfrac{s_{\mathrm{p}}}{\sqrt{n_i}}}$$

has a t_{n-r} distribution. Hence, a two-sided confidence interval for the ith mean, μ_i, has endpoints

Confidence limits for μ_i based on the one-way model

$$\boxed{\bar{y}_i \pm t\frac{s_{\mathrm{p}}}{\sqrt{n_i}}} \tag{7.14}$$

where the associated confidence is the probability assigned to the interval from $-t$ to t by the t_{n-r} distribution. This is exactly formula (6.20) from Section 6.3, except that s_{p} has replaced s_i and the degrees of freedom have been adjusted from $n_i - 1$ to $n - r$.

In the same way, for conditions i and i', the variable

$$T = \frac{\bar{y}_i - \bar{y}_{i'} - (\mu_i - \mu_{i'})}{s_{\mathrm{P}}\sqrt{\dfrac{1}{n_i} + \dfrac{1}{n_{i'}}}}$$

has a t_{n-r} distribution. Hence, a two-sided confidence interval for $\mu_i - \mu_{i'}$ has endpoints

Confidence limits for $\mu_i - \mu_{i'}$ based on the one-way model

$$\boxed{\bar{y}_i - \bar{y}_{i'} \pm t s_{\mathrm{P}}\sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}} \qquad \textbf{(7.15)}$$

where the associated confidence is the probability assigned to the interval from $-t$ to t by the t_{n-r} distribution. Display (7.15) is essentially formula (6.35) of Section 6.3, except that s_{P} is calculated based on r samples instead of two and the degrees of freedom are $n - r$ instead of $n_i + n_{i'} - 2$.

Of course, use of only one endpoint from formula (7.14) or (7.15) produces a one-sided confidence interval with associated confidence corresponding to the t_{n-r} probability assigned to the interval $(-\infty, t)$ (for $t > 0$). The virtues of formulas (7.14) and (7.15) (in comparison to the corresponding formulas from Section 6.3) are that (when appropriate) for a given confidence, they will tend to produce shorter intervals than their Chapter 6 counterparts.

Example 3
(*Example 1 revisited*)

Confidence Intervals for Individual, and Differences of, Mean Concrete Compressive Strengths

Return to the concrete strength study of Armstrong, Babb, and Campen. Consider making first a 90% two-sided confidence interval for the mean compressive strength of an individual concrete formula and then a 90% two-sided confidence interval for the difference in mean compressive strengths for two different formulas. Since $n = 24$ and $r = 8$, there are $n - r = 16$ degrees of freedom associated with $s_{\mathrm{P}} = 581.6$. So the .95 quantile of the t_{16} distribution, namely 1.746, is appropriate for use in both formulas (7.14) and (7.15).

Turning first to the estimation of a single mean compressive strength, since each n_i is 3, the plus-or-minus part of formula (7.14) gives

$$t\frac{s_{\mathrm{P}}}{\sqrt{n_i}} = 1.746\frac{581.6}{\sqrt{3}} = 586.3 \text{ psi}$$

So ± 586.3 psi precision could be attached to any one of the sample means in Table 7.7 as an estimate of the corresponding formula's mean strength. For

example, since $\bar{y}_3 = 4{,}527.3$ psi, a 90% two-sided confidence interval for μ_3 has endpoints

$$4{,}527.3 \pm 586.3$$

that is,

$$3{,}941.0 \text{ psi} \quad \text{and} \quad 5{,}113.6 \text{ psi}$$

In parallel fashion, consider estimation of the difference in two mean compressive strengths with 90% confidence. Again, since each n_i is 3, the plus-or-minus part of formula (7.15) gives

$$ts_{\mathrm{P}}\sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}} = 1.746(581.6)\sqrt{\frac{1}{3} + \frac{1}{3}} = 829.1 \text{ psi}$$

Thus, ± 829.1 psi precision could be attached to any difference between sample means in Table 7.7 as an estimate of the corresponding difference in formula mean strengths. For instance, since $\bar{y}_3 = 4{,}527.3$ psi and $\bar{y}_7 = 1{,}551.3$ psi, a 90% two-sided confidence interval for $\mu_3 - \mu_7$ has endpoints

$$(4{,}527.3 - 1{,}551.3) \pm 829.1$$

That is,

$$2{,}146.9 \text{ psi} \quad \text{and} \quad 3{,}805.1 \text{ psi}$$

Table 7.7
Concrete Formula Sample Mean Strengths

Concrete Formula	Sample Mean Strength (psi)
1	5,635.3
2	5,753.3
3	4,527.3
4	3,442.3
5	2,923.7
6	3,324.7
7	1,551.3
8	2,390.7

The use of $n - r = 16$ degrees of freedom in Example 3 instead of $n_i - 1 = 2$ and $n_i + n_{i'} - 2 = 4$ reflects the reduction in uncertainty associated with s_{P} as an

estimate of σ as compared to that of s_i and of s_{p} based on only two samples. That reduction is, of course, bought at the price of restriction to problems where the "equal variances" model is tenable.

7.2.2 Intervals for General Linear Combinations of Means

There is an important and simple generalization of the formulas (7.14) and (7.15) that is easy to state and motivate at this point. Its most common applications are in the context of factorial studies. But it is pedagogically most sound to introduce the method in the unstructured r-sample context, so that the logic behind it is clear and is seen not to be limited to factorial analyses. The basic notion is that μ_i and $\mu_i - \mu_{i'}$ are particular linear combinations of the r means $\mu_1, \mu_2, \ldots, \mu_r$, and the same logic that produces confidence intervals for μ_i and $\mu_i - \mu_{i'}$ will produce a confidence interval for any linear combination of the r means.

That is, suppose that for constants $c_1, c_2, \ldots, c_r$, the quantity

A linear combination of population means

$$L = c_1\mu_1 + c_2\mu_2 + \cdots + c_r\mu_r \tag{7.16}$$

is of engineering interest. (Note that, for example, if all c_i's except c_3 are 0 and $c_3 = 1$, $L = \mu_3$, the mean response from condition 3. Similarly, if $c_3 = 1$, $c_5 = -1$, and all other c_i's are 0, $L = \mu_3 - \mu_5$, the difference in mean responses from conditions 3 and 5.) A natural data-based way to approximate L is to replace the theoretical or underlying means, μ_i, with empirical or sample means, $\bar{y}_i$. That is, define an estimator of L by

A linear combination of sample means

$$\boxed{\hat{L} = c_1\bar{y}_1 + c_2\bar{y}_2 + \cdots + c_r\bar{y}_r} \tag{7.17}$$

(Clearly, if $L = \mu_3$, then $\hat{L} = \bar{y}_3$, while if $L = \mu_3 - \mu_5$, then $\hat{L} = \bar{y}_3 - \bar{y}_5$.)

The one-way model assumptions make it very easy to describe the distribution of $\hat{L}$ given in equation (7.17). Since $E\bar{y}_i = \mu_i$ and $\operatorname{Var}\bar{y}_i = \sigma^2/n_i$, one can appeal again to Proposition 1 of Chapter 5 (page 307) and conclude that

$$\begin{aligned} E\hat{L} &= c_1E\bar{y}_1 + c_2E\bar{y}_2 + \cdots + c_rE\bar{y}_r \\ &= c_1\mu_1 + c_2\mu_2 + \cdots + c_r\mu_r \\ &= L \end{aligned}$$

and

$$\begin{aligned} \operatorname{Var}\hat{L} &= c_1^2\operatorname{Var}\bar{y}_1 + c_2^2\operatorname{Var}\bar{y}_2 + \cdots + c_r^2\operatorname{Var}\bar{y}_r \\ &= c_1^2\frac{\sigma^2}{n_1} + c_2^2\frac{\sigma^2}{n_2} + \cdots + c_r^2\frac{\sigma^2}{n_r} \end{aligned}$$

$$= \sigma^2 \left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_r^2}{n_r} \right)$$

The one-way model restrictions imply that the $\bar{y}_i$ are independent and normal and, in turn, that $\hat{L}$ is normal. So the standardized version of $\hat{L}$,

$$Z = \frac{\hat{L} - E\hat{L}}{\sqrt{\operatorname{Var} \hat{L}}} = \frac{\hat{L} - L}{\sigma \sqrt{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_r^2}{n_r}}} \tag{7.18}$$

is standard normal. The usual manipulations beginning with this fact would produce an unusable confidence interval for L involving the unknown parameter σ. A way to reason to something of practical importance is to begin not with the variable (7.18), but with

$$T = \frac{\hat{L} - L}{s_{\mathrm{P}} \sqrt{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_r^2}{n_r}}} \tag{7.19}$$

instead. The fact is that under the current assumptions, the variable (7.19) has a t_{n-r} distribution. And this leads in the standard way to the fact that the interval with endpoints

Confidence limits for a linear combination of means

$$\boxed{\hat{L} \pm t s_{\mathrm{P}} \sqrt{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_r^2}{n_r}}} \tag{7.20}$$

can be used as a two-sided confidence interval for L with associated confidence the t_{n-r} probability assigned to the interval between $-t$ and t. Further, a one-sided confidence interval for L can be obtained by using only one of the endpoints in display (7.20) and appropriately adjusting the confidence level upward by reducing the unconfidence in half.

It is worthwhile to verify that the general formula (7.20) reduces to the formula (7.14) if a single c_i is 1 and all others are 0. And if one c_i is 1, one other is -1, and all others are 0, the general formula (7.20) reduces to formula (7.15).

Example 4

Comparing Absorbency Properties for Three Brands of Paper Towels

D. Speltz did some absorbency testing for several brands of paper towels. His study included (among others) a generic brand and two national brands. $n_1 = n_2 = n_3 = 5$ tests were made on towels of each of these $r = 3$ brands, and the numbers of milliliters of water (out of a possible 100) not absorbed out of a

Example 4 (*continued*)

graduated cylinder were recorded. Some summary statistics for the tests on these brands are given in Table 7.8. Plots (not shown here) of the raw absorbency values and residuals indicate no problems with the use of the one-way model in the analysis of the absorbency data.

One question of practical interest is "On average, do the national brands absorb more than the generic brand?" A way of quantifying this is to ask for a two-sided 95% confidence interval for

$$L = \mu_1 - \frac{1}{2}(\mu_2 + \mu_3) \tag{7.21}$$

the difference between the average liquid left by the generic brand and the arithmetic mean of the national brand averages.

With L as in equation (7.21), formula (7.17) shows that

$$\hat{L} = 93.2 - \frac{1}{2}(81.0) - \frac{1}{2}(83.8) = 10.8 \text{ ml}$$

is an estimate of the increased absorbency offered by the national brands. Using the standard deviations given in Table 7.8,

$$s_{\mathrm{P}}^2 = \frac{(5-1)(.8)^2 + (5-1)(.7)^2 + (5-1)(.8)^2}{(5-1)+(5-1)+(5-1)} = .59$$

and thus

$$s_{\mathrm{P}} = \sqrt{.59} = .77 \text{ ml}$$

Now $n - r = 15 - 3 = 12$ degrees of freedom are associated with s_{P}, and the .975 quantile of the t_{12} distribution for use in (7.20) is 2.179. In addition, since $c_1 = 1$, $c_2 = -\frac{1}{2}$, and $c_3 = -\frac{1}{2}$ and all three sample sizes are 5,

$$\sqrt{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \frac{c_3^2}{n_3}} = \sqrt{\frac{1}{5} + \frac{\left(-\frac{1}{2}\right)^2}{5} + \frac{\left(-\frac{1}{2}\right)^2}{5}} = .55$$

Table 7.8
Summary Statistics for Absorbencies of Three Brands of Paper Towels

Brand	i	n_i	$\bar{y}_i$	s_i
Generic	1	5	93.2 ml	.8 ml
National B	2	5	81.0 ml	.7 ml
National V	3	5	83.8 ml	.8 ml

So finally, endpoints for a two-sided 95% confidence interval for L given in equation (7.21) are

$$10.8 \pm 2.179(.77)(.55)$$

that is,

$$10.8 \pm .9$$

i.e.,

$$9.9 \text{ ml} \quad \text{and} \quad 11.7 \text{ ml} \tag{7.22}$$

The interval indicated in display (7.22) shows definitively the substantial advantage in absorbency held by the national brands over the generic, particularly in view of the fact that the amount actually absorbed by the generic brand appears to average only about 6.8 ml (= 100 ml − 93.2 ml).

Example 5

A Confidence Interval for a Main Effect in a 2^2 Factorial Brick Fracture Strength Study

Graves, Lundeen, and Micheli studied the fracture strength properties of brick bars. They included several experimental variables in their study, including both bar composition and heat-treating regimen. Part of their data are given in Table 7.9. Modulus of rupture values under a bending load are given in psi for $n_1 = n_2 = n_3 = n_4 = 3$ bars of $r = 4$ types.

Table 7.9
Modulus of Rupture Measurements for Brick Bars in a 2^2 Factorial Study

i, Bar Type	% Water in Mix	Heat-Treating Regimen	MOR (psi)
1	17	slow cool	4911, 5998, 5676
2	19	slow cool	4387, 5388, 5007
3	17	fast cool	3824, 3140, 3502
4	19	fast cool	4768, 3672, 3242

Notice that the data represented in Table 7.9 have a 2×2 complete factorial structure. Indeed, returning to Section 4.3 (in particular, to Definition 5, page 166),

Example 5 (*continued*)

it becomes clear that the fitted main effect of the factor Heat-Treating Regimen at its slow cool level is

$$\frac{1}{2}(\bar{y}_1 + \bar{y}_2) - \frac{1}{4}(\bar{y}_1 + \bar{y}_2 + \bar{y}_3 + \bar{y}_4) \tag{7.23}$$

But the variable (7.23) is the $\hat{L}$ for the linear combination of mean strengths μ_1, μ_2, μ_3, and μ_4 given by

$$L = \frac{1}{4}\mu_1 + \frac{1}{4}\mu_2 - \frac{1}{4}\mu_3 - \frac{1}{4}\mu_4 \tag{7.24}$$

So subject to the relevance of the "equal variances, normal distributions" description of modulus of rupture for fired brick clay bodies of these four types, formula (7.20) can be applied to develop a precision figure to attach to the fitted effect (7.23).

Table 7.10 gives summary statistics for the data of Table 7.9. Using the values in Table 7.10 leads to

$$\begin{aligned}\hat{L} &= \frac{1}{2}(\bar{y}_1 + \bar{y}_2) - \frac{1}{4}(\bar{y}_1 + \bar{y}_2 + \bar{y}_3 + \bar{y}_4)\\ &= \frac{1}{4}\bar{y}_1 + \frac{1}{4}\bar{y}_2 - \frac{1}{4}\bar{y}_3 - \frac{1}{4}\bar{y}_4\\ &= \frac{1}{4}(5{,}528.3 + 4{,}927.3 - 3{,}488.7 - 3{,}894.0)\\ &= 768.2 \text{ psi}\end{aligned}$$

and

$$\begin{aligned}s_{\mathrm{P}} &= \sqrt{\frac{(3-1)(558.3)^2 + (3-1)(505.2)^2 + (3-1)(342.2)^2 + (3-1)(786.8)^2}{(3-1)+(3-1)+(3-1)+(3-1)}}\\ &= 570.8 \text{ psi}\end{aligned}$$

Table 7.10
Summary Statistics for the Modulus of Rupture Measurements

i, Bar Type	$\bar{y}_i$	s_i
1	5,528.3	558.3
2	4,927.3	505.2
3	3,488.7	342.2
4	3,894.0	786.8

Further, $n - r = 12 - 4 = 8$ degrees of freedom are associated with s_p. Therefore, if one wanted (for example) a two-sided 98% confidence interval for L given in equation (7.24), the necessary .99 quantile of the t_8 distribution is 2.896. Then, since

$$\sqrt{\frac{\left(\frac{1}{4}\right)^2}{3} + \frac{\left(\frac{1}{4}\right)^2}{3} + \frac{\left(-\frac{1}{4}\right)^2}{3} + \frac{\left(-\frac{1}{4}\right)^2}{3}} = .2887$$

a two-sided 98% confidence interval for L has endpoints

$$768.2 \pm 2.896(570.8)(.2887)$$

that is,

$$291.1 \text{ psi} \quad \text{and} \quad 1{,}245.4 \text{ psi} \tag{7.25}$$

Display (7.25) establishes convincingly the effectiveness of a slow cool regimen in increasing MOR. It says that the differences in sample mean MOR values for slow- and fast-cooled bricks are not simply reflecting background variation. In fact, multiplying the endpoints in display (7.25) each by 2 in order to get a confidence interval for

$$2L = \frac{1}{2}(\mu_1 + \mu_2) - \frac{1}{2}(\mu_3 + \mu_4)$$

shows that (when averaged over 17% and 19% water mixtures) the slow, cool regimen seems to offer an increase in MOR in the range from

$$582.2 \text{ psi} \quad \text{to} \quad 2{,}490.8 \text{ psi}$$

7.2.3 Individual and Simultaneous Confidence Levels

This section has introduced a variety of confidence intervals for multisample studies. In a particular application, several of these might be used, perhaps several times each. For example, even in the relatively simple context of Example 4 (the paper towel absorbency study), it would be reasonable to desire confidence intervals for each of

$$\mu_1, \mu_2, \mu_3, \mu_1 - \mu_2, \mu_1 - \mu_3, \mu_2 - \mu_3, \text{ and } \mu_1 - \frac{1}{2}(\mu_2 + \mu_3)$$

Since many confidence statements are often made in multisample studies, it is important to reflect on the meaning of a confidence level and realize that it is attached to one interval at a time. If many 90% confidence intervals are made,

the 90% figure applies to the intervals **individually**. One is "90% sure" of the first interval, *separately* "90% sure" of the second, *separately* "90% sure" of the third, and so on. It is not at all clear how to arrive at a reliability figure for the intervals jointly or simultaneously (i.e., an a priori probability that all the intervals are effective). But it is fairly obvious that it must be less than 90%. That is, the **simultaneous or joint confidence** (the overall reliability figure) to be associated with a group of intervals is generally not easy to determine, but it is typically less (and sometimes much less) than the *individual* confidence level(s) associated with the intervals one at a time.

There are at least three different approaches to be taken once the difference between simultaneous and individual confidence levels is recognized. The most obvious option is to make individual confidence intervals and be careful to interpret them as such (being careful to recognize that as the number of intervals one makes increases, so does the likelihood that among them are one or more intervals that fail to cover the quantities they are meant to locate).

A second way of handling the issue of simultaneous versus individual confidence is to use very large individual confidence levels for the separate intervals and then employ a somewhat crude inequality to find at least a minimum value for the simultaneous confidence associated with an entire group of intervals. That is, if k confidence intervals have associated confidences $\gamma_1, \gamma_2, \ldots, \gamma_k$, the **Bonferroni inequality** says that the simultaneous or joint confidence that all k intervals are effective (say, γ) satisfies

The Bonferroni inequality

$$\boxed{\gamma \geq 1 - \big((1-\gamma_1) + (1-\gamma_2) + \cdots + (1-\gamma_k)\big)} \qquad \textbf{(7.26)}$$

(Basically, this statement says that the joint "unconfidence" associated with k intervals $(1-\gamma)$ is no larger than the sum of the k individual unconfidences. For example, five intervals with individual 99% confidence levels have a joint or simultaneous confidence level of at least 95%.)

The third way of approaching the issue of simultaneous confidence is to develop and employ methods that for some specific, useful set of unknown quantities provide intervals with a known level of simultaneous confidence. There are whole books full of such simultaneous inference methods. In the next section, two of the better known and simplest of these are discussed.

Section 2 Exercises

1. Return to the situation of Exercise 1 of Section 7.1 (and the pressure/density data of Example 1 in Chapter 4).

(a) Individual two-sided confidence intervals for the five different means here would be of the form $\bar{y}_i \pm \Delta$ for a number Δ. If 95% individual confidence is desired, what is Δ? If all five of these intervals are made, what does the Bonferroni inequality guarantee for a minimum joint or simultaneous confidence?

(b) Individual two-sided confidence intervals for the differences in the five different means

would be of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$ for a number Δ. If 95% individual confidence is desired, what is Δ?

(c) Note that if mean density is a linear function of pressure over the range of pressures from 2,000 to 6,000 psi, then $\mu_{4000} - \mu_{2000} = \mu_{6000} - \mu_{4000}$, that is $L = \mu_{6000} - 2\mu_{4000} + \mu_{2000}$ has the value 0. Give 95% two-sided confidence limits for this L. What does your interval indicate about the linearity of the pressure/density relationship?

2. Return to the tilttable testing problem of Exercise 2 of Section 7.1.

(a) Make (individual) 99% two-sided confidence intervals for the four different mean tilttable ratios for the four vans, μ_1, μ_2, μ_3 and μ_4. What does the Bonferroni inequality guarantee for a minimum joint or simultaneous confidence for these four intervals?

(b) Individual confidence intervals for the differences between particular pairs of mean tilttable ratios are of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$ for appropriate values of Δ. Find values of Δ if individual 99% two-sided intervals are desired, first for pairs of means with samples of size 4 and then for pairs of means where one sample size is 4 and the other is 5.

(c) It might be of interest to compare the average of the tilttable ratios for the minivans to that of the full-size vans. Give a 99% two-sided confidence interval for the quantity $\frac{1}{2}(\mu_1 + \mu_2) - \frac{1}{2}(\mu_3 + \mu_4)$.

3. Explain the difference between several intervals having associated 95% individual confidences and having associated 95% simultaneous confidence.

7.3 Two Simultaneous Confidence Interval Methods

As Section 7.2 illustrated, there are several kinds of confidence intervals for means and linear combinations of means that could be made in a multisample study. The issue of individual versus simultaneous confidence was also raised, but only the use of the Bonferroni inequality was given as a means of controlling a simultaneous confidence level.

This section presents two methods for making a number of confidence intervals and in the process maintaining a desired simultaneous confidence. The first of these is due to Pillai and Ramachandran; it provides a guaranteed simultaneous confidence in the estimation of all r individual underlying means. The second is Tukey's method for the simultaneous confidence interval estimation of all differences in pairs of underlying means.

7.3.1 The Pillai-Ramachandran Method

One of the things typically of interest in an r-sample statistical engineering study is the **estimation of all r individual mean responses** $\mu_1, \mu_2, \ldots, \mu_r$. If the individual confidence interval formula of Section 7.2,

$$\bar{y}_i \pm t\frac{s_{\mathrm{P}}}{\sqrt{n_i}} \tag{7.27}$$

is applied r times to estimate these means, the only handle one has on the corresponding simultaneous confidence is given by the Bonferroni inequality (7.26). This fairly crude tool says that if $r = 8$ and one wants 95% simultaneous confidence, individual "unconfidences" of $\frac{.05}{8} = .00625$ (i.e., individual confidences of 99.375%) for the eight different intervals will suffice to produce the desired simultaneous confidence.

Another approach to the setting of simultaneous confidence limits on all of $\mu_1, \mu_2, \ldots, \mu_r$ is to replace t in formula (7.27) with a multiplier derived specifically for the purpose of providing an exact, stated, *simultaneous* confidence in the estimation of all the means. Such multipliers were derived by Pillai and Ramachandran, where either all of the intervals for the r means are two-sided or all are one-sided. That is, Table B.8A gives values of constants k_2^* such that the r two-sided intervals with respective endpoints

P-R two-sided simultaneous 95% confidence limits for r means

$$\bar{y}_i \pm k_2^* \frac{s_{\mathrm{P}}}{\sqrt{n_i}} \tag{7.28}$$

have simultaneous 95% confidence as intervals for the means $\mu_1, \mu_2, \ldots, \mu_r$. (These values k_2^* are in fact .95 quantiles of the *Studentized maximum modulus distributions*.)

Table B.8B gives values of some other constants k_1^* such that if for each i from 1 to r, an interval of the form

P-R one-sided simultaneous 95% confidence intervals for r means

$$\left(-\infty, \bar{y}_i + k_1^* \frac{s_{\mathrm{P}}}{\sqrt{n_i}}\right) \tag{7.29}$$

or of the form

P-R one-sided simultaneous 95% confidence intervals for r means

$$\left(\bar{y}_i - k_1^* \frac{s_{\mathrm{P}}}{\sqrt{n_i}}, \infty\right) \tag{7.30}$$

is made as a confidence interval for μ_i, the simultaneous confidence associated with the collection of r one-sided intervals is 95%. (These k_1^* values are in fact .95 quantiles of the *Studentized extreme deviate distributions*.)

In this book, the use of r intervals of one of the forms (7.28) through (7.30) will be called the P-R method of simultaneous confidence intervals. In order to apply the P-R method, one must find (using interpolation as needed) the entry in Tables B.8 in the column corresponding to the number of samples, r, and the row corresponding to the degrees of freedom associated with s_{P}, namely $\nu = n - r$.

