9

Evaluation of measurement processes

Determining the capability of a measurement system is an important aspect of most process and quality improvement efforts. (Burdick *et al.*, 2003, p. 342)

Overview

The measure phase in a Six Sigma process improvement project is of crucial importance. It is natural therefore that teams working on projects need to be confident that the measurement processes they employ are sound and effective. Discussion of measurement processes has been postponed until this stage in the book so that use can be made of concepts from statistical models, of knowledge of designed experiments, and of components of variance in particular.

First the measurement of continuous variables will be considered. Following the introduction of the concepts of bias, linearity, repeatability and reproducibility, reference will be made to the problem of inadequate measurement units and how it manifests itself in control charts for variability. Gauge repeatability and reproducibility studies will be described and associated indices of measurement system performance introduced. Finally some reference will be made to scenarios involving attribute data.

9.1 Measurement process concepts

9.1.1 Bias, linearity, repeatability and reproducibility

A measurement system may be defined as 'the collection of instruments or gages, standards, operations, methods, fixtures, software, personnel, environment and assumptions used to quantify a unit of measure or fix assessment to the feature characteristic being measured; the

Six Sigma Quality Improvement with Minitab, Second Edition. G. Robin Henderson.

complete process used to obtain the measurement' (Automotive Industry Action Group (AIAG), 2002, p. 5).

Imagine that we wish to determine the concentration of mercury (Hg) in seawater in order to monitor effluent from a chemical manufacturing plant. Suppose, too, that we have available a batch of seawater with known mercury concentration of 30 ng/ml. There are three measurement systems available – Acme, Brill and Carat. Fifty determinations of mercury concentration were made with each system. Figure 9.1 displays the data in the form of fitted normal curves. The mean result from Acme was 30.10, which is close to the true value of 30. The mean from Brill was 29.89, which is also close to the true value. However, Carat gave a mean of 28.92: a good deal further from the true value than the means for the other two measurement systems.

The Carat result appears to show *bias*. Bias 'is the difference between the true value (reference value) and the observed average of measurements on the same characteristic on the same part' (AIAG, 2002, p. 49). Thus the bias for Carat is given by:

$$\text{Bias} = \text{Observed average value} - \text{Reference value} = 28.92 - 30.00 = -1.08.$$

This indicates that Carat yields measurements that, on average, are 1.08 ng/ml on the low side when the measurement system is used on seawater with mercury concentration of 30 ng/ml. The reader is invited to confirm that the bias for Acme and Brill is 0.10 and -0.11, respectively.

One way to deal with bias is by *calibration*, which may involve the adjustment of a gauge to account for the difference between the observed average value for a standard and the true reference value for that standard – which we have just defined as bias.

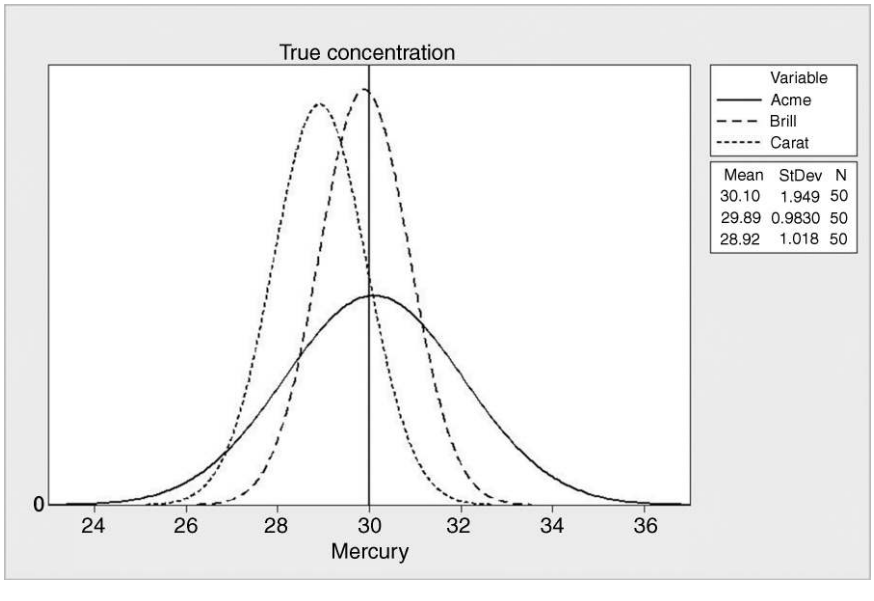

Figure 9.1 Distribution of mercury measurements.

One desirable property of gauges is good *repeatability*. Repeatability is defined as 'the variation in measurements obtained with *one measurement instrument* when used several times by *one appraiser* while measuring the identical characteristic on the *same part*' (AIAG, 2002, p. 52).

One could quote the standard deviation of such a set of measurements as a measure of variation. Historically, 5.15 times the standard deviation was used as an index of repeatability (AIAG, 2002, p. vi). The reason for this is that the interval from $\mu - 2.576\sigma$ to $\mu + 2.576\sigma$, with a range of 5.15σ, accounts for 99.00% of a normal distribution. The interval from $\mu - 3\sigma$ to $\mu + 3\sigma$, with a range of 6σ, accounts for 99.73% of a normal distribution. Minitab uses 6 times the standard deviation as the default.

Thus, adopting the Minitab default, Acme, Brill and Carat have estimated repeatability of 6×1.949, 6×0.983 and 6×1.018, i.e. 11.7, 5.9 and 6.1, respectively. Thus, if the bias of Carat could be removed by calibration, then it would be on a par with Brill in terms of repeatability.

Another desirable property of gauges is good *reproducibility*. Reproducibility is 'the variation in the average of the measurements made by *different appraisers* using the *same measuring instrument* when measuring the identical characteristic on the *same part*' (AIAG 2002, p. 53).

The upper panel in Figure 9.2 shows the distributions of measurements of the standard mercury solution where reproducibility is relatively good. All three operators, Edward, Fiona and George, employed the Brill system to make 50 measurements of the standard – the display shows normal distributions fitted to the data – and their means, indicated by fulcrums, were 29.8, 30.2 and 30.0 respectively with range 0.4. The lower panel in Figure 9.2 shows a situation where the reproducibility is relatively poor. All three operators, Una, Veronica and Walter, also employed the Brill system to make 50 measurements of the standard – the display shows

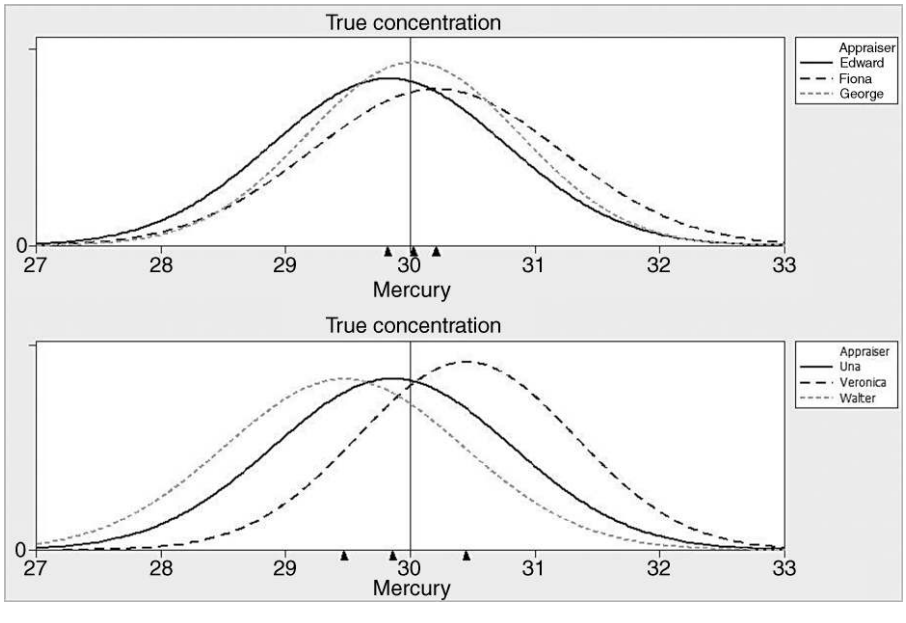

Figure 9.2 Two groups of operators with different levels of reproducibility.

normal distributions fitted to the data – and their means, indicated by fulcrums, were 29.8, 30.4 and 29.5 respectively with range 0.9.