Example 6
(*Example 3 revisited*)

Simultaneous Confidence Intervals for Eight Mean Concrete Compressive Strengths

Consider again the concrete strength study of Armstrong, Babb, and Campen. Recall that tests on $n_i = 3$ specimens of each of $r = 8$ different concrete formulas gave $s_{\mathrm{P}} = 581.6$ psi. Using formula (7.27) and remembering that there are $n - r = 16$ degrees of freedom associated with s_{P}, one has endpoints for 95% two-sided intervals for the formula mean compressive strengths

$$\bar{y}_i \pm 2.120 \frac{581.6}{\sqrt{3}}$$

that is,

$$\bar{y}_i \pm 711.9 \text{ psi} \tag{7.31}$$

In contrast to intervals (7.31), consider the use of formula (7.28) to produce $r = 8$ two-sided intervals for the formula mean strengths with *simultaneous* 95% confidence. Table B.8A shows that $k_2^* = 3.099$ is appropriate in this application. So each concrete formula mean compressive strength, μ_i, should be estimated using

$$\bar{y}_i \pm 3.099 \frac{581.6}{\sqrt{3}}$$

that is,

$$\bar{y}_i \pm 1{,}040.6 \text{ psi} \tag{7.32}$$

Expressions (7.31) and (7.32) provide two-sided intervals for the eight mean compressive strengths. If one-sided intervals of the form $(\#, \infty)$ were desired instead, consulting the t table for the .95 quantile of the t_{16} distribution and use of formula (7.27) shows that the values

$$\bar{y}_i - 1.746 \frac{581.6}{\sqrt{3}}$$

that is,

$$\bar{y}_i - 586.3 \text{ psi} \tag{7.33}$$

are individual 95% lower confidence bounds for the formula mean compressive strengths, μ_i. At the same time, consulting Table B.8B shows that for

Example 6 (*continued*)

simultaneous 95% confidence, use of $k_1^* = 2.779$ in formula (7.30) is appropriate, and the values

$$\bar{y}_i - 2.779\frac{581.6}{\sqrt{3}}$$

that is,

$$\bar{y}_i - 933.2 \text{ psi} \tag{7.34}$$

are simultaneous 95% lower confidence bounds for the formula mean compressive strengths, μ_i.

Comparing intervals (7.31) with intervals (7.32) and bounds (7.33) with bounds (7.34) shows clearly the impact of requiring simultaneous rather than individual confidence. *For a given nominal confidence level, the simultaneous intervals must be bigger (more conservative) than the corresponding individual intervals*.

It is common practice to summarize the information about mean responses gained in a multisample study in a plot of sample means versus sample numbers, enhanced with "error bars" around the sample means to indicate the uncertainty associated with locating the means. There are various conventions for the making of these bars. When looking at such a plot, one typically forms an overall visual impression. Therefore, it is our opinion that error bars derived from the P-R simultaneous confidence limits of display (7.28) are the most sensible representation of what is known about a group of r means. For example, Figure 7.11 is a graphical representation of the eight formula sample mean strengths given in Table 7.7 with $\pm$1,040.6 psi error bars, as indicated by expression (7.32).

When looking at a display like Figure 7.11, it is important to remember that what is represented is the precision of knowledge about the *mean* strengths, rather than any kind of predictions for individual compressive strengths. In this regard, the similarity of the spread of the samples on the side-by-side dot diagram given as Figure 7.1 and the size of the error bars here is coincidental. As sample sizes increase, spreads on displays of individual measurements like Figure 7.1 will tend to stabilize (representing the spreads of the underlying distributions), while the lengths of error bars associated with means will shrink to 0 as increased information gives sharper and sharper evidence about the underlying means.

In any case, Figure 7.11 shows clearly that the information in the data is quite adequate to establish the existence of differences in formula mean compressive strengths.

7.3.2 Tukey's Method

A second set of quantities often of interest in an r-sample study consists of the **differences in all $\frac{r(r-1)}{2}$ pairs of mean responses μ_i and $\mu_{i'}$**. Section 7.2 argued

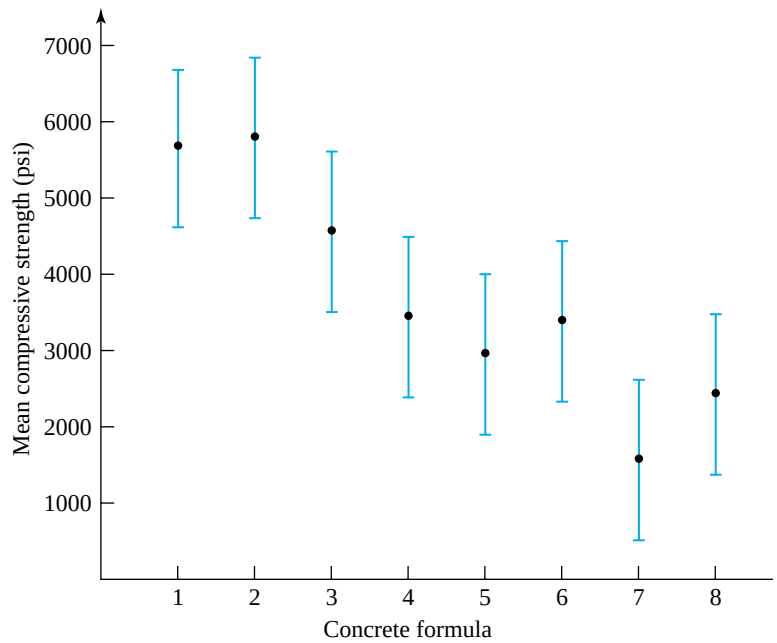

Figure 7.11 Plot of eight sample mean compressive strengths, enhanced with error bars derived from P-R simultaneous confidence limits

that a single difference in mean responses, $\mu_i - \mu_{i'}$, can be estimated using an interval with endpoints

$$\bar{y}_i - \bar{y}_{i'} \pm t s_{\mathrm{P}} \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}} \tag{7.35}$$

where the associated confidence level is an individual one. But if, for example, $r = 8$, there are 28 different two-at-a-time comparisons of underlying means to be considered (μ_1 versus μ_2, μ_1 versus $\mu_3, \ldots, \mu_1$ versus μ_8, μ_2 versus $\mu_3, \ldots$, and μ_7 versus μ_8). If one wishes to guarantee a reasonable simultaneous confidence level for all these comparisons via the crude Bonferroni idea, a huge individual confidence level is required for the intervals (7.35). For example, the Bonferroni inequality requires 99.82% individual confidence for 28 intervals in order to guarantee simultaneous 95% confidence.

A better approach to the setting of simultaneous confidence limits on all of the differences $\mu_i - \mu_{i'}$ is to replace t in formula (7.35) with a multiplier derived specifically for the purpose of providing exact, stated, simultaneous confidence in the estimation of all such differences. J. Tukey first pointed out that it is possible to provide such multipliers using quantiles of the *Studentized range distributions*.

Tables B.9A and B.9B give values of constants q^* such that the set of two-sided intervals with endpoints

Tukey's two-sided simultaneous confidence limits for all differences in r means

$$\bar{y}_i - \bar{y}_{i'} \pm \frac{q^*}{\sqrt{2}} s_{\mathrm{P}} \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}} \tag{7.36}$$

has simultaneous confidence at least 95% or 99% (depending on whether $Q(.95)$ is read from Table B.9A or $Q(.99)$ is read from Table B.9B) in the estimation of all differences $\mu_i - \mu_{i'}$. If all the sample sizes $n_1, n_2, \ldots, n_r$ are equal, the 95% or 99% nominal simultaneous confidence figure is exact, while if the sample sizes are not all equal, the true value is at least as big as the nominal value.

In order to apply Tukey's method, one must find (using interpolation as needed) the column in Tables B.9 corresponding to the number of samples/means to be compared and the row corresponding to the degrees of freedom associated with s_{P}, (namely, $\nu = n - r$).

Example 6 (*continued*)

Consider the making of confidence intervals for differences in formula mean compressive strengths. If a 95% two-sided individual confidence interval is desired for a *specific* difference $\mu_i - \mu_{i'}$, formula (7.35) shows that appropriate endpoints are

$$\bar{y}_i - \bar{y}_{i'} \pm 2.120(581.6)\sqrt{\frac{1}{3} + \frac{1}{3}}$$

that is,

$$\bar{y}_i - \bar{y}_{i'} \pm 1{,}006.7 \text{ psi} \tag{7.37}$$

On the other hand, if one plans to estimate *all* differences in mean compressive strengths with *simultaneous* 95% confidence, by formula (7.36) Tukey two-sided intervals with endpoints

$$\bar{y}_i - \bar{y}_{i'} \pm \frac{4.90}{\sqrt{2}}(581.6)\sqrt{\frac{1}{3} + \frac{1}{3}}$$

that is,

$$\bar{y}_i - \bar{y}_{i'} \pm 1{,}645.4 \text{ psi} \tag{7.38}$$

are in order (4.90 is the value in the $r = 8$ column and $\nu = 16$ row of Table B.9A.)

In keeping with the fact that the confidence level associated with the intervals (7.38) is a simultaneous one, the Tukey intervals are wider than those indicated in formula (7.37).

The plus-or-minus part of display (7.38) is not as big as twice the plus-or-minus part of expression (7.32). Thus, when looking at Figure 7.11, it is not necessary that the error bars around two means fail to overlap before it is safe to judge the corresponding underlying means to be detectably different. Rather, it is only necessary that the two sample means differ by the plus-or-minus part of formula (7.36)—1,645.4 psi in the present situation.

This section has mentioned only two of many existing methods of simultaneous confidence interval estimation for multisample studies. These should serve to indicate the general character of such methods and illustrate the implications of a simultaneous (as opposed to individual) confidence guarantee.

One final word of caution has to do with the theoretical justification of all of the methods found in this section. It is the "equal variances, normal distributions" model that supports these engineering tools. If any real faith is to be put in the nominal confidence levels attached to the P-R and Tukey methods presented here, that faith should be based on evidence (typically gathered, at least to some extent, as illustrated in Section 7.1) that the standard one-way normal model is a sensible description of a physical situation.

Section 3 Exercises

1. Return to the situation of Exercises 1 of Sections 7.1 and 7.2 (and the pressure/density data of Example 1 in Chapter 4).

(a) Using the P-R method, what Δ can be employed to make two-sided intervals of the form $\bar{y}_i \pm \Delta$ for all five mean densities, possessing simultaneous 95% confidence? How does this Δ compare to the one computed in part (a) of Exercise 1 of Section 7.2?

(b) Using the Tukey method, what Δ can be employed to make two-sided intervals of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$ for all differences in the five mean densities, possessing simultaneous 95% confidence? How does this Δ compare to the one computed in part (b) of Exercise 1 of Section 7.2?

2. Return to the tilttable study of Exercises 2 of Sections 7.1 and 7.2.

(a) Use the P-R method of simultaneous confidence intervals and make simultaneous 95% two-sided confidence intervals for the four mean tilttable ratios.

(b) Simultaneous confidence intervals for the differences in all pairs of mean tilttable ratios are of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$. Find appropriate values of Δ if simultaneous 99% two-sided intervals are desired, first for pairs of means with samples of size 4 and then for pairs of means where one sample size is 4 and the other is 5. How do these compare to the intervals you found in part (b) of Exercise 2 of Section 7.2? Why is it reasonable that the Δ's should be related in this way?

7.4 One-Way Analysis of Variance (ANOVA)

This book's approach to inference in multisample studies has to this point been completely "interval-oriented." But there are also significance-testing methods that are appropriate to the multiple-sample context. This section considers some of these and the issues raised by their introduction. It begins with some general comments regarding significance testing in r-sample studies. Then the one-way analysis of variance (ANOVA) test for the equality of r means is discussed. Next, the one-way ANOVA table and the organization and intuition that it provides are presented. Finally, there is a brief look at the one-way random effects model and ANOVA-based inference for its parameters.

7.4.1 Significance Testing and Multisample Studies

Just as there are many quantities one might want to estimate in a multisample study, there are potentially many issues of statistical significance to be judged. For instance, one might desire p-values for hypotheses like

$$\mathrm{H}_0\colon \mu_3 = 7 \tag{7.39}$$

$$\mathrm{H}_0\colon \mu_3 - \mu_7 = 0 \tag{7.40}$$

$$\mathrm{H}_0\colon \mu_1 - \frac{1}{2}(\mu_2 + \mu_3) = 0 \tag{7.41}$$

The confidence interval methods discussed in Section 7.2 have their significance-testing analogs for treating hypotheses that, like all three of these, involve linear combinations of the means $\mu_1, \mu_2, \ldots, \mu_r$.

In general (under the standard one-way model), if

$$L = c_1\mu_1 + c_2\mu_2 + \cdots + c_r\mu_r$$

the hypothesis

$$\mathrm{H}_0\colon L = \# \tag{7.42}$$

can be tested using the test statistic

$$T = \frac{\hat{L} - \#}{s_{\mathrm{P}}\sqrt{\dfrac{c_1^2}{n_1} + \dfrac{c_2^2}{n_2} + \cdots + \dfrac{c_r^2}{n_r}}} \tag{7.43}$$

and a t_{n-r} reference distribution. This fact specializes to cover hypotheses of types (7.39) to (7.41) by appropriate choice of the c_i and #.

But the significance-testing method most often associated with the one-way normal model is not for hypotheses of the type (7.42). Instead, the most common method concerns the hypothesis that all r underlying means have the same value. In symbols, this is

$$\mathrm{H}_0\colon \mu_1 = \mu_2 = \cdots = \mu_r \tag{7.44}$$

Given that one is working under the assumptions of the one-way model to begin with, hypothesis (7.44) amounts to a statement that all r underlying distributions are essentially the same—or "There are no differences between treatments."

Hypothesis (7.44) can be thought of in terms of the simultaneous equality of $\frac{r(r-1)}{2}$ pairs of means—that is, as equivalent to the statement that simultaneously

$$\mu_1 - \mu_2 = 0, \quad \mu_1 - \mu_3 = 0, \quad \ldots, \quad \mu_1 - \mu_r = 0,$$
$$\mu_2 - \mu_3 = 0, \quad \ldots, \quad \text{and} \quad \mu_{r-1} - \mu_r = 0$$

And this fact should remind the reader of the ideas about simultaneous confidence intervals from the previous section (specifically, Tukey's method). In fact, one way of judging the statistical significance of an r-sample data set in reference to hypothesis (7.44) is to apply Tukey's method of simultaneous interval estimation and note whether or not all the intervals for differences in means include 0. If they all do, the associated p-value is larger than 1 minus the simultaneous confidence level. If not all of the intervals include 0, the associated p-value is smaller than 1 minus the simultaneous confidence level. (If simultaneous 95% intervals all include 0, no differences between means are definitively established, and the corresponding p-value exceeds .05.)

We admit a bias toward estimation over testing per se. A consequence of this bias is a fondness for deriving a rough idea of a p-value for hypothesis (7.44) as a byproduct of Tukey's method. But a most famous significance-testing method for hypothesis (7.44) also deserves discussion: the one-way analysis of variance test. (At this point it may seem strange that a test about means has a name apparently emphasizing variance. The motivation for this jargon is that the test is associated with a very helpful way of thinking about partitioning the overall variability that is encountered in a response variable.)

7.4.2 The One-Way ANOVA *F* Test

The standard method of testing the hypothesis (7.44)

$$\mathrm{H}_0\colon \mu_1 = \mu_2 = \cdots = \mu_r$$

of no differences among r means against

$$\mathrm{H}_a\colon \text{not } \mathrm{H}_0 \tag{7.45}$$

is based essentially on a comparison of a measure of variability among the sample means to the pooled sample variance, s_p^2. In order to fully describe this method some additional notational conventions are needed.

Repeatedly in the balance of this book, it will be convenient to have symbols for the summary measures of Section 3.3 (sample means and variances) applied to the data from multisample studies, *ignoring the fact that there are r different samples involved*. Already the unsubscripted letter n has been used to stand for $n_1 + n_2 + \cdots + n_r$, the number of observations in hand ignoring the fact that r samples are involved. This kind of convention will now be formally extended to include statistics calculated from the n responses. For emphasis, this will be stated in definition form.

Definition 3
(A Notational Convention for Multisample Studies)

In multisample studies, symbols for sample sizes and sample statistics appearing without subscript indices or dots will be understood to be calculated from all responses in hand, obtained by combining all samples.

So n will stand for the total number of data points (even in an r-sample study), $\bar{y}$ for the grand sample average of response y, and s^2 for a grand sample variance calculated completely ignoring sample boundaries.

For present purposes (of writing down a test statistic for testing hypothesis (7.44)), one needs to make use of $\bar{y}$, the grand sample average. It is important to recognize that $\bar{y}$ and

The (unweighted) average of r sample means

$$\boxed{\bar{y}_{\cdot} = \frac{1}{r}\sum_{i=1}^{r} \bar{y}_i} \qquad \textbf{(7.46)}$$

are not necessarily the same unless all sample sizes are equal. That is, when sample sizes vary, $\bar{y}$ is the (unweighted) arithmetic average of the raw data values y_{ij} but is a weighted average of the sample means $\bar{y}_i$. On the other hand, $\bar{y}_{\cdot}$ is the (unweighted) arithmetic mean of the sample means $\bar{y}_i$ but is a weighted average of the raw data values y_{ij}. For example, in the simple case that $r = 2$, $n_1 = 2$, and $n_2 = 3$,

$$\bar{y} = \frac{1}{5}(y_{11} + y_{12} + y_{21} + y_{22} + y_{23}) = \frac{2}{5}\bar{y}_1 + \frac{3}{5}\bar{y}_2$$

while

$$\bar{y}_{\cdot} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{4}y_{11} + \frac{1}{4}y_{12} + \frac{1}{6}y_{21} + \frac{1}{6}y_{22} + \frac{1}{6}y_{23}$$

and, in general, $\bar{y}$ and $\bar{y}_{\cdot}$ will not be the same.

Now, under the hypothesis (7.44), that $\mu_1 = \mu_2 = \cdots = \mu_r$, $\bar{y}$ is a natural estimate of the common mean. (All underlying distributions are the same, so the data in hand are reasonably thought of not as r different samples, but rather as a single sample of size n.) Then the differences $\bar{y}_i - \bar{y}$ are indicators of possible differences among the μ_i. It is convenient to summarize the size of these differences $\bar{y}_i - \bar{y}$ in terms of a kind of total of their squares—namely,

$$\sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2 \tag{7.47}$$

One can think of statistic (7.47) either as a weighted sum of the quantities $(\bar{y}_i - \bar{y})^2$ or as an unweighted sum, where there is a term in the sum for each raw data point and therefore n_i of the type $(\bar{y}_i - \bar{y})^2$. The quantity (7.47) is a measure of the **between-sample variation** in the data. For a given set of sample sizes, the larger it is, the more variation there is between the sample means $\bar{y}_i$.

In order to produce a test statistic for hypothesis (7.44), one simply divides the measure (7.47) by $(r-1)s_{\mathrm{P}}^2$, giving

One-way ANOVA test statistic for equality of r means

$$\boxed{F = \frac{\dfrac{1}{r-1}\displaystyle\sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2}{s_{\mathrm{P}}^2}} \tag{7.48}$$

The fact is that if $\mathrm{H}_0\colon \mu_1 = \mu_2 = \cdots = \mu_r$ is true, the one-way model assumptions imply that this statistic has an $F_{r-1,\,n-r}$ distribution. So the hypothesis of equality of r means can be tested using the statistic in equation (7.48) with an $F_{r-1,\,n-r}$ reference distribution, where large observed values of F are taken as evidence against H_0 in favor of H_{a}: not H_0.

Example 7
(*Example 1 revisited*)

Returning again to the concrete compressive strength study of Armstrong, Babb, and Campen, $\bar{y} = 3{,}693.6$ and the 8 sample means $\bar{y}_i$ have differences from this value given in Table 7.11.

Then since each $n_i = 3$, in this situation,

$$\begin{aligned}\sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2 &= 3(1{,}941.7)^2 + 3(2{,}059.7)^2 + \cdots \\ &\quad + 3(-2{,}142.3)^2 + 3(-1{,}302.9)^2 \\ &= 47{,}360{,}780 \text{ (psi)}^2\end{aligned}$$

Example 7 (*continued*)

Table 7.11
Sample Means and Their Deviations from $\bar{y}$ in the Concrete Strength Study

i, Formula	$\bar{y}_i$	$\bar{y}_i - \bar{y}$
1	5,635.3	1,941.7
2	5,753.3	2,059.7
3	4,527.3	833.7
4	3,442.3	−251.3
5	2,923.7	−769.9
6	3,324.7	−368.9
7	1,551.3	−2,142.3
8	2,390.7	−1,302.9

In order to use this figure to judge statistical significance, one standardizes via equation (7.48) to arrive at the observed value of the test statistic

$$f = \frac{\frac{1}{8-1}(47{,}360{,}780)}{(581.6)^2} = 20.0$$

It is easy to verify from Tables B.6 that 20.0 is larger than the .999 quantile of the $F_{7,16}$ distribution. So

$$p\text{-value} = P[\text{an } F_{7,16} \text{ random variable} \geq 20.0] < .001$$

That is, the data provide overwhelming evidence that $\mu_1, \mu_2, \ldots, \mu_8$ are not all equal.

For pedagogical reasons, the one-way ANOVA test has been presented after discussing interval-oriented methods of inference for r-sample studies. But if it is to be used in applications, the testing method typically belongs chronologically before estimation. That is, the ANOVA test can serve as a *screening device* to determine whether the data in hand are adequate to differentiate conclusively between the means, or whether more data are needed.

7.4.3 The One-Way ANOVA Identity and Table

Associated with the ANOVA test statistic is some strong intuition related to the partitioning of observed variability. This is related to an algebraic identity that is stated here in the form of a proposition.

Proposition 1

For any n numbers y_{ij}

One-way ANOVA identity

$$(n-1)s^2 = \sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2 + (n-r)s_{\mathrm{P}}^2 \tag{7.49}$$

or in other symbols,

A second statement of the one-way ANOVA identity

$$\sum_{i,j}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \tag{7.50}$$

Proposition 1 should begin to shed some light on the phrase "analysis of variance." It says that an overall measure of variability in the response y, namely,

$$(n-1)s^2 = \sum_{i,j}(y_{ij} - \bar{y})^2$$

can be partitioned or decomposed algebraically into two parts. One,

$$\sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2$$

can be thought of as measuring variation between the samples or "treatments," and the other,

$$(n-r)s_{\mathrm{P}}^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

measures variation within the samples (and in fact consists of the sum of the squared residuals). The F statistic (7.48), developed for testing $\mathrm{H}_0\colon \mu_1 = \mu_2 = \cdots = \mu_r$, has a numerator related to the first of these and a denominator related to the second. So using the ANOVA F statistic amounts to a kind of <u>an</u>alyzing <u>o</u>f the raw <u>va</u>riability in y.

In recognition of their prominence in the calculation of the one-way ANOVA F statistic and their usefulness as descriptive statistics in their own right, the three sums (of squares) appearing in formulas (7.49) and (7.50) are usually given special names and shorthand. These are stated here in definition form.

Definition 4

In a multisample study, $(n-1)s^2$, the sum of squared differences between the raw data values and the grand sample mean, will be called the **total sum of squares** and denoted as *SSTot*.

Definition 5

In an unstructured multisample study, $\sum n_i(\bar{y}_i - \bar{y})^2$ will be called the **treatment sum of squares** and denoted as *SSTr*.

Definition 6

In a multisample study, the sum of squared residuals, $\sum(y - \hat{y})^2$ (which is $(n-r)s_{\mathrm{P}}^2$ in the unstructured situation) will be called the **error sum of squares** and denoted as *SSE*.

In the new notation introduced in these definitions, Proposition 1 states that in an unstructured multisample context,

A third statement of the one-way ANOVA identity

$$\boxed{SSTot = SSTr + SSE} \tag{7.51}$$

Partially as a means of organizing calculation of the F statistic given in formula (7.48) and partially because it reinforces and extends the variance partitioning insight provided by formulas (7.49), (7.50), and (7.51), it is useful to make an **ANOVA table**. There are many forms of ANOVA tables corresponding to various multisample analyses. The form most relevant to the present situation is given in symbolic form as Table 7.12.

The column headings in Table 7.12 are Source (of variation), Sum of Squares (corresponding to the source), degrees of freedom (corresponding to the source), Mean Square (corresponding to the source), and F (for testing the significance of the source in contributing to the overall observed variability). The entries in the Source column of the table are shown here as being Treatments, Error, and Total. But the name Treatments is sometimes replaced by Between (Samples), and the

Table 7.12
General Form of the One-Way ANOVA Table

ANOVA Table (for testing H_0: $\mu_1 = \mu_2 = \cdots = \mu_r$)				
Source	*SS*	*df*	*MS*	*F*
Treatments	*SSTr*	$r-1$	$SSTr/(r-1)$	$MSTr/MSE$
Error	*SSE*	$n-r$	$SSE/(n-r)$	
Total	*SSTot*	$n-1$		

name Error is sometimes replaced by Within (Samples) or Residual. The first two entries in the *SS* column must sum to the third, as indicated in equation (7.51). Similarly, the Treatments and Error degrees of freedom add to the Total degrees of freedom, $(n-1)$. Notice that the entries in the df column are those attached to the numerator and denominator, respectively, of the test statistic in equation (7.48). The ratios of sums of squares to degrees of freedom are called mean squares, here the mean square for treatments (*MSTr*) and the mean square for error (*MSE*). Verify that in the present context, $MSE = s_{\mathrm{P}}^2$ and *MSTr* is the numerator of the F statistic given in equation (7.48). So the single ratio appearing in the F column is the observed value of F for testing $\mathrm{H}_0\colon \mu_1 = \mu_2 = \cdots = \mu_r$.