A measurement system needs to have *stability*. 'Stability (or drift) is the total variation in the measurements obtained with a measurement system *on the same master or parts* when measuring *a single characteristic over an extended time period*' (AIAG 2002, p. 50).

Control charts may be used to monitor the stability of a measurement system. If an individual measurement of a standard or master part is made at regular intervals then an individuals or X chart may be used, with the centre line set at the reference value for the standard. If samples of measurements of a standard or master part are made at regular intervals then a mean or Xbar chart may be used with the centre line set at the reference value for the standard. A signal of special cause variation on the charts could indicate the need for calibration of the measurement system.

A final useful property of measurement systems is linearity. Linearity is an indication that 'gauge response increases in equal increments to equal increments of stimulus, or, if the gauge is biased, that the bias remains constant throughout the course of the measurement process' (NIST/SEMATECH, 2005). Readers should note that definition of linearity given by AIAG is strictly a definition of nonlinearity. Minitab provides a procedure for formally assessing linearity of a measurement system.

Suppose that the measurement systems Brill and Carat could be used with seawater samples containing up to 50 ng/ml of mercury. Standard solutions with known concentrations 10, 20, 30, 40 and 50 ng/ml are available. Five determinations of mercury concentration are made using each of the two systems. The data are tabulated in Table 9.1 and available in Mercury.MTW.

Figure 9.3 displays the data in the form of a scatterplot. The mean determination for each standard solution from each of the two systems is plotted against the true value. With no bias the means would equal the corresponding true values so the line with equation $y = x$ has been added to the diagram in order to indicate the ideal situation. The circular symbols for the Brill system are all close to the ideal line, suggesting no major bias. On the other hand, the triangular symbols for the Carat system diverge further from the ideal line as mercury concentration increases. This indicates that the Carat system is not exhibiting linearity, i.e. with Carat the bias changes with the true (reference) value.

The data in Table 9.1 are summarized in Table 9.2. The bias for each of the measurement systems is also given. The reader is invited to verify some of the bias values. There is no apparent pattern in the Brill bias values, but the Carat bias values are all negative and their magnitude generally increases as the concentration of the standard increases.

Minitab provides a formal analysis under **Stat > Quality Tools > Gage Study > Gage Linearity and Bias Study...**. The dialog box is shown in Figure 9.4. The use of the phrase 'Part numbers' reflects the widespread use of measurement process evaluation in the automotive industry. We enter **Part numbers:** 'Standard No.', **Reference values:** 'True Concentration' and **Measurement data:** 'Brill Estimate'. No entry was made in the **Process variation:** window, an option that will not be considered in this book. Information on the measurement system used etc. may be entered under **Gage Info...**, and under **Options...** one may enter a title and choose to estimate the repeatability standard deviation either using sample ranges or sample standard deviations. The default sample range method was selected, and the resulting output is shown in Figure 9.5.

For the standard solution with concentration 10 ng/ml the Brill system gave the five estimates 10.00, 10.52, 11.11, 10.30 and 11.53. The corresponding deviations from the

Table 9.1 Data from linearity investigation.

Standard no.	True concentration	Brill estimate	Carat estimate
1	10	10.00	9.04
1	10	10.52	9.16
1	10	11.11	10.59
1	10	10.30	8.85
1	10	11.53	10.58
2	20	20.76	19.45
2	20	21.61	17.92
2	20	18.22	18.51
2	20	20.44	20.97
2	20	20.46	20.48
3	30	28.95	28.42
3	30	29.46	28.36
3	30	30.52	30.46
3	30	30.69	26.76
3	30	30.05	30.30
4	40	40.70	37.54
4	40	39.67	40.02
4	40	38.35	38.37
4	40	39.39	37.48
4	40	40.51	37.86
5	50	50.37	50.00
5	50	50.69	48.24
5	50	51.31	47.73
5	50	50.67	48.79
5	50	47.61	47.29

true value of 10 are 0.00, 0.52, 1.11, 0.30 and 1.53. The mean of these gives the bias of 0.692. The five deviations (solid circles) and their mean (solid square) are plotted against the reference value of 10. Similar plotting has been done for the other four standard solutions. Thus the solid square symbols constitute the scatterplot of bias versus reference value.

The five values of bias, together with their overall mean, are shown in the bottom right hand corner of the display together with corresponding *P*-values. (The *P*-values arise from *t*-tests being performed on each sample of bias values, with null hypothesis that the mean is zero and alternative hypothesis that the mean is nonzero.) None of these are less than 0.05, so there is no evidence of bias for the Brill measurement system. Further discussion of the output in this case is unnecessary.

The corresponding output for the Carat system is shown in Figure 9.6. Here we have evidence of bias since three of the *P*-values are less than 0.05. The regression line fitted to the scatterplot of bias versus reference value has a slope that differs significantly from zero at the 5% level of significance (*P*-value of 0.029 quoted in the top right of the display). Thus we have evidence of nonlinearity here, i.e. that for the Carat system the bias is not constant but is related to the reference value. In summary, Acme is inferior to both Brill and Carat because of its inferior repeatability (see Figure 9.1). However, Carat is inferior to Brill because of its bias and nonlinearity.

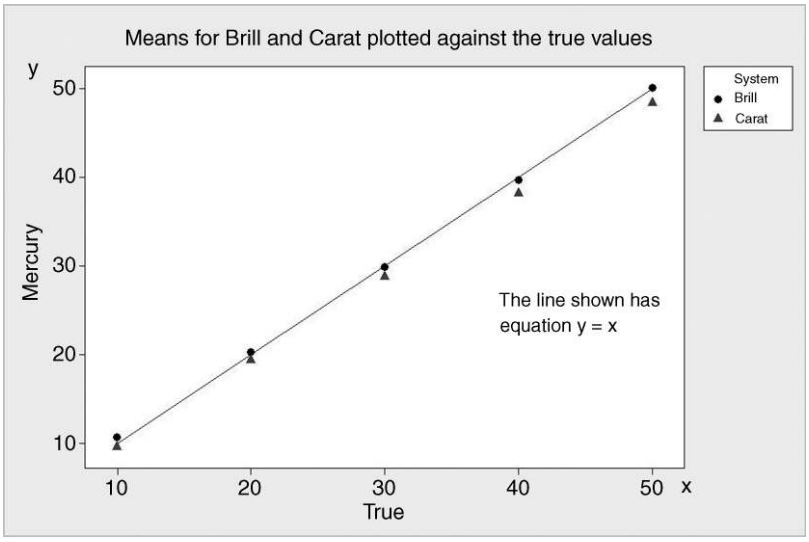

Figure 9.3 Scatterplot of means versus true values.

Table 9.2 Summarized data with bias values.

True	Brill mean	Brill bias	Carat mean	Carat bias
10	10.692	0.692	9.644	−0.356
20	20.298	0.298	19.466	−0.534
30	29.934	−0.066	28.860	−1.140
40	39.724	−0.276	38.254	−1.746
50	50.130	0.130	48.410	−1.590

9.1.2 Inadequate measurement units

Consider now some length (mm) data collected for random samples of four rods taken at 15-minute intervals from a production line. The measurement system used gives a digital display of length to two decimal places. Xbar and R charts of the data are displayed in Figure 9.7. There is no evidence from the charts of any special causes affecting the process – it appears to be behaving in a stable and predictable manner. The same data, rounded to one decimal place, give the Xbar and R charts shown in Figure 9.8. With the rounded data there are a number of signals of special cause variation on the charts. The data for both sets of charts are available in Inadequate.MTW.