Example 7 *(continued)*

Consider once more the concrete strength study. It is possible to return to the raw data given in Table 7.1 and find that $\bar{y} = 3{,}693.6$, so

$$\begin{aligned} SSTot &= (n-1)s^2 \\ &= (5{,}800 - 3{,}693.6)^2 + (4{,}598 - 3{,}693.6)^2 + (6{,}508 - 3{,}693.6)^2 \\ &\quad + \cdots + (2{,}631 - 3{,}693.6)^2 + (2{,}490 - 3{,}693.6)^2 \\ &= 52{,}772{,}190 \text{ (psi)}^2 \end{aligned}$$

Further, as in Section 7.1, $s_{\mathrm{P}}^2 = 338{,}213.1 \text{ (psi)}^2$ and $n - r = 16$, so

$$SSE = (n-r)s_{\mathrm{P}}^2 = 5{,}411{,}410 \text{ (psi)}^2$$

And from earlier in this section,

$$SSTr = \sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2 = 47{,}360{,}780$$

Then, plugging these and appropriate degrees of freedom values into the general form of the one-way ANOVA table produces the table for the concrete compressive strength study, presented here as Table 7.13.

Table 7.13
One-Way ANOVA Table for the Concrete Strength Study

ANOVA Table (for testing $\mathrm{H}_0\colon \mu_1 = \mu_2 = \cdots = \mu_8$)				
Source	*SS*	*df*	*MS*	*F*
Treatments	47,360,780	7	6,765,826	20.0
Error	5,411,410	16	338,213	
Total	52,772,190	23		

Example 7 *(continued)*

Notice that, as promised by the one-way ANOVA identity, the sum of the treatment and error sums of squares is the total sum of squares. Also, Table 7.13 serves as a helpful summary of the testing process, showing at a glance the observed value of F, the appropriate degrees of freedom, and $s_{\mathrm{P}}^2 = MSE$.

The computations here are by no means impossible to do "by hand." But the most sensible way to handle them is to employ a statistical package. Printout 1 shows the results of using MINTAB to create an ANOVA table. (The routine under MINITAB's "Stat/ANOVA/One-way" menu was used.)

Printout 1 ANOVA Table for a One-Way Analysis of the Concrete Strength Data

```
One-way Analysis of Variance

Analysis of Variance for strength
Source     DF        SS        MS        F        P
formula     7  47360781   6765826    20.00    0.000
Error      16   5411409    338213
Total      23  52772190
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level       N      Mean     StDev  -----+---------+---------+---------+-
1           3    5635.3     965.6                           (---*----)
2           3    5753.3     432.3                            (---*---)
3           3    4527.3     509.9                     (---*----)
4           3    3442.3     356.4              (----*---)
5           3    2923.7     852.9           (---*----)
6           3    3324.7     353.5             (----*---)
7           3    1551.3     505.5  (----*---)
8           3    2390.7     302.5       (----*---)
                                   -----+---------+---------+---------+-
Pooled StDev =    581.6              1600      3200      4800      6400
```

You may recall having used a breakdown of a "raw variation in the data" earlier in this text (namely, in Chapter 4). In fact, there is a direct connection between the present discussion and the discussion and use of R^2 in Sections 4.1, 4.2, and 4.3. (See Definition 3 in Chapter 4 and its use throughout those three sections.) In the present notation, the coefficient of determination defined as a descriptive measure in Section 4.1 is

The coefficient of determination in general sums of squares notation

$$R^2 = \frac{SSTot - SSE}{SSTot} \tag{7.52}$$

(Fitted values for the present situation are the sample means and SSE is the sum of squared residuals here, just as it was earlier.) Expression (7.52) is a perfectly general recasting of the definition of R^2 into "SS" notation. In the present one-way context, the one-way identity (7.51) makes it possible to rewrite the numerator of

the right-hand side of formula (7.52) as *SSTr*. So in an unstructured r-sample study (where the fitted values are the sample means)

The coefficient of determination in a one-way analysis

$$\boxed{R^2 = \frac{SSTr}{SSTot}} \tag{7.53}$$

That is, the first entry in the *SS* column of the ANOVA table divided by the total entry of that column can be taken as "the fraction of the raw variability in y accounted for in the process of fitting the equation $y_{ij} \approx \mu_i$ to the data."

Example 7 (*continued*)

In the concrete compressive strength study, a look at Table 7.13 and equation (7.53) shows that

$$R^2 = \frac{SSTr}{SSTot} = \frac{47{,}360{,}780}{52{,}772{,}190} = .897$$

That is, another way to describe these data is to say that differences between concrete formulas account for nearly 90% of the raw variability observed in compressive strength.

So the ANOVA breakdown of variability not only facilitates the testing of H_0: $\mu_1 = \mu_2 = \cdots = \mu_r$ but it also makes direct connection with the earlier descriptive analyses of what part of the raw variability is accounted for in fitting a model equation.

7.4.4 Random Effects Models and Analyses (*Optional*)

On occasion, the r particular conditions leading to the r samples in a multisample study are not so much of interest in and of themselves, as they are of interest as representing a wider set of conditions. For example, in the nondestructive testing of critical metal parts, if $n_i = 3$ mechanical wave travel-time measurements are made on each of $r = 6$ parts selected from a large lot of such parts, the six particular parts are of interest primarily as they provide information on the whole lot.

In such situations, rather than focusing formal inference on the particular r means actually represented in the data (i.e., $\mu_1, \mu_2, \ldots, \mu_r$), it is more natural to make inferences about *the mechanism that generates the means* μ_i. And it is possible, under appropriate model assumptions, to use the ANOVA ideas introduced in this section in this way. The balance of this section is concerned with how this is done.

The most commonly used probability model for the analysis of r-sample data, where the r conditions actually studied represent a much wider set of conditions

of interest, is a variation on the one-way model of this chapter called the **one-way random effects model**. It is built on the usual one-way assumptions that

$$y_{ij} = \mu_i + \epsilon_{ij} \tag{7.54}$$

Random effects model assumptions

where the ϵ_{ij} are iid normal $(0, \sigma^2)$ random variables. But it doesn't treat the means μ_i as parameters/unknown constants. Instead, the means $\mu_1, \mu_2, \ldots, \mu_r$ are treated as (unobservable) *random variables* independent of the ϵ_{ij}'s and themselves iid according to some normal distribution with an unknown mean μ and unknown variance σ_τ^2. The random variables μ_i are now called **random (treatment) effects**, and the variances σ^2 and σ_τ^2 are called **variance components**. The objects of formal inference become μ (the mean of the random effects) and the two variance components σ^2 and σ_τ^2.

Example 8

Magnesium Contents at Different Locations on an Alloy Rod and the Random Effects Model

Youden's *Experimentation and Measurement* contains an interesting data set concerned with the magnesium contents of different parts of a long rod of magnesium alloy. A single ingot had been drawn into a rod of about 100 m in length, with a square cross section about 4.5 cm on a side. $r = 5$ flat test pieces 1.2 cm thick were cut from the rod (after it had been cut into 100 bars and 5 of these randomly selected to represent the rod), and multiple magnesium determinations were made on the 5 specimens. $n_i = 10$ of the resulting measurements for each specimen are given in Table 7.14. (There were actually other observations made not listed in Table 7.14. And some additional structure in Youden's original data

Table 7.14
Measured Magnesium Contents for Five Alloy Specimens

Specimen 1	Specimen 2	Specimen 3	Specimen 4	Specimen 5
76	69	73	73	70
71	71	69	75	66
70	68	68	69	68
67	71	69	72	68
71	66	70	69	64
65	68	70	69	70
67	71	65	72	69
71	69	67	63	67
66	70	67	69	69
68	68	64	69	67
$\bar{y}_1 = 69.2$	$\bar{y}_2 = 69.1$	$\bar{y}_3 = 68.2$	$\bar{y}_4 = 70.0$	$\bar{y}_5 = 67.8$
$s_1 = 3.3$	$s_2 = 1.7$	$s_3 = 2.6$	$s_4 = 3.3$	$s_5 = 1.9$

will also be ignored for present purposes.) The units of measurement in Table 7.14 are .001% magnesium.

In this example, on the order of 8,300 test specimens could be cut from the 100 m rod. The purpose of creating the rod was to provide secondary standards for field calibration of chemical analysis instruments. That is, laboratories purchasing pieces of this rod could use them as being of "known" magnesium content to calibrate their instruments. As such, the practical issues at stake here are not primarily how the $r = 5$ particular test specimens analyzed compare. Rather, the issues are what the overall magnesium content is and whether or not the rod is consistent enough in content along its length to be of any use as a calibration tool. A random effects model and inference for the mean effect μ and the variance components are quite natural in this situation. Here, σ_τ^2 represents the variation in magnesium content among the potentially 8,300 different test specimens, and σ^2 represents measurement error plus variation in magnesium content within the 1.2 cm thick specimens, test location to test location.

When all of the r sample sizes n_i are the same (say, equal to m), it turns out to be quite easy to do some diagnostic checking of the aptness of the normal random effects model (7.54) and make subsequent inferences about μ, σ^2, and σ_τ^2. So this discussion will be limited to cases of equal sample sizes.

As far as investigation of the reasonableness of the model restrictions on the distribution of the μ_i and inference for μ are concerned, a key observation is that

$$\bar{y}_i = \frac{1}{m}\sum_{j=1}^{m}(\mu_i + \epsilon_{ij}) = \mu_i + \bar{\epsilon}_i$$

(where, of course, $\bar{\epsilon}_i$ is the sample mean of $\epsilon_{i1}, \ldots, \epsilon_{im}$). Under the random effects model (7.54), these $\bar{y}_i = \mu_i + \bar{\epsilon}_i$ are iid normal variables with mean μ and variance $\sigma_\tau^2 + \sigma^2/m$. So normal-plotting the $\bar{y}_i$ is a sensible method of at least indirectly investigating the appropriateness of the normal distribution assumption for the μ_i. In addition, the fact that the model says the $\bar{y}_i$ are independent normal variables with mean μ and a common variance suggests that the small-sample inference methods from Section 6.3 should simply be applied to the sample means $\bar{y}_i$ in order to do inference for μ. In doing so, the "sample size" involved is the number of $\bar{y}_i$'s—namely, r.

Example 8 *(continued)*

For the magnesium alloy rod, the $r = 5$ sample means are in Table 7.14. Figure 7.12 gives a normal plot of those five values, showing no obvious problems with a normal random effects model for specimen magnesium contents.

To find a 95% two-sided confidence interval for μ, we calculate as follows (treating the five values $\bar{y}_i$ as "observations"). The sample mean (of $\bar{y}_i$'s) is

$$\bar{y}_{.} = \frac{1}{5}\sum_{i=1}^{5}\bar{y}_i = 68.86$$

Example 8
(continued)

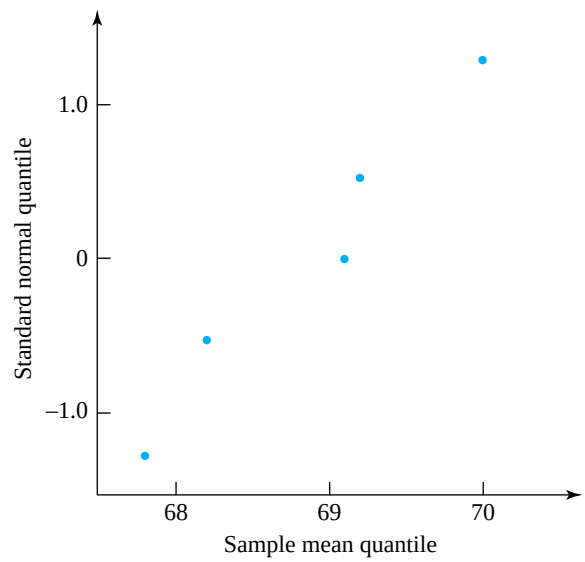

Figure 7.12 Normal plot of five specimen mean magnesium contents

and the sample variance (of $\bar{y}_i$'s) is

$$\frac{1}{5-1}\sum_{i=1}^{5}(\bar{y}_i - \bar{y}_.)^2 = .76$$

so that the sample standard deviation (of $\bar{y}_i$'s) is

$$\sqrt{\frac{1}{5-1}\sum_{i=1}^{5}(\bar{y}_i - \bar{y}_.)^2} = .87$$

Applying the small-sample confidence interval formula for a single mean from Section 6.3 (since $r - 1 = 4$ degrees of freedom are appropriate), a two-sided 95% confidence for μ has endpoints

$$68.86 \pm 2.776\frac{.87}{\sqrt{5}}$$

that is,

$$67.78 \times 10^{-3}\% \quad \text{and} \quad 69.94 \times 10^{-3}\%$$

These limits provide a notion of precision appropriate for the number $68.86 \times 10^{-3}\%$ as an estimate of the rod's mean magnesium content.

It is useful to write out in symbols what was just done to get a confidence interval for μ. That is, a sample variance of $\bar{y}_i$'s was used. This is

$$\frac{1}{r-1}\sum_{i=1}^{r}(\bar{y}_i - \bar{y}_{\cdot})^2 = \frac{1}{m(r-1)}\sum_{i=1}^{r} m(\bar{y}_i - \bar{y})^2 = \frac{1}{m(r-1)}SSTr = \frac{1}{m}MSTr$$

because all n_i are m and $\bar{y}_{\cdot} = \bar{y}$ in this case. But this means that under the assumptions of the one-way normal random effects model, a two-sided confidence interval for μ has endpoints

Balanced data confidence limits for the overall mean in the one-way random effects model

$$\boxed{\bar{y}_{\cdot} \pm t\sqrt{\frac{MSTr}{mr}}} \tag{7.55}$$

where t is such that the probability the t_{r-1} distribution assigns to the interval between $-t$ and t is the desired confidence. One-sided intervals are obtained in the usual way, by employing only one of the endpoints in display (7.55).

7.4.5 ANOVA-Based Inference for Variance Components (*Optional*)

Turning attention to the variance components in the random effects model (7.54), first note that as far as diagnostic checking of the assumption that the ϵ_{ij} are iid normal variables and inference for $\sigma^2 = \text{Var}\,\epsilon_{ij}$ are concerned, all of the methods of Section 7.1 remain in force. If one thinks of holding the μ_i fixed in formula (7.54), it is clear that (conditional on the μ_i) the random effects model treats the r samples as random samples from normal distributions with a common variance. So before doing inference for σ^2 (or σ_τ^2 for that matter) via usual normal theory formulas, it is advisable to do the kind of sample-by-sample normal-plotting and plotting of residuals illustrated in Section 7.1. And if it is then plausible that the ϵ_{ij} are iid normal $(0, \sigma^2)$ variables, formula (7.10) of Section 7.1 can be used to produce a confidence interval for σ^2, and significance testing for σ^2 can be done based on the fact that $r(m-1)s_{\text{P}}^2/\sigma^2$ has a $\chi^2_{r(m-1)}$ distribution.

Inference for σ_τ^2 borrows from things already discussed but also provides a new wrinkle or two of its own. First, significance testing for

$$\text{H}_0\colon \sigma_\tau^2 = 0 \tag{7.56}$$

is made possible by the observation that if H_0 is true, then (just as when $\text{H}_0\colon \mu_1 = \mu_2 = \cdots = \mu_r$ in the case where the μ_i are not random effects but fixed parameters) the $n = mr$ observations are all coming from a single normal distribution. So

ANOVA test statistic for $H_0\colon \sigma_\tau^2 = 0$ in the one-way random effects model

$$\boxed{F = \frac{MSTr}{MSE}} \tag{7.57}$$

has an $F_{r-1,\,n-r}$ distribution under the assumptions of the random effects model (7.54) when the null hypothesis (7.56) holds. Thus, the same one-way ANOVA F test used to test $\mathrm{H}_0\colon \mu_1 = \mu_2 = \cdots = \mu_r$ when the means μ_i are considered fixed parameters can also be used to test $\mathrm{H}_0\colon \sigma_\tau^2 = 0$ under the assumptions of the random effects model.

As far as estimation goes, it doesn't turn out to be possible to give a simple confidence interval formula for σ_τ^2 directly. But what can be done in a straightforward fashion is to give both a natural ANOVA-based single-number estimate of σ_τ^2 and a confidence interval for the ratio σ_τ^2/σ^2. To accomplish the first of these, consider the mean values of random variables $MSTr$ and MSE $(= s_{\mathrm{P}}^2)$ under the assumptions of the random effects model. Not too surprisingly,

$$E(MSE) = E s_{\mathrm{P}}^2 = \sigma^2 \tag{7.58}$$

(After all, s_{P}^2 has been used to approximate σ^2. That the "center" of the probability distribution of s_{P}^2 is σ^2 should therefore seem only reassuring.) And further,

$$E(MSTr) = \sigma^2 + m\sigma_\tau^2 \tag{7.59}$$

Then, from equations (7.58) and (7.59),

$$\frac{1}{m}\big(E(MSTr) - E(MSE)\big) = \sigma_\tau^2$$

or

$$E\,\frac{1}{m}(MSTr - MSE) = \sigma_\tau^2 \tag{7.60}$$

So equation (7.60) suggests that the random variable

$$\frac{1}{m}(MSTr - MSE) \tag{7.61}$$

is one whose distribution is centered about the variance component σ_τ^2 and thus is a natural ANOVA-based estimator of σ_τ^2. The variable in display (7.61) is potentially negative. When that occurs, common practice is to estimate σ_τ^2 by 0. So the variable actually used to estimate σ_τ^2 is

An ANOVA-based estimator of the treatment variance

$$\boxed{\hat{\sigma}_\tau^2 = \max\left(0,\ \frac{1}{m}(MSTr - MSE)\right)} \tag{7.62}$$

Facts (7.58) and (7.60), which motivate this method of estimating σ_τ^2, are important enough that they are often included as entries in an Expected Mean Square column added to the one-way ANOVA table when testing $\mathrm{H}_0\colon \sigma_\tau^2 = 0$.

Although no elementary confidence interval for σ_τ^2 is known, it is possible to give one for the ratio σ_τ^2/σ^2. A basic probability fact is that under the assumptions of the random effects model (7.54),

$$F = \frac{\dfrac{MSTr}{\sigma^2 + m\sigma_\tau^2}}{\dfrac{MSE}{\sigma^2}}$$

has an $F_{r-1,\,n-r}$ distribution. Some algebraic manipulations beginning from this fact show that the interval with endpoints

Confidence limits for σ_τ^2/σ^2 in the one-way random effects model

$$\boxed{\frac{1}{m}\left(\frac{MSTr}{U \cdot MSE} - 1\right) \quad \text{and} \quad \frac{1}{m}\left(\frac{MSTr}{L \cdot MSE} - 1\right)} \tag{7.63}$$

can be used as a two-sided confidence interval for σ_τ^2/σ^2, where the associated confidence is the probability the $F_{r-1,\,n-r}$ distribution assigns to the interval (L, U). One-sided intervals for σ_τ^2/σ^2 can be had by using only one of the endpoints and choosing L or U such that the probability assigned by the $F_{r-1,\,n-r}$ distribution to (L, ∞) or $(0, U)$ is the desired confidence.

Example 8 *(continued)*

Consider again the measured magnesium contents for specimens cut from the 100 m alloy rod. Some normal plotting shows the "single variance normal ϵ_{ij}" part of the model assumptions (7.54) to be at least not obviously flawed. Sample-by-sample normal plots show fair linearity (at least after allowing for the discreteness introduced in the data by the measurement scale used), except perhaps for sample 4, with its five identical values. The five sample standard deviations are roughly of the same order of magnitude, and the normal plot of residuals in Figure 7.13 is pleasantly linear. So it is sensible to consider formal inference for σ^2 and σ_τ^2 based on the normal theory model.

Table 7.15 is an ANOVA table for the data of Table 7.14. From Table 7.15, the p-value for testing $\mathrm{H}_0\colon \sigma_\tau^2 = 0$ is the $F_{4,45}$ probability to the right of 1.10. According to Tables B.6, this is larger than .25, giving very weak evidence of detectable variation between specimen mean magnesium contents.

The *EMS* column in Table 7.15 is based on relationships (7.58) and (7.59) and is a reminder first that $MSE = s_{\mathrm{P}}^2 = 6.88$ serves as an estimate of σ^2. So multiple magnesium determinations on a given specimen would be estimated to have a standard deviation on the order of $\sqrt{6.88} = 2.6 \times 10^{-3}\%$. Then the expected mean squares further suggest that σ_τ^2 be estimated by

$$\hat{\sigma}_\tau^2 = \frac{1}{10}(MSTr - MSE) = \frac{1}{10}(7.58 - 6.88) = .07$$

Example 8
(continued)

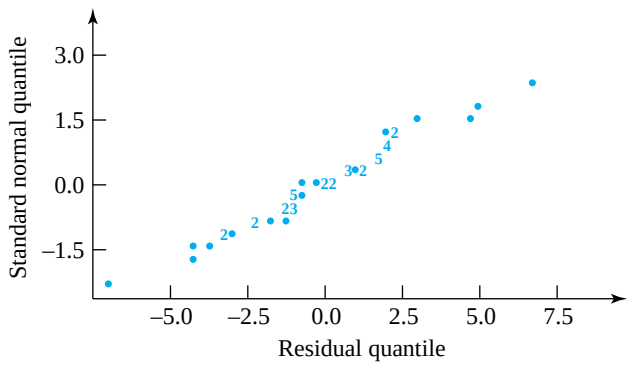

Figure 7.13 Normal plot of residuals for the magnesium content study

Table 7.15
ANOVA Table for the Magnesium Content Study

ANOVA Table (for testing $H_0: \sigma_\tau^2 = 0$)

Source	*SS*	*df*	*MS*	*EMS*	*F*
Treatments	30.32	4	7.58	$\sigma^2 + 10\sigma_\tau^2$	1.10
Error	309.70	45	6.88	σ^2	
Total	340.02	49			

as in equation (7.62). So an estimate of σ_τ is

$$\sqrt{.07} = .26 \times 10^{-3}\%$$

That is, the standard deviation of specimen mean magnesium contents is estimated to be on the order of $\frac{1}{10}$ of the standard deviation associated with multiple measurements on a single specimen.

A confidence interval for σ^2 could be made using formula (7.10) of Section 7.1. That will not be done here, but formula (7.63) will be used to make a one-sided 90% confidence interval of the form (0, #) for σ_τ/σ. The .90 quantile of the $F_{45,4}$ distribution is about 3.80, so the .10 quantile of the $F_{4,45}$ distribution is about $\frac{1}{3.80}$. Then taking the root of the second endpoint given in display (7.63), a 90% upper confidence bound for σ_τ/σ is

$$\sqrt{\frac{1}{10}\left(\frac{7.58}{\left(\frac{1}{3.80}\right)6.88} - 1\right)} = .56$$

The bottom line here is that σ_τ is small compared to σ and is not even clearly other than 0. Most of the variation in the data of Table 7.14 is associated with the making of multiple measurements on a single specimen. Of course, this is good news if the rod is to be cut up and distributed as pieces having known magnesium contents and thus useful for measurement instrument calibration.

Section 4 Exercises

1. Return to the situation in Exercises 1 of Sections 7.1 through 7.3 (and the pressure/density data of Example 1 in Chapter 4).

(a) In part (b) of Exercise 1 of Section 7.3, you were asked to make simultaneous confidence intervals for all differences in the $r = 5$ mean densities. From your intervals, what kind of a p-value (small or large) do you expect to find when testing the equality of these means? Explain.

(b) Make an ANOVA table (in the form of Table 7.12) for the data of Example 1 in Chapter 4. You should do the calculations by hand first and then check your arithmetic using a statistical computer package. Then use the calculations to find both R^2 for the one-way model and also the observed level of significance for an F test of the null hypothesis that all five pressures produce the same mean density.

2. Return to the tilttable study of Exercises 2 of Sections 7.1 through 7.3.

(a) In part (b) of Exercise 2 of Section 7.3, you were asked to make simultaneous confidence intervals for all differences in the $r = 4$ mean tilttable ratios. From your intervals, what kind of a p-value (small or large) do you expect to find when testing the equality of these means? Explain.

(b) Make an ANOVA table (in the form of Table 7.12) for the data of Exercise 2 of Section 7.1. Then find both R^2 for the one-way model and also the observed level of significance for an F test of the null hypothesis that all four vans have the same mean tilttable ratio.

3. The following data are taken from the paper "Zero-Force Travel-Time Parameters for Ultrasonic Head-Waves in Railroad Rail" by Bray and Leon-Salamanca (*Materials Evaluation*, 1985). Given are measurements in nanoseconds of the travel time (in excess of 36.1 μs) of a certain type of mechanical wave induced by mechanical stress in railroad rails. Three measurements were made on each of six different rails.

Rail	Travel Time (nanoseconds above 36.1 μs)
1	55, 53, 54
2	26, 37, 32
3	78, 91, 85
4	92, 100, 96
5	49, 51, 50
6	80, 85, 83

(a) Make plots to check the appropriateness of a one-way random effects analysis of these data. What do these suggest?

(b) Ignoring any possible problems with the standard assumptions of the random effects model revealed in (a), make an ANOVA table for these data (like Table 7.15) and find estimates of σ and σ_τ. What, in the context of this problem, do these two estimates measure?

(c) Find and interpret a two-sided 90% confidence interval for the ratio σ_τ/σ.

4. The following are some general questions about the random effects analyses:

(a) Explain in general terms when a random effects analysis is appropriate for use with multisample data.

(b) Consider a scenario where $r = 5$ different technicians employed by a company each make

$m = 2$ measurements of the diameter of a particular widget using a particular gauge in a study of how technician differences show up in diameter data the company collects. Under what circumstances would a random effects analysis of the resulting data be appropriate?

(c) Suppose that the following ANOVA table was made in a random effects analysis of data like those described in part (b). Give estimates of the standard deviation associated with repeat diameter measurements for a given technician (σ) and then for the standard deviation of long-run mean measurements for various technicians (σ_τ). The sums of squares are in units of square inches.

ANOVA Table

Source	*SS*	*df*	*MS*	*F*
Technician	.0000136	4	.0000034	1.42
Error	.0000120	5	.0000024	
Total	.0000256	9		

7.5 Shewhart Control Charts for Measurement Data

This text has repeatedly made use of the phrase "stable process" and emphasized that unless data generation has associated with it a single, repeatable pattern of variation, there is no way to move from data in hand to predictions and inferences. The notion that "baseline" or "inherent" variation evident in the output of a process is a principal limitation on system performance has also been stressed. But no tools have yet been presented that are specifically crafted for evaluating the extent to which a data-generation mechanism can be thought of as stable, or for determining the size of the baseline variation of a process.

W. Shewhart, working in the late 1920s and early 1930s at Bell Laboratories, developed an extremely simple yet effective device for doing these jobs. This tool has become known as the *Shewhart control chart*. (Actually, the nonstandard name *Shewhart monitoring chart* is far more descriptive. It also avoids the connotations of automatic/feedback process adjustment that the word *control* may carry for readers familiar with the field of engineering control.)

This section and the next introduce the topic of Shewhart control charts, beginning here with charts for measurement data. This section begins with some generalities, discussing Shewhart's conceptualization of process variability. Then the specific instances of Shewhart control charts for means, ranges, and standard deviations are considered in turn. Finally, the section closes with comments about the place of control charts in the improvement of modern industrial processes.

7.5.1 Generalities about Shewhart Control Charts

Stability of an engineering data-generating process refers to a consistency or repeatability over time. When one thinks of empirically assessing the stability of a process, it is therefore clear that samples of data taken from it at different points in time will be needed.

Example 9

Monitoring the Lengths of Sheets Cut on a Ream Cutter

Shervheim and Snider worked with a company on the cutting of a rolled material into sheets using a ream cutter. Every two minutes they sampled five consecutive sheets and measured their lengths. Part of the students' length data are given in Table 7.16, in units of $\frac{1}{64}$ inch over a certain reference length.

One of the goals of the study was to investigate the stability of the cutting process over time. The kind of multisample data the students collected, where the samples were separated and ordered in time, are ideal for that purpose.

Table 7.16
Lengths of 22 Samples of Five Sheets Cut on a Ream Cutter

Sample	Time	Excess Length
1	12:40	9, 10, 7, 8, 10
2	12:42	6, 10, 8, 8, 10
3	12:44	11, 10, 9, 5, 11
4	12:46	10, 9, 9, 8, 7
5	12:48	7, 5, 11, 9, 5
6	12:50	9, 9, 10, 7, 9
7	12:52	10, 8, 6, 11, 8
8	12:54	7, 10, 8, 8, 9
9	12:56	10, 9, 9, 5, 12
10	12:58	8, 10, 6, 8, 10
11	1:00	8, 10, 4, 7, 8
12	1:02	8, 10, 10, 6, 9
13	1:04	10, 8, 6, 7, 10
14	1:06	8, 6, 10, 8, 8
15	1:08	13, 5, 8, 8, 13
16	1:10	10, 4, 9, 10, 8
17	1:12	7, 7, 9, 7, 8
18	1:14	9, 7, 7, 9, 6
19	1:16	5, 10, 5, 8, 10
20	1:18	9, 6, 8, 9, 11
21	1:20	6, 10, 11, 5, 6
22	1:22	15, 3, 7, 9, 11

Data (like those in Table 7.16) collected for purposes of assessing process stability will often be r samples of some fixed sample size m, lacking any structure except for the fact that they were taken in a particular time order. So Shewhart control

charting is at home in this chapter that treats inference methods for unstructured multisample studies.

Shewhart's fundamental qualitative insight regarding variation seen in process data over time is that

Shewhart's partition of process variation

$$\begin{array}{c}\text{Overall process}\\ \text{variation}\end{array} = \text{baseline variation} + \begin{array}{c}\text{variation that}\\ \text{can be eliminated}\end{array} \tag{7.64}$$

Shewhart conceived of **baseline variation** as that which will remain even under the most careful process monitoring and appropriate physical interventions—an inherent property of a particular system configuration, which cannot be reduced without basic changes in the physical process or how it is run. This is variation due to **common (universal) causes** or **system causes**. Other terms used for it are **random variation** and **short-term variation**. In the context of the cutting operation of Example 9, this kind of variation might be seen in consecutive sheet lengths cut on a single ream cutter, from a single roll of material, without any intervening operator adjustments, following a particular plant standard method of machine operation, etc. It is variation that comes from hundreds of small unnameable, unidentifiable physical causes. When only this kind of variation is acting, it is reasonable to call a process "stable."

The second component of overall process variation is **variation that can potentially be eliminated** by appropriate physical intervention. This kind of variation has been called variation due to **special** or **assignable causes**, **nonrandom variation**, and **long-term variation**. In the sheet-cutting example, this might be variation in sheet length brought about by undesirable changes in tension on the material being cut, roller slippage on the cutter, unwarranted operator adjustments to the machine, eccentricities associated with how a particular incoming roll of material was wound, etc. Shewhart reasoned that being able to separate the two kinds of variation is a prerequisite to ensuring good process performance. It provides a basis for knowing when to intervene and find and eliminate the cause of any assignable variation, thereby producing process stability.

Shewhart control charts

Shewhart's method for separating the two components of overall variation in equation (7.64) is graphical and based on the following logic. First, periodically taken samples are reduced to appropriate summary statistics, and the summary statistics are plotted against time order of observation. To this simple time-plotting of summary statistics, Shewhart added the notion that lines be drawn on the chart to separate values that are consistent with a "baseline variation only" view of process performance from those that are not. Shewhart called these lines of demarcation **control limits**. When all plotted points fall within the control limits, the process is judged to be stable, subject only to chance causes. But when a point falls outside the limits, physical investigation and intervention is called for, to eliminate any assignable cause of variation. Figure 7.14 is a plot of a generic control chart for a summary statistic, w. It shows upper and lower control limits (*UCL* and *LCL*), some plotted values, and one "out of control" point.

There are any number of charts that fit the general pattern of Figure 7.14. For example, common possibilities relevant in the sheet-cutting case of Example 9 include control charts for the sample mean, sample range, and sample standard

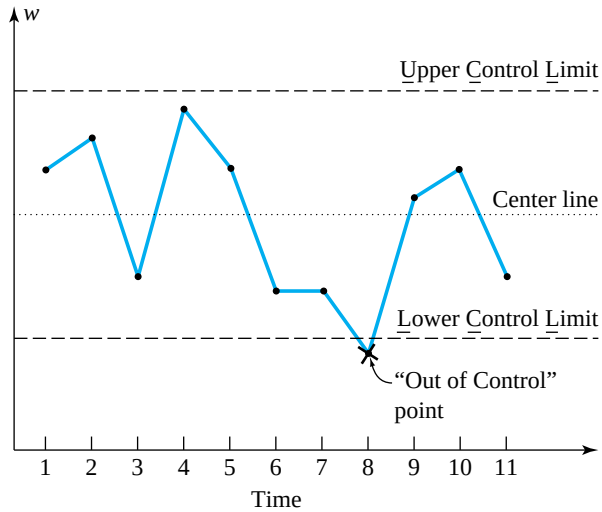

Figure 7.14 Generic Shewhart control chart for a statistic w

deviation of sheet lengths. These will presently be discussed in detail. But first, some additional generalities still need to be considered.

Setting control limits

For one thing, there remains the matter of how to set the position of the control limits. Shewhart argued that probability theory can be applied and appropriate stable-process/iid-observations distributions developed for the plotted statistics. Then small upper and lower percentage points for these can be used to establish control limits. As an example, the central limit material in Section 5.5 should have conditioned the reader to think of sample means as approximately normal with mean μ and standard deviation $\sigma/\sqrt{m}$, where μ and σ describe individual observations and m is the sample size. So for plotting sample means, the upper and lower control limits might be set at small upper and lower percentage points of the normal distribution with mean μ and standard deviation $\sigma/\sqrt{m}$, where μ and σ are a process mean and short-term standard deviation, respectively.

"Standards given" contexts

Two different circumstances are possible regarding the origin of values for process parameters used to produce control limits. In some applications, values of process parameters (and therefore, parameters for the "stable process" distribution of the plotted statistic) and thus control limits are provided from outside the data producing the charted values. Such circumstances will be called "**standards given**" situations. For emphasis, the meaning of this term is stated here in definition form.

Definition 7

When control limits are derived from data, requirements, or knowledge of the behavior of a process that are outside the information contained in the samples whose summary statistics are to be plotted, the charting is said to be done with **standards given**.

For example, suppose that in the sheet-cutting context of Example 9, past experience with the ream cutter indicates that a process short-term standard deviation of $\sigma = 1.9$ ($\frac{1}{64}$ in.) is appropriate when the cutter is operating as it should. Further, suppose that legal and other considerations have led to the establishment of a target process mean of $\mu = 10.0$ ($\frac{1}{64}$ in. above the reference length). Then control limits based on these values and applied to data collected tomorrow would be "standards given" control limits.

"Standards given" charting and hypothesis testing

One way to think about a "standards given" control chart is as a graphical means of repeatedly testing the hypothesis

$$\mathrm{H}_0\text{: Process parameters are at their standard values} \tag{7.65}$$

When a plotted point lies inside control limits, one is directed to a decision in favor of hypothesis (7.65) for the time period in question. A point plotting outside limits makes hypothesis (7.65) untenable at the time represented by the sample.

Retrospective contexts

In contrast to "standards given" applications, there are situations in which no external values for process parameters are used. Instead, a single set of samples taken from the process is used to both develop a plausible set of parameters for the process and also to judge the stability of the process over the period represented by the data. The terms **retrospective** or **"as past data"** will be used in this text for such control charting applications.

Definition 8

When control limits are derived from the same samples whose summary statistics are plotted, the charting is said to be done **retrospectively** or "**as past data**."

In the context of Example 9, control limits derived from the data in Table 7.16 and applied to summary statistics for those same data would be "as past data" control limits for assessing the cutting process stability over the period from 12:40 through 1:22 on the day the data were taken.

Retrospective charting and hypothesis testing

A way of thinking about a retrospective control chart is a graphical means of testing the hypothesis

$$\mathrm{H}_0\text{: A single set of process parameters was acting throughout the time period studied} \tag{7.66}$$

When a point or points plot outside of control limits derived from the whole data set, the hypothesis (7.66) of process stability over the period represented by the data becomes untenable.

7.5.2 "Standards Given" $\bar{x}$ Control Charts

The single most famous and frequently used Shewhart control chart is the one where sample mean measurements are plotted. Control charts are typically named

by the symbols used for the plotted statistics. So the following discussion concerns **Shewhart $\bar{x}$ charts**. In using this terminology (and other notation from the statistical quality control field), this text must choose a path through notational conflicts that exist between the most common usages in control charting and those for other multisample analyses. The options that will be exercised here must be explained.

Notational conventions for $\bar{x}$ charting

In the first place, to this point in Chapter 7 (also in Chapter 4, for that matter) the symbol y has been used for the basic response variable in a multisample statistical engineering study, $\bar{y}_i$ for a sample mean, and $\bar{y}_{\cdot}$ and $\bar{y}$ for unweighted and weighted averages of the $\bar{y}_i$, respectively . In contrast, in Chapters 3 and 6, where the discussion centered primarily on one- and two-sample studies, x was used as the basic response variable and $\bar{x}$ (or $\bar{x}_i$ in the case of two-sample studies) to stand for a sample mean. Standard usage in Shewhart control charting is to use the x and $\bar{x}$ ($\bar{x}_i$) convention, and the precedent is so strong that this section will adopt it as well. In addition, historical momentum in control charting dictates that rather than using $\bar{x}_{\cdot}$ notation,

Average sample mean (quality control notation)

$$\boxed{\bar{\bar{x}} = \frac{1}{r}\sum_{i=1}^{r}\bar{x}_i} \tag{7.67}$$

is used for the average of sample means. But this "bar bar" or "double bar" notation is used in this book *only* in this section.

Something must also be said about notation for sample sizes. It is universal to use the notation n_i for an individual sample size. But there is some conflict when all sample sizes n_i have a common value. The convention in this chapter has been to use m for such a common value and n for $\sum n_i$. Standard quality control notation is to instead use n for a common sample size. In this matter, we will continue to use the conventions established thus far in Chapter 7, believing that to do otherwise invites too much confusion. But the reader is hereby alerted to the fact that the m used here is usually going to appear as n in other treatments of control charting.

Having dealt with the notational problems, we turn to the making of a "standards given" Shewhart $\bar{x}$ chart based on samples of size m. An iid model for observations from a process with mean μ and standard deviation σ produces

$$E\bar{x} = \mu \tag{7.68}$$

and

$$\sqrt{\operatorname{Var}\bar{x}} = \frac{\sigma}{\sqrt{m}} \tag{7.69}$$

and often an approximately normal distribution for $\bar{x}$. The fact that essentially all of the probability of a normal distribution is within 3 standard deviations of its mean

led Shewhart to suggest that given process standards μ and σ, $\bar{x}$ chart control limits could be set at

"Standards given" control limits for $\bar{x}$

$$LCL_{\bar{x}} = \mu - 3\frac{\sigma}{\sqrt{m}} \quad \text{and} \quad UCL_{\bar{x}} = \mu + 3\frac{\sigma}{\sqrt{m}} \tag{7.70}$$

Additionally, he suggested drawing a *center line* on an $\bar{x}$ chart at the standard mean μ.

Limits (7.70) have proved themselves of great utility even in cases where m is fairly small and there is no reason to expect a normal distribution for observations in a sampling period. Formulas (7.68) and (7.69) hold regardless of whether a process distribution is normal, and the 3-sigma (of the plotted statistic $\bar{x}$) control limits in display (7.70) tend to bracket most of the distribution of $\bar{x}$ under nearly any circumstances. (Indeed, a crude but universal analysis, based on a probability version of the Chebyschev theorem stated in Section 3.3 for relative frequency distributions, guarantees that limits (7.70) will bracket at least $\frac{8}{9}$ of the distribution of $\bar{x}$ in any stable process context.)

Example 9 *(continued)*

Consider the use of process standards $\mu = 10$ and $\sigma = 1.9$ in $\bar{x}$ charting based on the data given in Table 7.16 (recall the values there are in units of $\frac{1}{64}$ in. over a reference length). With these standard values for μ and σ, since the $r = 22$

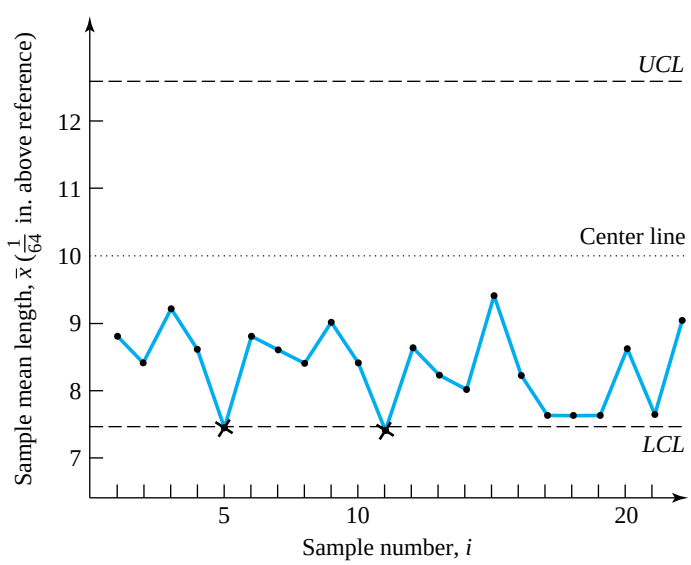

Figure 7.15 "Standards given" Shewhart $\bar{x}$ control chart for cut sheet lengths

samples are all of size $m = 5$, formulas (7.70) indicate control limits

$$UCL_{\bar{x}} = 10 + 3\frac{1.9}{\sqrt{5}} = 12.55 \quad \text{and} \quad LCL_{\bar{x}} = 10 - 3\frac{1.9}{\sqrt{5}} = 7.45$$

along with a center line drawn at $\mu = 10$. Table 7.17 gives some sample-by-sample summary statistics for the data of Table 7.16, including the sample means $\bar{x}_i$. Figure 7.15 is a "standards given" Shewhart $\bar{x}$ chart for the same data.

Figure 7.15 shows two points plotting below the lower control limit: the means for samples 5 and 11. But it is perfectly obvious from the plot what was going on in the data of Table 7.16 to produce the "out of control" points and corresponding debunking of hypothesis (7.65). Not one of the $r = 22$ plotted

Table 7.17
Sample-by-Sample Summary Statistics for 22 Samples of Sheet Lengths

i, Sample	$\bar{x}_i$	s_i	R_i
1	8.8	1.30	3
2	8.4	1.67	4
3	9.2	2.49	6
4	8.6	1.14	3
5	7.4	2.61	6
6	8.8	1.10	3
7	8.6	1.95	5
8	8.4	1.14	3
9	9.0	2.55	7
10	8.4	1.67	4
11	7.4	2.19	6
12	8.6	1.67	4
13	8.2	1.79	4
14	8.0	1.41	4
15	9.4	3.51	8
16	8.2	2.49	6
17	7.6	.89	2
18	7.6	1.34	3
19	7.6	2.51	5
20	8.6	1.82	5
21	7.6	2.70	6
22	9.0	4.47	12
	$\sum \bar{x} = 183.4$	$\sum s = 44.41$	$\sum R = 109$

Example 9 (*continued*)

sample means lies at or above 10. If an average sheet length of $\mu = 10$ was truly desired, a simple adjustment was needed, to increase sheet lengths roughly

$$10 - \bar{\bar{x}} = 10 - 8.3 = 1.7 \left(\tfrac{1}{64} \text{ in.}\right)$$

The true process mean operating to produce the data was clearly below the standard mean.

7.5.3 Retrospective $\bar{x}$ Control Charts

Retrospective (or "as past data") control limits for $\bar{x}$ come about by replacing μ and σ in formulas (7.70) with estimates made from data in hand, under the provisional assumption that the process was stable over the period represented by the data. That is, in calculating such estimates, a single set of parameters is presumed to be adequate to describe process behavior during the study period. Notice that supposing process stability the present situation is exactly the one met in the ANOVA material of Section 7.4 under the hypothesis of equality of r means. So one way to think about a retrospective $\bar{x}$ chart is as a graphical test of the constancy of the process mean over time. Further, the analogy with the material of Section 7.4 suggests natural estimates of μ and σ for use in formulas (7.70).

In Section 7.4, $\bar{y}$ was used to approximate a hypothesized common value of $\mu_1, \mu_2, \ldots, \mu_r$. In the present notation, this suggests replacing μ in formulas (7.70) with $\bar{\bar{x}}$. Regarding an estimate of σ for use in formulas (7.70), analogy with all that has gone before in this chapter suggests s_{P}. And indeed, s_{P} is a perfectly rational choice. But it is not one that is commonly used. Historical precedent/accident in the quality control field has made other estimates much more widely used. These must therefore be discussed, not so much because they are better than s_{P}, but because they represent standard practice.

The most common way of approximating a supposedly constant σ in control charting contexts is based on probability facts about the range, R, of a sample of m observations from a normal distribution. It is possible to derive the probability density for R defined in Definition 8 in Chapter 3 (see page 95), supposing m iid normal variables with mean μ and standard deviation σ are involved. That density will not be given in this book. But it is useful to know that the mean of that distribution is (for a given sample size m) proportional to σ. The constant of proportionality is typically called d_2, and in symbols,

$$ER = d_2\sigma \tag{7.71}$$

or equivalently,

$$\sigma = \frac{ER}{d_2} \tag{7.72}$$

Values of d_2 for various m are given in Table B.2. (Return to the comments preceding Proposition 1 in Section 3.3 and recognize that what was cryptic there should now make sense.)

Statements (7.71) and (7.72) are theoretical. The way they find practical relevance is to think that under the hypothesis that the process standard deviation is constant, the sample mean of sample ranges

Average sample range

$$\boxed{\overline{R} = \frac{1}{r}\sum_{i=1}^{r} R_i} \tag{7.73}$$

can be expected to approximate the theoretical mean range, ER. That is, from statement (7.72), it seems that

A range-based estimator of σ

$$\boxed{\hat{\sigma} = \frac{\overline{R}}{d_2}} \tag{7.74}$$

is a plausible way to estimate σ. On theoretical grounds, $\overline{R}/d_2$ is inferior to s_{P}, but it has the weight of historical precedent behind it, and it is simple to calculate (an important virtue before the advent of widespread computing power).

A second estimator of σ with quality control origins comes about by making the same kind of argument that led to statistic (7.74), beginning not with R but instead with s. That is, the fact that it is possible to derive a χ^2_{m-1} probability density for $(m-1)s^2/\sigma^2$ if s^2 is based on m iid normal (μ, σ^2) random variables has been used extensively (beginning in Section 6.4) in this text. That density can in turn be used to find a theoretical mean for s. As it turns out, although $Es^2 = \sigma^2$, the theoretical mean of s is not quite σ, but rather a multiple of σ (for a given sample size m). The constant of proportionality is typically called c_4, and in symbols,

$$Es = c_4\sigma \tag{7.75}$$

or equivalently,

$$\sigma = \frac{Es}{c_4} \tag{7.76}$$

It is possible to write out an explicit expression for c_4, namely

$$c_4 = \sqrt{\frac{2}{m-1}}\left(\frac{\Gamma\left(\frac{m}{2}\right)}{\Gamma\left(\frac{m-1}{2}\right)}\right)$$

Values of c_4 for various m are given in Table B.2. From that table, it is easy to see that as a function of m, c_4 increases from about .8 when $m = 2$ to essentially 1 for large m.

The practical use made of the theoretical statements (7.75) and (7.76) is to think that the sample average of the sample standard deviations

Average sample standard deviation

$$\boxed{\bar{s} = \frac{1}{r}\sum_{i=1}^{r} s_i} \tag{7.77}$$

can be expected to approximate the theoretical mean (sample) standard deviation Es, so that (from statement (7.76)) a plausible estimator of σ becomes

A standard deviation-based estimator of σ

$$\boxed{\hat{\sigma} = \frac{\bar{s}}{c_4}} \tag{7.78}$$

(It is worth remarking that $\bar{s}$ is not the same as s_{p}, even when all sample sizes are the same. s_{p} is derived by averaging sample variances and then taking a square root. $\bar{s}$ comes from taking the square roots of the sample variances and then averaging. In general, these two orders of operation do not produce the same results.)

In any case, commonly used retrospective control limits for $\bar{x}$ are obtained by substituting $\bar{\bar{x}}$ given in formula (7.67) for μ and either of the estimates of σ given in displays (7.74) or (7.78) for σ in the formulas (7.70). Further, an "as past data" *center line* for an $\bar{x}$ chart is typically set at $\bar{\bar{x}}$.

Example 9 *(continued)*

Consider retrospective $\bar{x}$ control charting for the ream cutter data. Using the column totals given in Table 7.17, one finds from formulas (7.67), (7.73), and (7.77) that

$$\bar{\bar{x}} = \frac{183.4}{22} = 8.3$$

$$\bar{R} = \frac{109}{22} = 4.95$$

$$\bar{s} = \frac{44.41}{22} = 2.019$$

Then, consulting Table B.2 with a sample size of $m = 5$, $d_2 = 2.326$, so an estimate of σ based on $\overline{R}$ is (from expression (7.74))

$$\frac{\overline{R}}{d_2} = \frac{4.95}{2.326} = 2.13$$

Also, Table B.2 shows that for a sample size of $m = 5$, $c_4 = .9400$, so an estimate of σ based on $\bar{s}$ is (from expression (7.78))

$$\frac{\bar{s}}{c_4} = \frac{2.019}{.94} = 2.15$$

(Note that beginning from the standard deviations in Table 7.17, $s_{\mathrm{p}} = 2.19$, and clearly $s_{\mathrm{p}} \neq \bar{s}$.)