Wheeler and Lyday (1989, pp. 3–9) refer to the problem of inadequate measurement units or inadequate discrimination due to a measurement unit that is too large. They state that the problem of inadequate discrimination 'begins to affect the control chart when the measurement unit exceeds the process standard deviation'. For the first pair of charts above the measurement unit was 0.01 mm whereas for the second the measurement unit was 0.1 mm.

C1	C2	C3	C4	C5
Standard No.	True Concentration	Brill Estimate	Carat Estimate	
1	10	10.00	9.04	
1	10	10.52	9.16	
1	10	11.11	10.59	
1	10	10.30	8.85	
1	10	11.53	10.58	
2	20	20.76	19.45	
2	20	21.61	17.92	

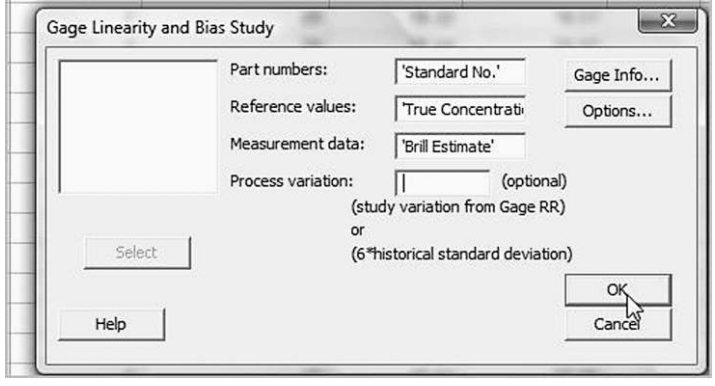

Figure 9.4 Dialog for gage linearity and bias analysis of the Brill system.

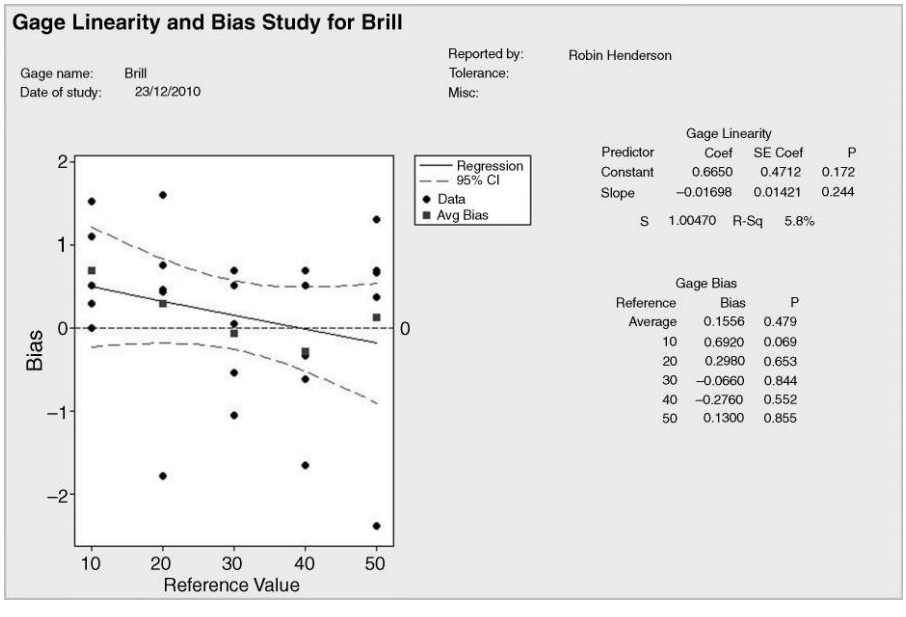

Figure 9.5 Brill linearity and bias analysis.

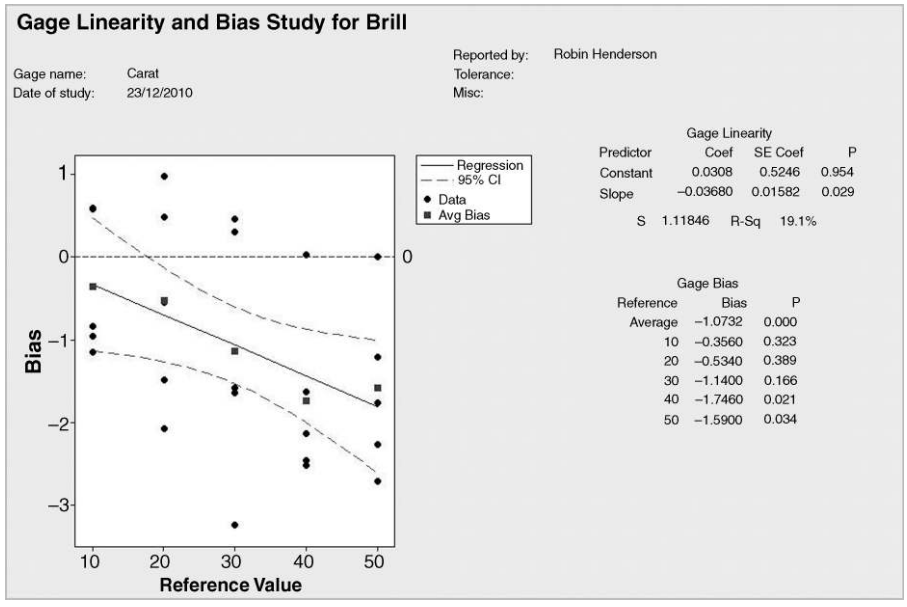

Figure 9.6 Carat linearity and bias analysis.

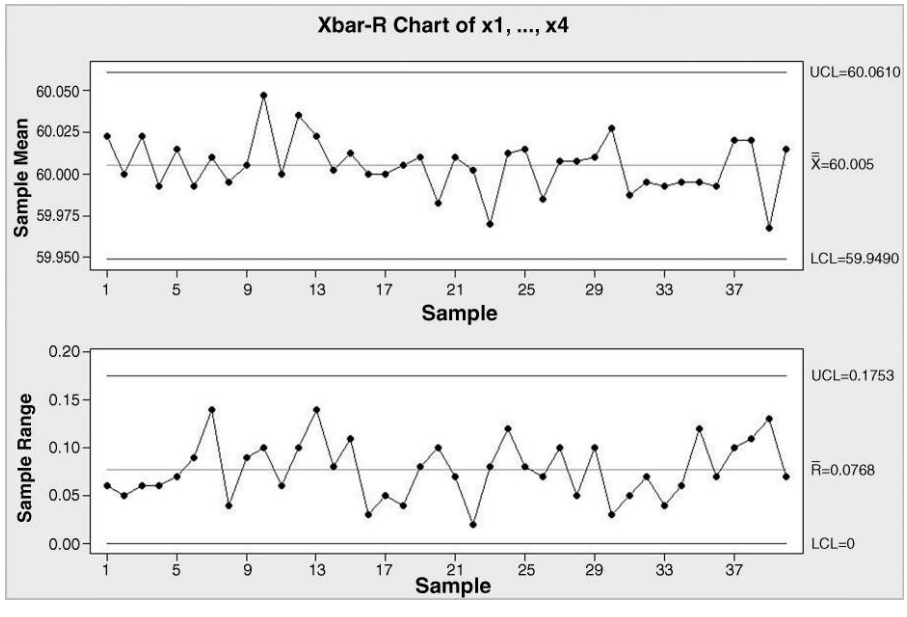

Figure 9.7 Control charts for rod length data.