Using (for example) statistic (7.74), one is thus led to substitute 8.3 for μ and 2.13 for σ in "standards given" formulas (7.70) to obtain the retrospective limits

$$LCL_{\bar{x}} = 8.3 - 3\frac{2.13}{\sqrt{5}} = 5.44 \quad \text{and} \quad UCL_{\bar{x}} = 8.3 + 3\frac{2.13}{\sqrt{5}} = 11.16$$

Figure 7.16 shows an "as past data" Shewhart $\bar{x}$ control chart for the ream cutter data, using limits based on $\overline{R}$.

Notice the contrast between the pictures of the ream cutter performance given in Figures 7.15 and 7.16. Figure 7.15 shows clearly that process parameters are not at their standard values, but Figure 7.16 shows that it is perhaps plausible to think of the data in Table 7.16 as coming from *some* stable data-generating mechanism. The observed $\bar{x}$'s hover nicely (indeed—as will be argued at the end of the next section—perhaps too nicely) about a central value, showing no "out of control" points or obvious trends. That hypothesis (7.66) is at least approximately true is believable on the basis of Figure 7.16.

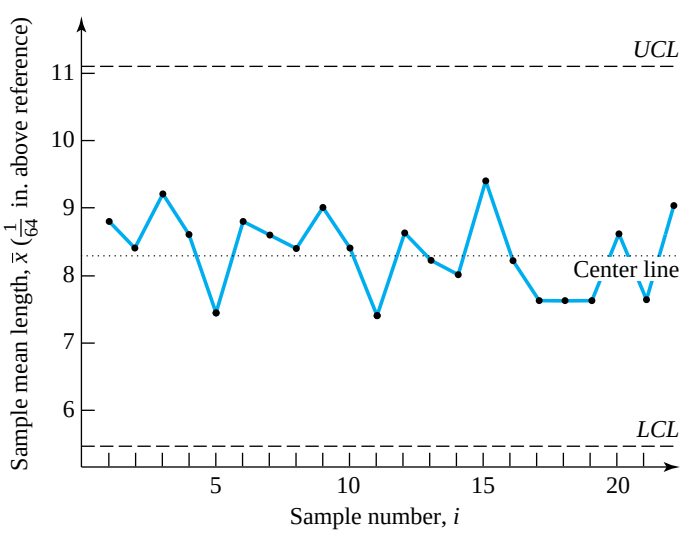

Figure 7.16 Retrospective Shewhart $\bar{x}$ control chart for cut sheet lengths

Several comments should be made before turning to a discussion of other Shewhart control charts for measurements. First, note that what is represented on an $\bar{x}$ chart is behavior (both expected and observed) of sample means, *not* individual measurements. It is unfortunately all too common to see engineering specifications (which refer to individual measurements) marked on $\bar{x}$ control charts either in place of, or in addition to, proper control limits. But how *sample means* compare to specifications for *individual measurements* tells nothing about either the stability of the process as represented in the means or the acceptability of individual measurements according to the stated engineering requirements. It is simply bad practice to mix (or mix up) control limits and specifications.

Control limits for $\bar{x}$ versus specifications for x

A second comment has to do with the fairly arbitrary choice of 3-sigma control limits in formulas (7.70). A legitimate question is, "Why not 2-sigma or 2.5-sigma or 3.09-sigma limits?" There is no completely convincing theoretical answer to this question. Indeed, arguments in favor of other multiples than 3 for use in formulas (7.70) are heard from time to time. But the forces of historical precedent and many years of successful application combine to make the use of 3-sigma limits nearly universal.

As a final point regarding $\bar{x}$ charts, the basic "standards given" formulas for control limits (7.70) are sometimes combined with formula (7.74) or (7.78) for estimating σ, and $\bar{\bar{x}}$ is put in place of μ to obtain formulas for retrospective control limits for $\bar{x}$. For example, using the estimate of σ in display (7.74), one obtains the formulas

$$LCL_{\bar{x}} = \bar{\bar{x}} - 3\frac{\overline{R}}{d_2\sqrt{m}} \quad \text{and} \quad UCL_{\bar{x}} = \bar{\bar{x}} + 3\frac{\overline{R}}{d_2\sqrt{m}} \tag{7.79}$$

In fact, it is standard practice to use the abbreviation

$$A_2 = \frac{3}{d_2\sqrt{m}}$$

and rewrite the limits in formulas (7.79) as

Range-based retrospective control limits for $\bar{x}$

$$\boxed{LCL_{\bar{x}} = \bar{\bar{x}} - A_2\overline{R} \quad \text{and} \quad UCL_{\bar{x}} = \bar{\bar{x}} + A_2\overline{R}} \tag{7.80}$$

Values of A_2 are given along with the other control chart constants in Table B.2. It is worthwhile to verify that the use of formulas (7.80) in the context of Example 9 produces exactly the retrospective control limits for $\bar{x}$ found earlier.

The version of retrospective $\bar{x}$ chart limits related to the estimate of σ in display (7.78) is

$$LCL_{\bar{x}} = \bar{\bar{x}} - 3\frac{\bar{s}}{c_4\sqrt{m}} \quad \text{and} \quad UCL_{\bar{x}} = \bar{\bar{x}} + 3\frac{\bar{s}}{c_4\sqrt{m}} \tag{7.81}$$

It is also standard practice to use the abbreviation

$$A_3 = \frac{3}{c_4\sqrt{m}}$$

and rewrite the limits in display (7.81) as

Standard deviation-based retrospective control limits for $\bar{x}$

$$\boxed{LCL_{\bar{x}} = \bar{\bar{x}} - A_3\bar{s} \quad \text{and} \quad UCL_{\bar{x}} = \bar{\bar{x}} + A_3\bar{s}} \tag{7.82}$$

Values of A_3 are given in Table B.2.

7.5.4 Control Charts for Ranges

The $\bar{x}$ control chart is aimed primarily at monitoring the constancy of the average process response, μ, over time. It deals only indirectly with the process short-term variation σ. (If σ increases beyond a standard value, it will produce $\bar{x}_i$ more variable than expected and eventually trigger an "out of control" point. But such a possible change in σ is detected most effectively by directly monitoring the spread of samples.) Thus, in applications, $\bar{x}$ charts are almost always accompanied by companion charts intended to monitor σ.

The conceptually simplest and most common Shewhart control charts for monitoring the process standard deviation are the ***R* charts**, the charts for sample ranges. In their "standards given" version, they are based again on the fact that it is possible to find a probability density for R based on m iid normal (μ, σ^2) random variables. Using this density, not only is it possible to show that $ER = d_2\sigma$ but the standard deviation of the probability distribution can be found as well. It turns out (for a given m) to be proportional to σ. The constant of proportionality is called d_3 and is tabled for various m in Table B.2. That is, for R based on m iid normal observations,

$$\sqrt{\text{Var } R} = d_3\sigma \tag{7.83}$$

Although the information about the theoretical distribution of R provided by formulas (7.71) and (7.83) is somewhat sketchy, it is enough to suggest possible "standards given" 3-sigma (of R) control limits for R. A plausible *center line* for a "standards given" R chart is at $ER = d_2\sigma$, and (using formula (7.83)) control limits are

$$LCL_R = ER - 3\sqrt{\text{Var } R} = d_2\sigma - 3d_3\sigma = (d_2 - 3d_3)\sigma \tag{7.84}$$

$$UCL_R = ER + 3\sqrt{\text{Var } R} = (d_2 + 3d_3)\sigma \tag{7.85}$$

The limit indicated in formula (7.84) turns out to be negative for $m \leq 6$. For those sample sizes, since ranges are nonnegative, no lower control limit is used. Formulas

(7.84) and (7.85) are typically simplified by the introduction of yet more notation. That is, standard quality control usage is to let

$$D_1 = (d_2 - 3d_3) \quad \text{and} \quad D_2 = (d_2 + 3d_3)$$

and rewrite formulas (7.84) and (7.85) as

"Standards given" control limits for R

$$\boxed{LCL_R = D_1\sigma \quad \text{and} \quad UCL_R = D_2\sigma} \tag{7.86}$$

Like the other control chart constants, D_1 and D_2 appear in Table B.2. Note that for $m \leq 6$, there is no tabled value for D_1, as no lower limit is in order.

Example 9 *(continued)*

Consider a "standards given" control chart analysis for the sheet length ranges given in Table 7.17, using a standard $\sigma = 1.9$ ($\frac{1}{64}$ in.). Since samples of size $m = 5$ are involved, Table B.2 shows that $d_2 = 2.326$ and $D_2 = 4.918$ are appropriate for establishing a "standards given" control chart for R. The center line should be drawn at

$$d_2\sigma = 2.326(1.9) = 4.4$$

and the upper control limit should be set at

$$D_2\sigma = 4.918(1.9) = 9.3$$

(Since $m \leq 6$, no lower control limit will be used.) Figure 7.17 shows a "standards given" control chart for ranges of the sheet lengths. It is clear from the figure that

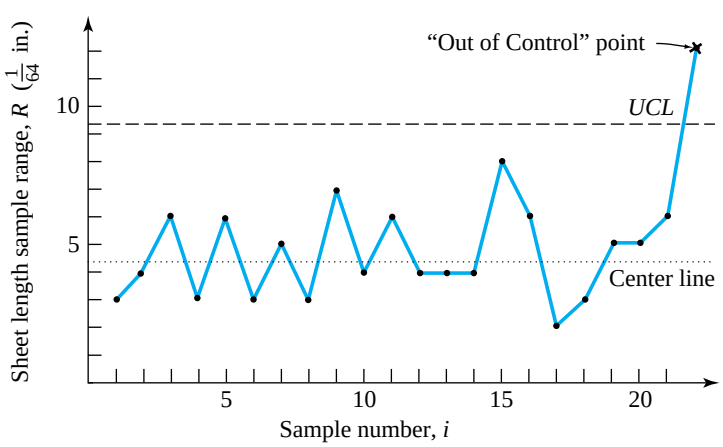

Figure 7.17 "Standards given" Shewhart R chart for cut sheet lengths

for the most part, a constant process standard deviation of $\sigma = 1.9$ is plausible, except for the clear indication to the contrary at sample 22. The 22nd observed range, $R = 12$, is simply larger than expected based on a sample of size $m = 5$ from a normal distribution with $\sigma = 1.9$. In practice, it would be appropriate to undertake a physical search for the cause of the apparent increase in process variability associated with the last sample taken.

As was the case for $\bar{x}$ charts, combination of formulas for the estimation of (supposedly constant) process parameters with the "standards given" limits (7.86) produces retrospective control limits for R charts. For example, basing an estimate of σ on $\overline{R}$ as in display (7.74), leads (not too surprisingly) to a retrospective *center line* for R at $d_2(\overline{R}/d_2) = \overline{R}$ and retrospective control limits

$$LCL_R = \frac{D_1\overline{R}}{d_2} \quad \text{and} \quad UCL_R = \frac{D_2\overline{R}}{d_2} \tag{7.87}$$

The abbreviations

$$D_3 = \frac{D_1}{d_2} \quad \text{and} \quad D_4 = \frac{D_2}{d_2}$$

are commonly used, and limits (7.87) are written as

Retrospective control limits for R

$$\boxed{LCL_R = D_3\overline{R} \quad \text{and} \quad UCL_R = D_4\overline{R}} \tag{7.88}$$

Values of the constants D_3 and D_4 are found in Table B.2.

Example 9 *(continued)*

For the ream cutter data, $\overline{R} = \frac{109}{22}$, so retrospective control limits for ranges of the type (7.88) put a center line at

$$\overline{R} = 4.95$$

and since for $m = 5$, $D_4 = 2.114$,

$$UCL_R = 2.114\left(\frac{109}{22}\right) = 10.5$$

Look again at Figure 7.17 and note that the use of these retrospective limits (instead of the $\sigma = 1.9$ "standards given" limits of Figure 7.17) does not materially alter the appearance of the plot. The range for sample 22 still plots above the upper control limit. It is not plausible that a single σ stands behind all of the 22

Example 9 (*continued*)

plotted ranges (not even $\sigma \approx \overline{R}/d_2 = 2.13$). It is pretty clear that a different physical mechanism must have been acting at sample 22 than was operative earlier.

For pedagogical reasons, $\bar{x}$ charts were considered first before turning to charts aimed at monitoring σ. In terms of order of attention in an application, however, R (or s) charts are traditionally (and correctly) given first priority. They deal directly with the baseline component of process variation. Thus (so conventional wisdom goes), if they show lack of stability, there is little reason to go on to considering the behavior of means (which deals primarily with the long-term component of process variation) until appropriate physical changes bring the ranges (or standard deviations) to the place of repeatability.

7.5.5 Control Charts for Standard Deviations

Less common but nevertheless important alternatives to range charts are control charts for standard deviations, s. In their "standards given" version, ***s* charts** are based on the fact that it is possible to find both a mean and variance for s calculated from m iid normal (μ, σ^2) random variables. We have already used the fact that $Es = c_4\sigma$. And it turns out that

$$\sqrt{\operatorname{Var} s} = \sqrt{1 - c_4^2}\,\sigma \tag{7.89}$$

Then formulas (7.75) and (7.89) taken together yield "standards given" 3-sigma control limits for s. That is, with a *center line* at $c_4\sigma$, one employs the limits

$$\begin{aligned} LCL_s &= c_4\sigma - 3\sqrt{1 - c_4^2}\,\sigma = \left(c_4 - 3\sqrt{1 - c_4^2}\right)\sigma \\ UCL_s &= c_4\sigma + 3\sqrt{1 - c_4^2}\,\sigma = \left(c_4 + 3\sqrt{1 - c_4^2}\right)\sigma \end{aligned}$$

Standard notation is to let

$$B_5 = \left(c_4 - 3\sqrt{1 - c_4^2}\right) \quad \text{and} \quad B_6 = \left(c_4 + 3\sqrt{1 - c_4^2}\right)$$

so, ultimately, "standards given" control limits for s become

"Standards given" control limits for s

$$\boxed{LCL_s = B_5\sigma \quad \text{and} \quad UCL_s = B_6\sigma} \tag{7.90}$$

As expected, the constants B_5 and B_6 are tabled in Table B.2. For $m \le 5$, $c_4 - 3\sqrt{1 - c_4^2}$ turns out to be negative, so no value is shown in Table B.2 for B_5, and no lower control limit for s is typically used for such sample sizes.

Example 9 *(continued)*

Returning once more to the ream cutter example of Shervheim and Snider, consider the monitoring of σ through the use of sample standard deviations rather than ranges, based on a standard of $\sigma = 1.9$ ($\frac{1}{64}$ in.). Table B.2 with sample size $m = 5$ once again gives $c_4 = .9400$ and also shows that $B_6 = 1.964$. So an s chart for the data of Table 7.16 has a center line at

$$c_4\sigma = (.94)(1.9) = 1.79$$

and an upper control limit at

$$UCL_s = B_6\sigma = 1.964(1.9) = 3.73$$

and, since the sample size is only 5, no lower control limit.

Figure 7.18 is a "standards given" Shewhart s chart for the s values given in Table 7.17. The story told by Figure 7.18 is essentially identical to that conveyed by the range chart in Figure 7.17. Only at sample 22 does the hypothesis that $\sigma = 1.9$ become untenable, and the need for physical intervention is indicated there.

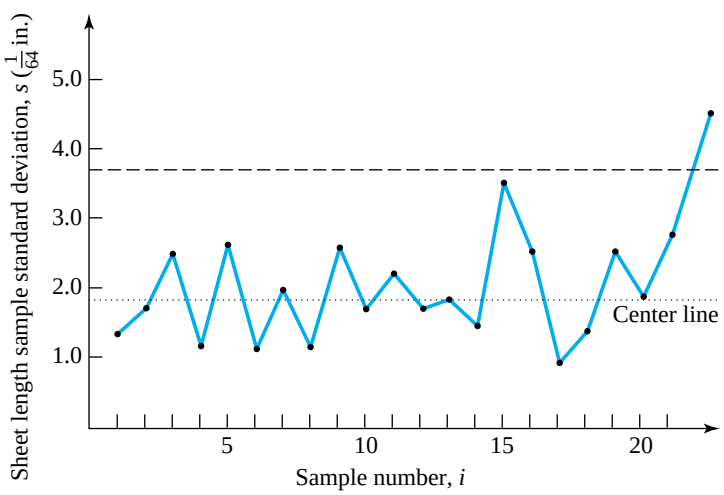

Figure 7.18 "Standards given" s chart for cut sheet lengths

As was the case for $\bar{x}$ and R charts, retrospective control limits for s can be had by replacing the parameter σ in the "standards given" limits (7.90) with any appropriate estimate. The most common way of proceeding is to employ the estimator $\bar{s}/c_4$ and thus end up with a retrospective *center line* for an s chart at $c_4(\bar{s}/c_4) = \bar{s}$ and retrospective control limits

$$LCL_s = \frac{B_5\bar{s}}{c_4} \quad \text{and} \quad UCL_s = \frac{B_6\bar{s}}{c_4} \tag{7.91}$$

And using the abbreviations

$$B_3 = \frac{B_5}{c_4} \quad \text{and} \quad B_4 = \frac{B_6}{c_4}$$

the retrospective limits (7.91) are written as

Retrospective control limits for s

$$\boxed{LCL_s = B_3\bar{s} \quad \text{and} \quad UCL_s = B_4\bar{s}} \tag{7.92}$$

Values of B_3 and B_4 are given in Table B.2.

Example 9 *(continued)*

For the ream cutter data, $\bar{s} = \frac{44.41}{22} = 2.02$, so retrospective control limits for standard deviations of the type (7.92) put a center line at

$$\bar{s} = 2.02$$

and, since $B_4 = 2.089$ for $m = 5$,

$$UCL_s = 2.089\left(\frac{44.41}{22}\right) = 4.22$$

Look again at Figure 7.18 and verify that the use of these retrospective limits (instead of the $\sigma = 1.9$ "standards given" limits) wouldn't much change the appearance of the plot. As was the case for the retrospective R chart analysis, these retrospective s chart limits still put sample 22 in a class by itself, suggesting that a different physical mechanism produced it than that which led to the other 21 samples.

Ranges are easier to calculate "by hand" than standard deviations and are easier to explain as well. As a result, R charts are more popular than s charts. In fact, R charts are so common that the phrase "$\bar{x}$ and R charts" is often spoken in quality control circles in such a way that the $\bar{x}/R$ pair is almost implied to be a single inseparable entity. However, when computational problems and conceptual understanding are not issues, s charts are preferable to R charts because of their superior sensitivity to changes in σ.

A useful final observation about the s chart idea is that for r-sample statistical engineering studies where all sample sizes are the same, the "as past data" control limits in display (7.92) can provide some rough help in the model-checking activities of Section 7.1 (in reference to the "single variance" assumption of the one-way model). $B_3\bar{s}$ and $B_4\bar{s}$ can be treated as rough limits on the variation in sample standard deviations deemed to be consistent with the one-way model's single variance assumption.

Example 10
(*Example 1 revisited*)

s Chart Control Limits and the "Equal Variances" Assumption in the Concrete Strength Study

In the concrete compressive strength study of Armstrong, Babb, and Campen, the $r = 8$ sample standard deviations based on samples of size $m = 3$ given in Table 7.3 (page 450) have $\bar{s} = 534.8$ psi. Then for $m = 3$, $B_4 = 2.568$, and so

$$B_4\bar{s} = 2.568(534.8) = 1{,}373 \text{ psi}$$

The largest of the eight values s_i in Table 7.3 is 965.6, and there are thus no "out of control" standard deviations. So as in Section 7.1, no strong evidence against the relevance of the "single variance" model assumption is discovered here.

7.5.6 Control Charts for Measurements and Industrial Process Improvement

The $\bar{x}$ and R (or $\bar{x}$ and s) control chart combination is an important engineering tool for the improvement of manufacturing processes. U.S. companies have trained literally hundreds of thousands of workers in the making of Shewhart $\bar{x}$ and R charts over the past few years, hoping for help in meeting the challenge of international competition. The record of success produced by this training effort is mixed. It is thus worth pausing briefly to reflect on what aid the tools of this section can and cannot rationally be expected to provide in the effort to improve industrial processes.

Out-of-control signals must produce action

In the first place, warnings of assignable variation provided by Shewhart control charts are helpful in reducing the variation of an industrial process only to the extent that they are acted on in a timely and competent fashion. If "out of control" signals don't lead to appropriate physical investigation and action to eliminate assignable causes, they contribute nothing toward improved process behavior. If workers collect data to be archived away on $\bar{x}$ and R chart forms and do not have the authority, skills, or motivation to intervene intelligently when excess process variation is indicated, they are engaged in a futile activity.

Control charts can prevent over-adjustment

Control charts can signal the need for process intervention. But perhaps nearly as important is the fact that they also tell a user when not to be alarmed at observed variation and give in to the temptation to adjust a stable process. This is the other side of the intervention coin. Inadvisably adjusting an industrial process that is subject only to common or random causes degrades its behavior rather than improves it. Rational use of Shewhart control charts can help prevent this possibility.

Control charts help maintain current process best performance

It is also important to say that even when properly made and acted on, Shewhart control charts can do only so much towards the improvement of industrial processes. They can be a tool for helping to reduce variation to the minimum possible for a given system configuration (in terms of equipment, methods of operation, etc.). But once that minimum has been reached, all that Shewhart charting does is to help

maintain that configuration's best performance—to maintain the "baseline variation only" situation corresponding to the status quo way of doing things.

Control charts are not directly tools for innovation

In a modern world economy, however, companies cannot hope to be leaders in their industries by being content simply to maintain stable, status quo methods of operation. Instead, ways must be found for improving beyond today's methods for tomorrow. This requires thought and, often, engineering experimentation. The philosophies and methods of experimental design and engineering data collection and analysis discussed in this book have an important role in that search for improvement beyond today's best industrial methodology. But the particular role of control charting in such efforts is only indirect. By using control charts and bringing a current process to stability, a basis or foundation for improvement through experimentation and reconfiguration is provided. Indeed, it can be argued fairly convincingly that unless an existing process is repeatable, there is no sensible way of evaluating the impact of experimental changes made to it, trying to find tomorrow's improved version of the process. It is important to realize, however, that the Shewhart control charts provide only the foundation rather than the necessary subject matter expertise or statistical tools needed to guide the experimental search for improved ways of doing things.

Section 5 Exercises

1. The following are some data taken from a larger set in *Statistical Quality Control* by Grant and Leavenworth, giving the drained weights (in ounces) of contents of size No. $2\frac{1}{2}$ cans of standard grade tomatoes in puree. Twenty samples of three cans taken from a canning process at regular intervals are represented.

Sample	x_1	x_2	x_3	Sample	x_1	x_2	x_3
1	22.0	22.5	22.5	11	20.0	19.5	21.0
2	20.5	22.5	22.5	12	19.0	21.0	21.0
3	20.0	20.5	23.0	13	19.5	20.5	21.0
4	21.0	22.0	22.0	14	20.0	21.5	24.0
5	22.5	19.5	22.5	15	22.5	19.5	21.0
6	23.0	23.5	21.0	16	21.5	20.5	22.0
7	19.0	20.0	22.0	17	19.0	21.5	23.0
8	21.5	20.5	19.0	18	21.0	20.5	19.5
9	21.0	22.5	20.0	19	20.0	23.5	24.0
10	21.5	23.0	22.0	20	22.0	20.5	21.0

(a) Suppose that standard values for the process mean and standard deviation of drained weights (μ and σ) in this canning plant are 21.0 oz and 1.0 oz, respectively. Make and interpret "standards given" $\bar{x}$ and R charts based on these samples. What do these charts indicate about the behavior of the filling process over the time period represented by these data?

(b) As an alternative to the "standards given" range chart made in part (a), make a "standards given" s chart based on the 20 samples. How does its appearance compare to that of the R chart?

Now suppose that no standard values for μ and σ have been provided.

(c) Find one estimate of σ for the filling process based on the average of the 20 sample ranges, $\overline{R}$, and another based on the average of 20 sample standard deviations, $\bar{s}$. How do these compare to the pooled sample standard deviation (of Section 7.1), s_{P}, here?

(d) Use $\bar{\bar{x}}$ and your estimate of σ based on $\overline{R}$ and make retrospective control charts for $\bar{x}$ and R.

What do these indicate about the stability of the filling process over the time period represented by these data?

(e) Use $\bar{\bar{x}}$ and your estimate of σ based on $\bar{s}$ and make retrospective control charts for $\bar{x}$ and s. How do these compare in appearance to the retrospective charts for process mean and variability made in part (d)?

2. A manufacturer of U-bolts collects data on the thread lengths of the bolts that it produces. Nineteen samples of five consecutive bolts gave the thread lengths indicated the accompanying table (in .001 in. above nominal).