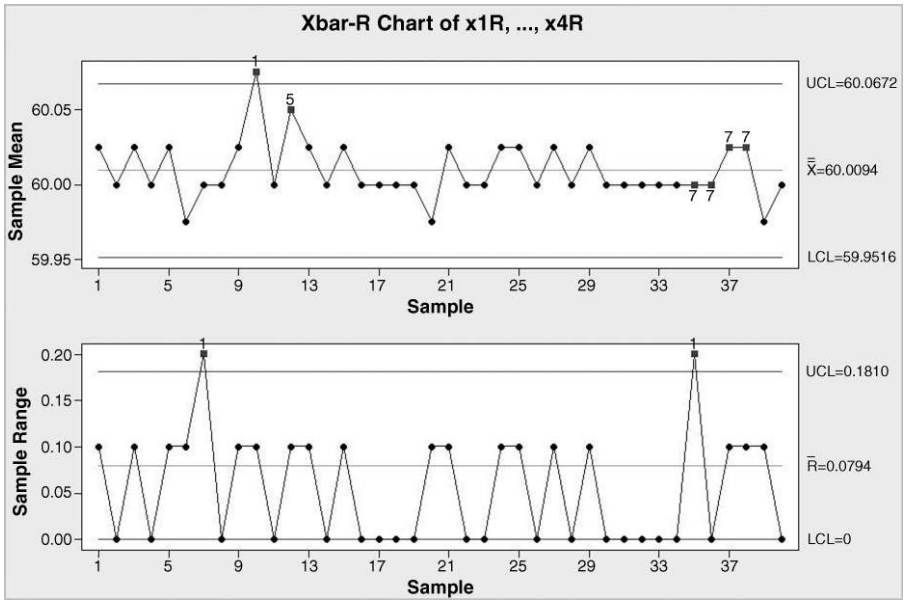

Figure 9.8 Control charts for rounded rod length data.

The process standard deviation estimated from the first range chart is approximately 0.04 mm, so for the second set of charts the measurement unit exceeds the estimated process standard deviation. As a result a process that was actually stable and predictable appeared to be subject to special cause variation.

In terms of a control chart for sample ranges, based on samples of up to 10 measurements, a measurement unit in excess of the process standard deviation generally leads to the occurrence of five or fewer values of range within the control limits. Wheeler and Lyday (1989, p. 8) comment:

The measurement unit borders on being too large when there are only 5 possible values within the control limits on the range chart. Four values within the limits will be indicative of inadequate measurement units, and fewer than four values will result in appreciable distortion of the control limits.

In the above example the fact that there are only two values of range within the control limits with the rounded data signals the inadequacy of the measurement unit. In some cases the problem of inadequate measurement units may simply be a result of failure to record enough significant figures from the measuring tool. In other cases it may be the result of a measurement system being incapable of detecting the variation in the product characteristic being measured.

9.2 Gauge repeatability and reproducibility studies

Gauge repeatability and reproducibility (R&R) studies may be used to estimate the components of variation that contribute to the overall variation in the measurements obtained from a measurement process. As a consequence, such studies enable the discriminatory power

of a measurement process to be assessed. Czitrom (1997, p. 3) wrote: 'A gage study is performed to characterize the measurement system in order to help identify areas of improvement in the measurement process and to ensure that the process signal is not obscured by noise in measurement.' In performing such studies it is usual to assume that the gauge is both bias-free and performing in a stable manner.

As an example we consider again a data set used in Chapter 8 giving the height (mm) of a set of 10 bottles measured twice using a height gauge by each of a group of three trainee Six Sigma Green Belts at Ardagh Glass Ltd., Barnsley. On each occasion the operators of the gauge measured the bottles in random order, and when measuring a bottle for the second time the operators were unaware of any previous measurements. The data are displayed in Table 9.3, available (in stacked format) in Heights.xls and reproduced by permission of Ardagh Glass Ltd., Barnsley.

Once a gauge R&R experiment has been completed, it can be informative to display the data before formal analysis is carried out. One form of display that may be used is the multi-vari chart. However, Minitab provides a special type of run chart for use with data from gauge R&R experiments. It is available using **Stat** > **Quality Tools** > **Gage Study** > **Gage Run Chart...**. The dialog is shown in Figure 9.9. Specify **Part Numbers:** Bottle, **Operators:** Operator and **Measurement data:** Height. (**Trial numbers:** and **Historical mean:** are optional.) Under **Gage Info...** details of the measurement tool used, study date etc. may be inserted and **Options...** enables a customized title for the chart to be created if desired.

The chart is displayed in Figure 9.10. Each panel of the display corresponds to a bottle. The two measurements on a bottle obtained by an operator are plotted and linked by a line segment. Note that Minitab arranges the operators in alphabetical order in the display as indicated by the key in the box to the right of the chart. Horizontal or near horizontal segments indicate good repeatability. Widely separated segments in panels would indicate poor reproducibility. Scrutiny of the chart suggests that Paul had the best performance in terms of repeatability.

Formal analysis involves estimation of the components of variance. The variance components obtained in Section 8.1.4 using ANOVA are shown in Panel 9.1. (It was concluded that the operator–bottle component of variation was zero.)

Table 9.3 Bottle height data.

	Neil		Lee		Paul	
Bottle no.	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
1	214.82	214.83	214.84	214.89	214.84	214.84
2	214.61	214.64	214.65	214.62	214.60	214.58
3	214.52	214.53	214.51	214.57	214.51	214.51
4	214.57	214.61	214.61	214.62	214.61	214.61
5	214.64	214.64	214.64	214.65	214.64	214.65
6	214.72	214.73	214.72	214.73	214.72	214.72
7	214.60	214.61	214.63	214.63	214.61	214.62
8	214.73	214.75	214.74	214.76	214.73	214.73
9	214.70	214.67	214.70	214.67	214.66	214.68
10	214.80	214.78	214.81	214.81	214.78	214.79

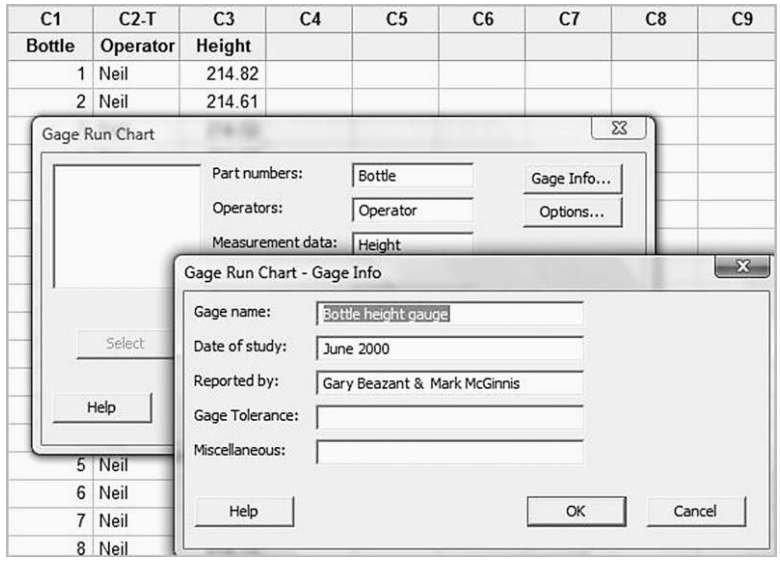

Figure 9.9 Dialog for creation of gauge run chart.

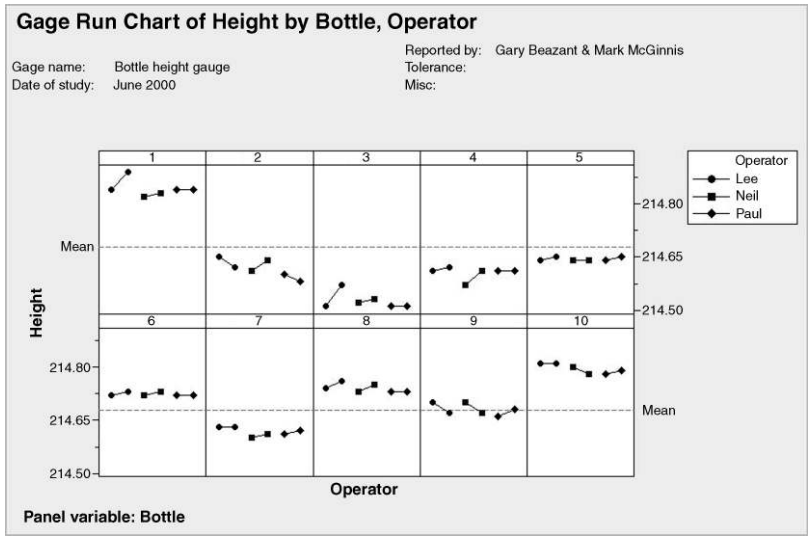

Figure 9.10 Gauge run chart.

Source	Estimated Value
Bottle	0.00934
Operator	0.00009
Error	0.00023

Panel 9.1 Components of variance for height gauge R&R study.

The total variation observed in measurements is partitioned into a component due to part-to-part variation and a component due to measurement system variation:

$$\sigma^2_{\text{Tot}} = \sigma^2_{\text{Part}} + \sigma^2_{\text{MS}}.$$

The measurement system variation is also referred to as total gauge R&R and is partitioned into a repeatability component (i.e. a component due to the gauge or measurement tool) and a reproducibility component (i.e. a component due to the operators, the users of the gauge or measurement tool):

$$\sigma^2_{\text{MS}} = \sigma^2_{\text{Repeat}} + \sigma^2_{\text{Reprod}}.$$

The reproducibility component is simply the variation due to operator if there is no operator–part interaction. When there is an operator–part interaction then the reproducibility component is given by

$$\sigma^2_{\text{Reprod}} = \sigma^2_{\text{Oper}} + \sigma^2_{\text{Oper}\times\text{Part}}.$$

The sources of variation and these formulae have been tabulated in Table 9.4. The variance components obtained directly from the ANOVA have been entered in bold, and the reader should check the calculation of the other entries in the column. The final column gives the standard deviations.

One widely used index in the evaluation of measurement processes is the proportion of the total variation that may be attributed to the measurement system (repeatability and reproducibility) variation. Variation is measured by standard deviation in this context, and it is usual to express the proportion as *percentage gauge R&R* (%R&R):

$$\begin{aligned}\%\text{R\&R} &= \frac{\sigma_{\text{MS}}}{\sigma_{\text{Tot}}} \times 100 \\ &= \frac{0.0179}{0.0983} \times 100 = 18\%.\end{aligned}$$

for the bottle height measurement system.

Table 9.4 Sources of variation for height gauge R&R study.

Source	Symbols and formulae	Variance component	Standard deviation
Part-to-part	σ^2_{Part}	**0.009 34**	0.0966
Operator	σ^2_{Oper}	**0.000 09**	0.0095
Operator × Part	$\sigma^2_{\text{Oper}\times\text{Part}}$	**0.000 00**	0.0000
Reproducibility	$\sigma^2_{\text{Reprod}} = \sigma^2_{\text{Oper}} + \sigma^2_{\text{Oper}\times\text{Part}}$	0.000 09	0.0095
Repeatability	σ^2_{Repeat}	**0.000 23**	0.0152
Total gauge R&R	$\sigma^2_{\text{MS}} = \sigma^2_{\text{Repeat}} + \sigma^2_{\text{Reprod}}$	0.000 32	0.0179
Total	$\sigma^2_{\text{Tot}} = \sigma^2_{\text{Part}} + \sigma^2_{\text{MS}}$	0.009 66	0.0983

The following guidelines for the acceptance of %R&R are given:

- Under 10% – generally considered to be an acceptable measurement system.
- 10% to 30% – may be acceptable based upon importance of application, cost of measurement device, cost of repair, etc.
- Over 30% – considered to be not acceptable – every effort should be made to improve the measurement system. (AIAG, 2002, p. 77)

Thus the bottle height measurement process falls into the middle of these three categories. These guidelines suggest that the measurement system could be improved. However, these guidelines are somewhat arbitrary and 'excessively conservative' according to Wheeler's (2003) critical review of them.

The determination of %R&R may be done directly in Minitab using **Stat** > **Quality Tools** > **Gage Study** > **Gage R&R Study (Crossed)...**. The study is said to be crossed since every bottle was measured by every operator. The dialog is basically the same as in Figure 9.9; the default **Method of Analysis**, **ANOVA**, was accepted, as were the defaults under **Options...**. In addition to the analysis in the Session window, a number of displays are provided – see Figure 9.11. (If desired the six plots may be displayed separately via **Options...**.)

The height by bottle and height by operator plots are main effects plots that enable the mean measurements for each bottle and for each operator respectively to be compared. The operator–bottle interaction plot gives a visual indication of the presence or absence of an interaction effect. The Xbar chart by operator displays the mean result for each bottle for each operator. However, unlike a process monitoring control chart of means where signals of possible special cause variation are generally bad news, in this scenario the more points that plot outwith the chart limits the better the measurement system – it is desirable that a measurement system signals differences between parts! The R chart by operator highlights

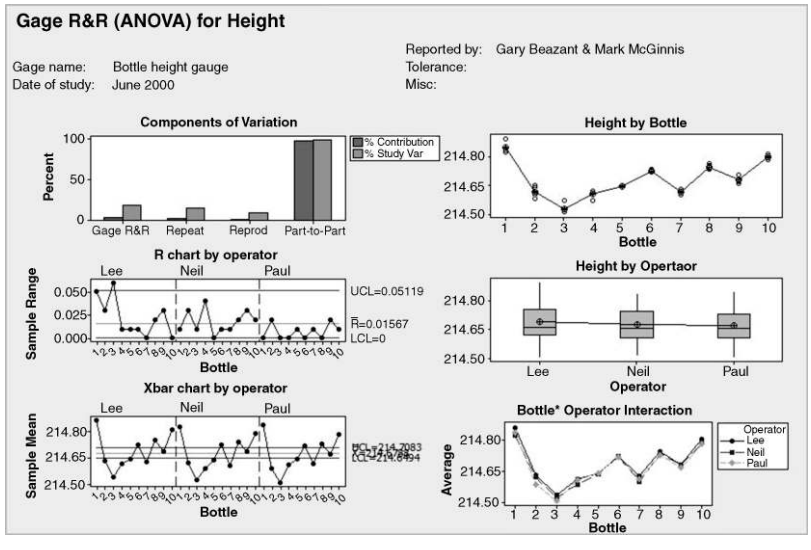

Figure 9.11 Displays from gauge R&R.

a suspicion that arose from scrutiny of the run chart in Figure 9.10, i.e. that Paul had the best repeatability. The signal for Lee provides evidence that he has significantly worse repeatability than the other two operators. Thus there may be a training issue that could be addressed.

The bar chart shows the components of variance and the matching standard deviations expressed as a percentage of total variance and total standard deviation, respectively. The second bar from the left represents the %R&R index. The ANOVA part of the Session window output is displayed in Panel 9.2. Minitab carries out the full ANOVA, with an interaction term. If the *P*-value for interaction exceeds 0.25 then the interaction term is removed from the model and the ANOVA re-calculated. (If desired the default value of 0.25 may be changed under **Options...**.) The ANOVA adopted is then used to compute the components of variance (as displayed in Panel 9.1) and to perform the calculations displayed in Table 9.4. The remaining portion of Session window output is shown in Panel 9.3.

The Total Gage R&R, expressed as a percentage of what Minitab refers to as study variation, and given as 17.96% in Panel 9.3, is the %R&R computed earlier as 18% from Table 9.4. (The reader should note that the percentages in the %Study Var column in the output do not sum to 100 whereas those in the %Contribution column do – this is because variances are additive but standard deviations are not.) At the foot of Panel 9.3 we are informed that the number of distinct categories that can be reliably differentiated by the measurement process is 7. This is another widely used index of measurement system performance and it is generally recommended that it should be 5 or more (AIAG, 2002, p. 45). The number of distinct categories is the rounded value of the discrimination ratio. Wheeler and Lyday (1989, pp. 54–59) give a detailed discussion of the discrimination ratio and its interpretation.