Sample	Thread Lengths	$\bar{x}$	R	s
1	11, 14, 14, 10, 8	11.4	6	2.61
2	14, 10, 11, 10, 11	11.2	4	1.64
3	8, 13, 14, 13, 10	11.6	6	2.51
4	11, 8, 13, 11, 13	11.2	5	2.05
5	13, 10, 11, 11, 11	11.2	3	1.10
6	11, 10, 10, 11, 13	11.0	3	1.22
7	8, 6, 11, 11, 11	9.4	5	2.30
8	10, 11, 10, 14, 10	11.0	4	1.73
9	11, 8, 11, 8, 10	9.6	3	1.52
10	6, 6, 11, 13, 11	9.4	7	3.21
11	11, 14, 13, 8, 11	11.4	6	2.30
12	8, 11, 10, 11, 14	10.8	6	2.17
13	11, 11, 13, 8, 13	11.2	5	2.05
14	11, 8, 11, 11, 11	10.4	3	1.34
15	11, 11, 13, 11, 11	11.4	2	.89
16	14, 13, 13, 13, 14	13.4	1	.55
17	14, 13, 14, 13, 11	13.0	3	1.22
18	13, 11, 11, 11, 13	11.8	2	1.10
19	14, 11, 11, 11, 13	12.0	3	1.41
$\sum\bar{x} = 212.4$		$\sum R = 77$		$\sum s = 32.92$

(a) Compute two different estimates of the process short-term standard deviation of thread length, one based on the sample ranges and one based on the sample standard deviations.

(b) Use your estimate from (a) based on sample standard deviations and compute control limits for the sample ranges R, and then compute control limits for the sample standard deviations s. Applying these to the R and s values, what is suggested about the threading process?

(c) Using a center line at $\bar{\bar{x}}$, and your estimate of σ based on the sample standard deviations, compute control limits for the sample means $\bar{x}$. Applying these to the $\bar{x}$ values here, what is suggested about the threading process?

(d) A check of the control chart form from which these data were taken shows that the coil of the heavy wire from which these bolts are made was changed just before samples 1, 9, and 16 were taken. What insight, if any, does this information provide into the possible origins of any patterns you see in the data?

(e) Suppose that a customer will purchase bolts of the type represented in the data only if essentially all bolts received can be guaranteed to have thread lengths within .01 in. of nominal. Does it appear that with proper process monitoring and adjustment, the equipment and manufacturing practices in use at this company will be able to produce only bolts meeting these standards? Explain in quantitative terms. If the equipment was not adequate to meet such requirements, name two options that might be taken and their practical pros and cons.

3. State briefly the practical goals of control charting and action on "out of control" signals produced by the charts.

4. Why might it well be argued that the name *control* chart invites confusion?

5. What must an engineering application of control charting involve beyond the simple naming of points plotting out of control if it is to be practically effective?

6. Explain briefly how a Shewhart $\bar{x}$ chart can help reduce variation in, say, a widget diameter, first by signaling the need for process intervention/adjustment and then also by preventing adjustments when no "out of control" signal is given.

7.6 Shewhart Control Charts for Qualitative and Count Data

The previous section discussed Shewhart $\bar{x}$, R, and s control charts, treating them as tools for studying the stability of a system over time. This section focuses on how the Shewhart control charting idea can be applied to attributes data (i.e., counts).

The discussion begins with p charts. Next u charts and their specialization to the case of a constant-size inspection unit, the c charts, are introduced. Finally, consideration is given to a number of common nonrandom patterns that can appear on both variables control charts and attributes control charts. Possible physical causes for them and some formal rules that are often recommended for automating their recognition are discussed.

7.6.1 *p* Charts

This text has consistently indicated that measurements are generally preferable to attributes data. But in some situations, the only available information on the stability of a process takes the form of qualitative or count data. Consideration of the topic of control charting in such situations will begin here with ***p* charts** for cases where what is available for plotting are sample fractions, $\hat{p}_i$. The most common use of this is where $\hat{p}_i$ is the fraction of a sample of n_i items that is nonconforming according to some engineering standard or specification. So this section will use the "fraction nonconforming" language, in spite of the fact that $\hat{p}_i$ can be the sample fraction having any attribute of interest (desirable, undesirable, or indifferent).

The probability facts supporting control charting for the fraction nonconforming are exactly those used in Section 6.5 to develop inference methods based on $\hat{p}$. That is, if a process is stable over time, each $n_i\hat{p}_i$ is usefully modeled as binomial (n_i, p), where p is a constant likelihood that any sampled item is nonconforming. (This section will explicitly allow for sample sizes n_i varying in time. Charts for measurements are almost always based on fairly small but constant sample sizes. But charts for attributes data typically involve larger sample sizes that sometimes vary.)

As in Section 6.5, a binomial model for $n_i\hat{p}_i$ leads immediately to

$$E\hat{p}_i = p \tag{7.93}$$

and

$$\sqrt{\operatorname{Var}\hat{p}_i} = \sqrt{\frac{p(1-p)}{n_i}} \tag{7.94}$$

But then formulas (7.93) and (7.94) suggest obvious "standards given" 3-sigma control limits for the sample "fraction nonconforming" $\hat{p}_i$. That is, if p is a standard

likelihood that any single item is nonconforming, then a "standards given" p chart has a *center line* at p and control limits

"Standards given" p chart control limits

$$LCL_{\hat{p}_i} = p - 3\sqrt{\frac{p(1-p)}{n_i}} \tag{7.95}$$

$$UCL_{\hat{p}_i} = p + 3\sqrt{\frac{p(1-p)}{n_i}} \tag{7.96}$$

In the event that formula (7.95) produces a negative value, no lower control limit is used.

Example 11

p Chart Monitoring of a Pelletizing Process

Kaminski, Rasavahn, Smith, and Weitekamper worked on the same pelletizing process already used as an example several times in this book. (See Examples 2 (Chapter 1), 14 (Chapter 3), 4 (Chapter 5), and 18 (Chapter 6).) Extensive data collection on two different days led the students to establish $p = .61$ as a standard rate of nonconforming tablets produced by the process, when run under a shop standard operating regimen. On a third day, the students took $r = 25$ samples of $n_1 = n_2 = \cdots = n_{25} = m = 30$ consecutive pellets at intervals as they came off the machine and plotted sample fractions nonconforming $\hat{p}_i$, on a "standards given" p chart made with $p = .61$. Their data are given in Table 7.18.

For samples of size $n_i = m = 30$, 3-sigma "standards given" p chart control limits are, from formulas (7.95) and (7.96),

$$LCL_{\hat{p}_i} = .61 - 3\sqrt{\frac{(.61)(1-.61)}{30}} = .34$$

$$UCL_{\hat{p}_i} = .61 + 3\sqrt{\frac{(.61)(1-.61)}{30}} = .88$$

and a center line at .61 is appropriate. Figure 7.19 is a "standards given" p chart for the data of Table 7.18.

Four $\hat{p}_i$ values plot below the lower control limit in Figure 7.19, and the $\hat{p}_i$ values run consistently below the chart's center line. These facts make untenable the hypothesis that the pelletizing process was stable at the standard value of 61% nonconforming on the day these data were gathered. In this example, points plotting "out of control" on the low side are an indication of process *improvement*. They nevertheless represent a circumstance warranting physical attention to determine the physical cause for the reduced fraction defective and possibly to learn how to make the improvement permanent.

Example 11 (*continued*)

Table 7.18
Numbers and Fractions of Nonconforming Pellets in 25 Samples of Size 30

i, Sample	$n_i\hat{p}_i$, Number Nonconforming	$\hat{p}_i$
1	13	.43
2	12	.40
3	9	.30
4	15	.50
5	17	.57
6	13	.43
7	20	.67
8	18	.60
9	18	.60
10	16	.53
11	15	.50
12	17	.57
13	15	.50
14	20	.67
15	10	.33
16	12	.40
17	17	.57
18	14	.47
19	16	.53
20	10	.33
21	14	.47
22	13	.43
23	17	.57
24	10	.33
25	12	.40
	$\sum n_i\hat{p}_i = 363$	

To make retrospective limits for a p chart, one must settle on a method of estimating the (supposedly constant) process parameter p. Here the pooling idea introduced in the two-sample context of Section 6.5 can be used. That is, as a direct extension of formula (6.71) of Section 6.5, let

Pooled estimator of a common p

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2 + \cdots + n_r\hat{p}_r}{n_1 + n_2 + \cdots + n_r} \tag{7.97}$$

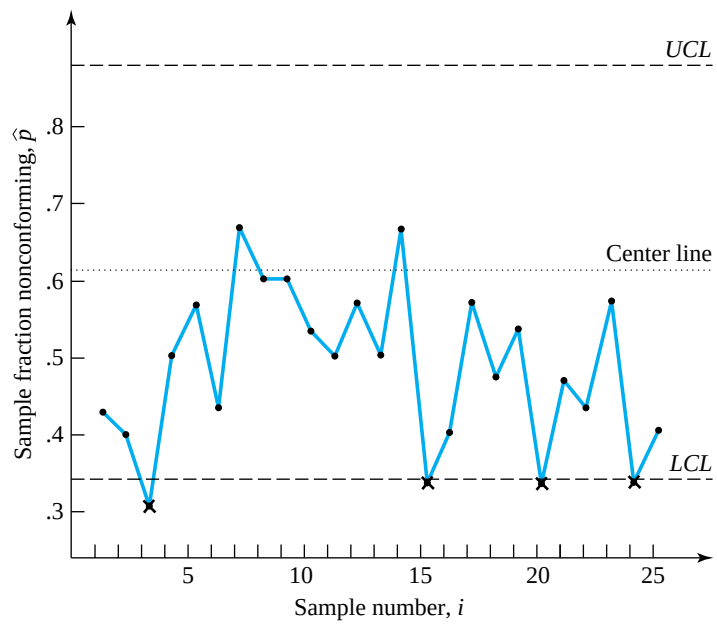

Figure 7.19 "Standards given" p chart for nonconforming pellets

($\hat{p}$ is the total number nonconforming divided by the total number inspected. When sample sizes vary, it is a weighted average of the $\hat{p}_i$.)

With $\hat{p}$ as in formula (7.97), an "as past data" Shewhart p chart has a *center line* at $\hat{p}$ and

Retrospective p chart control limits

$$LCL_{\hat{p}_i} = \hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n_i}} \tag{7.98}$$

$$UCL_{\hat{p}_i} = \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n_i}} \tag{7.99}$$

As in the "standards given" context, when formula (7.98) produces a negative value, no lower control limit is used for $\hat{p}_i$.

Example 11 (*continued*)

In the pelletizing case, the total number nonconforming in the samples was $\sum n_i \hat{p}_i = 363$. Then, since $mr = 30(25) = 750$ pellets were actually inspected

Example 11 (*continued*)

on the day in question,

$$\hat{p} = \frac{363}{750} = .484$$

So a retrospective 3-sigma p chart for the data of Table 7.18 has a center line at $\hat{p} = .484$ and, from formulas (7.98) and (7.99),

$$LCL_{\hat{p}_i} = .484 - 3\sqrt{\frac{(.484)(1 - .484)}{30}} = .21$$

$$UCL_{\hat{p}_i} = .484 + 3\sqrt{\frac{(.484)(1 - .484)}{30}} = .76$$

Figure 7.20 is a retrospective p chart for the situation of Kaminski et al. All points plot within control limits on Figure 7.20. So although it is not tenable that the pelletizing process was stable at $p = .61$ over the study period, it is completely plausible that it was stable at some value of p (and $\hat{p} = .484$ is a sensible guess for that value).

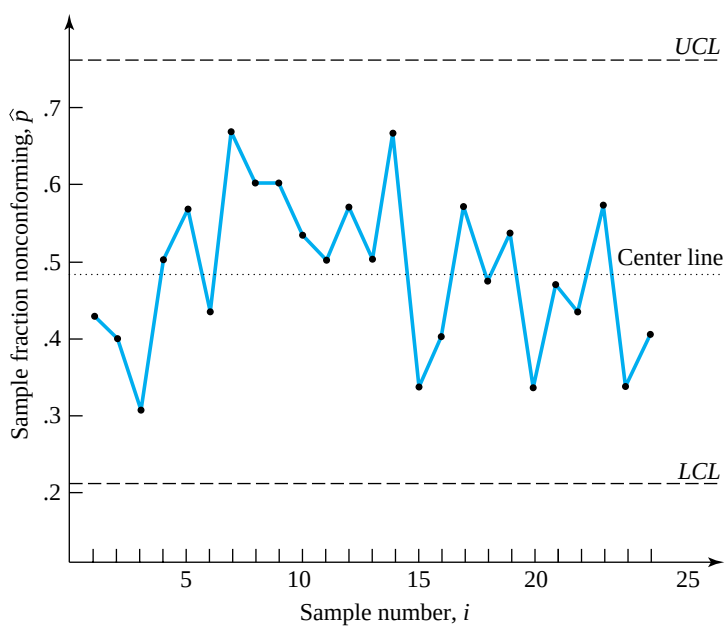

Figure 7.20 Retrospective p chart for nonconforming pellets

Because of the inherent limitations of categorical data in engineering contexts, little more will be said in this book about formal inference based on sample fractions

beyond what is in Section 6.5. For example, formal significance tests of equality of r proportions, parallel to the tests of equality of r means presented in Section 7.4, won't be discussed. However, the retrospective p chart can be interpreted as a rough graphical tool for judging how sensible the hypothesis H_0: $p_1 = p_2 = \cdots = p_r$ appears.

7.6.2 *u* Charts

Section 3.4 introduced the notation $\hat{u}$ for the ratio of the number of occurrences of a phenomenon of interest to the total number of inspection units or items sampled in contexts where there may be multiple occurrences on a given item or inspection unit. The most common application of ***u* charts** based on such ratios is that of nonconformance to some engineering standard or specification. This section will use the terminology of "nonconformances per unit" in spite of the fact that $\hat{u}$ can be the sample occurrence rate for any type of phenomenon (desirable, undesirable, or indifferent).

The theoretical basis for control charting based on nonconformances per unit is found in the Poisson distributions of Section 5.1. That is, suppose that for some specified inspection unit or unit of process output of a given size, a physically stable process has an associated mean nonconformances per unit of λ and

X_i = the number of nonconformances observed **on k_i units** inspected at time i

Then a reasonable model for X_i is often the Poisson distribution with mean $k_i\lambda$. The material in Section 5.1 then says that both $EX_i = k_i\lambda$ and $\text{Var}\, X_i = k_i\lambda$.

But notice that if $\hat{u}_i$ is the sample nonconformances per unit observed at period i,

Rate plotted on a u chart

$$\boxed{\hat{u}_i = \frac{X_i}{k_i}}$$

so Proposition 1 in Chapter 5 (page 307) can be applied to produce a mean and standard deviation for $\hat{u}_i$. That is,

$$\begin{aligned} E\hat{u}_i &= E\frac{X_i}{k_i} = \frac{1}{k_i}EX_i = \frac{1}{k_i}(k_i\lambda) = \lambda \\ \text{Var}\,\hat{u}_i &= \text{Var}\,\frac{X_i}{k_i} = \frac{1}{k_i^2}\text{Var}\,X_i = \frac{1}{k_i^2}(k_i\lambda) = \frac{\lambda}{k_i} \end{aligned} \tag{7.100}$$

so

$$\sqrt{\text{Var}\,\hat{u}_i} = \sqrt{\frac{\lambda}{k_i}} \tag{7.101}$$

The relationships (7.100) and (7.101) then motivate "standards given" 3-sigma control limits for $\hat{u}_i$. That is, if λ is a standard mean nonconformances per unit, then a "standards given" u chart has a *center line* at λ and

"Standards given" u chart control limits

$$LCL_{\hat{u}_i} = \lambda - 3\sqrt{\frac{\lambda}{k_i}} \tag{7.102}$$

$$UCL_{\hat{u}_i} = \lambda + 3\sqrt{\frac{\lambda}{k_i}} \tag{7.103}$$

The difference in formula (7.102) can turn out negative. When it does, no lower control limit is used.

Another matter of notation must be discussed at this point. λ is the symbol commonly used (as in Section 5.1) for a Poisson mean, and this fact is the basis for the usage here. However, it is more common in statistical quality control circles to use c or even c' for a standard mean nonconformances per unit. In fact, the case of the u chart where all k_i are 1 is usually referred to as a ***c* chart**. The λ notation used here represents the path of least confusion through this notational conflict and thus c or c' will not be used in this text. However, be aware that at least in the quality control world, there is a more popular alternative to the present λ convention.

When the limits (7.102) and (7.103) are used with nonconformances per unit data, one is essentially checking whether the prespecified λ is a plausible description of a physical process at each time period covered by the data. Often, however, there is no obvious standard occurrence rate λ, and u charting is to be done retrospectively. The question is then whether or not it is plausible that some (single) λ describes the process over all time periods covered by the data. What is needed in order to produce retrospective control limits for such cases is a way to use the $\hat{u}_i$ to make a single estimate of a supposedly constant λ. This text's approach to this problem is to make an estimate exactly analogous to the pooled estimate of p in formula (7.97). That is, let

Pooled estimator of a common λ

$$\hat{\lambda} = \frac{k_1\hat{u}_1 + k_2\hat{u}_2 + \cdots + k_r\hat{u}_r}{k_1 + k_2 + \cdots + k_r} \tag{7.104}$$

$\hat{\lambda}$ is the total number of nonconformances observed divided by the total number of units inspected. Then combining formula (7.104) with limits (7.102) and (7.103), a retrospective 3-sigma u chart has a *center line* at $\hat{\lambda}$ and

Retrospective u chart control limits

$$LCL_{\hat{u}_i} = \hat{\lambda} - 3\sqrt{\frac{\hat{\lambda}}{k_i}} \tag{7.105}$$

$$UCL_{\hat{u}_i} = \hat{\lambda} + 3\sqrt{\frac{\hat{\lambda}}{k_i}} \tag{7.106}$$

As the reader might by now expect, when formula (7.105) gives a negative value, no lower control limit is employed.

Example 12
(*Example 13, Chapter 3, revisited—see page 110*)

u Chart Monitoring of the Defects per Truck Found at Final Assembly

In his book *Statistical Quality Control Methods*, I. W. Burr discusses the use of u charts to monitor the performance of an assembly process at a station in a truck assembly plant. Part of Burr's data were given earlier in Table 3.19. Table 7.19 gives a (partially overlapping) $r = 30$ production days' worth of Burr's data. (The values were extrapolated from Burr's figures and the fact that truck production through sample 13 was 95 trucks/day and was 130 trucks/day thereafter. Burr gives only $\hat{u}_i$ values, production rates, and the fact that all trucks produced were inspected.)

Consider the problem of control charting for these data. Since Burr gave no figure λ for the plant's standard errors per truck, this problem will be approached as one of making a retrospective u chart. Using formula (7.104), and the column totals from Table 7.19,

$$\hat{\lambda} = \frac{\sum X_i}{\sum k_i} = \frac{6{,}078}{3{,}445} = 1.764$$

So an "as past data" u chart will have a center line at 1.764 errors/truck. From formulas (7.105) and (7.106), for the first 13 days (where each k_i was 95),

$$LCL_{\hat{u}_i} = 1.764 - 3\sqrt{\frac{1.764}{95}} = 1.355 \text{ errors/truck}$$

$$UCL_{\hat{u}_i} = 1.764 + 3\sqrt{\frac{1.764}{95}} = 2.173 \text{ errors/truck}$$

On the other hand, for the last 17 days (during which 130 trucks were produced each day),

$$LCL_{\hat{u}_i} = 1.764 - 3\sqrt{\frac{1.764}{130}} = 1.415 \text{ errors/truck}$$

$$UCL_{\hat{u}_i} = 1.764 + 3\sqrt{\frac{1.764}{130}} = 2.113 \text{ errors/truck}$$

Example 12 (*continued*)

Table 7.19
Numbers and Rates of Nonconformances for a Truck Assembly Process

i, Sample	Date	k_i, Trucks Produced	$X_i = k_i\hat{u}_i$, Errors Found	$\hat{u}_i$, Errors/Truck
1	11/4	95	114	1.20
2	11/5	95	142	1.50
3	11/6	95	146	1.54
4	11/7	95	257	2.70
5	11/8	95	185	1.95
6	11/11	95	228	2.40
7	11/12	95	327	3.44
8	11/13	95	269	2.83
9	11/14	95	167	1.76
10	11/15	95	190	2.00
11	11/18	95	199	2.09
12	11/19	95	180	1.89
13	11/20	95	171	1.80
14	11/21	130	163	1.25
15	11/22	130	205	1.58
16	11/25	130	292	2.25
17	11/26	130	325	2.50
18	11/27	130	267	2.05
19	11/29	130	190	1.46
20	12/2	130	200	1.54
21	12/3	130	185	1.42
22	12/4	130	204	1.57
23	12/5	130	182	1.40
24	12/6	130	196	1.51
25	12/9	130	140	1.08
26	12/10	130	165	1.27
27	12/11	130	153	1.18
28	12/12	130	181	1.39
29	12/13	130	185	1.42
30	12/16	130	270	2.08
		$\sum k_i = 3{,}445$	$\sum X_i = 6{,}078$	

Notice that since k_i appears in the denominator of the plus-or-minus part of control limit formulas (7.102), (7.103), (7.105), and (7.106), the larger the inspection effort at a given time period, the tighter the corresponding control limits. This is perfectly logical. A bigger "sample size" at a given period ought to make the

corresponding $\hat{u}_i$ a more reliable indicator of λ, so less variation of $\hat{u}_i$'s about a standard or estimated common value is tolerated.

Figure 7.21 is a retrospective u chart for the data of Table 7.19. The figure shows that the data-generating process can in no way be thought of as stable or subject to only random causes. There is too much variation in the $\hat{u}_i$ to be explainable as due only to small unidentifiable causes. Some of the variation can probably be thought of in terms of a general downward trend, perhaps associated with workers gaining job skills. But even accounting for that, there is substantial erratic fluctuation of the $\hat{u}_i$—which couldn't fit between control limits no matter where they might be centered. These data simply represent a real engineering process that, according to accepted standards, is not repeatable enough to allow (without appropriate sleuthing and elimination of large causes of variation) anything but "one day at a time" inferences about its behavior.

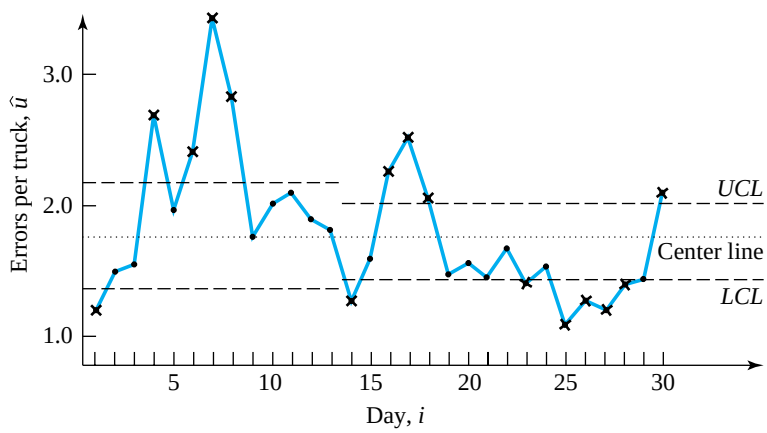

Figure 7.21 Retrospective u chart for truck assembly errors

This book has had little to say about formal inference from data with an underlying Poisson distribution. But retrospective u charts like the one in Example 12 can be thought of as rough graphical tests of the hypothesis $H_0: \lambda_1 = \lambda_2 = \cdots = \lambda_r$ for Poisson-distributed $X_i = k_i \hat{u}_i$.

7.6.3 Common Control Chart Patterns and Special Checks

Shewhart control charts (both those for measurements and those for attributes data) are useful for reasons beyond the fact that they supply semiformal information of a hypothesis-testing type. Much important qualitative information is also carried by **patterns** that can sometimes be seen in the charts' simple plots. Section 3.3 included some comments about engineering information carried in plots of summary statistics against time. Shewhart charts are such plots augmented with control limits. It is thus

appropriate to amplify and extend those comments somewhat, in light of the extra element provided by the control limits.

What is expected if a process is stable?

Before discussing interesting possible departures from the norm, it should probably be explicitly stated how a 3-sigma control chart is expected to look if a process is physically stable. One expects (tacitly assuming the distribution of the plotted statistic to be mound-shaped) that

1. most plotted points will lie in the middle, (say, the middle $\frac{2}{3}$) of the region delineated by the control limits around the center line,
2. a few (say, on the order of 1 in 20) points will lie outside this region but inside the control limits,
3. essentially no points will lie outside the control limits, and
4. there will be no obvious trends in time for any sizable part of the chart.

That is, one expects to see a random-scatter/white-noise plot that fills, but essentially remains within, the region bounded by the control limits. When something else is seen, even if no points plot outside the control limits, there is reason to consider the possibility that something in addition to chance causes is active in the data-generating mechanism.

Cyclical patterns on a control chart

Cyclical (repeated "up, then back down again") **patterns** sometimes show up on Shewhart control charts. Such behavior is not characteristic of plots resulting from a stable-process data-generating mechanism. When it occurs, the alert engineer will look for identifiable physical causes of variation whose effects would come and go on about the same schedule as the ups and downs seen on the chart. Sometimes cyclical patterns are associated with daily or seasonal variables like ambient temperature effects, which may be largely beyond a user's control. But at other times, they have to do with things like different (rotating) operators' slightly different methods of machine operation, which can be mostly eliminated via standardization, training, and awareness.