Gage R&R Study - ANOVA Method

```
Gage R&R for Height

Gage name:       Bottle height gauge
Date of study:   June 2000
Reported by:     Gary Beazant & Mark McGinnis
Tolerance:
Misc:
```

Two-Way ANOVA Table With Interaction

Source	DF	SS	MS	F	P
Bottle	9	0.506302	0.0562557	273.430	0.000
Operator	2	0.003863	0.0019317	9.389	0.002
Bottle * Operator	18	0.003703	0.0002057	0.863	0.621
Repeatability	30	0.007150	0.0002383		
Total	59	0.521018			

```
Alpha to remove interaction term = 0.25
```

Two-Way ANOVA Table Without Interaction

Source	DF	SS	MS	F	P
Bottle	9	0.506302	0.0562557	248.797	0.000
Operator	2	0.003863	0.0019317	8.543	0.001
Repeatability	48	0.010853	0.0002261		
Total	59	0.521018			

Panel 9.2 ANOVA tables from Minitab gauge R&R analysis.

Gage R&R

Source	VarComp	%Contribution (of VarComp)
Total Gage R&R	0.0003114	3.23
Repeatability	0.0002261	2.34
Reproducibility	0.0000853	0.88
Operator	0.0000853	0.88
Part-To-Part	0.0093383	96.77
Total Variation	0.0096497	100.00

Source	StdDev (SD)	Study Var (6 * SD)	%Study Var (%SV)
Total Gage R&R	0.0176462	0.105877	**17.96**
Repeatability	0.0150370	0.090222	15.31
Reproducibility	0.0092346	0.055408	9.40
Operator	0.0092346	0.055408	9.40
Part-To-Part	0.0966347	0.579808	98.37
Total Variation	0.0982327	0.589396	100.00

Number of Distinct Categories = 7

Panel 9.3 Session window output from gauge R&R analysis.

Gage R&R Study (Crossed) would be used when each part is measured on more than one occasion by each operator. Minitab also provides, via **Stat** > **Quality Tools** > **Gage Study** > **Gage R&R Study (Nested)...**, analysis for situations in which each operator measures a specific set of parts unique to that operator. Skrivanek (2009) gives an example of the evaluation of the measurement system used to test the hardness of a reinforced plastic part of a prosthetic device. The test involved subjecting a sample of randomly selected parts to a specified force. A durometer was then used to measure the depth of the indentation in the material created by the force and the score recorded. The parts for the devices are supplied in lots of 50. Since the sampled parts are effectively destroyed during measurement, a nested analysis is appropriate. Three appraisers were selected at random. Five lots were selected to represent the full range of the manufacturing process. The data are available in Hardness.MTW and are reproduced by permission of MoreSteam.com LLC.

A gauge run chart may be created, and the nested analysis also provides a number of displays as in Figure 9.11 for the crossed case. Using the defaults for the nested analysis, the Session window output in Panel 9.4 was obtained. We have a percentage gauge R&R of 26.97 and five distinct categories, indicating a measurement process with marginal performance.

Source	StdDev (SD)	Study Var (6 * SD)	%Study Var (%SV)
Total Gage R&R	0.93095	5.5857	26.97
Repeatability	0.93095	5.5857	26.97
Reproducibility	0.00000	0.0000	0.00
Part-To-Part	3.32415	19.9449	96.30
Total Variation	3.45205	20.7123	100.00

Number of Distinct Categories = 5

Panel 9.4 Session window output from Gage R&R analysis (nested).

Minitab also enables factors other than part and operator to be taken into account in assessing measurement system performance via **Stat > Quality Tools > Gage Study > Gage R&R Study (Expanded)...**. Further information may be found at http://www.minitab.com/en-US/training/articles/articles.aspx?id=8900. Further general advice and guidance on the design and analysis of measurement systems capability may be found in the review paper by Burdick *et al.* (2003) and the book by Burdick *et al.* (2005).

9.3 Comparison of measurement systems

Situations arise where it is desirable to compare the performance of two measurement systems. For example, a supplier and customer may measure parts using their own systems and it may be of mutual benefit to compare the measurements obtained by each on a set of parts, or a new system might be under consideration for purchase and it could be of interest to compare its performance with that of the system currently in use.

The worksheet Outside_Diameters.MTW contains the diameters of 30 parts measured in random order by one appraiser, using both system A and system B. A useful initial display is a scatterplot of the measurement obtained from system B plotted against that obtained from system A. In the ideal situation both systems would be bias-free and the two measurements would be identical. Thus it can be informative to plot the line $y = x$ on the scatterplot. Once the basic scatterplot has been created the line may be added by right-clicking the graph and using **Add > Calculated Line...**. Under **Coordinates** selection of the column containing the results from the first system as both the **Y column:** and the **X column:** yields the required line. The result is displayed in Figure 9.12.

The mouse pointer is shown located at the point corresponding to the first part. The point lies above the line, an indication that the measurement on that part from system B (47.956)

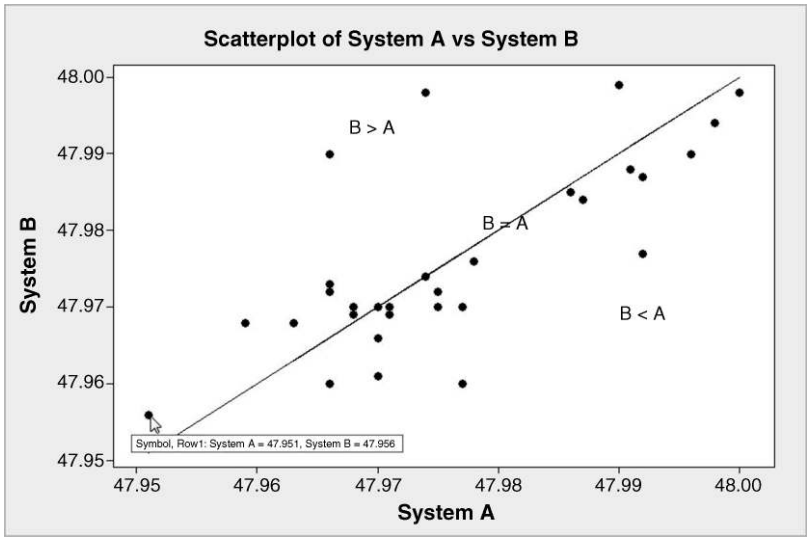

Figure 9.12 Scatterplot of measurements by two systems.

exceeds that from system A (47.951). Two points fall on the line, indicating that for parts 3 and 14 both systems gave the same measurement. Nearly twice as many points lie below the line as lie above it, suggesting the possibility of some relative bias for the two systems. Bland and Altman (1986) recommend plotting the difference between the two measurements on a part (A − B) versus the average of the two measurements (A + B)/2. They state: 'The plot of difference against mean also allows us to investigate any possible relationship between the measurement error and the true value. We do not know the true value, and the mean of the two measurements is the best estimate we have.' Such a plot is widely referred to as a Bland–Altman plot. They also advise display of the differences in a histogram – see Figures 9.13 and 9.14.

Bland–Altman plots are not available directly in Minitab. They may be created by adding reference lines to a scatterplot of difference against average or using a Minitab macro. Details of how to access and run the macro, which was created by Eli Walters, are given in Chapter 11 and at http://www.minitab.com/en-US/support/answers/answer.aspx?id=2504.

The mean of the differences is 0.000 10 and a reference line indicates this value, rounded to two decimal places. This value is close to zero and a formal t-test of the null hypothesis that the mean difference is zero yields a P-value of 0.951, so there is no evidence of relative bias for the two methods. Upper and lower limits of agreement, labelled ULA and LLA respectively, are also shown. The standard deviation of the differences is 0.008 89 and the limits are

$$\begin{aligned}&\text{Mean} \pm 1.96 \times \text{Standard deviation}\\ &= 0.000\,10 \pm 1.96 \times 0.008\,89\\ &= (-0.017\,32, 0.017\,52),\end{aligned}$$

or (− 0.017, 0.018) to three decimal places. If it is reasonable to consider the differences to be normally distributed then the calculated limits of agreement would be expected to contain

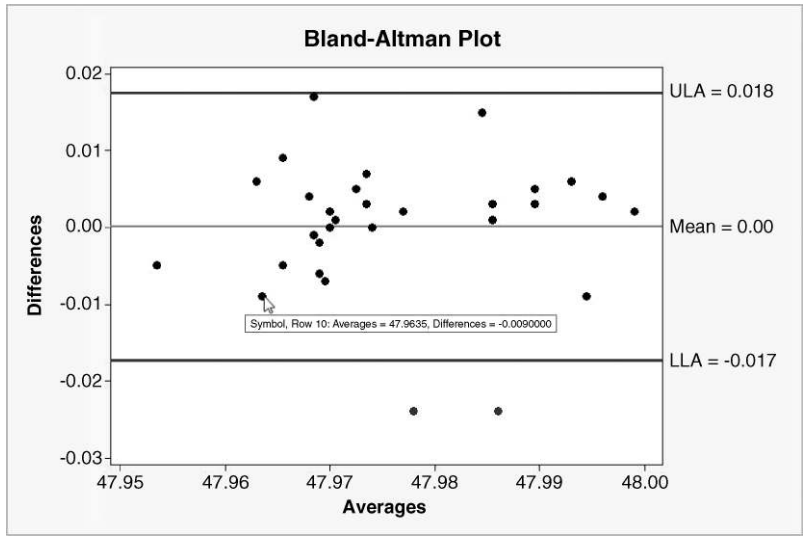

Figure 9.13 Bland–Altman plot of diameter measurements.

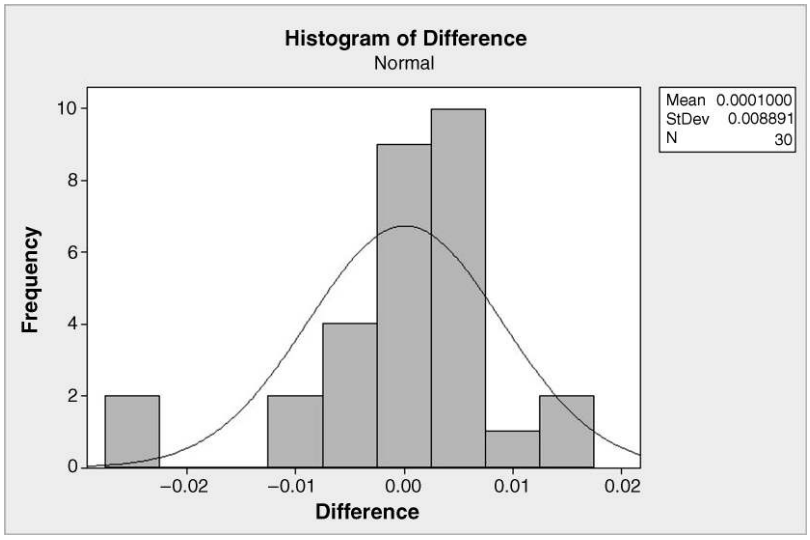

Figure 9.14 Histogram of the differences.

around 95% of the plotted points – in this case 93% of points fall between the limits. The relatively poor fit of the normal curve in Figure 9.14 casts doubt on this assumption – doubt that is confirmed by a normal probability plot. Ideally we would wish see the points in the Bland–Altman plot form a randomly distributed horizontal band of uniform width. A horizontal band implies that there is no relationship between difference in measurements and the size of the component, represented by the average of the two measurements. Correlation may be used to test formally for the presence of such a relationship. Here the correlation between difference and average is 0.015 with *P*-value 0.939, so there is no evidence of a linear relationship.

Setting aside the reservation about normality, the fundamental question is whether or not the limits of accuracy imply satisfactory repeatability for the two systems. Provided that differences in the range −0.017 to 0.018 are not of manufacturing or functional significance, the two measurement systems could be used interchangeably to make diameter measurements on the components. In the follow-up example at the end of the chapter an approximate method for the computation of confidence intervals for the limits of agreement is given. Further guidance is provided in the paper by Bland and Altman (1986) that is freely available at http://www-users.york.ac.uk/∼mb55/meas//ba.htm.

9.4 Attribute scenarios

Specimens of silicon are inspected before dispatch to a customer. Twenty-five specimens, which have been classified as either accept or reject by senior inspectors, were assessed and classified by two trainee inspectors. The data are available in Silicon.MTW. Use of **Stat** > **Quality Tools** > **Attribute Agreement Analysis...** provides informative analyses and displays. A portion of the data and the dialog are shown in Figure 9.15. Column C1 contains the reference number of the specimens, C2 the classification made by the experts (the

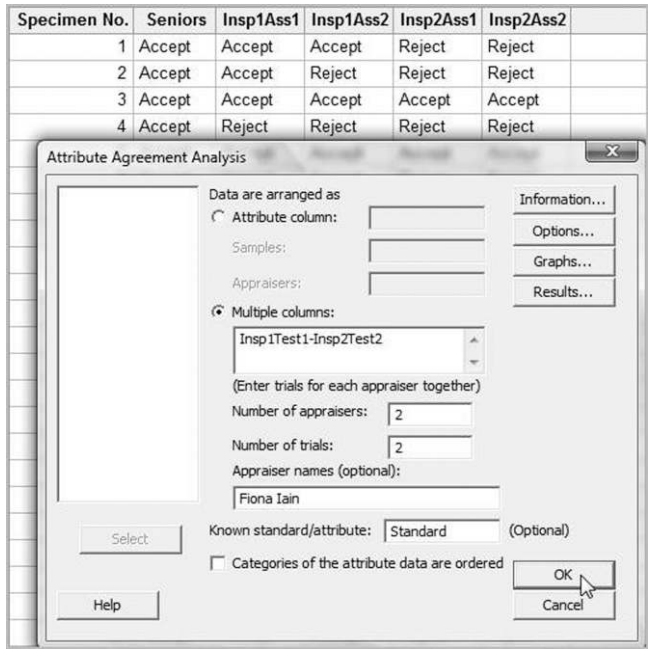

Figure 9.15 Data and dialog for attribute agreement analysis.

standard), C3 the first classification made by the first trainee, C4 the second classification made by the first trainee, and C5 and C6 contain the classifications made by the second trainee. For example, the senior inspectors deemed that the second specimen was acceptable; the first trainee deemed it as acceptable on the first assessment but as a reject on a second assessment, whereas the second trainee deemed it a reject on both his assessments.

The left-hand chart in Figure 9.16, labelled Within Appraisers, gives the proportion of assessments for each assessor where their first assessment agrees with their second. For example, Fiona's first assessment was the same as her second for 21 of the 25 specimens. This corresponds to 84% agreement, indicated by the solid circle. The line segment gives a confidence interval for the true proportion, for the first appraiser, of 64% to 95%. Similarly, Iain had 92% agreement with 95% confidence interval 74% to 99%. The corresponding Session window output for this first display is shown in Panel 9.5.

Fleiss' kappa statistic ranges from -1 to $+1$. The higher the value of kappa, the stronger is the agreement between the ratings. When kappa $= 1$ there is perfect agreement. When kappa $= 0$, then agreement is equivalent to that expected by chance. Minitab Help on this topic states that, as a general guide, kappa values less than 0.7 indicate that the measurement system requires improvement, while kappa values in excess of 0.9 are desirable. Dunn (1989, p. 37) discusses the benchmarks for the evaluation of kappa values displayed in Table 9.5. These benchmarks were proposed by Landis and Koch (1977), although Dunn comments that 'In the view of the present author these are too generous, but any series of standards such as these are bound to be subjective'.