Too much variation on a control chart

Again, the expectation is that points plotted on a Shewhart control chart should (over time) pretty much fill up but rarely plot outside the region delineated by control limits. This can be violated in two different ways, both of which suggest the need for engineering attention. In the first place, more variation than expected (like that evident on Figure 7.21), which produces multiple points outside the control limits, is often termed **instability**. And (after eliminating the possibility of a blunder in calculations) it is nearly airtight evidence of one or more unregulated process variables having effects so large that they must be regulated. Such erratic behavior can sometimes be traced to material or components from several different suppliers having somewhat different physical properties and entering a production line in a mixed or haphazard order. Also, ill-advised operators may overadjust equipment (without any basis in control charting). This can take a fairly stable process and make it unstable.

Too little variation on a control chart

Less variation than expected on a Shewhart chart presents an interesting puzzle. Look again at Figure 7.16 on page 507 and reflect on the fact that the plotted $\bar{x}$'s

on that chart hug the center line. They don't come close to filling up the region between the control limits. The reader's first reaction to this might well be, "So what? Isn't small variation good?" Small variation is indeed a virtue, but when points on a control chart hug the center line, what one has is *unbelievably* small variation, which may conceal a blunder in calculation or (almost paradoxically) unnecessarily large but nonrandom variation.

In the first place, the simplest possible explanation of a plot like Figure 7.16 is that the process short-term variation, σ, has been overestimated—either because a standard σ is not applicable or because of some blunder in calculation or logic. Notice that using a value for σ that is bigger than what is really called for when making the limits

$$LCL_{\bar{x}} = \mu - 3\frac{\sigma}{\sqrt{m}} \quad \text{and} \quad UCL_{\bar{x}} = \mu + 3\frac{\sigma}{\sqrt{m}}$$

will spread the control limits too wide and produce an $\bar{x}$ chart that is insensitive to changes in μ. So this possibility should not be taken lightly.

Systematic differences and too little variation on a control chart/ stratification

A more subtle possible source of unbelievably small variation on a Shewhart chart has to do with the (usually unwitting) mixing of several consistently different streams of observations in the calculation of a single statistic that is naively thought to be representing only one stream of observations. This can happen when data are being taken from a production stream where multiple heads or cavities on a machine (or various channels of another type of multiple-channel process) are represented in a regular order in the stream. For example, items machined on heads 1, 2, and 3 of a machine might show up downstream in a production process in the order 1, 2, 3, 1, 2, 3, 1, 2, 3, etc. Then, if there is more difference between the different types of observations than there is within a given type, values of a single statistic calculated using observations of several types can be remarkably (excessively) consistent.

Consider, for example, the possibility that a five-head machine has heads that are detectably/consistently different. Suppose four of the five are perfectly adjusted and always produce conforming items and the fifth is severely misadjusted and always produces nonconforming items. Although 20% of the items produced are nonconforming, a binomial distribution model with $p = .2$ will typically overpredict the variation that will be seen in $n_i\hat{p}_i$ for samples of items from this process. Indeed, samples of size $m = 5$ of consecutive items coming off this machine will have $\hat{p}_i = .2$, always. Clearly, no $\hat{p}_i$'s would approach p chart control limits.

Or in a measurement data context, with the same hypothetical five-head machine, consider the possibility that four of the five heads always produce a part dimension at the target of 8 in. (plus or minus, say, .01 in.), whereas the fifth head is grossly misadjusted, always producing the dimension at 9 in. (plus or minus .01 in.). Then, in this exaggerated example, naive mixing together of the output of all five heads will produce ranges unbelievably stable at about 1 in. and sample means (of five consecutive pieces) unbelievably stable at about 8.2 in. But the super-stability is not a cause for rejoicing. Rather it is a cause for thought and investigation that could well lead to the physical elimination of the differences between the various mechanisms producing the data—in this case, the fixing of the faulty head.

The possibility of unnatural consistency on a Shewhart chart, brought on by more or less systematic sampling of detectably different data streams, is often called **stratification** in quality control circles. Although there is presently no way of verifying this suspicion, some form of stratification may have been at work in the production of the ream cutter data of Shervheim and Snider and the $\bar{x}$ chart in Figure 7.16. For example, multiple blades set at not quite equal angles on a roller that cuts sheets (as sketched in Figure 7.22) could produce consistently different consecutive sheet lengths and unbelievably stable $\bar{x}$'s. Or even with only a single blade on the cutter roller, regular patterns in material tension, brought on by slight eccentricities of feeder rollers, could also produce consistent patterns in consecutive sheet lengths and thus too much stability on the $\bar{x}$ chart.

Changes in level

Other nonrandom patterns sometimes appearing on control charts include both gradual and more sudden **changes in level** and unabated **trends** up or down. Gradual changes in level can sometimes be traced to machine warm-up phenomena, slow changeovers in a raw material source, or introduction of operator training. And phenomena like tool wear and machine degradation over time will typically produce patterns of plotted points moving in a single direction until there is some sort of human intervention.

Bunching

The terms **grouping** and **bunching** are used to describe irregular patterns on control charts where plotted points tend to come in sets of similar values but where the pattern is neither regular/repeatable enough to be termed cyclical nor consistent enough in one direction to merit the use of the term *trend*. Such grouping can be brought about (for example) by calibration changes in a measuring instrument and, in machining processes, by fixture changes.

Runs

Finally, **runs** of many consecutive points on one side of a center line are sometimes seen on control charts. Figure 7.15, the "standards given" $\bar{x}$ chart for the sheet-length data on page 502, is an extreme example of a chart exhibiting a run. On "standards given" charts, runs (even when not accompanied by points plotting outside control limits) tend to discredit the chart's center line value as a plausible median for the distribution of the plotted statistic. On $\bar{x}$ charts, that translates to a discrediting of the target process mean as the value of the true process mean, thus indicating that the process is misaimed. (In the sheet-length situation of Figure 7.15, average sheet length is clearly below the target length.) And on a p or u chart, it

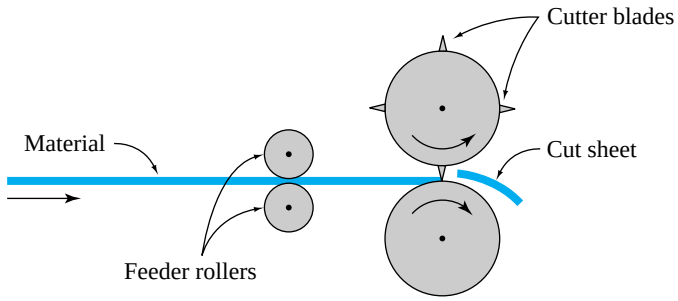

Figure 7.22 Schematic of a roller cutter

indicates the inappropriateness of the supposedly standard rate of nonconforming items or nonconformances. On retrospective control charts, runs on one side of the center line are usually matched by runs on the other side, and one of the earlier terms (cycles, trends, or grouping) can typically be applied in addition to the term *runs*.

In recognition of the fact that the elementary "wait for a point to plot outside of control limits" mode of using control charts is blind to the various interpretable patterns discussed here, a variety of **special checks** have been developed. To give the reader the flavor of these checks for unnatural patterns, two of the most famous sets are shown in Tables 7.20 and 7.21. Besides many other different sets appearing in quality control books, companies making serious use of control charts often develop their own collections of such rules. The two sets given here are included more to show what is possible than to advocate them in particular. The real bottom line of this discussion is simply that when used judiciously (overinterpretation of control chart patterns is a real temptation that also must be avoided), the qualitative information carried by patterns on Shewhart control charts can be an important engineering tool.

Table 7.20
Western Electric Alarm Rules (from the *AT&T Quality Control Handbook*)

- A single point outside 3-sigma limits
- 2 out of any 3 successive points outside 2-sigma limits on one side of the center line
- 4 out of any 5 successive points outside 1-sigma limits on one side of the center line
- 8 consecutive points on one side of the center line

Table 7.21
Alarm Rules of L. S. Nelson (from the *Journal of Quality Technology*)

- a single point outside 3-sigma limits
- 9 points in a row on one side of the center line
- 6 points in a row increasing or decreasing
- 14 points in a row alternating up and down
- 2 out of any 3 successive points outside 2-sigma limits on one side of the center line
- 4 out of any 5 successive points outside 1-sigma limits on one side of the center line
- 15 points in a row inside 1-sigma limits
- 8 points in a row with none inside 1-sigma limits

Section 6 Exercises

1. The accompanying data are some taken from *Statistical Quality Control Methods* by I. W. Burr, giving the numbers of beverage cans found to be defective in periodic samples of 312 cans at a bottling facility.

Sample	Defectives	Sample	Defectives
1	6	11	7
2	7	12	7
3	5	13	6
4	7	14	6
5	5	15	6
6	5	16	6
7	4	17	23
8	5	18	10
9	12	19	8
10	6	20	5

(a) Suppose that the company standard for the fraction of cans defective is that $p = .02$ of the cans be defective on average. Use this value and make a "standards given" p chart based on these data. Does it appear that the process fraction defective was stable at the $p = .02$ value over the period represented by these data?

(b) Make a retrospective p chart for these data. What does this chart indicate about the stability of the canning process?

2. The accompanying table lists some data on outlet leaks found in the first assembling of two radiator parts, again taken from Burr's *Statistical Quality Control Methods*. Each radiator may have several leaks.

Date	Number Tested	Leaks
6/3	39	14
6/4	45	4
6/5	46	5
6/6	48	13
6/7	40	6
6/10	58	2
6/11	50	4
6/12	50	11
6/13	50	8
6/14	50	10
6/17	32	3
6/18	50	11
6/19	33	1
6/20	50	3
6/24	50	6
6/25	50	8
6/26	50	5
6/27	50	2

(There were 841 radiators tested and a total of 116 leaks detected.) Make a retrospective u chart based on these data. What does it indicate about the stability of the assembly process?

3. In a particular defects/unit context, the number of standard size units inspected at a given opportunity varies. With

$$X_i = \text{the number of defects found on sample } i$$
$$k_i = \text{the number of units inspected at time } i$$
$$\hat{u}_i = X_i / k_i$$

the following were obtained at eight consecutive periods:

i	1	2	3	4	5	6	7	8
k_i	1	2	1	3	2	1	1	3
$\hat{u}_i$	0	1.5	0	.67	2	0	0	.33

(a) What do these values suggest about the stability of the process?

(b) Suppose that from now on, k_i is going to be held constant and that standard quality will be defined as a mean of 1.2 defects per unit. Compare 3-sigma Shewhart c charts based on

$k_i = 1$ and on $k_i = 2$ in terms of the probabilities that a given sample produces an "out of control" signal if

(i) the actual defect rate is standard.

(ii) the actual defect rate is twice standard.

4. Successive samples of carriage bolts are checked for length using "a go–no go" gauge. The results from ten successive samples are as follows:

Sample	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Sample Size	30	20	40	30	20	20	30	20	20	20
Nonconforming	2	1	5	1	2	1	3	0	1	2

What do these values indicate about the stability of the bolt cutting process?

5. Why is it essential to have an operational definition of a nonconformance to make effective practical use of a Shewhart c chart?

6. Explain why too little variation appearing on a Shewhart control chart need not be a good sign.

Chapter 7 Exercises

1. Hoffman, Jabaay, and Leuer did a study of pencil lead strength. They loaded pieces of lead of the same diameter (supported on two ends) in their centers and recorded the forces at which they failed. Part of their data are given here (in grams of load applied at failure).

4H lead	H lead	B lead
56.7, 63.8, 56.7	99.2, 99.2, 92.1	56.7, 63.8, 70.9
63.8, 49.6	106.0, 99.2	63.8, 70.9

(a) In applying the methods of this chapter in the analysis of these data, what model assumptions must be made? Make three normal plots of these samples on the same set of axes and also make a normal plot of residuals for the one-way model as means of investigating the reasonableness of these assumptions. Comment on the plots.

(b) Compute a pooled estimate of variance based on these three samples. What is the corresponding value of s_{P}?

(c) Use the value of s_{P} that you calculated in (b) and make (individual) 95% two-sided confidence intervals for each of the three mean lead strengths, $\mu_{4\mathrm{H}}$, μ_{H}, and μ_{B}.

(d) Use s_{P} and make (individual) 95% two-sided confidence intervals for each of the three differences in mean lead strengths, $\mu_{4\mathrm{H}} - \mu_{\mathrm{H}}$, $\mu_{4\mathrm{H}} - \mu_{\mathrm{B}}$, and $\mu_{\mathrm{H}} - \mu_{\mathrm{B}}$.

(e) Suppose that for some reason it is desirable to compare the mean strength of B lead to the average of the mean strengths of 4H and H leads. Give a 95% two-sided confidence interval for the quantity $\frac{1}{2}\left(\mu_{4\mathrm{H}} + \mu_{\mathrm{H}}\right) - \mu_{\mathrm{B}}$.

(f) Use the P-R method of simultaneous confidence intervals and make simultaneous 95% two-sided confidence intervals for the three mean strengths, $\mu_{4\mathrm{H}}$, μ_{H}, and μ_{B}. How do the lengths of these intervals compare to the lengths of the intervals you found in part (c)? Why is it sensible that the lengths should be related in this way?

(g) Use the Tukey method of simultaneous confidence intervals and make simultaneous 95% two-sided confidence intervals for the three differences in mean lead strengths, $\mu_{4\mathrm{H}} - \mu_{\mathrm{H}}$, $\mu_{4\mathrm{H}} - \mu_{\mathrm{B}}$, and $\mu_{\mathrm{H}} - \mu_{\mathrm{B}}$. How do the lengths of these intervals compare to the lengths of the intervals you found in part (d)?

(h) Use the one-way ANOVA test statistic and assess the strength of the evidence against $\mathrm{H}_0\colon \mu_{4H} = \mu_H = \mu_B$ in favor of H_a: not H_0. Show the whole five-step format.

(i) Make the ANOVA table corresponding to the significance test you carried out in part (h).

(j) As a means of checking your work for parts (h) and (i) of this problem, use a statistical package to produce the required ANOVA table, F statistic, and p-value.

2. Allan, Robbins, and Wyckoff worked with a machine shop that employs a CNC (computer numerically controlled) lathe in the manufacture of a part for a heavy equipment maker. Some summary statistics for measurements of a particular diameter on the part for 20 hourly samples of $m = 4$ parts turned on the lathe are given here. (The means are in 10^{-4} in. above 1.1800 in. and the ranges are in 10^{-4} in.)

Sample	**1**	**2**	**3**	**4**	**5**
$\bar{x}$	9.25	8.50	9.50	6.25	5.25
R	1	2	2	8	7

Sample	**6**	**7**	**8**	**9**	**10**
$\bar{x}$	5.25	5.75	19.50	10.0	9.50
R	5	5	1	3	1

Sample	**11**	**12**	**13**	**14**	**15**
$\bar{x}$	9.50	9.75	12.25	12.75	14.50
R	6	1	9	2	7

Sample	**16**	**17**	**18**	**19**	**20**
$\bar{x}$	8.00	10.0	10.25	8.75	10.0
R	3	0	1	3	0

(a) The midspecification for the diameter in question was 1.1809 in. Suppose that a standard σ for diameters turned on this machine is 2.5×10^{-4} in. Use these two values and find "standards given" control limits for $\bar{x}$ and R. Make both $\bar{x}$ and R charts using these and comment on what the charts indicate about the turning process.

(b) In contrast to part (a) where standards were furnished, compute retrospective or "as past data" control limits for both $\bar{x}$ and R. Make both $\bar{x}$ and R charts using these and comment on what the charts indicate about the turning process.

(c) If you were to judge the sample ranges to be stable, it would then make sense to use $\bar{R}$ to develop an estimate of the turning process short-term standard deviation σ. Find such an estimate.

(d) The engineering specifications for the turned diameter are (still in .0001 in. above 1.1800 in.) from 4 to 14. Supposing that the average diameter could be kept on target (at the midspecification), does your estimate of σ from part (c) suggest that the turning process would then be *capable* of producing most diameters in these specifications? Explain.

3. Becker, Francis, and Nazarudin conducted a study of the effectiveness of commercial clothes dryers in removing water from different types of fabric. The following are some summary statistics from a part of their study, where a garment made of one of $r = 3$ different blends was wetted and dried for 10 minutes in a particular dryer and the (water) weight loss (in grams) measured. Each of the three different garments was tested three times.

100% Cotton	Cotton/Polyester	Cotton/Acrylic
$n_1 = 3$	$n_2 = 3$	$n_3 = 3$
$\bar{y}_1 = 85.0$ g	$\bar{y}_2 = 348.3$ g	$\bar{y}_3 = 258.3$ g
$s_1 = 25.0$ g	$s_2 = 88.1$ g	$s_3 = 63.3$ g

(a) What restrictions/model assumptions are required in order to do formal inference based on the data summarized here (if information on the baseline variability involved is pooled and the formulas of this chapter are used)? Assume that those model assumptions are a sensible description of this situation.

(b) Find s_{P} and the associated degrees of freedom.

(c) What does s_{P} measure?

(d) Give a 90% lower confidence bound for the mean amount of water that can be removed from the cotton garment by this dryer in a 10-minute period.

(e) Give a 90% two-sided confidence interval for comparing the means for the two blended garments.

(f) Suppose that all pairs of fabric means are to be compared using intervals of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$ and that simultaneous 95% confidence is desired. Find Δ.

(g) A partially completed ANOVA table for testing $H_0: \mu_1 = \mu_2 = \mu_3$ follows. Finish filling in the table then find a p-value for a significance test of this hypothesis.

ANOVA Table

Source	*SS*	*df*	*MS*	*F*
	24,787			
	132,247			

4. The article "Behavior of Rubber-Based Elastomeric Construction Adhesive in Wood Joints" by P. Pellicane (*Journal of Testing and Evaluation*, 1990) compared the performance of $r = 8$ different commercially available construction adhesives. $m = 8$ joints glued with each glue were tested for strength, giving results summarized as follows (the units are kN):

Glue (i)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
$\bar{y}_i$	1821	1968	1439	616	1354	1424	1694	1669
s_i	214	435	243	205	135	191	225	551

(a) Temporarily considering only the test results for glue 1, give a 95% lower tolerance bound for the strengths of 99% of joints made with glue 1.

(b) Still considering only the test results for glue 1, give a 95% lower confidence bound for the mean strength of joints made with glue 1.

(c) Now considering only the test results for glues 1 and 2, assess the strength of the evidence against the possibility that glues 1 and 2 produce joints with the same mean strength. Show the whole five-step significance-testing format.

(d) What model assumptions stand behind the formulas you used in parts (a) and (b)? In part (c)?

For the following questions, consider test results from all eight glues when making your analyses.

(e) Find a pooled sample standard deviation and give its degrees of freedom.

(f) Repeat parts (a) and (b) using the pooled standard deviation instead of only s_1. What extra model assumption is required to do this (beyond what was used in parts (a) and (b))?

(g) Find the value of an F statistic for testing $H_0: \mu_1 = \mu_2 = \cdots = \mu_8$ and give its degrees of freedom. (*Hint:* These data are balanced. You ought to be able to use the $\bar{y}$'s and the sample variance routine on your calculator to help get the numerator for this statistic.)

(h) Simultaneous 95% two-sided confidence limits for the mean strengths for the eight glues are of the form $\bar{y}_i \pm \Delta$ for an appropriate number Δ. Find Δ.

(i) Simultaneous 95% two-sided confidence limits for all differences in mean strengths for the eight glues are of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$ for a number Δ. Find Δ.

5. Example 7 in Chapter 4 treats some data collected by Kotlers, MacFarland, and Tomlinson while studying strength properties of wood joints. Part of those data (stress at failure values in units of psi for four out of the original nine wood/joint type combinations) are reproduced here, along with $\bar{y}$ and s for each of the four samples represented:

		Wood Type	
		Pine	*Oak*
Joint Type	*Butt*	829	1169
		596	
		$\bar{y} = 712.5$	$\bar{y} = 1169$
		$s = 164.8$	
	Lap	1000	1295
		859	1561
		$\bar{y} = 929.5$	$\bar{y} = 1428.0$
		$s = 99.7$	$s = 188.1$

(a) Treating pine/butt joints alone, give a 95% two-sided confidence interval for mean strength for such joints. (Here, base your interval on only the pine/butt data.)

(b) Treating only lap joints, how strong is the evidence shown here of a difference in mean joint strength between pine and oak woods? (Here use only the pine/lap and oak/lap data.) Use the five-step format.

(c) Give a 90% two-sided confidence interval for comparing the strength standard deviations for pine/lap and oak/lap joints.

Consider all four samples in the following questions.

(d) Assuming that all four wood type/joint type conditions are thought to have approximately the same associated variability in joint strength, give an estimate of this supposedly common standard deviation.

(e) It is possible to compute simultaneous 95% lower (one-sided) confidence limits for mean joint strengths for all four wood type/joint type combinations. Give these (based on the P-R method).

(f) Suppose that you want to compare butt joint strength to lap joint strength and in fact want a 95% two-sided confidence interval for

$$\frac{1}{2}(\mu_{\text{pine/butt}} + \mu_{\text{oak/butt}}) - \frac{1}{2}(\mu_{\text{pine/lap}} + \mu_{\text{oak/lap}})$$

Give such a confidence interval, again making use of your answer to (d).

6. In an industrial application of Shewhart $\bar{x}$ and R control charts, 20 successive hourly samples of $m = 2$ high-precision metal parts were taken, and a particular diameter on the parts was measured. $\bar{x}$ and R values were calculated for each of the 20 samples, and these had

$$\bar{\bar{x}} = .35080 \text{ in.} \quad \text{and} \quad \overline{R} = .00019 \text{ in.}$$

(a) Give retrospective control limits that you would use in an analysis of the $\bar{x}$ and R values.

(b) The engineering specifications for the diameter being measured were .3500 in. $\pm$ .0020 in. Unfortunately, even practicing engineers sometimes have difficulty distinguishing in their thinking and speech between specifications and control limits. Briefly (but carefully) discuss the difference in meaning between the control limits for $\bar{x}$ found in part (a) and these engineering specifications. (To what quantities do the two apply? What are the different purposes for the two? Where do the two come from? And so on.)

7. Here are some summary statistics produced by Davies and Sehili for ten samples of $m = 4$ pin head diameters formed on a type of electrical component. The sampled components were groups of consecutive items taken from the output of a machine approximately once every ten minutes. The units are .001 in.

Sample	$\bar{x}$	R	s	Sample	$\bar{x}$	R	s
1	31.50	3	1.29	6	33.00	3	1.41
2	30.75	2	.96	7	33.00	2	.82
3	29.75	3	1.26	8	33.00	4	1.63
4	30.50	3	1.29	9	34.00	2	.82
5	32.00	0	0	10	26.00	0	0

Some summaries for the statistics are

$$\sum \bar{x} = 313.5 \quad \sum R = 22 \quad \text{and} \quad \sum s = 9.48$$

(a) Assuming that the basic short-term variability of the mechanism producing pin head diameters is constant, it makes sense to try to quantify it in terms of a standard deviation σ. Various estimates of that σ are possible. Give three such possible estimates based on $\overline{R}$, $\bar{s}$, and s_{P}.

(b) Using each of your estimates from (a), give retrospective control limits for both $\bar{x}$ and R.

(c) Compare the $\bar{x}$'s and R's given above to your control limits from (b) based on $\overline{R}$. Are there any points that would plot outside control limits on a Shewhart $\bar{x}$ chart? On a Shewhart R chart?

(d) For the company manufacturing these parts, what are the practical implications of your analysis in parts (b) and (c)?

8. Dunnwald, Post, and Kilcoin studied the viscosities of various weights of various brands of motor oil. Some summary statistics for part of their data are given here. Summarized are $m = 10$ measurements of the viscosities of each of $r = 4$ different weights of Brand M motor oil at room temperature. Units are seconds required for a ball to drop a particular distance through the oil.

10W30	SAE 30	10W40	20W50
$\bar{y} = 1.385$	$\bar{y}_2 = 2.066$	$\bar{y}_3 = 1.414$	$\bar{y}_4 = 4.498$
$s_1 = .091$	$s_2 = .097$	$s_3 = .150$	$s_4 = .204$

(a) Find the pooled sample standard deviation here. What are the associated degrees of freedom?

(b) If the P-R method is used to find simultaneous 95% two-sided confidence intervals for all four mean viscosities, the intervals produced are of the form $\bar{y}_i \pm \Delta$, for Δ an appropriate number. Find Δ.

(c) If the Tukey method is used to find simultaneous 95% two-sided confidence intervals for all differences in mean viscosities, the intervals produced are of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$, for Δ an appropriate number. Find Δ.

(d) Carry out an ANOVA test of the hypothesis that the four oil weights have the same mean viscosity.

9. Because of modern business pressures, it is not uncommon for standards for fractions nonconforming to be in the range of 10^{-4} to 10^{-6}.

(a) What are "standards given" 3-sigma control limits for a p chart with standard fraction nonconforming 10^{-4} and sample size 100?

(b) If p becomes twice the standard value (of 10^{-4}), what is the probability that the scheme from (a) detects this state of affairs at the first subsequent sample? (Use your answer to (a) and the binomial distribution for $n = 100$ and $p = 2 \times 10^{-4}$.)

(c) What does (b) suggest about the feasibility of doing process monitoring for very small fractions defective based on attributes data?

10. Suppose that a company standard for the mean number of visual imperfections on a square foot of plastic sheet is $\lambda = .04$.

(a) Give upper control limits for the number of imperfections found on pieces of material .5 ft × .5 ft and then 5 ft × 5 ft.

(b) What would you tell a worker who, instead of inspecting a 10 ft × 10 ft specimen of the plastic (counting total imperfections on the whole), wants to inspect only a 1 ft × 1 ft specimen and multiply the observed count of imperfections by 100?

11. Bailey, Goodman, and Scott worked on a process for attaching metal connectors to the ends of hydraulic hoses. One part of that process involved grinding rubber off the ends of the hoses. The amount of rubber removed is termed the skive length. The values in the accompanying table are skive length means and standard deviations for 20 samples of five consecutive hoses ground on one grinder. Skive length is expressed in .001 in. above the target length.

Sample	$\bar{x}$	s	Sample	$\bar{x}$	s
1	−.4	5.27	11	−2.2	5.50
2	0.0	4.47	12	−5.2	2.86
3	−1.4	3.29	13	−.8	1.30
4	1.8	2.28	14	.8	2.68
5	1.4	1.14	15	−2.0	2.92
6	0.0	4.24	16	−.2	1.30
7	−.4	4.39	17	−6.6	2.30
8	1.4	4.51	18	−1.0	4.21
9	.2	4.32	19	−3.2	5.76
10	−3.2	2.05	20	−2.4	4.28
				−23.4	69.07

(a) What do these values indicate about the stability of the skiving process? Show appropriate work and explain fully.

(b) Give an estimate of the process short-term standard deviation based on the given values.

(c) If specifications on the skive length are ±.006 in. and, over short periods, skive length can be thought of as normally distributed, what does your answer to (b) indicate about the

best possible fraction (for perfectly adjusted grinders) of skives in specifications? Give a number.

(d) Based on your answer to (b), give control limits for future control of skive length means and ranges for samples of size $m = 3$.

(e) Suppose that hoses from all grinders used during a given shift are all dumped into a common bin. If upon sampling, say, 20 hoses from this bin at the end of a shift, the 20 measured skive lengths have a standard deviation twice the size of your answer to (b), what possible explanations come to mind for this?

(f) Suppose current policy is to sample five consecutive hoses once an hour for each grinder. An alternative possibility is to sample one hose every 12 minutes for each grinder.
(i) Briefly discuss practical trade-offs that you see between the two possible sampling methods.
(ii) If in fact the new sampling scheme were adopted, would you recommend treating the five hoses from each hour as a sample of size 5 and doing $\bar{x}$ and R charting with $m = 5$? Explain.

12. Two different types of nonconformance can appear on widgets manufactured by Company V. Counts of these on ten widgets produced one per hour are given here.

Widget	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Type A Defects	4	2	1	2	2	2	0	2	1	0
Type B Defects	0	2	2	4	2	4	3	3	7	2
Total Defects	4	4	3	6	4	6	3	5	8	2

(a) Considering first total nonconformances, is there evidence here of process instability? Show appropriate work.

(b) What statistical indicators might you expect to observe in data like these if in fact type A and B defects have a common cause mechanism?

(c) **(Charts for Demerits)** For the sake of example, suppose that type A defects are judged twice as important as type B defects. One might then consider charting

$$\begin{aligned} X &= \text{demerits} \\ &= 2(\text{number of A defects}) \\ &\quad + (\text{number of B defects}) \end{aligned}$$

If one can model (number of A defects) and (number of B defects) as independent Poisson random variables, it is relatively easy to come up with sensible control limits. (Remember that the variance of a sum of independent random variables is the sum of the variances.)
(i) If the mean number of A defects per widget is λ_1 and the mean number of B defects per widget is λ_2, what are the mean and variance for X? Use your answers to give "standards given" control limits for X.
(ii) In light of your answer to (i), what numerical limits for X would you use to analyze these values "as past data"?

13. **(Variables Versus Attributes Control Charting)** Suppose that a dimension of parts produced on a certain machine over a short period can be thought of as normally distributed with some mean μ and standard deviation $\sigma = .005$ in. Suppose further that values of this dimension more than .0098 in. from the 1.000 in. nominal value are considered nonconforming. Finally, suppose that hourly samples of ten of these parts are to be taken.

(a) If μ is exactly on target (i.e., $\mu = 1.000$ in.), about what fraction of parts will be nonconforming? Is it possible for the fraction nonconforming ever to be any less than this figure?

(b) One could use a p chart based on $m = 10$ to monitor process performance in this situation. What would be "standards given" 3-sigma control limits for the p chart, using your answer from part (a) as the standard value of p?

(c) What is the probability that a particular sample of $m = 10$ parts will produce an "out of control" signal on the chart from (b) if μ remains at its standard value of $\mu = 1.000$ in.? How does this compare to the same probability

for a 3-sigma $\bar{x}$ chart for $m = 10$ set up with a center line at 1.000? (For the p chart, use a binomial probability calculation. For the $\bar{x}$ chart, use the facts that $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{m}$.)

(d) Compare the probability that a particular sample of $m = 10$ parts will produce an "out of control" signal on the p chart from (b) to the probability that the sample will produce an "out of control" signal on the ($m = 10$) 3-sigma $\bar{x}$ chart first mentioned in (c), supposing that in fact $\mu = 1.005$ in. What moral is told by your calculations here and in part (c)?

14. The article "How to Use Statistics Effectively in a Pseudo-Job Shop" by G. Fellers (*Quality Engineering*, 1990) discusses some applications of statistical methods in the manufacture of corrugated cardboard boxes. One part of the article concerns the analysis of a variable called box "skew," which quantifies how far from being perfectly square boxes are. This response variable, which will here be called y, is measured in units of $\frac{1}{32}$ in. $r = 24$ customer orders (each requiring a different machine setup) were studied, and from each, the skews, y, of five randomly selected boxes were measured. A partial ANOVA table made in summary of the data follows.

ANOVA Table

Source	*SS*	*df*	*MS*	*F*
Order (setup)	1052.39			
Error				
Total	1405.59	119		

(a) Complete the ANOVA table.

(b) In a given day, hundreds of different orders are run in this plant. This situation is one in which a random effects analysis is most natural. Explain why.

(c) Find estimates of σ and σ_τ. What, in the context of this situation, do these two estimates measure?

(d) Find and interpret a two-sided 90% confidence interval for σ and then the ratio σ_τ/σ.

(e) If there is variability in skew, customers must continually adjust automatic folding and packaging equipment in order to prevent machine jam-ups. Such variability is therefore highly undesirable for the box manufacturer, who wishes to please customers. What does your analysis from (c) and (d) indicate about how the manufacturer should proceed in any attempts to reduce variability in skew? (What is the big component of variance, and what kind of actions might be taken to reduce it? For example, is there a need for the immediate purchase of new high-precision manufacturing equipment?)

15. The article "High Tech, High Touch" by J. Ryan (*Quality Progress*, 1987) discusses the quality enhancement processes used by Martin Marietta in the production of the space shuttle external (liquid oxygen) fuel tanks. It includes a graph giving counts of major hardware nonconformances for each of 41 tanks produced. The accompanying data (see next page) are approximate counts read from that graph for the last 35 tanks. (The first 6 tanks were of a different design than the others and are therefore not included here.)

(a) Make a retrospective c chart for these data. Is there evidence of real quality improvement in this series of counts of nonconformances? Explain.

(b) Consider only the last 17 tanks. Does it appear that quality was stable over the production period represented by these tanks? (Make another retrospective c chart.)

(c) It is possible that some of the figures read from the graph in the original article may differ from the real figures by as much as, say, 15 nonconformances. Would this measurement error account for the apparent lack of stability you found in (a) or (b) above? Explain.

Tank	Nonconformances	Tank	Nonconformances
1	537	19	157
2	463	20	120
3	417	21	148
4	370	22	65
5	333	23	130
6	241	24	111
7	194	25	65
8	185	26	74
9	204	27	65
10	185	28	148
11	167	29	74
12	157	30	65
13	139	31	139
14	130	32	213
15	130	33	222
16	267	34	93
17	102	35	194
18	130		

16. Kaminski, Rasavahn, Smith, and Weitekamper worked with the same pelletizing machine referred to in Examples 2 (Chapter 1), 14 (Chapter 3), and 18 (Chapter 6). They collected process monitoring data on several different days of operation. The accompanying table shows counts of nonconforming pellets in periodic samples of size $m = 30$ from two different days. (The pelletizing on day 1 was done with 100% fresh material, and on the second day, a mixture of fresh and reground materials was used.)

(a) Make a retrospective p chart for the day 1 data. Is there evidence of process instability in the day 1 data? Explain.

(b) Treating the day 1 data as a single sample of size 750 from the day's production of pellets, give a 90% two-sided confidence interval for the fraction nonconforming produced on the day in question.

(c) In light of your answers to parts (a) and (b), explain why a process being in control or stable does not necessarily mean that it is producing a satisfactory fraction of conforming product.

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

(d) Repeat parts (a) and (b) for the day 2 data.

(e) Try making a single retrospective control chart for the two days taken together. Do points plot out of control on this single chart? Explain why this does or does not contradict the results of parts (a), (b), and (d).

(f) Treating the data from days 1 and 2 as two samples of size 750 from the respective days' production of pellets, give a two-sided 98% confidence interval for the difference in fractions of nonconforming pellets produced on the two days.

17. Eastman, Frye, and Schnepf counted defective plastic bags in 15 consecutive groups of 250 coming off a converting machine immediately after a changeover to a new roll of plastic. Their counts are as follows:

Sample	Nonconforming	Sample	Nonconforming
1	147	9	0
2	93	10	0
3	41	11	0
4	0	12	0
5	18	13	0
6	0	14	0
7	31	15	0
8	22		

Is it plausible that these data came from a physically stable process, or is it clear that there is some kind of start-up phenomenon involved here? Make and interpret an appropriate control chart to support your answer.

18. Sinnott, Thomas, and White compared several properties of five different brands of 10W30 motor oil. In one part of their study, they measured the boiling points of the oils. $m = 3$ measurements for each of the $r = 5$ oils follow. (Units are degrees F.)

Brand C	Brand H	Brand W	Brand Q	Brand P
378	357	321	353	390
386	365	303	349	378
388	361	306	353	381

(a) Compute and make a normal plot for the residuals for the one-way model. What does the plot indicate about the appropriateness of the one-way model assumptions?

(b) Using the five samples, find s_{P}, the pooled estimate of σ. What does this value measure? Give a two-sided 90% confidence interval for σ based on s_{P}.

(c) Individual two-sided confidence intervals for the five different means here would be of the form $\bar{y}_i \pm \Delta$, for an appropriate number Δ. If 90% individual confidence is desired, what value of Δ should be used?

(d) Individual two-sided confidence intervals for the differences in the five different means would be of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$, for a number Δ. If 90% individual confidence is desired, what value of Δ should be used here?

(e) Using the P-R method, what Δ would be used to make two-sided intervals of the form $\bar{y}_i \pm \Delta$ for all five mean boiling points, possessing simultaneous 95% confidence?

(f) Using the Tukey method, what Δ would be used to make two-sided intervals of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$ for all differences in the five mean boiling points, possessing simultaneous 99% confidence?

(g) Make an ANOVA table for these data. Then use the calculations to find both R^2 for the one-way model and also the observed level of significance for an F test of the null hypothesis that all five oils have the same mean boiling point.

(h) It is likely that the measurements represented here were all made on a single can of each brand of oil. (The students' report was not explicit about this point.) If so, the formal inferences made here are really most honestly thought of as applying to the five particular cans used in the study. Discuss why the inferences would not necessarily extend to all cans of the brands included in the study and describe the conditions under which you might be willing to make such an extension. Is the situation different if, for example, each of the measurements comes from a different can of oil, taken from different shipping lots? Explain.

19. Baik, Johnson, and Umthun worked with a small metal fabrication company on monitoring the performance of a process for cutting metal rods. Specifications for the lengths of these rods were 33.69 in. $\pm$ .03 in. Measured lengths of rods in 15 samples of $m = 4$ rods, made over a period of two

days, are shown in the accompanying table. (The data are recorded in inches above the target value of 33.69, and the first five samples were made on day 1, while the remainder were made on day 2.)

Sample	Rod Lengths	$\bar{x}$	R	s
1	.0075, .0100			
	.0135, .0135	.01113	.0060	.00293
2	−.0085, .0035			
	−.0180, .0010	−.00550	.0215	.00981
3	.0085, .0000			
	.0100, .0020	.00513	.0100	.00487
4	.0005, −.0005			
	.0145, .0170	.00788	.0175	.00916
5	.0130, .0035			
	.0120, .0070	.00888	.0095	.00444
6	−.0115, −.0110			
	−.0085, −.0105	−.01038	.0030	.00131
7	−.0080, −.0070			
	−.0060, −.0045	−.00638	.0035	.00149
8	−.0095, −.0100			
	−.0130, −.0165	−.01225	.0070	.00323
9	.0090, .0125			
	.0125, .0080	.01050	.0045	.00235
10	−.0105, −.0100			
	−.0150, −.0075	−.01075	.0075	.00312
11	.0115, .0150			
	.0175, .0180	.01550	.0065	.00297
12	.0020, .0005			
	.0010, .0010	.00113	.0015	.00063
13	−.0010, −.0025			
	−.0020, −.0030	−.00213	.0020	.00085
14	−.0020, .0015			
	.0025, .0025	.00113	.0045	.00214
15	−.0010, −.0015			
	−.0020, −.0045	−.00225	.0035	.00155
	$\bar{\bar{x}} = .00078$	$\overline{R} = .0072$	$\bar{s} = .00339$	

(a) Find a retrospective center line and control limits for all 15 sample ranges. Apply them to the ranges and say what is indicated about the rod cutting process.

(b) Repeat part (a) for the sample standard deviations rather than ranges.

The initial five samples were taken while the operators were first learning to cut these particular rods. Suppose that it therefore makes sense to look separately at the last ten samples. These samples have $\bar{\bar{x}} = -.00159$, $\overline{R} = .00435$, and $\bar{s} = .001964$.

(c) Both the ranges and standard deviations of the last ten samples look reasonably stable. What about the last ten $\bar{x}$'s? (Compute control limits for the last ten $\bar{x}$'s, based on either $\overline{R}$ or $\bar{s}$, and say what is indicated about the rod cutting process.)

As a matter of fact, the cutting process worked as follows. Rods were welded together at one end in bundles of 80, and the whole bundle cut at once. The four measurements in each sample came from a single bundle. (There are 15 bundles represented.)

(d) How does this explanation help you understand the origin of patterns discovered in the data in parts (a) through (c)?

(e) Find an estimate of the "process short-term σ" for the last ten samples. What is it really measuring in the present context?

(f) Use your estimate from (e) and, assuming that lengths of rods from a single bundle are approximately normally distributed, compute an estimate of the fraction of lengths in a bundle that are in specifications, if in fact $\mu = 33.69$ in.

(g) Simply pooling together the last ten samples (making a single sample of size 40) and computing the sample standard deviation gives the value $s = .00898$. This is much larger than any s recorded for one of the samples and should be much larger than your value from (e). What is the origin of this difference in magnitude?

20. Consider the last ten samples from Exercise 19. Upon considering the physical circumstances that produced the data, it becomes sensible to replace the control chart analysis done there with a random effects analysis simply meant to quantify

the within- and between-bundle variance components.

(a) Make an ANOVA table for these ten samples of size 4. Based on the mean squares, find estimates of σ, the standard deviation of lengths for a given bundle, and σ_τ, the standard deviation of bundle mean lengths.

(b) Find and interpret a two-sided 90% confidence interval for the ratio σ_τ/σ.

(c) What is the principal origin of variability in the lengths of rods produced by this cutting method? (Is it variability of lengths within bundles or differences between bundles?)

21. The following data appear in the text *Quality Control and Industrial Statistics* by A. J. Duncan. They represent the numbers of disabling injuries suffered and millions of man-hours worked at a large corporation in 12 consecutive months.

Month	**1**	**2**	**3**	**4**	**5**	**6**
Injuries	11	4	5	8	4	4
10^6 man-hr	.175	.178	.175	.180	.183	.198

Month	**7**	**8**	**9**	**10**	**11**	**12**
Injuries	9	12	2	6	6	7
10^6 man-hr	.210	.212	.210	.211	.195	.200

(a) Temporarily assuming the injury rate per man-hour to be stable over the period studied, find a sensible estimate of the mean injuries per 10^6 man-hours.

(b) Based on your figure from (a), find "control limits" for the observed rates in each of the 12 months. Do these data appear to be consistent with a "stable system" view of the corporation's injury production mechanisms? Or are there months that are clearly distinguishable from the others in terms of accident rates?

22. Eder, Williams, and Bruster studied the force (applied to the cutting arm handle) required to cut various types of paper in a standard paper trimmer. The students used stacks of five sheets of four different types of paper and recorded the forces needed to move the cutter arm (and thus cut the paper). The data that follow (the units are ounces) are for $m = 3$ trials with each of the four paper types and also for a "baseline" condition where no paper was loaded into the trimmer.

No Paper	Newsprint	Construction	Computer	Magazine
24, 25, 31	61, 51, 52	72, 70, 77	59, 59, 70	54, 59, 61

(a) If the methods of this chapter are applied in the analysis of these data, what model assumptions must be made? With small sample sizes such as those here, only fairly crude checks on the appropriateness of the assumptions are possible. One possibility is to compute residuals and normal-plot them. Do this and comment on the appearance of the plot.

(b) Compute a pooled estimate of the standard deviation based on these five samples. What is s_{P} supposed to be measuring in the present situation?

(c) Use the value of s_{P} and make (individual) 95% two-sided confidence intervals for each of the five mean force requirements $\mu_{\text{No paper}}$, $\mu_{\text{Newsprint}}$, $\mu_{\text{Construction}}$, μ_{Computer}, and μ_{Magazine}.

(d) Individual confidence intervals for the differences between particular pairs of mean force requirements are of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$, for an appropriate value of Δ. Use s_{P} and find Δ if individual 95% two-sided intervals are desired.

(e) Suppose that it is desirable to compare the "no paper" force requirement to the average of the force requirements for the various papers. Give a 95% two-sided confidence interval for the quantity $\mu_{\text{No paper}} - \frac{1}{4}(\mu_{\text{Newsprint}} + \mu_{\text{Construction}} + \mu_{\text{Computer}} + \mu_{\text{Magazine}})$.

(f) Use the P-R method of simultaneous confidence intervals and make simultaneous 95% two-sided confidence intervals for the five mean force requirements. How do the lengths of these intervals compare to the lengths of the intervals you found in part (c)? Why is it sensible that the lengths should be related in this way?

(g) Simultaneous confidence intervals for the differences between all pairs of mean force requirements are of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$, for an appropriate value of Δ. Use s_P and find Δ if Tukey simultaneous 95% two-sided intervals are desired. How does this value compare to the value you found in part (d)?

(h) Use the one-way ANOVA test statistic and assess the strength of the students' evidence against H_0: $\mu_{\text{No paper}} = \mu_{\text{Newsprint}} = \mu_{\text{Construction}} = \mu_{\text{Computer}} = \mu_{\text{Magazine}}$ in favor of H_a: not H_0. Show the whole five-step format.

(i) Make the ANOVA table corresponding to the significance test you carried out in part (h).

23. Duffy, Marks, and O'Keefe did some testing of the 28-day compressive strengths of 3 in. × 6 in. concrete cylinders. In part of their study, concrete specimens made with a .50 water/cement ratio and different percentages of entrained air were cured in a moisture room and subsequently strength tested. $m = 4$ specimens of each type produced the measured strengths (in 10^3 psi) summarized as follows:

3% Air	6% Air	10% Air
$\bar{y}_1 = 5.3675$	$\bar{y}_2 = 4.9900$	$\bar{y}_3 = 2.9250$
$s_1 = .1638$	$s_2 = .1203$	$s_3 = .2626$

(a) Find the pooled sample standard deviation and its associated degrees of freedom.

Use your answer to part (a) throughout the rest of this problem.

(b) Give a 99% lower confidence bound for the mean strength of 3% air specimens.

(c) Give a 99% two-sided confidence interval for comparing the mean strengths of 3% air and 10% air specimens.

(d) Suppose that mean strengths of specimens for all pairs of levels of entrained air are to be compared using intervals of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$. Find Δ for Tukey simultaneous 99% two-sided confidence limits.

(e) A partially completed ANOVA table for testing H_0: $\mu_1 = \mu_2 = \mu_3$ follows. Finish filling in the table, then find a p-value for an F test of this hypothesis.

ANOVA Table

Source	*SS*	*df*	*MS*	*F*
Total	14.1608			

24. Davis, Martin, and Poppinga used a ytterbium argon gas laser to make some cuts in stainless steel-316. Using 95 mJ/pulse and 20 Hz settings on the laser and a 15.5 mm distance to the steel specimens (set at a 45° angle to the laser beam), the students made cuts in specimens using 100, 500, and 1,000 pulses. (Although this is not absolutely clear from the students' report, it seems that four specimens were cut using each number of pulses.) The depths of cut the students measured were then as follows:

100 Pulses	500 Pulses	1000 Pulses
7.4, 8.6, 5.6, 8.0	24.2, 29.5, 26.5, 23.8	33.4, 37.5, 35.9, 34.8

(a) If the methods of this chapter are applied in the analysis of these three samples, what model assumptions must be made? Compute residuals and normal plot them as something of a check on the reasonableness of these assumptions. Comment on the appearance of the plot.

(b) Compute a pooled estimate of the standard deviation based on these three samples. What is s_P supposed to be measuring in the present situation?

(c) Make (individual) 95% two-sided confidence intervals for each of the three mean depths of cut, μ_{100}, μ_{500}, and μ_{1000}.

(d) Confidence intervals for the differences between particular pairs of mean depths of cut are of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$, for a number Δ.

Find Δ if individual 95% two-sided intervals are desired.

(e) Suppose that it is desirable to compare the per pulse change in average depth of cut between 100 pulses and 500 pulses to the per pulse change in average depth of cut between 500 pulses and 1,000 pulses. Give a 90% two-sided confidence interval for the quantity

$$\frac{1}{400}\left(\mu_{500}-\mu_{100}\right)-\frac{1}{500}\left(\mu_{1000}-\mu_{500}\right)$$

(You will need to write this out as a linear combination of the three means before applying any formulas from Section 7.2.) Based on this interval, does it appear plausible that the depth of cut changes linearly in the number of pulses over the range from 100 to 1,000 pulses? Explain.

(f) Use the P-R method of simultaneous confidence intervals and make simultaneous 95% two-sided confidence intervals for the three mean depths of cut. How do the lengths of these intervals compare to the lengths of the intervals you found in part (c)? Why is it sensible that the lengths should be related in this way?

(g) Simultaneous confidence intervals for the differences between all pairs of mean depths of cut are of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$, for a number Δ. Find Δ if Tukey simultaneous 95% two-sided intervals are desired. How does this value compare to the one you found in part (d)?

(h) Use the one-way ANOVA test statistic and assess the strength of the evidence against $H_0: \mu_1 = \mu_2 = \mu_3$. Show the whole five-step format.

(i) Make the ANOVA table corresponding to the significance test you carried out in part (h).

25. Anderson, Panchula, and Patrick tested several designs of "paper helicopters" for flight times when dropped from a point approximately 8 feet above the ground. Four different helicopters were made and tested for each design. Some summary statistics for the tests on four particular designs are given next. (The units are seconds.)

Design #1	Design #2	Design #3	Design #4
$n_1 = 4$	$n_2 = 4$	$n_3 = 4$	$n_4 = 4$
$\bar{y}_1 = 1.640$	$\bar{y}_2 = 2.545$	$\bar{y}_3 = 1.510$	$\bar{y}_4 = 2.600$
$s_1 = .096$	$s_2 = .426$	$s_3 = .174$	$s_4 = .168$

(a) Find a pooled estimate of σ in the one-way model. What does this quantity measure in the present context?

(b) Give 95% two-sided confidence limits for the mean flight time of helicopters of Design #1.

(c) P-R simultaneous two-sided 95% confidence limits for all mean flight times of the designs are of the form $\bar{y}_i \pm \Delta$. Find Δ.

(d) Give 95% two-sided confidence limits for the difference in mean flight times of helicopters of Designs #1 and #2.

(e) Tukey simultaneous two-sided 95% confidence limits for all differences in mean flight times of the designs are of the form $\bar{y}_i - \bar{y}_{i'} \pm \Delta$, for a number Δ. Find Δ.

(f) Based on your answer to part (e), do you believe that there are "statistically significant"/"statistically detectable" differences among these four designs in terms of mean flight times? Explain.

(g) Do a formal significance test of $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$. Show the whole five-step format.

(h) As a matter of fact, the four designs considered here were Design #1, 2 in. wings and 1 in. body; Design #2, 4 in. wings and 1 in. body; Design #3, 2 in. wings and 3 in. body; Design #4, 4 in. wings and 3 in. body. So the quantity

$$\frac{1}{2}\left(\mu_1+\mu_3\right)-\frac{1}{2}\left(\mu_2+\mu_4\right)$$

is a measure of the effect of changing from 2 in. wings to 4 in. wings. Give 95% two-sided confidence limits for this quantity.