The second chart displayed in Figure 9.16, labelled Appraisers vs. Standard, displays the proportions, with 95% confidence intervals, for agreement between appraisers and

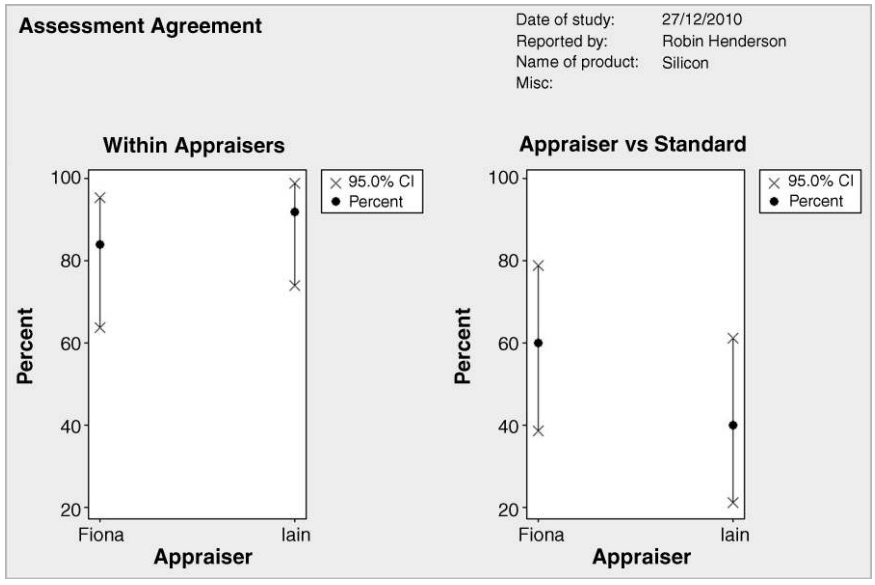

Figure 9.16 Displays from attribute agreement analysis.

Within Appraisers

Assessment Agreement

Appraiser	# Inspected	# Matched	Percent	95% CI
Fiona	25	21	84.00	(63.92, 95.46)
Iain	25	23	92.00	(73.97, 99.02)

Matched: Appraiser agrees with him/herself across trials.

Fleiss' Kappa Statistics

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)
Fiona	Accept	0.652778	0.2	3.26389	0.0005
	Reject	0.652778	0.2	3.26389	0.0005
Iain	Accept	0.750000	0.2	3.75000	0.0001
	Reject	0.750000	0.2	3.75000	0.0001

Panel 9.5 Session window output for attribute agreement analysis.

Table 9.5 Benchmarks for the evaluation of kappa values.

Value of kappa	Strength of agreement
0.00	Poor
0.01–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.60–0.80	Substantial
0.81–1.00	Almost perfect

Each Appraiser vs Standard

```
Assessment Agreement

Appraiser  # Inspected  # Matched  Percent      95% CI
Fiona               25         15    60.00  (38.67, 78.87)
Iain                25         10    40.00  (21.13, 61.33)

# Matched: Appraiser's assessment across trials agrees with the known standard.


Assessment Disagreement

           # Reject /           # Accept /
Appraiser      Accept  Percent      Reject  Percent  # Mixed  Percent
Fiona               4    22.22           2    28.57        4    16.00
Iain               13    72.22           0     0.00        2     8.00


# Reject / Accept:  Assessments across trials = Reject / standard = Accept.
# Accept / Reject:  Assessments across trials = Accept / standard = Reject.
# Mixed: Assessments across trials are not identical.


Fleiss' Kappa Statistics

Appraiser  Response      Kappa  SE Kappa         Z  P(vs > 0)
Fiona      Accept     0.255952  0.141421   1.80986     0.0352
           Reject     0.255952  0.141421   1.80986     0.0352
Iain       Accept    -0.129079  0.141421  -0.91273     0.8193
           Reject    -0.129079  0.141421  -0.91273     0.8193
```

Panel 9.6 Session window output for attribute agreement analysis.

the assessment made by the experts recorded in the column named Seniors. For 15 of the 25 specimens Fiona agreed with the experts on both tests. Thus her proportion for agreement with the standard was 60%, with 95% confidence interval 39% to 79%. Iain's proportion was 40%, with 95% confidence interval 21% to 61%. The corresponding Session window output for the second display is shown in Panel 9.6.

In addition to the agreement proportions and confidence intervals shown in the graphical display, an analysis of disagreement with the standard is given in the Session window output for each appraiser. In the case of Fiona, there were four specimens where she made the decision to reject on both tests, when the standard was to accept. The shorthand '# Reject/Accept' indicates the number of specimens for which the decision made by the appraiser was reject across all trials given that the standard specified the specimen as accept. The '/' symbol may be interpreted as the phrase 'given that' since what is being quoted are estimates of conditional probabilities. There were two specimens where she made the decision to accept on both tests when the standard was to reject. There were four specimens where she made the decision to reject on one of the tests and to accept on the other, regardless of the standard. Minitab refers to these cases as being mixed. Both kappa values for Fiona are less than 0.7, thus providing further evidence of an inadequate measurement system.

Windsor (2003) reports a case study of the measurement phase of a Six Sigma Black Belt project for which the type of analysis described above led to annual savings of $400,000. Some refer to attribute agreement analysis as 'attribute gage R&R (short method)'. He concludes that 'an attribute gage R&R can normally be performed at very low cost with little impact on the process' and that 'significant benefits can be gained from looking at even our most basic processes'.

Minitab also provides **Attribute Gage R&R (Analytic Method)...**. This may be used to assess the performance of measurement systems such as 'go/no-go' plug gauges used to measure the dimensions of machined components. The method will not be considered in this book. Further information may be obtained from Minitab Help or from AIAG (2002, pp. 125–140).

9.5 Exercises and follow-up activities

1. The worksheet Bands.MTW gives data for a bias and linearity study of a micrometer system carried out by a manufacturer of band saw blades. In the study five measurements were made of the width of each of ten reference pieces of steel band with known width of 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50 mm. Use Minitab to assess the bias and linearity of the system.

2. If possible, obtain a digital micrometer and a set of ten parts – diameters of coins of the same denomination may be measured, for example. Set up a gauge R&R study involving at least two operators and at least two measurements by each operator of each part. Plan your experiment carefully after some initial trials with the equipment and the chosen parts. The author has used measurements of coin diameter with training groups in the past. Use Minitab to assess the performance of the measurement process.

3. The worksheet ARGageR&R.MTW contains data form a gauge repeatability and reproducibility study of a digital micrometer carried out by the author and his wife (who had never used a digital micrometer before the day on which the experiment was carried out). The measurements were of the heights of cylindrical wooden beads.

   (i) Display the data in a gauge run chart.

   (ii) Assess the performance of the measurement system.

   (iii) Unstack the data and assess each operator separately.

4. Bland and Altman (1986) provide data on the peak expiratory flow rates (PEFR) for 17 subjects measured using two different meters – LWPFM and SWPFM – which are provided in PEFR.MTW and reproduced by permission of *The Lancet*. Display the data as for the example given earlier and confirm that the limits of agreement are -78 to 74.

   The authors state that the standard error of the limits is given approximately by $\sqrt{3s^2/n}$, where s is the standard deviation of the differences and n is the number of parts measured. Calculate approximate 95% confidence intervals for the limits of agreement.

5. Following further training, the appraisers Fiona and Iain, referred to in Section 9.3, assessed the same 5 specimens twice in a second attribute agreement study. The data are available in Reassess.MTW. Perform an attribute agreement analysis and comment on the effectiveness of the further training.